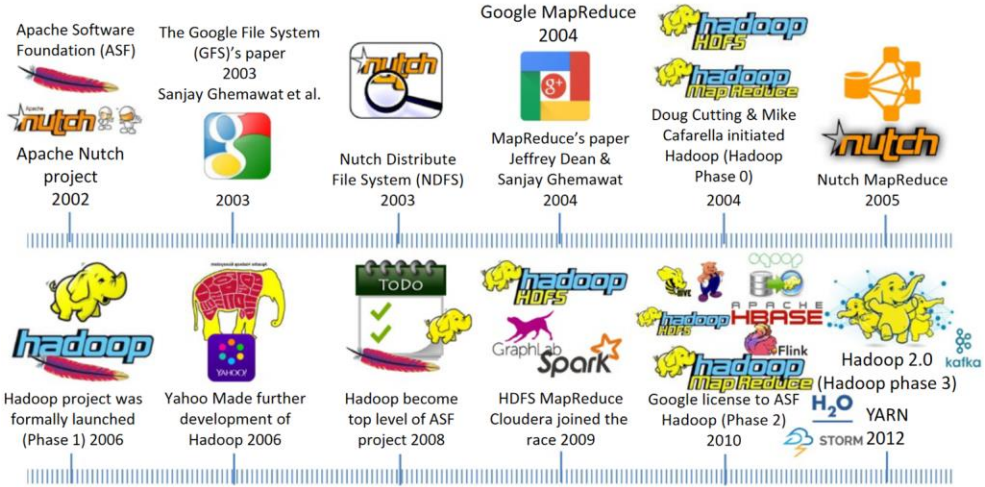
Büyük Veri Analizine Giriş

HADOOP Ekosistemi

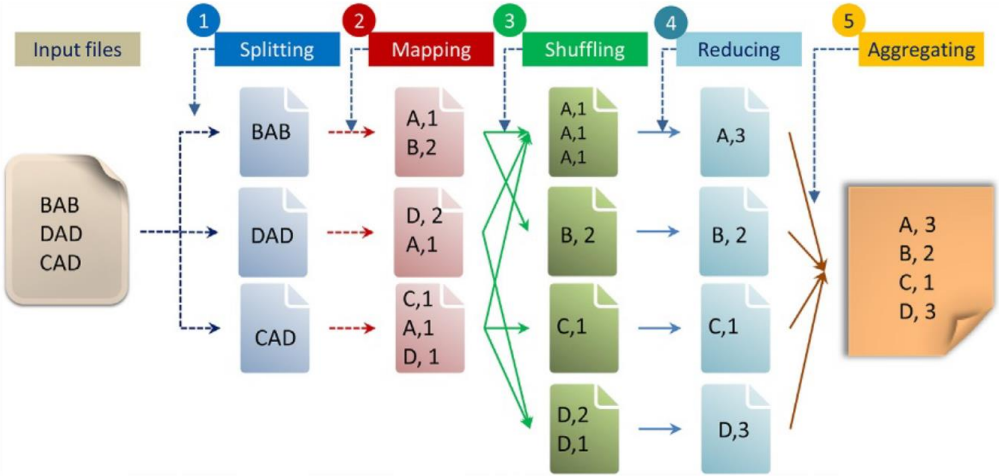
HADOOP

- Dosya depolama işlevi için **HDFS**
- Problemi dağıtık hale getirmek için **haritalama (map)**
- Paralel işleme işlevi için **azaltma (reduce)**
- Temelde Google tarafından geliştirilen **MapReduce** işleme modelini ve dağıtılmış dosya sistemini (**GFS**) kullanır.



MapReduce

- **MapReduce**, Hadoop Dosya Sisteminde (HDFS) depolanan büyük verilere erişmek için kullanılan Hadoop çerçevesindeki bir programlama modelidir.
- MapReduce, petabaytlarca veriyi daha küçük parçalara bölerek ve bunları Hadoop sunucularında **paralel olarak işleyerek** eşzamanlı işlemeyi kolaylaştırır. Sonunda, birleştirilmiş bir çıktıyı uygulamaya geri döndürmek için birden çok sunucudan tüm verileri toplar.



Mahout

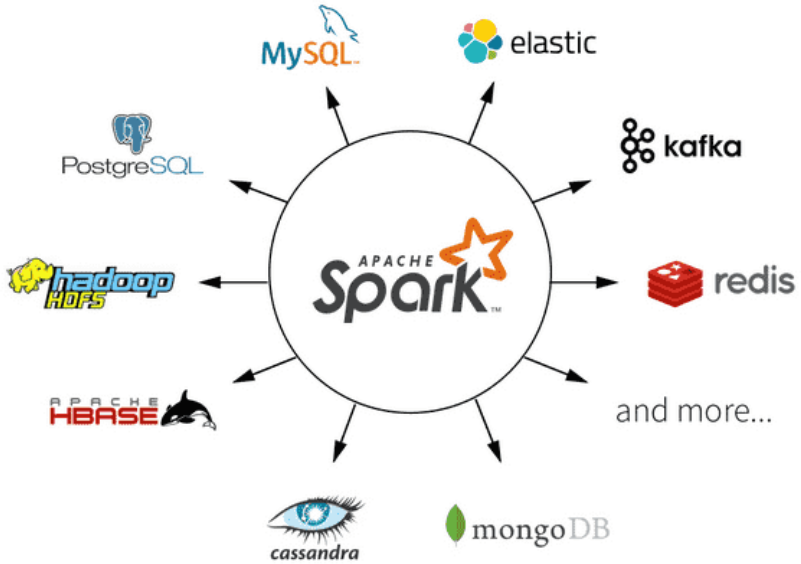
- Mahout'un orijinal amacı, teoride tüm ML algoritmalarını veya tekniklerini kapsayan Java tabanlı bir ML kitaplığı oluşturmaktı, ancak pratikte esas olarak üç tür ML algoritmasını işleyebilir:
  - **İşbirlikçi filtreleme (Collaborative filtering)**
  - **Kümeleme (Clustering)**
  - **Sınıflandırma (Classification)**
- Mahout kullanışlı bir ML kütüphanesi değildir. Hadoop makine öğrenme algoritmalarının iş yükleri için çok yavaştır.
- **MapReduce tabanlı sistemler ML algoritmaları için yetersiz kalmaktadır.**
- Bu sorun Hama, Storm, Spark ve Flink gibi ücretsiz yeni kütüphanalerin geliştirilmesine yol açtı.

SPARK

- Spark, UC Berkeley RAD Lab (AMP Lab) tarafından geliştirilmiştir.
- **MapReduce** modelininin ML algoritmalarında **yetersiz kalması sonucunda** geliştirilmiştir.
- Yinelemeli algoritma gerçekleştirme gibi belirli iş yükü türleri için **MapReduce'dan** 10-20 kat **daha hızlı çalışmaktadır.**
- Apache Spark teknolojisinde veriler bellek içinde saklanmaktadır ve yeniden kullanım sırasında I/O işlemi gerektirmez.
- **Scala, Python(PySpark) ve Java** dilleri için API desteği sağlar.
- Kendisine ait bir dosyalama sistemi bulunmamaktadır.
- **HDFS, GFS ve Cassandra** gibi pek çok dosya sistemiyle çalışabilir.
- Apache Spark 1.3 ile **Spark Sql** yapısı çıkmıştır.

* https://polen.itu.edu.tr/bitstream/11527/12992/1/10112653.pdf

SPARK



tracking_sales

product_name	product_price_per_unit	units_ordered
Desktop Computer	$500.00	5
Monitor	$200.00	5
Telephone	$150.00	2
Telephone	$150.00	3
Chair	$100.00	1

	product_name	product_price_per_unit	units_ordered	revenue
0	Desktop Computer	$500.00	5	2500.0
1	Monitor	$200.00	5	1000.0
2	Telephone	$150.00	2	300.0
3	Telephone	$150.00	3	450.0
4	Chair	$100.00	1	100.0

* https://datatofish.com/sql-to-pandas-dataframe/
https://databricks.com/spark/about

Analiz Türleri

- Analiz türleri 4 ayrı sınıfa ayrılmıştır :
  - **Tanımlayıcı Analiz (Descriptive Analytics)**
    - Geçmişte neler olduğunu analiz eder.
  - **Teşhis Analizi (Diagnostic Analytics)**
    - Geçmişte bir şeyin neden olduğunu anlamanıza yardımcı olur.
  - **Tahmine Dayalı Analiz (Predictive Analytics)**
    - Gelecekte gerçekleşmesi en muhtemel olanı tahmin eder.
  - **Öngörücü Analiz (Prescriptive Analytics)** :
    - Bu sonuçları etkilemek için yapabileceğiniz eylemleri önerir

Tanımlayıcı Analiz (Descriptive Analytics)

- Tanımlayıcı analitik, **tarihsel verilerin toplandığı, düzenlendiği ve daha sonra kolayca anlaşılabilecek bir şekilde sunulduğu**, yaygın olarak kullanılan bir veri analizi biçimidir.
- Tanımlayıcı analitik, yalnızca bir işte ne olduğuna odaklanır ve diğer analiz yöntemlerinden farklı olarak, bulgularından **çıkarımlar veya tahminler yapmak için kullanılmaz**.
- Tanımlayıcı analitik, daha ziyade, daha sonraki analizler için verileri bilgilendirmek veya hazırlamak için kullanılan **temel bir başlangıç noktasıdır**.
- Bulguları sunmak için çizgi, pasta ve çubuk grafikler gibi **görsel araçlar** kullanılır.
- Analizler geniş bir işletme kitlesi tarafından **kolayca anlaşılabilir** olmalıdır.

* https://studyonline.unsw.edu.au/blog/descriptive-predictive-prescriptive-analytics
https://www.logianalytics.com/predictive-analytics/comparing-descriptive-predictive-prescriptive-and-diagnostic-analytics/

Teşhis Analizi (Diagnostic Analytics)

- Teşhis analizi, **"Neden oldu?"** sorusuna yanıt vermek için verileri analiz eden bir analiz türüdür.
- Prosedürlerle tanımlanır, örneğin **veri madenciliği, veri keşfi ve korelasyon tespiti**.
- Davranışların ve olayların nedenlerini anlamak için verileri derinlemesine analiz eder.
- Yapay zeka destekli yazılım ve insan alanı bilgisinin bir kombinasyonuna dayanan teşhiş analizi, **“Neden oldu?”** Sorusunu yanıtlamanın en etkili yoludur.

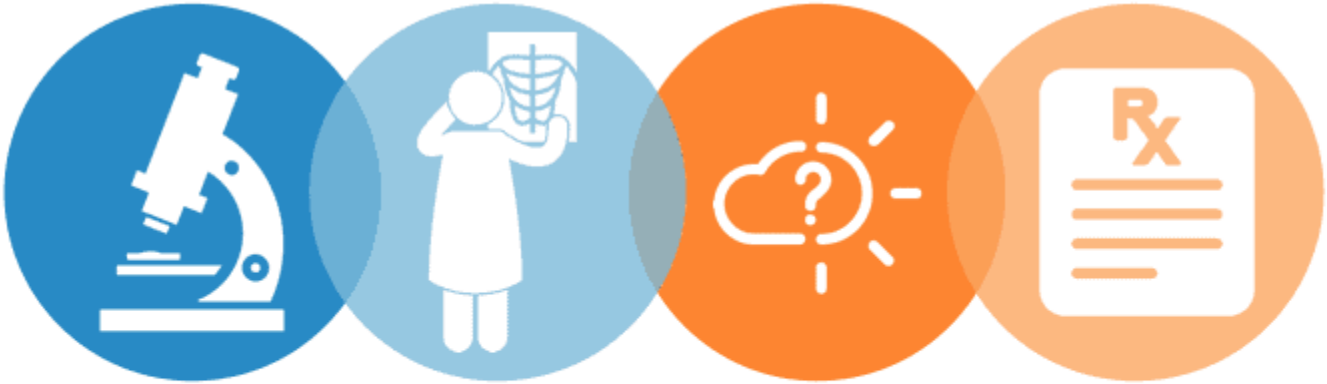
Tahmine Dayalı Analiz (Predictive Analytics)

- Tanımlayıcı analitik ve teşhiş analitiği geçmiş verilere odaklanırken, tahmine dayalı analitik, adından da anlaşılacağı gibi, **gelecekte neler olabileceğini tahmin etmeye** ve anlamaya odaklanır.
- **Geçmiş verilere bakarak** veri modellerini ve eğilimleri analiz etmek ve gelecekte neler olabileceğini tahmin etmek amaçlanır.
- Bunu yaparken, **gerçekçi hedefler belirleme, etkili planlama, performans beklentilerini yönetme ve risklerden kaçınma** dahil olmak üzere bir işletmenin birçok sorununu çözebilir.
- Tahmine dayalı analiz, **olasılıklara** dayanır.
- Veri madenciliği, **istatistiksel modelleme ve makine öğrenimi algoritmaları (sınıflandırma, regresyon ve kümeleme teknikleri)** gibi çeşitli teknikler kullanarak, gelecekteki olası sonuçları ve bu olayların olasılığını tahmin etmeye çalışır.

Öngörücü Analiz (Prescriptive Analytics)

- Tanımlayıcı analiz size ne olduğunu söyler ve tahmine dayalı analitik size ne olabileceğini söyler, öngörücü analiz ise **size ne yapılması gerektiğini söyler.**
- Bu metodoloji, iş analizi sürecinin üçüncü, **son ve en ileri aşamasıdır.**
- İşletmeleri eyleme çağıran, çalışanların kendilerine sunulan verilere dayanarak mümkün olan en iyi kararları vermelerine yardımcı olan aşamadır.
- Öngörücü analiz, bir şeyin ne, ne zaman ve daha da önemlisi **neden olabileceğini tahmin eder**. Her bir seçeneğinin olası sonuçlarını değerlendirdikten sonra, hangi kararların gelecekteki fırsatlardan en iyi şekilde yararlanacağına veya gelecekteki riskleri azaltacağına ilişkin önerilerde bulunulabilir.

Analiz Türleri



Descriptive
Explains what happened.

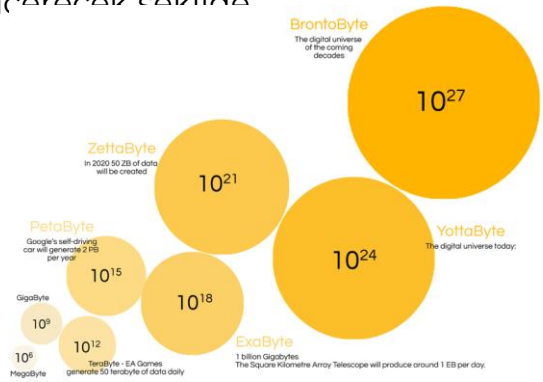
Diagnostic
Explains why it happened.

Predictive
Forecasts what might happen.

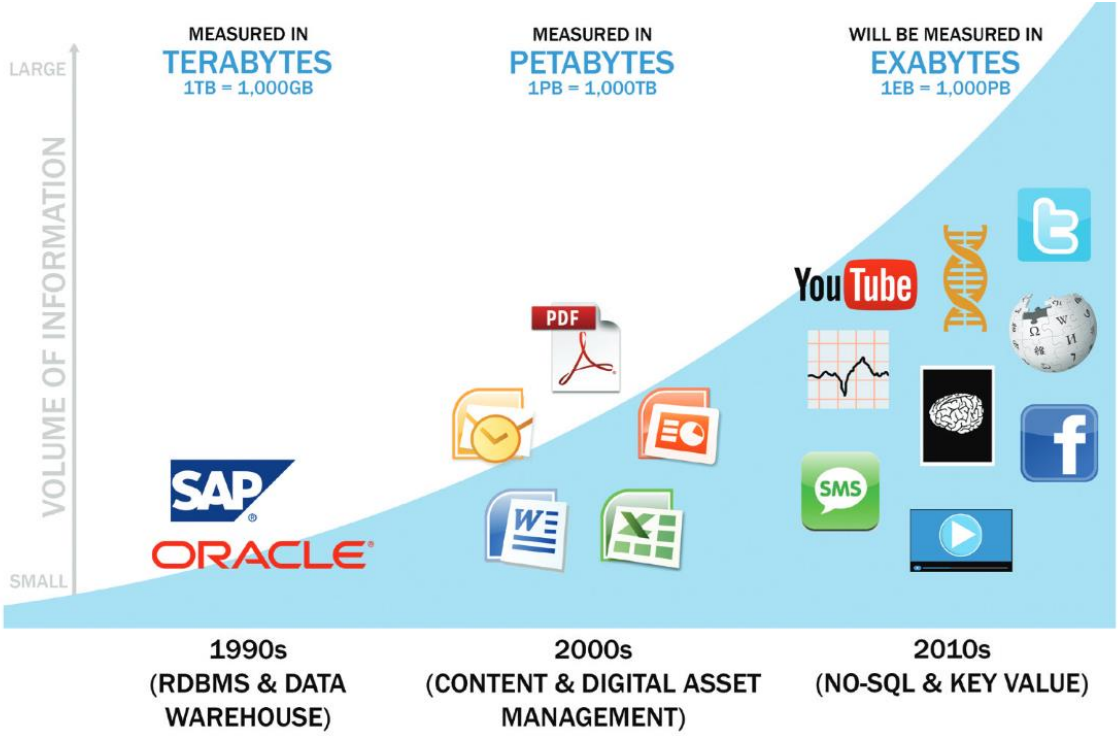
Prescriptive
Recommends an action based on the forecast.

Büyük Veri Kaynakları

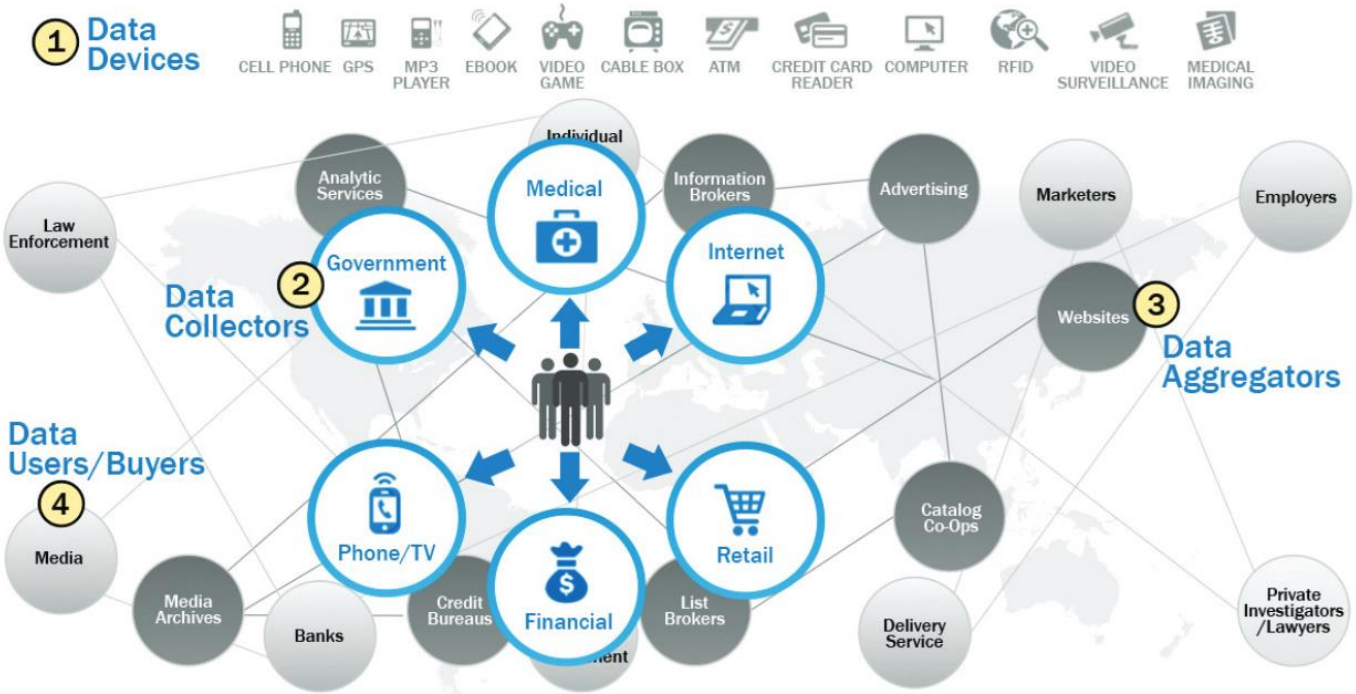
- 1990'larda bilgi hacmi genellikle terabaytlarla ölçülürdü.
  - Çoğu kuruluş, verileri satırlar ve sütunlar halinde(**yapılandırılmış**) analiz etti.
  - Bilgi depolarını yönetmek için ilişkisel veritabanlarını ve veri ambarlarını kullandılar.
- Takip eden on yıl, bu tür bilgileri yönetmek için farklı türde veri kaynaklarının ortaya çıkmasına neden oldu ve verilerin boyutu artmaya başladı. **(petabayt ölçekler)**
- 2010'larda, kuruluşların yönetmeye çalıştığı bilgiler, diğer birçok veri türünü içerecek şekilde genişledi. Bu dönemde herkes ve her şey **dijital bir ayak izi** bırakmaktadır.



Büyük Veri Kaynakları



Büyük Veri Ekosistemi



Büyük Veri Analizi Örneği

Büyük Veri, satış ve pazarlama analitiğini geliştirmek için birçok fırsat sunar. Bunun bir örneği ABD perakendecisi Target'tır. Target'in istatistikçileri, tüketici satın alma davranışını analiz ettikten sonra, perakendecinin üç ana yaşam olayı durumundan çok para kazandığını belirledi.

- **Evlilik**, insanların birçok yeni ürün satın alma eğiliminde olduğu zaman
- **Boşanma**, insanlar yeni ürünler satın aldığında ve harcama alışkanlıklarını değiştirdiğinde
- **Hamilelik**, insanların satın alacak birçok yeni şeyi olduğunda ve bunları satın almak için aciliyetleri olduğunda

- Target, bu yaşam olaylarının en kazançlısının üçüncü durum olduğunu belirledi: hamilelik
- Bu durum, alışveriş yapanlardan toplanan veriler sayesinde tespit edildi sonrasında hangi müşterilerinin hamile olduğunu tahmin eden bir model oluşturuldu.

* https://bigdatacenter.gazi.edu.tr/wp-content/uploads/buyuk_veri_ve_acik_veri_analitigi.pdf

Kaynaklar

- Big data: principles and paradigms. Buyya, Rajkumar, Rodrigo N. Calheiros, and Amir Vahid Dastjerdi, eds. Morgan Kaufmann, 2016.
- Data science and big data analytics: discovering, analyzing, visualizing and presenting data. Wiley, 2015.