Categorizing documents by topics in

Text analysis

⇒ Categorizing documents by topics in Text analysis typically involves using techniques like natural language processing and Machine learning algorithms.

⇒ These algorithms can analyze the content of documents to determine their main themes or topics based on the words and phrases used.

- Common methods include.

(1) Topic modelling:- It's type of statistical model for discovering topics that occur in a collection of document

↳ Latent Dirichlet Allocation (LDA) is a popular topic modelling technique that assumes each document is a mixture of topics. ~~and~~ ~~each word's presence~~

⇒ The topics produced by topic modeling techniques are clusters of similar words

Catagorizing documents by topics in
Text analysis

⇒ Catagorizing documents by topics in Text analysis typically Involves using techniques like natural language processing and Machine learning algorithms.

⇒ These algorithms can analyze the content of documents to determine their main themes or topics based on the words and phrases used.

- Common methods include.

(1) Topic modelling:- It's type of statistical model for discovering topics that occur in a collection of document

↳ Latent Dirichlet Allocation (LDA) is a popular topic modelling technique that assumes each document is a mixture of topics. ~~and each word's presence~~

⇒ The topics produced by topic modeling techniques are clusters of Similar words.

→ It is used for finding Semantic structure of text body

⇒ Latent Dirichlet Algorithm iteratively updates is a flexible generative probabilistic model.

⇒ It is a unsupervised model.

⇒ "LDA assumes each document is a set of topics, and that each word in the document is attributable to one of the document's topics."

⇒ Initially, the model assigns words in the documents to topics.

⇒ LDA iterates to improve the assignment of words to topics ~~based on probabilities~~ until it converges to a stable solution.

⇒ During iterations, it adjusts the assignment of words to topics based on probabilities.

⇒ The Output of LDA is typically a set of topics, each represented by a list of words that are most likely to occur within that topic. also for each document, the probability of belonging to Various topics.

(2) Clustering Algorithms.

⇒ These algorithms group documents into clusters based on their similarity in terms of word usage.

⇒ K-means clustering is commonly used for this purpose

(3) Deep Learning Approaches.

Deep learning models, particularly neural networks, can also be used for document categorization.

eg: Recurrent neural networks (RNNs)
convolutional neural networks (CNNs)

Determining sentiments in Text analysis

It is the process of computationally identifying and categorizing opinions from piece of text, and determine whether the writer's attitude towards a particular topic or the product is positive, negative, neutral.

For this ~~we~~ the ~~fol~~ we are using following approaches:-

- Lexicon-Based approaches

These methods use sentiment lexicons, which are dictionaries containing words labeled with their sentiment (positive, negative or neutral). The sentiment of a document is determined by aggregating the sentiments of its constituent words.

- Machine Learning-based approaches

Machine learning techniques, like classification algorithms, (eg: Support vector machines, Naive Bayes, Neural n/ws) can be used to classify text into sentiments.

⇒ By applying these techniques, sentiment analysis can provide valuable insights to public opinion, customer feedback, social media trends and more. It helps to organizations to make decisions and understand the sentiment of their audience.

Gaining Insights in Text Analysis

⇒ Gaining insights from text analysis involves extracting meaningful information, patterns, and knowledge from textual data to make decisions and to understand underlying trends.

[Note:- you can write the steps in life cycle also]

Module : IV

Hadoop Ecosystem

⇒ Apache Hadoop is an opensource frame work intended to make interaction with big data easier

⇒ Hadoop ecosystem is a platform or suit which provides Various services to solve the bigdata problems.

⇒ It includes Various Commercial tools and solutions.

Following are the components that collectively form a Hadop System

* HDFS (Hadoop Distributed System)

* YARN - Yet another Resource Negotiator)

* MAPREDUCE - (Pg-ing based Data processing)

* PIG, HIVE: (Query based processing data services)

* HBASE (No SQL DB)

* Sqoop (Data Exchange)

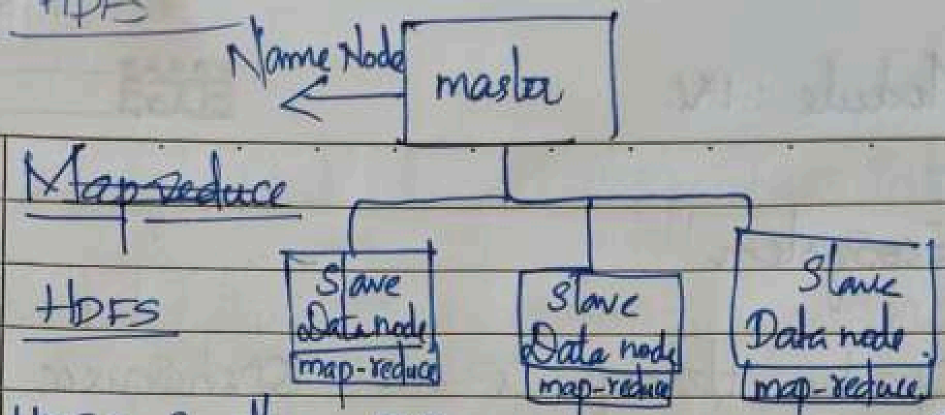
* Flume (log control details)

* Mahout (Machine Learning Libraries)

* Zookeeper (Cluster Manager)

* Oozie (Job scheduling)

HDFS



HDFS is the primary or major component of Hadoop ecosystem and it's Responsible for storing large datasets of structured and unstructured data across Various nodes and thereby maintaining the metadata in the form of log files. HDFS has Name node and Data node. Name node that guides the datanode as well as stores metad
Data node act as slave nodes are mainly used to store the data.

Map Reduce

MapReduce is a processing technique and a program model for distributed computing based on java.

It can process Vast amount of data in parellel across distributed computing environments

The map reduce algorithm contains 2 important tasks.

1. Map
2. Reduce.

Map stage.

The map / mapper's job is to process the input data. Generally, i/p data is in the form of file or directory and is stored in the Hadoop file System (HDFS). The i/p file is passed to the mapper function line by line The mapper processes the data and creates several small chunks of data.

Reduce stage

This stage is the combination of shuffle and Reduce stage.

The reducer's job is to process the data that comes from mapper

After processing, it produces a new set of output, which will be in the HDFs

Map reduce phases.

(1) Map phase.

The first phase is map reduce is known as map, during which the dataset file is divided into multiple splits.

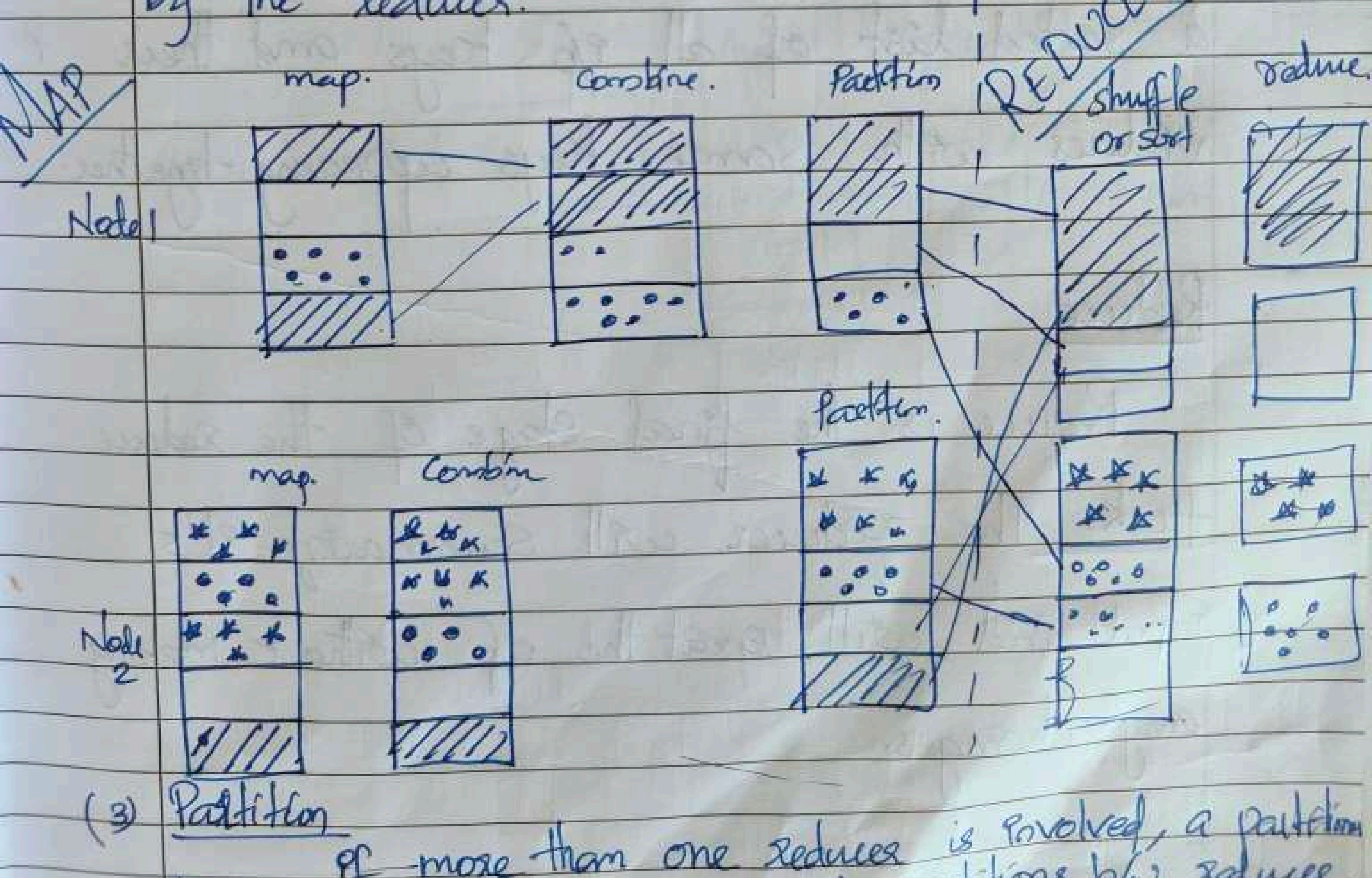
Each split is passed into its constituent records as key-value pair.

The key is usually the ordinal position of record, and the value is actual record.

The mapper processes each key-value pairs as per user logic and further generates a key-value pair as o/p.

(2) Combine phase

Generally, the o/p of the map function is handled directly by the reduce function. How ever map tasks and reduce tasks mostly run ove different nodes. But, there is a combine function that Summarize a mapper's o/p before it gets by the reduces.



(3) Partition

of more than one reduces is involved, a partition divides o/p. from mapper. into partitions b/w reducer.

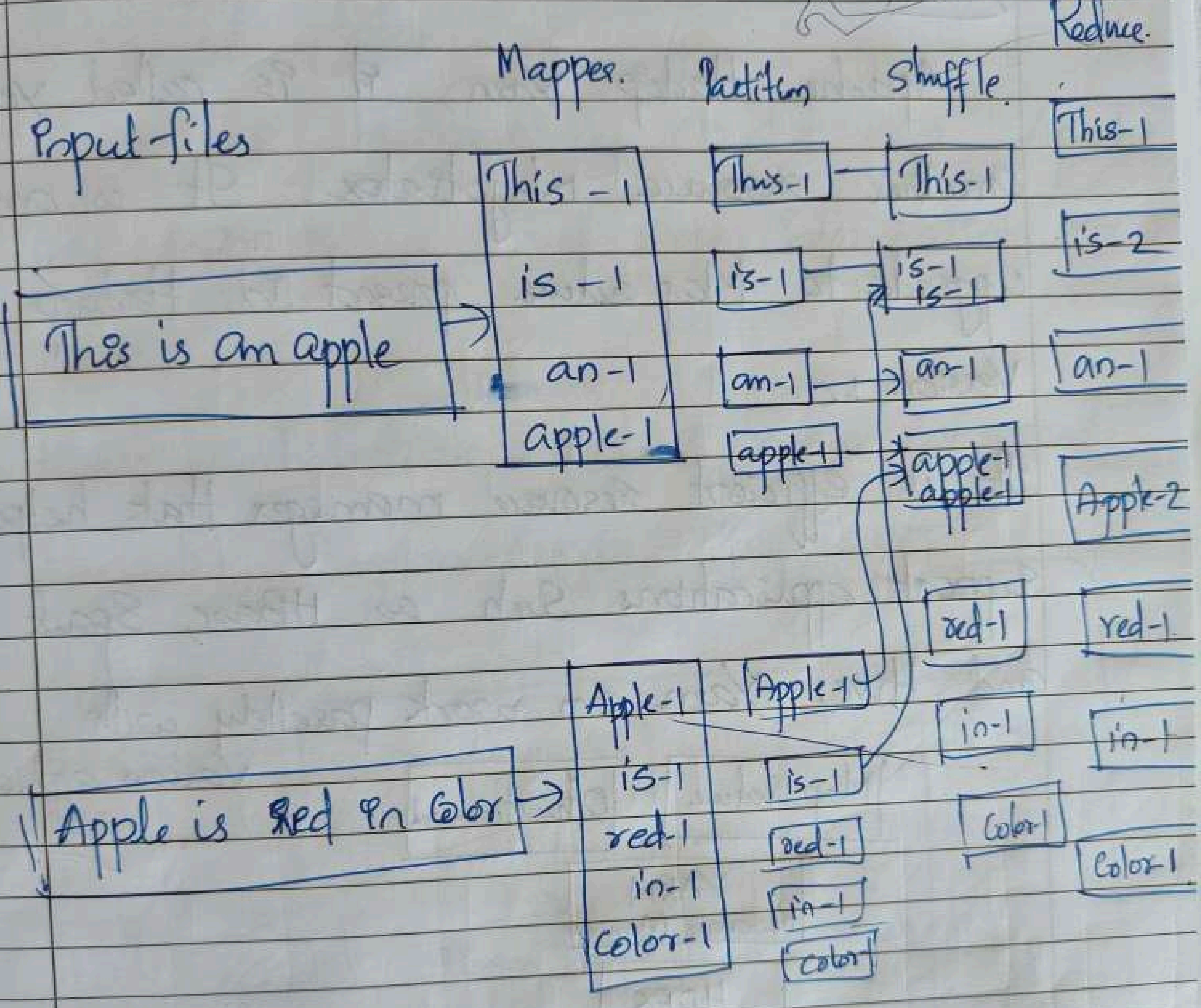
Shuffle and Sort

During first stage of reduce task, o/p from all partitioners is copied across the n/w to nodes running the reduce task. This is known as Shuffling.
Then it automatically groups and sorts according to the keys. So that o/p contains a sorted list of all i/p keys and their values with same keys appearing together

Reduce

Reduce is the final stage of the reduce task. The reducer will summarize it's input and will emit the o/p without making any changes.

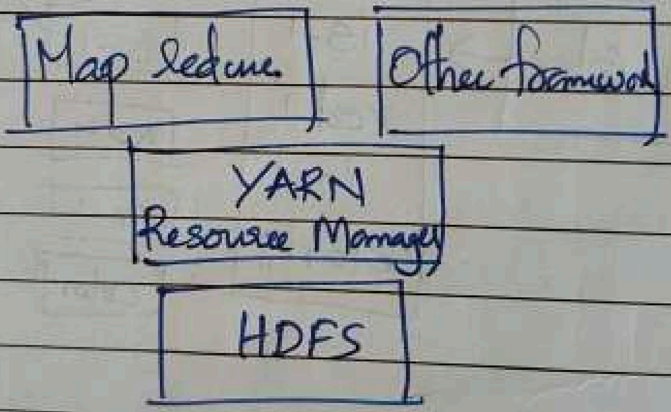
Example


Mapper.
Partition
Shuffle.
Reduce.
Input-files
This is an apple
This -1
is -1
an-1
apple-1
This-1
is-1
an-1
apple-1
This-1
is-1
is-1
an-1
apple-1
apple-1
This-1
is-2
an-1
Apple-2
red-1
red-1
in-1
in-1
Color-1
Color-1
Apple is Red in Color
Apple-1
is-1
red-1
in-1
color-1
Apple-1
is-1
red-1
in-1
color

Yarn

Apache Hadop Yarn, It is called yet another resource Negotiator. It is an upgrade to Map reduce present in Hadoop Version 1.0.

It is efficient resource manager that helps support applications such as HBase, Spark and Hive. Yarn can work parellely with Various applications



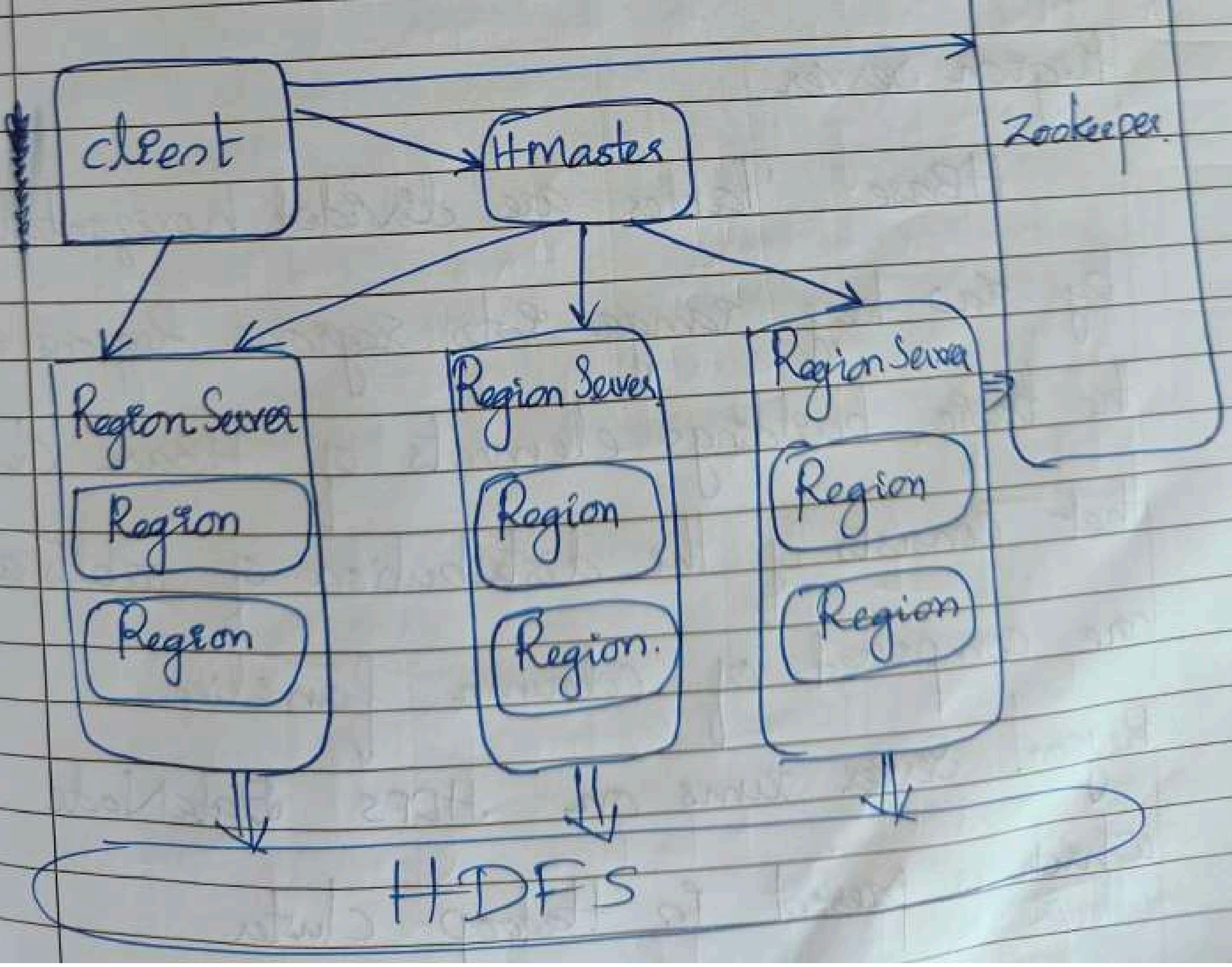
HBASE

HBASE is a top-level apache project written in java which fulfills the needs to read and write data in realtime.

⇒ It provides a simple interface to the distributed data.

⇒ It can be accessed by Apache Hive, Apache pig, Map Reduce, and store information in HDFS.

⇒ Hbase architecture has 3 main components Hmaster, Region Server, Zookeeper.



Hmaster

The Implementation of Master Server in HBase is Hmaster. It is a process in which regions are assigned to region Server as well as DDL (Create, delete table operations It monitor all region Server Instances

Region Server

HBase Tables are divided horizontally by row key range into regions. Regions are the basic buildings elements of HBase cluster that consists of the distribution of tables and are comprised of coloumn families. Region Server runs on HDFS DataNode which is present in Hadoop cluster.

PIG

Pig was basically developed by yahoo which works on a 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 activities of map-reduce are take care of. After processing pig stores the result in HDFS.

HIVE

⇒ With the help of SQL methodology and interface HIVE performs reading and writing of large datasets.

⇒ It's query language is called HQL (Hive Query Language).

⇒ It is highly Scalable as it allows real-time processing and Batch processing. Also all the SQL datatypes are supported by Hive, thus making the query processing easier.

⇒ Similar to query processing frame works, HIVE too comes with two components: JDBC Drivers and HIVE Command line.

⇒ JDBC, along with the ODBC drivers work on establishing the data storage permissions and connection whereas HIVE Command line helps in processing of queries.

Mahout

Mahout provides an environment for creating machine learning applications which are scalable. Machine learning algorithms allows us to build self-learning machine that evolve by itself without being explicitly programmed. Based on user behaviour data patterns and past experiences it makes important future decisions. We can call it a descendant of AI.

Mahout does?

i) Collaborative filtering.

Mahout mines user behaviour, their patterns and their characteristics based on that it predicts and make recommendations to the users. The typical use case is e-commerce website.

ii) Clustering

It organizes a similar group of data together like articles can contain blogs, news research papers etc.

iii) Classification

It means classifying and categorizing data into various sub departments like articles can categorized into blogs, news, essay, research papers etc.

iv) Frequent items set missing

Here Mahout checks, which objects are likely to be appearing together and make suggestions, if they are missing. for eg: cell phone and cover are bought together in general. so if you search for a cell phone, it will recommend you cover and cases.

Analytics- Technology and Tools: In-Database Analytics

What is In-Database Analytics

- In-database analytics is a technology that allows data processing to be conducted within the database by building analytic logic into the database itself. Doing so eliminates the time and effort required to transform data and move it back and forth between a database and a separate analytics application.
- An in-database analytics system consists of an enterprise data warehouse (EDW) built on an analytic database platform. Such platforms provide parallel processing, partitioning, scalability and optimization features geared toward analytic functionality.
- In-database analytics allows analytical data marts to be consolidated in the enterprise data warehouse.
- Data retrieval and analysis are much faster and corporate information is more secure .
- in-database analytics streamlines the analytics process, enhances performance, reduces complexity, and enables real-time or near-real-time insights generation.
- Companies use in-database analytics for applications requiring intensive processing – for example, fraud detection, credit scoring, risk management

Common examples of in-database analytics solutions include:

1. **SQL-Based Analytics:** Utilizing SQL queries with advanced analytical functions directly within the database system.
2. **Database-specific libraries:** Some databases offer libraries or extensions for machine learning, statistical analysis, and predictive modeling.
3. **Integrated Analytics Platforms:** Specialized analytical platforms or appliances that tightly integrate analytics and database capabilities for high-performance analytics.

SQL Essentials

1 Joins

- Joins in SQL are powerful operations used to combine rows from two or more tables based on related columns between them.
- They enable the retrieval of data from multiple tables simultaneously by establishing relationships between these tables.

There are different types of joins in SQL:

1.Inner Join:

- Returns rows when there is a match in both tables based on the join condition.
- SELECT * FROM table1 INNER JOIN table2 ON table1.column = table2.column;

2. Outer Join:

- Returns all rows when there is a match in either the left or right table. If there is no match, NULL values are included for columns from the opposite table.
- SELECT * FROM table1 FULL JOIN table2 ON table1.column = table2.column;

3. Left Join:

- Returns all rows from the left table and matching rows from the right table. If there is no match, NULL values are included for columns from the right table.
- SELECT * FROM table1 LEFT JOIN table2 ON table1.column = table2.column;

4.Right (Outer) Join:

- Returns all rows from the right table and matching rows from the left table. If there is no match, NULL values are included for columns from the left table.
- SELECT * FROM table1 RIGHT JOIN table2 ON table1.column = table2.column;

5.Self Join:

- When a table is joined with itself, typically used when the table contains hierarchical data or references to itself.
- SELECT e1.name, e2.name FROM employees e1 INNER JOIN employees e2 ON e1.manager_id = e2.employee_id;

2.Set Operations

- Set operations in databases are used to perform operations like union, intersection, and difference on the result sets of SQL queries.
- These operations allow data professionals to combine and manipulate data in various ways.

These set operations are handy for various scenarios in data analytics:

1.Data Integration: When combining data from multiple sources, UNION and UNION ALL help merge datasets with or without duplicates.

2.Data Validation: INTERSECT can be used to check for overlapping records between different datasets, ensuring data consistency.

3.Data Cleansing: EXCEPT or MINUS can identify data discrepancies or missing records between two datasets.

4.Data Manipulation: Set operations enable data professionals to filter and manipulate datasets in complex ways based on set theory principles.

Primary set operations:

1.UNION:

Combines the result sets of two or more SELECT statements into a single result set. It removes duplicates by default.

```
SELECT column1 FROM table1

UNION

SELECT column1 FROM table2;
```

2.INTERSECT:

Returns rows that appear in both result sets of two SELECT statements.

```
SELECT column1 FROM table1

INTERSECT

SELECT column1 FROM table2;
```

3.EXCEPT or MINUS:

Returns distinct rows from the first SELECT statement that are not present in the second SELECT statement.

```
SELECT column1 FROM table1

EXCEPT

SELECT column1 FROM table2;
```

In-Database Text Analysis

- In-database text analysis refers to performing text processing, search, and analysis directly within a database system.
- It involves using the database's capabilities to handle and analyze textual data, enabling various text-related operations without needing to extract data to external tools or platforms.
- This approach is particularly beneficial for managing and analyzing large volumes of textual data efficiently.

Here are key components and techniques involved in in-database text analysis:

1. **Full-Text Search:** Many database systems offer built-in full-text search capabilities. These functionalities allow users to perform keyword-based searches, find specific phrases or words within text fields
2. **Text Indexing:** Databases can create indexes specifically optimized for textual data, enabling faster search and retrieval operations on large volumes of text.
3. **Text Processing Functions:** Database systems may provide functions or extensions for text processing tasks, such as tokenization (splitting text into tokens/words), normalization (converting text to a standard form),
4. **Text Mining and Analytics:** In-database text analysis can include mining insights from text data, such as identifying trends, patterns, or associations within textual information.

Data Privacy and Ethics

- Data privacy and ethics are fundamental aspects of handling, managing, and utilizing data responsibly.

Privacy Landscape

1. **Protection of Personal Information:**

   Data privacy refers to the protection of sensitive and personally identifiable information (PII) of individuals. This includes names, addresses, social security numbers, health records, financial information, etc.

2. **Legal Compliance:**

   Adherence to data privacy laws and regulations, such as the GDPR (General Data Protection Regulation) in the European Union or CCPA (California Consumer Privacy Act) in California, which outline rules for collecting, storing, processing, and sharing personal data.

3. **Consent and Transparency:**

   Ensuring individuals are informed about how their data is collected, used, and shared. Obtaining explicit consent before collecting and processing their data.

4. **Data Security Measures:**

   Implementing robust security measures like encryption, access controls, data anonymization, and regular security audits to protect against unauthorized access, breaches, or data leaks.

Rights and Responsibilities

Rights and responsibilities in data privacy and ethics are crucial aspects that both individuals and organizations need to understand and uphold.

Rights:

1. **Right to Privacy:** Individuals have the right to control their personal data, including how it's collected, used, stored, and shared.
2. **Right to Access:** Individuals have the right to access their own data that's held by organizations and understand how it's being used.
3. **Right to Correction:** Individuals can request corrections or updates to inaccurate or outdated personal data.
4. **Right to Erasure (Right to be Forgotten):** Individuals can request the deletion or removal of their personal data under certain circumstances, especially if it's no longer necessary or if consent is withdrawn.
5. **Right to Data Portability:** Individuals have the right to obtain and reuse their personal data for their purposes across different services.
6. **Right to Consent:** Individuals have the right to give informed consent for the collection and processing of their data. Organizations must obtain clear and explicit consent for data usage.

Responsibilities:

1. **Data Protection and Security:** Organizations have a responsibility to implement robust data protection measures, ensuring the confidentiality, integrity, and security of individuals' data against unauthorized access, breaches, or misuse.
2. **Compliance with Regulations:** Organizations must comply with data protection laws and regulations applicable to their operations, including GDPR, CCPA, and other regional or industry-specific regulations.

3. **Transparency and Accountability:** Organizations should be transparent about their data practices, informing individuals about how their data is used and handled. They must also be accountable for their data handling practices.
4. **Data Minimization:** Collect and retain only necessary and relevant data. Avoid excessive collection or storage of personal information that isn't essential for business purposes.
5. **Ethical Use of Data:** Use data ethically and responsibly, avoiding discriminatory practices, biases, or unethical exploitation of personal information.
6. **Respecting Individuals' Rights:** Respect individuals' rights regarding their data, including providing access, facilitating corrections, honoring deletion requests, and ensuring data portability.

Emerging Technologies in Data Privacy

1. **Homomorphic Encryption:** Allowing computations to be performed on encrypted data without decrypting it, preserving data privacy during computations.
2. **Zero Trust Architecture:** Operating on the principle of "never trust, always verify," where access controls are continuously evaluated based on various factors like device health, user behavior, etc.
3. **Privacy-Preserving Technologies:** Differential privacy, federated learning, and secure multi-party computation, enabling data analysis while preserving individual privacy.

**********END OF MODULE 4**********