* + VERİ BİLİMCİ : Bir Kodun/Sistemin Arka Planda Nasıl Çalıştığını Bilmeli

Katma Değer Üreten Çözümler , çözümleri verimli bir şekilde uygulayabilir, yazılım mühendisliği, istatistik / makine öğrenimi ve uygulama alanında yeterli bilgiye sahip bir **uygulayıcıdır**.

Veri nedir? Bilgi hammadesi alt düzeyi daha değerli üst düzeye transfer etme çabası

**Veri** (Data): ham veri, kod, imge, ses,

**Enformasyon** (information): Organize edilmiş, yapısal hale getirilmiş, yorumlanmış, özetlenmiş veri,

**Bilgi (knowledge):** durum (case), kural (rule), süreç, model,

**Uzmanlık (expertise)**: Hızlı ve doğru öneri, açıklama, sonucu doğrulama (justification of result) ve mantığa bürüme (reasoning),

**Yeterlilik (capability):** Örgütlenmiş uzmanlık, bilgi dağarcığı, entegre edilmiş destek sistemi, merkezi yeterlilik.

* Veri bilginin ham hali anlamına geldiği için işlenmesi ve anlamlandırılması gerekir.
* Veriyi tanımlayan veriye ise ‘**meta veri**’ denir.

Veri tabanları:

* Mysql: Çok iyi şekilde özelleştirilmektedir, pek çok veri tabanının kolay bir şekilde yönetilmesini sağlar.
* Oracle : Büyük miktarda veriyi çok kullanıcı tarafından yönetilen bir ortamda, depolanmasına ve erişilmesini sağlar. Kurumsalda çok yaygın, kolay ve esnek, Büyük verileri hızlı bir şekilde işlem boyutunun yüksekliğine bakmadan işleyebilme özelliğine sahiptir.
* PostgreSQL: sql sorgu dilini direkt destekler, Güvenli, performansı yüksek, geniş bir veri tabanı sistemi, Oldukça güvenli ve hatasız, açık kaynak kodlu,
* SQLLİTE : Açık kaynak kodu, C C++,ise sunucu yapısı ve yapılandırma gereksinimi olmamasıdır., Dosya yapısı ile çalışır, hızlı

Büyük veri: Çok büyük miktarda verinin yaratılması ve kullanılabilmesinden dolayı ortaya çıkmıştır.nicelik ve kalitesine bakılır. Miktarı büyük veriler Amaç için toplanmayan verilerdir büyük veriyle karıştırma.

Büyük Veri: Bir veri amaç ve hedefler doğrultusunda toplanır

Miktarı Büyük Veri: Toplanmış büyük miktarda veri

5V Prensibi

**Variety (Çeşitlilik):** Üretilen veriler genel olarak yapısal olmadığı ve bir çok farklı ortamdan elde edilen veri olduğu için bütünleşik ve birbirlerine dönüştürülebiliyor olmaları gerekmektedir.

**Velocity (Hız):** Big data üretimi ile hızla veriler inanılmaz boyutlara ulaşmaktadır. Hızlı büyüyen veri, o veriye muhtaç olan işlem sayısının ve çeşitliliğinin de aynı hızda artması sonucunu ortaya çıkartmaktadır ve hem yazılımsal hem de donanımsal olarak yoğun

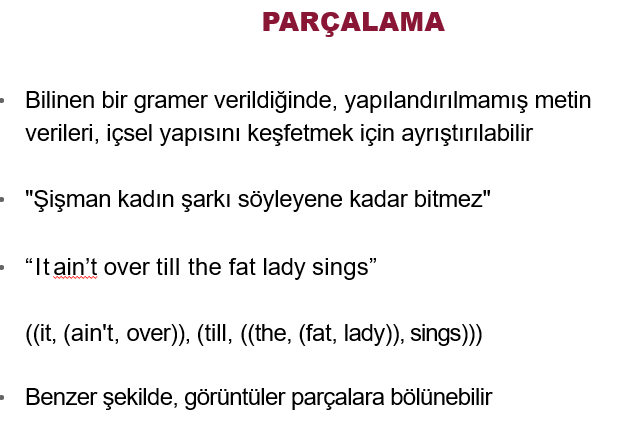
**Volume (Veri Büyüklüğü):**  ileride bu veri yığınları ile nasıl başa çıkacağımızı iyi düşünmemiz ve planlarımızı bu doğrultuda yapmamız gerekmektedir.

**Verification (Doğrulama):** verilerin güvenli olup olmadığını kontrol etmemiz gerektiği durumlar var Bu veri doğru kişiler tarafından görülebilir veya saklı kalması gerekiyor olabilir.

**Value (Değer):**  büyük verinin üretimi ve işlenmesi katmanlarında elde edilen verilerin şirketimiz için artı değer sağlıyor olması gerekiyor.

YAPILANDIRILMIŞ VERİLER :listeler ,n × m dizileri ,hiyerarşiler, (ör. organizasyon şeması),Ağlar,(ör. seyahat rotaları, hypertext = bağlantılar) ,Genel veri değişim biçimleri:XML, JSON

YAPILANDIRILMAMIŞ VERİLER :Metin, Görüntüler, Video, Ses, Genellikle, ör. ayrıştırma dili, görüntüleri bölümlere ayırma vb. kullanılarak yapılandırılabilir.



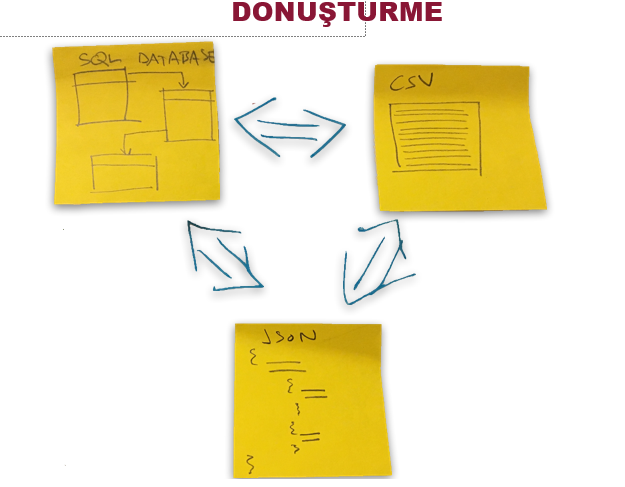
VERİ TÜRLERİ:

Veri çerçeveleri, veri dizileri,

diziler: Diziler Oluşturma, Yeniden Şekillendirme ,Dilimleme, Sayısal İşlemler

VERİ DÖNÜŞTÜRME:

Dönüştürme:



Diğer tür dönüştürme:

İçerik dönüştürme: içerik dönüşümleri, veri parçalama, vs

FİLTRELEME:

* Veri bölmesi: altkümeleme, sütun satır,
* Eksik değerler, birçok analiz yöntemi için önemli olabilir.
* Basit bir yol ile eksik girdileri olan tüm kayıtları filtrelemek iyi olabilir. Birçok veri kaybı oluşur.
* eksik girişlerin yerine "sahte" verileri girmek: ortalama, mod, makine öğrenmesi
* Veriler "rastgele tamamen eksikse", eksik öğeler içeren satırları silmek güvenlidir, tabii ki veri miktarı azalacaktır.
* Eksiklik özel bir duruma bağlıysa, satırları silmek sonuçları saptırabilir.

GÖSTERİMz

Çeşitli gösterim türleri vardır:

* + Grafiksel Olmayan vs Grafiksel
  + Tek değişkenli, İki değişkenli ve Çok değişkenli

Hepsini Kullanabilmeliyiz

Temel hedefler:

* Hatalı verileri kaldırmak veya düzeltmek
* İstediğimiz hipotezleri formüle edebilmek
* Uygun analiz yöntemlerini seçimimetin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

  Açıklama otomatik olarak oluşturuldu

Şüpheli verilerin araştırılması ve düzeltilmesi / kaldırılması gerekir

Kötü/Hatalı/Şüpheli veriler yanıltıcı olabilir

bilinçli bir analiz yapmak gerekiyor

İstatiksel özet: sürekli: ortalama, medyan, min, maks kategorik: değerler kümesi, mod

metin, diyagram, ekran görüntüsü, paralel içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, çiçek içeren bir resim

Açıklama otomatik olarak oluşturuldu

renk ve boyut kazandırılarak daha fazla değişken eklenebilir

TUFTE, 1983 : Temel 2 Kuralı

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, yazı tipi, ekran görüntüsü, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu

Acı örnek

D1-D2

Veri Madenciliği

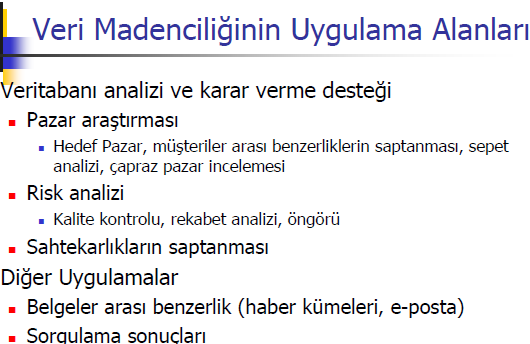
veri madenciliği bilgi keşfi işleminin aşamalarından biridir. veriyi belli bir modele

uydurur. Sorgulama ya da basit istatistik yöntemler veri madenciliği değildir.

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bilgi Keşfinin Aşamaları: Uygulama alanını inceleme, Amaca uygun veri kümesi oluşturma, Veri ayıklama ve önişleme, Veri azaltma ve veri dönüşümü, Veri madenciliği tekniği seçme, Veri madenciliği algoritmasını seçme, Model değerlendirme ve bilgi sunumu, Bulunan bilginin yorumlanması

Veri madenciliği amacı : aradığımız veri mevcutsa sonuçlarını anlamak metin, web sitesi, web sayfası, çevrimiçi reklamcılık içeren bir resim

Açıklama otomatik olarak oluşturuldu

Modelleri:

􀂄 Sınıflandırma (Classification): Veriyi öncedenbelirlenmiş sınıflardan birine dahil eder.

􀂄 Eğri uydurma (Regression): Veriyi gerçel değerli bir fonksiyona dönüştürür.

􀂄 Zaman serileri inceleme (Time Series Analysis): Zaman içinde değişen verinin değerini öngörür.

􀂄 Aykırılıkları Belirleme (Outlier Analysis): Verinin geneline uymayan nesneleri belirleme

􀂄 Demetleme (Clustering): Benzer verileri aynı grupta toplama

􀂄 Özetleme (Summarization): Veriyi altgruplara ayırır. Her altgrubu temsil edecek özellikler bulur.

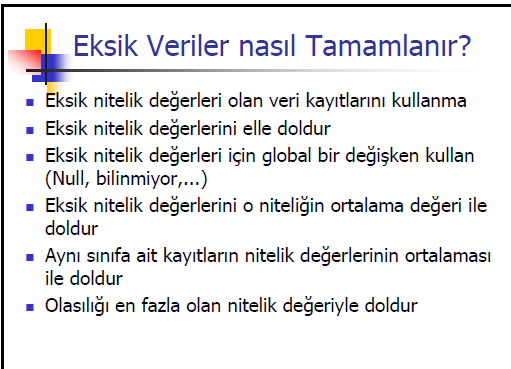
􀂄 İlişkilendirme kuralları (Association Rules)

􀂄 Sıralı dizileri bulma (Sequence Discovery): Veri içinde sıralı örüntüler bulmak için kullanılır.

Veri Ambarı: Amaca yönelik, Birleştirilmiş, Zaman değişkenli, Değişken değil

Verinin Gürültülü Olma Nedenleri: Eksik veri kayıtları, Gürültülü (hatalı) veri kayıtları, Tutarsız veri kayıtları,

Veri önişleme: veri temizleme, birleştirme, dönüşüm, azaltma

 daire, ekran görüntüsü, sanat içeren bir resim

Açıklama otomatik olarak oluşturuldu

Topluluklar nasıl bulunur:

Minimum kesim yöntemi

Hiyerarşik kümeleme

Girvan-Newman algoritması: toplulukların "arasında" olma olasılığı en yüksek olan kenarlara odaklanır. Bir düğüm çifti arasında birden fazla en kısa yol varsa, tüm yolların toplam ağırlığı birliğe eşit olacak şekilde her yola eşit ağırlık atanır.

Modülerlik maksimizasyonu

İstatiksel sonuç

Klika dayalı yöntemler

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu