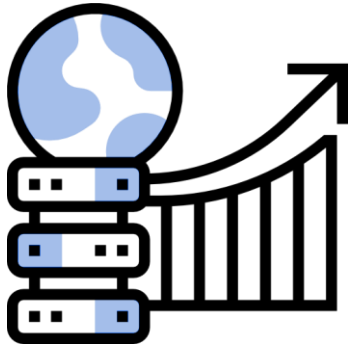
Büyük Veri Analizine Giriş

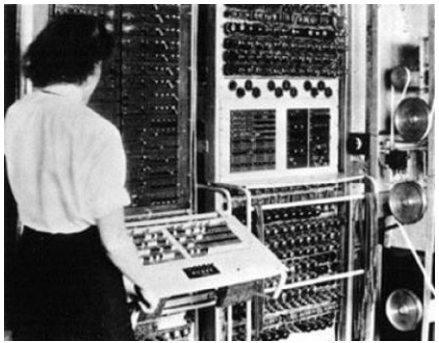
Büyük Veri Nedir?

- Genel bir tanımı yoktur.
  - Büyük Verinin popüler bir tanımı, verilerin üç ana özelliğine dayanmaktadır. (3V's of data).
  - Hacim, hız ve çeşitlilik (volume, velocity, and variety)
  - Bu tanım günümüzde yeterli değildir.
  - Günümüzdeki tanımlar (4V , 6V, 9V) ?



Büyük Verinin Temelleri

- **1663 -** John Graunt İngiltere'de veba sırasında ölüm oranlarını ve varyasyonlarını kaydederek ilk halk sağlığı kayıtları koleksiyonunu yayınladı.
- **1865 -** Richard Millar Devens, "iş zekası" terimini ilk kez kullandı. Bugün anladığımız gibi iş zekası, verileri analiz etme ve ardından eyleme dönüştürülebilir bilgiler sunmak için kullanma sürecidir.
- **1928 -** Fritz Pfleumer, bilgileri teyp üzerinde saklamanın bir yolunu icat etti.
- **1943 -** Birleşik Krallık, İkinci Dünya Savaşı sırasında Nazi kodlarını deşifre eden teorik bir bilgisayar ve ilk veri işleme makinelerinden birini üretti.



Colossus, dünyanın ilk kısmen programlanabilen dijital elektronik bilgisayarı

* https://whatis.techtarget.com/feature/A-history-and-timeline-of-big-data

Büyük Verinin Temelleri

- **1959 -** IBM'de programcı ve yapay zekanın öncüsü olan Arthur Samuel, makine öğrenimi (ML) terimini ortaya attı.
- **1965 -** ABD, milyonlarca vergi iadesini ve parmak izini manyetik bantta saklamak için ilk veri merkezi binalarını inşa etmeyi planladı.
- **1969 -** TCP/IP protokollerini içeren ilk geniş alan ağı olan ARPANET oluşturuldu. Bu, günümüzde internetinin temelini oluşturmaktadır.
- **1990 -** Tim Berners-Lee ve Robert Cailliau, World Wide Web'i geliştirdiler ve veriye çok daha yaygın ve kolay bir erişimin olduğu internet çağı başladı.

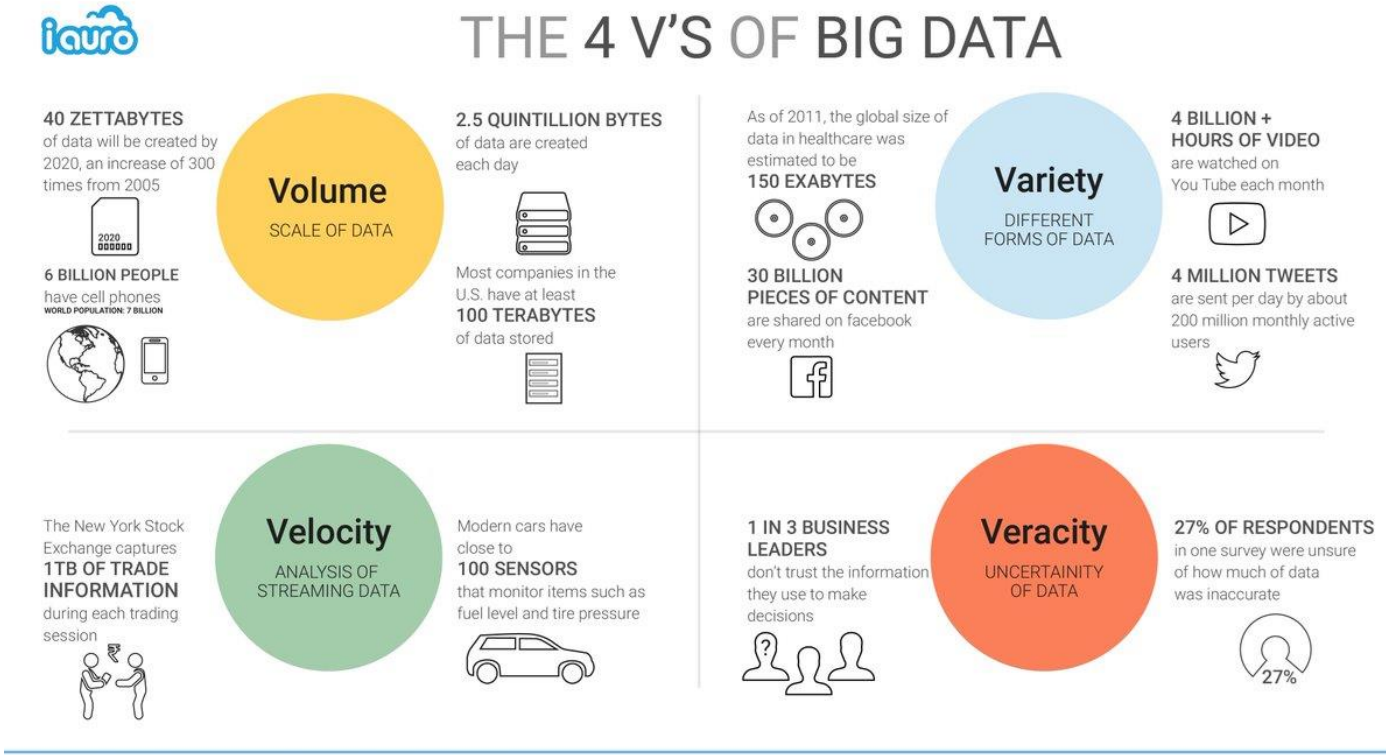
*Tim Berners-Lee CERN laboratuvarlarında **HTML** işaretleme dilini geliştirerek World Wide Web (**www**) olarak tanımlanan bilgi paylaşım sistemini kurdu ve aynı zamanda ilk web tarayıcısı yazılımını geliştirdi.*

* https://whatis.techtarget.com/feature/A-history-and-timeline-of-big-data
https://vizyonergenc.com/icerik/world-wide-web-nedir

Büyük Verinin Farklı Tanımları

- **Douglas Laney — 3V tanımı**
  - Büyük veri tanımlarından ilki 3V tanımı olarak adlandırılır.
    1. **Hacim (Volume)**
    2. **Hız (Velocity)**
    3. **Çeşitlilik (Variety)**
- **IBM — 4V tanımı**
  - IBM, 3V notasyonunun üstüne “Doğruluk” (Veracity) özelliğini ekleyerek yeni bir tanım oluşturdu.
    1. **Hacim (Volume) ->** Veri ölçeği anlamına gelir.
    2. **Hız (Velocity) ->** Akan verilerin analizini inceler.
    3. **Çeşitlilik (Variety) ->** Farklı veri biçimlerini ifade eder.
    4. **Doğruluk (Veracity) ->** Verilerin belirsizliğini ifade eder.

Büyük Verinin Farklı Tanımları



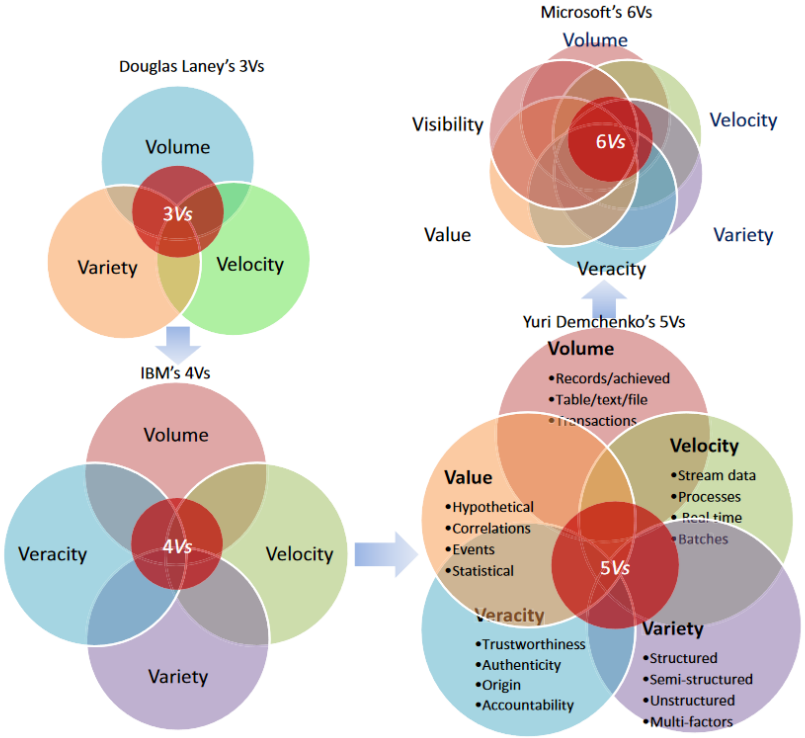
Reference : http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Büyük Verinin Farklı Tanımları

- **MICROSOFT — 6V tanımı**
  - Microsoft iş değerini en üst düzeye çıkarmak adına, 3V notasyonunun üstüne doğruluk (veracity), değişkenlik (variability) ve value (değer) özelliklerini ekleyerek 6V tanımını oluşturdu.

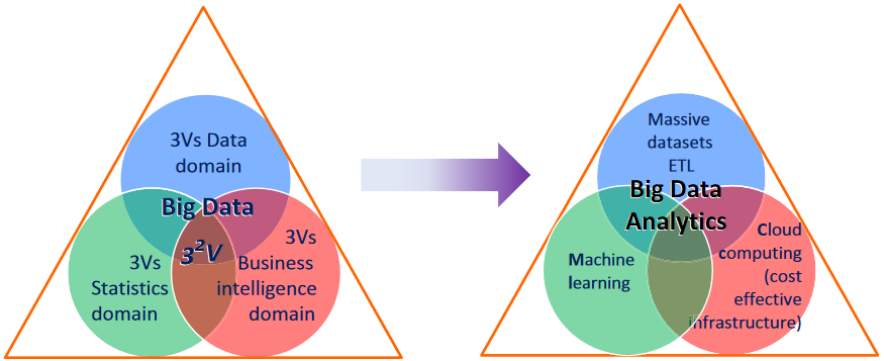
    1. **Hacim (Volume) ->** Veri ölçeği anlamına gelir.
    2. **Hız (Velocity) ->** Akan verilerin analizini inceler.
    3. **Çeşitlilik (Variety) ->** Farklı veri biçimlerini ifade eder.
    4. **Doğruluk (Veracity) ->** Verilerin belirsizliğini ifade eder.
    5. **Değişkenlik (Variability) ->** Veri setinin karmaşıklığını ifade eder.
    6. **Değer (Value) ->** Analitik çözümün ele alması gereken amaç, senaryo veya iş sonucunu ifade eder.

Büyük Verinin Farklı Tanımları

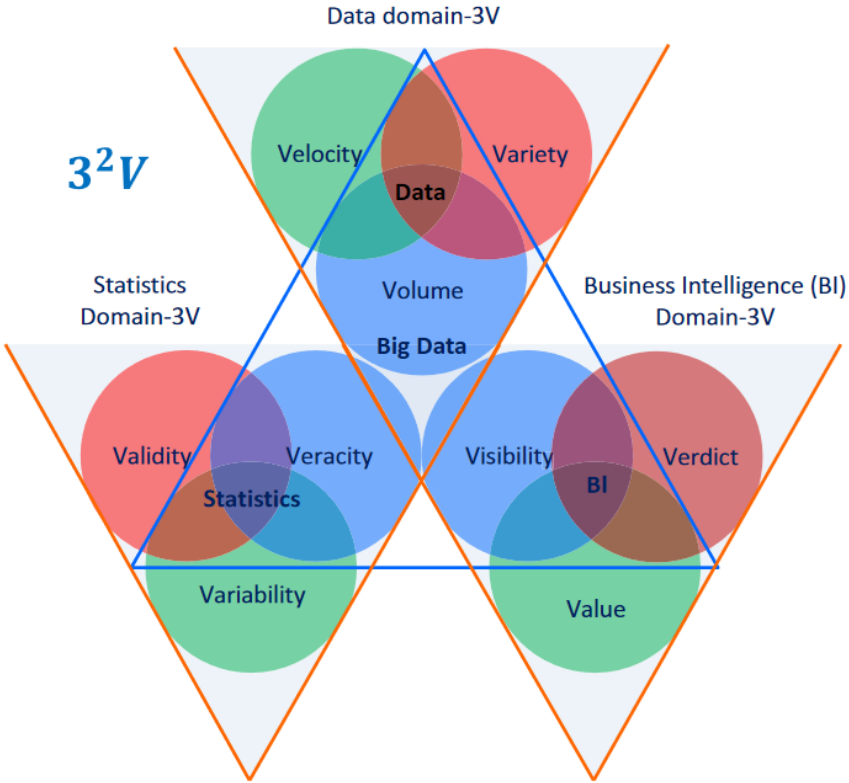


3V'DEN 9V'A BÜYÜK VERİ TANIMLAMA

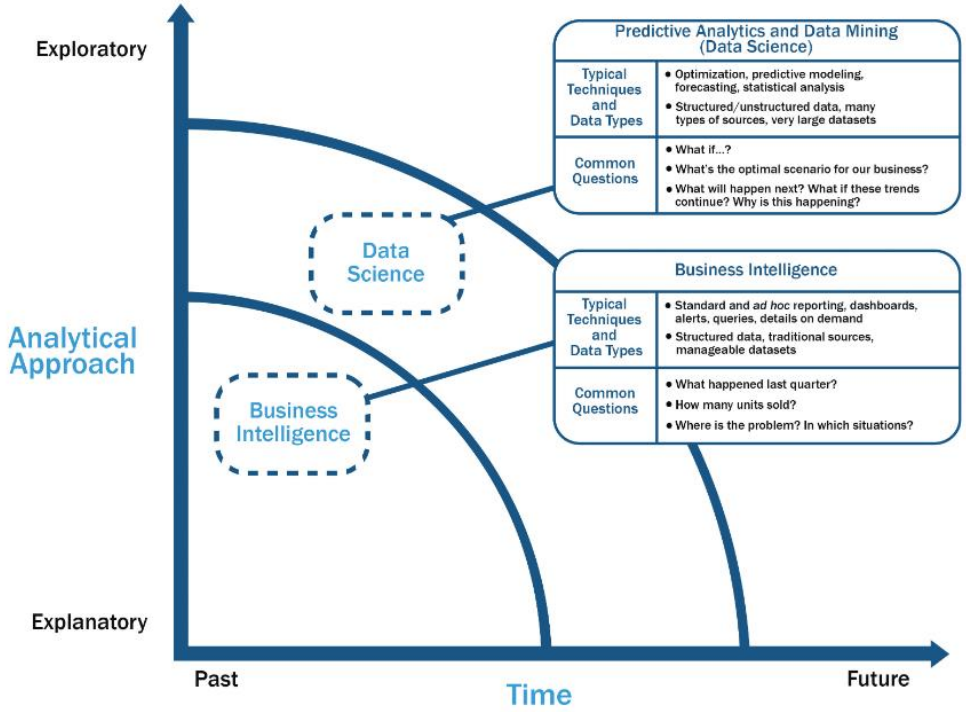
- Büyük veri tanımı 3 ana başlık altında toplanır.
  - **Veri alanı (Data domain)**
  - **İş zekası alanı (Business intelligence domain)**
  - **İstatistiksel alan (Statistical domain)**



3V'DEN 9V'A BÜYÜK VERİ TANIMLAMA



Veri Bilimi - İş Zekası Karşılaştırması



Veri Yapıları

- Birden çok biçimde bulunabilirler :

  a. **Yapılandırılmış veriler** (structured data)

     i. Yapılandırılmış veriler, genellikle bir veritabanında sütunlar ve satırlarla temsil edilen tablo verileridir.

  b. **Yarı yapılandırılmış veriler** (semi-structured data)

     i. Yarı yapılandırılmış veriler, Yapılandırılmış verilerden (ilişkisel veri tabanı) oluşmayan ancak yine de bazı yapıları olan bilgilerdir. (JSON,XML)

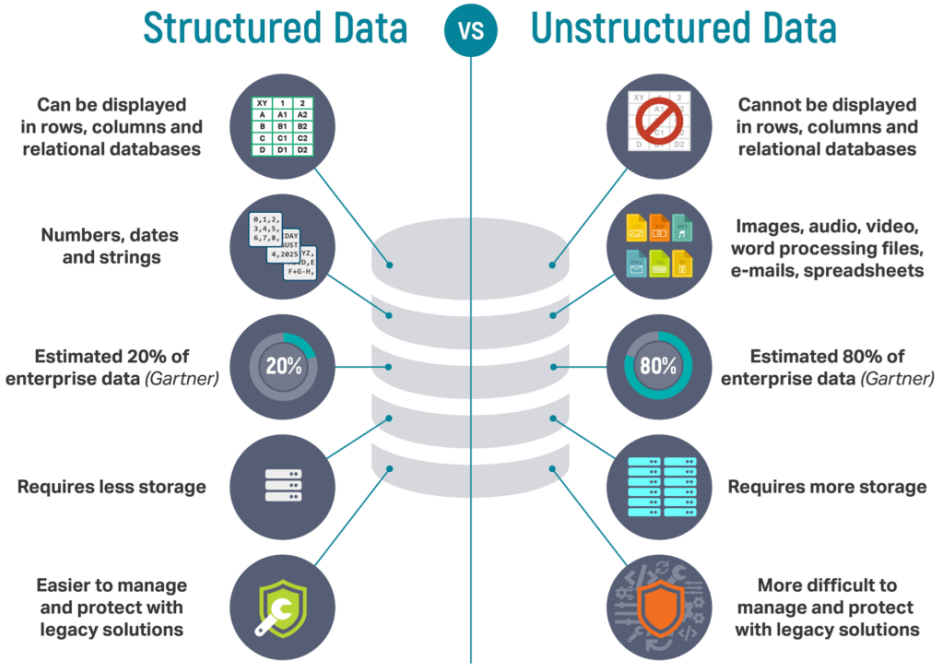
  c. **Yapılandırılmamış veriler** (unstructured data)

     i. Yapılandırılmamış veriler, önceden tanımlanmış bir şekilde organize edilmeyen veya önceden tanımlanmış bir veri modeline sahip olmayan bilgilerdir.

- Büyük Verilerin çoğu yapılandırılmamış veya yarı yapılandırılmıştır.

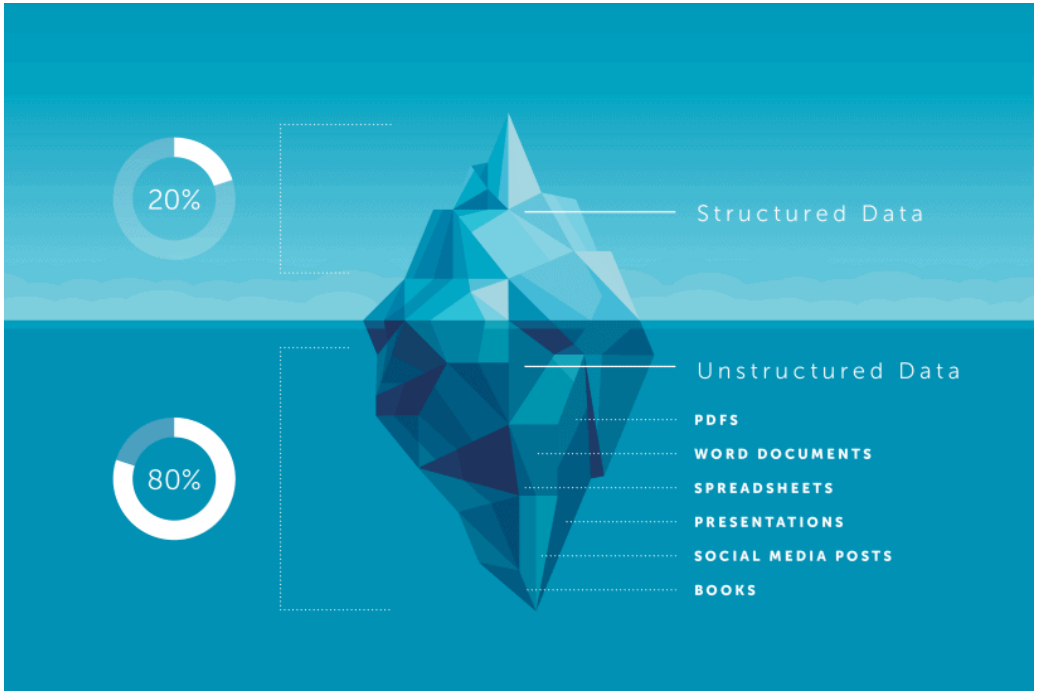
* https://dogus.com.tr/big-data-turleri-nelerdir/#

Veri Yapıları



* https://ichi.pro/tr/buyuk-veri-nedir-ve-bugunun-icgorulerini-nasil-guclendiriyor-131510281516444

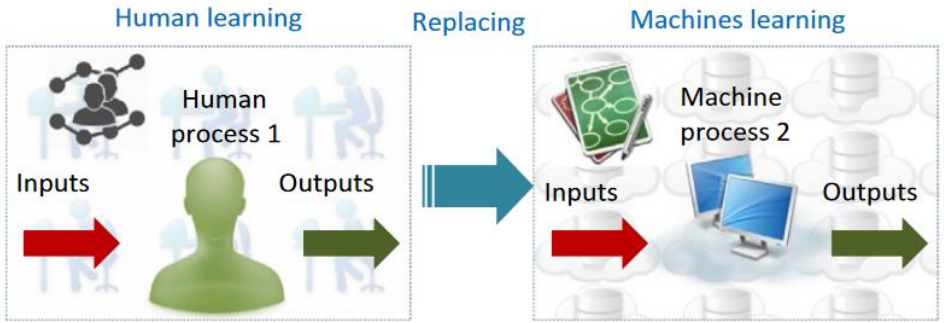
Veri Yapıları



* https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-

Büyük Veri Analizi ve Makine Öğrenmesi

- Arthur Samuel'e göre, ML'nin orijinal tanımı **“bilgisayarlara açık bir şekilde programlanmadan öğrenme yeteneği veren bir çalışma alanı”** dır.
- Makine öğrenmesi veriyi bilgiye dönüştürme sürecidir
- Makine öğrenmesinde nihai amaç karmaşık görevleri yerine getirmede insan yetkinliğine ulaşabilmektir.



Büyük Veri Analizi

- Büyük veri analizinde hesaplama desteği açısından, büyük miktarda veriyi makul bir sürede işleyebilen **dört ana mimari model** vardır.

1. **Analitik Devasa Paralel işleme (MPP) veritabanları** (ör. Greenplum, Netezza,Amazon Redshift)
2. **Bellek içi (in-memory) veritabanları** (Oracle Exalytics, HANA,Spark)
3. **MapReduce işleme modeli** (Hadoop ve Google Dosya Sistemi (GFS) gibi platformlar)
4. **Toplu Eş zamanlı Paralel (BSP) sistemler** (Apache HAMA,Giraph)

HADOOP

- **Hadoop büyük veri analizinde birçok analist ve yazılımcı için ilk tercih olmaktadır. Bunun nedenleri aşağıdaki gibi sıralanabilir :**

1.) Açık kaynaklı bir platformdur ve Java ile programlanmıştır.

2.) Yatay olarak ölçeklenebilir ve güvenilirdir. Donanım arızalarını tolere eder.

3.) Hataya dayanıklı bir sistemdir.

4.) Büyük veri miktarlarını depolamak ve işlemek için pratik bir platformdur.

5.) Çeşitlendirilmiş veri kaynakları için en uygunudur.

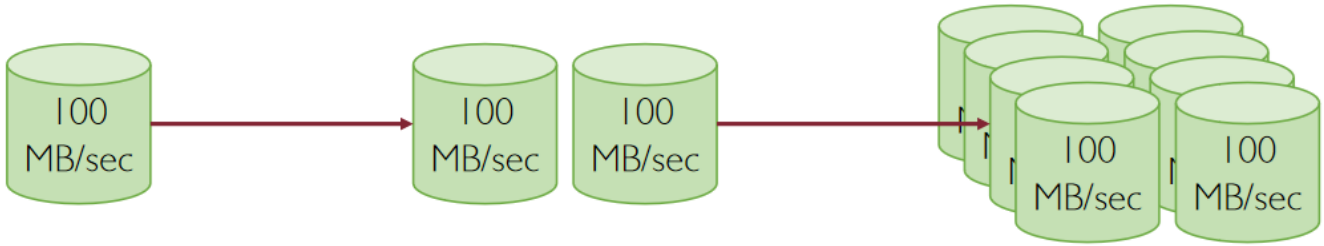
Dikey Ölçekleme

- **Mevcut olan sistemin özelliklerini, daha iyi donanımlar satın alarak genişletmeyi hedefler.**

- **Artıları**
  - Gerçeklenmesi ve uygulanması kolay bir çözümdür.
- **Eksileri**
  - Yeni donanımlar satın olmak maliyeti artırır.
  - Pahalı bir çözümdür.
  - İyileştirme fiziksel olarak sınırlıdır

Yatay Ölçekleme

- **Mevcut olan sistemin özelliklerini, problemin dağıtık çözümlenebilir hale getirilmesi ile aynı güce sahip ucuz donanımlar ile paralel çalışarak genişletmeyi hedefler.**
- **Artıları**
  - Esneklik (sadece gerektiğinde yeni diskler ekleyin)
  - Maliyeti düşüktür.
- **Eksileri**
  - Paralel çalışmayı yönetmek zordur.



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.