Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

AY: 2024-25

Class:	BE	**Semester:**	VII
Course Code:		**Course Name:**	Big Data Analytics

Name of Student:	Hitesh . A. Moota
Roll No. :	32
Assignment No.:	01
Title of Assignment:	
Date of Submission:	
Date of Correction:	

Evaluation

Performance Indicator	Max. Marks	Marks Obtained
Completeness	5	3
Demonstrated Knowledge	3	2
Legibility	2	1
Total	10	6

Performance Indicator	Exceed Expectations (EE)	Meet Expectations (ME)	Below Expectations (BE)
Completeness	5	3-4	1-2
Demonstrated Knowledge Legibility	3	2	1
Legibility	2	1	0

Checked by

Name of Faculty :

Signature : 21/10/24

Date :

Hitesh A. Mhatre

32
AIDS

BDA

Assignment - 01

Q1) What are 3V's of big data? Give two case studies indicating respective V's with justification

→ The 3V's of big data are volume, velocity & variety. These characteristics help define the challenges & opportunities, associated with managing & analyzing large & complex datasets

1. Volume :- Refers to the sheer amount of data generated & collected

2. Variety :- Pertains to the speed at which data is generated processed & public opinion

Volume Justification :-
Twitter generates an enormous amount of daily data, with hundreds of millions of tweets being posted. The large volume of data requires significant storage & processing capabilities

Velocity Justification :-
Tweets are generated & need to be processed in real time to provide timely insights into trending topics & user sentiment. The speed at which this data is produced & consumed is critical for the platform's Data Integration

CASE STUDY : Heathcare Data Integration (variety)
Scenario : A healthcare organization integrates various types of data from different sources, including patient records, lab results, imaging data & wearables device data, to provide comprehensive patient care.

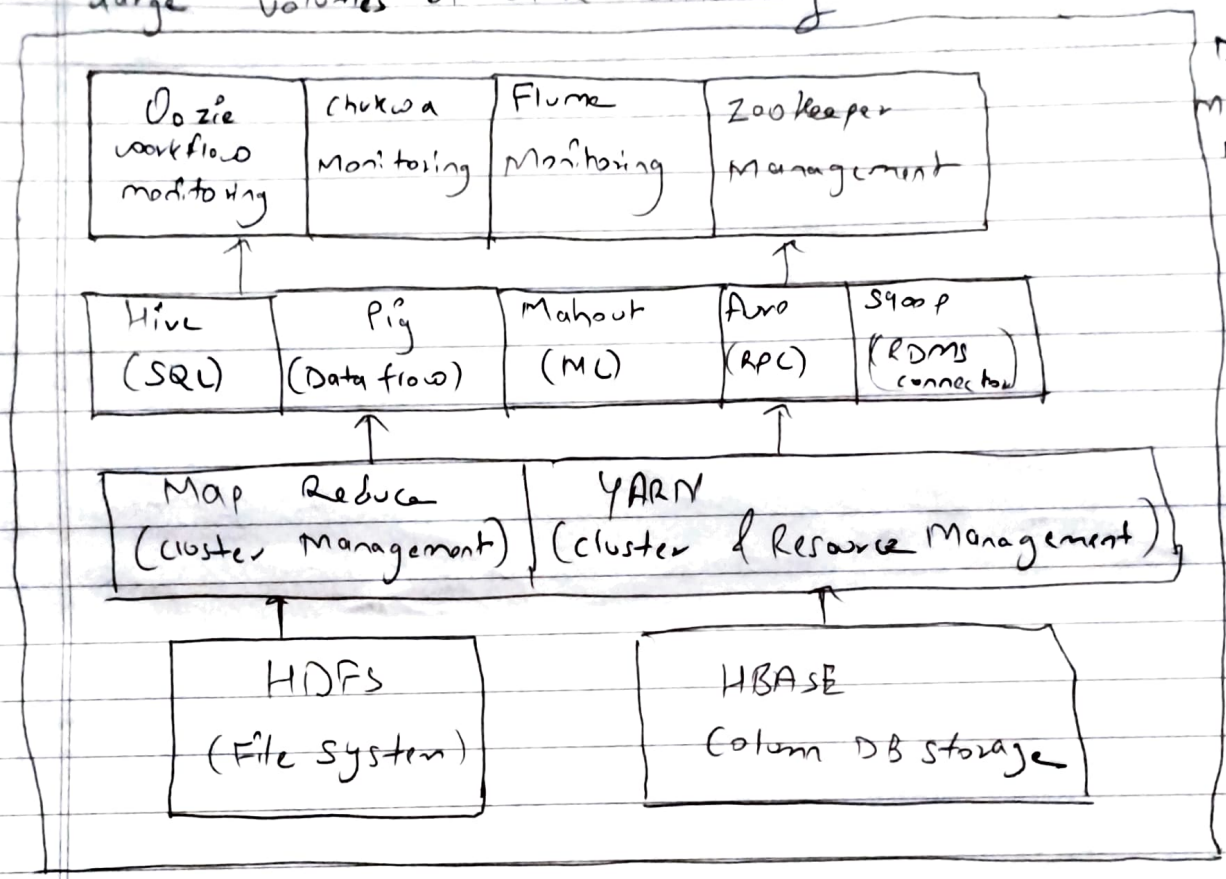
Variety justification:-

The data comes in various forms - structured data from Electronic health records (EHR), unstructured data from doctor's notes, semi structured data from lab results, & real time data from wearable devices. The ability to handle & integrate these diverse data types is essential for providing holistic patient insights & personalized care.

In both case studies, the respective V's highlight the primary challenges & considerations in managing big data effectively.

Q2 Explain Hadoop Ecosystem with its components in detail

→ The Hadoop Ecosystem is a suite of tools & technologies designed to store, process & analyze large volumes of data efficiently



Components :-

1. Hadoop common

- It is a programming model for file System & OS level abstractions & a set of common java libraries

(2) HDFS

- A DFS that stores data across multiple machines to ensure reliability & high availability

(3) YARN

- A resource management layer for Hadoop. It allocates system

resources to the various applications running on the Hadoop cluster.

4. Map reduce :-
It is a programming model for processing large datasets with a distributed algorithm on a Hadoop cluster.

Hadoop Ecosystem tools :-

(1) Pig - It is a high level platform for creating Mapreduce used with Hadoop. It uses a scripting language called pig latin.

(2) Hive - It is a data warehousing solution that provides a SQL-like interface to query data stored in Hadoop. It translates SQL queries into map Reduce jobs.

(3) Sqoop - It is a tool designed for efficiently transferring bulk data between Hadoop & structured data stores such as relational databases.

(4) Flume - A distributed, reliable & available service for efficiently collecting, aggregating & moving large amounts of log data.

(5) Oozie - It is a workflow Scheduler system to manage Hadoop jobs. It allows users to define a sequence of jobs that need to be executed in a specific order.

(6) Zookeeper - It is a centralized service for maintaining configuration information, naming, providing distributed synchronisation & providing group services.