**BÜYÜK VERİ KONUSUNA KAVRAMSAL GİRİŞ**

**(Aralık-2016)**

**1. Büyük Veri Nedir?**

*Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...*

*Dan Ariely (Professor at MIT)*

Büyük verinin herkes tarafından kabul edilmiş olan tam bir tanımı yoktur. Büyük verinin karakteristik özelliklerini Doud Laney 3V ile açıklamıştır. 3V’li betimleme en çok kullanılan betimlemedir:

Volume (Miktar): Büyük veride işlenecek veri miktarı çok fazladır.

Velocity (Hız): Kaynaklardan veriler çok büyük hızlarda gelmektedir.

Variety (Değişkenlik): Verinin niteliği çok değişik olabilmektedir. (Örneğin resim, yazı, ses, görüntü vs.)

3V’ye daha sonra bir 3V daha eklenmiş (veracity, validity, volatily), 6V haline getirilmiştir.

**1) Miktar (Volume)**

Büyük verinin en önemli özelliği yüksek volümlü olmasıdır. Eskiden büyük miktarda verilerle karşılaşmak çok zordu. Büyük miktarda veriler operatörler tarafından günlerce, aylarca hatta yıllarca elle tek tek giriliyordu. Ancak artık verilerin elde ediliş biçimleri değişmiştir. Şöyle ki: Veriler -sosyal medya uygulamalarında olduğu gibi- milyonlarca kişi tarafından oluşturulabilmektedir. Veriler çeşitli aygıtlar tarafından otomatik biçimde elde edilebilmektedir. (Örneğin kameralar, turnike geçiş sistemleri, MR cihazları, GPS aygıtları, Arama motorları, Bilet reservasyon siteleri vs.) Verinin elde ediliş biçiminin değişmesi miktarın büyümesinde önemli rol oynamıştır. Zaten «büyük veri» kavramı bu gelişmelerden sonra ortaya çıkmıştır. Eskiden büyük verilerle sık karşılaşılmıyordu. Zaten onu analiz edebilecek donanımsal yazılımsal alt yapı da yeterli değildi. Büyük verileri analiz etmek için gerçekten bugünkü gibi gelişmiş bilgisayar donanımlarına ve yzılımlarına gerek vardır.

**2) Oluşma Hızı (Velocity)**

Büyük verinin diğer bir karakteristik özelliği bunların çok hızlı bir biçimde oluşup duruma göre kaybolmasıdır. Örneğin, 2012 yılı itibarıyla:

- Youtube’a bir dakikada 48 saatlik video yükleniyor.

- Dakikada 571 yeni web sayfası oluşturuluyor.

- Facebook’a her gün 100 terabyte bilgi aktarılıyor. (Bugün kişisel bilgisayarlarımızdaki en büyük diskler 2 terabyte)

- Tweeter’da her gün 175 milyon tweet atılıyor.

- Facebook’ta her ay 30 milyar içerik paylaşılıyor.

- 2014 yılı itibarıyla WhatsApp’ta bir günde 64 milyar mesaj atılıyor.

Bir günde GSM operatörlerindeki mesajları, MRI cihazlarında yapılan tetkikleri, GPS’lerden atılan konum bilgilerini, turnikelerden geçenlerin sayılarını, bilet rezervasyonu yapanların oluşturdukları verileri, otoyollara giriş yapan araçları, Google’da aranan içerikleri düşünün… Bu veriler çok hızlı oluşuyor. Bazıları hemen güncelliğini yitiriyor…

**3) Çeşitlilik (Variety)**

Büyük veriyi oluşturan verilerin formatları çok değişkenlik göstermektedir. Örneğin bu veriler:

- Yazı biçiminde (text) bulunuyor olabilir.

- Resim biçiminde bulunuyor olabilir (bmp, jpeg, gif, pif vs.)

- Video biçiminde bulunuyor olabilir (mpeg4, avi, flv vs.)

- Kompozit biçimde bulunyor olabilir (pdf, doc vss.)

- Binary biçimde bulunuyor olabilir.

Büyük verinin bu çeşitliliği onun ilişkisel veritabanlarında saklanmasını zorlaştırabilir ve işlenmesini karmaşık hale getirebilir. Bu nedenle büyük verilerin işlenmesinde yoğun olarak başka veritbanı modelleri kullanılmaktadır. Büyük verilerin depolanmasında en çok tercih edilen veritabanı modeli NoSQL’dir. En çok kullanılan NoSql tarzı veritabanı yazılımlarından biri MongoDB’dir. MongoDB kursumuzun konuları içerisindedir.

**4) Doğruluk (Veracity)**

Büyük veriyi oluşturan veriler yanlı (biased) olabilir, bozulmuş (abnormal) olabilir, gürültü (noise) içeriyor olabilir. Çok miktarda verinin hızlı bir biçimde toplanması sırasında verilerin bir kısmının sağlıklı olmayabileceği veri işlemesi sırasında akılda tutulmalıdır.

**5) Geçerlilik (Validity)**

Büyük veriyi oluşturan verilerin geçerli ve tam olup olmadığı durumaa göre değişebilmektedir.

**6) Kalıcılık (Volatility)**

Büyük veriyi oluşturan verilerin büyük kısmı kalıcı değildir. İlgili sistem tarafından oluşturulur. O andaki gereksinimi gerçek zamanlı olarak karşılar ve önemini yitirir. Önemini yitiren veriler tamamen çöp haline gelmez. Onlar aanaliz edilerek çeşitli korelasyonlar, bağıntılar ve içgörüler elde edilebilir. Geçmiş veri geleceğe ışık tutabilir.

**2. Büyük Veri Analizi (Big Data Analytics) Nedir?**

Büyük verinin analizi bilgi parçalarının içindeki ve arasındaki ilişkileri görmek ve anlamakla ilgilidir. Büyük veri analizinde nedensellik yerine korelasyonlar aranır. Yani neden yerine ne yakalanmaya çalışılmaktadır. Örneğin ilk büyük veri uygulamalarından biri olan Farecast uçak biletininin en ucuza alınabileceği zamanı kestirmeye çalışıyordu. Bu sistem uçak bilet fiyatlarının neden ve nasıl dalgalandığını açıklamaya çalışmıyordu. Verileri konuşturarak doğrudan kestirimde bulunmayı amaçlıyordu.

Büyük veri incelemesinin en önemli noktalarından biri örnekleme (sampling) yerine tüm verinin (yani ana kütlenin hepsinin) dikkate alınmasıdır. Örnekleme ana kütlenin çok büyük olmasından dolayı onu inceleyecek olanak olmaması nedeniyle başvurulan bir yöntemdir. Artık pek çok durum için örneklemeye gerek kalmamıştır. Artık tüm veri analiz edilebilir.

**3. Büyük Veri Analizinde Kullanılan Araçlar ve Platformlar**

Büyük veri analizinde pek çok araç ve platform kullanılmaktadır. Burada en önemli birkaçı üzerinde duracağız.

**3.1. Hadoop**

Hadoop büyük veri alanında en çok kullanılan platformdur. Açık kaynak kodludur (dolayısıyla bedavadır). Hadoop büyük veriyi cluster’lara ayırarak daağıtık biçimde (yani birrden fazla bilgisayarda) saklayablmektedir. Hadoop dağıtık (distributed) çalışan bir sistemdir ve çok güçlü bir hesaplama gücü vardır. Hadoop esnek bir sistemdir. Hata toleransı yüksektir. Örneğin cluster’ları oluşturan bilgisayarlardan birinde bir sorun çıksa bile sistem çalışmaya devam eder. Hadoop ölçeklenebilir (scalable) bir sistemdir. Ölçeklendirilebilir demek verinin boyutu arttıkça ek cluster’lar oluşturularak sistemde bir değişiklik yapmadan sistemin devamının sağlanması demektir. Hadoop kursumuzun konuları içerisinde ele alınmaktadır.

**3.2. R Programlama Dili**

R istatistiksel ve matematiksel veri analizi için tasarlanmış «domain specific» bir programlama dilidir. R açık kaynak kodlu (dolayısıyla bedava) bir dil ve ortamdır. R büyük veri analizinde en yoğun kullanılan araçlardan biridir. R ile çok büyük veriler üzerinde arama, karşılaştırma gibi işlemler hızlı bir biçimde yapılabilmektedir. R’ın onlarca paketi vardır. Bu paketler sayesinde her türlü analizler, istatistiksel hipotez testleri, yapısal eşitlik modellemesi vs. yapılabilmektedir

**3.3. Google Big Query**

Big Query ile petabyte’lar mertebesinde veriler üzerinde çok hızlı analizler yapılabilmektedir. Platform Google’a aittir ve belli bir ücret ödenerek kullanılabilmektedir. Sistemi kullanmak için bir veritabanı yöneticisine gereksinim yoktur. Sistemi kısa bir eğitimle kullanmak mümkündür.

**3.4. Amazon Web Services**

Amazon firması tarafından geliştirilmiş çok yönlü bir platformdur.Amazon Web Services ücretlidir. Arka planda Hadoop’u da kullanmaktadır. Yüksek hızlı, ölçeklenebilir bir analiz platformu sunmaktadır. Makine öğrenmesi (machine learning) için de araçları bulunmaktadır.