SAKARYA ÜNİVERSİTESİ

Veri Madenciliği Uygulamaları

Hafta 3

Yrd. Doç.Dr. Nilüfer YURTAY

Veri Ambarları ve OLAP(On Line Analytical Processing)

3.1 Giriş

İlişkisel veritabanlarının boyutları artık bir çok orta-büyük seviyeli uygulamalarda onlarca gigabytetan terabyte'lara ulaşan boyutlarıyla raporlamayı imkansız bir hale getiriyor, üstelik bu raporlar genelde "elimizdeki ürünler ve onların **satış** fiyatları" gibi kolay raporlar olamıyor. Verinin ilişkisel yapısı nedeniyle karışıklığının yanında, boyutunun büyüklüğü nedeniyle doğan performans kaybı tahammül edilemez seviyelere ulaşabiliyor. Bir rapor almak için saatlerce bekleyen firmalar oldukça çoktur.

Günümüzde bir kurumun operasyonel seviyedeki hizmetlerini sunabilmesi için bilgi sistemleri vazgeçilmez olarak kabul edilmektedir. Çünkü bilgi işlem hizmetleri aksadığında kurum çalışamaz hale gelir..! Yöneticilerin karar verme sürecinde bilgi sistemlerinin rolü tahmin edilenden çok daha fazladır. Bu konu yavaş yavaş kabul görmektedir. Bu kabulü hızlandırmak için Veri ambarı ve OLAP kavramlarını incelemeliyiz..

Kurum yöneticilerinin karar faaliyetlerinde doğru ve zamanında bilgi önemlidir. Bu bilgi gerçekte kurumun işleyişi sırasında toplanan verilerde mevcuttur. **Karar destek sistemleri**, kurum içi ve dışı verilerin, karar verme sürecinde kullanılabilecek bilgiye dönüştürülmesiyle ilgilenir.

Bir kuruma ait veri değişik kaynaklarda bulunabilir. Bunların kolay ulaşım için tek bir havuzda toplanması istenir. Ayrıca kurumun operasyonel işlemlerini gerçekleştirdiği **OLTP (OnLine Transaction Processing)** sistemler (veritabanları) bilgi toplama üzerine (kayıt ekleme, çıkarma, silme ki bunlar hareket/Transaction olarak bilinir) uzmanlaşmıştır. OLTP sistemlerin karar destek faaliyetlerinde kullanılması performans açısından tavsiye edilmez. OLTP sistemlerden, karar destek faaliyetlerinde kullanılacak verilerin, (denormalizasyon gibi) performans kazandırıcı değişimlerden sonra, bu tek havuza toplanması gerekir. Ek olarak, OLTP sistemlerde, verilerin geçmiş halleri tutulmayabilir. Aslında karar verme açısından verideki değişim, yani verinin tarihsel değişimi de önemlidir. Bunların da bu tek havuzda tutulması gerekecektir. İşte bu gibi nedenlerle oluşturulan bu havuza **veri ambarı (data warehouse)** diyoruz. Verilerin ilgili kaynaklardan çekilip veri ambarına aktarılması **ETL (Extract Transform Load)** olarak bilinir. Bu iş için Microsoft DTS (Data Transformation Services) aracı kullanılabilir.

Veri ambarındaki verilerin karar destek faaliyetlerinde kullanılması aşağıdaki şekillerde olabilir.

- Sorgulama ve raporlama
- OLAP
- Veri madenciliği

Bu listede ilginç olanlar OLAP ve veri madenciliğidir.

Veri madenciliği (Data Mining), istatistiksel bazı yöntemlerin yardımıyla veri içinde gizli olan desenlerin ortaya çıkarılması ve bu desenlerin geleceği tahmin etmekte kullanılmasıdır. Hemen bir örnek verelim.. ASKİ'de çalışıyoruz. Şimdiye kadar tespit edilen kaçak su kullanan aboneleri veri madenciliği uygulamasına veriririz. Uygulama bu abone grubundaki gizli bağıntıları/desenleri bulur. Mesela su tüketim eğilimleri, borçlarını geciktirme süreleri vs.. Bundan sonra elimizdeki 1 milyon aboneye bu deseni uygularız ve muhtemel kaçaksu abonelerini buluruz...

OLAP (OnLine Analytical Processing) için üzerinde görüş birliğine varılan ortak özellik çok boyutlu veri analizidir (MultiDimensional analyzing). Çok boyutlu veri analizinde, veri değişik boyutlardan incelenir. Veri ve boyutları birlikte, küp olarak adlandırılır. Mesela satış verisinin, zaman, ürün ve bölge boyutlarından bakılarak değişimleri incelenebilir. Bu boyutlarda istenilen ayrıntı ve özet seviyesine çıkılabilir. Böylece değişimin sebebi daha iyi anlaşılabilir. Burada bir resim, çizim sağlayamıyorum. Ama biraz hayalgücüye bu sorunu aşabiliriz...

OLAP : İlişkisel veri tabanının aksine veriyi tekrarlayarak, mümkün olduğunca az ilişki ile (dolayısıyla çok daha fazla veri alanı ile) veriyi depolayan, bu şekilde veriye erişim hızımızı cok büyük ölçüde arttıran yapılardır ve de yukarıya baktığımızda, flat file mantığı ile uyuşan yapıdır.

Bir çok makalede OLAP sözcüğü ile bütünleşmiş küp kelimesi karşımıza çıkar. OLAP yapısının küpler ile ifade edilmesinin sebebi, aynen geometrik bir küp gibi kenarlara, boyutlara, her birim hacminde bir dataya sahip olmasıdır. Dolayısı ile nasıl ki rübik küpünü her çevirişinizde farkli bir yüzey, farklı bir data (aynı verilerden elde edilen farklı sonuclar) ile karşılaşır iseniz, OLAP yapısındaki datanıza da her farklı açıdan bakışınızda çok farklı sonuçlara ulaşırsınız.

Örneğin, x adlı içeceğimizin yaz aylarındaki turistik bölgelerdeki satış miktarlarının hava sıcaklığı ve gün dağılımlı (hafta içi hafta sonu) olarak gidişatını görmek istiyorsunuz. Ya da y adlı ürünüzü alan kişilerin a ürününü mü daha çok tercih ettiğini yoksa b ürününü mü daha çok tercih ettiğini merak ediyorsunuz. Soru örnekleri çoğaltılabilir. Önemli olan beklentilerinize yönelik olarak kübünüzü kurabilmektir.

Bu teknolojiyi en çok kullanan şirketler büyük alışveriş mağazaları (hatta bu konuda öncü olan ve bu sayede de fazlasıyla büyüyen wallmart) tır. Alışveriş mağazalarının bizlere dağıttığı üyelik / club / vs.. kartlarının da esas amacı bu tarz datalara ulaşabilmektir.

Bir OLAP küpü üzerinde şu işlemler yapılabilir:

Dice(Çevir) Satış verisinin bölge-zaman yüzünü incelerken, ürün-zaman yüzüne geçebiliriz.
Slice(Dilimle): Bütün aralığı değil de belirli bir aralığı seçebiliriz. Mesela son 1 yıla ait dilim..
Drill Down: Ayrıntı seviyesinde alta in. Mesela yıl bazından ay bazına geç.
Drill Up: Ayrıntı seviyesinde yukarı çık. Mesela şehir bazından bölge bazına çık.

Buradan OLAP küplerinin sadece 3 boyutlu olabileceğini çıkarmamalıyız. Daha az veya çok boyut da olabilir.

OLAP küpleri, tutuldukları yerlere göre farklı isimler alırlar. Eğer küpler çok boyutlu veritabanında tutuluyorsa **MOLAP (Multidimensional OLAP)**, web üzerinden erişiliyorsa **WOLAP (Web OLAP)**, uç birimde tutuluyorsa **DOLAP (Desktop OLAP)**, ilişkisel veritabanında tutuluyorsa **ROLAP (Relational OLAP)** adını alır[1].

Özetleyecek olursak OLAP'ın özellikleri şu şekildedir:

- Çok boyutlu inceleme özelliğine sahip olması.

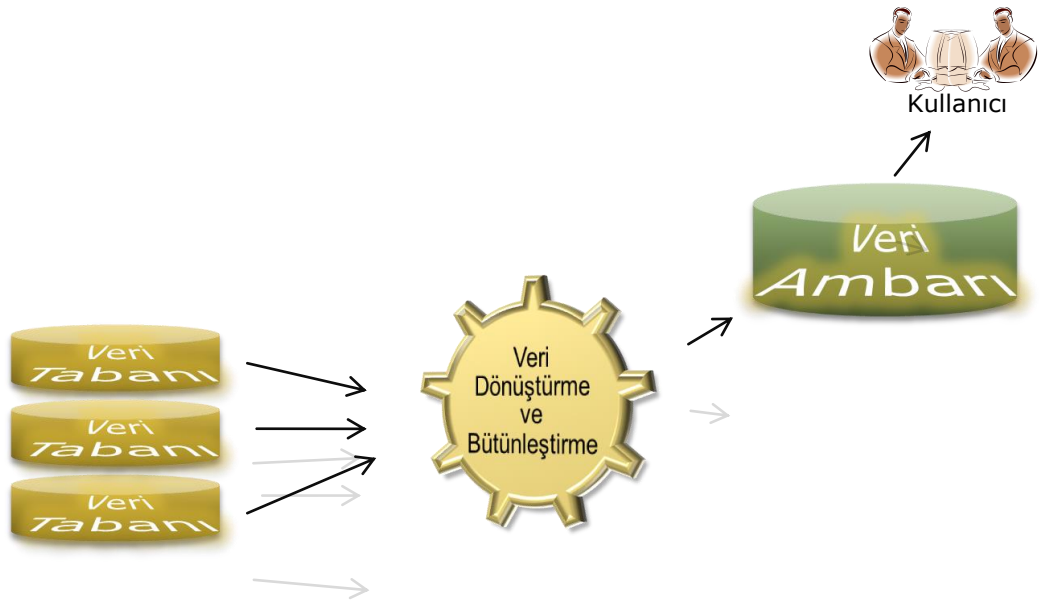
[1] http://olapreport.com/Architectures.htm

- Şeffaflık
- Erişilebilirlik
- Her seviyede sorgulama için aynı performansı gösterebilme özelliği
- Server - Client yapısında olması
- Çoklu kullanıcı desteği
- Esnek raporlanabilme
- Boyutlar ve gruplandırmalarda sınırların bulunmaması
- ve daha bir çok özellik...

3.2 Veri Ambarı Mimarisi

Verilerin oluşumundan ,kullanıcıya ulaşıncaya kadar olan sürece ait karekteristikler şunlardır.

1. Alınan verilerin dönüştürülmesi,
2. Veri ambarının oluşturulması
3. Kullanıcının erişimi,sunum

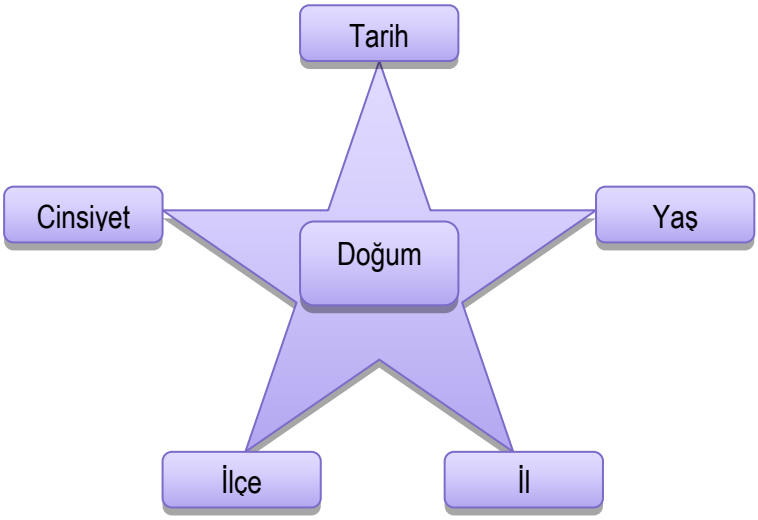


Şekil 3.1 Veri Ambarı Mimarisi

3.3 Veri Ambarı Veri Modeli

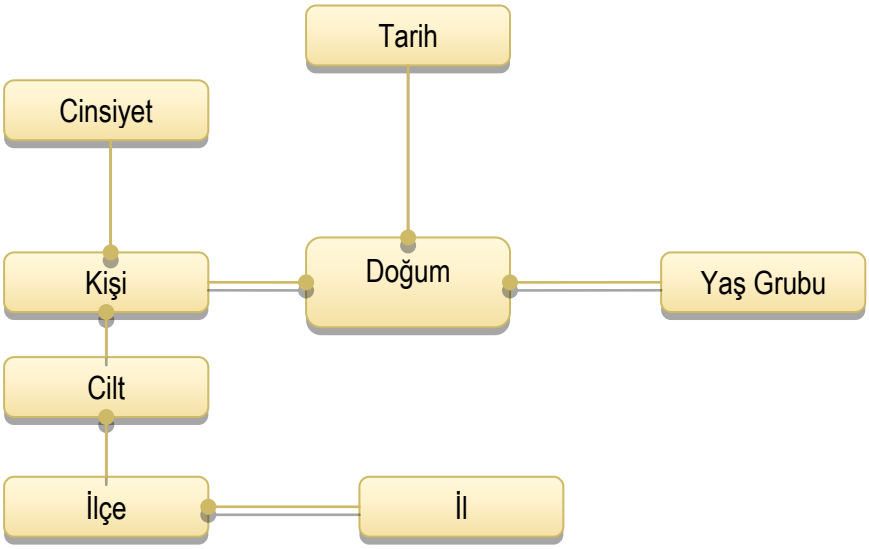
Veri ambarları uygulama programlarının veri modellerinden farklılık gösterir. Veri ambarları veriyi en kısa sürede çözümleyerek kullanıcıya ulaştırmayı amaçlar. Kurumların ihtiyaçlarına göre boyutsal model olarak ele alınır. Bir veri ambarında birden fazla boyut yer alabilir. Veri Ambarlarını oluşturan boyut tabloları, hiyerarşik bir şekilde veri tabanlarındaki tabloların temsili şeklinde yapılandırılmışlardır. Veri Ambarları çok boyutlu bütünleşik yapılara göre düzenlenmektedir. Farklı boyutların toplanması ile de bir küp yapısı oluşturulmuş olmaktadır. Veri küplerinde mutlaka bir boyut, zaman boyutu olmadır. Küp kavramı normalde 3 boyutu ifade etsede 3 den fazla olan boyuta sahip küp yapılarına da sıkça rastlanmaktadır.Bu şekilde oluşan veri modellerine veri küpü ,yıldız şema,Snowflake Şema veya Hybrid Şema olarakta adlandırılır.

Yıldız şema yapısı en sık tercih edilen şema modelidir. Merkezde konumlandırılmış bir fact table ve onun etrafını sarmış dimension table'ların bir yıldız şeklini anımsatmasından dolayı da Star Scheme (Yıldız Şema) adını almıştır. Bu şema yapısında fact table 'a keylerle bağlanmış olan dimension tablelar demormalize edilmiştir.

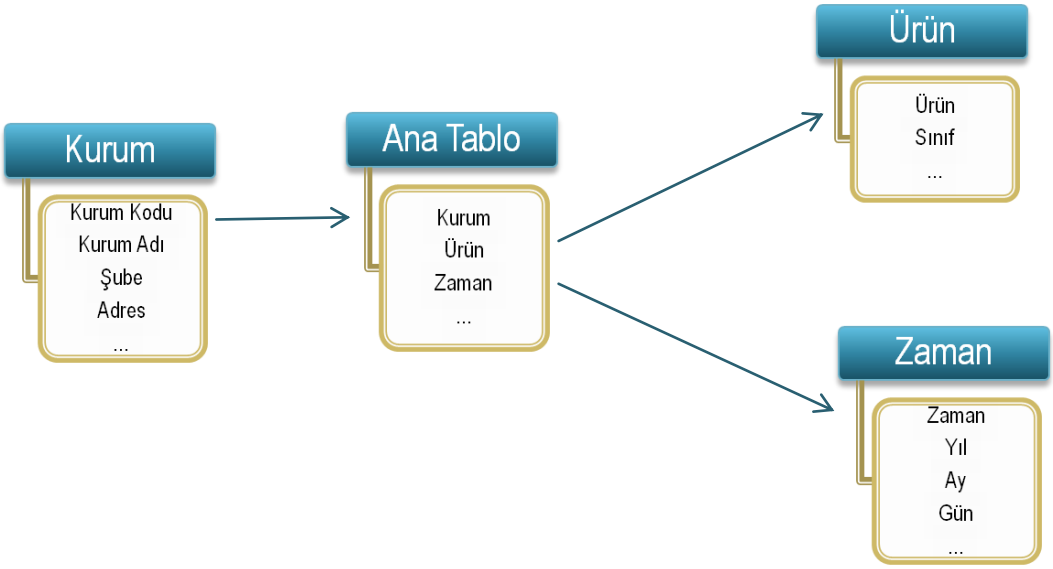


Şekil 3.2 Yıldız Şema

Snowflake şema yapısındaysa; merkezde konumlandırılmış bir fact table ve onun etrafını sarmış dimension table'lara bağlanmış başka dimension table'ların kar tanelerine benzer bir şekil oluşturmasından dolayı Snowflake Schema denmiştir. Bu şema yapısında fact table 'a keylerle bağlanmış olan dimension tablelar normalizasyon kurallarına göre dizilmişlerdir, zaten dimensionlara bağlı başka dimensionlar olması da normalizasyon kurallarına uymak içindir. Bu şema yapısının kurulması çok zaman alması bilgiye ulaşılmasını zorlaştırması nedeniyle de pek fazla tercih edilmemektedir.



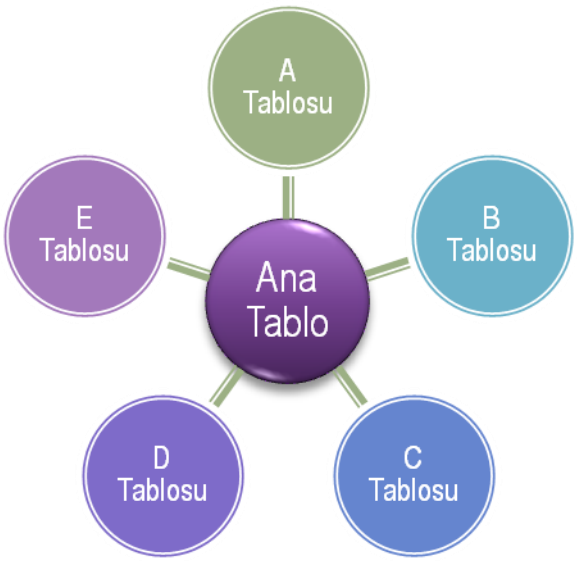
Şekil 3.3 Snowflake şema



Şekil 3.4 Tipik bir çok boyutlu model

3.4 Verinin Dönüşümü ve Oluşturulması

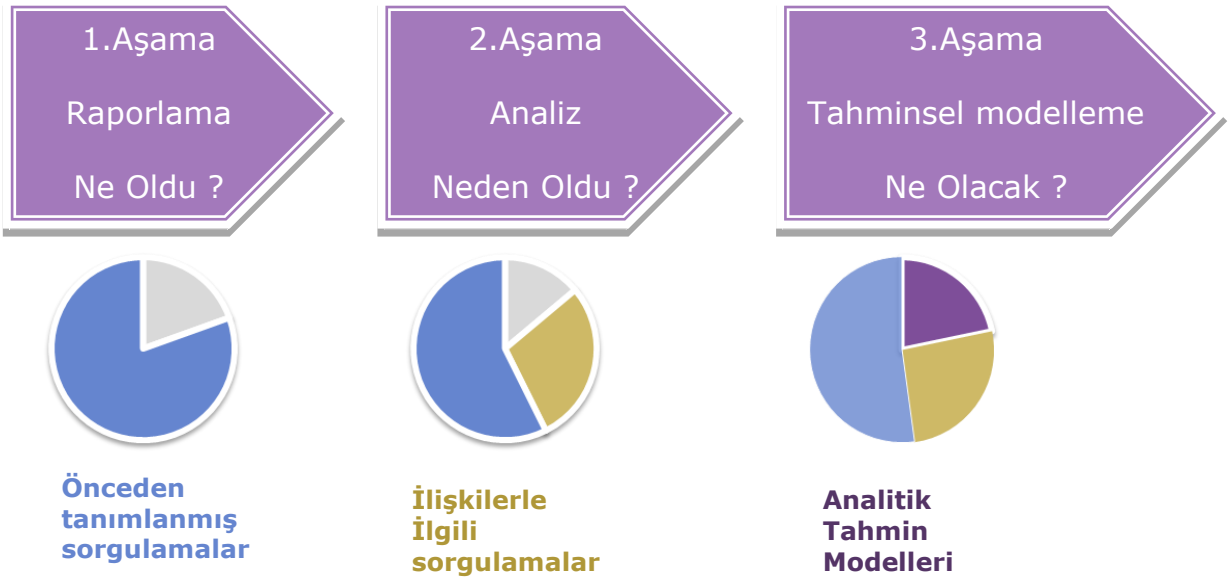
Aynı veriler farklı biçimde yer alabilir yada veriler üzerinde bozulmalar söz konusu olabilir. Bu durumlarda öncelikle verilerin iyileştirilmesi ve düzenlenmesi gerekir. Bu düzenlemeler ve özetlemeler aynı zamanda bütünleştirme aşamasının da gerçeklenmesini sağlar. Bu düzenlemeler aynı verilerin farklı ifade edilme biçimlerini de ortadan kaldırır. Dönüştürülen veri farklı bir fiziksel ortamlarda oluşturulan veri ambarlarına aktarılır. Aktarılan bu veriler üzerinde sadece sorgular çalıştırılarak kullanıcıların karar verme sürecine destek olurlar.



Şekil 3.5 Veri Dönüşümü

3.5 Veri Ambarında Sorgulama

Veri ambarları önceleri raporlama amaçlı sorgulamalar için kullanılıyordu. Sonra analiz aşamasına, daha sonrada tahminsel modellemeler yapma aşamasına geçmek mümkün oldu.



Şekil 3.6 Veri Ambarında Sorgulama

Raporlama aşamasında yapılan sorgulamalar :

- Geçen X aylık dönemde, müşteri ilişkilerinde en karlı dağıtım kanalı hangisidir?
- Ürün ve müşteri grupları için, aylık ve haftalık satış miktarları ne kadardır?
- Kimlerin bakiyeleri belirli bir miktarın altında?

Analiz aşamasında yapılan sorgulamalar :

Zaman içinde gelinen bu aşama, olaylar arasında bağlantı kurmak, birbirleriyle ilişkili ve birbirini etkileyen değişkenleri ortaya çıkarmak şeklinde gerçekleşmekte.

Örneğin, cinsiyet ile satın alma tutumu arasında nasıl bir ilişki olduğu, kadınların satın alma tutumlarının erkeklerden farklı olmasının sebeplerinin, belirli bölgelerde satışların artmasının veya azalmasının sebeplerinin neler olduğu türünde sorulara cevap bularak, ilişkilerin analiz edilmesini gerektiren sorgulamalardır.

Veri ambarında tutulan veriler, müşteriye ait tüm verileri, müşterinin işletmeyle olan tüm ilişkisini detaylı bir şekilde gösteren verilerdir. Ne almış, ne zaman almış, ne kadar ödemiş -kredi kartı müşterisi ise -kredi kartının limiti nedir, ödemesini peşin mi yoksa taksitle mi ve hangi yöntemle -şubeden mi, İnternetten mi, çağrı merkezinden mi- yapıyor, gibi soruların cevabı olan verilerdir. Verilerin müşteri bazında tutulması enformasyonel bilgi kaynağı olmasını sağlar.

Tahminsel Modelleme aşamasında yapılan sorgulamalarda ise yapılacak en kritik sorgulamalardan biri şudur:

"Önümüzdeki günlerde, haftalarda hangi müşterilerimizi kaybetme riskimiz var ve bu riskin derecesi nedir?"

Bu sorgulama sonucu elde edilen "çıkarımsal bilgi", "knowledge" olarak ifade edilen, keşifsel ve proaktif bir bilgidir.

Bu bilgiyi elde etmekteki amaç, kaybetme riski olan müşteriyi ayırdıktan sonra hangileri için bunun engellenmesi gerektiğini belirlemektir. Yani en azından, kaybetme riskini azaltmak için bir eylem planı, bir strateji geliştirmek ve proaktif davranmak için kullanılacak bir bilgidir.

Bu aşama, özellikle, **veri ambarlarının** klasik veri tabanı yaklaşımından farkını en belirgin biçimde ortaya koyan aşamadır. Veri ambarları aslında ileri düzey kullanıcılar ve tahmin yapma ihtiyacı bulunan, stratejik kararları verenler için, modelleme yapabilmek ve eyleme yönelik bilgiler elde etmek açısından önemli fırsatlar sunmaktadır. Bu aşamada, veri madenciliğinin tahmin modelleri de devreye girmekte, müşterilerin tutum ve davranışları hakkın da geleceğe yönelik tahminler yapılabilmektedir.

3.6 OLAP

OLAP araçları, her iş kullanıcısının kolaylıkla kullanabileceği yapısı ile veriye çok boyutlu erişimi sağlamaktadır.

OLAP araçları ile;

- **E**n çok kâr getiren müşterilerim kimlerdir?
  (Bayi ve perakendeci bazında.)
- **E**n kârlı ürünlerim nelerdir?
  (16 mm I. Sn. Kayın masif parke gibi.)
- **H**angi işletme ya da mağazamda, en çok hangi saat ve günlerde hareketlilik olmaktadır?

gibi sorulara cevap bulunabilmektedir.

Örnek(OLAP Raporlama)

X Firması Ürünlerinin tüketiciye etkin ve verimli şekilde ulaştırılması için satış, pazarlama ve lojistik hizmetler sunan bir dağıtım şirketidir. Binlerce Bayii tarafından kullanılan ve alternatif bir tahsilat aracı olan Paynet, üzerindeki milyonlarca işlem ile büyük bir veri kaynağı oluşturuyordu. Bu verilerin farklı boyutları ile üstelik kümülatif olarak raporlanması konvansiyonel raporlama teknikleri ile geliştirilmesi vakit alıcı oluyorudu. Ayrıca verileri farklı boyutları ile görebilmek ve yorumlayabilmek isteyen müşteriler esneklik açısından tatmin olmuyorlardı.

Çözüm olarak Microsoft SQL Server Reporting Services uygulaması gerçekleştirdi. Uygulamanın en kritik fazı müşteri tarafındaki raporlanma ihtiyaçlarının en iyi şekilde analiz edilip , doğru OLAP küplerinin saptanması oldu. OLAP küplerinin; uygulamanın veri mimarisi anlamında çerçevesini çiziyor olması nedeniyle, doğru şekilde tasarlanması projenin kısa dönemdeki başarısını ve uzun dönemdeki ölçeklenebilme becerisini belirledi.

- OLAP uygulaması sayesinde; Büyük miktarda verinin anlık olarak raporlanması
- Müşterinin teknik desteğe ihtiyaç duymadan bilgiyi farklı boyutları ile değerlendirmesi
- Yeni rapor ihtiyaçlarının hızlı bir şekilde karşılanması
- Farklı ortamlara XML ile kolay veri aktarımı
- Excel ve pivot tabloları ile entegrasyon
- Kolayca PDF ve Excel'e export mümkün oldu.

Kaynaklar

1. Özkan,Y.,"Veri Madenciliği Yöntemleri ",Papatya Yayıncılık,2008
2. Silahtaroğlu,G.,"Veri Madenciliği ",Papatya Yayıncılık,2008