USN □□□□□□□□□□ 21CS71

Seventh Semester B.E./B.Tech Degree Examination, Dec.2024/Jan.2025
Big Data Analytics

Time: 3 hrs. Max. Marks: 100

Note: ***Answer any FIVE full questions, choosing ONE full question from each module.***

Important Note : 1. On completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages.
2. Any revealing of identification, appeal to evaluator and /or equations written eg, 42+8 = 50, will be treated as malpractice.

Module-1

1 a. Discuss the evolution of Big Data. **(06 Marks)**
b. Explain the characteristics of Big Data. **(04 Marks)**
c. Explain Data Architecture Design, with a neat diagram. **(10 Marks)**

OR

2 a. Explain Analytics Scalability to Big Data and Massive parallel processing platforms. **(12 Marks)**
b. Explain Big Data Analytics applications with one case study. **(08 Marks)**

Module-2

3 a. List and explain the core components of Hadoop. **(10 Marks)**
b. Explain Hadoop Distributed File System. **(10 Marks)**

OR

4 a. Define MapRedeuce Frame work and its functions. **(06 Marks)**
b. Explain steps on the request to MapReduce and the types of process in MapReduce. **(10 Marks)**
c. Explain in brief on Flume Hadoop Tool. **(04 Marks)**

Module-3

5 a. Explain about No SQL datastore and its characteristics. **(10 Marks)**
b. Describe the principle of working of the CAP theorem. **(10 Marks)**

OR

6 a. Demonstrate the working of key- value store with an example. **(10 Marks)**
b. Describe the features of MongoDB, and its industrial application. **(10 Marks)**

Module-4

7 a. Explain the process in MapReduce when client submitting a job, with a neat diagram. **(10 Marks)**
b. Explain Hive Integration and workflow steps involved with a diagram. **(10 Marks)**

OR

8 a. Using HiveQL for the following :
i) Create a table with partition
ii) Add, rename and drop a partition to a table. (10 Marks)

b. What is PIG in BigData? Explain the feature of PIG. (10 Marks)

Module-5

9 a. Explain linear and non-linear relationship with essential graphs in machine learning. (10 Marks)

b. Write the block diagram of text mining process and explain its phases. (10 Marks)

OR

10 a. With a neat diagram, write the steps in K-means clustering. (10 Marks)

b. Explain the purpose of web usage analytics and the significance of web graphs. (10 Marks)

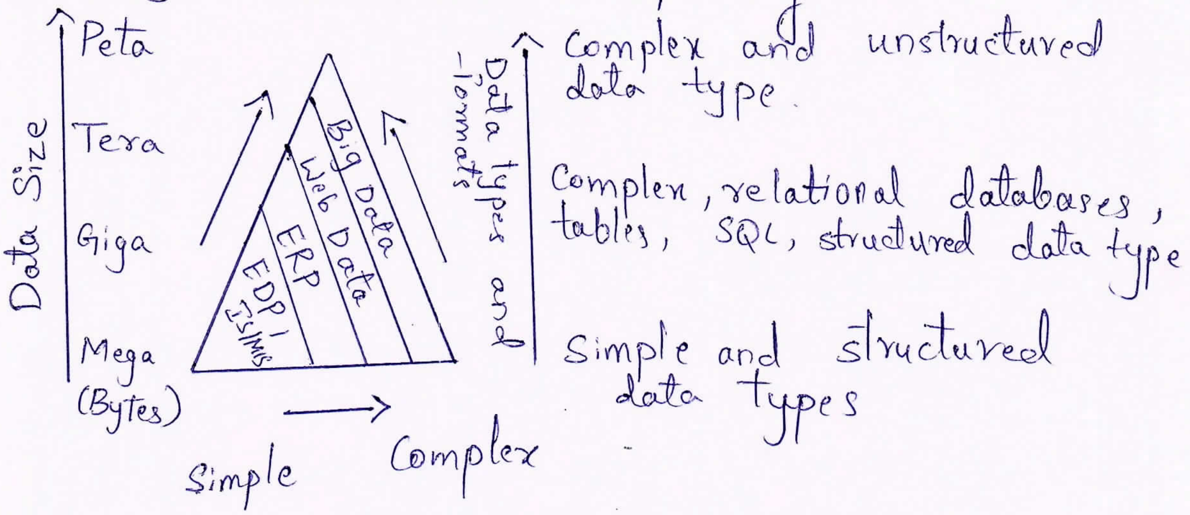
* * * * *

Dec - Jan 2024-25

Big Data Analytics (21CS71)

Module 1.

1. a. Discuss the evolution of Big Data


Data Size
Peta
Tera
Giga
Mega
(Bytes)
Big Data
Web Data
ERP
EDP / Islms
Simple
Complex
Data types and formats
Complex and unstructured data type.
Complex, relational databases, tables, SQL, structured data type
Simple and structured data types

Raise in the technology has led to production and storage of voluminous amounts of data. Earlier megabytes were used but nowadays petabytes are used for processing, analysis, discovering new facts and generation new knowledge. Conventional systems for storage, processing, analysis pose challanges in large growth in volume of data. Faster generation of data need quickly processing, analyzing and usage The above diagram shows data usage and growth. As size and complexity increase the proportion of unstructured data type also increase.

An example of Traditional tool for structured data storage and querying in RDBMS

Volume, velocity and variety (3Vs) of data need the usage of number of programs and tools for analyzing and processing at a very high speed.

b Explain the characteristics of Big Data

The characteristics of Big Data are as follows:- i) Volume
ii) Velocity
iii) Variety
iv) Veracity

i) Volume :- The phrase 'Big Data' contains the term Big, which is related to size of data, size defines the amount or quantity of data, which is generated from applications

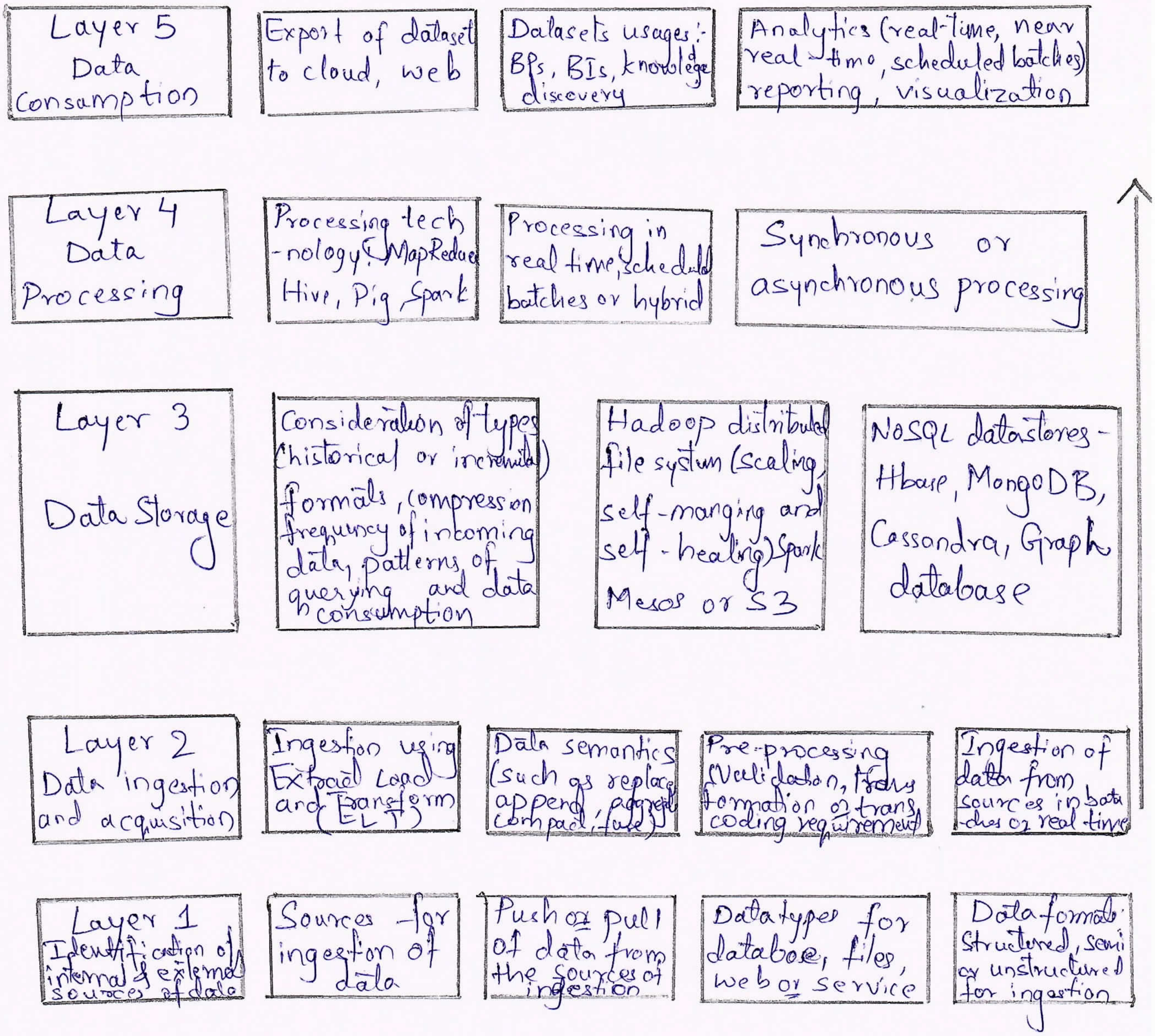
ii) Velocity :- It refers to the speed of generation of data. It is a measure of how fast the data generates and processes. To meet the demands and the challenges of processing Big Data, the velocity of generation of data plays a crucial role.

iii) Variety :- Big data comprises a variety of data. Data is generated from multiple sources in a system. This introdues variety in data and therefore introduces 'complexity'. The variety is due to availability of large number of heterogeneous platforms in industry This means that the type of which Big Data belongs to it also an important characteristic that needs to known for proper processing of data.

iv) Veracity :- It is also considered to take into account the quality of data captured which can vary gently, affecting its accurate analysis.

The 4Vs (i.e, volume, velocity, variety and veracity) data needs for mining, discovering patterns, business intelligence, AI, Machine Learning and data visualization tools.

c. Explain Data Architecture, with a neat diagram



Layer 5 Data Consumption	Export of dataset to cloud, web	Datasets usages:- BPs, BIs, knowledge discovery	Analytics (real-time, near real-time, scheduled batches) reporting, visualization	
Layer 4 Data Processing	Processing technology: MapReduce Hive, Pig, Spark	Processing in real time, scheduled batches or hybrid	Synchronous or asynchronous processing	
Layer 3 Data Storage	Consideration of types (historical or incremental) formats, compression frequency of incoming data, patterns of querying and data consumption	Hadoop distributed file system (scaling, self-manging and self-healing) Spark Mesos or S3	NoSQL datastores - Hbase, MongoDB, Cassandra, Graph database	
Layer 2 Data ingestion and acquisition	Ingestion using Extract Load and Transform (ELT)	Data semantics (such as replace append, aggregate compact, fuse)	Pre-processing (Validation, transformation or transcoding requirement)	Ingestion of data from sources in batches or real time
Layer 1 Identification of internal & external sources of data	Sources for ingestion of data	Push or pull of data from the sources of ingestion	Datatypes for database, files, web or service	Data formats: Structured, semi or unstructured for ingestion

L1 considers the following aspects in design:—

- Amount of data needed at ingestion layer.
- Push from L1 or pull by L2 as per the mechanism for the usages.
- Source data types: Database, files, web or service.
- Source formats i.e., semi-structured, unstructured or structured

L2 considers the following aspects:—

- Ingestion and ETL process either in real time, which means store and use the data as generated or in batches.

L3 considers the following aspects:—

- Data storage type, format, compression, incoming data frequency, querying patterns and consumption requirements for L4 or L5
- Data storage using Hadoop distributed file system

L4 considers the following aspects:—

- Data processing software such as MapReduce, Hive, Pig, Spark, Mahout, Spark Streaming
- Processing in scheduled bataches or real time or hybrid.

L5 considers the consumption of data for the following:

- Data Integration.
- Datasets usages for reporting and visualization
- Analytics BPs, BIs, knowledge discovery.
- Export of datasets to cloud, web or other system.

2 a. Explain Analatics Scalability to Big Data and Massive Parallel processing platforms

Scalability enables increase or decrease in capacity of data storage, processing and analytics it is the capability of a system to handle workload as per the magnitude of the work.

Vertical Scalability - Means scaling up the given systems resources and increasing the system analytics, reporting and visualization capabilities.

Scaling Up - Designing the algorithm according to the architecture that uses resources efficiently.

Horizontal Scalability - It refers to the number of system working in coherence and scaling out the workload.

Processing different datasets of large dataset deploys horizontal Scalibility. Scaling out means using more resources and distributing the processing and storage tasks in parallel.

Massievly Parallel Processing Platforms :-

Many programs are so large and complex that it is impractical or impossible to execute them on a single computer system, especially in limited computer memory.

Parallelization of tasks can be done at several levels : i) distributing separate tasks on separate threads on the same CPU

ii) distributing separate tasks onto separate CPU on the same computer.

iii) distributing separate tasks onto separate computer.

* Distributed Computing Model :—

This model uses cloud, grid or clusters which process and analyze big and large datasets on distributed computing nodes connected by high speed networks.

Big data processing uses a parallel, scalable and no-sharing program model, such as MapReduce for computation on it.

* Volunteer Computing - It is distributed computing paradigm which uses computing resources of the volunteers. Volunteers are organisation or members who own personal computers.

2 b. Explain Big Data Analytics applications with one Case Study.

Big Data Analytics in health care use the following data sources.

i) clinical records
ii) pharmacy records
iii) electronic medical records
iv) diagnosis logs and notes
v) additional data such as deviations from person usual activities.

Value - based and customer - centric healthcare:- It means cost effective patient care by improving healthcare quality using latest knowledge, usages of electronic health and medical records and improving coordination.

Health Internet of Things - create unstructure data. The data enables the monitoring of the devices data for patient parameters.

* Cloud Computing :— One of the best approach for data processing is to perform parallel and distibuted computing in a cloud computing environment.

Cloud Computing features are :—

i) on - demand service

ii) resource pooling.

iii) scalability

iv) accountability

v) broad network access

* Grid and cluster computing :—

Grid computing refers to distributed computing in which a group of computers from several locations are connected with each other to achieve a common task.

Features of Grid Computing.

i) Grid computing similar to cloud computing is scalable.

ii) Grid computing also forms a distributed network for resource integration.

Cluster computing — It is group of computers connected by a network. The group works together to accomplish the same task

Prevention of fraud, waste and abuse

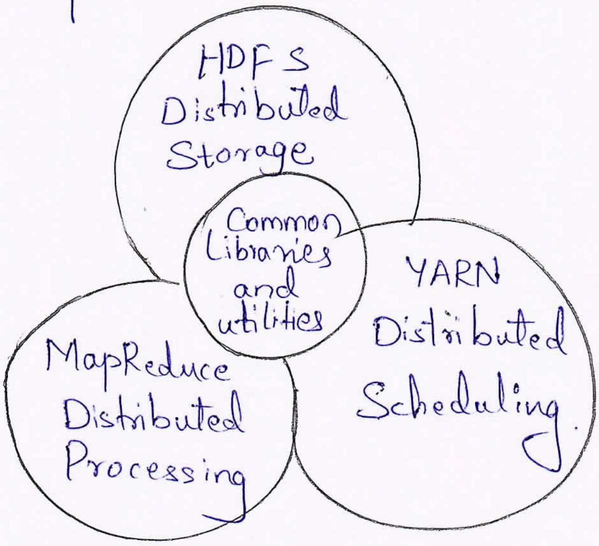
Big Data predictive analytics and help resolve excessive or duplicate claims in a systematic manner. The analytics of patient records and billing help in detecting anomalies such as overutilization of services in short intervals, different hospitals in different locations.

Patient real-time monitoring

Uses machine learning algorithms which process real-time events. They provide physicians the insights to help them make life-saving decisions and allow for effective interventions. The process automation sends the alerts to care providers and informs them instantly about changes in the condition of a patient.

Module - 2.

3 a. List and explain the core components of Hadoop.


HDFS Distributed Storage
Common Libraries and utilities
YARN Distributed Scheduling
MapReduce Distributed Processing

The Hadoop Core components of the framework are:-

1) Hadoop Common - The Common module contains the libraries and utilities that are required by the other modules of Hadoop.

Example:- Hadoop common provides various components and interface for distributed file system and general input/output.

2) Hadoop Distributed file System (HDFS) - A Java based distributed file system which can store all kinds of data on the disks at the clusters.

3) Map Reduce v1 - Software programming model in Hadoop 1 using Mapper and Reducer. The v1 processes large sets of data in parallel and in batches

3. b. Explain Hadoop Distributed File System.

The Hadoop Distributed File System (HDFS) was designed for Big Data Processing, the design assumes a large file write-once / read-many model that enables other optimization

The important aspects of HDFS are as follows :-

* The write-once / read-many design is intended to facilitate streaming reads.

* Files may be appended, but random seeks are not permitted. There is no caching of data.

* Coveraged data storage and processing happen on the same server nodes.

* Moving computation is cheaper than moving data.

* A reliable file system maintain multiple copies of data across the cluster.

* A specialized file system is used which is not designed for general use.

HDFS Components

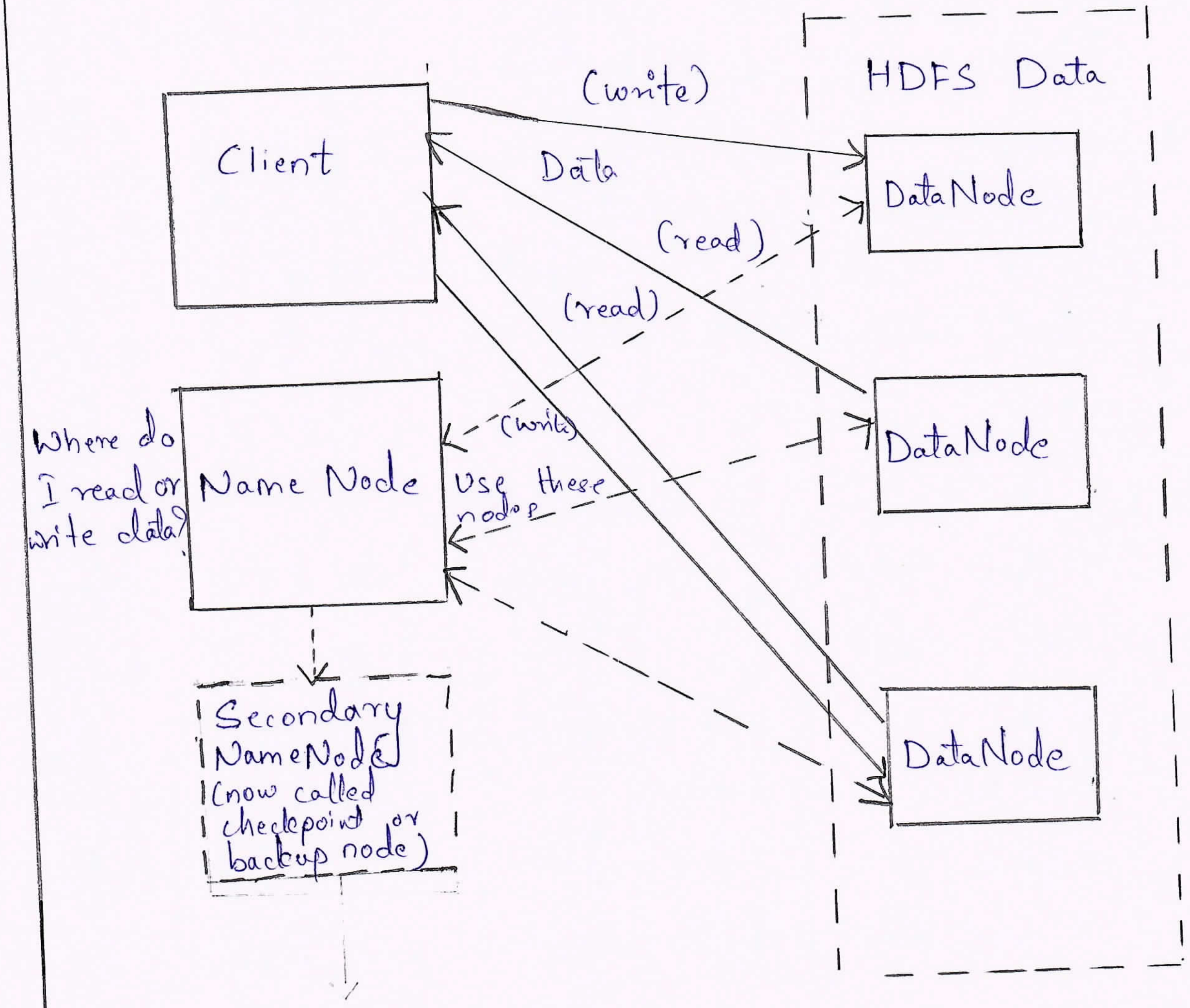
* The design of HDFS is based on two types of Nodes :- aNameNode and Multiple Datanode.

4) YARN - Software for manging resources for computing. The user application tasks or sub-tasks run in parallel at the Hadoop, uses scheduling and handles the requests for resources in distributed running of the tasks.

5) Map Reduce v2 - Hadoop 2 YARN — Based system for parallel processing of large datasets and distributed processing of application tasks.

6.) Spark - It is open source, cluster-computing framework of Apache Software Foundation Hadoop deploys data at disks.

- * Spark provisions in memory analytics, it enables OLAP and real time processing.
- * It does faster processing of Big Data.
- * Spark has been adopted by large organisations such as Amazon, eBay and Yahoo.



* For minimal Hadoop installation, there needs to be single Name Node daemon and single DataNode daemon running on atleast one machine

* File system namespace operations such as opening, closing and renaming files and directories. are all managed by Name Node.

* It determines the mapping of blocks to DataNodes and Handles Data Node Failures.

* The Slave are responsible for serving read and write request from the file system to the clients.

* The Name node manages block creation, deletion and replication

* Example - Client / Name Node / Data Node interaction is provided where client writes data, it frist communicates with the NameNode and request to create file.

* Datanode will store data
* Data blocks are replicated after they are written to the assigned node.
* After Data Node acknowledges that the file block replication is complete the client closes the file and informs the NameNode that operation is complete.

4 a. Define MapReduce Frame work and its functions

* MapReduce provides two important functions. The distribution of job based on client application task or user query
* The processing tasks are submitted to the Hadoop
* Daemon refers to highly dedicated program that runs in the background in a system.
  Example - MapReduce in Hadoop system.
* MapReduce runs as per assigned Job by JobTracker.

5 a. Explain about No SQL datastore and its characteristics

* A new category of data stores is NoSQL data stores.

* NoSQL is an altogether new approach of thinking about databases such as schema flexibility, simple relationships, dynamic schemas, auto sharding, replication.

* Issues with NoSQL data stores are lack of standardization

* No SQL is a class of non-relational data storage system with flexible data model.

The characteristics of NoSQL data store are as follows:=

* It is a class of non-relational data storage with flexible data model.

* Example of NoSQL data - architecture pattern of datasets are key-value pairs, name/value pairs, Coloumn family Big-data store

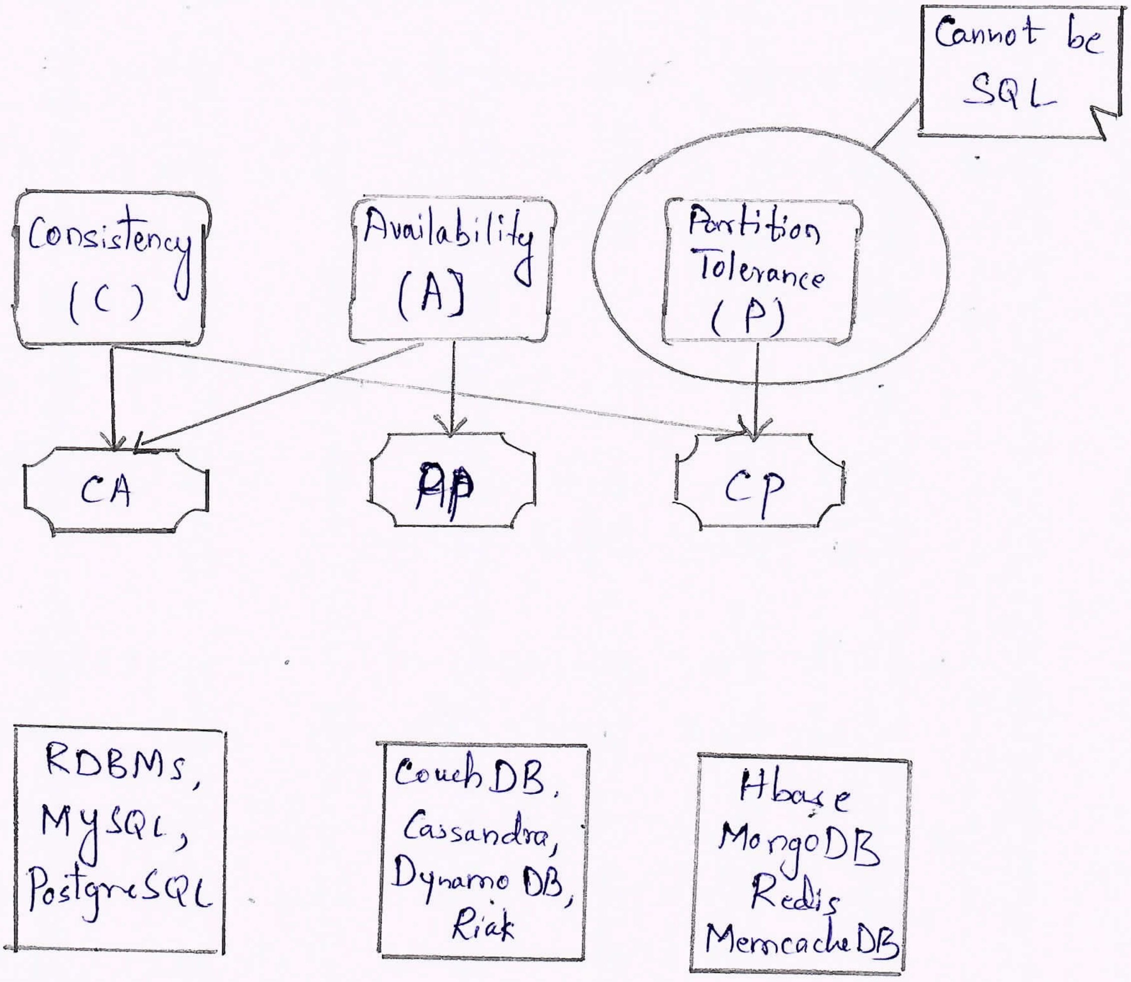
* NoSQL not necessarily has fixed schema such as table, do not use the concept of Joins

* Data written at one node can be replicated to multiple node.

* Example on NoSQL Data store are :-

1. Apache's HBase.
2. Apache's MongoDB.
3. Apaches Cassandra.

5. b. Descibe the principle of working of the CAP theorem.



- The Input data is in the form of an HDFS file. The output of the task also gets stored in HDFS

- Compute nodes and storage nodes are the same at a cluster.

- They are running on same set of nodes.

- High effeciency due to reduction in network traffic accross the cluster.

- A user application specifies locations of the input / output data and translates into map and reduc functions

- A job does implementation of appropriate interface and / or abstract - classes.

- The Hadoop client then submits the job and configuration to the Job Tracker, which then assumes the responsiblity of distributing the software configuration.

- The diagram shows MapReduce process when a client submits a job and the succeding action by the Job Tracker and Task Tracker.

- Map Reduce consists of a single master Job Tracker and one slave Task Tracker per cluster node.

* The master is responsible for scheduling the component tasks in a job onto slaves.

* The data for a MapReduce task is initally at input files.

* The input files typically reside in the HDFS

* The files may be line-based log files, binary format.

* The MapReduce framework operates entirely on key, value pairs

* The framework views the input to the task as a set of pairs and produce a set of pair as the output of the task.

* Among C, A and P, two are at least present for the application / service / process.

* Consistency means all copies have the same value like in traditional DBs

* Availability means atleast one copy is available in case a partition becomes inactive or fails.

* Brew's CAP theorem demonstrates that any distributed system cannot guarantee C, A and P together

  - Consistency - All nodes observes the same data at same time

  - Availability - Each request receives a response on success / failure.

  - Partition Tolerance - The system continues to operate as a whole even in case of message loss, node failure or node not reactable.

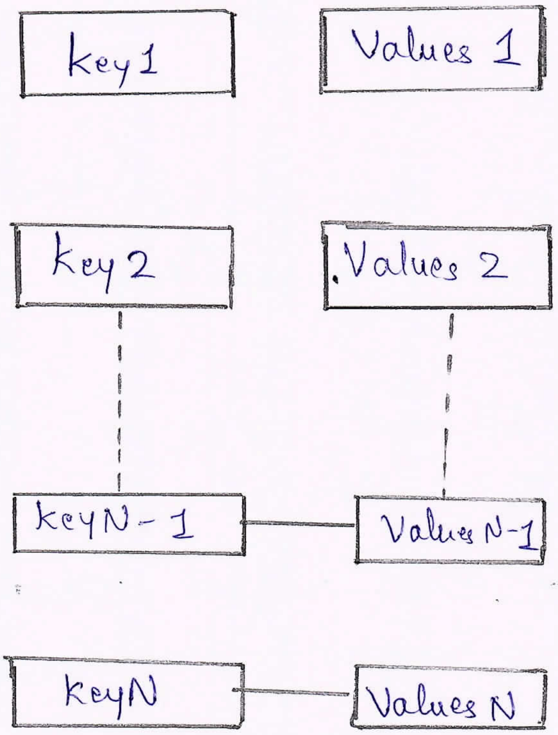
* The CAP theorem implies for a network partition system, the choice of consitency and availability are mutually exclusive.

* CA means consitency and availability

* AP means availability and partition tolerance * CP means consitency and partition.

6 a. Demonstrate the working of key-value store with an example.

The simplest way to implement a schema-less data store is to use key-value pairs



Key	Value.
"Ashish"	"Category: Student; Class: B.Tech; Semester: VII; Branch: Engineering; Mobile: 9434123456"
"Mayuri"	"Category: student; class: M.tech; Mobile: 8888823456"

Number of key-values pair, N can be a very large number.

6. b Describe the features of MongoDB, and its industrial application

MongoDB is an open source DBMS. MongoDB programs create and manage database.

Features of MongoDB are as follows:-

* MongoDB data store is a physical container for collections. Each DB gets its own set of files on the file system.

* Collection stores a number of MongoDB documents. of the collection are schema-less. Thus it is possible to store documents of varying structures in a collection.

* Document model is well defined. Structure of document is clear, Document is the unit of storing data in a MongoDB database. Documents are analogous to the records of RDBMS table, insert, delete and update operations can be performed on collection.

* MongoDB is a document data store in which one collection holds different documents. storage is one in JSON-style documents.

* Storing of data is flexible and data consists of JSON-like documents, fields can vary from document to document & data structure can be changed over-time

* Data store characteristics are high performance, scalability and flexibility.
* Data retrieval is fast in key-value pairs data store.
* A simple string called key maps to large data string or BLOB.
* Key value store access to primary key for accessing the values.

Advantages of key-value store.

* A query just requests the value and returns the value as a single item. Values can be of any data type.
* Key-value store is eventually consistent
* Key-value data store may be heirarchial
* Returned values on queries used to convert into lists, table-coloumns, data-frame fields and coloumns

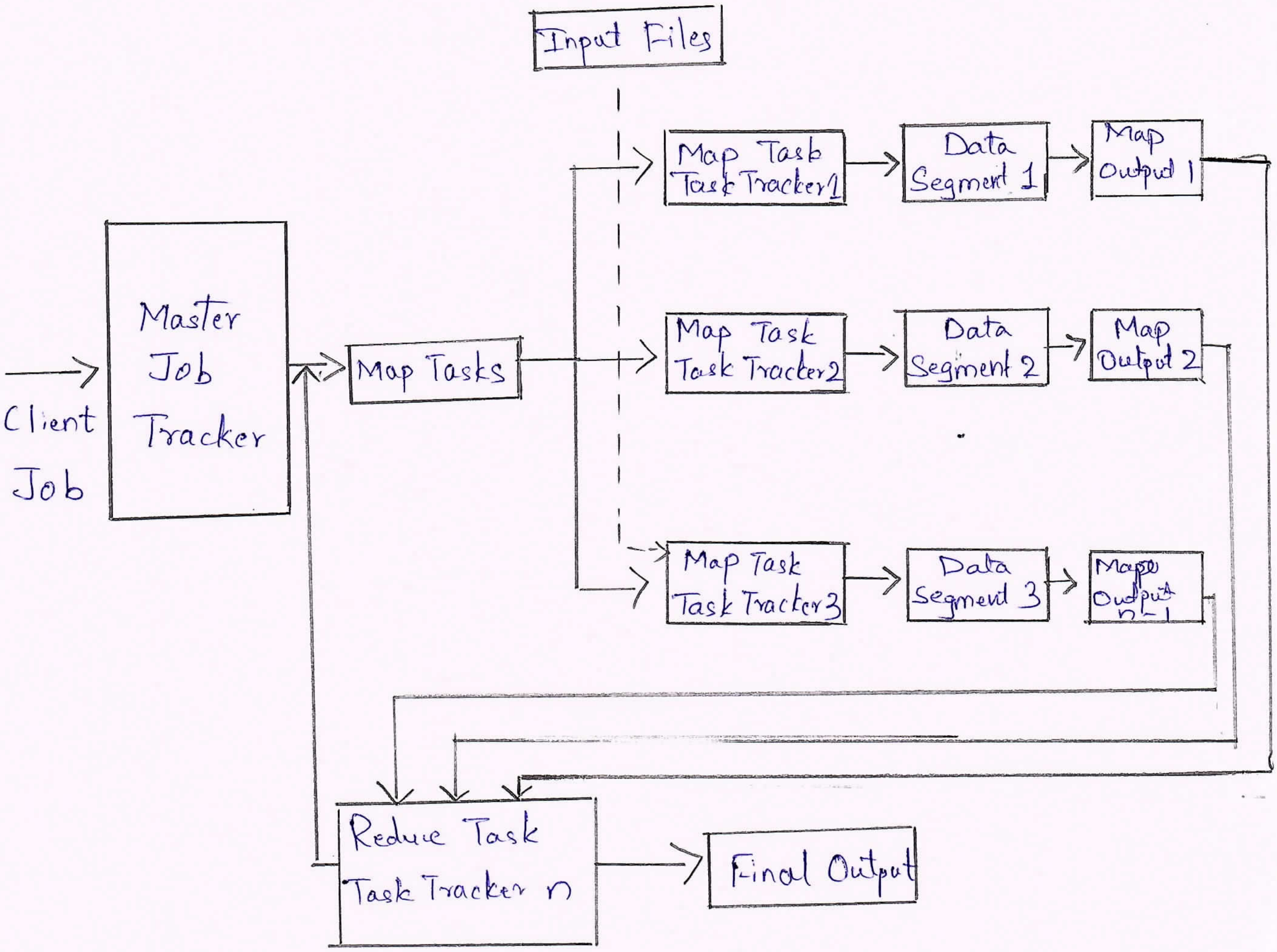
Limitations of key value store architectural pattern :-

* No indexes are maintained on values, ~~thus~~ a subset of values is not ~~sh~~ searchable.
* Maintaining unique value as keys may become more difficult when the volume of data increases.

7 a) Explain the process in MapReduce when client submitting a job, with a neat diagram

* ~~b. Explain steps on the request to MapReduce and types of process in MapReduce.~~

* Big data processing employs the MapReduce programming model.
* A job means a MapRedue program
* Each job consists of several smaller units called MapRedue tasks.
* A software execution framework in MapReduce programming defines the parallel tasks.



* A client node submits a request of an application to the Job Tracker.

* The job execution is controlled by two types of processes in MapReduce.

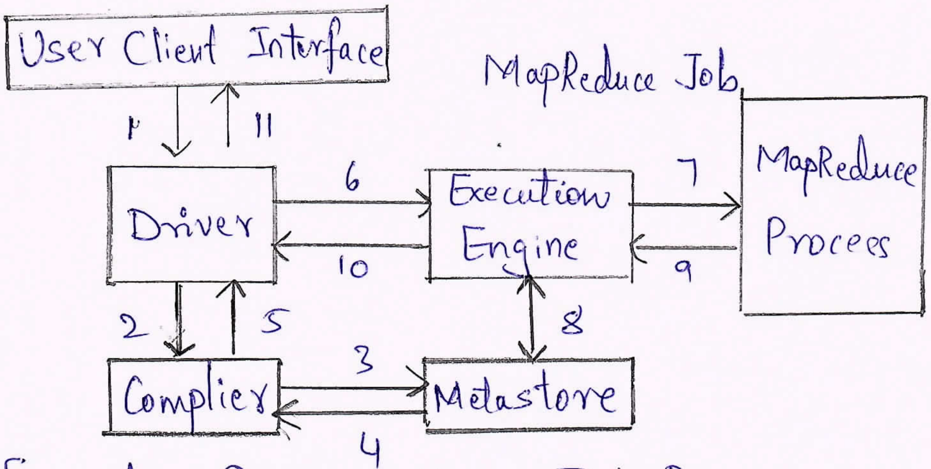
1. The Mapper deploys map tasks on the slots. Map tasks assign to those nodes where the data for application is stored.

2. The Hadoop system sends the Map and Reduce job to the appropriate servers in the cluster.

* Job execution is controlled by two types of proces in MapReduce.

1. A single master process called Job Tracker
2. The second is number of subordinate processes called Task Taskers.

7 b. Explain Hive Integration and work flow steps, involved with a diagram.



1. Execute Query - Hive Interface sends a query to Database Driver to execute the query.

2. Get Plan - Driver sends the query compiler that parses the query to check the syntax and query plan or requirement of the query

3. Get Metadata - Compiler sends metadata request to Metaslore

4. Send Metadata - Metastore sends metadata as a response to compiler.

5. Sends Plan - Compiler checks the requirement and resends the plan to driver. The parsing and compiling of the query is complete at this place

6. Execute Plan - Driver sends the execute plan to execution engine.

7. Execute Job - The Execution engine sends the job to Job Tracker, which is in Name node and assigns this job to Task Tracker, which is in data node then query executes the Job.

8. Metadata Operations - Execution engine can execute the met data operations with Metastore.

9. Fetch Result - Execution engine receives the results from Data Nodes.

10. Send Results - Execution engine sends the result to Driver.

11. Send Results - Driver sends the results to Hive Interface

- Storing of documents on disk is in BSON serialization format.
- Querying, indexing, and real time aggregation allows accessing and analyzing the data efficiently.
- No complex joins.
- Distributed DB makes availability high, and provides horizontal scalability.

The typical MongoDB application are content management and delivery systems, mobile application user data management, gaming, e-commerce, analytics archiving and logging.

8 a. Using HiveQL for the following.

i) Create a table with partition.

Hive organizes tables into partitions. Table partitioning refers to dividing the table data into some parts based on the values of particular set of coloumns. This is because SELECT is then from the smaller number of coloumn fields.

Table Partitioning

Create a table with Partition using commond:

```
CREATE [EXTERNAL] TABLE <table name>(<coloumn
name 1> <data type 1>, ........)

PARTITIONED BY (< coloumn name n> <datatype n>
[COMMENT <coloumn comment>], ......);
```

ii) Add, rename and drop a partition to a table

Add

Add a Partition in the existing Table using the following command:

```
ALTER TABLE <table name> ADD [IF NOT EXISTS] PARTITION
partition_spec [LOCATION 'location 1'] partition_spec
[LOCATION 'location 2'] .....;

partition_spec :(p_column = p-col_value, p-coloumn =
    p-col-value...).
```

Rename

Rename a Partition in the existing Table using the following command:

```
ALTER TABLE <table name> PARTITION partition_spec
RENAME TO PARTITION partition_spec;
```

drop

Drop a Partition in the existing Table using the following command :-

```
ALTER TABLE <table name> DROP [IF EXISTS]
PARTITION partition_spec, PARTITION
partition_spec;
```

8 b. What is PIG in Big Data? Explain the feature of PIG.

Apache developed Pig, which:

- Is an abstraction over MapReduce.
- Is an execution framework for parallel processing.
- Reduces the complexities of writing a MapReduce program.
- Is mostly used in HDFS environment.

Features of PIG are as follows

* Apache PIG helps programmers write complex data transformation using scripts. Pig Latin Language is very similar to SQL and posses a rich set of built in operators such as group, join, filter, limit, order by, parallel, sort and split.

* Create user defined functions to write custom functions which are not available in Pig.

* Process any kind of data, structured, semi-structured or unstructured data coming from various sources.

* Reduces the length of codes using multi-query approach.

* Handles incosistent schema in case of unstructured data as well.

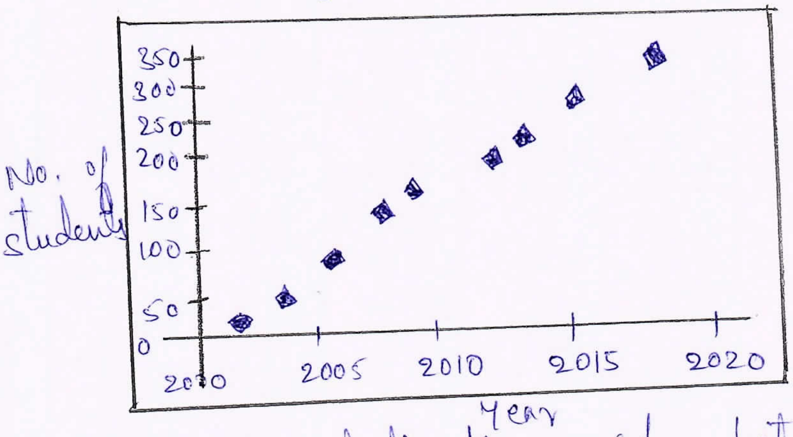
* Extracts the data, performs operations on that data and dumps the data in the required format in HDFS.

* Performs automatic optimization of tasks before execution

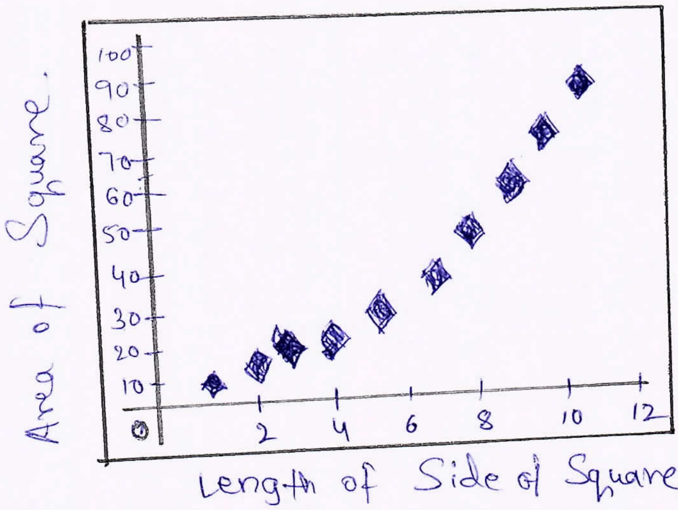
* Programmers and developers can concentrate on the whole operation without a need to create mapper and reduce tasks separately

Module - 5.

9 a. Explain linear and non-linear relationship with essential graphs in machine learning.

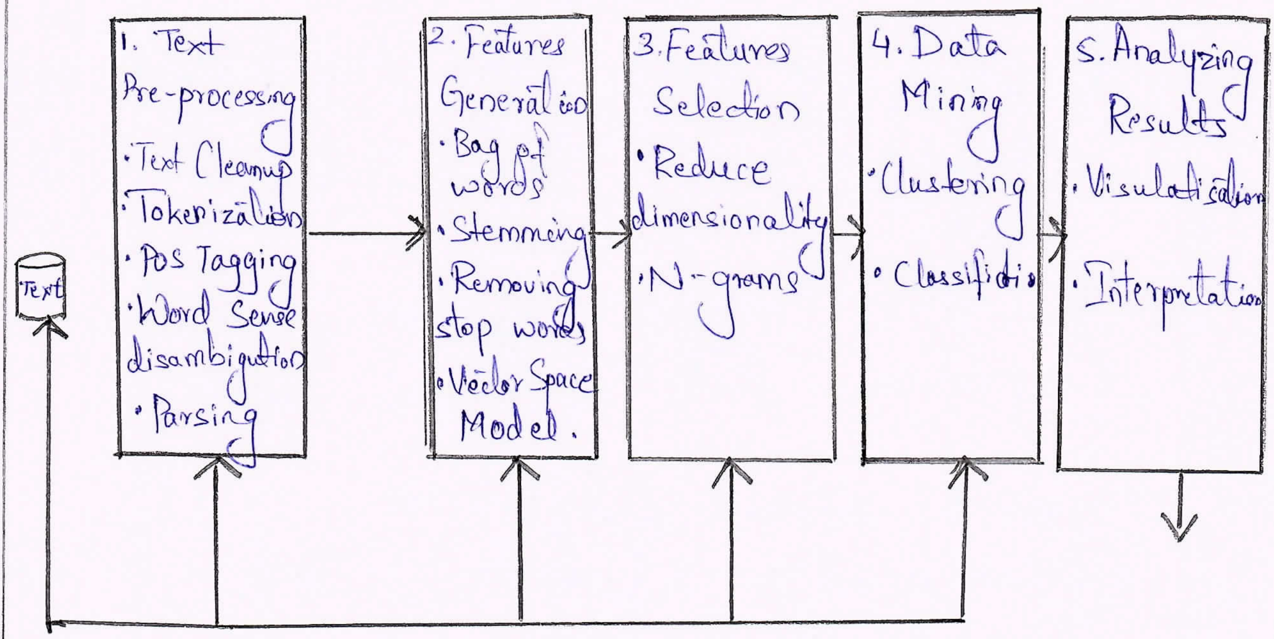


- A linear relationship exists between two variables, say x & y when a straight line can fit on a graph with at least some reasonable degree of accuracy
- The above figure shows a scatter plot which fits a linear relationship between the number of students opting for computer courses in years between 2000 and 2017
- A linear relationship can be positive or negative.
- A positive relationship implies if one variable increases in value, the other also increases in value



- A non linear relationship is said to exist between two quantitative variables where a curve can be used to fit data points.
- The side of a square and its area are not linear
- They have Qudratic relationship
- If the side of square doubles, then its area increases four times

8 b. Write the block diagram of text mining process and explain its phases.



Phase 1 :- Text pre-processing – enables Syntatic / Semantic text analysis

* Text cleanup – It is a process of removing unnecessary or unwanted information.

* Tokenization – It is a process of splitting the cleanup text into tokens using white spaces and punctuation marks as delimiters

* POS & tagging – It is a method that attempts labeling of each token.

* Word sense disambiguation – which identifies the sense of a word used in a sentence

* Parsing – It is a method which generates a parse-tree for each sentence.

Phase 2: Features Generation - It is a process which first defines features.

* Bag of words - Order of word is not that important for certain applications.
* Stemming - identifies a word by its root.
* Removing stop words from the feature space - they are common words, unlikely to help text mining.
* Vector Spae Model - Is an algebraic model for representing text documents as vector of identifiers.

Phase 3: Features Selection -

* Dimensionality reduction - Feature selection is one of the methods of division
* N - gram evaluation - finding the number of consecutive words of intrest and extract them.
* Noise detection and evaluation of outliners - The identification of unusual or suspicious items, events or observations from the data set.

Phase 4: Data mining techniques

* It enable insights about the structured database that resulted from previous phases.

  Examples of techniques are :-

  1. Unsupervised learning
  2. Supervised learning
  3. Identifying evolutionary patterns

Phase 5: Analysis results

i) Evaluate the outcome of complete process

ii) Visualization