### **Theory**

**What is Big Data?**

* The term Big Data refers to a huge volume of data that can not be stored, processed by any traditional data storage or processing units.
* Big Data is generated at a very large scale and it is being used by many multinational companies to process and analyse in order to get insights and improve the business of many organisations.

**How is Big Data generated?**

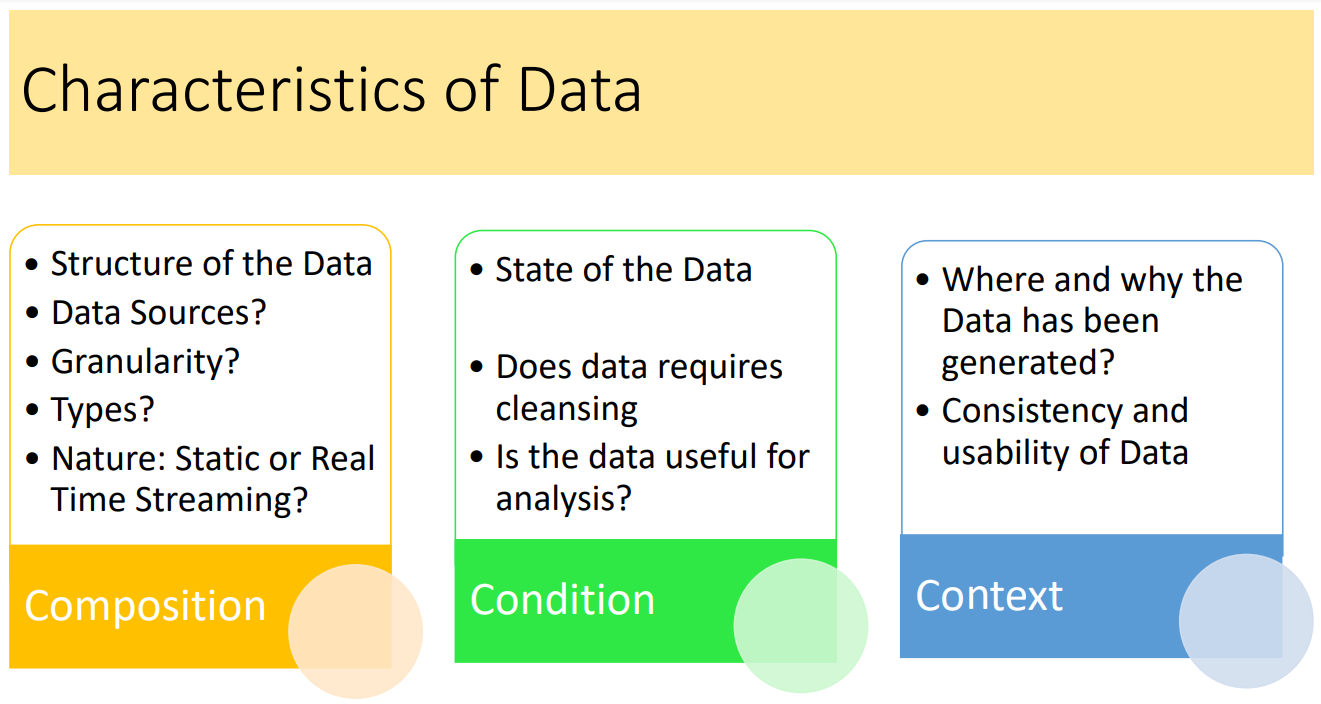
* With the development and increase of apps and social media and people and businesses moving online, there’s been a huge increase in data.
* Devices and sensors automatically generate diagnostic information on a daily basis.

**Why Big Data?**

* For businesses, Big Data helps to bring efficiency, quality, and personalized products and services, resulting in improved levels of customer satisfaction and profit.
* Enhancing operational Efficiency, Reducing cost & time, innovating new products & services and optimizing existing services.

**Big Data Characteristics**

* **Huge volume of data:** Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.
* **Complexity of data types:** Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.
* **Speed of new data creation and growth:** Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.



**Challenges of Big Data**

* The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

**Data Structures / Types of Big Data**

Big data can come in multiple forms, including structured, Semi-structured and unstructured data such as financial data, text files and multimedia files.

1. **Structured data**

* Data containing a defined data type, format, and structure (that is, transaction data, OLAP data cubes, traditional RDBMS, CSV files, and even simple spreadsheets).
* The structured data is stored, processed and retrieved in a fixed format.
* Such data can be stored in Excel, database etc.

1. **Semi-structured data**

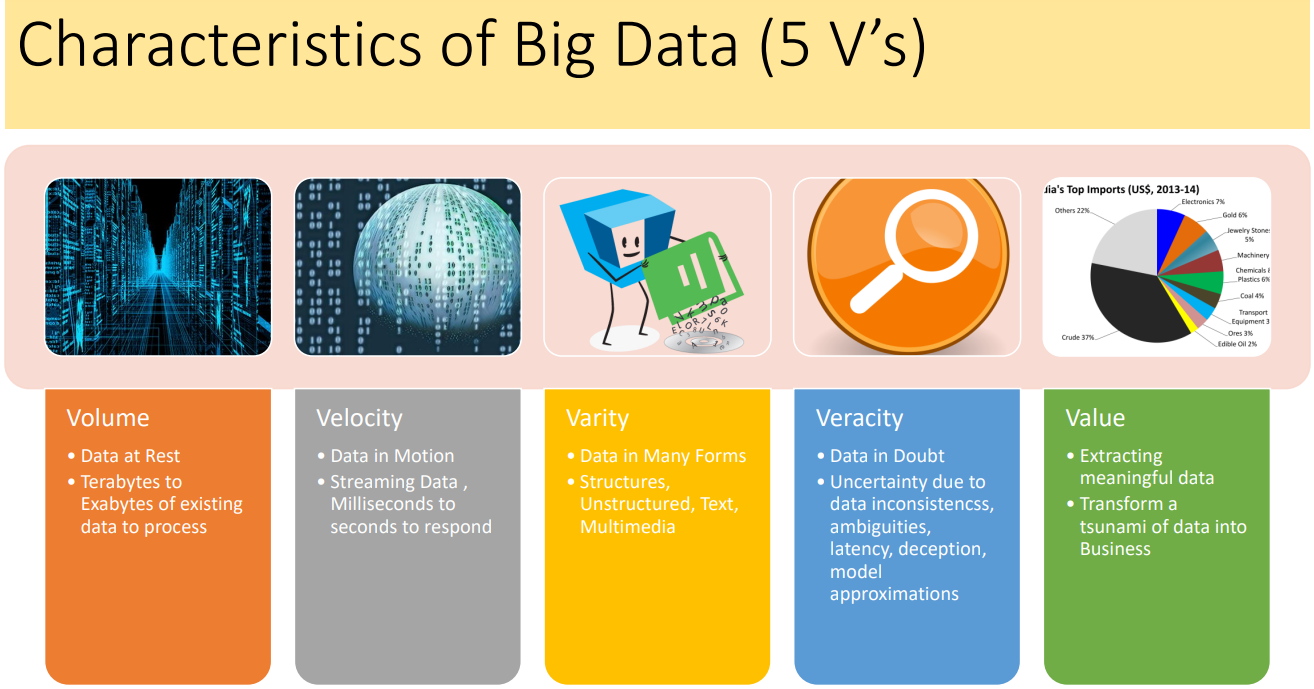
* Semi structured data pertains to the data containing both formats that is structured and unstructured
* To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data
* Examples, XML data, Json Data files.

1. **Quasi-structured data**

* Textual data with erratic data formats that can be formatted with effort, tools, and time (for instance, web clickstream data that may contain inconsistencies in data values and formats).

1. **Unstructured data**

* Data that has no inherent structure, which may include text documents, PDFs, images, email and video.
* Analysis of unstructured data is time consuming.



**Volume**

* Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, sensors, images and video.
* A particular data can actually be considered as a Big Data or not, depending upon the volume of data.

**Velocity**

* The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands.
* Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

**Variety**

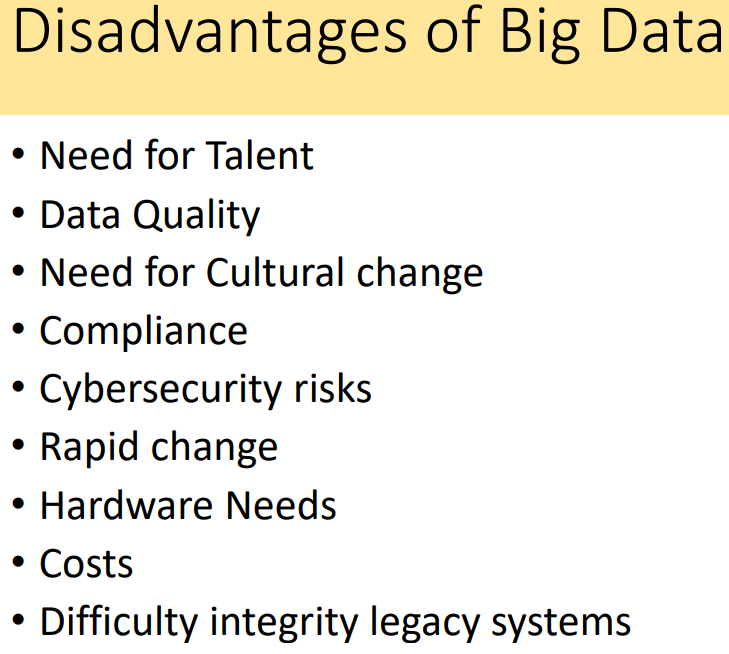
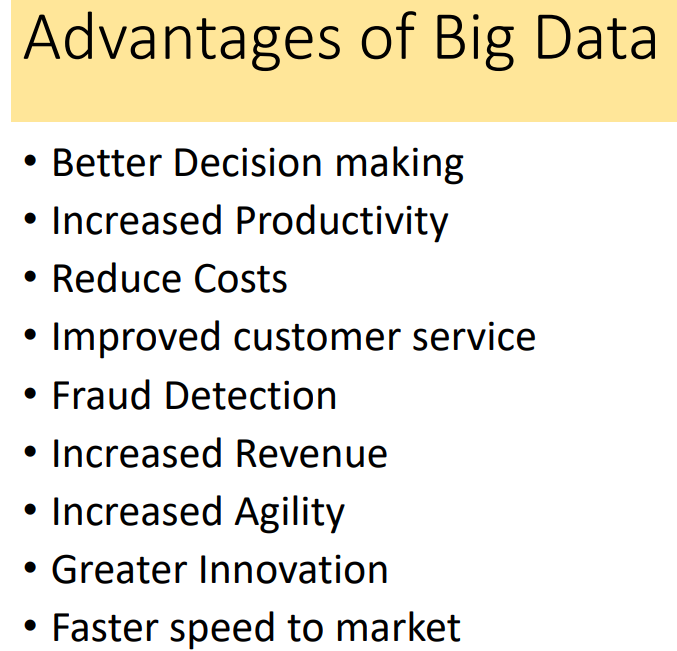
* Variety refers to heterogeneous sources and the nature of data, both structured and unstructured.
* This variety of unstructured data poses certain issues for storage, mining and analysing data.

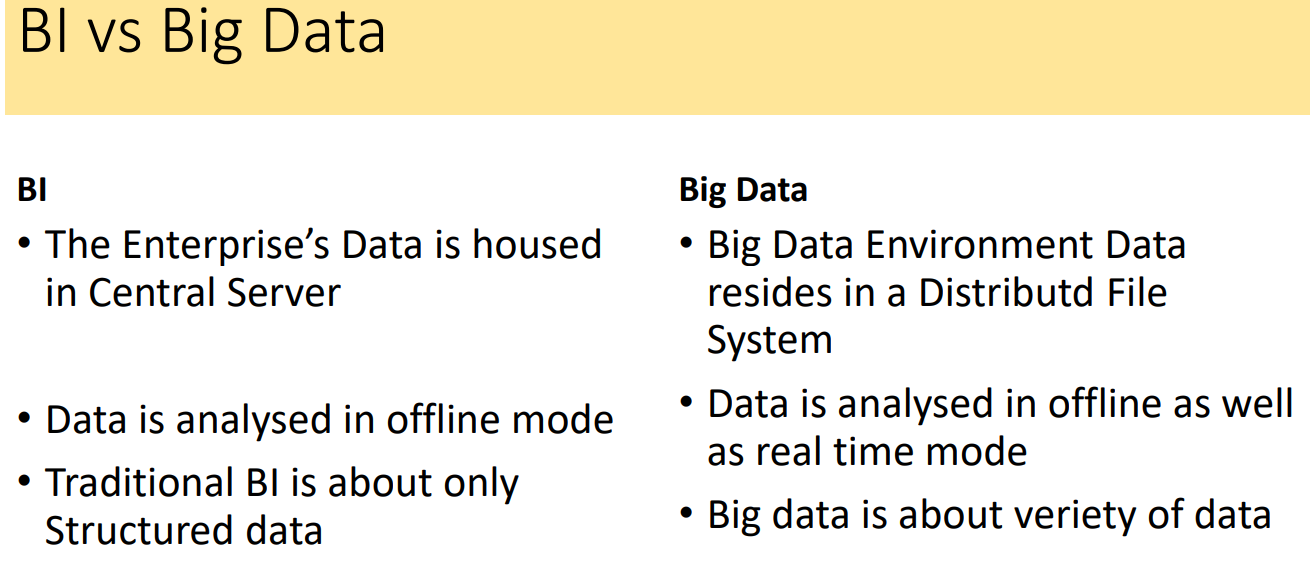
**Veracity**

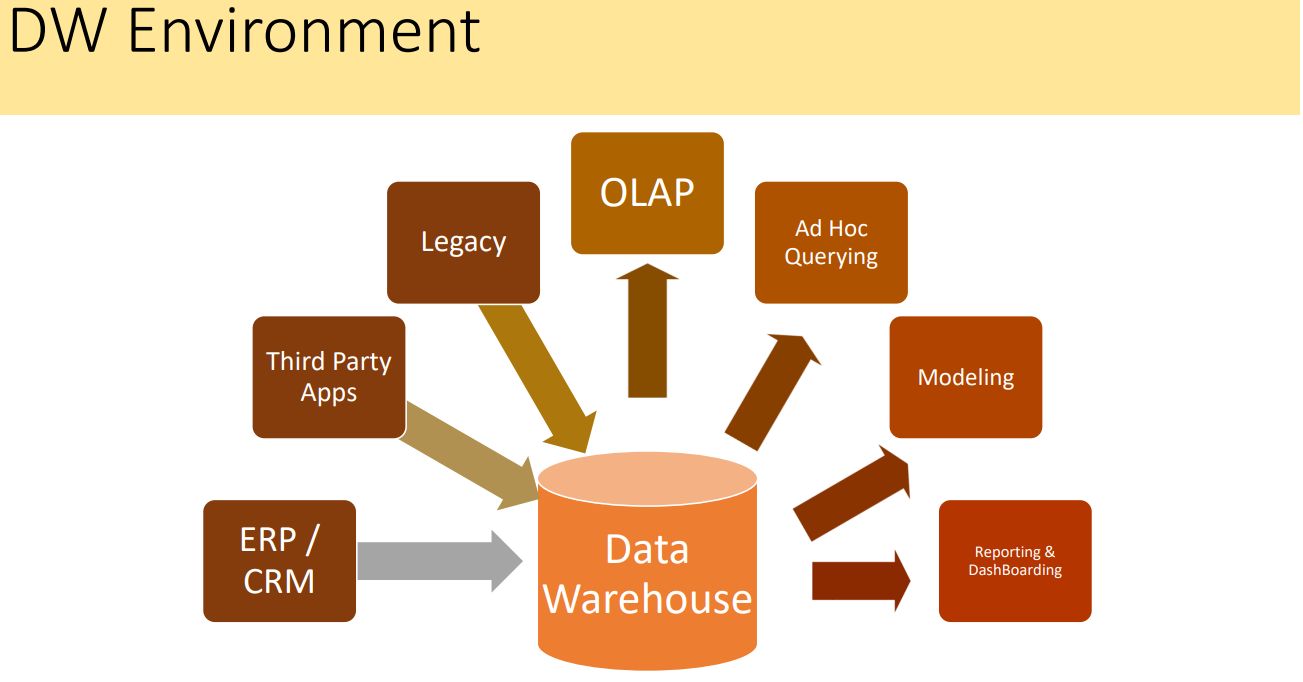
* Veracity basically means the degree of reliability that the data has to offer. Since a major part of the data is unstructured and irrelevant
* Uncertainty due to data inconsistencies, ambiguities, latency, deception, model approximations

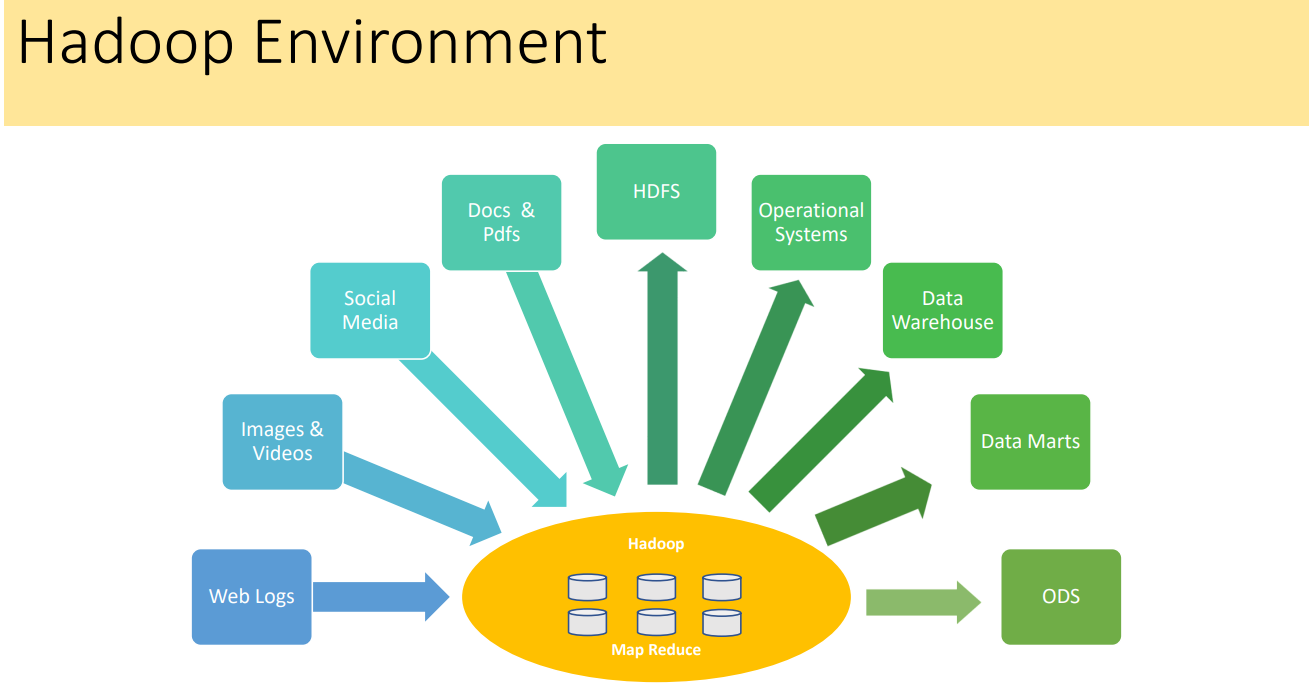
**Value**

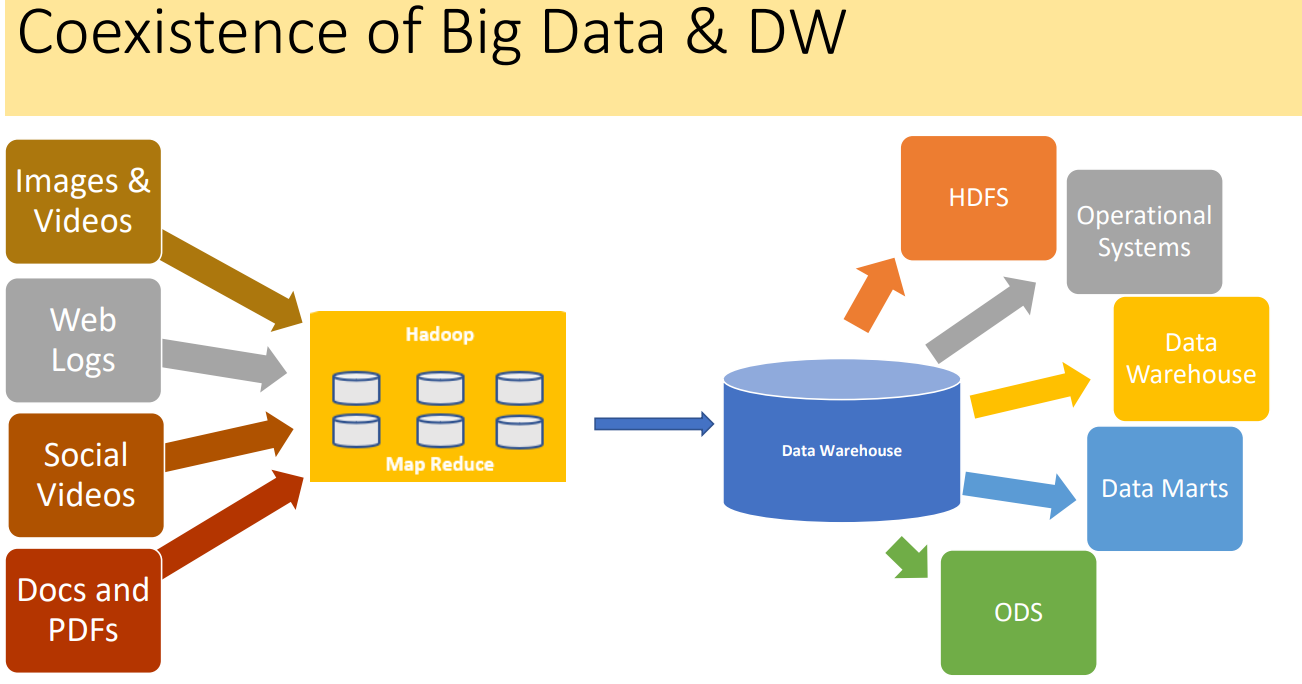
* Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable data that needs to be stored, processed, analyzed to find insights.





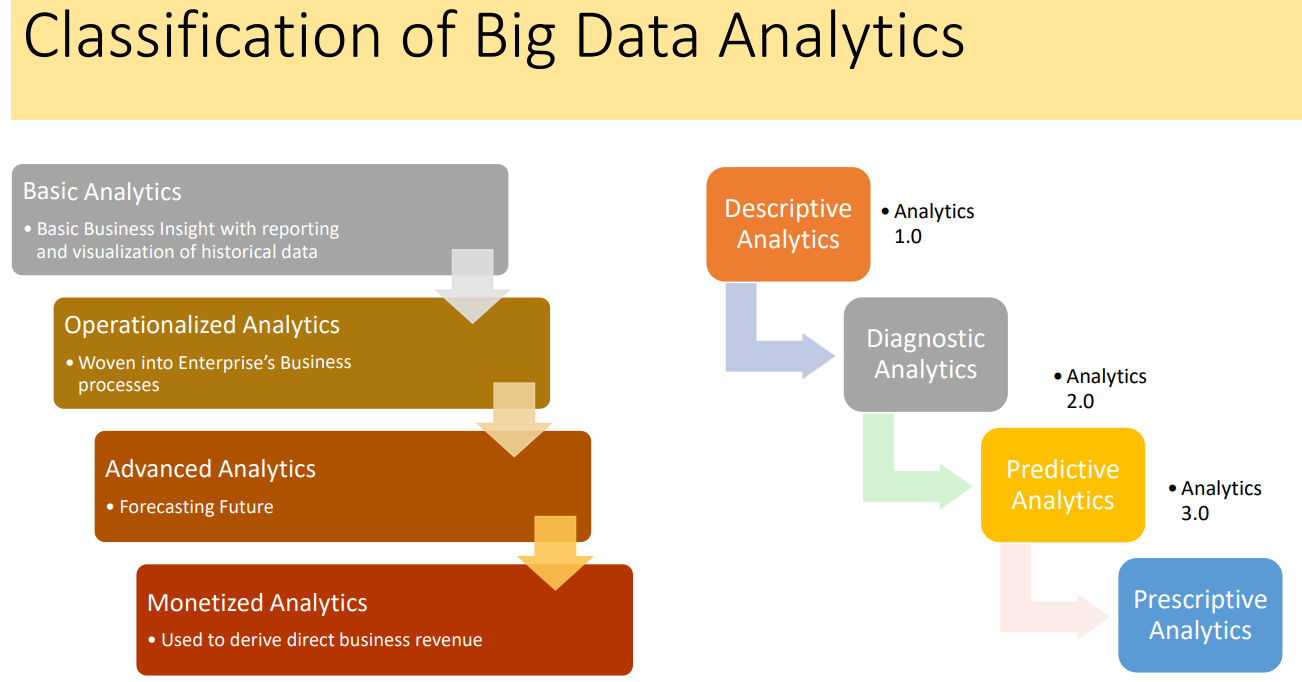






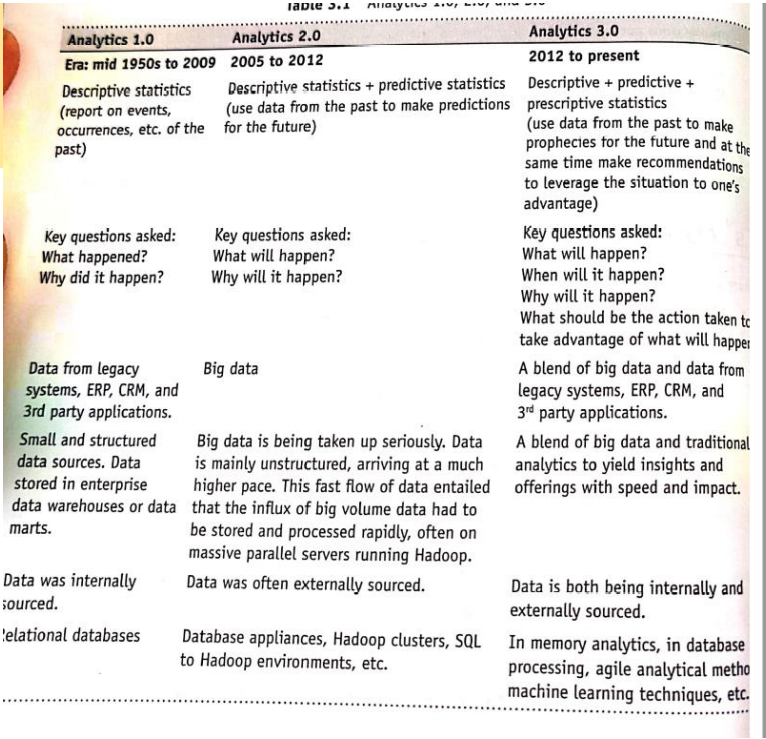
**What is Big Data Analytics?**

* Technology enabled Analytics
* Competitive advantage
* Reacher / Deeper business insight
* Real-Time Analytics
* Collaboration with data scientist and business users
* Working with huge datasets
* Better and faster decisions in real time
* Move code to data with greater speed & efficiency



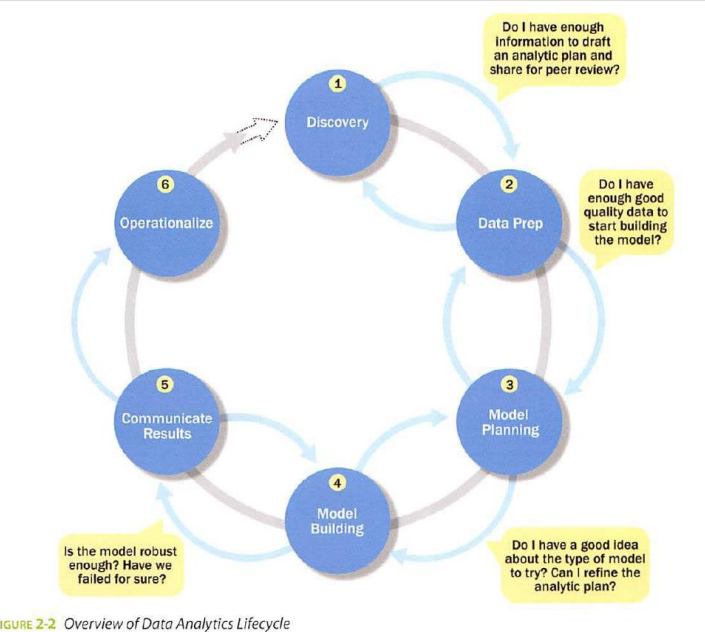
**Importance of Big Data Analytics**

* **Cost reduction:** Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
* **Faster, better decision making:** With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they’ve learned.
* **New products and services:** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers’ needs.



**Data Analytics Life Cycle**

The Data Analytics Lifecycle is designed specifically for Big Data problems and data science projects. The life cycle has six phases, and project work can occur in several phases at once.



**Discovery**: This phase involves understanding the business problem and context. It includes identifying key stakeholders, learning the business domain, framing the problem, and developing an initial hypothesis. It also involves determining potential data sources and assessing the necessary infrastructure.

**Data Preparation**: In this phase, raw data is cleaned and transformed to ensure it is suitable for analysis. This includes setting up an analytic sandbox, performing ETLT (Extract, Transform, Load, Test), data conditioning, and exploring the data through surveys and visualizations using tools like Hadoop and OpenRefine.

**Model Planning**: The focus here is on selecting the right analytical model based on the data and problem. Data exploration, variable selection, and model choice are done in this phase. Tools like R and SAS/ACCESS are commonly used for planning.

**Model Building**: During this phase, models are built and tested. Key activities include evaluating the model’s accuracy on test data, ensuring domain relevance, avoiding major errors, and refining the model as needed. Commercial tools like SAS and open-source tools like Python are used.

**Communicate Results**: After building the model, the results are communicated to stakeholders. This involves assessing the statistical significance of the results, documenting key findings, and presenting insights that can inform business decisions.

**Operationalize**: In this final phase, the model is deployed in a controlled, real-world setting through a pilot project. Continuous monitoring is set up to track model accuracy, and adjustments or retraining are done if performance drops.

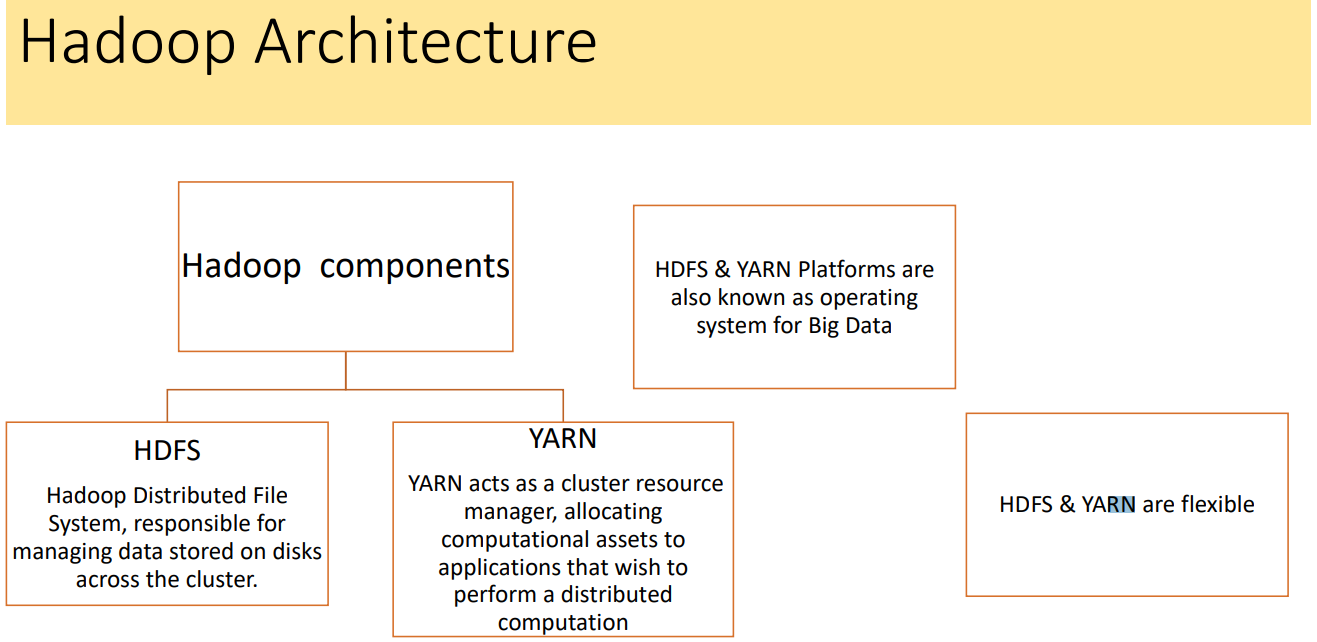
A **distributed system** must meet the following requirements

• Fault Tolerance: If a component fails, it should not result in the failure of the entire system. The system should gracefully degrade into a lower performing state. If a failed com‐ ponent recovers, it should be able to rejoin the system.

• Recoverability: In the event of failure, no data should be lost.

• Consistency: The failure of one job or task should not affect the final result

• Scalability: Adding load (more data, more computation) leads to a decline in performance, not failure; increasing resources should result in a proportional increase in capacity



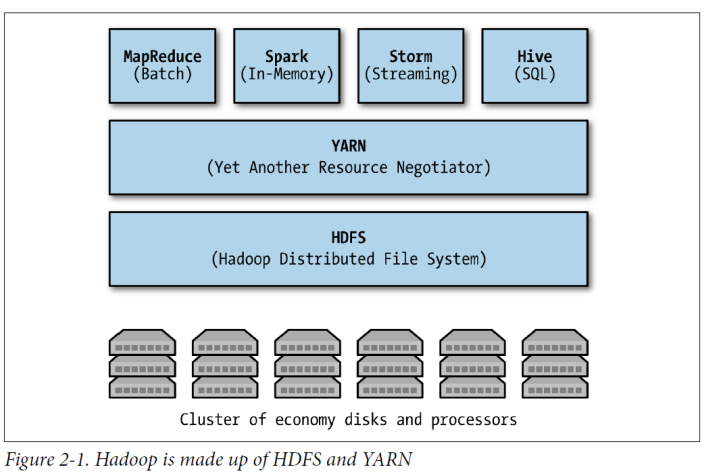
**Hadoop Architecture**

• HDFS and YARN work in concert to minimize the amount of network traffic in the cluster primarily by ensuring that data is local to the required computation.

• Duplication of both data and tasks ensures fault tolerance, recoverability, and consistency.

• Moreover, the cluster is centrally managed to provide scalability and to abstract low level clustering programming details.

• Together, HDFS and YARN are a platform upon which big data applications are built; perhaps more than just a platform, they provide an operating system for big data.



• Like any good operating system, HDFS and YARN are flexible.

• Other data storage systems aside from HDFS can be integrated into the Hadoop

framework such as Amazon S3 or Cassandra.

• Alternatively, data storage systems can be built directly on top of HDFS to provide more

features than a simple file system.

• For example, HBase is a columnar data store built on top of HDFS and is one the most

advanced analytical applications that leverage distributed storage.

• In earlier versions of Hadoop, applications that wanted to leverage distributed computing

on a Hadoop cluster had to translate user-level implementations into MapReduce jobs.

• However, YARN now allows richer abstractions of the cluster utility, making new data

processing applications for machine learning, graph analysis, SQL-like querying of data, or

even streaming data services faster and more easily implemented.

• As a result, a rich eco‐ system of tools and technologies has been built up around

Hadoop, specifically on top of YARN and HDFS

**Hadoop HDFS**

• Data is stored in a distributed manner in HDFS. There are two components of HDFS - name node and data node. While there is only one name node, there can be multiple data nodes.

• Hadoop enables you to use commodity machines as your data nodes. This way, you don’t have to spend millions of dollars just on your data nodes. However, the name node is always an enterprise server.

**Features of HDFS**

• Provides distributed storage • Can be implemented on commodity hardware • Provides data security • Highly fault-tolerant - If one machine goes down, the data from that machine goes to the next machine

**Master and Slave Nodes**

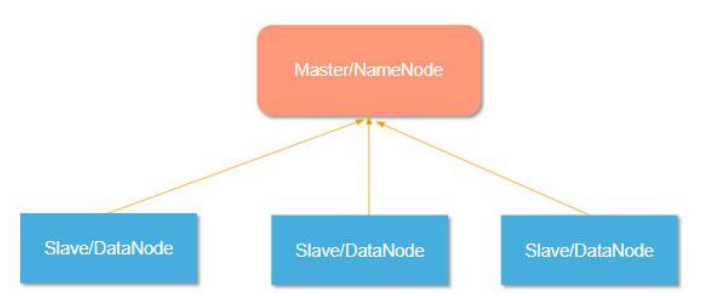
• Master and slave nodes form the HDFS cluster. The name node is called the master, and the data nodes are called slaves.

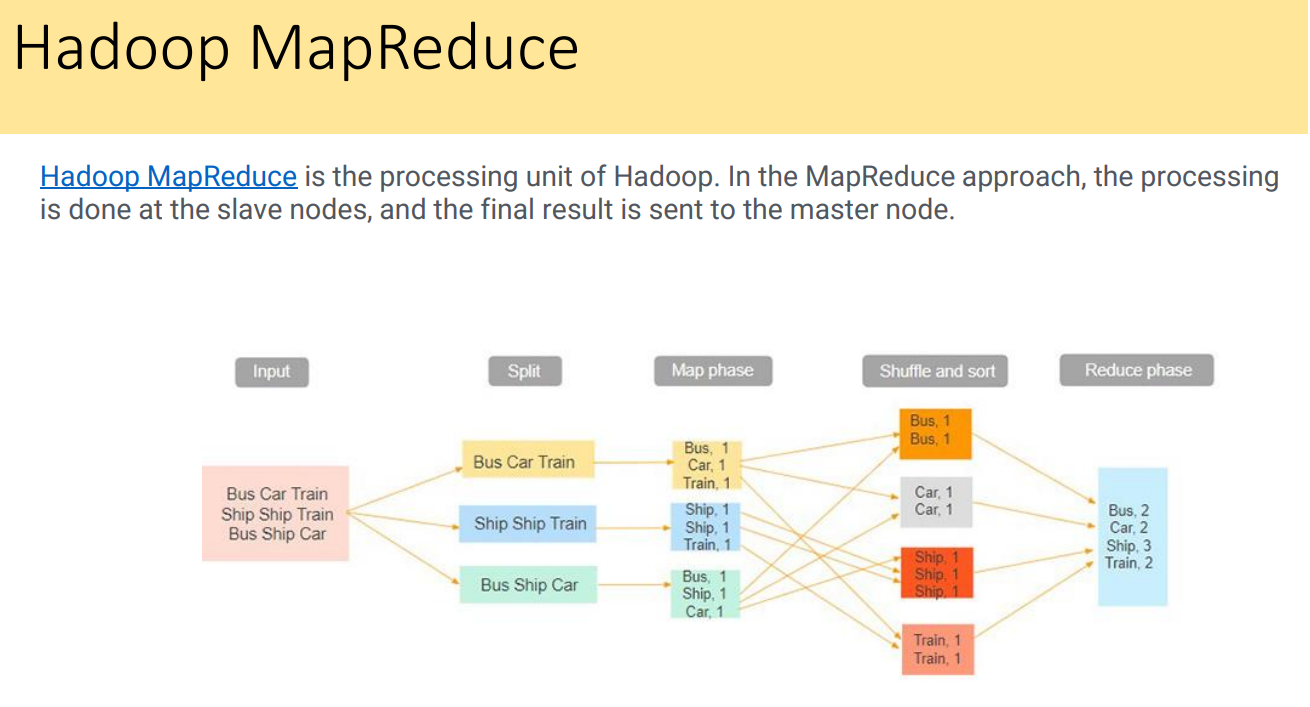
• The name node is responsible for the workings of the data nodes. It also stores the Metadata.

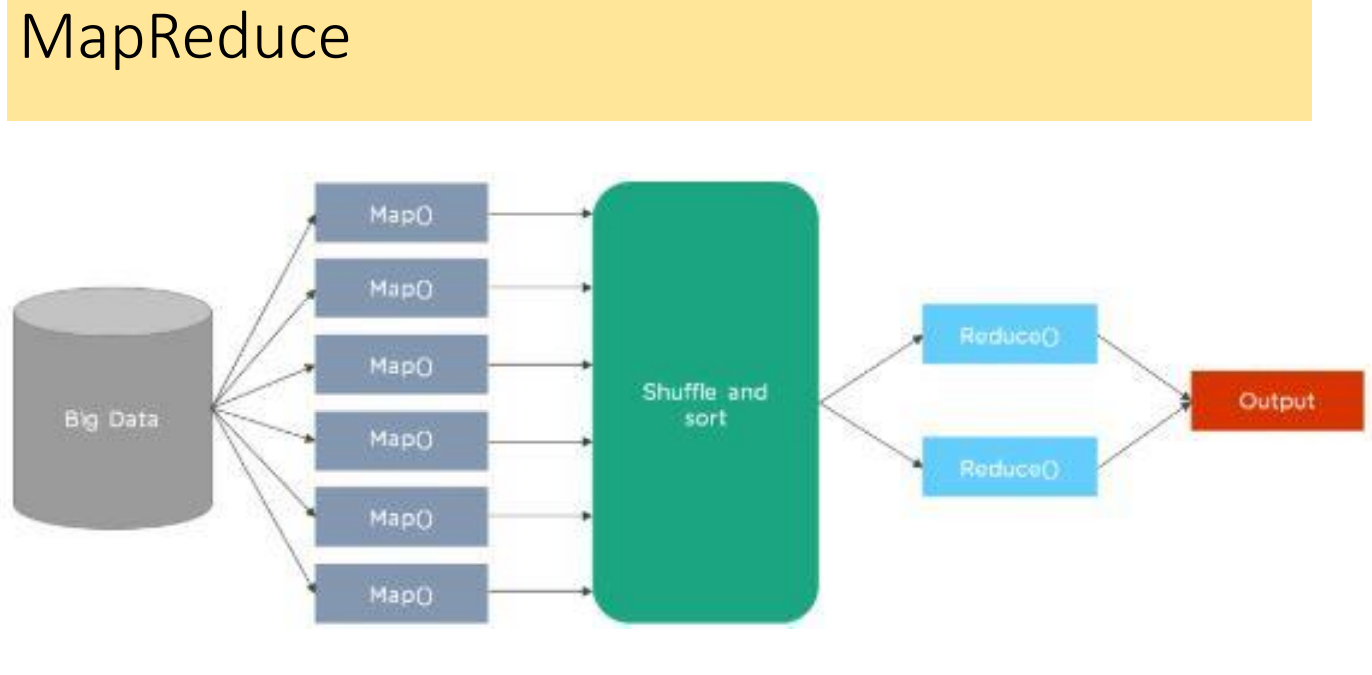
• The data nodes read, write, process, and replicate the data. They also send signals,

known as heartbeats, to the name node. These heartbeats show the status of the data

node.







### **MCQ**

1. What does the term "Big Data" refer to?

a) A small set of data used for analysis  
b) A large volume of data that traditional data storage cannot handle  
c) Data that is manually processed  
d) Data that only deals with numbers and figures  
**Answer: b) A large volume of data that traditional data storage cannot handle**

2. Why is Big Data important for businesses?

a) It helps to reduce the need for employee training  
b) It allows businesses to derive insights and gain a competitive advantage  
c) It decreases the use of technology in businesses  
d) It creates new competitors in the market  
**Answer: b) It allows businesses to derive insights and gain a competitive advantage**

3. How does Big Data contribute to business improvement?

a) By reducing the volume of data in the company  
b) By driving efficiency, quality, and personalized products and services  
c) By reducing the company's storage costs  
d) By simplifying all business processes  
**Answer: b) By driving efficiency, quality, and personalized products and services**

4. Which of the following technologies are part of Big Data Analytics?

a) Machine learning, data mining, and statistics  
b) Cloud computing  
c) Virtual reality  
d) Blockchain  
**Answer: a) Machine learning, data mining, and statistics**

5. In which way does Big Data benefit scientific efforts?

a) It enables investigations with richer results and deeper insights  
b) It simplifies research by focusing only on structured data  
c) It limits the scope of research to only small datasets  
d) It eliminates the need for researchers to analyze data manually  
**Answer: a) It enables investigations with richer results and deeper insights**

6. What makes Big Data different from traditional data?

a) It is mainly focused on structured data  
b) It deals with smaller datasets  
c) It includes both structured and unstructured data, with real-time feeds and queries  
d) It is generated by a single device or source  
**Answer: c) It includes both structured and unstructured data, with real-time feeds and queries**

7. How do credit card companies use Big Data?

a) To reduce the number of transactions per user  
b) To monitor every purchase and identify fraudulent activities  
c) To prevent customers from making purchases  
d) To limit the number of credit card users  
**Answer: b) To monitor every purchase and identify fraudulent activities**

8. How do mobile phone companies use Big Data?

a) To track customers who don’t pay bills on time  
b) To analyze calling patterns and predict potential defections  
c) To reduce network coverage  
d) To prevent calls to rival networks  
**Answer: b) To analyze calling patterns and predict potential defections**

9. What is a primary product for companies like LinkedIn and Facebook?

a) Data storage  
b) Customer support services  
c) Data itself  
d) Software tools for analysis  
**Answer: c) Data itself**

10. What is one defining characteristic of Big Data in terms of volume?

a) It consists of thousands or millions of rows  
b) It can contain billions of rows and millions of columns  
c) It is stored in a single server  
d) It is easy to analyze manually  
**Answer: b) It can contain billions of rows and millions of columns**

11. Which of the following is a key feature of the complexity of Big Data?

a) Data only comes in a single format  
b) Data is always structured  
c) Big Data includes a variety of new data sources, formats, and structures  
d) Data is stored without any analysis  
**Answer: c) Big Data includes a variety of new data sources, formats, and structures**

12. How fast is Big Data generated?

a) It is generated slowly and over a long period of time  
b) It is generated at a rapid pace with high velocity and near real-time analysis  
c) It is only created in batches  
d) It is only generated once a year  
**Answer: b) It is generated at a rapid pace with high velocity and near real-time analysis**

13. What is a challenge when dealing with Big Data?

a) Only structured data can be used  
b) Storing it on a traditional hard drive  
c) Analyzing vast amounts of unstructured data to extract meaningful patterns  
d) Reducing the amount of data generated  
**Answer: c) Analyzing vast amounts of unstructured data to extract meaningful patterns**

14. Which of the following describes the speed aspect of Big Data?

a) Data is processed slowly and in batches  
b) It involves near real-time analysis with rapid data ingestion  
c) Data is only analyzed at the end of each month  
d) Data is stored without being analyzed for a long time  
**Answer: b) It involves near real-time analysis with rapid data ingestion**

15. Why is Big Data considered a valuable asset for organizations?

a) Because it is easy to process and interpret  
b) It creates new business opportunities and allows for better decision-making  
c) It replaces all other types of data storage systems  
d) It eliminates the need for human employees  
**Answer: b) It creates new business opportunities and allows for better decision-making**

1. What is required for data to be loaded into a data warehouse?

a) Data needs to be unstructured and dynamic  
b) Data needs to be well understood, structured, and normalized with appropriate data type definitions  
c) Data must only be in text format  
d) Data should be collected from a single source  
**Answer: b) Data needs to be well understood, structured, and normalized with appropriate data type definitions**

2. What are departmental warehouses and local data marts?

a) Systems used for external data storage  
b) Systems created by business users to accommodate their need for flexible analysis  
c) Tools used for managing security and encryption of data  
d) Tools for reducing data redundancy  
**Answer: b) Systems created by business users to accommodate their need for flexible analysis**

3. Once data is in the data warehouse, how is it typically used?

a) It is stored but not processed  
b) It is read by additional applications across the enterprise for BI (Business Intelligence) and reporting purposes  
c) It is ignored until further notice  
d) It is only used by the IT department  
**Answer: b) It is read by additional applications across the enterprise for BI and reporting purposes**

4. What happens at the end of the workflow involving data in the data warehouse?

a) Data is discarded  
b) Data is archived for future use  
c) Analysts get data provisioned for their downstream analytics  
d) Data is converted into a new format for storage  
**Answer: c) Analysts get data provisioned for their downstream analytics**

5. Which of the following is a driver of Big Data?

a) Printed newspapers  
b) Mobile devices providing geospatial location data and metadata about usage  
c) Small-scale data entry forms  
d) Simple spreadsheet files  
**Answer: b) Mobile devices providing geospatial location data and metadata about usage**

6. What type of information does medical data, such as genomic sequencing and diagnostic imaging, represent in the context of Big Data?

a) Traditional data used for administrative purposes  
b) Non-traditional sources contributing to Big Data  
c) Data that does not require storage  
d) Data used only in government agencies  
**Answer: b) Non-traditional sources contributing to Big Data**

7. How do video surveillance systems contribute to Big Data?

a) By storing only video footage  
b) By providing thousands of video cameras across a city, generating large amounts of data  
c) By limiting the number of video cameras used  
d) By compressing data into small files  
**Answer: b) By providing thousands of video cameras across a city, generating large amounts of data**

8. How do smart devices contribute to Big Data?

a) By collecting data in a structured format only  
b) By providing sensor-based data from smart grids, buildings, and other infrastructures  
c) By reducing the need for additional data collection  
d) By limiting the variety of data collected  
**Answer: b) By providing sensor-based data from smart grids, buildings, and other infrastructures**

9. What are nontraditional IT devices, such as RFID readers and GPS systems, used for in Big Data?

a) They are used for traditional data processing  
b) They are used to collect and process real-time data from various sources  
c) They are used for storage of unprocessed data  
d) They are used exclusively for data visualization  
**Answer: b) They are used to collect and process real-time data from various sources**

10. What kind of data is collected by RFID readers, GPS navigation systems, and seismic processing in Big Data applications?

a) Real-time sensor-based data  
b) Data collected only from online sources  
c) Historical data from traditional databases  
d) Data limited to textual information  
**Answer: a) Real-time sensor-based data**

1. What is the first requirement for Big Data technologies?

a) High-speed internet  
b) Cheap and abundant storage  
c) High processing speed  
d) Specific data formats  
**Answer: b) Cheap and abundant storage**

2. What is needed to help with the quicker processing of Big Data?

a) Increased storage capacity  
b) Faster processors  
c) More security features  
d) Simplified data formats  
**Answer: b) Faster processors**

3. Which of the following is an example of an affordable, open-source, distributed Big Data platform?

a) Oracle  
b) Hadoop  
c) Microsoft Excel  
d) PostgreSQL  
**Answer: b) Hadoop**

4. What type of processing is required to handle Big Data effectively?

a) Sequential processing  
b) Parallel processing  
c) Single-threaded processing  
d) Manual processing  
**Answer: b) Parallel processing**

5. Which of the following Big Data technologies focuses on high connectivity and high throughput?

a) Cloud computing  
b) Low-latency processing  
c) Clustering virtualization  
d) Low bandwidth networks  
**Answer: c) Clustering virtualization**

6. What is the role of cloud computing in Big Data technologies?

a) It helps in high-speed data processing  
b) It provides flexible resource allocation arrangements  
c) It stores data locally in physical servers  
d) It simplifies the use of traditional databases  
**Answer: b) It provides flexible resource allocation arrangements**

7. Which of the following is NOT a characteristic required for Big Data technologies?

a) High connectivity  
b) High throughput  
c) Low latency  
d) Expensive hardware  
**Answer: d) Expensive hardware**

8. What is emphasized for Big Data processing instead of low latency?

a) Low-cost infrastructure  
b) High connectivity and high throughput  
c) Real-time data analysis  
d) Data compression techniques  
**Answer: b) High connectivity and high throughput**

9. What is the benefit of using parallel processing for Big Data?

a) It helps to reduce the need for multiple data sources  
b) It accelerates data processing by handling multiple tasks simultaneously  
c) It stores data more efficiently  
d) It simplifies the data visualization process  
**Answer: b) It accelerates data processing by handling multiple tasks simultaneously**

**What is the primary focus of Unit II syllabus mentioned in the document?**

* A) Data Storage Techniques
* B) Building Data Products at Scale with Hadoop
* C) Machine Learning Algorithms
* D) Data Visualization Methods
* **Answer:** B) Building Data Products at Scale with Hadoop

**Which chapter covers the topic of MapReduce with Python?**

* A) Chapter 1
* B) Chapter 2
* C) Chapter 3
* D) Chapter 4
* **Answer:** C) Chapter 3

**Which of the following is a topic included in the syllabus references?**

* A) Introduction to Java
* B) Data Analytics with Hadoop
* C) SQL for Data Science
* D) Cloud Computing Fundamentals
* **Answer:** B) Data Analytics with Hadoop

**In which section is 'Interactive Spark with PySpark' discussed?**

* A) Chapter 1
* B) Chapter 3
* C) Chapter 2
* D) Chapter 4
* **Answer:** D) Chapter 4

**What concept is NOT covered in Chapter 2 according to the syllabus?**

* A) Concepts of Operating System for Big Data
* B) Hadoop Architecture
* C) Advanced Machine Learning Techniques
* D) Working with Distributed File System
* **Answer:** C) Advanced Machine Learning Techniques

**Who are the authors of "Data Analytics with Hadoop"?**

* A) V.K Jain and Khanna
* B) Benjamin Bengfort and Jenny Kim
* C) Subhashini Chellappan and Seema Acharya
* D) Prof Bharati Bhole
* **Answer:** B) Benjamin Bengfort and Jenny Kim

**Which of the following frameworks is discussed in the document for data processing?**

* A) TensorFlow
* B) Hadoop Streaming
* C) Apache Kafka
* D) Scikit-learn
* **Answer:** B) Hadoop Streaming

**What is the primary function of Hadoop MapReduce?**

* A) Data storage
* B) Data processing
* C) Data visualization
* D) Data security  
   **Answer:** B) Data processing .

**Which of the following best describes the role of the Mapper in MapReduce?**

* A) Combine data
* B) Filter data
* C) Process input data and generate key-value pairs
* D) Store data  
   **Answer:** C) Process input data and generate key-value pairs .

**In the Word Count example presented in the document, what is the output of the Mapper?**

* A) The total number of words
* B) Key-value pairs where the key is the word and the value is the count (1)
* C) A sorted list of words
* D) The number of unique words  
   **Answer:** B) Key-value pairs where the key is the word and the value is the count (1) .

**Which component of Hadoop is responsible for the distributed storage of data?**

* A) YARN
* B) MapReduce
* C) HDFS
* D) Hive  
   **Answer:** C) HDFS .

**What does HDFS stand for?**

* A) High Data File System
* B) Hadoop Distributed File System
* C) Hadoop Dynamic File System
* D) High Definition File System  
   **Answer:** B) Hadoop Distributed File System .

**Which programming language is commonly used for writing MapReduce jobs?**

* A) Java
* B) Python
* C) Ruby
* D) C++  
   **Answer:** A) Java .

**What is the output of the Reducer in MapReduce?**

* A) Input data
* B) Final processed key-value pairs
* C) Error logs
* D) Initial dataset  
   **Answer:** B) Final processed key-value pairs .