

VERİ MADENCİLİĞİ

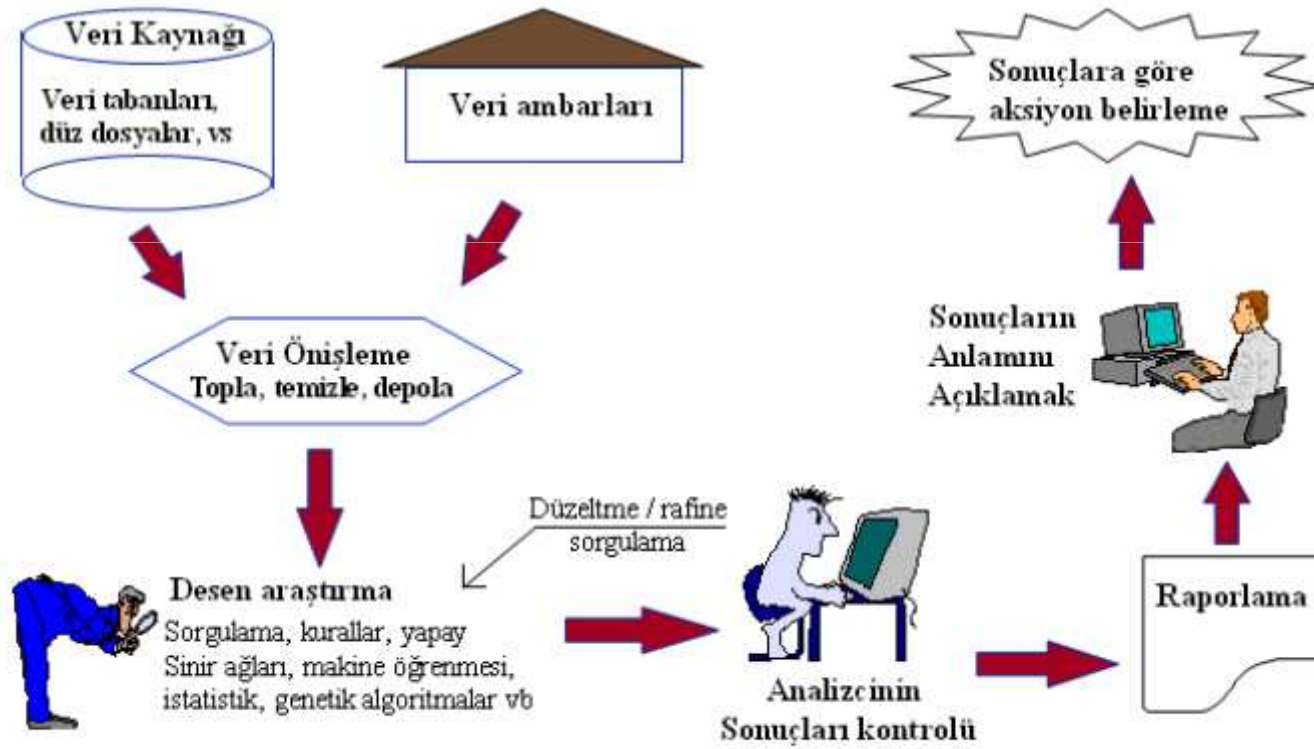
VERİ TABANLARINDA BİLGİ KEŞFİ

Veri Tabanlarında Bilgi Keşfi, veriden faydalı bilginin keşfedilmesi sürecinin tamamına atıfta bulunmakta ve veri madenciliği bu sürecin bir adımına karşılık gelmektedir.

VERİ TABANLARINDA BİLGİ KEŞFİ

VTBK, verinin nasıl depolanıp erişileceğinden, algoritmaların devasa veri setlerine nasıl ölçeklenebileceğine ve hala etkin olarak çalışmalarına, sonuçların nasıl yorumlanabileceği ve görselleştirilebileceğine ve bütün insan-makine interaksiyonunun kullanışlı olarak nasıl modellenip, desteklenebileceğine olmak üzere veriden bilginin keşfinin tüm süreçleri üzerine odaklanır

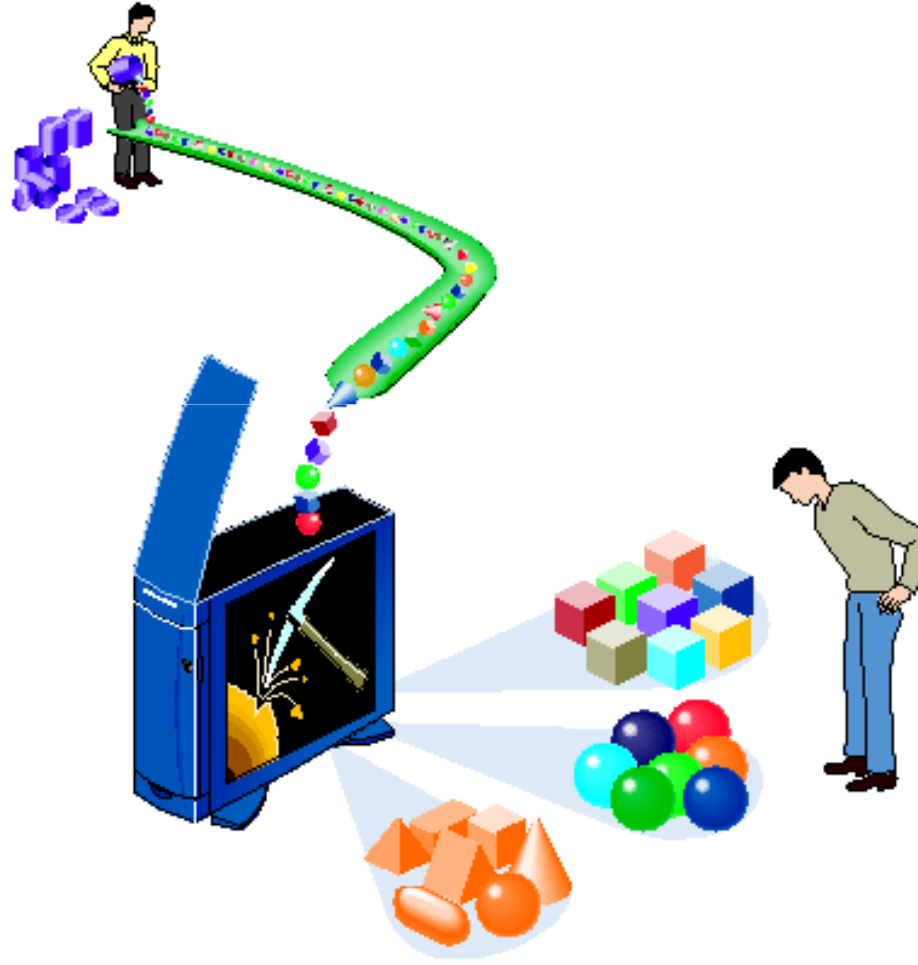
VERİ TABANLARINDA BİLGİ KEŞFİ



BILGI KEŞFİNİN AŞAMALARI

- Uygulama alanını inceleme : Konuyla ilgili bilgi ve uygulama amaçları
- Amaca uygun veri kümesi oluşturma: Veri seçme
- Veri ayıklama ve ön işleme : işlemin %70'lik bölümünü oluşturur
- Veri azaltma ve veri dönüşümü : incelemede gerekli boyutları (özellikleri) seçme, boyutlar arası ilişkiyi belirleme, boyut azaltma,
- Veri madenciliği tekniği seçme
- Sınıflandırma, eğri uydurma, bağıntı kuralları, demetleme
- Veri madenciliği algoritmasını seçme
- Model değerlendirme ve bilgi sunumu
- Bulunan bilginin yorumlanması

VERİ MADENCİLİĞİ NEDİR?



VERİ MADENCİLİĞİ NEDİR?

Net bir tanımı olmamakla birlikte var olan tanımlar da çok büyük farklılık içermiyor.

- Aktif araştırma alanlarından biri olan veri tabanlarında bilgi keşfi (vtbk) disiplini, çok büyük hacimli verileri tam ya da yarı otomatik bir biçimde analiz eden yeni kuşak araç ve tekniklerin üretilmesi ile ilgilenen son yılların gözde araştırma konularından biridir. Veri madenciliği, önceden bilinmeyen, veri içinde gizli, anlamlı ve yararlı örüntülerin büyük ölçekli veritabanlarından otomatik biçimde elde edilmesini sağlayan vtbk süreci içinde bir adımdır.
- Veri madenciliği büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanarak aranmasıdır.
- Veri madenciliği, büyük hacimli veri yığınları içerisinde karar alabilmek için potansiyel olarak faydalı olabilecek uygulanabilir ve anlamlı bilgilerin çıkarılmasına verilen addır. Veri madenciliği geniş anlamda veri analiz teknikleri bütünüdür ve tek başına bir çözüm değildir. Mevcut problemleri çözmek, kritik kararları almak veya geleceğe yönelik tahminleri yapmak için gerekli olan bilgileri elde etmeye yarayan bir araçtır. Ortaya çıkarılması hedeflenen bilgiler; üstü kapalı çok net olmayan önceden bilinmeyen, daha önce keşfedilmemiş ancak potansiyel olarak kullanışlı anlamlı ve kritik bilgilerdir.
- Varlığıyla bulunduğu yerde bir anlam ifade eden verilerin bir disiplin altında toplanarak belirli bir teknik veya teknikler kullanılarak işlenmesiyle artıklarından arındırılması sonucunda anlamlı bilgileri gün yüzüne çıkaran ve geleceğe yön verebilmemizi sağlayan ve bu sonuçlara matematiksel ve istatistiksel yöntemlerle mutlaklık kazandıran bir süreçtir

VERİ MADENCİLİĞİ NEDİR?

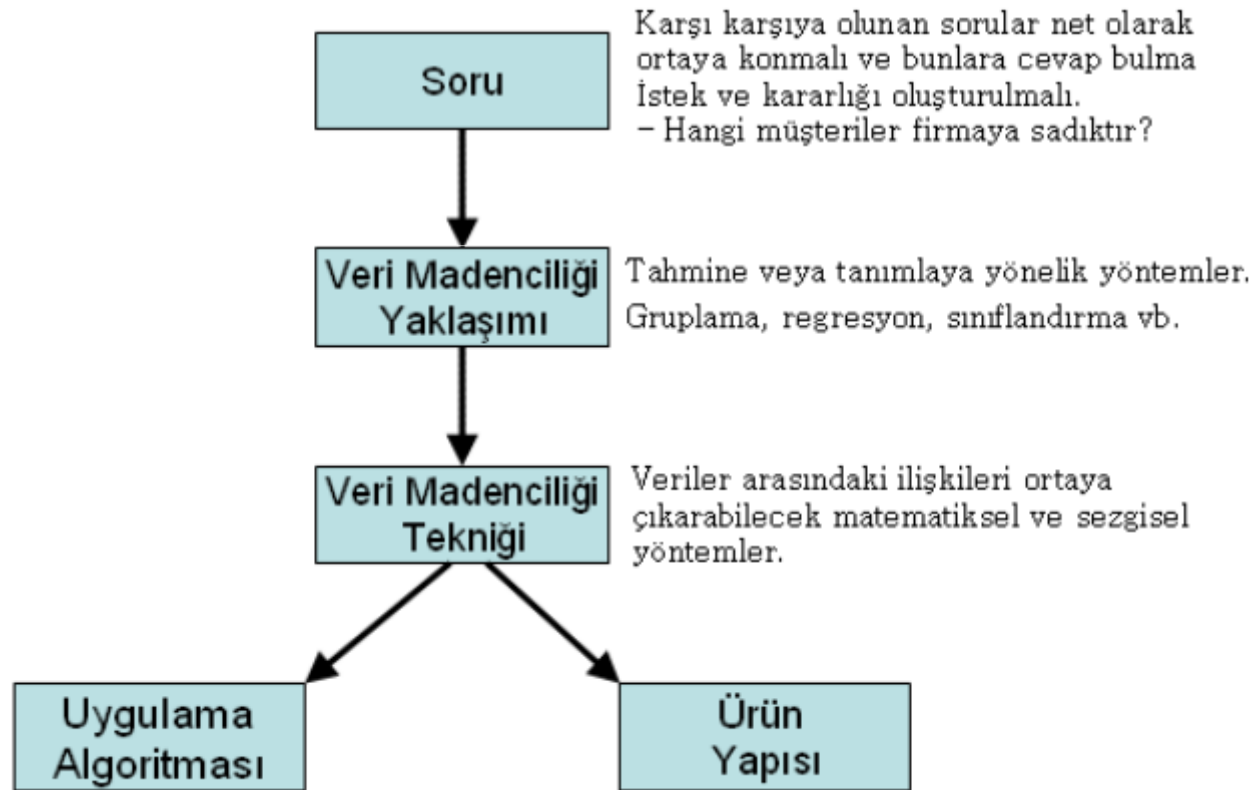
- En temel ifadesiyle büyük veri kümelerinde bulunan ve insan algısının çok uzun zaman ve klasik yöntemlerle çok fazla çaba gerektirerek fark edebileceği faydalı veri paternlerinin bu kümelerden ayıklanmasıdır.
- Veri madenciliği, pek çok analiz aracı kullanımıyla veri içerisinde örüntü ve ilişkileri keşfederek, bunları geçerli tahminler yapmak için kullanan bir süreçtir.
- Veri madenciliği, en basit tanımıyla, veri tabanlarındaki ilişkili örüntüleri OTOMATİK olarak belirlemektir.
- Veri madenciliği veri kümesi içerisinde keşfedilmemiş örüntüleri bulmayı hedefleyen tekniklerdir.
- En temel özelliklerinden biri araştırmacının analizi yönlendirmesinden ziyade verinin analizi ve analisti yönlendirmesidir.

VERİ MADENCİLİĞİ NEDİR?

Veri madenciliği henüz gelişmekte olan bir alan olduğu için pek çok farklı tanım yapmak mümkündür.

Buna karşın, Veri madenciliğinin tek hedefi düşük düzeyde enformasyon sağlayan veriden, yüksek düzeyde kıymetli bilgiyi açığa çıkarmaktır.

VERİ MADENCİLİĞİN YAPISI



VERİ MADENCİLİĞİ CEVAPLAYABİLECEĞİ SORULAR NELERDİR?

VERİ MADENCİLİĞİ CEVAPLAYABİLECEĞİ SORULAR

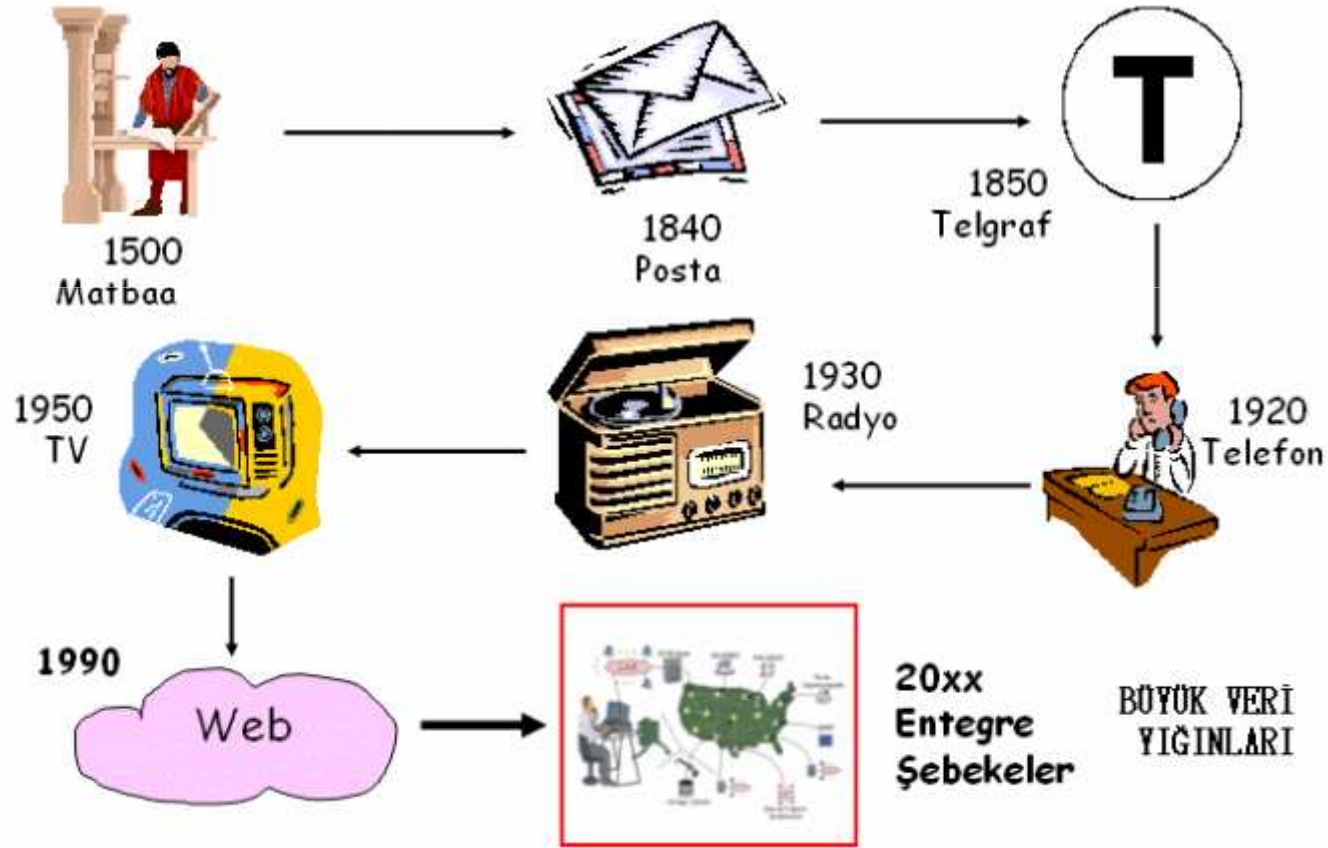
- Bilinmeyen bilgiyi ortaya çıkarmak için çalışır
- Bu sene potansiyel en iyi 10 müşterimiz kimler olacaktır sorusuna cevap arar
- Gelecek 5 sene şirketimiz hangi alanda büyümeli/büyüyebilir
- Bu hastanın hastalığı ne olabilir
- Kanseri erken teşhis edebilir miyim

VERİ MADENCİLİĞİNİN KULLANIM ALANLARI

- Finans Sektörü
- Haberleşme Sektörü
- Sağlık Sektörü
- Devlet Uygulamaları

Büyük hacimde veri olan her yerde veri madenciliği kullanılabilir.

VERİ MADENCILIĞININ TARİHÇESİ



VERİ MADENCİLİĞİNİN TARİHÇESİ

1950'ler

İlk bilgisayarlar

(Sayımlar için bilgisayarlar kullanılıyor)

1960'lar

Veri koleksiyonları

Veritabanı yaratımı

(Hiyerarşik ve ağ modelleri)

VERİ MADENCİLİĞİNİN TARİHÇESİ

1970'ler

İlişkisel veri modeli
İlişkisel VTYS uygulamaları

1980'ler

İlişkisel VTYS yaygınlaşıyor
Uygulamaya yönelik VTYS
(Mekansal, Bilimsel, Mühendislik,
vs.)

VERİ MADENCİLİĞİNİN TARİHÇESİ

1990'lar

Günlük işlemlerden derlenen büyük miktarda verinin nasıl değerlendirilebileceği sorgulanmaya başlıyor.

Bu noktada bahsedilmesi gereken birkaç önemli olay sözkonusu:

1989, KDD (IJCAI)-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısı

1991, KDD (IJCAI)-89'un sonuç bildirgesi sayılabilecek

'Knowledge Discovery in Real Databases:

A Report on the IJCAI-89 Workshop'

makalenin KDD ile ilgili temel tanım ve kavramları ortaya koyması

VERİ MADENCİLİĞİNİN TARİHÇESİ

1992, Veri Madenciliği konusunda ilk yazılımın geliştirilmesi

1995, 1. Uluslararası Bilgi Keşfi ve Veri Madenciliği Konferansı'nın (KDD-95) açılış konuşması

2000'ler

Veri Ambarları,

Veri Madenciliği yaygınlaşıyor. (Son 10 yılda veri depolama ünitelerinin fiyatlarında sürekli bir düşüş var.)

VERİ AMBARININ TANIMI

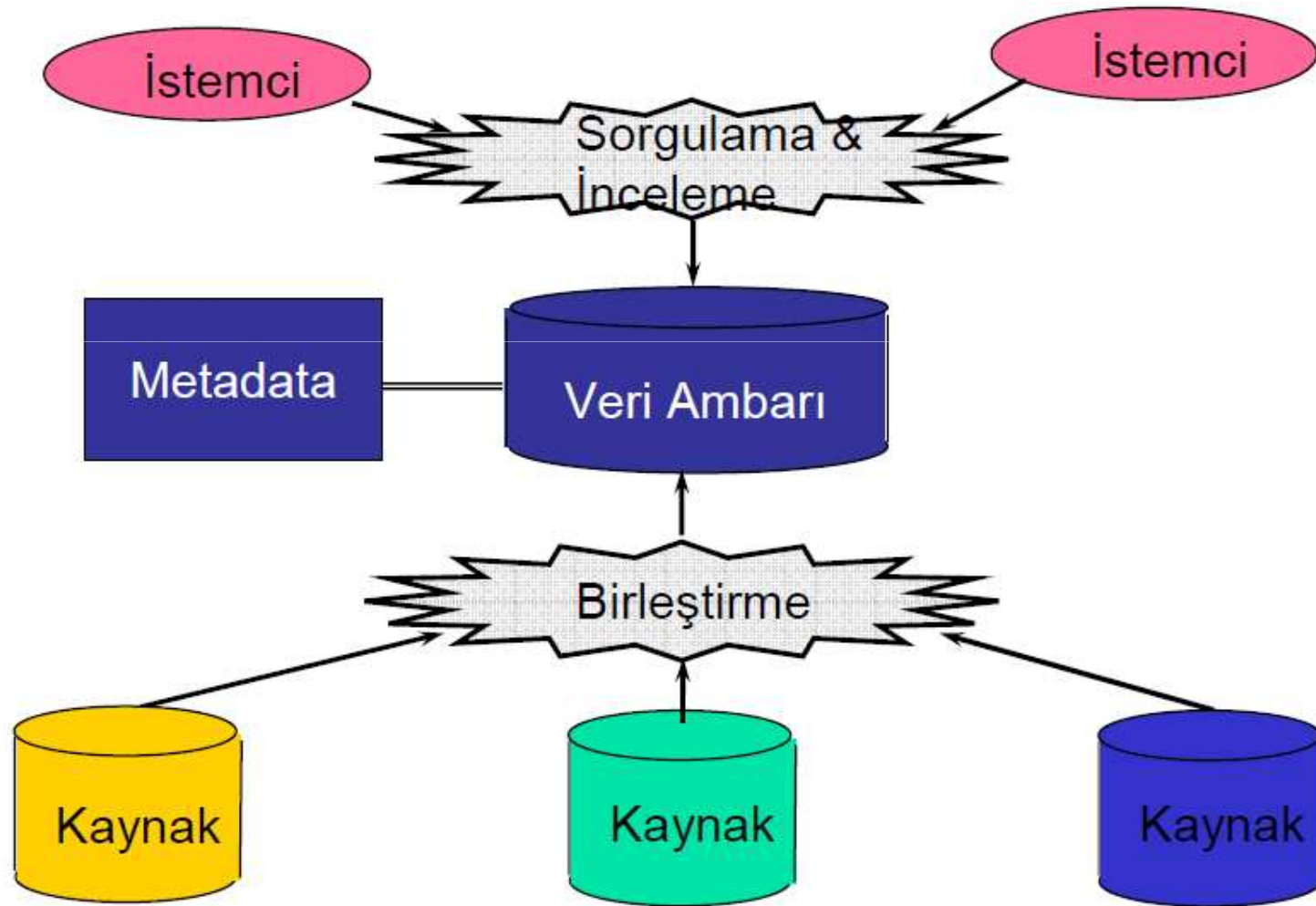
- Veri Ambarları, Veri Madenciliği ile eşanlı olarak anılan ve Veri Madenciliği sürecinin gerçekleştirildiği veriyi sağlayan özel bir veri tabanıdır.
- Tanım olarak Veri Ambarı, pekçok farklı kaynaktan ve genellikle de farklı yapıda verinin depolandığı ve hepsinin de aynı birleşik çatı altında kullanılmasının ümit edildiği yapılardır.
- Ayrıca, Veri Ambarı pek çok farklı kaynaktan elde edilen, farklı formatta veriyi aynı çatı altında analiz etme imkanı tanır.
- Veri ambarcılığı ise veri ambarını kurma ve kullanma sürecine denmektedir

Bir veri ambarı çoklu kaynaklardan (farklı veritabanlarından) gelen dataların toplanmasıyla oluşur. Buradan gelen datalar önce bazı aşamalardan geçer:

- data cleaning
- data integration
- data transformation
- data loading
- data refreshing (periyodik olarak)

Bir çok yerde şubesi olan bir şirketiniz olduğunu düşünün. Bu şirketin İstanbul, İzmir, Ankara şubelerinin veritabanları yukarıdaki aşamalardan geçer ve veri ambarı oluşturulur.

VERİ AMBARI MİMARİSİ



VERİ AMBARININ ÖZELLİKLERİ

Kurumun operasyonel veri tabanından ayrı olarak yapılandırılmış bir karar destek veritabanıdır.

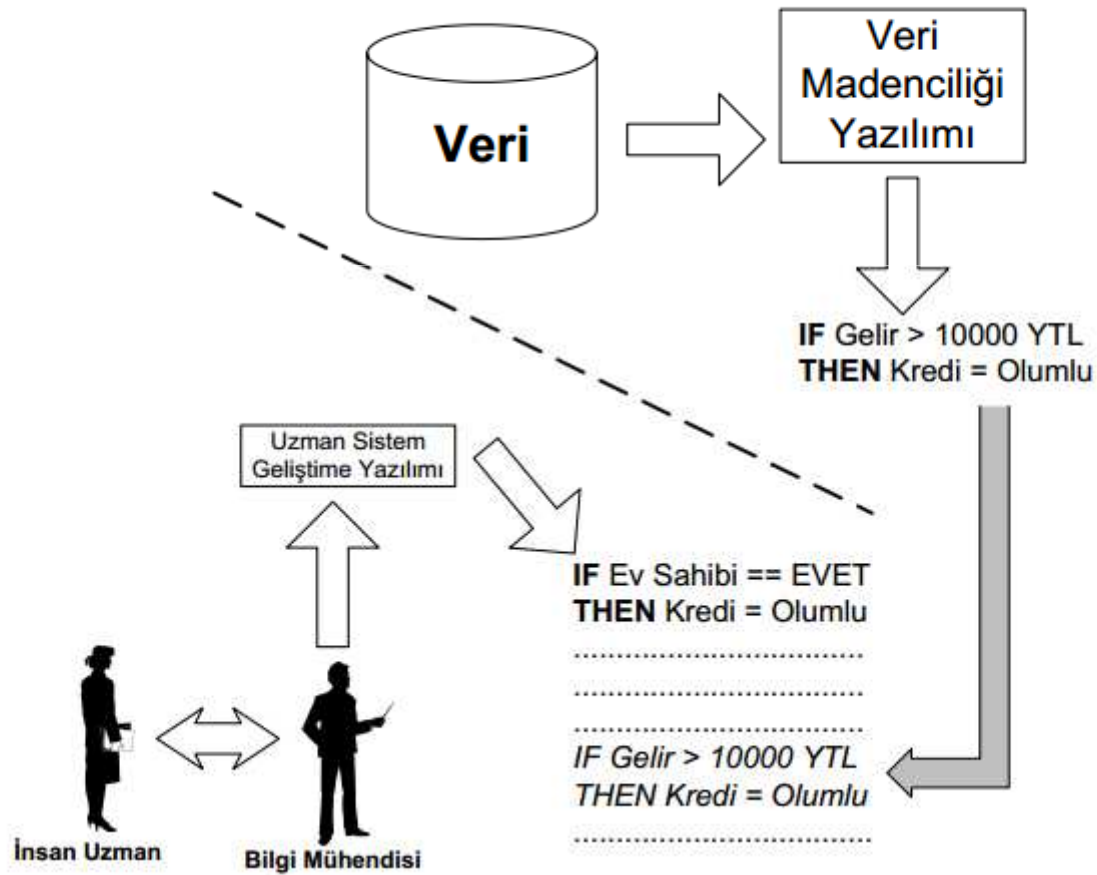
Konsolide edilmiş, tarihi verileri sağlayarak bilgi işlemeyi destekler.

Bir veri ambarı, yöneticilerin karar verme sürecine destek olmak üzere verinin

- Konuya yönelik
- Entegre
- Zamana bağlı değişen (zamana endeksli)
- Değişmez

halini sunmaktadır.

VERİ MADENCİLİĞİ ÖRNEK



SONUÇ

- Bir veri madenciliği çalışması için öncelikle çok miktarda kaliteli veri gerekir.
- Amaç bu veri içinde saklı, gelecekle ilgili tahmin yapmakta kullanılabilecek kural ve bağıntıların çıkarılmasıdır.
- Böyle bir çalışmanın başarılı olması için uygulama konusundaki uzmanların veri tabanları ve veri madenciliği konusundaki uzmanlarla beraber çalışması gerekir.
- Çalışma uzun sürebilir; zaman ve sabır gerekir.