

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

106252

İŞ AKLI ve VERİ AMBARI PROJELERİNDE VERİ
KALİTESİ ve VERİ TEMİZLEME TEKNİKLERİ
SİMÜLASYON UYGULAMASI

Bilgisayar Mühendisi, Sedat YAZICI

F.B.E. Bilgisayar Mühendisliği Anabilim Dalında
Hazırlanan

YÜKSEK LİSANS TEZİ

Tez Danışmanı : Prof. Dr. M. Yahya KARSLIGİL

TEZ YÜKSEK ÖĞRETİM KURULU
DOKÜMANTASYON MERKEZİ

Doç. Dr. Selim AKYOKUŞ Selim Akyokus

Doç. Dr. Oya Kalipsiz Oya Kalipsiz

İSTANBUL, 2001

İÇİNDEKİLER

	Sayfa
KISALTMA LİSTESİ.....	i
ŞEKİL LİSTESİ.....	iii
ÇİZELGE LİSTESİ.....	v
ÖNSÖZ.....	vi
ÖZET.....	vii
ABSTRACT.....	viii
1. GİRİŞ.....	1
1.1 Amaç ve Kapsam.....	1
1.2 Tez Konusu ile İlgili Önceki Çalışmalar.....	2
1.3 Kullanılan Yöntemler.....	3
1.4 Geliştirilen ve Önerilen Teknik ve Algoritmalar ile Simülasyon Yapan Uygulama Programı.....	4
2. TANIMLAR, KAVRAMLAR VE VERİ ERİŞİM TEKNOLOJİLERİNİN TARİHÇESİ.....	5
2.1 “Sistem”, “Bilgi Sistemi” ve “Uygulama” Tanımları.....	5
2.2 Bilgi Nedir? Veri / Bilgi / Anlamlandırılmış Bilgi (Data / Information / Knowledge) İlişkisi.....	5
2.2.1 İşlenmiş Veri Olarak: Bilgi.....	6
2.2.2 Belirsizliğin Tersi Olarak: Bilgi.....	7
2.2.3 Anlamlı Bir Sinyal Olarak: Bilgi.....	7
2.3 “Veri Yapısı” Tanımı.....	7
2.4 “Veri Sözlüğü” Tanımı.....	7
2.5 Bilgi Stratejisi Nedir?.....	8
2.6 Verinin Krallığı.....	9
2.7 Bilgiye Erişimin Tarihi.....	10
2.7.1 Veri Yönetimi.....	10
2.7.2 Dördüncü Kuşak Diller (4GLs) ve Bilgi Erişimi.....	11
2.7.3 Bilgi Merkezi.....	11
2.7.4 Bilgi Mühendisliği ve Bilgisayar Destekli Yazılım Mühendisliği (CASE) Araçları.....	12
2.7.5 Bilgi Ambarı.....	12
2.7.6 İstemci / Sunucu Mimari.....	12
2.7.7 Veri Ambarları ve Data Mart’lar.....	13
2.7.8 Veri Madenciliği.....	13
2.7.9 Bir Veri Ambarının Yaratılması.....	14
2.7.10 Veri Modeli ve “Meta Data” Yaratılması.....	14
2.7.11 Veri-Zaman Ölçeklendirmesi (Benchmarking).....	15

2.7.12	Verinin Temizlenmesi ve Dönüştürülmesi.....	15
2.7.13	Kullanıcıların Eğitimi.....	15
2.7.14	Veri Ambarına Yatırım.....	15
2.7.14.1	Veri Aktarımı.....	15
2.7.14.2	Veri Kalitesi.....	16
2.7.14.3	İş Akı.....	16
2.7.14.4	Uzaktan erişim.....	16
2.7.14.5	Metodoloji	16
2.7.14.6	Uygulama Jeneratörleri.....	16
2.7.14.7	Veri Havuzu (Repository), Sözlük (Dictionary) veya Klasör (Directory)....	17
2.7.14.8	Veri Modelleme.....	17
2.7.14.9	Veri Madenciliği.....	17
2.7.14.10	Arayazılım (Middleware).....	17
2.7.14.11	İlişkisel VTYS (Veri Tabanı Yönetimi Sistemi).....	17
2.7.14.12	Replikasyon	18
2.7.14.13	Performans Modelleme ve Ölçümü.....	18
3.	VERİ AMBARI , DATA MART, VERİ MADENCİLİĞİ.....	19
3.1	Veri Ambarı Kavramı ve Tarihi Gelişimi.....	19
3.1.1	En Kısa Şekliyle Kavramsal Tanımlar.....	19
3.2	Veri Ambarı: Geçici Olmayan Beklentiler?.....	20
3.3	Analiz, Veri Toplama.....	20
3.4	Çok Boyutlu Veri Görünümü.....	21
3.5	İş Sistemleri ve Veri Ambarları Arasındaki Farklar.....	21
3.6	Veri Ambarının Uygulamaya Alma (Implimentation).....	22
3.7	Data Mart ve Veri Madenciliği Seçeneği.....	22
3.8	Veri Ambarı, Data Mart ve Veri Madenciliği Karşılaştırması.....	23
3.9	Veri Ambarı / Data Mart Oluşturma.....	23
3.10	Diğer Veri Ambarı / Data Mart Kriterleri.....	24
3.11	Veri Madenciliği Nedir?.....	26
3.11.1	“Ad Hoc” Sorgulama.....	27
3.11.2	Online Analitik İşlem (OLAP).....	27
3.11.3	Veri Madenciliği	28
3.12	Veri Madenciliği ile İlgili Sorulacak 20 (10 + 10) Soru.....	29
3.12.1	Top 10, Veri Madenciliği: “İŞ” Soruları	29
3.12.2	Top 10, Veri Madenciliği: “TEKNİK” Sorular.....	30
3.12.3	Veri Madenciliği Nasıl Yapılıyor?.....	30
3.13	Karar Destek ve Veri Ambarı İlişkisi.....	31
3.13.1	Planlama : Tamın & Bütçe.....	31
3.13.2	Veri Ambarının Değişen Rolü.....	33
3.13.3	İş Akılının Değişen Rolü.....	33
3.13.4	Sınırları Zorlama ve İşbirliği Dili.....	34
3.14	Veri Ambarı Tipleri.....	34
3.15	Dijital İş ve Dijital Zeka.....	36
3.16	Data Mart’lardan Veri Ambarı Doğar mı ?.....	38
3.16.1	Supermartlar.....	39

3.17	Operasyonel Veri Ambarı (OVA).....	41
3.17.1	İşlem (Transaction) Versiyonu.....	41
3.17.2	Anlık (Snapshot) Versiyonu.....	41
3.18	Bir Veri Ambarı Projesine Girerken Nelere Dikkat Edilmelidir?.....	42
3.19	Veri Ambarı ve Web.....	42
3.20	xOLAP	43
4.	“İŞ KURALLARI SİSTEMİ” YARATMA.....	44
4.1	İş Mantığı	44
4.2	E / R Modelleme İş Kurallarında Kullanılabilir mi?	44
4.3	Bir “İş Kuralları Sistemi” Yaratma.....	45
4.3.1	“İş Kuralları Sistemi” Nedir ?	45
4.3.2	İş Kuralları Metodolojisinin Tekillliği	46
4.3.3	Önemli Avantajları.....	46
4.4	“İş Kuralı” nedir?	46
4.5	Değişik Tipteki İş Kuralları	46
4.6	Bütün İş Kuralları Veri Hakkında mıdır?	47
4.7	“İş Kuralları Sistemi” Metodolojisinin Fazları.....	48
4.8	Kural Analizinin Başlangıcı	48
4.9	Kural-Kalıplarına Giriş	49
4.10	Kural-Kalıp Analizi.....	49
4.11	Kural Zenginleştirme Mantıksal Veri Modeli	50
4.12	Toplam Tasarım.....	52
5.	CRM (CUSTOMER RELATIONSHIP MANAGEMENT - MÜŞTERİ İLİŞKİLERİ YÖNETİMİ)	53
5.1	Müşteri İlişkileri Yönetimi Nedir?	53
5.2	“Müşteri Deneyim Döngüsü”nü Anlamak.....	54
5.3	Müşteri Verisi Entegrasyonu: Efektif CRM'in Zorunlu Komponenti.....	56
5.3.1	Gerçekten Müşteri Odaklı olmak.....	56
5.3.2	Gerçek-zamanlı CDI (Müşteri Verisi Entegrasyonu)	56
5.3.3	Hızlı Veri Değerlendirme.....	57
5.3.4	Müşteri Verisi Entegrasyonun CRM için Gerçekleştirilmesi.....	57
5.4	CRM ve Veri Ambarcılığının Büyüme Aşamaları.....	57
5.4.1	DW/DSS Tip # 1: Raporlama.....	58
5.4.2	DSS/DW Tip # 2: Analiz.....	59
5.4.3	DSS/DW Tip # 3: Geleceği Tahmin.....	59
6.	İŞ AKLI YATIRIMLARININ FİNANSAL BOYUTU (ROI)	61
6.1	İş Akli (BI) Projelerinde Yatırımların Geri Dönüşümünün (ROI) Hesaplanması.....	61
6.1.1	ROI Metoduna Genel Bakış.....	61
6.1.2	Bir BI Projesinin Maliyeti Nasıl Tahmin Edilir?	61
6.1.2.1	ROI Hesaplama Örneği.....	62
6.1.3	ROI Duyarlılık Analizinin Gerçekleştirilmesi.....	63
6.1.4	İş Akli (BI) Projelerinin Mali Olmayan Kriterleri.....	64

6.2	Metadata ROI.....	64
7.	META DATA.....	67
7.1	CWM (Common Warehouse Metamodel-Ortak Ambar Metamodeli) Nedir?	67
7.1.1	CWM Takımı	67
7.1.2	CWM Mimarisi	68
7.1.3	CWM Kullanımı	68
7.2	Metadata Nedir?	69
7.3	(Jenerik) Metadata Ambarı	71
7.3.1	Meta Data Ambar Komponentleri	71
7.3.2	Jenerik Ambar Modelinin Mantıksal Görünümü	73
8.	KAYNAK, VERİ VE UYGULAMA ENTEGRASYONU	76
8.1	Kaynak Entegrasyonu.....	76
8.2	Dahili & Harici Verinin Entegrasyonu.....	76
8.2.1	Dahili & Harici Veri Entegrasyon Senaryoları.....	76
8.3	Veri Entegrasyonu : Veri Ambarının Temeli.....	77
8.4	Verinin, Kesinlik, Genişlik ve Derinliği.....	78
8.4.1	Kesinlik (Precision)	78
8.4.2	Genişlik (Breadth)	79
8.4.3	Derinlik (Depth)	79
8.5	Veri Entegrasyonunda Ambar Hatalarından Sakınma.....	80
8.6	ERP (Kurumsal Kaynak Planlama) Projelerinde “Veri Bütünlüğü”	81
8.7	e-iş Projelerinde “Veri Bütünlüğü”	81
8.7.1	B2C’de “Veri Bütünlüğü”	82
8.7.2	B2B’de “Veri Bütünlüğü”	82
8.8	Uygulama Entegrasyonu.....	83
8.8.1	RDF, Meta Data ve EAI.....	83
8.8.2	EAI için Bilgi Adalarının Birleştirilmesi.....	84
8.8.3	EAI için Tasarım Kalıpları: Veri Seviyesi & Metod Seviyesi.....	85
9.	VERİ KALİTESİ.....	88
9.1	Veri Kalitesi Konsepti.....	88
9.2	Veri Tutarsızlığı.....	90
9.3	Veri Kalitesizliğinin Sebepleri.....	92
9.4	Veri Kalitesinin Ölçülmesi.....	94
9.4.1	Önemli Verinin Belirlenmesi.....	94
9.4.2	Veri Kalitesini Ölçmek İçin Bir Yöntem Geliştirme.....	94
9.4.3	Ölçüm Metodu Tabanlı Toleransların Geliştirilmesi (Alan İstatistikleri Analizi - Analyze Domain Statistics)	95
9.4.4	Kalite Güvenliğini Otomatikleştirme.....	97
9.5	EDQM (Enterprise Data Quality Management – Kurumsal Veri Kalitesi Yönetimi)	97

9.6	EDQM Değeri Nasıl Ölçülür?	98
9.6.1	Yanlışlar (Mistakes)	98
9.6.2	Eşsesliler (Homonyms)	98
9.6.3	Standart Eksikliği (Lack of Standards)	99
9.6.4	Yasal Zorunluluklar (Legal Entities).....	99
9.6.5	Kayıp/Gizli Veri (Missing/Invisible Data).....	99
9.6.6	Hayali Veri (Phantom Data)	100
9.7	Veri Temizleme Araçları için Kritik Faktörler.....	100
9.8	Veri Kalitesi ve Veri Temizleme Araçları.....	100
9.9	Veri Temizleme Araçlarının Tarihçesi (Jenerasyonları)	101
9.9.1	Birinci Jenerasyon (Nesil) Veri Temizleme Araçları.....	101
9.9.2	İkinci Jenerasyon (Nesil) Veri Temizleme Araçları.....	102
9.10	Veri Kalitesi Garantileme.....	102
9.11	Veri Kalitesinin Gelir Üretimi ve Maliyet Yönetimi Üzerine Etkileri.....	103
9.12	Veri Kalitesi : 2000 Amerika Seçimleri (Florida Eyaleti).....	103
10.	VERİ TEMİZLEME (DATA CLEANSING)	105
10.1	Gerçek Dünya Verisi Kalitesizdir.....	105
10.2	Müşteri Verisindeki Problem.....	106
10.3	Veri Temizleme Genel Prosesi.....	107
10.4	Basit Veri Silme Metodu.....	107
10.5	Anlamlandırılmış Bilgi Tabanı (Knowledge Base)	109
10.6	Veri Temizleme Bilimi.....	109
10.7	İsim ve Adres Verisi Temizleme Analizi.....	112
10.8	Diğer Veri Kalitesi ve Temizleme Metodolojisi.....	112
10.9	Veri Temizleme Araçları Seçim Kriterleri.....	113
11.	ETL (SEÇME DÖNÜŞTÜRME YÜKLEME)	115
11.1	Bir BI Ortamında Kullanılan Veri Toplama Teknikleri ve Anahatları.....	115
11.2	Veri Ambarı İşlemleri ve ETL İşleme.....	117
11.2.1	Klasik Sistemlerden İndirme Bölgesine (Landing Pad)	117
11.2.2	İndirme Bölgesinden Taktiksel Veri Deposuna.....	119
11.2.3	Kurumsal Veri Ambarı Geliştirme.....	120
11.3	Müşteri Verisinin Kalitesi ve ETL.....	122
11.4	Parçalama (Parsing): Verinin Ne Olduğunu Biliyormusunuz?	123
11.5	Doğrulama: Verinin Doğru Olduğunu Nasıl Biliriz?	123
11.6	Standartlaştırma: Veri, Eşleştirme İçin Hazır mı?	124
11.7	Zenginleştirme: Her Veri İhtiyaç Duymalı mı?	125
11.8	Eşleştirme: Verinizdeki Tekrarlanmış Kayıtları Bulabiliyor musunuz?.....	127
11.9	Konsolidasyon: Veri Eşleştirildi, Sonrasında Ne Yapılacak?.....	129
11.10	Eşleştirme ve Konsolidasyon Çözümleri: Nasıl bir araya getirecek?.....	130
11.10.1	Eşleştirme Sistemleri.....	130
11.10.2	Konsolidasyon Sistemleri.....	132
11.11	Veri Kalitesi Yazılımı ve ETL Çözümleri Arasındaki Farklar.....	132
11.12	ETL Araçlarının Tarihsel Gelişimi.....	133

11.13	ETL Araçlarının Ölçeklenebilirliği (Scalability)	134
12.	ETL SİMÜLASYON UYGULAMASI.....	135
12.1	“YıldızETL” Uygulaması Mantıksal (E/R) Modeli.....	135
12.2	“YıldızETL” Uygulaması SQL Cümlecikleri.....	136
12.2.1	Tablo Yaratma SQL Cümlecikleri.....	136
12.2.2	Index Yaratma SQL Cümlecikleri.....	138
12.3	“YıldızETL” Uygulaması Genel Anlatımı.....	139
12.4	“YıldızETL” Uygulaması “Seçme Sihirbazı (Extractor Wizard)” Modülü...	141
12.5	“YıldızETL” Uygulaması (Cleanser & Transformer Wizard)” Modülü.....	149
13.	SONUÇLAR, TARTIŞMA KONULARI ve ÖNERİLER.....	153
KAYNAKLAR.....		161
ÖZGEÇMİŞ.....		167



KISALTMA LİSTESİ

4GL	Fourth Generation Language
API	Application Program Interface
B2B	Business-to-Business
B2C	Business-to-Customer
BI	Bussiness Intelligence
BT	Bilişim Teknoloji(ler)i
CASE	Computer Aided Software Engineering
CDI	Customer Data Integration
CIO	Chief Information Officer
CEO	Chief Executive Officer
CRM	Customer Relationship Management
CWM	Common Warehouse Metamodel
CWMI	Common Warehouse Metadata Initiative
DBA	DataBase Administrator
DDL	Data Definition Language
DFS	Data Federation Systems
DM	Data Mining
DOLAP	Desktop OLAP
DQAS	Data Quality Assurance Services
DSS	Decision Support Software
DW	Data Warehousing
EAI	Enterprise Application Integration
EDQM	Enterprise-wide Data Quality Management
EIS	Executive Information Systems
ERP	Enterprise Resource Planning
HOLAP	Hybrid OLAP
IE	Information Engineering
IQ	Information Quality
IRR	Internal Rate of Return
ISO 3166	Country Codes
ISO 4217	Currency Codes
IT	Information Technology
JCL	Job Control Language
LAN	Local Area Network
MDC	Meta Data Coalition
MERNİS	MERkezi Nüfus İşleri Sistemi
MOLAP	Multidimensional OLAP
MOF	Meta Object Facility
MOM	Message-Oriented Middleware
NPV	Net Present Value
OIM	Open Information Model
OVA	Operasyonel Veri Ambarı
ODS	Operational Data Store
OLAP	OnLine Analytical Processing
OMG	Object Management Group
PDCA	Plan-Do-Check-Act
POS	Point-of-Sale
RDF	Resource Description Framework
RFM	Recency, Frequency, and Monetary value

ROI	Return On Investment
ROLAP	Relational OnLine Analytical Processing
SCM	Supply Chain Management
SIC	Standard Industry Classification (Code)
SME	Subject-Matter Expert(s)
SQL	Structured Query Language
SSN	Social Security Number
TDS	Tactical Data Store
TP	Transaction Processing
TQM	Total Quality Management
UML	Unified Modeling Language
USPS	U.S. Postal Service
VTYS	Veri Tabanı Yönetin Sistemi
W3C	World Wide Web Committee
WOLAP	Web OLAP
XL	Microsoft Excel
XML	Extensible Markup Language
XMI	XML Metadata Interchange



ŞEKİL LİSTESİ

Şekil 2.1	Veri-Bilgi-Anlamlandırılmış bilgi ilişkisi	6
Şekil 2.2	Değer piramidi.....	6
Şekil 2.3	Kurumsal bilgi akış şeması.....	10
Şekil 3.1	Sunucu-tabanlı yaklaşım.....	25
Şekil 3.2	İstemci-tabanlı yaklaşım	25
Şekil 3.3	Data mart uygulama opsiyonu-1	25
Şekil 3.4	Data mart uygulama opsiyonu-2	26
Şekil 3.5	“Ad hoc query” sonuç kümesinin gösterimi	27
Şekil 3.6	OLAP analizi sonuç kümesinin gösterimi	28
Şekil 3.7	Kullanıcı katılımına göre veri madenciliği	28
Şekil 3.8	BI teknolojileri değer ve amaç şeması	29
Şekil 3.9	Veri ambarı uzayı	32
Şekil 3.10	Dijital iş komponentleri	37
Şekil 3.11	Değişik noktalarda değişik veri yapılarının (star join) bulunması	40
Şekil 4.1	“İş Kuralları Sistemi” için muhtemel “Yüksek-Seviyeli İçeriksel Mimari”.	45
Şekil 4.2	“İş Kuralları Sistemi” nasıl inşa edilir?	46
Şekil 4.3	Kural sınıflandırması	47
Şekil 4.4	“İş Kuralları Sistemi” metodolojisinin fazları.....	48
Şekil 4.5	Örnek “Kural-Kalıp” tablosu	49
Şekil 4.6	Bir kurallar koleksiyonu	49
Şekil 4.7	Uygun olmayan “Kural-Kalıp” tablosu	50
Şekil 4.8	Kural-kalıp 1.....	50
Şekil 4.9	Kural-kalıp 2.....	50
Şekil 4.10	Kural-kalıp 3.....	51
Şekil 4.11	Temel E/R modeli	51
Şekil 4.12	Örnek “Kural-zenginleştirme” veri modeli.....	51
Şekil 5.1	Müşteri Deneyim Döngüsü	55
Şekil 5.2	DSS / DW kullanımının gelişimi	58
Şekil 5.3	Raporlama uygulamaları ve sorular	58
Şekil 5.4	Analiz uygulamaları ve sorular	59
Şekil 5.5	Geliştirme ve Tahmin Yetenekleri.....	59
Şekil 5.6	Tahminci uygulamalar ve tahminci sorular.....	60
Şekil 6.1	Meta data yönelimli iş kullanıcısı arayüzü – DSS internet sitesi	65
Şekil 6.2	Meta data yönelimli iş kullanıcısı arayüzü – Araştırma sonuçları.....	66
Şekil 7.1	Meta Data Ambar Komponentleri.....	71
Şekil 7.2	Jenerik Ambar Modeli’nin mantıksal model görünümü	73
Şekil 8.1	"Göreceli (Relatif) Veri Doğruluğu" grafiği	78
Şekil 8.2	"Göreceli (Relatif) Doğruluk Derinliği" grafiği	80
Şekil 9.1	Veri tutarlılığı ve veri tekrarlama ilişkisi.....	90
Şekil 9.2	Veri tutarlılığı ve veri bütünlüğü ilişkisi.....	91
Şekil 9.3	Grafik 1.....	95
Şekil 9.4	Grafik 2 ve Grafik 3.....	96
Şekil 10.1	Veri nitelikleri temizleme metodolojisi tasarım aşamaları	112
Şekil 11.1	Klasik sistemlerden indirme bölgesine (Landing Pad)	118
Şekil 11.2	İndirme bölgesinden Taktiksel Veri Deposuna.....	119
Şekil 11.3	Taktiksel veri deposundan, kurumsal veri ambarına veri dönüşüm prosesi	121
Şekil 11.4	Kurumsal veri ambarından data marta veri aktarma	122
Şekil 12.1	YıldızETL simülatöründe kullanılan tabloların mantıksal (E/R) modeli.....	135
Şekil 12.2	YıldızETL simülasyon uygulaması ana ekranı.....	140

Şekil 12.3	YıldızETL sistem performansı izleme ve sistem bilgisi elde etme ekranı...	141
Şekil 12.4	YıldızETL simülasyon uygulaması ExtractorWizard modülü açılış ekranı..	142
Şekil 12.5	YıldızETL uygulaması ExtractorWizard modülü birinci kulakçık (page-tab 1).....	143
Şekil 12.6	YıldızETL uygulaması ExtractorWizard modülü seçilecek tabloların eklenmesi.....	143
Şekil 12.7	YıldızETL uygulaması ExtractorWizard modülü ikinci kulakçık (page-tab 2)	144
Şekil 12.8	YıldızETL uygulaması ExtractorWizard modülü üçüncü kulakçık (page-tab 3)	145
Şekil 12.9	YıldızETL uygulaması ExtractorWizard modülü dördüncü kulakçık (page-tab 4)	145
Şekil 12.10	YıldızETL uygulaması ExtractorWizard modülü beşinci kulakçık (page-tab 5)	146
Şekil 12.11	YıldızETL uygulaması ExtractorWizard modülü altıncı kulakçık (page-tab 6)	147
Şekil 12.12	YıldızETL uygulaması modülü dinamik oluşturulan SQL cümlecği görünümü.....	148
Şekil 12.13	YıldızETL uygulaması dinamik oluşturulan SQL cümlecği çalıştırma sonucu.....	148
Şekil 12.14	YıldızETL, Cleanser & Transformer Wizard modülü birinci adımı (Investigation)	149
Şekil 12.15	YıldızETL, Cleanser&Transformer Wizard modülü ikinci adımı (Standardization)	150
Şekil 12.16	YıldızETL, Cleanser&Transformer Wizard modülü üçüncü adımı (Matching)	150
Şekil 12.17	YıldızETL, Cleanser&Transformer Wizard modülü üçüncü adımı (Finalization)	151
Şekil 12.18	YıldızETL, Cleanser&Transformer Wizard modülü son adımı (Finalization)	152

ÇİZELGE LİSTESİ

Çizelge 3.1	Veri ambarı, data mart ve veri madenciliği karşılaştırması.....	23
Çizelge 3.2	Veri ambarı, data mart karşılaştırması.....	23
Çizelge 3.3	xOLAP terimleri listesi.....	43
Çizelge 4.1	“İş Kuralları Sistemi” prototipinin oluşturulması için genel yaklaşım adımları.....	52
Çizelge 6.1	Örnek XYZ şirketi için BI Gider Tablosu	62
Çizelge 6.2	XYZ'nin BI uygulamasını <u>kullanması</u> durumunda 3 yıllık maliyet dağılımı.....	63
Çizelge 6.3	XYZ'nin BI uygulamasını <u>kullanmaması</u> durumunda 3 yıllık maliyet dağılımı.....	63
Çizelge 6.4	Olasılık Yüzdeleri	64
Çizelge 8.1	Dahili & Harici veri entegrasyonu senaryoları özet tablosu.....	77
Çizelge 8.2	EAI oluşturmak için gerekli ürün tipleri.....	85
Çizelge 9.1	Veri kalitesizliğinin sebepleri	92
Çizelge 9.2	Veri kalitesizliğinin sebeplerinin derecelendirilmesi.....	93
Çizelge 10.1	Veri Temizleyici (DataCleanser) için örnek eşleştirme kayıtları.....	109
Çizelge 11.1	ETL Fonksiyonları Tablosu.....	116
Çizelge 11.2	“Klasik sistem-İndirme bölgesi” kalite ölçüm kriterleri.....	118

ÖNSÖZ

En başta ve öncelikle danışman hocam Sayın Prof. Dr. M. Yahya KARSLIGİL'e yüksek lisans çalışmam programı boyunca beni yönlendirdiği, bilgisini ve değerli görüşlerini bana aktardığı için çok teşekkür ederim. Ayrıca hem benim için bir örnek olmaları, hem de tavsiyeleri ile bana katkıda bulunan YTÜ Bilgisayar Bilimleri ve Mühendisliğinde görev yapan sevgili hocalarıma da teşekkürü bir borç bilirim.

Tez konusu olarak incelediğim “Veri Temizleme (Data Cleansing)” konusunu bana önerdiği için; I-BimSA A.Ş. “Business Intelligence” departmanı çözüm uzmanı sayın Nebahat DEMİRSOY'a ve ayrıca sevgili arkadaşım sayın Göksel ÇELİK'e teşekkür ederim.

Tez çalışmam boyunca desteğini esirgemeyen, beni anlayan ve yüreklendiren müstakbel eşim Reyhan GÜRBOSTAN'a, ayrıca yardımlarından ötürü değerli mesai arkadaşlarım sayın Oğuz AKBABA ve Erol ULUER'e; çalışmam boyunca zaman konusunda bana gösterdikleri esneklikten ötürü sayın amirlerime teşekkürü bir borç bilirim.

Hayatım boyunca desteklerini her zaman hissettiğim ve hissedeceğim sevgili aileme de teşekkür ederim.

ÖZET

Müşterilerinin, ürünlerin, tedarikçiler ve adreslerinin bilgisine yüksek oranda hakim olan şirketler başarılı stratejiler ve taktikler geliştirebilir. Bu bilgiler en temel seviyede gelir artırıcı girişimlerde ve maliyet yönetimi (yatırımların geri dönüşümü-ROI) girişimlerinde önemli etkiye sahiptirler.

Bununla birlikte, tam bilgilenme için veri kalitesi ön şart olarak karşımıza çıkar. Sadece müşteriler, ürünler, tedarikçiler, adres bilgileri ve bunların aralarındaki ilişkiler hakkındaki *doğru ve tam* bilgiler; kuruma başarılı olması için veri ambarcılığı projelerinde gerekli olan iş aklını kazandıracaktır.

Bugünkü iş dünyasında veriler, devamlı olarak farklı sistemler arasında akmaktadır. İşletmeye veriler, veri giriş sistemlerinden ve yapılaşmamış internet cevapları şeklinde sel gibi akmaktadır. Bir sistemden diğerine taşınmakta, iş kuralları sistemi içinde kalmak suretiyle kurumsal veri entegrasyon uygulaması (EAI) çözümlerine ulaşmadan önce ise, oldukça geniş bir şekilde yayılmaktadır. Ve veri işletme dışına entegre edilmiş bir şekilde, extranet ve Internet gibi tedarikçilere ve ortaklara ulaşılan sistemler aracılığıyla, seyahatine devam etmektedir.

Bir başka deyişle, toplam kalite yönetimi (TQM) konseptiyle bakıldığında, veri kalitesinde ve özellikle toplam sistem kalitesinde düşüşe yol açan, verideki ve sistemlerdeki karmaşıklık, her geçen gün artmaktadır.

Bu tezde, “YıldızETL” isimli bir ETL (veri seçme, dönüştürme ve yükleme) simülasyon uygulaması geliştirilmiştir. Bu yazılım uygulamasında, veri temizleme prosesinde kullanılan temel teknik ve algoritmalar, uygulamanın modüllerine geniş bir şekilde dağıtılmış olarak uygulanmıştır. Bu uygulamada müşteri ilişkileri yönetimi (CRM) için önemli olan müşteri/tedarikçi, ürün ve adres bilgilerine yoğunlaşmıştır. Sonuç olarak seçme modülü oldukça başarılıdır ve hem veri tabanı hem de ortam bağımsızdır. Temizleme modülü, güvenilir ve güçlü veri kütüphanelerine ihtiyaç duyan veri-referanslı bir araç olarak, doğru bir şekilde çalışmaktadır. Dönüştürme modülü de başarılı olarak fonksiyonunu yerine getirmekte, ancak pazardaki mevcut ürünler gibi eşleştirme konusunda geliştirmeye ihtiyaç duymaktadır.

Anahtar Kelimeler : Veri temizleme, veri kalitesi, Seçme Dönüştürme Yükleme (ETL), Kurumsal Uygulama Entegrasyonu (EAI), iş kuralları sistemi

ABSTRACT

Companies that know the most about their customers, products, vendors and locations could be able to formulate the most successful strategies and tactics. This information has a significant impact on the bottom line in both revenue-generating and cost administration initiatives (ROI).

However, data quality is a pre-requisite for being fully informed. Only *accurate, complete* data about customers, products, vendors and locations, and their relationships (CRM) with each other, will yield the enterprise business intelligence needed to win in data warehousing projects.

In the business world today, data moves constantly among diverse systems. It streams into the corporation through the data entry systems and unstructured Web responses. It flows from one corporate system to another, while keeping it stayed within the “business rules system” as disseminated more widely than ever before by enterprise application integration (EAI) solutions. And integrated data travels outside the corporation to vendors and partners who interact wit it through extranets and the Internet.

In other words, every day there is an increase in data and system complexity, which can lead to degradation in data quality and overall system quality, considering the total quality management (TQM) concept.

In this thesis, an ETL (Extract Transform Load) simulation tool, named as “YıldızETL”, is developed. The techniques and algorithms are applied in data cleansing proces which is widely distributed in modules of this software application. This applicaton is primarily concentrated on customer/vendor, product and address data for customer relationship management (CRM) concerns. As a result, extraction module is very succesful and it is both database and platform independent. Cleansing module works properly as a data-referenced tool that needs reliable and powerful data libraries. Transformation module is also succesfull but it must be improved in matching like the ones available in the market.

Keywords: Data cleansing, data quality, Extract Transform Load (ETL), Enterprise Application Integration (EAI), business rules system

1. GİRİŞ

Bu bölümde; tez konusu ile ilgili önceki çalışmaların içeriği, kullanılan yöntemler, önceki çalışmalar, geliştirilen ve önerilen teknik ve algoritmalar ile simülasyon yapan uygulama programı özet bir şekilde anlatılacaktır.

1.1 Amaç ve Kapsam

“Bugün ve gelecekte, bilgi birikimlerini rakiplerinden daha hızlı yönetmeyi ve bunu nasıl yapacağını bilen şirketler başarıyı yakalayabilecek olanlardır. Bu, sürekli yeni bilgileri edinme meselesi demek değildir. Bu, mevcut işten gerekli bilgiyi çekip çıkarabilme yeteneğidir. Bu yetenek eğilimleri görmek, kavramak ve bunların gereklerini rakiplerden önce yerine getirmektir.” (Grandy, 1997).

Bu söz; Louis V. Gerstner Jr.’a aittir. 1997 yılı Ekim ayında “Securities Industry Association” adlı kuruluştaki “IBM Chairman and CEO.” ünvanıyla yaptığı konuşmadan alınmıştır. Bilginin rekabetçi bir avantaj yakalamadaki anahtar rolünü en güzel şekilde ifade eden sözlerden biri olarak kabul edilmektedir.

Geleceğin giderek bugün olduğu baş döndürücü gelişmelerin yaşandığı günümüzde IT sektörü özellikle kendi yarattığı “internet ekonomisi” veya yaygın deyimle “yeni ekonomi” için önceleri veriyi saklayıp işlemeye yönelik tasarlanan klasik (legacy) sistemleri ile artık yetmemeye başlamıştı. Bu geleneksel yapıdan sonra doğup gelişen “Veri ambarı”, “Karar Destek Sistemi”, “CRM” ve “OLAP” kavramları ile yeni açılımlar sunan sektör; ekonominin web ortamına taşınmasıyla artık yeni bir kavramla yoluna devam etmeye karar verdi. Bu kavram yakın geleceğin en sık duyulacak ve uygulanacak olan “Digital Business” veya “Business Intelligence (İş Aklı)”, “E-Business Intelligence (E-İş Aklı)” kavramlarıydı. E-İş Aklı, E-iş ile İş Aklının kesişimi ile anlam kazanmış olup, İş Aklı uygulamalarının web ortamına uzanışını ifade etmektedir (Ilıkkın, 2000).

Verinin bilgiye, bilginin iş kararlarına ve bu mekanizmanın tüm iş ortamına yayılarak sonuçta bir iş aklına (Business Intelligence - BI) nasıl dönüştüğünü inceleyen bu akademik çalışmanın özel inceleme alanı “Data Cleansing” yani “(Kaynaktaki) verinin temizlenmesi”dir.

İş aklı (BI), veriyi kesin, bilgi temelli ve harekete geçirir bir hale dönüştürüyor. Müşterilerin veya potansiyel (internet-gezgini) müşterilerin sayıları günümüzde oldukça yüksek rakamlara tekabül etmektedir. Bu şartlar altında şirketlerin müşteri eğilimlerini bulmalarına, müşteri bağlılığı yaratmalarına, kişiye-özel hizmet sunmalarına, tedarikçi ilişkilerini arttırmalarına finansal riski düşürmelerine, yeni satış fırsatlarını ortaya çıkarmalarına olanak sağlamayı amaçlıyor. Kısacası BI; şirketlerin rekabetçi avantajlarını arttırıyor ve bu avantajları sürekli tutmayı hedefliyor.

Gerçekten bilginin stratejik kullanımı olan “dot.com” dünyasında da başarının yada başarısızlığın en kritik faktörlerinden birisi de bu. “E-iş” ile “İş Aklı”nın kesişiminin giderek anlam kazanması ile BI çözümleri web ortamına taşınmış ve literatüre “E-Business Intelligence” kavramı girmiştir artık.

Tez konusunun detay incelemesine girmeden önce bazı temel terim ve kavramların tanımlamaların yapılması gereklidir.

1.2 Tez Konusu ile İlgili Önceki Çalışmalar

İlk çalışmalar daha çok toplu iş (batch) prosesleri üzerine yoğunlaşmış, ortamdan bağımsız çözümler nasıl ortaya konulur, bunların geliştirme stratejisi konu edilmiştir. Artık konu ile ilgili çalışmalar sürekli devam etmekte olup, iş anlamı (semantik), iş kuralları ve meta data üzerine yoğunlaşmıştır.

Bu yeni konu ve problemler için, oldukça yüksek seviyeli “yapay zeka” ve “bulanık mantık” teknikleri kullanıldığı için AR-GE yatırımlarının büyük ve maliyetli olduğu görülmektedir. Tezin ilerleyen bölümlerinde başka araştırmacıların yaptığı çalışmalar kaynak gösterilerek şekil, grafik ve çizelgeler ile de desteklenerek detaylı şekilde anlatılmaya çalışılmıştır.

Ülkemizde henüz yüksek kapasiteli iş aklı ve veri ambarcılığı uygulamaları fazla bulunmadığından bu konu ile özellikle son on yılda üniversitelerde yüksek lisans ve doktora tezi çalışmalarında nadiren çalışma bulunmaktadır. Bu açıdan özetleri (abstract, YÖK ve ULAKBİM üzerinden) incelenen birkaç tez çalışması ise, kısmen değinilsede bu kapsam ve içerikte değildi.

1.3 Kullanılan Yöntemler

Veri erişiminin tarihçesinden başlayarak günümüz teknolojilerine kadar getirip bunlar hakkında detaylı bilgiler verilmiştir. Bu teknolojiler Data Mart, Veri Ambarcılığı, Veri Madenciliği, İş Akıllı, (BI) CRM (Müşteri İlişkileri Yönetimi), İş Kuralları ve Kural-Tabanlı Sistemler, Veri Kalitesi, İş Akıllı yatırımlarının geri dönüşümü, ETL araçları gibi ana başlıklar değişik bölümlerde ayrı-ayrı ve kimi zaman da bütünleşik bir şekilde incelenmiştir.

İncelenen araştırma çalışmaları genelde istatistiki ve grafiksel olarak sunulmaya çalışılmıştır. Sürekli olarak her konuda sınıflandırma yapılabilmeye çalışılmış, incelen teknik ve algoritmalar veya uyarlama ve uygulama çalışmaları adım-adım şeklinde maddesel bir yapıda sunulmuştur. Çeşitli senaryolar kullanarak veya varsayımlar ortaya koyarak konu örneklendirmeye çalışılmıştır. Kullanılan teknik ve yöntemler bir uygulama yazılımıyla görsel hale getirilmiş ve bir uygulama geliştirilmek suretiyle elle tutulur bir şekil almıştır. Gerçek veri ambarı ortamı ve verisi sağlanamadığından ve ayrıca tekil bir birey olarak oldukça büyük ve mali açıdan çok yüksek maliyetli olan bir böyle uygulamanın anacak simülasyonu yapılabilmektedir.

1.4 Geliştirilen ve Önerilen Teknik ve Algoritmalar ile Simülasyon Yapan Uygulama Programı

Uygulamada (YıldızETL) kullanılan tablolar, yani veri tabanı yapısı, *ERwin Designer* (Computer Associates, Inc.) aracı ile tasarlanmıştır. *Delphi 5* (Inprise Corp.) uygulama geliştirme aracı ile kodlanmıştır. Wizard (sihirbaz) mantığı ve teknolojisi kullanılarak tam bir araç (tool) yapısı sağlanmaya çalışılmıştır.

Veri kalitesi ve temizlenmesi uygulamalarında veri ve yapısı çok iyi biliniyorsa sonucu yüksek başarı sağlanabilecek bir uygulama ortaya çıkarmak mümkündür. Uygulamanın seçme modülü (extraction) oldukça başarılıdır ve hem veri tabanı hemde platform bağımsızdır. Temizleme modülü, güvenilir ve güçlü veri kütüphanelerine ihtiyaç duyan veri-referanslı bir araç olarak, doğru bir şekilde çalışmaktadır. Dönüştürme modülü de başarılı olarak fonksiyonu yerine getirmekte, ancak pazardaki mevcut ürünler gibi eşleştirme konusunda geliştirmeye ihtiyaç duymaktadır. İngilizce tabanlı ve kabul gören 900 civarında ve maliyeti yüksek algoritmaların satıldığı düşünülür ise bir de Türkçe dil problemleri eklenirse çok-çok zor olan ve yerel olarak bir çok araştırma ve tez konusu ortaya çıkacaktır. Veri temizliğinde

amaç kötü, kalitesiz veriden kurtulabilmek ile beraber zenginleştirme (enhancement, enrichment) fonksiyonlarının da önemi unutulmamalıdır.



2. TANIMLAR, KAVRAMLAR VE VERİ ERİŞİM TEKNOLOJİLERİNİN TARİHÇESİ

Bu bölümde tez konusu ile ilgili terminolojik jargonlar veriden, bilgiye ve sisteme kadar oldukça geniş bir çerçevede tanımlanmaya çalışılacaktır.

2.1 “Sistem”, “Bilgi Sistemi” ve “Uygulama” Tanımları

Senn (1984) sistemi şöyle tanımlamaktadır. Bir **sistem**, basitçe, interaktif şekilde aynı amaç için bir araya gelmiş komponentler kümesidir (Senn, 1984).

Sistemler açık ve kapalı sistemler olarak ikiye ayrılabilir. Çevresi ile interaktif bir ilişkide bulunan yani girişleri olan ve çıkışlar üreten sistemler **açık sistem** olarak adlandırılır. Tam aksine, çevresi ile hiçbir interaktif ilişkisi olmayan sistemler ise **kapalı sistem** olarak adlandırılır.

Bilgi sisteminin amacı; bilgi işlemek, organizasyon verilerini tutan dosyaların bakımlarını yapmak, bilgi, rapor ve sonuçlar üretmektir. Bir bilgi sistemi, alt sistemlerden oluşur. Bu alt sistemler yazılım, donanım, veri depoları ve veri tabanlarıdır. Özel ekipman, program, dosya ve prosedürlerden oluşan, özel alt sistemler kümesinin oluşturduğu bilgi sistemine ise **uygulama** (application) denmektedir.

2.2 Bilgi Nedir? Veri / Bilgi / Anlamlandırılmış Bilgi (Data/Information/Knowledge) İlişkisi

Gordon, S.R. ve Gordon, J.R. (1996) çiftine göre **bilgi**, içerik ve perspektif açısından yapılan çalışmanın şekline bağlı olmak üzere oldukça değişik anlamlar içeren kompleks bir konsepte sahiptir.

Bu konseptin zenginliğini göstermek üzere bilgiyi üç ana kategoride tanımlamak gerekir.

1. İşlenmiş veri olarak: Bilgi
2. Belirsizliğin tersi olarak: Bilgi
3. Anamlı bir sinyal olarak: Bilgi

2.2.1 İşlenmiş Veri Olarak Bilgi

İngilizce dilbilgisi kurallarına göre veri (data) çoğul, bilgi (information) ise tekil bir kelimedir.

Tanımsız kullanımları ve uygulamaları olan ham gerçekler genellikle **veri (data)** olarak tanımlanır. (Gordon, 1996)

Bilgi (information), “*seçimleri etkileyen işlenmiş veri*” olarak tanımlanır (Gordon, 1996). İşlemekten kasıt, formatlama, filtreleme, özetleme vb. işlemlerdir.

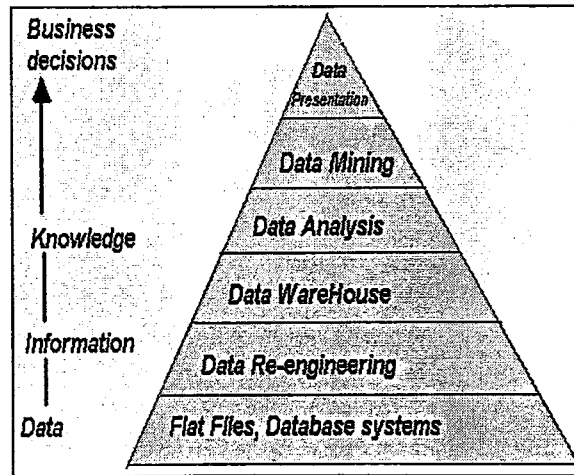
Anlamlandırılmış bilgi (knowledge), bilgiden türetilen, çıkartılan anlam olarak tanımlanır (Gordon, 1996). Anlamlandırılmış kelimesi yerine “kullanılabilir” kelimesi de ikame edilebilir.

Bu ilişki Şekil 2.1’deki gibi de gösterilebilir. Veri, önce bilgi, sonuçta da anlamlandırılmış bilgi olarak zenginleşmiş bir anlam içermektedir.



Şekil 2.1 Veri-Bilgi-Anlamlandırılmış bilgi ilişkisi (Gordon, 1996)

Değer Piramidi (Value Pyramid) olarak ifade edilen diyagramla biraz daha konuya yoğunlaşılır ise yukarıda farklı biçimde ifade edilen ilişki daha anlaşılır hale gelecektir.



Şekil 2.2 Değer piramidi (www.planetebi.com)

Şekil 2.2’de görüldüğü üzere veriden başlayan değer piramidinin en tepesinde yani tam olarak kendi başına pek fazla anlam ifade etmeyen salt bilgi sonuçta, iş kararlarına dönüşmektedir.

Eskiden veriye düz dosyalar ve veri tabanlarında ulaşılır iken, şimdi veri; değişim mühendisliği, veri ambarı, veri analizi, veri madenciliği ve veri sunumu adımlarından geçerek bir iş kararına dönüşmektedir.

2.2.2 Belirsizliğin Ters Olarak Bilgi

Belirsizlik ile ilgili şu alıntı belirsizlik ve bilgi arasındaki ilişkiyi en iyi şekilde anlatır.

“Ne kadar az bilgi, o kadar çok belirsizlik” (Arrow, 1984).

Organizasyonlarda, işletmelerde yöneticiler; bilgiyi belirsizliklerin azaltılması anlamında düşünmektedirler. Çünkü, yöneticiler karar verme sürecinde, çeşitli alternatifleri ve bunların sonuçlarını değerlendirirler. Bu şekilde yaptıkları takdirde kararın kalitesi ve karar verme sürecinin etkinliği artacaktır.

2.2.3 Anlamlı Bir Sinyal Olarak : Bilgi

Yöneticiler aynı zamanda rolleri gereği bir iletişimcidirler. Yani bilgi sonucu karar üreten (jeneratör) ve bilgi alan (receiver) bir iletişim cihazı gibidirler. Sinyalleri organize eden raporları veya verileri astlarından alırlar. Bu üçüncü tip bilgi tanımlama aynı zamanda, iletildiği anda verilen sinyali değerlendiren bir yönetici ister.

2.3 “Veri Yapısı” Tanımı

Veri yapısı (data structure), birbiri ile ilişkisi olan elemanların oluşturduğu veri elemanları kümesidir ve sistemi teşkil eden komponentlerden her birini tanımlamak üzere kullanılan bir yapıdır. Veri akışı , veri depolaması buna bir örnektir (Senn, 1984).

2.4 “Veri Sözlüğü” Tanımı

Sistem içerisindeki tüm veriler veri elemanlarından oluşur.

Birinci tanım: Veri elemanları da bir grup teşkil ederek veri-yapısı teşkil ederler. 2.3 nolu madde de belirtilen bu tanımlardan hareketle **veri sözlüğü**, en basit tanımıyla, sistem içindeki veri elemanlarının listesidir (Senn, 1984).

İkinci tanım: Bir diğer tanıma göre ise, ki bu daha yeni bir tanımlama olarak karşımıza çıkmaktadır. **Veri sözlüğü** veritabanı strüktürü (yapısı) hakkında veri depolayan ve metadata (hamveri olarak türkçeleştirilse de tam karşılığını vermediğinden metadata olarak kullanılacaktır) bilgisinin saklandığı yapı olarak tanımlamak mümkündür (Korth ve Silberschatz, 1991). Metadata kavramı oldukça önemli ve anahtar bir konudur. Bu konu ilerleyen bölümlerde oldukça geniş bir şekilde açıklanacaktır.

2.5 Bilgi Stratejisi Nedir?

Distefano (Distefano, 2000), efektif bir bilgi stratejisinin; birleştirici bir stratejinin, e-iş ve bilgi varlığını birbirine bağlayan bir vizyonu ortaya koyması gerektiğini ifade etmektedir. Kurumsal bilginin (enterprise information) kapsama alanı, birleştirici bir kurumsal varlık (asset) olarak yönetilmesine bağlıdır. Ancak bu şekilde kurumun misyonunu başarıyla yerine getirmesini sağlar.

Bilgi stratejisi; tüm kurum organizasyon ve lokasyonlarında işlenen sistem verisinden, müşteri verilerine, internetteki verilere ve karar destek verilerine kadar, tüm iş verilerini kapsamı gereklidir. Bu strateji, kurumun gelecekteki bilgi varlığını gösteren bir yol haritası şeklinde de tanımlanabilir. Bu bilgi varlığı tıpkı finansal sistemlerdeki duran varlıklar, dönen varlıklar gibi aynı zamanda finansal bir varlık da ifade etmektedir.

Bilgi stratejilerinin uygulanmasında iki farklı perspektif dengeli olarak uygulanmalıdır. Öncelikle, birinci iş olarak, bilgi stratejisinin tasarım ve yaratılması müşteri değeri noktasında bir altyapı üzerine kurulmalıdır. İkinci olarak da, bilgi stratejisinin operasyonel işlemleri ve rekabetçi avantajları bilgi seviyesinin yükseltilmesi ile daha ileri noktalara götüreceği, yani bilgiye ivme, hareket kazandıracağından böyle bir yapıya geçilmesi gerekliliği unutulmamalıdır.

Bilginin değeri, geleneksel finansal hesaplama yöntemleriyle, örneğin yatırımların geri dönüşümü (ROI-Return On Investment), kurumların yapmış oldukları performans değerlendirmelerinde oransal olarak tespit edilebilmiştir. Bu hesaplamalar oldukça zor olsa da, gelecekte şirketlerin bilgi sermayesi (information capital) de şirketlerin gerçek piyasa değeri hesaplanırken toplam tutara girecek bir finansal varlık/büyükölçölük olarak hak ettiği yeri alacaktır.

2.6 Verinin Krallığı

Orijinal Computerworld dergisinin 27 Mart 2000 de yayınladığı araştırma sonuçlarına göre (Langan, 2000), IT satınalma kararlarında “*kurumsal veritabanı projeleri*” bir numaralı öncelik olarak ortaya çıkmış, veya başka bir deyişle “Data is The King” başlığı dergi kapağı olmuştur. Amerikalı IT yöneticilerinin %95 oranında “*kurumsal veritabanı projeleri*” satın alacaklarını, yani 2000 ve 2001 yıllarında uygulamaya sokacakları belirtilmiştir.

Kurumsal kavramı, Amerikan şirketlerin bu ölçekteki projeleri tercih etme sebebi, oldukça fazla ülkede konuşlanmış bulunan dağınık-uluslararası şirket yapıları ve bu şirketleri tek bir kurum (enterprise) olarak değerlendirmenin gerekliliğini ortaya koymak için kullanılan bir kavram olarak ortaya çıkmaktadır. Oldukça büyük olan bu şirketler, şu anda, oldukça büyük olan mevcut yapılarını muhafaza ederek; mevcut iş modellerinin e-iş modellerine çevirmekle karşı-karşıya kalmış bulunmaktadır.

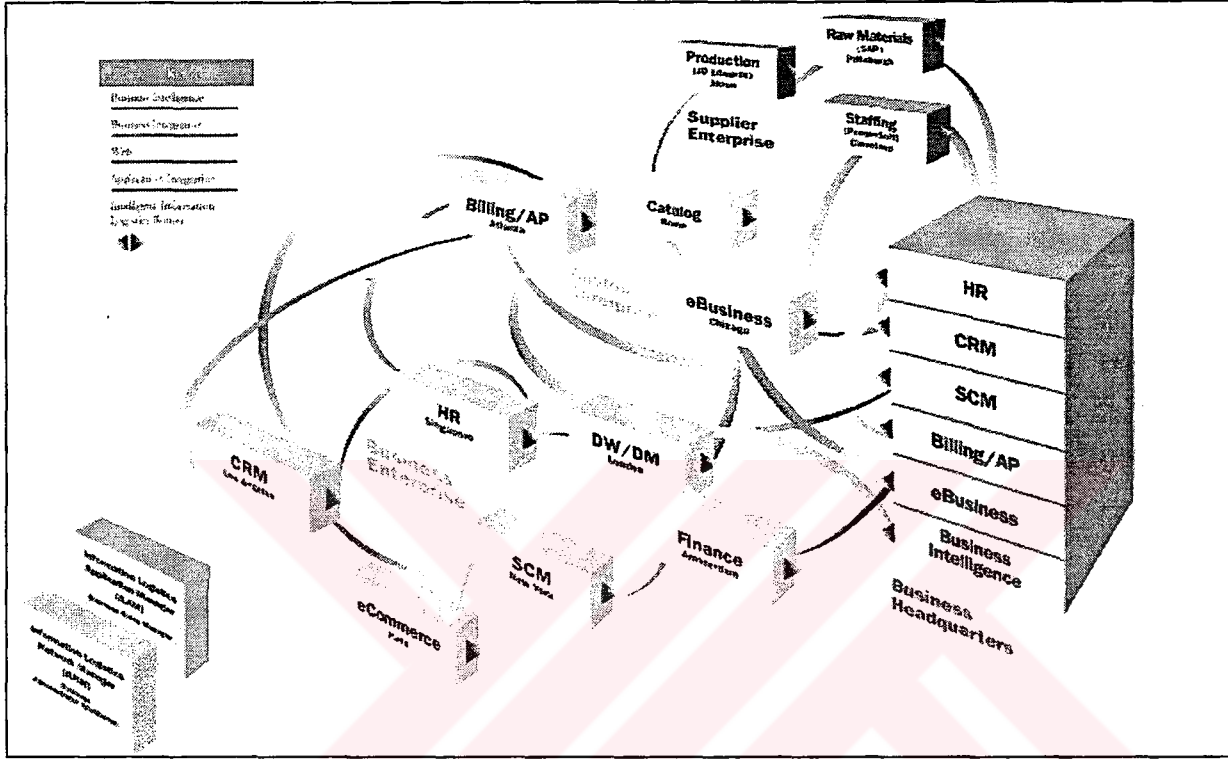
Yine Gartner Group tarafından yapılan bir araştırmaya göre, bu şirketler en popüler, en yaygın şekilde kullanılan 14 büyük ilişkisel veritabanı ürünlerinden birini ve birkaçını kullanmaktadırlar.

Computerworld dergisinin yaptığı araştırmada ortaya çıkan bir gerçek de, mevcut bilinen iş modellerinin e-iş ortamına kaymasındaki en büyük sebep yine veri ve bununla ilgili yaşanan problemlerdir. Veri ile ilgili problemler (veri kalitesi bölümünde oldukça geniş olarak incelenmiştir) çözüldüğü takdirde bu geçişin % 90 oranında başarılmış olacağı ortaya çıkmıştır.

Peki bu işlem nasıl başarılabilecek ? Kurumsal verinin tek bir bütün olarak görülebilmesi ve işlemlenebilmesi için, “*e-veri motoru (e-data engine)*” olarak adlandırılan yeni sınıf ürünler geliştirilmeye başlanmıştır. Şu anda bu ürünlerdeki hız, büyüyebilirlik ve kapsama alanı problemleri çözümlenmiş durumdadır. E-veri motorları yukarıda belirtilen 14 popüler veritabanının tümünü desteklemektedir.

2.7 Bilgiye Erişimin Tarihi

Hammaddenin işlenmesinden, elektronik ticarete uzanan yapıyı ifade eden bir uygulamanın şekli (Şekil 2.3) ile başlar isek bilgiye erişimin konseptinin daha anlaşılır olmasını sağlayabiliriz.



Şekil 2.3 Kurumsal bilgi akış şeması (www.d2k.com/images/centerspread_3g.jpg)

Bilgi işlem çalışmalarının daha ilk günlerinde, operasyonel veri üzerine yoğunlaşma olmuştur (Sterling Software, 1998). **Operasyonel veri**, şirketin günlük işlerini yürütmek için kullandığı veridir.

İlk yapılan uygulamalardaki teknolojiler genellikle siparişlerin işlenmesi, muhasebe ve resmi defter kayıtlarının takip edilmesi, üretim kontrol gibi uygulamalar üzerine yoğunlaşmış idi. Bilgiye erişimin tarihini hem fonksiyonalite hemde teknolojik kronoloji de dikkate alınarak inceler isek bu maddeyi daha alt maddelere ayırarak yapmak daha uygun olacaktır.

2.7.1 Veri Yönetimi

1970'li yıllarda, veri sağlanmasına yaklaşımlar genelde öncelikli olarak operasyonel rapor ihtiyaçları tarafından yönlendirilmekte idi. Veri yönetiminde verimlilik birincil baskın

yoğunlaşma faktörü idi. Düz dosya yapıları (flat files) yerini daha efektif kullanılan dosya organizasyon ve yapılarına bırakmıştır. Bunlar IMS, ISAM ve VSAM dosya yapılarıdır.

Kullanıcılar bilgisayarın gücünün farkına vardıkça, daha çok veri istemeye, talep etmeye başladılar. İş idarecileri artık bilgisayardan işlem seviyesinde raporlar isteyebiliyorlardı artık. Kullanıcıların sürekli yeni bir uygulama talebinde bulunma süreci başladı. Bu istekler, kullanıcılardan ihtiyaçlarını anlamak için onlarla görüşen sistem analizcisi kavramını ortaya çıkardı. Programlar genelde COBOL programlama diliyle yazılıp, derlendi ve test edildi. Derlemek için JCL (Job Control Language) ortaya çıktı.

2.7.2 Dördüncü Kuşak Diller (4GLs) ve Bilgi Erişimi

Biriken ve sürekli büyüyen, istek yığınları, kullanımı daha kolay yeni programlama dillerinin geliştirilmesine sebep oldu. Dördüncü kuşak diller olarak ifade edilen bu diller programcılarının ve sofistike son kullanıcıların kendi üretim sistemlerinden çok daha kısa sürede, istedikleri verileri alabildikleri programların yazılmasını sağladı.

4GL dilleri oldukça büyük bir üretkenlik gelişmesi sağlamış olsa da, bu olağanüstü büyük kullanıcı isteği birikimi, başka çözümlere de gereksinim olduğunu çok kısa sürede orta çıkardı.

2.7.3 Bilgi Merkezi

1980'li yıllar boyunca, çok önemli ve büyük gelişmeler oldu. Bunların her biri, doğru bilgiyi, doğru kişilere, doğru zamanda ulaştırmayı amaçlıyordu. Karar Destek Yazılımı (DSS-Decision Support Software) ve Yönetim Bilgi Sistemleri (EIS-Executive Information Systems) şemsiye altında oldukça güçlü veri alma araçları ortaya çıktı. Böylece bilişim (IT) departmanları daha fazla servis yönelimli yerler halini aldı.

Aynı zamanda, çok önemli bir yenilik ortaya çıktı. Bu yenilik, ilişkisel teknolojinin pratik uygulaması idi. İlişkisel veri yapıları yerlerini “ad hoc” denilen anlık (doğaçlama veya irticalen) oluşturulan yapılara veya karar destek veri erişimine bıraktı. Çünkü bu yenilik, İngilizceye çok benzer bir yapıya sahip olması sebebiyle, sıradan yani programcı olmayan kişilerin de kullanımı için kolaylaştırıcılık sağlayan yeni bir yapıyı getirmişti. Tabi bu durum,

programcılarının yükünü hafifletme ve anlık ihtiyaçları kullanıcıların kendi başlarına çözebilmelerine olanak tanımış bulunuyordu artık.

2.7.4 Bilgi Mühendisliği ve Bilgisayar Destekli Yazılım Mühendisliği (CASE) Araçları

Bilgi Mühendisliği ve Bilgisayar Destekli Yazılım Mühendisliği (CASE) araçları programlar içinde yeni yapılar yaratmaya teşvik etti. Bilgi Mühendisliği ile, yeniden yapılandırılmış veri ve veri uygulama sistemleri ile, ki bunlar ortak veri standardı oluşturdular, verinin tek bir ortak kaynaktan paylaşılıp kullanılmasını sağlayan entegre bir uygulama ortamı oluşması sağlandı. CASE araçları sistem geliştiricilerin uygulamaları modellemelerine ve bu yapıyı kullanarak uygulama geliştirmelerine olanak sağladı.

Bir çok sebepten dolayı, CASE araçları ve Bilgi Mühendisliği birlikte uzun süre devam edemedi. Veri modelleme, veri yönetimi açılarından bakıldığında temel metodların, tekniklerin ve araçların kullanımı devam etti. Özellikle, CASE teknolojisinin modelleme komponentleri, veri ambarlarına geçişte bir temel teşkil etmiştir.

2.7.5 Bilgi Ambarı

Tüm uygulamaların ve veri tabanlarının mimari tasarımını değiştirmektense, bilgi ambarı teknolojisi denen bağımsız bir kurumsal-çaplı veritabanının ilişkisel teknoloji kullanılarak inşa edilmesi ve kurumsal veri standartlarına dönüştürülmesi süreci başladı. Periyodik olarak, veri operasyonel veri depolarından seçilerek alınıp, bir takım veri dönüşüm tekniklerinden geçirilerek bir bilgi ambarında tekrar depo edilmesi işlemi yapılmaktaydı.

Bilgi ambarı bilişimde karşılan problemlerin bir kısmını adresleyip çözüm olabilse de, yaratılması ve koordine bir şekilde yönetilmesi oldukça pahalı idi.

2.7.6 İstemci / Sunucu Mimari

İstemci/Sunucu, halen kısmen varlığı devam etse de “mainframe” olarak tabir edilen ana bilgisayarların sonunu getirmek, yeni teknolojiler getirmek üzere çıkmıştır. İş istasyonları ve LAN olarak tabir edilen yerel alan ağlarında çalıştırılmak üzere her uygulama yeniden yazılacaktı. Bununla beraber, bu anabilgisayarların ve üzerlerindeki uygulamaların oldukça

yüksek hacimli bir dönüşümü idi. Anabilgisayar çok-katmanlı mimaride baş aktör olarak görülmeye başlandı ve akabinde ise veri ambarı doğdu.

2.7.7 Veri Ambarları ve Data Mart'lar

Yüksek seviyede tanımlamak gerekirse, bir **veri ambarı**, tüm iş sistemlerinin ayrı bir veri sözlüğüdür. Farklı sistemleri bir araya getirir. Böylece şirket, bilgilerini yaratma, görüntüleme ve analiz etme yeteneği kazanılmış olur.

Bill Inmon, veri ambarı teknolojisinin babası (gurusu) sayılan kişi, 1990'lı yılların başında veri ambarcılığının formalize edilmesini sağlamıştır. Bu formalizasyon, kurumsal-çaplı veritabanlarını, üretim veya daha genel anlamıyla operasyonel veritabanlarından ayırmıştır. Aynı yıllarda Microsoft firmasının Access raporlama aracının çıkması ile de yöneticiler kolaylıkla standart tabir edilen raporları kendi başlarına hazırlamaya başladılar.

Bununla beraber, bütçe ve kaynak kısıtlamalarından ötürü, bir çok organizasyon bu kadar büyük çapta veri tabanlarını yaratamadılar. Bunun yerine, ya entegre bir veri ambarı projesini ertelediler veya daha basit olan “data mart” dediğimiz uygulamayı tercih ettiler. Data mart, departman seviyesinde bir uygulama veya konu spesifik bir veritabanıdır. Data mart, bir departmanın spesifik sorgulama ihtiyaçlarının kümesini veya oradaki fonksiyonel son kullanıcıların ihtiyaçlarını, desteklemek amacıyla tasarlanır. Bir grup üzerine yönlendirilmiş sorgulamalar aşağı yukarı benzer yoğunlukta ve tahmin edilebilir bir performansa sahiptir. Yani tasarımı çok kötü değilse önemli performans problemleri yaşanmaz.

2.7.8 Veri Madenciliği

Veri ambarı fenomeninin arkasında standart rapor çalıştırma ve yukarıda bahsedilen yoğun talep birikimi iki ana sebep olarak yer alıyor ise de, gerçek rekabetçi avantaj aslında veri madenciliğindedir.

Veri madenciliği, sofistike veri analizi araçlarının kullanımını ve zaman ölçeklendirilmesi yapılmış verinin kullanımını içerir. Zaman ölçeklendirilmesi yapılmış (benchmarked) veri, zaman boyutuna sahiptir, örneğin; gün, ay ve çeyrek yıl (üç aylık) periyotları gibi. Veri ambarları bu tip verileri destekler, çünkü bildiğimiz üretim verisi gibi tek boyutlu değil, çok boyutlu olarak kendini gösterir. Bu zaman elemanı, iş analizcilerine veriyi çekip çıkarma yada veri madenciliğini sağlar. Eğilimleri ortaya çıkarabilir ve iş sonuçlarının analiz edilip

grafiklendirilmelerini sağlar. Veri madenciliğinin faydasının daha iyi anlaşılabilmesi için çok bilinen iki örnek verelim:

Büyük perakende mağazlarından biri bira (yüksek karlı bir ürün) ve çocuk bezi (düşük karlı bir ürün) arasındaki satınalma ilişkisini ortaya çıkardı. Kesinlikle böyle bir ilişki kullanıcılar tarafından ne bekleniyordu, ne de test edilmişti, ama veri madenciliğinin kullanımı sonucu bulundu. Ya netice? Çocuk bezlerinin ve biraların mağazalarda sadece yerlerinin değiştirilerek yakınlaştırılması ile bazı ürün-birleştirme promosyon çalışmalarının pazarlama gücünü arttırdığı ve tek ürün satmak yerine bu şekilde daha çok ürünün satılıp gelir artışının olabileceği gerçeği ortaya çıkmış oldu.

Büyük kredi kartı şirketlerinden biri, çalınan kredi kartları ile 1 dolar ve altında satılan ürünlerin (örnek gazolin) satışları arasında bir ilişki olduğunu buldu. Tekrarlamak gerekirse, bu ilişkide bir önceki örnekteki gibi ne beklenenen, ne de test edilen bir durum idi. Ya netice? Çalınan kredi kartlarının tespiti ve potansiyel suçlardan korunma için yeni güvenlik kriterleri ile, bu çeşit küçük ve hızlı satınalmalara dikkat edilmesi gerekliliğine dikkat çekiyorlar.

2.7.9 Bir Veri Ambarının Yaratılması

Bir veri ambarı yaratılması süreci, yapılacak çalışmanın hedeflerinin tanımlanması ve kısa veya orta vadeli bir proje olduğuna karar verilmesi ile başlar. Uygun yazılım araçlarının araştırılması ve belirlenmesi gereklidir. Ayrıca maliyet ve etkinin izlenmesi ve kurumsal çaplı bir modelin desteklenmesi için tepe yöneticilerinden birinde bu işten sorumlu olması gereklidir. Yönetim desteği birinci şarttır, bu durum bildiğimiz toplam kalite metodolojisi içinde geçerli bir durumdur.

2.7.10 Veri Modeli ve “Meta Data” Yaratılması

Birleşik bir mantıksal veri modelinin tasarlanması ve inşa edilmesi gereklidir. Kurumsal çaplı bir model hem müşteri hem ürün gibi tüm sistemlerin verilerini kombine bir yapı içine koymalıdır. Meta data kesinlikle tanımlanmalıdır. Çünkü meta data veri ambarında neyin kayıt edilip saklandığını tanımlayan yegane yapıdır.

2.7.11 Veri-Zaman Ölçeklendirmesi (Benchmarking)

Daha önce de belirtildiği üzere, zaman faktörü, verinin ambara akıtılmasındaki birincil kriterdir. İş analizi yapıldığında hangi zaman aralıklarının işimiz için uygun olduğuna karar vermemiz gereklidir. Ayrıca bu bilginin de meta data içinde yeralması gerekir.

2.7.12 Verinin Temizlenmesi ve Dönüştürülmesi

Veri temizlenmesi prosesi, çeşitli işlemlerin yapıldığı bir süreçler bütünüdür. Eksik veya kötü veri ile uğraşılması ve veri tutarlılığı kontrollerinin yapılması gibi işlemler ile uğraşılması gereklidir. Veri dönüştürülmesi esnasında, kodlama birliğinin sağlanması için ve iş terimlerinin doğru ve ikame edilebilir tekil bir yapıya kavuşturulabilmesi için “look-up” dediğimiz kod referans tablolarına ihtiyaç vardır.

2.7.13 Kullanıcıların Eğitimi

Veri ambarı kullanıcıları mevcut kullanımda olan veri erişim araçlarının nasıl kendi işlemlerinde kullanılabileceklerini ve elde ettikleri verinin genel olarak ne anlama geldiğini bilmeleri gerekmektedir. Yüksek seviyeli analizler için ise; son kullanıcıların sonuçları en iyi şekilde değerlendirebilmesi için istatistik alt yapısı olan iş analizcilerinden oluşması gereklidir. Kullanıcı eğitimi bu yeteneklerin oluşmasına ve/veya kullanılmasına yardımcı olur.

2.7.14 Veri Ambarına Yatırım

Bütünleşik bir veri ambarı girişiminde için, hem geliştirme hem de işleme alma süreçlerini desteklemek için çok önemli teknolojik yatırımlara kaynak ayrılması gereklidir. Bu maddenin alt maddeleri tam bir veri ambarı girişimi ve çalışması için gerekli olan ürün, tekniklerin ve teknolojilerin özet bir tanımın içermektedir. Bunlar aşağıdaki gibidir:

2.7.14.1 Veri Aktarımı

Mevcut çalışmakta olan operasyonel veri dosyalarından ve dış veri kaynaklarından verinin seçilmesi ve dönüştürülmesi, verinin iyileştirilmesi ve veri ambarı veya data mart'a yüklenmesidir. Bu ürünler veri ambarının ilk inşası ve devam eden kullanımında orijinal veri kaynaklarından verinin güncellemesi ve yenilenmesi yani ambarın sürekli beslenme işlevini

yerine getirirler. Veri aktarım araçları, veri modelleme ve veri kalite araçları ile beraber kullanıldığında başarılı sonuçlar üretir.

2.7.14.2 Veri Kalitesi

Verinin veri ambarındaki standartlara uyup uymadığını belirler. Bu araçlar filtreleme, temizleme ve dönüşüm ile veriyi arzu edilen spesifikasyonlara getirirler. Bu araçlar ayrıca bazı veri eşleme ve birleştirme tekniklerini de kullanırlar.

2.7.14.3 İş Akı

Kompleks iş kurallarını anlamak zorunda kalmadan, veri yapılarını ve doğal sorgulama dillerini (örneğin SQL) bilmeden veri ambarı veya data mart'lara erişime olanak sağlayan bir kullanıcı arayüzü sağlar.

2.7.14.4 Uzaktan erişim

Uzak lokasyonlardan, örneğin internet veya intranet imkanlarını da kullanarak, kullanıcılara sorgulamaların iletilmesi, sonuçların veri formu veya rapor formatında dağıtılması gibi olanakları destekler. Sorguların iletilmesinden ziyade, sonuçların dağıtılıp ilgili kişilere ulaştırılması daha öncelikli ve daha önemlidir. Aslında bu özellikleri destekleyen yazılımlar tüm ürünlerde gömülüdür, neredeyse standart halini almıştır.

2.7.14.5 Metodoloji

Bir veri ambarı veya data mart çalışmasındaki uygulanabilecek yaklaşımlar; her spesifik işlem veya adımda gerekli olabilecek proje yönlendirme ve yönetim metotlarının çeşitlerini sunar.

2.7.14.6 Uygulama Jeneratörleri

Günlük veri işleme mantığını, veri tabanı veya dosya yönetimini, hatta raporlamayı da içine alan tam bir uygulama üretir. Bazı uygulama jeneratörleri model-tabanlıdır, bazıları ise değildir. Bunların bir kısmı standart (örneğin COBOL) dediğimiz dillerde kod üretse de bir kısmı ise kendine özel (proprietary) derleyici ve yorumlayıcı dillere ait kod üretir.

2.7.14.7 Veri Havuzu (Repository), Sözlük (Dictionary) veya Klasör (Directory)

Meta data (veri hakkında veri) dediğimiz, veri karakteristiklerini (uzunluk, tip), eksiksiksiz veri tanımlamalarını, para birimi, verinin; kaynağı, sahibi, kullanım yeri ve veriyi anlamaya yetecek diğer tüm özellikleri depo eder. Bazı araçlar hem profesyonel bilişimci hemde son kullanıcıların iş amacına yönelik içerikleri görmelerini sağlamıştır. İşe özel bilgileri son kullancılara, teknik bilgileri bilişimcilere göstermek ve en başta bu mantığa uygun hem iş hem de teknik bilgileri içerecek tasarım yapmak gereklidir.

2.7.14.8 Veri Modelleme

Grafiksel olarak veri ambarındaki verinin yapısını ve ilişkilerini gösterir. Genellikle, mantıksal bir modelin başlatılmasını sağlar. Önerilen/düşünülen veri ambarı veya data mart'ın fiziksel tasarımını bu modelden geçerek yapma yeteneğine, uzun yıllardır bir çok veri modeli aracı sahiptir.

2.7.14.9 Veri Madenciliği

İş analizcilerinin, daha önceki örneklerde belirtilen tarzdaki, ilişkileri ve eğilimleri araştırmalarını sağlar. Tüm sorgulamalar hipotez-tabanlıdır. Bu sorgulamalar, üstü kapalı veya derinine inmeden bir çırpıda bulunamayacak olan, yani aşık olmaya ilişkin ilişkileri, ortaya çıkarmak için kullanılırlar.

2.7.14.10 Arayazılım (Middleware)

İki veya daha fazla yazılım ürününün arasındaki akışı düzenleyen arayüz yazılımıdır. Veri ambarı ortamında, özellikle veritabanı özel (proprietary) ise, arayazılım bu özel veritabanına erişimi sağlayacak ve özel sorgulamalar yapılabilmesini sağlayabilecek bir üründür. Mevcut her yazılım aracının, kendine özel bir erişim arayüzü yazmasına gerek kalmadan bir çok araca tek bir ortak erişim metodu işlevsel olarak, tek bir arayazılım tarafından sağlanır, ekstra çıkacak programcı iş yükünü azaltıp, maliyetlerin düşmesini sağlar.

2.7.14.11 İlişkisel VTYS (Veri Tabanı Yönetimi Sistemi)

Verinin ve dosyaların yönetimini sağlar. Veri, satırlar ve sütunlardan oluşan tablolar şeklinde organize edilmiştir. Tablolar ise birbiriyle ortak veriler aracılığıyla ilişkilendirilmiştir.

2.7.14.12 Replikasyon

Verinin kaynak ve hedef noktaları arasında, senkronizasyonu muhafaza edilerek, bir ortam veya platformdan diğerine kopyalanmasını sağlar. Replikasyon yazılımının tipik kullanımı için; verinin üretim sistemlerinden veri ambarına kopyalanması veya bir veri ambarından diğerine yada bir çok data martlara kopyalanması işlemleri örnek olarak verilebilir.

2.7.14.13 Performans Modelleme ve Ölçümü

Veri merkezinin, veri setlerinin yerleşikliğinin organizasyonunun yapılması, verinin yedeklenmesi, depolanması ve problem esnasında kurtarılmasını sağlar. Veri modelleme ve ölçüm araçları öncelikle veri merkezi yöneticilerinin kullanımı içindir.



3. VERİ AMBARI , DATA MART, VERİ MADENCİLİĞİ

Bu bölümde; veri ambarı, data mart ve veri madenciliği konuları incelenmiştir. Veri ambarının kavramsal seviyede çeşitli tanımları yapılmış, tarihçesi incelenmiştir. İş sistemleri ve veri ambarları arasındaki farklar incelenmiştir.

3.1 Veri Ambarı Kavramı ve Tarihi Gelişimi

Veri ambarı kavramı (aynı isim altında olmamasına rağmen) çok uzun bir geçmişe sahip değildir (Porter ve Rome, 1994). Karar yapıcılarının enformasyonel ihtiyaçları sürekli mevcuttur (Shah ve Milstein, 1997). Bir veri ambarı, karar vericilerin desteğinde zamanında ve sağlıklı, entegre, zaman değişimli, uçucu olmayan veri birikimidir (Cobb, 1996; McClanahan, 1996).

Bir veri ambarı nedir? Stanford Üniversitesine göre;

“Bir veri ambarı, sorgulamalar ve analizler için hazır olan bir entegre bilgi havuzudur. Veri ve bilgiler üretildikleri çok değişik kaynaklardan seçilirler. Bu çok değişik kaynaklardan orijinal olarak gelen bilgi üzerinde, çok daha kolay ve çok daha efektif bir şekilde sorgulamaların çalıştırılmasını sağlar (www.datawarehousing.com/whatis.asp).

3.1.1 En Kısa Şekliyle Kavramsal Tanımlar

- **Veri ambarı** — “Aktif iş sistem dosyalarından ayrı olarak, analizde kullanılmak için oluşturulan iş bilgisi havuzudur (Silvon Software, 2000a).
- **Data mart** — İş-yoğunluklu veri ambarıdır. Genellikle özel bir iş alanına (örneğin satış) yoğunlaşılır.
- **Veri madenciliği** — Veri içindeki kalıpları belirtmek için tasarlanan analiz teknikleridir. Data mart ve veri ambarındaki içerikler uygulanabilir.
- **OLAP (Anlık Analiz İşlem)** — Çok boyutlu verilerin interaktif analizini destekleyen araçlar ve uygulamalardır.

3.2 Veri Ambarı: Geçici Olmayan Beklentiler?

Sayıları giderek artan organizasyonlar, veri ambarcılığının sofistike veri analiz seviyesine ve bilgi tabanlarının kontrol imkanına olanak vereceğini ümit etmektedirler. Kesinlikle veri ambarcılığı, veri yoğunluğu fazla olan departman veya şirketler için zorunludur (Silvon Software, 2000b).

İş Akıllı (BI) sistemleri kurumun iyi kararlar üretme yeteneğinin geliştirilmesi için tasarlanırlar. İş Akıllı, bilginin kurumun en üst seviyesi olan tepe yönetiminden, en alt seviyesi olan bilgi işçilerine kadar her seviyede ve yeteri miktarda erişilebilir olmasını sağlar. Çok iyi dizayn edilmiş iş akıllı sistemleri; veri ambarcılığı, on-line analitik işleme (OLAP), işlemci / sunucu mimarisi, internet ve intranetler yeni ilişkisel veri tabanı yönetim sistemleri (RDBMS) , paralel işlemciler ve gelişmiş mikrosesler gibi teknolojilerden yararlanır. Bu teknolojilerden başarılı bir İş Akıllı sistemi için en önemli araçlar, veri ambarı (veya konu spesifik data mart) ve OLAP motoru teknolojileridir.

3.3 Analiz, Veri Toplama

1970'li yıllardan beri, şirketler ana iş fonksiyonlarını bizzat yürütmesi veya yardımcı olması sebebiyle bilgi teknolojilerine güvenmişlerdir. Bu gün veri ambarcılığı ve İş Akıllı sistemleri; kompleks, çok yönlü, çok boyutlu zaman serili analizlerini desteklemekte, hatta sadece veriyi toplayıp, gözetlemenin çok ötesine taşımaktadır.

Eski BT (Bilgi teknolojisi) sistemlerinin ürettikleri bilginin sınırlı değere sahip olmasında iki ana sebep mevcuttur. Birinci eksiklik bu eski teknolojilerin ürettikleri rapor çıktılarını içerir, ikinci sınırlama ise veri erişimiyle ilgilidir.

Birinci eksikliği inceleyelim. Eski teknolojiler ile kurulmuş sistemler “ad hoc” sorgulamaları (örneğin “what if “-“ eğer...?”) desteklemez. “Ad hoc” sorgulamalar insanın düşmesine benzer. Bir durum seçilir ve analiz edilir, daha sonra soru üretilir. Böylece daha çok bilgiye ihtiyaç duyulduğundan bu sorgulamalar ile o bilgilere erişilebilir. Önceki BT sistemleri; BT personelinin direk müdahalesi ile ancak, standart ve sabit raporlar üretilir.

İkinci eksiklik; eski BT sistemlerinin veriye erişimi idi . Tam bir iş bakışı kurabilmek, oturtabilmek için, tek bir dosya kaynağını karar-yapıcılarının erişimine açmak oldukça

yetersizdir. Gerekli bilgi ancak tüm organizasyonun her tarafına dağılmış ve kullanılmakta olan değişik veri dosyalarından elde edilebilir.

3.4 Çok Boyutlu Veri Görünümü

Veri ambarcılığı tüm veri dosyalarını merkezi bir veri havuzuna koyarak işlemler üzerine bina edilmiş çok boyutlu bir iş görünümü yaratır. Bunu yapmak için, veri ambarcılığı bağlantıları ve ilişkilerini “ denormalizasyon ve özetleme “ denen bir prosesi kullanılır.

Veri ambarcılığı ve uygulama data mart teknolojisine ilaveten OLAP motoru da başarılı bir İş Akli sistemi için ana teknolojidir. Bir OLAP motoru, aslında bir veri ambarının nasıl yapıldığını anlayan, optimize edilmiş bir sorgulama jeneratörüdür. “What if“ sorgulamaları dediğimiz, değişik koşul ve önermelerde nasıl sonuçların çıkacağını ortaya koyan yapıyı kullanarak doğru bilginin veri ambarından alınabilmesini sağlar.

3.5 İş Sistemleri ve Veri Ambarları Arasındaki Farklar

Veri ambarcılığı konsepti 1980’lerin sonunda başladı. Prensip, o zaman ve hala, mümkün mertebe daha fazla bilgi, mümkün mertebe daha hızlı iş ve çok düşük hata oranlarına sahip olmaktır.

“Veri ambarı”, iş sistemleri tarafından toplanan faydalı bilginin bir havuzda toplanması mantığıdır . Buradaki fark , veri ambarının kolay kullanım erişimi için dizayn edilmesidir. Bu fark iki sabit problemin çözülmesini sağlar. Bunlar:

- *İş karar – yapıcılarının kurumsal bilgiye zamansal erişim yeteneğinin kazandırılması.* — Çoğu operasyonel sistem yüzlerce tablo ve binlerce olan isimlerinden oluşan bir tasarıma sahiptir. Bu tasarım kullanıcılar açısından bakıldığında onlara karanlık bir yapıdır. Yapılan araştırmalar neticesinde eğer bir veri tabanında 22 tablodan daha fazla tablo var ise kullanıcılar verileri nereye konumlandıracağı noktasında zorluklar yaşamaktadır. İşte karmaşa bu noktadadır.

- *Sonuçların tutarlılığı*— Eğer herkes aynı veri kaynağını kullanıyorsa , rapor sonuçları senkronize edilmelidir. Sonuç tutarsızlıkları , iş (transaction) sistemlerinin doğal yapısından kaynaklanmaktadır. Bu online sistemlerde her yeni kayıt eklendiğinde bilgi içeriği değişir.

Örneğin; bir patron ve birde satış müdürü sistemden satış raporu almış olsun. Müdürün raporu sabah saat 10:00 da alınmış ve belirlenen kotaların % 25 altında olduğunu göstersin. Patron ise saat 12:15 te raporu çalıştırmış olsun. Satış eğilimlerinin tartışılacağı toplantının da saat 13:00 de yapılması planlanmış olsun. Saat 23:00'te de 50 milyon dolar tutarında bir satış aktivitesi sisteme işlenmiş olsun. Satış müdürünün yapacağı sunumda (ki hazırlanması için önceden rapor alması mantıklı) satışta problem olduğu ve kotaların altına düşme sebeplerini sıralanırken, patron ise %75 kota fazlasını gösteren rakamlara bakmaktadır. Bu ise başka bir karmaşayı göstermektedir.

3.6 Veri Ambarını Uygulamaya Alma (Implimentation)

Teoride geleneksel veri ambarı, her kullanıcının isteyeceği verinin tutulduğu, merkezi olarak yürütülen bir veri saklama yeteneğidir. Bu sebepler kaçınılmaz olarak teknoloji yönetimidir. Maliyet çok yüksektir (3 ile 5 milyon dolar arası), devreye alma süresi uzundur (12 ile 18 ay arası). Bu kadar büyük kurumsal çaplı uygulamalar finans ve zaman maliyeti yüzünden yüksek başarısızlık oranlarına sahiptir.

3.7 Data Mart ve Veri Madenciliği Seçeneği

Bunun yerine data mart'lar tercih edilmektedir. Bu; bir uygulamalar veri ambarı ile aynı fonksiyonları (kolay kullanıcı erişimi, veri senkronizasyonu, vb) sunsa da , genelde departman ve bölümdeki spesifik iş problemleri amaçlanmıştır. Maliyetler çok daha küçük (tipik olarak 500,000 doların altında, bir veri ambarı maliyetinin %10'undan az), devreye alma süresi ise 6 aydan (tecrübeli bir uyarlamacı düşünüldüğünde) azdır. Bir data mart MS Excel, Lotus 1-2-3 gibi elektronik tablo programlarında oldukça sofistike araçlara kadar kullanıcı erişimini destekler. Yakın zamanda veri madenciliği, veri ambarı karşısında bir diğer seçenek olmuştur. Aslında bu, kullanıcıların anlamakta zorluk çektiği bir teknolojidir. Veri ambarı/ data mart iş karar-yapıcıları tarafından gerçekleştirilen bilgi analiz problemlerinin bir çözümüdür. Veri madenciliği ise operasyonel iş sistemleri içindeki örneklemelerin (pattern), kalıpların temsil edilmesi metodudur. Bir veri madenciliği çok uygulaması hacimli işlem verisinin toplanmasıdır. Veri madenciliği sistemi belirli bir kalıp bulabilmek için elekten geçirir. Veri madenciliği büyük bir ekseriyetle perakendeciler tarafından satış noktalarındaki satın alma örneklemelerini bulmak için kullanılır. Şu anda veri madenciliği hala emekleme aşamasındadır. Oldukça pahalıdır ve yüksek nitelikli kullanıcılar ve araçlar istemektedir.

3.8 Veri Ambarı, Data Mart ve Veri Madenciliği Karşılaştırması

Yukarıdaki bilgileri tek bir tablo altında toplar isek, çok daha net olarak farklılıklar gözlemlenebilecektir (Çizelge 3.1).

Çizelge 3.1 Veri ambarı, data mart ve veri madenciliği karşılaştırması (Silvon Software)

Kriter	Veri ambarı	Data mart	Veri madenciliği
Harcama-maliyet	3-5 milyon USD	≤ 0.5 milyon USD	Çok pahalı
Zaman	12-18 ay	6 ay	Daha çok yeni,
Karlılık, fayda oranı	% 300	% 650	Tahmini çok zor

Bir başka ve detaylı karşılaştırma ise veri ambarı ve data mart arasında yapılmıştır (Çizelge 3.2).

Çizelge 3.2 Veri ambarı, data mart karşılaştırması (Strange, Ağustos 1997)

Kriter	Veri ambarı	Data Mart
Odak	• Uygulama bağımsız • Merkezi, paylaşıma açık • Karşılıklı iş hattı, kurumsal	• Özel uygulama gereksinimi • İş hattı departmantal veya kullanıcıya özel • İş sürecine bağımlı
Verilerin Durumu	• Geçmişten gelen veri • Rasgele veri dağılımı	• Ayrıntılı • Özetlenmiş veri
Konular	• Çoklu konu alanları	• Tek konu • Çok parçalı konu alanları • Operasyonel kaynaklardan enstantaneler
Veri Kaynakları	• Çoklu • Operasyonel veri, harici veri	• Birkaç • Operasyonel, harici veri • OLTP veritabanı enstantaneleri
Uyarılma Süresi	• İlk aşamada 9-18 ay (iki yada üç konu alanında) • Birkaç aşamada kurulum	• 4-12 ay
Özellikler	• Esnek • Uzun vadeli, stratejik • Veri temelli	• Sınırlı • Kısa ömürlü, taktik • Proje temelli

3.9 Veri Ambarı / Data Mart Oluşturma

Bir veri ambarı / data mart oluşturma adımları aşağı yukarı aynıdır. Burada pazarın %75'lik kısmını oluşturan data mart'ların oluşumuna konsantre olunmuştur.

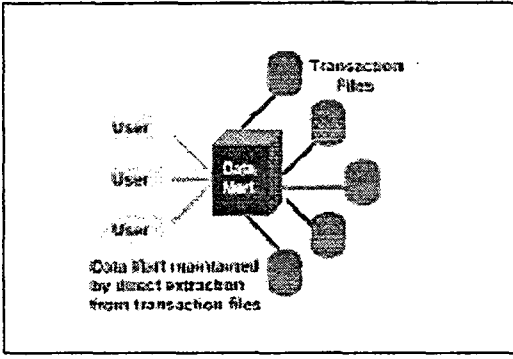
- Öncelikle “sürücü” bir kullanıcı belirlemek (Yani tek bir sorumlu). Teknoloji grupları tarafından sürdürülen data mart projeleri nadiren başarılı olmuştur.
- Kullanıcıların ihtiyaçları ve isteklerinin neler olduğuna karar vermek.
- Veri kaynağının nerede konuşlandırılacağını belirlemek.
- Data mart, veri tabanı yönetim sistemini (DBMS) tasarlamak.
- Kaynak veri dosyalarından gerekli verileri seçmek.
- Hedef data mart için kabul edilebilir bir formatta kaynak veriyi “dönüştürmek/ yeniden formatlamak”.
- Data mart’a yüklemek.

Bazı yerlerde, bu adımlar boyunca, kaynak iş dosyalarında önemli bir sayıda hatayla karşılaşılır. Bunlar “veri temizleme(data cleansing)” araçları ile çözümlenebilir.

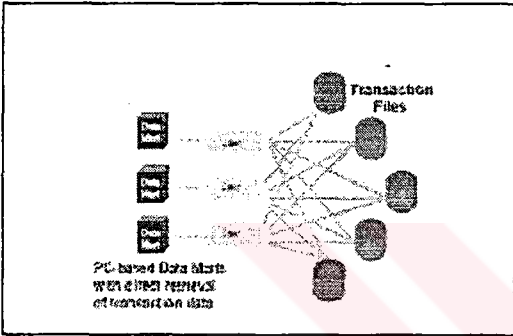
3.10 Diğer Veri Ambarı / Data Mart Kriterleri

Veri ambarcılığı giderek hareketlenen, kalabalıklaşan bir pazardır. Bu yüzden doğru varar verebilmek için aşağıdaki hususlar dikkate alınmalıdır.

- **Ölçeklenebilirlik** ana konudur. İlk kurulduğunda, data mart’ın kullanımı ve büyüklüğü geometrik olarak artan yeni kullanıcı lisansı ve veri saklama ucuzdur. Burada gözden kaçan nokta, bu ucuzluğun yanı sıra son step olan yükleme (loading) süresidir. Veri tabanları saklama konusunda ölçeklenebilir. Sistemler ise çok daha kullanıcı için ölçeklenebilir. Operasyonel çerçevede data mart’ın güncellemesi için müsait midir, bu ölçeklenebilir mi? Çoğu firmaların, en fazla 8-10 saatlik bir data mart güncelleme operasyon çerçevesine sahiptir. Eğer veri hacimleri artarsa, data mart daha fazla veriyi aynı zaman çerçevesi içinde güncelleyebilecek mi? İşte ölçeklenebilirlikten kasıt budur.
- **Uygulamaya alma** diğer önemli konudur. Bazı firmalar “tam uygulamaya alma” servisi sunarken bazıları ise sadece ürünlerini satıp, kendi uzmanlık tecrübenizce sizi baş başa bırakırlar veya danışman kiralamak zorunda kalırsınız. Uyarlamacının, endüstrinizdeki tecrübesinden emin olunmalıdır. Data mart teknolojisini bilmek savaşının ancak yarısıdır. Endüstriyi bilmek ise, başarılı bir uygulamaya alma için diğer %50lik kısmı tamamlar.
- **Ürün mimarisi** önemlidir. Data mart veya OLAP satıcıları iki ana yaklaşım sunmaktadır:



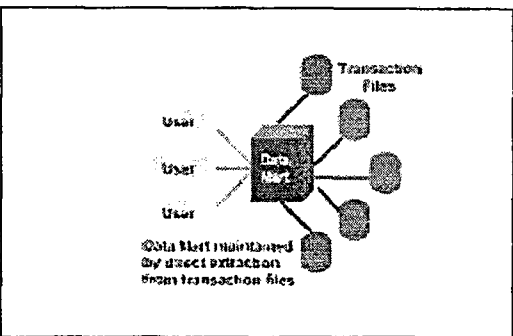
Şekil 3.1 Sunucu-tabanlı yaklaşım (Silvon Software, 2000)



Şekil 3.2 İstemci-tabanlı yaklaşım (Silvon Software, 2000)

PC-tabanlı yaklaşımda, her kullanıcının lokal veri tabanını ne zaman güncellediği problemi yaşanır. Tekrarlamak gerekirse, yüzlerce değişik rapor sunucu ile karşı karşıya kalırız. Eğer raporların senkronize olması ana kriter ise, sunucu-tabanlı yaklaşım yani kullanıcının direkt eriştiği data mart yaklaşımı tercih edilmelidir.

- Data mart uygulamaya alma kendi açısından iki seçeneğe sahiptir.

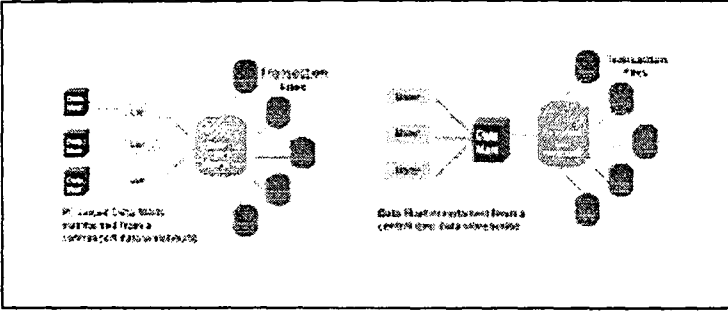


Şekil 3.3 Data mart uygulama opsiyonu-1 (Silvon Software, 2000)

– **Bir Sunucu-tabanlı yaklaşım.** Belli düzenli aralıklarla iş sistemlerinden verinin seçilerek data mart'ın yüklemesi yapılır. Ardından data mart sunucusunda saklı veriyi analiz etmek için kullanıcılar kendi işlemci (PC) yazılımlarını çalıştırırlar.

– **Bir istemci PC-tabanlı yaklaşım.** PC-tabanlı yaklaşım veri analizi yazılımını ve çok boyutlu data mart'ın her ikisini de aynı PC'ye koyar. Kullanıcılar veriyi direkt olarak iş sisteminden seçip alır ve PC tabanlı analiz sistemine yükler.

– İş sistemlerinden direk seçimler ile data mart yaratmak ve yaşatmak.



– Önce kurumsal bir veri ambarı yaratmak, sonra bu merkezi veri ambarında lokal data martlar seçip bunları yaratmak.

Şekil 3.4 Data mart uygulama opsiyonu-2 (Silvon Software, 2000)

Yukarıdaki grafiklerde görüldüğü üzere, merkezi kontrollü veri ambarı / işlemci yaklaşımı, tek bir data mart ile veya PC-tabanlı data martların merkezi veri ambarından seçimlere dayandırıldığı bir başka yapıyla configure edilebilir.

Aynı yaklaşımlar ve faydalar burada uygulanabilir. Tek bir data mart ile, tüm kullanıcılar aynı veriye bakabilirler.

- Aynı zaman içinde, aynı veri sunumuna herkes bakabiliyor mu?
- Data mart'lardaki tüm güncellemeler, merkezi ambara senkronize yapılabiliyor mu? Veya tüm ilişkisel veriler aynı anda güncellenebiliyor mu?

Bu ikisi genelde bir politika kriteridir. Merkezi bir ambar BT merkezi tarafından uygulamaya alınırken, data mart iş kullanıcıları tarafından uygulamaya alınır. Data mart güncellemeleri merkezi ambar güncellemeleriyle bağımlı olmalıdır. Merkezi veri ambarında tüm güncellemeler yukarıdan aşağı tüm data martlara iletilmelidir. Merkezi veri ambarları, oldukça komplike ve oldukça teknoloji ağırlıklı çözümlerdir.

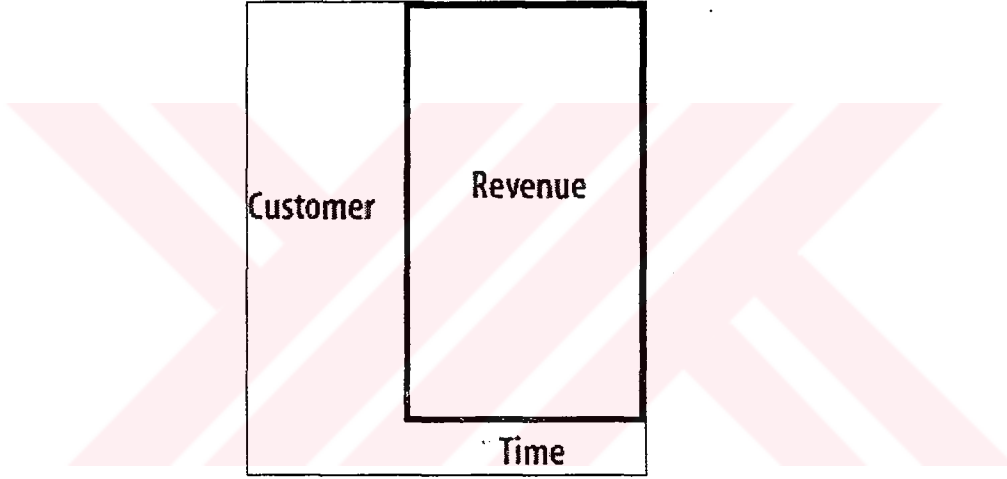
3.11 Veri Madenciliği Nedir?

Son yıllarda, veri madenciliği teknikleri şirketler tarafından müşterilerinin demografik profillerini anlamak ve kişisel özelliklerine uygun ürün ve çözümler sunmak amacıyla kullanıldı. Anlamlı bilgilere ulaşabilmek amacıyla, oldukça kompleks algoritmalar şeklinde çok değişik veri madenciliği teknikleri yazılımlara gömülmüştür. Son-kullanıcı veri madenciliği uygulamaları mevcut ise de, kullanıcılar tarafından genellikle anlaşılamadığından tüm organizasyon çapında yaygın olarak kullanılamamaktadır. Veri madenciliğinin

kapasitesini anlamının ve anlatmanın tek yolu, diğer İş Akı (BI) teknolojileri ile karşılaştırmaktır (Wu, 2000).

3.11.1 “Ad Hoc” Sorgulama

“Ad Hoc” sorgulama uygulaması ile; kullanıcılar, talep anında, veriye erişim yeteneğine sahiptirler. Ne sorarlarsa, onu alırlar. Örneğin; “Bu yıl boyunca, her bir müşteriden ne kadar gelir elde edilmiştir?” şeklinde bir soruya cevap verecek ad hoc sorgu bir kullanıcı tarafından yaratılır ve çalıştırılırsa, sorgulanma sonucunda müşteri adı ve bu yıl elde edilen gelir bilgisini gösterir. Şekil 3.5’de bu “ad hoc sorgu” sonucunun kümesel gösterimi görülmektedir.



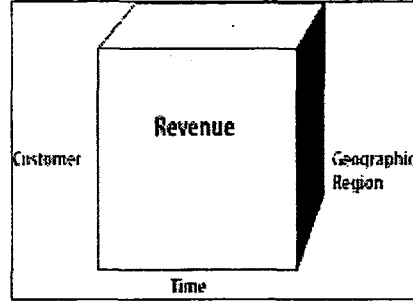
Şekil 3.5 “Ad hoc query” sonuç kümesinin gösterimi (Wu, DM Review, August 2000)

Müşteri bazında gelir toplanır ise yeni bir soru türetilir: “ Bu yıl ne kadar gelir elde edildi?”. Buna ilaveten şu sorular gelebilir: "En yüksek gelir hangi müşteriden elde edildi?" ve “ En düşük gelir miktarı hangi müşteriden elde edildi?” soruları cevaplanabilir. Bu sorgulama sonuçları, bir çok soruyu adreslese de, bu İş Akı (BI) teknolojisi, gizli örnekleme kalıplarını ve gizli ilişkileri belirleyemez. Kullanıcının isteği bu yıl müşterilerden ne kadar elde edildiği idi. Sonuçta aldığı bilgi yukarıdaki örneklerde görüldüğü gibi türetilse de bu eksendedir. Ne eksik ne de fazla.

3.11.2 Online Analitik İşlem (OLAP)

OLAP uygulamaları, kullanıcılara manuel olarak bilgiyi araştırma, özet ve detay bilgiyi analiz etme imkanını sağlar. Örneğin; “ Bu yılın her bir çeyreğinde coğrafi ve müşteri bazında elde

edilen gelir neydi?” sorusuna cevap verecek bir OLAP analizini bir kullanıcı yaratsa ve çalıştırırsa bu analizin sonuçları coğrafi bölge, müşteri ismi, çeyrekler bilgilerini içerecektir.



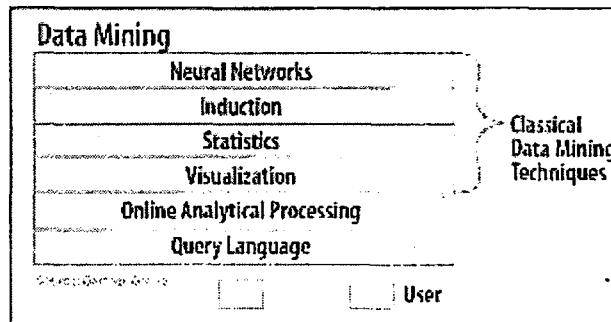
Şekil 3.6 OLAP analizi sonuç kümesinin gösterimi (Wu, DM Review, August 2000)

OLAP , sadece istenen veriler içindeki örneklemeleri, kalıpları ön plana çıkarabilir. Gerisi, yani OLAP analizi neticesinde ön plana çıkarılan trendleri ve örneklemeleri belirlemek kullanıcıya bırakılmıştır. Bu İş Akıllı (BI) teknolojisi de gizli örnekleme kalıplarını veya gizli ilişkileri belirleyemez.

3.11.3 Veri Madenciliği

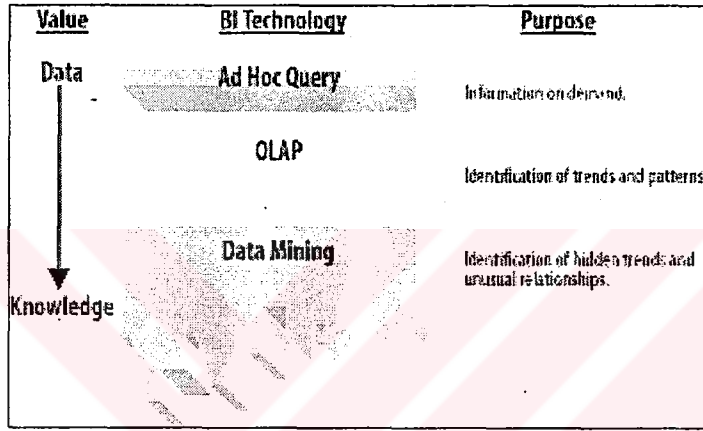
Veri madenciliği, bir veri popülasyonunda çok değişik tekniklerle anlaşılabilir. Gizli ve faydalı bilgiyi seçip alabilen, bir İş Akıllı (BI) teknolojisi olarak en iyi tanımlanabilir. Veri madenciliği gizli eğilimleri ve örnekleri çok geniş miktardaki verinin içinden keşfetmemize imkan sağlar (Brand ve Gerritsen, 1998).

Veri madenciliğinin bir başka tanımı ise, çok miktardaki veriyi iş avantajı için seçme, inceleme ve modelleme sürecidir (Timur, 2000). Çok çeşitli şekillerde veri madenciliği ve teknikleri uygulanabilir. Şekil 3.7, “kullanıcı katılımı”sırasına göre veri madenciliği tekniklerini göstermektedir.



Şekil 3.7 Kullanıcı katılımına göre veri madenciliği (Wu, DM Review, August 2000)

Ad hoc sorgulama uygulamaları bir veri tabanındaki veri değerini yüzeysel eşelemek olarak tabir edilir ise, OLAP kullanıcılara daha derinden ve anlamlı veriye ulaşma imkanı sağlar. Bununla beraber, veri madenciliği ise, daha derinleri kazar ve kullanıcılara gizli eğilimleri ve ilişkileri bu derinliklerde keşfetmelerine olanak sağlayan bilgiyi sunar. Veri madenciliğinin, bir ad hoc sorgulama ve bir OLAP analizi ile kombinasyonu ise, oldukça güçlüdür. Analiz edilen veri hakkında ve üzerinde ne tip aksiyonlar gerçekleştirebileceği hakkında anlamlı bilgi sağlar. Şekil 3.8’de İş Akıllı (Business Intelligence) Teknolojilerini birinci bölümde anlatılan değer piramidi mantığını da kullanarak, amaçları ile birlikte görülmektedir.



Şekil 3.8 BI teknolojileri değer ve amaç şeması (Wu, DM Review, August 2000)

3.12 Veri Madenciliği ile İlgili Sorulacak 20 (10 + 10) Soru

Sorulacak sorular işe özel sorular ve teknik sorular olarak 10 + 10 olmak üzere 20 adettir. Soru detaylarına girmeden; bu sorular kurulacak sistemin kalitesi, mimarisi, ölçeklenebilirliği, kullanımı ve sonuçtaki faydaları açısından önemlidir ve mutlaka bu temel kavramlarla ilgili daha alt sorular türetilmeli ve de kullanılmalıdır. Daha alt seviyedeki soru detaylar tez konusunun dışında olduğu için belirtilmemiştir (Parsaye, 1998).

3.12.1 Top 10, Veri Madenciliği : “İŞ” Soruları

İş kullanıcılarına sorulacak olan kurulacak sistemin kalitesi, kullanımı ve sonuçtaki faydaları açısından, 10 temel soru; (Information Discovery, Inc., 2001);

Soru 1: İş faydaları

Soru 2: Teknik birikim (know-how)

Soru 3: Sonuç anlaşılabilirlik ve bilgi açıklamaları

Soru 4: Takip eden sorular (follow-up questions)

Soru 5: İş kullanıcıları

Soru 6: Doğruluk, tamlık, tutarlılık

Soru 7: Artısal analiz

Soru 8: Veri yönetimi

Soru 9: Entegrasyon

Soru 10: Destek personeli

3.12.2 Top 10, Veri Madenciliği: “TEKNİK” Sorular

Teknik kullanıcılara sorulacak olan kurulacak sistemin mimarisi, ölçeklenebilirliği ve gücü açısından, 10 temel soru;

Soru 1: Mimari

Soru 2: Gerçek veriye erişim

Soru 3: Performans ve ölçeklenebilirlik

Soru 4: Çok tablolu veritabanları

Soru 5: Çok boyutlu analiz

Soru 6: Teşhis edilen şablonları tipleri ve sınıfları

Soru 7: Sistem girişi, insiyatifi

Soru 8: Veri tiplerinin işlenmesi

Soru 9: Veri bağımlılıkları ve hiyerarşileri

Soru 10: Esneklik ve gürültü duyarlılığı

3.12.3 Veri Madenciliği Nasıl Yapılıyor ?

Veri madenciliği çözümünü gerçekleştirirken, ilk önce iş problemini tanımlamamız ve veri madenciliğini kullanacağımız alanları tespit etmemiz gerekmektedir. Daha sonra veri ambarından bu probleme yönelik bilgiler alınır. Bir sonraki aşamayı ise veri temizlenmesi ve kalitenin artırılmasıdır. Yani verinin temizlenmesi, entegrasyonu, tekrarların ayıklanması, eksik bilgilerin tamamlanması ve benzeri süreçlerdir. Sonraki aşamada bu temizlenen bilgi, veri madenciliği araçlarıyla işlenmesi ve kampanya nalizi, müşteri profillemesi, müşteri karlılık analizi gibi değişik iş çözümlerinde kullanılmaktadır.

Örneğin; müşteri sizi bırakmadan önce onun davranışlarından sizi bırakmaya yakın olup olmadığını tahmin ederek, buna yönelik kampanyalar veya kişiye özel teklifler

geliştirebiliyorsunuz. Bir diğer konu olan çapraz ve arttırımlı satış sayesinde, oluşturulan müşteri profillerine göre kişiye özel ürün tercihlerini tespit etme şansı oluşuyor. Böylece pazara yeni çıkaracağınız ürünün müşteri portföyünüzdeki alıcısını belirleyebiliyorsunuz.

3.13 Karar Destek ve Veri Ambarı İlişkisi

Karar-yapıcılığın değeri, analizlerden ve alternatif kararların değerlendirilmesi için, gerekli simülasyon adımlarından ve karar etkisinin neler oluşturacağından gelmektedir (Tanler, 2000).

Veri ambarı ve iş aklı uygulamalarının en önemli eksiği, "karar uygulama" için sağladıkları destektir. **Karar uygulama**, kurumun içindeki tüm insanların, çoğunlukla dışındakilerinde katıldığı , işbirliği için de olduğu bir prosestir. Kararlar üzerinde anlaşılmalı ve kesinlikle aksiyon üstlenen herkese iletilmelidir. Karar uygulamasına destek verebilmek için, bilgi veri ambarına kaynaklar olarak hizmet veren operasyonel ve finansal sistemlere geri döndürülmelidir.

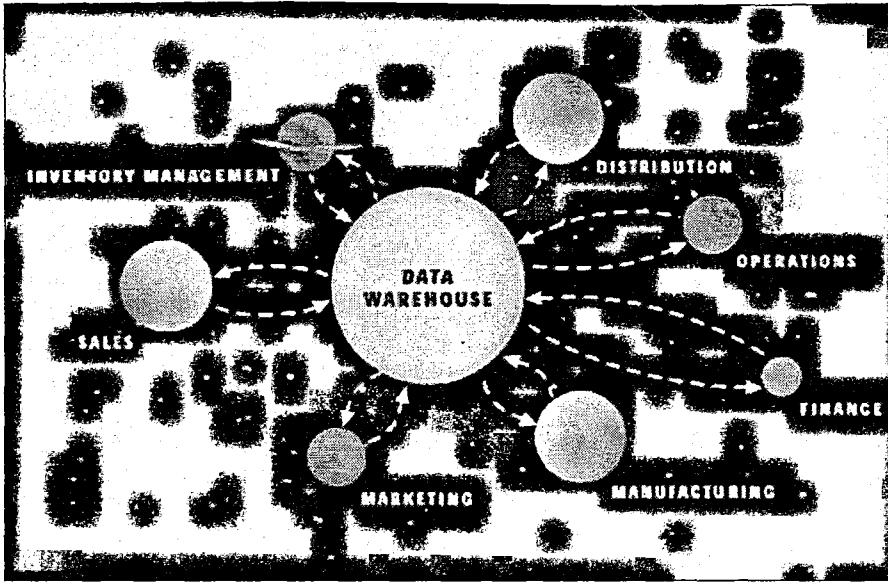
Veri ambarı tanımı birde bu açıdan yapıldığında:

"Veri ambarı karar yapmanın tek ve entegre kaynağıdır."

Veri ambarları için bir sonraki en büyük şey, karar uygulaması için zorunlu olan işbirlikçi (collaborative) planlama fonksiyonlarına verilecek destektir. Tahminlerde görülecek ciddi değişiklikler, her iş yönetim disiplinini etkileyecektir. **Tahmin** departmanlar arası işbirliğinin dilidir.

3.13.1 Planlama : Tamin & Bütçe

İş planlaması iç-içe geçmiş iki aktiviteye sahiptir. Tahmin gelecekteki taleplerin projeksiyonu iken, bütçesi ise gider satırlarındaki her bir parçanın konsolidasyonudur. Şekil 3.9'da görüldüğü üzere; tahmin bütçeleme için gerekli gelirin ihtimalini belirlemek için kullanılır, bununla beraber, tahminler ayrıca satınalma, üretim, envanter yönetim, lojistik vb. operasyonel sistemler için de kritiktir.



Şekil 3.9 Veri ambarı uzayı (Tanler, 2000)

Bütçeler kurumun finansal planıdır. Finansal plan kabul edildiği zaman, artık kurumun her aşamada performans ölçümünün görülebileceği birebir bağlı kalınması gereken bir yapı ortaya çıkmış demektir. Gelirin, birim satışlarının, fiyat ve iskontoların bir fonksiyonu olduğu düşünülürse bütçeleme projesinin ana girişi tahmin talebidir. Tahminler geniş bir iş operasyonları çerçevesine destek vermek için de gereklidir. Tahmin prosesi, geçmiş tarihli eğilimler, pazar faktörleri gelecekteki satışları projeksiyonlarının yer aldığı subjektif kararları da içerir.

Bütçeleme ve tahmin arasındaki bir diğer fark da, güncelleme sıklığıdır. İdeal olarak bütçeler yıllık yaratılır ve yıl boyunca değişmeden kalır. Pratikte, işteki çok önemli değişimler yaşanır bunlar bütçeye yansıtılır. Türkiye, Arjantin gibi ülkelerde yaşanan ekonomik krizlerin devlet ve özel sektör bütçelerine etkisi gibi tahminler yıl boyunca düzenli olarak yaratılır. Tahmin doğruluğunu daha ileriye götürmek, tahmin sıklığını artırmakla olur.

Bir karar destek aracı olarak bütçe, giderlerin kontrolü için bir çerçeve oturtur ve gelir hedeflerinin gerçekleşmemesi durumunda gider indirimleri yapılacak alanları belirler. Tahminin değeri, organizasyonun her gün aldığı operasyonel kararlara sağladığı destektedir. Tahmindeki en kritik gereksinim, önemli gelir değişimlerinin ön uyarı sistemi ile desteklenmesidir. Tahmin hatası, gerçek performansın, bir önceki tahminle karşılaştırılması ile sağlanır.

Niçin bu ayırım yapılmaktadır? Bugünkü iş dünyasının doğası, şirketlerin dinamik olarak operasyonel etkilerini iletmemeleri gerekliliğini doğruluyor. Ne sadece veri ambarcılığı, ne de

Internet, kurumlar arası (inter-enterprise) tahmin geliştirme konsensüsü üzerindeki işbirliği anahtar rolü oynayabilir. Aslında, her ikisi de buna katkı sağlamaktadır.

3.13.2 Veri Ambarının Değişen Rolü

Talep tahmin prosesi için en iyi veri kaynağı, veri ambarıdır. Veri ambarındaki veri, geçmişteki eğilimleri doğru yansıtabilmek için oldukça ciddi temizleme işlemine tabi tutulmuştur. Ayrıca bu veri sebep-ve-sonuç analizine güçlü bir taban teşkil edebilecek şekilde çok fazla sayıda kaynaktan alınıp bir araya getirilmiştir. Sofistike istatistik tahmin modelinin zayıf yanı, veri güvenliği zafiyetidir.

Bir çok sebepten ötürü, bütçeleme prosesi veri ambarının dışında bırakılır. Bununla beraber veri ambarı ve iş akıllı satış tahminlerinde rol oynamalıdır.

Bill Inmon'un orijinal veri ambarı tanımının bir kısmı şöyleydi: (Inmon, 1992-93)

". . . Yönetimlerin karar yapma sürecinde gerekli olan konu tabanlı, entegre, zaman-değişimli, uçucu olmayan (nonvolatile) bir veri toplama işlemi."

Bill Inmon uçucu olmayan (nonvolatile) kelimesini özenle seçip kullanmıştır. Bu ve birçok sebepten ötürü bu tanımlama oldukça rağbet görmekte ve bir çok kaynakta referans gösterilmektedir.

3.13.3 İş Aklının Değişen Rolü

İş aklının amacı, iş yöneticilerinin ardı-arkası gelmeyen soru yumaklarına cevap vermektir. Çünkü bir soruya verilen cevap hemen yeni bir soru üretmektedir. Her yeni soru, yeni ve daha kapsamlı analizlere ihtiyaç duymaktadır. Bir çok iş akıllı araçları sorgu jenerasyonu ve "zayıf (light)" analiz kapasiteli raporlara yoğunlaşmaktadır.

Cevaplanacak en zor sorular: "Niçin?", "Eğer..?" Bu soruları cevaplamak gereklidir. Fakat oldukça sofistike veri modelleme ve projeksiyon araçları gerektirmektedir.

İleri iş akıllı teknolojisi, müşteri davranışı ve pazar segmentasyonu gibi konular için gereklidir. Talep tahminleri iş planlaması için zorunludur. "Eğer..? (what if ...?)" sorularını adresler.

Tahminleme istatistiksel modelleme ve projeksiyon teknolojisi kadar istatistik tahminlerinin üzerine çıkabilme yeteneği gerektirir. Basit eğilimlerin projeksiyonlu istatistiksel tekniklerden ileri seviye rastgele modellere kadar bir çok teknik mevcuttur. Uygun istatistiksel model seçimi dikkatli bir analiz çalışması ister. İstatistiksel tahmin faydalıdır, fakat yönetimin gözden geçirme süreci ile pekiştirilmelidir. Çünkü bir çok istatistik metodu, gelecekteki neler olabileceğini birincil endikatör varsayarak geçmiştekilere bakarak karar verir. İnsan zekası (yönetim) gerektirir.

Tahmin, iş süreçlerinde kurum içinden ve dışından herkesin içine dahil edilmesi gereken bir prosestir. İş aklı bu proses adımlarını destekler. Bu adımlar:

1. Analiz
2. Modelleme
3. Gözden geçirme
4. Tahmin versiyonlarının uygulanması

3.13.4 Sınırları Zorlama ve İşbirliği Dili

İnternet teknolojileri, uygulamaları işten-müşteriye (B2C-business to customer), (B2B-business to business) uygulamalarına kadar oldukça hızlı gelişmektedir.

İş aklı web üzerinde dağıtımın kolaylaştırılması için optimize edilmelidir, ayrıca güvenlik ve ölçülebilirlik de oldukça önemli kriterler arasında yer almaktadır. Kesinlikle veri ambarcılığı ve iş aklının sınırlarının zorlanması şirket operasyonlarının nasıl işleyeceğini belirleyen kurallarının değişimine ve hızlı yeni dünyaya ayak uydurabilmesine potansiyel teşkil etmektedir.

Bir sonraki en önemli şey, veri ambarcılığı ve iş aklının iş yönetim prosesleri ile entegrasyonudur. İnternet teknolojileri; yeni bir global, departmanlar arası ve artan kurumlar arası işbirliğinin dili tahmin talebidir.

3.14 Veri Ambarı Tipleri

Nerde, geçmişi tutulması ve entegrasyonu gerekli, yüksek miktarda veri varsa orada veri ambarcılığı, uygulama sahası bulunur (Inmon, 2000b). Bunun manası veri ambarcılığı bankalar, sigorta şirketleri, üreticiler, perakendeciler, kamu kurum ve kuruluşları ve telekom

şirketleri içindir. Fakat bir çok durum ve örnekte veri ambarı sonuçta kendi karakteristiğine sahiptir.

İnsan kaynakları veri ambarı belkide en tekil (nevi şahsına münhasır) veri ambarıdır. Bir çok veri ambarı, yüksek miktardaki verilerin yönetimi için vardır. Perakendeciler, telekomunikasyoncular, bankacılar ve finans sektörünün hepsi yüksek miktarda veri içerir ve çok büyük rakamlı işlemler ihtiva eder. Fakat insan kaynakları veri ambarı bu ambarlar kadar yüksek miktarda veriyi içermez, çünkü işlem sayısı çok azdır. Ancak bir insan kaynakları veri ambarcılığında, çalışanlar merkezli uzun ve devam eden kayıtlar tutulmaktadır.

Sigortacılık ortamı da, bir iş ünitesinin tamamlanmasının oldukça uzun bir zamanı gerektirmesi açısından tekil bir örnektir. Bir poliçenin ödenmesinin bitmesi onlarca yıl sürebilmektedir. Bu ortam tıpkı ağır çekimli bir film gibidir. Bir sonraki poliçenin tamamlanması 20 yıl belki de ömür boyu olacaktır. Burada süre on-yıllar ile ifade edilmektedir. Buradan gerçekten önemli bir zaman aralığı operasyonel sistem ortamı nitelikleri ve veri ambarı ortamında görülmektedir. Bu aralık ise başka hiç bir şeyde görülmez.

Perakendecilik ambarı ve telekomünikasyon ambarları ise aksine verinin oldukça sık tekrarlanmasıyla yönetimidir. Bu verinin yönetilmesi için en azından iki ortak yol bulunmaktadır.

1. Birincisi, veri ambarına yükleme yapıldığı anda yaratılan bütünlük (aggregate) / profil kayıtlarının oluşturulmasıdır. Detay verinin veri ambarına yüklenmesi yerine, veri özetlenir ve bütünleşik hale getirilerek ambarı konur ise ambarda kaplayacağı hacim düşecektir. Bunun dezavantajlarından biri, bu bütünleştirme ve profillemeye aşamasında ciddi bir detay veri kaybının oluşmasıdır. Yerden ve gereksiz işlemlerden tasarruf etmek amacıyla perakende / telekom veri ambarları genellikle “sarmalanmış özet veri (rolling summary data)” yaratır. Sarmalanmış özet veri yapısında veri bir veritabanından diğerine zaman-bağımlı şekilde birleştirilir. Saatlik veriler gün sonunda günlük veri tabanına yüklenir. Günlük veriler hafta sonunda haftalık veri tabanına yüklenir. Haftalık veri ay sonunda aylık veri tabanına yüklenir ve böyle devam eder.

2. Bir diğer popüler alternatif saklama şekli ise, çok büyük miktardaki verinin işlendiği perakende / telekom veri tabanlarında görülen yüklü verinin alternatif yığınlarda saklanmasıdır. Detay verinin alternatif yerlerde saklanması, organizasyona en düşük maliyetle en detaylı verinin saklanması yeteneğini sağlar. Sadece aktif verinin işlem amacıyla disklerde saklanması sonucu, veri ambarı maliyetleri düşer, performans ise ciddi oranda artar.

E-ticaret veri ambarları oldukça yüksek miktarda veri içerir. E-ticaret ve perakende / telekom veri ambarları arasındaki fark, e –ticarete bütünleştirme ve profillemelerin oldukça istenilen bir durum olmasıdır. Klikleme verisinin arkasında bir dünya vardır, yani, kurumun tabloları ile interaktif verilerin oluşturduğu dünyadır.

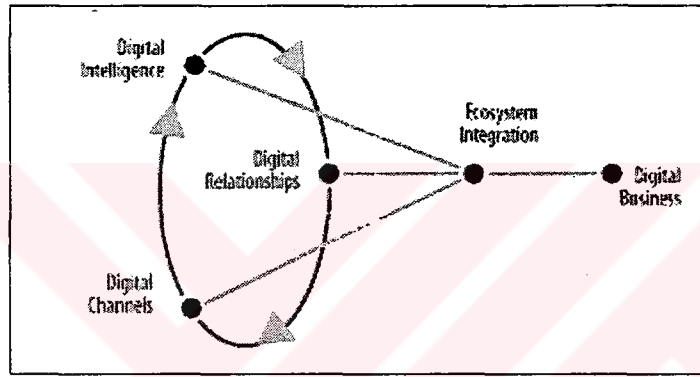
Bazı mühendislik ortamlarında, örneğin analog ortamdan veri alma ve bunu dijital ortama aktarma esnasında, tonlarca veri üretilir. Bu ortamda, yürütmenin esası, veri azaltmadan geçmektedir ve verilerin tümü veri ambarlarına kadar olan operasyonel sistemlerde vardır. Bu kadar veri en basit şekli ile bile uyumlu değildir ve bunlar veri ambarında tutulmamalıdır. Burada saklanması gereken veriler aykırılıklardır. Tolerans sınırlar içindeki değerler veri ambarında tutulmaz ve bunlar ebediyyen kaybolur.

“Star Join” diye tabir edilen veri ambarı temel mantığı bile tasarım açısından kendine has bazı özelliklere sahiptir. Tipik bir “star join”, “faktör (fact)” ve “boyut (dimension)” tablolarını içerir. Boyut tabloları, faktör tablolarına göre daha az veri içerir. Faktör tabloları daha çok veri sıklıkları içerir. Enterasan olanı, faktör tabloları genellikle numerik veri içerir. Boyut tabloları ise alfabetik veri içerir. Faktör tablolarında bulunan alfabetik tek veri anahtar (key) olarak kullanılan verilerdir. Bu anahtarlar alfabetik olabilir, diğer durumlarda faktör tabloları nereyse tamamen salt nümerik veri içerir.

3.15 Dijital İş ve Dijital Zeka

Hemen hemen tüm işletmelerde, tipik olarak iki ana komponent bulunur (Montgomery, 2000). Birinci komponent, bir ürünün veya satışı ile ilgili tüm aktiviteleri içeren **satış zinciri**; ikincisi ise, servis veya ürünün teminin gerçekleşmesi ve alıcıya sağlanmasının garanti edilmesi süreçlerini içeren tüm aktivitelerin içerildiği **tedarik zinciri** komponentidir.

Bir çok firma, satış zincirine destek veren elektronik ticaret sistemleri kurmak için zaman ve para harcarlar. Ayrıca, ürün ve servislerin bir zaman mentalitesi kullanılarak zamanında iletilmesini sağlamak için tedarik zincirleri kuruldu. Elektronik iş (e-business) elektronik ticaret sistemlerinin tedarik zinciri ile entegrasyonudur. Bu entegrasyonun etrafında rakipler, müşteriler, düzenleyiciler, tedarikçiler, ortaklar vb. bir çok bağımsız nesne (entity) mevcuttur. Bu kombinasyon ekosistem olarak adlandırılır. Ağlandırılmış ekonomide bu bir şirketin başarı ve başarısızlığının çizildiği sınırları belirler. Bir işletmenin dijital yetenekleri (dijital zeka, dijital kanallar, dijital ilişkiler eko sistemi içerir) bu eko sistemi optimize edeceği bir yol sunar. Sadece bu noktada bir şirket için gerçek manada bir dijital işten söz etmek mümkündür.



Şekil 3.10 Dijital iş komponentleri (Montgomery, DM Review, 2000)

Gerçek dijital iş, Şekil 3.10'da görüldüğü üzere dört ana komponentin (yeteneğinden) oluşur:

1. **Dijital Kanallar:** Dijital kanallar, dijital bilgilendirme, interaktivite, işlem ilişkilendirme ve entegrasyon için bir çok vasıtaya sahiptir. Müşteriler, satıcılar, ortaklar ve çalışanlara bir çok bağlantılar mevcuttur. İşten-işe, işten-tüketiciye, tüketiciden-tüketiciye kanalları içerir.
2. **Dijital Akıl:** Veri ambarcılığı iş akıllı, gelişmiş veri madenciliği, sinir ağları ve portal yeteneklerinin bir sonucudur. Eğilim detayları, tutum ve davranışlar, olaylar ve ilişkiler iş açısından hazır olarak erişilebilir.
3. **Dijital İlişkiler:** Bu müşteri ilişki yönetiminin dijital iş ile kombinasyonudur.
4. **Ekosistem Entegrasyonu:** Tüm yeteneklerin birleştirilmesi, dijital işte sonuncu ana komponenttir. Ayrıca, eski sistemleri ve dış sistemleri birbirine bağlamaya olan ihtiyaç, bir şirketin dijital iş gibi davranabilmesi için kritik bir gereksinimdir.

Gerçek bir iş oluşturmak için tüm yeteneklerin entegrasyonu gereklidir. Dijital zeka sistemi inşa edilirken, bu dört değişik ana tip gerçekleştirme yaklaşımlarına dahil edilmelidir. Bunları ve ayrıca şirketin hangi pozisyonda olmak istediğini anlamak önemlidir.

Tip 1 – Veri Ambarı Uygulayıcıları: Veri ambarı uygulayıcılarındaki fikir, bilginin erişilebilir lokasyona bir şekilde konumlanmasıdır, iş kullanıcıları bir şekilde kullanılabilir hale getireceklerdir. Bu yaklaşım tipik olarak dijital bir işi destekleyecek çerçeveyi sunmaz.

Tip 2 – İş Akıllı (IB) Uygulayıcıları: İş akıllı uygulayıcıları kullanıcı görüşmeleri yapar ve onların ihtiyaçlarını baz alan stratejileri geliştirir. Şirketin etrafındaki ekosisteme özel limitli dış kaynaklı bilginin seviyesi yükseltilir.

Tip 3 – BI Endüstri Ekspertleri: BI endüstri uzmanları iki-parçalı çözüm sağlarlar. Birincisi endüstrinin durumuna göre neye yoğunlaştırılması gerektiği ile kullanıcı problemlerinin giderilmesi için çözümün belirlenmesini içerir, ikincisi ise potansiyel gelişen pazarlar konusunu içerir. Bu çeşit bir uygulama için şirketin endüstrideki pozisyonunun lider konumda tutulabilmesi için ekspertler için dikey (pazar) uygulamalarında uzmanlık gereksinimi vardır.

Tip 4 – Gelişen Pazarlar Ekspertleri: Bu tip bir çözüm, endüstri eğilimlerini anlama ve gelişmiş veri madenciliği temelli bir geliştirmeye olanak verir. Gelişmiş veri madenciliği; sinir ağları ve tahminci analizlerin kullanımı demektir. Algoritma-tabanlı veri madenciliği normalde kişilere açıkça görülmeyen gelecekteki tendleri hissetme ve ortaya koyma yeteneği kazandırır.

Organizasyonun analiz edilmesi, dijital zeka sistemlerini gelecekteki başarısını sağlayacak gerçek bir dijital işi destekleyecek, kritik faktörleri, belirli bir etkinlik seviyesini göz önüne alarak oluşturulup oluşturulmadığını ortaya çıkarır.

3.16 Data Mart'lerden Veri Ambarı Doğar mı ?

Data martlar işe özgü yapıları nedeniyle kolay ölçeklenemezler. Fakat iyi modelleme yapıldığı takdirde kolaylıkla veri ambarlarına dönüştürülebilirler. Yapılan bir araştırmada (Gartner Group, 1997) şu beş şeye dikkat edilmesi gerekliliği ortaya konmuştur.

1. Data mart ekibiniz ile veri ambarı kardolarınızı ayrı çalıştırın.
2. Veri ambarı kurulumlarında data mart uygulama deneyiminizi kullanın.
3. Taktik çözümlerde en acil iş konularına karşı duyarlı olun.
4. Veri kaynaklarının sayısını 3 ile sınırlayınız.
5. Data mart'larınızın hızla çoğalmasını engelleyin.

Tam oluşturulmamış bir veri ambarı yerine bir data mart ne zaman kurulmalıdır? Aslında asıl sorun kaynaktır. En önemli kaynak, geliştirmeyi kurulumu yönlendiren para kaynağıdır (Michaels, 1997).

3.16.1 Supermartlar

Herhangi bir data martın, faktör tabloları setinden oluştuğu kabul edilir. Faktör tablolarının her bir çoklu alanlardan oluşan boyutsal anahtar (key) komponentlerini (foreign-keyler ile join edilmiş yapılar) içerir. Faktör tablolarının hepsi sıfır veya daha fazla faktör tablosunun boyutsal anahtar her bir boyutsal tablonun ana (primary) anahtarlarıdır. Bu yapıya yıldız birleştirme (**star join** veya **snowflake structure**) denmektedir.

Bill Inmon'a göre (Inmon, 2000a, DM Review May), Faktör tablolarının herhangi bir sayıdaki ilişkisel boyutlarla kuplesinden oluşan bir yıldız birleştirme, kullanıcı ihtiyaçlarını en iyi şekilde karşılayan verinin optimal denormalizasyonunu gösterir.”

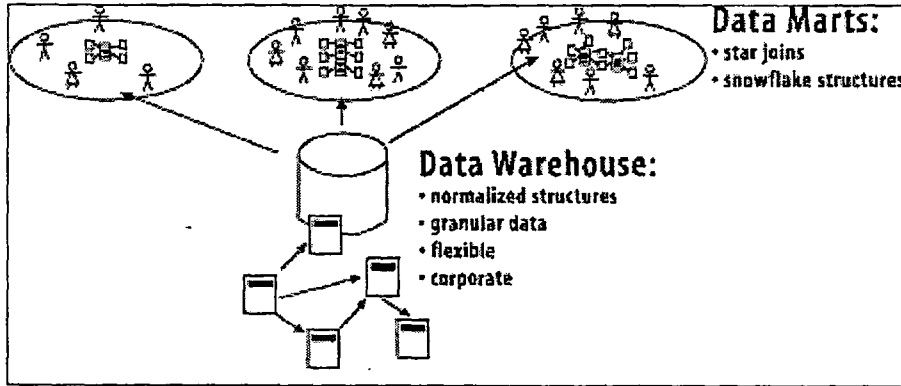
Bir yıldız birleştirme (star join), bir şekil ve içerikler kümesi ile bir grup kullanıcı için optimal iken diğer bir yıldız birleştirme diğer bir şekil ve içerikler kümesi ile diğer bir grup için optimaldir. Bunun için bazı sebepler şunlardır:

1. Verinin dizilimi (sequencing)
2. Veri tanımları
3. Parçalı yapı ve büyüyebilirlik (granularity)
4. Coğrafya
5. Ürünler
6. Veri ilişkiler
7. Zaman
8. Teknolojiler

9. Veri Kaynakları

10. Yönetim

Aynı işe değişik yollarla bakmak, değişik departmanlar için çok değişik yıldız-birleştirme (star-join) yapıları ile sonuçlanır. Bu durum Şekil 3.11’de görülebilir.



Şekil 3.11 Değişik noktalarda değişik veri yapılarının (star join) bulunması
(Inmon, DM Review, May 2000)

Bir uyumlu (conformed) boyut, kendisine bağlanabilen her bir faktör tablosunun aynı şeyi gösterdiği bir boyutsal yapıdır. Genel olarak, bu bir uyumlu boyutun her bir data martta kullanılabileceği anlamına gelir. Uyumlu boyutsal yapı oluşturmak çok önemli bir adımdır. En iyi uyumlu boyutlar parçalı yapının (granular) en yüksek seviyede olduğu yerde tanımlanabilir. Uyumlu boyutları en az bir veri ambarı anonim anahtarına sahip olmalıdır, ODS'lerdeki (Operational Data Store - Operasyonel Veri Kaynağı) anahtar alanlar gibi değil.

Supermart tüm veri ambarının komple bir alt kümesidir. Her bir supermart, uyumlu boyutların paylaşıldığı benzer tablolar ailesidir. Sonuç olarak tüm veri ambarı şunun içindir: Standardize edilmiş faktör tabloları ve uyumlu boyutsal tablolar tabanlı güçlü bir mimari ile bir araya getirilmiş farklı uygulanmış supermartların bir koleksiyonudur veri ambarı (Kimball, 1998a).

Eğer ihtiyaçlar belirlenmeden oluşturulan bir alt yapı (infrastructure) ile boyutsal modellemeye üretilen çözümler uymaz.

3.17 Operasyonel Veri Ambarı (OVA)

Eski, bağımsız OVA (ODS - Operasyonel Veri Kaynağı) direkt olarak kurumsal veri ambarına getirilmelidir. Kurumsal veri ambarı operasyonel sorgulama ve raporlama ihtiyaçlarını özetleyecektir. Bu yapıya "operasyonel veri ambarı" denmektedir (Kimball, 1998b). Her operasyonel veri ambarı verinin farklı modellenmiş iki versiyonuna ihtiyaç duyar:

1. İşlem (transaction) versiyonu ve
2. Periyodik anlık (snapshot) versiyonu

İşlem ve anlık versiyonları, Uzakdoğu Asya tabiriyle, bir operasyonel veri ambarının "yin ve yang" ıdır. İşlem versiyonu bize tüm ihtimalleri ile detaylı bir durum görünümü sunmakta, anlık versiyonu ise bize belirli bir andaki kurumsal durumu çok hızlı görmemizi sağlamaktadır. Her ikisi de gereklidir çünkü tüm perspektifleri düz ve zıt şekillerde ortaya koyacak kombine basit yapı mevcut değildir. Beraber kullanıldığında, işlem ve anlık versiyonları, tüm, anlık bir iş durumu görünümünü destekler ki bu da tümleşik bir veri ambarının bir parçasıdır.

3.17.1 İşlem (Transaction) Versiyonu

Operasyonel sistemin en basit görünümü tekil işlem seviyesindedir. Çoğunlukla ve de genellikle, anahtar-tipi bilgi işlemsel faktör kayıtlarına eklenir. Operasyonel veri ambarına işlem versiyonlu bir yapı inşa ettiğinizde, özetlenmiş veri ambarı üzerinde yapamadığınız bir çok güçlü analizleri yapma imkanına kavuşmuş olursunuz.

3.17.2 Anlık (Snapshot) Versiyonu

Değişik tipteki işlemlerin kar ve gelir üzerindeki etkisini yorumlamak için gerekli mantıksal yapı oldukça komplikedir. Bu ikileme cevap ikinci, ayrı bir faktör tablosu yaratmaktır ki, işlem tablosunun bir eşleneği olan bu tabloya "anlık (snapshot) tablosu" denmektedir. Bu tabloya kayıtlar genelde belirli periyotlarda alınan raporlar sonrasında eklenir ki bu da aylıktır genelde.

3.18 Bir Veri Ambarı Projesine Girerken Nelere Dikkat Edilmelidir ?

Standart soruların cevapları ile büyük başarılar elde edilse idi, hiçbir firma bu tür konulara bu derece kapsamlı yatırımlar yapmazdı. Eğer kurum olarak bir veri ambarı projesine başlanıyor ise çalışılacak firma veya firmalara aşağıdaki sorular sorulmalıdır (Kıral, 1997).

1. Veri ambarı oluşturmak için gereken metodoloji ve çatı (framework) yapıları nelerdir?
2. Veri ambarı projesine temelde hangi noktadan başlıyorlar?
3. Veri ambarı üzerinde tutulması gereken verinin detay ve tarihsel derinlikleri üzerinde sınırları var mı?
4. Sorgulamalarda vardıkları karmaşıklık düzeyi nedir? Sorguların oluşturulmasında bir sınırları var mıdır?
5. Bu konudaki yerel ve global referansları nelerdir?
6. Data mart ve kurumsal veri ambarı üzerindeki yaklaşımları nelerdir?
7. Proje tanımında çizdikleri sınırlar nelerdir?

Bu sorulara verilen cevaplar analiz edilerek işe başlanmalı ve sonuçta iş, bu sorulara en uygun çözümleri yerine getirecek firmaya verilmelidir.

3.19 Veri Ambarı ve Web

İnternet ve veri ambarcılığının her ikisi de IT sektöründe 1990'lı yılların sonunda en sıcak iki konuydu (Craig, 1997). Veri ambarcılığı internet ile aynı noktaya geliyor. Bu aynı noktaya gelme yeni bir karar destek araç kategorisi ile sonuçlanıyor: “web-imkanlı veri ambarı”. Bir veri ambarının web ortamına konmasının müthiş faydaları bulunmakla birlikte, ancak beraberinde bir çok son derece teknik ve organizasyonel problemler getirmektedir. Bunlar:

1. Güvenlik politikası
2. Projenin iş değeri (intranet, internet, extranet)
3. Etki değerlendirmesi
4. Kurulum ve yönetim

Kullanıcıların oldukça talepkar olduğu unutulmamalıdır. Bu yüzden verinin son kullanıcılara nasıl ulaştırılacağı ciddi manada değerlendirilmelidir. Kullanıcı bir internet tarayıcı ile

başlamalı, web sitesine konumlanmalı ve oradaki sorgulama veya raporu çalıştırmalıdır. Bu **pull technology** dediğimiz klasik örnektir. Bir çok zamanlar vardır ki kullanıcının bir isteği dahi olmadan bilginin kendisine otomatik ulaştırılmasıdır. Bunu yapmak için şirketler **push modeller** denen raporları tetikleyen yapıları kurmalıdır.

Veri ambarının web ortamına konulmasından dolayı sistem yönetimi, özellikle ağı (network) ve değişik sunucuları (load balacing - yük dağıtım ve dengesi) yönetmek için iyi yönetim araçlarına ihtiyaç vardır.

3.20 xOLAP

"OLAP" terimi artık kanıksanmıştır ve hiç kimse eski tabir olan "çok boyutluluk" (multidimensionality) terimini kullanmamaktadır (Parkes, 1998). Aslında "online" kelimesinin ne anlama geldiği konusunda da tam uzlaşma yoktur. Bu durum yıllar önce Web'in mantıksal tanımında da aynıydı. Bugün gelinen noktayı hepimiz bilmekteyiz. Sonra ROLAP (relational OLAP) çıktı. ROLAP metadata kullanarak ilişkisel tabloların çok boyutlu görünmesini sağlar. Peşinden diğer terimler kısaca "XOLAP" lar gelir. DOLAP veya "desktop OLAP", istemci tarafındaki veri dönüşümleri ve işlenmesini (RDBMS için front-end) gerçekleştirilir. MOLAP veya çok boyutlu (multidimensional) OLAP, ROLAP'a alternatif olarak çıktı. "Kural-tabanlı oluşturulmuş veri-küpleri üzerinde sorgulama imkanı verir. HOLAP veya "hybrid (birleşik) OLAP", ROLAP ve MOLAP'ın yeteneklerinin kombine edilmiş halidir. OLAP için bir diğer terim de bildiğimiz BI (iş akı)dır. Bilgi seviyesini daha yüksek seviyelere çıkartılarak daha iyi kararlar yapabilmek için kullanıcılara verilen bir perspektif nosyonudur. Bir diğer terim ise adı üstünde WOLAP veya "web OLAP"tır. Yani web üzerindeki OLAP'tır. Tüm OLAP terimleri özetle aşağıdaki tablodadır.

Çizelge 3.3 xOLAP terimleri listesi (Parkes, 1998, DBMS January 98)

OLAP	OnLine Analytical Processing
ROLAP	Relational OLAP
DOLAP	Desktop OLAP
MOLAP	Multidimensional OLAP
HOLAP	Hybrid OLAP
WOLAP	Web OLAP
BI	Business Intelligence

4. “İŞ KURALLARI SİSTEMİ” YARATMA

Bu bölümde; “İş Kuralları Sistemi” yaratma konusu veri-temelli olarak anlatılmıştır. İş mantığından başlayarak E/R modelleri ve “Kural-kalıp (rule-pattern)” modellerine kadar ayrıntılı bir şekilde incelenecektir.

4.1 İş Mantığı

“En zor veri entegrasyon problemi nedir?” sorusunun cevabı “İş Mantığı” olacaktır (McGoveran, 2000). Entegrasyon teknolojisi son 10 yılda çok büyük mesafe aldı. Başarılı uygulama entegrasyonu, kurumlar-arası (B2B) veya kurum-müşteri arası (B2C) e-ticaret uygulama entegrasyonundaki teknik problemler çözülmüştür. Sonuçta şu üç ana vazifenin yerine getirilmesine gerekmektedir :

1. Bağlantı
2. Zamanında veriyi yakalama
3. Veriyi anlama

4.2 E / R Modelleme İş Kurallarında Kullanılabilir mi ?

E/R modelleme *gerçekte* neyi garantiler? E/R (Entity / Relationship) modelleme; veri elemanları arasında “one-to-one”, “many-to-one”, “many-to-many” şeklinde tanımlanan ilişkileri bir diyagramda gösterme tekniğidir. Aslen bir E/R modeli sadece mantıksal bir modeldir. Güçlü E/R araçları, örneğin CA (Computer Associates) firmasının “ERwin Designer” adlı aracı, ilişkilerden, kısıtlama, anahtarlama kadar mantıksal modeli veri tanımlama diline (DDL-Data Definition Language) çevirerek fiziksel gerçekleştirmeyi sağlayabilmektedir. Ayrıca yazmış olduğun örnek uygulamanın tablo yapıları da yine bu E/R modelleme aracı (ERwin Designer) ile tasarlanmıştır. E/R modelleme başlangıç aşamasında yararlı bir teknik olsa da, aşağıdaki problemlerden ötürü iş kuralları sistemi yaratmak için uygun değildir:

1. E/R modelleme tam değildir.
2. E/R modelleme tekil gösterime sahip değildir
3. Bir çok gerçek veri E/R modellemenin desteklemediği “many-to-many” ilişkiye sahiptir.

4. Bir çok büyük E/R modeli idealdir, fakat gerçek değildir.
5. E/R modelleri nadiren gerçek veri modelidir.
6. E/R modelleri; SAP ERP uygulaması gibi 10,000 den fazla tablo içeren sistemlerde çok karışık diagramlar üretmektedir.

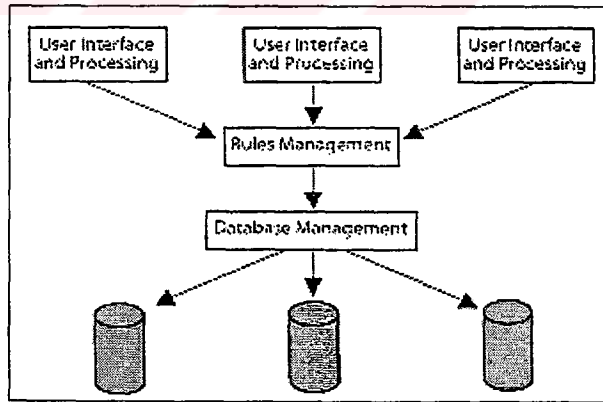
Sonuç olarak, normal veri işleme (transaction) sistemlerinde ve kısıtlı veri temizleme aktivitelerinde kullanılsa da veri ambarı iş kuralları uygulama ortamları için yetersiz kalmaktadır.

4.3 Bir “İş Kuralları Sistemi” Yaratma

Bugün, rekabet artık müşteriden bir klikleme kadar uzaktadır. Internet bazı pazarları yükseltse de diğerlerini de en azından sarsmaktadır (Von Halle, 2001a).

4.3.1 “İş Kuralları Sistemi” Nedir ?

En basit tanımıyla, “kurallar”ın (mantıksal hatta fiziksel olarak) ayrıştırıldığı, veri depoları, arayüzler ve uygulamalar arasında kurulabilen otomatik işleyen sisteme iş kuralları sistemi denir. Bu konseptin en basit gösterimi Şekil 4.1’de görülmektedir.

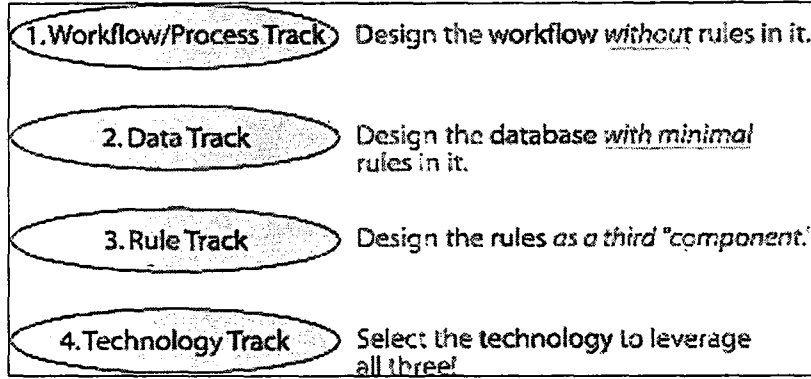


Şekil 4.1 “İş Kuralları Sistemi” için muhtemel “Yüksek-Seviyeli İçeriksel Mimari”
(Date, 2000, Wesley Longman, Inc.)

İş kuralları perspektifi, organizasyon düşünme veya karar-yapım kapasitesi üzerine yoğunlaşır. Kurallar; bilgi veya tercihlerin formal bir gösterimidir.

4.3.2 İş Kuralları Metodolojisinin Tekillliği

İş kuralları sistemi geliştirme metodolojisinin tekillliği, sistem geliştirme yaklaşımını dört ayrı fakat birbiriyle entegre katmanlardan oluşmasına dayanır. Şekil 4.2 incelendiğinde her bir katmanın görevleri açıkça görülmektedir.



Şekil 4.2 "İş Kuralları Sistemi" nasıl inşa edilir? (Date, 2000, Wesley Longman, Inc.)

4.3.3 Önemli Avantajları

Özetle avantajlar şunlardır :

1. Bir iş kuralları sistemi hızlı geliştirebilir.
2. Bir iş kuralları sisteminde bir kural çok kısa sürede ve de kolaylıkla değiştirilebilir
3. Bir iş kuralları sistemi giderek artan küçük parçacıklar şeklinde nakledilebilir.

4.4 "İş Kuralı" nedir?

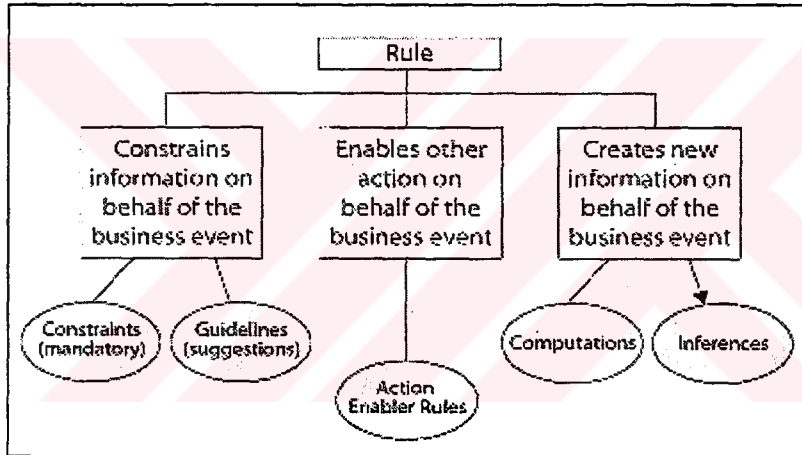
Bugün itibarıyla; "Kural" veya "İş kuralı" tabirleri için endüstriyel standartta kabul edilmiş bir tanım mevcut değildir (Business Rules Group - www.businessrulesgroup.org). İş perspektifine sadık kalabilmek için, iş kuralları iş kullanıcıları tarafından (programcılar değil) yazılmalı, anlaşılmalı ve implimentasyon ya da progama dilinden bağımsız kılınmalıdır.

4.5 Değişik Tipteki İş Kuralları

Üstte de belirtildiği gibi evrensel bir iş kuralları klasifikasyon tanımı veya şeması mevcut değildir. Buna rağmen; iş kuralları dünyasını üç ana kategoride tanımlayabiliriz :

1. **Terimler** (veri konsept ve detaylarını tanımlama)
2. **Olgular** (veri, kısıtlamalar ve kısıtlamalar arasındaki ilişkileri tanımlama)
3. **Kurallar:** Beş alt tipi vardır.
 - 3.1. Zorunlu kısıtlamalar
 - 3.2. Kısıtlamalar (iş olayları tarafından tetiklenir)
 - 3.3. Klavuzlar (iş olayları tarafından tetiklenir)
 - 3.4. Aksiyon tetikleyiciler (sınırlamalar dışındaki aksiyonlar)
 - 3.5. Hesaplama ve Çıkarımlar (mevcuttan yeni bilgi üretimi)

Terimler ve olaylar mantıksal model ve fiziksel veri tabanı için temel teşkil edecektir. Kuralların üçüncü sınıflandırması iş kuralları yaklaşımının püf noktasıdır. Şekil 4.3'de görüldüğü üstteki listeden de anlaşıldığı üzere kurallar beş ana alt tipe ayrılabilir.



Şekil 4.3 Kural sınıflandırması (Date, 2000, Wesley Longman, Inc.)

“**Zorunlu kısıtlama**” yapılan işlemin rededilmesini sağlayan bir kuraldır. “**Klavuz**” ise işlemi reddetmez ancak istenmeyen durumda sadece uyarıda bulunur. Çünkü mantıken de bir klavuz ancak yönlendirir, engelleme yapmaz. “**Hesaplama ve çıkarımlar**” mevcut bilgiden istifade ederek yeni bilgiler yaratır.

4.6 Bütün İş Kuralları Veri Hakkında mıdır?

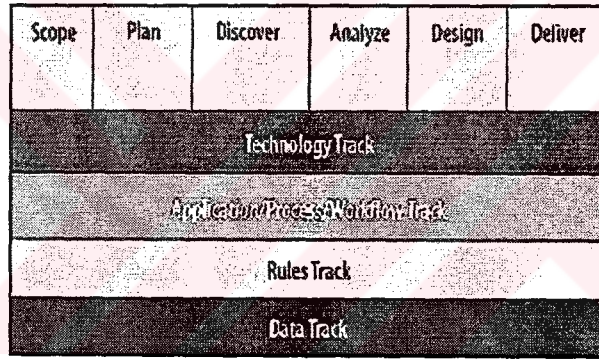
Bu bölümde anlatılmakta olan tüm iş kuralları, veri hakkındadır. Yani **terimler** veri konsept ve detaylarını tanımlarken, **olgular** veriler arası ilişkileri, **kısıtlama** ve **klavuzlar** veri değerlerini test ederler. **Hesaplamalar** yeni bir değer elde ederken, **çıkarımlar** bir sonuç verisine ulaşır, aksiyonu başlatacak veriyi önceden yakalarlar. Unutulmamalıdır ki ortak

kurallar ve veriler ortak bilgi ile sonuçlanır. Ortak bilgi ise daha iyi öğrenen bir organizasyon yaratır.

$$(Ortak) \text{ Kurallar} + (Ortak) \text{ Veri} = (Ortak) \text{ Bilgi} = \text{Akıllı Öğrenen Organizasyon}$$

4.7 “İş Kuralları Sistemi” Metodolojisinin Fazları

Bir iş kuralları yaklaşımı, tıpkı diğer system geliştirme metodolojileri gibi altı konsepti içeren değişik fazları içermektedir (Von Halle, 2001b). Bunlar Şekil 4.4’de görüldüğü gibi ölçek, plan, kesit, analiz, tasarım ve gerçekleştirme. Bu her adım ve gerçekleştirmeler görüldüğü gibi dört ana katmandan (veri, kurallar, Uygulama / Proses / İş akışı ve Teknoloji katmanlarından) oluşur.



Şekil 4.4 “İş Kuralları Sistemi” metodolojisinin fazları
(Von Halle, 2001, DM Review February 2001)

4.8 Kural Analizinin Başlangıcı

Kurallar bir iş bilgi sisteminin üçüncü boyutu olarak düşünülebilir (Von Halle, 2001b),. BT tarihinde geriye doğru bakarsak birinci boyut proses boyutudur. İkinci boyut ise veri boyutudur. Sistemi veri ve prosesin birleşimi olarak görmeye başladık. Eğer kuralları tam olarak algılayabilirsek kural boyutu diğerleri ile aynı değildir. Aslında iki boyutlu normal yapıya, kurallar ile bir derinlik boyutu (üçüncü boyut) getirilmektedir. Bir mantıksal model aşağıdaki *üç anahtar bilgiyi* içerir:

1. İlki, mantıksal kural kalitesi veya semantik bütünlük için terim ve olguların tanımlanmış ve analiz edilmiş kümesidir.
2. İkincisi, kural-bağımlılık ilişkilerinin gösterildiği tablo veya diyagram.
3. Üçüncüsü, veri veya bilgi referanslarının ilişkilerini gösteren tablo veya diyagram

4.9 Kural-Kalıplarına Giriş

Bir **kural kalıbı** (rule pattern), içindeki her bir atomik (tekil) kuralın kayıt edildiği ve yönetildiği bir kural reçetesi olarak düşünülebilir.

Örneğin Şekil 4.5 bir kural-kalıp tablosudur. En soldaki kolonlar (**IF**) durumsal gösterge kolonları, en sağdaki kolon (**THEN**) ise sonuç kolonudur. Uygun bir kural-kalıp tablosu her atomik kural için sadece bir sonuç içerir ve hiç bir durumsal kolonu *NULL* olamaz. Şekil 4.6 bu duruma uygundur.

Rule ID	IF Guardian.credit-rating-code	AND Guardian.special-deal-flag	THEN Guardian.good-credit-flag
Test rule 1	= A	= No	= Yes
Test rule 2	= B	= Yes	= Yes

Şekil 4.5 Örnek “Kural-Kalıp” tablosu (Von Halle,2001, DM Review April 2001)

Örnek tabloda birinci kolon Rule ID bilgisini, ikinci ve üçüncü kolonlar durumsal bilgiyi (IF), dördüncü kolon ise kuralın işletilmesi sonucu oluşan sonuç bilgisini saklar. Kural-kalıp analizi ardında yatan mantığın detaylarını inceleyelim.

4.10 Kural-Kalıp Analizi

Kural kalıplarının nasıl analiz edildiğini anlamak için Şekil 4.6'yı inceleyelim. Basit olması açısından sadece çıkarımlar (interference) gösterilmiştir.

Rule ID	Rule Classification	Rule
R2	Inference	If Guardian.credit-rating-code = A then Guardian.good-credit-flag = yes
R9	Inference	If guardian.credit-rating-code not = A and guardian.special-deal-flag = yes then guardian.good-credit-flag = yes
R10	Inference	If guardian.credit-rating-code = B then guardian.good-credit-flag = yes
R11	Inference	If guardian.credit-rating-code = A, then guardian.good-credit-flag = no
R12	Inference	If guardian.credit-rating-code = A and Business-Party.ID in Park-Ranger then Guardian.good-credit-flag = yes

Şekil 4.6 Bir kurallar koleksiyonu (Von Halle, 2001, DM Review April 2001)

Eğer Şekil 4.6'daki kuralların tümü tek bir kural-kalıp tablosunda birleştirilirse Şekil 4.7'deki gibi bir tablo ortaya çıkar. Şekil 4.7'de tüm kurallar aynı sonucu paylaşmakta fakat bazı

durumsal kolonlar NULL bilgi içeriyor. Çünkü bazı durumlar tüm kurallara uygulanmıyor ve ortaya uygun olmayan bir kural-kalıp tablosu çıkıyor. Şekil 4.7 indirgenerek Şekil 4.8, 4.9 ve 4.10 şeklinde üç ayrı uygun kural kalıbı elde edilebilir.

Rule ID	IF Guardian.credit-rating-code	AND Guardian.special-deal-flag	AND Business-Party.ID	THEN Guardian.good-credit-flag
R2	= A			= Yes
R9	Not = A	= Yes		= Yes
R10	= B			= Yes
R11	= A			= No
R12	= A		In Park Ranger	= Yes

Şekil 4.7 Uygun olmayan “Kural-Kalıp” tablosu (Von Halle, 2001, DM Review April 2001)

Kural analizi için, kural kalıpları tek yöntemdir. Eksik kısıtlama kontrolleri de eklenerek kural tamlığı zenginleştirilmek istenirse "Kural-zenginleştirme mantıksal modeli" uygulanmalıdır.

4.11 Kural Zenginleştirme Mantıksal Veri Modeli

Bir kural zenginleştirme mantıksal modeli bildiğimiz geleneksel veri modelinin doğal bir dönüşümdür. Veri analizi ve modelleme bir iş kuralları sistemi için geleneksel modellemekten iki farkla ayrılır.

Birincisi, taşınabilir alanı veri yapılarını temel alır, taşınamaz alanı ise kuralların çalıştırılması sonucu somutlaşan kalıcı bilgi halini alır. *İkincisi*; kuralların toplanması, bir araya getirilmesidir.

Rule ID	IF Guardian.credit-rating-code	THEN Guardian.good-credit-flag
R2	= A	= Yes
R10	= B	= Yes
R11	= A	= No

Şekil 4.8 Kural-kalıp 1 (Von Halle, 2001, DM Review April 2001)

Rule ID	IF Guardian.credit-rating-code	AND Guardian.special-deal-flag	THEN Guardian.good-credit-flag
R9	Not = A	= Yes	= Yes

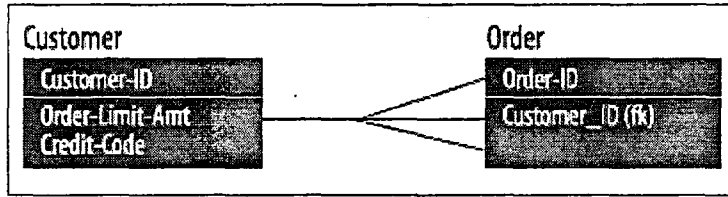
Şekil 4.9 Kural-kalıp 2 (Von Halle, 2001, DM Review April 2001)

Mantıksal modelleme metodolojisinin, iş kuralları yaklaşımı ile arasındaki fark; mantıksal modelin özellikleri veri bütünlüme kısıtlarını (data integrity constraints) da içine alarak kural analizini tamamlayabilmesidir.

Rule ID	IF	AND	THEN
R12	Guardian.credit-rating-code = A	Business-Party.ID In Park Ranger	Guardian.good-credit-flag = Yes

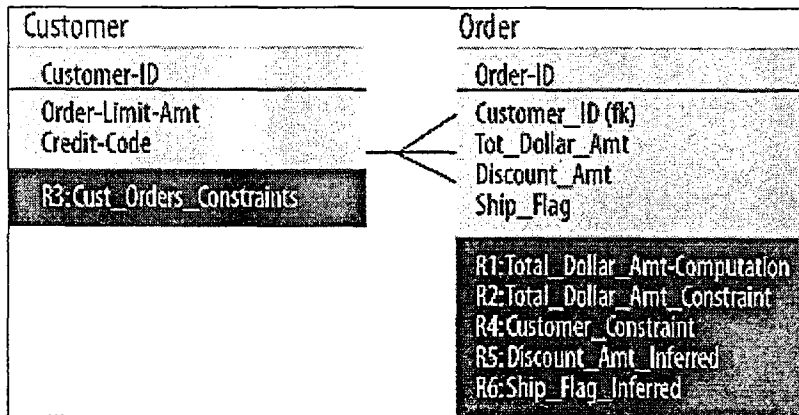
Şekil 4.10 Kural-kalıp 3 (Von Halle, 2001, DM Review April 2001)

Şekil 4.10'daki temel (en basit) tablo-ilişki modelini ele alalım. Şekil 4.11'deki gibi E/R modelleme ile Şekil 4.10'i görsel hale getirelim.



Şekil 4.11 Temel E/R modeli (Von Halle, 2001, DM Review April 2001)

Şimdi de Şekil 4.12'deki kural-zenginleştirme mantıksal veri modelini inceleyelim. Kural çalıştırılmadan altta koyu olarak yazılmış "R" li değerleri (R1, R2,... R6) bilemeyiz ve hesaplayamayız da. Bunlar temel bilginin değil, anlamlı bilginin parçalarıdır. Anlamlı bilgi "Tanımlar" bölümünde (Bölüm 2) de detaylı anlatıldığı gibi bilginin uygulama mantığı sonucu bize yansıyan bilgidir (information ---> knowledge). Sonuç olarak kural zenginleştirme modeli temel bilgi yönetiminden ziyade anlamlı bilgi yönetimine daha yakındır.



Şekil 4.12 Örnek "Kural-zenginleştirme" veri modeli
(Von Halle, 2001, DM Review April 2001)

Kuralların sadece anlam / mantık kalitesinden başka işletilme (çalıştırılma) sırasını da analiz etmek, bağımlılık hiyerarşisini ortaya çıkarmak gerekir. Bir algoritma gibi kuralların belirli sırada çalıştırılmaları gerekir.

Genel bir bakışla, kurallar temelde iki farklı yeteneğe sahiptirler (Von Halle, 2001c).

Birincisi; **veri-yönelimli kural yeteneği** (data-oriented rules). Genelde uygulama, kurallar arşivine bilgi geçirir. Kuralların uygulamaya dokunduğu noktalarda işletilmesi yeteneği bunun zıddıdır.

Karşılaştırma yapılırsa, **servis-yönelimli kural yeteneği**, proses-yönelimi ağırlıklıdır. Kurallarla iletişim veya veri paylaşımı için bir nesne modeline sahiptir. Kural işletiminde sadece veri vardır, kurallar yoktur. Kurallar, kural servisi tarafından. **“IF/ THEN”** yapısı kullanılabilir. Tasarım çalışmasında efor; kuralları, kümeler veya gruplar halinde hiyerarşik bir yapıya sokabilmek ve çalıştırılma sırası algoritmalarını oluşturup sisteme tanıtmada işlerinde harcanır.

4.12 Toplam Tasarım

Bir çok yönüyle bir iş kuralları sistemi tasarımı herhangi bir sistem tasarımıyla aynıdır. İş kuralları yaklaşımı tamamen bir veri yönlendirme prosesi ve kural yönlendirme entegrasyonudur (Von Halle, 2001d). Herhangi bir iş kuralları ürünüyle böyle bir prototip için uygulanması gereken adımlar Çizelge 4.1’de özetlenmiştir :

Çizelge 4.1 “İş Kuralları Sistemi” prototipinin oluşturulması için genel yaklaşım adımları
(Von Halle, 2001d, DM Review May 2001)

Adım 1:	Kural teknolojisi kullanan bir ürünü seçerek başla.
Adım 2:	Bir VTYS ürünü seç.
Adım 3:	VTYS’deki genel veri bütünlüğü kurallarını zorla.
Adım 4:	VTYS’deki toplam ve grupları zorla.
Adım 5:	Diğer kuralları (olgu, aksiyon, tetikleyiciler, kompleks kuralları) zorla.
Adım 6:	Ekran / sayfa tasarımlarını yapıp onları yarat.
Adım 7:	Kuralları değiştir.
Adım 8:	Elde edilen tecrübe değerlendirmesini yap.

5. CRM (CUSTOMER RELATIONSHIP MANAGEMENT - MÜŞTERİ İLİŞKİLERİ YÖNETİMİ)

CRM tanım ve içeriği açısından oldukça karıştırılan bir konu olduğundan, bu bölümde önce ne olmadığı tarif edilecek, daha sonrası ise ne olduğu örneklemelerle açıklanacaktır.

5.1 Müşteri İlişkileri Yönetimi Nedir?

CRM için "müşteri veri tabanı organizasyonu" ya da "müşteri veri ambarı" denilebilir. Bazılarına göre "kampanya yönetimi" bazılarına göre de "müşterilere ürünlerin sunulduğu ve satıldığı bir internet sitesi"dir. CRM ticari strateji olarak açıklanabilir (Bingle, 2001).

Öyleyse CRM nedir? CRM en basit tanımıyla, değişimdir. Müşteri merkezli olmak için organizasyonlar kendilerini değiştirmelidir. Bu yeni bir yazılım yüklemek kadar kolay değildir. Kontrolü ve yükü azaltıp dağıtarak şirketin diğer alanlarındakilerle bilgi paylaşımını sağlamaktır.

CRM yarışı ortadan kaldırmaktadır. Bu organizasyonu geliştirir. Organizasyon kültürünün gelişmesi bireyseliği azaltması, işbirlikçiliğin artmasını sağlar. Bunların hepsi müşterileri daha verimli yönetmek için organizasyonları değiştirmektir.

CRM vizyonla başlar. Bu vizyon üst yönetim tarafından inşa edilmelidir ve onların tüm desteği alınmalıdır. Vizyon, organizasyon ile haberleşme içindedir ve organizasyonun her alanında vardır. Tepe yönetici şirketlerde bu konuda birinci olmalıdır, diğer organizasyonel kadro onu takip etmelidir. Bu diğer önemli girişimler gibi bir stratejidir.

Stateji CRM'in ilk basamağıdır. Şu önemli nokta açığa çıkmaktadır ki; CRM girişimleri teknolojisiz olamaz. Bu işin yönlendirmesi ile yapılmalıdır. CRM girişiminin başarısı için bu çok önemlidir. Eğer teknolojinin yönlendirmesi ile yapılırsa bu oldukça pahalı olur ve verimli olmaz. İşin kendisi teknolojiden stratejik olarak çok öndedir.

Sadece teknoloji yönlendirmesi ile implemente edilmeye çalışılan CRM girişimleri milyonlarca dolarlık zararlar ile sonuçlanmıştır. Niçin? Sebep iş perspektifi yerine teknoloji yönlendirmesi esas alınmalıdır.

CRM muhakkak şirket kültürünü, organizasyonun genel davranışını değiştirecektir. Bildiğimiz gibi insan psikolojisi her zaman değişimi reddeder. Bu sebeple, uygulanacak CRM stratejisinde bu önemli nokta kesinlikle atlanmamalıdır. Değişimle nasıl başa çıkılacağı efektif bir şekilde kararlaştırılmalıdır. Bu değişimden hoşlanmayan veya zarar görecekt kişilerin tepkisiyle karşılaşılacak, hatta işini, mevcut gücünü veya kariyerlerini kaybedeceği endişesini taşıyan insanların “iş yokuşa sürme” diyebileceğimiz davranışları görülebilecektir.

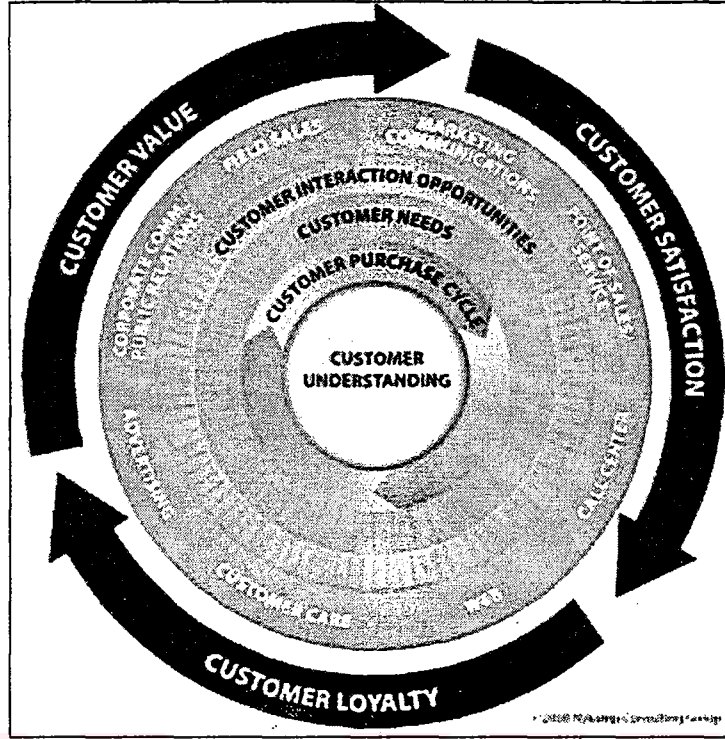
CRM' in değişim olduğu her zaman akılda tutulmalı ve buna göre hareket edilmelidir. İç haberleşmenin CRM için kritik bir başarı faktörü olduğu da unutulmamalıdır. Teknolojinin sadece gerekli olduğunda kullanılması gerektiği de başarı için bir diğer faktördür.

5.2 “Müşteri Deneyim Döngüsü”nü Anlamak

Müşteri memnuniyetini, bağlılığını ve değeri maksimize etme olgusu, müşterilerin mal ve hizmetleri nasıl bulduklarını yani deneyimlerini en ince detaylarıyla anlama ihtiyacını ortaya çıkarır ki bu da müşteri ile ilişkiyi çok iyi geliştirmenin gerektiğini gösterir. Bir müşteri deneyim döngüsü geliştirme işi aşağıdaki altı adımı içerir (Nykamp ve McEachern, 2000).

Adım 1: Müşteriyi Anlama: Bütün CRM girişimlerinin tüm alt yapı ve tüm prosesler müşterileri anlama üzerine bina edilip, buna göre takip edilmelidir. Detaylı müşteri profili, segmentasyon ve segment analizi bunu anlamaya yardımcı olacaktır.

Adım 2: Müşteri Satınalma Süreci : Her bir müşteri segmentinin satınalma süreci tanımlarını içerir. Tüm faktörler ürün kategorisine ve müşteri segment karakteristiklerine göre çok değişiklikler gösterir.



Şekil 5.1 Müşteri Deneyim Döngüsü (Nykamp ve McEachern, 2000, DM Review June 2000)

Adım 3: Müşteri İhtiyaçları: Üçüncü adım segmentin akıllığını arttırmak için satınalma sürecinin her adımında belirlenen müşteri ihtiyaçlarını içerir. Her segment tamamen değişik ihtiyaçlara ve değişik satınalma süreçlerine sahiptir.

Adım 4: Müşteri Etkileşim İmkanları: Dördüncü adım, müşteri etkileşimini tanımlayan iç ve dış imkanlarını, tüm kanallar ve medyalar arası etkileşimin her seviyesindeki karşılığını içerir.

Adım 5: Müşteri Deneyim Senaryoları : Bu adım yüksek öncelikli etkileşim imkanlarının her birini içermektedir.

Adım 6: İmplementasyon ve Ölçüm: Prosesteki bulguların gerçekleştirildiği ve ölçüldüğü aşamadır.

5.3 Müşteri Verisi Entegrasyonu: Efektif CRM'in Zorunlu Komponenti

Az bilgi tehlikeli bir şeydir. Farklı bir görüşe, satışa veya geleceğe mal olabilir. Müşterilerin tercihleri veya satın alma tarihçeleri hakkında eksik bilgi ve çerçeve sunarlar (McClain, 2000).

CRM, DW ve ERP için geliştirilen tüm uygulamalar; tam müşteri veri entegrasyonu (CDI - Customer Data Integration) isterler, bu oldukça zor fakat hayli değerlidir.

5.3.1 Gerçekten Müşteri Odaklı olmak

Herkesin söylediği, fakat içini tam olarak dolduramadığı kavram: “**müşteri odaklılık**”, “**müşteri merkezli olma**”. Müşteriler hakkında ne kadar çok şey bilinirse, onlara o kadar daha kolay istenen mal ve hizmetler sunulabilir.

Müşterinin organizasyon ile etkileşimlerinin bütünsel bir resmi olmaksızın, maksimum seviyede sonuçlar üretmek imkansızdır. Özellikle internet dramatik şekilde müşteri beklenti ve taleplerini artırmıştır.

Kafa-kafaya tabir ettiğimiz rekabet ortamında, müşteri-odaklı şirketler ve sonrasındaki bulguları kendi planlama, pazarlama ve satış eforlarında kullanırlar.

Kurumsal-odaklı yapı'dan, müşteri-odaklı model içeren bir yapıya geçiş operasyonu; şirketler için oldukça önemli bir yatırımı temsil etmektedir.

5.3.2 Gerçek-zamanlı CDI (Müşteri Verisi Entegrasyonu)

Şirketler, link atamalarını, tekrarlayan ve doğru olmayan girişlerin veri tabanındaki doğruluk seviyesini artırmak için yeni veri-eşleme teknolojisi kullanırlar. Örneğin “John Smith” ve “J. Smith” ve “J.B Smith” hepsi aynı kişi, üç farklı görünüm olduğu halde; böyle yapmakla, şirket aniden üretim ve postalama maliyetlerini düşürür ve ayrıca muhtemel başarı oranını ilerletmiş olur. Öncelikli işlem budur.

İsim ve adres bilgileri arasında kullanılan bağlantılar ile, şirket diğer veri tabanlarındaki bu isimler ve adreslerin tutarlılığını yakalayabilir. Şirketin dört, beş veya on tane veri tabanı da olsa hepsi artık sanki tek bir, güçlü kaynak olarak efektif şekilde işleyecektir.

Sonra, başarı şansını ilerletmek, şirket linklerini kullanarak çok hızlı bir şekilde demografik ve hayat standardı bilgilerini internet yardımıyla değerlendirir.

5.3.3 Hızlı Veri Değerlendirme

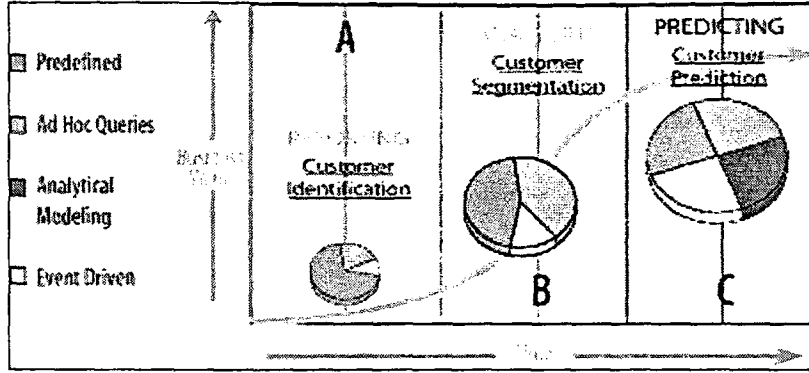
Online, gerçek- zamanlı veri değerlendirme kapasitesi şirketlerin çok daha iyi kararları daha hızlı oluşturmalarına imkan verir. Ayrıca değişen pazar şartlarında hızlılık ve esneklik sağlar.

5.3.4 Müşteri Verisi Entegrasyonun CRM için Gerçekleştirilmesi

Eğer gerçek bir müşteri ilişkileri yönetimi, ileri bakan şirketlerin hedefi ise, öyleyse gerçek-zamanlı müşteri verisi entegrasyonu açıkça yoğunlaşmaları gereken alanlardır. Bu, müşteriye sunulan hizmet ve servisleri geliştirecek, müşteri bağlılığını ve tercih yüzdesini artırıp, her bir müşteri için “yaşamsal değeri”ni kurum lehine maksimize ettirecektir.

5.4 CRM ve Veri Ambarcılığının Büyüme Aşamaları

Şirketler DW (Veri ambarı) ve DSS (Karar Destek Sistemleri) uygulamalarından hemen geleceği tahmin ve strateji geliştirme tahminlerine yardım etmelerini beklememelidiler. DW / DSS araçlarını müşteri ilişkileri yönetimi (CRM) için kullanma bir öğrenme prosesidir. Bir aşamada öğrenilen bilgi, sonraki daha gelişmiş analiz ve uygulamalar için temel teşkil edecektir. DSS / DW kullanımının CRM gelişiminin grafiksel gösterimi Şekil 5.2’de görülmektedir.

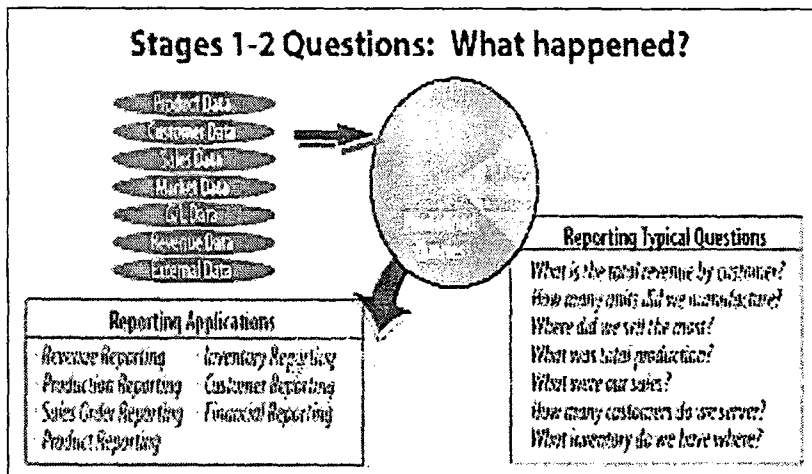


Şekil 5.2 DSS / DW kullanımının gelişimi CRM (McClain, 2000, DM Review June 2000)

5.4.1 DW/DSS Tip # 1: Raporlama

Birinci tip DSS, "önceden tanımlanmış sorgulamalar (*predefined queries*)" ın çokluğu ile karakterize edilebilir. Bu tipteki raporlama sistemleri genelde müşteriler veya pazar hakkında en sık sorulan sorulara cevap teşkil ederek sadece işin belirli bir alanına yönelik "küpler / şekiller / grafikler" biçiminde komple bir çizgisel bir set sunar. Şekil 5.2' nin A bölümü kullanım tipleri ve işsel değerinin dağılımını göstermektedir.

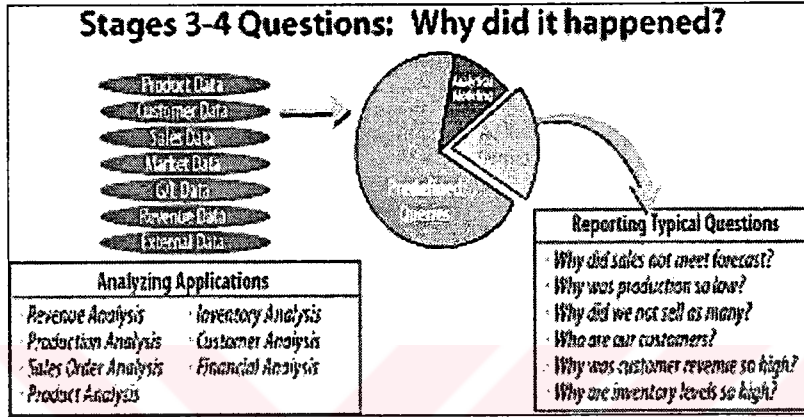
Şekil 5.3 ise DW / DSS rapor tiplerinin kullanımının dağılımını göstermektedir. Bunlar çoğunlukla raporlama, bazı ad hoc sorgulamaları ve nadiren bazı analiz uygulamalarını içerir.



Şekil 5.3 Raporlama uygulamaları ve sorular (McClain, 2000, DM Review June 2000)

5.4.2 DSS/DW Tip # 2: Analiz

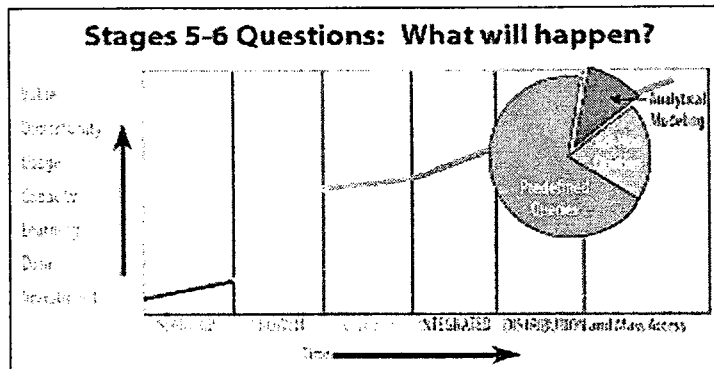
Şimdi de DSS' in ikinci tipinin (analiz) ad hoc sorgulamalarını, inceleyeceğiz. Bu aşama "niçin oldu ?" sorusuna yoğunlaşmaktadır (bkz Şekil 5.2, Bölüm B). Şekil 5.4'de görüldüğü üzere, müşteri bilgisinin kullanımı, segmentlere ayırma yeteneğini ve müşterileri aksiyonlarıyla birlikte analiz etme yeteneğini oldukça hızlandırır.



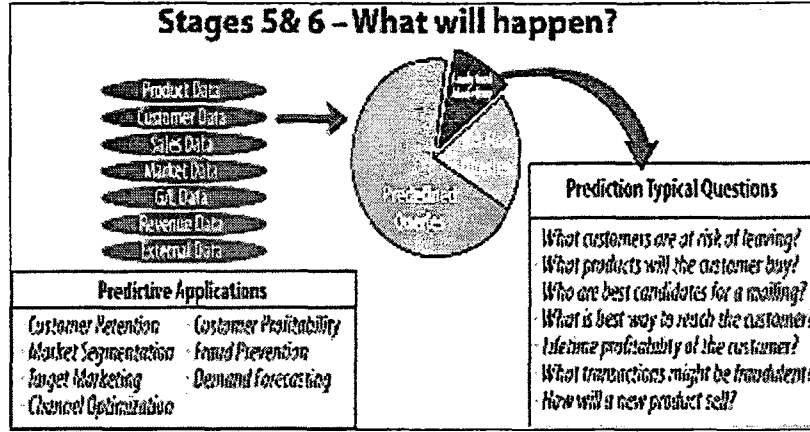
Şekil 5.4 Analiz uygulamaları ve sorular (McClain, 2000, DM Review June 2000)

5.4.3 DSS/DW Tip # 3: Geleceği Tahmin

"Analitik Modelleme" yeteneğine sahip güçlü bir veri ambarı, tıpkı kehanette bulunurcasına "Ne olacak ?" sorusuna cevap verebilecek yüksek yetenekleri sunar (bkz. Şekil 5.2, Bölüm C). Kurum şu anda "Gelecekte ne olacak?" tipindeki sorulara cevap verebilecek konuma geldi. Bu tip veya karakterdeki bir DSS ve DW uygulaması Şekil 5.5 ve 4.6'da gösterilmiştir.



Şekil 5.5 Geliştirme ve Tahmin Yetenekleri (McClain, 2000, DM Review June 2000)



Şekil 5.6 Tahminci uygulamalar ve tahminci sorular (McClain, 2000, DM Review June 2000)

Veri Ambarı ve CRM'in başarılı kullanımı için özet olarak üç madde söylenebilir:

1. Geçmişini bilmek
2. Mevcut durumu analiz etmek ve
3. (Proses ve modelleri ölçerek ve rafine ederek yüksek doğruluk oranı ile) Geleceği öğrenmek

6. İŞ AKLI YATIRIMLARININ FİNANSAL BOYUTU (ROI)

Bu bölümde; iş akıllı uygulaması yatırımlarının geri dönüşümün, yani bütçelenip maliyet-fayda düzleminde rakamsal değerlendirmeye nasıl ulaşılacağı anlatılacaktır.

6.1 İş Akıllı (BI) Projelerinde Yatırımların Geri Dönüşümünün (ROI) Hesaplanması

Organizasyonları yönetenler sürekli olarak her bir seçeneğin diğeri ile maliyet-fayda karşılaştırmasını yaparlar. Artan bir şekilde bu yöneticiler, proje yöneticilerine iş akıllı (BI- Business Intelligence) projelerini gerçekleştirmesinin maliyet-fayda değerlendirmesini sormaktadırlar. Çeşitli finansal metodlar uygulanarak bu değerlendirme yapılabilir. Örneğin; Dahili Geri Dönüşüm (IRR - Internal Rate of Return), Net Aktif Değer (NPV – Net Present Value), geri ödeme periyodu ve Yatırımların Geri Dönüşümü (ROI - Return on Investment). Her bir metodun kendine has faydaları olsa da, genel olarak kabul gören finansal değerlendirme metodu ROI olarak karşımıza çıkmaktadır (Wu, 2000).

6.1.1 ROI Metoduna Genel Bakış

ROI; kişilerin BI (iş akıllı) projesinin maliyetlerinin üzerine finansal faydaları koyarak, finansal bir ölçüm olarak sayısallaştırmalarını sağlar. Bu hesaplama işlemi her BI projesinden elde edilen tasarrufların bir yatırıma bölünmesi ile indirgenmiş ve projelendirilmiş sabit akışlarından elde edilir. Bu finansal ölçüm ve varyasyonlar; BI projesinin maliyetinin tahminini, mevcut raporlama ortamının bakım maliyetini (geçiş süresine kadar) ve geri ödeme periyodunun kabul edilebilir zaman eksenini içermektedir.

6.1.2 Bir BI Projesinin Maliyeti Nasıl Tahmin Edilir ?

Bir İş Akıllı (BI) uygulamasının devreye alınmasının toplam maliyeti nedir? Bu maliyet iki ana kategoriye ayrılabilir:

1. Direkt Maliyetler
2. Endirekt Maliyetler

Direkt maliyetler nakit hareketleridir. Yazılımın satınalma fiyatı, bakım veya destek ücreti, devreye alma giderleri (örneğin dahili ve harici işçilik giderleri), ilave donanım ve yazılım gereksinimleri ve kullanıcı eğitimleri vb. bu kapsamda değerlendirilir.

Endirekt maliyetlerin belirlenmesi veya sayısallaştırılması, kolay değildir. Çünkü genelde BI uygulamasının devreye alınmasından sonra oluşurlar, bu yüzden genelde BI uygulamasının devreye alınma işlemi toplam maliyet hesaplamasında etken bir faktör olarak ele alınmaz. Lakin, bu endirekt maliyetler toplam maliyetin önemli bir kısmına tekabül etmektedir. İstemci makine veya terminallerinin güncellenmesi, ağ iletişiminin güncellenmesi, destek yazılımının güncellenmesi, ilave dahili kullanıcı desteği ve başlangıç eğitimi vb.nin arkasından gelecek diğer kullanıcı eğitimleri de yine bu kapsamdadır. Eğer organizasyon bütçe içinde kalmak isterse, toplam devreye alma maliyetini anlamak ve de aşmamak oldukça önemli hale gelmektedir. Bu durum ayrıca BI devreye alma fazlarının belirlenmesine de yardımcı olur. Yazılım maliyetleri, bakım ve destek ücretleri ve ortalama konfigürasyon maliyeti projenin maliyet komponentlerini oluşturur ki, bu bilgi dış kaynaklardan elde edilebilir. Bir BI uygulamasının tahmini maliyeti ROI hesaplamasında kullanılacaktır.

6.1.2.1 ROI Hesaplama Örneği

XYZ şirketinin kurumsal çapta ad hoc sorgulama ve raporlama araçlarına hizmet edecek bir BI uygulaması yapmak istediğini varsayalım. XYZ şirketinin mevcut raporlama ortamının ihtiyacı aylık 1200 saatlik bir desteğe ihtiyaç duyduğunu kabul edelim. Mevcut raporlama ortamını, yeni bir BI uygulaması ile değiştirmek aşağıdaki giderleri beraberinde getirir (bkz. Çizelge 6.1).

Çizelge 6.1 Örnek XYZ şirketi için BI Gider Tablosu (Wu, 2000, DM Review, July 2000)

Giderler	
Donanım	\$250,000
Yazılım	\$400,000
İşçilik	\$750,000
Toplam	\$1,400,000

Varsayımlar :

1. BI yazılım bakım maliyeti her yıl için = % 10.
2. Her bir elemanın yıllık çalışma saati = 2000 saat.
3. İşçilik ortalama saati = 75\$.
4. XYZ'nin yatırım kazancı = % 9.
5. ROI hesaplaması için alınan zaman periyodu = 3 yıl

Çizelge 6.2 XYZ'nin BI uygulamasını kullanması durumunda 3 yıllık maliyet dağılımı
(Wu, 2000, DM Review, July 2000)

	Yıl 0	Yıl 1	Yıl 2	Yıl 3
Donanım	\$250,000			
Yazılım	\$400,000	\$40,000	\$40,000	\$40,000
İşçilik	\$750,000	\$375,000	\$375,000	\$375,000
Toplam	\$1,400,000	\$415,000	\$415,000	\$415,000

Çizelge 6.3 XYZ'nin BI uygulamasını kullanmaması durumunda 3 yıllık maliyet dağılımı
(Wu, 2000, DM Review, July 2000)

	Yıl 0	Yıl 1	Yıl 2	Yıl 3
İşçilik	\$1,080,000	\$1,080,000	\$1,080,000	\$1,080,000
Toplam	\$1,080,000	\$1,080,000	\$1,080,000	\$1,080,000
Net Tasarruflar (implimentasyon yılı)		\$665,000	\$665,000	\$665,000
%9 ile Hesaplanmış Net Tasarruflar		\$1,683,311		
ROI-Yatırımın Geri Dönüşümü				
%9 ile Hesaplanmış Net Tasarruflar		\$1,683,311		
BI Yatırımı		\$1,400,000		
% ROI-Yatırımın Geri Dönüşümü Yüzdesi		% 120		

Sonuçtaki; dönüşüm yüzdesinin sıfırdan büyük olması, BI projesinin uygulamaya alınmasının mali açıdan organizasyona yarar getireceğini göstermektedir.

6.1.3 ROI Duyarlılık Analizinin Gerçekleştirilmesi

Bir duyarlılık analizi nümerik olarak yaratılan bir çok senaryonun ve proje sonuçlarına etkisinin değerlendirilmesidir. Her bir senaryoya bir olasılık atanır ve en uygun mali ölçüm uygulanır. Eğer atanan yüzde değeri %50'den daha düşük bir başarı yüzdesi ile sonuçlanır ise, proje tekrar değerlendirilmelidir ve kritik olan başarı faktörünü belirleme ihtiyacı ortaya çıkarılmalıdır. Böylece başarısız olacak bir projenin de önüne geçilebilir.

Çizelge 6.4 Olasılık Yüzdeleri (Wu, 2000, DM Review, July 2000)

Senaryo	% Olasılık	Kar/(Zarar)	Ağırlıklı Sonuç
Başarılı	60%	\$1,683,311	\$1,009,987
Kısmen başarılı	30%	\$ -	\$ -
Başarısız	10%	\$(1,400,000)	\$(140,000)
Toplamlar	100%		\$869,987

6.1.4 İş Aklı (BI) Projelerinin Mali Olmayan Kriterleri

ROI hesaplama metodu; bir BI projesini değerlendirmek için sadece bir ölçüm değeri sunar iken, aslında mali olmayan fakat dikkate alınması gereken başka kriterler de mevcuttur. Bu mali olmayan kriterler veya kalitatif analiz; geliştirilmiş bilgi ayrımı, geliştirilmiş bilgi erişimi, anlamlı bilginin akışı vb. kriterleri içerebilir. Finansal ölçümler ile bu şekildeki kalitatif analizleri dengeleyen, bir balans mekanizmasının inşası oldukça önem arz etmektedir.

Özetle; bir BI projesinin ölçülebilir (**kantitatif**) ve ölçülemeyen (**kalitatif**) değerleri, proje başlatılmadan değerlendirmelidir. Üst yönetimden BI projesi için gerekli destek ve de kullanıcının işi içine sokulması da, başarılı bir uygulama ortaya çıkması için gerekli hususlardır. Organizasyonun içinden gerekli destek ve katılım yok ise bu BI projesi için ROI hesaplama metodunu kullanmanın bir faydası olmayacaktır ve finansal kriter olarak görülmeyen bu tip kriterler başarılı bir projeyi dahi, mali açıdan zarara uğratabilecektir.

6.2. Metadata ROI

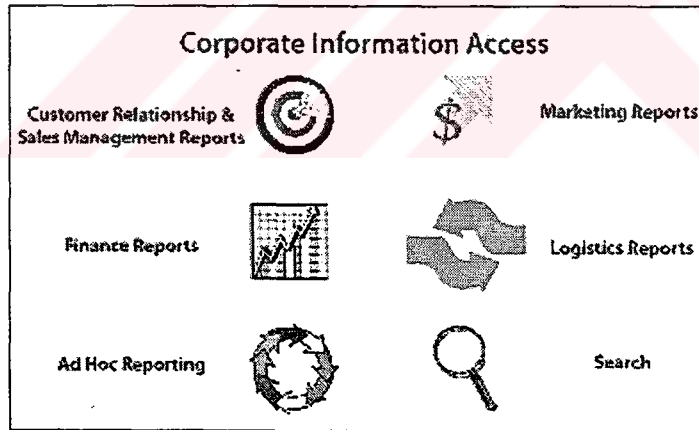
Bir meta data ambarı ciddi bir şekilde karar destek sistemlerinin geliştirme maliyetlerini (giderlerin düşmesi) azaltır ve yeni değişen karar destek sistemlerinin daha hızlı uygulamaya alınıp, pazarda uygulama sürecini (gelirlerin artışı) hızlandırır (Marco, 2000a). Meta data bunu teknik “**etki analiz raporları**”nı kullanarak realize eder. Bu etki analiz raporları kendi iç ortamlarında değişiklik önermelerini de, etkilerini de sorguladıklarından; karar destek uygulama geliştiricilerine oldukça ciddi katkıda bulunurlar.

Karar destek sistemleri verilerini işin gerçekte yürütüldüğü operasyonel sistemlerden toplarlar. Bir meta data ambarı, geliştirme zamanını, veri dönüşüm kurallarının çatısını oluşturacak süreyi, veri kaynaklarını belirlenme sürecini, BT (Bilişim Teknolojileri) sistemlerindeki veri içeriğinin belirlenme süresini ciddi manada düşürür. Bu çok kritiktir,

çünkü meta data ambarı olmaz ise dönüşüm kuralları sadece çalışanların hafızasında kalır ve de kalıcı olmaz, bu da kurumsallaşma açısından en istenmedik durumlardan biridir.

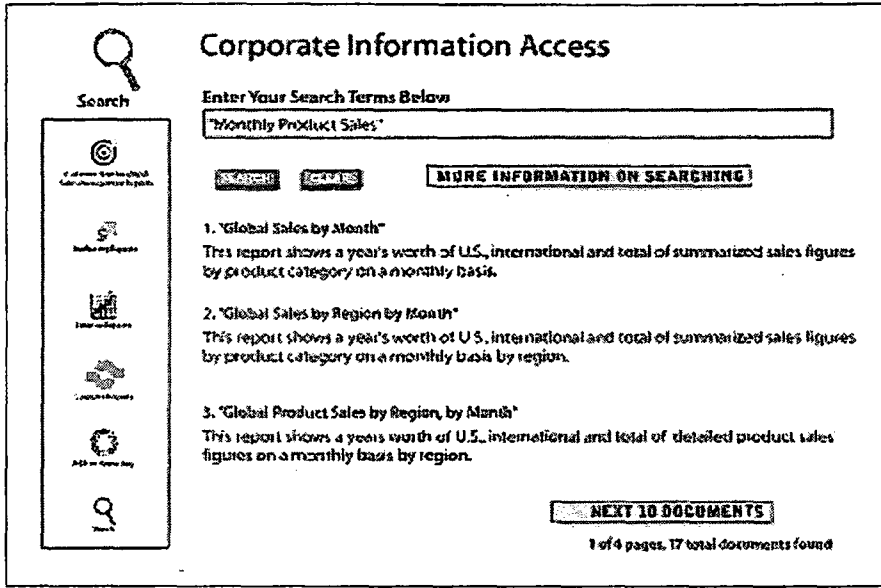
Örneğin varsayalım ki tüm sistemdeki uygulamaların hepsinde "müşteri no" alanının uzunluğu 20-byte alfanümerik iken, 30-byte alfanümerik yapılması istenmektedir. Kurumsal çapta bir etki analiz raporu, "müşteri no" alanının uzunluğundan etkilenebilecek tüm sistemleri, tabloları, dosyaları ve bunların içinde yer aldıkları alanları (domain) gösterecektir. Bu rapor çok net bir şekilde 30-byte alfanümerik değere dönüşmeyecek alanları da gösterir.

Aslında meta data BT sistemleri ile kullanıcılar arasında semantik bir ara seviyedir (Marco, 2000b). Daha açık şekliyle, meta data teknik terminolojiyi iş terminolojisinde kullanılan bildik kelimelere dönüştürür. Şekil 6.1 web-yetenekli bir karar destek sistemini göstermektedir. Görüldüğü üzere, bu ön yüz tamamen iş kullanıcısının düşünme mantığına göre dizayn edilmiştir. Buradaki mantık, iş kullanıcılarının istedikleri bilgiye hızlı ve kendi anlayabilecekleri formatta ulaşabilmeleridir.



Şekil 6.1 Meta data yönelimli iş kullanıcısı arayüzü – DSS internet sitesi (Marco, 2000, DM Review October 2000)

İş kullanıcıları DSS (karar destek) internet sitesinin arama sayfasına ulaştıklarında (bkz Şekil 6.2), kullanıcının örneğin "aylık ürün satışları"nı içeren karar destek raporları içinden arama yapabilmesidir. İşte bu nokta; meta datanın devreye girdiği noktadır. Meta data ambarında, karar destek raporlarının iş terimleri kullanılarak hazırlanmış meta data bilgisi bulunacaktır. Böylece iş kullanıcısı iş terimleriyle "aylık satışları" raporunu meta data terimleri ile yazılmış meta data bilgileri içinden arayabilecektir. Bu meta data aramasının sonuçları Şekil 6.2'de görülmektedir.



Şekil 6.2 Meta data yönelimli iş kullanıcısı arayüzü – Araştırma sonuçları (Marco, 2000, DM Review June 2000)

Meta data bilgisinin karar destek raporuna entegre edilmesi, karar destek sistemi içindeki veriyi çok daha değerli kılacaktır ve aynı zamanda karar yapımının doğruluğunu da oldukça artıracaktır.

7. META DATA

Bu bölümde; en basit tanımıyla “veri hakkında veri” olarak tanımlanan meta data konusu en ince detayları incelenecektir.. Sektöreki en büyük yazılım firmalarının oluşturduğu bu standart OMG tarafından geliştirilmiş olan CWM anlatılacaktır.

7.1 CWM (Common Warehouse Metamodel-Ortak Ambar Metamodeli) Nedir?

OMG'nin (Object Management Group- Nesne Yönetim Grubu) CWMI'ı (Common Warehouse Metada Initiative - Ortak Ambar Metadata Girişimi) geliştirme amacı, veri ambarcılığı araçları ile metadata ambarları arasında kolay metadata alış verişini dağıtık heterojen ortamlarda sağlamaktır (<http://www.cwmforum.org/about.htm>).

CWM (Common Warehouse Model - Ortak Ambar Modeli), metadata'yı paylaşabilen veri ambarı ve iş akıllı araçları tarafından sağlanan açık bir standarttır. Ortak Ambar Modelini anlamak için, öncelikle metadata'nın yer değiştirmesinin iş akıllı için niçin kritik olduğunun anlaşılması gereklidir.

Önerilen standart, bir nesne modeli sağlar. Nesne modeli, Uygulama Programı Arayüzü (API - Application Program Interface), alış-veriş formatları ve seçip alma (extraction), dönüştürme (transformation), taşıma (transportation), yükleme (loading), bütünleştirme (intergration) ve ayrıca veri ambarı projeleri analiz fazlarını da içeren çok geniş kapsamlı metadata servislerini destekler (Seeley, 2001).

CWMI'nin kökleri 1997'de OMG'nin standartlaştığı UML'e (Unified Modeling Language - Tekilleştirilmiş Modelleme Dili) dayanır. UML ile birlikte, OMG, MOF'u (Meta Object Facility) da standartlaştırmıştır. Basitçe MOF, OMG için temel dağıtık metadata yönetim mimarisidir.

7.1.1 CWM Takımı

IBM, Unisys, NCR, Hyperion, Oracle, UBS AG, Genesis Development Corporation, Dimension EDI gibi firmalar bir araya gelerek CWM'i (Ortak Ambar Metamodel)

yaratmışlardır. Bu birliktelik, Deere Company ve Sun Microsystems tarafından da desteklenmiştir.

7.1.2 CWM Mimarisi

İşbirliği (collaboration) iki anahtar ihtiyaca sahiptir. İlk olarak işbirlikçiler dilin yazım kuralı (syntax) uymak zorundadırlar. İkinci olarak, haberleşebilmelidirler. CWM haberleşme için bir yazım kuralı (syntax) oluşturarak başlar. Böylece satıcılar bağımsız bir şekilde ürünlerini metadata seviyesinde entegre edebilirler. Bu, jenerik veri ambarı mimarisinin metadata modelini tanımlar. Bu tanımlama veri ambarcılığı örnekleri için kurallar da ihtiva eder. Bunu UML içinde “temel sınıflar kümesi (set of core classes)” tanımlama işlemini yaparak becerir.

UML, nesne yönelimli sistem modelleme için grafiksel bir dildir. Bu tanımlamalar ile tanımlanmış sınıflar CWM’in içerisinde tanımlanmış tüm sınıfların atalarıdır (ancestor class). Temel CWM sınıfları (Boyut, Dönüşüm ve Eşleme modelleri olarak) alt sınıflara ayrılmışlardır.

İkinci olarak, işbirlikçilerin biri, diğeri ile haberleşebilmek zorundadır. CWM, metadata’yı XMI ile yer değiştirir. XMI (XML Metadata Interchange- XML Metadata Alış-verişi) dizi haline getirilmiş alış-veriş modelleri için geliştirilmiştir. XML (Extensible Markup Language - Büyütülebilir Artış Dili) WWW Koalisyonu (W3C) tarafından geliştirilmiştir ve XMI da yine OMG tarafından geliştirilmiştir.

CWM’nin veri ambarcılığı ve iş akıllı ürünleri için en önemli hedefi, web tarayıcılar için kullanılan HTML’nin gerçekleştirdiğidir.

7.1.3 CWM Kullanımı

CWM, tasarımından yapısına, işletiminden bakımına kadar, veri ambarcılığı yaşam döngüsündeki tüm aşamaları etkiler. Bunun faydaları yalnızca servis sağlayıcılar ve satıcılar için değil aynı zamanda bilgi teknolojisi yöneticileri ve de son kullanıcılar içindir. Şimdi CWM’in veri ambarcılığı ortamındaki bir değeri nasıl taşıdığını inceleyelim.

Ham veri operasyonel ortamdan çekip çıkarılır, tutarlı formata dönüştürülür ve ambara yüklenir. Verinin kaynaktan hedefe hareket prosesi, CWM dönüştürme modeli olarak

tanımlanır. Bu model kaynak ve hedef veri parametrelerine ek olarak dönüşüm mantığını da içermektedir. Ambar, dönüşüm modeli ile tanımlanmış düzenli prosesler halinde başlatılabilir. CWM, Mayıs 2000 de OMG'ye sunulmuştur.

7.2 Metadata Nedir?

Metadata'nın "**Metadata veri hakkında veridir**" şeklindeki klişeleşmiş tanımı gerçekte yeterli bir tanımlama değildir. *Meta* Yunancadan gelir. Anlamı *yanında* veya *sonra* şeklindedir. "*Meta-*" ön-ekini kullanmak nesneye, "taban nesneden sonraki" veya "yüksek bir aşaması" anlamını yükler. Tabanı aşmaktan daha kapsamlı bir anlamı vardır (Giovinazzo, 2000).

Terminolojik olarak, metadata normal veriden daha kapsamlı bir anlam ihtiva eder. Metadata veri parçaları için sadece format ve isim sağlamaz: Gerçekte var olan veri parçaları için genel içeriği sağlar. Bu iş aklı için kritiktir. Dönüştürülmeden ve temizlemeden sonra bu veri, karar deste araçları ve bağımlı datamartlar tarafından ambara geçirilir. Veri bir prosesden diğerine geçirilirken aynı zamanda verinin içeriği – metadata - paylaşılabilir zorundadır.

Metadata, veri ambarcılığında sonu olmayan bir tartışma konusudur (Dyché, 2000). Veri tanımlamaları ve yönlendirilebilir bilgiyi sağlayarak metadata veri ambarlarının kullanım kolaylığını getirir. Metadata Koalisyonu (MDC) olarak adlandırılan, çeşitli veri ambarcılığı satıcılarından oluşan bir grup, metadata saklama modelini yakın zamanda tanımlamıştır.

Üç temel metadata tipi vardır:

1. **İş metadatası** – Aynı zamanda son kullanıcı metadatası olarak da bilinir. İş metadatasının örneği SIC (Special Industry Code - Özel Endüstri Kodu) olabilir. İş metadatası; veri kaynak sistemleri tanımlamalarını, verilerin eşanlamlılarını, veri erişim kısıtlamalarını, geçerli değerleri ve çok önemli olan iş kurallarını da içerebilir.
2. **Veritabanı metadatası** – Özellikle veri tabanı nesnelerinin adları, veri tabanları, tablolar ve kolonlar. Örneğin, "ürün" tanımlaması, iş metadatası olarak kabul edilebilir. Ürün tablosunun açıklayıcı notları veri tabanı metadatasının işleri olacaktır. Veri tabanı metadatası kaynak sistem eşleme bilgisi gibi, idari bilgiyi de içerir.

3. **Uygulama metadatası** – Uygulama metadatası son kullanıcılar uygulama araçlarındaki bazı terimleri ve fonksiyonları açıklar. Bunu yaparak uygulama seviyesindeki terminoloji karmaşasından kaçınmak amaçlanmaktadır. Örneğin, aylık satış raporu. Aylık satış raporu ikonu aylık ürün satış listesini çalıştırır. Buna satış müdürleri, kendileri için tanımlanmış olan erişim hakları ile erişebilir. Aylık satış raporu örneğin bir önceki ayın son gününden sonraki 3 gün içinde erişilebilir kılınmıştır.

Ambarın içereceği iki tip metadata vardır:

1. **Teknik metadata:** Ambarın büyümesi ve süregelen bakımı için son derece kritiktir. Teknik metadata olmadan karar destek sisteminin zaman tüketiminin hesaplanması, iş analizinin ve değişikliklerin yapılması çok zordur.
2. **İş metadatası:** Veri ambarı ile iş kullanıcıları arasındaki bağlantıdır. İş kullanıcıları temelde idareciler veya iş analistleri gibi teknik eğitimi az kullanıcılardır. Onlar teknik terminoloji yerine iş terminolojisi ile hazırlanmış karar destek sistemlerine ihtiyaç duyarlar. İş metadatası; raporlardaki, sorgulamalardaki iş terimlerini, ambardaki verileri, verinin yerini, verinin güvenilirliğini, verinin içeriğini, veri üzerinde hangi dönüşüm kurallarının uygulandığını, verinin hangi klasik (legacy) sistemde kaynak olarak bulunduğunu sergiler.

İki genel tipte metadata kaynağı vardır.

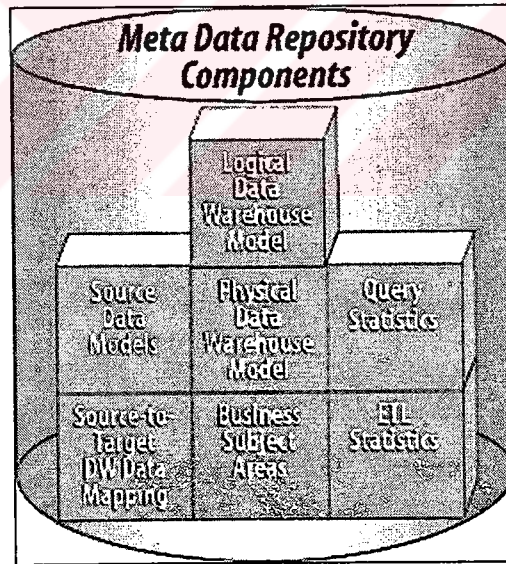
1. **Formal metadata kaynağı:** Bu tipteki metadata kaynakları tartışılmaktadır. Dokümanite edilmiş ve karar yapıcılar tarafından kabul edilmiş olmalıdır. Bu formal metadata kaynakları hem teknik hem de iş metadatası toplarlar.
2. **Formal olmayan (informal) metadata kaynağı:** Kolektif bilgi içerir. Politikalar ve ana noktalar standartlaşmamıştır. Bu bilginin her biti formal metadata kaynaklarında değerlidir. İş ile ilgili eğilimleri ihtiva eden çoğu değerli bilgileri bünyesinde barındırır. İş metadatasının çoğunlukla formal olmadığını vurgulamak önemlidir.

7.3 (Jenerik) Metadata Ambarı

Metadata ambar modeli ortamın merkezidir (Jennings, 2000). Burası, operasyonel kaynak sistemlerinin entegre edildiği, veri ambarı, seçip alma, dönüşüm ve yükleme (ETL) prosesleri, iş değerlendirmeleri, raporlar ve operasyonel istatistiklerin yapıldığı yerdir.

7.3.1 Meta Data Ambar Komponentleri

Şekil 7.1’de görüldüğü üzere metadata ambar veri modelinin 7 standart komponent vardır. İş ihtiyaçlarına göre model; silinerek, güncellenerek veya yeni komponent ilave edilerek revize edilebilir. Eğer şirket, tekrarlanan veriyi kaynak sistemden silmek için veri silme araçları kullanıyorsa, modelin ETL komponent alanı uygun bilgiyi almak için revize edilebilir. Alternatif olarak, eğer sorgulama istatistikleri almak konu dışı sayılırsa veya ortamdaki başka anlamların (OLAP araçları, metadata tabakası, veritabanı) içerisine yerleştirilirse, bu komponent modelden silinebilir.



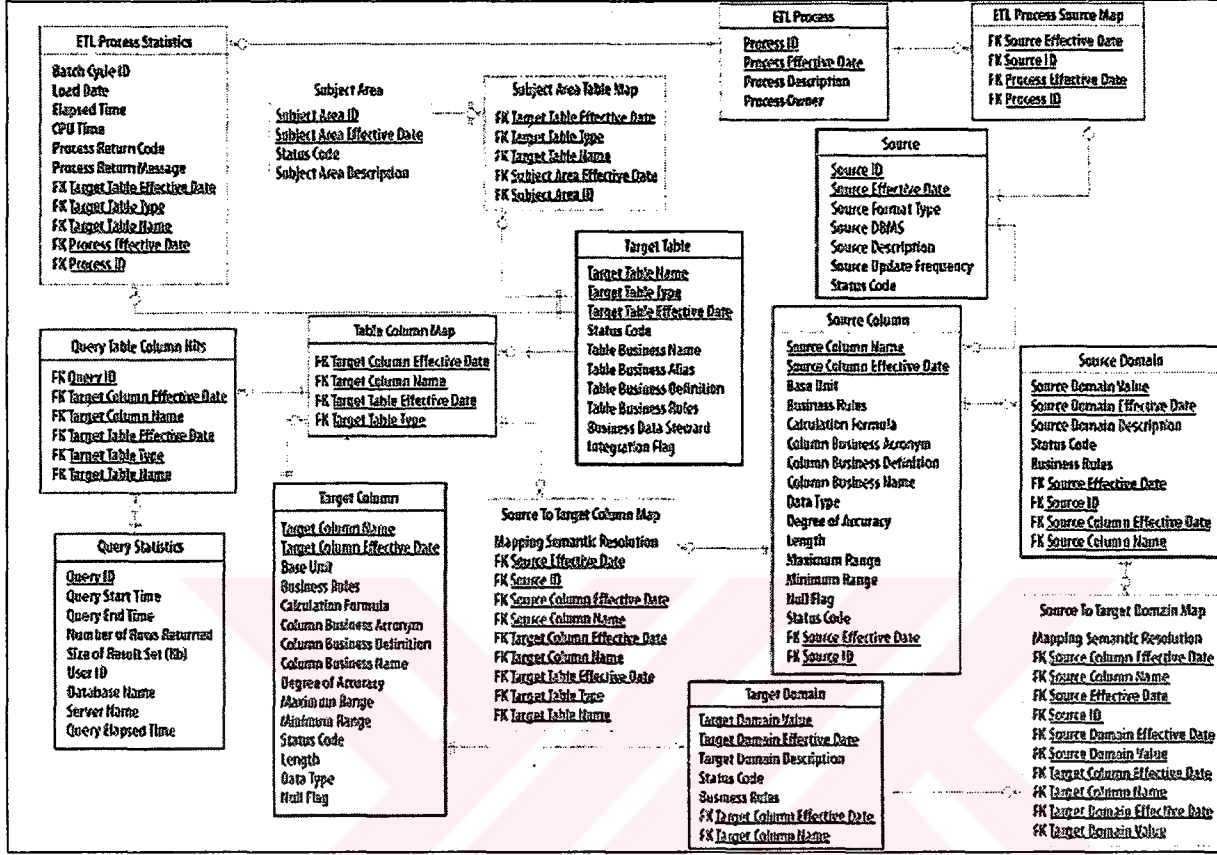
Şekil 7.1 Meta Data Ambar Komponentleri (Jennings, 2000, DM Review October 2000)

1. Jenerik metadata ambar modelinin ilk komponent alanı mantıksal veri ambar modeli hakkında bilgi içerir. Mantıksal veri ambar modeli, tüm özel iş konusu alanları için iş başlıklarını, başlıklar arasındaki ilişkileri ve her hedef modelden oluşan başlıkların niteliklerini içerir.

2. İkinci komponent, hedef modelin fiziksel veri ambarı hakkında detaylı bilgi içerir. Mantıksal veri ambar modeli, fiziksel veri ambarı boyutsal modeli olacak şekilde değişim süreci gecirmek zorundadır.
3. Jenerik modelin 3. komponenti, veri ambarını besleyen kaynaklar hakkında fiziksel bilgi içerir. Operasyonel bilgi kaynak veri tabanlarından, dosya seçip alımlarından, internet ve diğer formatlardan kaynaklanabilir.
4. Jenerik modelin 4. Komponenti, kaynaktan hedefe veri ambarı eşlemeleridir. Bu komponent tablo/kolon çapraz-referans eşlemesi ile operasyonel kaynak sistem ve hedef fiziksel veri ambar modeli arasındaki anlamsal eşleme çözümünü saklar. Bu metadata ambarının başlıca amaçlarından birisidir.
5. Jenerik modelin 5. Komponenti, iş konusu alanlarıdır. Burada fiziksel veri ambarı tablolarının mantıksal gruplaması saklanır.
6. Jenerik modelin 6. Komponenti, ETL istatistikleridir. Bu komponent içerisinde, tekil seçip alımlardaki, dönüşümdeki ve yükleme proseslerindeki bilgi yakalanır. Bu bilgi, gelişme prosesini tespit için, veritabanı güçlendirmeleri için, hata izolasyonu ve ambar için diğer optimizasyon prosedürlerini tespit için kullanılabilir.
7. Jenerik metadata ambar modelinin 7. ve sonuncu komponenti, sorgulama istatistikleridir. Veri ambarı veritabanına karşı yapılan her sorgulama hakkındaki bilgi, bu komponent içerisinde saklanır.

7.3.2 Jenerik Ambar Modelinin Mantıksal Görünümü

Jenerik meta data ambar modelinin, mantıksal model görünümü aşağıdaki gibidir (Şekil 7.2).



Şekil 7.2 Jenerik Ambar Modeli'nin mantıksal model görünümü (Jennings, 2000, DM Review October 2000)

Gerçek veri ambar modelinde, hem mantıksal hem de fiziksel bilgi ihtiva eden dört tablo vardır :

1. Hedef Tablo (Target Table)
2. Hedef Kolon Tablosu (Target Column Table)
3. Hedef Kolon Eşleme Tablosu (Target Column Map Table)
4. Hedef Alan Tablosu (Target Domain Table)

Hedef Tablo (Target Table) ambar içerisindeki Hedef Tablo Tipi (Target Table Type) kolonlarını kullanarak tabloların hem mantıksal hem de fiziksel görünümünü saklayabilir.

Tablo Kolonu Eşleme Tablosu (Target Column Map Table) basitçe mantıksal ve/veya fiziksel tablolar ile onların birleştirilmiş kolonları arasında çapraz-referans sağlar.

Hedef Kolon Tablosu (Target Column Table) içerisindeki alanlar hem teknik hem de iş ile ilgili bilgi içerir.

Hedef Alan Tablosu (Target Domain Table) sadece kodları ihtiva eden kolonlar için değerlerin tanımlamalarını içerir. Her olası alan değeri veya kolon için gerekli kod bu tablo içerisinde saklanır. Örneğin, ülke kodu kolonu Birleşik Devletler için USA ve Kanada için CAN, v.b alan değerlerine sahiptir.

Veri ambarı için gerekli bilgiyi besleyen operasyonel sistemler hakkında fiziksel bilgi aşağıdaki üç tabloda saklanır:

1. Kaynak Tablo (Source Table)
2. Kaynak Kolon Tablosu (Source Column Table)
3. Kaynak Alan Tablosu (Source Domain Table)

Kaynak Tablo (Source Table) kaynak sistem veritabanı veya ambarın hedef tablolarına bilgi toplamak için seçilmiş dosyalar hakkında bilgi içerir. Kaynak ID kolonu, kaydın belirli sistemini tekil bir şekilde belirlemek için kullanılır (örneğin; ERP, Sipariş Yönetimi). Kaynak Format Tipi, sunucu/veritabanı, dizin/dosya veya elektronik tablolamalarda (örneğin Microsoft XL spread sheet) olduğu gibi kaynak bilginin kategorisini belirlemek üzere bir yöntem sağlar.

Kaynak Kolon (Source Column Table) ve **Kaynak Alan** (Source Domain Table), Hedef Komponent Tabloları, Hedef Kolon ve Hedef Alan içerisinde bulunan aynı tanımlamaları takip eder. Bu işlemi sadece kaynağı hedefteki yerine koyarak yapar.

Kaynaktan-Hedef Kolon (Source-to-Target Column Table) ve **Alan Eşleme** tabloları (Domain Map Table), seçip alma, dönüştürme ve yükleme (ETL) prosesleri için gerekli tasarım stratejisini operasyonel kaynaklarını besleyici ambar tablolarına bağlamak suretiyle inşa edilmesini sağlarlar.

Üç tablo - ETL Proses, ETL Proses Kaynak Eşleme ve ETL Proses İstatistikleri (ETL Process, ETL Process Source Map and ETL Process Statistics) - ambar tabloları ve operasyonel kaynakları, onlara erişen ve yükleme prosesleri hakkında istatistikleri alan prosedürler için eşleme yaparlar.

ETL Proses Tablosu ambar ortamında ETL prosedürlerini tekil bir şekilde teşhis etmek için proses ID kolonu içerir.

ETL Proses İstatistik Tablosu hem ambar tabloları için yönlendirmeyi hem de belirli ardışıl döngü yüklemesi için, prosesin tablolarını ve ayrıca tarih/zaman istatistiklerini içerir. Basit ETL prosesi, bir veya birden fazla hedef ambar tablosunu ilgili iş ihtiyaçları ve/veya kaynakları yükleyebilir.

ETL Proses Kaynak Eşleme tabloları ETL prosesleri ve operasyonel sistemler veya seçilmiş dosyalar arasında çapraz-referans sağlar.

Konu Alanı ve Konu Alanı Tablo Eşleme tabloları, mantıksal gruplanmış son kullanıcı veya veri ambarı tablolarının iş değerlendirmesini sağlar. Örneğin; satış, finans, insan kaynakları.

Konu Alanı tablosu tablo gruplarını tekil bir şekilde belirleme için Konu Alanı ID' si kolonu içerir. Tablo grupları iş perspektifinde düşünülmüştür. **Konu Alanı Tablo Eşleme tablosu**, konu alanı için basitçe ambar tablolarını(mantıksal ve fiziksel) gruplar.

Sorgulama İstatistik (Query Statistics) tabloları ambara karşılık yapılan sorgulama sonuçları hakkında çeşitli bilgiler içerir.

Sorgulama Tablosu Kolonu Ulaşımlar tablosu (Query Table Column Hits Table) veri tabanı sorgulamaları ile ambar tabloları ve kolonları arasında çapraz refererans sağlar.

8. KAYNAK, VERİ VE UYGULAMA ENTEGRASYONU

Bu bölümde; veri entegrasyonu veri ambarının temeli olan kaynak entegrasyonu, veri entegrasyonu ve uygulama entegrasyonu konuları ele alınacaktır.

8.1 Kaynak Entegrasyonu

“ Kaynak entegrasyonu, veri ambarcılığının temel problemlerinden biridir. Kaynak entegrasyonu istenen uygulamalar için gerekli tasarım ve bakım için gerekli iki kritik faktör, alanın (domain) içerik modellemesi ve konsept gösteriminin bir sebebe bağlanmasıdır. ”
(Calvanese, Giacomo, Lenzerini, Nardi. ve Rosati, 1998).

8.2 Dahili & Harici Verinin Entegrasyonu

Dahili veri şirketin kendi sistemlerinde toplanıp, yönetilmelidir (Inmon, 2001). Dahili veri, veri ambarını oluşturur. Harici veri şirket dışından operasyonel işlemler sonucu toplanır. Aradaki fark verinin içeride veya dışarıda işlenmesidir. Veri ambarcılığında dahili ve harici verinin harmanlanması önemlidir. Yöneticilerin şirketin kritik başarı faktörlerini analiz edeceği zaman, harici ve dahili verilerde ki bilgileri, özellikle rakamları, birbiriyle karşılaştırmaları oldukça yararlıdır. Böylece hem dahili hem de harici verileri kıyaslayabilen yöneticiler daha doğru bir bakış açısı kazanırlar ve daha başarılı kararlara imza atabilirler. Genelde harici veri dahili veriyle uyumsuzluk gösterir ve harmanlanması zordur. Özetle, eğer dahili veri, harici veri ile de aynı ortamda bulunuyor ise o oramdaki veri ambarı tamdır.

8.2.1 Dahili & Harici Veri Entegrasyon Senaryoları

Veri tiplerini harmanlamak neden çok zordur? Çünkü ne anahtar bilgilerin; ne içerikleri, ne anlamları, ne de yapıları birbirine uyar. Bu durumu açıklamak için üç değişik senaryoyu kurgulayalım. Bu senaryoların özetlenmiş şekli Çizelge 8.1’de gösterilmiştir.

Çizelge 8.1 Dahili & Harici veri entegrasyonu senaryoları özet tablosu (Inmon, 2001)

Senaryo	Format	Nasıl çalışır?
1	İdeal formatlar, harici veriler yapısaldır.	Harici anahtarlar dahili anahtarlar ile çalışır.
2	Harici veri ham veridir.	Harici veri dahili veri ile çalışacak şekilde dönüştürülür (transformasyon).
3	Kaba denebilecek veri, daha iyi veri ile eşleştirilebilir.	- Harici veri, dahili veriden daha kabadır: Posta kodu (harici) & Kişisel (dahili) - Dahili veri, harici veriden daha kabadır: Şehir (harici) & Eyalet (dahili)

8.3 Veri Entegrasyonu : Veri Ambarının Temeli

Tarihin hiç bir anında bugünkü kadar hızlı, çok yönlü bir hayat ve rekabet yaşanmamıştır. Değişim doğrusal olmayan bir şekilde artmaktadır. Ne kadar bilirsek, o kadar daha çok bilmeye ihtiyaç duyuyoruz (Rolleigh ve Thomas, 2001). Veri entegrasyonundaki asıl hedef, veri kalitesini yükseltmektir. Kaliteli veri şu üç kritik prosesin bir fonksiyonudur:

- Operasyonel sistemlerden seçilen verinin temizlenmesi ve konsolidasyonu :** Her bir müşterinin işlem tarihçesinin yakınlığı, sıklığı ve parasal değeri (RFM); pazarlama ve satış personeli için son derece kritiktir. Çünkü RFM müşterinin gelecekteki davranışlarını ve bu yüksek tahmin endikatörlerini baz alarak "müşteri segmentasyonu" oluşturabilmeye imkan verir.
- Her bir müşteri kaydının posta, demografik, jeodemografik, hayat standardı ve fizikografik veri elemanları ile zenginleştirilmesi :** Bu tip veriler, müşteri profillerininin yapısallaştırılması ve modellenmesiyle müşteri davranışlarının tahminine imkan verir.
- Efektif metadata :** Metadata, veri ve geçirdiği prosesler hakkındaki bilgidir. Her bir prosesin sonuçlarını ve de önceki projelerdeki veri kalitesi eğilimlerini belirleme fonksiyonlarını da içerebilir. Aslında metadata olmaksızın, veri ambarının ne içeriği ve ne de kalitesinin analiz edilmesi imkansızdır.

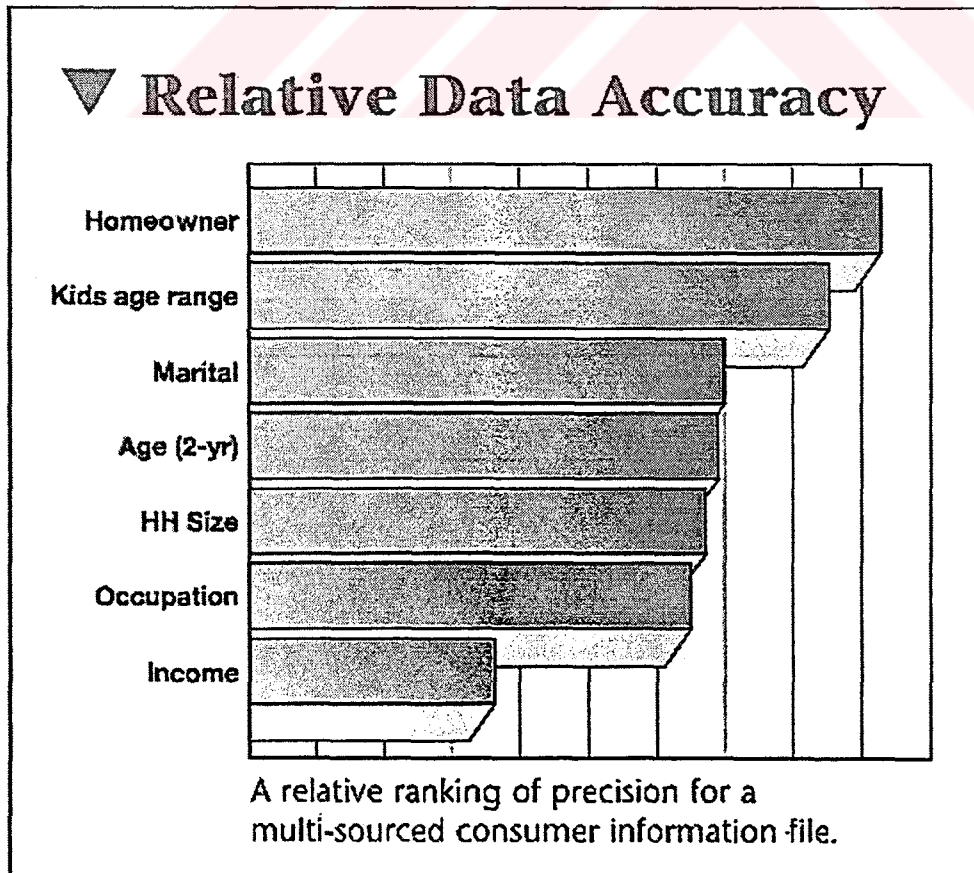
8.4 Verinin, Kesinlik, Genişlik ve Derinliği

Harici veri elemanları satın alınacağı zaman; fiyatından başka, içerdiği veri kalitesinin de değerlendirilmesi önemlidir. Daha iyi anlamak için veri kalitesi elemanlarını kavramak ve şu üç determinantı maksimize etmeye çalışmak gerekir:

1. Kesinlik (Precision)
2. Genişlik (Breadth)
3. Derinlik (Depth)

8.4.1 Kesinlik (Precision):

"Göreceli (relatif) Veri Doğruluğu" grafiğinde (Şekil 8.1) çok-kaynaklı tüketici dosyasındaki ortak elemanlar için, veri kesinliğinin tipik göreceli ölçeklenmesi görülmektedir. Çok derleyici modelinden doğru olmayan veya eksik olan verileri ayarlama yararlanır. Ve bir "çok-kaynaklı dosya", herhangi bir veri elemanının doğru okunması ihtimalini artırır. Kesinlik elde etmek için anahtarlar; **güvenirlilik** (reliability) ve **doğrulanmışlık** (validity) dır.



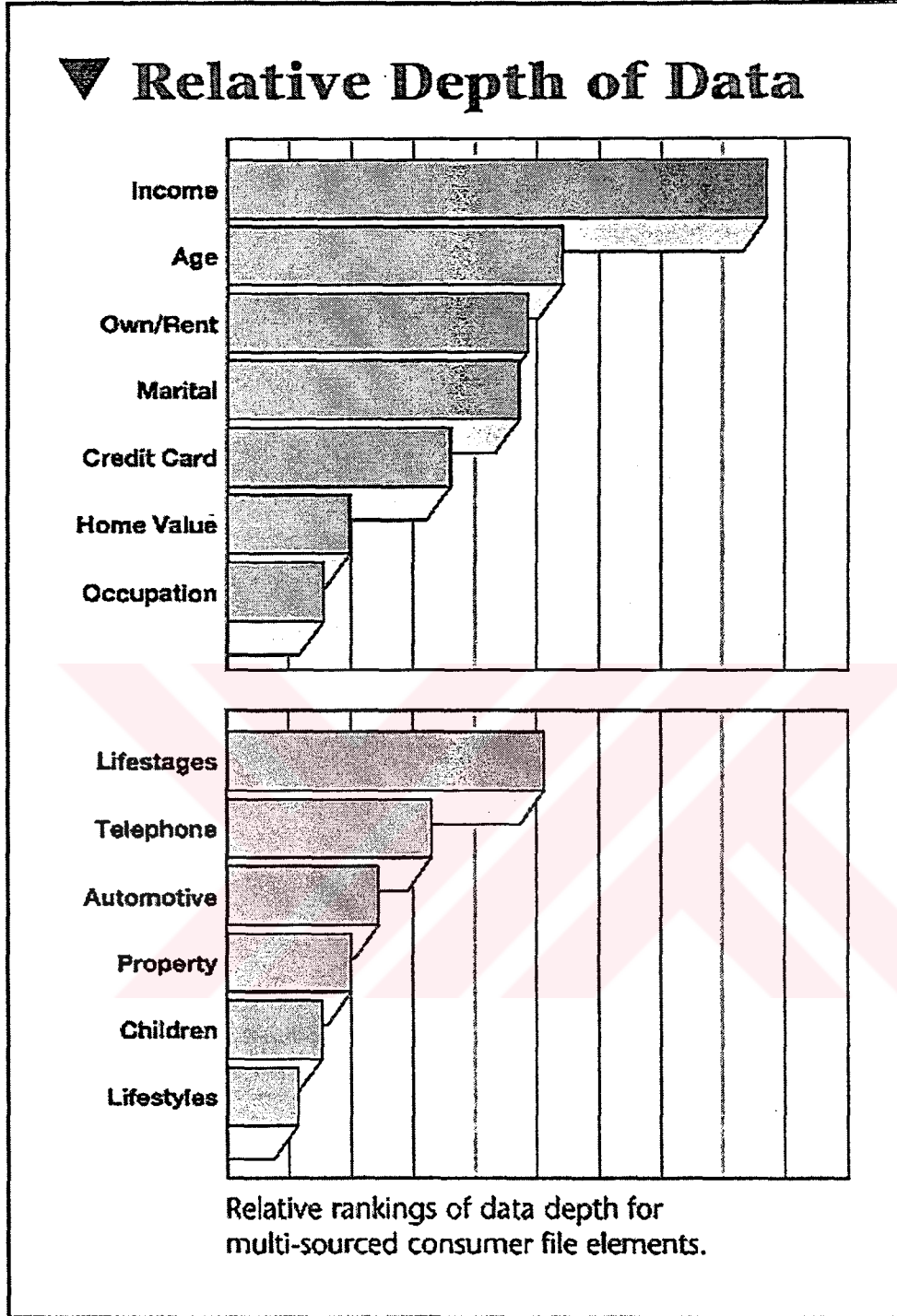
Şekil 8.1 "Göreceli (Relatif) Veri Doğruluğu" grafiği (Axicom Corp., www.axiom.com)

8.4.2 Genişlik (Breadth):

Zenginleştirilmiş verideki ikinci kalite determinantı, genişliktir. Genişlik, bir veri elemanından elde edilebilecek örtüşme bilgisidir (ayrıca "eşleme oranları" olarak da bilinir). Eşlemeye hazır olan kayıt sayısı ve eşleme işleminin karmaşıklığı olarak da tanımlanabilir.

8.4.3 Derinlik (Depth):

Bazı uygulamalar kapsam derinliğine ihtiyaç duyarlar, bununla her bir tekil isim için ne kadar veri kapsandığının tanımı yapılır. Derinlik, veri elemanı ve kaynağına göre değişkendir ve tipik derinlik ölçeklendirmesi "Göreceli (Relatif) Doğruluk Derinliği" grafiğinde (Şekil 8.2) gösterilmiştir. Çok-kaynaklı ortamlar, bir çok dosyanın kombine edildiği yerler, veri derinliğinin güçlü artışa sahip olduğu yerler olarak ortaya çıkmaktadır.



Şekil 8.2 "Göreceli (Relatif) Doğruluk Derinliği" grafiği (Axicom Corp., www.acxiom.com)

8.5 Veri Entegrasyonunda Ambar Hatalarından Sakınma

Görüldüğü üzere, veri entegrasyon proseslerinin doğru çözümlenmesi veri ambarı çalışmasının başarıyla sonuçlanması ihtimalini artırır. Bu doğrular:

1. İş ihtiyaçlarına uyarlanmış sofistike standardizasyon, birleştirme/silme yetenekleri ile doğru olmayan, tutarsız ve hatta kayıp verilerin düzeltilebilmesi sağlanır.
2. Veri bütünleştirme veri garantisi için temel yetenekler oluşturabilme ile ileriye götürülebilir.
3. Müşteri verileri, harici veriler çok ani bir şekilde zenginleştirilip, güçlendirilebilir.
4. Veri hataları minimize edilebilir- hatta tamamen elimine bile edilebilir- eğer elde uygun araçlar ve problemlerin belirlenmesi için prosedürler kullanılırsa.
5. Ambarın tasarımı esnasında, pazarlamacıların verileri nasıl kullanacağına dair tüm ihtimaller değerlendirilmelidir. Bu tasarım; verinin nasıl tanımlanacağı, nasıl saklanacağı, hatta güçlü analitik araçlarla (örneğin OLAP) nasıl kullanacağını etkileyecektir.
6. Veri inşa edildikten sonra, artık düzenli bir şekilde güncellenmeli, konsolide edilmeli ve özetlenmelidir. Hatta belirli aralıklarla (periyodik) harici veriler uygulamak suretiyle zenginleştirilmelidir.

8.6 ERP (Kurumsal Kaynak Planlama) Projelerinde “Veri Bütünlüğü”

Bir ERP uygulaması devreye alma çalışmaları sırf, doğru verinin eksikliğinden ötürü beklenen faydaları sunamamaktadır (Jones, www.eaijournal.com).

ERP planlama ve çalıştırma sistem mantığına temel giriş olarak kabul edilen veri dosyalarındaki hatalar, tüm ERP proseslerinde potansiyel hatalar oluşturacaktır. Bu dosyalardaki hatalar, temizlenmesi en güç hatalardır. Kabul edilebilir seviyeye getirmek, ortalama 4-6 ay sürmektedir. Veri bütünleştirilmesi, gözardı edilebilir ise yeni ERP sisteminin, eski doğru olmayan planlama verileri aynı olacak ve kayıplardan oluşan eski sistemdeki sonuçlara benzer sonuçlar üretilmiş olacaktır.

8.7 e-iş Projelerinde “Veri Bütünlüğü”

"E-iş" uygulamaları inşa edilirken, en alt seviyede doğru olmayan verilerin oluşturacağı etkinin minimize edilmesi için şirketlerde strateji geliştirilmelidir. Ayrıca, organizasyon iş ortakları ile olan ilişkilerindeki veri tutarsızlıklarının etkisini de azaltmalıdır. Bunlar müşteriler veya tedarikçilerde olabilir (Hammer, 2000).

8.7.1 B2C’de “Veri Bütünlüğü”

Web üzerinden bir şeyler almaya çalışanların pek çok sebebi vardır. Bu kişiler ayrıca ürünler, mevcudiyetleri ve fiyatları hakkında doğru bilgiler ve basit bir satınalma şekli isterler. İstemedikleri tek şey; ürünün artık olma~~ya~~acağı veya daha kötüsü zamanında teslim edilemeyecek olmasıdır.

“www.abc.com” veya “www.def.com” diye iki e-komisyon sitesi ele alalım. Her ikisi de üreticilerden gelen ürünleri satsınlar. Genelde aynı ürünü üreten birden fazla firma bulunur. İdeal olarak e-komisyon firmaları hiç bir envanter tutmamalılar veya çok-çok az stok ile çalışmalıdırlar. Bir başka deyişle en fazla satılan ürünleri, zamanında hatta daha önce müşteriye ulaştırabilmek için bunları stoklarında garantilemelidirler. Bu tip ürünlerin çoğu için mevcudiyetinin takibi ihtiyacı doğmaktadır. Bu tip ürünlerden biri sipariş edildiğinde şu üç temel yaklaşımdan birini uygulamaya koyabilir:

1. En iyiyi varsayıp, siparişi almak.
2. Üreticilerden gelen gerçek zamanlı ilgili bilgiyi erişim sağlamaya çalışmak.
3. Doğru olmama ihtimali de olsa zamanında bir veri önbellege yaratmak, burada zorunlu bilgileri saklamak

8.7.2 B2B’de “Veri Bütünlüğü”

Veri akışı perpektifinden bakıldığında, B2B önemli ölçüde B2C teknolojisinden farklıdır. B2C bir e-komisyon firması için üreticilerini veri tabanından gelen bilgiyi elde etme ve siparişi ile ilgili güncelleme bilgilerini gönderme işlemler için çift yönlü iletişime ihtiyaç duyar. Bununla birlikte, B2B de, “kaskat-tetikleyiciler” tipi, bir çalışma söz konusudur. Her bir işlem diğerini tetikler ve bu şekilde devam eder.

Bir çok durumda, üreticilerin müşterileri ile bir ilişkisi vardır ve risk oluşabilecek alanlarda karar verebilmek veya riski göğüsleyebilmek için, bu ilişkiyle elde edilen müşteri tarihçesinden yararlanabilir. Eğer B2B vizyonu gerçekleştirilebilir ise, o üreticinin (tedarikçinin) müşteri sayısında, doğal olarak da ürün ve hizmetlerin sayısında exponansiyel bir artış olacaktır. Çok yüksek hız ve çok düşük stokla çalışmada oluşabilecek en küçük hata bile, çok kritik durumlara sebep olmaktadır. Bu yüzden B2B kullanan şirketlerin gelir tahmini yapabilmeleri oldukça zor olmaktadır.

8.8 Uygulama Entegrasyonu

Mimari bir ana kriterdir. Manuel kodlama veya veri dönüşüm programları ile üretilmiş verilerde bile kaynak ve hedefler arasında direkt bağlantı için zorunlu kriter mimaridir (Data Junction, Corp., 1999).

Mimari, proseslerin segmentasyonu ve ETL (Extract Transform Load – Seçme Dönüştürme Yükleme) proseslerinin gerçekleştiği lokasyon kontrolü imkanına ihtiyaç duyar. İster tüm ETL proseslerini tek bir ana makinadan toplu olarak yapalım, ister birden fazla makina da hatta paralel yapalım, yük dengelemesi ve diğer hususlar için bir mimariye ihtiyacımız vardır.

8.8.1 RDF, Meta Data ve EAI

RDF (Kaynak Tanımlama Çatısı), XML'in bir parçasıdır. Uygulamalar arası veri değişimini ve ortak çalışmayı sağlar (Linthicum, 2000a). Dünya WWW komisyonu (W3C) değişik kaynakların tek bir meta data ile ortak çalışmasını sağlamak üzere RDF'i geliştirmiştir.

Bunun için RDF, XML ile kullanılır. Buradaki fark veya amaç; XML'den taşınan veri için ortak bir format kullanmak için istifade edilirken, XML'in üst katmanlarında yer alan RDF ise veri kategorisi tanımlamak için kullanılır. XML verisi RDF formatında olduğu zaman, uygulamalar verinin kim tarafından gönderildiğini bilmeksizin veriyi bilebilir veya anlayabilir. RDF; meta data, veri arama, kataloglama ve ilişkilendirme gibi birçok alanda (EAI dahil) uygulanabilir.

Bugün bir çok EAI projesi, ambar (repository) konseptinden istifade eder. Bu ambarlar veri elemanları, girişler, prosesler, çıkışlar ve uygulamalar arası ilişkiler gibi bilgileri içerebilir. Ambar, EAI mimarisi için gerekli tüm bilgileri ve programcı için kurumsal yapıda herhangi bir bilgi parçacığını nereye koyacağını ve bağlantı sağlayacağını belirten tüm bilgileri sağlamalıdır. Hatta zorunlu durumlarda EAI için ana klasör (veya dosya) olmalıdır. Gelecekteki uygulama geliştirmeleri için kesinlikle referans olarak ana klasördür.

Ambarlar aslında meta data bilgisini kaydetmeden ibaret değildir. Metadata'yı, EAI çözümünün kalbi olarak tanımlamak mümkündür, çünkü hem veri hemde iş model bilgilerini birlikte içermektedir. Bir ambarı (repository) hayata geçirmek için üç temel adım vardır:

1. Verinin belirlenmesi,
2. Verinin kataloglanması,
3. Kurumsal meta datamodelinin oluşturulması.

Kurumsal meta data model, ambar baz alınarak, kurum içindeki çeşitli veri depolarının entegrasyonu için ana-klavuz olarak kullanılmaktadır. Üç tip ambar vardır:

1. Mantık Ambarı uygulama geliştirme için
2. Nesne Ambarı uygulama geliştirme için
3. Meta data Ambarı EAI & e-iş için

Hali hazırda "de facto standard" ve "ambar" modeli mevcut değildir, fakat kurumsal ambar ve birleşik kurumsal meta data modeli yaratmak, mükemmel bir entegrasyon mimarisi için ilk adımdır. Ancak, bu gerçekleştirilmesi oldukça zor bir durumdur.

8.8.2 EAI için Bilgi Adalarının Birleştirilmesi

Fortune 1000 olarak adlandırılan dünyanın en büyük 1000 şirketinin CIO'ları (Bilgi İşlem Organizasyonu Yöneticileri) arasında yapılan araştırma sonucu, karşılaştıkları en büyük sorunun kurumsal bilginin değişik parçalarının nasıl birleştirileceğidir (Stonebrake, 2000).

Bilgi ada veya adacıklarının bağlanabilirlikleri yeterli değildir. Çünkü, bunlar birlikte çalışmaları için başlangıçta dizayn edilmemişlerdir. Bilgi adalarının iki özelliği buna mani olmaktadır:

1. **Anlam Uyuşmazlığı:** Her bir ada (örneğin, uygulama) kendi kurumsal nesnelerinin anlam bilgisine sahiptir, örneğin müşteriler, satışlar ve gelirler.
2. **Veri Tekrarlaması:** Her bir adanın verisi diğer adalardaki verilerle kısmen çakışabilir, yani tekrarlanabilir. Bu kısmi tekrarlamaya veri bütünlüğü açısından çok ciddi problemler doğurabilir.

Bilgi adaları birleştirmesindeki ana sebepler şunlardır:

1. Daha iyi karar yapabilmek için uygulama kullanıcılarına tekil bir bilgi görünümü sağlamak,
2. Çapraz-sistem tutarlığına erişmek.

EAI çözümleri oluşturmak için kullanılan çeşitli ürün tipleri Çizelge 8.2’de görüldüğü gibi kategorize edilebilir.

Çizelge 8.2 EAI oluşturmak için gerekli ürün tipleri (Stonebrakre, 2000).

MOM (Message-Oriented Middleware)	Yayın ve abonelik sistemleri
TP (Transaction Processing)	Uygulama sunucuları, veri ambarları, data martlar
DFS(Data Federation Systems)	Mantıksal entegrasyon sistemi

Burada iki grup teknoloji kullanılmaktadır:

1. **İtme (Push) Teknolojileri:** örneğin, MOMS, TPs, Application Servers.
2. **Çekme (Pull) Teknolojileri:** örneğin, DFSs, Gateways, Distributed Database.

İtme teknolojisinin iki ciddi eksikliği vardır:

1. Birleştirilecek lokal yapıların uyumlu oldukları varsayılır (Daha öncede belirtildiği üzere, kurumsal nesneleri oluşturan bilgi adaları nadiren uyumluluk gösterir).
2. Dağıtık veri tabanı ürünleri, sorgulamaları çalıştırmak için merkezi optimasyona ihtiyaç duyar (Merkezi kaynak ataması, lokal kaynaklardaki DBA'lere verilen yetkilendirmeyi etkisiz kılar).

Bu tip problemleri çözmek için, DFS (Data Federation System-Veri Birleştirme Sistemi) denilen yeni tip bir ürün kullanılır. DFS, sistemlerin bağımsız olarak birleştirilmesi, mantıksal bütünlüğe kavuşması için tasarlanır. DFS ve dağıtık veri tabanları, veri seviyesinde entegrasyonu gerçekleştirirler.

8.8.3 EAI için Tasarım Kalıpları: Veri Seviyesi & Metod Seviyesi

Tasarım kalıpları arayüzlerin tanımlanması, katagorize edilmesi ve tekrar kullanımını sağlayabilmek için oluşturulan bir yaklaşım metodudur. Böylece uygulamadan-uygulamaya kurulacak iletişim en üst seviyeye çıkarılabilir. Pratik amaçlar için iki temel kalıp söz konusudur (Linthicum, 2000b) :

1. Veri seviyesi
2. Metod seviyesi

Farklarını incelemeyen önce, kalıpların tasarımlarda görülen hataların bir çeşit tanımını yaptığını belirtmemiz gerekir. Her bir kalıp üç-parçalı şematik bir yapıya sahiptir:

1. **İçerik**, problem oluşan noktaları veya durumları tanımlar veya kalıp için doğrulamadan doğru şekilde nasıl geçileceğini tanımlar.
2. **Problem**, kural tarafından adreslenen yani konteks içerisinde zaman-zaman oluşan sorunlardır.
3. **Çözüm**, problemin iyi kanıtlanmış ve tutarlı önermelerle giderilmesi

Veri seviyesi tasarım kalıpları, tipik veri depoları arasında verinin taşınması prosesidir. Yani bilginin bir veri tabanından seçilmesi, gerekiyorsa işlenmesi ve bir diğer veri tabanında güncellenmesi olarak da tanımlanabilir. Veri seviyesi şu tasarım kalıplarını içermektedir:

1. İlişkisel,
2. Nesne-yönelimi,
3. Hiyerarşik veya,
4. Çok boyutlu kalıplar.

Metod seviyesi arayüz tasarımı kalıpları, kurum içinde mevcut olan iş mantığının paylaşımıdır. İki temel yaklaşım mevcuttur. Tek bir ortak fiziksel sunucu (örneğin uygulama sunucusu) üzerinde, ortak uygulama sunucuların ortak bir kümesi yaratılabilir veya tam tersi dağıtık nesneler kullanarak, mevcut uygulamaların daha önceden içinde olduğu bilinen metodların paylaşımı yaratılabilir. Metod seviyesi şu tasarım kalıplarını içermektedir:

1. Nesne yönelimli,
2. İşlemsel veya,
3. Prosedürel kalıplar.

EAI, meta datadan çok yüksek oranda istifade eder. Daha da detaylandırırsak, meta data EAI'nın damarları arasında dolaşan onu besleyen, güçlü olmasını sağlayan kan gibidir (Moragenthal, 1999).

Meta data ambarını kurumsal-çapta hayata geçirmek için "klasör (directory) servisleri" çoğunlukla gözardı edilen ihtimallerdendir. Organizasyonel meta data olarak kabul edilen "çalışanlar, telefon numaraları, fiziksel donanım ve güvenlik" bilgilerine dağıtık erişim sağlarlar. Geleneksel olarak, klasör servisi "sarı sayfalar" olarak tabir edilen elektronik sayfa

teknolojisinde kullanılır. Bununla beraber, klasör servisleri dağıtık ve hiyerarşik meta data verilerinin yakalanması ve sunulması uygulamaları için idealdir. Maalesef, klasör servislerinde arayüz problemi sıkça yaşanır. Üç yaygın veya gelişmekte olan arayüzler şunlardır:

- Lightweight Directory Access Protocol (LDAP)
- NetWare Directory Services (NDS)
- Active Directory Service Interface (ADSI).

LDAP açık protokole sahiptir ve en alt seviyede diğer ikisi için ortak fonksiyonel paydayı destekleyebilir. Fakat şu anda LDAP'nin kurumsal çaptaki bir meta data servisi fonksiyonunu destekleyebilecek güçte sağlam olup olmadığı bilinmemektedir.

Klasör servisleri arayüzü veritabanı-yönelimli metodolojilerinin yerini alabilirler. Bu arayüz MDC (Meta Data Coalition - www.mdcinfo.com) tarafından geliştirilen OIM (Open Information Model - Açık Bilgi Modeli) tabanlıdır. OIM, UML (Unified Markup Language), XML ve SQL kullanılarak veri-merkezli meta data değişim spesifikasyonu sağlar.

9. VERİ KALİTESİ

Bu bölümde; veri kalitesi ve tutarlılığı arasındaki ilişkileri, veri tekrarlamaları, veri anormallikleri ve veri kalitesizliği yaratan diğer problemler incelenmiştir.

9.1 Veri Kalitesi Konsepti

Veri kalitesi (veya bilgi kalitesi) çok değişik şekillerde tanımlanabilir. **Veri kalitesi** sürekli bir şekilde müşteri ihtiyaçlarına hitab eden bilgi olarak tanımlanabilir. Kendine özgü olarak, mantıksal ve tutarlı sırada bilginin düzenlenmesi prosesi olarak da tanımlanabilir. Burada amaç veriyi kuvvetlendirmek, zenginleştirmek, temizlemek ve birden fazla aynı amaca hizmet eden veriyi tek bir kayda indirmek (No, 2000).

Diğer bir tanımlama da Larry English (English, 2000a, DM Review January) tarafından yapılmıştır: “Bilgi kalitesi, sadece veri kalitesini ölçmek değildir aynı zamanda kalitesiz bilginin iş maliyetlerini ölçmektir. Ayrıca, bilgi kalitesini artırmak sadece veri temizleme için değildir, proses içerisindeki tekrerr eden yapılarıdaki temel sebepleri bulmak ve düzeltmek içindir.”

Veri kalitesinin önemine değinen bir Amerikan amirali ise şu sözü söylemiştir:

“Bir doğru ölçüm, bin uzmanın fikrinden daha değerlidir.” (Forino, 2000a)

(Grace Hopper, Amiral, US Donanması)

Mükemmel denebilecek bir IQ (Information Quality) çözümü, yine Larry English (English, 2000b, DM Review October 2000) , tarafından aşağıdaki kutucukta gösterilen bir formüller yaklaşım ile ifade edilebilir.

IQ Çözümü = Eğitilmiş, kalifiye insanlar +

Tanımlanmış iş hedefleri (temel nedenler anlaşılmış ve çözümlenmiş) +

Tanımlanmış IQ prosesleri +

Kendini ispatlamış araçlar (ürünler)

Bu eşitlikten de anlaşılacağı üzere, araçlar (ürünler) tek başına çözüm değildir. Ancak, maalesef yazılım araçlarının çözüm olarak kabul edilmesi oldukça yaygın olan yanlış bir kanıdır.

Şirket kültürü açısından bakıldığında veri kalitesi uygulamasına başlanması, kültürel bir değişim (Culture Change) veya bilinen kavramı ile değişim yönetiminin temelini ifade etmektedir, ki bu da şu şekilde formülize edilebilir (Lori Silverman, Progress Software).

$$C = D \times V \times F > R$$

C = Change (Değişim)

D = Dissatisfaction of product (ürün memnuniyetsizliği)

V = Vision (Vizyon)

F = First steps reaching to vision (Vizyona erişimdeki ilk adımlar)

R = Resistance (Direnç)

Değişime gösterilen direnç, değişim veya çarpım işlemine giren diğer unsurlar aşamaz ise veri kalitesi çalışmalarının getireceği kurumsal (EDQM alt başlığında görüleceği üzere) değişim perspektiflerinin başarısız açılımları oluşacaktır. Başarı her zaman tercih edilendir. Değişime karşı direnci, en basit uygulamalarımızın (stok kodu tanıtım ekranı gibi) yeni teknolojiler ile yazılmış halinde bile son kullanıcılardan görmekteyiz.

English (English, 1998a) düşük kaliteli verinin maliyetini ve onun organizasyona etkileri üzerinde çalışmıştır. Kaliteden kaynaklanan ve veri kalite problemlerine çözüm olması bakımından, eğitimi savunan bir değer biçme testini formüle etmiştir. (English, 1998b; English, 1998c). Bununla birlikte, English (English, 1998a), Strehlo (Strehlo, 1996) ve Cipriano (Cipriano, 1995) gibi bir çok kişi veri kalitesine erişebilecek çeşitli yolların tamamını ortaya koyamamışlardır. Bence her şeyin başı olduğuna inandığım eğitim, veri kalitesizliğini minimize edebilecek çözümlerin sadece bir parçasıdır. Tabiidir ki, minimize etmenin içerisinde başka ölçütler de vardır.

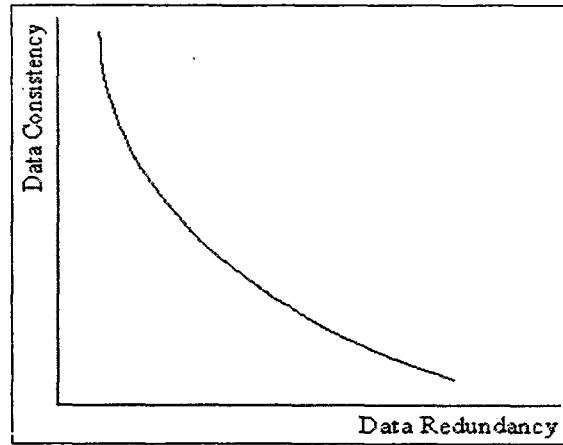
Delone ve McLean'ın araştırmalarından (Delone ve McLean, 1992), ortaya çıkan yedi önemli nokta şöyledir:

1. Bilginin doğruluğu,
2. Çıktı zamanlaması,
3. Güvenilirlik,
4. Tamlık,
5. Aşıkılık/Anlaşılabilirlik,
6. Kesinlik ve
7. Doğruluk.

Veri ambarındaki veri kalitesinin başarısında oldukça önemi olan iki önemli nokta, veri temizliği ve veri kalitesi arttırmalıdır (English, 1996). Şu anda dünya çapındaki şirketler/organizasyonlar veri bütünlüğü açısından çok önemli olan uygulamanın kritik noktalarına odaklanmışlardır (Jordan, 1997). Veri tekrarlarını ortadan kaldırarak veya kontrol ederek veri yoğunluğuna erişebiliriz. Birleştirilmiş iyi veri ile birlikte veri yönetimi veri doğruluğunu bir üst seviyeye çıkaracaktır. Veri ambarcılığının büyümesi 6 milyar dolarlık pazardaki, verilerini aktif bir şekilde yönetmek isteyen organizasyonlar olduğunu göstermektedir (Trillium Software System-White Paper EN TRIL0200, 1999).

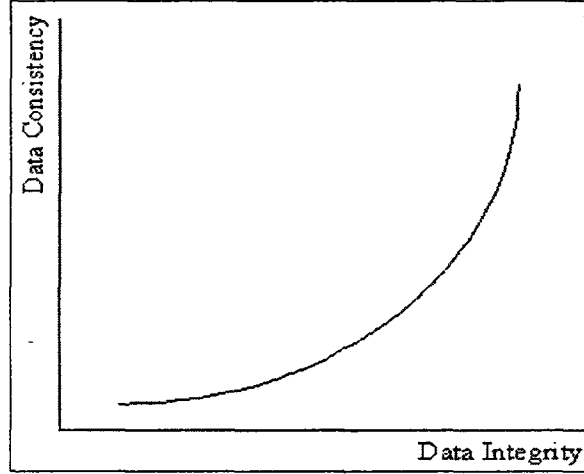
9.2 Veri Tutarsızlığı

Veri tutarsızlığı veri tabanı içerisinde aynı verinin farklı versiyonlarının olması durumunda oluşur (Awad ve Gotterer, 1992). Şekil 9.1'de görüldüğü gibi, veri tutarlılığı ve veri tekrarlaması arasında ters orantı vardır. Veri tekrarlaması az olursa veri kalitesi yüksek olacak demektir.



Şekil 9.1 Veri tutarlılığı ve veri tekrarlaması ilişkisi (Rudra ve Yeo, 1999)

Veri tekrarlamasını kontrol ve elimine ederek veri tutarlılığına ulaşabiliriz. Birleştirilmiş veri yönetimi veri bütünlüğünü bir üst seviyeye taşıyacaktır. Diğer yandan, veri tutarlılığı ve bütünlüğü arasında doğru orantı vardır (Şekil 9.2).



Şekil 9.2 Veri tutarlılığı ve veri bütünlüğü ilişkisi (Rudra ve Yeo, 1999)

Firth'in (Firth, 1997) veri kirliliği nedenselliği (Çizelge 9.1) Rudra, A. ve Yeo, E.' nin araştırmalarında kullanıldı ve veri yöneticileri için anket olarak değiştirildi. İnceleme hem açık uçlu yani cevabı tartışmaya açık, hem de tartışmasız sorular ile yapıldı. Savunan kişilere beş noktalı skala üzerinde (düşük mevkiden yükseğe) veri kirliliğinin sebep olduğu bazı önemli noktalar soruldu. Alınan sonuçları değerlendirmek için nicel analiz kullanıldı. Avusturalya'nın 50 büyük işyerindeki 500 çalışanı hedeflendi.

Çizelge 9.1 Veri kalitesizliğinin sebepleri (Firth, 1997)

Ne zaman / Nerede	Niçin
1. Sistem değişimleri, yeni şekle dönüşümler veya yeniden yapılanmalar	- Değiştirme proseslerinde yetersiz veri kalitesi testleri. - Değiştirme programlarının ortaya çıkardığı yeni hatalar.
2. Heterojen sistem entegrasyonları	- Çapraz sistemlerde değişken veya tutarsız veri. - Entegrasyonda yetersiz veri kalitesi testleri.
3. Heterojen sistemlerin post-entegrasyonu	- Veri çapraz sistemlerde çelişkili veya tutarsızdır. - Zor farkedilen düşük kaliteli veri yeni senaryoların oluşmasına yol açar.
4. Üretim yazılımı	- Yazılım ihtiyaçları tamamlanmış veya geliştirme sürecindeki hatalar. - Yazılım mühendisliğinin veya üretim kontrolünün olmaması.
5. Veri tabanı tasarımı	- Kayıt ve alan tanımlamaların çok dağınık, yapısız veya normalleştirilmemiş olması. - Şema eksiklikleri, yetersiz kontroller ve bütünleştirme kuralları
6. Veri yaşlandırma	Şirket verinin yaşını takip edemez veya güncelleme veya veriyi zenginleştirmek için programa sahip değildir.
7. Müşteri Reaksiyonu / Reaksiyonsuzluğu	- Veri hiçbir zaman tam yakalanmaz. - Müşteri formları kötü bir şekilde tasarlanmıştır veya müşteriyi reaksiyon vermesi için motive edilmemektedir.
8. Sahtekarlık	Fiziksel ve mantıksal sistem güvenliği savsaklanmıştır veya telfi eden kontroller yoktur.
9. Sistemi uluslararasılaştırma	Kısmen kaplamak veya yorum veya kodların kullanımı semboller, formatlar ulusal farklılıklar içermektedir.
10. Giriş hatası	- Sistem giriş metodları kötü bir şekilde tasarlanmıştır veya otomatik onaylamalar yoktur. - İnsan hataları kolayca ele alınmamıştır.
11. İş kuralları	Sistem ihtiyaçları elverişli değildir veya mevcut iş kuralları veriye cevap vermemektedir.
12. Politika ve planlama	Veri kalitesi yönetimine karşı olan yönetim duyarsızlığı.

Araştırma sonucu veri tabanlarının ortalama kurulma süresi 37 ay, ortalama kullanıcı sayısı 36 çıkmıştır.

9.3 Veri Kalitesizliğinin Sebepleri

Çizelge 9.2, Rudra A. ve Yeo E. tarafından yapılmış olan anketin sonuçlarını göstermektedir.

Çizelge 9.2 Veri kalitesizliğinin sebeplerinin derecelendirilmesi (Rudra ve Yeo, 1999)

Veri kirliliğinin sebebi	Ortalama puan	Standart Sapma
Hiç değiştirilmeyen veri	3.0	1.41
Heterojen sistem bütünleştirmesi	2.6	0.79
Yönetimin politika ve planlama eksikliği	2.6	0.98
Eski veri ile başa çıkma yeteneksizliği	2.4	1.13
Veri tabanı tasarımındaki hatalar	2.4	1.51
Mali problemler	2.3	0.76
Giriş hatası	2.3	1.25
Tutarsız yorumlara sebep olan sistem lokalizasyonu veya kodların, sembollerin ve formatların kullanımı	2.1	0.79
Güvenlik ve hile	2.0	1.41
Yetersiz yazılım geliştirme	1.9	0.69
Yetersiz sistem dönüştürme, geçiş veya yeniden yapılanma	1.7	0.82

Ortalama puanlara göre, araştırma sonuçlarından veri ambarında veri kirliliğine sebep olan üç temel neden aşağıdakiler gibi gözükmektedir:

1. Hiç değiştirilmeyen veri
2. Heterojen sistem bütünleştirmesi
3. Yönetimin politika ve planlama eksikliği

Kirlenmenin temel sebebini bulmak için, uygulanan proses, anlaşılmış veya kontrol edilmiş olmak zorundadır. Bu görev İngilizce de **Gemba**, Japoncada ise **KAIZEN** (KAI = Değişim, ZEN = Daha iyi için) prensipleri olarak adlandırılır (English, 2000c). Aslında, Gemba bütün bilginin kaynağıdır. Örneğin bir hipermarkette Gemba, her bir POS cihazıdır.

9.4 Veri Kalitesinin Ölçülmesi

Gerçekte, veri kalitesini ölçmek ve yönetmek için güvenilir ve yerinde bir yöntem yoktur (Forino, 2000b). Verinin ne kadar iyi veya kötü olduğunu söyleyecek ölçümlere ihtiyaç duyulur. Tesis edilen tanımlamalar ve iş kuralları ile verinin ne kadar uygun gerçeklikte olduğunu gösteren istatistiksel raporlara ihtiyaç duyulur. Ankette, veri kalitesi aşağıdaki yöntemlerle ölçülmüştür:

1. Verinin bütünlüğü (Completeness),
2. Veri girişinin tutarlılığı (Consistency),
3. Verinin doğruluğu (Accuracy),
4. Anahtar verilerin tekilliği (Uniqueness of primary keys),
5. Veriyi destekleyen iş kurallarının sürekliliği (Durability of business rules).

Veri kalitesi ölçme ile ilgili diğer yaklaşım (Bongiovanni, 2000) dört aşamadan oluşmaktadır. Aşağıdaki alt bölümler, veri kalitesi ile ilgili problemlerin önceden (yani proaktif bir şekilde, tıpkı koruyucu hekimlik gibi) yakalanıp, grafiksel olarak raporlayan bu metodu detaylı bir şekilde açıklamaktadır.

9.4.1 Önemli Verinin Belirlenmesi

Önemli verinin saptanması görevi aslında oldukça basittir. İlk adım, organizasyon için konu içeriği uzmanlarının (SMEs - subject master experts) belirlenmesi ile ilgilidir. Sonuçta elde edilen, organizasyon içerisindeki en önemli boyutların ve gerçeklerin listesidir.

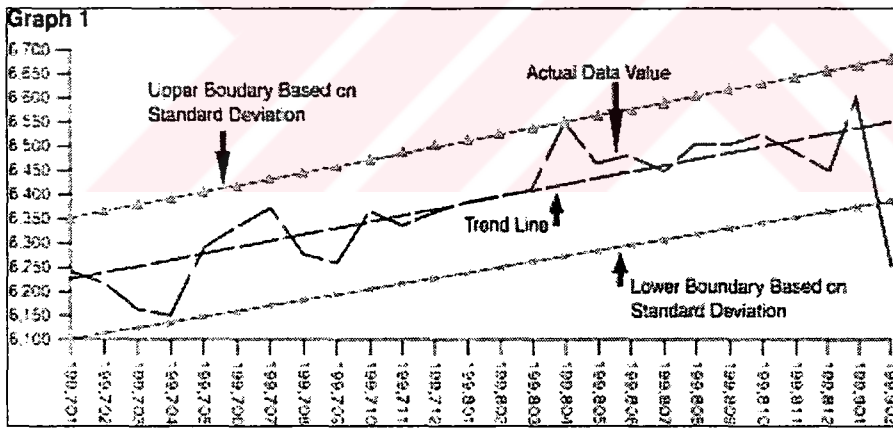
9.4.2 Veri Kalitesini Ölçmek İçin Bir Yöntem Geliştirme

Önemli veri gerçekleri ve boyutları listesinin saptanmasından sonra, Bongiovanni veri kalitesini ölçmek için bir yöntem geliştirmiştir. Bongiovanni, kaliteyi ölçmek için en iyi yöntemin kaynaktan hedefe doğru, istatistiksel olarak geçerli örneklerin kontrolünün mümkün olduğunu tesbit etmiştir.

9.4.3 Ölçüm Metodu Tabanlı Toleransların Geliştirilmesi (Alan İstatistikleri Analizi - Analyze Domain Statistics)

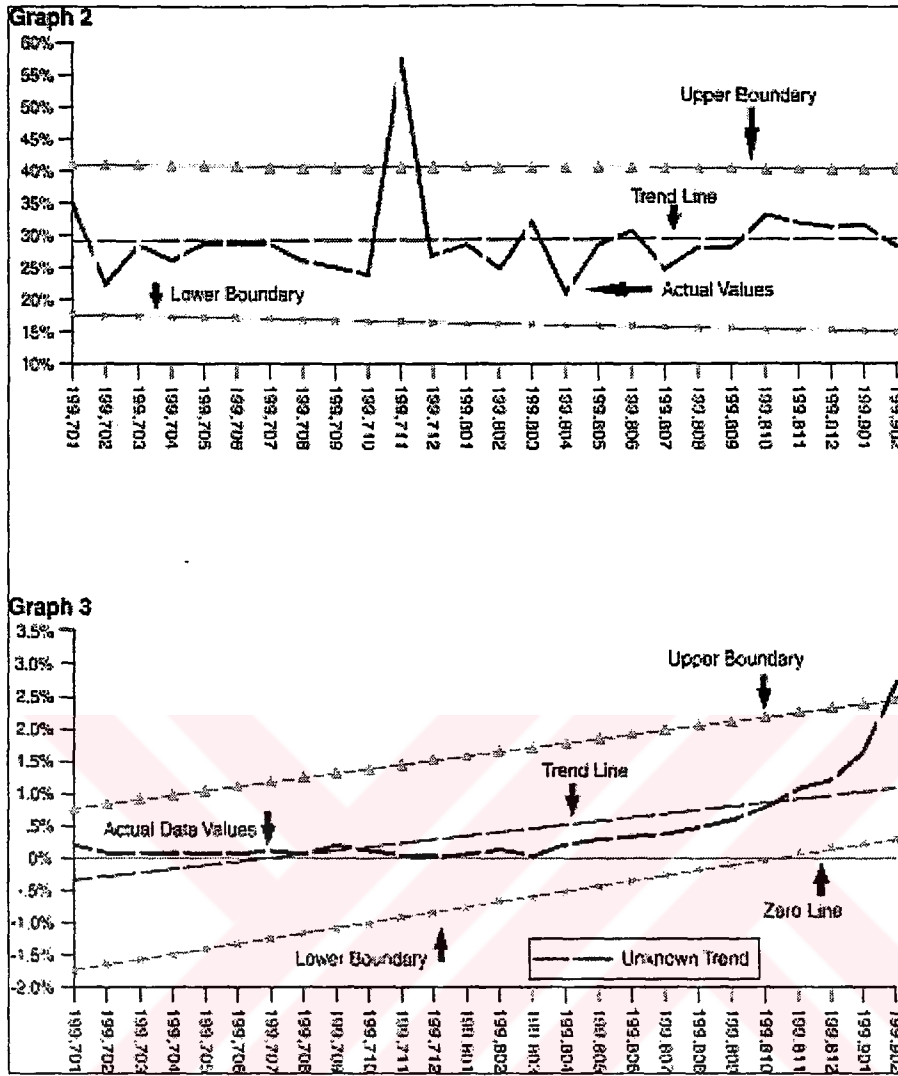
Alan istatistikleri (Domain statistics); özel alanlar içerisindeki veriden türetilen istatistiklerdir. Böylece verinin farklı periyotlar arasındaki değerleri mukayese edilebilir. Bu eğilim analizini kullanırken, veri içerisinde oluşan değişikliklerin grafiğini çizebiliriz. Bu değişiklikler alan istatistikleri analizi için hayati önem ihtiva etmektedir. Bongiovanni; çıkarılan, şekil değiştiren ve yüklenen prosetteki verinin tutarlı olduğundan emin olmak istedi. Bongiovanni bunu istatistiksel geçerli örnekler kullanarak saptadı.

Bu yaklaşımın arkasındaki teori şudur. En küçük seviyedeki doğal dengeye sahip bir iş bile ani ve dramatik değişiklikleri engelleme eğilimi gösterilmelidir. Böylece herhangi bir şey çok hızlı veya dramatik bir şekilde değiştiği zaman, eğer bu pozitif bir değişiklik olsa bile muhtemelen tesbit edilebilecektir. Ani ve dramatik değişiklikleri tesbit etmenin en iyi yöntemi, alan istatistiklerinin analizinin grafiğini çizmektir.



Şekil 9.3 Grafik 1 (Bongiovanni, 2000, DM Review June 2000, s:32)

Gerçek veri değerleri, Grafik 1' deki (Şekil 9.3) diğer çizgiler için temel oluşturur. Grafikte paralel bir şekilde bulunan diğer üç çizgi gerçek veri değerlerinden türetilmiştir. Merkezdeki çizgi gerçek veri değerleri için lineer regresyondur ve “**trend line** - eğilim çizgisi” olarak adlandırılır. Eğilim çizgisi önemlidir, çünkü alt ve üst sınırları tesbit edilmesine yardımcı olur. Standart sapmanın bazı çeşitlerinin kullanımı, ilerdeki sorgulamalarda gerekli olan verideki değişik tipteki sapmalar araştırabilir.



Şekil 9.4 Grafik 2 ve Grafik 3 (Bongiovanni, 2000, DM Review June 2000, s:71)

En basit ve en yaygın veri sapması sivri bir şekilde görülen kırılım noktalarıdır (extremum noktaları). Bu örnek, verinin alt ve üst sınırların dışına çıktığı basit bir örnektir. Eğer Grafik 2'yi (Şekil 9.4) incelersek, Kasım 1997'deki değer yüzdesi (bu noktada bilinmese de) dramatik boyutlara ulaşmış bir şekilde diğer aylardan farklılık göstermektedir. Bu nokta, dönüştürme probleminin, yükleme probleminin, veri kazanç probleminin sebebi olabilir veya aynı zamanda verinin doğru gösterimi de olabilir.

Grafik 3 (Şekil 9.4) oldukça değişik tipte bir problemi göstermektedir. Grafik, Nisan 1998'de başlayan bilinmeyen değerlerin sayısındaki artış eğilimini göstermektedir. Bununla birlikte artış Şubat 1999'a kadar herhangi bir şekilde, yeterli dramatiklik seviyesinde bir tepe yapmamıştır. Bu değer nasıl değiştiğine bağlı olan standart sapmanın nedenidir ve değişim aşamalı olduğunda, standart sapma azar azar değişimle büyür.

9.4.4 Kalite Güvenliğini Otomatikleştirme

Var olan problemlerin tespiti yerine, her boyut ve gerçek istatistik tesbit edilmek zorundadır. Bu toleransların kullanımı, aylık bazda alan istatistikleri değerlendirilerek veri ambarındaki kalitenin gösterilmesinden daha kolaydır. Toleranslar; tarihsel gerçeklere, uzman fikirlerine ve istatistiksel yöntemlere bağlıdır. Toleranslar tepe endikatörü olabilir veya onlar gerçek değerlere temel oluşturabilir. Gerçek değerler kullanılarak, Grafik 3’de gösterildiği gibi, problemi yakalamak mümkün olabilir. Alan istatistikleri analizi kullanılarak, bir organizasyon problemini tesbit edebilir ve oldukça faydalı bir şekilde onları önceden yakalayacak bir metod geliştirilebilir.

9.5 EDQM (Enterprise Data Quality Management – Kurumsal Veri Kalitesi Yönetimi)

Kurumsal Veri Kalitesi Yönetimi (EDQM) ile verinin doğru, zamana uygun, konuya uygun ve tutarlı olarak tüm organizasyon çapında ve organizasyon içindeki çoklu iş üniteleri arasında kalabilmesi sağlanır. Bu da doğru ve tutarlı veriden sağlam kararların oluşturulabilmesini sağlayabilecektir. Müşeri İlişkileri yönetimi (CRM), Tedarik Zinciri Yönetimi (SCM) ve Kurumsal Kaynak Planlama (ERP) uygulamaları çok müthiş bir hızla büyümüş ve çoğalmıştır. CRM, SCM ve ERP uygulamaların efektif olabilmesi için kullandıkları verilerin güvenli (kaliteli) olması gerekir (Chow, 2000).

Efektif EDQM uygulaması veri-temizliği (data-cleansing) maliyetlerini oldukça aşağılara çekmektedir. EDQM de tıpkı TQM (Total Quality Management – Toplam Kalite Yönetimi) gibi kurumsal çapta ve tek bir sistem gibi tasarlanarak geliştirilmesi gerektiğinden oldukça zor, pahalı ve belkide başarısızlık şansı yüksek bir uygulamadır.

Veri kalitesi değerlendirmesi prosesi yaparken, kabul edilebilir tüm veri aralıklarının kurumsal çapta bir veri tabanından gelmesi veya referans alınması gereklidir (Kachur, 2000). Yazılan örnek simülasyon uygulamasında bu tablo(lar) “**correct data library(s)**” veya “**cross-reference table(s)**” olarak kullanılmıştır. Bu değerler her bir alan (field veya domain) için tek-tek kararlaştırılır.

9.6 EDQM Değeri Nasıl Ölçülür?

İşletmeler "Veri Ambarcılığı"nı verilerini düzenli ve birleşik bir yapıya dönüştürecek bir değer, bir uygulama olarak kabullenmiş bulunmaktadır. EDQM niçin değerlidir ve ne tip problemler çözülür? Belkide veri kalitesinin değerini gösterebilmenin en iyi yolu; kötü verinin köklerini gözlemlemek ve organizasyonu etkileyebilecek bozuk veriyi incelemektir. Bunlar:

1. Yanlışlar (Mistakes)
2. Eşsesliler (Homonyms)
3. Standard eksikliği (Lack of standards)
4. Yasal zorunluluklar (Legal entities)
5. Kayıp/gizli veri (Missing/invisible data)
6. Hayali veri (Phantom data)

9.6.1 Yanlışlar (Mistakes)

Verideki yanlışların orijinin anlaşılması bakımından en kolay problemlerdendir. Bunlar:

1. Yazılış hataları
2. Baskı hataları
3. Sınır değeri taşmaları veya
4. Uyumsuz veri tipleri

9.6.2 Eşsesliler (Homonyms)

Yazılışları aynı fakat, anlamları farklı olan kelimelerdir. İçinde eşsesli bulunan metinlerin yanlış yorumlanması verinin doğruluğu üzerinde ciddi bir etkiye sahip olabilir. Öneğin anonim şeklinde söylenen aşağıdaki Türkçe manide “yavrular” kelimesi bir eşsesli örneğidir.

Ay yavrular, yavrular

Yuvada kuş yavrular

Ellerin derdi biter

Benim derdim yavrular.

Veya İngilizceden bir örnek, “No” kelimesinin veri olarak girildiğini varsayarsak aşağıdaki gibi değişik anlamlar görülebilir:

1. Olumsuz bir cevap, yani evetin karşıtı
2. Kuzey (North) kelimesinin kısaltması
3. 1, 2 veya 35 gibi bildiğimiz numara kısaltması

9.6.3 Standart Eksikliği (Lack of Standards)

Veri girişi sorumlulukları değişik insanlar ve üniteler arasında dağıtıldığı zaman varyasyonların oluşması, tıpkı aşağıdaki envanterden toplanan bilgilerin görünümünde olduğu gibi, kaçınılmazdır. Ürün (Product) ve lokasyon (Location) alanları örnek alındığında PC ve Notebook ürünleri aşağıdaki gibi farklı şekillerde (standartsız) girilmiş olabilir.

<u>Ürün</u>	<u>Lokasyon</u>
PC Bin	Location 223
Personal Computer	Shelf 223
Laptop	On the Shelf
Notebook	Yes

9.6.4 Yasal Zorunluluklar (Legal Entities)

Bir çok örnekte olduğu gibi, isimlendirme şekillerine ekleme ve çıkartmalar eldeki dökümanın yasal tanımını etkiler. Bir çok bankacılık ve finans uygulamaları oldukça karışık, isimlendirme şekillerine sahiptirler, fakat dökümanın yasal zeminde kalması ve kanunlara, yönetmeliklere uyabilmesi için bu isimlendirmelere uyulma zorunluluğu vardır.

9.6.5 Kayıp/Gizli Veri (Missing/Invisible Data)

Genellikle mevcut veri, uygun bir yapıya ve kabul edilebilir değerlere sahiptir, aslında doğru da gözükür, fakat; tanımlama ve ilişkilendirme mekanizmaları tarafından tanınmadığı için zayıf, kalitesiz veri dağları oluşmaktadır. Bu problem genellikle organizasyonun bilgisi dışında gerçekleşir. Dikkat edilmez ise personel veya uygulama veriyi doğru zanneder. Örneğin, "Cumhuriyet Cad. Türkiye" adresini ele alalım. Her ilde; hatta ilçe veya semtte neredeyse bir Cumhuriyet Caddesi bulunur.

9.6.6 Hayali Veri (Phantom Data)

Bir çok uygulamada, hayali veri (örneğin 99/99/99) bulunduğu alanda (field) geçerli bir veri olarak bulunmadığı tesbit edilir. Aynı şekilde, bir "flag" koyarak bu alandaki değerlerin hiç bir şekilde alınamayacağını belirtmek gerekir. Burada hayali veriye örnek olarak verdiğimiz tarih değeri sorunlu bir kayıttır ve geçerli bir kayıt olarak kabul edilemez.

9.7 Veri Temizleme Araçları için Kritik Faktörler

Adaptasyon yeteneği ve organizasyon içinde bilgi akümüülüzasyonu; "tekerliği yeniden icat etme" kadar zor bir şey değildir ve veri temizleme araçları için kritik faktörlerdir. Veri temizleme araçları:

1. Bir uygulamadan diğerine taşınabilmedir (portable).
2. En küçüğünden en büyüğüne tüm uygulamalar için ölçeklenebilmelidir (scalable) ve
3. Organizasyonun tümünde yeniden-kullanılabilme yeteneğine, önceki bilgi ve kuralların üzerine bina edilebilme yeteneklerine sahip olmalıdır (reusable).
4. Kolay uygulanabilmelidir.
5. Mevcut uygulama iş prosesleri ile kolay entegre olabilmelidir ve
6. Bilgi ve kuralların kolaylıkla bakımının, izlenmesinin ve ayarlanmasını sağlanması gerekmektedir.

9.8 Veri Kalitesi ve Veri Temizleme Araçları

Proaktif veri kalitesi girişimleri "Veri Giriş" safhasında başlar. "Veri giriş doğrulaması", kötü veriye karşı "**birinci savunma hattı**" olarak veri sınırlarını kontrol eden ve veri giriş prosesi esnasında bütün zorunlu alanların girilmesini kontrol edenler yer alır. "Veri seçme ve meta data araçları" ise "**ikinci savunma hattı**"nı teşkil ederler.

Bir çok yeni sistemde doğrulama kontrolleri yerleşik hale gelmiştir. Daha yeni jenerasyon çözümleri önceki alanlara girilen değerler üzerine bina edilen ve kabul edilebilir değer sınırlamalarını daraltan oldukça sofistike "*durumsal mantık*" kullanan çözümler olarak karşımıza çıkmaktadır.

Veri giriş doğrulamaları için, yanlışları yine de düzeltebilir, diyemeyiz. **"Spell checker"** kullanan kelime işlemciler kelimelerin doğru yazılıp yazılmadığını mükemmel kontrol etse de **grammar hatalarını yakalayamazlar**. Aynı şekilde veri giriş personelleri doğru alanlara doğru format ve aralıkta, ancak yanlış veri girerlerse bu hata asla uygulama tarafından yakalanamayacaktır. Bu yüzden ki; bu araçların kurumsal standartlar ile birlikte kullanılma zorunluluğu vardır. Bu standartlar veri giriş ve reddetme için kesin kabul edilebilir mekanizmalara müdahale ederler.

9.9 Veri Temizleme Araçlarının Tarihçesi (Jenerasyonları)

Veri temizleme araçları(jenerasyonları):

1. Birinci jenerasyon (nesil) veri temizleme araçları,
2. İkinci jenerasyon (nesil) veri temizleme araçları; olmak üzere iki ana jenerasyona ayrılır.

9.9.1 Birinci Jenerasyon (Nesil) Veri Temizleme Araçları

Bu birinci nesil araçlar verinin kaynak sistemden hedef sisteme aktarılırken veri parçalama, hata bulma ve düzeltme rutinleri gibi toplu iş (batch) prosesleri içermektedirler. Bu araçlar hataların oluşmasını engelleme ve kaynakta iken düzeltme proseslerini gerçekleştirerek koruyucu hekimlik gibi proaktif bir rol üstlenmek yerine, sadece hata bulma ve bunları düzeltme yeteneklerine sahiptirler.

Bir çok yönüyle kurumsal veri kalitesi yönetimi (EDQM), Toplam Kalite Yönetimine (TQM- Total Quality Management, önde gelenleri **Deming, Crosby ve Jurand**) benzemektedir. TQM hataların oluşmasını önlemeye dayanırken, EDQM ise veri tutarsızlıklarının doğasını anlar ve organizasyonun klasik (legacy) alt yapısındaki kaynağına değişim mühendisliği tekniklerini uygular. Hem kurumsal çapta hata oluşumunu yakalama ve yığınak yapma, hem de bu tip araçların başarılı uygulamasının oluşmasını sağlayacaktır.

Birinci nesil veri temizleme araçlarının şu eksiklikleri vardır:

1. Bir çok birinci nesil araçlar platform spesifiktir.
2. Bir çok araç uygulama ve temizleme mantığı kodun içine gömülüdür.
3. Bir çok araç danışmanlık yardımına ihtiyaç duymakta ve özel yardımsız sonuca ulaşamama ile karşı karşıya kalınmaktadır.

9.9.2 İkinci Jenerasyon (Nesil) Veri Temizleme Araçları

EDQM'i destekleyen yeni nesil değişim mühendisliği araçları piyasaya sürüldü. Bu araçlar birinci nesil araçlardan aşağıdaki özellikleri nedeniyle ayrılmaktadır:

1. Çok yönlü ve güçlü
2. Taşınabilir ve platform bağımsız
3. Standart tabanlı
4. Genişletilebilir ve uyarlanabilir
5. Kullanım kolaylığı

Deming'e göre: "Gerçek kalite hataların tespitinden ziyade, onları elimine edecek proseslerin geliştirilmesi ile ortaya çıkar. "Tıpkı TQM" in yanlış ve fazladan çalışmanın olmamasına yoğunlaşması gibi, EDQM de kurumsal çapta her yerde veri kalitesinin sağlanma garantisine yoğunlaşır.

Eğer organizasyon "kural tabanlı yaklaşım" konseptini kurumsal veri kalitesine uyar ise ekstra faydalar sağlayacaktır. Kuralların görülebilir ve uyarlanabilir olması şu iki şeyi sağlar:

1. Kurum içinde ne verisinin bulunduğu kapsamlı ve açıklıkta bilinmesi,
2. Basitleştirilmiş bakım.

Veri temizleme ile ilişkilendirilmiş iş kuralları da değişmelidir. Kurumsal bir yaklaşım kuralların bir kere uygulanıp "n" kez geliştirilmesine (utilization) imkan sağlamalıdır.

9.10 Veri Kalitesi Garantileme

Toplam Kalite Yönetiminin (TQM) hata bulma ve iyileştirme prosesleri üzerine bina edilmiştir. Veri kalitesi garantileme servisleri (DQAS - Data Quality Assurance Services) bilgi varlığının kontrolü ve iş aklı alt yapısının optimum değerini elde etmeyi sağlar. Veri kalitesi garantileme servisleri (DQAS) şunlardır :

1. Geliştirme ekiplerine ve iş kullanıcılarına denetleme, izleme ve onaylama yetkileri vererek veri entegrasyon döngüsünün ana hatlarında problem oluşumunu engellemek.

2. Geniş kapsamlı veri kalite problemleri ve iş kuralları zorlamalarının sınırlamaları belirlenerek, veri aktarım faaliyetlerinin maliyet indirimleri ile işe katkısını sağlamak.
3. Toplam kalite yaklaşımı baz alınarak iş aklı alt yapısının güçlü, yapısal ve tekrarlanabilir proselere dönüşmesinin sağlanması.
4. Veri kalitesi yatırımının orginazisyona katkısının ölçülebilmesi için finansal dönüşüm ölçüm kriterlerinin belirlenmesi.

9.11 Veri Kalitesinin Gelir Üretimi ve Maliyet Yönetimi Üzerine Etkileri

Bir şirkette, alt seviyede etkilenen iki aktivite vardır: “Gelir-üretici girişimler” ve “Maliyet - yönetimi girişimleri”. Veri kalitesi her ikisinin de başarısı için kritiktir. Değişik yönleri aşağıda listelenmiştir, fakat tez konusu kapsamında olmadığından detaylarına inilmemiştir (Atkins, 2000):

1. Müşteri ve ürün kanadı
2. Satış kanadı
3. Gelir kanadı
4. Satınalma kanadı
5. Envanter tedarikçileri ve Lokasyonlar Kanadı

9.12 Veri Kalitesi : 2000 Amerika Seçimleri (Florida Eyaleti)

Florida seçimleri herkesi bir menteşe gibi aynı noktaya çakılı hale getirmişti, fakat istatistiksel boyutu ile de ön plana çıktı. Burada, 2000 yılı başkanlık seçimlerine profesyonel veri yönetimi perspektifi ile bakacağız (Forino, 2001).

Florida seçimlerinde 25 oy kadar dar bir aralık George Bush ve Al Gore için seçimi çok kritik bir noktaya getirdi. Aniden tüm Amerikan halkı, hiç bir toleransa oy sayımlarında mücade edilmemesini istedi. Beş uzun hafta boyunca ulusun dikkati veri kalitesinin üzerindeydi. Çok güzel bir örnek olduğundan oluşan problemler ele alınmış ve aşağıdaki gibi listelenmiştir.

Problem # 1: *Oylama makinaları pek güvenilir değildi.*

Newsweek dergisine göre; Amerika genelinde seçmenlerin % 34'ü oylamada punch-card sistemine güveniyordu.

Öneri / Uyarlama: Bütünleşik uygulama sistemlerimizdeki veri giriş sistemlerimiz güvenli mi?

Problem # 2: *Eksik / gelemeyen oyların, özellikle ülke dışındaki askerlerin oylarının sayımı yapılmadı.*

Öneri / Uyarlama: Uzaktan otomatik veya personelimiz tarafından girilen (örneğin mobil satış otomasyonu) verilerin kaliteli veri olduğunun garantisini verecek ölçme sistemlerine sahip miyiz ?

Problem # 3: *Oy sayımında şüphe görüldüğünde ne tip standartlar veya iş kuralları takip edilmelidir?*

Öneri / Uyarlama: Veri girişi yapan personel, konusunda iyi eğitildi mi? Veri girişinde oluşabilecek hataları önleyecek sistemleri kurulu mu? Veri giriş personeli zamana ve veriye dikkat ederek başarılı sonuçlar üretecek girişler yapıyor mu?

Sonunda, referans olarak Amerikan Anayasası (Article 2, Section 1) alınmıştır. Anormallikler burada öngörülmüş ve bunların çözümü için uygun mekanizmalar tasarlanmıştır. Görüldüğü üzere, veri kalitesinde proaktiflik kesinlikle bir zorunluluktur.

10. VERİ TEMİZLEME (DATA CLEANSING)

“Gerçek dünya verisi kalitesizdir” yaklaşımı ile veri temizleme konusu incelenecektir. Veri kalitesi problemlerinin sebepleri, temizleme prosesinin tüm aşamaları incelenecektir. Ayrıca bölüm sonunda veri temizleme araçlarının seçim kriterleri de çok boyutlu olarak ele alınacaktır.

10.1 Gerçek Dünya Verisi Kalitesizdir

Büyük ticari ve/veya devlet organizasyonları için, veritabanlarındaki, bilgiyi birleştirmek ve ilişkilendirmek oldukça zor bir problemdir. Bu tip büyük veri ambarları, daima birden fazla tekrar eden kayıtlar içerir. Hatta aynı veri tüm kaynaklarda mevcut ise kaynak sayısı kadar tekrar etmesi demektir ki, n tane versiyonun tek ve doğru olacak olana dönüştürülmesi gereklidir.

Veri ambarı içerisindeki veri genel olarak onaylanmıştır, fakat ambardaki verinin temizliği görmezden gelinmiştir (Kimball, 1996). Veri bütünlüğü çözümünün planlanması yapılırken oluşan planlama eksikliğinin sebeplerinden birisi, bilgi işlem elemanlarının kalitesiz verinin işe etkisini yeterince önemsememelerindendir.

Temel problem farklı kaynaklardan gelen verinin farklı olması veya en basit şekliyle hatalı olmasıdır. Ana kalite problemi gerçek nesnenin farklı veri kayıtları tarafından modellenmesidir. Bir kaç faktörden oluşan bu probleme “*Nesne Belirleme Problemi (Object Identity Problem)*” denir (Galhardas, Florescu, Shasha ve Simon, 2000).

1. İlk olarak, veri *hatalar* içerebilir. Genellikle “John Smith” in “Joh Smith” olarak yazılmasında olduğu gibi eksik yazma ile bu problem oluşur.
2. İkinci olarak, veri farklı kaynaklardan geldiği zaman, farklı konvansiyonel isimlendirme kullanılmış olabilir. Örneğin, aynı müşteri farklı tablolarda “John Smith”, “Smith John” veya “J.Smith” gibi doğru, fakat çok az farklı versiyonlarıyla kaydedilmiş olabilir.

Ortak bilginin belirlenmesi ve veri setini temizleme/ayıklama prosesi “**Veri Temizleme (Data Cleansing)**” olarak adlandırılır (Electronic Digital Documents, Inc., 2001). Değişik endüstriler bu problemi şu terminolojiler ile kullanmaktadır : Birleştirme/Silme,

Duplikasyonu engelleme ve Kayıt ilişkilendirme (**Merge/Purge, De-Dupe and Record Linking**). Bu isimlendirmeler IT terminolojisinde kullanılsa da, asıl terminoloji “veri temizleme” dir.

Nesne Belirleme Probleminin düzeltilmesi, **veri temizleme araçları** (data cleaning tools) olarak adlandırılan yazılımlar seti ile garantilenebilir. Veri kalitesi meseleleri üç farklı bağlamda ortaya çıkar (Galhardas, Florescu, Shasha, Simon, Matsumoto ve Saita, 2001).

1. Herhangi bir kimse anormallikleri tek kaynaktan düzeltmeye kalkarsa (örneğin, dosyadan duplike kayıtların elimine etmeye çalışması durumunda), veya
2. Kötü yapılandırılmış veya yapılandırılmamış veri yapılandırılmış veri şekline dönüştürken
3. Herhangi birisi birden fazla kaynaktan gelen veriyi yeni tek bir kaynağa katmak isterken (örneğin, bir veri ambarı oluşturma)

10.2 Müşteri Verisindeki Problem

Bir şirket için belirlenmiş müşteri listesi genelde %90 oranında doğrudur. %10’ luk müşteri bilgisi hatalıdır. Bunun %5 i (toplam verinin % 0.5’i) düzeltilebilir adresler içerir. Müşterilerine önem veren şirketler, bu bilgilerin yüksek kaliteli olmasına da önem verirler. Fakat bir çok firma, tamamlanmamış, hatalı veya gereksiz müşteri bilgilerinden aşağıdaki sebeplerden dolayı sürekli problemler yaşar (Innovative Systems, Inc., 2001).

1. Müşterilerin isim ve adreslerinin değişik varyasyonları olabilir.
2. Veri girişi esnasında harfleri yanlış girme olabilir.
3. Bilgi yanlış alanlara girilmiş olabilir.
4. Firmanın değişik alanları farklı veri giriş standartlarına sahip olabilir.

Kusurlu veriyi silme zor olabilir, çünkü halledilmesi gereken bir çok olasılık ve doğru olmayan hatalı öğeler vardır. Bunlar:

1. İsimler tanımsız/geçersiz olabilir.
2. Aynı şey için farklı kısaltmalar kullanılmış olabilir. (örneğin, Av. veya Ave.)
3. Konu dışı terminoloji dahil edilmiş olabilir.
4. Adres veya posta kodu verisi yanlış olabilir.

10.3 Veri Temizleme Genel Prosesi

Tüm bu öğeleri hatasız bir şekilde tanımlamak ve standartlaştırmak ad ve adres kelimelerini ve örneklerini içeren kapsamlı bir “**veri kalite sözlüğü**”nden - *otomatik* - geçirmekle olur. Bir kez hatalı örnek (flawed pattern) tespit edildiğinde otomatik olarak standartlaştırılabilir ve “düzeltildi” - *veya ciddi hatalarda* - “tekrar incelenecek” şeklinde işaretlenebilir.

En basit veri temizleme şekilleri “*Duplike Ayıklama Sıralaması (Duplicate Elimination Sort)*” hesaplaması veya “*her kayıtın tamamen diğerleri ile karşılaştırılması*”dır. Bu iki strateji, var olan verinin toplam sıralamasının yapılmasını ve verinin bozulmamasını sağlar. Gerçek dünyadaki bir çok veri kümesinde, hem toplam sıralama prosesinin, hem de basit duplike yakalama prosesinin varlığı söz konusu değildir. Bu proses kabul edilebilir küçük hataların bulunduğu yani ortak veri çakışmalarının bulunmadığı bir veri kümesi için söz konusudur.

Örneğin, iki kaydın eşitliği *kurallar kümesinin* yerine, basitçe aritmetik eşitlik olarak tanımlanabilir. İki farklı kaydın (muhtemelen farklı iki veritabanından) aynı nesneyi ifade ettiği bilgisi *kullanıcı-tanımlı kurallar* ile belirlenir.

Veri Temizleyici (Data Cleanser - bu tipdeki ürünlerin genel adı), öncelikle tüm veri kümesini veriden türetilen kullanıcı tarafından tanımlanmış anahtarlar ile liste içerisindeki potansiyel kayıt eşlemelerini birbirine yakınlaştırmak amacı ile sıralar. Özenli ve sıkı çalışmalar ile, sonuca ulaşmayı hedefleyen farklı anahtarları içeren bağımsız sıralamaların *tek seviyeli geçiş (one-pass)* izin vermeyen sonuçların oluşmasını sağladığı görülmüştür. *Çoklu geçiş (multiple passes)* sonrasında, “*transitive closure*” hesaplaması sürekli bir doğruluk için etkinlik sağlanabilir.

Bundan çıkarılacak sonuç; **birçok tekil “ucuz” veri geçiş prosedürleri, tek bir pahalı veri geçişinden daha doğru sonuçlar yaratır.** Bu da iterasyon demektir.

10.4 Basit Veri Silme Metodu

İki veya daha fazla veritabanı verilmiş olsun, öncelikle onları ardışıl (sequential) kayıt listesi şeklinde birleştiririz ve *sıralanmış-komşular metodunu (sorted-neighborhood method)*u uyguluyoruz. Veri temizleme için sıralanmış-komşular metodu 3 aşamada özetlenebilir.

1. Anahtar oluşturma: Her kayıt için ilgili alanlardan veya alanların bir kısmından türetilerek anahtar oluşturulur.

2. **Veriyi sırala:** Veri listesi bu anahtar kullanılarak sıralanır.
3. **Birleştir (Merge):** Sabit-büyüküklüklü bir pencere (fixed-sized window) bu liste içerisinde hareket ettirilerek, uyan kayıtlar limitli bir bölgede mukayese edilir.

Şimdi kullanıcının aklında kalabilecek kural tipleri için aşağıdaki gibi örnek ele alalım.

İki ismin hemen hemen aynı şekilde yazıldığını varsayalım (ama tam aynı değil), ve bunlar aynı adrese sahip olsunlar. Onların aynı kişi olduğunu düşünelim. Diğer yandan, iki kaydın da aynı sosyal güvenlik numarasına (SSN - Social Security Number) sahip olduğunu düşünelim, fakat isim ve adresler tamamen farklı olsun. Adını değiştiren veya taşınan kişinin aynı kişi olduğunu düşünebilirdik veya kayıt farklı kişileri simgeler diyebilirdik. Ve sosyal güvenlik numarasının birisi için yanlış olduğunu düşünebilirdik. Başka extra bilgi olmaksızın ikincisini kabul edebiliriz.

Ne kadar fazla bilgi varsa, kayıtlardan o kadar çok çıkarımlar yapmak mümkündür. Örneğin, “Michael Smith” ve “Michele Smith” aynı adrese sahip olabilirdi ve onların adları makul bir şekilde yakın olabilirdi. Eğer cinsiyet ve yaş bilgisi varsa “Michael” ve “Michele”in evli veya kardeş olduğunu anlayabilirdik.

Doğrudan pazarlama için adreslemede yaygın bir şekilde bulunabilen hatalı kayıtlar aşağıdaki çizelgede (Çizelge 10.1) örnek olarak gösterilmektedir. Tablodaki ardışıl kayıt çiftleri **kural-tabanlı (rule-based)** bir eşitlik tarafından belirlenebilir.

Çizelge 10.1 Veri Temizleyici (DataCleanser) için örnek eşleştirme kayıtları (Electronic Digital Documents, Inc., 2001, <http://www.npsa.com/edd/>)

ID	Name (First, Initial, Last)	Address
334600443	Lisa Boardman	144 Wars St
334600443	Lisa Brown	144 Ward St.
525520001	Ramon Bonilla	38 Ward St.
525250001	Raymond Bonilla	38 Ward St.
0	Diana D. Ambrosion	40 Brik Church Av.
0	Diana A. Dambrosion	40 Brick Church Av.
0	Colette Johnen	600 113 th St. apt. 5a5
0	John Colette	600 113 th St. ap. 585
850982319	Ivette A Keegan	23 Florida Av.
950982319	Yvette A Kegan	23 Florida St.

10.5 Anlamlandırılmış Bilgi Tabanı (Knowledge Base)

Piyasadaki veri temizleme araçları genellikle 1.000.000’den fazla ad ve adres ilişkili kelime ve 100.000’den fazla kelime kalıbı içerir. Bu dev bir bilgi makinesi gibidir ve bu bilgiyi toplamak da çok zordur. Tüm tedarikçiler; hem ürün dökümanlarından, broşürlerinden hem de web sitelerinden anlaşıldığı üzere oldukça farklı şirketlerden veya ülkelerden 20-30 senede ancak bu kadar bilgiyi toplamışlar ve referans olarak kullanmak üzere bir “Anlamlandırılmış Bilgi Tabanı (Knowledge Base)” oluşturmuşlardır.

10.6 Veri Temizleme Bilimi

Veri temizleme bir çok şekilde olabildiği gibi, en önemli veri-temizleme örnekleri, önceki uygulama listelerindeki ürün tanımlaması ihtiyaçları müşteri, ürün, prosedür ve tanımlamaları gibi somut nesnelerden ortaya çıkmaktadır. Veri temizleme, kayıdı doğru veri ile güncellemekten çok daha öte bir anlam ifade eder, yani basit bir proses değildir. Ciddi veri temizlemeleri ayrıştırma (decomposing) ve tekrar birleştirme (reassembling) işlemlerini içerir.

Temizlemeyi altı adıma bölebiliriz:

1. Ögeleştirme (Elementizing),
2. Standartlaştırma (Standardizing),
3. Doğrulama (Verifying),
4. Eşleme (Matching),
5. Ev sahipliği (Householding),
6. Belgeleme (Documenting).

Bu altı adımı göstermek için aşağıdaki örnek adresi ele alalım.

Ralph B and Julianne Kimball Trustees for Kimball Fred C
Ste. 116
13150 Hiway 9
Box 1234 Boulder Crk
Colo 95006

Bu adres belki Address_1'den Address_5'e kadar beş değişik alanına girilmiş olabilir. Adreslerin parçalarının güvenilir bir sıralaması olmayabilir. Adresin önemli bir yeri yanlış girilmiş de olabilir.

Adım 1. Öğeleştirme (Elementizing): İlk adım bu adresi öğeleştirmek için temizlemektir: Öğeleştirme sonucu şu şekilde üretebilir:

Alıcı ilk adı	Addressee First Name(1):Ralph
Alıcı ikinci adı	Addressee Middle Initial(1):: B
Alıcı soyadı	Addressee Last Name(1): Kimball
Alıcı ilk adı	Addressee First Name(2): Julianne
Alıcı soyadı	Addressee Last Name(2):Kimball
Alıcı akrabalığı	Addressee Relationship: Trustees for
Akraba kişinin ilk adı	Relationship Person First Name: Fred
Akraba kişinin ikinci adı	Relationship Person Middle Name: C
Akraba kişinin soyadı	Relationship Person Last Name: Kimball
Sokak adres no	Street Address Number:13150
Sokak adı	Street Name: Hiway 9
Daire no	Suite Number: 116
Posta kutusu no	Post Office Box Number: 1234
Şehir	City: Boulder Crk
Eyalet	State: Colo
5 hanelik alan kodu	Five Digit Zip: 95006

Bu standart öğeler, adres analiz edildiğinde bulunanlara bağlıdır. Ralph ve Julianne, bazı öğelerin farklı tipte olması durumunda, özel olmaktan daha ziyade kurumsal da olabilirlerdi ve de farklı olurlardı.

Adım 2. Standartlaştırma (Standardizing): 2. adım öğelerin standartlaştırmasıdır. En azından 4 öge standart form içerisine konulmak zorundadır. Biz zaten 'Ste' yi çıkarmıştık. Çünkü onu 'suite' olarak onayladık. 'Hiway 9'un 'Highway 9' olabileceğinden dolayı bu geçici değişikliği yapmak, ve gerçek sokak adının 'Highway 9' olduğundan şüphelenebiliriz. Aynı zamanda 'Boulder Crk'i 'Boulde Creek' ve 'Colo' yu 'Colorado' olarak değiştirebiliriz.

Adım 3. Doğrulama (Verifying): 3. adım tutarlılığın ve ögelerin standartlaşmasının kontrol edildiği adımdır. Diğer bir deyişle, içerikte herhangi bir hata var mı? Açıkça olmamasına rağmen, adresin içinde göze batan içerik hataları vardır. 95006 Posta kodu içerisindeki Boulder Creek California’ dadır. Eyalet ismi California olarak değiştirilmelidir. California için verilmiş olan veri noktalarının üçünün ikisi eyalet ismini California’ ya değiştirmiştir. Bu kayıt muhtemelen gelecekteki kontrol için işaretlenmelidir. Örneğin, eğer “Ralph Kimball” veya “Julianne Kimball” için başka bir adres izine rastlanırsa, doğru eyalet kontrol edilebilir. Bu işlem Colaroda’ daki Boulder Creek’ in mantıklı bir şekilde keşfedilmesinden daha önemlidir. Şu anda ögeleştirdik, standartlaştırdık ve adres doğrulama kontrolünü yaptık, 5. ve 6. adımları gerçekleştirmek için iyi pozisyondayız.

Adım 4. Eşleme (Matching): Eşleme ya “Ralph Kimball” veya “Julianne Kimball”ı diğer müşteri kayıtlarından bulmakla ya da adresin tüm ögelerinin hemen hemen aynı olmasından emin olunması sonucu gerçekleşebilir. Değiştirilmiş adresler problemini farklı öge tiplerine “önceki” ve “aktif” şeklinde bir atama yapmak suretiyle çözmüş bulunmaktayız.

Adım 5. Ev sahipliği (Householding): Ev sahipliği “Ralph” ve “Julianne”nin evsahipliği ilişkisini kurar. Çünkü onlar aynı adresteler. Büyük bir binada, farklı dairelerde yaşayan insanları çıkarmak için dikkatli olmak zorundayız. “Ralph” ve “Julianne”nin evli olduğunu anlamak için daha başka dahili ve harici veri kaynaklarına sahip olmalıyız.

Adım 6. Belgeleme (Documenting): Veri temizlemenin 6. adımı metadata içerisindeki ögeleştirme, standartlaştırma, doğrulama, eşleme ve ev sahipliğinin sonuçlarının belgelenmesinden oluşur. Bu, sonradan yapılan temizleme olaylarında adreslerin daha iyi anlaşılmasına ve son kullanıcı uygulamalarının daha iyi segmentlenmesine ve müşteri verisinin daha iyi anlaşılmasına yardım eder.

Veri ambarı iyi veri değerlerini görmek için mükemmel bir yerdir. Veri temizleme araçları bilgi işlem elemanlarına temiz veri içerisindeki önemli noktaları açar. Güçlü veri temizleme sistemleri veriyi %100 oranında doğru bir şekilde belirlemeye çalışırlar.

10.7 İsim ve Adres Verisi Temizleme Analizi

Söz gelimi, Amerika (US) sokak adreslerini aşağıda verilen 7 öğeden tanımlanabilir:

Sokak no	Yön öneki	Sokak adı	Sokak tipi	Yön soneki	Ek isim	Dahili numara
123	N.	Main	Street	East	Suite	10

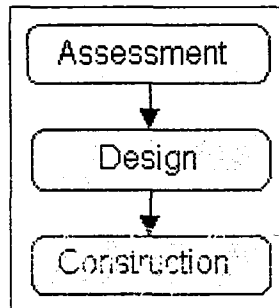
Basit bir adres, bu öğelerin hiç birini veya tümünü içerebilir. Örneğin : “123 N.Main Street East Suite 10” da olduğu gibi. Diğer yandan özel isim aşağıda görüldüğü üzere tanımlanabilir.

Önek	İlk ad	İkinci ad	Soyad	Sonek	Ek
Mr.	John	B.	Smith	Jr.	Attorney

Veri, iç kütüphanelerde (data library) saklanır. Örneğin, NAME PREFIX (Ad öneki) kütüphanesi sadece “Mr”. ve “Mister” değerlerini değil aynı zamanda bu öneklerin erkelere ait olduğunu da tutar. Her ikisinin de standart versiyonu “MR.” dir. NAME (Ad) kütüphanesi; isim bilgilerini, ilk ad ve de tahmini bir cinsiyet bilgisi ile birlikte içerir. STREET TYPE (Sokak tipleri) kütüphanesi, Amerikadaki tüm kabul edilmiş cadde tiplerini içerir. Posta servisleri (Court, Parkway, Street, Avenue) ve standartlaştırılmış versiyonları (CT, PKWY, ST, AVE ve BLVD) saklanır. Veri, veri-spesifik kurallar ile ipuçlarını belirleme ve anlamak için kullanılır.

10.8 Diğer Veri Kalitesi ve Temizleme Metodolojisi

Bu veri kalitesi temizleme metodolojisi (SAS Corp., 2001) aşağıdaki şekildeki aşamaları içerir. (Şekil 10.1):



Şekil 10.1 Veri nitelikleri temizleme metodolojisi tasarım aşamaları (SAS Corp., 2001)

Aşağıdakiler safhanın önemli yönlerini özetlemektedir..

- **Değerleme:** Bu aşama iş gereksinimleri hakkında araştırma bilgisi için kritiktir. Projenin oluşturulmuş iş analistleri takımı, ambar yöneticisi, veri güvenliği temsilcisi ve yönetim tarafından, özel veri ve amaçlar için anahtar iş ihtiyaçlarının belirlenmesi için tartışılması faydalıdır. Bu takım içerisinde projenin faaliyet alanı, ihtiyaç duyulan doğruluk ve uygulamalar belirlenir. Takım, verinin nasıl toplanacağını gözden geçirmelidir. Bu, nasıl tamamlanacağına veya başlanacak tutarsız veri kaynağına yardımcı olur. Veri büyük bir veri giriş takımı tarafından mı toplandı? Veri dünya çapında erişen kullanıcılar tarafından mı sunuldu? Veri belirli hedeflenmiş amaçlar için mi kullanılacak? Veya veri gelecekte şirket temellerinin ilk adımı mı olacak? Veri özel demografiyi hedefleyen doğrudan pazarlama kampanyası için mi kullanacak?
- **Tasarım:** Bu aşamada iş ihtiyaçlarına hitab eden veri tabanlarını, tabloları ve alanları belirleyen değerlendirme sonuçları kullanılır. Adres verisi, ad bilgisi, organizasyon bilgisi? Bunların kombinasyonları? Proje yöneticisi, ambar yöneticisi ve veri kalite uzmanı kullanılacak veri tabanını, tabloları ve alanları belirleyen stratejiyi geliştirirler. Onlar öz bilgi için (tüm bu araştırma sonuçları değerlendirme aşamasında oluşmuştur) odak noktasını ortaya çıkarırlar.
- **İnşa :** İnşa aşamasında, şu iki rol ile insanlar tanıştırılmalıdır. Veri kalitesi uzmanı ve ambar yöneticisi. Veri kalite uzmanı iş birimini bir üyesi iken, ambar yöneticisi ise bilgi işlem organizasyonun bir üyesidir.

10.9. Veri Temizleme Araçları Seçim Kriterleri

Veri temizliği, endüstriyel etkinlikte veri kalitesi yönetimi sonuçları ile, değişim ve yeniden yapılanma ile aynı anlama gelmektedir. Bir çok Fortune 1000 şirketi, müşterilerinin bütünleşmiş veri görüntüsüne erişebilmek için, var olan operasyonel sistemleri ve eski sistemleri içinde yöntemleri araştırmaktadırlar.

“En iyi deneme” için yapılan araştırmalarda bir çok organizasyon çabuk çözüm için sadece veri temizlemenin aşırı derecede kompleks ve endüstriyel açıdan geçerli sonuçlarının olması

gerektiğini keşfetmişlerdir. Aşağıdaki liste veri temizleme çözümlerinin seçimi için tavsiyeleri içermektedir.

1. Veri temizleme prosesi belirli projelerden bağımsız olarak seçilir. Özel projeler için geliştirilen kurallar ve iş mantıkları yeniden kullanılabilir. Bu veri kalitesi yönetiminin geniş anlamli yatırımlarının geliştirilmesine izin verecektir.
2. Prosesleri seçerken veri girişi anında temizliği yapmasına önem vermek gerekir. On-line uygulamalar, klasik sistemler ve operasyonel veri içersinde tüm veri tipleri araştırılmalıdır. Verinin kalitesini tesbit etmeye, verinin bozulmasını engellemeye ve var olan verinin temizlenmesine olanak sağlar.
3. Hem müşteri bilgisini hem de belirsiz adres verisini temizlemeye izin veren proseslere sahip temizleme araçları seçilmelidir. Proses, karışık örnekleri aynı isimden ve adres verisinden türetecek şekilde sofistike olmalıdır. İdeal çözüm, içerik duyarlı olmalı ve tüm veri tiplerinin analizini sağlayabilecek araştırma yeteneklerine sahip olmalıdır.
4. Ortam bağımsız bir proses seçilir. Gelecekteki veri kalitesi yönetimi, taşınabilirliği ve ölçeklenebilirliğine mevcut alt yapı sisteminin geleceği için ihtiyaç duyabilir. Eğer öğrenilenler başka ortamlarda tekrar kullanabilirse, kurumsal çapta veri temizleme prosesi geliştirebilecek yeteneğe sahip olunmuş demektir.

11. ETL (SEÇME DÖNÜŞTÜRME YÜKLEME)

Bu bölümde verinin, veri ambarına aktarılma prosesi ve bu prosesi en güzel şekilde tanımlayan ETL (Extract Transform Load) konusu ele alınmış ve ETL araçları incelenmiştir.

11.1 Bir BI Ortamında Kullanılan Veri Toplama Teknikleri ve Anahatları

Aşağıdaki adımlar (Bolten, 2000), veri ambarına veri toplama için kullanılan bazı teknikleri ve anahatları göstermektedir.

Adım 1: Eşleme Spesifikasyonunu Tanımla. Bu veri elemanları ve formatlarının detaylı olarak kaynak ve hedef tanımlarını içermelidir.

Adım 2: Iterasyon Konu Alanlarını Belirle. Bu ortamında mantıksal izolasyon ile tek bir veri akışı içine alınan veri elemanı gruplarına “iterasyon konu alanları” denir. Bunlar; kaynaklar ve hedefler arasında, ayrıca veri ambarları ve data martların kendi aralarında veri toplama için belirlenen alanların tümüdür.

Adım 3: Veri Toplama ETL Prosesini Tasarla. En üst seviyede, veri toplama prosesini tanımlamak için üç ana proses vardır. Bunlar; seçme, dönüşme ve yükleme. Her bir prosesin içinde, kesinlikle belirli fonksiyonların icra edilmesi gerekmektedir.

Bu ETL proseslerinin ihtiyaçlarını karşılamak amacıyla kullanılan tekrar-kullanılabilir fonksiyonların listesi Çizelge 11.1’de görülmektedir.

Çizelge 11.1 ETL Fonksiyonları Tablosu (Bolten, 2000, DM Review September 2000, s:22)

Seçme	Kaynak sistemden veriyi alma
Filtreleme	Gerekli veriyi filtreleme
Veri entegrasyon beslemesi	Aynı kaynak sistemden farklı veri tabloları ve dosyalarının entegrasyonu
Veri çatallaşması tesbiti	Kaynak sistemdeki değişikliklerin belirlenmesi ve veri ambarı için uygun olup olmadığına karar verilmesi
Veri transferi	Kaynak sistem ve veri ambarı ortamları arasında veri transferi
Veri entegrasyonu çapraz-besleyici	Farklı kaynak sistemlerden farklı dosyalar ve farklı tabloların entegrasyonu
Temizleme	Veriyi temizleme
Veri kalitesini düzeltme	Kalite ölçme ve düzeltme metodlarının veriye uygulanması
Zenginleştirme	Verinin anlamını bazı ilavelerle zenginleştirme
Referans oluşturma	Veri ve ilgili tablo arasında referans yaratma
Kur dönüşümü	Standart (ulusal) bir para birimine dönüştürme
Algoritmik fonksiyon	Veri üzerinde spesifik bir algoritmik fonksiyon çalıştırma
İnşa etme	Hedef kaydı yaratma
Yetki kontrol	Kaynak ve hedef yetki kontrollerini yapma
Yükleme	Veri ambarı tablosuna verinin transfer edilmesi
Yükleme-sonrası	Yükleme sonrası aktivitelerinin gerçekleştirilmesi (geçici tabloların silinmesi, indekslerin yeniden oluşturulması, vb.)
Arşivleme	Harici arşiv dosyasının oluşturulması
Yedekleme	Yedekleme dosyasının yaratılması
Meta data ve istatistikler	Meta data verisinin ve iterasyon konu alanlarının çalışmasını gözeten istatistiklerin güncellenmesi

Adım 4: Iterasyonu Uygula: Tasarım aşamasındaki fonksiyonlar şimdi ETL aracının içine transfer edilebilir.

Veri yüklenmesinin otomatik hale getirilmesi için, bir iterasyonun uygulanışı bir planlama mekanizması içine yerleştirilmelidir.

Bu yaklaşım, BI ortamları için yaratılan veri toplama alt sistemlerini destekleyecektir. Ayrıca gerçek zamanlı (örneğin, web-tabanlı) veri toplama içinde gerekebilir. Çok yoğun, gerçek-

zamanlı veriler için yeni ihtiyaçlar gerekebilir. Bu yapısal yaklaşımın, tasarım ve geliştirme için veri toplama alt sistemlerinde kullanılması oldukça yararlı olacaktır.

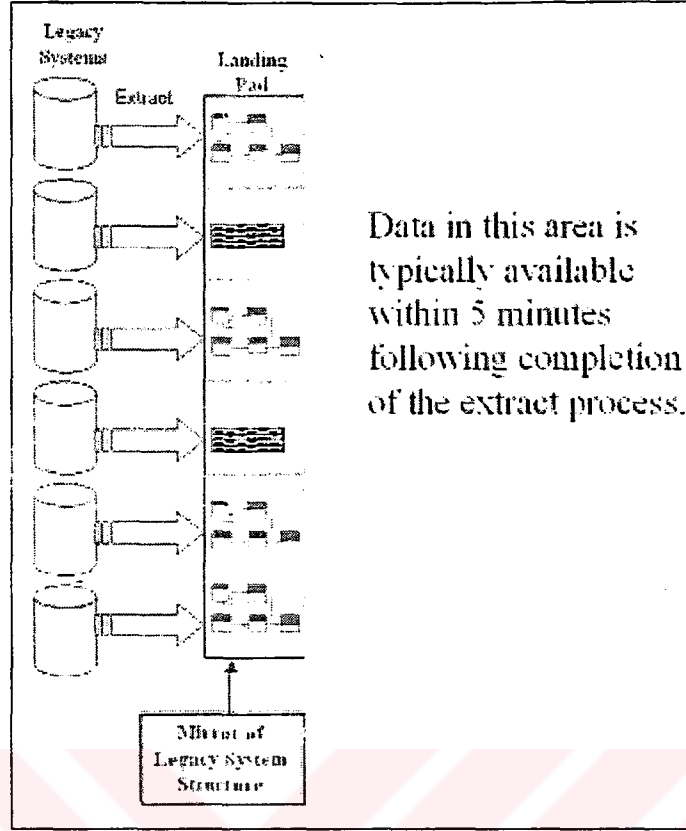
11.2 Veri Ambarı İşlemleri ve ETL İşleme

Birleşmiş ihtiyaç planlama (JRP) ve birleşmiş uygulama tasarımı (JAD) safhaları boyunca iş kuralları ve bunları destekleyecek veri elemanları belirlenir. Bill Inmon'un metodolojisi kullanılarak normalize edilmiş bir veri ambarından data martlar geliştirilebilir (Borland, R.,2000a). Normalize bir ambar inşa edebilmek için; tam ve doğru bir mantıksal model yaratılmalı, peşinden fiziksel modele çevrilmeli ve tasarımı ile proje uygulama planı oluşturulmalıdır. Böylece giderek büyüyen veri ambarı inşa etmiş oluruz. Aslında bir veri ambarı projesinde ana komponent proje ekibinin kendisidir. Ancak; ekip yetenekli, bilgili ve kendini sırf projeye adayacak profesyonellerden oluşturulmalıdır.

Önerilen bu modelde, veri ambarı geliştirme ve operasyonlarını sıralı ve fazlı bir mantıkla adresliyoruz. Klasik sistemlerden kullanıcı-erişebilirliği olan verinin taşınması adımı, ilk adımdır ve bu ETL'in "E"sidir. (İhtiyaçları tam karşılayan bir yazılım aracı setinin seçilmiş ve kullanıma hazır olduğunu varsayıyoruz.)

11.2.1 Klasik Sistemlerden İndirme Bölgesine (Landing Pad)

Şekil 11.1'de gösterildiği üzere, verinin özellikle klasik sistemlerden veri indirme bölgesindeki sunucularla taşınması önerilmektedir. Her bir klasik sistem verisi ayrı bir veri tabanına yüklenmeli ve her veri tabanının veri yapısı klasik sisteminkiyle tamamen aynı olmalıdır. Eğer kaynak sistem RDBMS ise ilişkisel model kullanılarak bu yapı oluşturulabilir, eğer RDBMS değil ise örneğin sıralı (sequential) bir yapı söz konusu ise indirme bölgesinin yapısı da aynı olmalıdır. Özetle, indirme bölgesi klasik sistemin bir kopyasıdır (mirroring).



Şekil 11.1 Klasik sistemlerden indirme bölgesine (Landing Pad)
(Borland, <http://www.datawarehouse.com/iknowledge>)

İndirme bölgesindeki veri, son kullanıcının erişimine açılmaz, sadece veri ambarı ekibinin kullanımına rezerve edilmiştir. Seçme prosesi, aşağıdaki ölçümleri (Çizelge 11.2) kalite kontrol amaçları için üretmelidir:

Çizelge 11.2 “Klasik sistem-İndirme bölgesi” kalite ölçüm kriterleri

Kayıt sayısı	Klasik kayıtlar = 3000 kayıt, Extracted rows = 3000 kayıt
Toplam byte sayısı	Seçilenlerin boyutu klasik sisteminkine eşit olmalı
Seçilen kolonlar için özet değerler (hesap bakiyesi, alınan parçalar vb.)	Bu değerler “view”lar yaratılarak seçilmiş verilerden bulunur (eğer SQL tabanlı bir veri tabanı mevcut ise).
Kayıt edilmeyen kayıtlar	Doğrulamadan geçirilmelidir.
Seçme ve yükleme zamanı	Performans için.

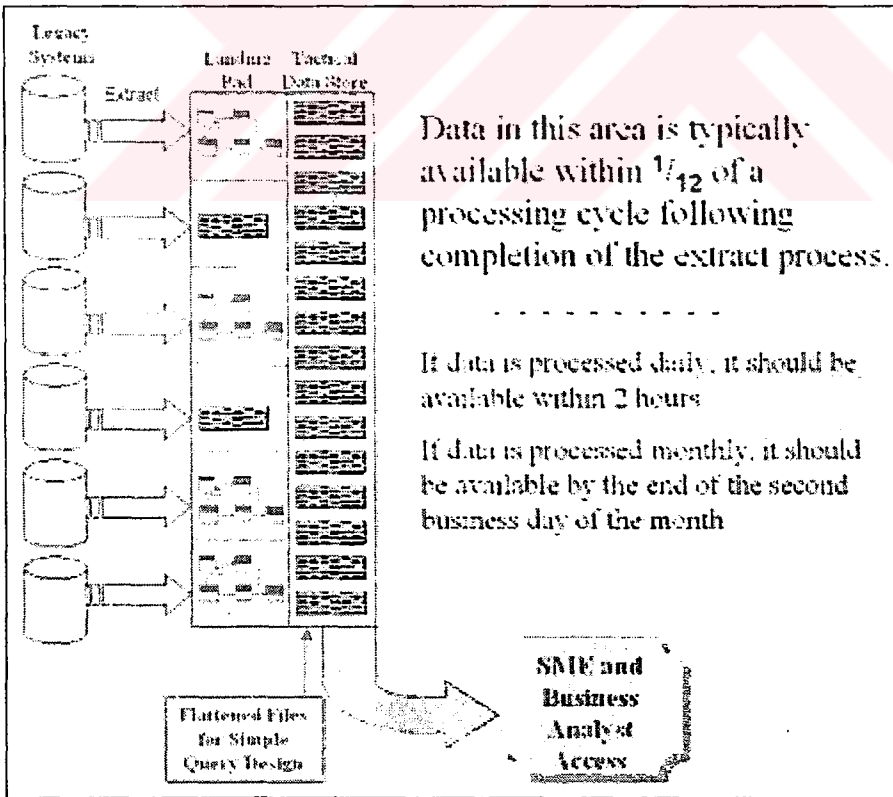
Tipik olarak eğer kayıt sayısı ve toplam değerleri her bir klasik sistem için yüzde 0.25 hatalı ise (%99.75 başarılı) yükleme prosesini başarılı sayabiliriz.

11.2.2 İndirme Bölgesinden Taktiksel Veri Deposuna

Bu adım verinin indirme bölgesinden taktiksel veri bölgesine taşınmasını gerçekleştirir. Bu bölge son kullanıcının erişim sağlayabileceği birinci seviyedir (Borland, 2000b). Şekil 11.2'de de görüldüğü taktiksel veri depoları düz dosyalar içerir. Bu dosyaların kullanılma amacı; hızlı sorgulama, hızlı sıralama ve kolay dönüştürme fonksiyonlarını kolayca destekleyebilmeleridir.

TDS (Taktiksel Veri Deposu) hazır olduğunda, bu bölgedeki veri artık konu alan uzmanlarına (SME) ve iş analistlerine hizmet verecek hale gelmiş demektir. İndirme bölgesinden TDS'e aktarma operasyonu toplam işlem döngüsünün 1/12 sine (8.3 %) denk gelecek bir şekilde bitirilmiş olmalıdır. Bunun manası, her gün veri yükleme işlemi yapılıyorsa veri indirme bölgesine veri geldikten sonra en geç 2 saat içinde ($1/12 \times 24$ saat = 2 saat) taktiksel veri deposuna aktarılmış olması gerekmektedir. Eğer yükleme işlemi aylık yapılıyorsa, gelecek ayın ikinci iş gününün sonunda hazır olması gerekmektedir.

$$(1/12) * 24 = 2 \text{ saat günlük, } [(1/12) * 24] * 30 = 60 \text{ saat} = 2.5 \text{ gün}$$



Şekil 11.2 İndirme bölgesinden Taktiksel Veri Deposuna
(Borland, <http://www.datawarehouse.com/iknowledge>)

Taktiksel veri deposu ayrıca bazı birleştirilmiş dosya yaratma, SME ekibine klasik sistemden ayrı bir şekilde daha geniş bir bakış açısı yakalama imkanı sağlar.

11.2.3 Kurumsal Veri Ambarı Geliştirme

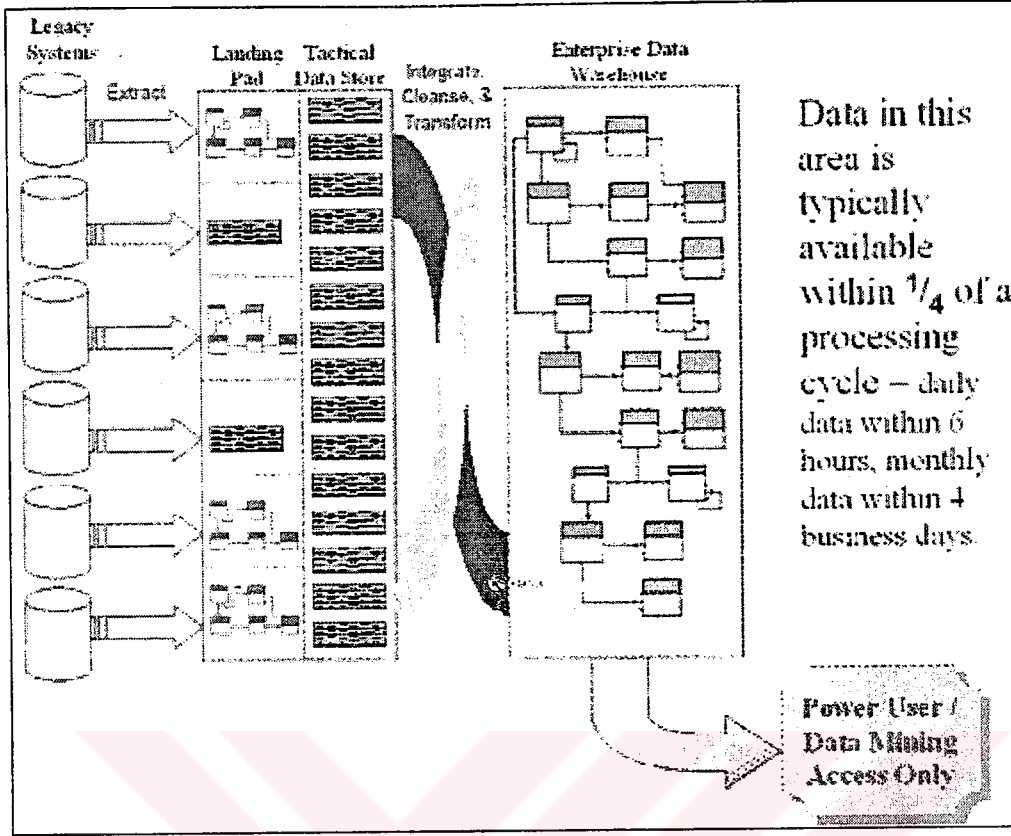
Şimdi de projenin gerçek çalışma dünyasına getirilmesine, yani başlangıçta inşa edilmesi planlanan normalize veri ambarının oluşturulması işleminin nasıl olacağına bakacağız. İlk etapta iki senaryo ihtimali oluşur: (Borland, 2000c)

Birinci senaryo: Kullanıcı grubu elde ettiği bilgiden heyecan duyacak ve klasik sistemlerden, operasyonel veri depolarından, taktiksel veri depolarından vb. veri erişim noktalarından daha geniş kapsamda ihtiyaçlar bildireceklerdir.

İkinci senaryo: İşler zamanla düzeldikçe, toplam başarının veri ambarından geçtiği görülecektir. Şimdi üst yönetim, veri ve iş mimarisi ekipleri; geliştirilmesine yardımcı oldukları iş nesne modelleri ve mantıksal veri modelleri üzerine bina edilmiş tamamen normalize edilmiş veri ambarlarını kullanımda, yani çalışır halde görmek isteyeceklerdir.

Kurumsal çapta bir veri ambarı hedef olmalı ve bu senaryolar paralel olarak cereyan ederken, projenin önceliği bu hedefi gerçekleştirmek için gerekenin yapılması olmalıdır.

Şekil 11.3 taktiksel veri deposundaki düz dosyalardan, normalize edilmiş yapıya sahip olan kurumsal veri ambarına aktarım için gerekli olan gerçek dönüştürme proseslerini şematik bir yapıda göstermektedir.

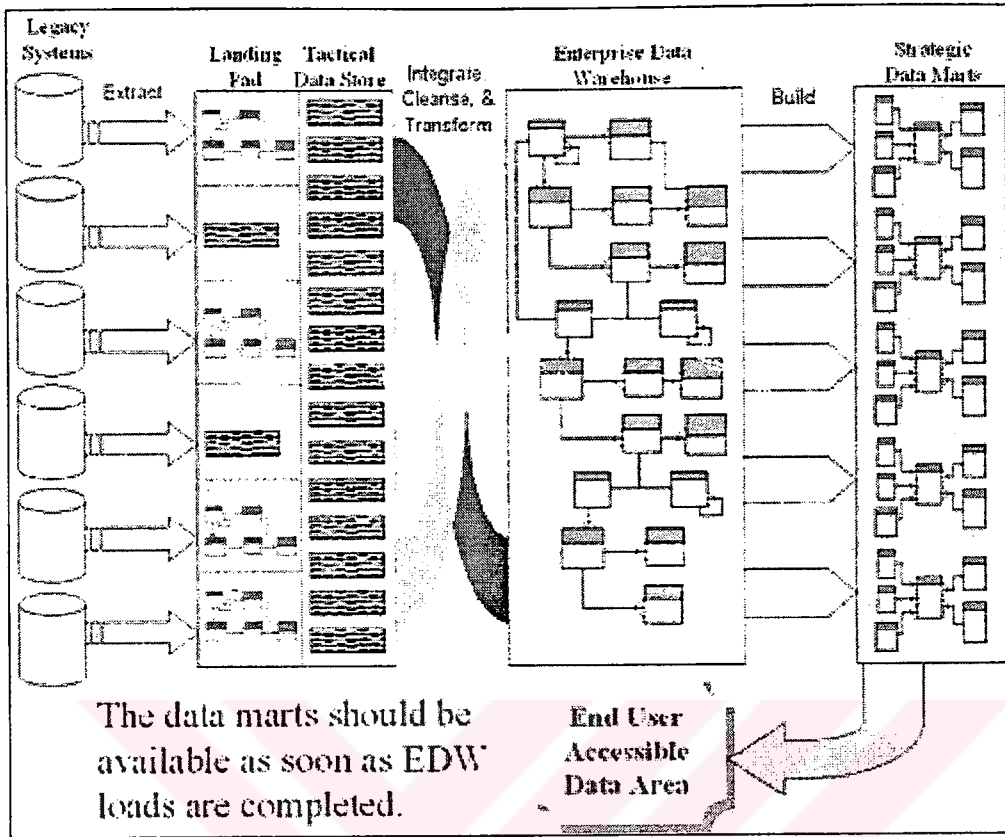


Şekil 11.3 Taktiksel veri deposundan, kurumsal veri ambarına veri dönüşüm prosesi
(Borland, <http://www.datawarehouse.com/iknowledge>)

Artık data martları oluşturma zamanıdır. Inmon metodolojisi kullandığımızdan; mini veri ambarı diyebileceğimiz konu spesifik, özel bilgi içeren data martları veri ambarından oluşturabiliriz.

Eğer veri ambarında yüksek kalitede veri olması bir seviyeye kadar kabul edilirse, mini ambar kabul edilen data martlarda ekstra mekanizma kullanmadan doğru sorgulama sonuçları elde edebiliriz.

Şekil 11.4 kurumsal veri ambarından data martlara veri taşınma prosesini göstermektedir.



Şekil 11.4 Kurumsal veri ambarından data marta veri aktarma
(Borland, <http://www.datawarehouse.com/iknowledge>)

11.3 Müşteri Verisinin Kalitesi ve ETL

Doğruluk, birçok veri ambarı projesinin başarısını bloke eden en büyük engellerden biridir. Aslında META GROUP araştırma şirketinin çeşitli anketlerine göre, veri ambarları uygulaması yapan şirketlerin karşılaştığı problemlerin başında "veri kalitesi" problemi gelmektedir (First Logic, Inc., 2001).

Veri kalitesinin en önemli alanlarından biri müşteri bilgisidir. En doğru bilgiye ulaşmak için, veri kalitesi aşağıdaki her bir kriter adımına dikkat ederek çalıştırarak gerekmektedir:

1. Seçme,
2. Dönüştürme,
3. Konsolidasyon ve
4. Bakım

Bir çok veri kalitesi ve konsolidasyon ürünleri aşağıdaki iki kategoriden birine düşer:

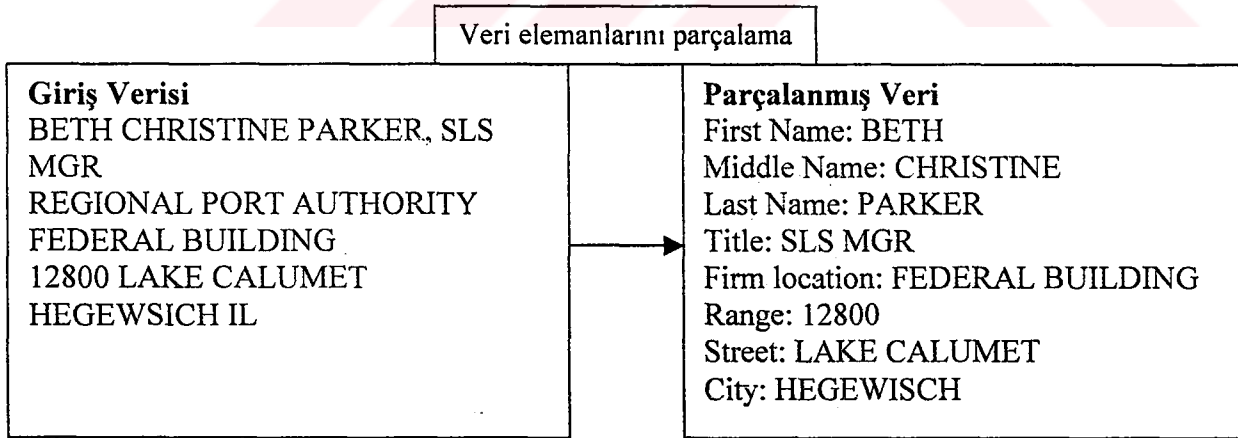
1. **veri-referanslı-çözümler**, sofistike algoritmalarla referans tabloları kombine edilir,
2. **veri-referanssız-çözümler**, sadece algoritmalara güvenilir.

Veri kalitesine üç aşamada erişilir:

1. Temizleme	Veri parçalanır, doğru hale getirilir, standart hale getirilir ve doğru eşleştirme için zenginleştirilir.
2. Eşleştirme	Benzer bilgilere ulaşabilmek için kaynak içinde veya kaynaklar arasında karşılaştırmalar yapılır.
3. Konsolidasyon.	Eşleştirilen veri elemanları konsolide edilir ve veri ambarına data marta veya diğer bir depolama alanına koyulur.

11.4 Parçalama (Parsing): Verinizin Ne Olduğunu Biliyormusunuz?

Veri temizleme de parçalama ilk kritik komponenttir. Parçalama, uzun karakter (string) veri yerine tekil komponentler(token) kullanarak karşılaştırma yapmaya imkan verdiğinden sonra yapılacak olan veri doğrulama, standartlaştırma ve eşleme işlemlerini kolaylaştırır.



11.5 Doğrulama: Verinin Doğru Olduğunu Nasıl Biliriz?

Veri parçalandıktan sonra, veri temizlemedeki ikinci aşama olan “doğrulama” gelmektedir. Müşteri bilgisi, temizleme ve doğrulamada en çok sıkıntı çekilen veri tipidir.

Elemanların doğrulanması

Parçalanmış veri		Doğrulanmış veri
First Name: BETH		First Name: BETH
Middle Name: CHRISTINE		Middle Name: CHRISTINE
Last Name: PARKER		Last Name: PARKER
Title: SLS MGR		Title: SLS MGR
Firm: REGIONAL PORT AUTHORITY		Firm: REGIONAL PORT AUTHORITY
Firm location: FEDERAL BUILDING		Firm location: FEDERAL BUILDING
Range: 12800		Address: 12800 SOUTH BUTLER DRIVE
Street: LAKE CALUMET		City: CHICAGO
City: HEGEWISCH		State: IL
State: IL		ZIP: 60633
		ZIP4: 2398.6
Solda görülen kayıta sokak (street) ve şehir (city) alanları sağda görüldüğü gibi düzeltilmiştir.		

Akıllıca ve güvenilir bir doğrulama için, ikincil bir güvenilir veri kaynağını (referans) kullanabilen bir yazılım kullanılmalıdır. Birçok durumda; doğrulama, eşleştirme için veri hazırlama amaçlı olarak kullanılır. Burada orjinal kayıtlar değişmeden kalır.

11.6 Standartlaştırma: Veri, Eşleştirme İçin Hazır mı?

Standartlaştırma, veri temizlemenin sıradaki prosesi, örneğin müşteri bilgisinin, tercih edilen ve tutarlı bir formata kavuşturulmasını sağlar. İkincil veri kaynakları ile entegre çalışan yazılım çözümleri (yani veri-referanslı-çözümler) daha etkin bir standartlaştırma yapar ve güvenle iş kararları oluşturulmasını sağlar. Örneğin, Amerika Posta İdaresi (USPS) ulusal adres bankasını kullanan çözümler, adres standartlaştırmasında başarılı olmuşlardır.

Elemanların standartlaştırılması

Doğrulanmış veri First Name: BETH Middle Name: CHRISTINE Last Name: PARKER Title: SLS MGR Firm: REGIONAL PORT AUTHORITY Firm location: FEDERAL BUILDING Address: 12800 SOUTH BUTLER DRIVE City: CHICAGO State: IL Zip: 60633 Zip4: 2398	Standartlaştırılmış veri Pre-name: Ms. First name: Beth First name match standards: Elizabeth, Bethany, Bethel Middle name: Christine Middle name match standard: Kristine Last name: Parker Title: Sales Mgr. Firm name: Regional Port Authority Firm location: Federal Bldg. Address: 12800 S Butler Dr City: Chicago State: IL Zip: 60633 Zip4: 2398
<i>Belirtilen kritere göre veri standartlaştırılır. Aynı kriterler eşleştirme safhasınd da rol oynayacaktır.</i>	

Eşleştirme standartları kullanılarak daha başarılı eşleştirme yapılabilmesi için, oldukça sofistike araçları ihtiva eden yazılımlar gereklidir. Bazı yazılımlar ön isim, isim, ünvan ve iş lokasyonlarını standartlaştırır ("Doctor"--"Dr.", "Junior"--"Jr.", "Floor"--"Flr.", vb.). Ayrıca güçlü isim bilgileri tablosu (örneğin 100.000 isim) ile *isimden cinsiyet tayini* gibi çok spesifik bilgilerle müşteri hakkında maksimum veriye sahip olup, onu tanımak bire-bir pazarlamada çok yüksek seviyede imkanlar tanır.

11.7 Zenginleştirme: Her Veri İhtiyaç Duymalı Mı?

Zenginleştirme, veri temizlemenin son adımıdır, yeni veri ekleme ve eksik veriyi tamamlama fonksiyonlarını kapsar. Yeni veri müşteriyi daha iyi tanıma ve gelecekteki satın alma davranışını tahmin etmeye yarar. Bu tip bilgi şunları içerebilir:

1. **Demokratik Veri:** Örneğin; yaşı, çocuk durumu, gelir eğitim seviyesi, ve evlilik durumu, satış hacmi, çalışan sayısı, vb.
2. **Coğrafi Veri:** Örneğin; apartman veya kat no, eksik adres elemanları, telefon numaraları, en / boy bilgisi, ilçe kodları ve siyasi bölgeler (Örneğin Marmara, Ege, vb).
3. **Davranış Verisi:** Örneğin; kredilendirme değeri, tercih edilen iletişim kanalları ve RFM (en son satınalma, ne kadar sıklıkla, kaç para) bilgileri.
4. **Fiziksel Grafik Verisi:** Örneğin; hobiler, ilgi alanları ve politik tercihler.
5. **Olay Yönetimli Veri:** Örneğin; evlilik, çocuğunun doğumu.
6. **Hesaplanan Veri:** Örneğin; kredi oranları, yaşam stili - standardı tahmini, ömür boyu değer hesaplamaları ve cevaplama (feedback) sayıları.



Kaydın zenginleştirilmesi

Standartlaştırılmış veri:

Pre-name: Ms.
First name: Beth
First name match standards: Elizabeth
Bethany, Bethel
Middle name: Christine
Middle name match standard: Kristine,
Cristine
Last name: Parker
Title: Sales Mgr.
Firm name: Regional Port Authority
Firm location: Federal Bldg.
Address: 12800 S Butler Dr
City: Chicago
State: IL
Zip: 60633
Zip4: 2398

Temizlenmiş ve zenginleştirilmiş veri (veri kaynağı 2):

Ms. Beth Christine Parker, Sales Mgr.
Match standards: FN: Elizabeth, Bethany,
Bethel;
MN: Kristine, Cristine
Reg Prt Auth
12800 S. Butler Dr.
Chicago IL 60633-2398
Phone: 708-555-1234
Fax: 708-555-5678
Gender: Strong, female
Account number; 00192744
SSN: 123-45-6789
Latitude: 41.662364
Longitude: -087.576274
MSA Code: 1600
Census block: 5104009
Cart: C004
DPBC: 02
LACS: F
County Code: 031
Lot: 0031
Lot order:A

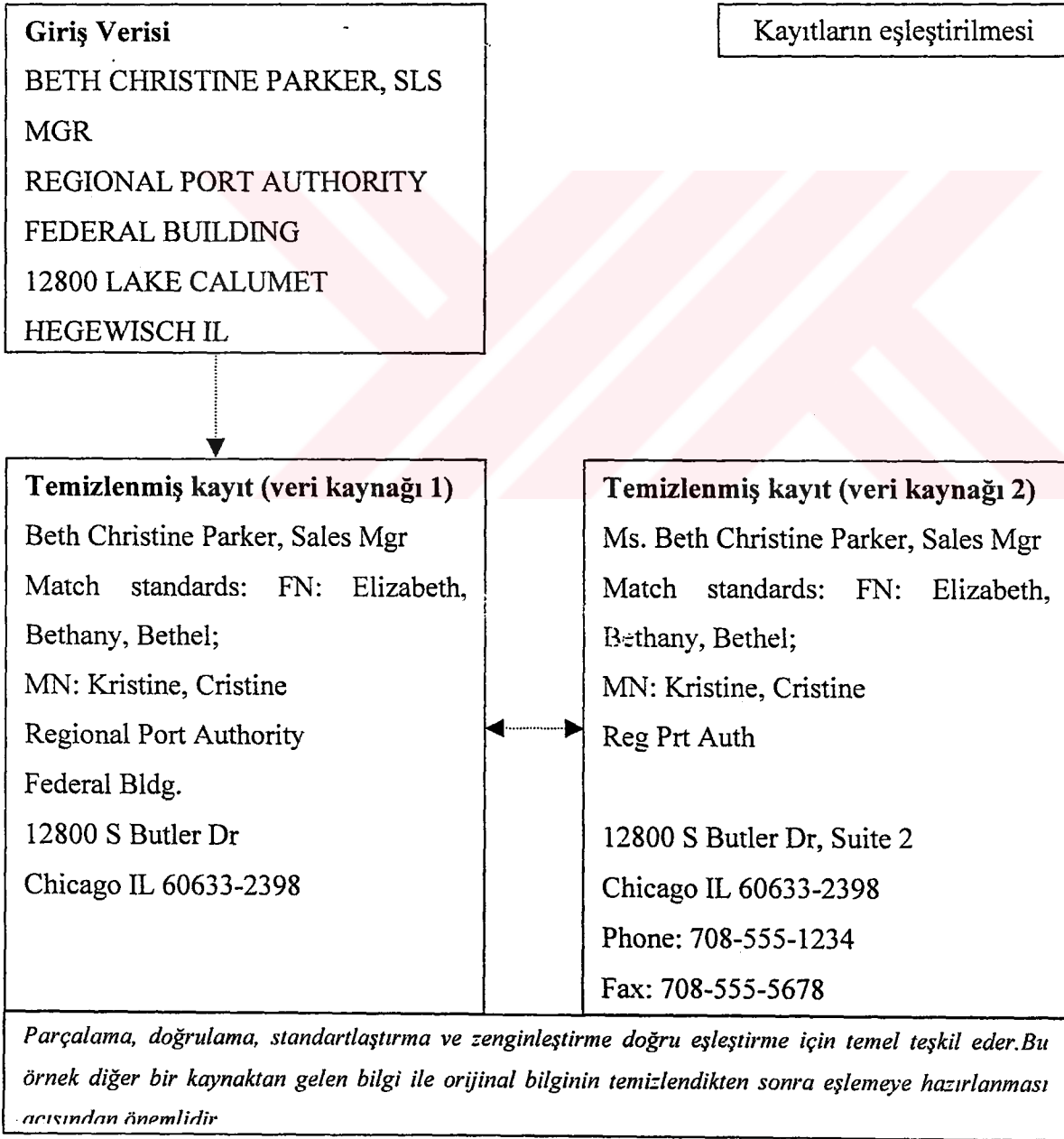
11.8 Eşleştirme: Veriinizdeki Tekrarlanmış Kayıtları Bulabiliyor musunuz?

Eşleştirme veri depoları içinde aynı verileri belirleme imkanı sağlar. Bu veri ambarcılığının kalbidir. Temizlenmiş bilgi ve eşleştirme standartları kullanarak her biri nesnel bilginin (örneğin müşteri, ürün vb.) çift veya tekrarlanmış kayıtlarını elimine ederek tek bir bütün bilgi olması için konsolide edebiliriz. Bu bize şu faydaları sağlar:

1. Müşteri veya ürünü tam olarak gösterir ve onların hakkında doğru veri üretir
2. Pazarlama promosyonlarının cevap oranlarını iyileştirir
3. Müşteriye gereksiz ve sıkıcı tanıtım yapılması, ürün sunulması riskini azaltır.
4. Doğru ve yeni bir müşteri portföyle eğilim ve kalıpları belirleyebiliriz.

Örneğin müşteri adı ve adres bilgilerinin eşleştirilmesi istenebilir. Veya adres değişiklikleri olabilir, ama sadece isim ve telefon bilgileri kesin eşleşsin de denebilir. Aşağıda işten-işe (B2B) verileri için örnekler verilmiştir:

1. Şirket birleştirmeleri elde edilmiş veriler veya anonim şirketlerin isim değişiklikleri
2. Bölümler, bayiler ve ortak şirketler arası ilişkiler
3. İş kısaltmaları (örneğin NASDAQ veya NYNEX)
4. baş harfler- tek tek okunan, bir kelimeler grubu veya unvanın ilk harfleri (örneğin AT&T veya CIA,FBI)



11.9 Konsolidasyon: Veri Eşleştirildi, Sonrasında Ne Yapılacak?

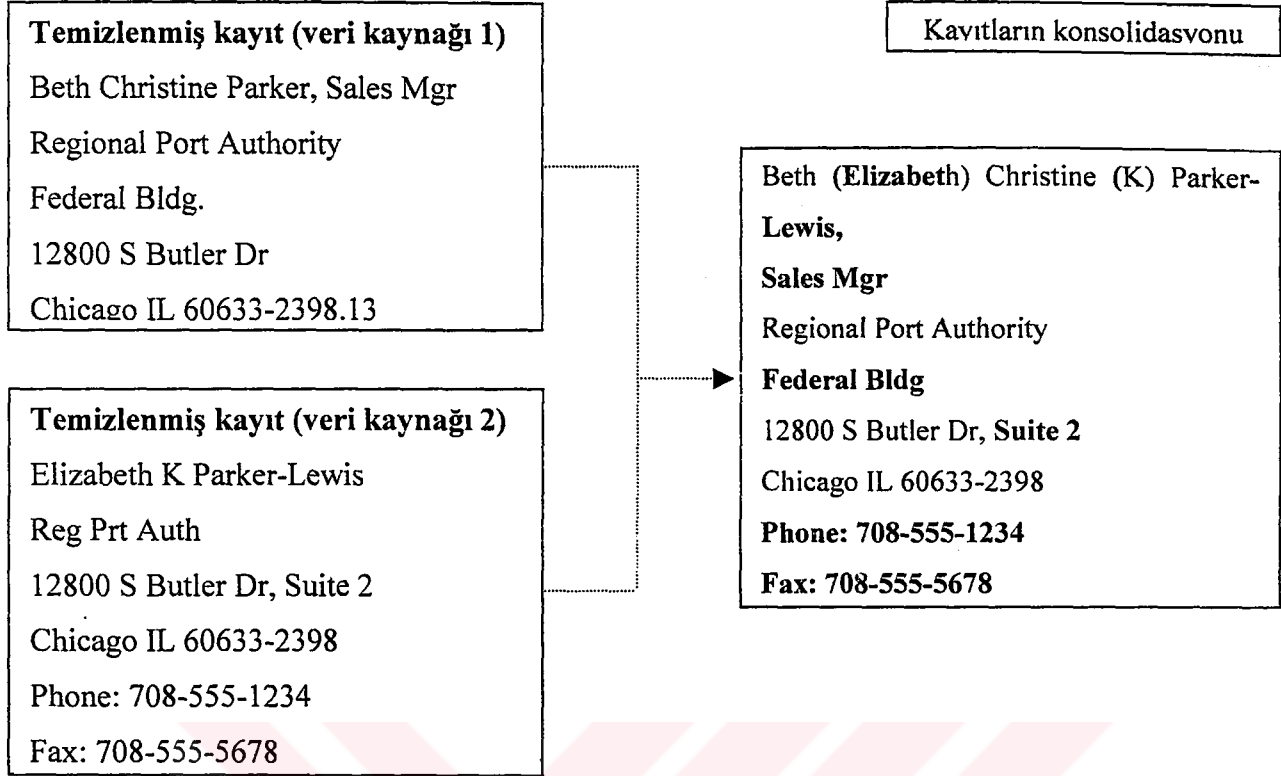
Veriler eşleştirildiğinde, müşteri kayıtları arasındaki ilişkiler belirlenebilir ve her biri tek bir konsolide görünüme dönüştürülebilir. Bu kritik komponent; başarı ile bire-bir pazarlama çalışmalarına uygulandığında, müşteriye anlama kapasitesinin artışı yaşanacaktır. Epsilon şirketinin yöneticilerinden Greg Holzwarth'a göre "Bire-bir pazarlama, organizasyonlara müşterilerle temas kurulan her noktada daha iyi hizmet sunabilme imkanı sağlar".

Konsolidasyon için iki metod kullanabiliriz, bunların her ikisi de bir çok veri ambarcılığı ve bire-bir pazarlama uygulamaları için zorunludur.

1. Birinci konsolidasyon prosesi, mevcut kaynaklarının hepsini kullanarak herhangi bir müşteri hakkındaki tüm verilerin kombine edilmesi
2. İkinci prosesi, müşteri ilişkisi belirleme- müşteriler arasındaki bağlantıları ortaya çıkarır.

Müşteri ilişki belirlemede ortak iki tip mevcuttur:

1. **Ev-lendirme (Householding):** Tipik olarak, ev-lendirme aynı adres ve soyisim içeren kayıtlarının arasında bağlantı kurulmasıdır, yani bu iki tüketici aynı evde oturmaktadır.
2. **İş gruplama:** Aynı bilgileri (örneğin şirket ismi, adres, departman, veya ünvan) paylaşan kayıtların gruplanmasıdır.



Gruplanmış veya ev-lendirilmiş kayıtların karakteristikleri ve satınalma alışkanlıkları belirlenerek özel teklifler ve daha iyi direkt pazarlama faaliyetleri yaratılabilir. Her bir müşteri verisi kombine edildiğinde, veri kaynakları ve spesifik veri alanları, arasındaki önceliklerin belirlenmesi gerekecektir. Örneğin bazı kayıtların daha tam bilgi içermesi, bazılarının ise daha az güvenilir bilgi içermesi sebebiyle güvenilir olanın önceliği daha fazla olmalıdır.

11.10 Eşleştirme ve Konsolidasyon Çözümleri: Nasıl bir araya getirecek?

Eğer eldeki veri parçalanmamış (unparsed) veriler içeriyorsa, birçok eşleme hatası veya yanlış konsolide edilmiş aslında aynı fakat ayrı olan müşteri kayıtları oluşması tehlikesi ile karşı karşıya kalınacaktır. Eşleştirme ve konsolidasyon çözümlerini ayrı ayrı inceleyelim.

11.10.1 Eşleştirme Sistemleri

Hali hazırda çok değişik eşleştirme sistemleri mevcuttur. Bunların her biri eşleştirme sonucuna değişik yollarla ulaşmaktadır.

1. Anahtar-kod eşleştirme: Bir veya birkaç alanda (field) ilk bir kaç karakteri kullanarak tekil karşılaştırma yapılır. Bu ilkel metod, nadiren pratik uygulamalarda kullanılır. Veri az ise uygundur, ama bir çok hata eşleştirme ile sonuçlanır.

2. Ses benzeşimi (Soundexing): Fonetik (söyleniş) benzerliklerini (örneğin "f" ve "ph" veya Quick ve Kwik) bulur. Genelde telefonla alınan bilgilerde oluşan hataların giderilmesi için kullanılır. Ses benzeşimi tekniği yeterli değildir, çünkü sadece fonetik hataların düzeltilmesini sağlar.

3. Benzerlik eşlemesi (Similaity): Aynı zamanda "bulanık" (fuzzy eşleştirmesi) de denir. Bu metod da, iki diskrit komponent arasında benzerlik eşlemesinin hesaplanması ile eşleştirme yapılabilir. Tam %100 eşleştirme beklenmemelidir; çünkü sadece yanlış yazım, fonetik, baskı ve transpozisyonel hataların düzeltilmesinde kullanılabilir. Geniş bir kitlenin en iyi eşleştirme metodu olarak kabul ettiği bir sistemdir. Özellikle standartlaştırılmayan (soyisimler, iş isimleri ve ev numaraları) bilgiler için değerli bir eşleştirme metodudur.

Örneğin; "1001 Rose St." adresi için :

1001 ve 101 benzerlik oranı : 85 %

1001 ve 1010 benzerlik oranı : 75 %

1001 ve 1025 benzerlik oranı : 50 %

Benzerlik eşlemesi bir alandaki tüm karkterleri ele alır ve her birinin pozisyonuna göre bir eşleme değeri verir. Üsteki örnekte benzerlik değerlerinden eşleme yüzdesi üretilmiştir.

4. Ağırlıklı eşleştirme: Ses benzeşimi ve benzerlik eşlemesi sistemlerinin birleştirilerek uygulandığı bir metoddur.

Kombine Yaklaşım - Benzerlik, ağırlık ve bazı amaca özel algoritmaların karışımı genelde en iyi çözümdür. Kombine yaklaşım kullanıldığı zaman, iş eşleştirme kurallarının esnek bir şekilde değiştirilmesine ihtiyaç duyulur. Örneğin, evinden taşınmış kişilerin bilgileri her zman güvensizdir. Adres kayıtlarına bakmak yerine, kombine yaklaşımdan istifade ile diğer alternatif alanları, (örneğin, sosyal güvenlik numarası-SSN, kredi kartı numaraları, hesap numaraları ve doğum tarihleri) kullanarak eşleştirme yapmak gerekir.

11.10.2 Konsolidasyon Sistemleri

Veri temizleme ve akabindeki eşleştirme proseslerinin her ikisi de tek bir amaca hizmet eder: "Doğru veri konsolidasyonu". Bunu inşa etmek için şunlara ihtiyaç duyulur:

1. Esnek konsolidasyon çözümünün operasyonel mevcut verinin kombine edilmesi ve
2. Sürekli akan veri besleme sistemlerinin bakımı

Başarısı kanıtlanmış bir konsolidasyon sisteminin ana komponentleri aşağıdakilere imkan vermelidir:

1. Gelen veri kaynaklarına öncelik verebilme
2. Alanlara öncelik verilebilmesi
3. Orjinal veri kaynaklarının muhafaza edilmesi (bakımı)

Her bir küçük başarıda, tecrübe, güven ve sürekli bir destek oluşacağından tüm bu metodların çok zor da olsa oluşturulup, uygulanabilmesi veri ambarcılığı veya iş aklı sistemlerini de nihayetinde başarılı kılar. Aslında bu her iş için de geçerlidir.

11.11 Veri Kalitesi Yazılımı ve ETL Çözümleri Arasındaki Farklar

En basit tanımıyla (No, 2000); ETL yazılımı, kayıtları bir veri kaynağından seçip alır, yeni formatlara dönüştürür ve hedef varış noktalarına yükler. Diğer taraftan; veri kalitesi yazılımı ise her bir tekil kaydın içeriğinin, doğru, güncel ve tutarlı bir şekilde sunabilmesinin garantilenmesi konusuna yoğunlaşır.

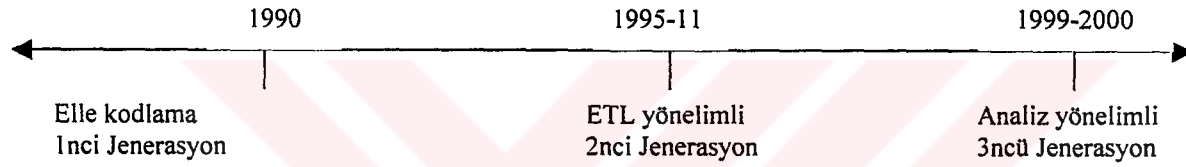
Unutulmamalıdır ki; ETL prosesleri, veri ambarı projelerindeki proseslerin %70'ini gerçekleştirdiğinden, bu harcanan zamanı minimuma indirir. Çünkü; en fazla zaman seçme, dönüştürme ve yükleme işlemlerine harcanır (Meyre, 2001). ETL yazılım araçlarının ana faydası, çok veri içeren klasik sistemler için ürettiği veri dönüştürme fonksiyonlarıdır (White, 2000). Eğer gerekirse üretilen kod uyarlanabilir ve hatta eldeki mevcut kaynak kod rutinleri ile zenginleştirilebilir. Fakat kod-üreten bu tip araçlardaki problem, programların toplu iş (batch) programları olması, karmaşık ve sayılarının da çok olması sebebiyle yönetim ve bakım zorluğudur. Çünkü her iş kuralı değişiminde veya iş mantığı değişiminde bu programların da değişmesi gerekir ki, bu oldukça zor olmaktadır.

Koruyucu hekimlik de olduğu gibi aşağıdaki sırada araçların kullanımı ile veri kalitesi problemleri daha oluşmadan çözülebilir.

1. Veri Profili belirleme araçları (Data profiling)
2. Veri Değişim Mühendisliği araçları (Data re-engineering)
3. ETL araçları

Bu üç tip aracın da hepsi, ancak ve ancak yukarıdaki sırada kullanılır ise veri kalitesinin düşürülmesi engellenebilir.

11.12 ETL Araçlarının Tarihsel Gelişimi



1990'lara kadar; organizasyonlar ETL program ihtiyaçlarını, veri aktarımı olarak (data migration adıyla) kabul ederek elle (manuel) analiz ve kodlama yapmışlardır. Bu tip veri aktarımı girişimleri genelde başarısızlıkla sonuçlanmıştır. Bu tip veri aktarımı prosesine **birinci-jenerasyon-yaklaşımı** denmektedir.

1990'ların ortalarında gelecek nesil çözümler, üreticiler tarafından pazara sunuldu. Bunlar ETL için otomatik program kodu üretmekteydiler. ETL üreticileri artık ETL-yönelimli veri aktarım yaklaşımını ortaya attılar. Otomatik veri aktarım kodu üretimi, hataların hızlı tesbitine katkıda bulunabilir. Buna **ikinci-jenerasyon-ETL-yönelimli veri aktarım yaklaşımı** denmektedir. Bu yaklaşım, klasik sistemler için istenen program değişikliklerini son derece azaltılmıştır.

Çok yakın zamanda, veri aktarım prosesini ilk defa otomatik hale getiren araçlar ortaya çıktı. Bu tip araçlara, **veri profileme araçları** denir. Veri analizini otomatikleştiren, uygun metodolojiler ile birleştirilen ve uygulayan bu araçlar, veri aktarımındaki riskleri radikal olarak düşürmektedir.

11.13 ETL Araçlarının Ölçeklenebilirliği (Scalability)

Her proje veya uygulama için ortam, ölçeklenebilirlik, vb. kriterler önemli olduğu gibi, ETL araçları için de, son derece önemli ve de zorunludur. En küçük uygulamadan (küçük bir data mart) en büyük uygulamaya (devasa ölçülerde real-time web-tabanlı veri ambarı) kadar ölçeklenebilecek bir yapı sunabilmek, hazırlayabilmek çok zor, fakat önemli bir kriterdir. ETL geliştirme ortamının kriterleri (Miller, 2000)

1. En uç seviyede ölçeklenebilirlik
2. Büyük çıktılar
3. Sağlam performans
4. Düşük sahip olma maliyeti

Kullanıcı arayüzü bu konuda önemli bir husustur. Kaynak ve hedef ortamları için $n \times (n - 1)$ değişik ekran arayüzü oluşturma ihtimali söz konusudur. Bilgiler dört ayrı kaynaktan gelsin, bu durumda 12 ($4 * 3$) değişik ekran söz konusu demektir. Bu da bir çeşit geliştirme performansı kriteridir, yani nasıl bir arayüz zaman ve değişim maliyetini azaltır diye bakmak gerekir.

Ölçeklenebilirlik kriterleri şunlardır (Shash, 2000):

1. Performans
2. Süreklilik
3. Bakım kolaylığı

ETL ortamının performansını artıracak ana kriterler aşağıdaki gibidir:

1. Veri akışını düşür
2. I / O eliminasyonu
3. Ortak ETL proses optimasyonu
4. Dar boğazların ve kaynakların dengelenmesi problemlerinin eliminasyonu
5. Ortak ölçeklenebilirliğinin belirlenmesi için kurallar koyulması
6. Veri paralelliğinin kullanımı

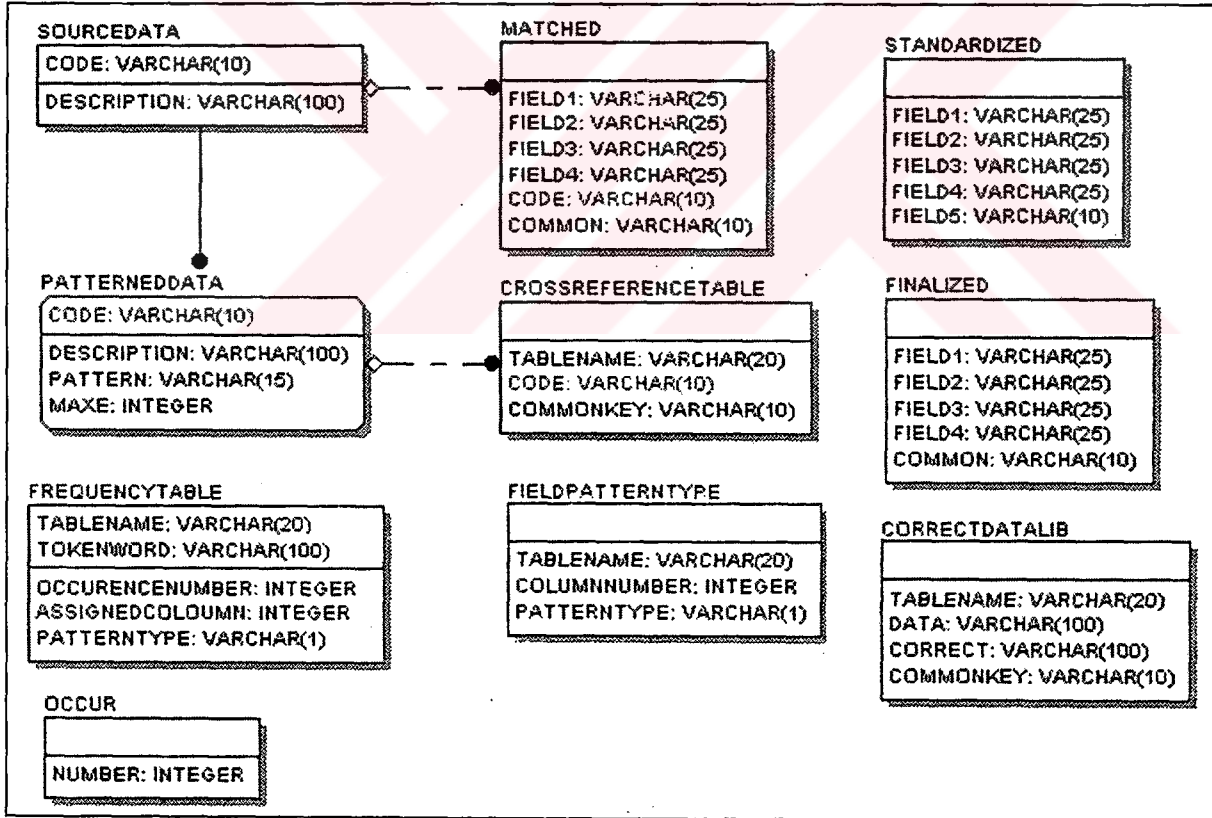
Paralel işlemci ve paralel programlama kullanan ETL üretici firma çok az (2-3) olduğundan en güçlü ölçeklenebilirlik tedariki çok düşük bir ihtimal ve pahalı bir tedarik gerektiriyor.

12. ETL SİMÜLASYON UYGULAMASI

Bu bölümde; ETL araçları birinci jenerasyon tipinde hazırlanmış olan örnek simülasyon uygulaması anlatılacaktır. Bu uygulamanın ismi “YıldızETL” olarak belirlenmiştir.

12.1 “YıldızETL” Uygulaması Mantıksal (E/R) Modeli

YıldızETL, uygulaması basit bir ETL ürünüdür, hatta simülatörüdür. Bu ürün geliştirilirken kullanılan ve program esanısında otomatik ve durumsal yaratılan (CREATE TABLE komutu ile) temel tablo yapıları Şekil 12.1’de gösterilmiştir. Bu tasarım, ERwin Designer (Computer Associates firmasının) E/Rmodelleme aracı ile hazırlanmış mantıksal bir modelledir. Tıpkı önceki bölümlerde anlatıldığı gibi mantıksal E/R modeli oluşturulmuş, tarif edilen metodolojiye uygun olarak fiziksel model ve fiziksel veritabanı/veri ambarı elde edilmiştir.



Şekil 12.1 YıldızETL simülatöründe kullanılan tabloların mantıksal (E/R) modeli

12.2 “YıldızETL” Uygulaması SQL Cümlecikleri

Fiziksel model oluşturmak için kullanılan SQL cümlecikleri aşağıdaki gibidir.

12.2.1 Tablo Yaratma SQL Cümlecikleri

ERwin Designer; aracı ile uygulama programında kullanılan TABLOlar ile ilgili üretilen SQL-scriptleri:

```
CREATE TABLE CORRECTDATALIB (
    TABLENAME      VARCHAR(20),
    DATA            VARCHAR(100),
    CORRECT          VARCHAR(100),
    COMMONKEY        VARCHAR(10) );
```

```
CREATE TABLE CROSSREFERENCETABLE (
    TABLENAME      VARCHAR(20),
    CODE            VARCHAR(10),
    COMMONKEY        VARCHAR(10) );
```

```
CREATE TABLE FIELDPATTERNTYPE (
    TABLENAME      VARCHAR(20),
    COLUMNNUMBER    INTEGER,
    PATTERNNTYPE     VARCHAR(1) );
```

```
CREATE TABLE FINALIZED (
    FIELD1           VARCHAR(25),
    FIELD2           VARCHAR(25),
    FIELD3           VARCHAR(25),
    FIELD4           VARCHAR(25),
    COMMON           VARCHAR(10) );
```

```
CREATE TABLE FREQUENCYTABLE (
    TABLENAME      VARCHAR(20),
    TOKENWORD        VARCHAR(100),
```


OCCURENCENUMBER	INTEGER,
ASSIGNEDCOLOUMN	INTEGER,
PATTERN TYPE	VARCHAR(1));

```
CREATE TABLE MATCHED (
  FIELD1          VARCHAR(25),
  FIELD2          VARCHAR(25),
  FIELD3          VARCHAR(25),
  FIELD4          VARCHAR(25),
  CODE            VARCHAR(10),
  COMMON          VARCHAR(10) );
```

```
CREATE TABLE OCCUR (
  NUMBER          INTEGER );
```

```
CREATE TABLE PATTERNEDDATA (
  CODE            VARCHAR(10) NOT NULL,
  DESCRIPTION     VARCHAR(100),
  PATTERN         VARCHAR(15),
  MAXE            INTEGER );
```

```
CREATE TABLE SOURCEDATA (
  CODE            VARCHAR(10) NOT NULL,
  DESCRIPTION     VARCHAR(100) NOT NULL );
```

```
CREATE TABLE STANDARDIZED (
  FIELD1          VARCHAR(25),
  FIELD2          VARCHAR(25),
  FIELD3          VARCHAR(25),
  FIELD4          VARCHAR(25),
  FIELD5          VARCHAR(10) );
```

12.2.2 Index Yaratma SQL Cümlecikleri

ERwin Designer; aracı ile uygulama programında kullanılan INDEXler ile ilgili üretilen SQL-scriptleri:

```
CREATE INDEX XIF3CROSSREFERENCETABLE ON CROSSREFERENCETABLE  
( CODE );
```

```
CREATE UNIQUE INDEX XPKFREQUENCYTABLE ON FREQUENCYTABLE  
(  
    TABLENAME,  
    TOKENWORD );
```

```
ALTER TABLE FREQUENCYTABLE  
    ADD PRIMARY KEY (TABLENAME, TOKENWORD);
```

```
CREATE INDEX XIF2MATCHED ON MATCHED  
( CODE );
```

```
CREATE UNIQUE INDEX RDB$PRIMARY5 ON PATTERNEDDATA  
( CODE );
```

```
CREATE INDEX XIF1PATTERNEDDATA ON PATTERNEDDATA  
( CODE );
```

```
ALTER TABLE PATTERNEDDATA  
    ADD PRIMARY KEY (CODE);
```

```
ALTER TABLE CROSSREFERENCETABLE  
    ADD FOREIGN KEY (CODE)  
        REFERENCES PATTERNEDDATA;
```

```
ALTER TABLE MATCHED
```

```
ADD FOREIGN KEY (CODE)
REFERENCES SOURCEDATA;
```

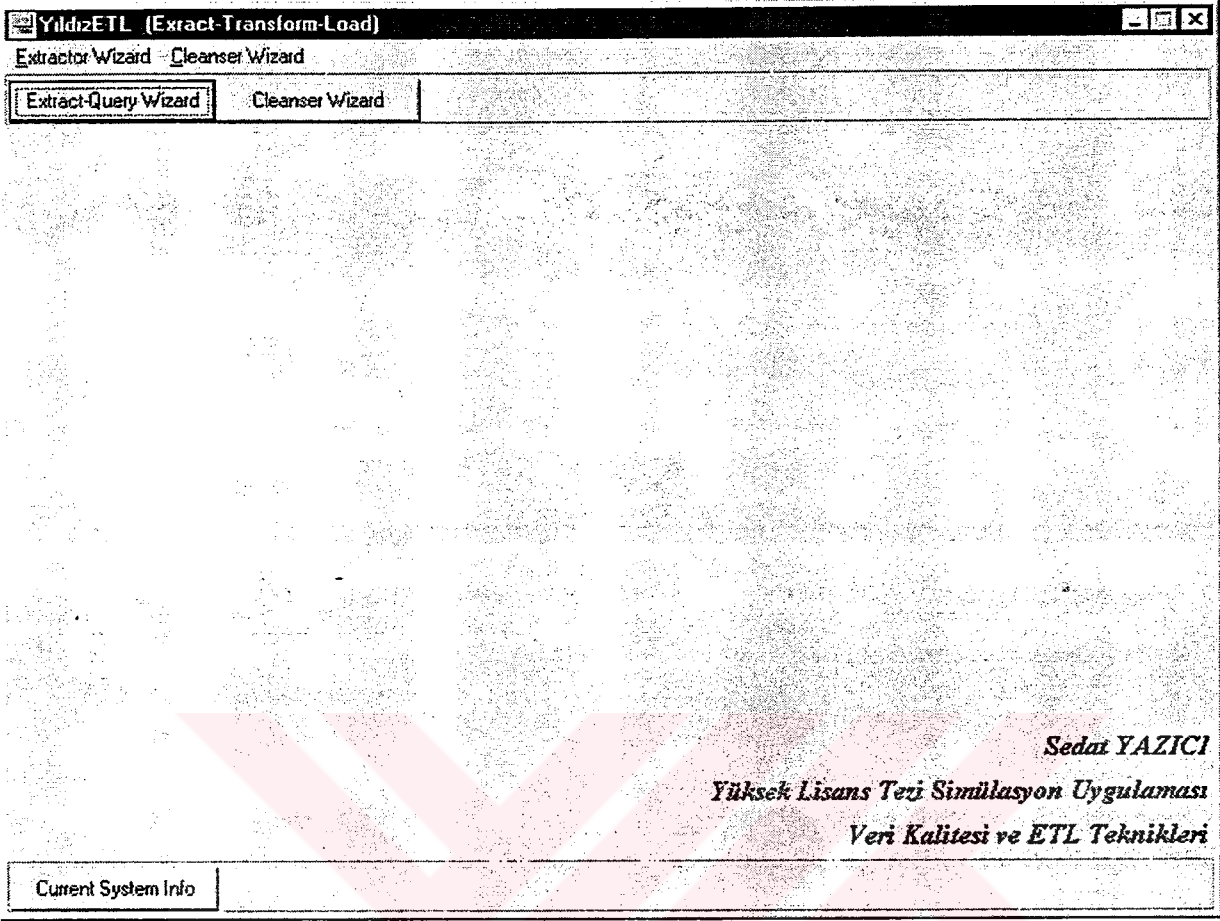
```
ALTER TABLE PATTERNEDDATA
ADD FOREIGN KEY (CODE)
REFERENCES SOURCEDATA;
```

```
CREATE UNIQUE INDEX XPKSOURCEDATA ON SOURCEDATA
( CODE );
```

```
ALTER TABLE SOURCEDATA
ADD PRIMARY KEY (CODE);
```

12.3 “YıldızETL” Uygulaması Genel Anlatımı

Bu örnek uygulama basit bir ETL ürünüdür, hatta simülatörüdür. Geliştirilen teknik ve algoritmaların simüle edildiği, uygulandığı bir ETL programıdır. Fonksiyon adımları olarak bakıldığında ETL’in temel fonksiyonlarının yerine getirir. Ancak, çok geniş kapsamlı veri sözlükleri (isimi adresi cinsiyet, kısaltmalar, vb.) kullanmamakta, “ses benzeşimi” ve “iş mantığı analizi” de yapmamaktadır. Ancak bunların her biri son derece ağır ve uzmanlaşma istediğinden çok uzun zamanlı uygulamalardır. Örnek vermek gerekirse iş aklı ürünleri üreten bir firma (Data Junction); kurumsal uygulama entegrasyonu (EAI) konusunda dünyanın en iyisi olduğunu iddia ederken, 700-adam yılda (fiilen 15 yılda) bu noktaya geldiğini belirtmektedir. EAI sadece konunun bir parçası olduğu düşünülür ise kapsam ve büyüklük daha da net anlaşılacaktır. Programın ilk açılan ekranı aşağıdaki (Şekil 12.2) gibidir.



Şekil 12.2 YıldızETL simülasyon uygulaması ana ekranı

Programın sistem performansı izleme ve sistem bilgisi elde etme ekranı aşağıdaki (Şekil 12.3) gibidir.

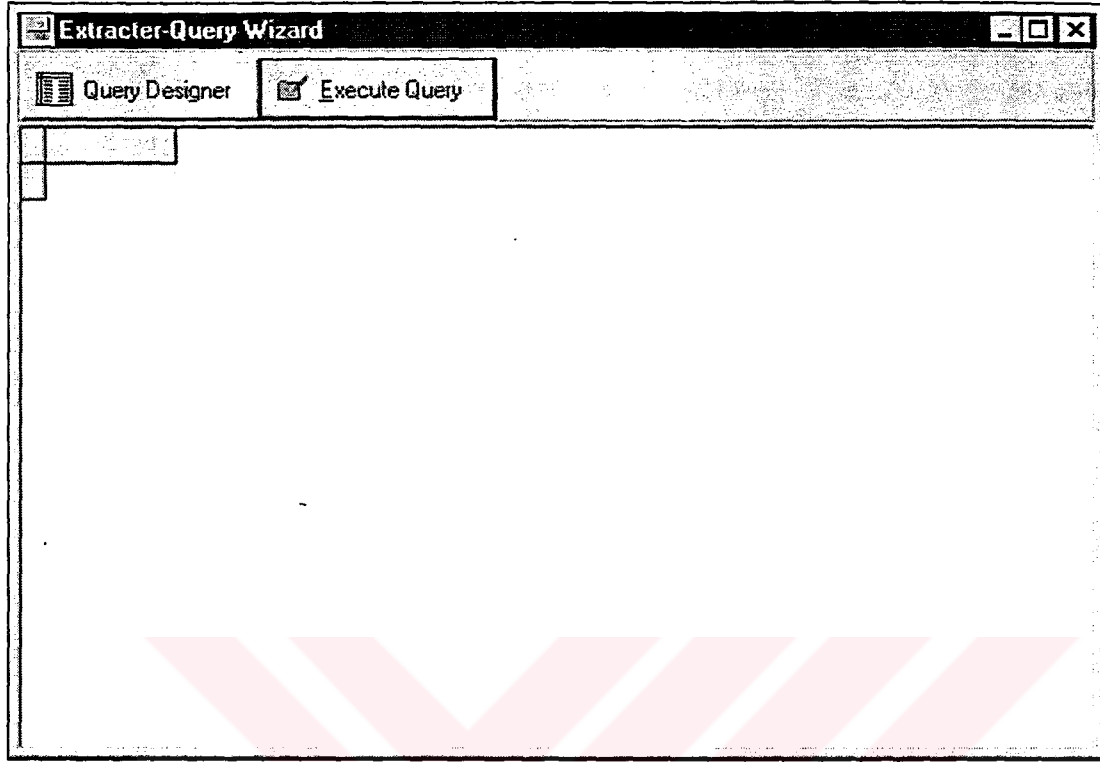
System Performans Info	
Volume Name	
Volume Serial No	1247-0EF8
File System Name	FAT32
File Drivers	A: C: D:
Processor Type	Pentium
OS Version	4.0 (1111. b)
Platform	Windows 95
Computer Name	SEDAT
FPU	Yes
User Name	SEDAT YAZICI
Company Name	ALFA
CD Serial ID	17597-DEM-0023207-47548
Bios Date	11/28/97
Bios Name	Phoenix
Bios Version	PhoenixBIOS 4.0 Release 5.12
BUS Type	ISA
CPU Type	Pentium
Machine Type	IBM PC/AT
Platform ID	1
Total Physical Memory	32.256 Kb
Availible Physical Memory	Kb
Total Virtual Memory	2.093.056 Kb
Availible Virtual Memory	2.032.512 Kb
Memory Loaded	100%
Cpu Usage	3%
Cpu	167 Mhz
	8%
☒ Tamam	

Şekil 12.3 YıldızETL sistem performansı izleme ve sistem bilgisi elde etme ekranı

12.4 “YıldızETL” Uygulaması “Seçme Sihirbazı (Extractor Wizard)” Modülü

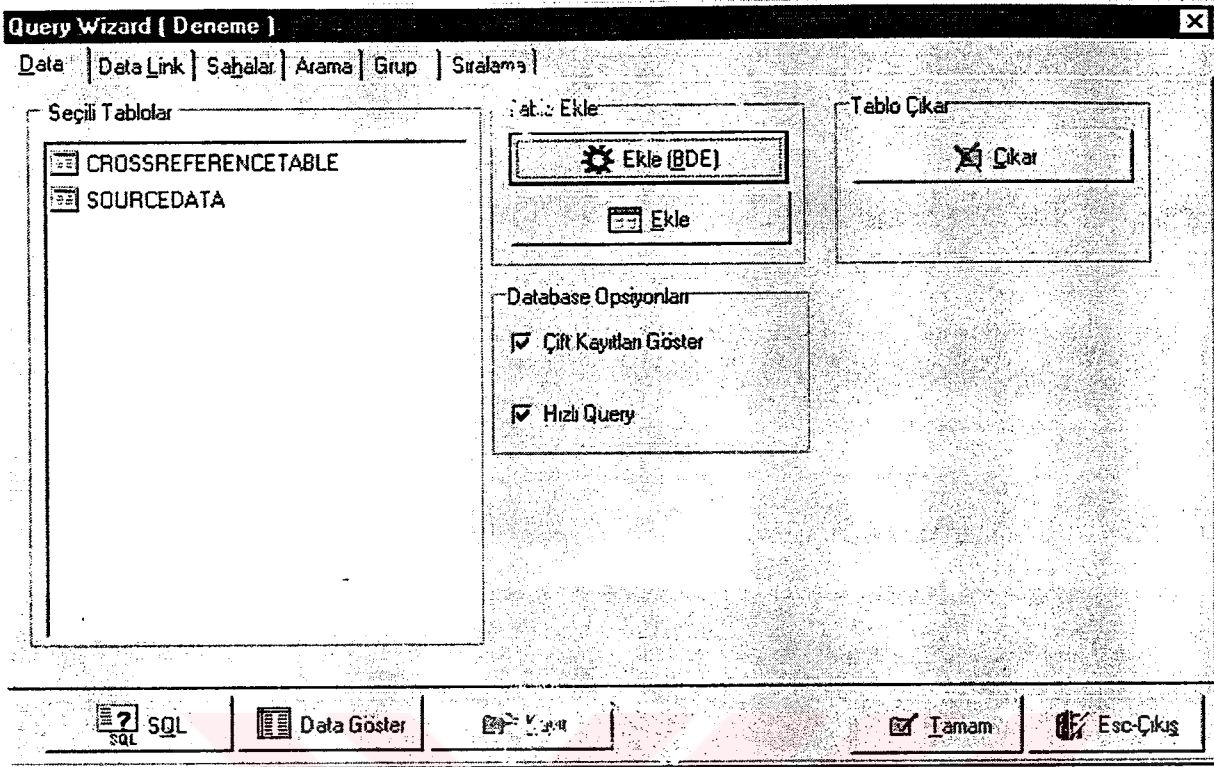
Veri ambarı veya iş akıllı uygulamaların ilk temel prosesi, operasyonel veri kaynaklarından (ODS-Operational Data Source) ilgili verinin seçilerek alınmasıdır. Gerek kullanım kolaylığı, gerekse modüler yapı sağlanabilmesi için sihirbaz (Wizard) mantığı ve şekli kullanılmıştır.

Uygulama, seçme (extraction) konusunda oldukça yeteneklidir. Çoklu-ortam veri tabanlarında problemsizdir. MVS mainframe CICS, WinNT ve AS/400, Interbase, Oracle ve bilumum masaüstü lokal veri tabanları ile çalışabilmektedir. Hem aracısız (native), hem de duruma göre aracılı (ODBC) kullanılabilmekte; dinamik SQL cümlecikleri oluşturup, verinin ambara alınmasını simüle etmektedir. Bu bir sihirbaz (Extractor Wizard) şeklinde yapılmıştır. Seçme sihirbazının (Extractor Wizard) ilk açılan ekranı aşağıdaki (Şekil 12.4) gibidir.



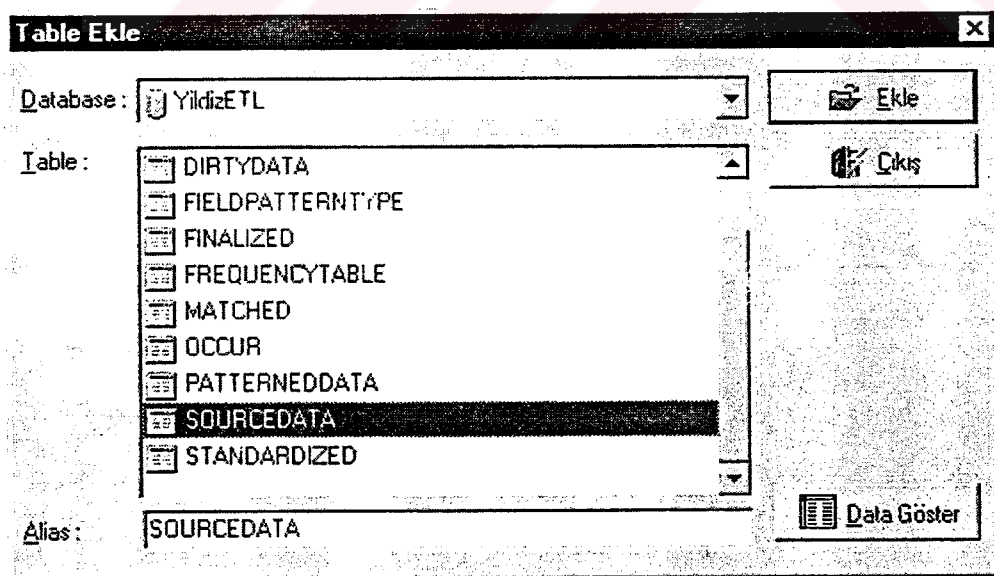
Şekil 12.4 YıldızETL simülasyon uygulaması ExtractorWizard modülü açılış ekranı

Bu ekranda seçme işlemini tanımlamak için “Query Designer” düğmesine basınız. Sırasıyla aşağıdaki ekranlar gelecektir.



Şekil 12.5 YıldızETL uygulaması ExtractorWizard modülü birinci kulakçık (page-tab 1)

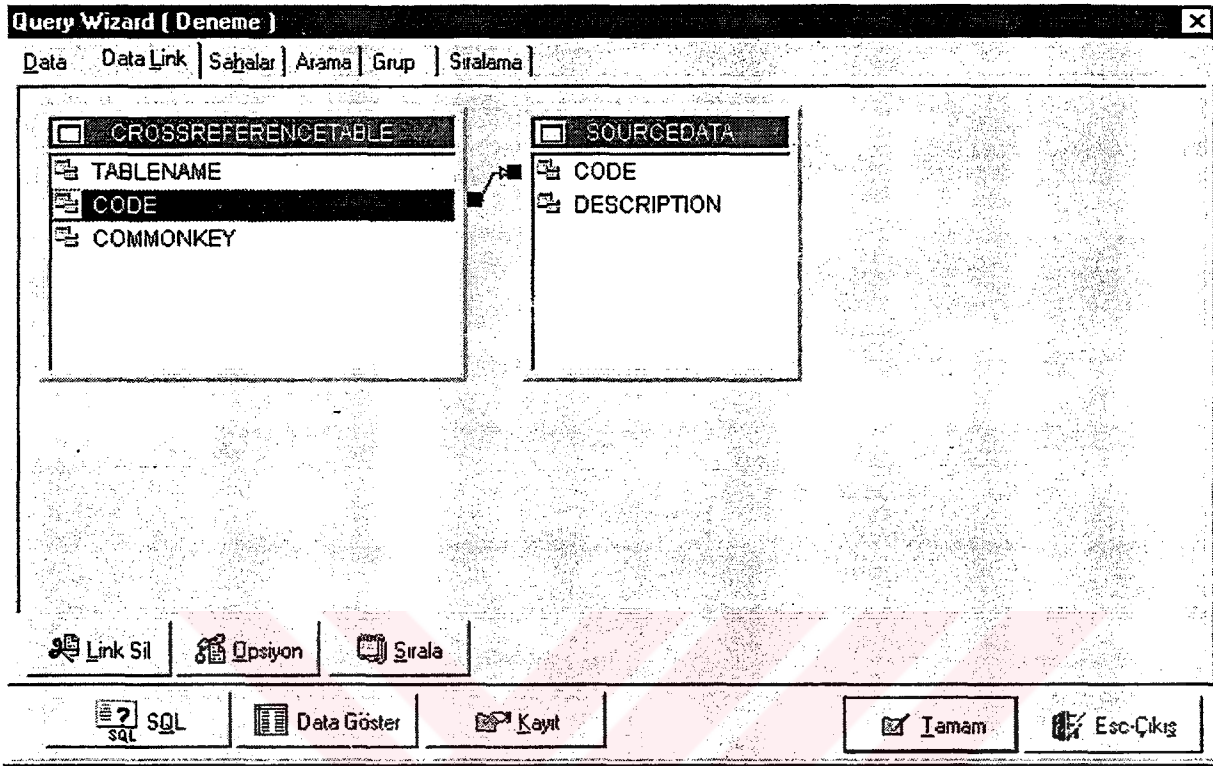
Bu ekranda istenen ortamdan istenen tablo seçilmektedir. “Ekle” düğmesine basıldığında Şekil 12.6’daki gibi bir ekran gelir.



Şekil 12.6 YıldızETL uygulaması ExtractorWizard modülü seçilecek tabloların eklenmesi

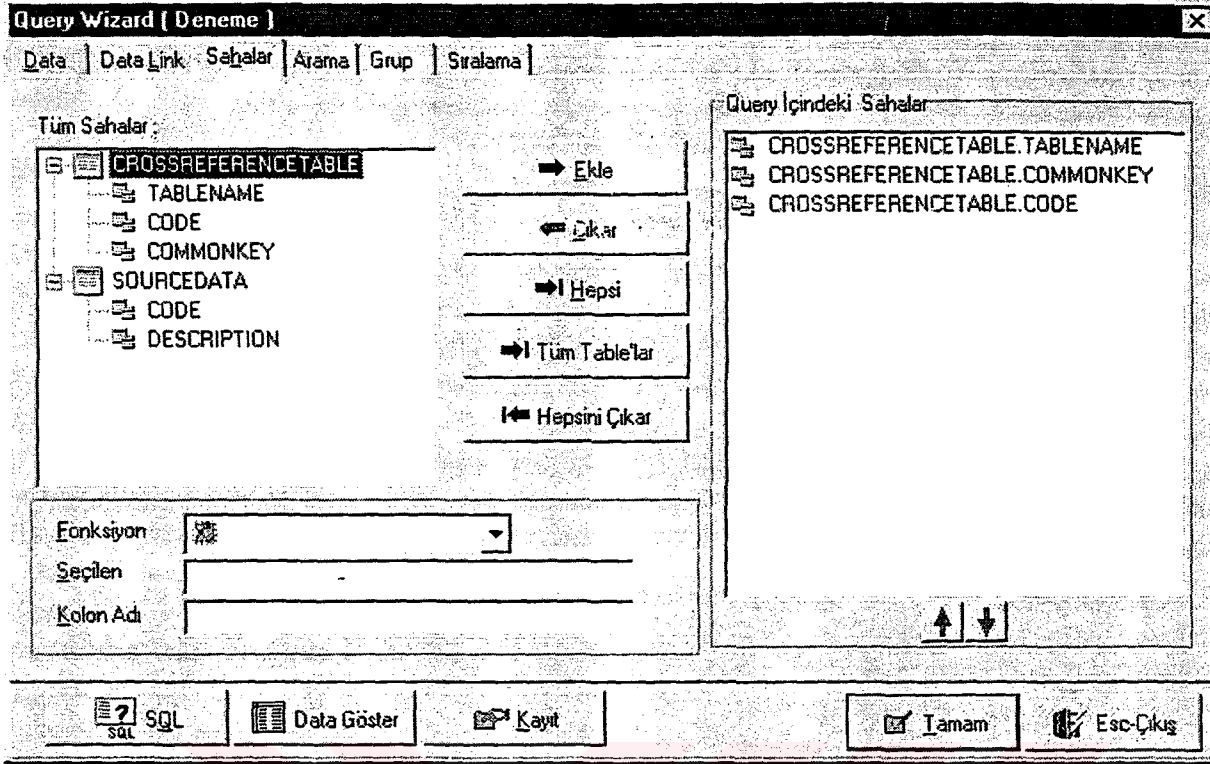
Üstteki ekranın üstünde, istenen veri tabanı seçilir. Seçili veri kaynağındaki “Tablo Listesi” den bir tablo ismi seçilerek “Ekle” tuşuna basılır. Bu ekranda tablo eklemekten başka istenen tablonun alanlarını parametrik tasarlanmış bir ekranda görmek için “Data Göster” tuşuna

basılması gereklidir. Seçilmiş tablolar arasında ilişki kurmak istersek ikinci kulakçıkta (Şekil 12.7) işlem yapmalıyız.



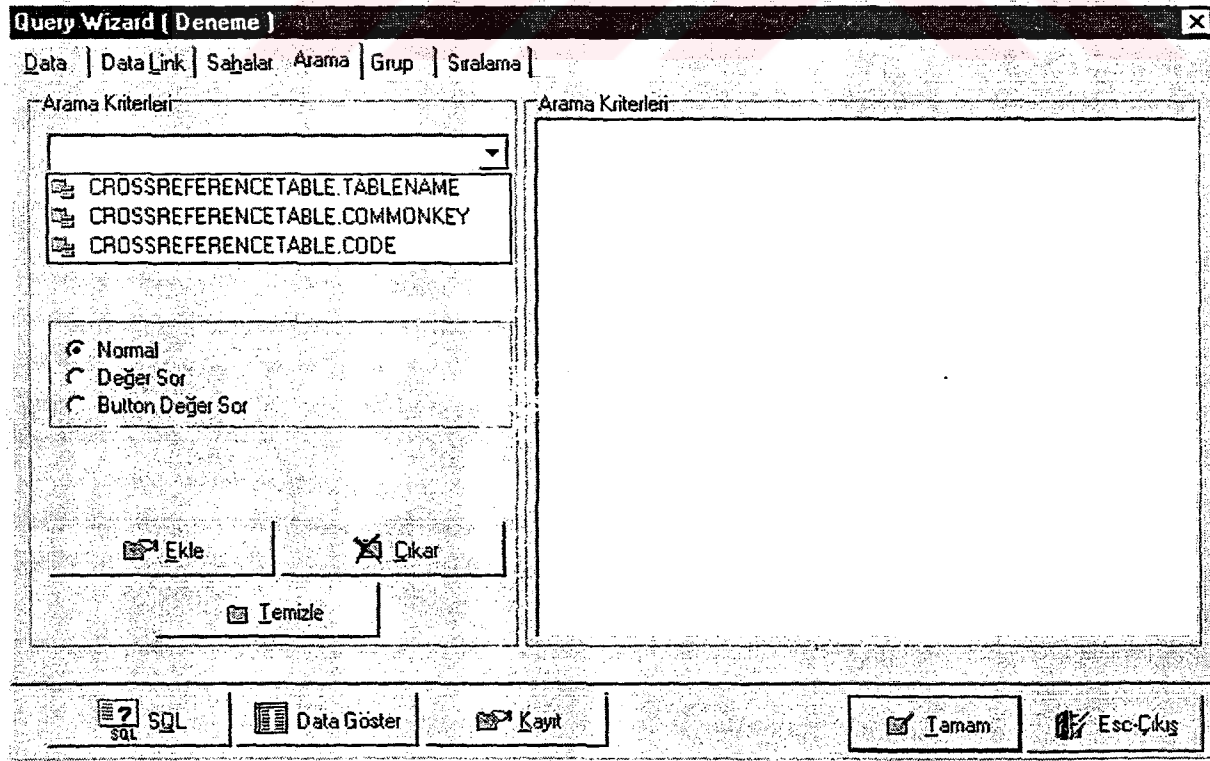
Şekil 12.7 YıldızETL uygulaması ExtractorWizard modülü ikinci kulakçık (page-tab 2)

Bir tablodaki bir alan (field veya domain) seçilerek, diğer bir tablodaki alanın sol üzerine bırakılmaktadır (drag-and-drop). Böylece oluşan dinamik SQL cümlesinin WHERE-cümlecisi oluşturulmuş olur. Sağlanan bu link silinebilir, hiç link yoksa WHERE-cümlecisi de olmaz.



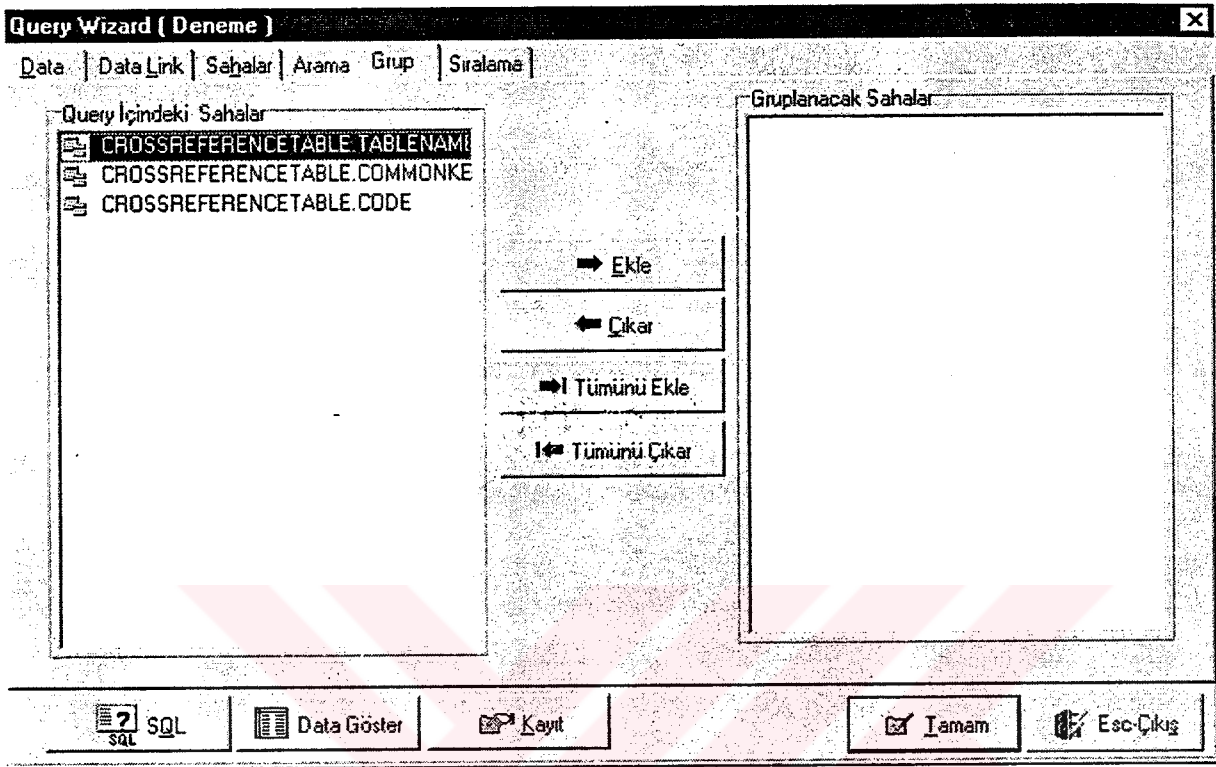
Şekil 12.8 YıldızETL uygulaması ExtractorWizard modülü üçüncü kulakçık (page-tab 3)

Dinamik yaratılan SQL cümlesinin SELECT-cümlecisi oluşturulmuş olur. İstenen tablolarının istenen alanları seçilerek sağdaki alan listesine aktarılır.



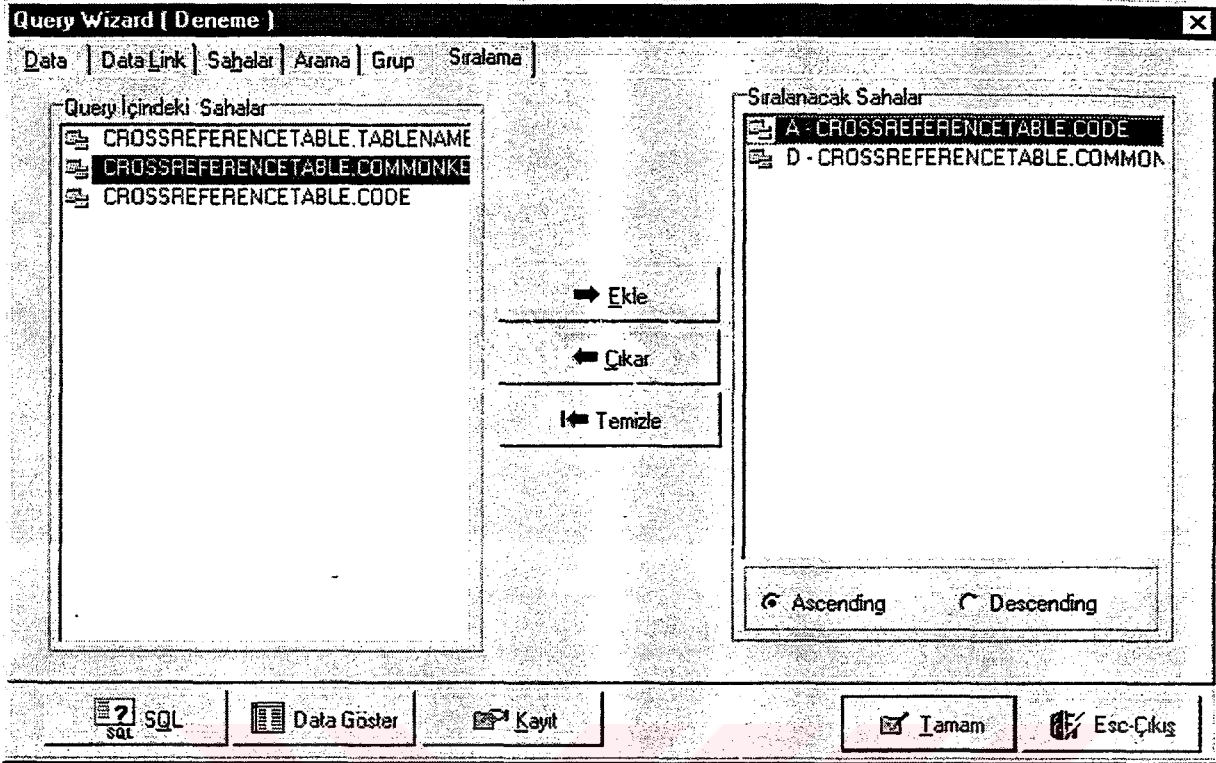
Şekil 12.9 YıldızETL uygulaması ExtractorWizard modülü dördüncü kulakçık (page-tab 4)

Dinamik yaratılan SQL cümlesinin WHERE-cümlecisi kısmına başka arama kriterleri eklenmek istenirse bu kulakçıkta gerekli işlemler ve seçimler yapılmalıdır.



Şekil 12.10 YıldızETL uygulaması ExtractorWizard modülü beşinci kulakçık (page-tab 5)

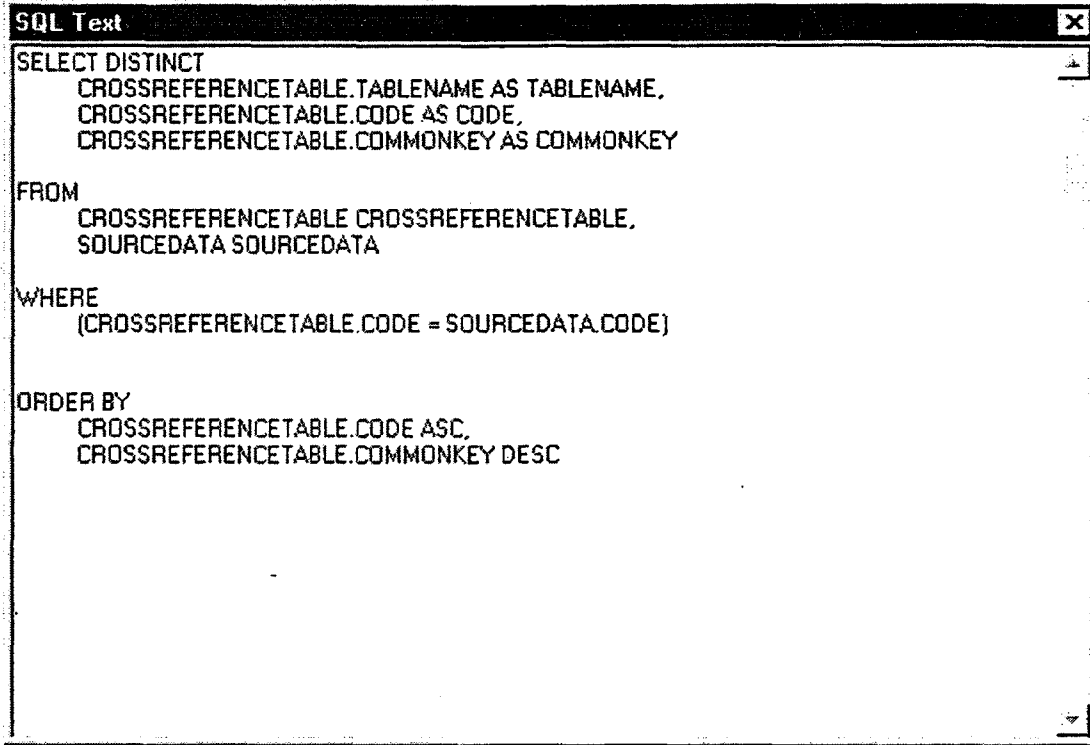
Dinamik yaratılan SQL cümlesinin GROUP BY-cümlecisi oluşturuluyor.



Şekil 12.11 YıldızETL uygulaması ExtractorWizard modülü altıncı kulakçık (page-tab 6)

Dinamik yaratılan SQL cümlesinin ORDER BY-cümlecisi oluşturuluyor. Sağda sıralanacak alanlar kısmında her bir alanın solunda “A-Ascending order” ve “D-Descending order” anlamında tek karakterlik bir açıklama karakteri söz konusudur.

Oluşturulan dinamik SQL cümlecisi Şekil 12.12’de görüldüğü gibidir.



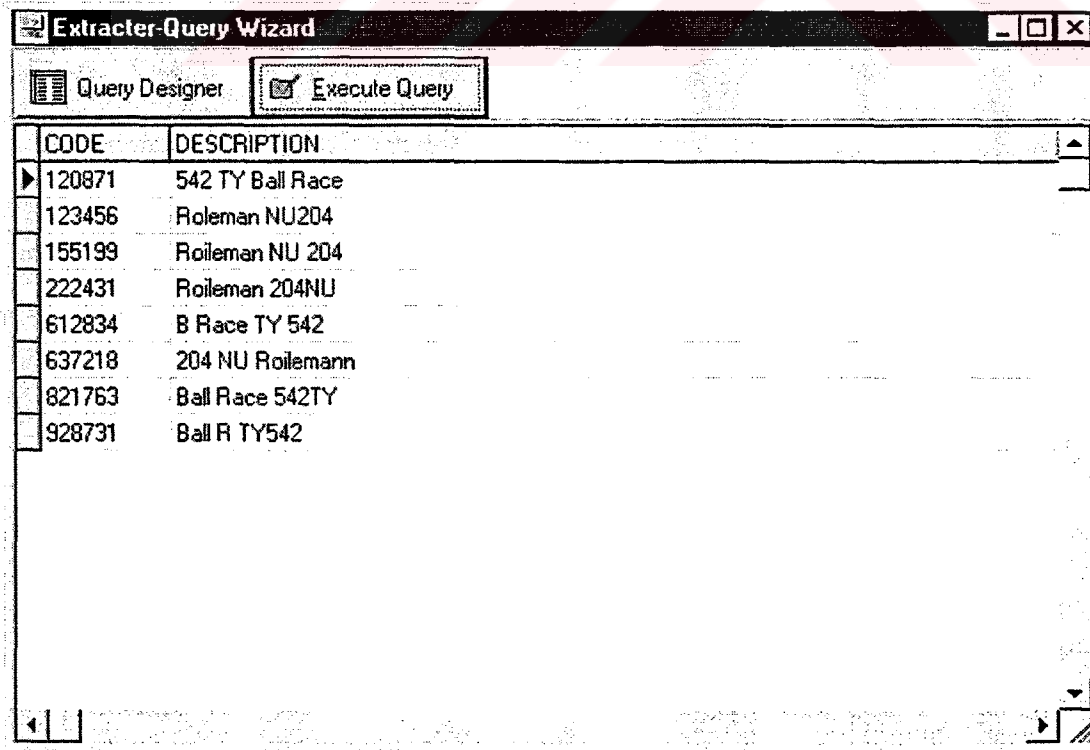
```

SQL Text
SELECT DISTINCT
  CROSSREFERENCETABLE.TABLENAME AS TABLENAME,
  CROSSREFERENCETABLE.CODE AS CODE,
  CROSSREFERENCETABLE.COMMONKEY AS COMMONKEY
FROM
  CROSSREFERENCETABLE CROSSREFERENCETABLE,
  SOURCEDATA SOURCEDATA
WHERE
  (CROSSREFERENCETABLE.CODE = SOURCEDATA.CODE)
ORDER BY
  CROSSREFERENCETABLE.CODE ASC,
  CROSSREFERENCETABLE.COMMONKEY DESC

```

Şekil 12.12 YıldızETL uygulaması modülü dinamik oluşturulan SQL cümleciği görünümü

Üstte görülen SQL cümleciği “Kayıt” tuşuna basılarak çalıştırılır. Aşağıdaki ekran görüntüsünde olduğu üzere oluşturulan sorgunun (query) sonucu listede gösterilmiştir. Bu seçme prosesinin bir başarılı simülasyonudur.



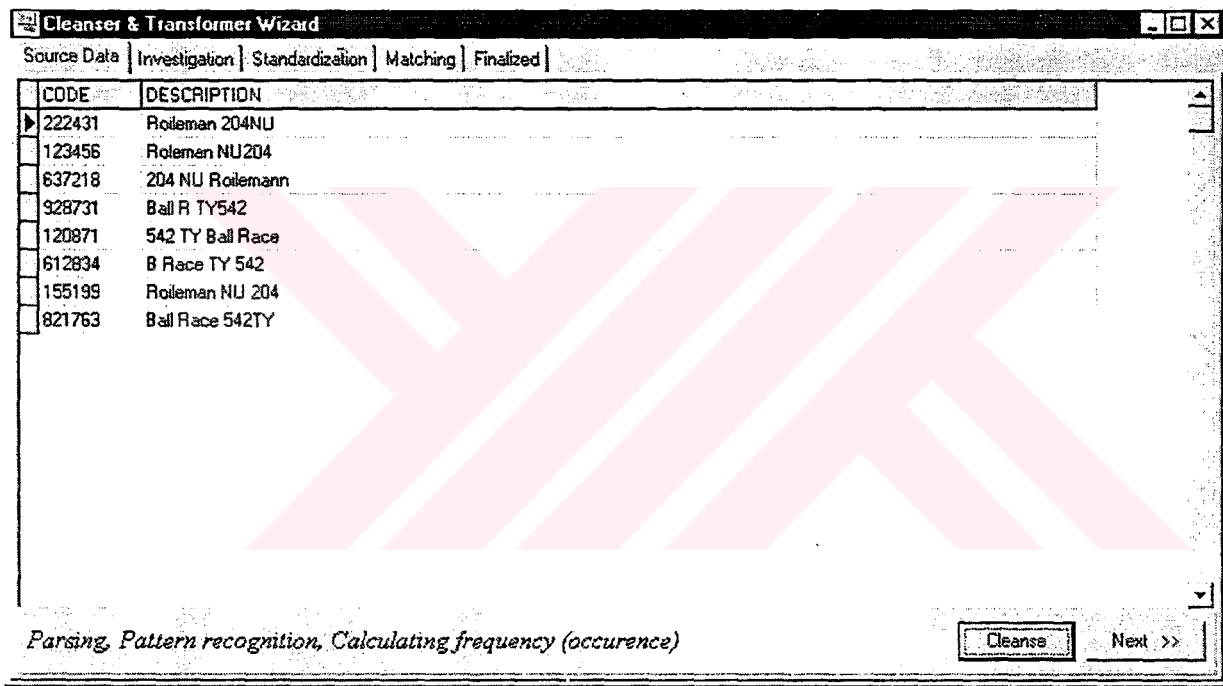
CODE	DESCRIPTION
120871	542 TY Ball Race
123456	Roileman NU204
155199	Roileman NU 204
222431	Roileman 204NU
612834	B Race TY 542
637218	204 NU Roilemann
821763	Ball Race 542TY
928731	Ball R TY542

Şekil 12.13 YıldızETL uygulaması dinamik oluşturulan SQL cümleciği çalıştırma sonucu

Seçme (ETL'in E'si) başarılı bir şekilde simüle edilmiştir. Şimdi Dönüştürme ve temizleme proseslerini inceleyelim.

12.5 “YıldızETL” Uygulaması (Cleanser & Transformer Wizard)” Modülü

Temizleme-Dönüştürme sihirbazı (Cleanser-Transformer Wizard) ise Extractor ile seçilen ve kalitesiz kabul edilen veri tablosu (SourceData) anahtar (key) ve tanım (description) alanları olarak ele alınır. Temizleme ve Dönüştürme sihirbazının ilk kulakçığı aşağıdaki şekilde (Şekil 12.14) gibidir.



Şekil 12.14 YıldızETL, Cleanser & Transformer Wizard modülü birinci adımı (Investigation)

Bu adımda öncelikle temizleme “Cleanser” düğmesine basılmalıdır. Bu sırada doğru bilgi içerdiği zorunlu olarak kabul edilen referans tablosundaki verilere göre ham verinin doğru karşılığı atanır. Tanım alanındaki her bir anlamlı kelime (token) ayrıştırılarak (parsing), kullanım sıklığı (frequency-occurrence), veri tipi, vb. bilgiler elde edilir. Daha sonra (investigation), bir önceki aşamada bulunan “pattern tag” dediğimiz işaretleyiciler (pattern recognition) sayesinde (maximum token sayısı + 1) adet alandan (field) oluşan standartlaştırma (standardization) tablosu dinamik SQL cümlesi (CREATE TABLE.....) ile oluşturulur.

“Next” düğmesine basıldığında gerçekleşen bu işlemlerin sonucunu, aşağıdaki şekildeki gibi (Şekil 12.15) görebiliriz.

Cleanser & Transformer Wizard							
Source Data Investigation Standardization Matching Finalized							
Code	Description	Pattern Tag	# of Tags	Table Name	Token Word	# of Occurrence	Pattern
120871	542 TY Ball Race	NCCC	4	SourceData	Roileman	2	C
123456	Roileman NU204	Cc	2	SourceData	204NU	1	c
155199	Roileman NU 204	CCN	3	SourceData	Roileman	1	C
222431	Roileman 204NU	Cc	2	SourceData	NU204	1	c
612834	B Race TY 542	CCCN	4	SourceData	204	2	N
637218	204 NU Roilemann	NCC	3	SourceData	NU	2	C
821763	Ball Race 542TY	CCc	3	SourceData	Roilemann	1	C
928731	Ball R TY542	CCc	3	SourceData	Ball	3	C
				SourceData	R	1	C
				SourceData	TY542	1	c
				SourceData	542	2	N
				SourceData	TY	2	C
				SourceData	Race	3	C

C Full character
c Alphanumeric (starting with numeric)

N Full numeric
C Alphanumeric (starting with character)

Frequency Pattern Table

Şekil 12.15 YıldızETL, Cleanser&Transformer Wizard modülü ikinci adımı (Standardization)

Parçalan ve elde edilen her bir token için; bu ekrandaki büyük “C” harfi “token”un tamamen alfabetik karakter ve “N” ise tamamen nümerik alanları göstermektedir. Küçük “c” harfi ise hem nümerik, hemde alfabetik karakter içeren token’lar için kullanılmıştır. Elde edilen bilgiler detayları ile birlikte dinamik olarak “Frequency pattern” tablosuna aktarılır. Algoritmik işlemlerin kabiliyetini ve olması gereken durum simüle edbilmek için sağ alt köşede iki düğme konmuştur.

Standartlaştırma adımı ekranı aşağıdaki şekildeki (Şekil 12.16) gibidir.

Cleanser & Transformer Wizard									
Source Data Investigation Standardization Matching Finalized									
FIELD1	FIELD2	FIELD3	FIELD4	FIELD5	TableName	Token Word	# of Occurrence	Pattern Type	Column #
Ball	Race	542	TY	120871	SourceData	Roileman	2	C	-1
Roileman		204	NU	123456	SourceData	204NU	1	c	-1
Roileman		204	NU	155199	SourceData	Roileman	1	C	-1
Roileman		204	NU	222431	SourceData	NU204	1	c	-1
B	Race	542	TY	612834	SourceData	204	2	N	-1
Roilemann		204	NU	637218	SourceData	NU	2	C	-1
Ball	Race	542	TY	821763	SourceData	Roilemann	1	C	-1
Ball	R	542	TY	928731	SourceData	Ball	3	C	-1
					SourceData	R	1	C	-1
					SourceData	TY542	1	c	-1
					SourceData	542	2	N	-1
					SourceData	TY	2	C	-1
					SourceData	Race	3	C	-1
					SourceData	B	1	C	-1
					SourceData	542TY	1	c	-1

Matching Frequency & Pattern Table Algoritmik Doğrusu

Şekil 12.16 YıldızETL, Cleanser&Transformer Wizard modülü üçüncü adımı (Matching)

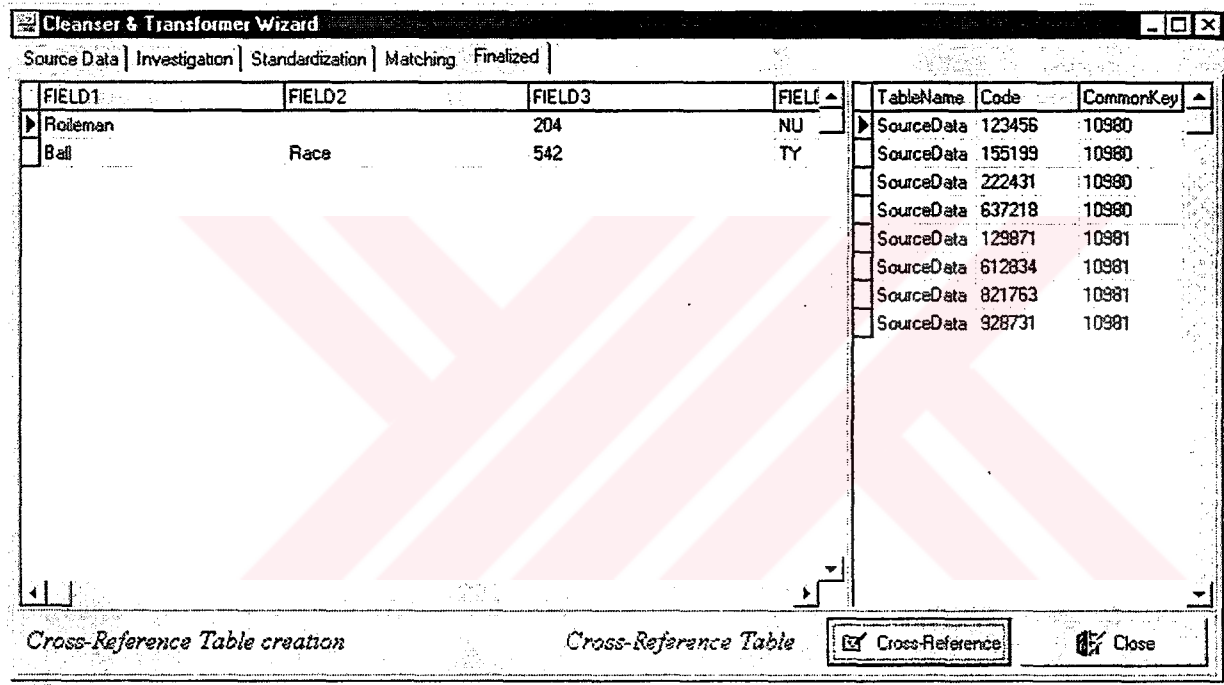
Veri eşleştirme (matching) yapabilmeye daha uygun bir yapıya geçmiş bulunmaktayız. Önce standartlaştırma yapılmalı, sonra bu veri üzerinden eşleştirmeye geçilmelidir. Standartlaştırma da field0-----fieldn kadar alana bilgiler aynı olacak şekilde taşınmaya çalışılır (Şekil 12.17).

Cleanser & Transformer Wizard							
Source Data Investigation Standardization Matching Finalized							
FIELD1	FIELD2	FIELD3		TableName	RawData	CorrectData	CommonKey
Roileman		204		SourceData	Roileman	Roileman	10980
Roileman		204		SourceData	Roileman	Roileman	10980
Roileman		204					
Roilemann		204					
Ball	Race	542					
B	Race	542					
Ball	Race	542					
Ball	R	542					

Consolidation, code unification Correct Data Library Next >>

Şekil 12.17 YıldızETL, Cleanser&Transformer Wizard modülü üçüncü adımı (Finalization)

Eşleştirilmiş bilgi ile ekran aşağıdaki gibidir. Eşleştirme esnasında anahtar alanları ortak anahtar alanlar ile eşleştirebilmek gerekmektedir. Böylece “code unification” dediğimiz prosesle; aynı olduğunu tahmin ettiğimiz bir kayıttan kaç tane olursa olsun (data redundancy veya duplication) tek bir kayıda indirgenmesi işleminin yapılabilmesi gerekmektedir. Ancak; uygulama matching (eşleme) konusunda zayıftır. Sadece ve sadece bu konu üzerinde dünyada bilinen 900 den fazla algoritma olduğu (tabiki İngilizce tabanlı eşleme) ve bunların da çok yüksek rakamlara (milyon dolarlara) kaynak kodları ile verilebildiği düşünülürse zorluk daha da anlaşılabilecektir.



Şekil 12.18 YıldızETL, Cleanser&Transformer Wizard modülü son adımı (Finalization)

Eşleştirilmiş (ki %100 eşleştirme çok zordur) kayıtlar kod birleştirme (consolidation) veya referans kullanma yardımıyla ortak anahtar (common key) kullanılarak tekil kayıtlara indirilmektedir (finalization-Şekil 12.18). Bunun yapılabilmesi çapraz-referans (cross-reference) tablosu oluşturulabilmesi demektir. Bu da, artık her yeni gelen bilginin sadece karşılığının burada olup olmadığını kontrol edecek ek programlar ile iteratif bir sistem kurulabileceğini gösterir. YıldızETL aracı, veri-referanslı bir araç fonksiyonlitesi göstermektedir. Artık veri ambara yüklenmeye hazırdır. “Correct data library” ve “cross-refernce table” iki önemli tablodur. Bunların (ve bu tip tabloların) beslenmesi oldukça zor, zahmetli bir iştir. Benzer ETL uygulamalarında 25-30 yıldır biriktirilen bilgiler kullanılmaktadır ve biriktirme sürekli devam etmektedir.

13. SONUÇLAR, TARTIŞMA KONULARI ve ÖNERİLER

Sonuç ve öneriler için; her bölümde incelenen konuları teker teker ele alıp, bunlar hakkında değerlendirmeler sunmak uygun olacaktır. En sonunda da bu değerlendirmeleri bütünleşik bir şekilde özetlemek gerekecektir.

Birinci bölümde; tez konusu ile ilgili önceki çalışmaların içeriği, kullanılan yöntemler, önceki çalışmalar, geliştirilen ve önerilen teknik ve algoritmalar ile simülasyon yapan uygulama programı özet bir şekilde anlatılmıştır.

İkinci bölümde; tez konusu ile ilgili terminolojik jargonlar veriden, bilgiye ve sisteme kadar oldukça geniş bir çerçevede tanımlanmaya çalışılmıştır. Sürekli anlamları karıştırılan veri, bilgi, anlamlandırılmış bilgi farklılıkları ve aralarındaki hiyerarşik ilişki tanımlanmış ve değer piramidi şeklinde bir üçgen şablonda simüle edilmiştir. İş akıllı ve veri ambarcılığının bugünkü teknolojiye nasıl ulaştığını “Bilgiye Erişimin Tarihçesi” başlığı ile, manuel ortamdaki internet uygulamalarına kadar her türlü teknoloji tarihi perspektifi içinde kronolojik sırada detaylı olarak aktarılmıştır.

Üçüncü bölümde; veri ambarı, data mart ve veri madenciliği konuları incelenmiştir. Veri ambarının kavramsal seviyede çeşitli tanımları yapılmış, tarihçesi incelenmiştir. İş sistemleri ve veri ambarları arasındaki farklar incelenmiştir. Örneğin, yapılan araştırmalar neticesinde eğer bir veritabanında 22 tablodan daha fazla tablo var ise kullanıcılar verileri nereye konumlandıracağı noktasında zorluklar yaşamaktadır. Bu durum kullanıcının değerlendirme kapasitesine uygun ve özet bir tarzda, verilerin iş sistemlerinden veri ambarına aktarılması gerekliliğini ortaya koymaktadır. Bu noktada yapılacak tasarım ve verinin sunum şeklinin önemi ortaya çıkmaktadır. Veri ambarı, data mart ve veri madenciliği uygulamaları; harcama-maliyet, zaman ve karlılık-fayda oranı şeklinde üç temel kriterde incelenmiştir. Ayrıca veri ambarı ve data mart uygulamaları arasındaki farklar özet bir tablo şeklinde sunulmuştur. Veri ambarı ve data mart oluşturma ve devreye alma metodolojileri incelenmiş, dikkat edilmesi gereken kriterler ayrıntıları ile açıklanmıştır. Veri madenciliği, OLAP ve bu konulardaki teknik ve işin kendisi ile ilgili cevaplanması gerekli 20 temel soru belirlenmiştir. Karar destek ve veri ambarı ilişkisi; planlama, bütçe tahmini, iş aklının sürekli değişen rolü, sınırları zorlama ve işbirliği dili açısından incelenmiştir. Dijital iş ve dijital zeka konseptleri tartışılmış ve bunun bu bölümde anlatılan diğer kavramlar ile ilişkileri sorgulanmıştır. Operasyonel Veri

Ambarı (OVA), işlem (transaction) ve anlık (snapshot) versiyonları ve hangi durumlarda hangi versiyonun kullanılması gerektiği ayrıntılandırılmıştır. Belkide en önemli sorularda biri diyebileceğimiz; “Bir veri ambarı projesine girerken nelere dikkat etmeliyiz?” sorusuna cevap verilemeye çalışılmıştır. Veri ambarcılığı ve web ilişkisi ise “yeni ekonomi”, “internet ekonomisi”, “e-iş” uygulamalarında ve veri ambarında dikkat edilmesi gerekli hususlar sınıflandırılmıştır. XOLAP başlığı altında ise en çok karıştırıldığı bilinen OLAP kavramları terminolojik açıdan incelenmiştir. Bu bölümde incelenen kavramsal nesnelerden hangisi gerçek hayatta çalışan gerçek bir proje halini alırsa alsın, öncelikle işin, kullanıcının ve yönetici ekibinin entegre bir şekilde düşünüp, tüm analiz ve tasarım çalışmasının buna göre yapılması gerektiği sonucuna varmaktayız. Bu sonuca giden aşamaların detayları ve dikkat edilmesi gerekli tüm hususlar bölüm içinde sırasıyla sunulmuştur.

Dördüncü bölümde; “İş Kuralları Sistemi” yaratma konusu veri-temelli olarak anlatılmıştır. İş mantığından başlayarak E/R modelleri ve “Kural-kalıp (rule-pattern)” modellerine kadar ayrıntılı bir şekilde incelenmiştir. İş kuralları sistemi, metodolojisi ve faydaları açısından irdelenmiştir. İş kuralı, sınıflandırması yapılmıştır. İş kuralı sisteminin metodolojisinin uygulaması ve uygulama fazları belirtilmiştir. Kural analizi, IF / THEN sorgulamaları ile iş kuralı yaratma, işletme ve sakıncaları hususları detayları ile anlatılmıştır. Kural analizi için, kural kalıpları tek yöntemdir. Eksik kısıtlama kontrolleri de eklenerek kural tamlığı zenginleştirilmek istenirse “Kural-zenginleştirme mantıksal modeli” uygulanmalıdır. Uygun kural kalıbı için; IF kısmına yazılan kolonlarda NULL bilgi, yani belirsiz bir kural olmamalıdır. Zenginleştirme hem tüm veri için geçerli bir son aşama prosesi olsa da kurallar içinde uygulanması zorunlu bir prosestir. Hem hatalar minimuma indirilir, hem de %100’e çok yakın bir veri kalitesine gidişin başarıyla sonuçlandırılmasına hizmet eder. Bölüm sonunda da toplam tasarımda yapılması gerekenler bir çizelge halinde özetlenmiştir. Bu çizelgede ürün seçiminden, veri ve kurallar üstündeki tecrübeye kadar yapılması gerekenler emirsel formda sunulmuştur.

Beşinci bölümde; müşteri ilişkileri yönetimi (CRM) konusu ele alınmıştır. CRM tanım ve içeriği açısından oldukça karıştırılan bir konu olduğundan, bu bölümde önce ne olmadığı tarif edilmiş, daha sonrası ise ne olduğu örneklemelerle açıklanmıştır. CRM en basit tanımıyla, değişimdir. Müşteri merkezli olmak için organizasyonlar kendilerini değiştirmelidir. Bu yeni bir yazılım yüklemek kadar kolay değildir. Kontrolü ve yükü azaltıp dağıtarak, şirketin diğer alanlarındakileri kişilerle bilgi paylaşımını sağlamaktır. Bu organizasyonu geliştirir.

Organizasyon kültürünün gelişmesi, bireyseliğin azalmasını ve işbirliğinin artmasını sağlar. Bunların hepsi müşterileri daha verimli yönetmek için organizasyonları değiştirmektir. “Müşteri Deneyim Döngüsü” grafiksel ve de yazılı ifadeler ile anlatılmış, müşteriye tanıma da nelere dikkat edilmesi gerektiği incelenmiştir. Müşteri verisinin entegrasyonu, CRM’in yani aslında uygulayan firmanın başarısını en yüksek seviyeye çıkarması, ayrıca bütünleşik bir bakış ve tutarlı sonuçlar üretilmesi yönünden oldukça önemli bir husustur. Veri ambarı ve CRM uygulamalarının çok hızlı büyüdüğü düşünülür ise geçmişini bilmek, mevcut durumu analiz etmek, proses ve modelleri ölçerek ve de rafine ederek yüksek doğruluk oranı ile geleceği tahmin etmek metodolojik yaklaşım olarak benimsenmelidir. Bir CRM uygulaması bu sonuçları desteklemelidir. Nasıl olması gerektiği ile ilgili örnekler grafikler ile bölüm içinde detaylı olarak anlatılmıştır.

Altıncı bölümde; iş aklı uygulaması yatırımlarının geri dönüşümün, yani bütçelenip maliyet-fayda düzleminde rakamsal değerlendirmeye nasıl ulaşılacağı anlatılmıştır. Finansal fayda metodlarından en uygun metod ROI (Return On Investment) olarak seçilmiştir. Direkt ve özellikle de endirekt maliyetlerin nasıl hesaplanacağı, nelere dikkat edileceği uygulama yılı artı üç yıllık bir sürede, bir BI (business intelligence) örnek projesi ele alınmış ve hesaplama tekniği açıklanmıştır. Sonuçta bulunan değerde ROI yüzdesi sıfırdan ne kadar yüksek ise uygulama kuruma o kadar fayda getirecektir. Diğer durumlarda zaten zarar getireceği düşünüldüğünden, projeye başlamama kararı çıkacaktır. Ayrıca duyarlılık analizi ile olasılıkların nasıl detaylandırılacağı konusu ve meta data ambarının maliyeti azaltmaya katkıları aktarılmıştır. Ortaya konan mali sonuçların yöneticiler sunumu ile ilgili, örnek BI uygulamaları da şekillerle anlatılmaya çalışılmıştır.

Yedinci bölümde; en basit tanımıyla “veri hakkında veri” olarak tanımlanan meta data konusu işlenmiştir. OMG (Object Management Group) tarafından geliştirilmiş olan CWM (Common Warehouse Metamodel) anlatılmıştır. Sektörel en büyük yazılım firmalarının oluşturduğu bu standart tüm meta data modelleri için kullanılmalıdır. UML grafiksel dizayn metodolojisi de yine geleceğin dünyasında ortak bir ortam, ürünler arası ortak bir iletişim dili olarak kabul görecektir. Ayrıca yine bu bölümde metadata detaylı olarak kelime kökünden, tiplerine kadar incelenmiş, karşılaştırmalar yapılmıştır. Jenerik bir meta data ambarı modeli önerilmiş, bu modelde önerilen tablolar kullanımları açıklanmıştır. Bu model projeden projeye aynen aktarılıp bir kaç küçük modifikasyonla uygulanabilecek durumdadır. Özetle, meta data modeli olmaksızın bir iş aklı ve veri ambarı uygulaması düşünülemez.

Sekizinci bölümde; kaynak entegrasyonu, veri entegrasyonu ve uygulama entegrasyonu konuları ele alınmıştır. Kaynak entegrasyonu dahili (kurum içi) ve harici (kurum dışı) olmak üzere ikiye ayrılmaktadır. Buradaki problem hem iç, hemde dış verinin kendi içinde doğru ve tutarlı olmasından öte, bu ikisinin de hem uyum hemde tutarlı olmasını sağlamak gerekir ki, bu durum çözülmesi ve sürekliliğinin sağlanması çok zor olan bir konudur. Veri entegrasyonu veri ambarının temelidir, hatta bir çeşit tanımı da diyebiliriz. Veri kalitesinin maksimize edilmesi için temizlenmesi, zenginleştirilmesi ve de konsolide (entegrasyon sonucu bütünleştirme) edilmesi gerekmektedir. Efektif meta data yönetimi de gereklidir. Veri entegrasyonunun tam sağlanabilmesi için hem teknik hemde iş meta data verileri meta data ambarında tutulmalı ve çok iyi işletilip yönetilebilmelidir. ERP, B2B, B2C örneklemeleri ve bunlar için gerekli entegrasyon işlevleri de incelenmiştir. Uygulama entegrasyonunun tüm diğer uygulamalar gibi kurumsal çapta düşünülüp planlanması gereklidir ki, buna “kurumsal uygulama entegrasyonu (EAI)” adı verilmektedir. Bu en önemli ve teknolojik yönden büyüyen pazarı temsil eden entegrasyon şeklidir. Bunun için meta data ambarı olmazsa olmazlardandır. Tekrar meta datanın önemine değinildi, çünkü meta data bir binadaki kolonlar hüviyetindedir. Veri ve model seviyesi olmak üzere iki metodoloji kullanılmaktadır. Eğer bir uygulama mantığı veya sadece bir yerden bir yere veri taşınması ve birleştirilmesi gibi içinde algoritmik bir fonksiyon içermeyen işlemler için “veri-seviyesi metodu” kullanılır. En etkin ve gerekli olanı açıkça anlaşıldığı “metod-seviyesi metodu”dur. Detay karşılaştırma bölümde anlatılmıştır. Lakin burada vurgulamak istediğimiz toplu (batch) işlerde veri-seviyesi metodunun ve de muhakkak uygulama mantığı da içereceğinden metod-seviyesi metodu ile birlikte kullanılmasının daha efektif sonuçlar ortaya koyacağıdır.

Dokuzuncu bölümde; veri kalitesi incelenmiştir. Veri kalitesi konsepti, kurumlara getireceği değişiklik ve bunun yaratacağı sorunlar ve bu konularda yapılmış araştırma sonuçları incelenmiştir. Bu sonuçlar; bilginin doğruluğu, çıktı zamanlaması, güvenilirlik, tamlık, ilişkisel bütünlük, aşıkılık, kesinlik ve doğruluk üzerinde yoğunlaşmıştır. Veri tutarlılığı ve veri tekrarlarının ters, lakin veri tutarlılığı ve veri bütünlüğünün ise doğru orantılı olduğu grafiklerle şekillendirilmiştir. Veri kalitesizliğinin on bir ana konu ve noktadan niçin kaynaklandığı, bir tablo şeklinde özetlenmiştir. Ayrıca yine bunların istatistiksel metodlar (ortalama puan, standart sapma) ile tablosu sunulmuştur. Buna göre en önemli üç sebep; hiç değiştirilmeyen veri, heterojen sistem bütünleştirilmesi ile yönetimin politika ve planlama eksikliğidir. Veri kalitesi kalitatif bir kavram gibi gözükse de ölçülebilmektedir. Bu ölçümde; verinin tamlığı-bütünlüğü, veri girişinin tutarlılığı, verinin doğruluğu, anahtar verilerin

tekilliği ve veriyi destekleyen iş kurallarının sürekliliği parametre olarak kullanılmaktadır. “Alan istatistikleri metodu”, çözüm olarak incelenmiştir. Bu metot tolerans (kabul edilebilir veri aralıkları ve dışındaki bölgeler) ölçümlerinin temel kabul edildiği istatistiksel bir metodudur. Grafikler yardımıyla detaylı bir şekilde anlatılmıştır. Veri kalitesini otomatikleştirme ve “kurumsal veri kalitesi yönetimi (EDQM)” kriterleri ile bunun ölçülmesi konusu işlenmiştir. Temel EDQM problemleri; yanlışlar, eşsesliler, standard eksikliği, yasal zorunluluklar, kayıp/gizli veri ve hayali veri olmak üzere 6 ana kategoride Türkçe ve İngilizce örneklemelerle incelenmiştir. Veri kalitesi, veri temizleme araçları (tarihçeleri dahil) analiz edilmiştir. Veri kalitesinin gelir ve maliyetler (giderler) üzerine etkileri dahi, mali perspektif ile bakabilmek için incelenmiştir. Meta data, aynı zamanda entegrasyon ve veri temizleme ve zenginleştirme araçları ile birlikte kullanılıp; müşteri, adres ve ürün kayıtları ne kadar kaliteli hale getirilirse, hem bir o kadar gelir artışı, hem de bir o kadar gider düşüşü de yaşanır ve kar maksimizasyonu dediğimiz ideal ekonomik göstergesi (Maximum Revenue = Minimum Cost) yakayabiliriz. Bu bölümde güncel fakat bir o kadar da önemli bir olay (Amerika, 2000 yılı başkanlık seçimleri, Florida Eyaleti) veri kalitesi için örneklerden biri olarak seçilmiştir. Problemler, sebepleri ve toplumsal/siyasi bu olayın veri kalitesi açısından iş hayatına uyarlanması ve de veri kalitesi merkezli yorumlanması oldukça yararlı sonuçlar ortaya koymuştur.

Onuncu bölümde; veri temizleme konusu incelenmiştir. “Gerçek dünya verisi kalitesizdir” yaklaşımı ile konu incelenmiştir. Veri kalitesi veya en basit şekliyle kabul edilebilir veri doğruluğu; problemin kaynaklandığı veri giriş noktalarında, yani kökünde, çözümlenmesi gerekir mantığı ile konuya yaklaşılmıştır. Lakin girilen veri her yönden doğru kabul edilsede eğer yanlış alanlara girildiyse bunu hiçbir uygulama yakalayamaz ve doğal olarak teorideki ideal çözümlerde de kalitesizlik oluşur. İşte bu yüzden gerçek dünya verisi % 1 seviyesinde de olsa (kabul edilebilir oran %0.025 tir) kalitesiz kabul edilmelidir. Veri temizlemenin yoğunlaşacağı konular tüm verilerden önce; müşteri, müşteri adres, ürün, ürün lokasyon şeklinde olmalıdır. Bunlar aslında ticaretin temel komponentleridir. Bunlar nakit gelir ve gideri belirleyen unsurlardır. Bir veri kalitesi sözlüğü (knowledge base repository) kullanarak, yani doğruluğu referans kabul eden tablolar kullanarak, özellikle kullanıcı tanımlı hem teknik hemde işsel kurallar tanımlayabilmek ve de iteratif olarak sürekli tekrarlanabilen bir metodolojik yaklaşımlar sentezi ile veri kalitesi problemlerini çözebilmek mümkündür. Temizleme prosesini ögeleştirme (elementizing), standartlaştırma (standardizing), doğrulama (verifying), eşleştirme (matching), ev sahipliği (householding) ve belgeleme (documenting)

olmak üzere altı değişik adımda inceleyebiliriz. Literatürden literatüre isimlendirmeler değişse de, konsept olarak aynı şeyi ifade etmektedirler. Ayrıca bölüm sonunda veri temizleme araçlarının seçim kriterleri de çok boyutlu olarak ele alınmıştır.

On birinci bölümde; verinin, veri ambarına aktarılma prosesi ve bu prosesi en güzel şekilde tanımlayan ETL (Extract Transform Load – Seçme Dönüştürme Yükleme) konusu ele alınmış ve bunu gerçekleştiren ETL araçları incelenmiştir. Öncelikle veri toplama teknikleri ve bu noktada yaşanan problemler incelenmiştir. ETL diğer bölümlerde aktarılan problemlere komple çözüm sağlayabilecek yegane araçlardandır. Fakat meta data, veri ve uygulama entegrasyonu, iş kurallarına uygun modelleme, vb. araçlar ile ortak kullanılmalıdır. Çünkü veri kalitesi maksimizasyonu son derece zor bir prosestir. İşte bu yüzden ki, milyarlarca dolarlık pazarı teşkil eden bu ürünlere şirketler milyonlarca dolar harcamaktadırlar. Maliyetleri düşürmesi ve daha bir çok sebepten dolayı internete yönelinmesinden ötürü de, veri hacmi çok büyümüş ve problem daha da işin içinden çıkılmaz hale gelmiştir. ETL veri-seçmeden, metadataya kadar bir çok fonksiyonu içinde barındıran bir fonksiyonlar kümesidir. İteratif olması gereken ETL prosesi çeşitli alt adımlardan oluşur. Bu araçların kullanımı ile veri kalitesi problemleri, koruyucu hekimlik de olduğu gibi daha oluşmadan çözülebilir. Bu araçlar; veri profili belirleme araçları, değişim mühendisliği araçları ve ETL araçları. Ayrıca yine bu bölümde ETL araçlarının tarihsel gelişimi (jenerasyonları), ölçeklenebilirliği ve seçim kriterleri de incelenmiştir.

On ikinci bölümde; yapılan örnek uygulama anlatılmıştır. Bu örnek uygulama basit bir ETL ürünüdür, hatta simülatörüdür. Geliştirilen teknik ve algoritmaların simüle edildiği, uygulandığı bir ETL programıdır. Fonksiyon adımları olarak bakıldığında ETL'in temel fonksiyonlarının yerine getirir. Ancak, çok geniş kapsamlı veri sözlükleri (isimi adresi cinsiyet, kısaltmalar, vb.) kullanmamakta, “ses benzeşimi” ve “iş mantığı analizi” de yapmamaktadır. Ancak bunların her biri son derece ağır ve uzmanlaşma istediğinden çok uzun zamanlı uygulamalardır. Kaldı ki;örnek vermek gerekirse iş akıllı ürünleri üreten bir firma, sadece kurumsal uygulama entegrasyonu (EAI) konusunda dünyanın en iyisi olduğunu iddia ederken 700-adam yılda (filen 15 yılda) bu noktaya geldiğini belirtmektedir. EAI sadece konunun bir parçası olduğu düşünülür ise kapsam ve büyüklük daha da net anlaşılacaktır. Uygulama matching (eşleme) konusunda zayıftır. Sadece ve sadece bu konu üzerinde dünyada bilinen 900 den fazla algoritma olduğu (tabiki İngilizce tabanlı eşleme) ve bunların da çok yüksek rakamlara (milyon dolarlara) kaynak kodları ile verilebildiği

düşünülürse zorluk daha da anlaşılacaktır. Uygulama, seçme (extraction) konusunda oldukça yeteneklidir. Çoklu-ortam veri tabanlarında problemsiz çalışmaktadır. MVS mainframe CICS, WinNT ve AS/400, Interbase, Oracle ve bilumum masaüstü lokal veri tabanları ile çalışabilmekte hem aracısız (native), hem de duruma göre aracılı (ODBC) kullanılabilmekte; dinamik SQL cümlecikleri oluşturup, verinin ambara alınmasını simüle etmektedir. Bu bir sihirbaz (Extractor Wizard) şeklinde yapılmıştır. Temizleme-Dönüştürme sihirbazı (Cleanser-Transformer Wizard) ise Extractor ile seçilen ve kalitesiz kabul edilen veri tablosu (SourceData) anahtar (key) ve tanım (description) alanları olarak ele alınıp tanım alanındaki her bir anlamlı kelime (token) ayrıştırılarak (parsing), kullanım sıklığı, veri tipi, vb. bilgiler elde edilir. Daha sonra (investigation), bir önceki aşamada bulunan “pattern tag” dediğimiz işaretleyiciler sayesinde (maximum token sayısı + 1) adet alandan (field) oluşan standartlaştırma (standardization) tablosu dinamik SQL cümlesi (CREATE TABLE.....) ile oluşturulur. Veri eşleştirme (matching) yapabilmeye daha uygun bir yapıya geçiş demektir. Eşleştirilmiş (ki çok zor, ulusal dil işin içine girince çok daha karışık olmakta) kayıtlar kod birleştirme veya referans kullanma yardımıyla ortak anahtar (common key) kullanılarak tekil kayıtlara indirilmektedir (finalization). Bunun yapılabilmesi çapraz-referans (cross-reference) tablosu oluşturulabilmesi demektir. Bu da, artık her yeni gelen bilginin sadece karşılığının burada olup olmadığını kontrol edecek ek programlar ile iteratif bir sistem kurulabileceğini gösterir.

Kodlama sistemlerinin önemi ve bu konuda yapılan bir diğer saha inceleme çalışması:

1978’den beri yürütülen MERNİS (Merkezi Nüfus İşleri Sistemi) projesi formları (Derince/Kocaeli) ilçe nüfus müdürlüğünden temin edilmiş ve uygulama, bu teze esas konular kapsamında incelenmiştir. Her ilçe müdürlüğü lokal Oracle veri tabanı kullanmaktadır. Hafta da bir ilçe müdürlüğü kendi yedeklemesini yapmakta ve 15 günlük yedeğini (verisini) il müdürlüğüne, il müdürlüğü de ayda bir bakanlığa (Ankara) göndermektedir. Yani veri tazelenme süresi bakanlık merkezli olarak bakıldığında online değil aylık ve offline’dır. Bazı kodlamalarda, ki kodlama en önemli konulardan biridir, uluslararası kodlamalar (cinsiyet : 341-erkek, 342-kadın, 343-belirsiz veya 11: İslam, 9-ülke kodu) kullanılsada örneğin adres verileri uygulaması yurt çapında henüz hayata geçirilmemiş bulunmaktadır. Yapılan görüşmelerde son derece yüksek (%15 lik) bir hata oranı olduğu, verilen kimlik numarası sisteminin yeterince işletilemediği belirtilmiştir. Daha kimlik verisini halletmeden adres verisine geçmenin uygun olmayacağını belirten uzmanlar; bu nokta da hata veya yanlış riskinin veya tutarsızlık oranının toplam da maksimum %25 i bulabilebileceği

vurgulanmıştır. Ayrıca zaten bu konuda herhangi bir kodlama çalışması yapılmadığını free-text yazılacağını aynı mahallede (ki maalesef kodlama sistemi henüz yok) 10,000 kişi oturduğu varsayılrsa ve de (bir çok sebepten ötürü) mahalle isim değişikliği yapılırsa, adres bilgilerinin teker teker bütün kimlik kayıtlarından güncellenmesi gerektiği bunun ise tutarsız kayıtlar yaratacağı daha başlanmadan ortaya konulabilmektedir. Nüfus uzmanı olan bu kişiler uzman bir BT personeli olmadıkları halde bunu dile getirmektedirler. Diyelim bir ilçede 35 mahalle var, (lokaldeki) mahalle tablosunda 35 kayıt olacaktır. Kod değil de isim değişse, kimlik kayıtlarında da sadece mahalle kodu bulunacağından hiçbir problem yaşanmayacaktır. Bunlar kayıt girişinde zorunlu alanlar olsa ve girişte doğrulamalar sağlıklı yapılırsa raporlamalar, analizler ve istatistikler çok sağlıklı sonuçlar üretilebilir hale gelir. Bu sağlıklı sonuçlar karar-yapıcıların sağlıklı ve de başarılı kararlar üretmelerini sağlayacak ve ulusal çapta planlama yapıp, çözüm geliştirme olanakları açacaktır (yatırımların geri dönüşümü, mali ve mali olmayan faydalar olarak da düşünülebilir).

Sonuç olarak; veri kalitesizliği en küçük uygulamadan en büyük uygulamaya kadar gerçek dünyada var olan bir problemdir. Iteratif olarak verileri sürekli temizlenmeli, zenginleştirilmeli ve bütünlük için entegre edilmelidir. İş kurallarına uygun, kodlama sistemi oturtulmuş, ortak bir meta data ambarı kullanan, veri giriş (kalitesizlik probleminin ana kaynağı) noktalarında doğrulamaları güçlü olan ve yanlış/hatalı kayıda müsaade etmeyen tıpkı koruyucu hekimlik gibi proaktif sistemler kurulur ise veri kalitesi %100'e yükseltilebilir.

KAYNAKLAR

Amit, R. ve Yeo, E., (1999), "Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia", IEEE 32nd International Conference on System Sciences, Hawaii, 1999,
<http://computer.org/proceedings/hicss/0017/0017012abs.htm>.

Arrow, K.J., (1984), The Economics of Information, 138, Harward University Press, Cambridge.

Awad, E.M. ve Gotterer, M.H., (1992), Database Management, Boyd & Fraser Publishing Company, Danvers, Massachusetts.

Bingle, M., (2001), "CRM-What is it ?", (12.02.2001)
<http://www.datawarehouse.com/iknowledge/articles/article.cfm?ContentID=1173>.

Bolten, R., (2000), "Data Population in a BI Environment", DM Review September 2000, s:22, 64.

Borland, R. (2000a), "Data Warehouse Operations Strategy and ETL Processing Part 1",
<http://www.datawarehouse.com/iknowledge/articles/article.cfm?ContentID=932>.

Borland, R. (2000b), "Data Warehouse Operations Strategy and ETL Processing Part 2",
<http://www.datawarehouse.com/iknowledge/articles/article.cfm>.

Borland, R. (2000c), "Data Warehouse Operations Strategy and ETL Processing Part 3",
<http://www.datawarehouse.com/iknowledge/articles/article.cfm?ContentID=973>.

Borland, R. (2000d), "Data Warehouse Operations Strategy and ETL Processing Part 4",
<http://www.datawarehouse.com/iknowledge/articles/article.cfm?ContentID=977>.

Business Rules Group (www.businessrulesgroup.org).

Calvanese, D., Giacomo, G.D., Lenzerini, M., Nardi D. ve Rosati, R., (1998), "Source Integration in Data Warehousing", IEEE 9th International Workshop on Database and Expert Systems Applications, <http://computer.org/proceedings/dexa/8353/83530192abs.htm>.

Chow, K.T., (2000), "Enterprise Data Quality: Covering all your Bases", DM Review December 2000.

Cipriano, F., (1995), The Impact of Information Systems on Quality Performance: An Empirical Study, International Journal of Operations and Production Management, Vol. 15(6), s:69-83.

Cobb, L, (1996), "Data Warehousing : The Latest Fad?", Management-Accounting-London, June 1996, Vol. 74(6), s:44-46.

Craig, R., (1997), "Data Warehousing and the Web", Oracle Magazine, September/October 97, s:11-12.

Data Junction, Corp., (1999), "Application Integration : Tactics to Data Transformation Nirvana", June 1999, www.datajunction.com.

Date, C.J., (2000), "What Not How: The Business Rules Approach to Application Development." Addison-Wesley Longman, Inc.

Delone, W.H. ve McLean, E. R., (1992), "Information Systems Success: The Quest for the Dependent Variable, Information Systems Research", 3:1, pp.60-95.

Distefano, J. (2000), "Information Strategy", DM Review, July 2001, s:28,45.

Dyché, J., (2000), "Data Warehouses, Metadata, and Middleware", eAI Journal September 2000, s:71-76.

Electronic Digital Documents, Inc., (2001), "Real-World Data is Dirty", <http://www.npsa.com/edd/>.

English, L.P. (1996), "Help for Data Quality", Information Week, <http://techweb.cmp.com/iw/600/00oldat.htm>.

English, L.P., (1998a), "Data Quality: Meeting Customer Needs", <http://www.pitneysoft.com/products/DQEng001.htm>.

English, L.P., (1998b), "DQ Point 6: Institute Training on Data Quality", DM Review May 98

English, L.P., (1998c), "The High Costs of Low Quality Data", DM Review, January 98.

English, L.P., (2000a), "KAIZEN and Information Quality", DM Review, January 2000.

English, L.P., (2000b), "Which Information Quality Tool is Best?", DM Review, October 2000.

English, L.P., (2000c), "Go to Gemba When Data Quality Problems Occur", DM Review May 2000, s:36,78.

First Logic, Inc., (2001), "Customer Data Quality : Building the foundation for a one-to-one customer relationship".

Firth, C., (1997), "When Do Data Quality Problems Occur?", <http://sunflower.singnet.com.sg/~cfirth/dq1.htm>.

Forino, R., (2000a), "Data e.Quality: The Data Quality Assesment, Part 1", DM Review August 2000.

Forino, R., (2000b), "Data e.Quality: The Data Quality Assesment, Part 2", DM Review September 2000.

Forino, R., (2001), "Data e.Quality: Election 2000", DM Review January 2001.

Galhardas, H., Florescu, D., Shasha, D., ve Simon, E., (2000), "An Extensible Framework for Data Cleaning", www.infira.fr/~galharda/cleaning.html (20 Nov 2000).

Galhardas, H., Florescu, D., Shasha, D., Simon, E., Matsumoto, J.P. ve Saita, C.A., (2001), "AJAX: An Extensible Data Cleaning Tool", www.infra.fr/~galharda/ajax-desc.html.

Giovinazzo, W.A., (2000), "Industry Standard-Introduction to CWM", Oracle Magazine July 2000.

Gordon, S.R. ve Gordon J.R. (1996), Information Systems-A Management Approach with Activities and Readings-International Edition, The Dryden Press, Texas.

Grandy, C., (1997), "Breakthrough Performance for Star Schemas", www.disc.com.

Hammer, K., (2000), "Data Integrity in e-Business",
<http://www.eaijournal.com/DataIntegration/DataIntegration.asp>.

<http://www.cwmforum.org/about.htm>.

Ilıkkın, A., (2000), "E-ticaret dünyasında başarının anahtarı : e-Business Intelligence", BThaber, Sayı :290, 16-22 Ekim 2000, s:29.

Information Discovery, Inc., (2001), "The Top 10 Data Mining Questions",
http://www.dmreview.com/portal_ros.cfm?NavID=92&WhitePaperID=68&PortalID=25&Topic=55.

Inmon, B., (1992-93), "Building the Data Warehouse", John Wiley & Sons.

Inmon, B., (2000a), "The problem with Dimensional Modeling", DM Review, May 2000, s:68-70.

Inmon, B., (2000b), "Different Data Warehouse Types", DM Review, July 2000, s:64-66.

Inmon, W.H., (2001), "Integrating Internal & External Data",
<http://www.eaijournal.com/DataIntegration/Integrating.asp>.

Innovative Systems, Inc., (2001), "How To Clean Dirty Customer Data Quickly And Economically", www.innovativesystems.net.

Jennings, M.F., (2000), "The Generic META DATA REPOSITORY", DM Review October 2000, s:55-58.

Jones, R., "Data Integrity – The Missing Element in ERP Success",
<http://www.eaijournal.com/DataIntegration/ShowMeTheData.asp>.

Jordan, A., (1997), "Warehouse Data Integrity: How Long and Bumpy the Road?",
<http://www.data-warehouse.com/resource/articles/jord8.htm>.

Kachur, R., (2000), "The Data Warehouse Diary: Data Quality Assessment for Data Warehouse Design", DM Review April 2000.

Kıral, C., 1997, "Bir veri ambarı projesine girerken nelere dikkat edilmelidir?", BT Haber, Sayı :113, s:62-63.

- Kimball, R., (1998a), "Bringing up Supermarts", DBMS January 1998, s: 47-53.
- Kimball, R., (1998b), "The Operational Data Warehouse", DBMS January 1998, s:14-16
- Kimball, R., (1996), "Dealing with Dirty Data", DBMS Sempter 1996,
<http://www.dbmsmag.com/9609d14.html>.
- Korth, H.F. ve Silberschatz (1991), Database System Concepts, Mc Graw-Hill, Amerika.
- Langan G. (2000), "The Langan Short of IT : Data is King !", DM Review, Mayıs.
http://www.dmreview.com/master_sponsor.cfm?NavID=68&EdID=2220.
- Linthicum, D.S., (2000a), "Creating Enterprise Metadata Repositories for EAI and e-Business", eAI Journal January 2000, s: 13-18.
- Linthicum, D.S., (2000b), "Data Level vs. Method Level The Design Patterns for EAI",
<http://www.eaijournal.com/DataIntegration/DataLevel.asp>.
- Marco, D., (1998), "Managing Meta Data", DM Review March 1998.
- Marco, D., (2000a), "Meta Data and Data Administration: Meta Data ROI: Making Your IT Department Better, Stronger, Faster", DM Review June 2000, s:44-45.
- Marco, D., (2000b), "Meta Data and Data Administration: Meta Data ROI: A Competitive Advantage", DM Review October 2000, s:32,34.
- McClanahan,D., (1996), "Making Sense of Enterprise Data" , Data Based Advisor, Nov 1996, Vol. 14 n11, s:76.
- McClain, D.S., (2000), "Customer Data Integration: The Essential Component of Effective CRM", DM Review June 2000, s:22,24,59.
- McGoveran, D., (2000). "Business Semantics", EAI Journal, October 2000, s:10.
- Meyer, S., (2001), "Which ETL Tool is Right for You?", DM Review March 2001.
- Michaels, D., (1997), "Data Mart / Data Warehouse Difference", Oracle Magazine, September/October 97, s:61.
- Miller, J.Y., (2000), "A Case for Scalable ETL", DM Review December 2000
- Montgomery, D., (2000), "Digital Business and Digital Intelligence", DM Review, s:67,87.
- Moragenthal, J., (1999), "Metadata : A Key to EAI",
<http://www.eaijournal.com/DataIntegration/MetaDataKey.asp>.
- No, A., (2000), "Frequently Asked Questions about Data Quality", DM Review November 2000.
- Nykamp, M. ve McEachern, C., (2000), "Customer Relationship Report: Understanding the Customer Experience Cycle", DM Review June 2000, s:38,67.

- Parkes, H.C., (1998), "What is the beef behind the OLAP blab?", DBMS, January 98, s:80.
- Parsaye, K., (1998), "A Characterization of Data Mining Technologies and Processes." The Journal of Data Warehousing, January 1998.
- Porter, J.D. ve Rome, J.J., (1994), "The Data Warehouse : 2 Years Later...Lessons Larned", Arizona State University, 1994 CAUSE Annual Conference, Orlando, FL.
- Rolleigh, L. ve Thomas, J., (2001), "Data Integration : TheWarekouse Foundation", DM Review, March 2001.
- SAS Corp., (2001), "Data Cleansing",
<http://www.sas.com/rnd/warehousing/quality/index.html>.
- Senn, J.A., (1984), Analysis and Design of Information Systems, Mc Graw-Hill, Newyork.
- Seeley, R., (2001), "Is CWMI the Holy Grail of Meta Data Standards?",
<http://www.eaijournal.com/DataIntegration/HolyGrail.asp>.
- Shah, A.D. ve Milstein B.M., (1997), "Data Warehousing: Practical Tips for Successful Implementation", <http://www.sw-expo.com/perfdev/data-warehousing.html>.
- Shash, F., (2000), "Managing Big Data : Building the Foundation for a Scalable ETL", DM Review December 2000.
- Silvon Software, (2000a), "Defining Data Warehousing-What is it and who needs it?".
- Silvon Software, (2000b), "A Top-Down View of Data Warehousing : A valuable toll in building business intelligence & analytic applications".
- Sterling Software, (1998),"VISION: Solutions for DataWarehousing", DM Review February 98,
<http://www.dmreview.com/portal.cfm?NavID=92&WhitePaperID=112&PortalID=10&Topic=5>.
- Strange, K., (1997), "Data mart'lardan veri ambarı doğarmı ?", Computerworld Türkiye, Ağustos 1997.
- Strehlo, K., (1996), "Data Warehousing: Avoid Planned Obsolescence", Datamation January 96, <http://www.datamation.com/PlugIn/issues/1996/jan15/01bsw200.html>.
- Tanler, B., (2000), "The Next Big thing For Data Warehouses", Dm Review, July 2000, s:60-64.
- Timur, Ö., (2000), "Veri Madenciliğinde Yeni Versiyon", InformationWeek Türkiye, Eylül 2000.
- Trillium Software System-White Paper EN TRIL0200, (1999), "Achieving Enterprise Wide Data Quality", www.trilliumsoft.com ve www.techguide.com.

Von Halle,B., (2001a), "Building a Business Rules System, Part 1", DM Review January 2001.

Von Halle,B., (2001b), "Building a Business Rules System, Part 2 of 5", DM Review February 2001.

Von Halle,B., (2001c), "Building a Business Rules System, Part 4", DM Review April 2001.

Von Halle,B., (2001d), "Building a Business Rules System, Part 5", DM Review May 2001.

Wu, J., (2000a), "Business Intelligence : What is Data Mining?", DM Review, August 2000.

Wu, J., (2000b), "Business Intelligence: Calculationg ROI for Business Intelligence Projects", DM Review Online, July 2000.

www.acxicom.com.

www.datawarehousing.com/whatis.asp.

White, C., (2000), "First Analysis", Intelligent Enterprise, June 5, 2000, Volume 3-Number 9.

ÖZGEÇMİŞ

Doğum tarihi	13.11.1972	
Doğum yeri	Artvin	
Lise	1986-1989	Ankara Laborant Meslek Lisesi
Lisans	1989-1995	Yıldız Üniversitesi Elektrik-Elektronik Fak. Bilgisayar Bilimleri ve Mühendisliği Bölümü
Yüksek Lisans	1995-2001	Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Programı

Çalıştığı kurumlar

1992-1995	Khan Bilgisayar Ltd.
1995-1996	İSKİ Bilgi İşlem D.Bşk.
1996-1997	Belbim A.Ş.
1997-1999	Alfa Yazılım Danışmanlık Ltd.
1999-2000	Askerlik hizmeti
2000-	I-BimSA A.Ş.

YÜKSEK LİSANS
DOKÜMANİSYON MERKEZİ