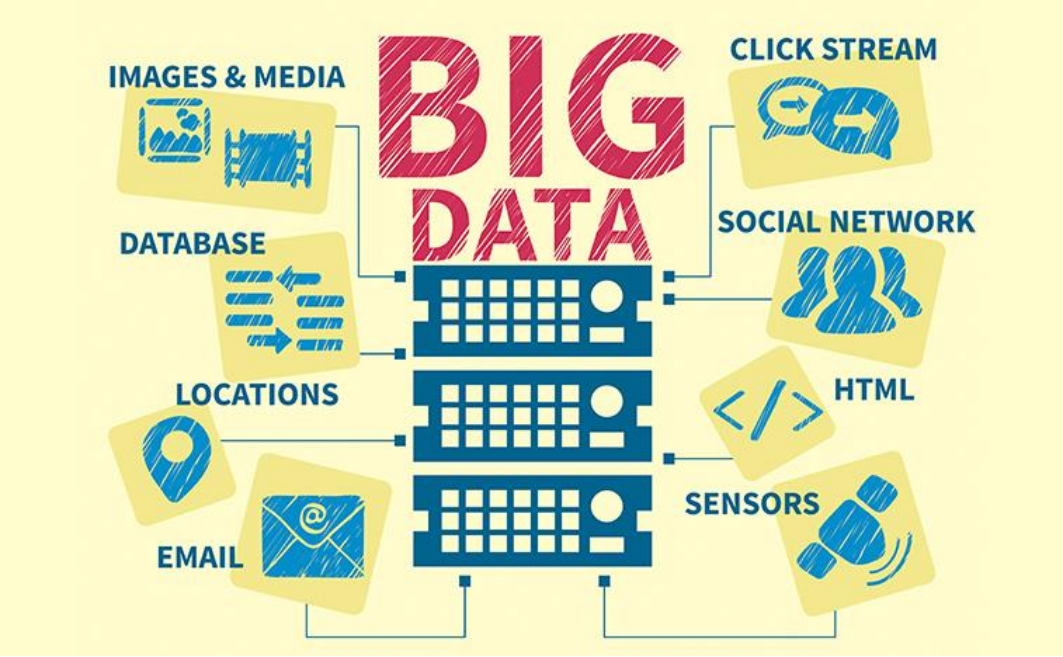
Big Data ve Data Science Nedir?

Big Data Nedir?

- Teknolojinin gün geçtikçe gelişen dünyasında veri tabanı sistemlerinde depolanan ve işlenen veri miktarı da artmıştır.
- Depolanan veri miktarı arttığı gibi, bu verinin çeşitliliği de artmıştır.

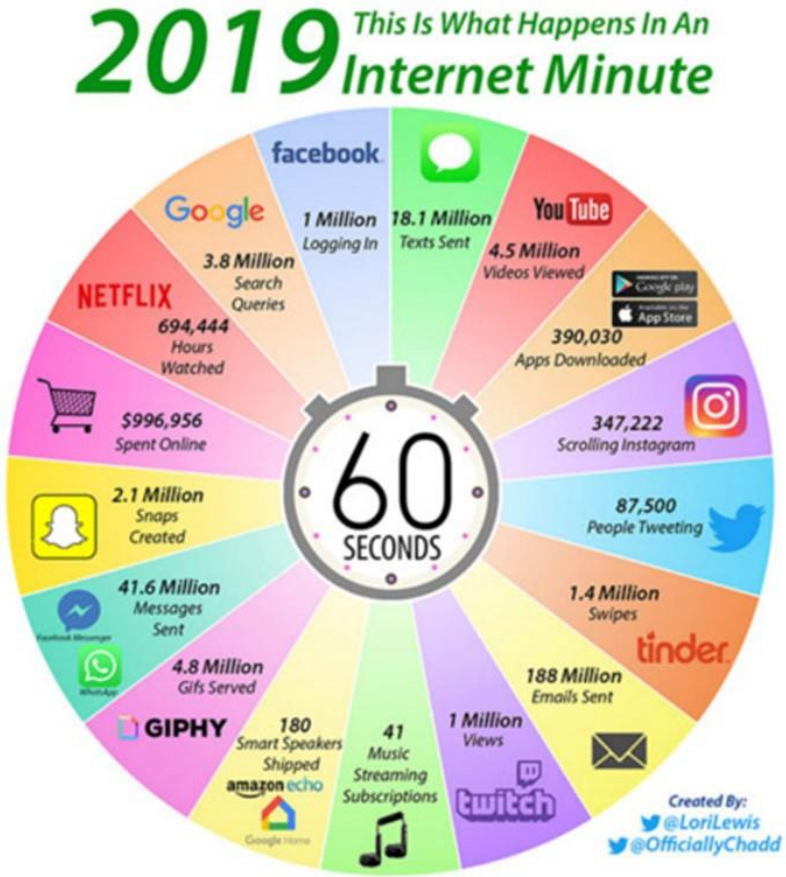


Big Data Nedir?

- Birçok kaynaktan farklı türde veriler üretilmeye başlanmıştır.
- (Internet of Things IOT) kavramı hayatımıza girdikten sonra verinin düzensizliği hızla artmıştır.
- 2000 yılında dünya üzerinde yer alan verinin sadece %20'si dijital ortamda tutulurken 2017 itibari ile bu oran %98'e ulaşmıştır.

Big Data Nedir?

- 1 Dakikada internette;
  - 3,8 Google araması
  - 18 Milyon Text Mesaj
  - 4,5 Youtube videosu
  - 88 Milyon Tweet
  - 188 Milyon E-mail



Big Data Nedir?

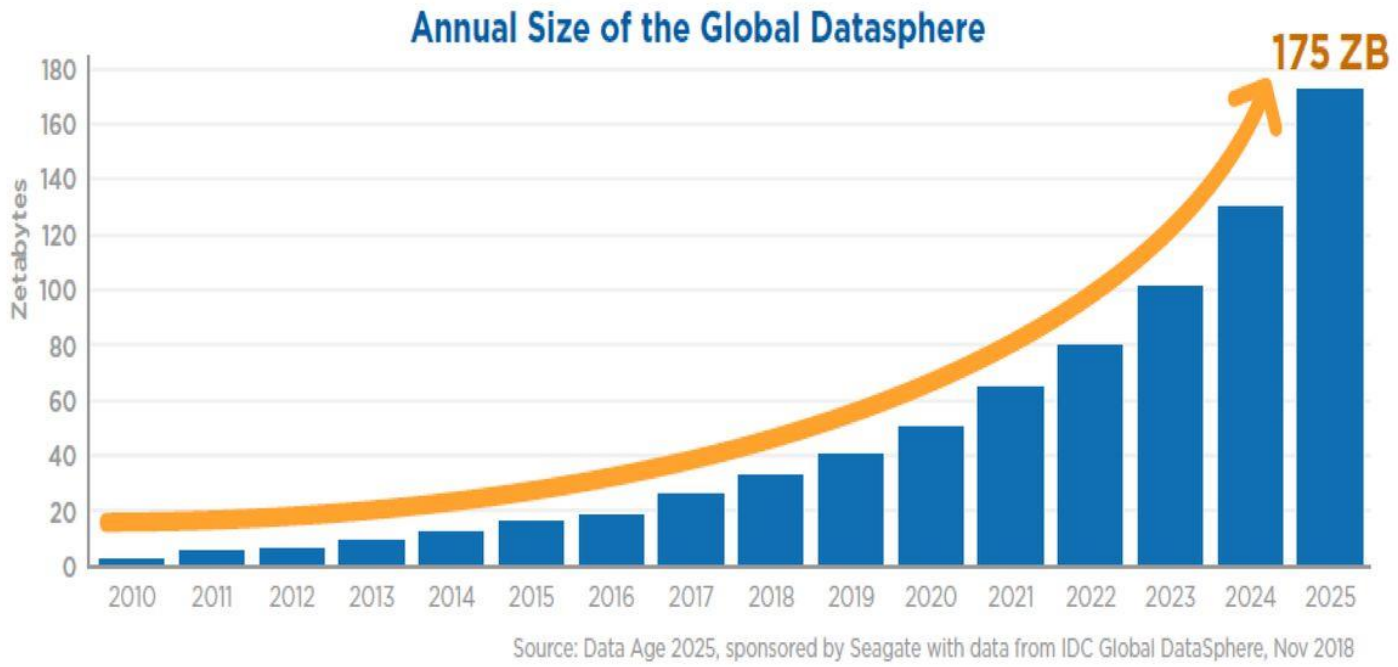
- Veri Boyutunu ölçmede kullanılan birimler şunlardır;

1 Bit = 1 or 0
1 Byte = 8 bit
1 Kilobyte = 1024 byte
1 Megabyte = 1024 kilobyte
1 Gigabyte = 1024 megabyte
1 Terabyte = 1024 gigabyte
1 Petabyte = 1024 terabyte

1 Exabyte = 1024 petabyte
1 Zettabyte = 1024 exabyte
1 Yottabyte = 1024 zettabute
1 Brontobyte = 1024 yottabyte
1 Geopbyte = 1024 brontobyte

Big Data Nedir?

- Dünya üzerinde toplam veri boyutu 2020 Yılında 44 Trilyon Gigabyte'a ulaşmıştır, yani 44 Zettabyte.



Big Data Nedir?

- Üretilen veri ise mesaj, fotoğraf, video, mp3, sunum, sensör verileri gibi birçok farklı varyasyondadır.
- Veri kaynakları olarak gösterilebilecek başlıklar ise;
  - Firma – Müşteri ilişkileri, Cep telefonları, bilgisayarlar, sensörler gibi internete bağlı her türlü cihaz, sosyal medya platformları

Big Data Nedir?

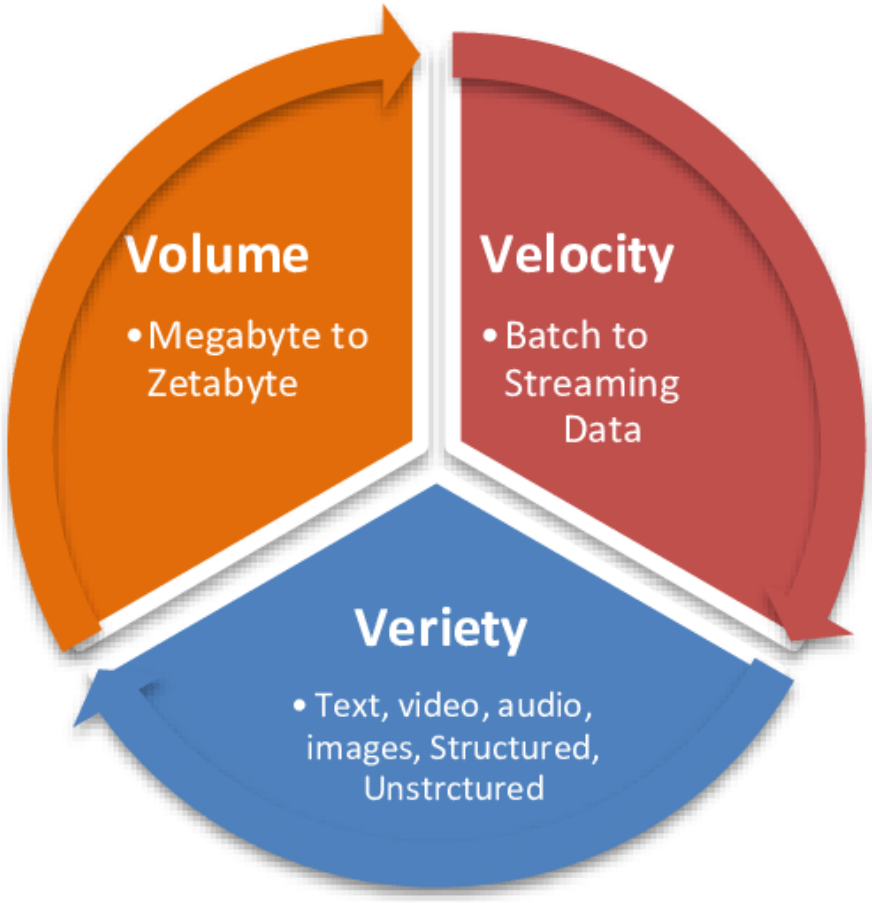
- Örnek ; E-ticaret sitesi.
  - Üyelik bilgileri düzenli veri.
  - Site içerisindeki hareketler düzensiz veri.
  - Üyelik bilgisi bir kez, Lokasyon bilgisi sürekli alınır.
  - Lokasyon bazlı kampanyalar Kullanıcıya önerilir.

Big Data Nedir?

- **Big data yani büyük veri mimarisi, boyutu ve çeşitliliği artan büyük öleçkli verinin performanslı ve düşük maliyetle tutulabilmesine imkan veren mimariye verilen isimdir.**

Big Data Nedir?

- Big data mimarisi 3V ile tanımlanır.
  - Volume (Hacim)
  - Velocity (Hız)
  - Variety (Çeşitlilik)



Big Data Nedir?

- Volume : Veri setinin büyüklüğünü tanımlar. Petabaytlar ile ifade edilir.
- Velocity : Veri setinin büyüme hızını tanımlar.
- Variety : Veri setinin yer alan verinin çeşitliğinini tanımlar.

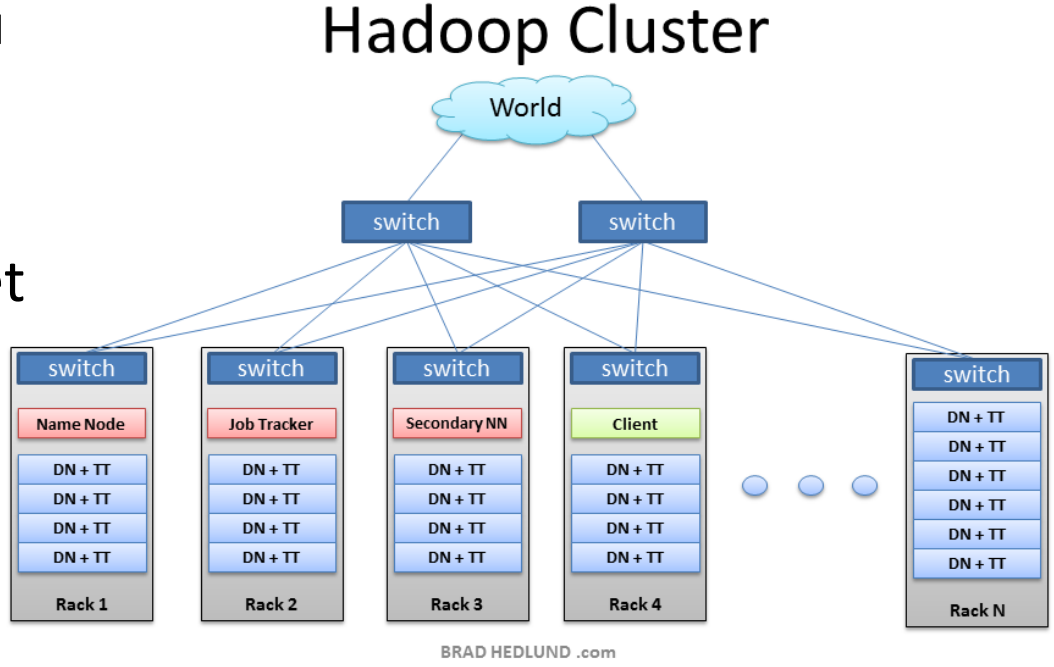
- Not: Bazı kaynaklarda 4’üncü bir V olarak Veracity(Gerçeklik) kavramı kabul edilir. Bu kavram verinin doğruluğundaki belirsizlik durumunu ifade eder.

Big Data Nedir?

- Big Data mimarisinin temelini cluster (küme) ve distributed systems (dağıtık mimari) oluşturur.
- Bu mimariler sayesinde veri geleneksel modellere göre daha düşük maliyet ve daha yüksek performans ile depolanır ve işlenir.
- Bu mimariler sayesinde büyük ölçekli sistemlerin hata durumlarından etkilenme handikapı en aza indirilir.

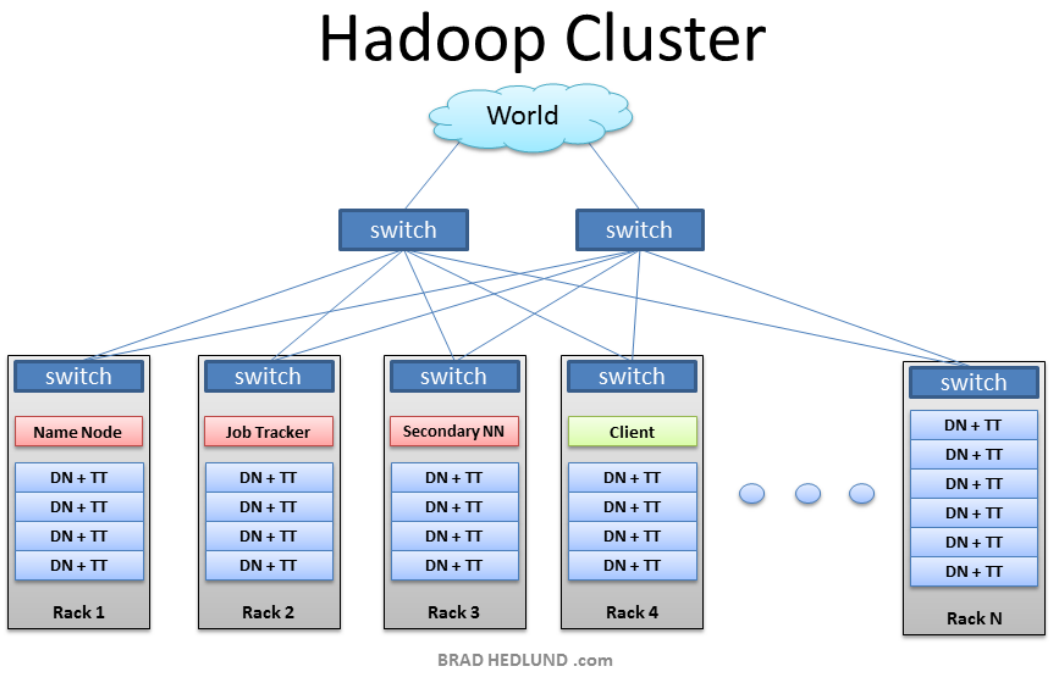
Big Data Nedir?

- **Cluster (Küme) mimarisi** ; birden çok sunucunun aynı amaç için hizmet verdiği sistemlerdir.
- Sistemlerin sunucu hatalarından etkilenmeden hizmet vermeye devam etmesi için tasarlanmış mimaridir.
- Cluster mimarisinde yeni kaynak eklemek için yeni sunucu ekleme mantığı var olan sunucu kaynaklarını büyütmeye oranla daha kolaydır.



Big Data Nedir?

- **Distributed (Dağıtık) Mimari** : Veri tutulurken ve işlenmesi aşamasında yapılacak işlemlerin belirli parçalara bölünerek Cluster'da yer alan Sunucular üzerine dağıtılmasıdır.
- Böylelikle Cluster üzerindeki sunucu kaynakları maksimum verimle kullanılır.
- Cluster'da yer alan her bir sunucuya **Node(düğüm)** denir.

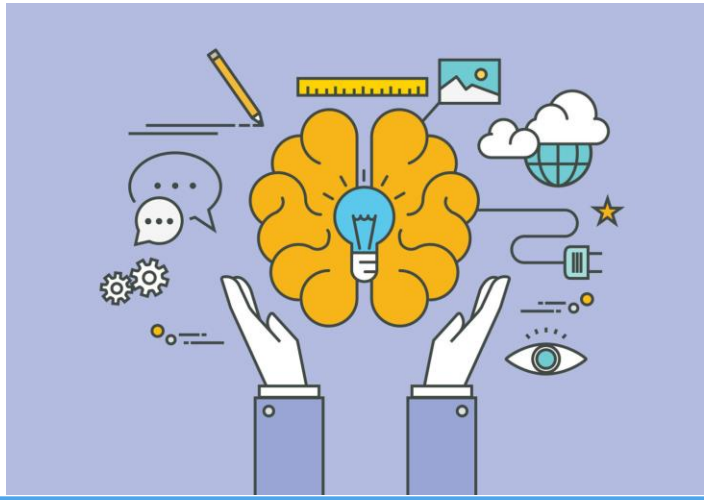


Big Data Administrator Kimdir?

- Big data ortamlarının yönetimini sağlar.
- Büyük verinin depolanacağı HDFS (Hadoop) altyapısını kurar ve yönetir.
- Farklı kaynaklardaki verinin Big data ortamlarına aktarılmasını sağlar.
- Big data ortamı ile ilgili kaynakların maksimum performans ve minimum maliyetle çalışmasını sağlar.
- Big data admin, Hadoop ekosistemi ve bu ekosistemde yer alan her türlü yazılımın kurulumu ve yönetiminden sorumludur.
- Linux işletim sistemleri konusunda temel seviyede bilgi sahibi olmalıdır. İyi bir linux bilgisi ise iyi bir big data admin olmanın en büyük adımlarındandır.

Data Science Nedir?

- Big data mimarisinde tutulan veriden anlamlı sonuçlar çıkarmak ise bu verinin tutulmasındaki temel amaçtır.
- Büyük veri ortamlarında tutulan verinin analiz edilip anlamlı sonuçlar ve öngörüler ortaya koyan bilim dalı ise Data Science yani veri bilimi olarak adlandırılmaktadır.
- Data science kısaca veriden bilginin çıkarılmasıdır.



Data Science Nedir?

- Veri bilimini değerli kılan özelliklerin başında ise eldeki veriden geleceğe yönelik tahminlerin çıkarılmasıdır.
- E-ticaret örneğine dönersek; geçmişte müşterinin aldığı ürünleri ve almasa bile ilgilendiği ürünlerin bilgilerini tutarak bu müşterinin bir sonraki ziyaretinde onun ilgilenebileceği ürünleri kendisine sunmak bir veri analiz işleminin sonucudur. Ve veriden belirli modeller ile bir tahmin çıkarılmasını gerektirir.

Data Science Nedir?

- Data Science'ın bileşenlerinden ilki istatistiktir.
- Çözmek istenilen problem istatiksel bir modele oturtulup bu problemin çözümü hedeflenir.
- İstatistiksel öğrenme teorisi altında yer alan belli algoritma ve modelleri kullanarak eldeki veri probleme uygun şekilde analiz edilir.

Data Science Nedir?

- ÖRNEK ; Altın ons fiyatının geçmiş verisinden faydalanarak 2 sene sonraki ons fiyatını tahmin etmek için İstatistiksel öğrenme teorisinde yer alan algoritmalardan lineer regresyon modeli kullanmaya karar verildik ve ons fiyatını etkileyen değişkenlerden birisini **faiz** diğerini ise dolar **kur bilgisi** olarak belirledik. Bu durumda ons fiyat için şu şekilde bir denklem kurduk;

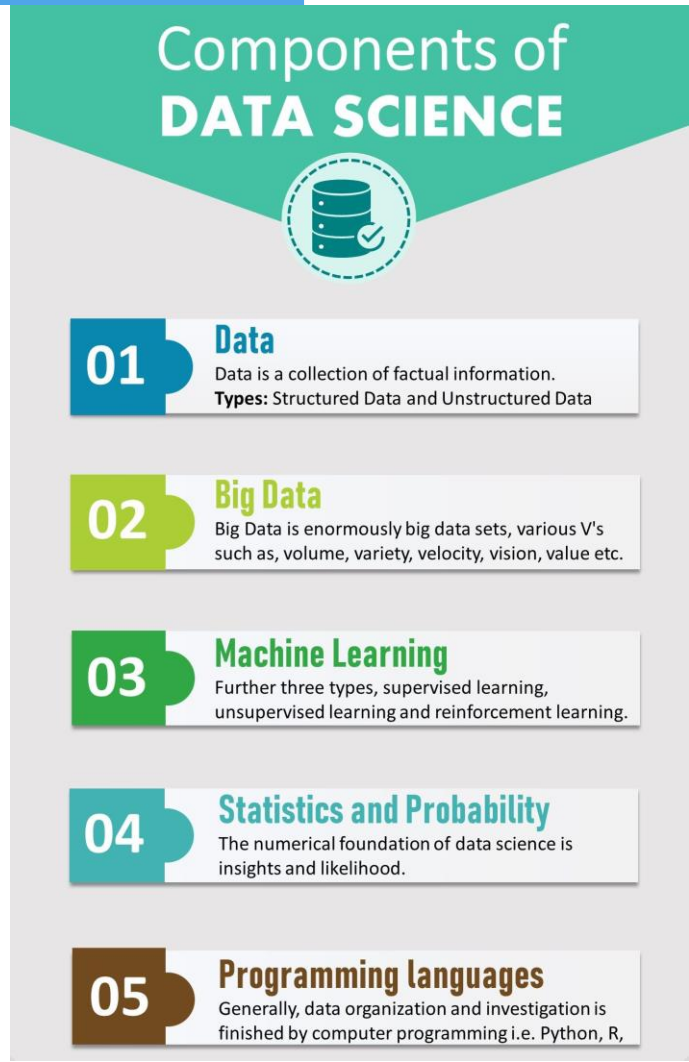
  Ons fiyat = sabit + Parametre1 X **Faiz** + Parametre2 X **Dolar Kuru** + Hata Payı

Data Science Nedir?

- Data science denince akla gelen makine öğrenmesi (Machine learning) kavramı ise bahsedilen bu istatiksel öğrenme teorisinin bilgisayarlara entegrasyonudur denebilir.
- Makine öğrenmesinde temel amaç elde bulunan geçmiş veri ile bilgisayarlara bir işin yapılmasını öğretmedir.
- Eldeki veri seti ile, oluşturulan modeller eğitilerek makinelere öğrenme becerisi programlama dilleri aracılığı ile kazandırılır.

Data Science Nedir?

- Data science'ın bir diğer bileşeni ise Bilgisayar teknolojileridir.
- Verinin tutulması, aktarımı sırasında bu alandan faydalanıldığı gibi verinin işlenmesi sırasında da başta programlama dilleri olmak üzere birçok bilgisayar biliminden faydalanılmaktadır.



Data Science Nedir?

- Bilgisayar teknolojileri veri biliminde aşağıdaki amaçlar için kullanılır;
  - Veriyi aktarmak, depolamak ve gerektiğinde erişebilmek.
  - Veriyi temizlemek ve yeni veri seti üretimi yapmak.
  - Veriyi görselleştirmek.
  - Veri üzerinden istatistiksel modeller çıkarmak.
  - Makine öğrenmesi ile algoritmaları bilgisayarların anlayacağı programlama kodlarına dökmek.
  - Modelleri veri ile eğitmek.
  - Modelleri dış dünyaya hizmet verecek şekle getirmek.

Data Science Nedir?

- Bu başlıklar haricinde Data science biliminde belirli ölçüde uğraşılan problemin ilgilendiği alan hakkında da bilgi sahibi olunması gerekir.
- Basit bir örnek ile açıklamak gerekirse bir forvet oyuncusunun attığı gol sayısının giydiği ayakkabının rengi ile ilgisinin olmadığını tespit etmek alan bilgisine girer.

Data Science Nedir?

- Data Science ile amaçlanan nedir?
  - Geçmişe dair anlamlı sonuçlar çıkarabilmek
  - İçinde bulunulan anın daha verimli geçirilmesini sağlayabilmek
  - Gelecekte gerçekleşecek senaryolar hakkında tahmine bulunabilmek

Data Scientist Kimdir?

- Veri Bilimci, bir programcıdan çok istatistik bilen, bir istatistikçiden çok programlama bilen kişiye denir.
- Veri Bilimi'nin bilgisayar bilimleri içinde yararlandığı tek şey programlama değil. Esasında, dağıtık mimarilerden, büyük veri teknolojilerine kadar bir çok başlık Veri Bilimi'nin kullandığı araç kümesine giriyor

Data Scientist Kimdir?

Bir veri bilimcinin uğraşması gereken işlerin bazıları;

- Zengin veri kaynakları bulup işlemek
- Donanım, yazılım ve bağlantı kısıtlarına rağmen büyük miktarda verileri yönetebilmek
- Veri kaynaklarını birleştirmek
- Veri kümeleri arasında tutarlılığı sağlamak
- Veriyi anlamlandıracak görselleştirmeleri yaratabilmek
- Matematik ve istatistik modeller oluşturabilmek
- Anlamı bulmak
- Anlamı diğer teknik insanlara ve olası olarak teknik olmayan insanlara sunabilmek

Data Scientist Kimdir?

- Nasıl Data Scientist Olunur. Neleri bilmek avantaj sağlar?
- Programming bilgisi : Örnek vermek gerekirse Java , python olabilir.
- Database ve SQL bilgisi : Oracle , Mysql , Postgresql
- Operating Sistem Bilgisi : Linux/ Unix te temel komutları bilmek son derece önemli , fazlası büyük avantaj
- Big DATA bilgisi : Başta Hadoop ecosistemi ,Presto , Hive , Hbase , Impala ,Pig vb.

Data Scientist Kimdir?

- Big Data Real Time işler : Gümünümüzün en önemli farklarından bir tanesi anlık işleri yakalama ve process edip analiz etme. SPARK , Kafka , Nifi .
- Istatistik bilgisi
- Algoritma bilgisi
- Alan bilgisi ve veriyi anlamlandırabilmek

Data Scientist Yıllık Ücretleri

- **Facebook** - $ 145,365
- **IBM** - $ 114,635
- **Microsoft** - $ 129,917
- **Uber** - $ 125,672
- **Airbnb** - $ 140,312
- **Google** - $ 140,212
- **Amazon** - $ 120,407
- **Twitter** - $ 142,665

Data Scientist Kimdir?

- Big Data Real Time işler : Gümünümüzün en önemli farklarından bir tanesi anlık işleri yakalama ve process edip analiz etme. SPARK , Kafka , Nifi .
- Istatistik bilgisi
- Algoritma bilgisi
- Alan bilgisi ve veriyi anlamlandırabilmek

Neler Öğreneceğiz?

- **Big Data Administrator Eğitimi ;**
- **Linux İşletim Sistemlerine Giriş ve Çok kullanılan komutlar.**
- **Big Data Nedir.**
- **Big Data Teknolojisinin Kullanım alanları.**
- **Hadoop File System Mimarisi**
- **Big Data Platformları**

Neler Öğreneceğiz?

- Big Data ile Sorgulama İşlemleri
- Big Data ile Veri Aktarım İşlemleri
- NoSQL Veri Tabanları
- Elasticsearch kurulumu ve ELK örnek uygulaması

Neler Öğreneceğiz?

- Data Scientist Eğitimi ;
- Python
- Veri İşlemleri
- Apache Spark
- Spark Makine Öğrenmesi
- Spark Streaming ve Structured Streaming İşlemleri
- Hadoop File System Mimarisi
- Veri Görselleştirme

Big Data Kullanım Alanları

- Big Data ve Data science teknolojileri veri ile ilgilenen her sektör ve kurum tarafından etkili biçimde kullanılmaktadır.
- Başta bankacılık ve finans sektörü olmak üzere, büyük teknoloji şirketleri, e-ticaret sisteleri, savunma sanayi ve otomotiv sektörleri gibi bir çok alanda bu teknolojiler kullanılmaktadır.

Big Data Kullanım Alanları

- Bankacılık sektörü : Fraud detection (Sahtecilik kontrolü)

Big Data Kullanım Alanları

- E-ticaret sektörü : Kullanıcı davranış analizi, market trendleri, kampanya önerileri, müşteri profiline uygun ürünler sunulması.

Big Data Kullanım Alanları

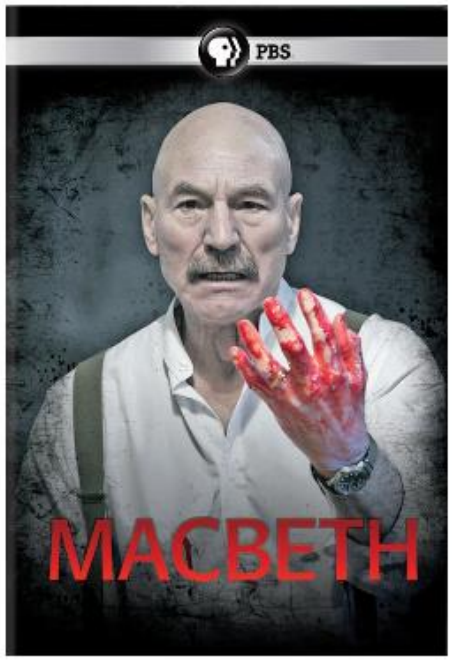
- Sağlık sektörü : Yeni tedavi yöntemleri bulunması, salgın hastalık kontrolü ve önleyici adımlar, Gen haritaları incelenerek kalıtsal hastalıklara tedavi arayışları.

Big Data Kullanım Alanları

- Sosyal Medya : Ücretsiz servisler ile kullanıcı kitlesini genişletip, tutulan kullanıcı verisinden kaynak sağlamaya yönelik şirket politikaları mevcuttur. Bu yüzden big data ve data science teknolojileri sosyal medya şirketleri tarafından yoğun şekilde kullanılır.

Big Data Kullanım Alanları

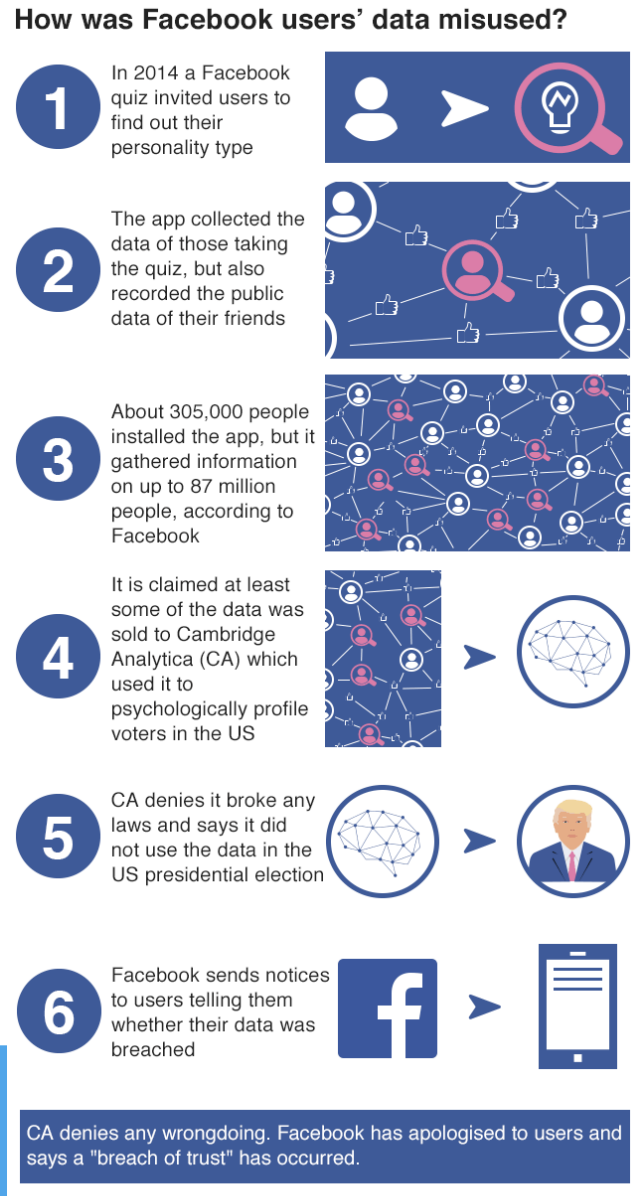
- Netflix : Online yayın platformu olan Netflix Big data mimarisini en etkili kullanan şirketlerden biridir. Kullanıcılarının hoşuna giden afiş rengini onlara göstermede dahi Big Data teknolojisinden faydalanır durumdadırlar.



NETFLIX

Big Data Kullanım Alanları

- Cambridge Analytica : Veri Analiz şirketi. Politik kampanyaların veri analiz sonuçları ile yönlendirilmesini sağlamıştır.



Big Data Kullanım Alanları

- Devletler : Savunma sanayi, sağlık alanları başta olmak üzere kurumların politika ve hedeflerin net şekilde ortaya koyulması için Devletler de veri biliminden istifade etmektedir.