Date __/__/__

Page

Assignment - I

1. Explain classification of big data

-> Big data can be broadly classified into structured, semi-structured and unstructured data.

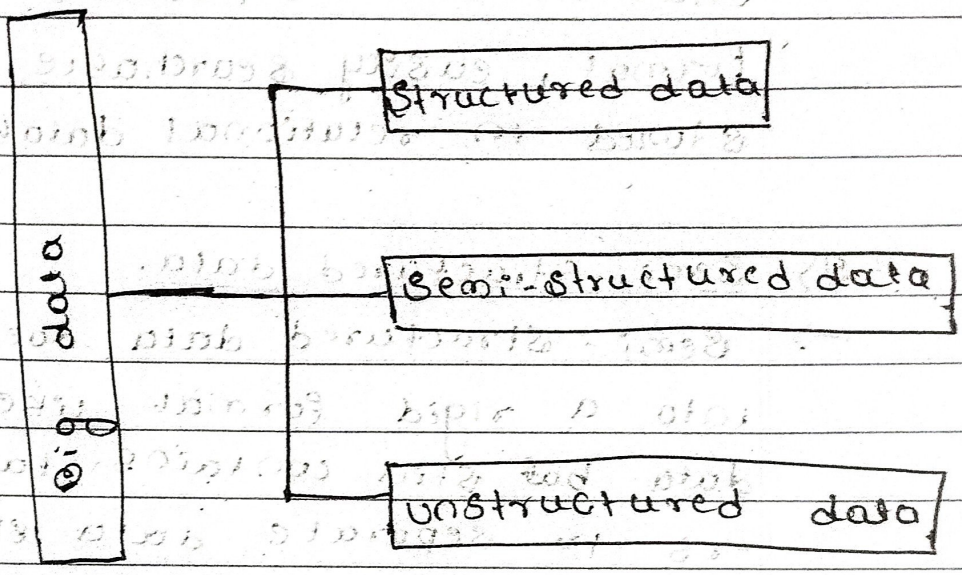


Fig: - Classification of Big data

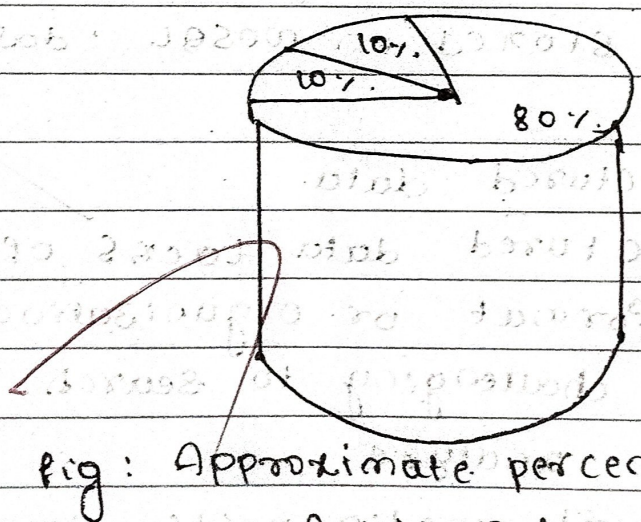


Fig: Approximate percentage distribution of big data

i) Structured data

-> Structured data is highly Organized

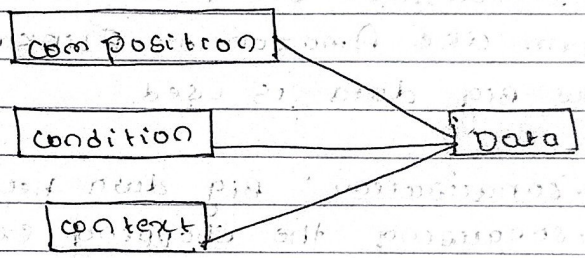
and in a way that makes it easily searchable databases.
Ex:- financial transactions, customer information, product details.
Characteristics: follows a predefined format, easily searchable, and typically stored in relational database.

2) Semi-structured data.
→ Semi-structured data does not fit into a rigid format like structured data but still contains tags or markers to separate data elements
Ex:- JSON data, XML files, CSV files
Characteristics:- more flexible than structured data, contains metadata and can be stored in NOSQL databases.

3) Unstructured data.
Unstructured data lacks of predefined format or organization, making it more challenging to search and analyse.
Ex:- Social media posts, images, videos, audio files, emails.
Characteristics:- Does not fit into traditional databases, requires advanced analytics tools and often contains rich, contextual information.

Q) List and explain characteristics of big data illustrate, consider by an example of E-commerce, how big data is used.

The characteristic of Big data



1: Composition: The composition of data deals with the structure of data, that is, the source of data, The granularity, The types and the nature of data as to weather it is static or real time streaming

2) Condition:- The condition of data deals with the state of data, that is" can one use this data as is for analysis?" or" Does it require.

Cleaning for further enhancement and enrichment?"

8) Context :- The context of data deals with "where has this data generated?" "how sensitive is this data?" "what are the events associated with this data?" and so on

Let's consider an e-commerce platform like Amazon or Flipkart.

How Big data is used

1) Personalization : Big data helps in personalizing the shopping experience for customers by analysing their browsing history, purchase patterns and search queries.

2) Inventory Management : E-commerce platforms use big data analytics to predict demand, manage inventory levels, and optimize supply chain operations.

3) Customer Segmentation : Big data enables segmentation of customers based on behaviors, preferences and demographics to target marketing campaigns effectively.

4) Fraud Detection :- Analysing Big data helps in detecting and preventing fraudulent transactions by identifying casual patterns.

Characteristics of Big data :-

i) Volume : The amount of data.

ii) Velocity : The speed at which data is generated and processed

iii) Variety : The type of data (structured, semi structured, unstructured)

iv) Veracity : The quality and accuracy of data.

* Big data often includes inaccurate, incomplete or inconsistent data. Ensuring data quality and filtering out noise is essential for reliable analytics

Value :-

The usefulness of data for decision making

Date / /
Page

Q 8) Discuss about importance of Big data Analytic

—> Big data analytics play a crucial role in todays data driven world Here are some important of Big data analytics

1.) Reactive - Business Intelligence :- what does Business intelligence (BI) help us with? It allows the business to make faster and better decision by providing the right information to right person at the right time in the right format.
It is about analysis of the past or historical data and then displaying the finding of the analysis or reports in the form of enterprise dashboards, alerts notifications, etc. It has support for both pre-defined specified reports as well as ad hoc quering

2) Relative - Big data Analytics :- Here The analysis is done on huge data sets but the approach is still reactive as it is still based on

Date / /
Page

static data

3) proactive analytics : The analytics is done on huge datasets, it support futuristic decision making by the use of data mining, predictive modeling, text mining, and statistical analysis, this analysis is not on big data as it still uses the traditional database management practices on big data and therefore has sever limitations on the storage capacity and the processing capability.

4) proactive - Big data analytics :- This is sieving through terabyte, petabytes, exabytes of information to filter out the relevant data to analyze, This also includes high performance analytics to gain rapid insights from big data and the ability to solve complex problems using more data

Q 4) Explain about big data and data warehouse coexistence.

Arka

Big data and data warehouses serve different purposes but can complement each other within an organization data architecture.

1) Data warehouses :-
* Store structured data
* optimised for SQL queries and business intelligence.
* used for reporting, dashboards and historical analysis.

2) Big data platforms :-
* Handle large volume of structured semistructured and unstructured data.
* support advance analytics, machine learning and real-time processing
* often use distributed computing frameworks like Hadoop or Spark

How they coexistence :-
1) Different Purposes :- Data warehouses are designed for structured data and support designed for structured data and support bussiness intelligence, reporting and analytics Big data solutions handle large volumes of unstructured and semi structured data for advanced analytics, machine learning and real-time processing

2) Data integration : Organisations can integrate data from big data platforms into their data warehouse for comprehensive analysis and reporting.

3) Data Lake Architecture : data lake can store big data, which can then be processed and loaded into a data warehouse for structured analytics

Q5) Explain about typical data warehouse and hadoop environment.

=> A typical data warehouse environment is a operational or transactional or day-to-day bussiness data is gathered from enterprise Resource planning systems, customer Relationship management (CRM)

legancy systems and several third party applications
* The data from these sources may differe in format [data could have been housed in any RDBMS such as oracle, MS SQL server, DB2, MySQL, and Teradata, and so on or in spreadsheet .. (.xls, xlsx, etc) or .csv or txt]
* Data may come from data sources located in the same geography or df geographics.
* this data is then integrated, cleaned up, transformed, and standardized through the process of Extraction, Transformation and loading (ETL).
The transformed data is then loaded into the enterprise data warehouse (available at the enterprise level) or data marts (available at the bussiness unit / functional unit level)
* A host of market leading bussiness intelligence and analytics tools are then used to enable decision making from the use of ad-hoc queries, SQL, enterprise dashboards

data mining etc.

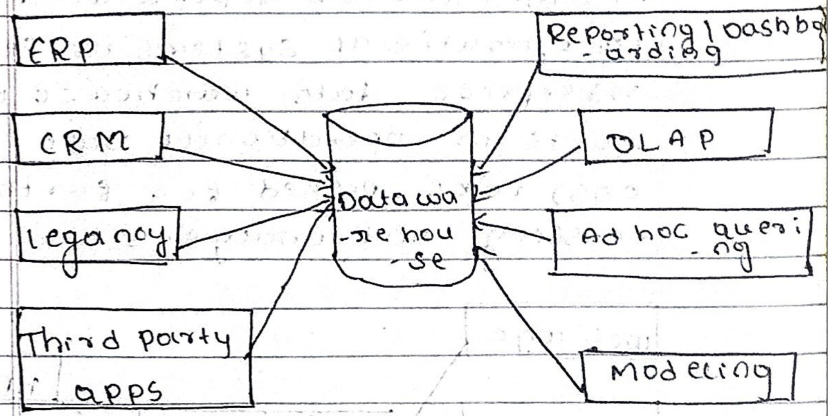


fig: A typical data warehouse Environment

hadoop environment :-
It is very different from the data warehouse environment and where exactly is this difference.

* The data source are quiet dispa-rate from web logs to images audios and videos to social media data to the various docs, pdfs etc. Here the data is focus is not just the data within the company's firewall but also data residing outside the company's firewall

The data is placed in Hadoop Distributed file system (HDFS). If need be this can be repopulated back to operational systems or fed to enterprise data warehouse or data marts or operational data store (ODS) to be picked for further processing and analysis.

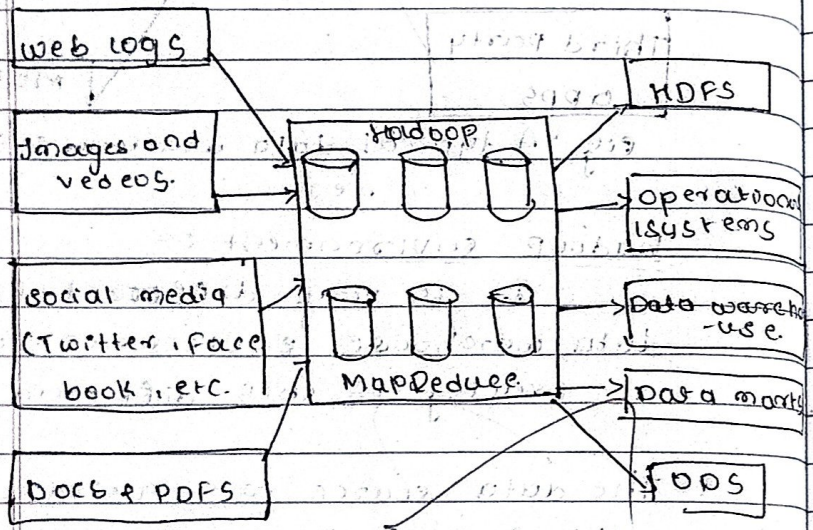


Fig: A typical Hadoop environment

Q6) Explain the technologies used in big data environment and about few top analytical tools.

=> Technologies used in Big data environment Big data environments utilize a variety of technologies to store, process and analyze large volumes of data, some key technologies include.

Storage Technologies:-

Hadoop Distributed file System (HDFS): A distributed file system for storing large amounts of data.

NOSQL Databases: Databases like HBase, Cassandra, MongoDB are designed for handling large volumes of unstructured and semi-structured data

processing Technologies:-

Map Reduce:- A programming model for processing data in parallel across a cluster of nodes.

Apache Spark: An in-memory data processing engine for fast data processing and analytics

Analytic technologies:

Date / /
Page

-> Machine learning :- libraries like Tensorflow, pytorch and scikit-learn are used for building machine learning models

-> Data mining :- Technologies like clustering decision trees, and association rule mining are used for discovering patterns in data.

Top analytical tools :-

1) Power BI
* Description : A business analytics service by microsoft for data visualization and reporting.
* use case : Data visualization, reporting and business intelligence.

2) Apache Spark.
* Description : An in memory data processing engine for fast data processing and analytics.
* use case : Real time data processing, machine learning and data analytics

4) Google Analytics
* Description : A web analytics service for tracking website traffic and behaviour
* use case : web analytics, user behavior analysis and marketing optimization

Q7) Explain about NOSQL and its types with an example.

-> NOSQL (NOT ONLY SQL)

NOSQL database are designed to handle large volumes of unstructured and semi-structured data. They offer flexible scheme designs, high scalability and high performance.

Few features of NOSQL databases are as follows.

1) They are open source
2) They are non-relational
3) They are distributed
4) They are schema-less

5) They are cluster friendly.

Types of NOSQL Databases :-

we have already stated that NOSQL data bases are non-relational. They can be broadly classified into the following

1) key value : it maintains a big hash table of keys and values.
for example : Dynamo, Redis, Riak

2) Sample key-value pair in key-value pair in key-value database

key	value
first name	Simmonds
last name	David

3) Document : it maintains data in collections constituted of documents
for example : MongoDB, Apache couch-chDB, couchbase, MarkLogic etc.

4) column : each storage block has data from only one column
for example : cassandra, HBase etc

5) Graph : They are also called network data base ; A Graph stores data in nodes for example : Neo4j, HyperGraphDB etc.

Module - 2

Q 1) Explain about Hadoop ? why Hadoop and history of Hadoop

-) Hadoop is an open source, distributed computing framework that enables the processing and storage of large volumes of data across a cluster of nodes. it was designed to handle massive amounts of data known as big data, provide a scalable, fault tolerant, and flexible solution for data processing and analysis

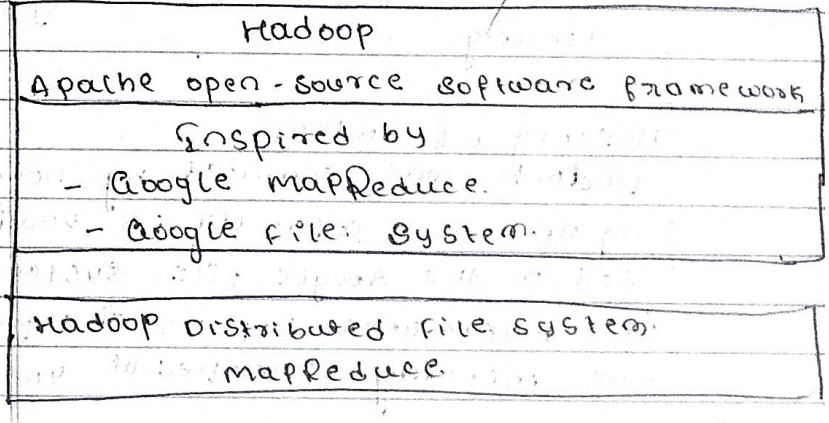


Fig :- Hadoop

Date __/__/__

Page _____

Why Hadoop ?

Hadoop was created to address the limitations of traditional data processing systems which were unable to handle the rapidly growing volumes of data. Hadoops distributed architecture allows it to scale horizontally. Add more nodes to the cluster as data volumes grow.

* Handle unstructured data: process and store data in various formats such as text, images, and videos.
* provide fault tolerance : Automatically replic data and tasks across nodes to ensure data availability and processing continuity.

History of Hadoop.

Hadoop was founded by Doug cutting and mike cafarella in 2005, based on the Google File System (GFS) and mapReduce papers. The project was initially developed at yahoo!

Arka

Date __/__/__

Page _____

and later become an Apache software function project

* 2004 : Doug cutting and mike capa-rella started working on Hadoop at yahoo!
* 2006 : Hadoop graduated from the Apache incubation and become a top level project
* 2011 : Apache Hadoop 1.0 was released, marking a significant milestone in the project's development

Q2) List Hadoop core components and explain with appropriate diagram

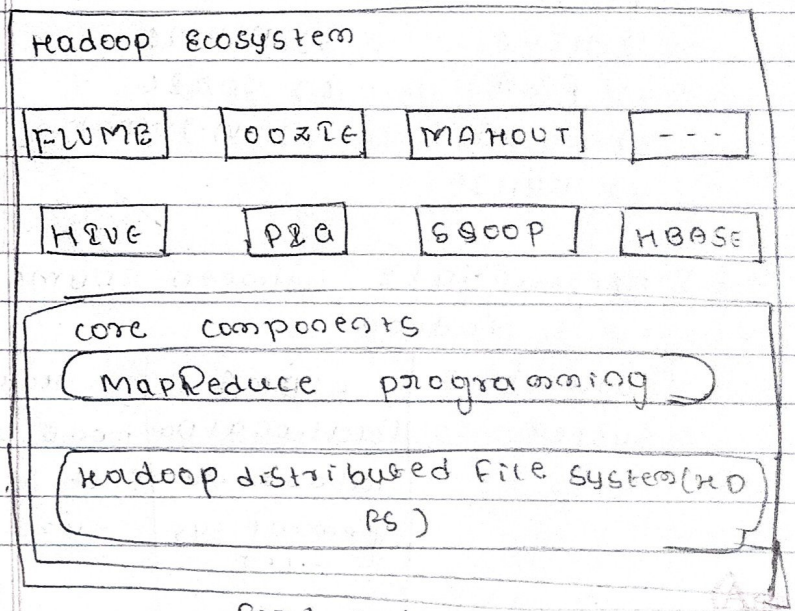


Fig : Hadoop components

Hadoop core components

1) HDFS
-> Storage component
-> Distributed data across several nodes
-> Natively redundant

2) MapReduce
-> Computational framework
-> Splits a task across multiple nodes
-> process data in parallel

-> Hadoop Ecosystem :- Hadoop Eco-system are support projects to enhance the functionality of Hadoop cost components. The Eco projects are as follows:

1) HIVE
2) PIG
3) SQOOP
4) HBASE
5) FLUME
6) OOZIE
7) MAHOUT

Q 3) Differentiate between RDBMS VS Hadoop

parameters	RDBMS	Hadoop
System	Relational Database management system	node based flat structure.
Data	Suitable for Structured data	Suitable for structured, unstructured data, supports variety of data formats in real time such as XML, JSON, text based flat file format etc
processing	OLTP	Analytical, big data processing
Choice	when the data needs consistent relationship	Big data processing which does not require any consistent relationship blw data
Choice.	when the data needs consistent relationship	Big data processing which does not require.

		any consiste -nt relations -hip blw data
processor	Needs expensi -ve hardware or night-end processors to store huge volumes of data	In a Hadoop cluster, a node requires only a processor, a net -work card, and few hard drive -rs cost around $4000 per ter ~~objects~~ - a bytes of storage

Q4) Explain HDFS and the features

Hadoop Distributed File System (HDFS):
HDFS is a distributed file system designed to store and manage target volume of data across a cluster of nodes it provides a sca -lable, fault-tolerant and flexible sln for data storage and proce -ssing

Key features of (HDFS) :-

→ HDFS is a distributed file system designed to store and manage large volume of data across a cluster of nodes it provides a scalable, fa -ult-tolerant and flexible soln. for data storage and processing

Key features of HDFS

* Distributed storage : HDFS stores data across multiple nodes in a cluster, allowing for horizontal scaling and increased storage capacity

* Fault Tolerance :- HDFS replica -tes data blocks across nodes to ensure data availability and dur -ability in case of node failure -es

* High throughput :- HDFS is opti -mized for high throughput data acess, making it suitable for large -scale data processing and analy -tics.

* Flexible data model. :- HDFS suppo

Date __/__/__

Page _____

- its a variety of data formats, including structured, semi-structured and unstructured data.

HDFS features.

* Data Replication :- HDFS replicates data blocks across nodes to ensure data availability and durability
* Block size :- HDFS stores data in fixed size blocks (typically 128 MB or 256 MB) which allows for efficient data processing and storage
* Data integrity :- HDFS uses checksums to ensure data integrity and detect corruption
* Scalability :- HDFS is designed to scale horizontally, allowing for the addition of new nodes as storage needs grow.