Chapter 1:

Big Data

Lebanese University

Faculty of Information 1

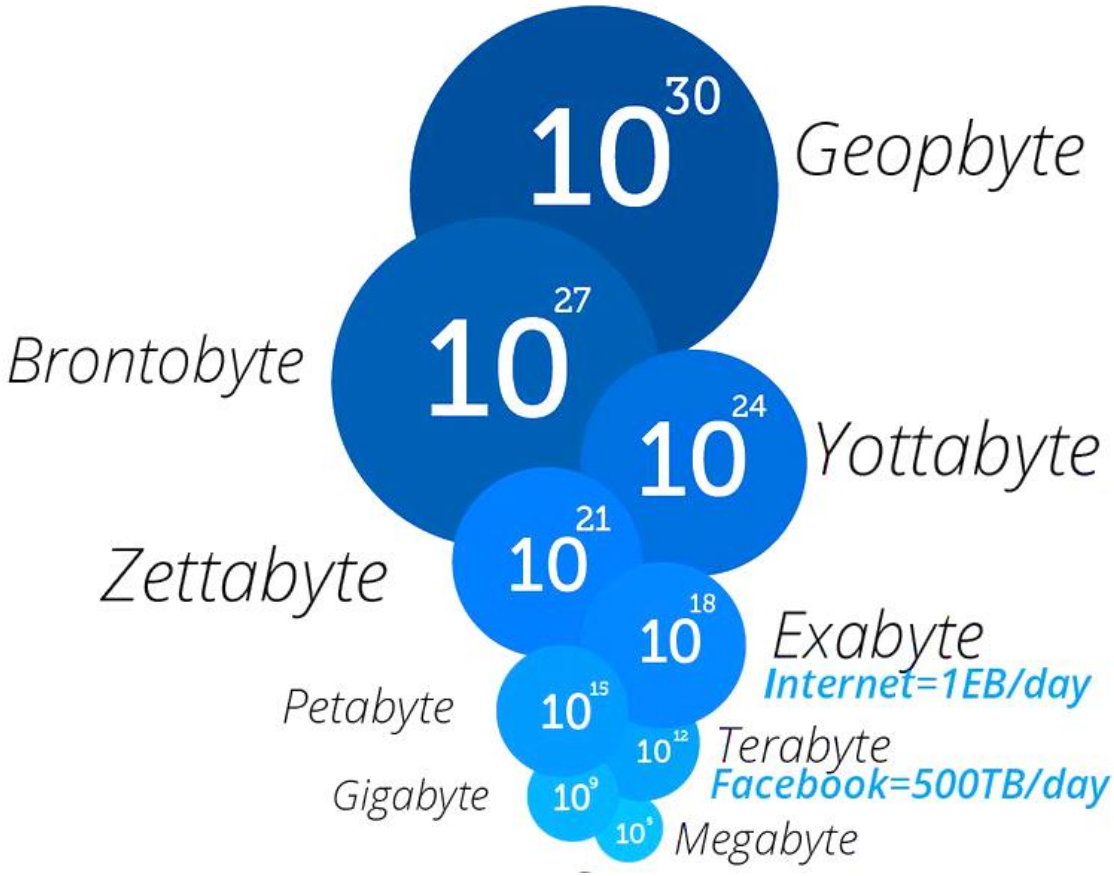
hussein.hazimeh@ul.edu.lb

Dr. Hussein Hazimeh

Outline

- **Data nowadays**
  - Data types, facts, data sources, business decisions, future of data size
- **Big data**
  - What's big data?, how big is the big data?, the Vs properties, big data challenges
- **Big data solutions**
  - Overview to Hadoop, big data landscape, big data companies

Data Unit Measures



Data Types

Structured Data

What you find in a DB
(typically)

Examples: RDMBS, spreadsheet

Unstructured Data

What you find in the 'wild'
(text, images, audio, video)

Exemples: email, documents, images, reports

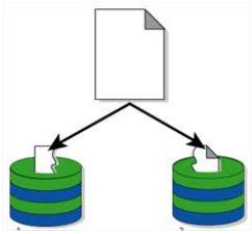
Semi Structured data: XML data

Challenges for Unstructured Data

How do you store Billions of Files?

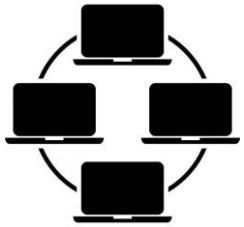
How long does it take to migrate 100's of TB's or data every 3-5 years

Data has no structure



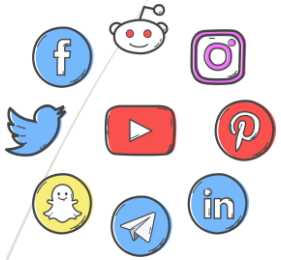
Data Redundancy

Data Backup



Resources Limitation

Sources of Data Generation



Social Media

Sensors

Cell Phones

GPS

Purchase

WWW

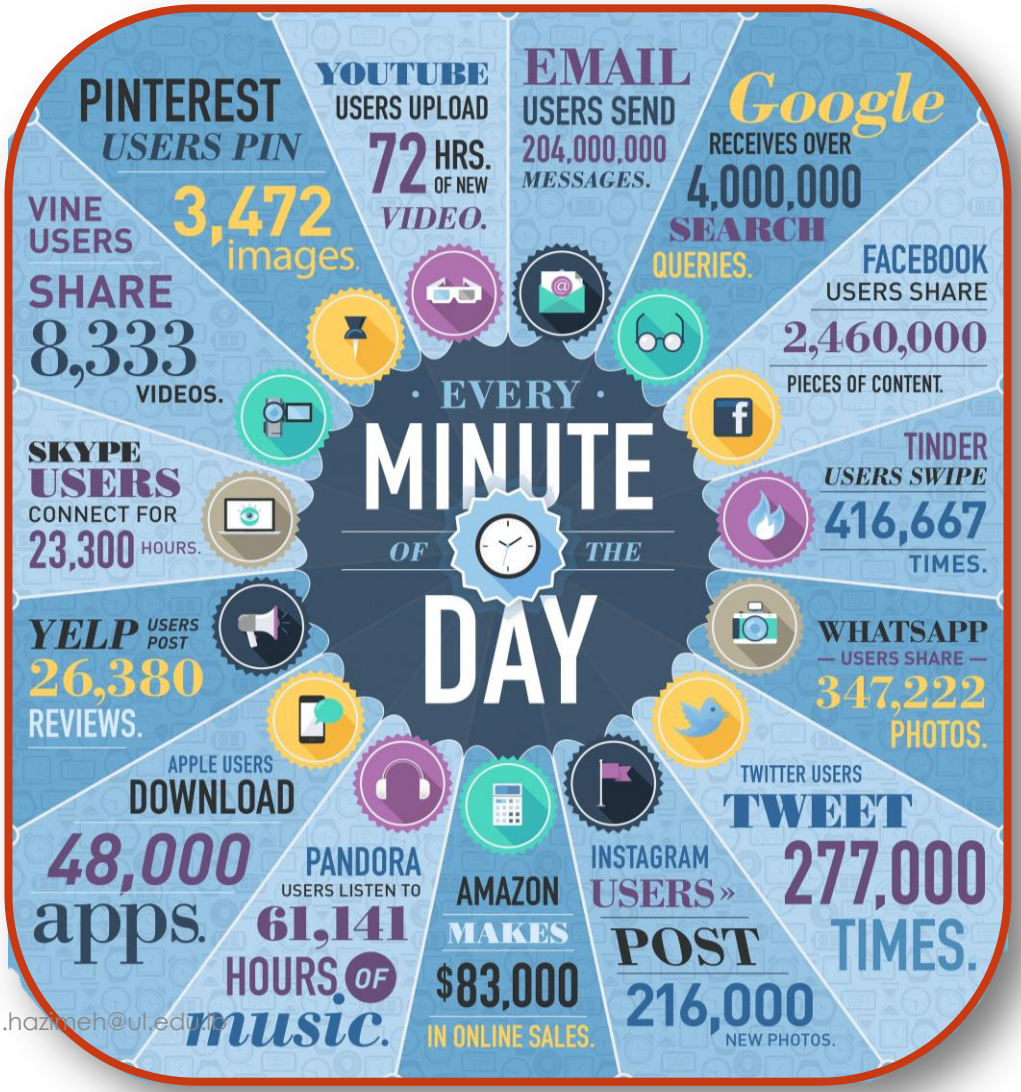
E-mails

Media streaming

Healthcare

IoT

Facts About Data



Facts About Data

70% of data is created by Individuals – but enterprises are responsible for storing and managing 80% of it.

52% of travelers use social media to plan for their vacations.

35% of purchases on Amazon are though recommendations

75% of what people watch on Netflix are recommendations.

Lake Of Data And Business Decisions

- **Business fact:**

1 in 3 business leaders are frequently making business decisions based on information they don t trust or don't have

Can Traditional DBMS Solve This ?

Size

RDBMS finds it challenging to handle huge data volumes.

It needs to add more central processing units (or CPUs) or more memory to the DBMS to scale up vertically.

Data types

RDB can't categorize unstructured data comes from social media (audio, video, texts, and emails).

Velocity

Also, "big data" is generated at a very high velocity. RDBMS lacks in high velocity because it's designed for steady data retention rather than rapid growth.

Cost

Even if RDBMS is used to handle and store "big data," it will turn out to be very expensive.

Arriving to… BIG DATA

What is Big data?

- **Big data** is a term that describes the large volume of data – both structured and unstructured – that generates on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

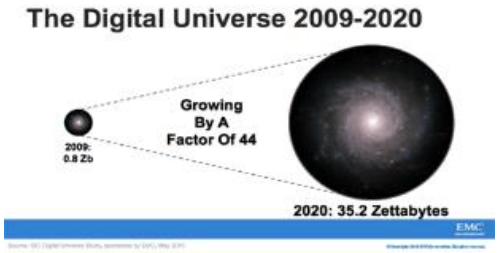
- **Big data** is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation

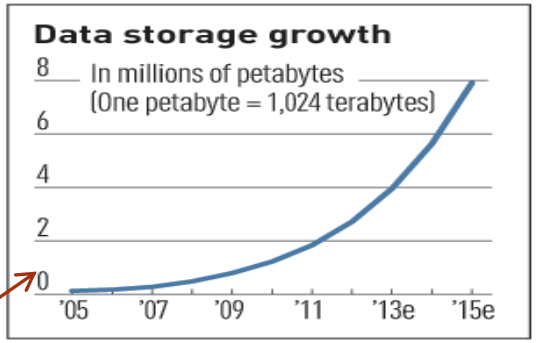
- **Big data** is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy.

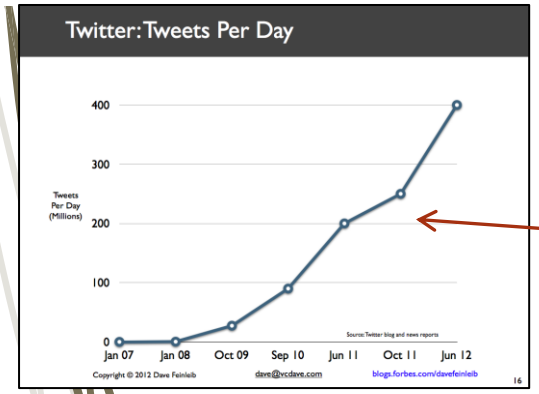
WIKIPEDIA
The Free Encyclopedia

Characteristics of Big Data: 1-Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially



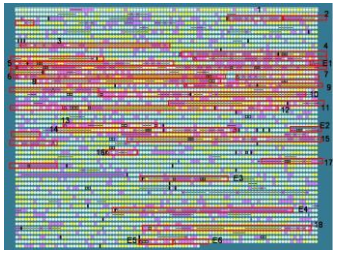


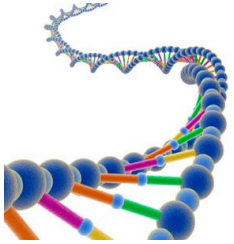


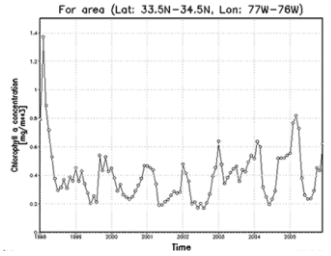
Exponential increase in collected/generated data

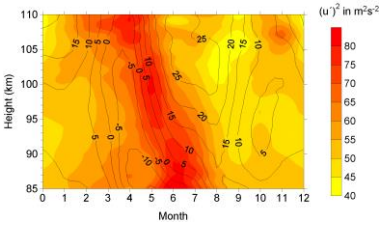
Characteristics of Big Data: 2-Complexity (Varity)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data







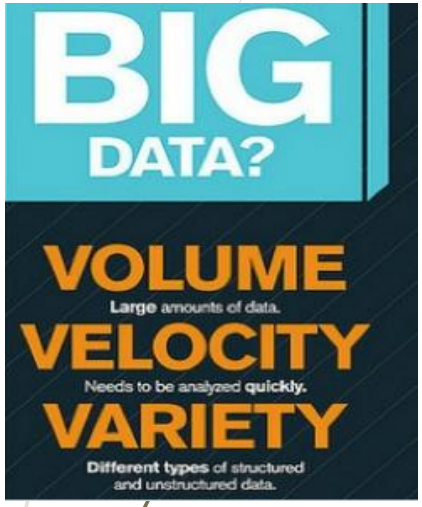


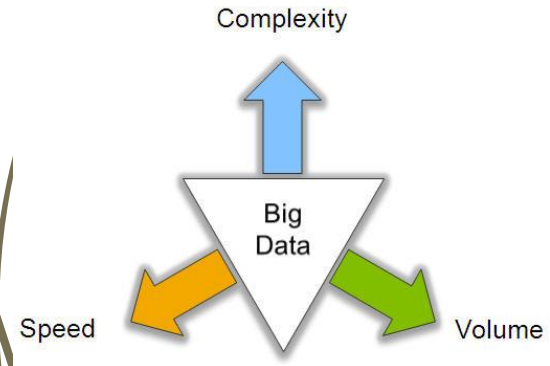
To extract knowledge➔ all these types of data need to linked together

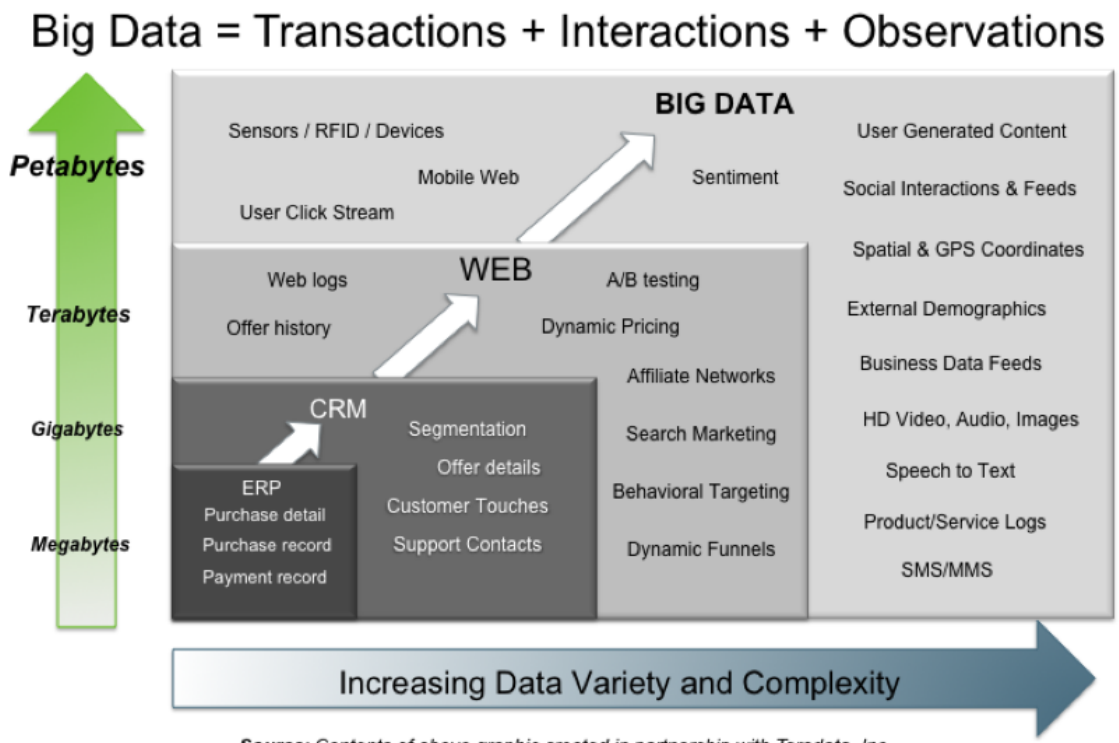
Characteristics of Big Data: 3-Speed (Velocity)

- Data is begin generated fast and need to be processed fast
- Les données sont générées rapidement et doivent être traitées rapidement
- Online Data Analytics/ Analyse de données en ligne
- Late decisions ➔ missing opportunities
- Décisions tardives ➔ occasions manquées
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase you like ➔ send promotions right now for store next to you
  - en fonction de votre emplacement actuel, de l'historique de vos achats, de ce que vous aimez ➔ envoi des promotions instantanés pour un magasin à côté de vous
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction
  - **Surveillance médicale**: des capteurs surveillent vos activités et votre corps ➔ toute mesure anormale nécessite une réaction immédiate

Big Data: 3V's

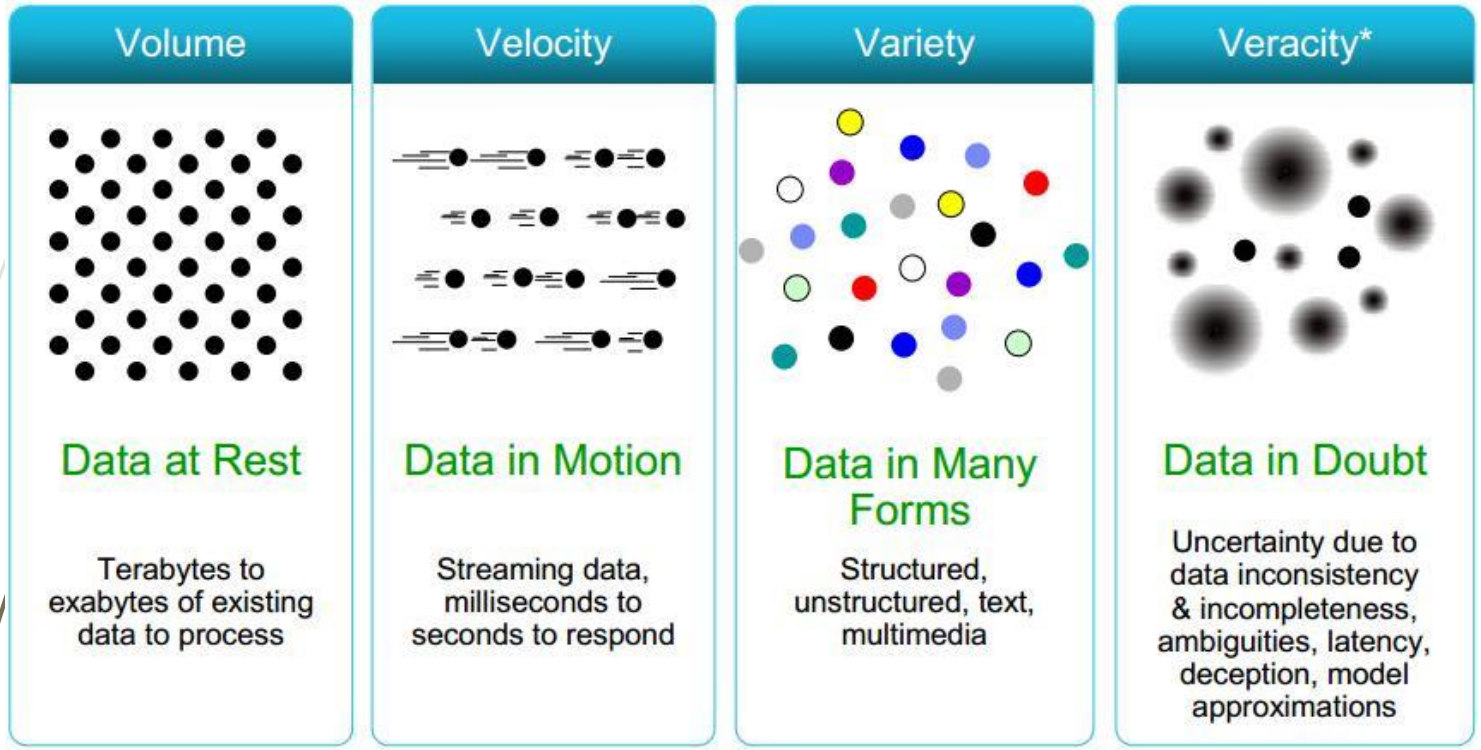




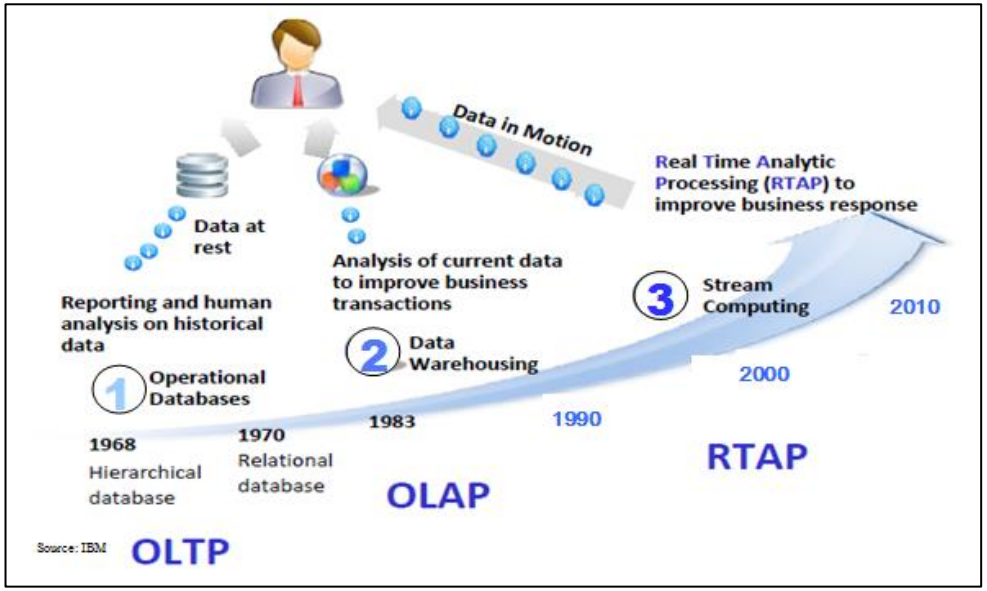


***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Some Make it 4V's



Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Big Data In Action

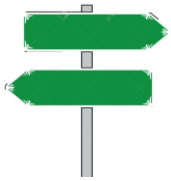
UPS stores a large amount of data – much of which comes from sensors in its vehicles - GPS

Data Analytics and Data Science

the world's largest operations research project

ORION (On-Road Integration Optimization and Navigation)

savings of more than 8.4 million



85 million miles off of daily routes

Saved $30 million/Day

Big Data In Action

Walmart

"We want to know what every product in the world is. We want to know who every person in the world is. And we want to have the ability to connect them together in a transaction."
-Neil Ashe, CEO of Global E-commerce at Walmart

Walmart collects 2.5 petabytes of information from 1 million customers from 6000 stores.

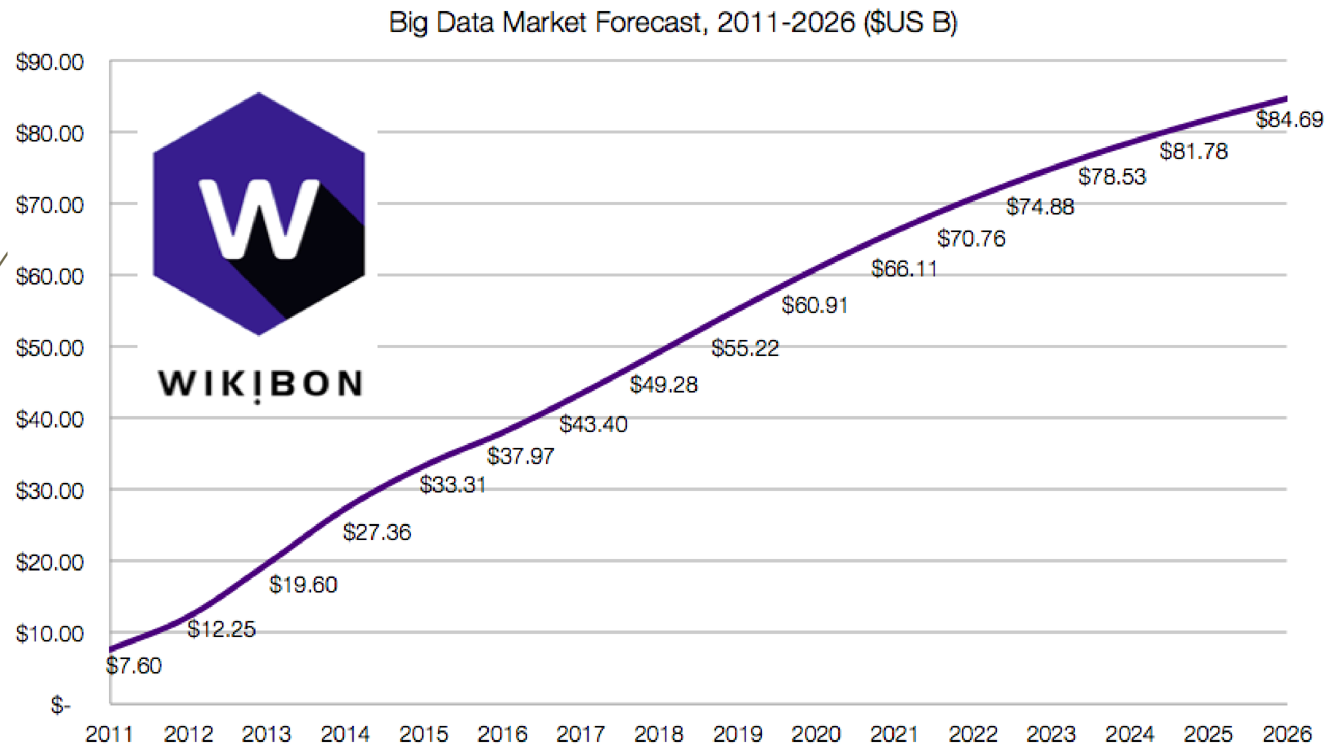
Big data System (Kosmix)

30% on their Online sales

Revenue got increased by 40%

Big Data Market Forecast

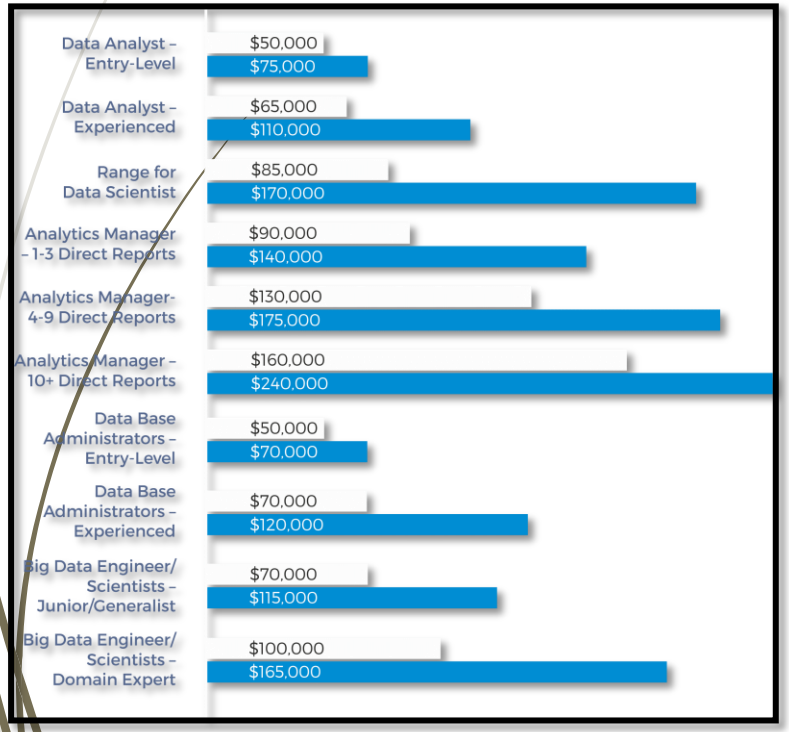
Only 27% of the executives surveyed described their Big Data initiatives as successful.

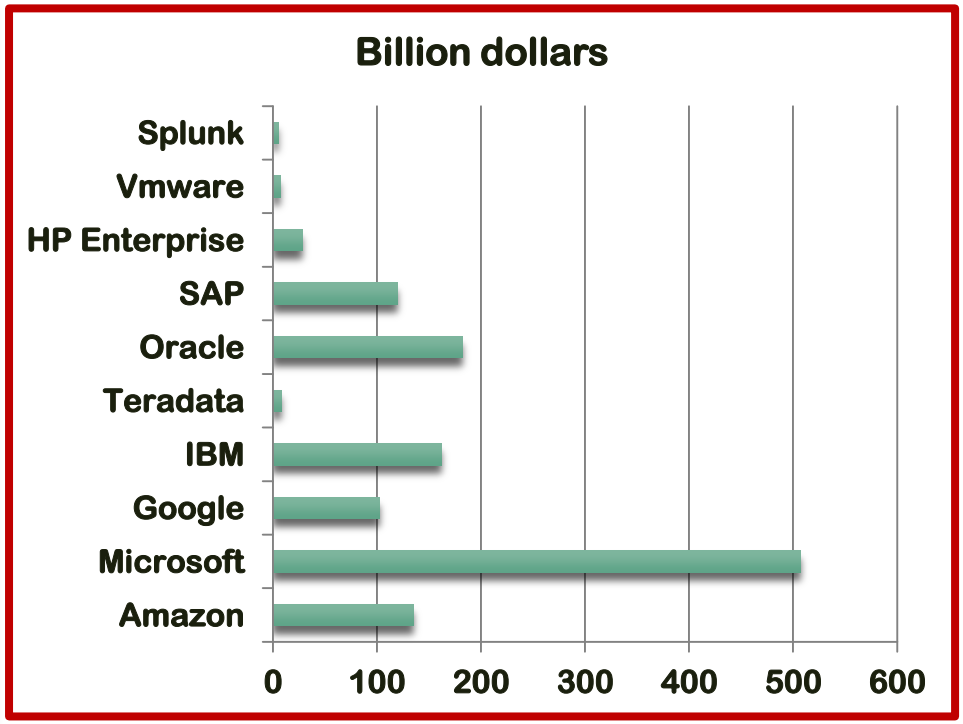


The "big data" market is expected to cross $80 billion by 2026.

Big Data Jobs Trend

- **IBM, Cisco and Oracle together advertised 26,488 open positions that required big data expertise in the last twelve months.**





The Jobs Landscape in 2022

emerging roles, global change by 2022

Top 10 Emerging

1. Data Analysts and Scientists
2. AI and Machine Learning Specialists
3. General and Operations Managers
4. Software and Applications Developers and Analysts
5. Sales and Marketing Professionals
6. Big Data Specialists
7. Digital Transformation Specialists
8. New Technology Specialists
9. Organisational Development Specialists
10. Information Technology Services

declining roles, global change by 2022

Top 10 Declining

1. Data Entry Clerks
2. Accounting, Bookkeeping and Payroll Clerks
3. Administrative and Executive Secretaries
4. Assembly and Factory Workers
5. Client Information and Customer Service Workers
6. Business Services and Administration Managers
7. Accountants and Auditors
8. Material-Recording and Stock-Keeping Clerks
9. General and Operations Managers
10. Postal Service Clerks

Source: Future of Jobs Report 2018, World Economic Forum

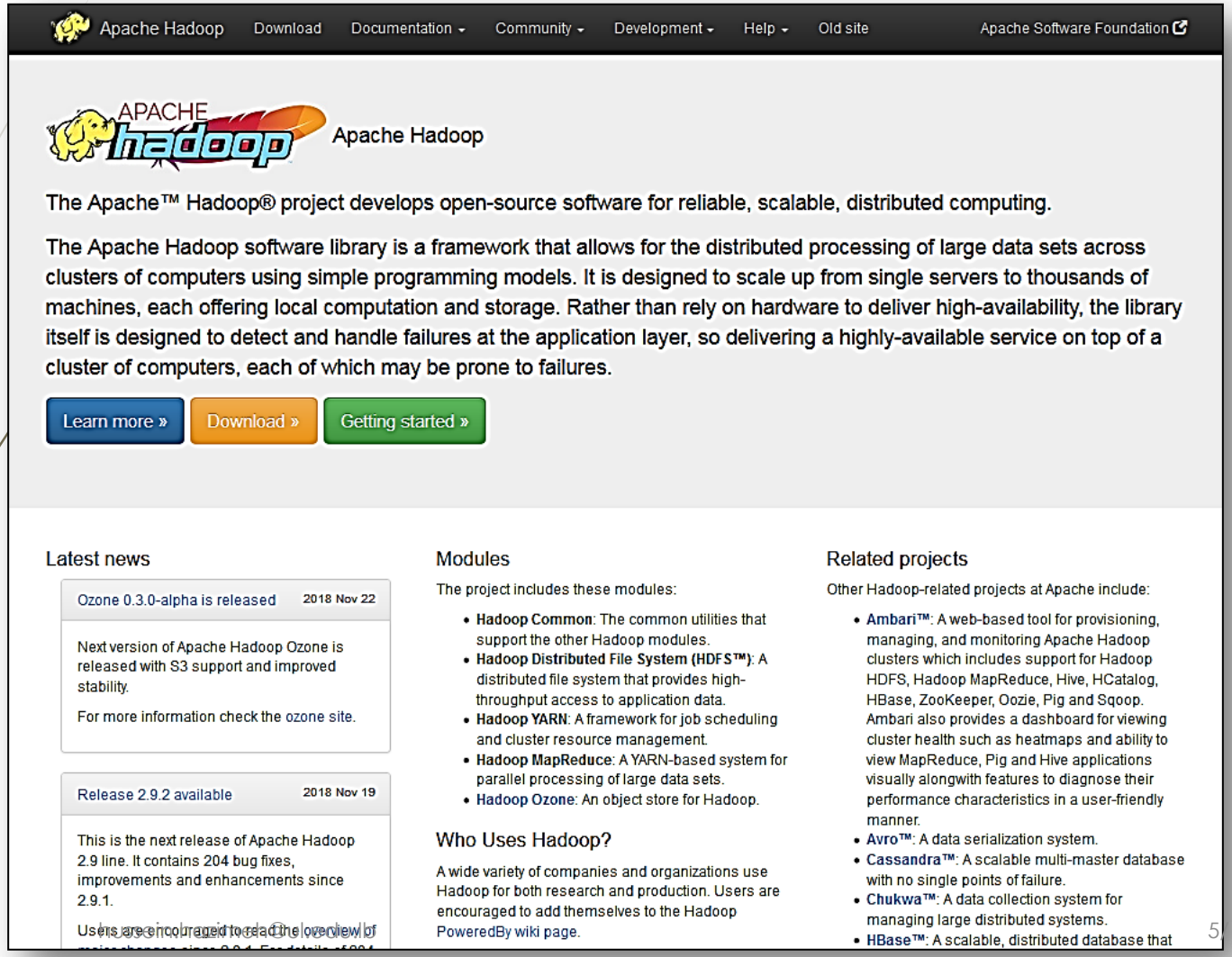
Big Data VS Data Science VS Data Analyst

- **Big Data specialist:**
  - working with unstructured data, storing, retrieving
  - Analytics, mathematics, statistical and computer skills
- **Data scientist:**
  - Searching for patterns and models in data
  - Try to predict forecast
  - Machine learning, statistical and strong mathematical skills
- **Data analyst:**
  - Try to analyze present situation
  - Business, statistical, and visualization skills

What is Hadoop?

- Is a solution for Big Data
- Is a big data analysis engine

What is Hadoop?



Apache Hadoop | Download | Documentation | Community | Development | Help | Old site | Apache Software Foundation

APACHE hadoop

Apache Hadoop

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Learn more » | Download » | Getting started »

Latest news

Ozone 0.3.0-alpha is released 2018 Nov 22

Next version of Apache Hadoop Ozone is released with S3 support and improved stability.

For more information check the ozone site.

Release 2.9.2 available 2018 Nov 19

This is the next release of Apache Hadoop 2.9 line. It contains 204 bug fixes, improvements and enhancements since 2.9.1.

Users are encouraged to read the overview of

Modules

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.
- **Hadoop Ozone**: An object store for Hadoop.

Who Uses Hadoop?

A wide variety of companies and organizations use Hadoop for both research and production. Users are encouraged to add themselves to the Hadoop PoweredBy wiki page.

Related projects

Other Hadoop-related projects at Apache include:

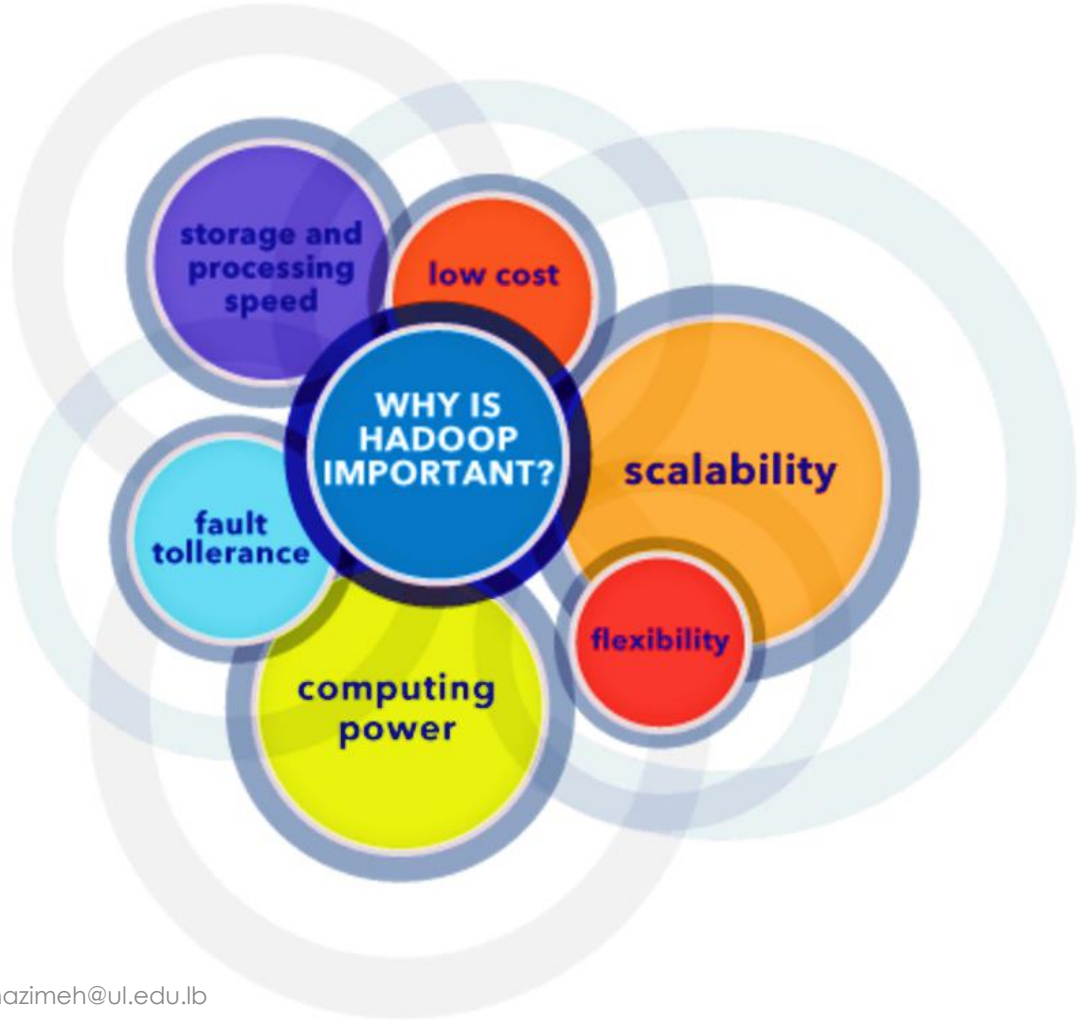
- **Ambari™**: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™**: A data serialization system.
- **Cassandra™**: A scalable multi-master database with no single points of failure.
- **Chukwa™**: A data collection system for managing large distributed systems.
- **HBase™**: A scalable, distributed database that

What is Hadoop?

- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

- Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

Why Hadoop Is Important ?



Why Hadoop Is Important ?

- **Ability to store and process huge amounts of any kind of data, quickly.**
  - **With data volumes and varieties constantly increasing, especially from social media and the Internet of Things (IoT), that's a key consideration.**

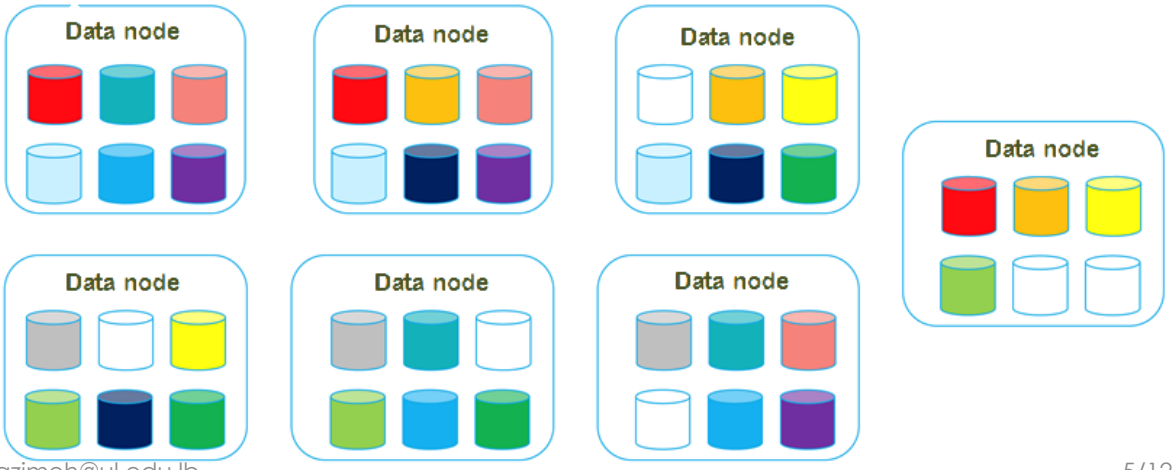
Why Hadoop Is Important ?

- Computing power
  - Hadoop's distributed computing model processes big data fast. The more computing nodes you use, the more processing power you have.



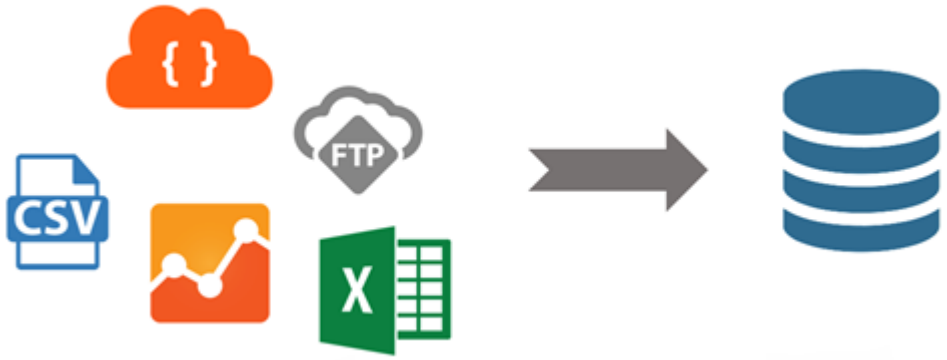
Why Hadoop Is Important ?

- **Fault tolerance**
  - Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. Multiple copies of all data are stored automatically.



Why Hadoop Is Important ?

- **Flexibility**
  - Unlike traditional relational databases, you don't have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.

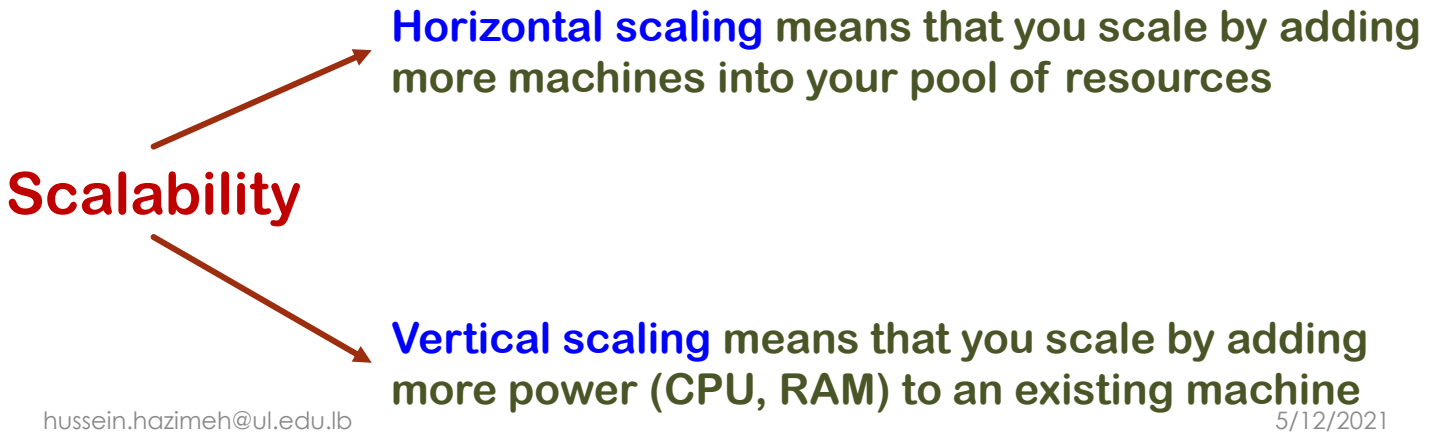


Why Hadoop Is Important ?

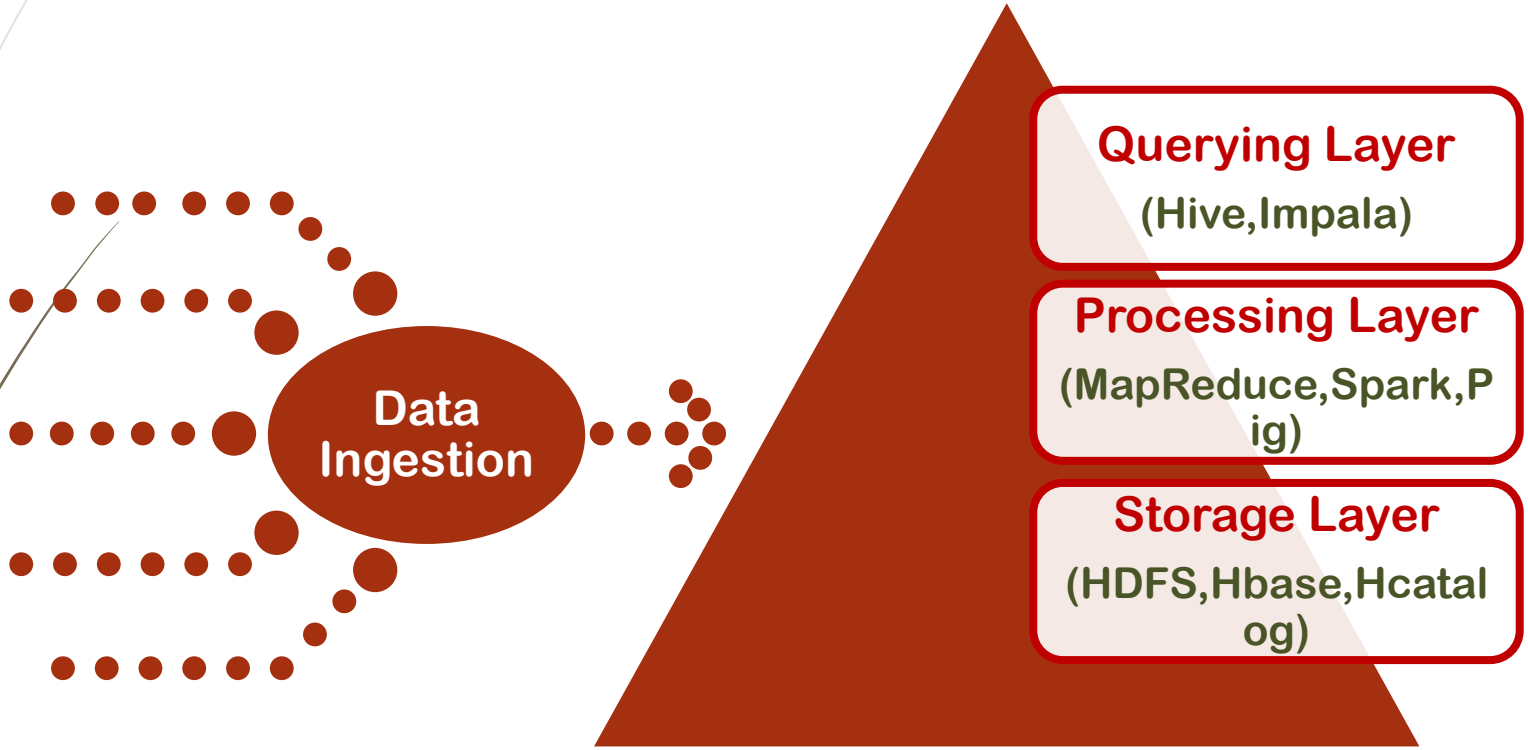
- **Low cost**
  - The open-source framework is free and uses commodity hardware to store large quantities of data.

Why Hadoop Is Important ?

- **Scalability**
  - You can easily grow your system to handle more data simply by adding nodes. Little administration is required.

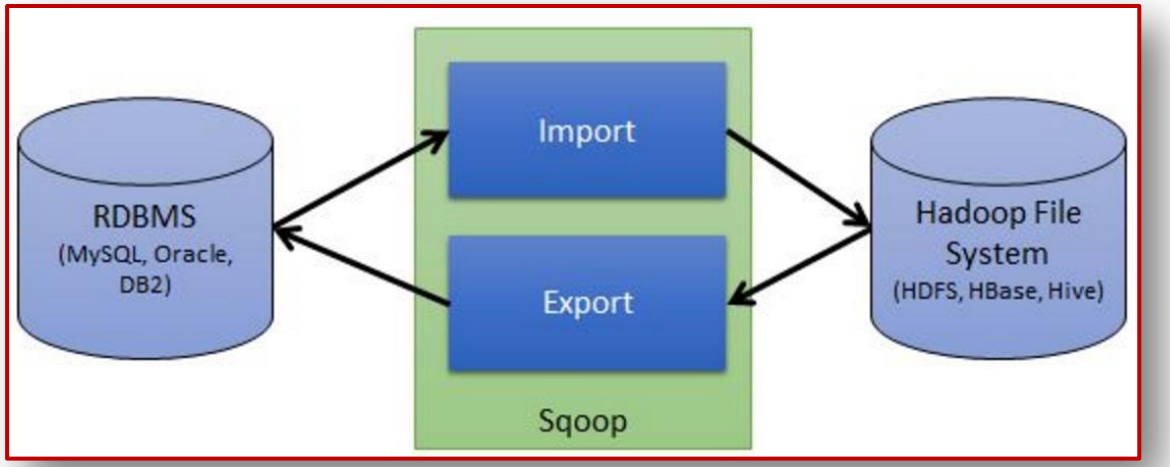


Hadoop Ecosystem



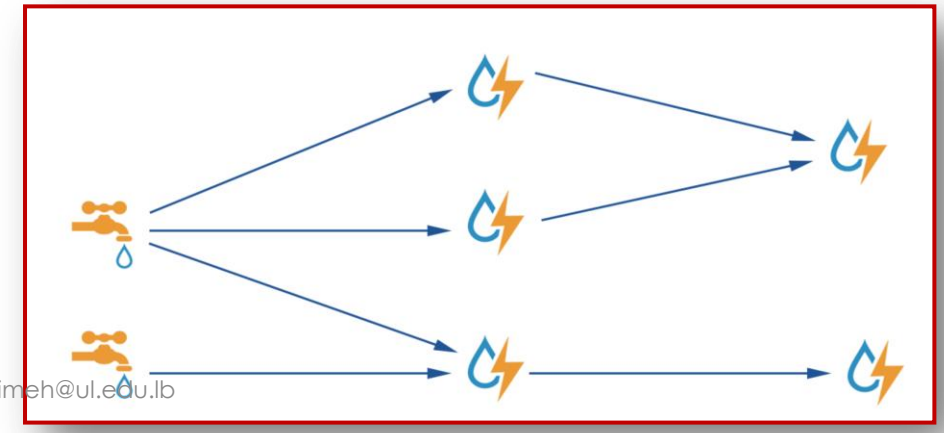
Hadoop: Data Ingestion

- **Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases.**



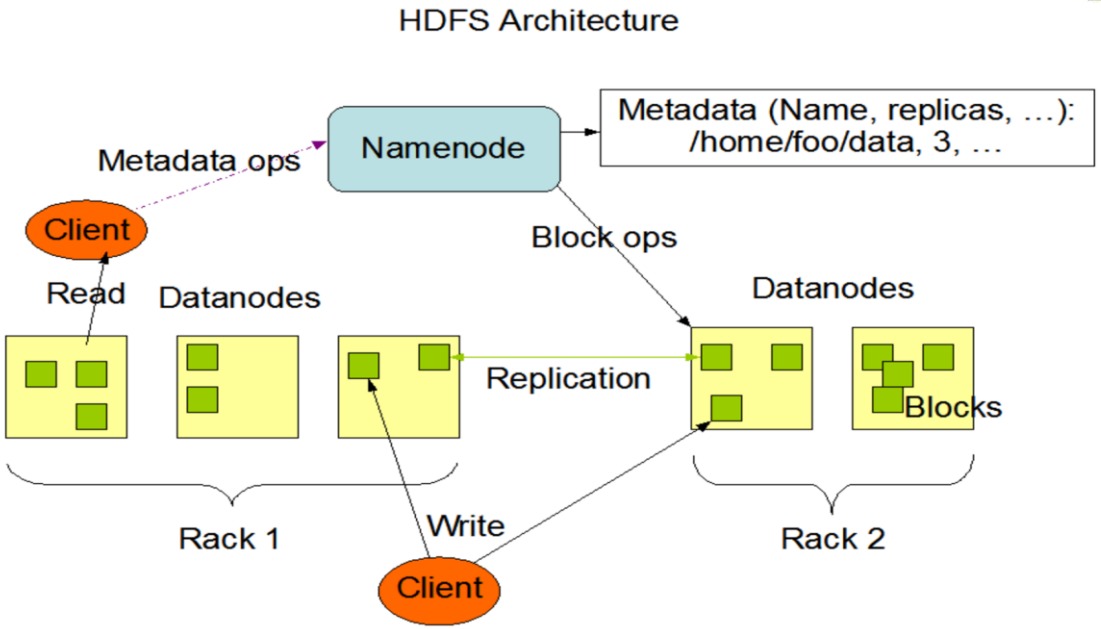
Hadoop: Data Ingestion

- **Storm** is real-time computation system. **Storm** makes it easy to reliably process unbounded streams of data, doing for real-time processing.
- A **Storm** topology consumes streams of data and processes those streams in arbitrarily complex ways, repartitioning the streams between each stage of the computation however needed.



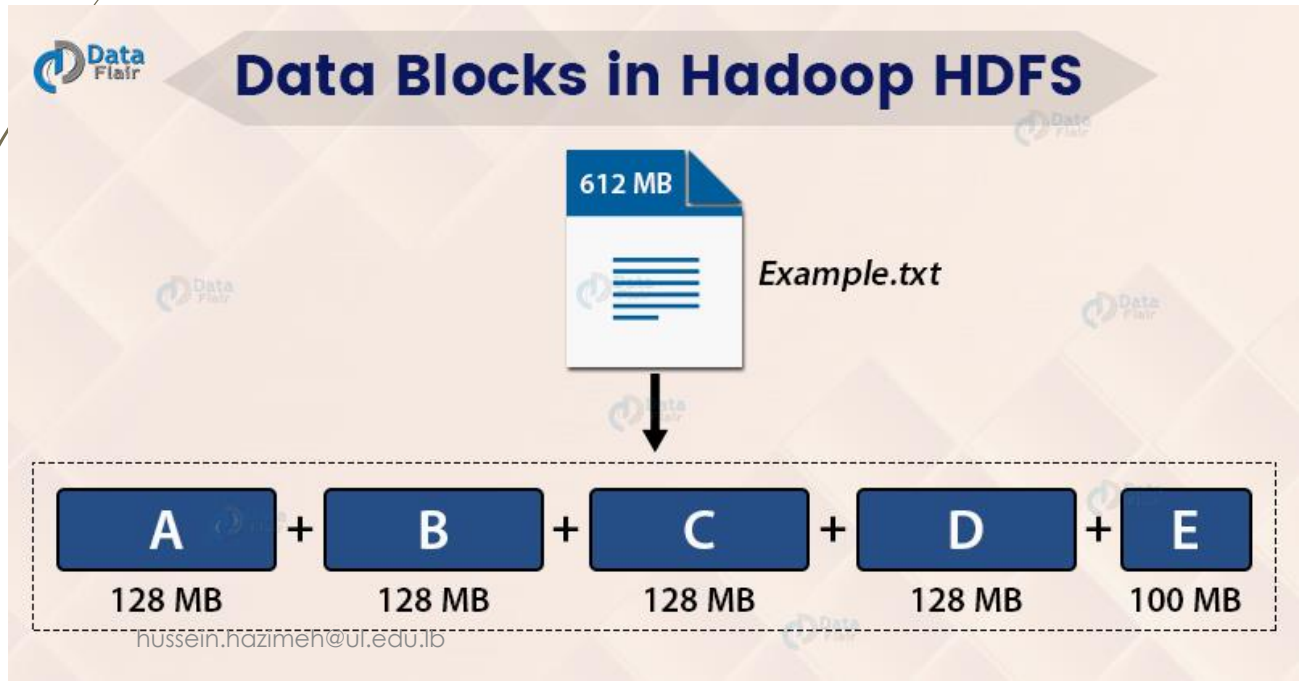
Hadoop: Data Storage Layer

- **Hadoop Distributed File System (HDFS) offers a way to store large files across multiple machines. Hadoop and HDFS was derived from Google File System (GFS) paper.**



Hadoop: Data Storage Layer

- **Hadoop Distributed File System (HDFS) offers a way to store large files across multiple machines. Hadoop and HDFS was derived from Google File System (GFS) paper.**



HDFS Architecture

- When a client wants to **write a file to HDFS**, it communicates to the NameNode for metadata. The Namenode responds with a number of blocks, their location, replicas, and other details. Based on information from NameNode, the client directly interacts with the DataNode.
- The Master node is the NameNode and DataNodes are the slave nodes.

HDFS Architecture

- What is HDFS NameNode?
- NameNode is the centerpiece of the Hadoop Distributed File System. It maintains and manages the **file system namespace** and provides the right access permission to the clients.
- The NameNode stores information about blocks locations, permissions, etc. on the local disk in the form of two files:
- **Edit log:** It contains all the recent changes performed to the file system namespace to the most recent Fsimage.

HDFS Architecture

- Functions of HDFS NameNode
  - It executes the file system namespace operations like opening, renaming, and closing files and directories.
  - NameNode manages and maintains the DataNodes.
  - It determines the mapping of blocks of a file to DataNodes.
  - NameNode records each change made to the file system namespace.
  - It keeps the locations of each block of a file.
  - NameNode takes care of the replication factor of all the blocks.
  - NameNode receives heartbeat and block reports from all DataNodes that ensure DataNode is alive.
  - If the DataNode fails, the NameNode chooses new DataNodes for new replicas.

HDFS Architecture

- What is HDFS DataNode?
  - DataNodes are the slave nodes in Hadoop HDFS. DataNodes are **inexpensive commodity hardware**. They store blocks of a file.
- Functions of DataNode
  - DataNode is responsible for serving the client read/write requests.
  - Based on the instruction from the NameNode, DataNodes performs block creation, replication, and deletion.
  - DataNodes send a heartbeat to NameNode to report the health of HDFS.
  - DataNodes also sends block reports to NameNode to report the list of blocks it contains.

HDFS Architecture

- What is Backup Node?
  - In Hadoop, Backup node keeps an **in-memory, up-to-date copy** of the file system namespace. It is always synchronized with the active NameNode state.

HDFS Architecture

- What are Blocks in HDFS Architecture?
  - Internally, HDFS split the file into block-sized chunks called a block. The size of the block is **128 Mb** by default. One can configure the block size as per the requirement.
  - For example, if there is a file of size 612 Mb, then HDFS will create four blocks of size 128 Mb and one block of size 100 Mb.
  - The file of a smaller size does not occupy the full block size space in the disk.
  - For example, the file of size 2 Mb will occupy only 2 Mb space in the disk.
  - The user doesn't have any control over the location of the blocks.
  - HDFS is highly fault-tolerant. **Now, look at what makes HDFS fault-tolerant.**

HDFS Architecture

- What is Replication Management?
  - For a distributed system, the data must be redundant to multiple places so that if one machine fails, the data is accessible from other machines.
  - In Hadoop, HDFS stores replicas of a block on multiple DataNodes based on the replication factor.
  - The replication factor is the number of copies to be created for blocks of a file in HDFS architecture.
  - If the replication factor is 3, then three copies of a block get stored on different DataNodes. So if one DataNode containing the data block fails, then the block is accessible from the other DataNode containing a replica of the block.
- If we are storing a file of 128 Mb and the replication factor is 3, then (3*128=384) 384 Mb of disk space is occupied for a file as three copies of a block get stored.

HDFS Architecture

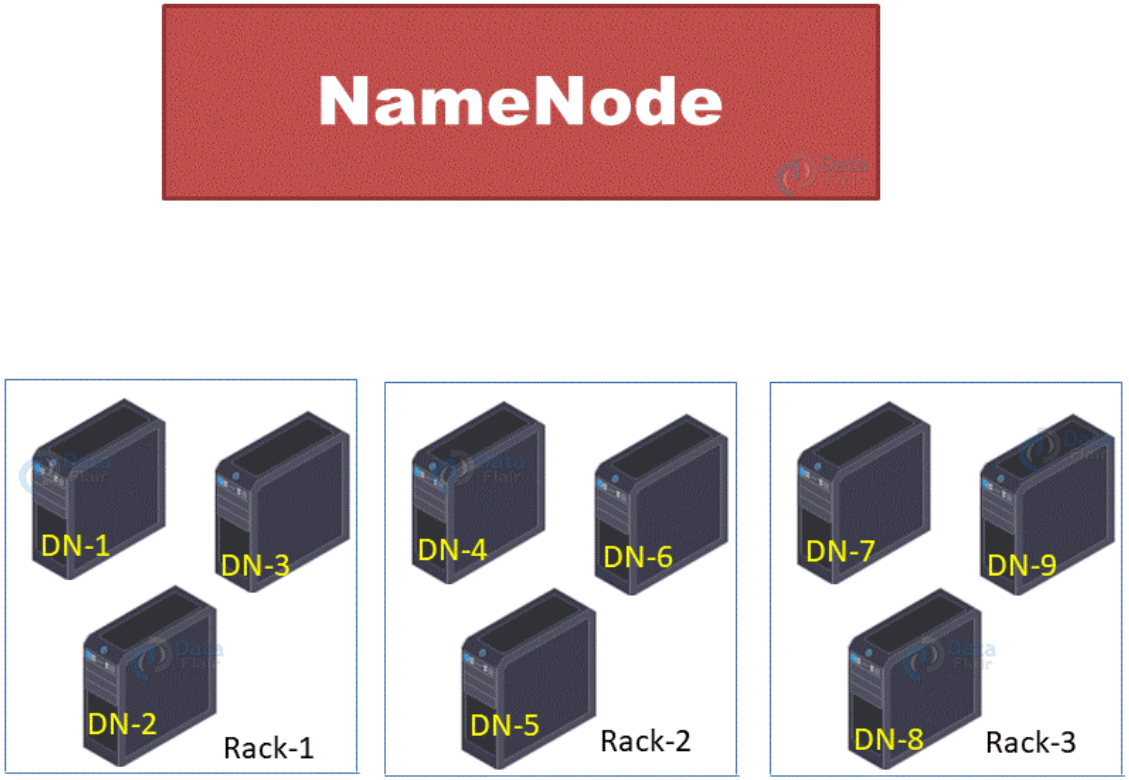
- What is Rack in HDFS Architecture?
- Let us now talk about how HDFS store replicas on the DataNodes? What is a rack?

- **Rack** is the collection of around 40-50 machines (DataNodes) connected using the same network switch. If the network goes down, the whole rack will be unavailable.

HDFS Architecture

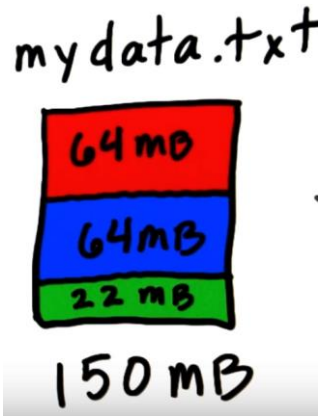
- **Rack Awareness** is the concept of choosing the closest node based on the rack information.
- To ensure that all the replicas of a block are not stored on the same rack or a single rack, NameNode follows a rack awareness algorithm to store replicas and provide latency and fault tolerance.
- Suppose if the replication factor is 3, then according to the rack awareness algorithm:
  - The first replica will get stored on the local rack.
  - The second replica will get stored on the other DataNode in the same rack.
  - The third replica will get stored on a different rack.

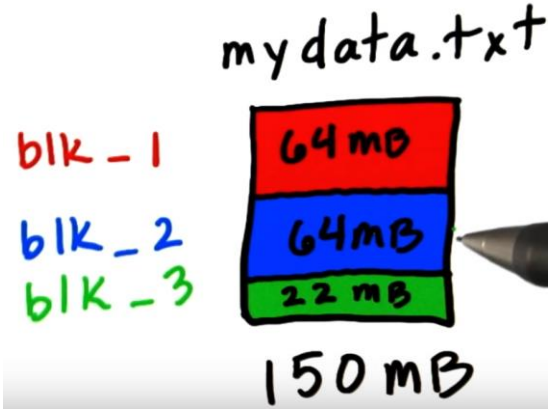
HDFS Architecture



Hadoop HDFS Architecture

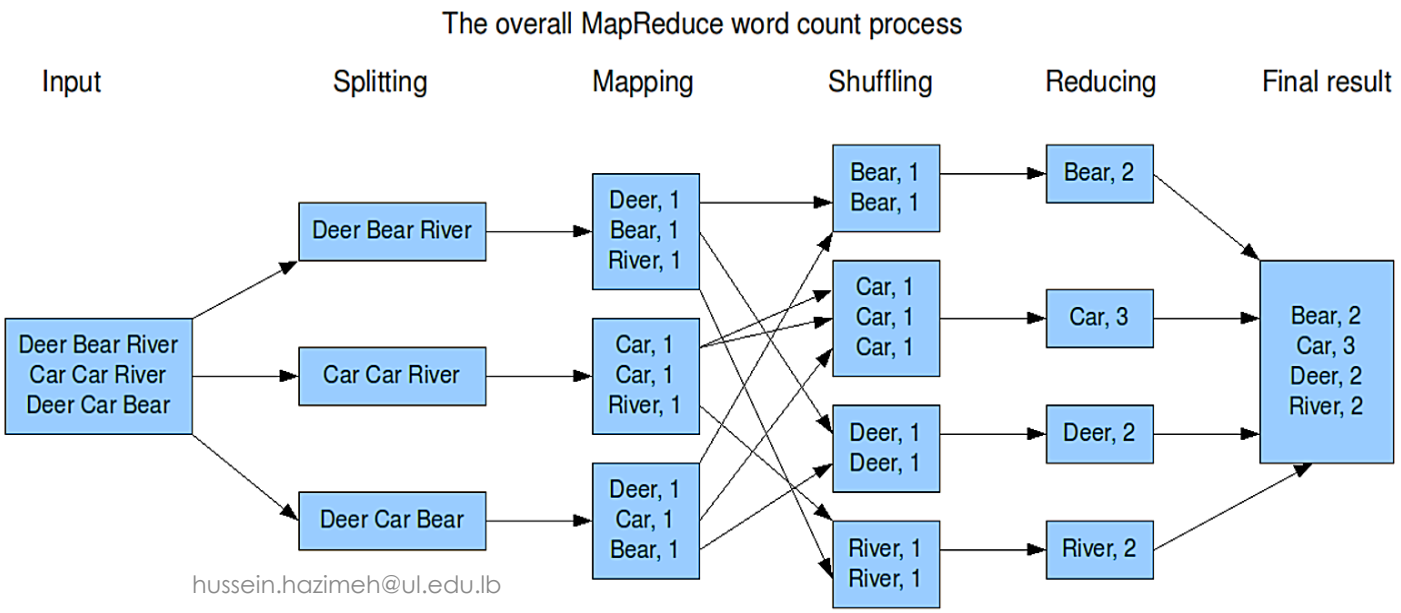
Hadoop: Data Storage Layer



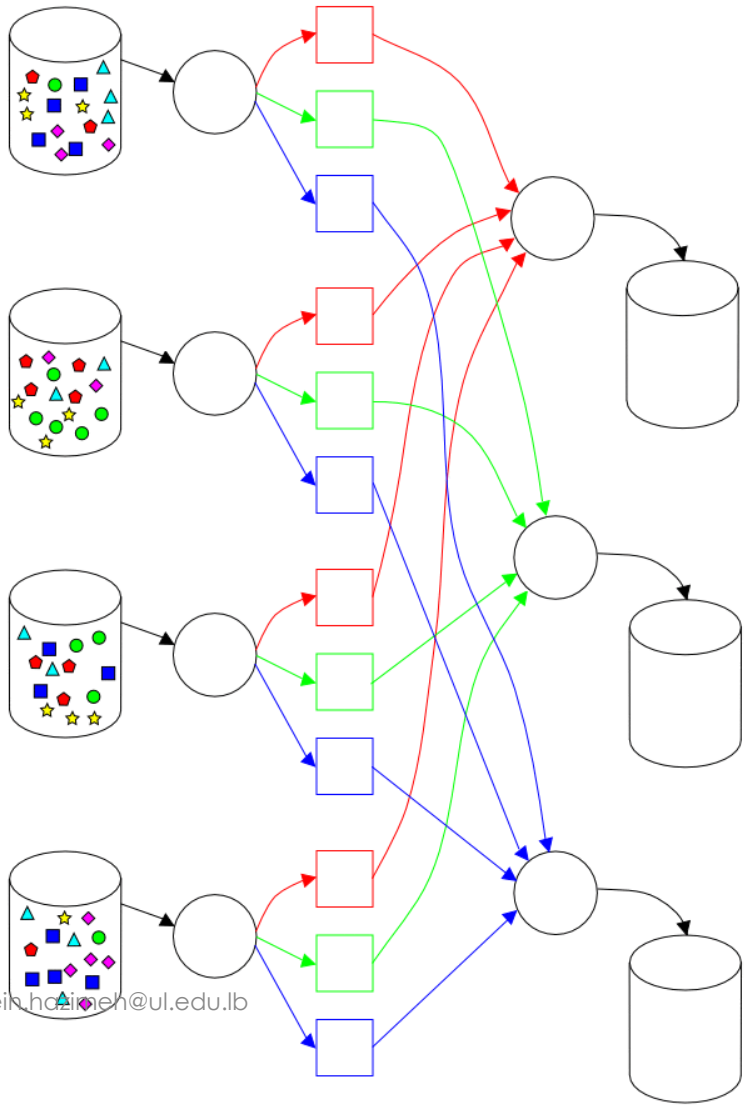


Hadoop: Data Processing Layer

- **MapReduce** is the heart of Hadoop. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster with a parallel, distributed algorithm.

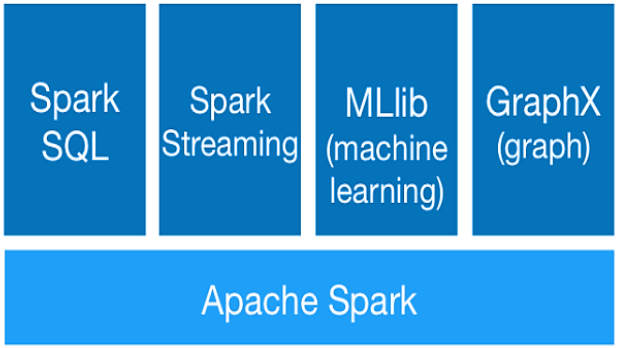


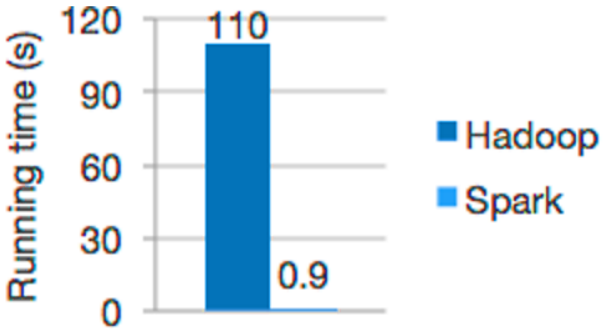
Hadoop: Data Processing Layer



Hadoop: Data Processing Layer

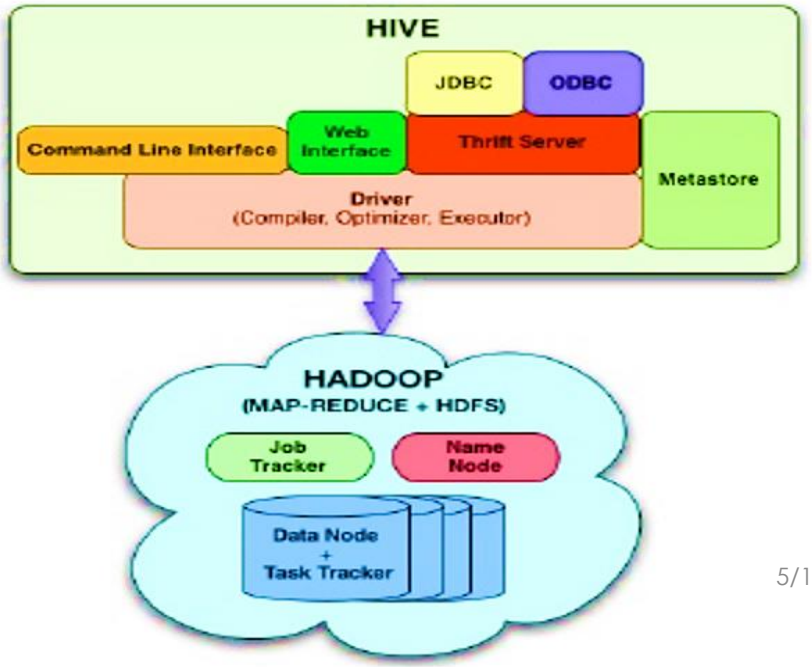
- **Spark:** In memory data analytics cluster computing framework originally developed in the AMPLab at UC Berkeley.
- Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.





Hadoop: Data Querying Layer

- **Hive:** A distributed data warehouse built on top of HDFS to manage and organize large amounts of data.
- **Hive** provides a query language based on SQL semantics (HiveQL) which is translated by the runtime engine to MapReduce jobs for querying the data.



Hadoop: Management Layer

- **Apachi Ambari:** intuitive, easy-to-use Hadoop management web UI.
- **Apache Ambari** was donated by Hortonworks team. It's a powerful and nice interface for Hadoop and other typical applications from the Hadoop ecosystem.