Data Handling& Analytics

Mrs.D.K.Magdum

Big Data

- **DATA**
- The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.
- **BIG DATA**
- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time.
- It is data with so large size and complexity that none of the traditional data management tools can store it or process it efficiently. Big data is also data but with huge size.

Examples of big data

The **New York Stock Exchange** generates about ***one terabyte*** of new trade data per day.

500+terabytes of new data get ingested into the databases of social media site **Facebook**, every day

Types Of Big Data

Following are the types of Big Data:

1. Structured
2. Unstructured
3. Semi-structured

- **Structured**
- Any data that can be stored, accessed and processed in the form of fixed format is termed as a ‘structured’ data.
- It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms.
- Example: The employee table in a company database will be structured as the employee details, their job positions, their salaries, etc., will be present in an organized manner.

Continued

Unstructured

- Unstructured data refers to the data that lacks any specific form or structure whatsoever.
- This makes it very difficult and time-consuming to process and analyze unstructured data.
- Email is an example of unstructured data. Structured and unstructured are two important types of big data.

Semi-structured

- Semi-structured data can contain both forms of data.
- It refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.

Additional Information

- Examples of unstructured data: **video, audio or image files, as well as log files, sensor or social media posts**.
- examples of semi-structured data sources are **emails, XML and other markup languages, binary executables, TCP/IP packets, zipped files, data integrated from different sources, and web pages, CSV files**.
- CSV- Common-separated Values
- XML - **extensible markup language**
- The Extensible Markup Language (XML) is **a simple text-based format for representing structured information: documents, data, configuration, books, transactions, invoices, and much more**.

Characteristics of Big Data- 3 ‘V’s of Big Data

Variety

- Variety of Big Data refers to structured, unstructured, and semi-structured data that is gathered from multiple sources.
- Data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data.

Velocity

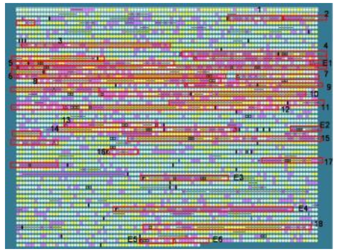
- Velocity essentially refers to the speed at which data is being created in real-time.
- In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

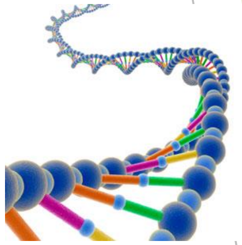
Volume

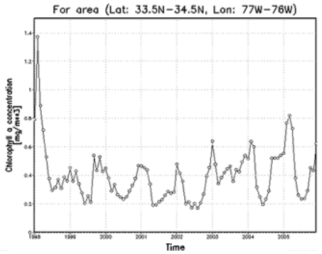
- Big Data indicates huge ‘volumes’ of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data are stored in data warehouses.

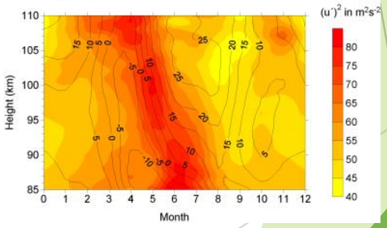
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)



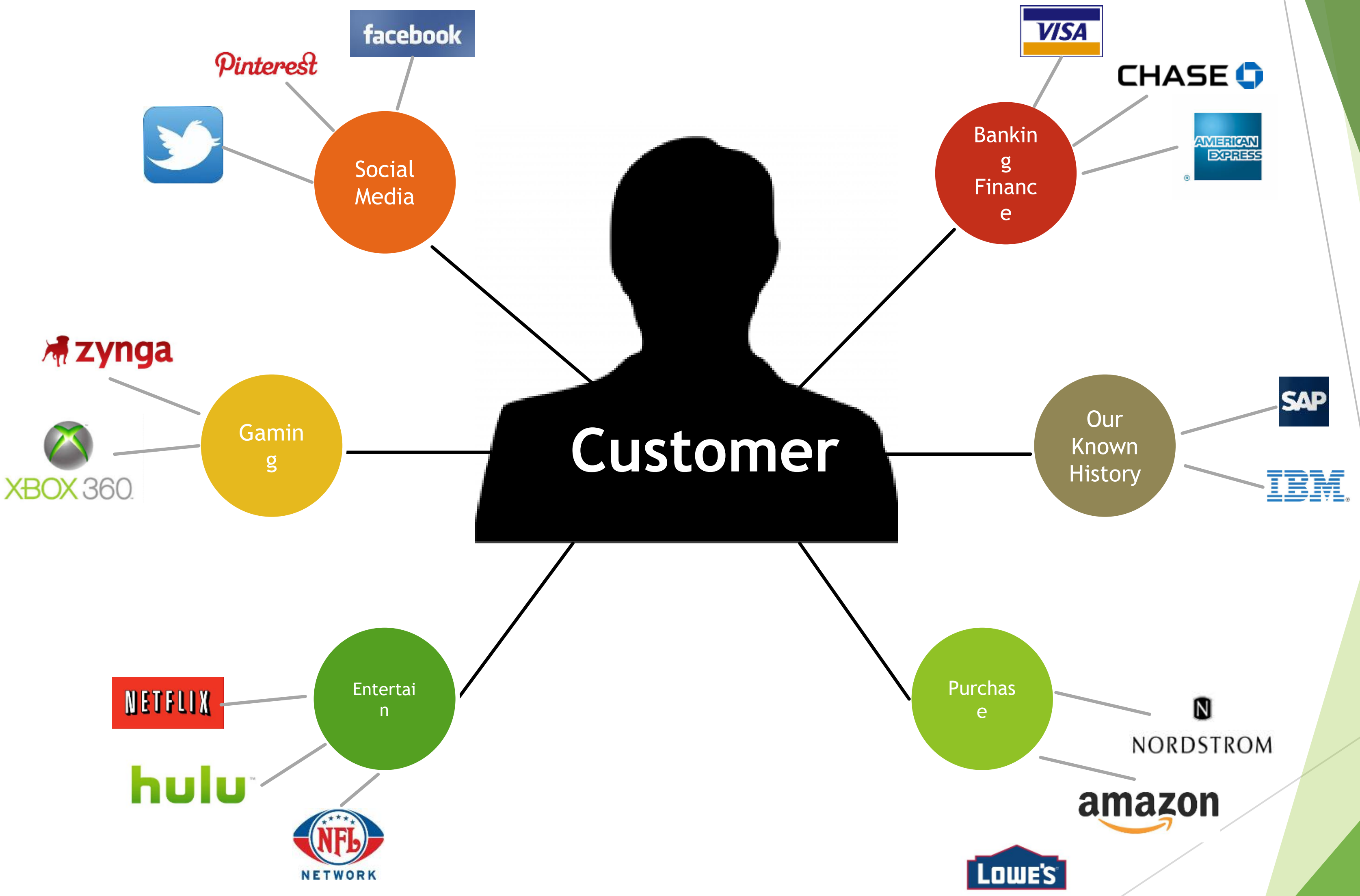



For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration [mg/m**3]
Time



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

A Single View to the Customer



Customer

Social
Media

Bankin
g
Financ
e

Gamin
g

Our
Known
History

Entertai
n

Purchas
e

Velocity (Speed)

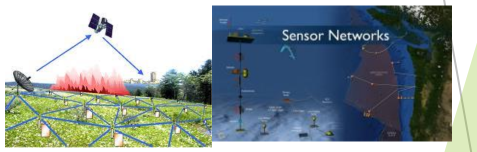
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- Examples
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Handling Big Data

- Big Data handling is the systematic organization, administration as well as governance of massive amounts of data.
- The process includes the management of both unstructured and structured data.
- **Big Data Technologies** is the utilized software that incorporates data mining, data storage, data sharing, and data visualization

Hadoop

- Apache Hadoop is an open-source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data.
- Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.
- Hadoop efficiently stores large volumes of data on a cluster of commodity hardware.
- Hadoop is not only a storage system but also a platform for processing large data along with storage.
- Hadoop supports coding in programming languages like Java, C, C++, Perl, Python, ruby, etc.
- Hadoop is an efficient framework because of running jobs on multiple machines simultaneously which is a parallel processing model.

continued

- Hadoop is apache open source frame work and a large-scale distributed batch processing infrastructure to process large amount of data.
- Apache is the most widely used web server software. Developed and maintained by Apache Software Foundation, Apache is an open source software available for free. It runs on 67% of all webservers in the world. It is fast, reliable, and secure

Hadoop framework consists three main core components.

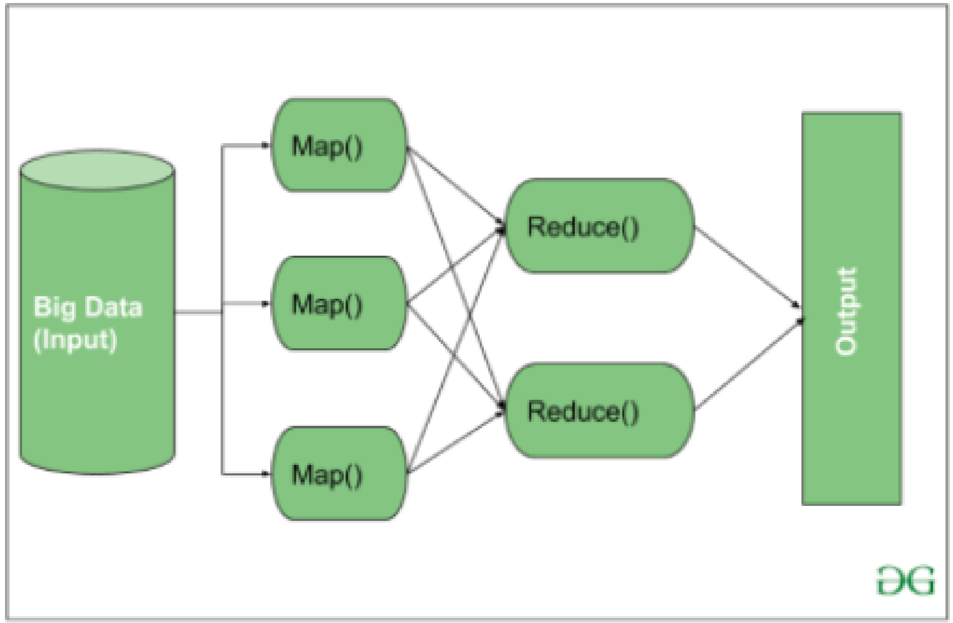
- **HDFS:** It is responsible for storing massive amount of data on the cluster and storage layer of Hadoop.
- **MAP Reduce:** It is responsible for processing massive amount of data on the cluster and data processing layer of Hadoop.
- **Yarn:** It is responsible for resource management.
- HDFS and Map reduce are its kernel for Hadoop.

3 Components

- **HDFS(Hadoop Distributed File System)**: HDFS is working as a storage layer on Hadoop. The data is always stored in the form of data-blocks on HDFS where the default size of each data-block is 128 MB in size which is configurable. HDFS has NameNode and DataNode
- **MapReduce**: MapReduce works as a processing layer on Hadoop. Map-Reduce is a programming model that is mainly divided into two phases Map Phase and Reduce Phase. It is designed for processing the data in parallel which is divided on various machines(nodes).
- **YARN(yet another Resources Negotiator)**: YARN is the job scheduling and resource management layer in Hadoop. The data stored on HDFS is processed and run with the help of data processing engines like graph processing, interactive processing, batch processing, etc. The overall performance of Hadoop is improved up with the Help of this YARN framework.

Hadoop – Architecture

- **MapReduce**
- MapReduce nothing but just like an Algorithm or a data structure that is based on the YARN framework. The major feature of MapReduce is to perform the distributed processing in parallel in a Hadoop cluster which Makes Hadoop working so fast.
- MapReduce has mainly 2 tasks which are divided phase-wise:
- In first phase, **Map** is utilized and in next phase **Reduce** is utilized.



Continued

- **HDFS -**Hadoop Distributed File System
- It is utilized for storage permission is a Hadoop cluster. It is mainly designed for working on commodity Hardware devices(inexpensive devices), working on a distributed file system design.
- HDFS is designed in such a way that it believes more in storing the data in a large chunk of blocks rather than storing small data blocks.
- NameNode(Master) and DataNode(Slave) are the two Data storage Nodes in HDFS.

NameNode: NameNode works as a Master in a Hadoop cluster that guides the Datanode(Slaves).

- Namenode is mainly used for storing the Metadata i.e. the data about the data. Meta Data can be the transaction logs that keep track of the user’s activity in a Hadoop cluster.
- Meta Data can also be the name of the file, size, and the information about the location
- Namenode instructs the DataNodes with the operation like delete, create, Replicate, etc.

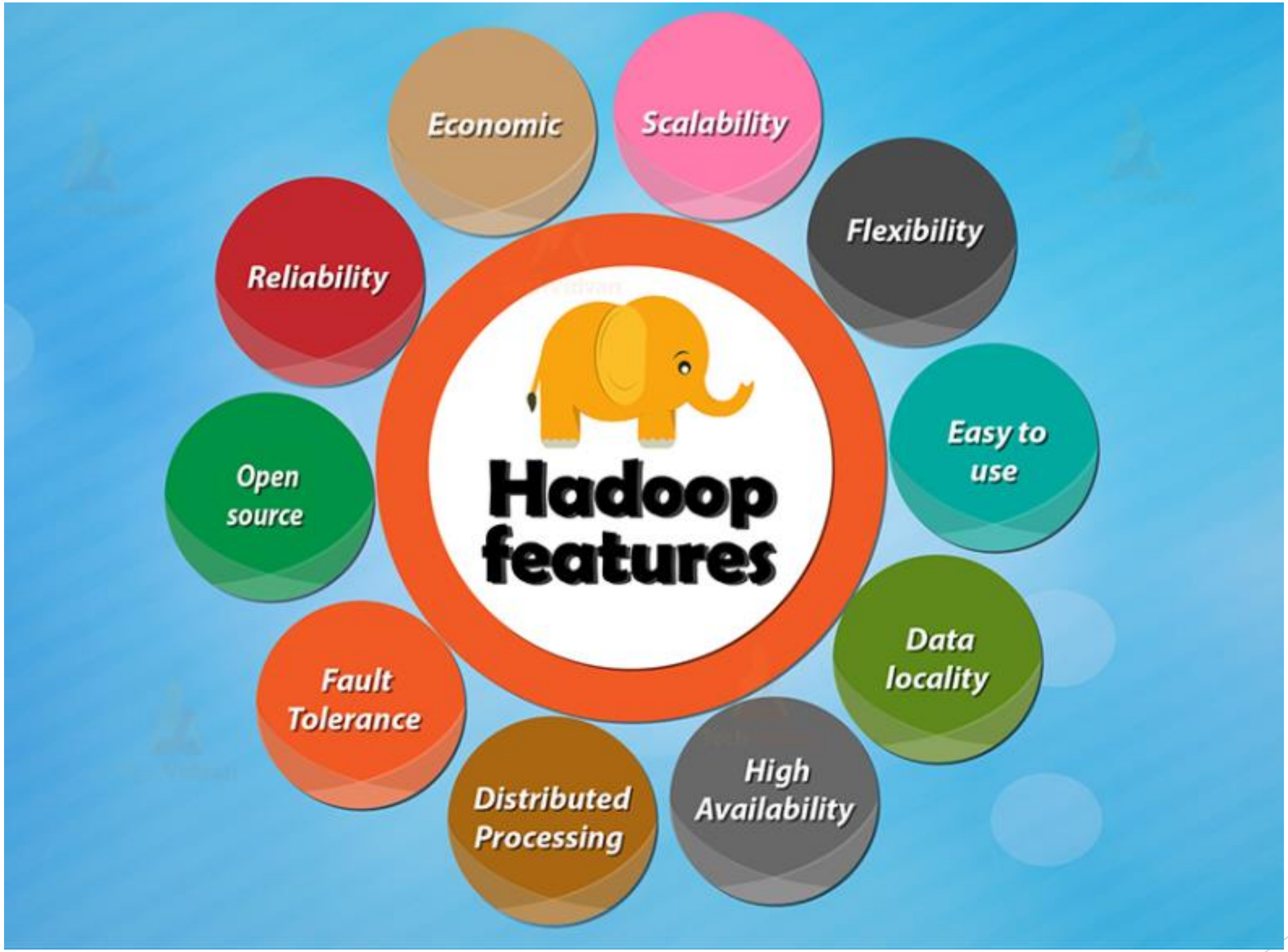
Continued

- **DataNode:** DataNodes works as a Slave
- DataNodes are mainly utilized for storing the data in a Hadoop cluster, the number of DataNodes can be from 1 to 500 or even more than that.
- The more number of DataNode, the Hadoop cluster will be able to store more data. So it is advised that the DataNode should have High storing capacity to store a large number of file blocks.

Continued

YARN(Yet Another Resource Negotiator)

- YARN is a Framework on which MapReduce works.
- YARN performs 2 operations that are Job scheduling and Resource Management.
- The Purpose of Job schedular is to divide a big task into small jobs so that each job can be assigned to various slaves in a Hadoop cluster and Processing can be Maximized.
- Job Scheduler also keeps track of which job is important, which job has more priority, dependencies between the jobs and all the other information like job timing, etc.
- And the use of Resource Manager is to manage all the resources that are made available for running a Hadoop cluster.


Hadoop features
Economic
Scalability
Flexibility
Easy to use
Data locality
High Availability
Distributed Processing
Fault Tolerance
Open source
Reliability

Hadoop Characteristics

- **Open Source**
- Hadoop is an open source project and its code can be modified according to business requirements.
- **Distributed Processing**
- As data is stored in a distributed manner in HDFS across the cluster and data is processed in parallel on a cluster of nodes.
- **Faster**
- Hadoop is extremely good at high-volume batch processing because of its ability to do parallel processing.
- Hadoop can perform batch processes multiple times faster than on single thread server or on the mainframe.

continued

Fault Tolerance

- The data sent to one individual node and the same data also replicates on other nodes in the same cluster.
- If the individual node failed to process the data, the other nodes in the same cluster available to process the data.

Reliability

- Due to data replication in the cluster, data is reliably stored on the cluster of machine despite machine failures.
- If the node failed to process the data, the data will be stored reliably due to this characteristic of Hadoop.

continued

High Availability

- Data is highly available and accessible despite hardware failure due to multiple copies of data.
- If the machine or hardware crashes, then data will be accessed from another path.

Scalability

- Hadoop is a highly scalable storage platform as it can store and distribute very large data sets across hundreds of systems/servers that operate in parallel.
- Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data processing.
- And also supports hardware horizontal scalability which can add the nodes during the processing without system downtime.

continued

Flexibility

- Hadoop manages data whether structured or unstructured, encoded or formatted, or any other type of data.
- Businesses can use Hadoop to derive valuable business insights from data sources such as social media, email conversations.
- Hadoop brings value to the table where unstructured data can be useful in the decision-making process.

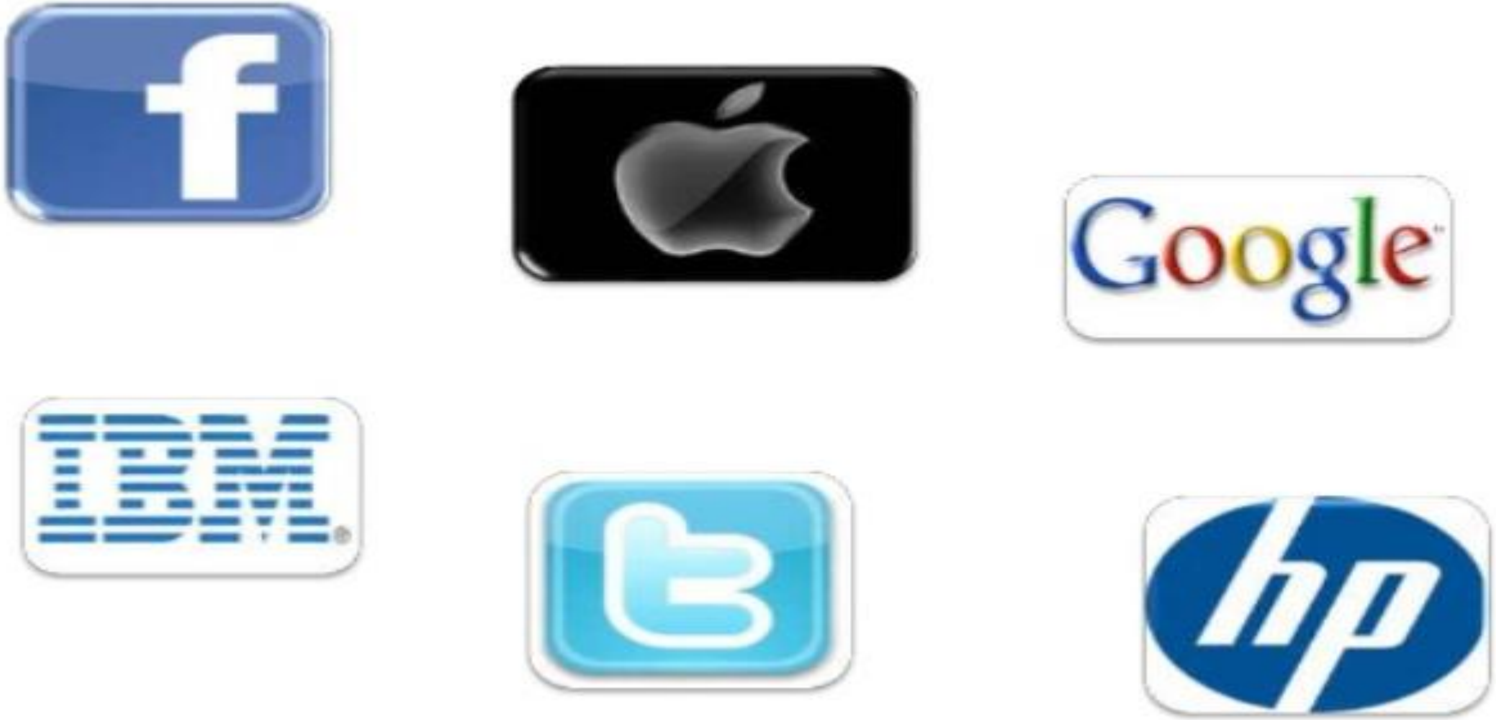
Economic/Cost-effective

- Hadoop offers a cost-effective storage solution for businesses exploding data sets.
- Hadoop is not very expensive as it runs on a cluster of commodity hardware.

Easy to use

- No need of the client to deal with distributed computing, the framework takes care of all the things. So, Hadoop is easy to use.

Famous Hadoop users



Additional Information

- A distributed file system (DFS) is **a file system with data stored on a server**. The data is accessed and processed as if it was stored on the local client machine.
- HDFS is **a distributed file system that provides access to data across Hadoop clusters**. A cluster is a group of computers that work together
- Multitenancy environment: multiple customers share the same application, in the same operating environment, on the same hardware, with the same storage mechanism. In virtualization, every application runs on a separate virtual machine with its own OS.
- A Multi-tier Architecture is **a software architecture in which different software components, organized in tiers (layers), provide dedicated functionality**.

Top 5 Cloud providers

Microsoft Azure

Google Cloud

Adobe
Creative Cloud

Cloud providers

- public cloud providers: **AWS, Microsoft and Google**.
- private cloud providers: **HP Data Centers, Microsoft, Elastra-private cloud, and Ubuntu**
- Hybrid cloud providers: **Microsoft**, **Vmware**, **IBM**, **Cisco**, **HP**

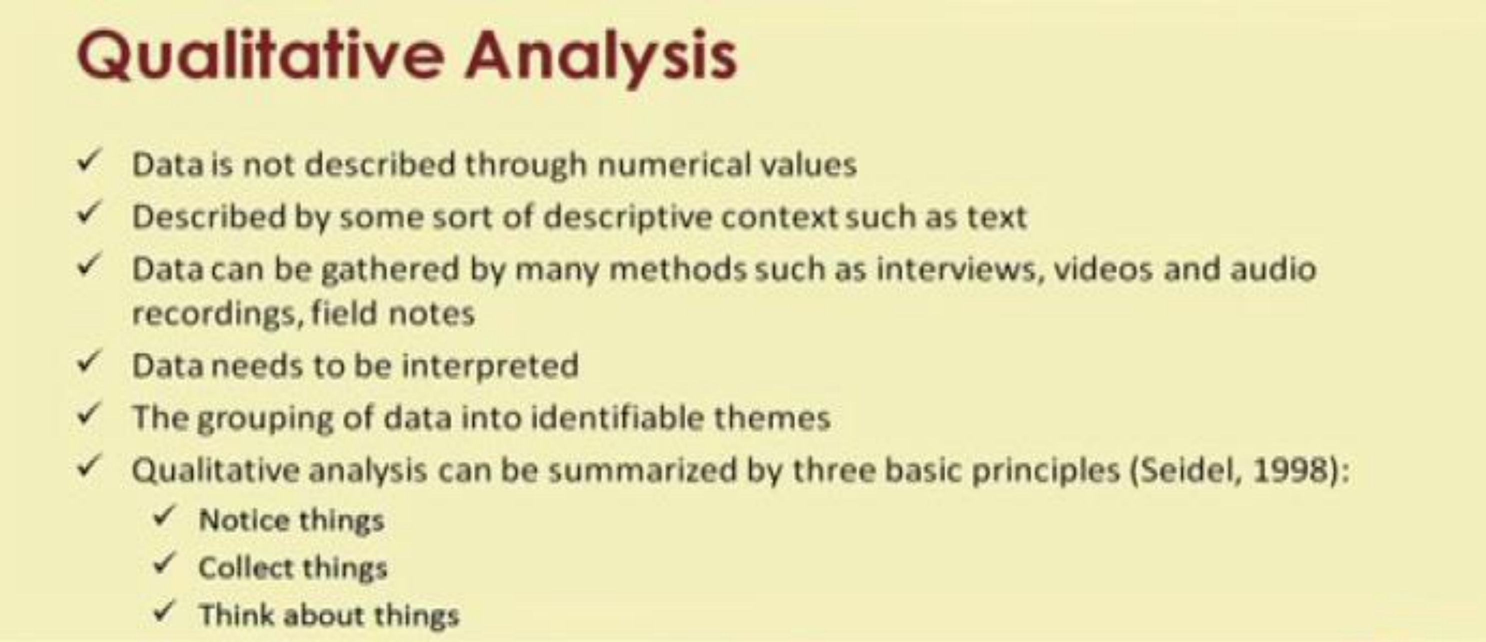
Data Analytics

- Data analytics is a process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software.
- It refers to the process and practice of analyzing data to answer questions, extract insights, and identify trends.
- This is done by using an array of tools, techniques, and frameworks that vary depending on the type of analysis being conducted.
- Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more informed business decisions and by scientists and researchers to verify or disprove the scientific models theories and hypothesis.

Types of Data Analysis

Two types of analysis

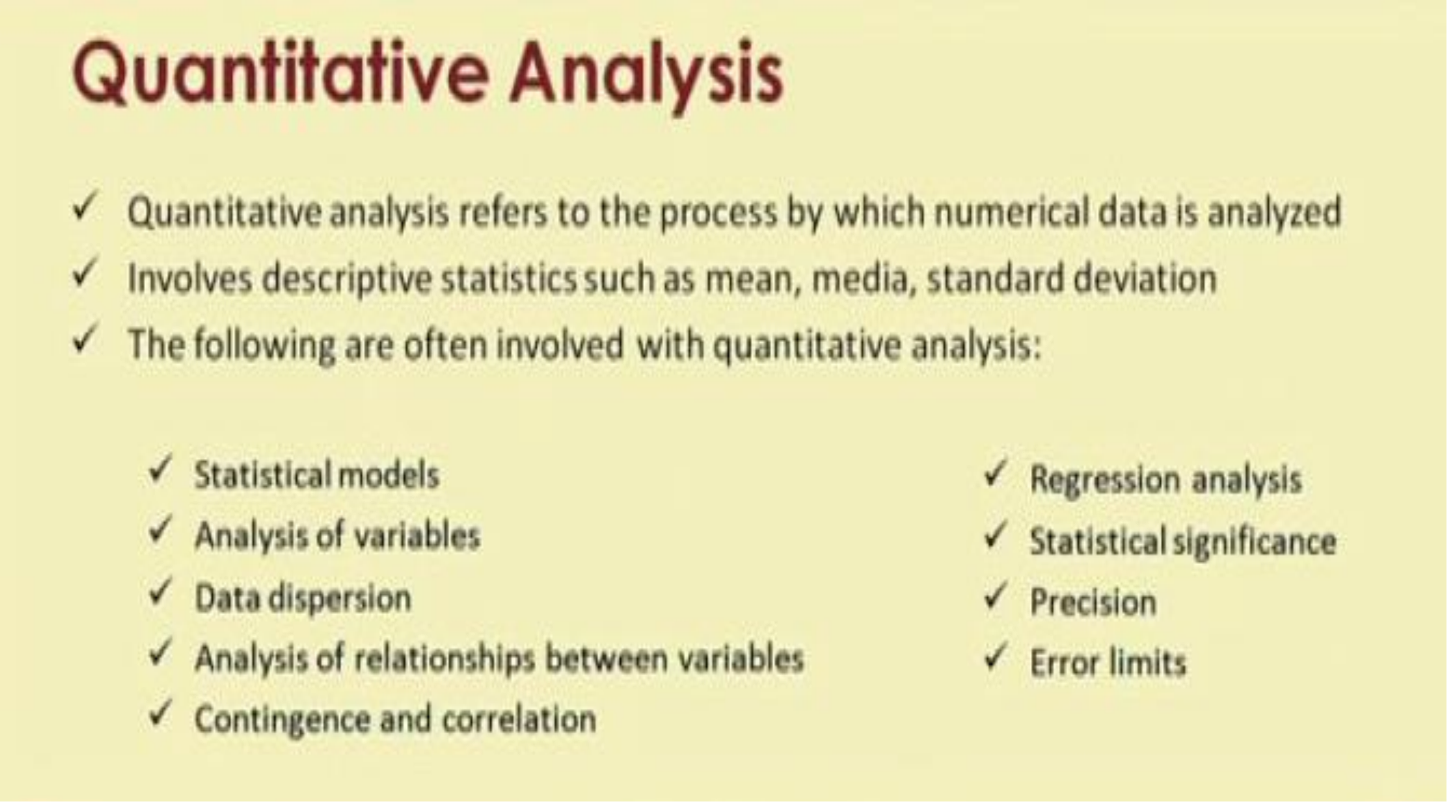
- Qualitative analysis**:** Deals with the analysis of data that are categorical in nature
- Quantitative analysis **:** Refers to the process by which numerical methods can be used, numerical data can be analyzed through quantitative analysis.



Qualitative Analysis

- ✓ Data is not described through numerical values
- ✓ Described by some sort of descriptive context such as text
- ✓ Data can be gathered by many methods such as interviews, videos and audio recordings, field notes
- ✓ Data needs to be interpreted
- ✓ The grouping of data into identifiable themes
- ✓ Qualitative analysis can be summarized by three basic principles (Seidel, 1998):
  - ✓ Notice things
  - ✓ Collect things
  - ✓ Think about things

continued



Quantitative Analysis

- ✓ Quantitative analysis refers to the process by which numerical data is analyzed
- ✓ Involves descriptive statistics such as mean, media, standard deviation
- ✓ The following are often involved with quantitative analysis:
  - ✓ Statistical models
  - ✓ Analysis of variables
  - ✓ Data dispersion
  - ✓ Analysis of relationships between variables
  - ✓ Contingence and correlation
  - ✓ Regression analysis
  - ✓ Statistical significance
  - ✓ Precision
  - ✓ Error limits

quantitative analysis

- Quantitative analysis is often associated with numerical analysis where data is collected, classified, and then computed for certain findings using a set of statistical methods.
- Data is chosen randomly in large samples and then analyzed.
- The advantage of quantitative analysis, the findings can be applied in a general population using research patterns developed in the sample.
- Quantitative analysis is generally concerned with measurable quantities such as weight, length, temperature, speed, width, and many more.
- The data can be expressed in a tabular form or any diagrammatic representation using graphs or charts.
- Quantitative data can be classified as continuous or discrete, and it is often obtained using surveys, observations, experiments, or interviews.

qualitative analysis

- Qualitative analysis is concerned with the analysis of data that cannot be quantified.
- This type of data is about the understanding and insights into the properties and attributes of objects (participants).
- qualitative analysis seeks to get a deeper understanding, the researcher must be well-rounded with whichever physical properties or attributes the study is based on.
- The typical data analyzed qualitatively include color, gender, nationality, taste, appearance, and many more as long as the data cannot be computed.
- Such data is obtained using interviews or observations.
- the data is collected in small, unrepresentative samples in an unstructured way

Comparison

Parameters	Qualitative Analysis	Quantitative Analysis
Definition	It is based on classification of objects (participants) according to properties and attributes	It is based on classification of data based on computable values
Data collection	data collected include color, race, religion, nationality, and many more	data is collected in large, representative samples that can generalize the entire population.
Research methodology	methodology is exploratory where the analysis seeks to get a deeper understanding of why a certain phenomenon occurs.	The methodology in quantitative analysis can be conclusive such as how much or how many times a certain phenomenon occurs not why it does occur.
research findings	are specific to the objects being studied and are not applicable on the general population	the findings can be applicable on the general population.
Method of data collection	researchers often ask open-ended questions, conduct interviews, and observations	researchers take measurements, conduct surveys, experiments and observations.

Types of Data Analytics

- **Predictive data analytics**
- Predictive analytics may be the most commonly used category of data analytics. Businesses use predictive analytics to identify trends, correlations, and causation.
- The category can be further broken down into predictive modeling and statistical modeling
- Predictive analytics looks forward to attempting to divine unknown future events or actions based on data mining, statistics, modeling, deep learning and artificial intelligence, and machine learning
- Predictive models are applied to business activities to better understand customers, with the goal of predicting buying patterns, potential risks, and likely opportunities.
- Example: Retail, health, weather forecasting, energy etc.

Prescriptive data analytics

- prescriptive analytics factors information about possible situations or scenarios, available resources, past performance, and current performance, and suggests a course of action or strategy. It can be used to make decisions on any time horizon, from immediate to long term.
- By considering all relevant factors, this type of analysis yields recommendations for next steps.
- Prescriptive analytics is where AI and big data combine to help predict outcomes and identify what actions to take.
- This category of analytics can be further broken down into *optimization* and *random testing*.

Diagnostic data analytics

- Diagnostic data analytics is the process of examining data to understand the cause and event or why something happened.
- Diagnostic data analytics help answer why something occurred.
- Techniques such as drill down, data discovery, data mining, and correlations are often employed.
- Drilling down involves focusing on a certain facet of the data or particular widget.
- In the discovery process, analysts identify the data sources that will help them interpret the results.
- Data mining is an automated process to get information from a massive set of raw data. And finding consistent correlations in your data can help you pinpoint the parameters of the investigation.
- Diagnostic data is data that is automatically recorded by infrastructure, vehicles, machines, software, and devices for the purposes of troubleshooting problems.

Descriptive data analytics

- Descriptive analytics is the process of using current and historical data to identify trends and relationships.
- It’s sometimes called the simplest form of data analysis because it describes trends and relationships but doesn’t dig deeper.
- Descriptive analytics is relatively accessible and likely something your organization uses daily.
- Basic statistical software, such as Microsoft Excel or data visualization tools, such as Google Charts and Tableau, can help parse data, identify trends and relationships between variables, and visually display information.
- Descriptive analytics is especially useful for communicating change over time.

Thank You