MODULE 4

Dealing with Human-Generated Data/Concept of Metadata: -

- Intentional data, photos, videos, record audio ,text on a social network,Like on Facebook. Web searches, Bookmarked web searches, Emails, Phone calls.
- All of these are kinds of data that do not exist until the person makes them happen, so these are records of human actions.
- This type of data requires a level of processing beyond the capabilities of relational database systems because the goal is to interpret the meaning and create human readable data.
- Hence the concept of Metadata came in to picture that is data about data.

Big Data: -

Big data is data that contains greater variety, arriving in increasing volumes and with more velocity.

Five V's of Big Data: -

Volume	Velocity	Variety	Veracity	Value
Scale of data	Speed of data	Diversity of data	Accuracy gained from data	Insights gained from data.

Big Data Physical Infrastructure	Big Data Security Infrastructure
The physical infrastructure for big data refers to the hardware and software components required to store, process, and analyze large volumes of data.	The security infrastructure for big data refers to the measures put in place to protect the data from unauthorized access, theft, or manipulation
This can include servers, storage systems, networking equipment, and data centers.	This can include firewalls, encryption, access controls, authentication mechanisms, and monitoring tools.
The goal of the physical infrastructure is to ensure that data is available, accessible, and reliable.	The goal of the security infrastructure is to ensure that the data remains confidential, integral, and available only to authorized users.
Physical infrastructure ensures that data is accessible and can be processed efficiently	Security infrastructure ensures that the data is safe from unauthorized access or manipulation.

Operations on Big Data Sources: -

- Incorporate all the data sources that will give a complete picture of the business and see how the data impacts the business.
- As the world changes, it is important to understand that operational data now has to encompass a broader set of data sources, including unstructured sources such as social media data in all its forms.
- It approaches to data management in the big data world, including document,graph, columnar, and geospatial database architectures.
- These are referred to as NoSQL, or not only SQL, databases.

Role of Structured and Unstructured Data: -

Structured Data	Unstructured Data
Structured data is data that is organized and stored in a specific format, such as in rows and columns in a database	Unstructured data, on the other hand, is data that is not organized in a specific format and does not have a pre-defined schema.
Structured data is easy to query, analyze, and use to make data-driven decisions.	Unstructured data is more difficult to analyze and use for decision-making than structured data.
Structured data is commonly used in business intelligence, financial analysis, and performance management.	Unstructured data provides valuable insights into customer sentiment, market trends, and emerging issues.
The role of structured data is to provide a clear picture of a specific aspect of a business or organization	The role of unstructured data is to provide a more holistic view of a business or organization

By combining structured and unstructured data, decision-makers can gain a more comprehensive view of their business or organization, leading to more informed and effective decision-making.

These unstructured data are typically used with non relational databases such as NoSQL databases and include the following structures:

Key-Value Pair Database	Document Database	Columnar Database	Graph Database	Spatial Database
KVP are used in lookup tables, hash tables, and configuration files	provide a technique for managing repositories of unstructured and semi-structured data such as text documents, web pages	efficient database structure that stores data in columns rather than rows	make use of graph structures with nodes and edges to manage and represent data	They are optimized to store and query geometric objects that can include points, lines, and polygons.

Data Services and Tools: -
Data services and tools refer to software and applications that are designed to help users manage, analyze, and extract insights from data.
There are various data services and tools available, and their use depends on the specific needs and requirements of the user. Some common data services and tools include:

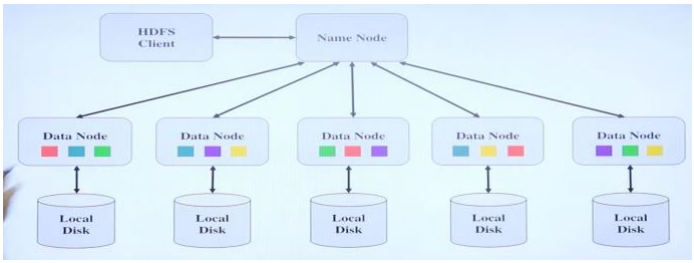
1. Business Intelligence
2. Data Warehousing Tools
3. Data Integration Tools
4. Data Analytics Tools
5. Data Visualization Tools
6. Data Governance Tools

Analytical Data Warehousing: -
Algorithms that support Big Data Analytics:

Sketching and Streaming	Dimensionality Reduction	Numerical Linear Algebra	Compressed Sensing	
Algorithm is used when streaming data from sensors	algorithms help to convert data that is highly dimensional into much simpler data	These algorithms are used when data includes large matrices	These algorithms are useful when the data is sparse or signal data from a streaming sensor are limited to a few linear or time-based measurements.	

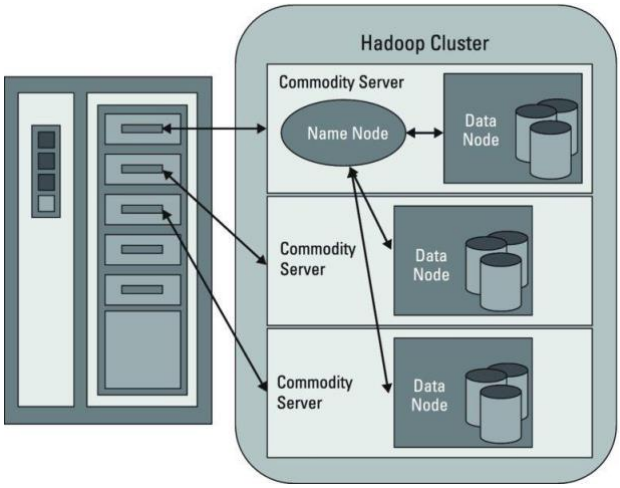
Hadoop: -

- Hadoop has emerged as one of the most important technologies for managing large amounts of unstructured data in an efficient manner because it uses distributed computing techniques.
- Hadoop enables you to use parallelization techniques to improve efficiency.
- It is an open-source community, codebase, and market for a big data environment that is designed, among other things, to parallel-execute code written to MapReduce.
- Text documents, ontologies, social media data, sensor data, and other forms of nontraditional data types can be efficiently managed in Hadoop.



Two key components to Hadoop are described here:
Hadoop Distributed File System (HDFS): A data storage cluster that is both highly reliable and low cost used to make it easy to manage related files across different machines.
MapReduce engine: Provides a way to distribute the processing of the algorithms across a large number of systems.

HADOOP CLUSTER: -



NameNodes	DataNodes
A Namenode is the master node in Hadoop that manages the metadata of the file system.	A Datanode is a slave node in Hadoop that stores the actual data in the HDFS.
The role of the NameNode is to keep track of where data is physically stored in the cluster	Of the two components, NameNodes have some intelligence, whereas Data Nodes are more simplistic.
To maintain this knowledge, the NameNode needs to understand which blocks on which data nodes make up the complete file.	They store and retrieve the data blocks in the local filesystem of the server.
The NameNode manages all access to the files, including reads, writes, creates, deletes, and replication of data blocks on the data nodes.	They also store the metadata of a block in the filesystem.
In addition, NameNode has the important responsibility of telling the Data Nodes if there is anything for them to do.	In addition, Data Nodes send reports to the NameNode about what blocks are available for file operations.

Motion Data	Streaming Data
Data in motion refers to the data that is actively moving or being transferred from one place to another.	Streaming data is a specific type of data in motion that refers to a continuous flow of data that is generated in real-time.
Data that is being transmitted over a network, sent between applications, or processed in real-time	Streaming data is typically unbounded and infinite in nature
Examples of data in motion include network traffic, sensor data from IoT devices, and data	Examples of streaming data include stock market data feeds, weather sensor data, and social

streams from social media platforms.	media updates.

Integration of Big Data with Traditional Data: -

- The integration of big data with traditional data involves combining the two types of data to gain insights and make better-informed decisions.
- Traditional data typically refers to structured data that can be easily processed using traditional database management systems, while big data includes both structured and unstructured data, such as social media feeds, sensor data, and log files.
- The integration of big data with traditional data can provide a more complete picture of a business or organization's operations and help identify patterns and trends that might not be visible with traditional data analysis methods.
- This integration can be achieved through various approaches, such as:

1. Data Warehousing
2. Data Lakes
3. Advanced Analytics
4. API's