Unit 3

(11) Diff btw Structured, unstructured and semi structured data

Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table	It is based on XML/ Resource Description framework (RDF)	It is based on character & binary data.
Version management	Versioning over tuples, rows, tables	Versioning over tuples or graph is possible	Versioned as a whole
Flexibility	It is schema dependent & less flexible	It is more flexible than structured data but less flexible than unstructured data.	It is more flexible & there is absence of schema.
Scalability	It is very difficult to scale DB schema	It is scaling is simpler than structured data	It is more scalable.
Robustness	very robust	New technology not very spread	—
Query performance	structured query allow complex joining	Queries over anonymous nodes are possible	only textual queries are possible

unstructured

Analytics of unstructured data

Unstructured data is data that doesn't have a fixed form or structure. Images, videos, audio files, text files, social media data, geospatial data, data from IoT devices, & surveillance data are example of unstructured data. About 80% - 90% of data is unstructured.

Unstructured data analysis is complex & requires specialized techniques, unlike structured data, which is straightforward to store & analyze.

Here is a quick glance at all of the unstructured data analysis techniques we discussed.

1. Keep the business objective(s) in mind.
2. Define metadata for faster data access.
3. Choose the right analytics techniques.
   a. Exploratory data analysis techniques
   b. Qualitative data analysis techniques
   c. Artificial Intelligence (AI) & machine learning (ML) techniques.
4. Identify the right data sources
5. Evaluate the technologies you'd want to use
6. Get real time data access
7. Store & integrate data using data lake
8. Wrangle the data to get the desired features.

Data analytics use cases.

To achieve the strategic objectives of the company a better data strategy will help how to use that data in an efficient manner. These data use cases which are identified in this process are called use cases. These use cases are the key data elements for the data projects ahead.

While developing a data strategy for a business, a few optimal numbers of use cases determined. Over a no. of use cases result in an unrealistic & cluttered data strategy which may be risky. These data use cases are different for every business which is actually driven by the business strategy

The main elements involved in designing & developing better data use cases are.

1. Data evaluation.
2. Identification of suitable analytical techniques & tools
3. Hierarchy of outcome metrics

Here are top five use cases for Data Analytics

① Security Intelligence :-

Data analytics are also deployed for improving security operations against hackers & cyber criminals. An IT department handles a large amount of data. Security is an important factor; many companies have use these analytics to help in obtaining better solutions to detect & prevent such attacks. User and Entity Behaviour analytics (UEBA), security information event management (SIEM) tools, & machine learning can be used to detect abnormalities, & unusual user activities.

② Customer Relationship Analytics:

One of the difficult tasks in marketing is to identify customers who are going to spend money consistently for a long time. This business insight will help the companies to gain such customers which will be a lifetime value for company. The data analytics examples in business include Telecommunications, Banking, Utilities & retail. Customer segmentation also helps in establishing potential marketing strategies.

③ Recommendation Engines:

You may have noticed "recommendations for you" on youtube, Spotify or other media services. These personalized recommendations are done using & help in improving the overall user experience. This can be a winning factor as there is a lot of competition in the entertainment & media sector, you can also see such recommendations while shopping online.

④ E-Commerce:

A system wide infrastructure with data analytics is a superior way to enhance the efficiency and performance of the business. The system metrics are used to track the performance of IT modules & the user logs. These will identify user behaviour in e-commerce sites. By using this data, retailers can gain insights that help in developing agile techniques, better business performance & profits.

⑤ Internet of things:

There are endless ways of applying data analytics to IOT solutions for eg:- almost every industry uses sensor data to have actionable insights as a measure of productive maintenance. Customer revenue may be tracked as a part of the security options of a device. Other examples can be logistic tracking, where the vehicles are tracked from time to time & rerouted in the case of bad weather or unforeseen circumstances that affect delivery time.

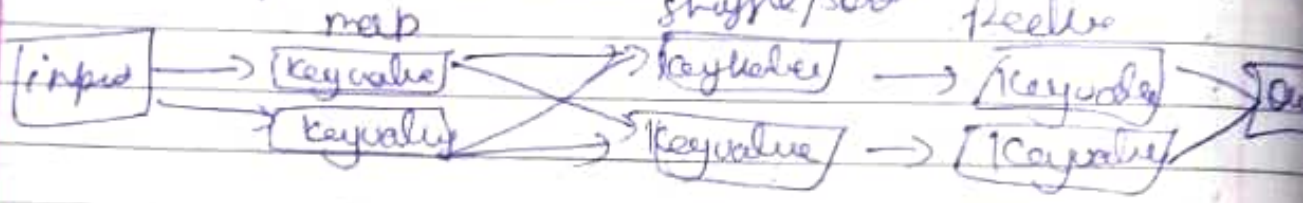
Mapreduce

It is one of the core component of hadoop which actually for the computation it has 2 components.

(1) Job tracker (2) Task Tracker

Job tracker ⇒ map reduce splits the jobs into various small small task & this is accomplished by the Job tra

Task tracker ⇒ It actually performs the computation if we have created a cluster of 50 nodes then the master node will have one job tracker & one Task tracker the slave nodes will have one task tracke So in all there will be one job tracker in the cluster & 50 task tracker.



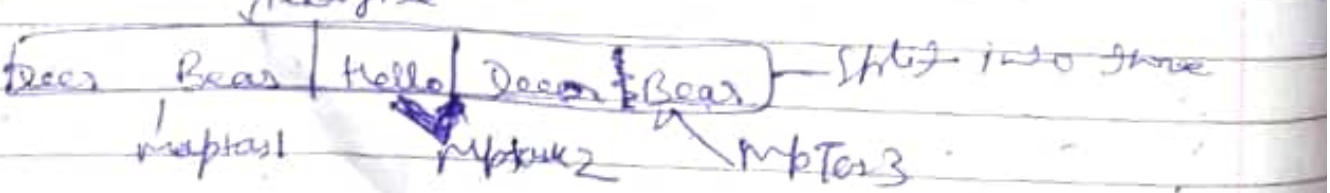
There are 3 functions inter the mapreduce.

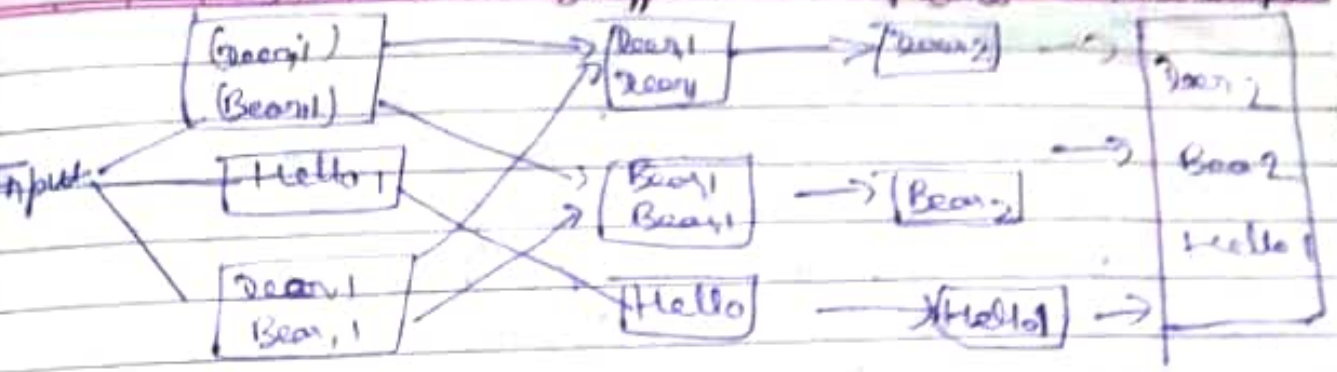
(1) Mapreduce → the data one computations is transfered to the mapreduce where a block of data is read & processed & produced key value pairs as intermediate output

(2) Shuffle or Sort ⇒ The intermediate output generated by map function is received by shuffle function where the data is shuffled or sorted based on their keys & the output is transferred to the reduce function

(3) Reduce → The final computations is performed by the reduce function & the output generated is combined together to form the final output

Eg





Benefits of map reduce

1) Parallel processing =) In map reduce job is divided among multiple nodes and each node works with a part of job simultaneously. So map reduce is based on divide and conquer paradigm which helps us to process the data using different machines as the data is processed by multiple machines instead of a single machine in parallel. The time taken to process the data gets reduced by tremendous amount.

2) Data locality =) In the traditional system data is sent to the processing unit & then it is processed but as the data grows & becomes very huge, having this amount of data to the processing unit posed the following issues.

a) moving huge data to processing unit is costly & degrades the network performance.

b) The master node get over burdent & may fail map reduce allows us to overcome the above issues by bringing the processing unit to the data is distributed among multiple nodes where each node processes the part of the data residing on it. This allows us to have the following advantage.

(1) To move processing unit to the data.

(2) The processing time is reduced as all the nodes are working with their part of the data in parallel.

(3) Every node gets the part of data to process & therefore there is no chance of a node getting over burdent.

Hadoop

Apache Hadoop is an Open source, Java based software platform that manages data processing & storage for big data applications. Hadoop works by distributing large data sets & analytics jobs across nodes in a computing cluster, breaking them down into smaller workloads that can be run in parallel. Hadoop can process structured & unstructured data & scale up reliably from a single server to thousand of machines.

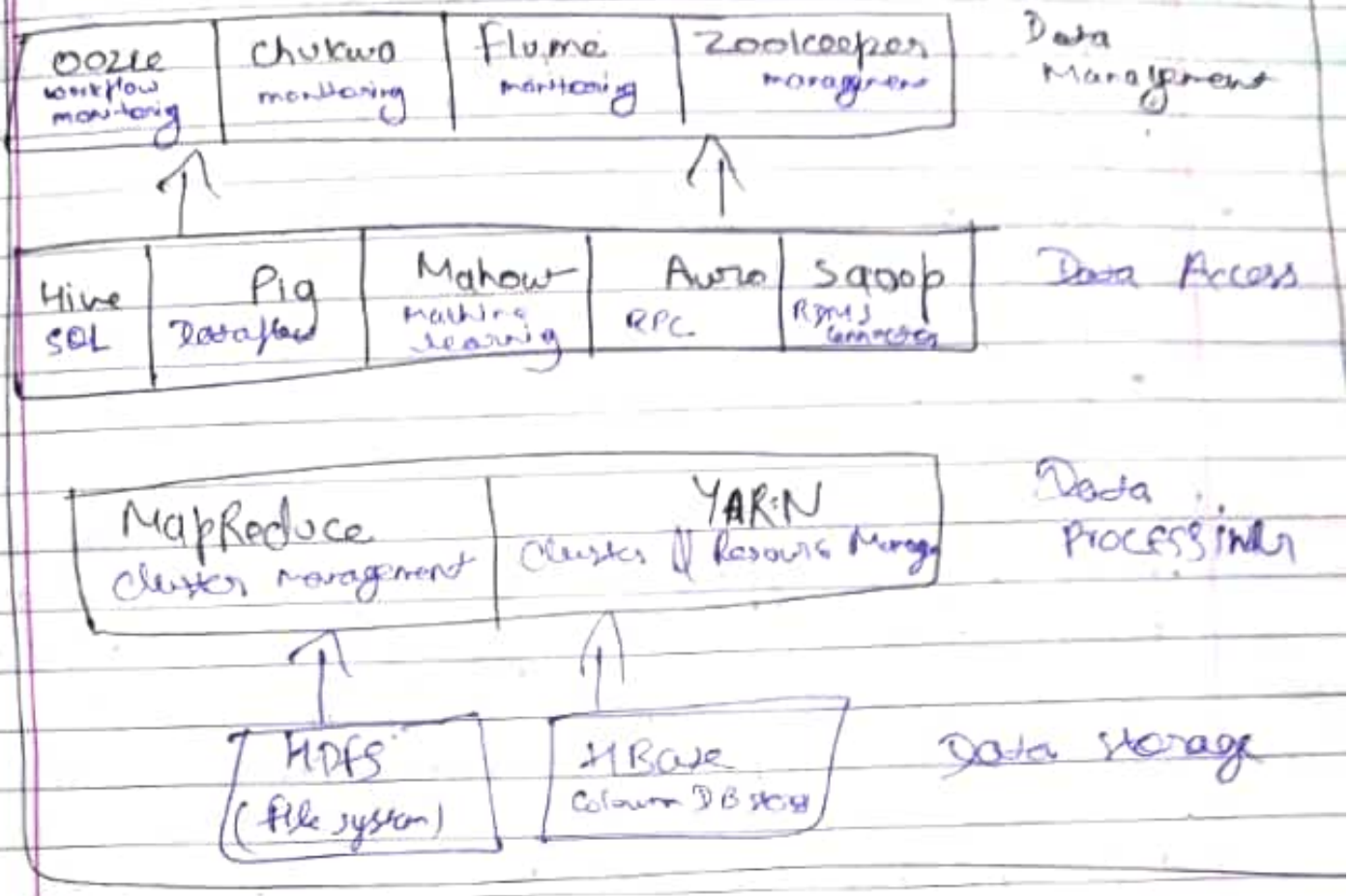
It is highly scalable.
It consists of 3 component

(1) Hdfs :- Reliable storage system with half of the world data stored in it
(2) Mapreduce :- layer consist of distributed processor
(3) yarn : layer consist of resource manager.

Hadoop Ecosystem

The Hadoop ecosystem is a framework that helps in solving big data problems. The core component of the Hadoop ecosystem is a Hadoop distributed file system. Hdfs is a distributed file system that has the capability to store a large stack of data sets. With the help of shell-commands HADOOP interactive with Hdfs. Hadoop Breaks up unstructured data & distributes it to different sections for Data analysis. The ecosystem provide many component & technologies that have the capability to solve complex business tasks.

Hadoop Ecosystem



Hdfs

- Hdfs is the primary or major component of Hadoop ecosystem and it's responsible for storing large data sets of structured or unstructured data across various nodes and thereby maintaining the metadata in the form of log files.
- Hdfs consist of 2 core component i.e
  (1) Namenode (2) Datanode
  - Namenode is the prime node which contains metadata (data about data) requiring comparatively fewer resources than the datanodes are commodity hardware in the distributed environment. undoubtedly, making Hadoop cost effective.
- Hdfs maintains all the coordination between the clusters & hardware thus working at the heart of the system.

Yarn

- yet another Resource Negotiator as the name implies YARN is the one who helps to manage the resources across the clusters. In short, it performs Scheduling & resource allocation for the Hadoop System.
- Consists of 3 major components i.e.
  1. Resource manager
  2. Node manager
  3. Application Managers.

- Resource manager has the privilege of allocating resources for the applications in a system whereas Node managers work on the allocating of resources such as CPU, memory bandwidth per machine & later on acknowledge resource managers. Application manager works as an interface between the resource manager & node manager & performs negotiations as per the requirement of the two.

MapReduce :-

- By making the use of distributed & parallel algorithms MapReduce makes it possible to carry over the processing's logic & helps to write applications which transform big data sets into a manageable one.
- MapReduce makes the use of 2 functions i.e.
  1. map() & Reduce() whose task is:
  1. Map() performs sorting & filtering of data & thereby organizing them in the form of group. Map generates a key value pair based result which is later on processed by the Reduce() method.
  2. Reduce(), as the name suggests does the summarization by aggregating the mapped data

In single Reduce() - takes the output generated by map() as input & combines those tuples into smaller sets of tuples.

PIG

- Pig was basically developed by yahoo which works on Pig latin language, which is Query based language similar to SQL.
- It is a platform for structuring the data flow, processing and analyzing huge data sets.
- Pig does the work of executing commands & in the background all the activities of mapReduce are taken care of. After the processing pig stores the result in HDFS.
- Pig latin language is specially designed for this framework which runs on Pig runtime. Just the way Java runs on the JVM.
- Pig helps to achieve ease of programming & optimization hence is a major segment of the hadoop ecosystem.

HIVE

- With the help of SQL, methodology & interface, HIVE performs reading & writing of large data sets. However its query language is called as HQL (Hive Query language)
- It is highly scalable as it allows real-time processing & batch processing both. Also, all the SQL datatypes are supported by hives thus making the query processing easier.
- Similar to the Query processing frameworks HIVE too comes with two components: JDBC drivers & HIVE command line.
- JDBC along with ODBC drivers work on establishing the data storage permissions & connection whereas HIVE command line helps in the processing of queries.

Mahout :-

- Mahout allows machine learnability to a system application. Machine learning as the name suggests helps the system to develop itself based on some patterns, user/environmental interaction or on the basis of algorithms.
- It provides various libraries or functionalities such Collaborative filtering, Clustering & Classification which are nothing but concepts of machine learning. It allows invoking algorithms as per our need with its own libraries.

Apache Spark :

- Its platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, & visualizations etc.
- It consumes in memory resource hence, thus being faster than the prior in terms of optimization.
- Spark is best suited for real time data whereas Hadoop is best suited for structured data or batch processing hence both are used in most of the companies interchangably.

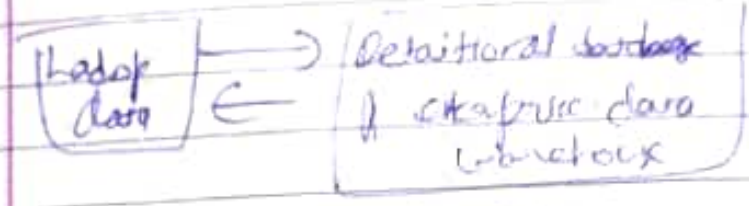
Apache H-Base :

- It a NoSQL database which support all kinds of data & thus capable of handling anything of Hadoop Database. It provides capabilities of Google's BigTable thus able to work on Big data sets effectively.
- At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short

quick span of time. At such times, HBase comes handy as it gives us a reliable way of storing limited data.

Sqoop

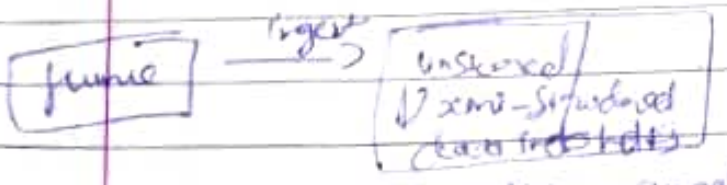
Sqoop is used to transfer data between Hadoop and external datastores such as relational databases and enterprise data warehouses. It imports data from external data stores into HDFS, Hive & HBase.



It can also duplicate machine generated data, which will then be sent to Sqoop. The Sqoop then goes to the Task manager, which in turn connects to enterprise data warehouse documents based system of RDBMS. It can map those task into Hadoop.

Flume

Flume is a data collection & ingestion tool & distributed service for collecting, aggregating & moving large amount of log data. It ingests online streaming data from social media, log files, web server into HDFS.

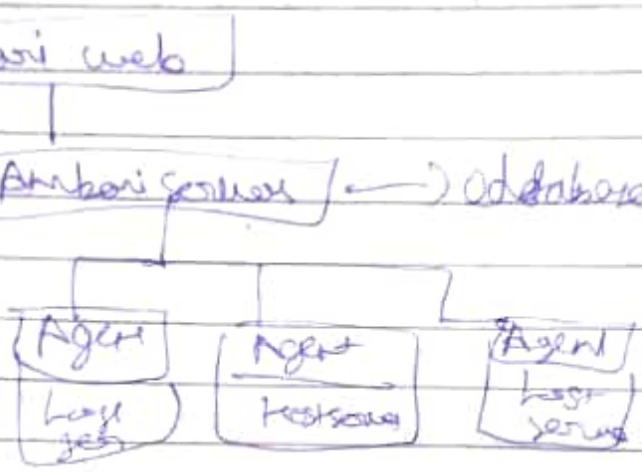


Data is taken from various sources, depending on your organization's needs. It then goes through the source, channel & sink. The sink feature ensure that everything is in sync with the requirements. Finally the data is dumped into HDFS.

Page No.:
Date:

Ambari

It is an open source tool responsible for keeping track of running applications & their statuses. Ambari manages, monitors, & provisions Hadoop clusters. Also it also provides a central management service to start, stop & configure Hadoop services.



Storm

The Storm is an engine that process real-time streaming data at a very high speed. It is written in Clojure. A storm can handle over 1 million jobs on a node in a fraction of a second. It is integrated with Hadoop to harness higher throughputs.

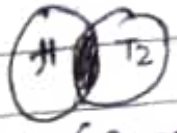
Now that we have looked at the various data wrapping tools & streaming services let us take a look at the security frames in the Hadoop ecosystem.

SQL Joins

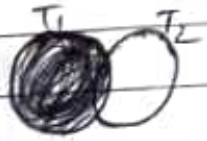
A Join Clause is used to combine rows from 2 or more tables based on related columns between them.

Types of SQL Joins

(Inner) Join: Returns records that have matching values in both tables.

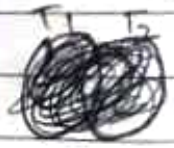


Left (OUTER) Join: Returns all records from the left table & the matched records from the right table



Right (Outer) Join: Return all records from the right table & the matched from the left table.

full (outer) join: Returns all records when there is a match in either left or right tables



SQL Set Operations

The SQL set Operations is used to combine the two or more SQL SELECT statements.

Types of Set Operators

(1) Union =>

- The SQL union operator is used to combine the result of two or more SQL SELECT queries.
- In the UNION operation, all the no of datatype of columns must be same in both tables on which UNION operation is being applied.
- The Union operation eliminates the duplicate rows from its resultset.

Eg SELECT column_name FROM table1
UNION
SELECT column_name from table2;

(2) Union All

Union all operation is equal to the UNION operation but it returns the set without removing duplicate & sorting the data.

Eg SELECT column_name FROM table1
UNION ALL
SELECT column_name FROM table2;

(3) Intersect

- It is used to combine two SELECT statements. The Intersect operation returns the common rows from both the SELECT statements.
- In the Intersect operators, the no of datatype & columns must be the same.
- It has no duplicates & it returns the data in ascending order by default.

Eg SELECT column_name from table1
INTERSECT
SELECT column_name FROM table2;

④ MINUS

- It combines the result of 2 SELECT statements.
- MINUS operator is used to display the rows which are present in the first query but absent in the 2nd query.
- It has no duplicates & data arranged in ascending order by default.

Eg SELECT Column_name from table1,
MINUS
SELECT Column_name from table2;