Teoride veri madenciliği bilgi keşfi işleminin aşamalarından biridir. Pratikte veri madenciliği ve bilgi keşfi eş anlamlı olarak kullanılır.

Veri madenciliği kendi başına oldukça büyük bir disiplin olmasına rağmen aslında veri tabanlarında bilgi keşfi adı verilen sürecin bir parçasıdır. Veri tabanından bilgi keşfi veriden yararlı bilgi çıkarma sürecidir. Veri tabanında bilgi keşfi, veri madenciliği ile paralel sürece sahiptir. Çeşitli şekillerde depolanan verileri bu süreçte veri madenciliği yaparak, bilgi yığınlarını anlamlandırmaya çalışır. Geleneksel sorgu ve raporlama araçlarının büyük veri yığınları karşısında yetersiz kalması hem veri madenciliği hem de veri tabanlarında bilgi keşfi uygulamalarını doğurmuştur.

Peki bilgi keşfinin aşamaları nelerdir? farklı veri kaynaklarımız var bu veri kaynakları sadece veri tabanları olmak zorunda da değil yani bir text olabilir webde bulunan bir kaynak bir video olabilir yani herhangi bir veri kaynağını alıyoruz bu verileri ilk olarak temizliyoruz, içerisinde bulunan hatalı verilerden arındırıyoruz ya da içerisinde eksik veriler varsa bu verileri düzenleyip tamamlıyoruz. Bunlar için çeşitli çözüm yöntemlerimiz var. Sonra bir veri ambarına yani işleyebileceğimiz hızlı bir veri ambarına aktarıyoruz farklı platformlarda olabilir ya da klasik bir veri tabanı da olabilir. Daha sonra burada istediğimiz, ulaşmaya çalıştığımız yani hedefe ihtiyaca yönelik verileri seçiyoruz. Bütün verileri almamız gerekmeyebilir, birbirini tekrar eden aynı şeyi ifade eden veriler olabileceği için bir tane veriyi alsak bizim için yeterli olacaktır. Yani ileriki basamağımızda ihtiyacımıza yönelik veriler elimizde olmuş olacak. Sonra bu veriler üzerinde çeşitli veri madenciliği algoritmalarını çalıştırıp bir örüntü oluşturuyoruz ve bu örüntüleri kullanarak bilgiyi elde ediyoruz.

Peki burada bahsettiğimiz üzerinde bu kadar oynadığımız veri kavramı ne oluyor? Veri ne demek? Veri büyüyüp küçülür mü? Veri nasıl büyür? Verinin küçüğü büyüğü olur mu ? Büyük veri nedir? Peki bu büyük veri ve verinin farkı nedir? Verinin büyük veriye olan yolculuğunu anlamak için ilk olarak verinin tanımına bakabiliriz TDK'ye göre veri: Bir araştırmanın, bir tartışmanın, bir muhakemenin temeli olan ana öge, muta, done demektir. Başka bir tanıma göreyse, veri varlığı bilinen işlenmemiş ham haldeki kayıtlar olarak adlandırılır, veriler işlenerek ve düzenlenerek bilgiye dönüşür. Yani tanımlara göre düşündüğümüzde tek bir veri tek başına anlamsız olabiliyor, mesela bir futbol oyuncusunun bir maçtaki doğru pas yüzdesinin yüzde 70 olması, tek başına bize bir hikâyenin tamamını anlatmıyor ama bu verilere tüm oyuncularının oranının verilerini eklersek ligdeki oyuncularının durumunu anlatan büyük bir veriye ulaşıyoruz peki bu veriyi anlamlandırmanın basamakları neler?

Daha kolay anlayabilmek için bilgi piramidinden bahsedelim. Bilgi piramidinin 4 basamağı var ve her bir basamaktan sonrakine ilerlerken önceki basamaktan anlayıp öğrendiklerimizle hareket ediyoruz. Piramidin en temelinde veri bulunmakta. Veri basit bir şekilde işlendiğinde 2.basamaga yani bilgi basamağına ulaşıyoruz bu bilgiler birikince birikim basamağına geçiyoruz yani bu tecrübe demek oluyor. Bu tecrübelerden faydalanmaya başladığımızdaysa son basamağa yani bilgeliğe ulaşıyoruz. Örneğin ilk defa dışarıya çıkan bir insan düşünün, gözleri ile gördüğü her şey veri sayılır. Karşıdan bu insana bir taş atıldığını düşünelim, görüntüden atılanın taş olduğunu anlaması bilgiyken bu taşın ona yaklaştığını anlaması birikim olur, eğer o taştan kaçmazsa kendisine isabet edeceğini bilmesi ise bilgelik oluyor.

Peki gelelim asıl konuya bu veriler nereden geliyor? Bugün evden kaçta çıkıp kaçta geldiğiniz, nerelere gittiğiniz. Konum bilginiz kapalı olsa bile operatör şirketleri tarafından anbean bilinmekte. Kredi kartınızdan yaptığınız tüm ödemelerinizde aldığınız ürünün gramına kadar tüm bilgiler

bankalarca biliniyor. Alışveriş yaparken tıkladığınız tüm ürünler ya da kullandığınız tüm siteler Google tarafından biliniyor yani dijitalleştirilip işlenebilecek olan her şey bir veri havuzu oluşturmaya başladı çünkü bu saydıklarımızın hepsi sizin hakkınızda bir bilgi içeriyor ve tüm bu bilgiler büyük veri adı altında artık bir yerlerde toplanıyor. Bu veri havuzu artık o kadar çok büyüdü ki eskiden insanlar sorduğu sorulara cevap almak için deneyler yaparken şimdilerdeyse bu büyük veri havuzuna sorular sormaya başladık yani işin bu noktasında büyük veri sorular sorup cevaplar alabileceğimiz bir kaynağa dönüştü, bu büyük veriden alınan cevaplar sayesinde biz olmasak bile bizim bilgilerimizle hareket edip hayatımıza etki edecek sistemler kurmuş oluyoruz.

Peki veri ile büyük veri arasındaki farka gelirsek, Pek çok alanda bilgileri depolayan büyük verinin birkaç özelliği var. Bunlardan ilki çeşitlilik yani (variety) dünya üzerinde toplanan ve büyük veriyi oluşturan bütün bilgiler tek bir forma ya da başka bir forma dönüşüyor olmalı bu dönüşüm aynı zamanda çok hızlı gerçekleşmeli ki sistemde doluluk meydana gelmesin işte bu da büyük verinin ikinci özelliği hız yani (velocity) büyük veri her gün içeriğini katlayarak artıyor depolama alanları bu bilgilerle sürekli doluyor bu bize 3.özelligi gösteriyor veri büyüklüğü yani (volume). Ayrıca biriken bu bilgilerin depolanmadan önce doğruluğundan emin olunması gerekiliyor ki burada da karşımıza 4.özellik çıkıyor doğrulama yani (verification) son olarakta bu verilerin depolandıktan sonra bir işimize yaraması gerekiyor en önemli özelliğimizde bu oluyor değer yani(value) işte büyük verinin bu özellikleri 5V olarak adlandırılıyor

Peki bu büyük veri nasıl oluşuyor, veri madenciliği dediğimiz yöntemle istatistik bilimi ve yapay zekalar kullanılarak toplam bilgiden yararlı olanlar alınarak müşteriler ve olası alıcılar için anlamlı bütünler halinde depolanıyor bir nevi ayıklama ve ayrıştırma işlemi gibi. Kitaplar makaleler veya haberler iki boyutlu ya da 3 boyutlu görseller fotoğraflar videolar ise görsel veri olarak depolanıyor. Örneğin internet mağazasından satın aldığınız bir kitaba göre sosyal ağlarda karşınıza çıkan kitap önerisi veya izlemiş olduğunuz bir film ya da youtube videosuna göre benzer filmler ve videoların önerilerde görülmesi bu konu için en iyi örneklerdendir. Marketlerde ürünlerin raflara dizilme sıralamasından reyon sıralamasına hatta ürün çeşidine göre ambalajların renklendirilmesi bile büyük veri ve veri madenciliği sayesindedir.

Peki veri tabanı ve veri ambarları? Veri ambarları veri tabanlarının bir parçası olarak düşünülmektedir. Literatürde çıkışına baktığımız zaman veri tabanıyla uğraşan insanların ilk başlarda uğraştığı ve oradan dallanmış bir kavramdır. Veri tabanı ile veri ambarının farklarını konuşacak olursak; veri tabanının tüm verileri tutmak gibi bir yapısı varken veri ambarıysa veri tabanının küçük bir versiyonu olarak düşünülebilir yani daha hızlı çalışan daha özelleştirilmiş ve nispeten daha az amaca yönelik veriler taşıyan küçük bir versiyonudur. Asıl önemli olan ise veri ambarının farklı veri kaynaklarından besleniyor olmasıdır. Örneğin bir şirketiniz var ve bu şirketin her alanı için ayrı bir şirket çalışıyor ulaşım için ayrı giyim için ayrı toplantılar için ayrı kullanılan eşyalar için ayrı veri kaynaklarımız var yani buna benzer çok fazla veri kaynağının birleştirilerek belli bir amaca yönelik olarak raporlar hazırlanmasını sağlayan, verinin daha verimli kullanılabilmesi için geliştirilen sistemlere veri ambarı diyoruz. Burada veri ambarına baktığımız zaman bir konunun özelleştirilme durumu var. Yani farklı kaynaklardan veriler gelmekte. Ayrıca zaman içerisinde değişiklikler ile birlikte değişmekte kendini günceller bir halde. Ek olarak aslında güncelleme işlemleri veri tabanında yapılmakta veri ambarını kimse gidip değiştirmez, bir güncelleme yapmaz veri ambarı kendini veri tabanına göre günceller. Ayrıca yönetim yani karar verme sürecinde kullanılan bir sistem, düzenektir. Veri ambarı denilen şeyde aslında bir veri tabanıdır ama özelleştirilmiş amaca yönelik bir veri tabanıdır.