7] a) List and discuss the classification of digital data.

Classification of digital data:-

→ Structured data
→ unstructured data
→ Semi structured data.

Structured data:-

Structured data generally resides in a relational database, and as a result, it is sometimes called as relational data. This type of data can be easily mapped into predesigned fields. For example, a database designer may set up fields for phone numbers, zip codes and credit card numbers that accept a certain number of digits. Structured data has been or can be placed in fields like these.

Unstructured data:-

Unstructured data is information that either doesnot have a predefined data model and/or does not fit well into relational database. Unstructured information is typically text heavy, but may contain data such as dates, numbers and facts, as well. Following are the tools used to manage unstructured data.

→ Bigdata tools:- Software like Hadoop can process stores of both unstructured and structured data that are extremely large.
→ Business intelligence software: It helps companies make sense of their structured and unstructured data for the purpose of making better business decisions.
→ Data integration tools: This tools combine data from disparate sources so that they can be analysed from a single application.
→ Search and indexing tools: These retrieve information from unstructured data files such as documents, web pages and photos.

Semistructured data:-

Semistructured data is information that doesn't reside in a relational database but that does have some organizational properties and make it easier to analyse. Examples of semistructured data include XML documents and NoSQL databases.

b) What is bigdata? Discuss the challenges in Bigdata.

Bigdata is the term for collection of data sets so large and complex that it becomes difficult to process using on-hand database system tools or traditional data processing applications.

Challenges in Bigdata:-

1. Storing exponentially growing huge datasets.
   → Data generated in past 2 years is more than the previous history in total. By 2020, total digital data will grow to 44 zettaBytes approximately. By 2020, about 1.7mb of new information will be created every second for every person.

2. processing data having complex structure

Structured	Semi-Structured	Unstructured
→ Organized data format	→ Partial Organized data	→ Unorganized data.
→ Data schema is fixed	→ Lacks formal structure of data model.	→ Unknown schema
→ Ex: RDBMs data.	→ Ex: XML etc	→ Ex: multimedia etc.

3. Processing data faster:-
   → The data is growing at much faster rate than that of disk read/write speed.
   → Bringing huge amount of data to computation unit becomes a bottle neck.

a) Discuss credit risk management.

Credit risk analysis focus on past credit behaviours to predict the likelihood that a borrower will default on any type of debt by failing to make payments which they obligated to do.

Credit risk principle: drive the business using the optimal balance of risk and reward

Traditionally, credit risk management was rooted in philosophy of minimizing losses. Now, a days, the credit risk managers and many other leaders have understood that there is acceptable level of risks that can be taken which increases profit.

Credit risk professionals are stake holders who address all aspects of business from finding new and profitable customers to maintaining and growing relationship with existing customers.

Four critical parts of credit risk frame.

1 Customer aquisition:-

From customer aquisition perspective, credit risk managers decide whether to extend credit and how much. Lacking any previous experience with the prospect, they depend heavily on third party credit reports and scores and may assist marketting organizations.

2. Account management:-

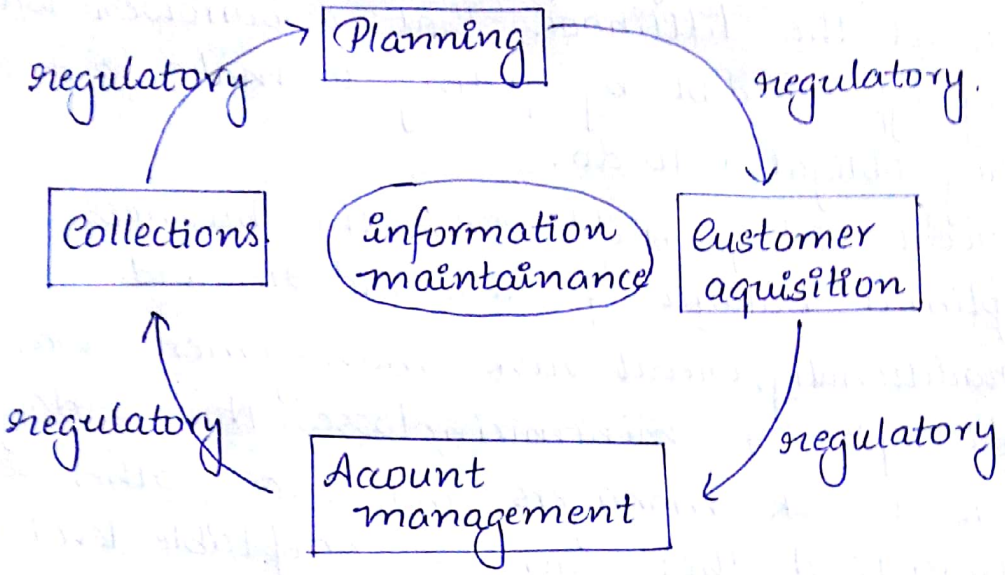
This requires periodic customer risk assessment that influence key decisions on credit line increases and decreases.

3. Collections:

Continuous monitoring of an existing profile can help credit risk managers to expect their losses and manage thiu collections better.

4 Planning:-

It is handled through bigdata Analytics.

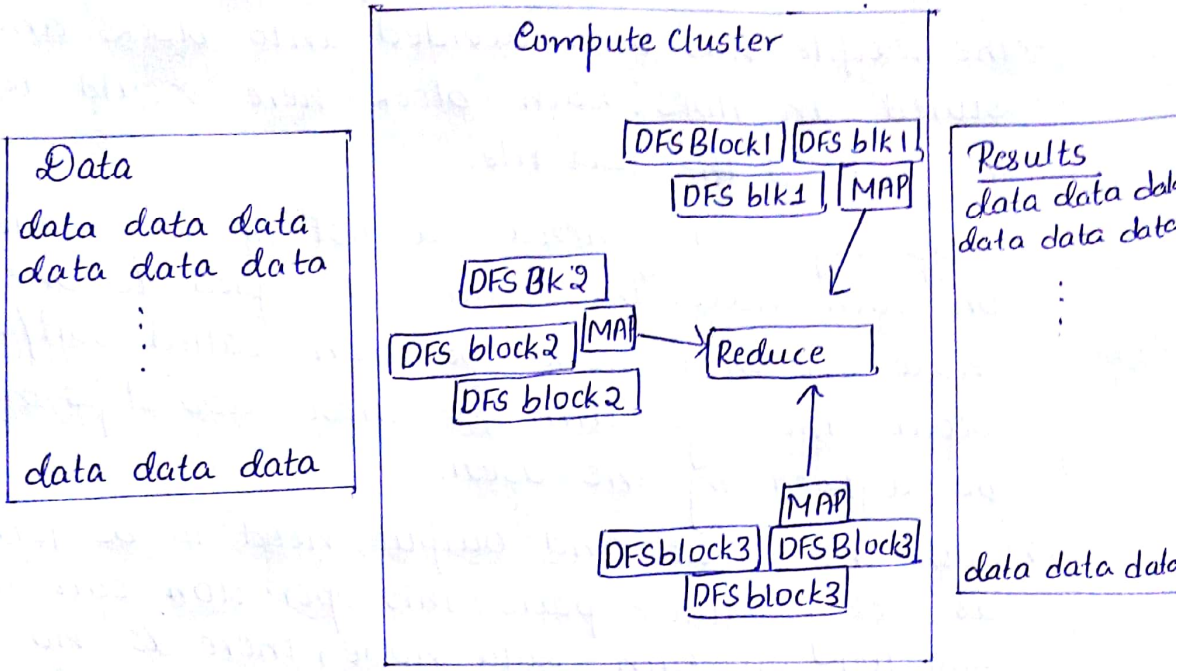


b) Explain the features and advantages of hadoop.

Two critical components of hadoop are:-

1. HDFS (Hadoop distributed file system):-
HDFS is the storage system for a Hadoop cluster. When data lands in the cluster, HDFS breaks it into pieces and distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of complete data set, and each piece of data is replicated on more than one server.

2. Map Reduce:- Because hadoop stores the entire data set in small pieces across a collection of servers, analytical jobs can be distributed, in parallel to each of the servers storing part of the data. Each server evaluates the questions against its local fragment simultaneously and reports its results back for collation into a comprahensive answer. Mapreduce is the agent that distributes the work and collects the results.



Advantages of Hadoop:-

→ Hadoop can manage large amount of complex data.
→ Hadoop processes variety and complex datas in a very small fraction of a second.
→ Hadoop doesnot require expensive, highly reliable hardware to run on.
→ Hadoop delivers a high throughput of data
→ Hadoop provides security to data by not allowing multiple writers and arbitrary file modifiers.

---

a) Write a neat diagram and explain processing data with hadoop.

Data processing in hadoop is done through Map reduce. We just need to include two functions. map() and reduce(). Lets take an example of processing large block of file. map() should be written inside the class that extends MapperClass and reduce should be written inside the class that extends Reducer class.

→ The textfile has been divided into blocks and stored in HDFS. Each block here would represent a part of the text file.

→ Map step will generate a list of key-value pairs on each node. These are all copied to one single node. On that node operation called sort/merge occurs. The algorithm for these two steps need to be defined by the user.

→ Both the input and output need to be formatted as <key, value> pairs. This operation can run in parallel on each data node, there is no interdependency in the inputs and outputs. Data is sorted by key. Key-Value pairs with same key are merged.

→ reduce() will run on each pair generated by sort/merge step. The reduce function will combine all the values for the samekey in some way.

→ The map method takes a key and a value. It processes the input and writes the output to a context object. The context object stores the output and is accessed by the rest of MapReduce system.

→ We should write a driver class whose main() method will point to our Mapper and Reducer.

→ The main method will take 2 string inputs, i.e. Input text file & Output filepath.

Running a MapReduce object.

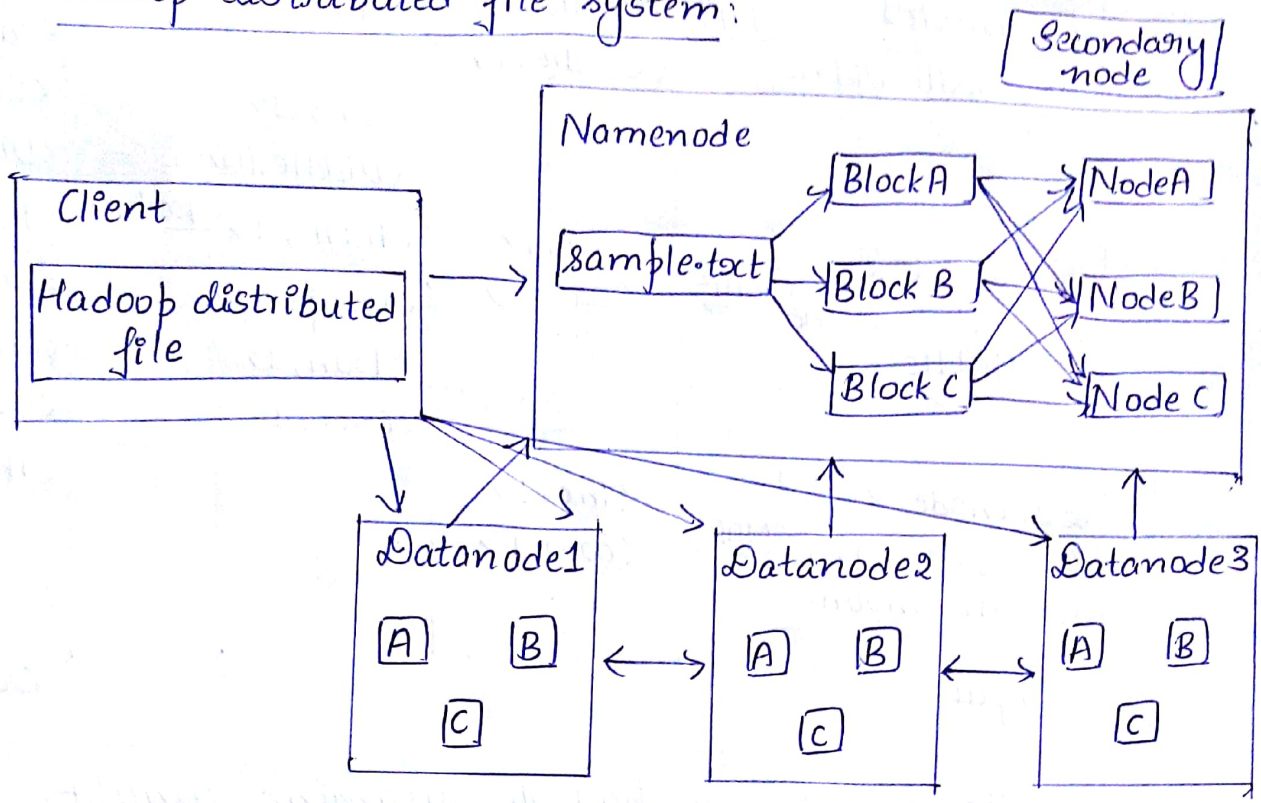
→ Build the JAR file containing all the code.

→ Include all Hadoop JARs as dependencies.

→ Run the job at the command line.

4 a) With a neat diagram explain the Hadoop distributed file system.

Hadoop distributed file system:



Hadoop distributed file system has a client, Namenode, Datanode, secondary node

Namenode:- When input data enters the cluster, it will be reaching namenode first. It is broken down into smaller blocks and then sent to the datanodes. The namenode performs operation such as read & write on the text file. It communicates or accesses datanodes using rockids. Namenode also keeps track of the datablock stored in datanode. Namenode is also called as masternode.

Datanodes: These are also called as slave nodes. It recieves the datablock from namenode for processing. Datablock send heartbeat signals to namenode to check if the connection is still there. Data nodes in a cluster communicate with each other if there is some parallel processing of data required to be done.

Secondary node:

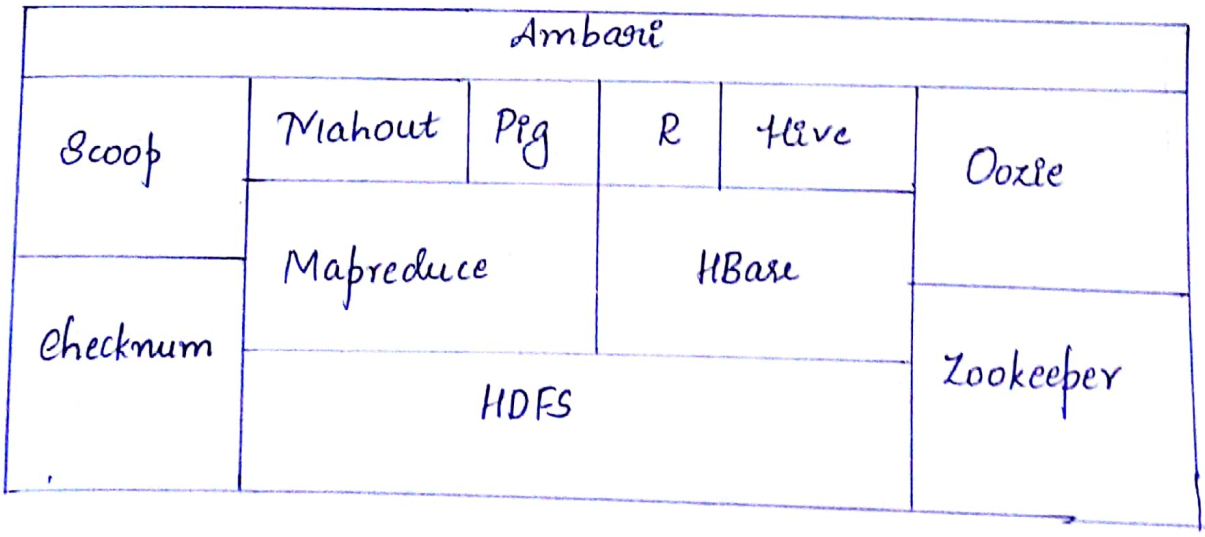
It is nothing but a namenode but namenode and secondary nodes are stored on different machines. If at all namenode fails or breaks, then secondary node takes up its place and performs namenode operation. The secondary node stores all the modified actions of name node in it.

---

b) Components of Hadoop ecosystem:-

i) HDFS
ii) HBase
iii) Mapreduce
iv) Mahout.
v) Pig
vi) R
vii) Hive
viii) Ambari
ix) Oozie
x) Zookeeper
xi) Scoop.
xii) Checknum.

Components.



i) HDFS:- It is used for storing data in its naive form.

ii) Pig:- It is used for data flow. Even if programer doesnot have any knowledge on mapreduce, pig automatically helps in doing it.

iii) Oozie:- Used for managing and performing all Apache Hadoop Systems.

iv) HBase:- Compares data with RDBMS and stores data in large tabled structures.

v) Zookeeper:- It is used for managing all distributed application system.

vi) Ambari:- It is used in web application for providing, managing data etc