

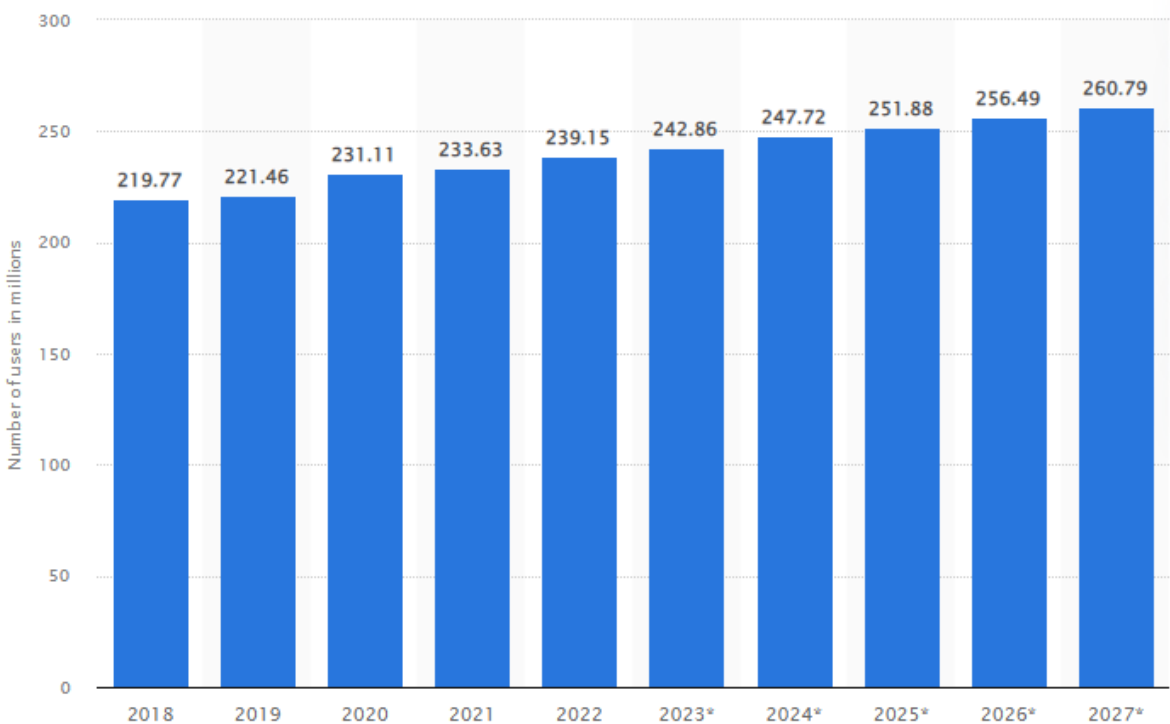
# Introduction

**"Data is the new currency, and I don't believe we are truly embracing it and all its possibilities." - ADEC Innovations CEO at UNEA-2.**

- Overview .....2
- Main Big Data Components .....3
- Types of Data.....4
- Why Big Data?.....7
- Examples Of Big Data .....8
- Characteristics Of Big Data .....9
- Big Data Processing .....12
- Big Data Analytics .....14
- Big Data Technologies: Big Data is not just Hadoop! .....15
- Big Data Programming Languages.....20
- Applications of Big Data .....20
- The Benefits of Big Data Analytics.....21
- Big Data Challenges .....21
- Big Data Project Sample.....21

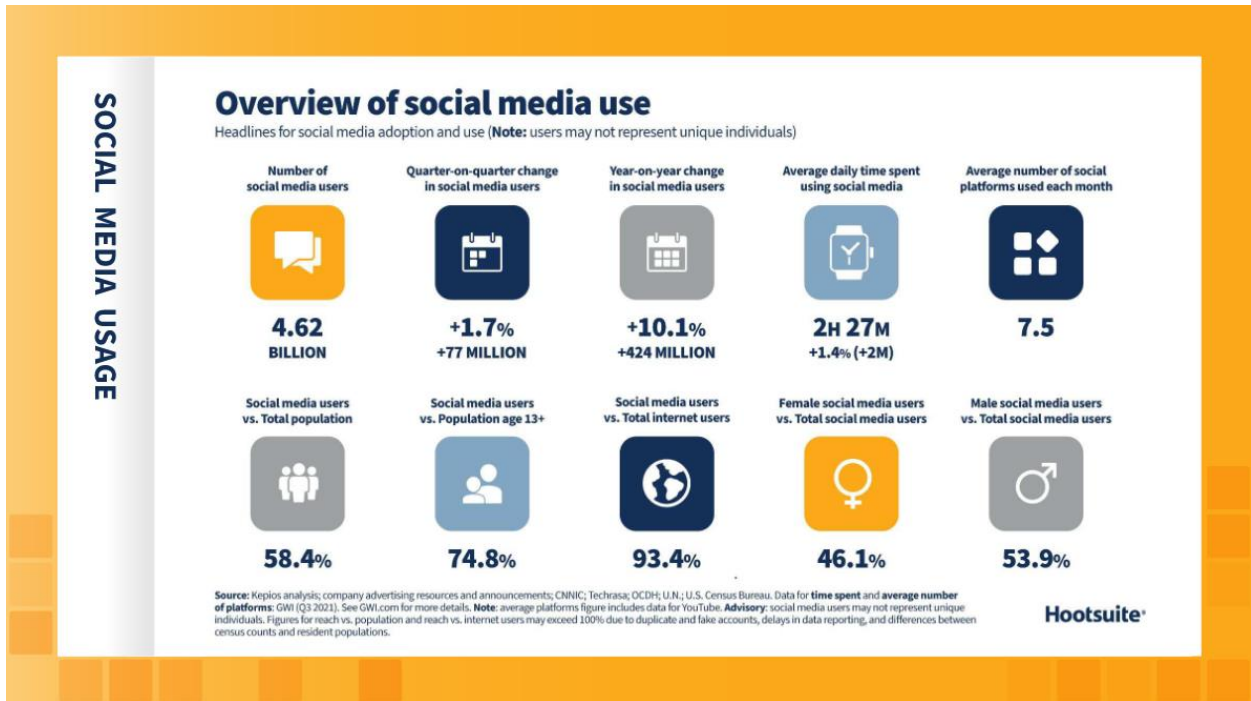
- **Overview**

- Nowadays, data is growing exponentially with time
- Traditional technologies cannot store it or process big data efficiently.
- Imagine the number of browsing, transactions on Amazon
- Data managements was limited to text (document management systems, i.e., search engine) and database management systems.
- Today, we web logs, web customer behaviors, etc.
- Number of FB users:



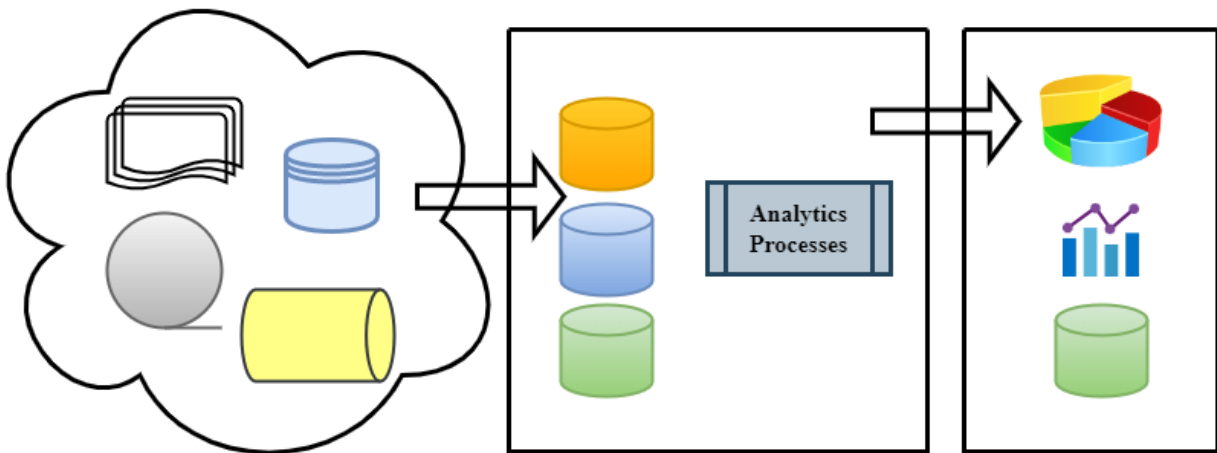
© Statista 2023

- Overview of social media use (hootsuite.widen.net)



- **Main Big Data Components**

- Data Sources
- Data Storage
- Data Processing
- Data Analytics



- **Types of Data**

- **Structured Data:**

- The data format is predetermined Any data stored in a database table (Exclude BLOBs):
    - Example: Employee table in a database.

Emp ID	FN	LN	Dept
1001	Mary	Paul	Marketing
1002	James	Bond	IT
1003	Michelle	Li	Support

- **Unstructured Data:**

- Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model
    - It is estimated that 80-90% of a company data is unstructured, and it is growing at high rate every year.
    - Example: Pictures, videos, sound files, and PDF documents
    - Storage solutions:

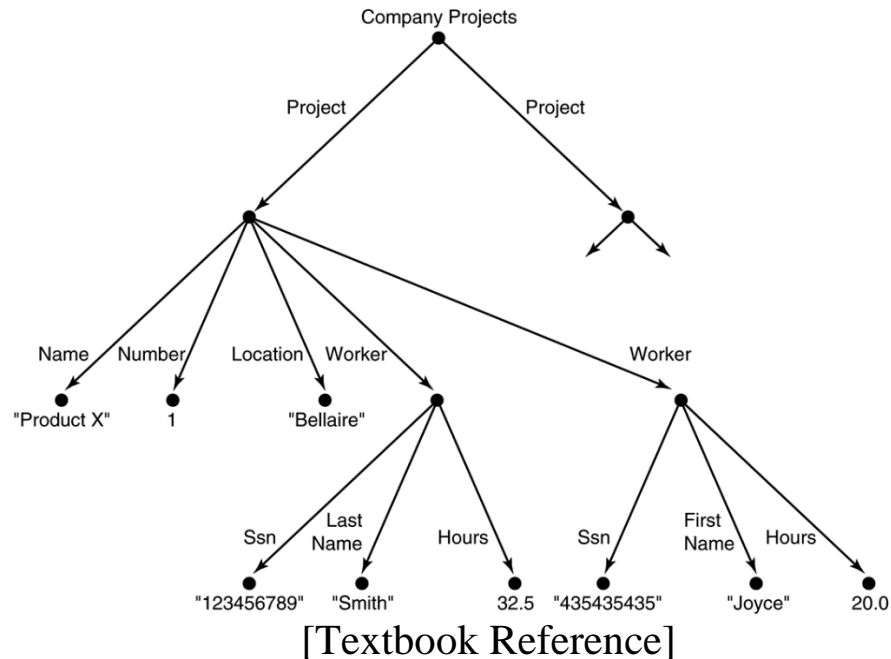
- **MongoDB:**

- An open source, JSON-like document-oriented database.
        - It is used to store both unstructured and semi-structured data
        - Example:

```
{  
  "_id" : EmpId("A1101"),  
  "FirstName" : "James",  
  "LastName" : "Berry",  
  "Age" : 51,  
  "Interests" : [ "Spinning", "Soccer" ]  
  "Address": {  
    "Street": "000 Leesburg Pike",
```

```
"City": "Falls Church",  
"State": "VA "  
"Zip": "22041",  
}  
}
```

- Big data graph database:
  - Apache Giraph
  - Amazon Neptune
  - Neo4j
  - ArangoGraph
- **Semi-structured Data:**
  - Semi-structured data is information that does not reside in a relational database but organized using meta tags that make it easier to analyze.
  - Easier to analyze than unstructured data
  - Semi-structured data may be displayed as a directed graph:
    - The labels or tags on the directed edges represent the schema names—the names of attributes, object types (or entity types or classes), and relationships.
    - The internal nodes represent individual objects or composite attributes.
    - The leaf nodes represent actual data values of simple (atomic) attributes.
    - Example:



- Example: Tweets organized by hashtags, Emails organized by Inbox, Draft, and Json or XML files.

○ Batch vs Real-Time/Stream Data

- It is imperative that you know the type of big data you are working with → TO choose the right data analytics algorithm for your problem.
- Batch Data:
  - Data is collected over a period of time such as a year and then it is fed into an analytics system
  - Use cases:
    - Applications where you don't need real-time analytics results
    - Examples: billing, payroll, overall performance of all locations of a company, etc.
- Stream Data:
  - Data comes at a high velocity, and you need to make critical decision right away: Intrusion Detection System

- You may not be able to collect all the data.
- You may not be able to get an exact solution to your problem → approximate solution would be enough
- Process Unbounded and Bounded Data
- Any kind of data is produced as a stream of events. Credit card transactions, sensor measurements, machine logs, or user interactions on a website or mobile application, all of these data are generated as a stream.
  
- Use Cases:
  - Applications that require analytics results in real time.
  - Examples: Log monitoring, fraud detection, analyze customer behavior, etc.
  
- **Why Big Data?**
  - “Big data” is **high-volume, velocity, and variety** information assets that demand **cost-effective**, innovative forms of **information processing** for enhanced insight and **decision making.**” - Gartner
  - Huge amount of data that cannot be processed using traditional methods:
    - How to store the data?
    - How to process the data?
  - Aren't we happy with RDBMS?
    - Yes, but only for transaction data management.
  - What if data is large?
    - Can we just distribute our RDBMS tables over multiple servers? → very expensive join operations.
  - RDBMS Limitations:
    - Oracle has hard limit of 1000 columns per table. If your split a table, a join will cost more.
    - MySQL has hard limit of 4096 columns per table

- Very expensive join operation for large tables
    - Cluster the table → still expensive joins
- Handling unstructured data?
  - RDBMS → BLOB
  - From 80% to 90% of data generated and collected by organizations is unstructured, and its volumes are growing rapidly — many times faster than the rate of growth for structured databases. (mongodb.com)

- **Examples Of Big Data**

- Social Media
- A single Jet engine can generate 10+terabytes of data in 30 minutes of flight time → Petabytes of data per day
- Financial Services
  - Fraud detection
    - Detect Banks monitor credit cardholders' purchasing patterns checking for fraudulent transactions.
  - Money laundering
- Healthcare:
  - To reduce the overall cost of healthcare
    - Prediction of epidemic outbreaks
    - Early symptom detection to avoid preventable diseases
    - Electronic health records management system
    - Prediction and prevention of serious medical conditions
- Digital Marketing
  - Bring product and services to customers.
- E-commerce:
  - Example of Amazon and alike.



- **Characteristics Of Big Data**

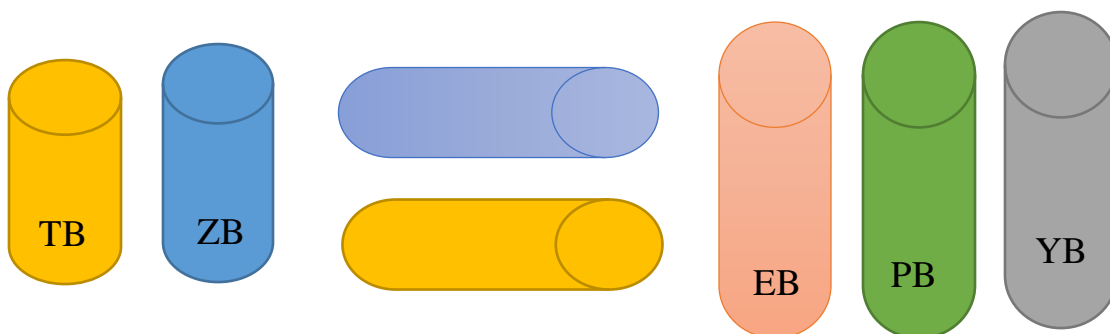
- The following are known as “Big Data Characteristics”:

- 3Vs, 4Vs, More Vs
- Let us look at the most 5 common Vs
  - **Volume:** Very Large Amount of Data
  - **Velocity:** Produce Data at Vary Fast Rate
  - **Variety:** Produce Data in Different Formats
  - **Veracity:** The correctness of Data
  - **Value:** Business value of the big data

- **Volume:**

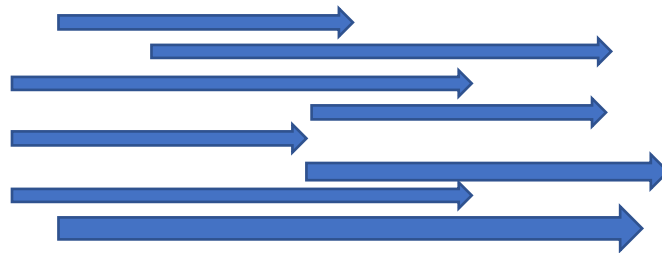
- How much Data is generated?
- Very Large Amount of Data: Terabytes (T) to Exabytes (E) of data to process

Name	Equal To	Size (In Bytes)
Bit	1 bit	1/8
Nibble	4 bits	1/2 (rare)
Byte	8 bits	1
Kilobyte (KB)	1024 bytes	1024
Megabyte	1, 024kilobytes	1, 048, 576
Gigabyte (GB)	1, 024 megabytes	1, 073, 741, 824
Terabyte (TB)	1, 024 gigabytes	1, 099, 511, 627, 776
Petabyte (PB)	1, 024 terabytes	1, 125, 899, 906, 842, 624
Exabyte (EB)	1, 024 petabytes	1, 152, 921, 504, 606, 846, 976
Zettabyte (ZB)	1, 024 exabytes	1, 180, 591, 620, 717, 411, 303, 424
Yottabyte (YB)	1, 024 zettabytes	1, 208, 925, 819, 614, 629, 174, 706, 176



○ **Velocity:**

- Produce Data at Vary Fast Rate
- Examples of applications:
  - Radio-frequency identification (RFID),
  - Global positioning system (GPS),
  - Number of tweets per second
  - Number of FB posts per second
  - Etc.
- Streaming Data: milliseconds to seconds to query



- Facebook users upload more than 900 million photos a day (<https://www.datacenterfrontier.com/>) →

**≈ 10,416,667 photo per second.**

- New **Tweets per second** (TPS) record: 143,199 TPS.  
Typical day: more than 500 million Tweets sent;  
average **5,700 TPS**. (blog.twitter.com)

○ **Variety:**

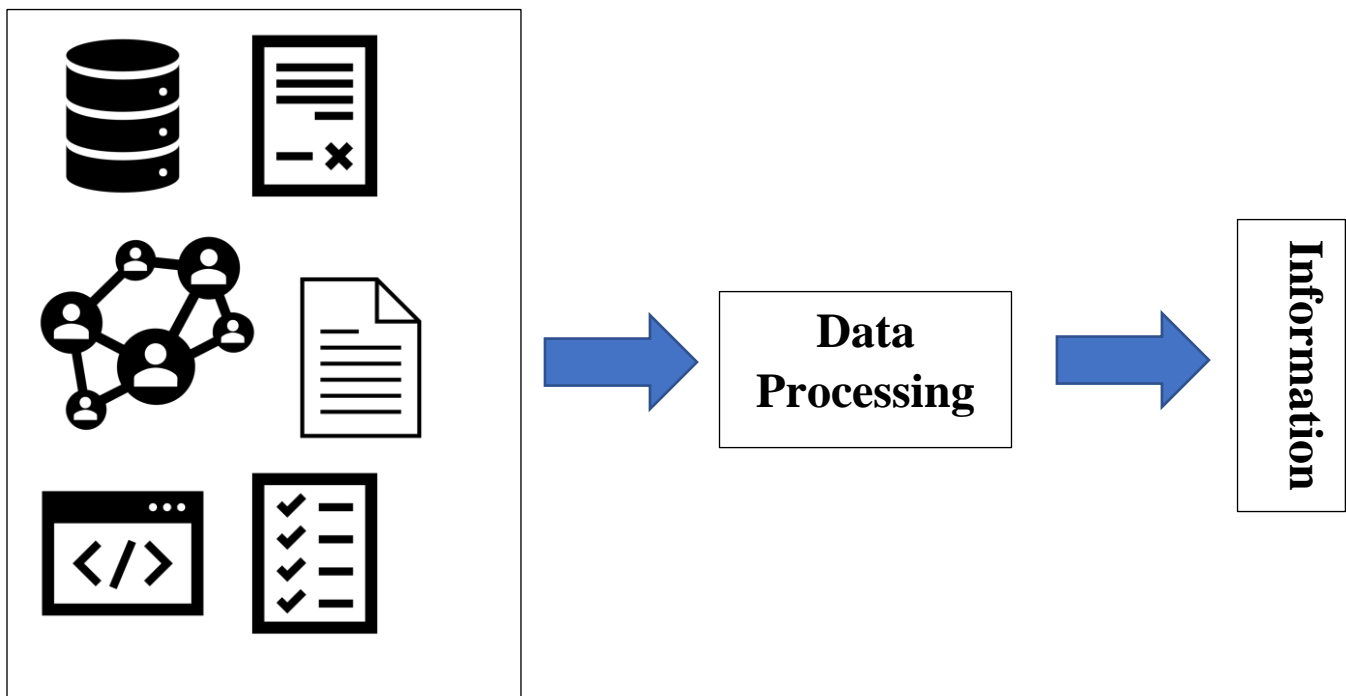
- Different forms of Data
- Produce Data in Different Formats:
  - Structured
  - Unstructured
  - semi-structured

○ **Veracity:**

- Quality and accuracy of data 1
- According to [www.merriam-webster.com](http://www.merriam-webster.com) :

- 1. conformity with truth or fact: ACCURACY
  - 2: devotion to the truth: TRUTHFULNESS
  - 3: power of conveying or perceiving truth
  - 4: something true
- Source of Veracity:
  - Bugs in Application
  - Noise: missing or incomplete data
  - Outliers or Anomaly
  - Duplication of Data
  - Users under different login information
- **Value:**
  - Extract the value of the Big Data using proper analytics to make the right business decision
  - Business growth
  - Discover new business opportunities
- **Other V's:**
  - **Viscosity:**
    - Complexity or degree of correlation
    - It describes the flow of the data: how quickly data is generated and how quickly that data moves.
  - **Variability:**
    - Inconsistency in data flow
    - Does the meaning of the data change over time?
  - **Volatility:**
    - How long does data need to be kept for?
    - Durability or how long-time data is valid and how long it should be stored
  - **Viability:**
    - It describes data activeness.
    - Capability to be live and active

- **Validity:**
  - Valid data is critical in making the right decisions
- **Visualization:**
  - Ways to convey results of your big data project
- **Big Data Processing**
  - Main objective but for large datasets:



- Cluster Computing:
  - NIH Biowulf Cluster started in 1999 with 40 "boxes on shelves".
- NoSQL Database:
  - Non-relational databases are in the form of four different types of stores:
    - key-value,

- column,
  - graph or document pairs.
- Columnar database:
  - Data is stored in columns rather than rows.
  - This implementation reduces the number of read data operations during query processing and provides high performance for large concurrent queries.
- Data Pipelines:
  - The main objective is to process raw data (i.e. structured, semi-structured, and unstructured data) and save it in data store, like a data lake or data warehouse, for analysis.
  - The process consists of converting, organized, and cleaning raw before storing it.
  - Three core steps make up the architecture of a data pipeline.
  - Pipeline Main Components:
    - Data ingestion:
    - Data Transformation:
      - It consists of a set of processes that converts the raw data into a format suitable for data repository.
      - For example, the processes transform nested fields in a JSON format to extract the key fields for analysis.
    - Data Storage:
      - The transformed data is stored to be used by consumers and subscribers.

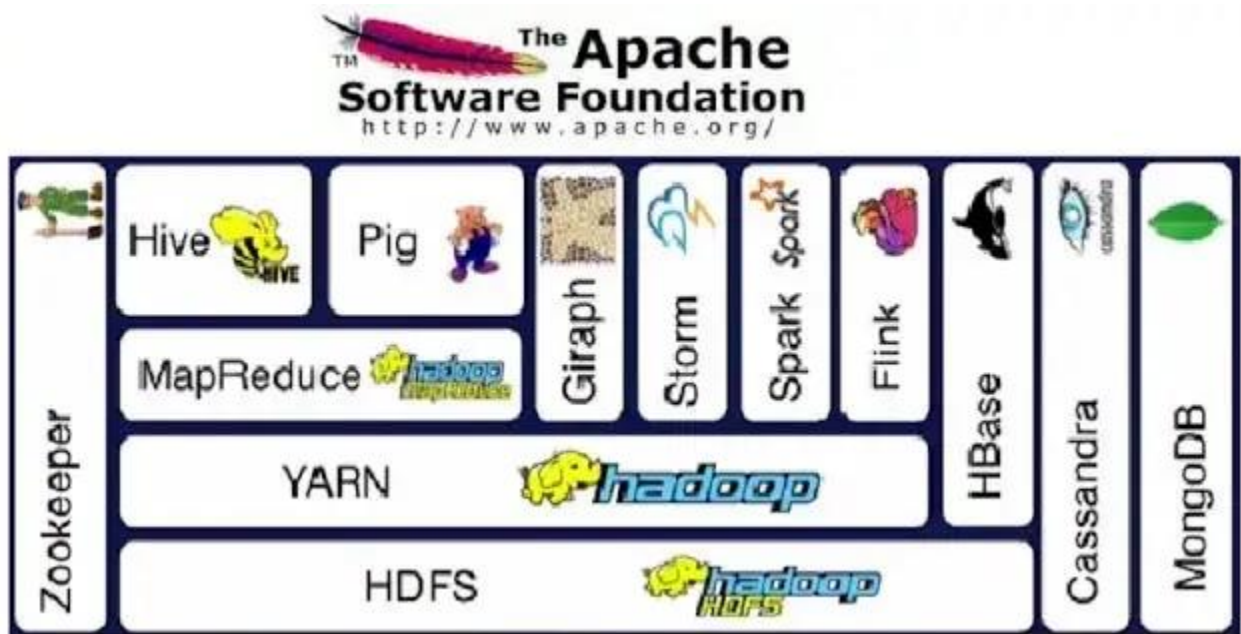


- **Big Data Analytics**

- Big Data analytics is a combination of several techniques and processing methods that collect, organize, and analyze large sets of data to discover patterns and other useful information.
- There are four types of Big Data Analytics:
  - **Predictive Analytics:**
    - It also known as “**What might happen in the future?**”
    - It uses prediction models
    - It analyses past data sets to predict future outcomes.
  - **Prescriptive Analytics:**
    - It is also known as “**What should we do next?**”
    - It works on a data set and determines what actions needs to be taken and manage and optimize the overall business performance- Email automation.
  - **Descriptive Analytics:**
    - Also known as “**What happened?**”
    - It analyzes the past data and provide a graphical representation to determine what happens and why.
  - **Diagnostic Analytics:**
    - It is also known as “**Why did this happen?**”
    - It usually comes after Descriptive Analysis.
    - It uses different techniques such as data mining, to provide insight into the causes of specific events.
    - For example, the analysis shows a decrease in revenue in a specific store, the diagnostic analysis reveals it is because of poor customer service.

-

- **Big Data Technologies: Big Data is not just Hadoop!**
  - Need specialized software tools and ways to store and analyze big data
  - How big data is managed?
    - Where is data stored?
      - Distributed, cloud based?
      - NoSQL databases
    - How is data processed?
      - Distributed processing?
    - Types of analytics
  - Hadoop Ecosystem:



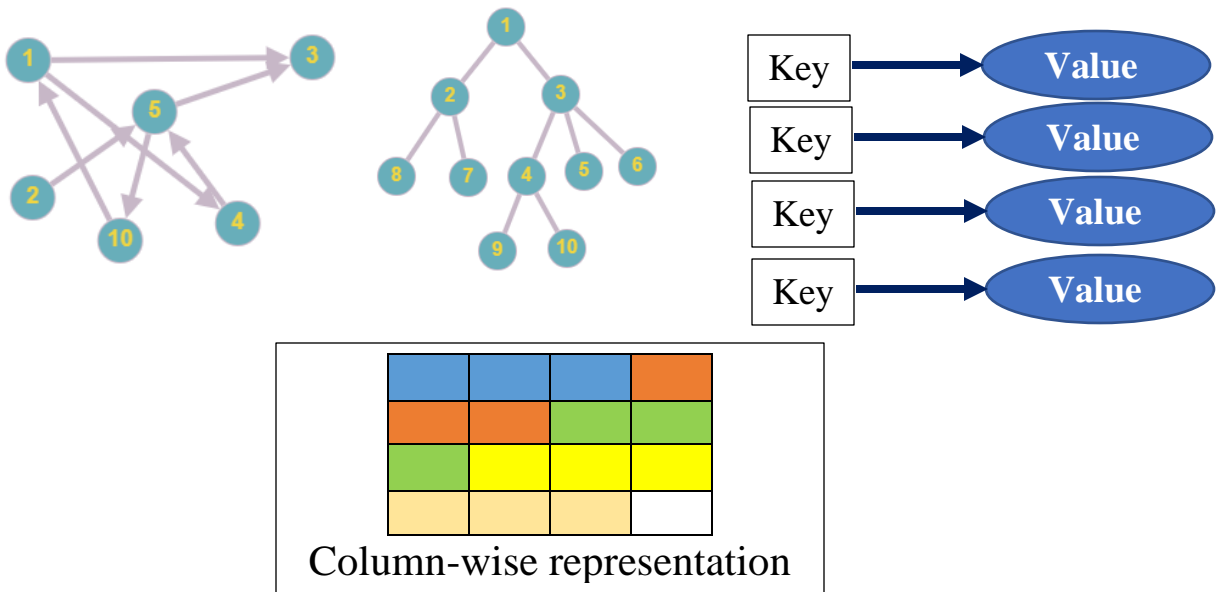
**Hadoop Ecosystem (credits Apache Foundation)**

- Hadoop main components:
  - HDFS, YARN, and MapReduce
    - HDFS (Hadoop Distributed File System):
      - A file management system

- YARN:
  - Resource Manager and Service Scheduler
- MapReduce:
  - Parallel Programming Model
- HIVE:
  - Database/Datawarehouse system using SQL-like query language called HiveQL
- PIG:
  - It uses a high-level scripting language for data analysis called: Pig Latin script language
- Graph-base Processing:
  - GIRAPH:
    - It is an iterative graph processing system
    - It is an open-source implementation of the Google proprietary Pregel.
  - GraphX:
    - It is a component in Spark for graphs and graph-parallel computation.
- STORM:
  - It is a task parallel, distributed data streaming technology.
- SPARK:
  - Spark is a data-parallel processing framework.
  - Spark workflows are designed in Hadoop MapReduce but runs 100 times faster than Hadoop in certain applications.
- FLINK:



- It a framework and distributed processing engine for both batch and stream data
  - It executes arbitrary dataflow programs in a data-parallel and pipelined manner
- NoSQL Database:
  - HBase, CASSANDRA, MongoDB, Etc.
- ZOOKEEPER:
  - Responsible for service management and computation state:
    - partition/worker mapping
    - global state: #superstep
- **Other Big Data Technologies:**
  - Python
  - R system
  - NoSQL Databases:
    - Also known as “not only SQL” or “non-SQL”
    - It is a non-relational DBMS, that does not require a fixed schema
    - Different types of NoSQL databases are:
      - document databases
      - key-value databases
      - wide-column stores
      - graph databases.



- Examples of column-based NoSQL databases include Cassandra, HBase, MongoDB, and Hypertable
- Apache Kafka:
  - It is a distributed streaming platform
  - It is a publish-subscriber based fault-tolerant messaging system capable of handling large volumes of data.
- Data Visualization:
  - **Tableau** is the most popular and efficient tool used in the business intelligence domain.
  - **Plotly** mainly used for making Graphs and associated components.
- Additional technologies:

Data Serving



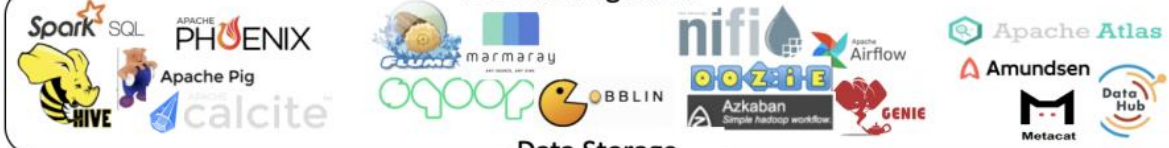
ML



Data Processing



Data Management







Data Storage



Foundational



- **Big Data Programming Languages**

			
Java	R	Python	Scala

- **Scala**

- Big Data in Healthcare
- It was created in 2003 to address issues with Java.
- It is a multi-paradigm programming language; Object oriented and functional Statically Typed Run on JVM
- It also supports concurrent and synchronized processing.
- Scalable and efficient in handling big data analytics

- **Other Programming Languages:**

- Julia, Go, Etc.

- **Applications of Big Data**

- Big Data in Healthcare
- Big Data in Education
- Big Data in E-commerce
- Big Data in Media and Entertainment
- Big Data in Finance
- Big Data in Travel Industry
- Big Data in Telecom
- Big Data in Automobile

- **The Benefits of Big Data Analytics**

- Customer Acquisition and Retention
- Focused and Targeted Promotions
- Potential Risks Identification
- Cost optimization
- Improve Efficiency

- **Big Data Challenges**

- Keeping the big data system cost in control: Data storage can become very expensive as data grow and processing time may become slow.
- Data Integration: Imagine you acquire a new company that has a completely different IT system.
- Expensive maintenance
- Which technology works best for your problem?

- **Big Data Project Sample**

