

Date: 5/12/25 IV B.Tech I sem Regular R-22 December-2025

Subject: Big Data Analytics (AI732PE)

Scheme of Evaluation

part-A

1) a) what is data architecture in big data systems? (2M)

Big Data Architecture is a framework and it is designed to handle the ingestion, processing, and analysis of data that is too large (or) complex for traditional database systems.

b) List any two data pre-processing techniques.

- 1) Data cleaning
- 2) Data integration
- 3) Data transformation
- 4) Data Reduction.

} (1M)

c) Define HDFS and its purpose (1M)

HDFS $\Rightarrow$ Hadoop distributed file system.

HDFS stores large volume of data (or) datasets by breaking them into blocks and then distribute to multiple machines/blocks. It can provide scalability, reliable and cost effective.

d) mention any two advantages using spark over HDFS.

- 1) faster processing speed
- 2) ease of use
- 3) In memory operation.
- 4) Real time data Analysis.

} (1M)

e) What is data Analytics life cycle? (1M)

data Analytics life cycle is a systematic process. It will help to Analyze the data from start to finish. like discovery of data

Data Preparation, model planning, model building, communication, and operationalization.

f) differentiate b/w descriptive Analysis & Predictive analysis.

descriptive data	Predictive data
1) using past data it can do Analysis 2) <u>Ex</u> : - sales report	1) It will ^{do} forecasting 2) <u>Ex</u> : Product recommendation

(1m)

g) define training data & testing data

Training data: In general the datasets are used to train a ML model.

the model learns the pattern; relationships and different features in the data set.

Test data: The test data is used to evaluate the performance of the trained ML/designed model. (1m)

h) what is the model evaluation and why it is important.

The model evaluation is used for observe the specific performance metrics. How the ML model performs on a new unseen data (or) information. (1m)

i) what are the advantages of visual Analytics

- 1) understand the complex data
- 2) faster
- 3) improve decision making
- 4) ability to identify uncover hidden patterns

(1m)

j) list any two charts used for visualization of categorical data

(1m)

- 1) bar charts, 2) pie charts, 3) line charts, scatter plots, Histograms

Part-B

2) a) Explain the concepts of data governance and data quality management.

→ Data Governance is an enterprise wide framework of policies, process standards, how data is managed and protected. (5M)

→ Data handling; security; privacy and accessibility.

→ Assigning the ownership and accountability for that data.

→ Data Quality is the ability of your data to serve its intended purpose based on factors such as accuracy, completeness, consistency, reliability and these factors that play a huge role in determining data quality.

2) b) Discuss the methods of handling duplicate and missing values in datasets.

Missing data in the training data set can reduce the power fit of a model or can lead to a biased model because we have not analysed the behaviour and relationship with other variables correctly. It can lead to wrong prediction or (5M) classification.

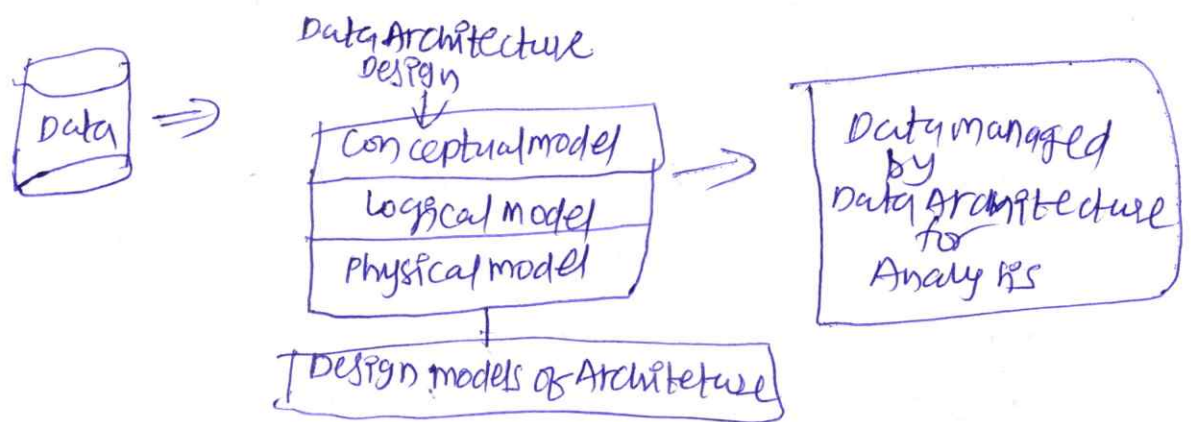
→ In statistics missing data (or) missing values occurs when no data value is stored for the variable in an observation.

Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

2) Duplicate data values: A data set may include data for entities which are duplicates of one another. It may happen when say the same person submits a form more than once.
→ It is a specialized data compression technique for eliminating duplicate copies of repeating data.

3) Discuss the role of ~~big~~ data architecture in big data analytics and list the main phase of data management cycle. (10m)

A) Data architecture design is set of standards which are composed of certain policies, rules, models and standards which manages, what type of data is collected, from where it is collected, the arrangement of collected data, storing data, utilizing data and securing data.



→ Data architecture design is important for creating a vision of interactions occurring b/w data systems.

1) Conceptual model:-

2) Logical model.

3) Physical model.

The main phase of the data management cycle is.

Data Acquisition / Ingestion:- collecting raw data from various sources into a central system.

- Data storage:- Storing data in suitable systems; optimized for different event processing needs.

Data processing:- transforming raw data into usable formats, involving cleaning, filtering and enriching.

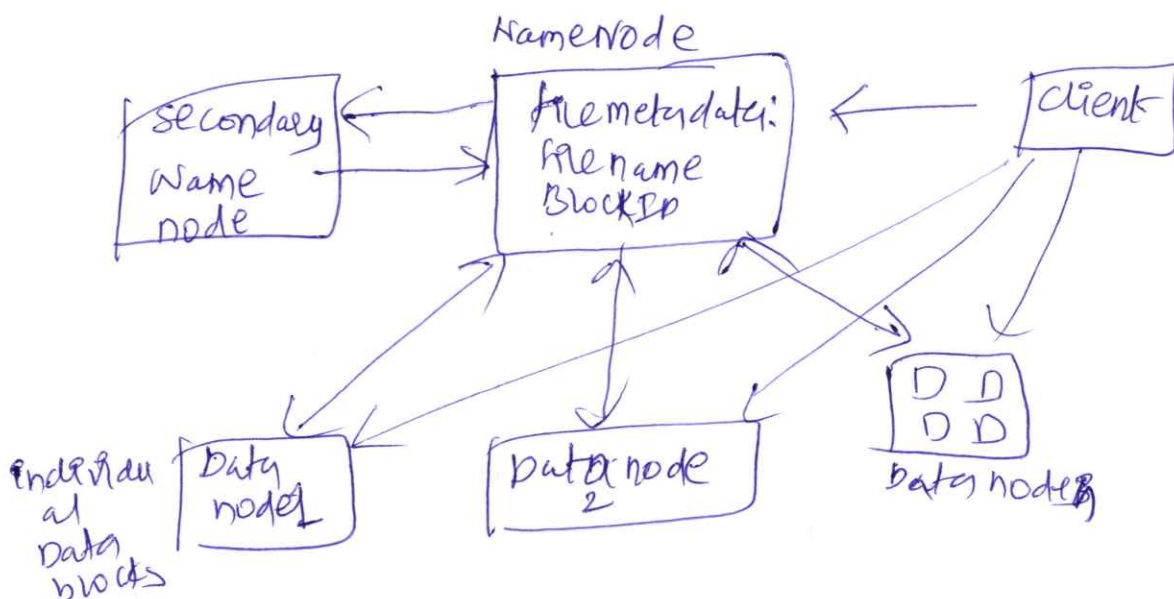
Data Analysis:- Applying tools and techniques to extract insights, discover patterns and generate reports.

Data Governance & Security:- managing access, ensuring compliance and maintaining data quality throughout life cycle.

Data Archiving & Deletion:- Moving aged or less used data to cheaper storage or deleting it securely.

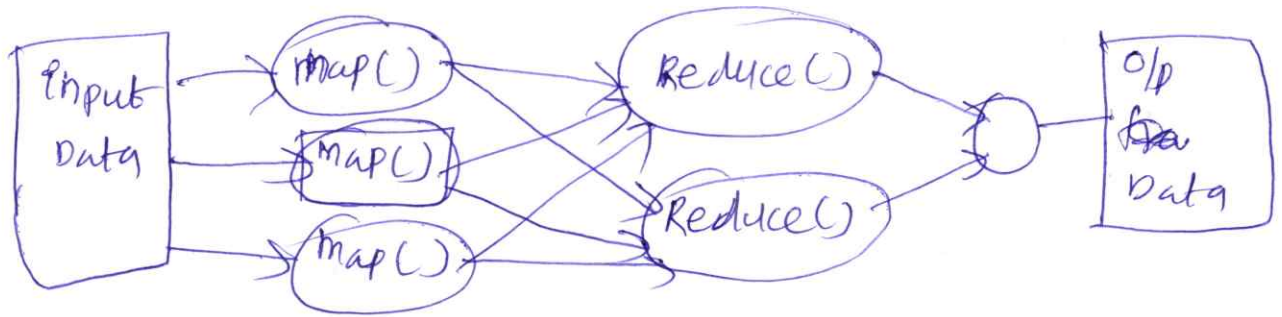
4) Explain the working of Hadoop and mapreduce framework with neat diagram. (5M)

→ Hadoop framework is there are 2 main layers
1) HDFS 2) Execution Engine



Hadoop architecture performance depends upon hard-drive through put and the network speed for the data transfer.

MapReduce is a software framework for easily writing application which process big amount of data in parallel on large clusters.



4) b) describe how data is transferred b/w nodes in spark.

Data transfer between nodes in Apache Spark, especially during operations like shuffles relies on several key mechanisms

1) serialization:- Before data can be sent across (5m)

the network, it must be converted from its in memory object representation into a byte stream.

a) Java Serializer b) Kryo Serializer.

2) network communication (RPC):- Spark executors communicate using Remote Procedure calls (RPC).

3) shuffle services:- for large scale data transfer during shuffle operations, spark utilizes a shuffle services.

4) Block manager:- each executor has a block manager responsible for storing and retrieving data blocks.

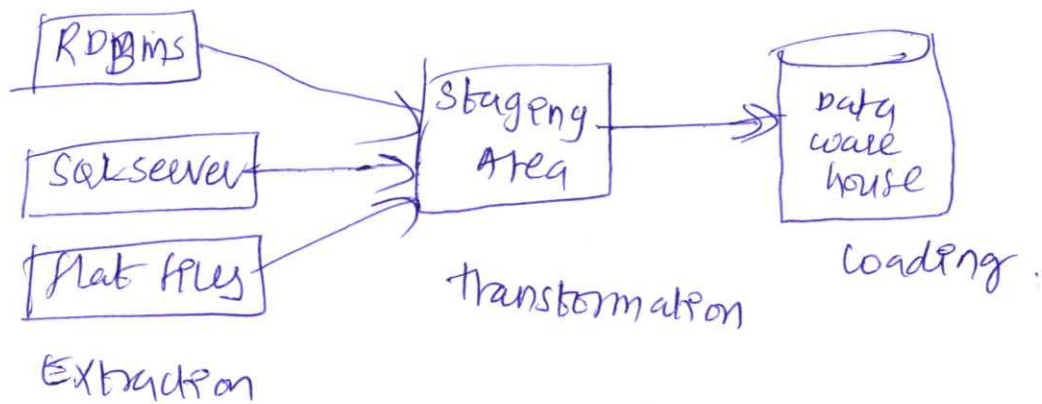
5) Discuss ETL process stages, Extraction, and transformation and loading with suitable examples. (10m)

Extract:- data is pulled from multiple, disparate sources, which can include databases (SQL, NoSQL) Flat files, CRM systems (like salesforce); ERP systems and even social media feeds

Transform:- This is where data quality and consistency are enforced.

a) cleaning b) standardization c) Aggregation/summarization d) derivation e) joining

Load:- The processed clean data is loaded into a target system.



EX:- E-commerce Retailer.

An online retailer wants to analyze customer purchasing behavior across their physical stores and websites.

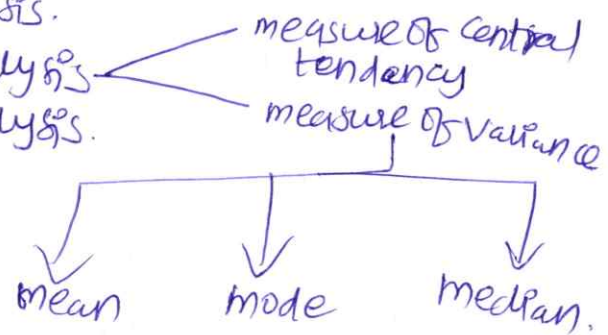
6) a) Explain the needs for statistical analysis in big data.

Statistical Analysis is a fundamental aspect of data. (5M)
Science that helps in enabling us to extract meaningful insights from complex datasets.

Majorly two basic statistical analysis.

1) Descriptive Statistical Analysis

2) Predictive Statistical Analysis.



mean:- $\bar{x} = \frac{\sum x}{n}$

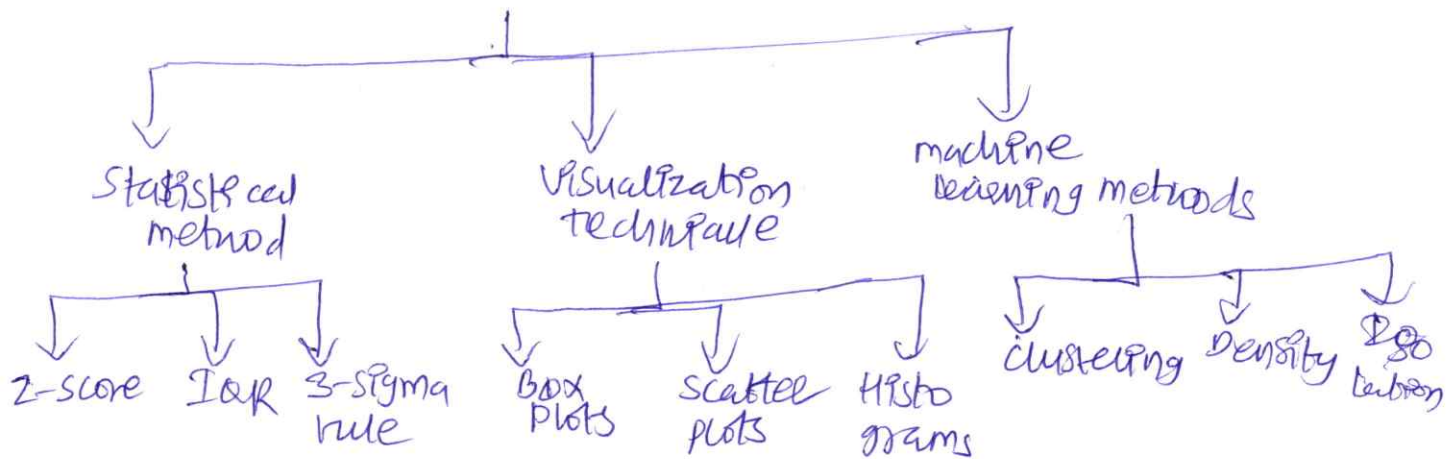
x = observation n = no. of terms

mode:- It is the value that has the highest frequency in the given data set.

median:- It is the middle value of the data set.

6) b) discuss the techniques for detecting and outliers elimination.

(5m)
detecting and eliminating outliers involve statistical (z-score, IQR) visualization (box plots, scatter plots) and ML methods (Isolation forest, LOF) to find extreme data points.



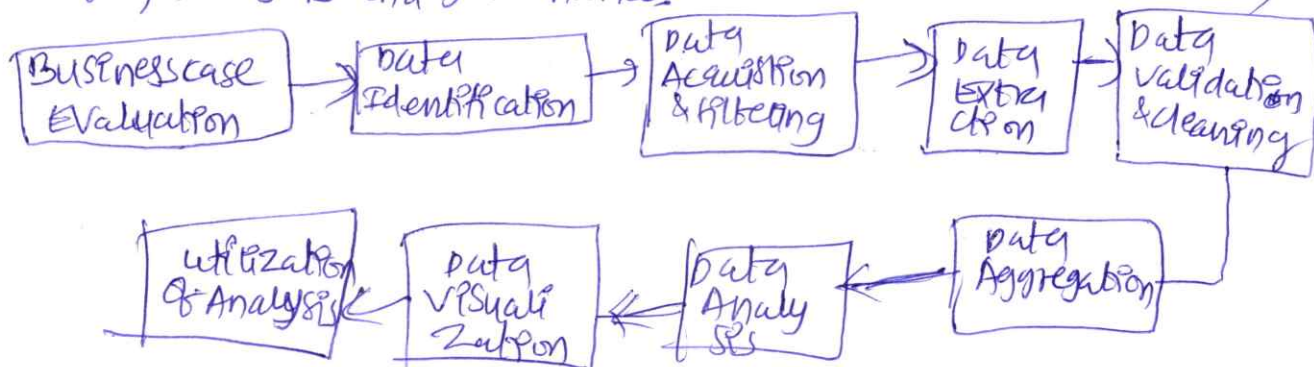
techniques to data elimination:-

- 1) Removal (deletion): delete outlier records, best for errors and small data sets.
- 2) Transformation: - Apply math functions to reduce outlier impact and normalized data.
- 3) Imputation: - Replace outlier with mean/median.

7) Illustrate the big data life cycle and explain each phase.

(10m)

big data Analytics is a field that involves managing the entire data life cycle, including data collection, cleaning, organisation, storage, Analysis and governance.



→ data identification phase focuses on identifying the necessary data sets and their sources for the analysis project.

→ data acquisition process entails gathering data from all of the sources mentioned in the previous phase.

→ data converting it into a format that the underlying big data

→ Data Aggregation and Representation phase focuses on combining multiple data sets to create a cohesive view.

8) a) Explain steps involved in building a Predictive model for datasets. (5m)

Predictive analytics is a distinct field within data analysis that utilizes various methods, statistics, machine learning algorithms and Artificial neural networks. to Predict potential future outcomes based on historical data.

Step 1:- Define the project goals.

Step 2:- collect the data in a single dataset

Step 3:- Prepare and Process data

Step 4:- select a suitable Predictive modelling technique

Step 5:- Build and train a Predictive data analytics model

Step 6:- Deploy and monitor the model.

8) b) compare supervised and unsupervised algorithms

supervised	unsupervised (5m)
Labeled data (I/P & O/P pairs)	unlabeled data (I/P only)
Predict outcomes, classify data	discover patterns
Requires human labeling for training data	Requires minimax or no human labeling

Classification, Regression

Generally less computationally complex

high accuracy

clustering, dimensionality reduction.

more computationally complex

Accuracy can be more difficult to quantify

9) Discuss cross validation and over-fitting issues in machine learning models. (10M)

cross validation is a technique used to check how well a machine learning model performs on unseen data while preventing overfitting.

- Splitting the data sets into several parts
- Training the model on some parts and testing it on the remaining part
- Repeating this resampling process multiple times by choosing different parts of the datasets.
- Averaging the results from each validation step to get the final performance.

overfitting:- overfitting occurs when a machine learning model learns the training data too well, including its noise and irrelevant details, rather than generalizing the underlying patterns.

This leads to high performance on the training data but poor performance on unseen new data.

10) a) Explain how visualization enhance data driven decision making (5M)

Data visualization enhances decision making by transforming complex numbers into easy to understand

charts and graphs.

→ It improves communication across teams & leads to faster.

→ allowing quick spotting of patterns and trends.

- 1) Identifying patterns & trends
- 2) communicating complex information
- 3) spotting outliers
- 4) comparing data
- 5) continuous improvement

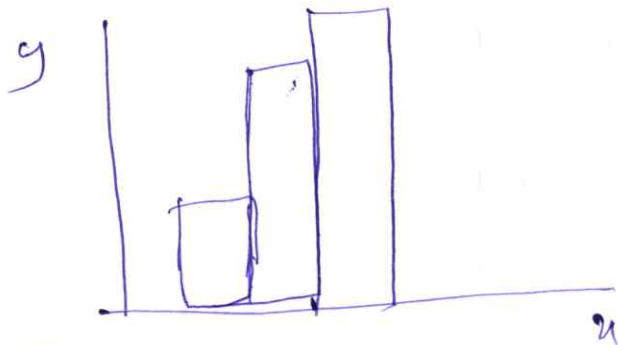
10) b) demonstrate the process of designing interactive dashboards using Qlik View or Tableau.

(5m)

→ Tableau is a data visualization tool that can be used by data analysts, scientists, statisticians etc.

→ To visualize the data and get a clear opinion based on the data analysis.

→ Tableau helps users understand data by presenting it in a visual format, such as charts, graphs and maps.



Qlik View:- In order to get started really quickly with your visualization, Tableau has a feature called "show me". It is located in the top right of the screen, and it shows the different types of visual Tableau offers.

11) describe various charts types and their suitability for different data scenarios in visualization (10m)

A primary goal of data visualization is to communicate information clearly and efficiently to users via the statistical graphics, plots, information graphics, tables and charts.

Types of Data Visualization:-

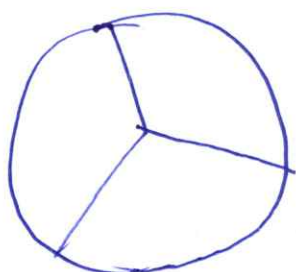
Bar charts:- ideal for comparing categorical data or displaying frequencies, bar charts offer a clear visual representation of values.

Line charts:- perfect for illustrating trends over time. Line charts connect data points to reveal patterns and fluctuations.

Pie charts:- efficient for displaying parts of a whole. Pie charts offer a simple way to understand proportions and percentages.

Scatter plots:- showcase relationship b/w two variables, identifying patterns and outliers through scattered data points.

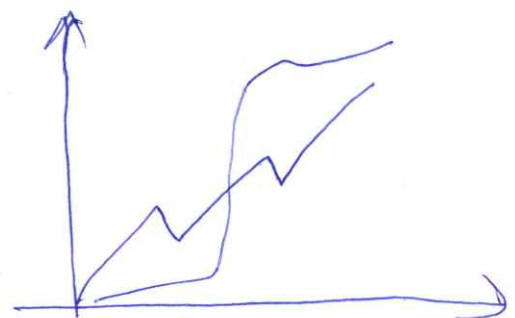
Histograms:- depicts the distribution of a continuous variable, providing insights into the underlying data patterns.



Pie charts



Bar charts



Line charts