SNS COLLEGE OF TECHNOLOGY
COIMBATORE-35

DEPARTMENT OF INFORMATION TECHNOLOGY

19ITE305 – BIG DATA ANALYTICS

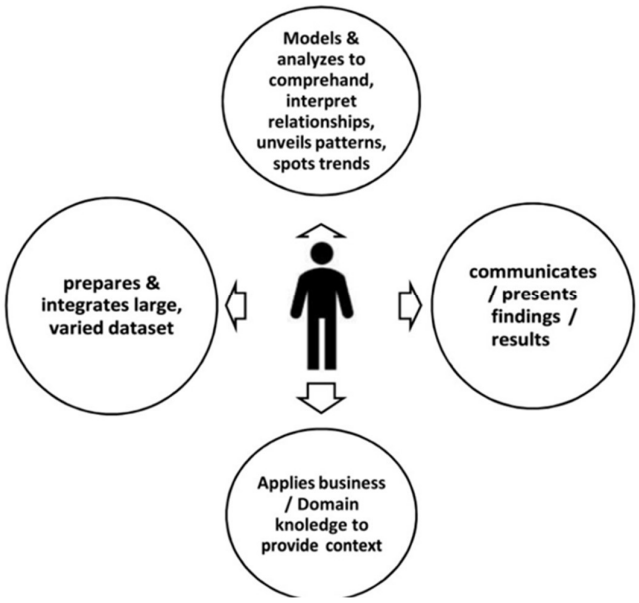
UNIT I: INTRODUCTION TO BIG DATA AND ANALYTICS
Topic 7: Terminologies used in Big Data Environments

Data Management: A data scientist employs several approaches to develop the relevant datasets for analysis. Raw data is just "RAW", unsuitable for analysis. The data scientist works on it to prepare to reflect the relationships and contexts. This data then becomes useful for processing and further analysis.

Analytical Techniques: Depending on the business questions which we are trying to find answers to and the type of data available at hand, the data scientist employs a blend of analytical techniques to develop models and algorithms to understand the data, interpret relationships, spot trends, and reveal patterns.

Business Analysis: A data scientist is a business analyst who distinguishes cool facts from insights and is able to apply his business expertise and domain knowledge to see the results in the business context.

Communicator: He is a good presenter and communicator who is able to communicate the results of his findings in a language that is understood by the different business stakeholders.



A big data environment refers to the infrastructure, tools, and technologies that are used to collect, store, process, and analyze large and complex data sets. It typically includes the following components:

1. **Data storage**: Large and complex data sets need to be stored in a way that allows for efficient processing and analysis. This can include data lakes, Hadoop Distributed File System (HDFS), and NoSQL databases.
2. **Data processing:** Data needs to be processed in a way that allows for efficient analysis and insights. This can include batch processing using Hadoop MapReduce, real-time processing using stream processing technologies, and machine learning and artificial intelligence (AI) algorithms.
3. **Data analysis**: Data needs to be analyzed in a way that allows for insights and decision-making. This can include SQL-based analysis, data visualization tools, and machine learning and AI algorithms.
4. **Data governance:** To ensure data is accurate, complete, consistent and compliant with regulations, organizations put in place Data Governance policies and procedures.
5. **Data security:** To protect sensitive information and prevent unauthorized access, organizations need to implement security measures such as data encryption, access controls, and incident response.
6. **Cloud Computing**: Many organizations are using cloud computing platforms, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP), to handle the scale and complexity of big data.
7. **Data ingestion and integration:** Data may be collected from a variety of different sources, and it may be necessary to clean, transform, and integrate this data before it can be analyzed.
8. **Data visualization:** Data visualization tools such as Tableau, QlikView and Power BI are used to transform large data sets into interactive and easy-to-understand visualizations.
9. **Data Governance:** To ensure data is accurate, complete, consistent and compliant with regulations, organizations put in place Data Governance policies and procedures.

- **Data quality:** To ensure data is fit for the intended purpose and meets the needs of the consumer, data quality assessments and management are performed.

Here are a few key terminologies used in big data environments:

1. **Hadoop:** An open-source software framework that enables the storage and processing of large and complex data sets on a distributed computing infrastructure.
2. **MapReduce:** A programming model that is used to process large data sets in parallel across a distributed computing infrastructure.
3. **NoSQL:** A type of database that is optimized for handling large and complex data sets, including unstructured and semi-structured data.
4. **Distributed computing:** A type of computing in which a task is divided among multiple computers that work together to complete the task.
5. **Cluster computing:** A type of distributed computing in which multiple computers work together as a single system.
6. **Cloud computing:** A type of computing in which resources, such as storage and processing power, are provided on-demand over the internet.
7. **Data lake:** A centralized repository that allows an organization to store all of their structured and unstructured data at any scale.

8. **Stream processing:** A type of data processing that allows for real-time analysis of data streams as they are generated.
9. **Machine learning:** A type of artificial intelligence that enables computers to learn and improve from data.

- **NoSQL database:** Non-Relational databases which are used to store unstructured data and handle high-velocity, high-variety and high-volume data.
- **Data pipeline:** A series of steps that data goes through from ingestion to storage, processing, and analysis.
- **Data Governance:** The management and control of data as it is collected, stored, and used. It includes data quality, data security, data privacy, data lineage and more.
- **Data Quality:** The degree to which data meets the needs of the consumer and is fit for the intended purpose.
- **Data Governance policies:** Rules, regulations and guidelines that organizations put in place to ensure that their data is accurate, complete, consistent and compliant with the regulations.