

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
SAYISAL YÖNTEMLER BİLİM DALI

**BİLGİ YÖNETİMİ ÇÖZÜMLERİ VE İŞ ZEKASI PROJELERİNDE
VERİ KALİTESİ YÖNETİMİ UYGULAMALARI**

Yüksek Lisans Tezi

SENEM DÜLGE

İstanbul, 2009

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
SAYISAL YÖNTEMLER BİLİM DALI

**BİLGİ YÖNETİMİ ÇÖZÜMLERİ VE İŞ ZEKASI PROJELERİNDE
VERİ KALİTESİ YÖNETİMİ UYGULAMALARI**

Yüksek Lisans Tezi

SENEM DÜLGE

Danışman: Doç. Dr. HAKAN YILDIRIM

İstanbul, 2009

Marmara Üniversitesi
Sosyal Bilimler Enstitüsü Müdürlüğü

Tez Onay Belgesi

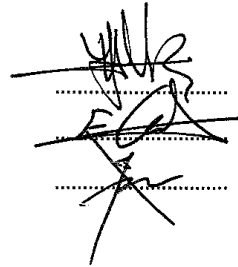
İŞLETME Anabilim Dalı SAYISAL YÖNTEMLER Bilim Dalı Yüksek Lisans
öğrencisi SENEM DÜLGE'nin BİLGİ YÖNETİMİ ÇÖZÜMLERİ VE İŞ ZEKASI
PROJELERİNDE VERİ KALİTESİ YÖNETİMİ UYGULAMALARI adlı tez çalışması
,Enstitümüz Yönetim Kurulunun 16.07.2009 tarih ve 2009-12/34 sayılı kararıyla ile
oluşturulan jüri tarafından oy birliği / oy çokluğu ile Yüksek Lisans Tezi olarak kabul
edilmiştir.

Öğretim Üyesi Adı Soyadı

İmzası

Tez Savunma Tarihi : 30.09.2009

- 1) Tez Danışmanı : DOÇ. DR. HAKAN YILDIRIM
2) Jüri Üyesi : YRD. DOÇ.DR. ELİF MAKBULE ÇEKİCİ
3) Jüri Üyesi : DOÇ. DR. TUNCAY CAN



ÖNSÖZ

Veri stratejik bir değerdir ve bugünün kurumlarında veri kalitesinin temini bir iş zorunluluğu haline gelmekte, kurumlar giderek artan bir şekilde veri kalitelerini arttırıcı çözümler bulma yoluna gitmektedirler. Veri kalitesi, güncel bir konu olmaktan öte organizasyonlarda devam eden iş süreçlerini hiç kuşkusuz etkileyecek bir yönetim felsefesidir. Bu gerçeğin farkındalığına varan kurumlar rekabette büyük ve vazgeçilmez avantajlar sağlayacaklardır.

Bu çalışma ile bilgi yönetimi çözümleri ve iş zekası projelerinde veri kalitesi uygulamaları anlatılmaya çalışılmıştır. Bu çalışmayı sonuçlandırmamda eşsiz görüşleri ile beni destekleyen, yönlendiren, bilgisini bana aktaran değerli hocam Doç. Dr. Hakan Yıldırım'a, hem benim için bir örnek olmaları hem de tavsiyeleri ile bana katkıda bulunan Marmara Üniversitesi Sayısal Yöntemler Bilim Dalı'nda görev yapan sevgili hocalarıma, iş zekası ve veri kalitesi konularında benimle deneyimlerini paylaşan SAS Türkiye Kıdemli Danışmanı ve Proje Yöneticisi Gönül Değirmendereli'ye, yardımlarından dolayı değerli mesai arkadaşım Görkem Şevik'e ve hiçbir zaman desteğini esirgemeyen arkadaşlarım Sultan Kılıç ve Emre Örs'e çok teşekkür eder, çalışmanın tüm ilgililere yararlı olmasını dilerim.

Hayatım boyunca desteklerini her zaman hissettiğim ve hissedeceğim sevgili aileme teşekkürlerin en büyüğünü bir borç bilirim.

İstanbul, 2009

Senem DÜLGE

GENEL BİLGİLER

İsim ve Soyadı	: Senem Dülge
Anabilim Dalı	: İşletme
Programı	: Sayısal Yöntemler
Tez Danışmanı	: Doç. Dr. Hakan Yıldırım
Tez Türü ve Tarihi	: Yüksek Lisans – Eylül, 2009
Anahtar Kelimeler	: Bilgi Yönetimi Çözümleri, İş Zekası, Veri Kalitesi, Veri Ambarı, Adres Verisi Standartlaştırma, Ayrıştırma, Tekilleştirme ve Zenginleştirme

ÖZET

Değerli bilgi; sahip olunan ya da olunması gereken, firmanın kar ve değer elde etmesine yarayan sektör, ürün, teknoloji ve organizasyon bilgileridir. Bilgi yönetimi, bu bilgileri yönetmenin ötesinde, sürecin doğru çözümlerle yönetimi ile ilgilidir. Bu süreç; doğru ve kaliteli veriyi elde etme, bilgi haline dönüştürme, bilgiyi geliştirme, koruma, kullanma, paylaşma ve bir iş zekasına dönüştürmeden oluşur. Yani bilgi yönetimi, ulaşılabilir kaliteli ve gerekli bilgileri; bu bilgilerin kullanımı için gerekli olan işlemlerin tanımlanmasını ve analizini kapsar. Yapılan herhangi bir analizin kalitesi ancak onu besleyen verilerin güvenilirliği ve kalitesi kadardır.

Bu çalışma ile ikinci bölümde bilişim kavramları özetlenmiş, üçüncü bölümde bilgi yönetimi çözümleri ve iş zekası anlatılmış, dördüncü bölümde veri kalitesi yönetimi ve uygulama prosesleri detaylı olarak incelenmiş, beşinci bölümde ise uygulama çalışması olarak örnek bir veritabanından alınan müşteri adres verilerine DataFlux ürünü ile veri kalitesi uygulama prosesleri gerçekleştirilmiş ve sonuçlar ortaya konulmuştur.

ABSTRACT

Data is a key strategic asset, and ensuring its quality has become a business imperative in today's enterprises and they are increasingly seeking solutions to improve their data quality. Spending the money, time and resources to collect massive volumes of data without ensuring its proper management is ineffective and often leads to failed business initiatives. The quality of any analysis is only as good as the data upon which it is based. All too often, companies are finding that a variety of analyses still yield faulty results, because the data is not of high quality. Data warehousing, data mining, marketing automation, and other data-driven solutions cannot result in success and deliver attractive "Returns on Investment" unless data quality issues have been resolved in a sufficient manner.

With this study, in the second part information concepts are summarized, in the third part information management solutions and business intelligence technologies are discussed, in the fourth part data quality management and application processes are analyzed and in the fifth part as an application, data quality application processes are applied to a sample of customer address data with DataFlux tool and the results are introduced and highlighted.

İÇİNDEKİLER

TABLO LİSTESİ	vi
ŞEKİL LİSTESİ	vii
KISALTMALAR VE SİMGELER.....	ix

GİRİŞ.....	1
------------	---

1. BİLİŞİM KAVRAMLARI	4
-----------------------------	---

1.1. Veri - Enformasyon – Bilgi.....	4
1.2. Veri Yapısı	5
1.3. Veri Tabanı	5
1.4. Veri Sözlüğü	5
1.5. Veri Modelleme	6
1.6. Data Mart’lar.....	7
1.7. Veri Ambarları	7

2. BİLGİ YÖNETİMİ ÇÖZÜMLERİ.....	10
----------------------------------	----

2.1. Bilgi Yönetiminin Tanımı.....	10
2.2. Bilgi Yönetim Süreçleri	11
2.3. İş Zekası	12
2.4. Veri Ambarı ve Karar Destek Sistemleri	15
2.5. Veri Ambarı ve Data Mart’lar.....	19
2.6. Çok Boyutlu Veri Görünümü - OLAP	20
2.7. Veri Entegrasyonu ve ETL (Çekme Dönüştürme Yükleme)	22
2.8. Veri Madenciliği	23

3. VERİ KALİTESİ YÖNETİMİ.....	27
--------------------------------	----

3.1. Veri Tekrarı, Veri Tutarsızlığı ve Veri Bütünlüğü	30
3.2. Veri Kalitesizliğinin Sebepleri.....	31
3.3. Veri Kalitesi Yönetiminde Karşılaşılabilecek Olası Veri Problemleri	33
3.4. Veri Kalitesinin Ölçülmesi	36
3.5. Veri Kalitesi Araçlarının Seçiminde Kritik Unsurlar	37

3.6. Veri Kalitesi Uygulama Süreci	40
3.6.1. Veri Profillemesi	41
3.6.2. Veri Standartlaştırma ve Düzeltme.....	45
3.6.3. Verinin Dönüştürülmesi ve Ayrıştırma (Parsing).....	48
3.6.4. Veri Zenginleştirme ve Veri Doğrulama	53
3.6.5. Veri Eşleştirme ve Tekilleştirme	54
3.6.6. Veri Kalitesinin Raporlanması ve İzlenmesi	55
4. UYGULAMA	56
4.1. Adres Verilerinin Profillenmesi.....	57
4.2. Adres Verilerinin Standartlaştırılması	64
4.2.1. “SEHIR” Alanı için Standartlaştırma Şeması ve Standartlaştırma Tanımı Yaratılması	65
4.2.2. “ADRES” Alanı için Standartlaştırma Şeması ve Standartlaştırma Tanımı Yaratılması	68
4.3. Adres Verilerinin Bileşenlerine Ayrıştırılması	73
4.3.1. “SEHIR” Alanının “Semt, İlçe, İl” Bileşenlerine Ayrıştırılması.....	73
4.3.2. Şehir Bilgisi Olmayan Fakat İlçe Bilgisi İçeren Kayıtlar İçin Şehir Bilgisi Bulunarak Adresin Zenginleştirilmesi	79
4.3.3. Arama Tablosu Aracılığıyla Ülke Alanı ‘Türkiye’ Girilmiş Olmasına Rağmen Yurtdışı Olan Adreslerin Tespit Edilmesi	81
4.3.4. “ADRES” Alanının Bileşenlerine Ayrıştırılması	82
SONUÇ	101
KAYNAKÇA	103

TABLO LİSTESİ

Tablo 3.1 : Veri Kalitesizliğinin Sebepleri	32
Tablo 3.2 : Veri Profilleme, İş Kuralları, Veri Standartlaştırma Kuralları Örneği	47
Tablo 3.3 : Adres Verisi Ayırıştırma Örneği I	52
Tablo 3.4 : Adres Verisi Ayırıştırma Örneği II	52
Tablo 3.5 : Adres Verisi Zenginleştirme Örneği	54
Tablo 4.1 : Adres Alanı Ayırıştırma İçin Türetilen İkinci Seviye Gramer Kuralları I	85
Tablo 4.2 : Adres Alanı Ayırıştırma İçin Türetilen İkinci Seviye Gramer Kuralları II	86
Tablo 4.3 : Adres Alanı Ayırıştırma İçin Türetilen İkinci Seviye Gramer Kuralları III	87
Tablo 4.4 : Adres Alanı Ayırıştırma İçin Türetilen İkinci Seviye Gramer Kuralları IV	88
Tablo 4.5 : Adres Alanı Ayırıştırma İçin Türetilen Ana Kategori Gramer Kuralları I	89
Tablo 4.6 : Adres Alanı Ayırıştırma İçin Türetilen Ana Kategori Gramer Kuralları II	90
Tablo 4.7 : Bileşenlerine Ayrılmış Örnek 20 Adet Adres Verisi I	96
Tablo 4.8 : Bileşenlerine Ayrılmış Örnek 20 Adet Adres Verisi II	97
Tablo 4.9 : Bileşenlerine Ayrılmış Örnek 20 Adet Adres Verisi III	98
Tablo 4.10 : Bileşenlerine Ayrılmış Örnek 20 Adet Adres Verisi IV	99

ŞEKİL LİSTESİ

Şekil 1.1 : Veri Ambarı Bileşenleri	9
Şekil 2.1 : İş Zekası Mimarisi	13
Şekil 2.2 : Veri Ambarı Yaratma Süreci	18
Şekil 2.3 : Veri Ambarı ve Data Mart	20
Şekil 2.4 : Çekme Dönüştürme Yükleme (ETL) Süreç Tasarımı.....	23
Şekil 2.5 : Veri Madenciliği Süreci	25
Şekil 3.1 : Veri Kalitesi Nedenleri Dağılımı	35
Şekil 3.2 : Veri Kalitesi Süreci.....	41
Şekil 3.3 : Veri Profillemeye Grafiksel Gösterim I.....	43
Şekil 3.4 : Veri Profillemeye Grafiksel Gösterim II	44
Şekil 3.5 : Veri Profillemeye Grafiksel Gösterim III.....	44
Şekil 3.6 : Veri Standartlaştırma Şeması Örneği.....	46
Şekil 3.7 : Veri Ayırıştırması Çalışmalarında Kullanılabilecek Veri Sözlüğü Örneği.....	49
Şekil 3.8 : Fonetik Kuralları Kütüphanesi Örneği.....	50
Şekil 3.9 : Gramer Kuralları Kütüphanesi Örneği.....	51
Şekil 4.1 : Profillenecek Müşteri Adres Verisi.....	57
Şekil 4.2 : Profillemede Seçilen Değişkenler ve Metrikler	59
Şekil 4.3 : Profilleme Raporu	60
Şekil 4.4 : Ülke Alanı Frekans Dağılımı	61
Şekil 4.5 : Şehir Alanı Veri Deseni Frekans Dağılımı	62
Şekil 4.6 : Şehir Alanı Veri Deseninden Yola Çıkılarak Veri Analizi.....	63
Şekil 4.7 : Profilleme Çalışması Yapılan Değişkenlere ait Özet Profilleme Raporu	64
Şekil 4.8 : ‘ADRES_VERI_FILTERED’ Tablosunda Element Analizi.....	65
Şekil 4.9 : SEHIR Alanı için Standartlaştırma Şeması Hazırlanması	66
Şekil 4.10 : SEHIR Alanı için Standartlaştırma Şeması	66

Şekil 4.11 : SEHIR Alanı için Standartlaştırma Tanımı	67
Şekil 4.12 : SEHIR Alanı Standartlaştırma	68
Şekil 4.13 : ADRES Alanı için Standartlaştırma Şeması Hazırlanması.....	69
Şekil 4.14 : ADRES Alanı için Standartlaştırma Şeması	70
Şekil 4.15 : ADRES Alanı için Standartlaştırma Tanımı	71
Şekil 4.16 : ADRES Alanı için Standartlaştırma İçin Kullanılan Regex	72
Şekil 4.17 : ADRES Alanı Standartlaştırma	72
Şekil 4.18 : ‘TRTUR Subdistrict District City’ Gramer Kuralları Kütüphanesi.....	75
Şekil 4.19 : PTT’den Alınan ‘Türkiye Adres ve Posta Kodları Özet’ Tablosu.....	76
Şekil 4.20 : ‘TRTUR Subdistrict District City’ Veri Sözlüğü	77
Şekil 4.21 : ‘Subdistrict District City’ Ayırıştırma Tanımı.....	78
Şekil 4.22 : Standartlaştırılmış SEHIR Alanının Bileşenlerine Ayırıştırılması	79
Şekil 4.23 : IL_ILCE_KODLARI Arama Tablosu	80
Şekil 4.24 : Arama Tablosu Aracılığıyla Şehir Bilgisinin Zenginleştirilmesi	80
Şekil 4.25 : Arama Tablosu Aracılığıyla Yurtdışı Adreslerinin Tespit Edilmesi	81
Şekil 4.26 : ADRES ve SEHIR Alanlarının Birleştirilmesi	82
Şekil 4.27 : ‘TRTUR OrgAddress’ Gramer Kuralları Kütüphanesi.....	91
Şekil 4.28 : ‘TRTUR Orgaddress’ Veri Sözlüğü	92
Şekil 4.29 : ‘Address Tez’ Ayırıştırma Tanımı	94
Şekil 4.30 : ADRES Alanının Bileşenlerine Ayırıştırılması	95

KISALTMALAR VE SİMGELER

BT : Bilgi Teknolojileri

TDWI : Veri Ambarı Enstitüsü (The Data Warehousing Institute)

BY : Bilgi Yönetimi

CRM : Müşteri İlişkileri Yönetimi (Customer Relationship Management)

ERP : Kurumsal Kaynak Planlama (Enterprise Resource Planning)

OLAP : Çok Boyutlu Veri Görselleme (On-line Analytical Processing)

ETL : Çekme-Dönüştürme-Yükleme (Extract-Transform-Load)

VKY : Veri Kalitesi Yönetimi

TKY : Toplam Kalite Yönetimi

Regex : Düzenli İfadeler (Regular Expression)

GİRİŞ

Günümüzde bilginin önemi giderek artmaktadır. Buna bağlı olarak yeni bir sektör (bilgi sektörü) ortaya çıkmıştır. Gelişmiş ekonomilerde bilgi sektörü en büyük sektör haline gelmiştir. Bu değişimi vurgulamak için "bilgiye dayalı ekonomi" terimi kullanılmaktadır. Bilgiyi toplamak, işlemek, düzenlemek, depolamak, bir yerden bir yere aktarmak ve bu bilgiye erişmek için kullanılan bilgi teknolojisi ekonomik yapıda bir dönüşüm yaratmıştır. Bilgi teknolojisi alanında yaşanan hızlı gelişmeler, günümüz işletmelerini rekabetin her an kıyasıya yaşandığı bir pazar yapısına götürmektedir. Bu yapı içinde "bilgi" karşımıza yükselen bir değer olarak çıkmakta ve işletmelerin rekabet gücünü doğrudan etkilemektedir.

İnternet teknolojileri ile beraber hızla değişen iş dünyası ve ekonomi bir çok farklı terim ve uygulamaları da beraberinde getirmiştir. Girişimciler hem bilgilerinin değerini, hem de bu bilgilerden en yüksek geri dönüşü alabilmek için bilgiyi nasıl yönetmeleri gerektiğini bilmenin artık bir zorunluluk olduğunun farkına varmışlardır. Bilginin kaybolmaması, boşa harcanmaması, doğru yönlendirilmesi ve üretken olabilmesi için "bilgi yönetimi" kavramı ortaya çıkmıştır. Bilgi yönetimi, organizasyonları bilgiye odaklanma konusunda yönlendirirken diğer yandan da beraberinde getirdiği yöntem ve tekniklerle, bilginin doğru yönlendirilmesi ve üretken olabilmesini sağlamalıdır.

Bilgi Yönetimin beraberinde getirmiş olduğu bu yeni kavramlardan birisi de "İş Zekası"dır. Verinin bilgiye, bilginin iş kararlarına ve bu mekanizmanın da tüm iş ortamına yayılarak bir iş zekasına dönüşmesini sağlayan çözümler, şirketlerin diğer bilgi sistemleri ile entegre çalışıp şirketlerin stratejilerini belirlerken ihtiyaç duydukları bilgilere tek bir kaynaktan ulaşabilmelerini sağlarlar. İş Zekası çözümleri günümüzde geniş bir yelpazeyi kapsayarak sorgulama, raporlama, istatistiksel analiz, tahmin yürütme ve veri analizi, veri madenciliği, veri kalitesi gibi alanlarda kullanılmaktadır.¹

¹ İş Zekası, 2004, <http://www.nexum.com.tr/tr/hizmetler/iszekasi.asp> (17 Temmuz 2007)

Bir kurumun öğrenme yetkinliği ve öğrendiklerini hayata geçirme yetkinliği gibi rekabet avantajı yakalamanın en önemli koşullarından biri, öğrenme yetkinliğine katkı yaratacak bilginin depolanması ve analiz edilebilir hale getirilmesidir ve bunun için temel şart kullanılacak verilerin kalitesi ve güvenilirliğidir.²

Yanlış ya da birbiri ile tutarsız veriler, kurumların bugünkü ve gelecekteki iş problemlerini anlamasına engel olmaktadır. Sağlıksız verinin farkına varılıp bir an önce önlem alınıp düzeltilmez ise; kurumlarda kazanç kaybı, operasyonel gecikmeler ve müşteri memnuniyetsizliği gibi birçok konuda olumsuz sonuçlar doğuracak sağlıksız kararların alınmasına sebep olacaktır.

PricewaterhouseCoopers tarafından “Global Data Management Survey”de kötü veri kalitesi yüzünden şirketlerin %75’inin net kar/zararında ciddi sarsıntı yaşadığı belirtilmiştir. The Data Warehousing Institute (TDWI) tarafından yapılan bir araştırmada, müşteri verisindeki kalite problemlerinin Amerika’da kurumlara yılda 600 Milyar doların üzerinde bir maliyet getirdiği saptanmıştır. Gartner, Inc., veri ambarı projelerinin %50’sinden fazlasının veri kalitesindeki problemleri giderememe nedeniyle başarısızlığa uğradığını belirtmiştir. Gartner, ayrıca kurumların işletme gelirlerinin %10 - %20’sinin veri kalitesi ile ilişkili problemlerin giderilmesi için harcandığı üzerinde durmaktadır.³

Bu çalışma ile birinci bölümde bilişim kavramları özetlenmiş, ikinci bölümde bilgi yönetimi çözümleri ve iş zekası anlatılmış, üçüncü bölümde veri kalitesi yönetimi ve uygulama süreçleri detaylı olarak incelenmiş, dördüncü bölümde ise uygulama çalışması olarak örnek bir veritabanından alınan müşteri adres verilerine DataFlux ürünü ile veri kalitesi uygulama süreçleri gerçekleştirilmiş ve sonuçlar ortaya konulmuştur.

² İş Zekası Çözümleri, (t.y.) <http://www.ironmanconsulting.com/v2/tr/3-4-is-zekasi-cozumleri.asp> (28 Temmuz 2007)

³ Total Data Quality Management, 2004, <http://mitiq.mit.edu/MITIQ/TDQM%20Workshop/Newzealandtr46.pdf> (2 Eylül 2007)

Doğru kararlara ve doğru analiz sonuçlarına ancak bu analizlere konu olan verinin anlaşılabilir, kendi içinde anlamlı ve kaliteli olması şartı ile ulaşılabilir. Verilerin kalitesinin artırılması ve varolan verilerden analiz yapılmasının olanaksız olduğu ortamlarda verilerin anlamlı bileşenlerine ayrıştırılması gerekmektedir. Adres verileri gibi birçok organizasyonda bir bütün halinde saklanan bilgiler bileşenlerine ayrıştırılmadan anlamlı sonuçlar elde edilemez ki günümüzde Türkiye İstatistik Kurumu tarafından yürütülen ve uzun süredir devam eden Adres Kayıt Sistemi projesinde dahi en önemli adım varolan adreslerin bileşenlerinin tespit edilmesi ve adreslerin doğruluğunun tespit edilebilmesi olmuştur. Bu tez çalışmasının temel amacı, hem organizasyonlara örnek teşkil etmesi hem de akademik olarak pratiksel bir ön çalışma olması nedeniyle örnek bir veritabanındaki adreslere Dataflux yazılımı aracılığıyla bileşenlerine ayrıştırma çalışması uygulanması ve bileşenlerine ayrıştırılan adresler üzerinden yapılacak her türlü analize ön hazırlık niteliği teşkil etmesidir.

BİLİŞİM KAVRAMLARI

1.1. Veri - Enformasyon – Bilgi

Veri, enformasyon ve bilgi sürekli karıştırılan ve birbirleri yerine kullanılan kavramlardır. Fakat dikkatli irdelendiklerinde, aralarındaki ayrım çizgisinin çok belirgin olduğu ortaya çıkmaktadır. Organizasyonel başarı ve başarısızlık; hangilerine ihtiyaç duyduğumuzu, hangilerine sahip olduğumuzu ve elimizdekilerle ne yapabileceğimizi veya yapamayacağımızı bilmemize bağlıdır. ⁴

Veri, enformasyonun hammaddesidir. Veri şekillendirilerek (sınıflandırma, özetleme, analiz etme ve yorumlama) enformasyon haline getirilir. Enformasyonu değerli hale getiren onun doğruluğu, ulaşılabilirliği, zamansızlığı, karar vermede etkili olması, tanımlanmış ortamda uygun ve faydalı olmasıdır. Enformasyon, verinin tersine yorum içerir ve bu yorum kişinin görüş ve deneyimleri ile birleşerek faaliyete imkan tanıyan bilgi haline gelir. Tüm bu açıklamaları basit bir örnekle tanımlayacak olursak;

- Veri: 25
- Verilerin içine konabileceği bir bağlam olan enformasyon: Hava sıcaklığı 25 derecedir.
- Verilerden ve enformasyondan çıkarılmış bir sonuç olan bilgi: Kayak gezisini ertelememiz gerekir. ⁵

⁴ Thomas H. Davenport ve Laurence Prusak, **Working Knowledge: How Organizations Manage What They Know**, 2. Baskı, Boston: Harvard Business School Press, 2000, s. 25-31, 112-117

⁵ R. E. Bohn, **Measuring and Managing Technical Knowledge**, 1. Baskı, Boston: Sloan Management Review, 1994 s. 48-52

1.2. Veri Yapısı

Veri yapısı, birbiri ile ilişkisi olan elemanların oluşturduğu veri elemanları kümesidir ve sistemi teşkil eden bileşenlerden her birini tanımlamak üzere kullanılan bir yapıdır. Veri akışı, veri depolaması buna bir örnektir. ⁶

1.3. Veri Tabanı

Veri tabanı, en geniş anlamıyla; birbiriyle ilişkili verilerin tekrara yer vermeden, çok amaçlı kullanımına olanak sağlayacak şekilde depolanması olarak tanımlanabilir. Enformasyon alanında en çok bibliyografik veri tabanları kullanılmaktadır. Bunun dışında sayısal, tam metin, adres, sözlük, kaynak veri tabanları da bulunmaktadır. ⁷

Bilgisayar terminolojisinde, sistematik erişim imkanı olan, yönetilebilir, güncellenebilir, taşınabilir, birbirleri ile arasında tanımlı ilişkiler bulunabilen bilgiler kümesidir.

1.4. Veri Sözlüğü

Veri sözlüğü, veritabanı ile ilgili bilgilerin yine veritabanında tablolar ve görüntüler halinde saklanmasıyla oluşur. Veritabanı ile ilgili bu bilgiler kullanıcılar, haklar, veritabanı nesneleri, tablo kısıtlamaları vb. bilgilerinden oluşur. Veri sözlüğü veritabanının en önemli bölümüdür. Veritabanı hakkındaki tüm bilgiler veri sözlüğünden komutlar yazarak öğrenilebilir.

⁶ James A. Senn, **Analysis and Design of Information Systems**, 1. Baskı, New York: McGraw Hill, 1984 s. 54-60, 175-178

⁷ Uzm.Vet.Het. Ebru Soyuyüce ve diğerleri, “Bilgisayara Dayalı İndeksleme” **Kütüphane-Enformasyon Arşiv Alanında Yeni Teknolojiler Ve Türkmarc Sempozyumu**, İstanbul: Türk Kütüphaneciler Derneği İstanbul Şubesi, 1991, s.194-199

Veri sözlüğü, veritabanı kurulurken standart olarak oluşturulur ve yapılan değişiklikler otomatik olarak veri sözlüğüne yansıtılır. Örneğin bir kullanıcı bir tabloya bir alan eklerse, veri sözlüğünde o tabloya ilişkin yapı kullanıcı fark etmeden değiştirilir böylece veri sözlüğünden veritabanı hakkında sürekli olarak güncel ve sağlıklı bilgi almak mümkün olur. Veri sözlüğü içerisindeki bilgiler komutlarla sorgulanıp görülebilir. Fakat hiçbir kullanıcı veri sözlüğünü değiştiremez, ekleme yapamaz ve kayıt silemez. Veri sözlüğü tabloları için görüntüler oluşturulmuştur. Tüm kullanıcılar hakları olduğu müddetçe bu görüntülerden bir ya da birkaçını sorgulayabilirler. ⁸

1.5. Veri Modelleme

Veri Modelleme, bir işletmenin, kurumun hatırlamaya değer bulduğu verilerin şekil ve metin olarak ifade edilmesidir. Diğer bir deyişle bir işletmede teknik ve teknik olmayan herkesin bilişim ihtiyaçlarını ifade etmeye çalışırken birbirini anlamada kullanabileceği görsel bir iletişim dilidir. Yazılım geliştirmenin en önemli süreçlerinden biri olan veri modelleme bilişim ihtiyaçlarının keşfedilmesi ve herkesin anlayabileceği bir şekilde belgelenmesi işlemidir. Bilişim ihtiyaçları, veriler ve işletme ihtiyaçlarını destekleyen işletme kurallarıdır. Bir veri modeli herhangi bir işletmenin veya bir yazılımın karmaşık bilişim ihtiyaçlarının tümünü yeterince ifade edebilmek için kullanılabilecek bir araçtır. Bir bilişim sistemi başlıca 3 ihtiyacı karşılar; çeşitli verilerin saklanması, işlenmesi ve görüntülenmesi. Bilişim sistemlerinin temelinde veri yer almaktadır. Veri Tabanı sistemleri ise en basit ifadeyle; verinin saklanması ve işlenmesi ile ilgili olarak geliştirilen genel amaçlı çeşitli yazılımlardır. Kısaca bilginin işlenmemiş hali olarak tanımlanan verinin modellenmesi herhangi bir bilişim sistemi geliştirmede neredeyse işin yarısını oluşturur. ⁹

⁸ Bilgisayar-Yazılım-ORACLE, (t.y.) <http://www.textara.com/oracle-data-base-veri-tabani-yolu-nedir-komutlar-ornek?page=0%2C8> (4 Eylül 2007)

⁹ Veri Modelleme, (t.y.) http://tr.wikipedia.org/wiki/Veri_modelleme (4 Eylül 2007)

1.6. Data Mart'lar

İçerdiği veri açısından veri ambarlarının bir alt kümesi olan, bir ve birden fazla olabilecek mantıksal yapılara data mart adı verilir. Data mart'ın amacı sadece belirli bir alan hakkında bilgi içermektir. Böylece herhangi bir raporlama programı data mart'ı sorgulayıp veriyi hızlı bir şekilde alabilir, gerekli işlemlerden geçirildikten sonra kullanıcılar raporlarını etkin bir şekilde oluşturabilir. Data mart departmansal bazda bir veri depolayıcısı konumundayken veri ambarları bütün şirketin verilerinden sorumludur. Data mart belirli bir konuya odaklı ya da departman düzeyindeki bir veri deposudur. Sözelimi pazarlama ya da muhasebe gibi. Bazı şirketler genel ve bütünsel bir veri ambarı kullanmak yerine departman düzeyindeki veriyi kullanmak isterler. Bazı durumlarda data mart'lar veri ambarından analiz yolu ile elde edilmiş parçalardan ya da bu tür farklı parçaların birleşiminden oluşturulmaktadır. Çok basit anlamda orta büyüklükteki bir organizasyonda tek bir veri ambarı ve aynı zamanda birden fazla data mart bulunur.

1.7. Veri Ambarları

Veri ambarı, bir işletmenin ya da kuruluşun değişik birimleri tarafından toplanan bilgilerin, gelecekte kullanılabilecek ya da değerlendirilebilecek olanlarının arka planda üst üste yığılarak birleştirilmesinden oluşan büyük çaplı bir veri deposudur.

Karar destek sistemlerinde yer alan veriler çeşitli inceleme ve araştırmalardan geçirilerek, idarecilerin gelecekte kurumun verimliliğini artırmasını, geleceğe yönelik kurum politikalarının belirlenmesini ve buna benzer yönetsel kararların alınmasını kolaylaştırır.¹⁰

Veri ambarları, ilişkili verilerin sorgulanabildiği ve analizlerin yapılabildiği bir depodur. Bir veri ambarı, analizler ve sorgular için kullanılabilir bütünleşmiş bilgi

¹⁰ Veri Ambarları, (t.y.) <http://www.ikademi.com/insan-kaynaklari-bilgi-sistemleri/1237-veri-ambarlari.html>
(17 Kasım 2007)

deposudur. Veri ve bilgiler, üretildiklerinde heterojen kaynaklardan elde edilirler. Veri ambarı, başlangıçta farklı kaynaklardan gelen verinin üzerinde daha etkili ve daha kolay sorguların yapılmasını sağlamaktadır.¹¹

Veri ambarları; konu tabanlı, bir bütün halinde entegre edilmiş, zaman değişkeni içeren, sürekli ve tutarlı verilerin toplanması sonucu organizasyonel karar destek sistemleridir.¹²

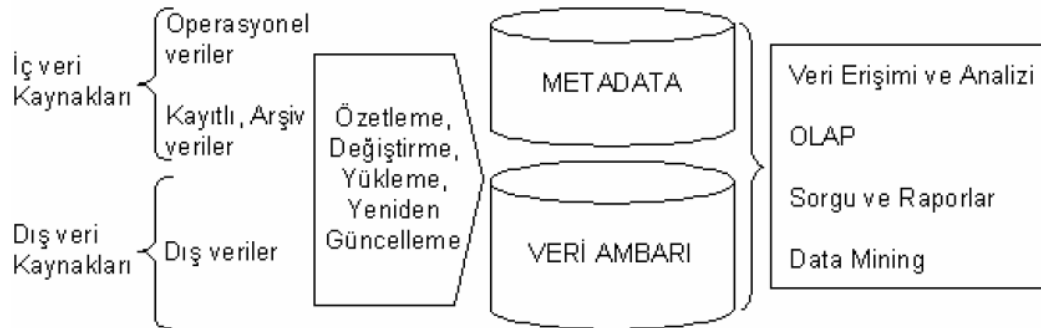
Veri ambarı bileşenleri ve fonksiyonları aşağıdaki gibi özetlenebilir:

- Değişik platformlar üzerindeki işletimsel uygulamalara ait verilere erişim ve gerekli verilerin bu platformlardan alınması,
- Alınan verilerin temizlenmesi, tutarlı duruma getirilmesi, özetlenmesi, birleştirilmesi ve birbirleriyle entegrasyonunun sağlanması,
- Dönüştürülen verilerin veri ambarı veya data mart ortamına dağıtımı,
- Gönderilen verilerin bir veri tabanında toplanması,
- Depolanan bilgi ile metadatada bulunan ilgili bilgilerin veri kataloğunda saklanması ve son kullanıcılara sunulması,
- Veri ambarı veya data mart'ta bulunan bilgileri uç kullanıcıların karar destek amaçlı kullanımının sağlanması.¹¹

Şekil 1.1 basit anlamda veri ambarı bileşenlerini göstermektedir.

¹¹ Ümit Burak Usgurlu, “Veritabanı, Veri Madenciliği, Veri Ambarı, Veri Pazarı”, (t.y) Başkent Üniversitesi, Bilgisayar Mühendisliği Yönetim Bilişim Sistemleri, <http://mail.baskent.edu.tr/~20394676/0302/bil483/HW2.pdf> (17 Kasım 2007)

¹² W.H. Inmon, **Building the Data Warehouse**, 4. Baskı, New York: Wiley, 2005, s. 41-49, 276-278, 421-429



Şekil 1.1 : Veri Ambarı Bileşenleri ¹¹

Veri ambarı oluşturmanın organizasyona sağlayacağı faydaları aşağıdaki şekilde sıralamak mümkündür :

- Veri ambarı veriye ulaşımı kolaylaştırır.
- Farklı uygulamaların verilerini birleştirir.
- Kaynak tek olduğu için farklı kullanıcıların aldığı raporlar aynı parametreler ile farklı sonuçlar üretmez.
- Veri ambarı yeni işler için kullanıcılara yol gösterir. ¹³

¹³ Veri Ambarı Nedir? , 2007, <http://veriambarinedir.blogspot.com/> (17 Kasım 2007)

2. BİLGİ YÖNETİMİ ÇÖZÜMLERİ

2.1. Bilgi Yönetiminin Tanımı

Günümüzde bilginin kolaylıkla ulaşılabilen bir değer haline gelmesi ve rekabet şartlarının ağırlaşması, işletmelerin bilgi teknolojilerini yoğun olarak kullanmasını gerektirmektedir. Bilgi yönetimi; her türlü bilginin üretilmesi, farklı bilgi ağları arasında bilginin transfer edilmesi, hazır bilgiye erişim, erişilmiş bilgilerin kolaylıkla yayılması ve bilginin sektörün ve şirketin ihtiyaçları doğrultusunda kullanılmasını sağlayan bir iş kolu olarak tanımlanabilir.¹⁴

Entelektüel sermaye, bilgi birikimi yönetimi gibi kavramlar giderek yaygınlaşırken, sürekli değişim içinde olan dijital çağa ve iş hayatındaki rekabet ortamına ayak uydurmak için artık bilgiyi en pragmatik yollarla, en verimli şekilde kullanmak bir zorunluluk haline gelmiştir.

İnternet teknolojileri ile beraber hızla değişen iş dünyası ve ekonomi bir çok farklı terim ve uygulamaları da beraberinde getirmiştir. Bilginin kaybolmaması, boşa harcanmaması, doğru yönlendirilmesi ve üretken olabilmesi için 'bilgi yönetimi' kavramı ortaya çıkmıştır. Geleneksel firmaların bir çoğu hala sahip oldukları bilgilerin değerinin ve kullanım alanlarının farkında değildirler. Oysa başarı, artık bugün bilginin kalitesi ile doğru orantılıdır.

Bir organizasyona yönelik Bilgi Yönetimi, organizasyonun kendisinin ve başkalarının deneyimlerinden bilgi elde etmesi ve bu bilginin organizasyonun misyonunu yerine getirmesi için kullanılması amacıyla gerçekleştirdiği faaliyetleri kapsar. Bu faaliyetler, teknoloji, organizasyonel yapılar ve varolan bilgi düzeyini yükseltme ya da yeni

¹⁴ http://www.meslekegitimleri.com/sau/01.hakkinda_bilgi.htm (22 Aralık 2007)

bilgiler yaratmaya yönelik stratejilerce yönlendirilir. Burada kritik olan nokta, ilgili sistemde yer alan unsurların (organizasyon, insan, bilgisayar ya da bütünleşik insan-bilgisayar sistemleri) öğrenme, problem çözme ve karar vermeye yönelik olarak doğru ve kaliteli bilgiye ulaşma, bilgiyi saklama ve bilgiden yararlanma becerilerinin geliştirilmesidir. Bilişim teknolojileri bilgi yönetimi için en elverişli araçları sağlıyor olsa da, unutulmaması gerekir ki; bilgi yönetimi salt teknoloji demek değildir, kurumların iş zekası yaratabilme becerileri bilgiden değer elde edilmesinde büyük önem taşımaktadır.¹⁵

2.2. Bilgi Yönetim Süreçleri

Bilgiyi yönetmenin nihai amacı örgütsel düzeyde performansı artırmaktır. Ayrıca, bilgi yönetiminin müşteri ihtiyaçlarının daha iyi anlaşılması, yenilik performansını geliştirilmesi, kaliteyi artırması ve maliyetleri düşürmesi gibi firmanın performansını artıracak farklı etkileri de vardır.¹⁶

Bir organizasyonun mutlaka gerçekleştirmesi gereken bilgi yönetim süreç adımları şunlardır;

1. Bilgi ihtiyacını belirleme
2. Kaliteli ve güvenilir bilgiyi elde etme ve yaratma
3. Bilgi stoğunu oluşturma
4. Bilgiyi organizasyona yayma

¹⁵ Todd R. Groff ve Thomas P. Jones, **Introduction to Knowledge Management: KM in Business**, 1. Baskı, New York: Butterworth-Heinemann, 2003, s. 52-57, 139

¹⁶ N. Zakaria, A. Amelinckx, ve D. Wilemon, (2004). "Working Together Apart? Building a Knowledge-Sharing Culture for Global Virtual Teams", *Creativity and Innovation Management*, Vol.13, No.1, s. 15-29.

Belirtilen bu süreçler uygulayıcı organizasyona bağlı olarak detaylandırılabilir, sayıları artırılabilir ya da bazı süreçler birleştirilerek sayıları azaltılabilir.^{17,18}

2.3. İş Zekası

Günümüzün başarılı şirketleri, kar marjlarının hızla eridiği zamanımızda, işini karlı olarak yapabilenlerdir. Globalizasyonun hızla arttığı bir iş dünyasında hem giderleri kontrol altında tutmak, hem de gelir fırsatlarını iyi değerlendirmek gerekmektedir. İş zekası, kısa zamana kadar sadece büyük firmaların ilgilendiği, biraz akademik, biraz geleceğe dair, pahalı bir alan olmasına rağmen günümüzde iş zekası çözümleri ve uygulamaları şirketler arasında oldukça yaygınlaşmıştır.

İş zekası çözümleri şirketlerin diğer bilgi sistemleri ile entegre çalışıp şirketlerin stratejilerini belirlerken ihtiyaç duydukları bilgilere tek bir kaynaktan ulaşabilmelerini sağlarlar. İş Zekası çözümleri günümüzde geniş bir yelpazeyi kapsayarak sorgulama, raporlama, istatistiksel analiz, tahmin yürütme ve veri analizi, veri madenciliği, veri kalitesi gibi alanlarda kullanılmaktadır.

İş zekası uygulamaları için öncelikli olarak bir kurumda, operasyonel veriye gereksinim vardır. Örneğin, her türlü müşteri bilgileri, her türlü üretim, dağıtım, muhasebe vb. verileri operasyonel veri olarak ele alınmaktadır. İş zekası, ancak operasyonel verinin üzerine kurgulanan bir çözümdür. Mevcut verilere bakıp bunları anlayarak geleceğe ilişkin belli mantıklı sonuçlar çıkarabilmek ve öngörüler, raporlamalar yapmak olanaklı olabilmektedir. Bir kurumun iş zekası çözümleri uygulamaya başlayabilmesi için yorumlanacak derecede veri elde etmesi gerekmektedir.

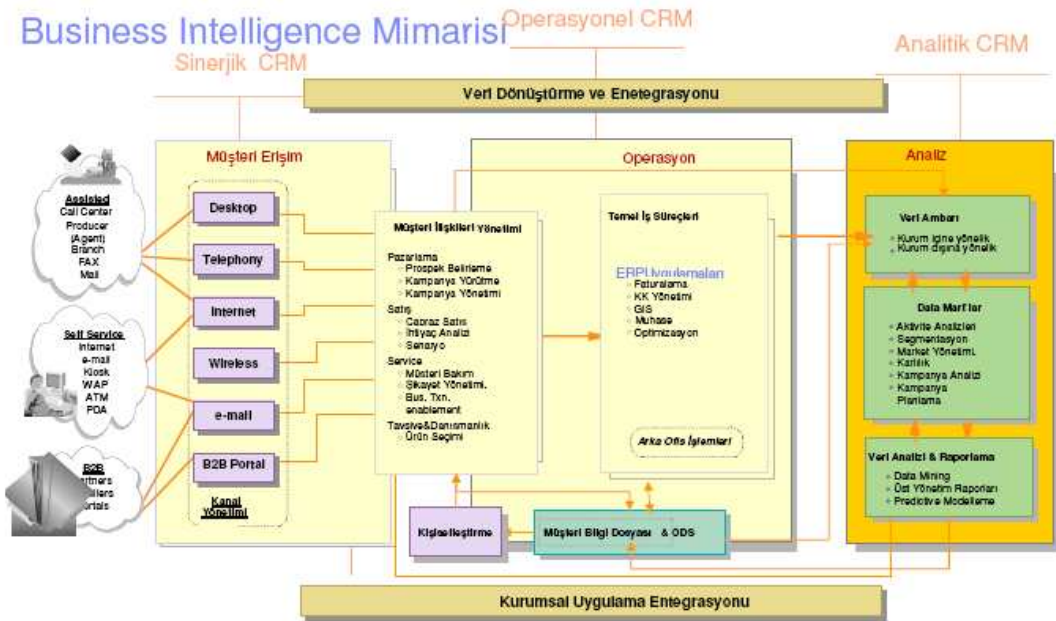
¹⁷ I. Nonaka ve H. Takeuchi, **The Knowledge Creating Company**, 1. Baskı, New York: Oxford University Press, 1995, s. 221-224, 292

¹⁸ Thomas A. Stewart, **Intellectual Capital: The New Wealth of Organizations**, 1. Baskı, New York: Doubleday Business, 1998, s. 255-261

CA- Computer Associates'in "Birikim Yönetimi Sistemini" kurma yolundaki olgunluk seviyesi ve dünyadaki örneklerde de görülen şu anda iş zekası uygulamalarının temel 4 adımı şu şekilde özetlenebilir:

- Veriye ulaşmak
- Verilerin temizlenmesi ve filtrelenmesi
- Filtrelenmiş bilginin kişiye özel yorumlanması
- Bilginin harekete geçirilmesi-hareket ve karar aşaması.

Şekil 2.1 İş Zekası Mimarisi resimlendirmektedir.



Şekil 2.1 : İş Zekası Mimarisi¹⁹

¹⁹ Kurumunuzu Akıllı Yazılımlar Yönetsin, 2002, http://www.btvizyon.com.tr/kulakcik.phtml?kulakcik_nox=53 (7 Ocak 2008)

İş zekası çözümleri her sektör için bir ihtiyaçtır. Sektör ne olursa olsun, büyük çaplı veri kümesinin içinden hedefe yönelik bilgiyi çekebilmek, raporlayabilmek, daha da ötesi geleceğe yönelik öngörüler yapabilmek asıl amaç olmalıdır. Bunları yaparken sadece kurumların koyduğu kurallar değil, görülemeyen ilişkileri de yakalayabilmek çok önemlidir. Gerçek iş zekası uygulamalarının içine yapay zeka tanımları da girmektedir. İş zekası günümüz şirketlerinde üst yönetimin en fazla dikkat ettiği katmanlardan biri olmuştur. Çok farklı sistemin entegre edilmesi, onlardan yararlanıp Bilgi Yönetim Sistemlerinin oluşturulması asıl gelinmesi gereken noktalardan biridir. Daha da fazlası, görülemeyen verinin içinden sistemin kendisinin belli bir akıl yürüterek sonuç üretmesi, proaktif olması kurumların iş zekası uygulamaları için hedef noktası olmalıdır. Farklı kaynaklardan gelen verilerin öncelikle bir merkezde toplanması gerekmektedir, bundan sonraki adım verileri yorumlayıp karar alma aşamasına gelme olarak açıklanabilir. Bu noktada da iki yaklaşım mevcuttur. Kuralları kullanıcının tanımladığı uzman sistemler ve kendisi kural yaratabilen yapay zeka sistemleri. Yapay zeka sistemleri için organizasyonun gelişmiş olması, doğru ve bol veri hacmine gerek vardır. Günümüzde henüz kurumlar, uzman sistemlerle çalışmayı tercih etmektedirler.

İşletmenin olgunluğuna göre iş zekası çözümleri uygulanabilmektedir. Henüz operasyonel sistemleri kurulma aşamasında olan bir şirkete iş zekası kurmak doğru olmayabilir. Bazı sektörlerde kurum hem üretici, hem perakendeci, hem bankacı olabilmekte ve kredi kartları sistemini, üretim sistemini ve perakende sistemini entegre çalıştırması gerekebilmektedir. İşte burada iş zekası çözümlerinden yararlanması gerekmektedir.²⁰

İş zekası araçları, bilgiyi verimli bir şekilde yöneterek firmalara rekabette proaktif davranabilme avantajı sunarlar. Kurumsal İş Zekası, özellikle teknik birikimi olmayan son kullanıcıları, günlük hayatta kullandıkları terimlerle bilgiye erişebilecekleri, eriştikleri bilgi

²⁰ Kurumunuzu Akıllı Yazılımlar Yönetsin, 2002, http://www.btvizyon.com.tr/kulakcik.phtml?kulakcik_nox=53
(7 Ocak 2008)

üzerinde analiz yapabilecekleri ve elde ettikleri sonuçları kurum içi ya da kurum dışındaki kullanıcılarla paylaşabilecekleri bir platform sunmaktadır.

Geçmişle karşılaştırıldığında bir kurumdaki bilgiye erişen iç ve dış kullanıcı sayılarının artması ve aynı zamanda anlık bilgiye erişim ihtiyacının ön plana çıkması, kurumsal iş zekası platformuna operasyonel iş zekası çözümlerinin de entegre olmasını gerektirmeye başlamıştır. Buna paralel olarak verinin bu kadar büyük kitleye ulaştırılması, daha düşük kullanıcı seviyelerine hitap edebilmek amacıyla verinin görselleştirilmesi gerekliliğini daha çok ortaya çıkarmıştır.

Bir kurumun iş zekasının ne olduğu ve ne kadarının kullandığı, ne kadar zekayı bilgisayar ortamında kurumun hizmetine sunduğu (yani ne kadar zekaya erişildiği) ölçülmelidir. Bu ölçüm olmadan kurumsal zekanın artırıldığını da söylemek inandırıcı olmaz. Bilgileri bir araya getirerek derli toplu halde kullanıma sunmak geçici bir zeka artışı gibi görülebilir. Bunun yanıltıcı olmaması gerekir. İş zekası yaratılması ve ölçülmesi sadece bilişim sistemlerini değil aynı zamanda sosyal bilimcileri de yakından ilgilendirmektedir. Bu konu günümüzde en önemli araştırma alanlarından birisi olarak karşımıza çıkmaktadır.²¹

2.4. Veri Ambarı ve Karar Destek Sistemleri

Günümüz şartlarında şirketlerin ve kurumların hayati işlevlerini yerine getirmede bilgi sistemleri vazgeçilmez düzeyde önemli hale gelmiştir. Kurumun yaşam sürecini devam ettirebilmesi için gerekli temel fonksiyonlarının otomatize edildiği uygulamalar, ve bu uygulamaların verilerinin tutulduğu veri tabanı yönetim sistemleri artık her kurumun gündelik yaşantısında vazgeçilmez parçalar haline gelmiştir. Bu gelişim kısa sürede

²¹ Ikujiro Nonaka, Georg von Krogh, ve Sven Voelpel, **Organizational Knowledge Creation Theory: Evolutionary Paths and Future Advances**, 1. Baskı, Boston: Organization Studies Press, 2006, s. 14-16

yapılamayacağı için, kurumun iş yapısı geliştikçe ve değıştikçe çeşitli uygulamalar bu yapıya eklenir. Öte yandan, bu uygulamalarda toplanan verilerin daha detaylı incelenmesi ve karar sürecine daha fazla katkı sağlaması için oluşturulan karar destek sistemlerine, Internet'in devreye girmesiyle kurumsal bilgi yapılarında oluşan "portal" ve benzeri "Internet bazlı" uygulamaları ve bu "Internet uygulamalarıyla" arka taraftaki ana uygulamaları birbiriyle bağlayan ara katman yazılımları da eklediğiniz zaman, oluşan ortam kolay kolay içinden çıkılamayacak bir karmaşı hale gelecektir.²²

Veri ambarının temel hedefi, organizasyon için "işlere yönelik gerçeklerin tek bir görüşünü" ortaya koyabilmektir. Bu amaca ulaşmak için çeşitli programlar aracılığıyla, operasyonel ortamdan, veri ambarı verisini barındıran veritabanından ve kullanıcılara veri sağlayan sistemlerden veriler çekip çıkarılır. Geleneksel operasyonel uygulamalar, karar desteğine yönelik bilgiler sağlamaktan uzaktır. Mevcut işlere yönelik bilişim sistemleri altyapısı ile, en iyi kararları verebilmek için gerekli olan bilgiler elde edilememektedir. Bu durumda, uzun vadeli düşünüp, resmin tamamını görmek isteyen yöneticiler "karar destek sistemleri"ne başvurmaktadır. Bu yöneticiler, örneğin belirli bir yerde bulunan bir depodan, geçen hafta A ürününden ne kadar çıktığını öğrenmek yerine, örneğin son 5 yılda belirli bir coğrafi bölgede yer alan depolardan çıkan A ürününün toplam sayısı ile ilgilenirler. Yöneticilerin ilgilendiği konuların başında, örneğin, "hangi ürünler zamanın %15'inden daha fazlasında stokta bulunamamıştır?", "hangi müşteriler, 2006 ve 2007 ile karşılaştırıldığında, 2008 yılında en fazla siparişi girmiştir?" türünden daha karmaşık, rutin olmayan sorular gelir. Bu tür istekler, bir arada bulunan, oldukça geniş tarihsel veriler içeren bir veritabanı ile sağlanabilir. Önemli olan, soruların yanıtlarının derhal alınmasından daha çok, gerçeği yansıtan, sağlıklı bilgilerin makul bir bekleme süresi sonunda elde edilebilmesidir.

Veri ambarını karakterize eden en temel özellik, operasyonel uygulamalarla karar destek sistemleri uygulamaları arasındaki farktır. Birbirinden farklı işlevler olan bu iki süreci ayırmakla, operasyonel ve karar destek sistemleri uygulamalarının kendilerinin en iyi

²² Veri Ambarı ve Karar Destek Sistemleri, (t.y.) <http://www.theturkler.com/makale/429/> (17 Kasım 2007)

halledebileceği konulara odaklaşması sağlanmış olur. Amaç, herkes için optimum olan bir veri ambarı oluşturabilmektir.²³

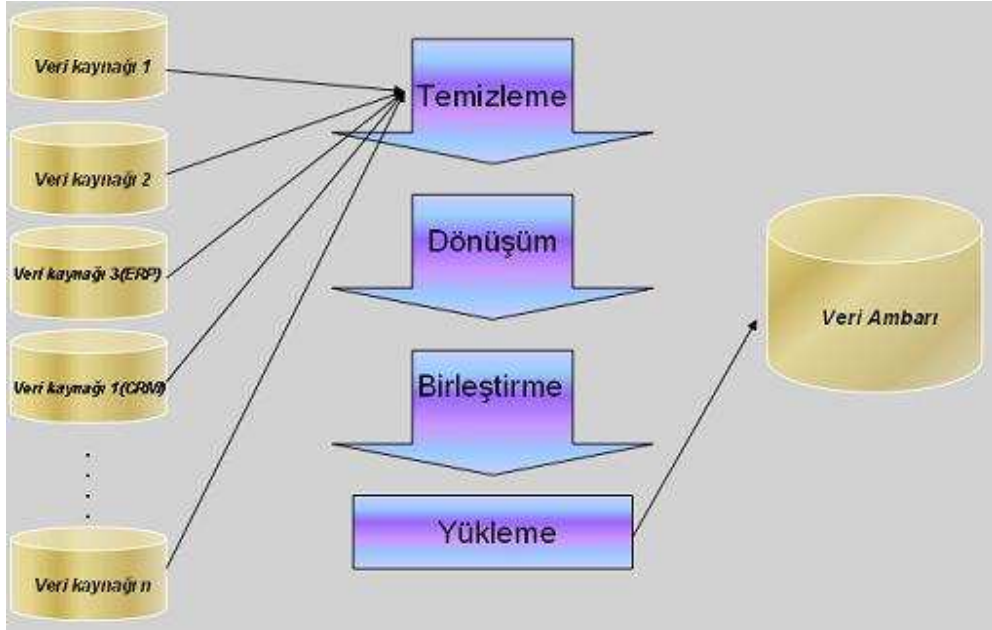
Veri ambarı çözümleri, detay verilerin stratejik ve anlık yönetim süreçlerinde doğru ve faydalı kullanımına yönelik çözümlerdir. Operasyonel veriler, tek bir yapıda birleştirilerek doğru ve hızlı analitik sorgulara hazır hale getirilir. Veri Ambarları aşağıdaki analiz ve uygulamalara kaynak oluşturmaktadır:

- Müşteri Analizleri
- Portal Uygulamalar
- Finansal Raporlama ve Analizler
- Tedarikçi Performans Analizleri
- Kurumsal Performans Yönetimi
- Varlık Yönetimi
- Risk Analizi
- Satış Analizi
- Operasyonel Analizler

Veri ambarlarında bulunan verilerin önemli özelliklerinden birisi de temiz ve dönüşümden geçmiş olmasıdır. Bu işlemlerden geçirilmesinin nedeni, veri ambarından bir veriye erişmek istediğimizde bu verinin hazır bir şekilde bulunması gerekliliğidir. Şekil 2.2 bu süreci göstermektedir.

²³ Veri ambarı (Datawarehousing) Temelleri, (t.y.)

http://www.bilisimrehber.com.tr/arastirma/tr_arastirma_veriambari_temelleri.phtml (17 Kasım 2007)



Şekil 2.2 : Veri Ambarı Yaratma Süreci ²⁴

Kendi içinde karar destek sisteminin başarıyla uygulayan Şişecam'ın Enformasyon Teknolojisi Müdürü Şahin Soysal, sistemin faydasını kendi uygulamalarından verdiği bir örnekle özetliyor: "Her yıl Eylül-Ekim aylarında, bir sonraki yılın bütçesini yaparız. Bütçe çalışmalarının çeşitli fazlarından en önemlisi satış bütçemizdir. Bütçeyi yaparken geçmiş verilere bakarız fakat, geçmiş yıllara çok farklı açılardan bakmak gerekiyor. 'Hangi pazarda, hangi ürünler, ne zaman nasıl satılmış?' gibi tüm parametreleri birbiriyle uyumlu bir şekilde eşleştirip sağlıklı bir bütçe yapabilmek için çok yetenekli araçlara ihtiyacınız var. Ürün, pazar, çevre koşulları, promosyonlar, tedarik planlarının hepsini muhtelif senaryolarla ilişkilendirip optimize etmek gerekiyor. Bunu da ancak, bir karar destek sistemi ve veri ambarı aracılığıyla yapmanız mümkün..." ²⁴

²⁴ Veri Ambarı (Datawarehousing) Temelleri, (t.y.)

http://www.bilisimrehber.com.tr/arastirma/tr_arastirma_veriambari_temelleri.phtml (17 Kasım 2007)

Karar verme sürecini işletme lehine önemli ölçüde kısaltan bu sistemler, sadece yöneticiler için değil, istediği bilgiyi anında sistemden çekebilen şirket çalışanlarının verimliliğinde de önemli bir rol oynamaktadır.²⁵

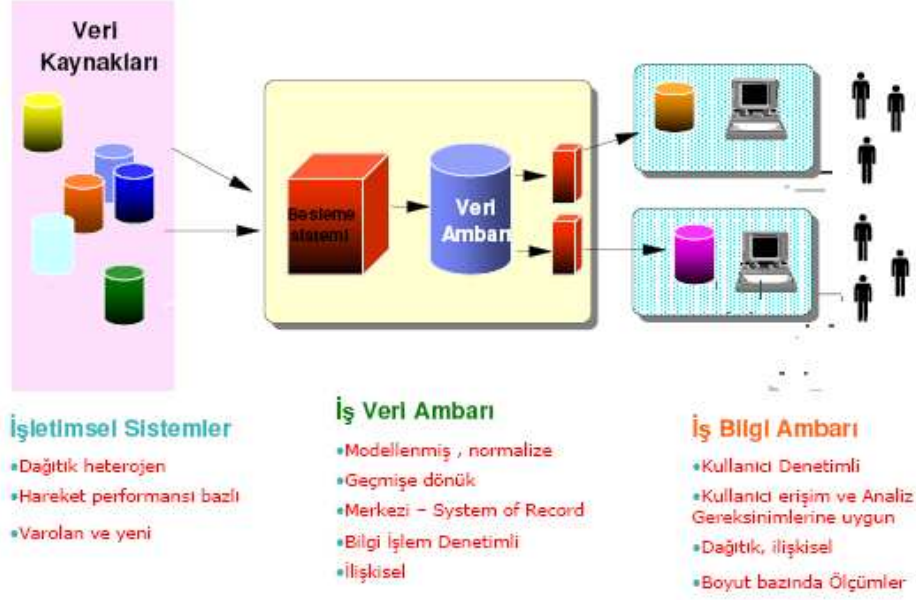
2.5. Veri Ambarı ve Data Mart'lar

Bill Inmon, veri ambarı teknolojilerinin gurusu sayılan kişi, 1990'lı yılların başlarında veri ambarcılığının formalize edilmesini sağlamıştır. Bill Inmon'un düşüncesi, her yöneticinin yapması gerekenin, tek bir sistemde toplanmış eski verilere ulaşmaktan çok, tüm sistemlerden gelen en son verilere ulaşması gerektiğiydi. Buna Veri Ambarı dedi. Zaman değişkenli ambardaki veri yalnızca doğru zaman aralığında geçerlidir. İstikrarlı gerçek zamanda güncelleşmeyen fakat işletimsel sistemlerde düzenli olarak yenilenebilir. Her zaman veriler veritabanına eklenebilir ve bütünlük her zaman sağlanmaktadır. Bu formalizasyon kurumsal çaplı veri tabanlarını, üretim veya daha genel anlamıyla operasyonel veri tabanlarından ayırmıştır.

Bununla beraber, bütçe ve kaynak kısıtlamalarından ötürü, birçok 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 ya da konuya özel bir veritabanıdır. Data mart, bir departmanın kesin ve net sorgulama ihtiyaçlarının kümesini veya oradaki fonksiyonel son kullanıcıların ihtiyaçlarını desteklemek amacıyla tasarlanmaktadır.²⁶

²⁵ Fazla Bilgi Eziyor, 2002 <http://www.coretech.com.tr/Templates/haberdetay.aspx?articleid=38&zoneid=18> (4 Şubat 2008)

²⁶ Data Mart vs Data Warehouse - The Great Debate, 2008 <http://opensourceanalytics.com/2006/03/14/data-mart-vs-data-warehouse-the-great-debate> (5 Şubat 2008)



Şekil 2.3 : Veri Ambarı ve Data Mart²⁶

2.6. Çok Boyutlu Veri Görünümü - OLAP

Veri ambarcılığı tüm veri dosyalarını merkezi bir veri havuzuna koyarak işlemler üzerine inşa edilmiş çok boyutlu bir iş görünümü yaratır. Bunu yapmak için, veri ambarcılığı bağlantılarına ve ilişkilerine “denormalizasyon ve özetleme” denen bir süreç uygulanır.

İlişkisel veri tabanlarının yaygınlığı ve sonrasında ortaya çıkan Veri Ambarlarının gelişmesi ile beraber, verilere daha hızlı şekilde erişme ve çok boyutlu analiz ihtiyaçları, bilim adamlarını ve yazılım şirketlerini, daha farklı yapılar geliştirmeye itmiştir. Bu amaçla geliştirilen bir teknoloji olan OLAP (On-line Analytical Processing), ilişkisel veri tabanları gibi, bilimsel temeller üzerine değil, OLAP ürünleri üreten firmaların desteğinde çıkan bir teknoloji olmuştur. Bir veri yapısının OLAP olarak nitelendirilebilmesi için 12 kural belirlenmiştir. Bu kurallar sırası ile şu şekildedir:

- Çok boyutlu inceleme özelliğine sahip olması,
- Şeffaflık,
- Erişilebilirlik,
- Her seviyede sorgulama için aynı performansı gösterebilme özelliği,
- İstemci-Sunucu yapısında olması,
- Sınırsız şekilde çapraz raporlama olanağının olması,
- En alt seviyedeki verilerin otomatik olarak ayarlanması,
- Her şarta uygun boyutlandırılabilirlik,
- Çok kullanıcı desteğinin olması,
- Her seviyede verilerin değiştirilebilir olması,
- Esnek raporlama özelliği,
- Boyut ve gruplamalarda sınır olmaması.

Zaman kazancının dışında OLAP, verilere çok boyutlu bakabilme özelliği, karmaşık hesaplamalar yapabilme ve zaman boyutunu beraberinde barındırabilme gibi çok önemli özellikleri de beraberinde getirmektedir.²⁷

OLAP ile veriye çok boyutsal görünüm sağlanmaktadır. İlişkisel OLAP (ROLAP), ilişkisel tablolar arasındaki boyutlar ve olguların oluşmasıdır. ROLAP yapısıyla yıldız şema gibi veri modelleme teknolojilerinin oluşturulmasıyla verinin ilişkisel görünümü elde edilir. Çok boyutsal OLAP (MOLAP) ise çoklu hiyerarşik yapıya sahip boyutlar ile veri küpleri elde edilerek her boyutu, veride bir alana karşılık gelen çok boyutlu bir küp olarak bakmayı ve incelemeyi sağlar. Böylece boyut bazında gruplama, boyutlar arasındaki korelasyonları inceleme ve sonuçları grafik, rapor gibi görsellik ön planda olarak sunma olanağı sağlar.²⁸

²⁷ OLAP-1, (t.y.), http://www.danismend.com/konular/bilgiveteknoyon/bilgi_olap1.htm (5 Şubat 2008)

²⁸ Doç. Dr. M. Ali Salahlı , **Veri Tabanı Yönetim Sistemleri Ders Notları**, 1. Baskı, Çanakkale: Çanakkale Onsekiz Mart Üniversitesi, 2006, s. 15-19

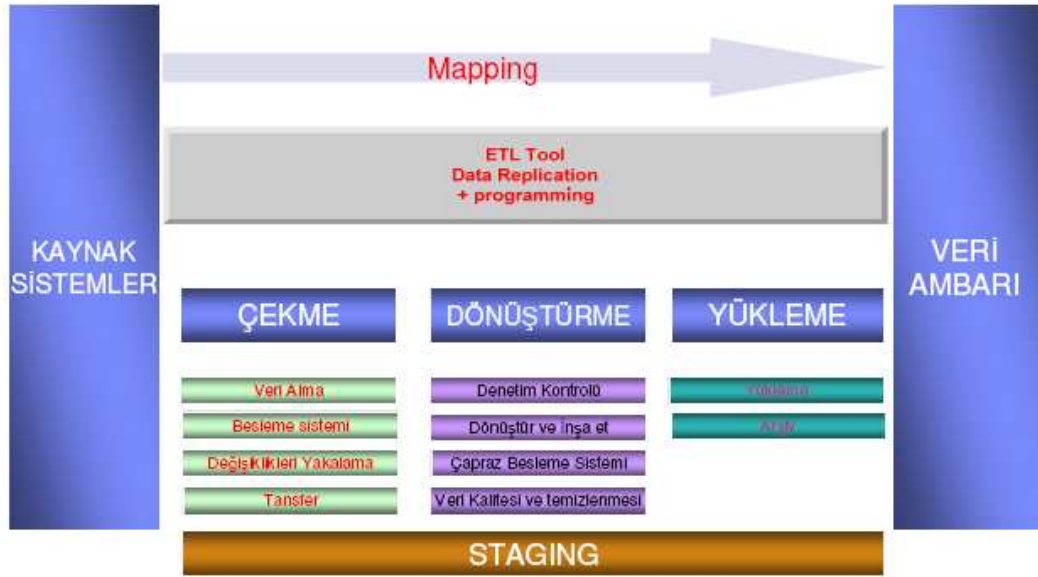
2.7. Veri Entegrasyonu ve ETL (Çekme Dönüştürme Yükleme)

Kurumların sahip olduğu veri kaynaklarının çeşitliliği gün geçtikçe artmaktadır. Bir kurumda operasyonel sistemlerde Oracle, SQL Server, Sybase, DB2, Access, Excel, DBF gibi farklı ortamlarda biriktirilen verilerle karşılaşmak çok beklenen bir durumdur ve bir çok organizasyonda bu veri kaynaklarının bir çoğunu aynı anda bulunmasına ek olarak, son yıllarda ön plana çıkan ve özellikle üzerinde planlama/bütçeleme gibi dikey uygulamalarla birlikte gelen MS Analysis Server, DB2 OLAP server, Essbase veri kaynakları ya da rekabet bilgileri için web'ten periyodik olarak kurum içine çekilen veri kaynakları gibi ortamlar, veri ambarı oluşturarak raporlama yapmak isteyen kurumlar için, verilerin entegrasyonu sırasında büyük sıkıntılara neden olmaktadır.

Bu aşamada kurumların seçmesi gereken 2 yol bulunmaktadır. Kurumlar, ya klasik ve geleneksel yöntemlerle her farklı veri kaynağına erişim sırasında bir takım programlar yazacak, ya da bu işler için özellikle geliştirilmiş, hem hayata geçirilmesi, hem bakımı çok daha kolay olan ETL yazılım araçları kullanacaklardır. Verilerin artması ve veri kaynaklarının çeşitlenmesi, özellikle 2. yöntem olan ETL yazılım araçlarının ve ETL süreçlerinin, kurumların değerlendirme süreçlerinde önceliklendirmesine neden olmaktadır. ETL (Extract-Transform-Load (Çekme-Dönüştürme-Yükleme)) işlemi, heterojen kaynaklarda bulunan verilerin çekilerek, gerekli dönüştürmeler, temizlemeler ve özetlemeler yapılarak veri ambarına yüklenmesi işlemidir. Örneğin, farklı kaynaklarda tutulan cinsiyet kodlarının tek kod yapısına dönüştürülmesi, 2000 yılından önce tutulan veriler için yıl hanesinin 4 basamağa dönüştürülmesi, veri tabanı kısıtlarına uymayan kayıtların ayıklanması, mantıksal hatalar içeren kayıtların temizlenmesi gibi aşamalar ETL'in dönüştürme kısmında yapılmaktadır.²⁹

²⁹ R. Borland, **Data Warehouse Operation Strategy and ETL Processing Part I**, 1. Baskı, Boston: MIT Press, 2000, s. 98-99

Veri Entegrasyonu süreçlerindeki ETL optimizasyonu sağlayabilen şirketler verilerinin geçerliliği, kalitesi, kullanılabilirliği, güvenilirliği ve bu verileri kullanarak uyumlu İş Zekâsı ortamı oluşturabilirler. Şekil 2.4 ETL süreçlerini göstermektedir.³⁰



Şekil 2.4 : Çekme Dönüştürme Yükleme (ETL) Süreç Tasarımı³⁰

2.8. Veri Madenciliği

Karmaşık ve dinamik bir ekonomi ortamında işletmelerin karar verme süreçlerindeki etkinlikleri, çok değişkenli büyük miktarlardaki veri kümelerinde saklı bulunan bilginin elde edilmesine ve işlenmesine bağlıdır. Bu noktada, doğru yanıtlandığında işletmelerin isabetli karar almalarında belirleyici rol oynayacak stratejik önemde sorular vardır. Örneğin;

- En iyi müşterilerimizi nasıl koruyabiliriz?

³⁰ R. M. Bolten, "Data Population in a BI Environment", **DM Review Magazine**, Eylül 2000, s. 15-17

- Hangi müşterilerin şirketi bırakıp rakiplere gitme olasılığı daha yüksektir?
- Müşterilerin gelecekte nasıl davranacağı doğru bir şekilde öngörülebilir mi?

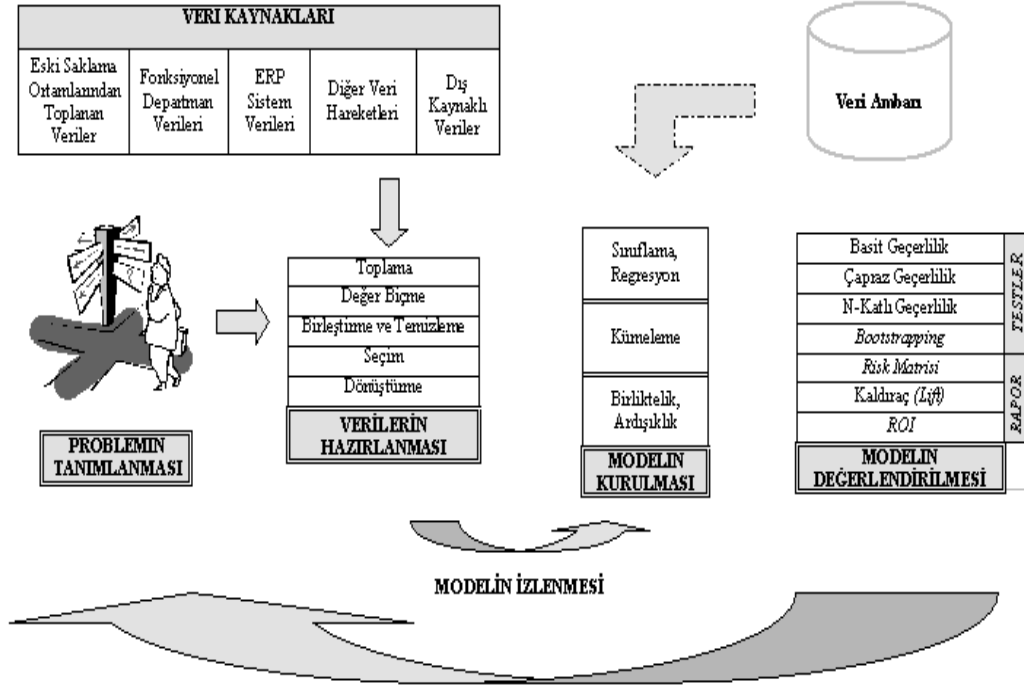
Veri madenciliği bu veri yığınlarının içinden işletme yöneticileri için en gerekli olanların seçilmesi, düzenlenmesi ve modellenmesi süreçlerini içerir. Bu noktada veri madenciliğini, karar verme mekanizmaları için yeni bilgiler üreten teknikler ve kavramlar bütünü olarak tanımlamak mümkündür.

Veri madenciliği, veri ambarlarındaki tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş verileri ortaya çıkarma, bunları karar vermek ve gerçekleştirmek için kullanma sürecidir. Bu tanımdan yararlanarak veri madenciliğinin aynı zamanda bir istatistiksel süreç olduğunu da söylemek mümkündür.³¹

Veri madenciliği süreci Şekil 2.5’ de gösterilmiştir. Veri madenciliği süreci, verileri veri ambarından alır, bunları derler, düzenler ve yorumlar.³²

³¹ A. Feelders, H. Daniels, ve M. Holsheimer, **Methodological and Practical Aspects of Data Mining, Information and Management**, 1.Baskı, New York: Wiley, 2000, s.37, 271-281

³² J. Han ve M. Kamber, **Data Mining: Concepts and Techniques**, 1. Baskı, Londra: Academic Press, 2001, s. 111-116, 175-180



Şekil 2.5 : Veri Madenciliği Süreci ³²

Veri madenciliğinin karar verici için olası yararları aşağıdaki gibi sıralanabilir:

- Mevcut müşterilerin karar verici tarafından daha iyi tanınmasını sağlayabilir.
- Özellikle finans sektöründe mevcut müşterileri bölümlere ayırıp, kredi risk davranış modelleri oluşturarak, yeni başvuruda bulunan müşterilere karşı riskin minimize edilmesini sağlayabilir.
- Mevcut müşterilerin ödeme performansları incelenip kötü ödeme performansı gösteren müşterilerin ortak özellikleri belirlenerek, benzer özelliklere sahip tüm müşteriler için yeni risk yönetim politikaları oluşturulabilir.
- En iyi müşteriler veya müşteri bölümlerinin bulunmasında kullanılabilir. Bulunan bu iyi müşteri bölümlerine yönelik yeni pazarlama stratejileri oluşturulabilir.

- Kuruluşlar tarafından düzenlenecek çeşitli kampanyalarda mevcut müşteri kitlesinin seçimi ve bu müşterilerin davranış özelliklerine yönelik kampanya şartlarının oluşturulması sağlanabilir.
- Bankacılık faaliyetlerinde, küçük işletmelere yönelik olarak makine ve ekipman satışı yapan dağıtıcı firmalarla ortak hareket ederek oluşturulacak satış paketleri ile pazarlama stratejileri geliştirilebilir.
- Mevcut müşteriler üzerinde firma ürünlerinin çapraz satış kapasitesinin artırılması sağlanabilir.
- Mevcut müşteriye tanıyarak kuruluşların müşteri ilişkileri yönetimlerinde düzenleme ve geliştirmeler yapılabilir. Bu sayede kuruluşun müşterilerini daha iyi tanıyarak müşteri gibi düşünme kapasitelerinin artırılması sağlanabilir.
- Günümüzde var olan yoğun rekabet ortamında kuruluşların hızlı ve kendisi için en doğru kararı almalarını sağlayabilir.
- Kuruluşlar veri analizi ile müşterilerini kişiselleştirilmiş ürün ve hizmetler hakkında bilgilendirebilirler.
- Veri madenciliği ile kuruluşların müşteriyle bütünleşmiş satış politikaları oluşturması sağlanabilir.³³

³³ Gökhan Silahtaroglu, **Kavram ve Algoritmalarıyla Temel Veri Madenciliği**, 1. Baskı, İstanbul: Papatya Yayıncılık Eğitim, 2008, s. 121-132

3. VERİ KALİTESİ YÖNETİMİ

Günümüz iş zekası hizmetleri, bu hizmet alanı ile geçmişe yönelik veri saklayan sistemlerdeki verinin kalitesini ve günümüz koşullarındaki işe yararlılığını ölçmeyi ve arttırmayı, anlamlı hale gelmesini sağlamayı ve müşterilerin hızlı ve esnek raporlama gereksinimlerini karşılamayı hedeflemektedir. Bu fonksiyonun gerçekleşmesi için özel yazılımlar kullanılmakta, çeşitli istatistiksel metotlar ile verilerin analizi gerçekleştirilmektedir. Veri Kalitesi Yönetimi ile, uygulamalardaki veriler firmaların talepleri doğrultusunda düzeltilmekte veya yeniden sisteme girilmektedir. Özellikle veri dönüşümünün yaşandığı platformlara sahip her organizasyonda eski sistemlerde var olan hataların giderilmesi veri kalitesinin temel amaçlarındandır.³⁴

Kurumlarda müşteriler hakkında ana iki başlık altında bilgi toplanabilir. Bunlardan birincisi, demografik bilgilerdir. Kişilerin yaşı, yaşadıkları yer, eğitim düzeyleri vs. bu başlık altında incelenebilir. Diğer ise kişilerin kurum ile ilişkileri başladıktan sonraki davranışlarını içerir. Kredi kartı harcamaları, alışveriş alışkanlıkları vs. bu başlıkta incelenebilir.

Kişilerin davranış bilgileri, eğer işlemlerin tutulduğu sistemler düzgün tasarlanmış ise, kolaylıkla analiz edilebilir ortamlara alınabilir ve üzerinde incelemeler yapılabilir. Ancak demografik bilgiler hem toplanması hem güncel tutulması en zor bilgilerdir. Bugün, finansal kurumlar başta olmak üzere bir çok kurum, üzerinde analiz yapacak, hatta bazı yasal gereksinimleri karşılayacak doğru kalitede veriye sahip bulunmamaktadır. Bu durumda veri kalitesinin önemi ve kurum bünyesindeki tüm analiz ve karar alma süreçlerine etkisi ön plana çıkmaktadır.

³⁴ Veri Analizi Hizmetleri, (t.y.), http://www.deloitte.com/dtt/section_node/0,1042,sid%253D8431,00.html (8 Mart 2008)

Veri kalitesi sürekli bir şekilde müşteri ihtiyaçlarına hitap eden bilgi olarak tanımlanabilir. Kendine özgü olarak mantıksal ve tutarlı sırada bilginin düzenlenmesi süreci olarak da veri kalitesi tanımını yapmak mümkündür. Amaç; veriyi kuvvetlendirmek, zenginleştirmek, temizlemek ve birden fazla aynı amaca hizmet eden veriyi tek bir kayda indirmektir.³⁵

Bilgi kalitesi, sadece veri kalitesini ölçmek değildir, aynı zamanda kalitesiz bilginin iş maliyetlerini de ölçmek demektir. Bilgi kalitesini artırmak sadece veri temizleme için değil, süreç içerisinde tekerrür eden yapılardaki olası sorunların sebeplerini bulmak ve düzeltmek içindir.³⁶

Aşağıda veri kalitesi yönetimi hakkında bu konuda uzman kişilerin yaklaşımlarına yer verilmiştir.

“Eğer herhangi bir veri kalitesi süreciniz yoksa, elinizdeki veriden değerli çıkarımlar sağlamaya başlayabileceğinizi düşünmeyin.”

David Loshin

“Veri kalitesi yönetimi anlayışını benimsemeyen kurumlar, işletme gelirlerinin %10 ile %20’den daha fazlasını veri kalitesi ile doğrudan veya dolaylı olarak ilişkili problemlerin giderilmesine harcamaktadırlar! “

Larry English

"Birçok kurum gerçek veri yönetimi prensiplerini anlamamaktadır. Veri kalitesi sadece veri temizleme anlamına gelmemektedir."

Larry English

³⁵ A. No, “Frequently Asked Questions about Data Quality”, **DM Review Magazine**, Kasım 2000, s. 18-21

³⁶ Larry English, “Which Information Quality Tool is Best?”, **DM Review Magazine**, Ekim 2000, s. 8-9

“Kötü veri kalitesi, arabanın öncamındaki kirlere benzer. Uzun bir süre hafif görüş kaybı ile yolunuza devam edebilirsiniz, fakat bir noktada ya durup camı temizlemeniz gerekir ya da herşeyi riske atmanız.”

Ken Orr

"CRM projelerinin başarısızlığındaki en önemli neden: Veriye önem verilmemektedir. Kurumlar sahip oldukları verilerin kalitesi ile ilgili – veriyi nasıl temizleyecekleri, nasıl temiz tutacakları, nerede depolayacakları ve hangi 3. parti verilerine ihtiyaç duyacakları-ayrıntılı bir anlayışa sahip olmalıdırlar. Çözüm: Bir veri kalitesi stratejisi belirleyin. CRM projelerine ayırdığınız zamanın yarısını veri ile ilgili süreçlere ayırın. "

Scott Nelson and Jennifer Kirby - "Seven Key Reasons Why CRM Fails"

Kurumlarda veri kalitesi sorununa neden olan başlıca etmenler aşağıdaki gibi sıralanabilir:

- Tamlık: Verideki eksik bilgiler
- Uygunluk: Standart dışı formatta saklanmış veriler
- Anlaşılabilirlik: Verinin kullanıcılar tarafından anlaşılır olması
- Tutarlılık: Birbiri ile çelişen bilgiler
- Doğruluk: Yanlış veya eskimiş bilgiler
- Tekerrür: Aynı bilginin farklı şekillerde tekrarlanması
- Bütünlük: Bilgilerin arasındaki ilişkilerin saptanamaması

Kalite sorununa neden olan etmenler göz önüne alınarak varolan verinin denetimi yapıldığında eldeki verinin durumu tespit edilmiş olur. Daha sonra, veri kalitesi kuralları tanımlanır ve firmanın ulaşmak istediği hedefler belirlenir. Örneğin "adres bilgisinde %90 doğruluk" gibi. Veri kalitesini iyileştirmek için planlar yapılır. Oluşturulan plan ve kurallar uygulanarak veri kalitesi iyileştirilir. Raporlar ile veri kalitesinin yeni durumu tespit edilir. Böylece veri kalitesi sürekli olarak izlenebilir. Bu süreç bir döngü halinde uygulandığı takdirde veri kalitesi kontrol altında tutulabilmektedir. Veri kalitesi yönetimi müşteri

ilişkileri, kaynak planlaması, veri ambarı ve iş zekası uygulamaları, veri eşleme, veri göçü gibi birçok konuda kullanılır.³⁷

Veri ambarları projelerinde oldukça önemli iki nokta, veri temizliği ve veri kalitesi artırma çalışmalarıdır. Veri tekrarlarının ortadan kaldırılarak, birleştirilmiş ve zenginleştirilmiş verilerle yapılan analizler şirketleri en etkin yoldan doğru sonuca götürmektedir.

Larry English, veri kalitesi çözümünü; eğitilmiş kalifiye insanlar, tanımlanmış iş hedefleri, tanımlanmış veri kalitesi süreçleri ve kendisini ispatlamış araçların kullanımının bir toplamı olarak tanımlamıştır. Buradan da anlaşılabacağı gibi sadece veri kalitesi denetimini yapabilen ve kaliteli veriye ulaşımı sağlayabilen yazılımların kullanımı organizasyonları kaliteli ve güvenilir veriye ve bununla beraber veri kalitesi yönetimine götürmemektedir. Şirket kültürü açısından bakıldığında veri kalitesi uygulamalarına başlanması, kültürel bir değişim veya bilinen kavramı ile değişim yönetiminin temelini ifade etmektedir.³⁸

3.1. Veri Tekrarı, Veri Tutarsızlığı ve Veri Bütünlüğü

Aynı verinin değişik yerlerde kopyalarının bulunması, bir yerde güncellenen verinin diğer yerde güncellenmemesi olasılığına ve bu durum da veri tutarsızlığına yol açar. Veritabanı sistemleri ve uygulamaların bir bütün olarak tasarlanmaması, alt sistemler arasında ilişkinin düzgün olarak kurulamaması ve birden çok uygulamada verilerin aynı veritabanı içinde ortak kullanılacak biçimde tasarlanması veri tekrarına neden olmaktadır.

³⁷ Thomas C. Redman, **Data Quality For The Information Age**, 1. Baskı, Norwood:Artech House Publishers, 1997, s. 101-111, 266-268, 299

³⁸ [Larry English](#), “Information Quality Management: The Next Frontier”, **DM Review Magazine**, Nisan 2000, s.10

Veri tabanları ya da veri ambarlarında veri tekrarı arttıkça, veri tutarsızlığı da artacaktır. Bir başka anlatımla, veri tutarlılığı ve veri tekrarı ters orantılı değişkenlerdir.

Veri tekrarlamasını kontrol ederek veya ortadan kaldırarak veri tutarlılığına ulaşılabilir. Birleştirilmiş veri yönetimi, veri bütünlüğünü bir üst seviyeye taşıyacaktır. Veri tutarlılığı ve bütünlüğü arasında doğru orantı vardır.³⁹

3.2. Veri Kalitesizliğinin Sebepleri

İşletmelerde veri kalitesizliğine yol açan sebepler Tablo 3.1’de açıklanmıştır.

³⁹ A. Rudra ve E. Yeo, “Issues in user perceptions of data quality and satisfaction in using a data warehouse-an Australian experience”, **Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, 2000**. Vol.1, Ocak 2000, s. 7

Tablo 3.1 : Veri Kalitesizliğinin Sebepleri ⁴⁰

Ne zaman / Nerede	Niçin
Sistem değişimleri, yeni şekle dönüşümler veya yeniden yapılanmalar	*Değiştirme süreçlerinde yetersiz veri kalitesi testleri *Değiştirme programlarının ortaya çıkardığı yeni hatalar
Heterojen sistem entegrasyonları	*Çapraz sistemlerde değişken veya tutarsız veri *Entegrasyonda yetersiz veri kalitesi testleri
Heterojen sistemlerin post-entegrasyonu	*Veri çapraz sistemlerde çelişkili ya da tutarsızdır. *Zor farkedilen düşük kaliteli veri yeni senaryoların oluşmasına yol açar.
Üretim yazılımı	*Yazılım ihtiyaçları tamamlanmış veya geliştirme sürecindeki hatalar. *Yazılım mühendisliğinin ya da üretim kontrolünün olmaması.
Veri tabanı tasarımı	*Kayıt ve alan tanımlamalarının çok dağınık, yapısız veya normalleştirilmemiş olması. *Şema eksiklikleri, yetersiz kontroller ve bütünleştirme kuralları.
Veri yaşlandırma	*Şirket verinin yaşını takip edemez veya güncelleme/zenginleştirme için gerekli programlara sahip değildir.
Müşteri reaksiyonu/reaksiyonsuzluğu	*Veri hiçbir zaman tam yakalanmaz. *Müşteri formları kötü bir şekilde tasarlanmıştır veya müşteriyi reaksiyon vermesi için yeterince teşvik etmemektedir.
Sahtekarlık	*Fiziksel ve mantıksal sistem güvenliği sağlanamamıştır veya kontroller yoktur.
Giriş hatası	*Sistem giriş metotları kötü bir şekilde tasarlanmıştır veya otomatik onaylamalar yoktur. *İnsan hatalarına sistemsel olarak ele alınacak şekilde tasarlanmamıştır.
İş kuralları	*Sistem ihtiyaçları elverişli değildir veya mevcut iş kuralları veriye cevap vermemektedir.
Politika ve planlama	*Veri kalitesi yönetimine karşı olan yönetim duyarsızlığı

⁴⁰ C. Firth, (1997), "When do Data Quality Problems Occur?", <http://sunflower.singnet.com.sg/~cfirth/dq1.htm> (13 Mart 2008)

Rudra, A. ve Yeo, E. veri kirliliğinin nedenselliğini araştırmışlardır. İnceleme, hem açık uçlu yani cevabı tartışmaya açık hem de tartışmasız sorular ile yapılmıştır. Sorulan kişilere beş noktalı ölçek üzerinde veri kirliliğinin sebep olduğu bazı noktalar sorulmuştur. Alınan sonuçlar nicel analiz yöntemiyle değerlendirilmiş ve ortalama puanlara göre veri ambarında veri kirliliğine sebep olan 3 temel neden aşağıdaki gibi bulunmuştur.⁴¹

- Zamanında güncellenmesi gerektiği halde güncellenmeyen/güncellenemeyen veriler
- Heterojen sistem bütünleştirilmesi
- Yönetimin planlama ve politika eksikliği

3.3. Veri Kalitesi Yönetiminde Karşılaşılabilecek Olası Veri Problemleri

Veri kalitesi yönetimi, verinin doğru, zamana uygun, konuya ve amacına uygun ve tutarlı olarak tüm organizasyon içindeki çoklu iş birimleri içerisinde standardının sağlanmasını amaçlamaktadır. Bu sayede doğru veriden doğru kararlar alınması ve organizasyonların stratejik değerlerini ve kararlarını doğru bir şekilde verebilmesi sağlanır. Müşteri İlişkileri Yönetimi, Tedarik Zinciri Yönetimi ve Kurumsal Kaynak Planlama gibi şirket uygulamalarının doğru sonuçlar verebilmesi için verilerin güvenilir ve kaliteli olması gerekmektedir.⁴²

Etkin veri kalitesi yönetimi, veri kalitesizliği maliyetini oldukça düşürmektedir. Veri Kalitesi Yönetimi de tıpkı Toplam Kalite Yönetimi gibi kurumsal çapta ve tek bir sistem gibi tasarlanarak geliştirilmesi gerektiğinden oldukça pahalı, zor ve yorucu bir yönetim felsefesidir.

⁴¹ Larry English, "Total Information Quality Management: A Complete Methodology for IQ Management", **DM Review Magazine**, 2003, s. 9

⁴² K.T. Chow, "Enterprise Data Quality: Covering All Your Bases", **DM Review Magazine**, Aralık 2000, s. 13-17

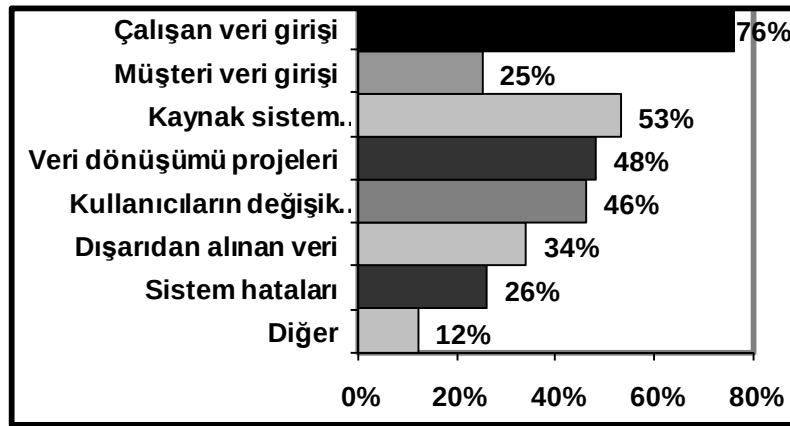
Veri kalitesinin deęerini gstermenin en iyi yollarından biri kt verinin kklerini gzlemlemek ve organizasyonu etkileyebilecek bozuk veriyi ortaya ıkarabilmektir. Veri kalitesi ynetiminde karřılařılabilecek olası veri problemleri ařaęıdaki gibi sıralanabilir:⁴³

- Yanlıřlar: Yazılıř hataları, baskı hataları, sınır deęeri tařmaları, uyumsuz veri tipleri vb.
- Eřsesliler: Yazılıřları aynı fakat anlamları aynı olan kelimeler. İerisinde eř sesli bulunan metinlerin veri kalitesi araları tarafından veya kullanıcılar tarafından yanlış yorumlanması verinin doęruluęu zerinde ciddi bir etkiye sahiptir. rneęin, bir isim alanında “Veli Ahmet řen” ifadesinin olması mřterinin Veli ve Ahmet diye iki isminin olduęu anlamına da velisinin isminin “Ahmet řen” olduęu anlamına da gelebilmektedir. Aynı řekilde adres verisi ierisindeki “Aralık, ıkmař” gibi isimler hem sokak belirtici olarak kullanılabilmekte (Ertuęrul ıkmař) hem de zel kurum isimleri veya apartman, site isimleri gibi bileřenlerin ismi (ıkmař Holding, Aralık Apartmanı) olarak kullanılabilmektedir.
- Standart Eksiklięi: Veri giriři sırasında en ok karřılařılan veri kalitesizlięine ve bozukluęuna yol aan sebeplerden birisi de belli bir standardın olmamasıdır. rneęin; adres verilerinde belirtilen “Sok., Sokak, Sokaęı, Sk.” veya aık alan sorularında belirtilen “Var, Evet, E, Yes, Y, Tamam vs.” gibi bu standart eksiklięine rnek olarak gsterilebilir.
- Yasal Zorunluluklar: İsimlendirme řekillerine ekleme ya da ıkartmalar eldeki verinin yasal amalarla kullanılması durumunda byk problemlere neden olmaktadır. Birok bankacılık ve finans uygulamaları olduka karıřık isimlendirme řekillerine sahiptirler ve dokmanın yasal erevelerde kalabilmesi iin bu isimlendirme řekillerine uyulması gerekmektedir. Bu zorunluluk řirketin dięer sistemlerinde depolanan verilerle uyumsuzluklara yol aabilmektedir.

⁴³ Michael L. Gonzales, President, The Focus Group, Ltd, “Enterprise Data Quality For Business Intelligence” [Brořr] 2003, s. 2-7

- **Kayıp/Doğrulanamaz/Hayali Veri:** Genellikle mevcut veri ilk bakışta, uygun bir yapıya ve kabul edilebilir değerlere sahiptir. Fakat tanımlama ve ilişkilendirme mekanizmaları tarafından tanınmadığı için zayıf, kalitesiz veri kümeleri olarak kalmaktadırlar. Organizasyonlar verilerinin genel bir analizini yaparak ve oransal doluluklara bakarak, verilerin doluluğunun her zaman doğrulanabilir doluluk olmadığını bilincine varmalıdırlar. Örneğin bir organizasyon Anne Kızlık Soyadı alanının doluluk oranına bakarak, müşterilerin %80'inin Anne Kızlık Soyadı alanının var olduğunu düşünebilir fakat içerideki verinin profillenmesi sonucu çıkabilecek ‘Yok, Mevcut değil, Bilinmiyor, Aynısı, Bulunamadı vs.’ gibi anlamsız verilerin elenmesi gerekliliği ve bunun sonrasında ancak gerçek doluluk ve doğruluğa ulaşılacağı unutulmamalıdır. Aynı şekilde adres alanından örnek verilmesi gerekirse, bir adresin “Mustafa Kemal Cad. İstanbul” olması bu adrese ulaşılabilir olduğunu göstermemektedir. Hemen hemen her ilçede, her semtte bir Mustafa Kemal Caddesi mevcuttur.

Kalitesiz verinin nedenleri ve bu sebeplerin oransal olarak dağılımı Şekil 3.1’de gösterilmiştir.⁴⁴



Şekil 3.1 : Veri Kalitesi Nedenleri Dağılımı⁴⁴

⁴⁴ *Data Quality and the Bottom Line*, 2002, <http://www.tdwi.org/research/display.aspx?ID=6589>, (8 Nisan 2008)

3.4. Veri Kalitesinin Ölçülmesi

Veri kalitesinin ne kadar iyi veya ne kadar kötü olduğunu söyleyecek ölçümler; tanımlamalar ve iş kuralları ile verinin ne kadar uygun gerçeklikte olduğunu gösteren istatistiksel raporlardır. Veri kalitesini ölçmek için kesin ve tek bir yöntem olmamakla beraber, mutlaka ölçümlenmesi gereken alanlar aşağıdaki gibidir: ⁴⁵

- Verinin bütünlüğü (Completeness)
- Veri girişinin tutarlılığı (Consistency)
- Verinin doğruluğu (Accuracy)
- Anahtar verilerin tekilliği (Uniqueness of primary keys)
- Veriyi destekleyen iş kurallarının sürekliliği (Durability of business rules)

Veri kalitesi ölçme genel yaklaşımlar çerçevesinde dört aşamadan oluşmaktadır: ⁴⁶

- **Önemli Verinin Belirlenmesi:** Stratejik öneme sahip olan veriler ve analiz, raporlama, modelleme ve anahtar performans göstergelerinin belirlenmesinde kullanılan verilerin saptanması ve veri kalitesinin bu veriler üzerinden doğru periyotlar ve doğru yöntemlerle gerçekleştirilmesi veri kalitesi yönetimi için ilk ve en önemli adımlardan biridir. Sonuçta elde edilen sonucun organizasyon içerisindeki en önemli boyutların ve gerçeklerin gerçek bir resminin olacağı gözden kaçırılmamalıdır.

- **Veri Kalitesini Ölçmek için bir Yöntem Belirleme:** Kaliteyi ölçmek için literatürde en iyi yöntemin, kaynaktan hedefe doğru, istatistiksel olarak geçerli örneklerin kontrolü olarak belirtilmiştir.

⁴⁵ R. Forino, "Data e.Quality: Election 2000", **DM Review Magazine**, Ocak 2001, s. 23-25

⁴⁶ Nat Bongiovanni, "Analyzing Domain Statistics", **DM Review Magazine**, Haziran 2000, s. 13-16

- Ölçüm Metodu Olarak Eğilim Yüzdeleri Analizinin (Trend Analizi) Kullanılması: Organizasyon içerisinde saklanan verilerin belli istatistiki değerlerinin baz kabul edilerek, değişen ve eklenen verinin de bu değerler çerçevesinde değerlendirilmesi yoluyla, yeni eklenen verinin tolerans sınırları içerisinde kalıp kalmadığının izlenmesi yöntemiyle yeni gelen veride sınırlar dışında değişikliklerin olup olmadığı saptanabilir. Bu yaklaşımın temelinde yatan teori; eğilim çizgisinden sapmaların yani tolerans sınırının dışında kalan değişikliklerin kolayca ve anında farkedilebilmesidir. Sonuçların grafik yöntemiyle görselleştirilmesi sonuçların takip edilebilirliği ve anlaşılabilirliğini artıracaktır.
- Kalite Güvenilirliğini Otomatikleştirme: Varolan problemlerin tespitine ek olarak her boyut için verilerin beklenen değerleri ve istatistikleri de belirlenir. Belirlenen tolerans değerlerine göre veri ambarlarındaki verilerin kalitesi belirlenen periyotlarla izlenebilir, bu süreç otomatikleştirilip kalite sorunlarına neden olacak etkenler ilk elden düzeltilebilir.

3.5. Veri Kalitesi Araçlarının Seçiminde Kritik Unsurlar

Stratejik kurumsal uygulamalar, müşteri ilişkileri yönetimi ve İş Zekası uygulamaları tarafından üretilen bilgilerin güvenilirliği, ancak onları besleyen verilerin güvenilirliği kadardır. Bu tür çözümleri gerçekleştirmiş birçok kurumda kullanıcılar veri ambarındaki verileri, doğrulukları ve tutarlılıklarına tam olarak güvenmedikleri için kullanmamaktadırlar. Sonuçta, büyük yatırımlara mal olmuş olan çok sayıda kurumsal bilgi teknolojisi projesi beklenen geri dönüşü sağlayamamaktadır.

Verilerin doğru olarak birleştirilmesini ve güvenilir bilginin sağlanmasını olumsuz yönde etkileyen pek çok faktör bulunmaktadır. Kurumlardaki sistemlerin ve standartların çeşitliliğinin artması, üçüncü parti verilerinin entegrasyonundaki zorluklar, tekrarlanan veriler ve uygulamalar, veri girişlerindeki hatalar ve eksiklikler bunlardan birkaçıdır. Çoğu durumda veri standartları ya yoktur ya da kurumsal boyutta uygulanmamaktadır.

Birçok kurum veri kalitesi probleminde habersizdir. Veriyi kullananlar, sonuçta elde edilen bilginin yanlış olduğu ortaya çıkana kadar verinin hatalı olduğunu fark etmemektedirler. Stratejik kurumsal projeler beklenen geri dönüşleri sağlayamadıkları zaman, sorunun çoğunlukla verideki eksiklikler, gereksiz tekrarlar ve tutarsızlıklardan kaynaklandığı ortaya çıkmaktadır.

Veri kalitesi ile ilgili sorunlara bir çözüm arayışının özellikle son dönemde arttığı gözlenmektedir. Bu konuda atılacak küçük adımlar daha önemli kazanımlara dönüşebilmektedir.

Veri Kalitesi Çözümü, kurumsal bilginin kalitesini artırarak, daha doğru, daha tutarlı, tekilleştirilmiş bilgilerle daha etkin kararların alınmasında kritik bir rol üstlenmektedir. Çözüm, verilerdeki tutarsızlıkların giderilmesini, farklı sistemler üzerine yayılmış verilerin etkin şekilde entegre edilmesini, aynı yapı ve formata getirilmesini, veriler içindeki tekrarların ortadan kaldırılarak “gerçeğin tek ve tutarlı versiyonu”nun elde edilmesini sağlamalıdır.⁴⁷

İyi bir çözümde kurumların verilerini analiz etmek, temizlemek, zenginleştirmek, tamamlamak ve birleştirmek üzere güçlü bir kurumsal platform sağlanmalı, bunun sonucunda kurumlarda standartlaşmış ve kalitesi artırılmış bir veri ortamı oluşturmalıdır. Veri kalitesini sağlamak üzere tutarlı olmayan, standartlara uymayan, tekrarlar içeren veriler üzerinde güçlü profillemeye, analiz etme, ayrıştırma (parsing), standartlaştırma, eşleştirme, kümeleme ve tekilleştirme rutinleri uygulanmalıdır.

Veri Kalitesi araçlarında aranması gereken özellikler aşağıdaki gibi sıralanabilir:⁴⁸

- Kurumsal verilerin kalitesinin denetlenmesi, takibi ve aktif olarak yönetilmesi: Veri Kalitesi Çözümü, veri kalitesinin kurumsal boyutta analiz edilmesini ve

⁴⁷ <http://www.sas.com> (8 Nisan 2008)

⁴⁸ **Knightsbridge Solutions LLC**, “White Paper - Worst Practices in Business Intelligence: Why BI Applications Succeed Where BI Tools Fail”, 1. Baskı, [Broşür] 2005

değerlendirilmesini sağlamaktadır. Profilleme işlemi potansiyel boşlukların olduğu alanların belirlenmesini ve bunların giderilmesi için yapılması gerekenleri ortaya çıkartmaktadır. Böylece kurumların geliştirilecek temel veri alanlarına ve iş süreçlerine odaklanmaları sağlanmış olmaktadır.

- Çoklu sistemler ve iş birimleri arasında dağılmış verilerin entegre edilebilmesi ve standart hale getirilebilmesi: Veri Kalitesi Çözümü ile kurumlar, veri kalitesini sağlayacak iş kurallarını veri kaynakları ve platformlar arasında ortak kullanabilmektedir. Uçtan uca denetim olanakları ile veri kalite algoritmaları ve çözümleri uygulanarak sistem bazında kalite sorunları ortaya çıkarılmaktadır. Standart ve özel süreçler yürütülerek farklı sistemler üzerindeki veriler tek ve doğru bir görünüm olarak düzenlenmekte ve standartlaştırılmaktadır.

- Veri düzeltme ve temizleme kurallarının tanımlanması, değişikliklere kolayca uyulanması ve verilerin temizlenmesi: Kolay kullanımlı ara yüzler, iş analistlerinin veri standartlaştırma ve iş kurallarını kolayca tanımlamalarını ve bu tanımların etkilerini gözlemlemelerini sağlamaktadır. Modern veri kalitesi araçları hem iş kullanıcılarının hem de teknik kullanıcıların verileri temizlemelerini, standartlaştırmalarını, entegre etmelerini ve verideki eksiklikleri tamamlamalarını olanaklı kılmaktadır.

- Hatalı ve bozuk veri girişinin önlenmesi: Verilerin sisteme girişi sırasında verideki problemlerin belirlenip giderilmesi, veri kalitesinin sağlanmasının en düşük maliyetli yöntemidir. Çözümün gerçek zamanlı veri kalitesini sağlayan bileşeni, veri kalitesi rutinlerinin hareket bazlı sistemler içinden kullanılmasını sağlamaktadır. Böylece hatalı veri girişi engellenebilmektedir.

- Karar vericilere güvenilecekleri bilginin sağlanması: Doğru veriler, tutarlı raporlama ve analitik sonuçları doğurur. Veri kalitesinin otomatik olarak ETL ve veri entegrasyonu süreçlerine entegre edilmesiyle, iş zekâsı ve analitik uygulamalar için kullanılan veriler doğru ve güncel olacaktır. Sonuç olarak karar vericiler kendilerine iletilen bilgilere güvenilecek ve güven içinde karar alabileceklerdir.

3.6. Veri Kalitesi Uygulama Süreci

Bir veri kalitesi aracı aşağıdaki veri kalitesi süreçlerini gerçekleştirebiliyor olmalıdır:⁴⁹

- Ölçümleme: Kalitesiz ve bozuk veriyi sayısal ve tıpsel olarak belirleyip ölçümleme.
- Analiz Edebilme: Kalitesiz verinin doğasını ve nedenlerini analiz edebilme.
- Standartlaştırma: Ölçüm değerlerinin ve veri değerlerinin normalize edilmesi, belirlenen iş kurallarına veya üçüncü parti referans verisine göre verilerin formatlanabilmesi.
- Düzeltme: Verilerin düzeltilmesi ve gerek kendi içinde belli algoritmalarla gerekse referans veri tabanları ile iletişim kurarak doğrulanabilmesi.
- Ayırıştırma: Veri elemanlarını veri yapılarına ve belli bölüm/alanlara ayırıştırabilme.
- Zenginleştirme: Ek bilgilerin varolan bilgiden çıkartılması ve anlamlandırılıp zenginleştirilmesi (Adresten Posta kodunun bulunması, isim-soyad'tan cinsiyet analizi, il kodu alanından eksik adres bilgilerinin doldurulması vs.).
- Eşleştirme: Çoklu kayıtların aynı ya da farklı veri tabanlarından bulunup belli algoritmalar kullanarak bu kayıtların tekilleştirilmesi.
- Birleştirme: Tek bir oluşum için çoklu veri kayıtlarının tekil bir kayıta ve tek bir veri kaynağında birleştirilebilmesi.
- Raporlama ve İzleme: Veri Kalitesi bilgisinin sürecin her aşamasında raporlanabilmesi ve izlenebilmesi.

Şekil 3.2. veri kalitesi sürecini göstermektedir.⁵⁰

⁴⁹ **Knightsbridge Solutions LLC**, “White Paper - A Practical Approach to Data Quality: Proven Formulas for Delivering Measurable Results”, 1. Baskı, [Broşür] 2006



Şekil 3.2 : Veri Kalitesi Süreci ⁵⁰

3.6.1. Veri Profilleme

Veri kalitesi metodolojisinde, bir kurumun veri kalitesini arttırmaya başlayacağı ilk nokta Veri Profilleme olarak nitelendirilmektedir. Veri Profilleme sırasında verilerin sayısallaştırılması ve analiz edilmesi mümkündür. Veri yapısı, içeriği ve diğer verilerle ilişkilerin saptanacağı ve veri kalitesi çalışmalarının ilk adımı olarak kabul edilen aşama profillemeye aşamasıdır. Bu sayede örneğin bir tablo alanı içindeki verilerin kendini ne kadar

⁵⁰ **Business Objects**, "White Paper - Data Quality First: It's Just Logical - The Importance of Trusted Information to Successful Data Integration and Business Intelligence", 1. Baskı, [Broşür] 2006

tekrar ettiđi, ne kadarının boş, ne kadarının dolu olduđu, ya da iki tablo arasındaki tutarsız bilgilerin oranı gibi bilgilere sayısal olarak erişmek ve gerektiğinde detaya inerek bunların hangi kayıtlardan geldiđini görmek mümkündür. Bazı veri kalitesi araçları, veri kalitesinin istenilenden daha bozuk olduđu durumlarda verilecek kurallarla, ilgili kişilere otomatik mail gönderme, ve kurumun sahip olduđu veri kalitesi indeksinin bir kontrol paneli (dashboard) ortamında takip edilmesi gibi özelliklere de sahiptir.⁵¹

Veri profillemeye sayesinde eldeki verinin durumu ve sorunları (eksiklik, standartlara uygunsuzluk, tutarsızlık gibi) gün ışığına çıkacak ve neyle uğraşıldığı detaylı olarak anlaşılabilecektir. Örneğin, bir banka müşteri veri tabanında yapılacak denetim sonucunda müşteriler hakkında olmazsa olmaz vergi numarası bilgisinin %60 oranında boş; telefon numaralarının da %40 oranında sadece 5 ve daha azı rakamdan oluşan yanlış bilgiden oluştuđu; birden fazla bulunan adres alanındaki bilgilerin de tutarsız doldurulduđu görülebilir. Bu ilk aşama, temizliğe başlamadan önceki durumu tespit etmek ve yol haritasını çizmekte bir temel teşkil etmek amacına hizmet edecektir.

Veri profillemeye araçları genellikle yapılan profillemeye sonuçlarını grafiksel olarak gösterebilmektedirler. Şekil 3.3 bir veri tabanındaki “Satışlar” tablosundaki değişkenlere ait kayıt sayısını ve her bir kaydın boş değerlerini göstermektedir.

⁵¹ **DataFlux Corporation**, “White Paper - Five Steps to More Valuable Enterprise Data”, 1. Baskı [Broşür] 2007



Şekil 3.3 : Veri Profillemeye Grafiksel Gösterim I

Veri profillemeye çalışmalarında profillemeye verinin deseni de analiz edilmelidir. Veri deseni, profillemeye alan sayısal olması gerekirken, karakter ifadeler girilmiş ise ya da tam tersi durumlarda analiz ettiğimiz verinin yapısal durumunu anlamamıza ve veri temizleme kurallarının netleşmesinde faydalı olacaktır. ⁵²

Şekil 3.4 örnek bir veri tabanındaki “Telefon” alanının veri desenini göstermektedir. Veri deseninden de görüldüğü gibi telefon alanı deseni standart değildir. Burada 9 herhangi bir sayısal değeri, “A” ise karakter değeri ifade etmektedir.

⁵² **Business Intelligence Network**, “Research Report - Data Profiling, Data Integration and Data Quality: The Pillars Of Master Data Management”, 1. Baskı [Broşür] 2007

Field: Phone				
Defined type: VARCHAR				
Defined length: 15 chars				
Column Profiling	Frequency Distribution	Frequency Distribution (Chart)		
Pattern	Frequency Distribution	Pattern Frequency Distribution (Chart)	Percentiles	Outliers
Pattern	Count	Percentage		
999-999-9999	3166	96.73		
(999)999-9999	42	1.28		
(999) 999-9999	34	1.04		
999 99 9999 999	20	0.61		
999 999 9999	5	0.15		
999-999-AAAA	2	0.06		
9-999-999-9999	2	0.06		
a	1	0.03		
99 99 9999 999	1	0.03		

Şekil 3.4 : Veri Profilleme Grafiksel Gösterim II

Şekil 3.5 örnek bir alanın istatistik analizini göstermektedir. Veri profilleme çalışmalarında dikkat edilmesi gereken en önemli istatistikler arasında boş değer sayısı, minimum değer, maksimum değer desen analizi ve frekans dağılımıdır.

Field: LoanAmount		
Defined type: double		
Defined length: 53 bit		
Column Profiling	Frequency Distribution	Frequency Distribution (Chart)
Pattern Frequency Distribution	Pattern Frequency Distribution (Chart)	Percentiles
Outliers		
Metric Name	Metric Value	
Data Type	double	
Primary Key Candidate	no	
Unique Count	1140	
Uniqueness	70.11	
Pattern Count	(not applicable)	
Minimum Value	-223000	
Maximum Value	9999999	
Minimum Length	(not applicable)	
Maximum Length	(not applicable)	
Null Count	2	
Blank Count	(not applicable)	
Actual Type	double	
Count	1628	
Data Length	53 bit	
Mean	114348.170972	
Median	4888499.5	
Mode	0	
Non-null Count	1626	
Nullable	YES	
Ordinal Position	7	
Decimal Places	0	
Standard Deviation	429438.361236	
Standard Error	10649.778281	

Şekil 3.5 : Veri Profilleme Grafiksel Gösterim III

Veri Profillemeye yazılımları bir kurumda, veriden sorumlu, verinin sahibi olan iş tarafındaki birim ya da departmanlar tarafından kullanılacak nitelikte olmalı. Bu da bu tür yazılım araçlarının, teknik birimlerden çok son kullanıcıların kullanabileceği bir arayüze sahip olmasını gerektirmektedir. Örneğin, bir veritabanında E-mail1 ve E-mail2 gibi 2 farklı alan olduğu halde bunlardan sadece 1 tanesinde e-mail adresi yazılıysa, diğer alan boş olsa bile “doğru bir veridir bu!” diyebilen bir kullanıcı bu kuralı kendisi tanımlayabilmelidir. Böyle bir kullanıcı proaktif yöntemlerle, veri kalitesini periyodik ve otomatize yöntemlerle takip edebilmeli ve gerektiğinde yine otomatik olarak uyarılar gönderebilmelidir. Örneğin böyle bir uyarı kurumun veri kalitesi standartları olarak nitelendirilebilecek bir limitin altında kalması durumunda ilgili kişilere bir mail gönderilmesi şeklinde hayata geçirilebilir. Bu tür kuralları verecek olan kullanıcılar veri kalitesinin artırma sürecinde temel olarak teknik ekiplerden çok iş tarafındaki kullanıcılar olarak değerlendirilmelidir.⁵³

3.6.2. Veri Standartlaştırma ve Düzeltme

Veri profillemeye sonuçları doğrultusunda iş kuralları çerçevesinde belli veri standartlaştırma ve düzeltme kuralları tanımlanır. Buradaki amaç, veri ambarındaki verilerin formatlarının, içeriklerinin ve anlamlarının belli teknik standartlara oturtulabilmesidir. Verinin doldurulması, belli standartlara oturtulması ve tekilleşmesi bu aşamada gerçekleşecektir. Veri düzeltme çalışmaları verinin kendi içerisinde tanımlanan kurallarla yapılabileceği gibi referans veri tabanları yardımıyla (Vergi kimlik numarası doğrulama ve yanlış algoritmadaki vergi numaralarının silinmesi) da yapılabilir.⁵⁴

⁵³ Larry English, “Information Quality in Reference Data”, **DM Review Magazine**, Mart 2006, s. 10

⁵⁴ **DataFlux Corporation**, “White Paper - Five Steps to More Valuable Enterprise Data”, 1. Baskı [Broşür] 2007

Veri standartlaştırma aşamasında bazı veri kalitesi araçları veri benzerliğine ve frekans dağılımına bakarak otomatik standartlaştırma şemaları hazırlayabilmektedirler. Unutulmaması gereken bir nokta, otomatik olarak hazırlanan şemalar frekans dağılımı baz alınarak belli algoritmalar kullanılarak yüzdesel benzerlik baz alınarak hazırlanan şemalar olacaktır. Bu durumda gözle kontrol gerekir ve gerekli yerlerde standartlaştırma şemasına eklemeler yapılacaktır. Şekil 3.6 örnek bir standartlaştırma şemasını göstermektedir. Sol taraftaki “data alanı” gerçek il bilgisini, sağ taraftaki “Standart” alanı da standart formattaki veriyi göstermektedir.

Scheme		Entries: 440
Data	Standard	
Ç KALE	ÇANAKKALE	
ÇANAK KALE	ÇANAKKALE	
ÇANAKKAL E	ÇANAKKALE	
Ç KALE	ÇANAKKALE	
ÇANK1R1	ÇANKIRI	
ÇANKIRI	ÇANKIRI	
ÇANK1R1	ÇANKIRI	
ÇORUM	ÇORUM	
ÇORUM	ÇORUM	
ÇORUM	ÇORUM	
İST	İSTANBUL	
İSTAN	İSTANBUL	
İSTANBUL	İSTANBUL	
İSTNBL	İSTANBUL	
İSTANBUL	İSTANBUL	
İST	İSTANBUL	
İST	İSTANBUL	
İSTAN	İSTANBUL	
İSTANBUL	İSTANBUL	
İSTNBL	İSTANBUL	
İ S T A N B U L	İSTANBUL	
İSTANBUL	İSTANBUL	
İ.S.T	İSTANBUL	
İ.S.T	İSTANBUL	
İ S T A N B U L	İSTANBUL	
İSTANBUL	İSTANBUL	
İST	İSTANBUL	
İSTANBUL	İSTANBUL	

Şekil 3.6 : Veri Standartlaştırma Şeması Örneği

Örnek bir veri tabanında “e-mail” alanına yapılan profillemeye sonuçları ve bu sonuçlar doğrultusunda alınan iş kararları ve standartlaştırma kuralları Tablo 3.2’deki gibidir.

Tablo 3.2 : Veri Profilleme, İş Kuralları, Veri Standartlaştırma Kuralları Örneği

E-MAIL Alanı	
PROFİLLEME SONUÇLARI	
	· İncelenen toplam kayıt sayısı: 9.127.712
	· Boş kayıt sayısı: 8.563.255
	· Dolu kayıt sayısı : 564.457
	· Farklı kayıt sayısı: 522.773
	· En çok girilmiş bilgi : Space %93,82
	· İkinci en çok Nokta (.) girilmiş %1,64
	· @(:) den sonra .com, .gov, .edu , .net gibi ifadelerden biri e-mail adreslerinde bulunmuyor.
	· “yok:yok” 5136 defa, “YOK:YOK” 4146 defa, “yoktur:yoktur” 567 defa girilmiş.
İŞ KURALLARI	
	· e-mail adreslerinde Türkçe karakter yer almamalı.
	· E-mail adresleri en az 8 karakter içermelidir.
	· İçerisinde mutlaka @ (:) bulunmalıdır.
	· @(:) den sonra en az bir nokta (.) olmalıdır.
	· E-mail adresleri küçük harf ile girilmelidir.
	· @(:) den önce tek başına “yok” ya da “YOK” kelimesi olmamalıdır.
	· İçerisinde “www” olmamalıdır.
STANDARTLAŞTIRMA ve DÜZELTME KURALLARI	
	· Tüm e-mail adresleri küçük harfe çevrilmeli.
	· Var olan Türkçe karakterler İngilizce karakterlere çevrilmeli.
	· Yukarıdaki iş kurallarına uymayan kayıtlar null (ya da boşluk) ile değiştirilmelidir.
	· E-Mail alanındaki ilk 4 karakteri www. olarak geçen bilgiler web adresi olarak işaretlenecektir.

3.6.3. Verinin Dönüştürülmesi ve Ayırıştırma (Parsing)

Veri tabanlarında bulunan bilgiler, her zaman kullanıma uygun ve belli iş ihtiyaçlarını ve analiz çalışmalarını yapabilecek nitelikteki veri birimlerine ayrılmış durumda değildir. Örnek olarak, adres verisi birçok firmada tüm adres verisini tek bir alanda içerecek şekilde tutulmaktadır. “ÇANAKKALE SERAMİK SİTESİ KÜÇÜK BLOKLAR B D KAPISI KAT: 3 D: 17 ETİLER İSTANBUL TÜRKİYE” adresinde belli bir il, ilçe, semt ya da daha küçük adres bileşenleri bazında analiz yapılması gerekliliğinde, kampanya çalışmaları ya da hilekarlık tespiti analizlerinde kurumlar bu adres verisinin bileşenlerini belirlemek ve standardizasyonunu sağlamak zorunluluğundadırlar. Bu ihtiyaçlar çerçevesinde Veri Kalitesi Yönetimi felsefesi doğrultusunda, veri alanlarının belli bileşenlere ayrıştırılması ve veri dönüştürülmesi ile yeni bilgi alanlarının ortaya çıkarılması sağlanmaktadır.

Organizasyonlarda görülen ve en çok veri ayrıştırmasına konu olan alanlar aşağıdaki gibidir: ^{55,56}

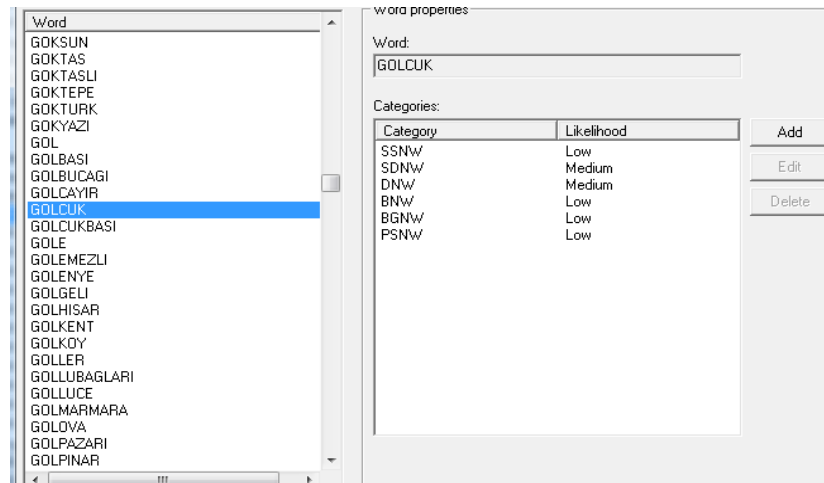
- Adres alanının ülke / il / ilçe / semt / cadde / sokak / site adı / bina adı / bina numarası / posta kodu gibi bileşenlere ayrıştırılması
- İsim Soyadı alanının ünvan/isim/soyad/ek bilgi/firma ismi/firma uzantısı bileşenlerine ayrıştırılması
- Adres alanlarının içerisinde bulunan telefon, firma adları gibi ek bilgilerin ayrıştırılması,
- E-mail alanlarının kullanıcı adı/domain bileşenlerine ayrıştırılması
- Telefon alanlarının ülke kodu/alan kodu/baz numara/dahili numara/ek bilgi bileşenlerine ayrıştırılması.

⁵⁵ **DataFlux Corporation**, “White Paper - Five Steps to More Valuable Enterprise Data”, 1. Baskı [Broşür] 2007

⁵⁶ **DataFlux Corporation**, “White Paper - Proving Value for Data Efforts: An ROI Framework for Data Governance and Data Quality Projects”, 1. Baskı [Broşür] 2007

Veri ayrıştırma çalışmalarında “Anlamlandırılmış Bilgi Tabanı” kullanılır. “Anlamlandırılmış Bilgi Tabanı”, genellikle ilgili veriye baz oluşturacak dev bir veri sözlüğü ve verilerin ilişkilerini belirleyecek anlamlandırılmış referans bir veri deposudur. Bu bilgi deposunda standartlaştırma şemaları, veri sözlüğü, gramer kuralları kütüphanesi, fonetik kurallar kütüphanesi bulunmaktadır. Kullanılan anlamlandırılmış bilgi deposu sayesinde eldeki veriler anlamlı bir şekilde standartlaştırılıp, ayrıştırılır ve anlamlı bilgi bileşenlerine dönüştürülür.⁵⁷

Veri sözlüğünde, analiz edilecek verideki bileşenleri oluşturacak her bir kelimenin hangi gruba girdiği belirtilir. Şekil 3.7’de “Gölcük” kelimesinin bir adres içerisinde hangi kategorilerde yer alabileceğini gösteren veri sözlüğü yapısı gösterilmektedir. “Gölcük” bir adres içerisinde hem ilçe, hem semt ismi olarak, hem cadde, sokak ismi olarak, hem site ismi, hem de bina ismi olarak kullanılabileceğinden bu kategorilere atanmıştır. Kategori seçiminde referans veri tabanlarından, eldeki verinin analizinden ve uzman kişilerin bilgilerinden yararlanılır. Adres gibi hemen hemen her kelimeyi içerebilecek ve belli bir standarda oturmamış alanlarda veri sözlüğü, gramer kütüphanesi, standartlaştırma şemaları hazırlamak oldukça zor, zaman alıcı ve kurumdan kuruma farklılık gösterebilecek bir çalışmadır.



Şekil 3.7 : Veri Ayrıştırması Çalışmalarında Kullanılabilecek Veri Sözlüğü Örneği

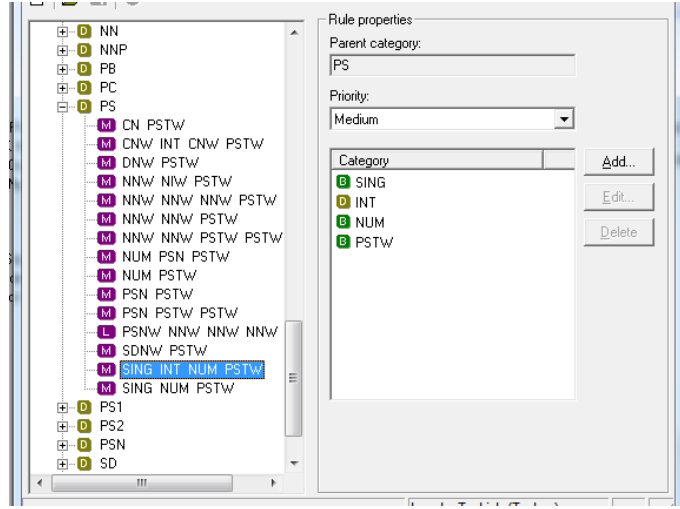
⁵⁷ Resources, (t.y.) <http://www.dataflux.com/Resources/> (8 Nisan 2008)

Fonetik kuralları kütüphanesinde, kullanılan dildeki bazı şekilsel bozulmalar ve birbiri yerine kullanılması muhtemel ses yanlışlarını içerebilecek kurallar tanımlanır. Örneğin, hastane ve hastahane ve hastane, pastahane ve pastane gibi “tahane” yerine “tane” kullanımı çok yaygındır. Veri yapısında ortak bir standardı yakalamak ve verileri düzgün dönüştürebilmek için fonetik kütüphaneleri oluşturulur. Şekil 3.8 örnek bir fonetik kuralları kütüphanesini göstermektedir.

[illegible]

Şekil 3.8 : Fonetik Kuralları Kütüphanesi Örneği

Veri ayrıştırmasında en önemli ve en zor tanımlardan bir tanesi, veri örtüsüne, veri dizilişine ve iş kurallarına bakılarak gramer kuralları kütüphanesinin oluşturulmasıdır. Gramer kuralları, veri sözlüğünde belirlenen kategorilerin hangi sırada gözlemlenebileceğini gösterir. Şekil 3.9, “Cadde (PS)” alanının hangi kategorilerden oluşabileceğini göstermektedir. Cadde ismi “259/1 Cadde”, “32. Cad.”, “Barbaros Caddesi”, “Mustafa Kemal Atatürk Cd.”, “M.Kemal Atatürk Cad.” gibi çok farklı şekilde kurallarla ya da kuralsızlıklarla ifade edilebilir. Cadde bilgisini adres alanından ayrıştırabilmek için “Cadde” ifadesinin hangi şekillerde yazılabileceğinin gramer kütüphanesinde tanımlanmış olması gerekmektedir.



Şekil 3.9 : Gramer Kuralları Kütüphanesi Örneği

Veri ayrıştırma ve dönüştürme aşamasında tanımlanan “Anlamlandırılmış Bilgi Tabanı” kullanılarak veriler anlamlı ve tanımlı alt bileşenlere ayrıştırılır. Tablo 3.3 ve 3.4 “Adres” alanının örnek bir veri tabanındaki standartlaştırılmış halini ve alt bileşenlerine ayrılmış parçalarını göstermektedir.

Tablo 3.3 : Adres Verisi Ayrıştırma Örneği I

ADRES (Standartlaştırılmış)	DALYAN MAH. 4206. SOK. NO: 5 ÇEŞME İZMİR TÜRKİYE 5328981236	ÇANAKKALE SERAMİK SİTESİ KÜÇÜK BLOKLAR B D KAPISI KAT: 3 D: 17 ETİLER TÜRKİYE	BÜYÜKDERE CAD. NO: 117 GAYRETTEPE İSTANBUL 82900	SABANCI CENTER KULE 2 KAT: 3 LEVENT İSTANBUL 34289
Organizasyon İsmi				
Mahalle	DALYAN MAH.			
Cadde			BÜYÜKDERE CAD.	
Sokak	4206. SOK.			
Site/Bina Grubu		ÇANAKKALE SERAMİK SİTESİ KÜÇÜK BLOKLAR B		SABANCI CENTER
Bina Adı		D KAPISI		
Bina Numarası	NO: 5		NO: 117	KULE 2
Kat		KAT: 3		KAT: 3
Daire		D: 17		
Semt	ÇEŞME	ETİLER	GAYRETTEPE	LEVENT
Belde				
Posta Kodu			82900	34289
İlçe				
İl	İZMİR		İSTANBUL	İSTANBUL
Ülke	TÜRKİYE	TÜRKİYE		
Ek Bilgi	5328981236			

Tablo 3.4 : Adres Verisi Ayrıştırma Örneği II

ADRES (Standartlaştırılmış)	ESKİŞEHİR YOLU BETA SİTESİ NO: 8/13 ÜMİTKÖY ANKARA TÜRKİYE	ÇUKUROVA ÜNİVERSİTESİ FEN EDEBİYAT FAK.KİMYA BÖLÜMÜ ADANA TÜRKİYE	HALASKARGAZI CAD. NO: 243 BÜRO 27 OSMANBEY ŞİŞLİ İŞYERİ	ORGANİZESANAYİ BÖLGESİ İSKENDERUN TÜRKİYE PK : 240
Organizasyon İsmi		ÇUKUROVA ÜNİVERSİTESİ FEN EDEBİYAT FAK.KİMYA BÖLÜMÜ		
Mahalle				ORGANİZESANAYİ BÖLGESİ
Cadde	ESKİŞEHİR YOLU		HALASKARGAZI CAD.	
Sokak				
Site/Bina Grubu	BETA SİTESİ			
Bina Adı				
Bina Numarası	NO: 8/13		NO: 243	
Kat				
Daire			BÜRO 27	
Semt	ÜMİTKÖY		OSMANBEY	
Belde				
Posta Kodu				
İlçe			ŞİŞLİ	İSKENDERUN
İl	ANKARA	ADANA		
Ülke	TÜRKİYE	TÜRKİYE		TÜRKİYE
Ek Bilgi			İŞYERİ	PK : 240

3.6.4. Veri Zenginleştirme ve Veri Doğrulama

Varolan veriden yeni anlamlar çıkartabilme, farklı kaynaklarda farklı kodlamalarla ya da anlamsız içeriklerle depolanan verilerin anlamlı bilgiler haline getirilmesi, organizasyonlara değer katan veri zenginleştirme süreci Veri Kalitesi Yönetimi felsefesinin önemli adımlarından biridir.⁵⁸

Organizasyonlarda depolanan bilgi genellikle tam doğru değil ya da güncel değildir. Herhangi bir alanda depolanan ve o alanın içeriği ile ilgili olmayan veriler, organizasyon için değer taşıyabilir ve bu bilgiler çoğunlukla gözden kaçırılabilir. Örneğin; bir müşterinin telefon bilgisine erişilmek istenildiğinde telefon bilgisi alanına bakılır fakat adres ya da isim soyad gibi farklı alanlara yazılmış olabilecek telefon numaraları genellikle gözden kaçmakta ve kurumlar için faydalı bir veri olmaktan uzaklaşmaktadır. Veri zenginleştirme çalışmaları sırasında bunun gibi verinin içinde gizli olan verinin ortaya çıkarılması ana amaçtır.

Veri zenginleştirilmesi ve doğrulanması çalışmalarına örnek olarak aşağıdaki aktiviteler verilebilir:

- İsim soyad bilgisinden yola çıkılarak cinsiyet analizi yapılması,
- Adres verisinden yola çıkılarak eksik olan adres bileşenlerinin doldurulması (Semt/İlçe bilgisine göre il bilgisinin bulunması ve il alanının zenginleştirilmesi gibi),
- Telefon verisinden il kodunun bulunarak il bilgisinin zenginleştirilmesi,
- Posta kodu verisinden referans veri tabanları aracılığı ile eksik adres verilerinin ve telefon alan kodlarının zenginleştirilmesi
- Referans il/ilçe/semt kodları tablosundan veride yanlış olan il kodlarının bulunması

⁵⁸ **DataFlux Corporation**, “White Paper - Data Management: Why Your Organization Must Understand and Achieve It”, 1. Baskı [Broşür] 2007

- Referans adres veri tabanları kullanılarak varolan adresin geçerli bir adres olup olmadığının doğrulanması
- Referans telefon alan kodları tablosu kullanılarak geçersiz telefon numaralarının belirlenmesi ve geçerli telefonların doğrulanması
- Adres verisinden ya da posta kodundan adres koordinatlarının belirlenmesi

Tablo 3.5 örnek bir adres verisi kullanılarak alt bileşenlerine ayrıştırılmış ve zenginleştirilmiş veri bileşenlerini göstermektedir. Posta kodu, İl, Ülke, Telefon ve Şahıs/Kurum bilgileri orijinal adresten türetilmiş ve zenginleştirilmiş alanlardır.

Tablo 3.5 : Adres Verisi Zenginleştirme Örneği

ORJİNAL ADRES VERİSİ	ARMAN MOBİLYA A.Ş. DALYAN MH. BARBAROS BULVARI 43/1 SK. 15/3 ÇEŞME 2347898
Organizasyon İsmi	ARMAN MOBİLYA A.Ş.
Mahalle	DALYAN MAH.
Cadde	BARBAROS BULVARI
Sokak	43/1 SOK.
Bina Numarası	15
Daire	2
Semt	ÇEŞME
Posta Kodu	35785
İl	İZMİR
Ülke	TÜRKİYE
Telefon	2322347898
Şahıs/Kurum	KURUM

3.6.5. Veri Eşleştirme ve Tekilleştirme

Veri kaynaklarında genellikle aynı müşteriye ait çoklu kayıtlar bulunabilmektedir. Bu da gereksiz veri hacmine ve müşteriye tam olarak tanıyamama sorununa yol açmaktadır. Aynı zamanda farklı müşteri kayıtlarında farklı bilgiler bulunabilir. Bir kayıta anne kızlık soyadı varken diğerinde olmaması, ya da TC kimlik numarası alanının bir kayıta dolu iken diğer kayıta boş olması gibi.

Çoklu kayıtların iş kuralları çerçevesinde belli eşleştirme kuralları ve algoritmaları kullanılarak tekilleştirilmesi; organizasyonların müşteri ile tek bir iletişim noktası

kurabilmesini, yüksek değerli müşterilerin belirlenebilmesini, müşteri düzeyinde risk seviyesi hesaplanabilmesini, analizlerin ve modellemelerin doğru yapılabilmesini ve verinin en doğru, etkin şekilde kullanabilmesini sağlamaktadır.⁵⁹

Bir kurumda müşterinin tek bir görünümü yoksa; müşteri değişen bilgilerini güncellemek ve sahip olduğu tüm ürünlerde aynı bilgilerle tanınmak için aynı güncellemeleri defalarca yapmak/yaptırmak zorunda kalır, önemli müşterilere kötü hizmet verilebilir, müşteri/ ürün ilişkileri tam olarak anlaşılamaz. Tekilleştirilmiş müşteri yapısı ile müşteri bilgisini tek bir kez günceller, güncellenen bilgi tüm ürün ve uygulamalarda paylaşılır, tüm servis noktalarında müşteriye aynı şekilde davranılır, ürün opsiyonları ve fiyatlama müşteri isteğine ve değerine göre belirlenebilir, personelin aynı verileri tekrar tekrar girmesine gerek kalmayacak altyapı kurulacağından personel maliyetleri azalır.⁶⁰

3.6.6. Veri Kalitesinin Raporlanması ve İzlenmesi

Veri kalitesi çalışmaları başlanılıp bitirilen bir projeden çok devamlı raporlanan, denetlenen ve takibi yapılan bir yönetim felsefesidir. Toplam kalite yaklaşımı baz alınarak iş zekası alt yapısının güçlü, yapısal ve tekrarlanabilir süreçlere dönüşmesi sağlanmalıdır.

Yapılan çalışmaların başarısını ölçmek, yeni çalışmalar planlamak, organizasyonun istediği veri kalitesi seviyesine ne kadar ulaştığının ölçülmesi, veri kalitesi yatırımının organizasyona katkısının ölçülmesi ve finansal ölçüm kriterlerinin belirlenmesi; veri kalitesi yönetiminin getirilerinin kurumsal boyutta farkındalığını sağlayan önemli kriterler arasındadır.⁶¹

⁵⁹ Amit Rudra ve Emilie Yeo, 1999, "Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia", <http://dblp.uni-trier.de/rec/bibtex/conf/hicss/RudraY99> (8 Nisan 2008)

⁶⁰ Larry English, "IQ and Data Warehousing: Trends and Best Practices", *DM Review Online*, Nisan 2006, <http://www.infoimpact.com/articles/2006-04 IQ in DW-Trends & Challenges.pdf> (13 Nisan 2008)

⁶¹ **Dataflux Corporation**, "White Paper - Data Monitoring: Add Controls to Your Data Governance Program", 1. Baskı [Broşür] 2006

4. UYGULAMA

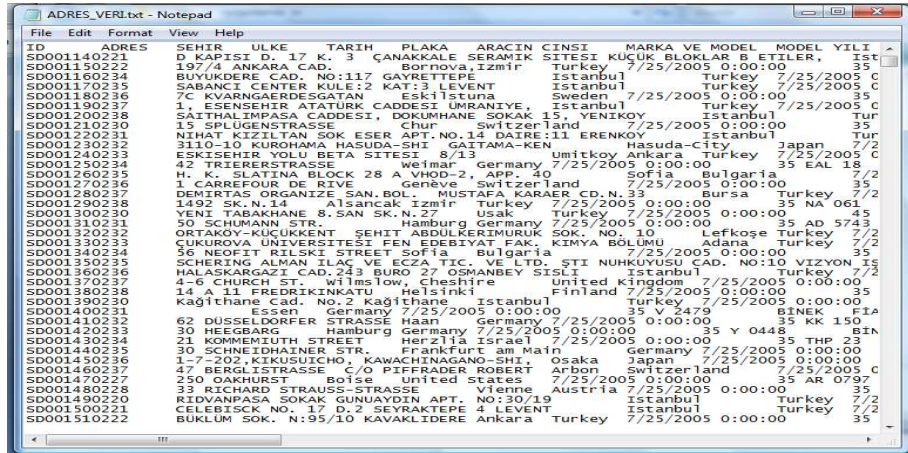
Veri Kalitesi uygulama çalışması olarak 2006 İstanbul Rallisine katılan kayıtlı potansiyal katılımcı adres verileri üzerinden adres verilerinin analiz edilmesi, profillenmesi, veri kalitesi kurallarının oluşturulması ve bu kurallar doğrultusunda veri temizlenmesi, standartlaştırma şemalarının oluşturulması ve adres bilgisinin standartlaştırılması, adres ayrıştırma kurallarının belirlenmesi, kullanılacak sözlüklerin ve gramer kütüphanelerinin yaratılması ve adresin bileşenlerine ayrıştırılması ele alınacaktır.

Uygulama DataFlux ürünü ile gerçekleştirilecek ve sonuçlar görsel olarak sunulacaktır. Dataflux ürününün seçilme sebepleri arasında kurulum ve kullanım kolaylığı, sonuçların görsel olarak ifade edilebilirliği, verilerin bileşenlerine ayrıştırma kurallarının kolayca tanımlanabilmesi, veri profillemesi, veri temizleme, veri ayrıştırma ve veri izleme fonksiyonlarının tek platformdan kullanılabilirliği sıralanabilir.

İş Zekası yazılım ve danışmanlık şirketi olan SAS'ın 2000 yılında bünyesine kattığı DataFlux, 2008 yılında dünyaca ünlü analist şirket Gartner'ın "2008 Magic Quadrant of Data Quality Tools" raporuna göre en etkili veri kalitesi yazılım çözümü olarak seçilmiştir. Veri profillemesi, veri kalitesi, veri entegrasyonu ve veri çoğaltma'dan oluşan DataFlux çözümleri, ellerindeki bilgi kaynaklarından anlaşılır değerler yaratmak isteyen kurumlara veri yönetimi çözümleri sunmaktadır. Dataflux'a rakip veri kalitesi çözümleri sunan ilk 4 şirket Business Objects, IBM, Informatica ve Trillium Software olarak sıralanabilir.

4.1. Adres Verilerinin Profillenmesi

Profilleme çalışması rasgele seçilen 3002 müşterinin “Adres, Şehir, Ülke, Tarih, Plaka, Aracın Cinsi, Araç Marka ve Modeli, Model Yılı, Özel Kişi, Tüzel Kişi bilgileri, Ad, Soyad, Firma Ünvanı, Görevi” verileri kullanılarak gerçekleştirilmiştir. Amaç veri kalitesi çalışmasına başlamadan önce eldeki veriyi tanımak ve oluşturulacak standartlaştırma ve ayrıştırma kurallarının doğruluğunu sağlayabilmektir. Şekil 4.1 veritabanından seçilen 3002 verinin metin dosyasına alınmış halini göstermektedir.



ID	ADRES	SEHIR	ULKE	TARİH	PLAKA	ARACIN CINSI	MARKA VE MODEL	MODEL YILI
SD001140221	D KAPISI D. 17 K. 3	ÇANAKKALE	SERAMİK	SİTESİ KÜÇÜK BLOKLAR 8 ETİZLER, İST				35
SD001150222	197/4 ANKARA CAD.	Bornova, İzmir	Türkey	7/25/2005	0:00:00			35
SD001160234	BUYUKDERE CAD. NO:117	GAYRETTEPE	Istanbul	Türkey	7/25/2005	0:00:00		35
SD001170235	SABANCI CENTER KULE:2 KAT:3	LEVANT	Istanbul	Türkey	7/25/2005	0:00:00		35
SD001180236	7C KVARNGERDEGATAN	ESKİSTUVA	Sweden	7/25/2005	0:00:00			35
SD001190237	1. ESENSEHIR ATATURK CADDESI	UMRANIYE	Istanbul	Türkey	7/25/2005	0:00:00		35
SD001200238	SAİTHALIMPASA CADDESI, DOKUMHANE SOKAK 15,	YENIKOY	Istanbul	Türkey	7/25/2005	0:00:00		35
SD001210239	15 SPLUGENSTRASSE	chur	Switzerland	7/25/2005	0:00:00			35
SD001220231	NIHAT KIZILTAN SOK ESER APT.NO.14	DAIRE:11 ERENKÖY	Istanbul	Türkey	7/25/2005	0:00:00		35
SD001230232	3110-10 KUROHAMA HASUDA-SHI	GAITAWA-KEN	Hasuda-City	Japan	7/25/2005	0:00:00		35
SD001240233	ESKİSEHIR YOLU BETA SİTESİ	8/13	Umitkoy Ankara	Türkey	7/25/2005	0:00:00		35
SD001250234	42 TREERERSTRASSE	Weimar	Germany	7/25/2005	0:00:00			35
SD001260235	H. K. SLATINA BLOCK 28 A	VHOD-2, APP. 40	Sofia	Bulgaria	7/25/2005	0:00:00		35
SD001270236	1 CARREFOUR DE RIVE	Genève	Switzerland	7/25/2005	0:00:00			35
SD001280237	DEMİRTAS ORGANİZE SAN.BOL.	MUSTAFA KARER CD.N.33	Bursa	Türkey	7/25/2005	0:00:00		35
SD001290238	1492 SK.N.14	Alsancak İzmir	Türkey	7/25/2005	0:00:00			35
SD001300239	YENİ TABAKHANE 8.SAN SK.N.27	Usak	Türkey	7/25/2005	0:00:00			35
SD001310231	50 SCHUMANN STR.	Hamburg	Germany	7/25/2005	0:00:00			35
SD001320232	ORTAKÖY-KÜÇÜKKENT ŞEHİT ABDÜLKERİMURUK SOK. NO. 10	Lefkoşe	Türkey	7/25/2005	0:00:00			35
SD001330233	CUKUROVA UNİVERSİTESİ FEN EDEBİYAT FAK. KİMYA BÖLÜMÜ	Adana	Türkey	7/25/2005	0:00:00			35
SD001340234	56 NEOFIT RILSKI STREET	Sofia	Bulgaria	7/25/2005	0:00:00			35
SD001350235	SCHERING ALMAN İLAÇ VE ECZA TİC. VE LTD. ŞTİ	NUHKUYUSU CAD. NO:10	VIZYON	Türkey	7/25/2005	0:00:00		35
SD001360236	HALASKARGAZI CAD.243	BURO 27 OSMANBEY SİSİ	Istanbul	Türkey	7/25/2005	0:00:00		35
SD001370237	4-6 CHURCH ST.	Wilmslow, Cheshire	United Kingdom	7/25/2005	0:00:00			35
SD001380238	14 A 11 FREDRIKKNKATU	Helsinki	Finland	7/25/2005	0:00:00			35
SD001390239	Kağıthane cad. No.2	Kağıthane	Istanbul	Türkey	7/25/2005	0:00:00		35
SD001400231	62 DÜSSELDORFER STRASSE	Haar	Germany	7/25/2005	0:00:00			35
SD001410232	30 HEEGBARG	Hamburg	Germany	7/25/2005	0:00:00			35
SD001420233	21 KORMEIMUTH STREET	Herzlia	Israel	7/25/2005	0:00:00			35
SD001430234	30 SCHNEIDHAINER STR.	Frankfurt am Main	Germany	7/25/2005	0:00:00			35
SD001440235	1-7-202, KIKUSUICHO,	KAWACHINAGANO-SHI,	Osaka	Japan	7/25/2005	0:00:00		35
SD001450236	47 BERGLISTRASSE	C/O PIRFRADER ROBERT	Arbon	Switzerland	7/25/2005	0:00:00		35
SD001460237	250 OAKHURST	Boise	United States	7/25/2005	0:00:00			35
SD001470227	33 RICHARD STRAUSS-STRASSE	Vienne	Austria	7/25/2005	0:00:00			35
SD001480228	RİDVANPASA SOKAK	GÜNÜAYDIN APT. NO:30/19	Istanbul	Türkey	7/25/2005	0:00:00		35
SD001490220	CELEBİSK NO. 17 D.2	SEYRAKTEPE 4 LEVENT	Istanbul	Türkey	7/25/2005	0:00:00		35
SD001500221	BUKLUM SOK. N:95/10	KAVAKLIDERE	Ankara	Türkey	7/25/2005	0:00:00		35
SD001510222								35

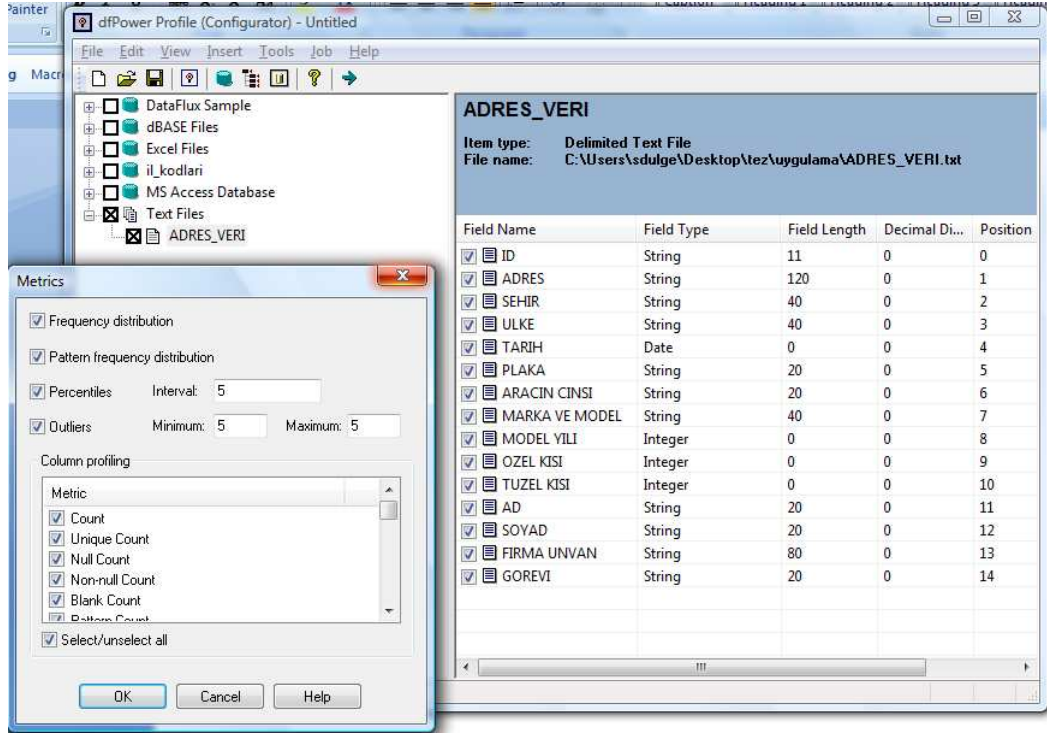
Şekil 4.1 : Profillenecek Müşteri Adres Verisi

DataFlux ürününün profillemeye modülü ile müşteri verileri profillenmiştir. Profillemeye aşamasında tüm değişkenler için hesaplanması istenen metrikler aşağıdaki gibidir:

- Frekans Dağılımı
- Veri Deseni Frekans Dağılımı
- Yüzdelik Değer (%5 aralıklarla)
- Uç Değerler (Minimum %5, Maksimum %5)

- Kayıt Sayısı
- Tekil Kayıt Sayısı
- Boş (null) Kayıt Sayısı
- Boş Olmayan Kayıt Sayısı
- Boş Kayıt Sayısı Yüzdesi
- Veri Deseni Kayıt Sayısı
- Minimum Değer
- Maksimum Değer
- Minimum Uzunluk
- Maksimum Uzunluk
- Ortalama
- Medyan
- Mod
- Standart Sapma
- Veri Tipi
- Veri Uzunluğu

Şekil 4.2 profilleme aşamasında seçilen değişkenleri ve metrikleri göstermektedir.



Şekil 4.2 : Profillemede Seçilen Değişkenler ve Metrikler

Profilleme sonucunda üretilen raporda her değişken için seçilen metrik değerler gösterilmektedir. Her değişken için frekans dağılımını, değişken profilini, veri deseni frekans dağılımını, yüzdelik değerleri, uç değerleri ayrı ayrı görmek mümkündür. Şekil 4.3. profilleme sonucu üretilen raporu göstermektedir.

Field Name	Ordinal Position	Count	Null Count	Blank Count	Minimum Value	Maximum Value	Mode	Pattern Count	Unique Count	Uniqueness	Primary Key
ID	1	3002	0	1	ID	TK151545508	(no data/ambig.)	3	3002	100.00	yes
ADRES	2	3002	0	118	"Emniyet Genel Müdürlüğü C	ŞİŞLİ İSTANBUL		2830	2855	95.10	no
SEHIR	3	3002	0	2117	.	İgel		57	223	7.43	no
ULKE	4	3002	0	2116	Albania	Uzbekistan		16	43	1.43	no
TARİH	5	3002	1173	(not applicable)	30.06.2005 00:00:00	10.08.2005 00:00:00	27.07.2005 00:00	(not applicable)	18	0.98	no
PLAKA	6	3002	0	1172	01 AJ 266	PLAKA		8	1831	60.99	no
ARACIN CINSI	7	3002	0	1172	ARACIN CINSI	ÇEKİCİ		10	14	0.47	no
MARKA VE MODEL	8	3002	0	1172	ALFA ROMEO 156	IVECO OTOTYOL		73	173	5.76	no
MODEL YILI	9	3002	1172	(not applicable)	0	2006	2005	(not applicable)	23	1.26	no
ÖZEL KİŞİ	10	3002	1172	(not applicable)	0	1	0	(not applicable)	2	0.11	no
TUZEL KİŞİ	11	3002	1172	(not applicable)	0	1	1	(not applicable)	2	0.11	no
AD	12	3002	0	2261	ABBAS	ŞÜKRÜ		49	433	14.42	no
SOYAD	13	3002	0	2261	ABRAVCI	ŞİRİN		23	559	18.62	no
FİRMA UNVAN	14	3002	0	1913	A.H.T. OTOTEM TURİZM RE	ŞERİF SEVEN İNŞ.NAK		152	153	5.10	no
GÖREVİ	15	3002	0	1173	ARAÇ KURTARMA	V.I.P. ARACI		12	22	0.73	no

Field: MARKA VE MODEL	
Defined type:	STRING
Defined length:	40 chars

Column Profiling	Frequency Distribution	Pattern Frequency Distribution	Percentiles	Outliers	Notes
------------------	------------------------	--------------------------------	-------------	----------	-------

Metric Name	Metric Value
Ordinal Position	8
Count	3002
Null Count	0
Blank Count	1172
Minimum Value	ALFA ROMEO 156
Maximum Value	IVECO OTOTYOL
Mode	
Pattern Count	73
Unique Count	173
Uniqueness	5.76
Primary Key Candidate	no

Şekil 4.3 : Profilleme Raporu

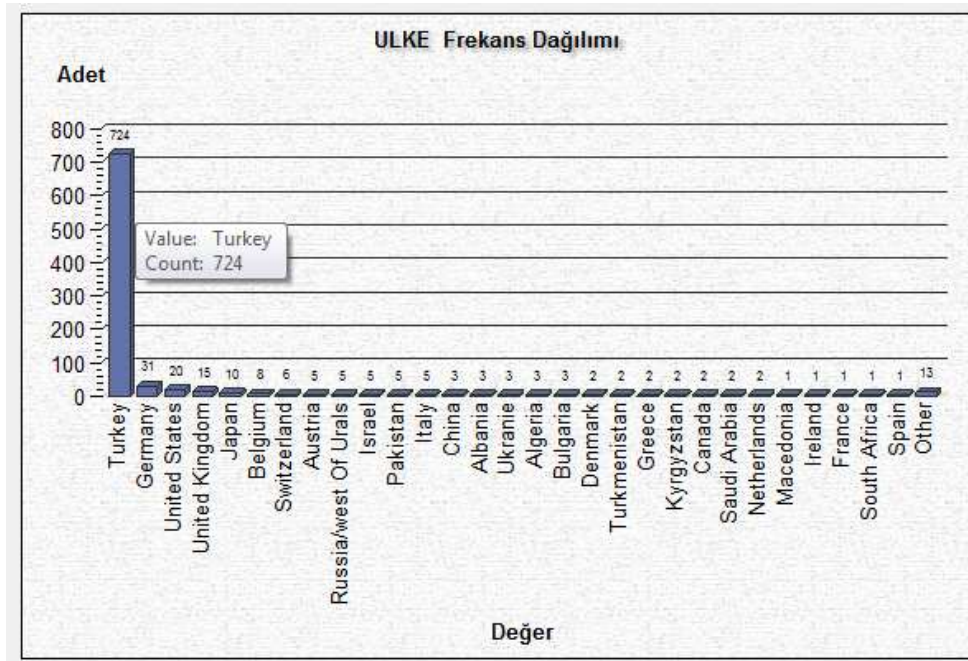
Örnek olarak “Ülke” değişkeni ele alınacak olursa;

- Boş kayıt sayısı : 2116
- Boş kayıt yüzdesi : %70
- Minimum değer : Albania
- Alan karakter olduğu için alfabetik olarak minimum değer seçilmiştir.
- Maksimum değer : Uzbekistan
- Veri deseni kayıt sayısı : 16 farklı veri deseni.
- Her karakter değer ‘A’ ile, her numerik değer ‘9’ ile ifade edilecek olursa 73

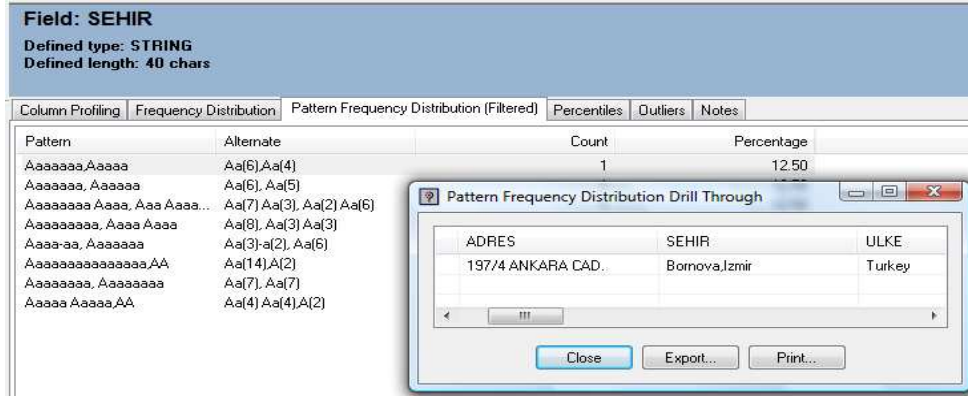
farklı desende veri mevcuttur. Örneğin; “HYUNDAI GETZ” veri deseni “AAAAAAA AAAA” yani “A(7) A(4)” şeklinde iken “BMW 520 I” veri deseni “AAA 999 A” yani “A(3) 9(3) A” şeklindedir. Veri deseni analizi, veri içinde anlamsız desenleri saptamamızda kolaylık sağlar. Plaka alanı ele alınacak olursa, sayısal değer içermeyen ve sadece karakterlerden ibaret olan bir veri deseni geçersiz bir kaydı işaret etmektedir.

- Tekil kayıt sayısı : 43 farklı değer içermektedir.
- Minimum karakter sayısı : 4
- Maksimum karakter sayısı : 20

Şekil 4.4 ‘Ülke’ değişkeni için frekans dağılımını göstermektedir. Müşterilerin 724 adedi (%24.12 oranında) Türkiye’de ikamet ederken, 31 adedi (%1.03) Almanya’da ikamet etmektedir. Burada dikkat edilmesi gereken nokta %70 müşteri için ülke alanının girilmemiş olduğudur. Tez kapsamındaki adres standartlaştırma ve ayrıştırma çalışmaları, sadece Türkiye adresleri için yapılacaktır. Ülke alanının boş olması bu adreslerin Türkiye adresi olmadığını göstermemektedir fakat dolu olup da Türkiye dışı bir ülke girilmiş olan adresler için standartlaştırma ve ayrıştırma çalışması yapılmayacaktır.



Şekil 4.4 : Ülke Alanı Frekans Dağılımı



Şekil 4.6 : Şehir Alanı Veri Deseninden Yola Çıkılarak Veri Analizi

Profilleme çalışması yapılan toplam 14 değişkene ait özet profilleme raporu Şekil 4.7’deki gibidir. Temizleme, standartlaştırma ve ayrıştırma yapılacak veriler seçilirken, Türkiye dışı adresler ve “ADRES, SEHIR, ULKE” alanlarının hepsi boş olan adresler çıkartılacaktır. 3002 adres verisinden analize konu olan veri adedi 2723 (%91)’tür.

Profilleme çalışmalarından elde edilen bir diğer sonuç ise, şehir alanının %70 kayıta boş olduğu ve çoğu kayıta şehir bilgisinin ADRES alanının içinde yazıldığıdır. Örneğin; “BÜYÜKDERE CANFES SK NO :86/2D :2 SARIYER İSTANBUL” ya da “FEVZİÇAKMAK MAH. DAĞYOLU C 2/1 SOK NO 1/42 ÇİĞLİ İZMİR” gibi. Adresi bileşenlerine ayırma çalışması kapsamında ADRES, SEHIR ve ULKE alanı tüm veriler için birleştirilecek ve ortak bir ayrıştırma kuralları kütüphanesi oluşturularak adresler bileşenlerine ayrıştırılacaktır.

DEĞİSKEN ADI	Kayıt Sayısı	Boş Kayıt Sayısı	Boş Kayıt Yüzdesi	Minimum Değer	Maksimum Değer	Mod	Veri Deseni Kayıt Sayısı	Tekil Kayıt Sayısı	Tekillik Oranı	Minimum Uzunluk	Maksimum Uzunluk	Ortalama	Medyan	Standart Sapma
AD	3002	2261	75%				49	433	14,42	2	20			
ADRES	3002	118	4%				2830	2855	95,1	1	120			
ARACIN CINSİ	3002	1172	39%				10	14	0,47	5	18			
FİRMA UNVAN	3002	1913	64%				152	153	5,1	6	80			
GOREVİ	3002	1173	39%				12	22	0,73	6	16			
MARKA VE MODEL	3002	1172	39%				73	173	5,76	3	22			
MODEL YILI	3002	1172	39%	0	2006	2005		23	1,26			2002,13	2004	46,92
ÖZEL KİŞİ	3002	1172	39%	0	1	0		2	0,11			0,40	0	0,49
PLAKA	3002	1172	39%				8	1831	60,99	5	12			
ŞEHİR	3002	2117	71%				57	223	7,43	1	26			
SOYAD	3002	2261	75%				23	559	18,62	2	20			
TARİH	3002	1173	39%	30.06.2005	10.08.2005	27.07.2005		18	0,98					
TÜZEL KİŞİ	3002	1172	39%	0	1	1		2	0,11			0,59	1	0,49
ULKE	3002	2116	70%				16	43	1,43	4	20			

Şekil 4.7 : Profillemeye Çalışması Yapılan Değişkenlere ait Özet Profillemeye Raporu

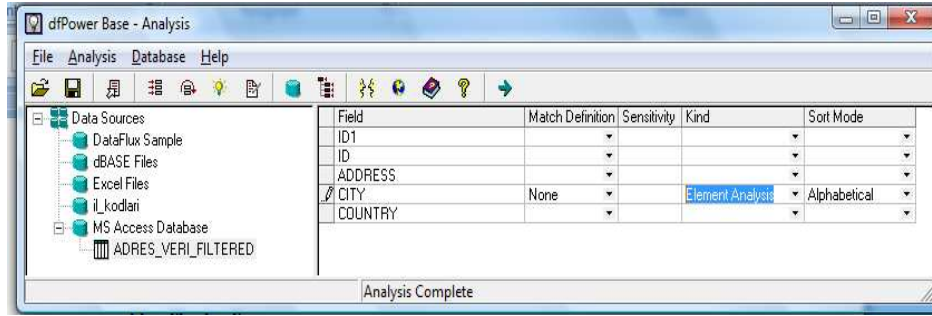
4.2. Adres Verilerinin Standartlaştırılması

Türkiye dışı adresler ve “ADRES, SEHIR, ULKE” alanlarının hepsi boş olan adresler çıkartıldığında 3002 adres verisinden analize konu olan 2723 (%91) adet veri elde edilmektedir.

2723 adet adres verisi bileşenlerine ayrıştırılmadan önce “ADRES” ve “SEHIR” alanları standartlaştırılacaktır. Bu amaçla her iki alan için ayrı ayrı standartlaştırma tanımı yapılacaktır.

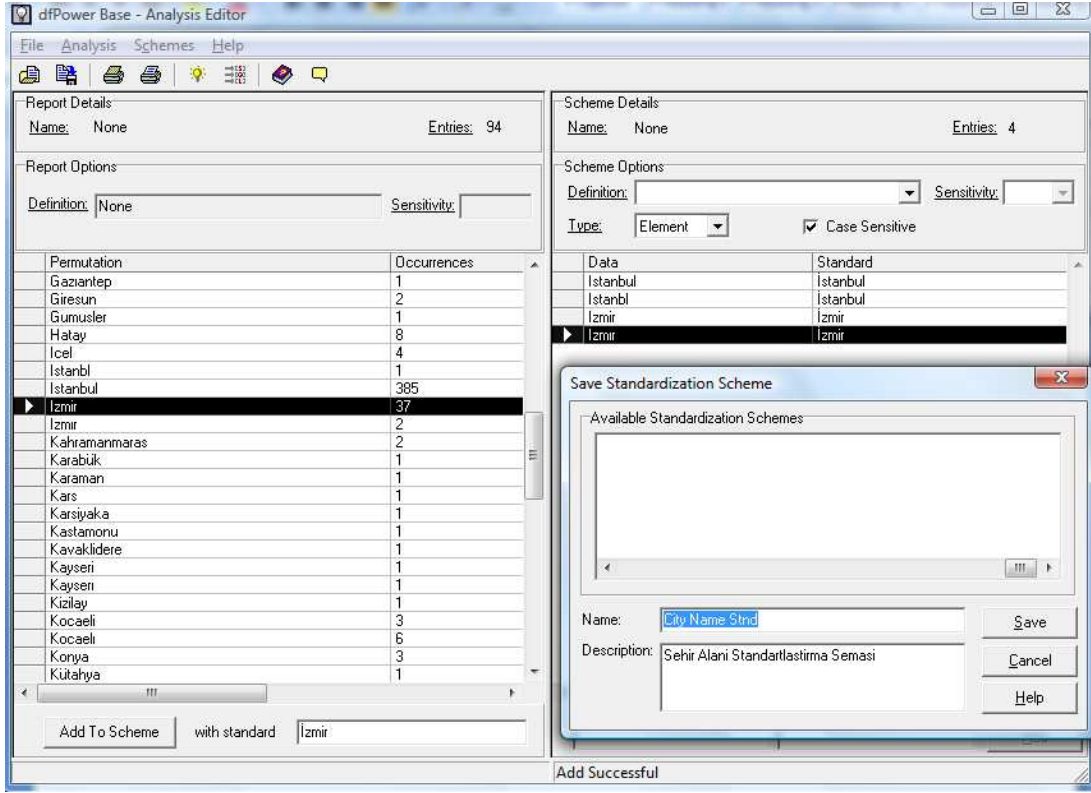
4.2.1. “SEHIR” Alanı için Standartlaştırma Şeması ve Standartlaştırma Tanımı Yaratılması

Şehir bilgisinin standartlaştırılması amacıyla araştırmaya konu olan 2723 veri Microsoft Access yolu ile seçilmiş, DataFlux Analiz modülü ile Element Analizi yapılmıştır.



Şekil 4.8 : ‘ADRES_VERI_FILTERED’ Tablosunda Element Analizi

Element analizi sonucunda üretilen raporda, ‘SEHIR’ alanının içindeki verilerin permütasyonlarını adetsel olarak görmek ve oluşan rapor kullanılarak standartlaştırma şeması oluşturmak mümkündür. Varolan 94 farklı değer için standartlaştırma şeması hazırlanmış ve ‘City Name Stnd’ ismi ile şema kaydedilmiştir. Şekil 4.9’da sol taraftaki pencere, oluşturulan Element Analizi Raporunu; sağ taraf ise Standartlaştırma Şemasını göstermektedir. Görüldüğü gibi 94 farklı değer için verilerin standart halleri girilmiştir. Örneğin; ‘İstanbul’ ‘İstanbul’ olarak, ‘İzmir’ ‘İzmir’ olarak standartlaştırılmıştır.



Şekil 4.9 : SEHIR Alanı için Standartlaştırma Şeması Hazırlanması

Oluşturulan standartlaştırma şemasından örnek olarak ‘ŞANLIURFA’ standardı için girilen farklı kombinasyonlar ve standart hali Şekil 4.10’da gösterilmiştir.

Data	Standard
Ş .URFA	ŞANLIURFA
Ş ANLIURFA	ŞANLIURFA
Ş URFA	ŞANLIURFA
Ş .URFA	ŞANLIURFA
ŞA NLIURFA	ŞANLIURFA
ŞANL1 URFA	ŞANLIURFA
ŞANL1URFA	ŞANLIURFA
ŞANLI URFA	ŞANLIURFA
ŞURFA	ŞANLIURFA

Şekil 4.10 : SEHIR Alanı için Standartlaştırma Şeması

SEHIR alanı için yapılması planlanan standartlaştırma ve temizleme kuralı şu şekildedir:

- Şehir bilgisi büyük harf olarak standartlaştırılacaktır.
- Üretilen standartlaştırma şeması kullanılacaktır.

Belirlenen standartlaştırma koşullarını sağlamak amacıyla ‘SEHIR ALANI STANDARTLASTIRMA’ tanımı yapılmıştır. Bu tanım içinde, ‘City Name Stnd’ şemasının kullanılacağı ve tüm değerlerin büyük harfe çevrileceği belirtilmiştir. Şekil 4.11 yaratılan standartlaştırma tanımını göstermektedir.

Standardization Definition: SEHIR ALANI STANDARTLASTIRMA Data Type: Subdistrict - District - City - Country

Parse Definition:
(None)

☐ Allow parsed input

Tokens

Token Name
☒ Subdistrict - District - City ...

☐ Use token separator: []

Regexlibs Schemes Casing

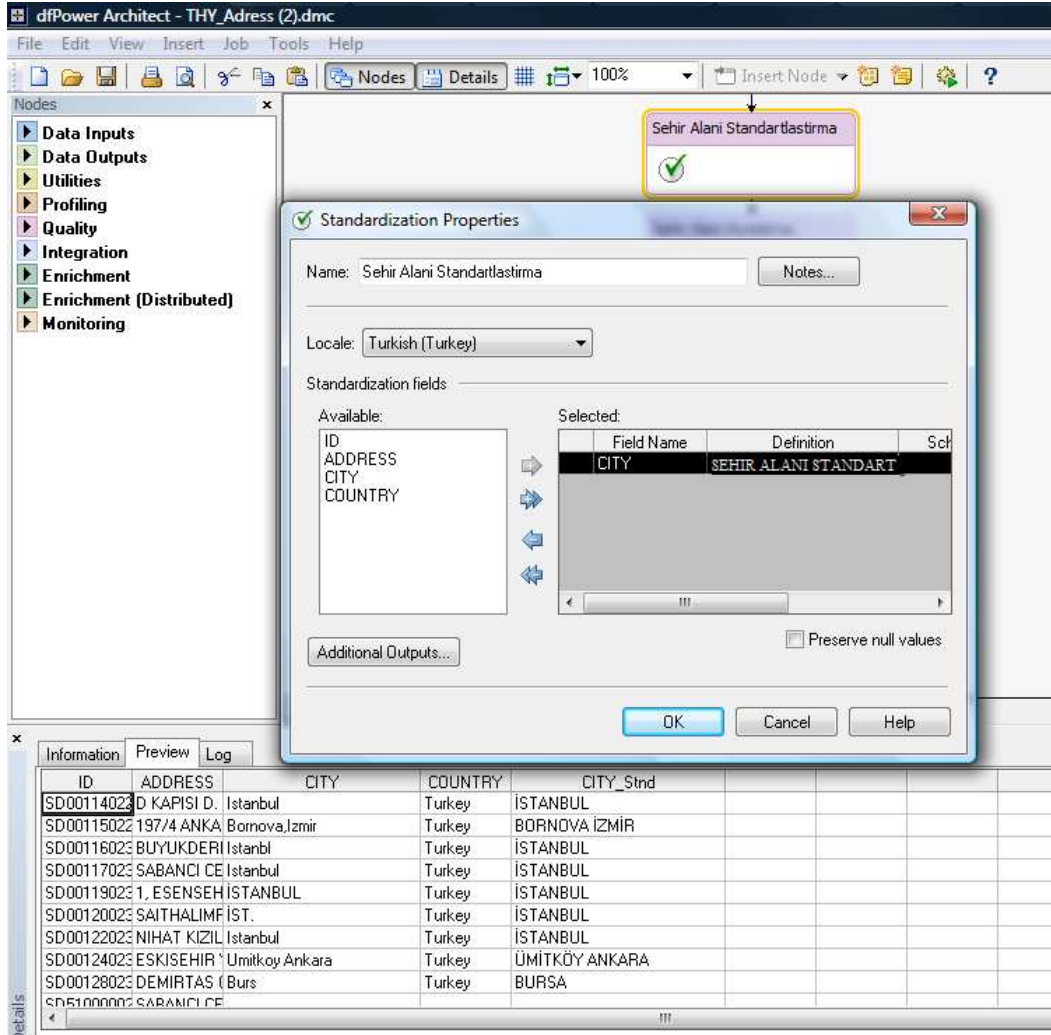
Scheme Type	Case Sensitive
City Name Stnd	Element

Add... Delete

Test... Batch Process...

Şekil 4.11 : SEHIR Alanı için Standartlaştırma Tanımı

Şekil 4.12 oluşturulan standartlaştırma tanımı kullanılarak ‘SEHIR’ alanı standartlaştırılmasını göstermektedir. ‘CITY’ orijinal veriyi, ‘CITY_Stnd’ ise standartlaştırılmış halini göstermektedir.



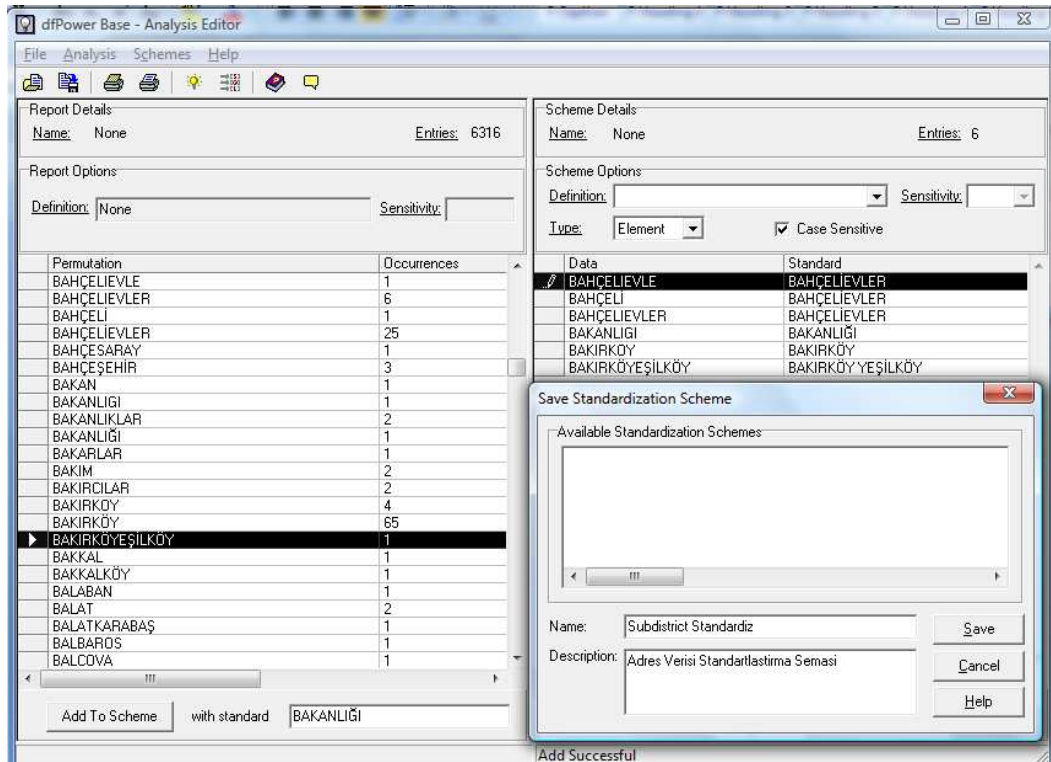
Şekil 4.12 : SEHIR Alanı Standartlaştırma

4.2.2. “ADRES” Alanı için Standartlaştırma Şeması ve Standartlaştırma Tanımı Yaratılması

ADRES alanı için standartlaştırma şeması yaratılması için ‘ADRES_VERI_FILTERED’ tablosunda Element Analizi yapılmış ve Element analizi sonucunda üretilen raporda varolan 6316 farklı değer kullanılarak standartlaştırma şeması hazırlanmış ve ‘Subdistrict Standardiz’ ismi ile şema kaydedilmiştir. Standartlaştırma

şeması hazırlanırken adres içerisinde geçen “256/B ya da 45/5” gibi sokak ya da cadde ismi belirtebilecek sayısal değerlere standartlaştırma uygulanmamıştır. Ayrıca adres içerisinde geçen cadde, sokak, bina, mahalle (Cad., Caddesi, Sk., Sok. Sokak, Apt., Mah., Mahallesi) gibi belirteçler bu aşamada standartlaştırılmamıştır. Genel çerçeve itibari ile adres verisi bütün halinde standartlaştırılıp bileşenlerine ayrıldıktan sonra bu belirteçlere ayrı bir standartlaştırma uygulanmasına karar verilmiştir. Buradaki sebep ise; bazı adreslerde belirteç olabilecek kelimelerin bina adı, site adı ya da sokak, cadde adı olarak da kullanıldığının saptanmasıdır. Örneğin; ‘GÜL SOKAK APT.’ ifadesi geçen bir adreste ‘SOKAK’ kelimesini ‘SOK.’ olarak standartlaştırmak hatalı olacaktır zira buradaki “GÜL SOKAK” apartmanın ismini belirtmektedir. Aynı şekilde “GÜMÜŞ CADDE SİTESİ” ifadesinde “GÜMÜŞ CADDE” site ismini belirtmektedir.

Şekil 4.13’te sol taraftaki pencere, oluşturulan Analiz Raporunu; sağ taraf ise Standartlaştırma Şemasını göstermektedir.



Şekil 4.13 : ADRES Alanı için Standartlaştırma Şeması Hazırlanması

Oluşturulan standartlaştırma şemasından örnek olarak ‘ŞİŞLİ’ semti için girilmiş olan farklı kombinasyonlar ve standart hali Şekil 4.14’te gösterilmiştir.

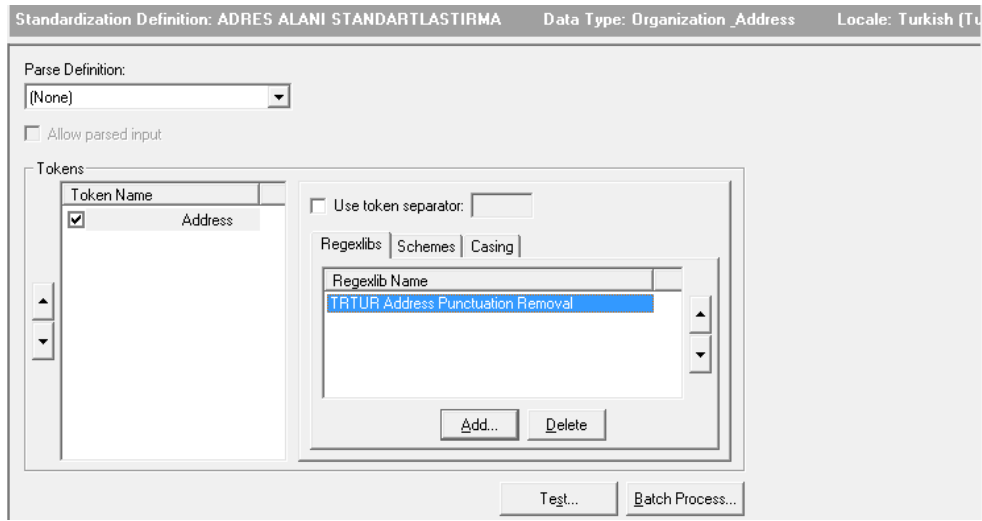
Data	Standard	
-ŞİŞL	ŞİŞLİ	
-ŞİŞLİ	ŞİŞLİ	
.ŞİŞLİ	ŞİŞLİ	
/ŞİŞ	ŞİŞLİ	
/ŞİŞL	ŞİŞLİ	
/ŞİŞLİ	ŞİŞLİ	
CŞİŞLİ	ŞİŞLİ	
NŞİŞLİ	ŞİŞLİ	
SİŞLİ	ŞİŞLİ	
SİŞLİ	ŞİŞLİ	
SİŞLİ	ŞİŞLİ	
SİŞLİ	ŞİŞLİ	
SİŞLİ	ŞİŞLİ	
SİŞLİ	ŞİŞLİ	
TŞİŞLİ	ŞİŞLİ	
ŞİŞLİ	ŞİŞLİ	
ŞİŞLİ	ŞİŞLİ	
ŞİŞLİ	ŞİŞLİ	
ŞYŞLY	ŞİŞLİ	
ŞİŞLİ	ŞİŞLİ	
ŞİŞLİ	ŞİŞLİ	
ŞİİLİ	ŞİŞLİ	
ŞİİŞLİ	ŞİŞLİ	
ŞİŞKLi	ŞİŞLİ	
ŞİŞKİ	ŞİŞLİ	
ŞİŞL	ŞİŞLİ	
ŞİŞL.	ŞİŞLİ	
ŞİŞLİ	ŞİŞLİ	

Şekil 4.14 : ADRES Alanı için Standartlaştırma Şeması

ADRES alanı için yapılması planlanan standartlaştırma ve temizleme kuralı şu şekildedir:

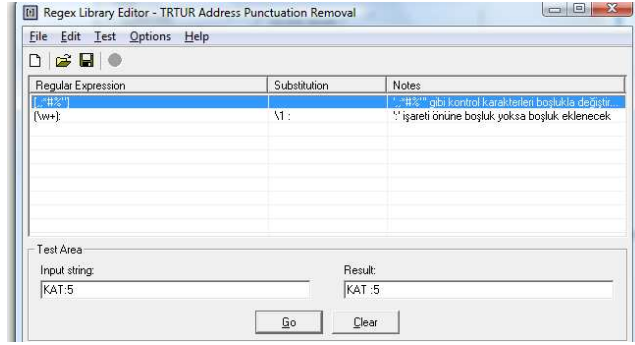
- Adres bilgisi büyük harf olarak standartlaştırılacaktır.
- Üretilen standartlaştırma şeması kullanılacaktır.
- ‘;,*#%”’ gibi kontrol karakterleri boşlukla değiştirilecektir. (Adres verisinde anlamsız girilen karakterlere rastlanmıştır ve bunlar temizlenecektir)
- ‘:’ işareti önüne boşluk yoksa boşluk eklenecektir. (Kat:5 şeklinde yazılmış bir veri KAT :5 haline dönüştürülecektir.)

Belirlenen standartlaştırma koşullarını sağlamak amacıyla ‘ADRES ALANI STANDARTLASTIRMA’ tanımı yapılmıştır. Bu tanım içinde, ‘Subdistrict Standardiz’ şemasının kullanılacağı, tüm değerlerin büyük harfe çevrileceği ve kontrol karakterlerinin yukarıdaki koşullarda belirtildiği gibi değiştirileceği hazırlanan bir düzenli ifadeler (regular expression: regex) şeması ile belirtilmiştir. Şekil 4.15 yaratılan standartlaştırma tanımını göstermektedir.



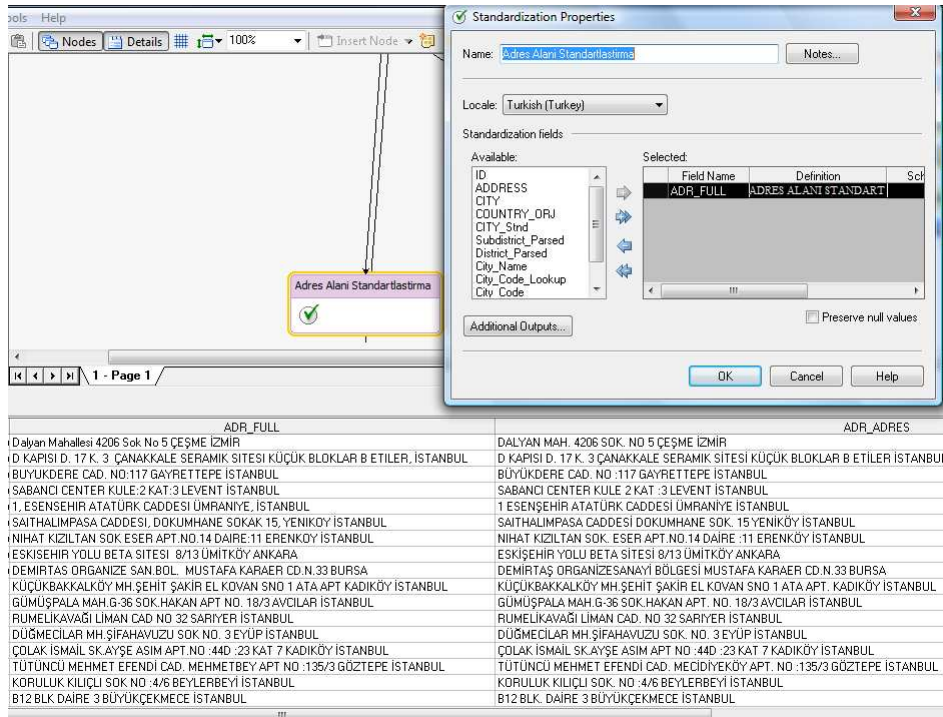
Şekil 4.15 : ADRES Alanı için Standartlaştırma Tanımı

Kontrol karakterleri ile ilgili belirtilen kuralların oluşturulması için yaratılan regex şeması Şekil 4.16'daki gibidir. Düzenli ifadeler (regex); veri standartlaştırma, temizleme, ayrıştırma ve eşleştirme aşamalarında karakter seviyesinde işlem ve dönüştürme yapmayı sağlar. Kelime ya da tümce bazında standartlaştırma yapmak için standartlaştırma şemaları kullanılmaktadır. Burada yazılan ifadeler PERL programlama dilinde yazılmıştır.



Şekil 4.16 : ADRES Alanı için Standartlaştırma İçin Kullanılan Regex

Şekil 4.17 oluşturulan standartlaştırma tanımı kullanılarak ‘ADRES’ alanı standartlaştırılmasını göstermektedir. ‘ADR_FULL’ standartlaştırılmamış ADRES verisi ile daha önceden standartlaştırılmış SEHIR verisinin birleştirilmiş halidir. ADR_ADRES ise birleştirilmiş adres ve şehir bilgisinin (ADR_FULL) standartlaştırılmış halini göstermektedir.



Şekil 4.17 : ADRES Alanı Standartlaştırma

4.3. Adres Verilerinin Bileşenlerine Ayrıştırılması

Adres verilerinin bileşenlerine ayrıştırılma çalışmasında ilk önce, standartlaştırılmış ADRES ve SEHIR alanları birleştirilecektir. Profilleme çalışmasından elde edilen sonuçlara göre %70 oranında SEHIR alanının boş olduğu ve boş olan adreslerde şehir bilgisinin ADRES alanının içerisinde yer alması ve her adres için ortak ayrıştırma kuralları uygulanmak istenmesi bu kararın verilmesine sebep olmuştur.

Profilleme çalışmaları sonucunda SEHIR alanının içerisinde semt/ilçe/il/ülke bilgisinin de girildiği gözlemlenmiştir. ‘Bornova,Izmir’, ‘Umitkoy Ankara’ gibi semt/ilçe ve il bilgisi ya da sadece ‘Çeşme’ gibi il bilgisi varolmayan ama sadece ilçe bilgisinin yer aldığı kayıtlara rastlanmıştır. Bu adresler için SEHIR alanı ilk önce semt, ilçe ve il bileşenlerine ayrıştırılacaktır. Daha sonra ise şehir bilgisi olmayan fakat ilçe bilgisini içeren kayıtlar için il bilgisi bulunacaktır.

4.3.1. “SEHIR” Alanının “Semt, İlçe, İl” Bileşenlerine Ayrıştırılması

Veri ayrıştırma çalışmalarında “Anlamlandırılmış Bilgi Tabanı” kullanılır. “Anlamlandırılmış Bilgi Tabanı”, ilgili veriye baz oluşturacak dev bir veri sözlüğü ve verilerin ilişkilerini belirleyecek anlamlandırılmış gramer kuralları deposudur. İlk önce ayrıştırma çalışmasına konu olan alanın analiz edilmesi ve veri içerisindeki mantıksal sıralamaların saptanması gerekmektedir. Örneğin; SEHIR alanı İl, İlçe ve Semt bileşenlerine ayrıştırılacaksa; bu bilgilerin ilgili alan içerisinde hangi sırayla tutulmuş olabileceği hem mantıksal kurallar ile hem de veri analizi ile belirlenmelidir. SEHIR alanının içerisinde sadece ‘İZMİR’ gibi şehir bilgisi olabileceği gibi ‘KARŞIYAKA İZMİR’ gibi ilçe-il sıralamasında bilgi de varolabilir ya da tam tersi bir kombinasyonla veri ‘İZMİR KARŞIYAKA’ şeklinde il-ilçe sıralamasında girilmiş olabilir. İlçe bilgisine yer

verilmeden ‘ÜMİTKÖY, ANKARA’ gibi semt-il bilgisi yer alabilir. Gramer kuralları kütüphanesi oluşturulurken bu kombinasyonların her birisinin belirlenmesi gerekmektedir.

SEHIR alanı ayrıştırma çalışmasına konu olacak ‘TRTUR Subdistrict District City’ gramer kuralları kütüphanesi oluşturulmuştur. Veri içerisinde belirlenen kombinasyonlar bu kütüphanede kurallar olarak belirtilmiştir. Olası veri kombinasyonları aşağıdaki gibi oluşturulmuştur:

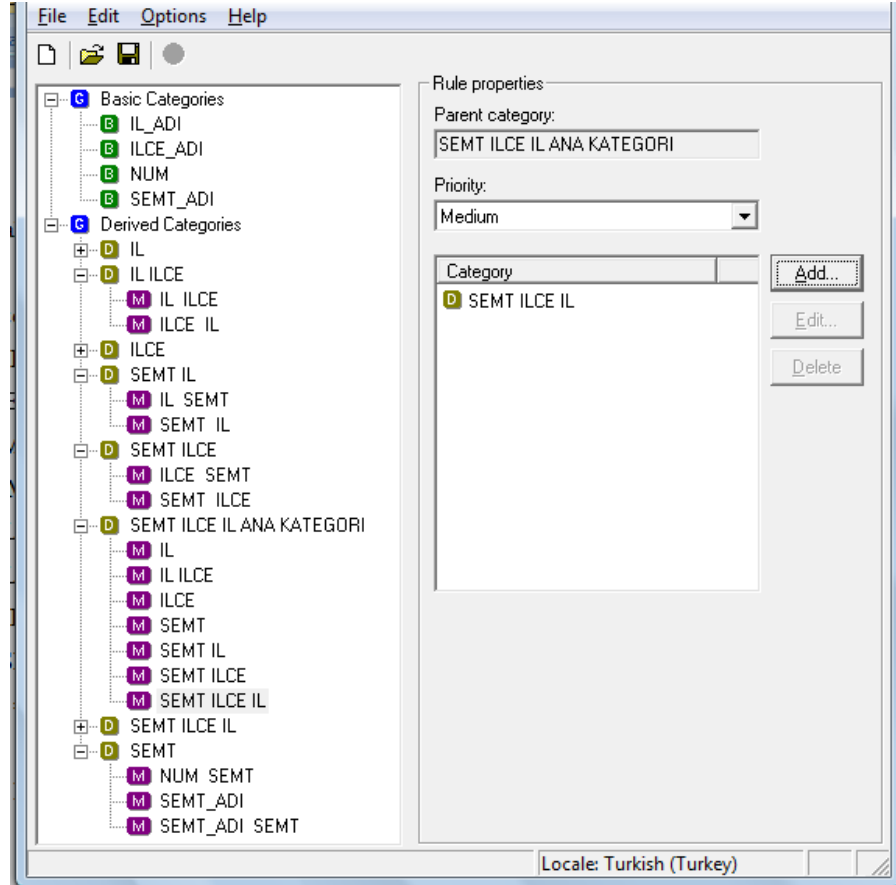
Temel Kategoriler :

- SEMT_ADI
- ILCE_ADI
- IL_ADI
- NUM (Sayısal değerler için. Örneğin; 4. Levent semti gibi)

Türetilen Kategoriler için veri sırası ve her birine örnek aşağıdaki gösterilmiştir:

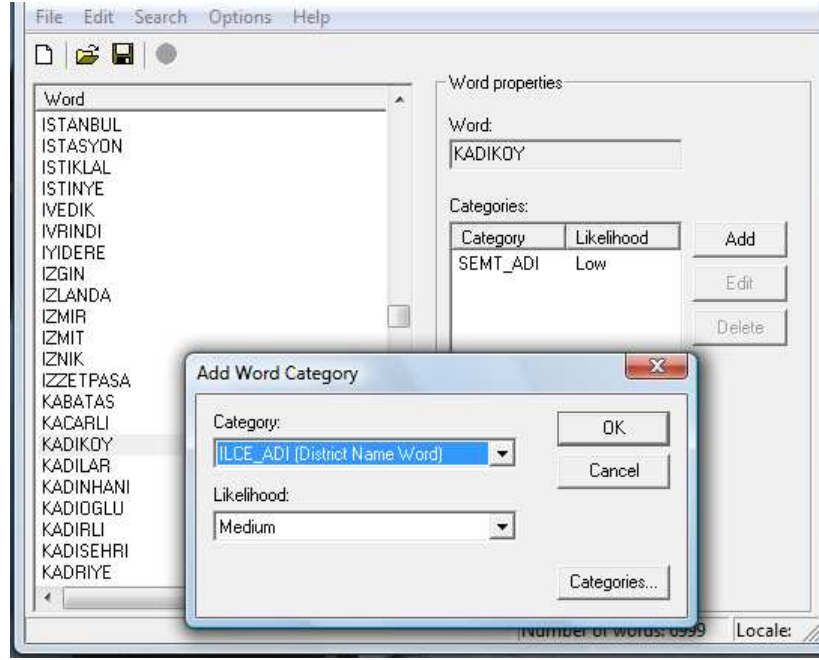
- IL = IL_ADI (İstanbul)
- ILCE = ILCE_ADI (Şişli)
- SEMT = SEMT_ADI (Nişantaşı)
NUM & SEMT_ADI (4. Levent)
- IL ILCE = IL & ILCE (İstanbul Şişli)
ILCE & IL (Şişli İstanbul)
- SEMT IL = IL & SEMT (İstanbul Nişantaşı)
SEMT & IL (Nişantaşı İstanbul)
- SEMT ILCE = ILCE & SEMT (Şişli Nişantaşı)
SEMT & ILCE (Nişantaşı Şişli)
- SEMT ILCE IL = SEMT & ILCE & IL (Nişantaşı Şişli İstanbul)
IL & SEMT & ILCE (İstanbul Nişantaşı Şişli)
- IL & ILCE & SEMT (İstanbul Şişli Nişantaşı)

Şekil 4.18 oluşturulan ‘TRTUR Subdistrict District City’ gramer kuralları kütüphanesini göstermektedir.



Şekil 4.18 : ‘TRTUR Subdistrict District City’ Gramer Kuralları Kütüphanesi

Gramer kuralları belirlendikten sonra, oluşturulan temel kategorilere (SEMT_ADI, ILCE_ADI, IL_ADI) giren kelimelerin belirleneceği bir veri sözlüğü tanımlanmalıdır. Analiz edilecek verideki bileşenleri oluşturacak her bir kelimenin hangi gruba girdiği veri sözlüğünde belirtilir. Veri sözlüğünün oluşturulmasında PTT web sayfasından indirilen ‘Türkiye Adres ve Posta Kodları Özet’ tablosu kullanılmıştır. Şekil 4.19 ‘Türkiye Adres ve Posta Kodları Özet’ tablosunun resmidir. Bu tablo mahalle ve köy seviyesine kadar Türkiye adres bileşenlerini içermektedir. Tablodan elde edilen il, ilçe ve semt bilgileri Türkçe

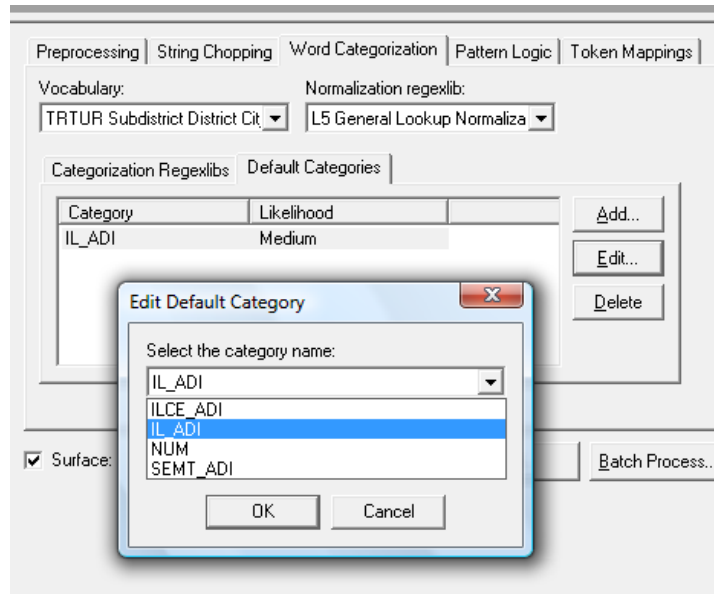


Şekil 4.20 : ‘TRTUR Subdistrict District City’ Veri Sözlüğü

SEHIR alanının semt, ilçe ve il bleşenlerine ayrılması için oluşturulan gramer kurallarının ve veri sözlüğünün kullanıldığı ‘Subdistrict – District – City’ ayrıştırma tanımı yaratılmıştır. Şekil 4.21 oluşturulan ayrıştırma tanımını göstermektedir. Ayrıştırma tanımında belirtilen kurallar aşağıdaki gibidir.

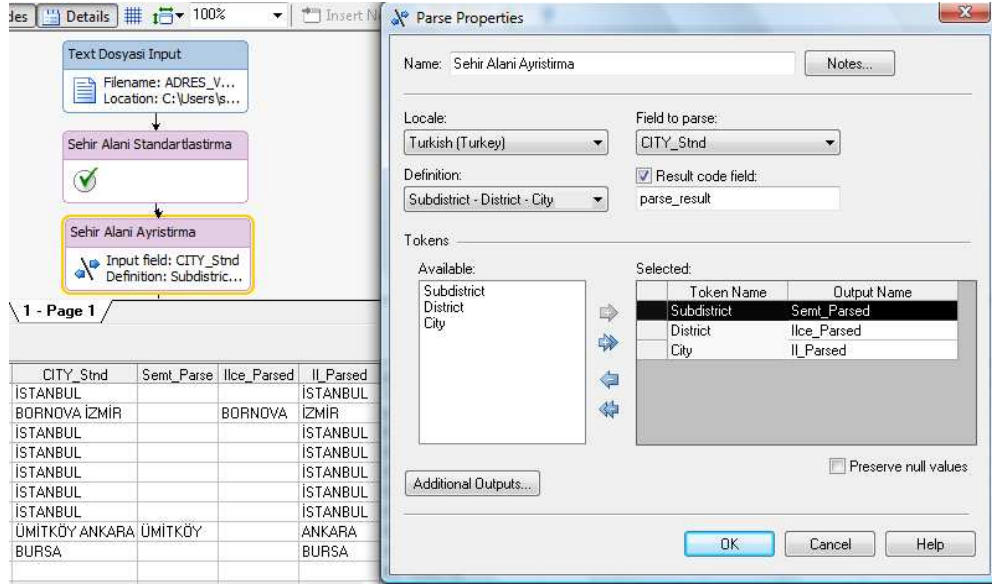
- Veri içerisindeki her kelimeyi ayırt edebilmek için; boşluk veya ‘_’ gibi ifadeler kelimeler arası ayraç olarak kabul edilecektir. Bunun için ‘TRTUR Subdistrict District City’ kelime ayırma kuralları tablosu (string chopping tablosu) kullanılmıştır.
- Veriler büyük harfe çevrilerek ve Türkçe karakterlerinden arındırılarak normalize edilecek ve veri sözlüğünde bu kelimelere karşılık gelen kategoriler aranacaktır. Bunun için, ‘L5 General Lookup Normalization’ düzenli ifadeler (regex) kuralları oluşturulmuştur.
- ‘TRTUR Subdistrict District City’ veri sözlüğü kullanılacaktır.

- ‘TRTUR Subdistrict District City’ gramer kuralları kütüphanesi kullanılacaktır.
- ‘Subdistrict’ belirteci Semt adını, ‘District’ belirteci İlçe adını, ‘City’ belirteci İl adını simgeleyecektir.



Şekil 4.21 : ‘Subdistrict District City’ Ayırıştırma Tanımı

Tanımlanan ayırıştırma tanımı kullanılarak; SEHIR alanı semt, ilçe ve il bileşenlerine ayrıştırılmıştır. Şekil 4.22, standartlaştırılmış CITY_Stnd alanının Semt_Parsed, Ilce_Parsed ve Il_Parsed alanlarına ayrıştırılmasını göstermektedir. Toplam 2723 adres verisi içerisinde 889 adet SEHIR alanının dolu olduğu kayıt vardır ve 16 adet kayıt için SEHIR alanı bileşenlerine ayrıştırıldığında ilçe ya da semt alanı dolu olan kayda rastlanmıştır.



Şekil 4.22 : Standartlaştırılmış SEHIR Alanının Bileşenlerine Ayırıştırılması

4.3.2. Şehir Bilgisi Olmayan Fakat İlçe Bilgisi İçeren Kayıtlar İçin Şehir Bilgisi Bulunarak Adresin Zenginleştirilmesi

PTT'den elde edilen il bilgisi ve ilçe bilgisi kullanılarak referans IL_ILCE_KODLARI tablosu hazırlanmıştır. Şekil 4.23 IL_ILCE_KODLARI tablosunu göstermektedir.

Daha önceki bölümde standartlaştırılan SEHIR bilgisinden ayrıştırılan 'Il_Parsed' alanı boş olup da 'Ilce_Parsed' alanı dolu olan kayıtlar için IL_ILCE_KODLARI tablosunda varolan ilçe bilgisine karşılık gelen il bilgisi bulunarak adres zenginleştirilmiştir. Şekil 4.24, Il_Parsed bilgisi boş olan ve Ilce_Parsed alanında ise 'Çeşme' olan bir adres verisi için il bilgisinin arama tablosu (lookup tablosu) aracılığıyla 'İZMİR' olarak DataFlux aracılığıyla oluşturulan bir iş (job) ile nasıl doldurulduğu gösterilmektedir.

1	ADANA	0	MERKEZ
1	ADANA	1	CEYHAN
1	ADANA	2	FEKE
1	ADANA	3	KARAIŞALI
1	ADANA	4	KARATAŞ
1	ADANA	5	KOZAN
1	ADANA	6	POZANTI
1	ADANA	7	SAİMBEYLİ
1	ADANA	8	TUFANBEYLİ
1	ADANA	9	YUMURTALIK
1	ADANA	10	ALADAĞ
1	ADANA	11	İMAMOĞLU
1	ADANA	12	SEYHAN
1	ADANA	13	YÜREGİR
2	ADYAMAN	0	MERKEZ
2	ADYAMAN	1	BESNİ
2	ADYAMAN	2	ÇELİKHAN
2	ADYAMAN	3	GERGER
2	ADYAMAN	4	ÇELİKPAZARI

Şekil 4.23 : IL_ILCE_KODLARI Arama Tablosu

The screenshot shows a data management tool interface with a workflow diagram on the left and an 'SQL Lookup Properties' dialog box on the right.

Workflow Diagram:

- Sehir Alani Ayristirma** (Input field: CITY_Stnd, Definition: Subdistric...)
- Sehir Lookup** (DSN: MS Access Dat..., Table: IL_KODLARI)
- ILCE ADI DOLU / IL ADI BOS** (Branch 2: Land all data first)
- Ilce Lookup** (DSN: MS Access Dat..., Table: IL_ILCE_KOD...)
- Veri Birlestirme**

SQL Lookup Properties Dialog Box:

- Name:** Ilce Lookup
- Lookup table:** IL_ILCE_KODLARI
- Field mappings:**
 - Input table:** ID, ADDRESS, CITY, COUNTRY_ORJ, CITY_Stnd
 - Lookup table:** IL_KODU, IL_ADI, ILCE_KODU, ILCE_ADI
- Expressions:** District_Parsed = ILCE_ADI
- Single matches only:** ☒
- Output fields:**
 - Available:** IL_KODU, IL_ADI, ILCE_KODU, ILCE_ADI
 - Selected:**

Field Name	Output Name
IL_ADI	City_Name
IL_KODU	City_Code_Lookup

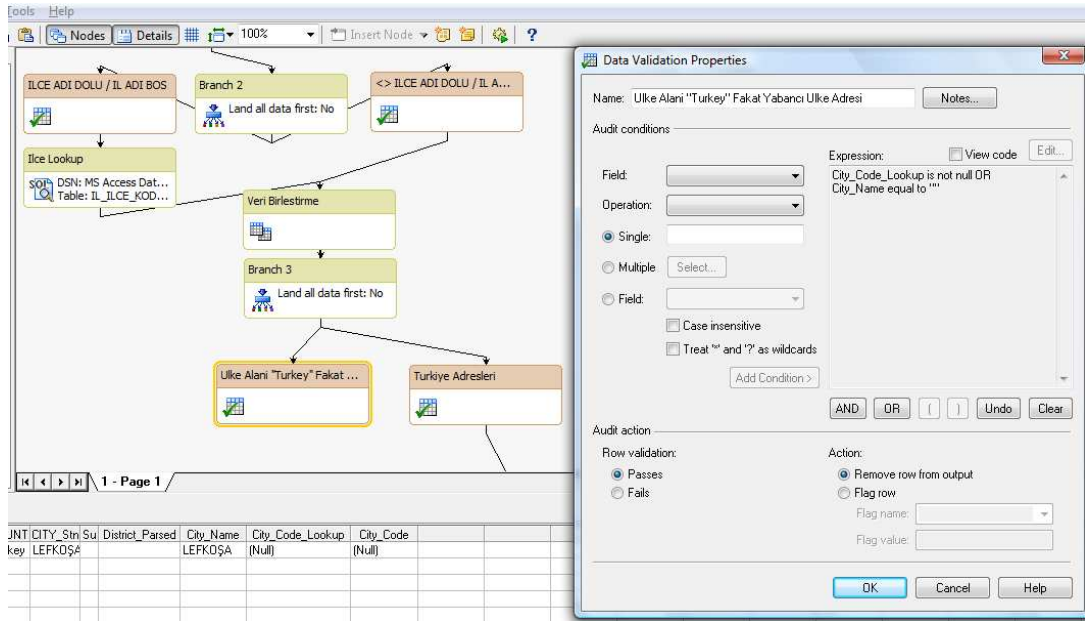
Data Table:

JNT	CITY_Stnd	Su	District_Parsed	City_Code	City_Name
key	ÇEŞME		ÇEŞME	(Null)	İZMİR

Şekil 4.24 : Arama Tablosu Aracılığıyla Şehir Bilgisinin Zenginleştirilmesi

4.3.3. Arama Tablosu Aracılığıyla Ülke Alanı ‘Türkiye’ Girilmiş Olmasına Rağmen Yurtdışı Olan Adreslerin Tespit Edilmesi

Şehir bilgisi zenginleştirilen kayıtlar ve diğer kayıtlar birleştirildikten sonra ülke alanında ‘Turkey’ geçmesine ve il alanı dolu olmasına rağmen IL_ILCE_KODLARI arama tablosundan bu il bilgisine denk gelen il kodu bulunamayan kayıtlar incelenmiş ve bu kayıtların yurtdışı adres kayıtları olduğu saptanmıştır. Şekil 4.25’te ülke alanı ‘Turkey’ olan, İl alanı ‘Lefkoşa’ olan ve dolayısıyla IL_ILCE_KODLARI tablosunda geçerli bir il kodu bulamayan 1 kayıt bulunmuştur. Bu kayıt adres ayrıştırma çalışmaları ‘Türkiye Adresleri’ kapsamında yapılacağı için kapsam dışı bırakılmıştır.



Şekil 4.25 : Arama Tablosu Aracılığıyla Yurtdışı Adreslerinin Tespit Edilmesi

4.3.4. “ADRES” Alanının Bileşenlerine Ayrıştırılması

‘SEHIR’ alanı standartlaştırılıp, bileşenlerine ayrıldıktan ve zenginleştirildikten ve arama tablosu aracılığıyla Türkiye dışı adresler elendikten sonra, ADRES ve elde edilen CITY_STND alanı birleştirilerek bütün bir adres verisi elde edilmiştir. Şekil 4.26’da birleştirilen ‘ADDRESS’, ‘CITY_STND’ ve ‘CITY_STND’ alanından zenginleştirilen ‘CITY_NAME’ alanının birleştirilmiş hali olan ADR_FULL alanı görülmektedir.

The screenshot shows the iText - THY Adres (2).dmc software interface. The main window displays a workflow diagram with nodes for 'Branch 3', 'Land all data first: No', 'Ulke Alanı "Türkiye" Fakat ...', 'Türkiye Adresleri', and 'Adres, Sehir Alanının Birles...'. An 'Expression Properties' dialog box is open, showing the 'Expression' tab with the following code:

```
string ADR_FULL  
if (len(ADDRESS) == 1) then ADDRESS = null  
if (isnull(City_Code)) then  
  ADR_FULL = 'ADDRESS' & " " & 'CITY_Stnd' & " " & 'City_Name'  
else  
  ADR_FULL = 'ADDRESS' & " " & 'CITY_Stnd'
```

The dialog box also has buttons for 'Functions...', 'Fields...', 'Save...', and 'Load...'. Below the dialog box, a table preview is visible with columns: ADDRESS, CITY_Stnd, City_Name, and ADR_FULL. The table contains 10 rows of data, including addresses from Dalman Mahallesi, D KAPISI D. 17 K. 3 ÇANAKKALE SERAMİK SİTESİ, and others.

Şekil 4.26 : ADRES ve SEHIR Alanlarının Birleştirilmesi

ADR_FULL alanına Bölüm 4.2.2’de oluşturulan ‘ADRES ALANLARI STANDARTLASTIRMA’ tanımı uygulanmış ve ADR_ADRES alanı oluşturulmuştur.

Şekil 4.17 oluşturulan standartlaştırma tanımı kullanılarak adres alanı standartlaştırmasını göstermektedir.

SEHIR alanının bileşenlerine ayrılması çalışmasında yapılmış olduğu gibi burada da ADRES verisi için ilk önce ayrıştırma çalışmasına konu olan alan ve adres bileşenlerinin ilgili alanda ne şekilde sıralanmış olduğu analiz edilmiştir. PTT'nin öngörmüş olduğu adres standardı baz alınmış ve bu kurala ek olacak ve veride varolan adres yazılışlarını da içinde barındıracak 'TRTUR OrgAddress' gramer kuralları kütüphanesi oluşturulmuştur. Veri içerisinde belirlenen kombinasyonlar bu kütüphanede kurallar olarak belirtilmiştir. Türkiye standartlarında adres yazımının belli bir standarda henüz daha oturmamış olması ve ele alınan veride çok fazla sayıda farklı adres yazımının olması dolayısıyla olası veri kombinasyonları çok sayıda olmaktadır. Bu çalışma büyük kurumların çok fazla adres verisini ayrıştırma çalışmalarında en önemli zaman alıcı faaliyetlerden biridir. Tez çalışması kapsamında zaman ve kaynak kısıtları nedeniyle sadece temel kategoriler ve türetilen temel kategorilerden bazıları tanımlanmış, tüm veri kombinasyonları ele alınmamıştır. Bu da adresin bileşenlerine ayrıştırılma başarısını belli oranda düşürecektir fakat unutulmaması gerekir ki burada tanımlanan kategoriler detaylı çalışmalara baz oluşturacak nitelikte ve geliştirilebilecek bir alt yapıyla tanımlanmıştır.

Temel Kategoriler :

- ORGANIZASYON_ADI
- ORGANIZASYON_BELIRTECI
- MAHALLE_ADI
- MAHALLE_BELIRTECI
- CADDE_ADI
- CADDE_BELIRTECI
- SOKAK_ADI
- SOKAK_BELIRTECI
- SITE_BINA_GRUBU_ADI

- SITE_BINA_GRUBU_BELIRTECI
- BINA_ADI
- BINA_ADI_BELIRTECI
- BLOK_ADI
- BLOK_ADI_BELIRTECI
- BINA_NUMARASI
- KAT_NUMARASI
- KAT_BELIRTECI
- DAIRE_NUMARASI
- DAIRE_BELIRTECI
- KOY_ADI
- KOY_BELIRTECI
- SEMT_ADI
- SEMT_BELIRTECI
- BELDE_ADI
- BELDE_BELIRTECI
- POSTA_KODU
- ILCE_ADI
- IL_ADI
- ULKE_ADI
- EK_BILGI

Türetilen 2. seviye ana kategoriler için veri sırası Tablo 4.1, 4.2, 4.3 ve 4.4'te gösterilmiştir. Türetilen değişkenler tekrarlanabilir yani kendi içlerinde tekrar kategori olarak kullanılabilir olarak oluşturulmuştur. Mesela ORGANIZASYON kategorisi üretilirken ORGANIZASYON kuralının içerisindeki her bir kuralı tekrar tekrar yazıp başına NUM (sayısal değerleri ifade eder) eklenmesi yerine 'NUM & ORGANIZASYON' kuralı eklenmiştir. Bu sayede 'AKGUL TİC. A.Ş.' ve '5. AKGUL TİC. A.Ş.' ve '4.

AKGUL KARDESLER HOLDING’ organizasyon ismi olarak ayırt edilebilir. NUM, ‘4, 2, 5675 gibi sayısal değerleri, SING ise ‘A-Z’ harfleri ifade etmektedir.

Tablo 4.1 : Adres Alanı Ayırıştırma İçin Türetilen İkinci Seviye Gramer Kuralları I

II. SEVİYE KATEGORİ İSMİ	KATEGORİ KURALI
ORGANİZASYON	ORGANİZASYON_ADI & ORGANİZASYON_BELİRTECI
	ORGANİZASYON_ADI
	NUM & ORGANİZASYON
	SING & ORGANİZASYON
	ORGANİZASYON_ADI & ORGANİZASYON_BELİRTECI & ORGANİZASYON_BELİRTECI
	ORGANİZASYON & SING
	ORGANİZASYON & SING & SING
MAHALLE	MAHALLE_ADI & MAHALLE_BELİRTECI
	MAHALLE_ADI
	MAHALLE_ADI & MAHALLE_ADI & MAHALLE_BELİRTECI
	MAHALLE_ADI & MAHALLE_ADI & MAHALLE_ADI & MAHALLE_BELİRTECI
	MAHALLE & MAHALLE_BELİRTECI
CADDE	CADDE_ADI & CADDE_BELİRTECI
	CADDE_ADI & CADDE_ADI & CADDE_BELİRTECI
	CADDE_ADI & CADDE_ADI & CADDE_ADI & CADDE_BELİRTECI
	CADDE_ADI & SING & CADDE_ADI & CADDE_BELİRTECI
	CADDE & CADDE_BELİRTECI
	NUM & CADDE
	SING & CADDE
SOKAK	SOKAK_ADI & SOKAK_BELİRTECI
	SOKAK_ADI & SOKAK_ADI & SOKAK_BELİRTECI
	SOKAK_ADI & SOKAK_ADI & SOKAK_ADI & SOKAK_BELİRTECI
	SOKAK_ADI & SING & SOKAK_ADI & SOKAK_BELİRTECI
	SOKAK & SOKAK_BELİRTECI
	NUM & SOKAK
	SING & SOKAK

Tablo 4.2 : Adres Alanı Ayrıştırma İçin Türetilen İkinci Seviye Gramer Kuralları II

II. SEVİYE KATEGORİ İSMİ	KATEGORİ KURALI
SITE_BINA_GRUBU	SITE_BINA_GRUBU_ADI & SITE_BINA_GRUBU_BELIRTECI
	SITE_BINA_GRUBU_ADI
	SITE_BINA_GRUBU_ADI & SITE_BINA_GRUBU_ADI & SITE_BINA_GRUBU_BELIRTECI
	SITE_BINA_GRUBU_ADI & NUM & SITE_BINA_GRUBU_ADI & SITE_BINA_GRUBU_BELIRTECI
	SITE_BINA_GRUBU_ADI & SING & SITE_BINA_GRUBU_ADI & SITE_BINA_GRUBU_BELIRTECI
	SITE_BINA_GRUBU & SITE_BINA_GRUBU_BELIRTECI
	SING & SITE_BINA_GRUBU_BELIRTECI
	NUM & SITE_BINA_GRUBU_BELIRTECI
	SING & SITE_BINA_GRUBU
	SING & SING & SITE_BINA_GRUBU
	NUM & SITE_BINA_GRUBU
BINA	BINA_ADI & BINA_ADI_BELIRTECI
	BINA_ADI & BINA_ADI & BINA_BELIRTECI
	BINA_ADI & NUM & BINA_ADI & BINA_BELIRTECI
	BINA_ADI & SING & BINA_ADI & BINA_BELIRTECI
	SING & BINA_BELIRTECI
	BINA & BINA_BELIRTECI
	SING & BINA
	SING & SING & BINA
	NUM & BINA
	NUM & SING & BINA
BLOK	BLOK_ADI & BLOK_ADI_BELIRTECI
	SING & BLOK
	SING & NUM & BLOK
	NUM & BLOK
	NUM & SING & BLOK

Tablo 4.3 : Adres Alanı Ayrıştırma İçin Türetilen İkinci Seviye Gramer Kuralları III

II. SEVİYE KATEGORİ İSMİ	KATEGORİ KURALI
KAT_NUMARASI DAIRE_NUMARASI BINA_NUMARASI	NUM
	SING & NUM
	SING & NUM & SING
	NUM & SING
	NUM & NUM
	NUM & NUM & SING
KAT	KAT_NUMARASI & KAT_ BELİRTECİ
	KAT_ BELİRTECİ & KAT_ NUMARASI
DAIRE	DAIRE_NUMARASI & DAIRE_ BELİRTECİ
	DAIRE_ BELİRTECİ & DAIRE_ NUMARASI
KOY	KOY_ADI & KOY_ BELİRTECİ
	KOY_ADI
	KOY & KOY_ BELİRTECİ
	NUM & KOY
	SING & KOY
	SING & SING & KOY
	NUM & SING & KOY
SEMT	SEMT_ADI & SEMT_ BELİRTECİ
	SEMT_ADI
	SEMT & SEMT_ BELİRTECİ
	NUM & SEMT
	SING & SEMT
	SING & SING & SEMT
	NUM & SING & SEMT
BELDE	BELDE_ADI & BELDE_ BELİRTECİ
	BELDE_ADI
	BELDE & BELDE_ BELİRTECİ
	SING & BELDE

Tablo 4.4 : Adres Alanı Ayrıştırma İçin Türetilen İkinci Seviye Gramer Kuralları IV

II. SEVİYE KATEGORİ İSMİ	KATEGORİ KURALI
POSTA_KODU	NUM
İLCE	İLCE_ADI
	NUM & İLCE_ADI
	SING & İLCE_ADI
	İLCE & İLCE_ADI
İL	İL_ADI
	SING & İL_ADI
	İL & İL_ADI
ULKE	ULKE_ADI
EK_BILGI	EK_BILGI

Adreslerin bileşenlerine ayrıştırılması için belirlenen temel kategoriler ve bu kategorilerden türetilen ikinci seviye kategoriler kullanılarak ‘Ana Kategori’ olarak ADRES oluşturulmuş ve bu kategorinin içerisinde belirlenen kurallar Tablo 4.5 ve 4.6’da gösterilmiştir.

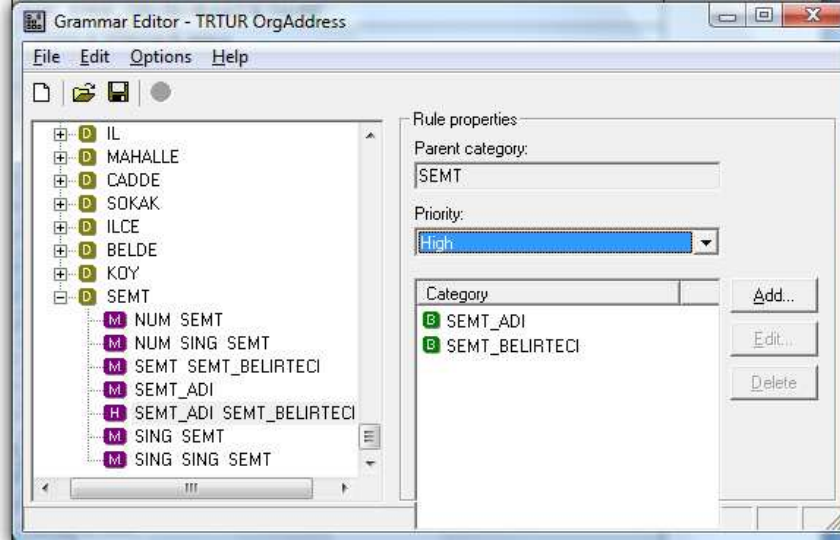
Tablo 4.5 : Adres Alanı Ayrıştırma İçin Türetilen Ana Kategori Gramer Kuralları I

ANA KATEGORİ İSMİ	KATEGORİ KURALI
ADRES	MAHALLE & CADDE & SOKAK & BINA & BINA_NUMARASI & KAT & DAIRE
	MAHALLE & CADDE & SOKAK & SITE_BINA_GRUBU & BINA_NUMARASI & KAT & DAIRE
	MAHALLE & CADDE & SOKAK & SITE_BINA_GRUBU & BINA_NUMARASI & DAIRE
	SITE_BINA_GRUBU & BLOK & BINA_NUMARASI
	MAHALLE & CADDE & SOKAK & BINA & BINA_NUMARASI & DAIRE
	CADDE & SOKAK & BINA & BINA_NUMARASI & DAIRE
	CADDE & SOKAK & BINA & BINA_NUMARASI & KAT
	MAHALLE & SOKAK & BINA & BINA_NUMARASI & KAT
	MAHALLE & CADDE & BINA & BINA_NUMARASI & KAT
	MAHALLE & CADDE & BINA_NUMARASI & KAT
	MAHALLE & CADDE & BINA_NUMARASI & DAIRE
	MAHALLE & SOKAK & BINA_NUMARASI & DAIRE
	MAHALLE & SOKAK & BINA_NUMARASI & KAT
	SOKAK & MAHALLE & BINA_NUMARASI & KAT
	CADDE & MAHALLE & BINA_NUMARASI & KAT
	CADDE & MAHALLE & BINA & BINA_NUMARASI & KAT
	SOKAK & MAHALLE & BINA & BINA_NUMARASI & KAT
	SOKAK & MAHALLE & BINA & KAT
	CADDE & MAHALLE & BINA & KAT
	MAHALLE & CADDE & SOKAK
	MAHALLE & CADDE
	MAHALLE & SOKAK
	MAHALLE & BINA & KAT & DAIRE
	MAHALLE & CADDE & SITE_BINA_GRUBU & BLOK & BINA & BINA_NUMARASI & KAT & DAIRE
	CADDE & SITE_BINA_GRUBU & BLOK & BINA & BINA_NUMARASI & KAT & DAIRE
	SOKAK & SITE_BINA_GRUBU & BLOK & BINA & BINA_NUMARASI & KAT & DAIRE
	SITE_BINA_GRUBU & BLOK & BINA & BINA_NUMARASI & KAT & DAIRE
	CADDE & SITE_BINA_GRUBU & BLOK & BINA_NUMARASI & KAT & DAIRE
	SOKAK & SITE_BINA_GRUBU & BLOK & BINA_NUMARASI & KAT & DAIRE
	MAHALLE & CADDE & BINA & BINA_NUMARASI
	SOKAK & SITE_BINA_GRUBU & BLOK & BINA_NUMARASI
	SITE_BINA_GRUBU & BINA_NUMARASI & KAT & DAIRE
	BINA & BINA_NUMARASI & KAT & DAIRE

Tablo 4.6 : Adres Alanı Ayrıştırma İçin Türetilen Ana Kategori Gramer Kuralları II

ANA KATEGORİ İsmi	KATEGORİ KURALI
ADRES	MAHALLE & CADDE & BINA_NUMARASI & BINA
	MAHALLE & SOKAK & BINA_NUMARASI
	MAHALLE & SOKAK & SITE_BINA_GRUBU & KAT
	SOKAK & BINA_NUMARASI & DAIRE
	BINA & BINA_NUMARASI & KAT
	SOKAK & BINA_NUMARASI & MAHALLE
	CADDE & MAHALLE & SOKAK & BINA_NUMARASI
	CADDE & SOKAK & BINA_NUMARASI
	CADDE & BINA_NUMARASI & MAHALLE
	CADDE & BINA & BINA_NUMARASI
	MAHALLE & SOKAK & SITE_BINA_GRUBU & BINA_NUMARASI & DAIRE
	MAHALLE & CADDE & BINA_NUMARASI & SITE_BINA_GRUBU & BLOK & DAIRE
	MAHALLE & CADDE & SOKAK & BINA_NUMARASI
	SITE_BINA_GRUBU & KAT & DAIRE
	CADDE & SOKAK & BINA
	SOKAK & BINA & BINA_NUMARASI
	SOKAK & BINA & BINA_NUMARASI & KAT
	ORGANİZASYON
	SEMT & ILCE
	SEMT & ILCE & IL
	SEMT & IL & ILCE
	ILCE
	ILCE & IL
	IL
	IL & ULKE
	IL & ILCE
	ORGANİZASYON & ADRES
	ADRES & ULKE
	ADRES & IL
	ADRES & ILCE
	ADRES & SEMT
	ADRES & BELDE
	ADRES & KOY
	ADRES & EK_BILGI
	ULKE & ADRES
	IL & ADRES

Şekil 4.27 oluşturulan ‘TRTUR OrgAddress’ gramer kuralları kütüphanesinde ‘SEMT’ bileşeni için tanımlanan kuralları göstermektedir. Bu kütüphanede yukarıda belirtilen tüm kurallar tanımlanmıştır.

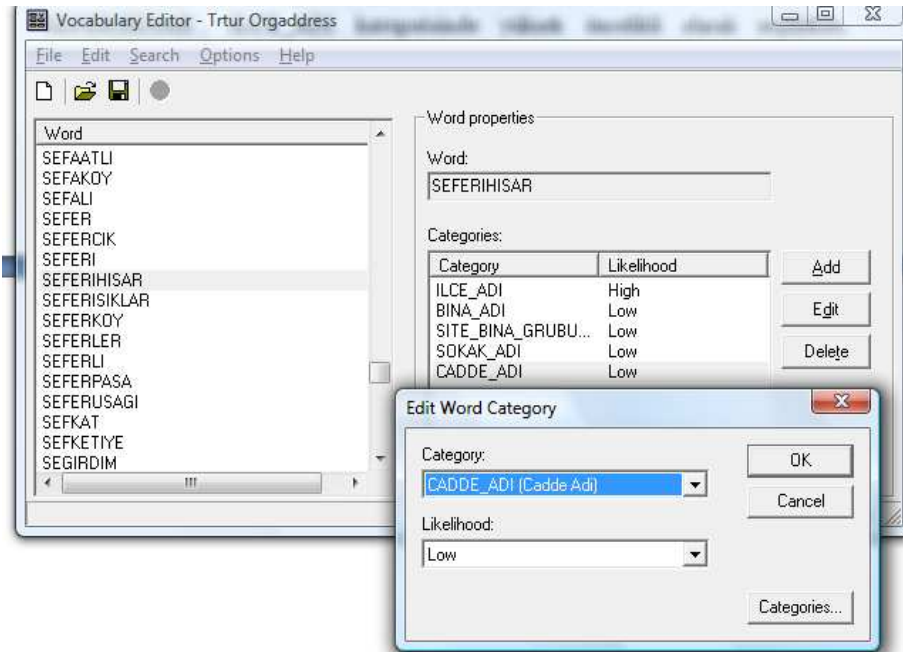


Şekil 4.27 : ‘TRTUR OrgAddress’ Gramer Kuralları Kütüphanesi

Gramer kuralları belirlendikten sonra, oluşturulan temel kategorilere (IL_ADI, ILCE_ADI, BINA_ADI, MAHALLE_BELIRTECI) konu olan kelimelerin belirleneceği bir veri sözlüğü tanımlanmıştır. Veri sözlüğünün oluşturulmasında il, ilçe, semt, belde, mahalle, köy isimleri PTT web sayfasından indirilen ‘Türkiye Adres ve Posta Kodları Özet’ tablosu kullanılarak elde edilmiştir. Tablodan elde edilen bilgiler Türkçe karakterlerinden arındırılmış ve ilgili kategorileri belirlenerek ‘TRTUR Orgaddress’ veri sözlüğü oluşturulmuştur.

Tez çalışmasına konu olan adresler sadece Türkiye adresi olduğu için ULKE_ADI kategorisine sadece ‘TÜRKİYE’ eklenmiştir. BINA_ADI, CADDE_ADI, SOKAK_ADI ve SITE_BINA_GRUBU_ADI kategorilerini oluşturacak Türkiye’deki adresler için bir veri tabanı bulunamamıştır. Bundan dolayı PTT’den alınan ilçe, semt, belde, mahalle, köy

isimlerine “düşük öncelik”le BINA_ADI, CADDE_ADI, SOKAK_ADI ve SITE_BINA_GRUBU_ADI kategorileri eklenmiştir. Veri sözlüğü oluşturulurken her kelime için belirlenen her kategoriye önceliklendirmek mümkündür. Örneğin; ‘SEFERIHISAR’, ILCE_ADI kategorisinde yüksek öncelikli olarak seçilirken, BINA_ADI, CADDE_ADI, SOKAK_ADI ve SITE_BINA_GRUBU_ADI kategorilerinde düşük öncelikli olarak belirtilmiştir. Şekil 4.28, ‘TRTUR Orgaddress’ veri sözlüğünü göstermektedir.



Şekil 4.28 : ‘TRTUR Orgaddress’ Veri Sözlüğü

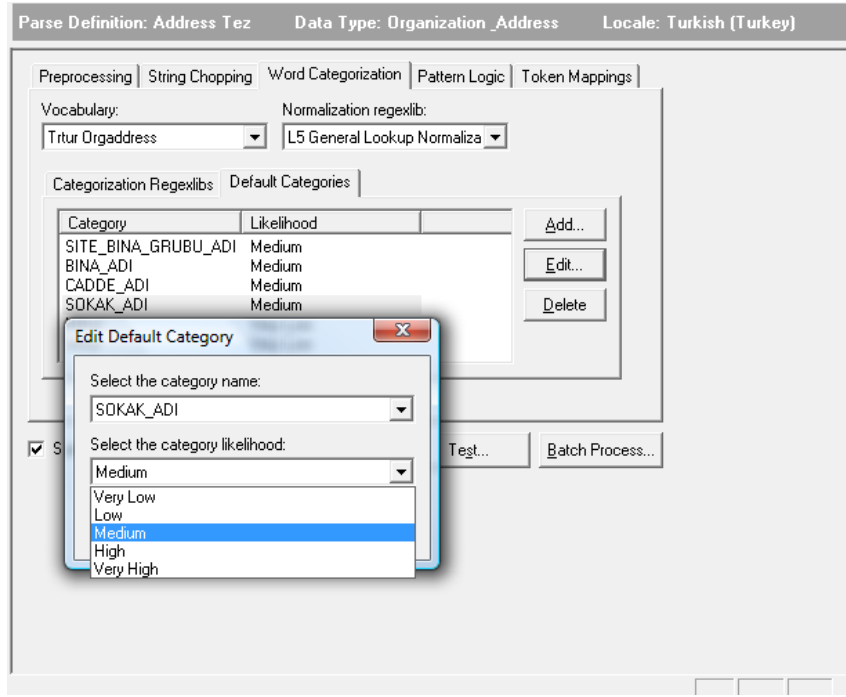
Veri Sözlüğü hazırlanırken belirteç niteliğinde olan CADDE_BELIRTECI, SOKAK_BELIRTECI, MAHALLE_BELIRTECI gibi kategorilere ilgili ‘Cad., Caddesi, Sokak, Sokağı, Sok., Mahallesi, Mahalle, Mah., Mh.’ gibi kelimeler eklenmiştir.

ADR_ADRES alanının bileşenlerine ayrıştırılması için oluşturulan gramer kurallarının ve veri sözlüğünün kullanıldığı ‘Address Tez’ ayrıştırma tanımı yaratılmıştır.

Şekil 4.29 oluşturulan ayrıştırma tanımını göstermektedir. Ayrıştırma tanımında belirtilen kurallar aşağıdaki gibidir.

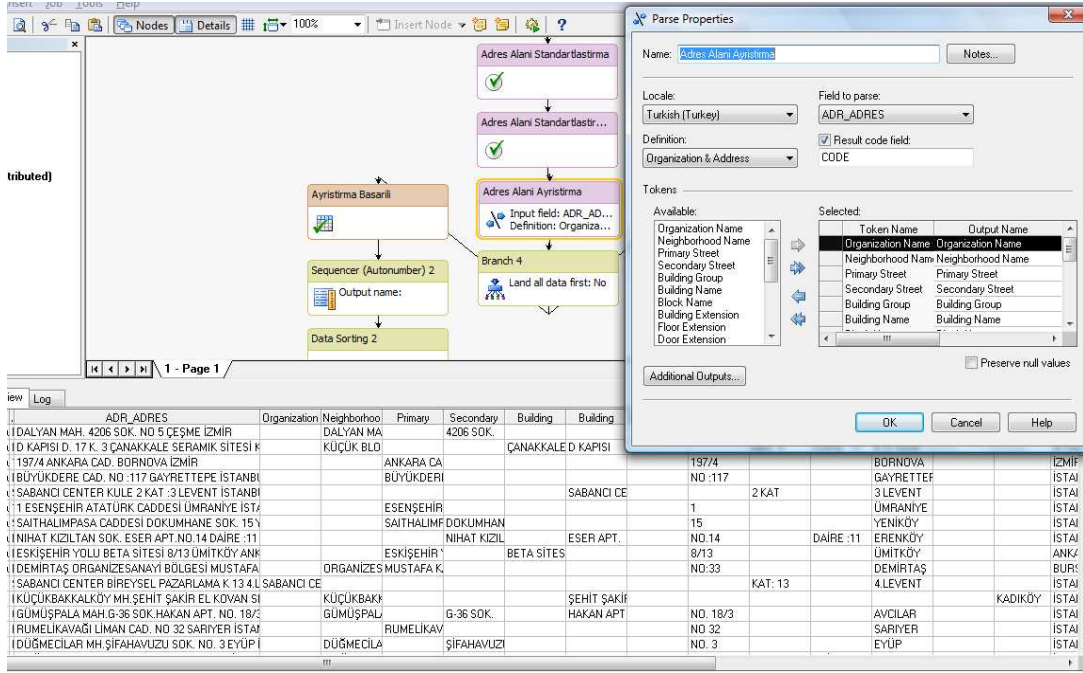
- Veri içerisindeki her kelimeyi ayırt edebilmek için; boşluk veya ‘_’ gibi ifadeler kelimeler arası ayraç olarak kabul edilecektir. Bunun için ‘TRTUR Subdistrict District City’ kelime ayırma kuralları tablosu (string chopping tablosu) kullanılmıştır.
- Veriler büyük harfe çevrilerek ve Türkçe karakterlerinden arındırılarak normalize edilecek ve veri sözlüğünde bu kelimelere karşılık gelen kategoriler aranacaktır. Bunun için, ‘L5 General Lookup Normalization’ düzenli ifadeler (regex) kuralları oluşturulmuştur.
- ‘N’, ‘K’ ya da ‘D’ harflerini takip eden “:” işareti ve sonrasında da bir sayı varsa bu harfler sırasıyla NO, KAT ve DAIRE olarak değiştirilecek ve sonrasına “:” işareti eklenecektir.
- ‘TRTUR Orgaddress’ veri sözlüğü kullanılacaktır.
- ‘TRTUR OrgAddress’ gramer kuralları kütüphanesi kullanılacaktır.
- ‘Organization Name’ belirteci ORGANIZASYON kategorisini,
‘Neighborhood Name’ belirteci MAHALLE kategorisini,
‘Primary Street’ belirteci CADDE kategorisini,
‘Secondary Street’ belirteci SOKAK kategorisini,
‘Building Group’ belirteci SITE_BINA_GRUBU kategorisini,
‘Building Name’ belirteci BINA kategorisini,
‘Block Name’ belirteci BLOK kategorisini,
‘Building Extension’ belirteci BINA_NUMARASI kategorisini,
‘Floor Extension’ belirteci KAT kategorisini,
‘Door Extension’ belirteci DAIRE kategorisini,
‘Village Name’ belirteci KOY kategorisini,
‘Subdistrict’ belirteci SEMT kategorisini,
‘Town’ belirteci BELDE kategorisini,
‘Postal Code’ belirteci POSTA_KODU kategorisini,

‘District’ belirteci ILCE kategorisini,
‘City’ belirteci IL kategorisini,
‘Country’ belirteci ULKE kategorisini,
‘Additional Info’ belirteci EK_BILGI kategorisini belirtmektedir.



Şekil 4.29 : ‘Address Tez’ Ayırıştırma Tanımı

Tanımlanan ayırıştırma tanımı kullanılarak; ADR_ADRES alanı bileşenlerine ayırıştırılmıştır. Şekil 4.30, ‘Address Tez’ ayırıştırma tanımı kullanılarak belirlenen bileşenlere ayırıştırma işini ve ayırıştırma sonucu oluşan alanları göstermektedir. Ayırıştırma kuralları kullanarak başarılı bir ayırıştırma yapıldıysa oluşturulan iş (job) sonucunda ‘CODE’ alanı üretilmektedir ve ayırıştırılan adreslerde bu CODE alanı ‘OK’ olarak belirtilmektedir.



Şekil 4.30 : ADRES Alanının Bileşenlerine Ayrıştırılması

Toplam 2722 adres verisi içerisinde 2176 (%80) adet adres başarılı bir şekilde ayrıştırılırken 546 (%20) adet adres bileşenlerine ayrıştırılamamıştır. Bu oran oluşturulan ayrıştırma tanımındaki gramer kurallarının ve veri sözlüğünün iyileştirilmesi yoluyla, ek kurallar tanımlanarak ya da standartlaştırma şemalarının geliştirilmesi yoluyla artırılabilir fakat gözden çıkarılmaması gereken bir nokta büyük hacimde veri tabanına sahip olan kurumlarda çok çeşitli adres yazım kombinasyonlarının olacağı ve tüm kuralları belirlemenin ve %100 ayrıştırma yapabilmenin bütçesel, sistemsal ve zamansal açıdan çok maliyetli olduğudur.

Tablo 4.7, 4.8, 4.9, 4.10 bileşenlerine ayrılmış örnek 20 adet adres verisini göstermektedir.

Tablo 4.7 : Bileşenlerine Ayrılmış Örnek 20 Adet Adres Verisi I

ADR_FULL	KARTALTEPE MAH. TERAKKİ CAD. NO. 22/5 LALE APT BAKIRKÖY İSTANBUL	ÖMERPAŞA SOK. NO 45/8 GÖZTEPE KADIKÖY İSTANBUL	TEYYARECİ SADIK SOK NO. 3 DAİRE 10 BAKIRKÖY İSTANBUL	ÇAMLIK CAD. İĞDE SOK NO :8/7 BAHÇELİEVLER İSTANBUL	BASIN EKSPRES YOLU MEHMETAKİF MAH. FEZA SOK. NO. 33 İKİTELLİ KÜÇÜKÇEKMECE İSTANBUL
Organizasyon					
Mahalle	KARTALTEPE MAH.	GÖZTEPE			MEHMETAKİF MAH.
Cadde	TERAKKİ CAD.			ÇAMLIK CAD.	BASIN EKSPRES YOLU
Sokak		ÖMERPAŞA SOK.	TEYYARECİ SADIK SOK.	İĞDE SOK.	FEZA SOK.
Site/Bina Grubu					
Bina	LALE APT.				
Blok					
Bina Numarası	NO. 22/5	NO 45/8	NO. 3	NO :8/7	NO. 33
Kat					
Daire			DAİRE 10		
Köy					
Semt				BAHÇELİEVLER	İKİTELLİ
Belde					
Posta Kodu					
İlçe	BAKIRKÖY	KADIKÖY	BAKIRKÖY		KÜÇÜKÇEKMECE
İl	İSTANBUL	İSTANBUL	İSTANBUL	İSTANBUL	İSTANBUL
Ek Bilgi					

Tablo 4.8 : Bileşenlerine Ayrılmış Örnek 20 Adet Adres Verisi II

ADR_FULL	NAMIKKEMAL MAH. ALTINDAĞ SOK NO 1/4 ÜMRANİYE İSTANBUL	İSTİNYE EMİRGAN CAD. MERTOĞLU APT NO.94/4 İSTİNYE İSTANBUL	CUMHURİYETİ CAD. BAHÇE SOK NO 3 TUZLA İSTANBUL	GÜRSEL MAH. MERCEK SOK NO. 10 KAT :2 KAĞITHANE İSTANBUL	OTTO VERSAND LMT. MERKEZ MAH.KOYALTI MEVKİİ SOK NO 24 İSTANBUL
Organizasyon					OTTO VERSAND LMT.
Mahalle	NAMIKKEMAL MAH.			GÜRSEL MAH.	MERKEZ MAH.
Cadde		İSTİNYE EMİRGAN CAD.	CUMHURİYETİ CAD.		
Sokak	ALTINDAĞ SOK.		BAHÇE SOK.	MERCEK SOK.	KOYALTI MEVKİİ SOK.
Site/Bina Grubu					
Bina		MERTOĞLU APT.			
Blok					
Bina Numarası	NO 1/4	NO.94/4	NO 3	NO. 10	NO 24
Kat				KAT 2	
Daire					
Köy					
Semt	ÜMRANİYE	İSTİNYE	TUZLA	KAĞITHANE	
Belde					
Posta Kodu					
İlçe					
İl	İSTANBUL	İSTANBUL	İSTANBUL	İSTANBUL	İSTANBUL
Ek Bilgi					

Tablo 4.9 : Bileşenlerine Ayrılmış Örnek 20 Adet Adres Verisi III

ADR_FULL	DENİZ CAD. FIRIN SK NO 8/16 ALTKAYNA RCA KADIKÖY İSTANBUL	DUGMECILER MH.BICAKCI SK KAMERHAN SİTESİ NO: :45 DR :08 EYÜP İSTANBUL	HARAÇCI BELDESİ HACIMAŞLI KÖY YOLU ÜZERİ NO. 48 ARNAVUTKÖY İSTANBUL	İNKILAP MAH. KÜÇÜKSU CD. NO. 149 KAPLANKE NT SİT. D BLK DAİRE 12 ÜMRANIYE İSTANBUL	LALEZAR SOK. 1 11 FENERBAHÇE İSTANBUL
Organizasyon					
Mahalle		DUGMECILER MH.	ARNAVUTKÖY	İNKILAP MAH.	
Cadde	DENİZ CAD.		HACIMAŞLI KÖY YOLU ÜZERİ	KÜÇÜKSU CD.	
Sokak	FIRIN SK	BICAKCI SK			LALEZAR SOK.
Site/Bina Grubu		KAMERHAN SİTESİ		KAPLANKE NT SİT.	
Bina					
Blok				D BLK.	
Bina Numarası	NO 8/16	NO : :45	NO. 48	NO. 149	11
Kat					
Daire		DR :08		DAİRE 12	1
Köy					
Semt		EYÜP		ÜMRANIYE	FENERBAHÇE
Belde			HARAÇCI BELDESİ		
Posta Kodu					
İlçe	KADIKÖY				
İl	İSTANBUL	İSTANBUL	İSTANBUL	İSTANBUL	İSTANBUL
Ek Bilgi	ALTKAYNA RCA				

Tablo 4.10 : Bileşenlerine Ayrılmış Örnek 20 Adet Adres Verisi IV

ADR_FULL	MERKEZ MH.29 EKİM CD.SEDİR SOK NO. 8 YENİBOSNA İSTANBUL	BARBAROS C.KAPLAN İŞHN NO 24/20 B EVLER İSTANBUL	İKİTELLİ ORGANİZE BÖLGE ESOT SİT S BLK. 5-6 İKİTELLİ İSTANBUL	NAR TEPE MH.ATATÜRK CD.AKDENİZ SK NO. 42 GÜRPINAR BÜYÜKÇEKMECE İSTANBUL	MOTIS LTD.STİ. BEY MAH. KAYACIK SOK.NEHİR İSHANI KAT:4/11 GAZİANTEP
Organizasyon					MOTIS LTD.STİ.
Mahalle	MERKEZ MH.			NAR TEPE MH.	BEY MAH.
Cadde	29 EKİM CD.	BARBAROS C.		ATATÜRK CD.	
Sokak	SEDİR SOK.			AKDENİZ SK	KAYACIK SOK.
Site/Bina Grubu			İKİTELLİ ORGANİZE BÖLGE ESOT SİT		NEHİR İSHANI
Bina		KAPLAN İŞHN			
Blok			S BLK.		
Bina Numarası	NO. 8	NO 24/20	05.Haz	NO. 42	
Kat					KAT :4/11
Daire					
Köy					
Semt	YENİBOSNA		İKİTELLİ		
Belde				GÜRPINAR	
Posta Kodu					
İlçe		B EVLER		BÜYÜKÇEKMECE	
İl	İSTANBUL	İSTANBUL	İSTANBUL	İSTANBUL	GAZİANTEP
Ek Bilgi					

Bileşenlerine ayrıştırılamayan adreslere örnek olarak aşağıdaki adresler gösterilebilir. Gösterilen örnek adreslere bakıldığında ya anlamsız kelimelerin girildiğini, ya gramerde tanımlanmayan bileşen sıralarının yer aldığını ya kelime tekrarları olduğunu ya da sözlükte tanımlanmayan kelimeler kullanıldığı görülmektedir.

- 50.YIL MAH. O CAD. NO 102 K 4 SULTANÇİFTLİĞİ GA İSTANBUL
Varolan adreste yer alan “GA” kelimesi sözlükte yer almamaktadır ve adres parse edilememiştir.
- SAN BİR BULVARI 4.BÖLGE NO. 58 HADIMKÖY BÜYÜKÇEKMECE HADIMKÖY İSTANBUL
Varolan adreste “Hadımköy” kelimesi iki defa yer almıştır ve bu şekilde bir gramer kuralı gramer kütüphanesinde yer almamaktadır.
- FEVZİPAŞA C N/137 FATİH İSTANBUL FATİH İSTANBUL
Varolan adreste “Fatih İstanbul” ifadesi iki defa yer almıştır ve bu şekilde bir gramer kuralı gramer kütüphanesinde yer almamaktadır.
- EMEKLİ İSTANBUL
Varolan adreste yer alan “Emekli” kelimesi sözlükte yer almamaktadır ve bir adreste yer alması beklenmeyen bir kelimedir, adres parse edilememiştir.
- TOTAL OİL BAYİİ YETMİŞBEŞİNCİYIL MAH.CUMHURİYET CADN 11 KÜÇÜKKÖY İSTANBUL
Varolan adreste yer alan “Cadn” kelimesi sözlükte yer almamaktadır ve yanlışlıkla Cadde ve No kısaltmasının birlikte yazılmış halidir. Standartlaştırma şemasına bu kelime eklendiği takdirde adres parse edilebilir duruma gelmektedir.
- ERNA MAŞ TİCARET SAN A.Ş. AVC İSTANBUL
Varolan adreste yer alan “Avc” kelimesi sözlükte yer almamaktadır, standartlaştırma şemasına bu kelime eklendiği takdirde adres parse edilebilir duruma gelmektedir fakat bu durumda da her “Avc” kısaltmasının Avcılar’ı tanımladığı bilinemeyeceği için başka adresleri bozma riski vardır. Bu tarz kesin olmayan durumlarda diğer adresleri bozma riskine karşı hiçbir düzeltme yapılmaması tercih edilmiştir.

SONUÇ

Yapılan herhangi bir analizin kalitesi ancak onu besleyen verilerin güvenilirliği ve kalitesi kadardır. Güvenilir olmayan veri ile çalışan şirketler yapılan çeşitli analizlerin yanlış sonuçlar ürettiğini fark etmektedirler. Bunun en önemli sebebi veri kalitesindeki yetersizliklerdir. Veri kalitesi ile ilgili sorunlara bir çözüm bulunmadıkça; veri ambarı, veri madenciliği, pazarlama otomasyonları ve diğer veriye dayalı çözümlerin başarısızlıkla sonuçlanması ve gerekli yatırım getirisini sağlayamamaları kaçınılmazdır.

Etkili bir veri kalitesi stratejisi, kurumun iş çevresini anlamasında, karlılığı artırmada, maliyet yaratan operasyonel etkinsizliklerin azaltılmasında ve müşteri memnuniyetini artırmada yardımcı olacaktır. Veri kalitesi yönetiminin temel amacı; ham veriyi tutarlı, doğru ve güvenilir bir kurumsal bilgi haline dönüştürmek için altyapının oluşturulmasıdır ve bu sayede işletmeler kurumsal rekabette önemli bir avantaj sağlayacaktır.

Veri Kalitesi tez çalışması olarak bir veritabanından alınan müşteri adres verileri üzerinden adres verilerinin analiz edilmesi, profillenmesi, veri kalitesi kurallarının oluşturulması ve bu kurallar doğrultusunda veri temizlenmesi, standartlaştırma şemalarının oluşturulması ve adres bilgisinin standartlaştırılması, adres ayrıştırma kurallarının belirlenmesi, kullanılacak sözlüklerin ve gramer kütüphanelerinin yaratılması ve adresin bileşenlerine ayrıştırılması ele alınmış ve toplam 2722 adres verisi içerisinde 2176 (%80) adet adres başarılı bir şekilde ayrıştırılırken 546 (%20) adet adres bileşenlerine ayrıştırılamamıştır.

Veri tabanlarında bir bütün olarak saklanan adres verilerinin “Organizasyon, Mahalle, Cadde, Sokak, Site/Bina Grubu, Bina, Blok, Bina Numarası, Kat, Daire, Köy, Semt, Belde, Posta Kodu, İlçe, İl, Ek Bilgi” gibi bileşenlerine ayrıştırılarak verilerin ayrı ayrı saklanması organizasyonlara aşağıdaki faydaları getirecektir:

- Kurumların kat/daire seviyesine kadar coğrafi bölge/alanlara göre müşteri hacmini belirlemelerini sağlamaktadır.
- Kat/Daire seviyesine kadar tüm adres bileşenleri bilindiği için aynı aile fertlerinin tespit edilebilmesi ve ailelere özel pazarlama kampanyalarının yapılabilmesine olanak sağlar.
- Adres verilerinin bir bütün halinde saklanması ve ayrıştırılmamış olması durumunda müşterinin hangi il, ilçe ya da mahalle gibi adreste ikamet ettiği kesin olarak bilinmemektedir. Bu durumda ise herhangi bir kampanya, dolandırıcılık, segmentasyon, çapraz satış ya da potansiyel müşteri analizlerinde bu veriler kullanılamamaktadır. Örneğin; belli bir mahallede oturan müşterilerin belli ürünleri daha çok tercih ettiği tespit edilebilirse bu mahalle müşterilerine özel çapraz satış kampanyaları düzenlenebilir ya da çeşitli pazarlama stratejileri geliştirilebilir.
- Adres verilerinin bir bütün halinde tutulması durumunda, müşterinin kurum ile tekrar temas kurması halinde müşteri temsilcilerinin hangi adres verisinin eksik olduğunu tespit etmeleri çok güçken; adresin bileşenlerine ayrıştırılmış halde tutulması hangi verinin eksik olduğu ve müşteriden bu bilginin talep edilmesi süreci çok daha kolay ve etkin olacaktır.
- Veri tabanlarında genellikle bir müşteri için yaratılmış bir çok müşteri kaydı olabilmektedir. Ad/Soyad ya da kimlik numarası gibi bilgiler ile birlikte bileşenlerine ayrıştırılmış adres verileri de kullanılarak veri tabanında bir müşteriye ait yaratılmış birçok kayıt tekilleştirilebilir ki bu da organizasyonların müşterisini tanıması ve operasyonel faaliyetlerinin azalmasını sağlayacaktır.
- Adres verilerinin bileşenlerine ayrıştırılması sonucu referans adres veritabanları kullanılarak bu adreslerin geçerli bir adres olup olmadığı saptanabilir ki bu da en önemlisi dolandırıcılık ihtimalinin önceden tespit edilebilmesi, sigortacılık gibi sektörlerde riskin hesaplanabilmesini sağlar. Aynı zamanda posta ile kampanya yapılması durumunda geçersiz adreslere posta gönderilmemesi, operasyonel maliyeti azaltacaktır.

KAYNAKÇA

Kitaplar

- Bohn, R. E., **Measuring and Managing Technical Knowledge**, 1. Baskı, Boston: Sloan Management Review, 1994
- Borland, R., **Data Warehouse Operation Strategy and ETL Processing Part I**, 1. Baskı, Boston: MIT Press, 2000
- Davenport, Thomas H. ve Prusak, Laurence, **Working Knowledge: How Organizations Manage What They Know**, 2. Baskı, Boston: Harvard Business School Press, 2000
- Feelders, A., Daniels, H., ve Holsheimer, M., **Methodological and Practical Aspects of Data Mining, Information and Management**, 1. Baskı, New York: Wiley, 2000
- Groff, Todd R. ve Jones, Thomas P., **Introduction to Knowledge Management: KM in Business**, 1. Baskı, New York: Butterworth-Heinemann, 2003
- Han, J. ve Kamber, M., **Data Mining: Concepts and Techniques**, 1. Baskı, Londra: Academic Press, 2001
- Inmon, W.H., **Building the Data Warehouse**, 4. Baskı, New York: Wiley, 2005
- Nonaka, Ikujiro, Krogh, Georg von, ve Voelpel, Sven, **Organizational Knowledge Creation Theory: Evolutionary Paths and Future Advances**, 1. Baskı, Boston: Organization Studies Press, 2006
- Nonaka, I. ve Takeuchi, H., **The Knowledge Creating Company**, 1. Baskı, New York: Oxford University Press, 1995
- Redman, Thomas C., **Data Quality For The Information Age**, 1. Baskı, Norwood: Artech House Publishers, 1997
- Salahlı, Doç. Dr. M. Ali Salahlı , **Veri Tabanı Yönetim Sistemleri Ders Notları**, 1. Baskı, Çanakkale: Çanakkale Onsekiz Mart Üniversitesi, 2006
- Senn, James A., **Analysis and Design of Information Systems**, 1. Baskı, New York: McGraw Hill, 1984

Stewart, Thomas A., **Intellectual Capital: The New Wealth of Organizations**, 1. Baskı, New York: Doubleday Business, 1998

Silahtaroğlu, Gökhan, **Kavram ve Algoritmalarıyla Temel Veri Madenciliği**, 1. Baskı, İstanbul: Papatya Yayıncılık Eğitim, 2008

Soyuyüce, Uzm.Vet.Het. Ebru ve diğerleri, "Bilgisayara Dayalı İndeksleme" **Kütüphane-Enformasyon Arşiv Alanında Yeni Teknolojiler Ve Türkmanc Sempozyumu**, İstanbul: Türk Kütüphaneciler Derneği İstanbul Şubesi, 1991

Sürekli Yayınlar

- Bolten, R. M., “Data Population in a BI Environment”, **DM Review Magazine**, Eylül 2000
- Bongiovanni, Nat, “Analyzing Domain Statistics”, **DM Review Magazine**, Haziran 2000
- Chow, K.T., “Enterprise Data Quality: Covering All Your Bases”, **DM Review Magazine**, Aralık 2000
- English, Larry, “Which Information Quality Tool is Best?”, **DM Review Magazine**, Ekim 2000
- English, Larry, “Information Quality Management: The Next Frontier”, **DM Review Magazine**, Nisan 2000
- English, Larry, “Information Quality in Reference Data”, **DM Review Magazine**, Mart 2006
- English, Larry, “Total Information Quality Management: A Complete Methodology for IQ Management”, **DM Review Magazine**, 2003
- Forino, R., “Data e.Quality: Election 2000”, **DM Review Magazine**, Ocak 2001
- No, A., “Frequently Asked Questions about Data Quality”, **DM Review Magazine**, Kasım 2000
- Rudra, A. ve Yeo, E., “Issues in user perceptions of data quality and satisfaction in using a data warehouse-an Australian experience”, **Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, 2000.**, Vol.1, Ocak 2000
- Zakaria, N., Amelinckx, A., ve Wilemon, D., (2004). “Working Together Apart? Building a Knowledge-Sharing Culture for Global Virtual Teams”, **Creativity and Innovation Management**, Vol.13, No.1

Diğer Kaynaklar

Business Intelligence Network, “Research Report - Data Profiling, Data Integration and Data Quality: The Pillars Of Master Data Management”, 1. Baskı [Broşür] 2007

Business Objects, “White Paper - Data Quality First: It’s Just Logical - The Importance of Trusted Information to Successful Data Integration and Business Intelligence”, 1. Baskı, [Broşür] 2006

DataFlux Corporation, “White Paper - Five Steps to More Valuable Enterprise Data”, 1. Baskı [Broşür] 2007

DataFlux Corporation, “White Paper - Proving Value for Data Efforts: An ROI Framework for Data Governance and Data Quality Projects”, 1. Baskı [Broşür] 2007

Dataflux Corporation, “White Paper - Data Monitoring: Add Controls to Your Data Governance Program”, 1. Baskı [Broşür] 2006

DataFlux Corporation, “White Paper - Data Management: Why Your Organization Must Understand and Achieve It”, 1. Baskı [Broşür] 2007

Gonzales, Michael L., President, The Focus Group Ltd, “Enterprise Data Quality For Business Intelligence”, [Broşür] 2003

Knightsbridge Solutions LLC, “White Paper - Worst Practices in Business Intelligence: Why BI Applications Succeed Where BI Tools Fail”, 1.Baskı, [Broşür] 2005

Knightsbridge Solutions LLC, “White Paper - A Practical Approach to Data Quality: Proven Formulas for Delivering Measurable Results”, 1. Baskı, [Broşür] 2006

Knightsbridge Solutions LLC, “White Paper - Worst Practices in Business Intelligence: Why BI Applications Succeed Where BI Tools Fail”, 1.Baskı, [Broşür] 2005

Knightsbridge Solutions LLC, “White Paper - A Practical Approach to Data Quality: Proven Formulas for Delivering Measurable Results”, 1. Baskı, [Broşür] 2006

English, Larry, “IQ and Data Warehousing: Trends and Best Practices”, *DM Review Online*, Nisan 2006, <http://www.infoimpact.com/articles/2006-04 IQ in DW-Trends & Challenges.pdf> (13 Nisan 2008)

Firth, C., (1997), "When do Data Quality Problems Occur? ", <http://sunflower.singnet.com.sg/~cfirth/dq1.htm> (13 Mart 2008)

- Rudra, Amit ve Yeo, Emilie, 1999, “Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia”, <http://dblp.uni-trier.de/rec/bibtex/conf/hicss/RudraY99> (8 Nisan 2008)
- Usgurlu, Ümit Burak, “Veritabanı, Veri Madenciliği, Veri Ambarı, Veri Pazarı”, (t.y) Başkent Üniversitesi, Bilgisayar Mühendisliği Yönetim Bilişim Sistemleri, <http://mail.baskent.edu.tr/~20394676/0302/bil483/HW2.pdf> (17 Kasım 2007) *Bilgisayar-Yazılım-ORACLE*, (t.y.) <http://www.textara.com/oracle-data-base-veri-tabani-yolu-nedir-komutlar-ornek?page=0%2C8> (4 Eylül 2007)
- <http://www.dataflux.com/Resources/> (8 Nisan 2008)
- http://www.meslekegitimleri.com/sau/01.hakkinda_bilgi.htm (22 Aralık 2007)
- <http://www.sas.com> (8 Nisan 2008)
- Data Mart vs Data Warehouse - The Great Debate*, 2008
<http://opensourceanalytics.com/2006/03/14/data-mart-vs-data-warehouse-the-great-debate> (5 Şubat 2008)
- Data Quality and the Bottom Line*, 2002,
<http://www.tdwi.org/research/display.aspx?ID=6589>, (8 Nisan 2008)
- Fazla Bilgi Eziyor*, 2002
<http://www.coretech.com.tr/Templates/haberdetay.aspx?articleid=38&zoneid=18> (4 Şubat 2008)
- İş Zekası*, 2004, <http://www.nexum.com.tr/tr/hizmetler/iszekasi.asp> (17 Temmuz 2007)
- İş Zekası Çözümleri*, (t.y.) <http://www.ironmanconsulting.com/v2/tr/3-4-is-zekasi-cozumleri.asp> (28 Temmuz 2007)
- Kurumunuzu Akıllı Yazılımlar Yönetsin*, 2002,
http://www.btvizyon.com.tr/kulakcik.phtml?kulakcik_nox=53 (7 Ocak 2008)
- OLAP-1*, (t.y.), http://www.danismend.com/konular/bilgiveteknoyon/bilgi_olap1.htm (5 Şubat 2008)
- Total Data Quality Management*, 2004,
<http://mitiq.mit.edu/MITIQ/TDQM%20Workshop/Newzealandtr46.pdf> (2 Eylül 2007)
- Veri Modelleme*, (t.y.) http://tr.wikipedia.org/wiki/Veri_modelleme (4 Eylül 2007)

Veri Ambarları, (t.y.) <http://www.ikademi.com/insan-kaynaklari-bilgi-sistemleri/1237-veri-ambarlari.html> (17 Kasım 2007)

Veri Ambarı Nedir? , 2007, <http://veriambarinedir.blogspot.com/> (17 Kasım 2007)

Veri Ambarı ve Karar Destek Sistemleri, (t.y.) <http://www.theturkler.com/makale/429/> (17 Kasım 2007)

Veri Ambarı (Datawarehousing) Temelleri, (t.y.)
http://www.bilisimrehber.com.tr/arastirma/tr_arastirma_veriambari_temelleri.phtml (17 Kasım 2007)

Veri Analizi Hizmetleri, (t.y.),
http://www.deloitte.com/dtt/section_node/0,1042,sid%253D8431,00.html (8 Mart 2008)