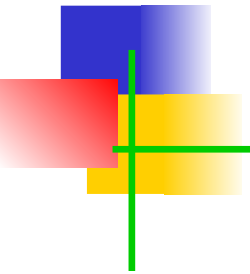
Büyük Veri Analizine Giriş
Büyük Veriye Genel Bakış



Dr. Süleyman Eken

Bilgisayar Mühendisliği
Kocaeli Üniversitesi

Sunum Planı

- Dağıtık sistemlere kısa temas
- Büyük veri nedir?
- Büyük veri ile neler yapılabilir?
- Büyük veri nasıl işlenir?
- Nereden başlamalı?
- Sertifika programları, iş ianları vs

Dağıtık Sistemlerin Tanımlanması

- *"Ağ üzerindeki bilgisayarlarda bulunan donanım veya yazılım bileşenlerinin yalnız mesaj göndererek haberleştikleri sistem."* [Coulouris]

- *"Dağıtık bir sistem, kullanıcılara tek bir sistem olarak görünen, bağımsız bilgisayarlar bütünüdür."* [Tanenbaum]

- Örnek: WWW, Intranet (organization), P2P sistemler (Napster gibi)
  - Bankalar (Bankamatikler)
  - Bilet rezervasyonu

DS Tanımının Getirdikleri

- Dağıtık sistemlerdeki bilgisayarlar ayrı kıtalar üzerinde, aynı bina veya aynı oda içerisinde bulunabilir. DS'in getirdikleri:
  - Birlikte ve birbirinden bağımsız çalışan sistemler
    - İşlerini birbirinden bağımsız yaparlar
    - *Aynı zamanda program çalıştırır, bütün bir işleme kaynağı gibi görünür, birlikte çalışırlar*
    - İşlemler mesaj alışverişiyle anlaşırlar.
  - Heterojen (çeşitlilik, farklılık): networks, donanım, os, iletişim vs
  - Bağımsız bozulma: biri bozulsa da diğerleri çalışmaya devam eder.

Kısaca zorluklar

- Heterogeneity (Çeşitlilik)
  - Çeşitli bileşenler birbiriyle uyumlu şekilde çalışabilmeli
- Distribution transparency (Dağınıklık şeffaflığı)
  - Dağınıklığın varlığı mümkün oldukça kullanıcıdan saklanmalı
- Fault tolerance (Hata payı)
  - Bir bileşenin bozulması (kısmi bozukluk) tüm sistemin bozulmasına sebep olmamalı
- Scalability (Ölçeklenebilirlik)
  - Sistem, artan kullanıcı sayısına rağmen verimli çalışmaya devam edebilmeli
  - Sisteme yeni kaynaklar eklenerek performans artışı sağlanabilmeli. (horizontal vs vertical scaling)

DS Özet

- Dağıtık sistemler her yerde bulunur.
- Internet, dünyanın her bir yanındaki kullanıcıların, her bir yandaki servislere erişimlerini sağlar.
- Kaynak paylaşımı dağıtık sistem kurmaya teşvik eden etmenlerin başta gelenidir.
- DS kurulumu birçok zorluğu beraberinde getirir:
  - Çeşitlilik, Açıklık, Güvenlik, Ölçeklenebilirlik, Hata denetimi, Birlikte çalışma, Şeffaflık.
- Dağıtık sistemler küreselleşmeyi sağlar:
  - Topluluk (Sanal takımlar, kuruluşlar, sosyal ağlar)
  - Science (e-Science) (Bilim)
  - Business (e-Bussiness) (İş)

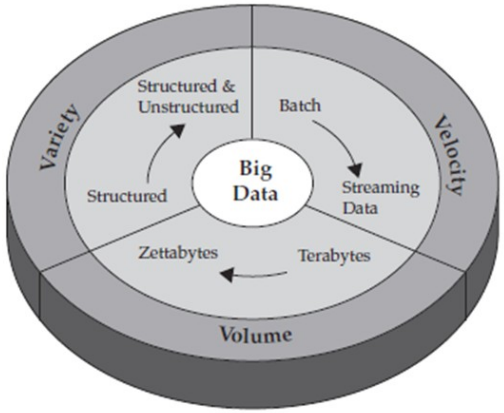
Büyük Veri Nedir?

- “Big Data” is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.

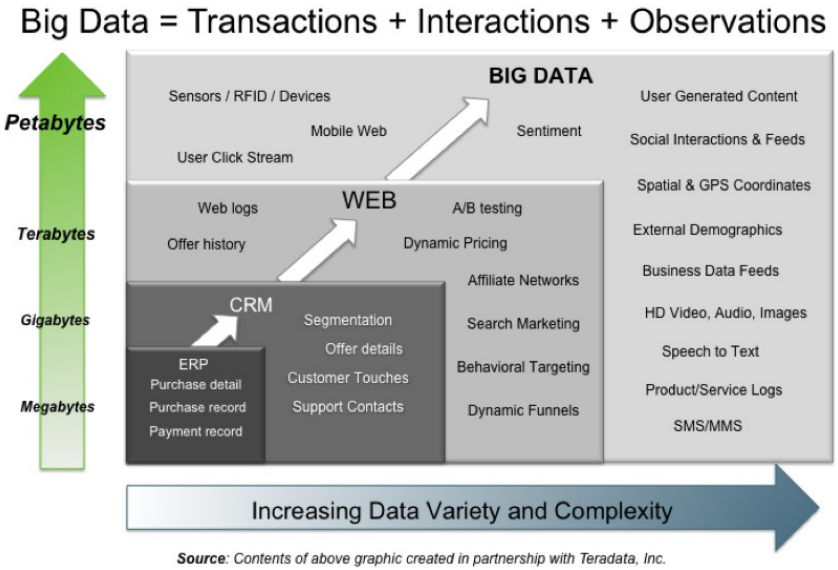
--McKinsey May 2011 article Big Data: The next frontier for innovation, competition, and productivity

Büyük Veri Karakteristikleri

- Volume (Hacim)
- Variety (Çeşitlilik)
- Velocity (Hız)
- Diğer V'ler
  - Veracity (Kalite)
  - Validity (Geçerlilik)
  - Variability (Tutarsızlık)
  - Value (Değer)



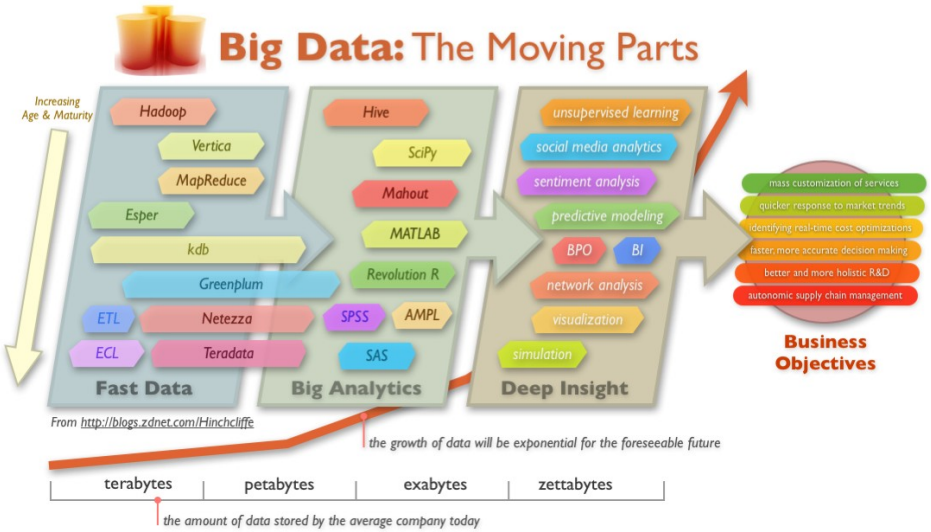
IBM characterizes Big Data by its volume, velocity, and variety—or simply, V³.

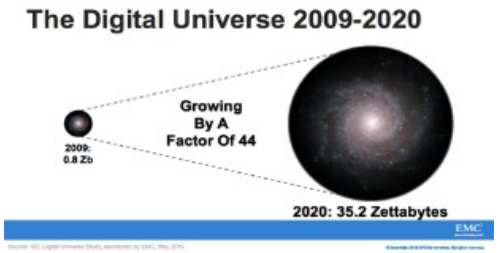


***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Büyük Ne Demek?

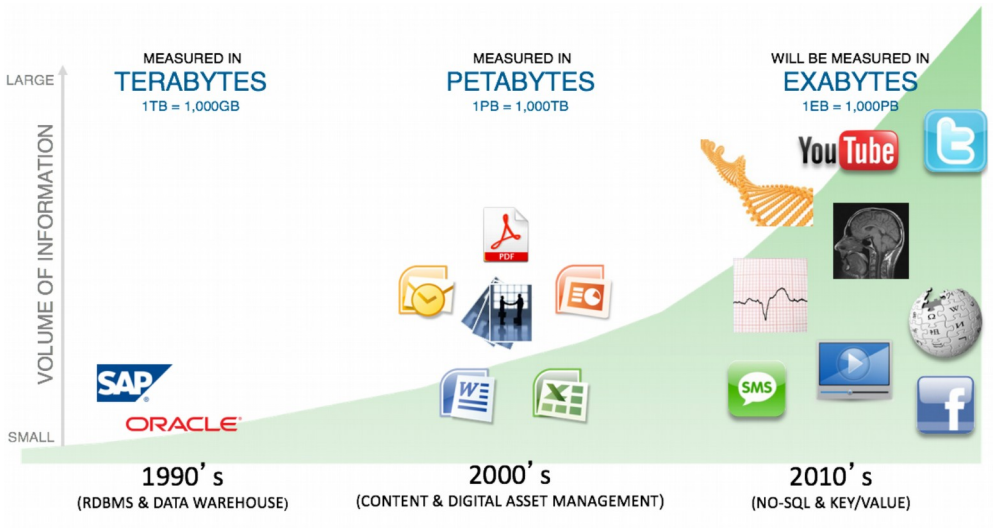
- 1 Bit = Binary Digit
- 8 Bits = 1 Byte
- 1000 Bytes = 1 Kilobyte
- 1000 Kilobytes = 1 Megabyte
- 1000 Megabytes = 1 Gigabyte
- 1000 Gigabytes = 1 Terabyte
- 1000 Terabytes = 1 Petabyte (2^50)
- 1000 Petabytes = 1 Exabyte (2^60)
- 1000 Exabytes = 1 Zettabyte (2^70)
- 1000 Zettabytes = 1 Yottabyte
- 1000 Yottabytes = 1 Brontobyte
- 1000 Brontobytes = 1 Geopbyte
  - Twitter generates more than 8 TB of data every day, Facebook 10 TB.
  - 8 zettabytes (ZB) by 2015 (zetta = 2^70 = 10^21)
  - 35 zettabytes (ZB) by 2020.
  - Her iki yılda ikiye katlanma on goruluyor.





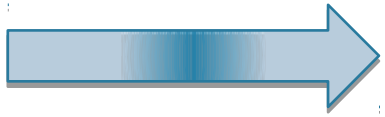
Büyük Veri Nasıl ve Kim Tarafından Üretilir?

- İnsanlar
  - Sosyal medya, web, bloglar, ...
- Sensörler / araçlar
  - Akıllı telefon (GPS), üretim makinaları/robotlar, akıllı sayaçlar (elektrik), arabalar (200+ sensör), ...
- Internet of Things (IoT) (30+ milyar by 2020)
- Web of Things



Büyük Veri Nasıl ve Kim Tarafından Üretilir?

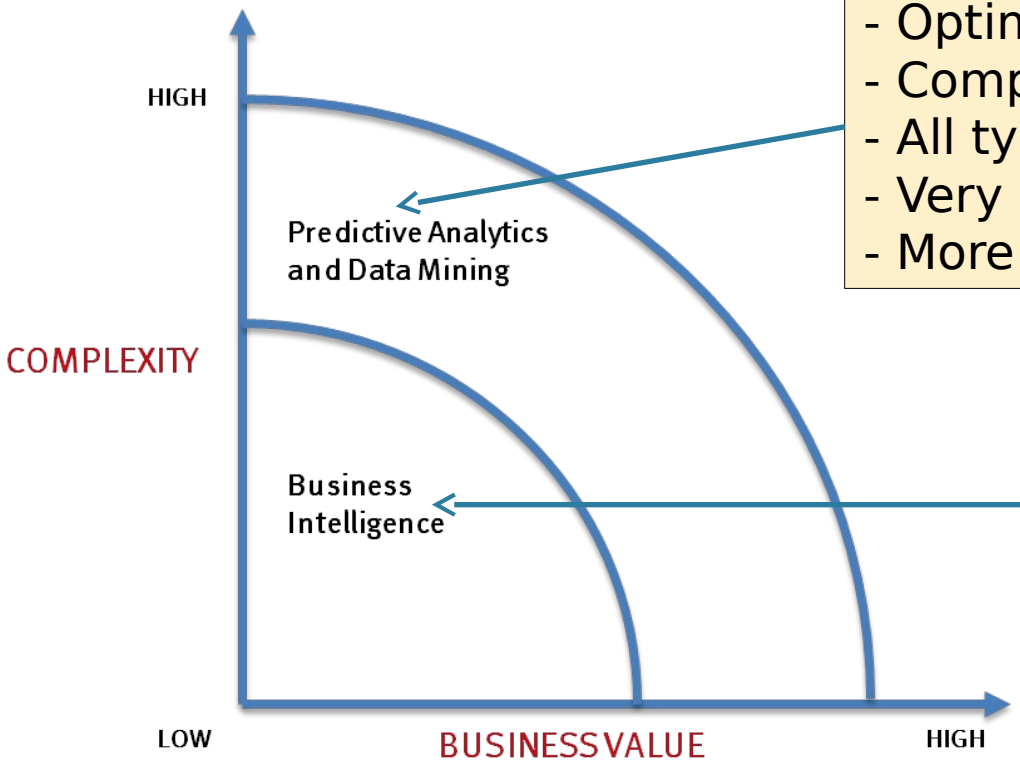
- Verinin üretilme ve kullanılma modeli değişti
- Eski model: birkaç şirket üretsin diğerleri kullansın


- Yeni model: hepimiz üreticiyiz, hepimiz tüketiciyiz



Büyük veri neyi yönlendiriyor?



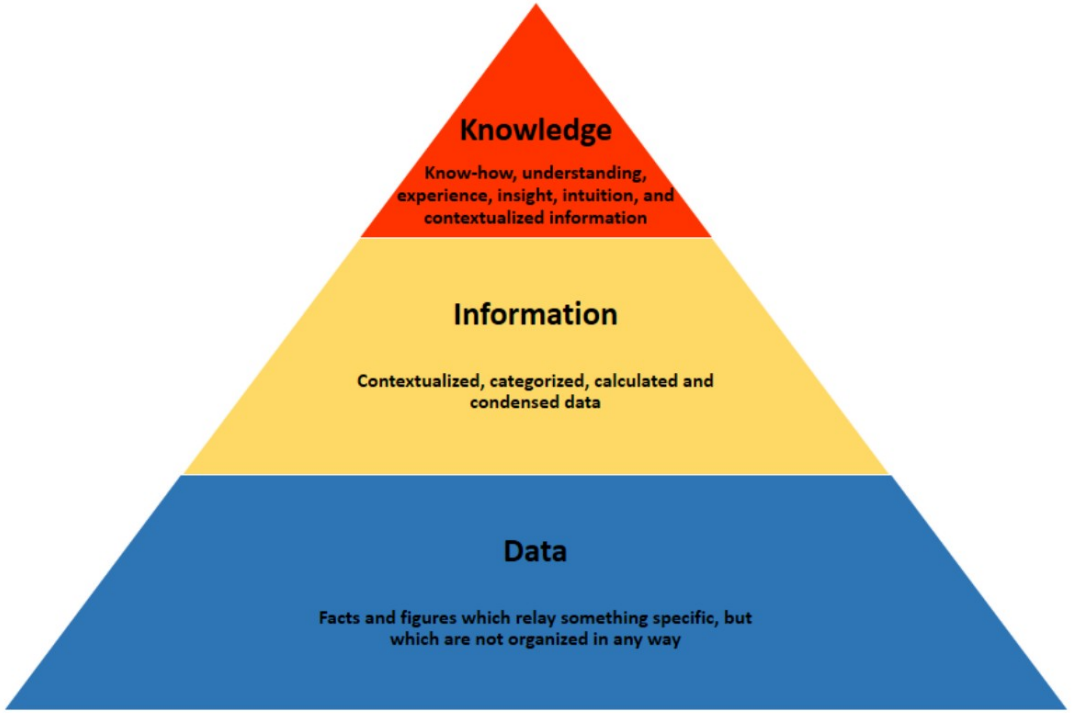
- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Büyük Veriyle Çalışmak Niye Zor?

- Klasik veri ambarlarından daha fazla veriye ihtiyaç duyulması
  - logs, Twitter feeds, blogs, customer surveys, ...
- Doğru soruların sorulmasının gerekliliği
  - Verinin kendi başınayken sessiz olması
- Doğru soruların sorulmasına konsantre olabilmek için gereken teknoloji ve organizasyon desteği
- Verinin sürekli olarak büyümeye devam etmesi

Data->Information->Knowledge



- DIKW piramidi, knowledge hiyerarşisi

Veri Mühendisliği / Veri Bilimi

- Data-->Information-->Knowledge
- Örüntü bulma (finding patterns)
- Sınıflandırma (classification)
- Tahmin (predicting)
- Veri madenciliği (data mining)
- İş zekası (business intelligence)
- Büyük veri analizi
- Veri saklanması (data storage, database, data warehouse)
- Bilinen yöntemleri dağıtık veri üzerinde paralel işleme
- Raporlama ve görselleştirme

Büyüyen Bu Verilerle Neler Yapılabilir?

- Bir çok alanda sınırsız sayıda şey...
  - Perakende Sektörü
  - Sosyal Medya
  - Müşteri İlişkileri
  - Telekom
  - Kişiler Arası İlişkiler
  - İstihbarat
  - Emniyet
  - Medya ve Eğlence Dünyası
  - Sağlık Hizmetleri
  - Yaşam Bilimleri
  - ...

Daha Fazla Ürün Satışı

- Yılın en çok alış veriş yapılan döneminde çok popüler olan bir bilgisayar oyununun yeni sürümü piyasaya çıkıyor olsun.
- Oyunu almak için gelen müşteriye ‘yok satmak’ mağazanın hem ciro hem de müşteri kaybetmesine neden olabilir.
- İnternet ve sosyal medya ortamlarında bu oyunla ilgili olarak yer alan bilgilerin analiz edilmesiyle oyunla ilgili tüketici eğilimleri tespit edilebilir.
- Mağazanın eğilimleri tahmin etmesi ve gelebilecek taleplere hazırlıklı olması satışlarını artıracaktır.
  - Örnek: Bu hafta sonu bu oyunun en az 2000 kopyesi satılabilir.

Daha Fazla Ürün Satışı

- İnsanlar yeni iPhone hakkında ne düşünüyor
  - Örneğin yeni iPhone hakkında atılan tweet'ler analiz edilip sınıflandırılabilir.
  - olumlu, nötr ve olumsuz
  - Analiz yapılırken duygular da (emoji) dikkate alınmalıdır.
    - Müthiş!
    - Müthiş ;-)

Müşteri Kaybetmemek

- Bir telekom firması kullanıcılarının davranışlarını yakından takip ediyor olsun.
- Herhangi bir kullanıcı rakip telekom firmalarından birinin web sitesini ziyaret ettiğinde
  - Operatör değiştirmek istiyordur.
  - Aslında sadece telefonunu değiştirmek istiyordur.
- Telekom firması bu durumu tespit edebilirse
  - Müşteri kaybına neden olabilecek sorunları giderebilir.
  - O müşterinin ilgilendiği telefonu içeren bir teklif sunabilir.

Bu da İlginizi Çekebilir Diyebilmek

- Sanal mağazalarda yapılan gezintileriniz kayıt altına alınmaktadır.
  - Çerezler yoluyla olduğu kadar analitik sunucularına veri gönderilmesiyle de yapılmaktadır.
- Müşteriye sunulabilecek öneriler
  - Benzer gruptaki (yaş, eğitim düzeyi vb.) müşterilerin ilgilendiği ürünler olabilir.
  - Aynı ürünle ilgilenen kişilerin ilgi duyduğu başka ürünler olabilir.

Bu da İlginizi Çekebilir Diyebilmek-2

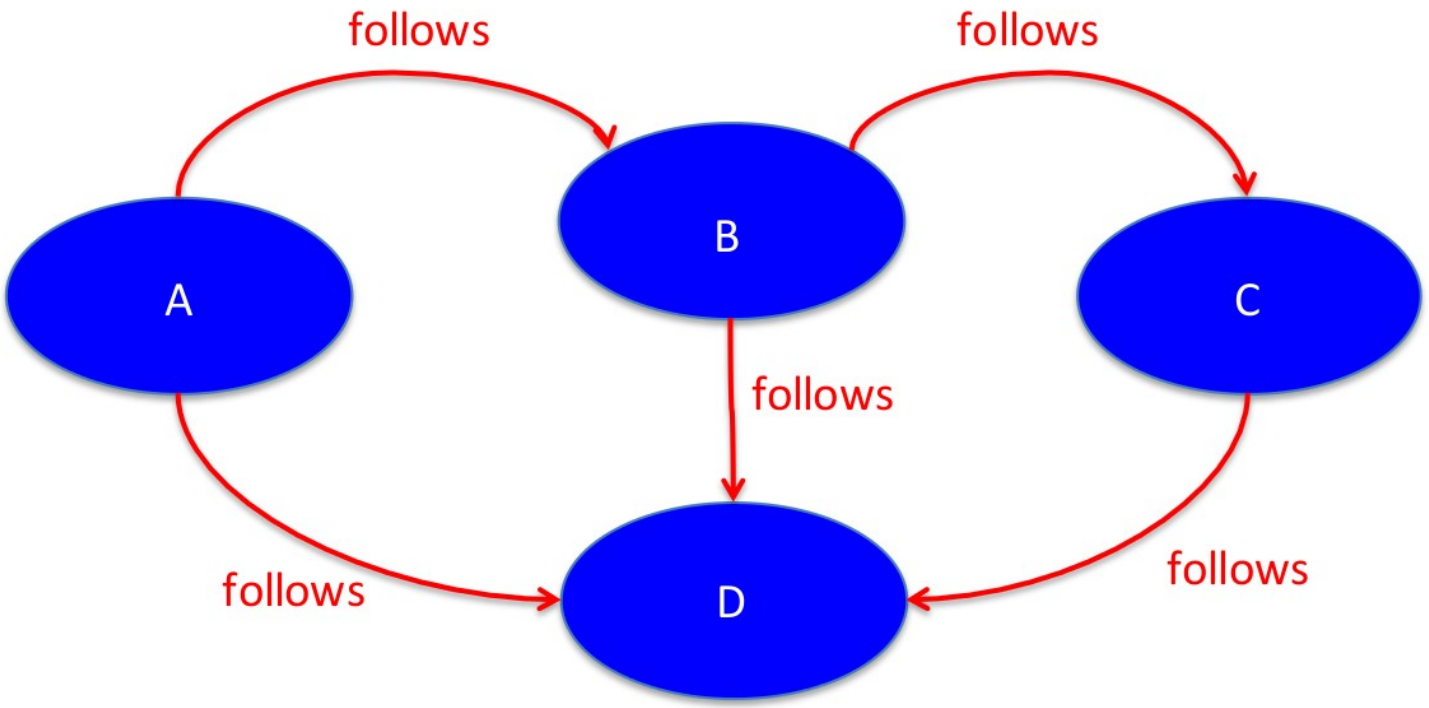
- Peki öneri ne zamana kadar geçerli olmalı?
- Elmas küpe almış (?) bir kullanıcıya başka ne önerilmeli?
  - Ayrıca, gerçekten satın alıp almadığını biliyor muyuz?
- Eğer aldıysa elmas küpe aldığı bu kişiye sürekli olarak hatırlatılmalı mı?
  - Ne kadar para harcadığını hatırlatarak o kişinin canını yakmak insani bir hareket mi?
- Ya da saçları seyrek olan bir kişiye sürekli saç çıkartan losyonların reklamını yapmak rencide edici olmaz mı?
  - Facebook profil fotoğrafımıza bakarak saçımızın seyrekliğini tespit ediyor olabilir mi?

Bu Kişiyi de Tanıyor Olabilirsiniz Diyebilmek

- LinkedIn müşterilerine temas kurmak isteyebilecekleri kişileri önermektedir.
- Olası yeni bağlantılar; ortak okul, iş, coğrafi konum ve var olan bağlantılar gibi faktörlere dayanılarak önerilir.
- Sonuç olarak:
  - Tanıyor olabileceğiniz kişiler özelliği LinkedIn'e yeni müşteriler kazandırmıştır.
  - Bu amaçla halihazırdaki kullanıcılara gönderilen iletilere onların siteyi ziyaret etmelerinde %30 daha fazla etkili olmuştur.
  - Benzer özelliği Facebook, Twitter ve Google+ da sunmaktadır.

İlişkilerin Keşfi

- Sosyal medyada kim kimi takip ediyor?

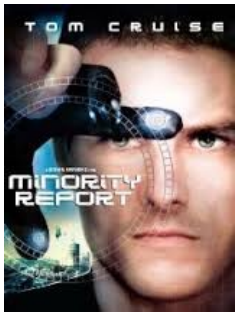


Terörist Ağların Takibi

- Ağ analizinde kullanılan çeşitli metrikler uzunca bir zamandır teorik olarak yerleşmiş haldedir. Örneğin;
  - Yoğunluk: Ağın gücü ya da etkinliği,
  - Çap: Bilginin ağdaki yayılma hızı,
  - Merkezilik: Kişinin popülaritesi,
  - Arasındalık: Köprü konumundaki kişiler,
  - Yakındalık: Diğer kişilere olan mesafe hakkında bilgi verebilir.
- Bu metriklerden yararlanarak terörist ağlar içindeki ilişki yapısı deşifre edilebilir.
- İstihbarat örgütleri normalde terörist ağları takip eder ve onu yok etmek yerine tesirsiz halde tutmaya çalışır.
- Enterne edilmesi halinde ağı çalışamaz hale getirecek teröristler hedeflenir.

Suçla Mücadele

- Minority Report filmi 2002 yılında gösterime girdiğinde suçu işlenmeden önce tahmin etmek ve önlemek belki de bir hayaldi.
- LAPD'nin deprem sonrası oluşabilecek artçı sarsıntıları tahmin eden modele suçla ilgili bilgileri beslemesiyle başlamıştır.
  - 80 yıl geriye giden 13 milyon suça ait veri hazır bulunmaktadır.



- Belirlenen pilot bölgelerde, örneğin yılın hangi günlerinde ne tür suçların işlenebileceği tahmin edilmeye çalışılmaktadır.
  - Polis devriyeleri bu tahminlere göre düzenlenmektedir.

Yeni Tedavilerin Geliştirilmesi

- Tedavilerin etkinliğini karşılaştırabilmek için karmaşık sorguların sorulabilmesi
- Benzer hastalıkları taşıyan hastaların durumlarının karşılaştırılması
- Hastalarla birlikte hastanelerde neler olup bittiğinin anlaşılabilmesi
- Uygulanan başarılı tedavilerdeki benzerliklerden yararlanarak yeni tedavilerin geliştirilmesi

Sağlıkta Şüpheli İlaç Harcamalarının Tespiti

- Bir erkeğe sezaryen yapıldığına dair fatura kesildiği ve bu sanal operasyonun maliyetinin devlete ödetildiği geçmişte görülmüştür.
  - Basit bir veritabanı sorgusuyla bu tip şüpheli durumlar tespit edilebilir.
- Daha karmaşık sorgular (daha nitelikli dolandırıcılıklar) büyük veri kullanarak cevaplanabilir.
  - Örneğin altçizge madenciliği yöntemiyle sürekli tekrarlanan örüntüler tespit edilebilir.
    - Bazı sağlık kurumlarındaki belli doktorların belli kişilere sürekli olarak bazı pahalı ilaçlar için reçete yazması gibi.

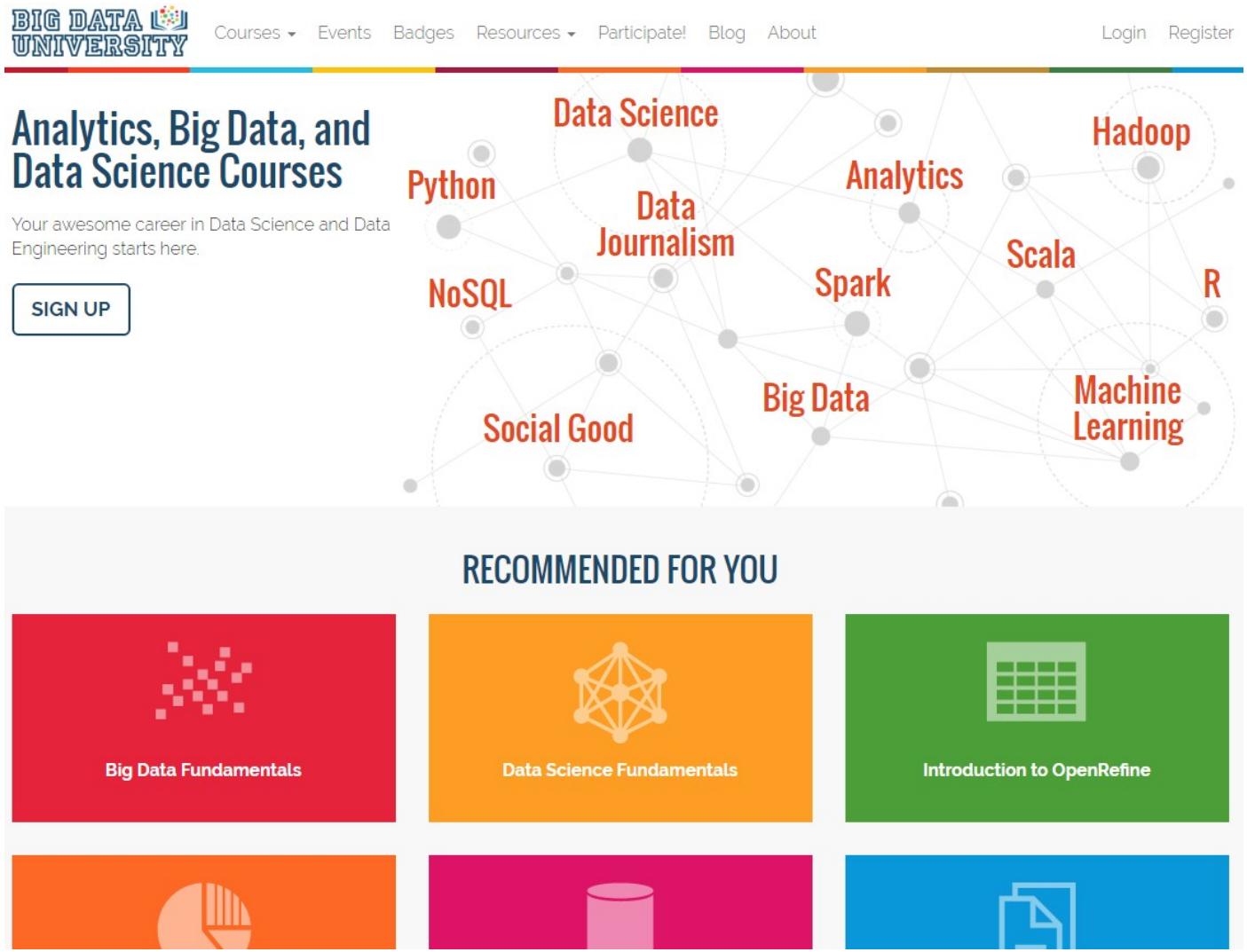
Human Genome Project

- 13 yılda 8 petabyte veri üretilmiştir.
- 30+ yazılım paketi geliştirilip, ilgili genomics verisiyle beraber herkesin kullanımına açılmıştır.
- Kanser ve başka ciddi hastalıklara neden olan hücresel mutasyonlar tespit edilmeye çalışılmaktadır.
- Bu tespitler ışığında yeni ilaçlar ve yöntemler geliştirilmektedir.
- Büyük veri uzmanı açısından bakıldığında olay çoğu zaman string-matching probleminden ibarettir.

Daha Büyük = Daha Akıllı?

- Evet
  - Verideki hatalar tolere edilebilir.
  - Makine öğrenmesi algoritmaları çok daha iyi çalışmaktadır.
- Ama
  - Daha fazla veri daha fazla hata da demektir.
  - Yeterince veri varsa herşey ispat edilebilir.
  - Doğru soruyu sormak için hala insana ihtiyaç duyulmaktadır.

Nereden Başlamalı?



İş İlanları

Sr Data Scientist
General Electric
Kocaeli(Gebze)

BAŞVUR

Qualifications:

GE is the world's Digital Industrial Company, transforming industry with software-defined machines and solutions that are connected, responsive and predictive. Through our people, leadership development, services, technology and scale, GE delivers better outcomes for global customers by speaking the language of industry.

-Master's Degree in a "STEM" major (Science, Technology, Engineering, Mathematics) plus 1 year analytics development for industrial applications in a commercial setting OR Ph.D. in a "STEM" major (Science, Technology, Engineering, Mathematics)
-Demonstrated skill in the use of one or more analytic software tools or languages (e.g., SAS, SPSS, R, Python)
-Demonstrated skill at data cleansing, data quality assessment, and using analytics for data assessment
-Demonstrated skill in the use of applied analytics, descriptive statistics, feature extraction and predictive analytics on industrial datasets
-Demonstrated skill at data visualization and storytelling for an audience of stakeholders

Machine Learning and Artificial Intelligence
Insider
İstanbul(Avr.)(Şişli)

BAŞVUR

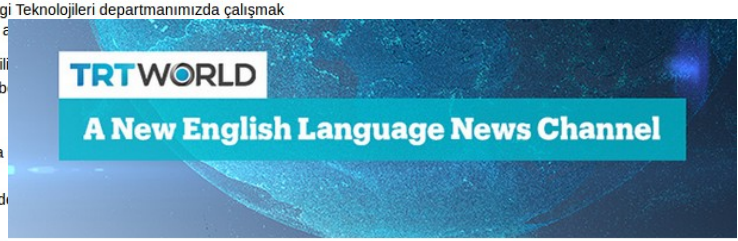
- Proficiency Java, Python or Scala
- 3+ years working as an engineer in a data focused team building machine learning algorithms and predictive models at large scale.
- Experience with big data architectures such as Lambda Architecture.
- Experience working with big data technologies (like Hadoop, Java Map/Reduce, Hive, Spark SQL)
- Experience in Machine Learning frameworks such as Scikit-learn, H2O, Spark MLLib, Prediction.IO
- Self-motivated; independent, organized and proactive; highly responsive, flexible, and adaptable when working across multiple teams.
- Strong SQL skills.
- Experience with recommendation systems is preferred.
- Experience with AWS is favorable.
- PHD is a plus.
- Experience in NoSQL databases (hbase, casandra or elasticsearch)

Genel Nitelikler

2008 yılında kurulan Markafoni.com, cazip indirim fırsatlarıyla son dönemin yerli ve yabancı marka koleksiyonlarını sunan, Türkiye'nin ilk ve öncü kampanyalı satış sitesidir. %90'a varan indirimlerden yararlanan Markafoni üyeleri, giyimden aksesuara, kozmetikten dekorasyon ve ev-yaşama kadar pek çok kategoride seçilmiş markalara ulaşabilir. 7,2 milyon üyesiyle Markafoni, ayda 21,6 milyon ziyaretçi ve 6,2 milyon tekil ziyaretçi almaktadır. Markafoni "VIP Müşteri Servisleri" ile müşteri ilişkilerine odaklanmış; başarılı pazarlama ve iletişim stratejileriyle pek çok ödül kazanmıştır. Birçok kategori altında seçilmiş markaları sunan Markafoni Grubu'nu Türkiye'nin ilk ve öncü kampanyalı satış sitesi Markafoni.com; dünya çapında ünlü parfüm ve kozmetik markalarının online kozmetik ve güzellik mağazası olan Misspera.com ve Türkiye'nin en büyük online ayakkabı mağazası olan Zizigo.com oluşturmaktadır. Markafoni Grubu, 2014 yılında Naspers'ın alt kuruluşu olan MIH Allegro BV tarafından satın alınmıştır. Naspers; internet, ödeme sistemleri, televizyon ve basılı medya alanlarında 130'dan fazla ülkede faaliyet göstermektedir.

İstanbul Maslak'ta bulunan merkez ofisimizde, Bilgi Teknolojileri departmanımızda çalışmak üzere aşağıdaki niteliklere sahip "Veri Madencisi" a

- Üniversitelerin Bilgisayar Mühendisliği veya ilgili
- Tercihen Data mining, Machine learning veya b seviyesinde çalışmaları olan
- En az 5 yıl bu alanda iş deneyimi olan
- Matlab ve R gibi data analiz tools kullanımında
- PostgreSQL RDBMS ve SQL deneyimi olan
- Java, Groovy, Scala veya Python tekonojilerind
- Spark, MLLib veya Scala deneyimi olan
- Agile software development (Pair programming
- İyi düzeyde İngilizce seviyesi olan



QUALIFICATIONS

- Bachelor's Degree in Computer Engineering, Industrial Engineering, Management Information Systems, Statistics, or similar fields
- At least 3 years experience in the Data Science or Analytics field.
- Write SQL queries to gather and combine data from multiple data sources
- Proven experience with data science/analytical tools and programming language, such as Python, Sas, R
- Experienced in using machine learning algorithms
- Demonstrated understanding of the news and media industry
- A passion and commitment to the TRT World spirit of journalism.
- Proficient in English

Sertifika Programları

HOW TO BECOME CERTIFIED

To become a Hortonworks Certified Professional, you need to earn at least one of our certifications:

HDPCD

HDP CERTIFIED DEVELOPER ›

for Hadoop developers using frameworks like Pig, Hive, Sqoop and Flume.

HDPCD-Spark

HDP CERTIFIED APACHE SPARK DEVELOPER ›

for developers responsible for developing Spark Core and Spark SQL applications in Scala or Python.

HDPCD-Java

HDP CERTIFIED JAVA DEVELOPER ›

for developers who design, develop and architect Hadoop-based solutions written in the Java programming language.

HDPCA

HDP CERTIFIED ADMINISTRATOR ›

for administrators who deploy and manage Hadoop clusters.

HCA

HORTONWORKS CERTIFIED ASSOCIATE ›

for an entry point and fundamental skills required to progress to the higher levels of the Hortonworks certification program.

Sertifika Programları-2

Cloudera Certified Professional (CCP)

The industry's most demanding performance-based certifications, CCP evaluates and recognizes a candidate's mastery of the technical skills most sought after by employers.

Certification Overview	Description	Exams	Register for Exam
CCP Data Engineer	CCP Data Engineers possesses the skills to develop reliable, autonomous, scalable data pipelines that result in optimized data sets for a variety of workloads.	1	**Register**
CCP Data Scientist	Named one of the **top five big data certifications**, CCP Data Scientists have demonstrated the skills of an elite group of specialists working with big data. Candidates must prove their abilities under real-world conditions, designing and developing a production-ready data science solution that is peer-evaluated for its accuracy, scalability, and robustness.	3	**Register for DS700** **Register for DS701** **Register for DS702**

Cloudera Certified Associate (CCA)

CCA exams test foundational skills and sets forth the groundwork for a candidate to achieve mastery under the CCP program

Certification Overview	Description	Exams	Register for Exam
CCA Spark and Hadoop Developer	CCP Data Engineers possesses the skills to develop reliable, autonomous, scalable data pipelines that result in optimized data sets for a variety of workloads.	1	**Register**
CCA Data Analyst	A CCA Data Analyst has proven their core analyst skills to load, transform, and model Hadoop data in order to define relationships and extract meaningful results from the raw input.	1	**Register**
Cloudera Certified Administrator for Apache Hadoop (CCAH)	Individuals who earn CCAH have demonstrated the core systems administrator skills sought by companies and organizations deploying Apache Hadoop.	1	**Register**

Referanslar

- T. White, “Hadoop The Definitive Guide: Storage and Analysis at Internet Scale”, allitebooks.com, 4th Edition, 2009
- D. Miner & A. Shook, “MapReduce Design Patterns”, O’Reilly, 2013
- G. Turkington, “Hadoop Beginner's Guide”, PACKT Publishing, 2013
- N. Marz & J. Warren, “Big Data: Principles and best practices of scalable real-time data systems”, Manning, 2015
- Fuat Akal, B3Lab Guz Okulu, Büyük Veri Uygulamaları (Slayt 16-29)