T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR VE BİLİŞİM MÜHENDİSLİĞİ ANABİLİM DALI
BİLİŞİM TEKNOLOJİLERİ PR. (YL) (UZAKTAN EĞİTİM)

VERİ MADENCİLİĞİ UYGULAMALARI 2. ÖDEV
BÜYÜK VERİ İLE VERİ MADENCİLİĞİ

Hazırlayan
SEDAT ÖZTÜRK
E235013168

Öğretim Üyesi
Prof. Dr. NİLÜFER YURTAY

NİSAN 2024

1. Giriş

Günümüzde meydana gelen gelişmelerin temelini veri, enformasyon ve bilgi oluşturmaktadır. Veri üzerinde işlem yapılmamış küçük bir bilgi parçacığıdır. Veriler tek başına bir anlam ifade etmezler. Enformasyon, karar vermek için değeri olan ve organize edilmiş verilerin özetlenmesiyle elde edilir. Bilgi de enformasyon verilerinin düzenlenmiş ve analiz edilmesi sonucu değer kazanmış haldir. Bu kavramlar hayatımızın her aşamasında karşımıza çıkar ve oldukça hızlı bir şekilde artarlar.

2. Büyük Veri nedir?

Büyük veri, çeşitli kaynaklardan edinilen devasa büyüklükteki verilerin toplanması anlamına gelir. Toplanan veriler, analiz edilerek farklı teknik ve yöntemler ile anlamlı hale getirilmesi için veriler sınıflandırılır ve işlenir. Büyük veriler akışı hiç kesilmeyen yeni kaynaklardan elde edilen çeşitli veri kümelerinden oluşur.

Büyük verinin temel bileşenleri 5V olarak da bilinenen; çeşitlilik (Variety), hacim (Volume), hız (Velocity), doğruluk (Veracity), değer (Value) özelliklerini bulunmaktadır.

- Çeşitlilik (Variety): Verilerin birbirinden farklı formatlara sahip olmasıdır. Yapılandırılmamış veri, metin, e-posta, fotoğraf, video ve ses gibi birçok farklı biçimde gelir. Yapılandırılmış veri RDBS de (ilişkisel veritabanı) bulunan tablo formatındaki verileri ifade etmektedir.

- Hacim (Volume): Verilerin miktarı için kullanılan terimdir. Yıllar geçtikçe verilerin hacmi hızla artmaktadır.

- Hız (Velocity): Gelişen teknolojiyle birlikte daha hızlı şekilde hareket eder. Hız oranının yükseldiği değerlendirilmektedir.

- Doğruluk (Veracity): Veri güvenliği en önemli konulardan biri olup sağlıklı ve verimli sonuç alabilmek için hatalı verilerin temizlenmesi gerekir.

- Değer (Value): İşlenen veriler anlamlı ve hizmet ettiği kuruma değer yaratmasıdır.

3. Büyük Veri Teknolojileri

Büyük veri teknolojileri verinin depolanması, veri madenciliği, veri analizi ve veri görselleştirme olmak üzere 4 farklı gruba ayrılır.

a. Veri Depolama:

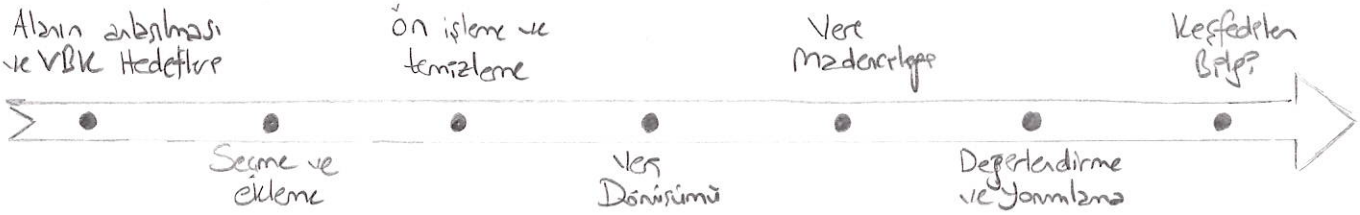
Veri Depolama, verilerin saklandığı ve yönetildiği teknolojileri kapsar. Zaman içerisinde bu verilerin saklamak büyük bir zorluk oluşturmuş ve saklanması için belli başlı veri tabanları ve dosya sistemleri geliştirilmiştir. Büyük veri depolama ve analiz etmek için Hadoop, Mongodb ve Cassandra sistemleri örnek gösterilebilir. Bu sistemler yüksek işlem gücü ve çok sayıda eş zamanlı görevleri yönetme yeteneğine sahiptirler.

b. Veri Madenciliği:

Veri Madenciliği, büyük ölçekli veri kümelerinde ihtiyaca yönelik verileri bulur, analiz eder ve işleyebilir anlamlandırır. Araştırmacılar büyük verilere araç kullanmaksızın analiz edilemeyeceğini görmüşler ve bilgi keşfi kavramı ortaya çıkmıştır. Bu kavram büyük miktarda veri içindeki değerli bilgileri bulma, analiz etme ve kullanma sürecini ifade eder.

Bilgi Keşif Kavramı:

Bilgi Keşfi kavramı 7 aşamadan oluşmaktadır.



- Uygulama alanının anlaşılması ve geliştirilmesi: Bu aşamada uygulama alanı anlamaya çalışılarak çeşitli hazırlıklar yapılır ve bilgi keşfinde ki amaç geliştirilir.

- Seçme ve ekleme: Veri tabanından alınan analiz ile ilgili verilerden, probleme ilişkin olan verilere seçme sürecidir.

- Ön işleme ve veri temizliği: Eksik, gürültülü ve tutarsız verilerin temizleme sürecidir.

- Veri Dönüşümü: Bu aşamada verinen uygun formlara dönüştürülüp veri madenciliğinde kullanılabilecek hale getirme sürecidir.

- Veri madenciliği: Hazırlanan veriler üzerinden amacına göre Veri Madenciliği Algoritmalarının uygulama sürecidir.

- Değerlendirme ve yorumlama: Bu aşamada oluşan örüntüler değerlendirilerek bir önce aşamada belirlenen amaç ile alaka kurulmalıdır ve elde edilen bilgi amaca yönelik kullanılmalıdır.

- Keşfedilen Bilgi: Veri madenciliği elde edilmiş bilginin kullanıcıya sunulmasıdır.

Presto, Apache spark, Apache Flink, Apache Hadoop veri madenciliği için kullanılan en popüler programlardır. Özellikle presto verileri ayrı bir analiz sistemine taşımaya gerek kalmadan depolandığı yerde sorgulanabilir.

Veri Madenciliği Yöntemleri:

Veri madenciliği yöntemlerine 3 grupta toplayabiliriz. Sınıflama ve regresyon, kümeleme ve birliktelik kuralları modellemedir.

Sınıflama ve Regresyon: Sınıflama verinin önceden belirlenen çıktılara uygun olarak ayrıştırılmasını sağlayan bir tekniktir. Çıktılar, önceden bilindiği için sınıflama, veri kümesini denetimli olarak öğrenir.

Modelde kullanılan başlıca teknikler: Karar Ağaçları, Yapay Sinir Ağları, Genetik Algoritmalar, K-En Yakın Komşu, Regresyon Analizi, Naïve-Bayes, Kaba Kümeler

Kümeleme: Nesneler benzerleriyle gruplama sürecine kümeleme denir. Kümeleme analizi, temel amacı nesnelere sahip oldukları karakteristik özellikler baz alarak gruplamak olan çok değişkenli teknikler grubudur.

Genel olarak başlıca kümeleme yöntemleri: Bölme yöntemleri, Hiyerarşik yöntemler, Yoğunluk tabanlı yöntemler, Izgara tabanlı yöntemler, Model tabanlı yöntemler

Birliktelik Kuralları: Büyük veri kümeleri arasında birliktelik ilişkilere bulurlar. Bu yöntem büyük miktardaki işlem kayıtlarından ilginç birliktelik ilişkileri keşfetmek ve şirketlerin kar alma işlemlerini daha verimli hale getirmektedir.

c. Veri Analizi:

Veri analizi, yararlı bilgilere keşfetmek, sonuç çıkarmak ve karar vermeyi desteklemek amacıyla verileri incelemek, temizlemek, dönüştürmek ve modellemek için kullanılan bir süreçtir. Veri analizi teknolojisinde öne çıkan isimler arasında Kafka, Splunk, Spark v.b. yazılım platformları yer alıyor.

d. Veri Görselleştirme:

Veri görselleştirmesi, verinin grafikler, tablolar veya haritalar sayesinde görsel bir dile dönüştürülmesi anlamına geliyor. Görselleştirmede amaç, istatistiksel ve değişen bilgilerinin klasik formatta sunulan kompleks verilere, kolay algılanabilecek grafik arayüzler ile rahat anlaşılır hale getirmektir. Bu sayede verilerdeki bilgiler, ortaya çıkan trendler ve düşüşler kolayca görülüyor ve yorumlanabiliyor. Bu teknolojiye, verilerin görselleştirilmesini sağlayan Tableau, Microsoft Power BI, Qlik Sense, Chart Blocks, Plotly, Looker v.b. yazılım programları, örnek olarak gösterilebilir.