VERİ, VERİ AMBARI ve OLAP

Dr.Günay TEMÜR

Veri Nedir?

- Nesneler ve nesnelerin niteliklerinden oluşan küme
  - *kayıt (record), varlık (entity), örnek (sample, instance) nesne için kullanılabilir.*
- Nitelik (attribute) bir nesnenin (object) bir özelliğidir.
  - *bir insanın yaşı, ortamın sıcaklığı…*
  - *boyut (dimension), özellik (feature, characteristic) olarak da kullanılır.*
- Nitelikler ve bu niteliklere ait değerler bir nesneyi oluşturur.

Nitelik (Attributes)

Nesne (objects)

Id	cinsiyet	Medeni Durum	Yıllık Gelir
1	E	Evli	125K
2	K	Bekar	80K
3	K	Evli	50K
4	E	Bekar	60K
5	E	Evli	70K
6	E	Bekar	30K
7	K	Evli	40K
8	E	Evli	100K
9	K	Evli	80K

Değer Kümeleri

- Nitelik için saptanmış sayılar veya semboller
- Nitelik & Değer Kümeleri
  - *aynı nitelik farklı değer kümelerinden değer alabilir*
    - ağırlık: kg, lb(libre, ağırlık ölçüsü)
  - *farklı nitelikler aynı değer kümesinden değer alabilirler*
    - ID, yaş: her ikisi de sayısal

İstatistiksel Veri Türleri

- **1- Nümerik Veriler :** Sayısal-Nümerik-Nicel Veriler de denmektedir. Boy, Yaş gibi süreklilik arz eden değerler Nümerik verilerdir. “Daha fazla” ifadesi ile kullanılabilirler. Sürekli ve süreksiz olarak iki başlıkta ele alınabilir:
  - *a) Sürekli Nümerik Veriler: Yaş, Sıcaklık*
  - *b) Aralıklı Nümerik Veriler (Interval): Çocuk Sayısı, Kaza Sayısı*
- **2-Nominal Veriler :** Kategorik bir veri çeşididir. “Daha fazla” ifadesi ile kullanılmazlar. İkiye ayrılır:
  - *a)Binary Veriler: Var-Yok, Kadın-Erkek, Hasta-Sağlıklı*
  - *b)İkiden Çok Kategorili: Medeni Durum-Renk-Irk-Şehir, İsim, Forma Numarası Örneğin forma numarası oyuncunun seviyesi ile ilgili bir bilgi içermez.*

İstatistiksel Veri Türleri

- **3-Ordinal Veriler :** Ordinal veriler de yine kategorik veri türündendir. Fakat değerleri arasında sıralı bir ilişki bulunmaktadır. “Daha fazla” ifadesi ile kullanılabilirler ancak nekadar daha fazla olduğunun ölçüsünü veremezler. Örneğim: Eğitim Düzeyi, Sosyoekonomik ölçek skorları gibi. Nominal veriler, ordinal verilere göre daha az bilgi taşırlar.
- **4-Ratio Veriler :** Nümerik verilere benzerler. 100 santigrat derece, 50 santrigat derecenin iki katı denilemez ama derece kelvine çevrilirse 60 kelvin 30 kelvinin 2 misli sıcak denilebilir. Oran verilebilir veri türlerine Ratio veriler denir. Burada kelvin derece ratio türünden bir değişken iken, santigrat ise nümerik veri türüne örnek olarak verilebilir.

Nitelik Türleri

- Belli aralıkta yeralan değişkenler (interval)
  - *sıcaklık, tarih*
- İkili değişkenler (binary)
  - *cinsiyet*
- Ayrık ve sıralı değişkenler (nominal, ordinal, ratio scaled)
  - *göz rengi, posta kodu*

Problem

- Gerçek uygulamalarda toplanan veri kirli
  - *eksik: bazı nitelik değerleri bazı nesneler için girilmemiş, veri madenciliği uygulaması için gerekli bir nitelik kaydedilmemiş*
    - *meslek = “ ”*
  - *gürültülü: hatalar var*
    - *maaş= “-10”*
  - *tutarsız: nitelik değerleri veya nitelik isimleri uyumsuz*
    - *yaş= “35”, d.tarihi: “03/10/2004”*
    - *önceki oylama değerleri: “1,2,3”, yeni oylama değerleri: “A,B,C”*
    - *bir kaynakta nitelik değeri ‘ad’, diğerinde ‘isim*

Verinin Gürültülü Olma Nedenleri

- Eksik veri kayıtlarının nedenleri
  - *Veri toplandığı sırada bir nitelik değerinin elde edilememesi, bilinmemesi*
  - *Veri toplandığı sırada bazı niteliklerin gerekliliğinin görülememesi*
  - *İnsan, yazılım ya da donanım problemleri*
- Gürültülü (hatalı) veri kayıtlarının nedenleri
  - *Hatalı veri toplama gereçleri*
  - *İnsan, yazılım ya da donanım problemleri*
  - *Veri iletimi sırasında problemler*
- Tutarsız veri kayıtlarının nedenleri
  - *Verinin farklı veri kaynaklarında tutulması*
  - *İşlevsel bağımlılık kurallarına uyulmaması*

Sonuç

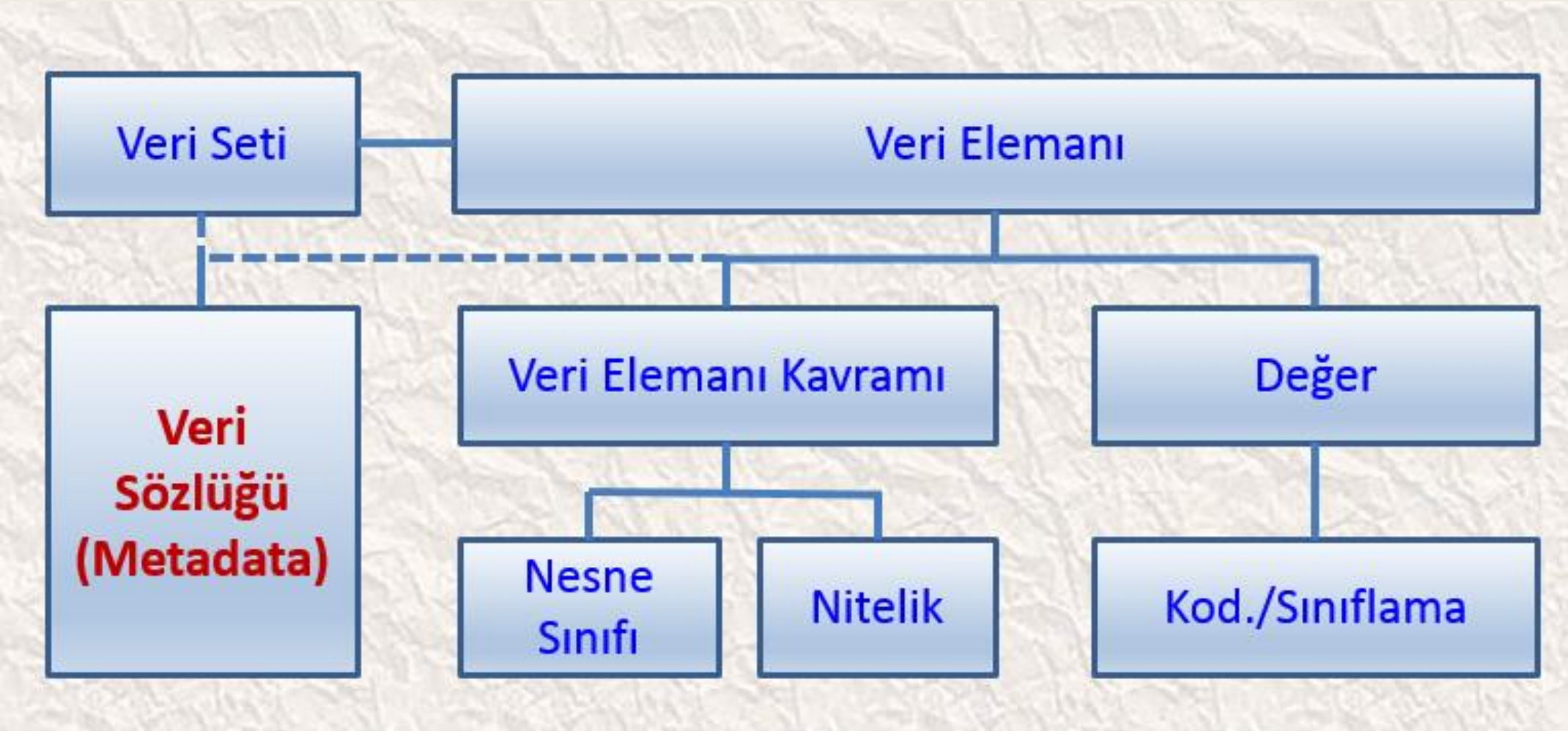
- Veri güvenilmez
  - *Veri madenciliği sonuçlarına güvenilebilir mi?*
  - *Kullanılabilir veri madenciliği sonuçları kaliteli veri ile elde edilebilir.*
- Veriniz eğer kaliteli ise Veri Madenciliği uygulamalarından beklenen bilgi daha yararlıdır.

Veri Önişleme

- Veri temizleme
  - *Eksik nitelik değerlerini tamamlama, hatalı veriyi düzeltme, aykırılıkları saptama ve temizleme, tutarsızlıkları giderme*
- Veri birleştirme
  - *Farklı veri kaynağındaki verileri birleştirme*
- Veri dönüşümü
  - *Normalizasyon ve biriktirme*
- Veri azaltma
  - *Sonuçları elde edilecek şekilde veri miktarını azaltma*

Veriyi Tanıma

Veriyi Tanıma



Veriyi Tanımlayıcı Özellikler

- Amaç: Veriyi daha iyi anlamak
  - *Merkezi eğilim (central tendency), varyasyon, yayılma, dağılım.*
- Verinin dağılım özellikleri
  - *Ortanca, en büyük, en küçük, sıklık derecesi.*
- Sayısal nitelikler -> sıralanabilir değerler
  - *verinin dağılımı*
  - *kutu grafiği çizimi ve sıklık derecesi incelemesi*

Merkezi Eğilimi Ölçme

- Ortalama
  - *Genel olarak ağırlıklı ortalama*
- Ortanca (Median)
  - *Veri sayısına göre; veri tek ise orta değer çift ise ortadaki iki değer.*
- Mod (Frekans)
  - *Veri içinde en sık görülen değer*

- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma

Veri Temizleme

- Gerçek uygulamalarda veri eksik, gürültülü veya tutarsız olabilir.
- Veri temizleme işlemleri
  - *Eksik nitelik değerlerini tamamlama*
  - *Aykırılıkların bulunması ve gürültülü verinin düzeltilmesi*
  - *Tutarsızlıkların giderilmesi*

Eksik Veri

- Eksik nitelik değerleri olan veri kayıtlarını kullanma
- Eksik nitelik değerlerini elle doldur
  - *Eksik nitelik değerleri için global bir değişken kullan (Null, bilinmiyor,…)*
  - *Eksik nitelik değerlerini o niteliğin ortalama değeri ile doldur*
  - *Aynı sınıfa ait kayıtların nitelik değerlerinin ortalaması ile doldur*
  - *Olasılığı en fazla olan nitelik değeriyle doldur*

Gürültülü Veri

- Ölçülen bir değerdeki hata
- Yanlış nitelik değerleri
  - *hatalı veri toplama gereçleri*
  - *veri girişi problemleri*
  - *veri iletimi problemleri*
  - *teknolojik kısıtlar*
  - *nitelik isimlerinde tutarsızlık*

Gürültülü Veri nasıl düzeltilir?

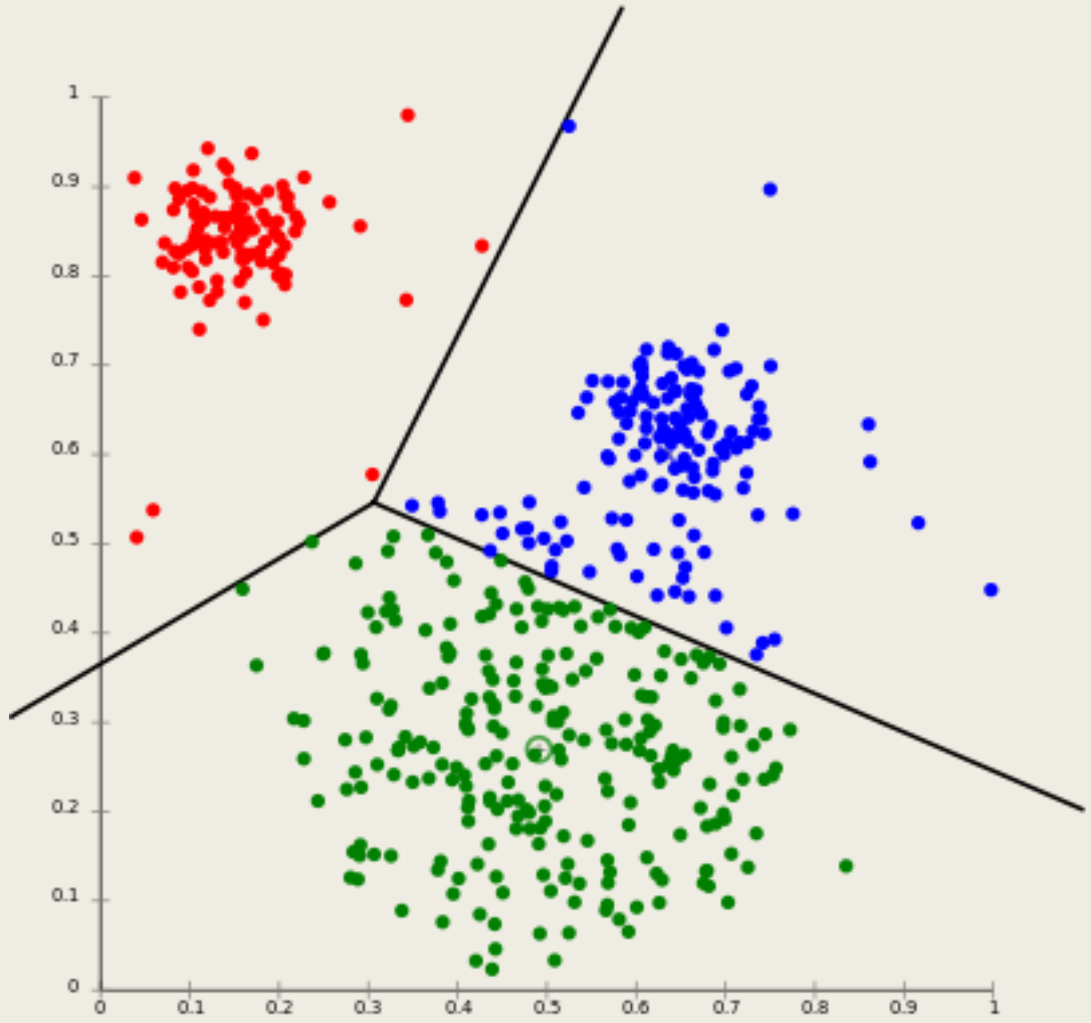
- Gürültüyü yok etme
  - *Bölmeleme*
    - *veri sıralanır, eşit genişlik veya eşit derinlik ile bölünür*
  - *Kümeleme*
    - *aykırılıkları belirler*
  - *Eğri uydurma*
    - *veriyi bir fonksiyona uydurarak gürültüyü düzeltir.*

Bölmeleme

- Veri sıralanır: 4, 8, 15, 21, 21, 24, 25, 28, 34
  - *Eşit genişlik: Bölme sayısı belirlenir. Eşit aralıklarla bölünür*
  - *Eşit derinlik: Her bölmede eşit sayıda örnek kalacak şekilde bölünür.*
    - *her bölme ortalamayla ya da bölmenin en alt ve üst sınırlarıyla temsil edilir.*

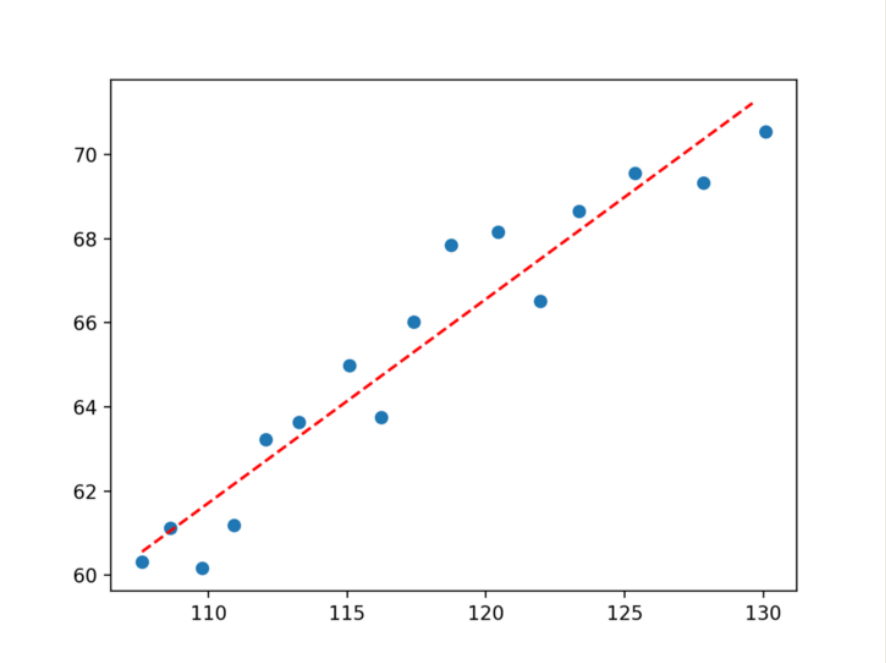
Kümeleme

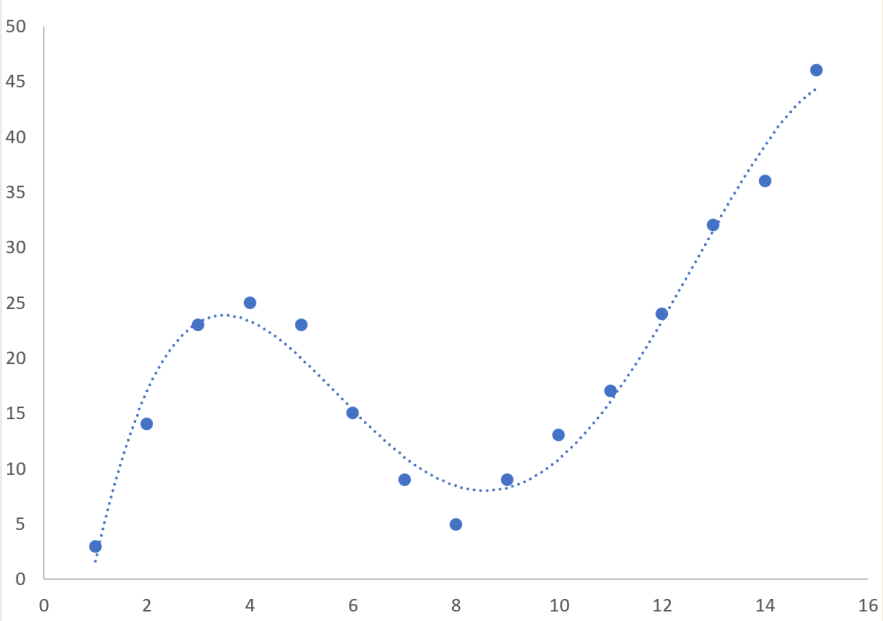
- Benzer veriler aynı kümede olacak şekilde gruplanır
- Bu kümelerin dışında kalan veriler aykırılık olarak belirlenir ve silinir.



Eğri Uydurma

- Veri bir fonksiyona uydurulur. Doğrusal eğri uydurmada, bir değişkenin değeri diğer bir değişken kullanılarak bulunabilir.





- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma

Veri Birleştirme

- Farklı kaynaklardan verilerin tutarlı olarak birleştirilmesi
- Şema birleştirilmesi
  - *Aynı varlıkların saptanması*
  - *meta veri kullanılır*
- Nitelik değerlerinin tutarsızlığının saptanması
  - *Aynı nitelik için farklı kaynaklarda farklı değerler olması*
  - *Farklı metrikler kullanılması*

Gereksiz Veri

- Bu veriler genel olarak farklı veri kaynaklarından verilerin birleşimi ile oluşan gereksiz verileri içerir.

- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma

Veri Dönüşümü

- Veri, veri madenciliği uygulamaları için uygun olmayabilir
- Seçilen algoritmaya uygun olmayabilir
  - *Veri belirleyici değil*
  - *Çözüm*
  - *Veri düzeltme*
    - *Bölmeleme*
    - *Kümeleme*
    - *Eğri Uydurma*
  - *Biriktirme*
  - *Genelleme*
  - *Normalizasyon*
  - *Nitelik oluşturma*

Normalizasyon

- min-max normalizasyon
- z-score normalizasyon
- ondalık normalizasyon

Nitelik Oluşturma

- Yeni nitelikler oluşturma
  - *orjinal niteliklerden daha önemli bilgi içersin*
    - *alan=boy x en*
  - *veri madenciliği algoritmalarının başarımı daha iyi olsun*

- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma

Veri Azaltma

- Veri miktarı çok fazla olduğu zaman veri madenciliği algoritmalarının çalışması ve sonuç üretmesi çok uzun sürebilir
  - *veriyi azaltma başarımı artırır*
  - *sonucun (nerdeyse) hiç değişmemesi gerekir*
- Veri azaltma
  - *nitelik birleştirme*
  - *nitelik azaltma*
  - *veri sıkıştırma*
  - *veri ayrıştırma ve kavram oluşturma*
  - *veri küçültme*
    - eğri uydurma • kümeleme • histogram • örnekleme

Veri Sıkıştırma

- Verinin boyutunu azaltır
  - *daha az saklama ortamı*
  - *veriye ulaşmak daha çabuk*
- Kayıplı ve kayıpsız veri sıkıştırma
  - *bazı yöntemler bazı veri tiplerine uygun*
  - *her veri tipi için kullanılan yöntemler de var*
- Eğer veri madenciliği yöntemi sıkıştırılmış veri üzerinde doğrudan çalışabiliyorsa elverişli

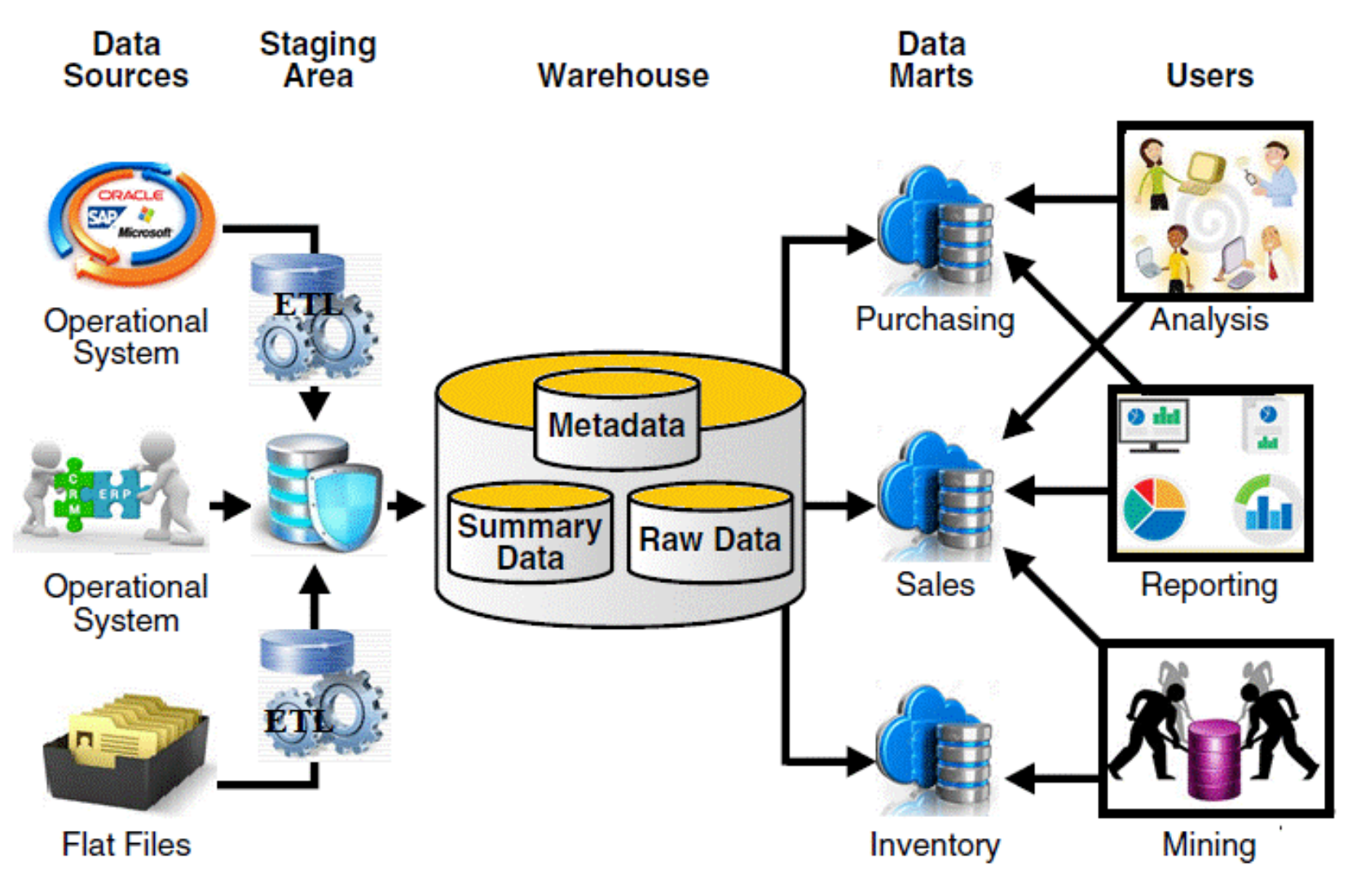
İçerik

- Veri Ambarı Nedir?
- Çok boyutlu veri modeli
- Veri ambarı mimarisi
- Veri ambarı uygulaması
- Veri ambarından veri madenciliğine

Veri Ambarı

- Veritabanı: Birbirleriyle ilişkili bilgilerin depolandığı alanlardır.
- Veri Ambarı: İlişkili verilerin sorgulandığı ve analizlerinin yapılabildiği bir depodur. Veri ambarı veritabanını yormamak için oluşturulmuştur. Bir veri ambarı ilgili veriyi kolay, hızlı, ve doğru biçimde analiz etmek için gerekli işlemleri yerine getirir. Veri ambarı, işlemsel sistemlerdeki veriyi kopyalayıp, karar verme işlemi için uygun formda saklar.

Veri Ambarı


Data Sources
Staging Area
Warehouse
Data Marts
Users
Operational System
ETL
Operational System
Metadata
Summary Data
Raw Data
Purchasing
Analysis
Sales
Reporting
Flat Files
ETL
Inventory
Mining

Veri Ambarı

- Data Mart: Veri ambarlarının alt kümeleridir. Veri ambarları bir iş probleminin tamamına yönelik bir bakış sağlarken, data mart'lar sadece belli bir kısma bakış sağlarlar. Veri pazarları ile veriye hızlı erişim sağlayabiliriz. İkinci olarak, verinin gruplanmamış yapıda olması ve farklı iş birimlerinin farklı verileri görmesidir. Bu da bize gereksiz bir iş yükü ve güvenlik sorununa neden olmaktadır. İşte tam bu noktada, veri pazarları konuya, bölümlere uygun, veri ambarının küçük bir kopyası halinde çözüm sunmaktadır.

Veri Ambarı

- Amaca yönelik
- Birleştirilmiş
- Zaman değişkenli
- Değişken değil