Big Data

What is Big Data?

- A collection of so large & complex data sets that become difficult to process on traditional database
- Big data is the data that exceeds the capacity of a traditional database
- challenges
    - data storage & management
    - increasing hardware requirement
    - performance issue

Big Data: 5 Vs

- Volume
- Velocity
- Variety
- Veracity
- Value

Big Data: Volume

- This is related to the amount/quantity of data
- Google
  - process 20 PB each day (2008)
    - 20 PB = 20 000 000 GB
    - 1 PB = 1000 TB, 1 TB = 1000 GB
  - crawl 20 billion pages a day (2012)
  - Search index 100+ PB (May/2014)
- Yahoo
  - 19 Hadoop clusters (we will talk in the course about Hadoop)
  - 600 PB of data (2015)

Big Data: Volume

- Facebook (2014)
  - 300 PB of data in Hive (data warehouse)
  - 600 TB every day
- Internet Archive (2014)
  - 400 billion Web pages
  - +10 PB

Example of Big Data

- Internet Archive Wayback machine
    - largest Web archive in the world
    - 20 years of Web archive
    - ~10 petabytes
- CommonCrawl
    - 7 years of Web crawling
    - available on Amazon S3 as part of the Amazon Public Datasets
    - > each crawl is about 100TB of data

Big Data: Variety

- data comes in different types
  - structured data
    - stored in columns and rows
  - unstructured data
    - email, photos, audios, videos, pdf, Web pages, …

Big Data: Velocity

- is related to the speed at which data is generated
- this is important
    - take actions with low/no latency
    - realtime response
- Examples:
    - On Google
        - 2.5 million queries per second
        - 20 million photos are viewed every second
    - Youtube:
        - 100 hours of videos is uploaded every minute
    - Twitter
        - 300,000 tweets every second

Big Data: Veracity

- is about data quality
- especially in the automated-decision making, where no human is involved
    - you want to make sure that your data & analysis derived from it, is correct

Big Data: Value

- Big data is no good, if nothing useful comes out of it
- Companies are making values from their big data
    - discover customer preferences
    - recommendation based on location & preferences

Value of Big Data

- To change big data into value, we need
  - infrastructure that supports
  - computation power
  - analytic tools that works at scale

Big Data Technologies

- Hadoop echo system
  - storage: Hadoop Distributed File System
  - processing: Hadoop MapReduce, PigLatin, Hive, Spark
- NoSQL databases
  - HBase, MongoDB, and Cassandra
- Data Streaming
  - process data as it comes
    - Kafka

Big Data Sources

- Internet
  - made it possible for every one to generate data;
    - web pages (billion of Web pages),
    - blogs (200 million),
    - emails (300 billion emails /day)
- Social media
  - facebook, youtube, instagram, twitter
- Smartphone
  - cameras, and location awareness (GPS)

The Need for Scalability

- Big Companies have a huge amount of data
  - no single computer can handle it
  - need a cluster of computers
- How many computers do modern services need?
  - Facebook has more than 60,000 servers
  - Google has more than 1 M servers
  - Intel has more than 100,000 servers
  - Microsoft has 200,000 servers

Scale Up

PC

Server

Cluster

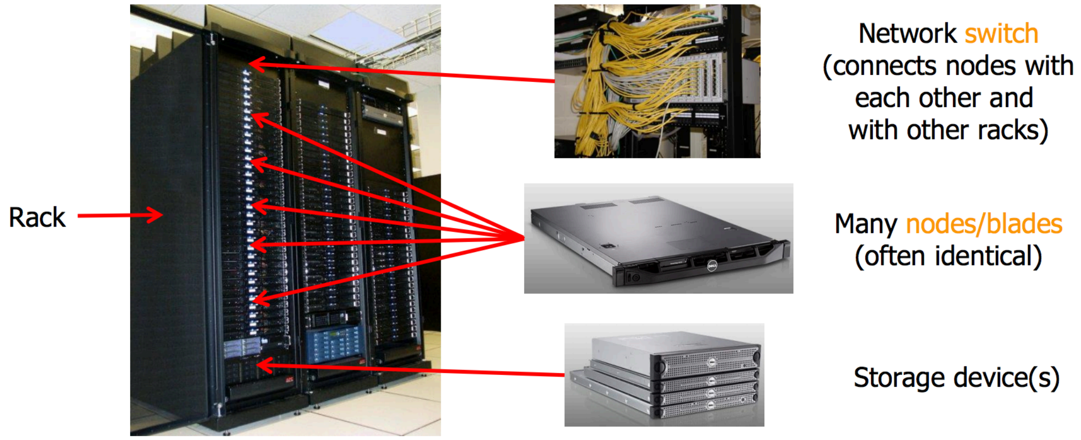
Data center

Clusters

- many similar machines, physically close to each other

- often special, standardized hardware (racks for example)

- usually owned by one organization



Power & Cooling

- Cluster needs a lot of power
    - Example:
        - 140 Watts per server
        - rack of 32 servers needs 4.5 KW
    - most of the power turns into heat
    - requires massive cooling