Unit VI

Unit III

3 Big Data Processing

3.1 : Big Data Analytics - Ecosystem and Technologies

Q.1 Explain Big data Ecosystem with suitable diagram.

☞ [SPPU: June-22, End Sem, Marks 7]

Ans. : Big data ecosystem is the comprehension of massive functional components with various enabling tools. Capabilities of the big data ecosystem are not only about computing and storing big data, but also the advantages of its systematic platform and potentials of big data analytics.

- Fig. Q.1.1 shows emerging big data ecosystem.

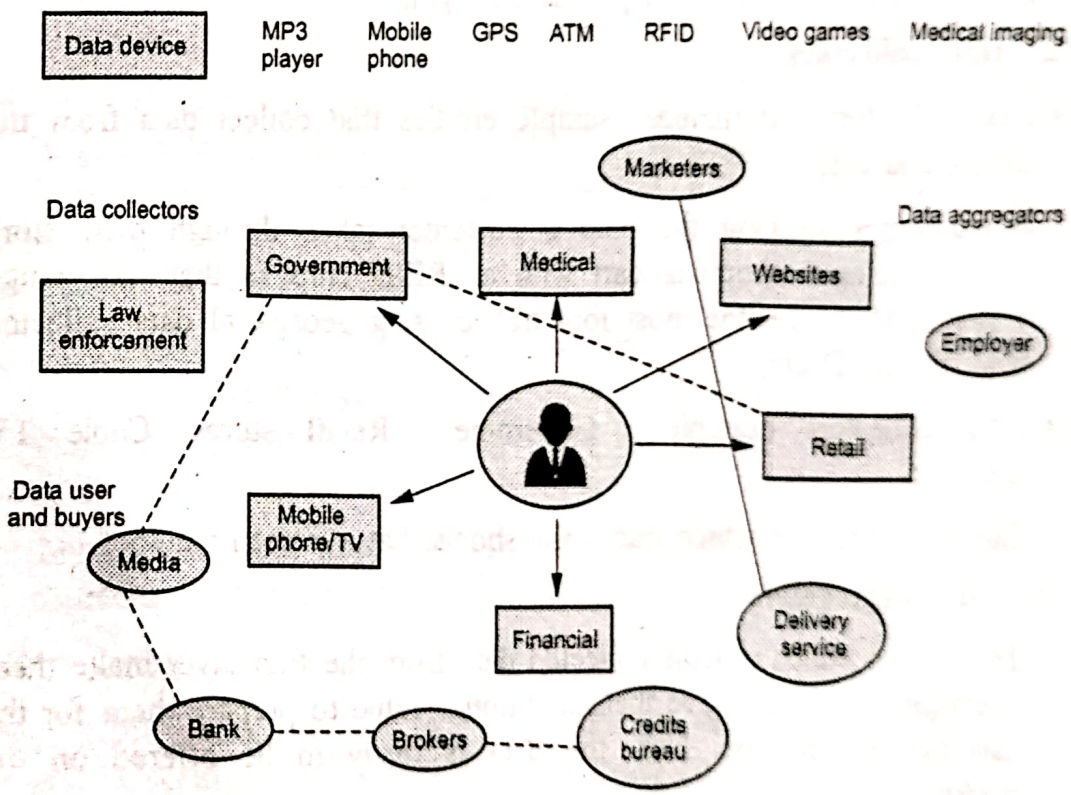


Fig. Q.1.1 Emerging big data ecosystem

1. Data device :

- Data device and sensor network gather data from various locations and continuously generate new data.
- **Example of data devices :** Playing games, smart phone and retail **shopping.**
- **Sensor data :** Growing network of sensor devices generate data based on monitoring environmental conditions, such as temperature, sound, pressure, power, water level etc.
- This data can have a wide range of practical applications if collected, aggregated, analyzed and acted upon. Examples include, water level monitoring, machine health monitoring and smart home monitoring.
- **Mobile networks :** Mobile network generates large number of data to share picture, video, audio file and text. These data is process at every mobile tower with associated demographics, location latencies etc. Sometime, mobile network may crash for large number of data movement takes place.
- Retail shopping loyalty cards records not just the amount an individual spends, but the location of stores that person visits, the kinds of product purchased, the store where goods are purchased most often and the combinations of product purchased together.

2. Data collectors :

- Data collectors : It includes sample entities that collect data from the device and user.
- Retail stores tracking the path a customer takes through their store while pushing a shopping cart with an RFID chip so they can gauge which products get the most foot traffic using geospatial data collected from the RFID chips.
- Data collectors example : Government, Retail stores, Cable TV provider.
- Cable TV provider which tracks the shows that a person watches.

3. Data aggregators :

- The entities which process collected data from the first layer make them understandable. They give them additional value to prepare them for the handing over process. Now the data is ready to be offered on the market.

- Typically, one of these data aggregations can transform and package the data as products to sell to list brokers that might want to generate marketing lists of people who may be good targets for specific ad campaigns.

4. Data users and buyers :

- These entities represent a group of the final layer from the Big Data ecosystem. This group has the final benefits from the collected and aggregated data offered by the data aggregations.
- Data users may want to track or prepare for natural disasters by identifying which areas a hurricane will affect first hand. It can be observed by tracking tweets about it or discussing it in social media.

3.2 : Introduction to Google File System

Q.2 Explain Google file system. ☞ [SPPU : June-22, End Sem, Marks 7]

Ans. : • Google file system is "a scalable distributed file system for large distributed data-intensive applications" created by Google. Initially used to store Google's search indexes and the crawling data, GFS is now mostly used to store user generated content.

- GFS was built primarily as the fundamental storage service for Google's search engine.
- GFS typically will hold a large number of huge files, each 100 MB or larger, with files that are multiple GB in size quite common. Thus, Google has chosen its file data block size to be 64 MB instead of the 4 KB in typical traditional file systems. The I/O pattern in the Google application is also special.
- Files are typically written once, and the write operations are often the appending data blocks to the end of files. Multiple appending operations might be concurrent. There will be a lot of large streaming reads and only a little random access.
- There is no data cache in GFS as large streaming reads and writes represent neither time nor space locality

GFS architecture :

- A GFS cluster consists of a single master and multiple chunk servers and is accessed by multiple clients.

Basic terms :

a. **Master** : Single, coordinates system-wide activities. Can have read-only 'Shadow' servers.

b. **Chunk** : 64 MB storage block representing a file or piece thereof.

c. **Chunkserver** : Many, stores chunks of data.

d. **Replica** : Either primary or secondary. A Chunkserver that replicates a given block.

e. **Client** : Runs tasks on data.

- It is easy to run both a chunkserver and a client on the same machine, as long as machine resources permit.
- Files are divided into fixed-size chunks. Each chunk is identified by an immutable and globally unique 64 bit chunk handle assigned by the master at the time of chunk creation.
- Chunkservers store chunks on local disks as Linux files and read or write chunk data specified by a chunk handle and byte range. For reliability, each chunk is replicated on multiple chunkservers.
- The master maintains all file system metadata. This includes the namespace, access control information, the mapping from files to chunks and the current locations of chunks.
- Clients interact with the master for metadata operations, but all data-bearing communication goes directly to the chunkservers.
- Neither the client nor the chunkserver caches file data. Fig. Q.2.1 shows GFS architecture.

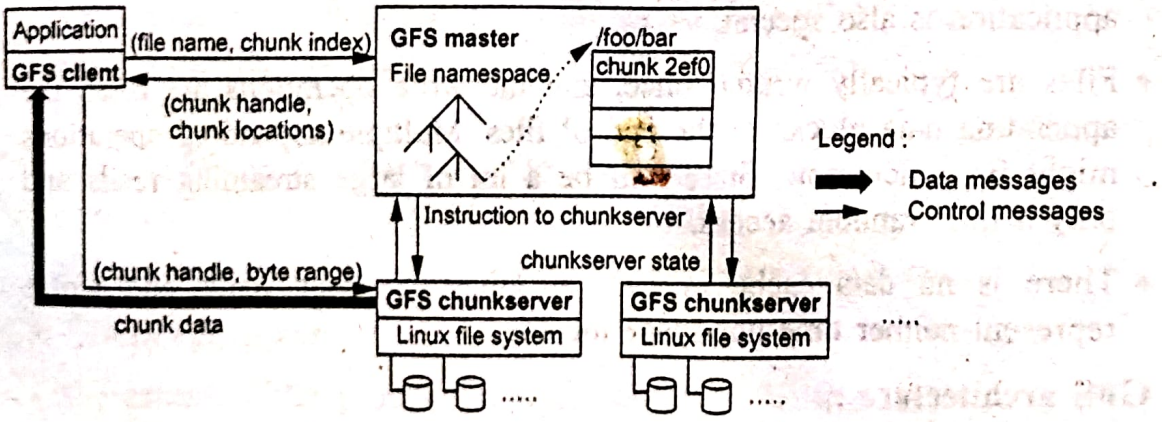


Fig. Q.2.1 GFS architecture

- Clients never read and write file data through the master. Instead, a client asks the master which chunkservers it should contact. It caches this information for a limited time and interacts with the chunkservers directly for many subsequent operations.
- First, using the fixed chunk size, the client translates the file name and byte offset specified by the application into a chunk index within the file. Then, it sends the master a request containing the file name and chunk index.
- The master replies with the corresponding chunk handle and locations of the replicas. The client caches this information using the file name and chunk index as the key. The client then sends a request to one of the replicas, most likely the closest one.
- The request specifies the chunk handle and a byte range within that chunk.
- Further reads of the same chunk require no more client-master interaction until the cached information expires or the file is reopened.
- In fact, the client typically asks for multiple chunks in the same request and the master can also include the information for chunks immediately following those requested. This extra information sidesteps several future client-master interactions at practically no extra cost.

Q.3 Explain Google file system and its advantages.

☞ **[SPPU : Dec.-22, End Sem, Marks 5]**

Ans. : Advantages :

1. GFS provides a location independent namespace.
2. GFS spreads file's data across storage servers, distributing read/writes.
3. GFS uses commodity machines, lowering infrastructure costs.
4. GFS stores minimal metadata in memory

Also Refer Q.2.

Q.4 Explain how Google file system solves big data processing challenges. ☞ **[SPPU : May-18, End Sem, Marks 4]**

Ans. : • Big data organizes and extracts the valued information from the rapidly growing, large volumes, variety forms and frequently changing data sets collected from multiple and autonomous sources in the minimal possible time, using several statistical and machine learning techniques.

- GFS is high performance file system on commodity hardware clusters for large scale data processing.
- Google offers big query to operate on Google big tables. New generation of query languages, examples are Google big query.
- Web log mining is the study of the data available in the web. This involves searching for the texts, words, and their occurrences.
- One example for web log mining is searching for the words, and their frequencies by Google big query data analytics use Google big query platform to run on the Google cloud infrastructure.
- Mapreduce framework has made complex large-scale data processing easy and efficient. MapReduce is inherently designed for high throughput batch processing of big data that take several hours and even days, while recent demands are more centered on jobs and queries that should finish in seconds or at most, minutes.

3.3 : Hadoop Architecture

Q.5 Explain Hadoop distributed file system.

☞ **[SPPU : Dec.-22, End Sem, Marks 8]**

Ans. : • Hadoop Distributed File System (HDFS) is a distributed file system inspired by GFS that organizes files and stores their data on a distributed computing system.

- The Hadoop core is divided into two fundamental layers : The MapReduce engine and HDFS.
- The MapReduce engine is the computation engine running on top of HDFS as its data storage manager.
- Hadoop is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It provide software framework for distributed processing of large datasets in real-time applications.
- Hadoop provides the basic platform for big data processing. The hadoop architecture have mainly two parts : Hadoop distributed File System (HDFS) and the MapReduce engine.
- HDFS is Distributed Files system designed to run on commodity hardware, which is highly fault- tolerant andscalable. Fig. Q.5.1 shows HDFS architecture.

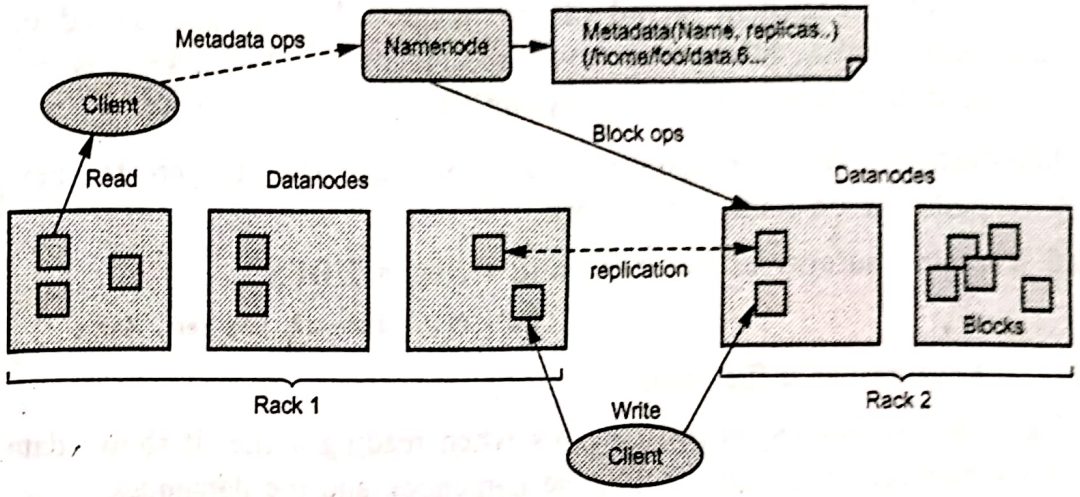


Fig. Q.5.1 Hadoop architecture

- Hadoop Distributed File System is a block - structured file system where each file is divided into blocks of a pre-determined size. These blocks are stored across a cluster of one or several machines.
- Apache Hadoop HDFS Architecture follows a Master/Slave Architecture, where a cluster comprises of a single NameNode (Master node) and all the other nodes are DataNodes (Slave nodes).
- To store a file in this architecture, HDFS splits the file into fixed-size blocks (e.g., 64 MB) and stores them on workers (DataNodes).
- HDFS can be deployed on a broad spectrum of machines that support Java. Though one can run several DataNodes on a single machine, but in the practical world, these DataNodes are spread across various machines.
- NameNode is the master node in the Hadoop HDFS Architecture that maintains and manages the blocks present on the DataNodes (slave nodes).
- NameNode is a very highly available server that manages the File System Namespace and controls access to files by clients.
- DataNodes are the slave nodes in HDFS. Unlike NameNode, DataNode is a commodity hardware, that is, a non-expensive system which is not of high quality or high-availability. The DataNode is a block server that stores the data in the local file ext3 or ext4.
- Journal is the modification log of image, which is available in local hosts native file system. Journal is updated for every client transaction.

- Checkpoint is persistent record of the image, which is also stored on local hosts native file system to enable recovery. NameNode is not allowed to update or modify Checkpoint file.
- Administrator or Checkpoint Node can demand to create new checkpoint file on startup, or restart.

Q.6 Explain anatomy of File read and write in HDFS.

☞ **[SPPU : June-22, End Sem, Marks 7]**

Ans. : Anatomy of a file read :

- Fig. Q.6.1 shows sequence of events when reading a file. It shows data flows between client and HDFS, the namenode and the datanodes.

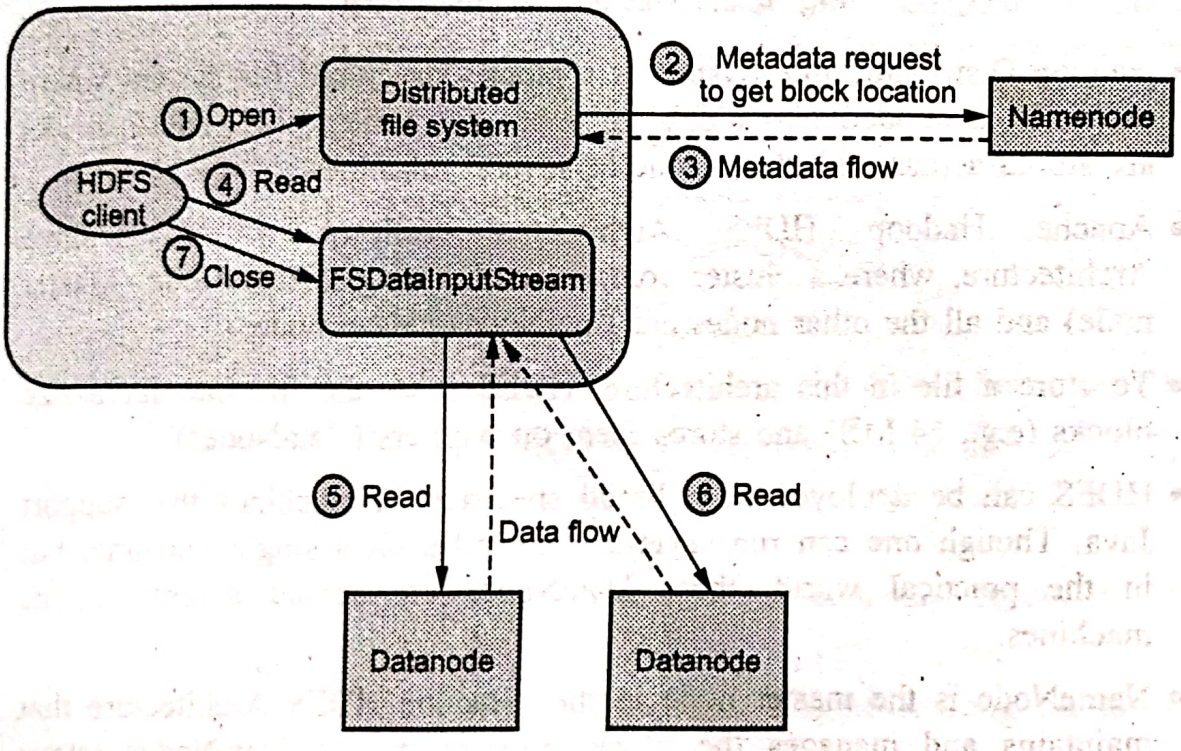


Fig. Q.6.1 Client reading data from HDFS

- The client opens the file it wishes to read by calling open() on the FileSystem object, which for HDFS is an instance of Distributed FileSystem (DFS). DFS calls the namenode, using RPC, to determine the locations of the blocks for the first few blocks in the file.
- For each block, the namenode returns the addresses of the datanodes that have a copy of that block. Furthermore, the datanodes are sorted according to their proximity to the client. If the client is itself a datanode, then it will read from the local datanode, if it hosts a copy of the block.

- The DFS returns an FSDataInputStream to the client for it to read data from. FSDataInputStream in turn wraps a DFSInputStream, which manages the datanode and namenode I/O.
- The client then calls read() on the stream. DFSInputStream, which has stored the datanode addresses for the first few blocks in the file, then connects to the first (closest) datanode for the first block in the file.
- Data is streamed from the datanode back to the client, which calls read() repeatedly on the stream. When the end of the block is reached, DFSInputStream will close the connection to the datanode, then find the best datanode for the next block. This happens transparently to the client, which from its point of view is just reading a continuous stream.
- Blocks are read in order with the DFSInputStream opening new connections to datanodes as the client reads through the stream. It will also call the namenode to retrieve the datanode locations for the next batch of blocks as needed. When the client has finished reading, it calls close() on the FSDataInputStream.
- During reading, if the DFSInputStream encounters an error while communicating with a datanode, then it will try the next closest one for that block.

Anatomy of a file write :

- Fig. Q.6.2 shows anatomy of a file write.

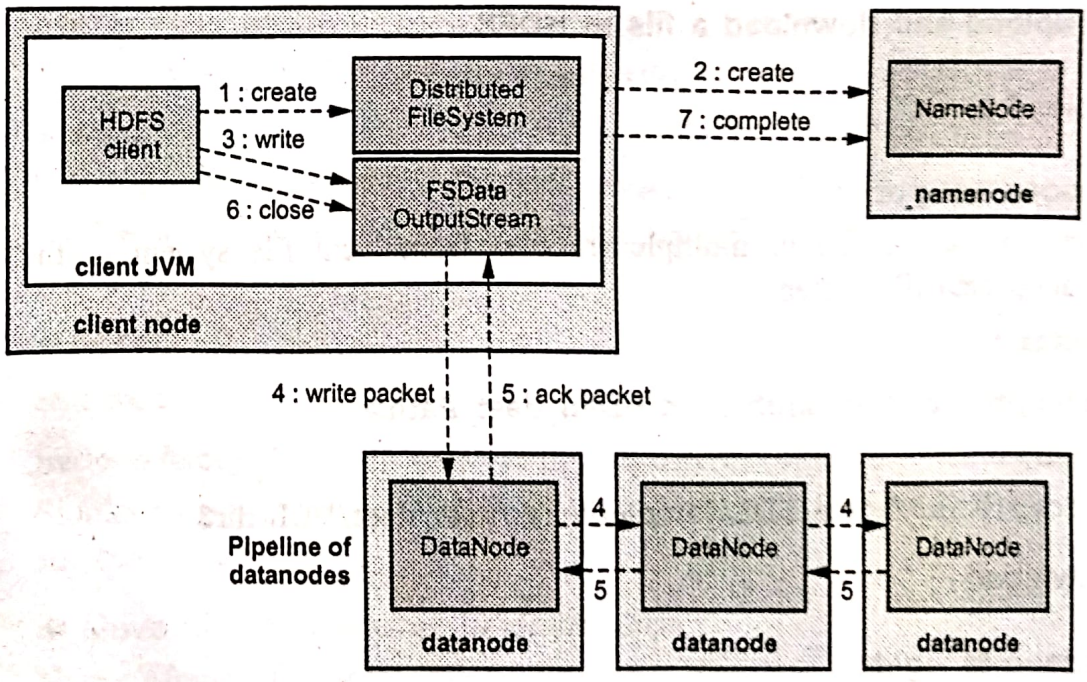


Fig. Q.6.2

1. The client calls create() on DistributedFileSystem to create a file.
2. An RPC call to the namenode happens through the DFS to create a new file.
3. As the client writes data, data is split into packets by DFSOutputStream, which is then writes to an internal queue, called data queue. Datastreamer consumes the data queue.
4. Data streamer streams the packets to the first DataNode in the pipeline. It stores packet and forwards it to the second DataNode in the pipeline.
5. In addition to the internal queue, DFSOutputStream also manages on "Ackqueue" of the packets that are waiting for acknowledged by DataNodes.
6. When the client finishes writing the file, it calls close() on the stream.

Q.7 Write and explain any two Hadoop shell commands.

☞ **[SPPU : June-22, End Sem, Marks 4]**

OR Write 5 Hadoop Shell commands.

☞ **[SPPU : Dec.-22, End Sem, Marks 5]**

Ans. : Simple shell commands can be written for Hadoop application by considering HDFS structure. Following are some of the frequent shell commands which can be used during hadoop operations.

1) Upload and download a file in HDFS

Upload :

hadoop fs - put

Copy single src file or multiple src files from local file system to the Hadoop data file system.

Syntax :

hadoopfs-put <localsrc> ... < HDFS_dest_Path>

Example :

hadoop fs_put/home/DDL/Samplefile.txt/user/user/DDL/dir3

Download

hadoop fs - get :

Copies /Downloads files to the local file system

syntax :

hadoop fs-get <hdfs_src> <localdst>

Example :

hadoop fs-get/user/DDL/dir3/Samplefile.txt/home/

2) See contents of a file

Same as unix cat command :

Syntax :

hadoop fs-cat <path(filename)>

Example :

hadoop fs-cat/user/DDL/dir1/abc.txt

3) Copy a file from source to destination

This command allow multiple sources as well in which case the destination must be a directory.

Syntax :

hadoop fs-cp <source> <dest>

Example :

hadoop fs - cp/user/DDL/dir1/abc.txt/user /user/DDL/dir2

Copy a file from/To Local file system to HDFS

copyFromLocal

Syntax :

hadoop fs-copy FromLocal<localsrc> URI

Example :

hadoop fs-coopyFromLocal/home/DDL/pqr.txt/user/DDL/pqr.txt

Similar to put command, except that the source is restricted to a local file reference.

copyToLocal

Syntax :

hadoop fs-copyTolocal[-ignorecrc][-crc] URI<localdst>

Similar to get command, except that the destination is restricted to a local file reference.

4) Move file from source to destination.

Note : Moving file across filesystem is not permitted.

Syntax :

hadoop fs-mv <src> <dest>

Example :

hadoop fs-mv/user/DDL/dir1/abc.txt/user/DDL/dir2

5) Remove a file or directory in HDFS

Remove files specified as argument. Deletes directory only when it is empty.

Syntax :

hadoop fs-rm <arg>

Example :

hadoop fs-rm /user/DDL/dir1/abc.txt

Recursive version of delete.

Syntax :

hadoop fs-rmr<arg>

Example :

hadoop fs-rmr/user/DDL/

Q.8 Explain MapReduce with proper diagram for word count example. ☞ **[SPPU : Dec.-19, 22, June-22, End Sem, Marks 7]**

Ans. : MapReduce is a computation that decomposes large manipulation jobs into individual tasks that can be executed in parallel across a cluster of servers. The results of tasks can be joined together to compute final results. (See Fig. Q.8.1 on next page)

Input (output of Map function)	Set of Tuples	(Bus,1), (Car,1), (bus,1), (car,1), (train,1), (car,1), (bus,1), (car,1), (train,1), (bus,1), (TRAIN,1),(BUS,1), (buS,1), (caR,1), (CAR,1), (car,1), (BUS,1), (TRAIN,1)
Output	Converts into smaller set of tuples	(BUS,7), (CAR,7), (TRAIN,4)

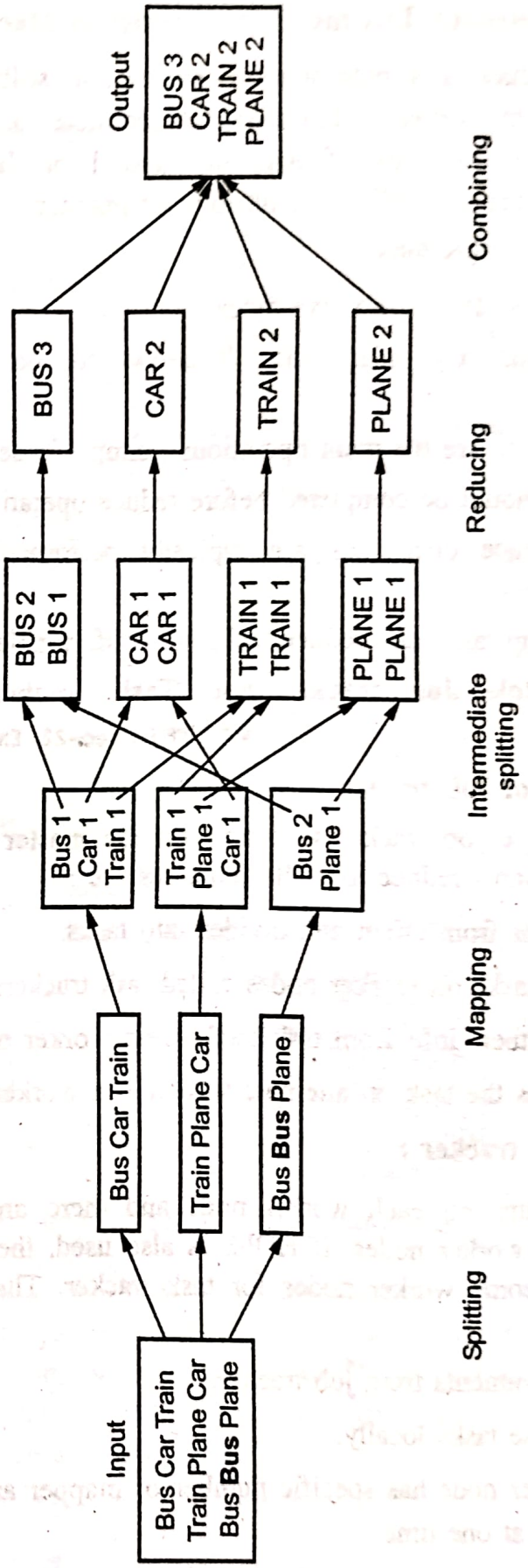

Output
BUS 3 CAR 2 TRAIN 2 PLANE 2
Combining
BUS 3
CAR 2
TRAIN 2
PLANE 2
Reducing
BUS 2 BUS 1
CAR 1 CAR 1
TRAIN 1 TRAIN 1
PLANE 1 PLANE 1
Intermediate splitting
Bus 1 Car 1 Train 1
Train 1 Plane 1 Car 1
Bus 2 Plane 1
Mapping
Bus Car Train
Train Plane Car
Bus Bus Plane
Splitting
Input
Bus Car Train Train Plane Car Bus Bus Plane

Fig. Q.8.1

Q.9 Define MapReduce. List the characteristics of MapReduce.

Ans.: • MapReduce is a programming model and software framework first developed by Google. Intended to facilitate and simplify the processing of vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner.

Characteristics of MapReduce :

1. Very large scale data : peta, exa bytes
2. Write once and read many data. It allows for parallelism without mutexes
3. Map and Reduce are the main operations : simple code
4. All the map should be completed before reduce operation starts.
5. Map and reduce operations are typically performed by the same physical processor.
6. Number of map tasks and reduce tasks are configurable.

Q.10 Explain Role Job tracker and Task tracker in Hadoop Architecture. ☞ **[SPPU : Dec.-22, End Sem, Marks 5]**

Ans. : Function of Job tracker :

- There is a single job tracker that runs on the master node. It is the driver for the map - reduce jobs. Its functions are :
  1. Accepts jobs from client and divides into tasks.
  2. Schedules tasks on worker nodes called task trackers.
  3. Keeps heartbeat info from task trackers on worker nodes.
  4. Reschedules the task on alternate worker if a worker fails.

Function of task tracker :

- Task tracker runs on each worker node and there are as many task trackers as the worker nodes. If HDFS is also used, then data nodes of HDFS also become worker nodes for task tracker. The functions of a task tracker are :
  a. Takes assignments from job tracker.
  b. Executes the tasks locally.
  c. Each worker node has specific number of mapper and reducer tasks it can take at one time.
  d. The tasks assigned are run in parallel.

e. Normally they can take more map jobs than reduce tasks.

f. Task tracker does a task attempt before executing task.

g. Task tracker may do multiple attempts before declaring a task as failed.

h. Task tracker maintains a connection with the task attempt called umbilical protocol.

i. Task tracker sends a regular heartbeat signal to job tracker indicating its status including available map and reduce tasks.

j. Task tracker runs each task attempt in a separate JVM. So even if the task has bad code due to which it fails, it will not cause task tracker to abort.

Q.11 Explain job execution in Hadoop with example.

☞ **[SPPU : May-18, End Sem, Marks 6]**

Ans. : • Fig Q.11.1 shows job execution sequence of Hadoop.

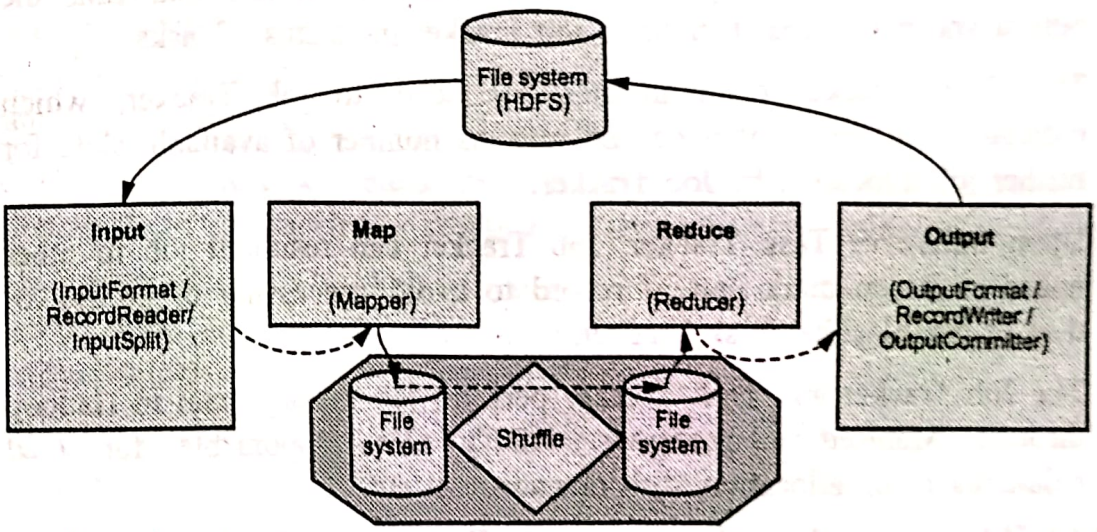


Fig. Q.11.1 Job execution sequence of Hadoop

• Daemon service in Hadoop environment is job Tracker, which is responsible for job execution and processing. For every hadoop cluster there is only one Job Tracker, run on JVM process. Its responsible for executing all job requests made by client application. The process of execution of job sequence is as follows :

• Job request is submitted by client.

• Job Tracker processes the job request

- Job Trackers communicates the NameNode for getting the data location in the cluster.
- Job Tracker finds Task Tracker modes in DataNodes and issues job to these nodes.
- Job Tracker notifies Job Tracker about the status of Job completion.

- Upon completion of Job, the Job Tracker update the status and client is signaled about the completion of processing.
- Task Tracker node accepts jobs from Job Tracker in cluster. Task Tracker maintains set of slots for maintaining the details of total number of tasks it can accept at given point of time. Following are the responsibilities of Task Tracker :
- Task Tracker creates separate JVM process for task, so the particular task can be isolated and failure of that does not affect the Task Tracker.
- Task Tracker monitors job execution created by it and maintains the output and exit codes. It notifies Job Tracker the status of tasks.
- The Task Tracker communicates heartbeats to job Tracker, which indicates availability of node and informs number of available slots for further job allocation by Job Tracker.
- Upon failure of Task Tracker, Job Tracker can resubmit job to other nodes or it can mark part of record to avoid processing of portion of data or can blacklist Task Tracker.
- The Job Tracker and Task Tracker perform job management in Hadoop through MapReduce processes. HDFS also responsible for load balancing block allocation, disk management etc.

Q.12 What are the advantages of Hadoop ? Explain Hadoop architecture and its components with proper diagram.

☞ **[SPPU : Dec.-18, End Sem, Marks 5]**

Ans. : Advantages :

1. Scalable : Hadoop cluster can be extended by just adding nodes in the cluster.
2. Cost effective : Hadoop is open source and uses commodity hardware to store data so it really cost effective as compared to traditional relational database management system.

3. Resilient to failure : HDFS has the property with which it can replicate data over the network
4. Hadoop can handle unstructured as well as semi-structured data.
5. The unique storage method of Hadoop is based on a distributed file system that effectively maps data wherever the cluster is located.

Also Refer Q.5.

Q.13 Explain Hadoop ecosystem with proper diagram.

Ans. : • Hadoop ecosystem is neither a programming language nor a service, it is a platform or framework which solves big data problems.

- The Hadoop ecosystem refers to the various components of the Apache Hadoop software library, as well as to the accessories and tools provided by the Apache Software Foundation for these types of software projects and to the ways that they work together.
- Hadoop is a Java - based framework that is extremely popular for handling and analysing large sets of data. The idea of a Hadoop ecosystem involves the use of different parts of the core Hadoop set such as MapReduce, a framework for handling vast amounts of data and the Hadoop Distributed File System (HDFS), a sophisticated file - handling system. There is also YARN, a Hadoop resource manager.
- In addition to these core elements of Hadoop, Apache has also delivered other kinds of accessories or complementary tools for developers.
- Some of the most well - known tools of the Hadoop ecosystem include HDFS, Hive, Pig, YARN, MapReduce, Spark, HBase, Oozie, Sqoop, Zookeeper, etc.
- Fig. Q.13.1 shows Apache Hadoop ecosystem.

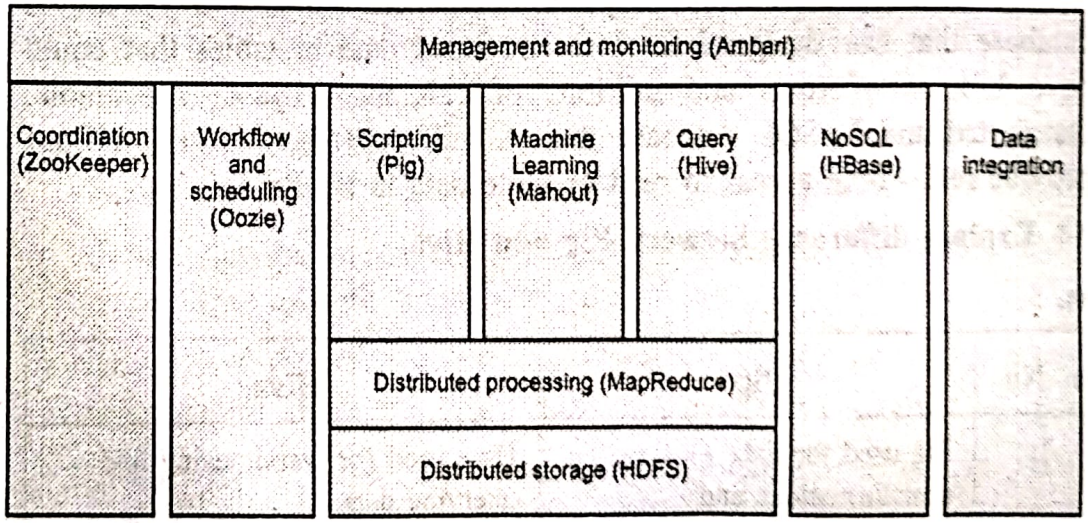


Fig. Q.13.1 Apache Hadoop ecosystem

- Hadoop Distributed File System (HDFS), is one of the largest Apache projects and primary storage system of Hadoop. It employs a NameNode and DataNode architecture. It is a distributed file system able to store large files running over the cluster of commodity hardware.
- YARN stands for Yet Another Resource Negotiator. It is one of the core components in open source Apache Hadoop suitable for resource management. It is responsible for managing workloads, monitoring and security controls implementation.
- Hive is an ETL and Data warehousing tool used to query or analyze large datasets stored within the Hadoop ecosystem. Hive has three main functions : Data summarization, query and analysis of unstructured and semi - structured data in Hadoop.
- Map - Reduce : It is the core component of processing in a Hadoop Ecosystem as it provides the logic of processing. In other words, MapReduce is a software framework which helps in writing applications that processes large data sets using distributed and parallel algorithms inside Hadoop environment.
- Apache Pig is a high - level scripting language used to execute queries for larger datasets that are used within Hadoop.
- Apache Spark is a fast, in - memory data processing engine suitable for use in a wide range of circumstances. Spark can be deployed in several ways, it features Java, Python, Scala and R programming languages and supports SQL, streaming data, machine learning and graph processing, which can be used together in an application.
- Apache HBase is a Hadoop ecosystem component which is a distributed database that was designed to store structured data in tables that could have billions of rows and millions of columns. HBase is scalable, distributed and NoSQL database that is built on top of HDFS. HBase provide real - time access to read or write data in HDFS.

Q.14 Explain difference between Pig and Hive.

Ans. :

Sr. No.	Pig	Hive
1.	Pig used for data transformations and processing.	Hive used for warehousing and querying data.

2.	Pig works on structured, semi - structured and unstructured data.	Hive works only on structured data.
3.	Pig does not support web interface.	Hive support web interface.
4.	Pig is a scripting platform that runs on Hadoop clusters, designed to process and analyze large datasets. Pig uses a language called Pig Latin, which is similar to SQL.	Hive is a data warehouse system used to query and analyze large datasets stored in HDFS. Hive uses a query language called HiveQL, which is similar to SQL.
5.	Pig support Avro file format.	Hive does not support Avro file format.
6.	Creating schema is not required to store data in Pig.	Hive supports schema.
7.	Pig loads data quickly.	Hive takes time to load but executes quickly.
8.	Pig works on the client - side of the cluster.	Hive works on the server - side of the cluster.
9.	Used for programming.	Used for reporting.

Q.15 Explain HBase with its architecture.

Ans. : • HBase is an open source, non - relational, distributed database modeled after Google's BigTable. HBase is an open source and sorted map data built on Hadoop. It is column oriented and horizontally scalable.

- It is a part of the Hadoop ecosystem that provides random real - time read/write access to data in the Hadoop file system. It runs on top of Hadoop and HDFS, providing Big Table - like capabilities for Hadoop.
- HBase supports massively parallelized processing via MapReduce for using HBase as both source and sink.
- HBase supports an easy - to - use Java API for programmatic access. It also supports Thrift and REST for non - Java front - ends.
- HBase is a column oriented distributed database in Hadoop environment. It can store massive amounts of data from terabytes to

petabytes. HBase is scalable, distributed big data storage on top of the Hadoop eco system.

- The HBase physical architecture consists of servers in a Master-Slave relationship. Typically, the HBase cluster has one Master node, called HMaster and multiple Region Servers called HRegionServer.
- Fig. Q.15.1 shows Hbase architecture.

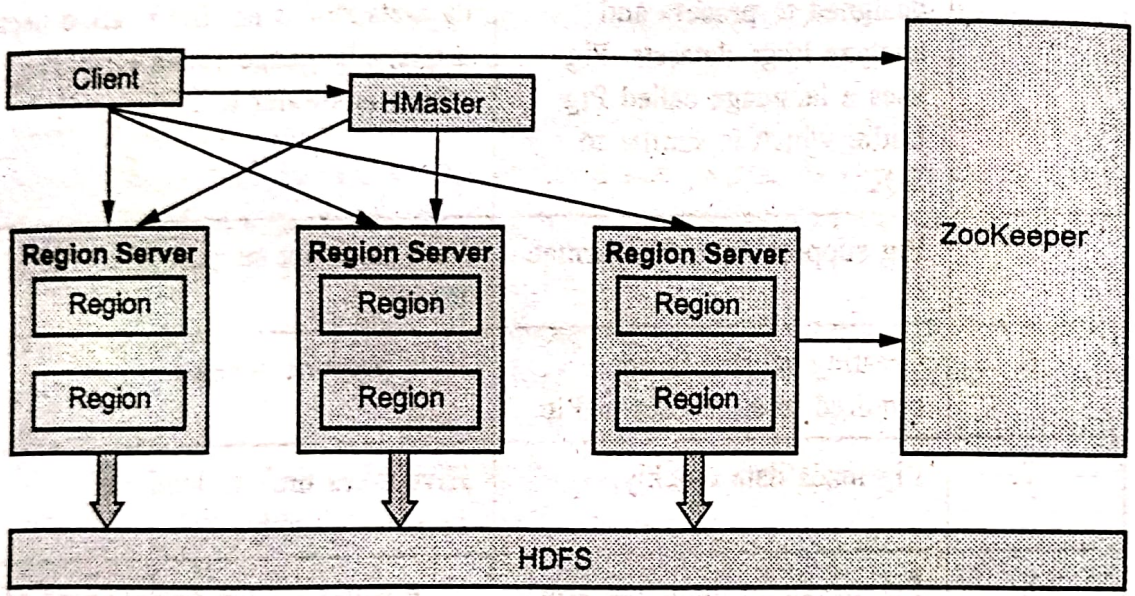


Fig. Q.15.1 Hbase architecture

- Zookeeper is a centralized monitoring server which maintains configuration information and provides distributed synchronization. If the client wants to communicate with regions servers, client has to approach Zookeeper.
- HMaster in the master server of Hbase and it coordinates the HBase cluster. HMaster is responsible for the administrative operations of the cluster.
- HRegions servers : It will perform the following functions in communication with HMaster and Zookeeper.
  1. Hosting and managing regions.
  2. Splitting regions automatically.
  3. Handling read and writes requests.
  4. Communicating with clients directly.
- **HRegions :** For each column family, HRegions maintain a store. Main components of HRegions are Memstore and Hfile

- Data model in HBase is designed to accommodate semi - structured data that could vary in field size, data type and columns.
- HBase is a column - oriented, non - relational database. This means that data is stored in individual columns, and indexed by a unique row key. This architecture allows for rapid retrieval of individual rows and columns and efficient scans over individual columns within a table.
- Both data and requests are distributed across all servers in an HBase cluster, allowing you to query results on petabytes of data within milliseconds. HBase is most effectively used to store non - relational data, accessed via the HBase API.

Q.16 Write short note on Mahout.

Ans. : • Mahout is an open source machine learning library from Apache written in java. It also supports a number of clustering algorithms like k - means, mean - shift and canopy.

- The primitive features of Apache Mahout include :
  1. The algorithms of Mahout are written on top of Hadoop, so it works well in distributed environment.
  2. Mahout uses the Apache Hadoop library to scale effectively in the cloud.
  3. Mahout offers the coder a ready - to - use framework for doing data mining tasks on large volumes of data.
  4. Mahout lets applications to analyze large sets of data effectively and in quick time.
  5. Includes several MapReduce enabled clustering implementations such as k - means, fuzzy k - means, Canopy etc.
  6. Supports Distributed Naive Bayes and Complementary Naïve Bayes classification implementations.
  7. Comes with distributed fitness function capabilities for evolutionary programming.
  8. Includes matrix and vector libraries.
- Mahout is an open source machine learning library built on top of Hadoop to provide distributed analytics capabilities. Mahout incorporates a wide range of data mining techniques including collaborative filtering, classification and clustering algorithms.

Q.17 Discuss with diagram overflow of programming model of MapReduce operation.

Ans.: • The computation takes a set of input key/value pairs, and produces a set of output key/value pairs. The user of the MapReduce library expresses the computation as two functions: Map and Reduce.

- Map, written by the user, takes an input pair and produces a set of intermediate key/value pairs. MapReduce library groups together all intermediate values associated with the same intermediate key " I" and passes them to the Reduce function.
- The Reduce function, also written by the user, accepts an intermediate key I and a set of values for that key. It merges together these values to form a possibly smaller set of values.
- Fig. Q.17.1 shows the overall flow of a MapReduce operation.

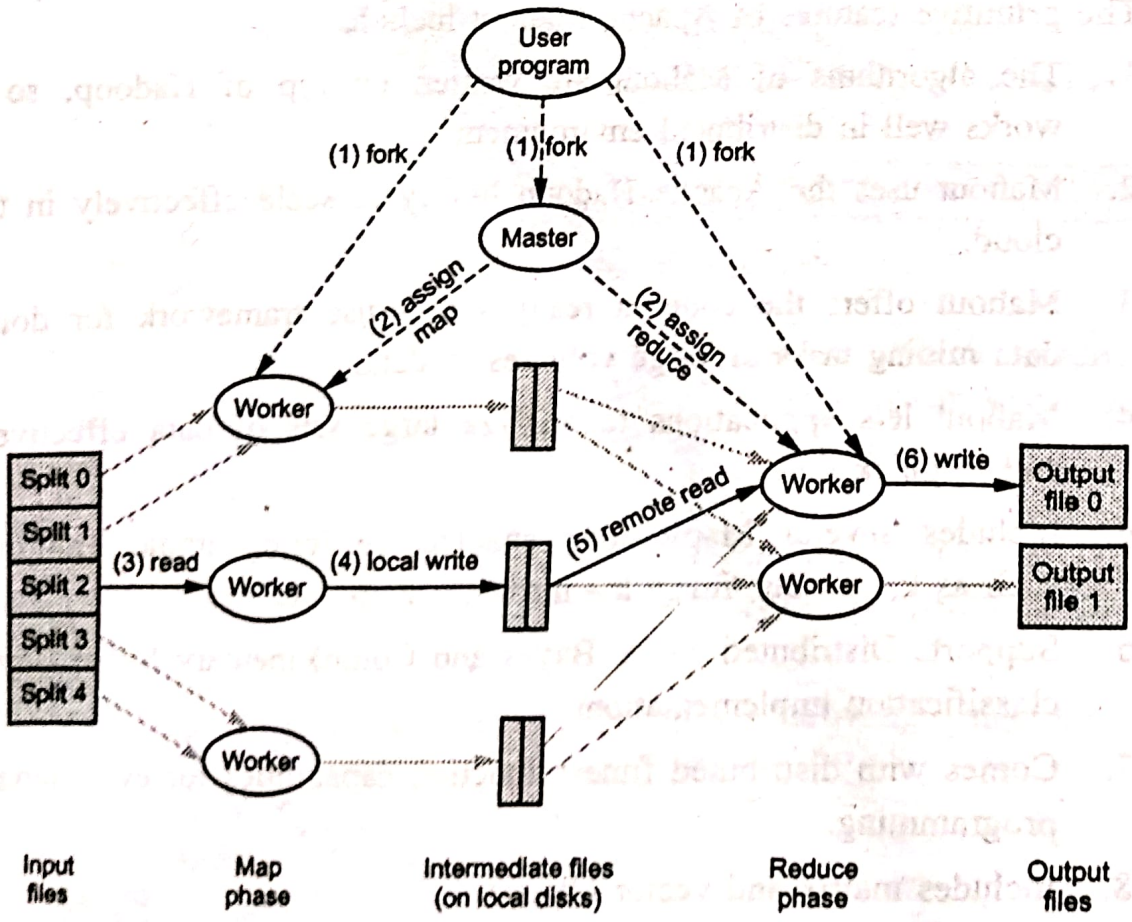


Fig. Q.17.1 MapReduce execution

- When the user program calls the MapReduce function, the following sequence of actions occurs :

1. MapReduce library in the user program first splits the input files into M pieces of typically 16 MB to 64 MB per piece. It then starts up many copies of the program on a cluster of machines.
2. One of the copies of the program is special i.e. master. The rest are workers that are assigned work by the master. There are M map tasks and R reduce tasks to assign. The master picks idle workers and assigns each one a map task or a reduce task.
3. A worker who is assigned a map task reads the contents of the corresponding input split. It parses key/value pairs out of the input data and passes each pair to the user-defined Map function. The intermediate key/value pairs produced by the Map function are buffered in memory.
4. Periodically, the buffered pairs are written to local disk, partitioned into R regions by the partitioning function. The locations of these buffered pairs on the local disk are passed back to the master, who is responsible for forwarding these locations to the reduce workers.
5. When a reduce worker is notified by the master about these locations, it uses remote procedure calls to read the buffered data from the local disks of the map workers. When a reduce worker has read all intermediate data, it sorts it by the intermediate keys so that all occurrences of the same key are grouped together.
6. The reduce worker iterates over the sorted intermediate data and for each unique intermediate key encountered, it passes the key and the corresponding set of intermediate values to the user's Reduce function. The output of the Reduce function is appended to a final output file for this reduce partition.
7. When all map tasks and reduce tasks have been completed, the master wakes up the user program. At this point, the MapReduce call in the user program returns back to the user code.

Q.18 Explain heartbeat mechanism in HDFS.

☞ **[SPPU : Dec.-18, End Sem, April-19, In Sem, Marks-5]**

Ans. : • Heartbeat is a signal indicating that it is alive. A DataNode sends heartbeat to Namenode and task tracker will send its heart beat to job tracker.

• Fig. Q.18.1 shows heartbeat mechanism.

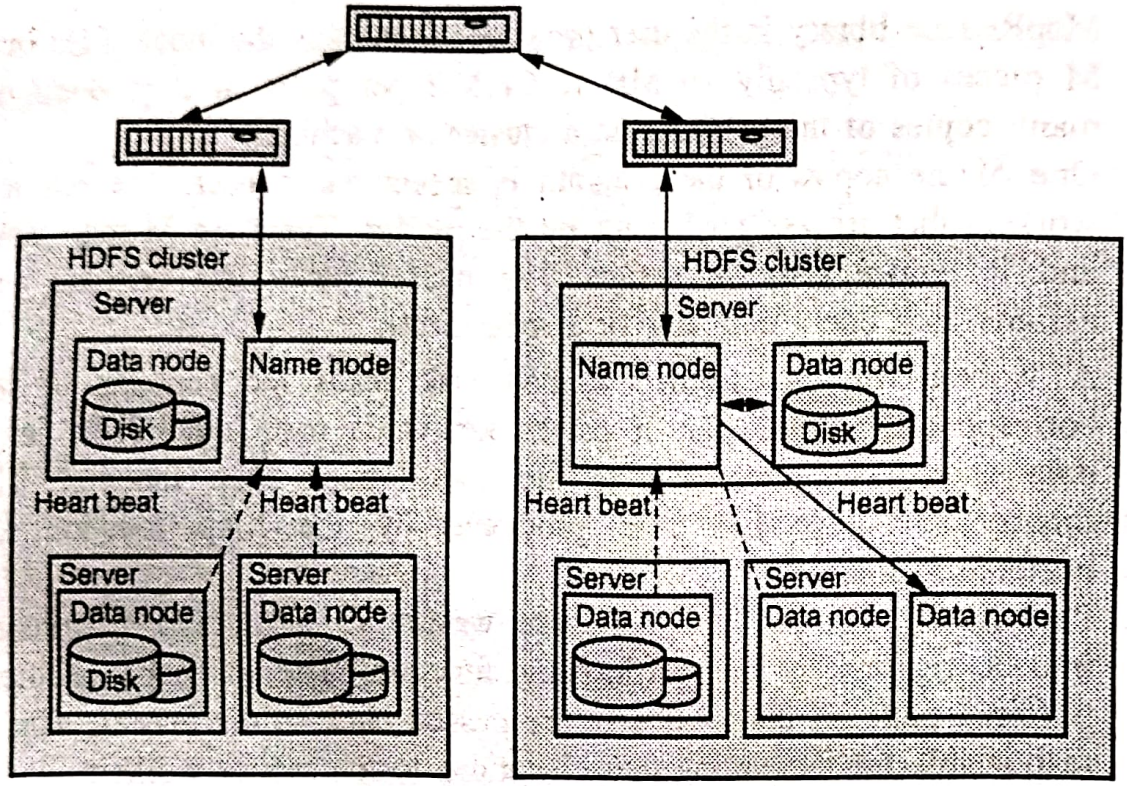


Fig. Q.18.1

- The connectivity between the NameNode and a DataNode are managed by the presistent heartbeats that are sent by the DataNode every three seconds.
- The heartbeat provides the NameNode confirmation about the availability of the blocks and the replicas of the DataNode.
- Additionally, heartbeats also carry information about total storage capacity, storage in use and the number of data transfers currently in progress. These statistics are by the NameNode for managing space allocation and load balancing.
- During normal operations, if the NameNode does not receive a heartbeat from a DataNode in ten minutes the NameNode, it considers that DataNode to be out of service and the block replicas hosted to be unavailable.
- The NameNode schedules the creation of new replicas of those blocks on other DataNodes.
- The heartbeats carry roundtrip communications and instructions from the NameNode, including commands to :

  a) Replicate blocks to other nodes.

  b) Remove local block replicas.

c) Re-register the node.

d) Shut down the node.

e) Send an immediate block report.

Q.19 What is the role of sorter, shuffler and combiner in Map reduces paradigm ? ☞ [SPPU : April-19, In Sem, Marks 4]

Ans. : • A Combiner, also known as a semi-reducer, is an optional class that operates by accepting the inputs from the Map class and thereafter passing the output key-value pairs to the Reducer class.

- The main function of a Combiner is to summarize the map output records with the same key. The output of the combiner will be sent over the network to the actual Reducer task as input.
- The process of transferring data from the mappers to reducers is known as shuffling i.e. the process by which the system performs the sort and transfers the map output to the reducer as input. So, shuffle phase is necessary for the reducers, otherwise, they would not have any input.
- Shuffle phase in Hadoop transfers the map output from Mapper to a Reducer in MapReduce. Sort phase in MapReduce covers the merging and sorting of map outputs.
- Data from the mapper are grouped by the key, split among reducers and sorted by the key. Every reducer obtains all values associated with the same key. Shuffle and sort phase in Hadoop occur simultaneously and are done by the MapReduce framework.

3.4 : Introduction to NOSQL

Q.20 Define NOSQL. Why NOSQL ? Explain features of NOSQL.

Ans. : • NoSQL means Not Only SQL, it solves the problem of handling huge volume of data that relational databases cannot handle. NoSQL databases are schema free and are non - relational databases. Most of the NoSQL databases are open source.

Why NoSQL ?

- It can handle large volumes of structured, semi-structured and unstructured data.
- Agile sprints, quick iteration and frequent code pushes.

- Object - oriented programming that is easy to use and flexible.
- Scale - out architecture,

Types of NoSQL Stores -

1. Column oriented (Accumlo, Cassandra, HBase)
2. Document Oriented (MpngoDB, Couchbase, Clusterpoint)
3. Key - value (Dynamo, MemcacheDB, Riak)
4. Graph (Allegro, Neo4j, OrientDB)_

- Features of NoSQL

1) **Multi-model :** The concept is to allow multiple data models in a single database.
2) **Distributed :** NoSQL databases use the shared - nothing architecture, implying that the database has no single control unit or storage.
3) **Eliminated downtime :** The data is maintained at various nodes owing to its architecture, the failure of one node will not affect the entire system.
4) **NoSQL databases** can process structured, semi-structured or unstructured data with the same ease, thereby increasing performance.
5) **High Scalability :** NoSQL databases use horizontal scaling and thus the data remains accessible even when one or more nodes go down.

Q.21 What is CAP theorem ? How it helps to big data ?

Ans.: • CAP theorem states that it is impossible for a distributed data store to simultaneously provide more than two out of the following three guarantees :

1. Consistency : Every read receives the most recent write or an error
2. Availability : Every request receives a response - without guarantee that it contains the most recent write
3. Partition tolerance : The system continues to operate despite an arbitrary number of messages being dropped by the network between nodes.

- The CAP theorem says : No distributed system can have these three properties.
- Consistency : In the shopping cart example, user should always be getting the previously added items. The shopping cart itself has to be consistent.

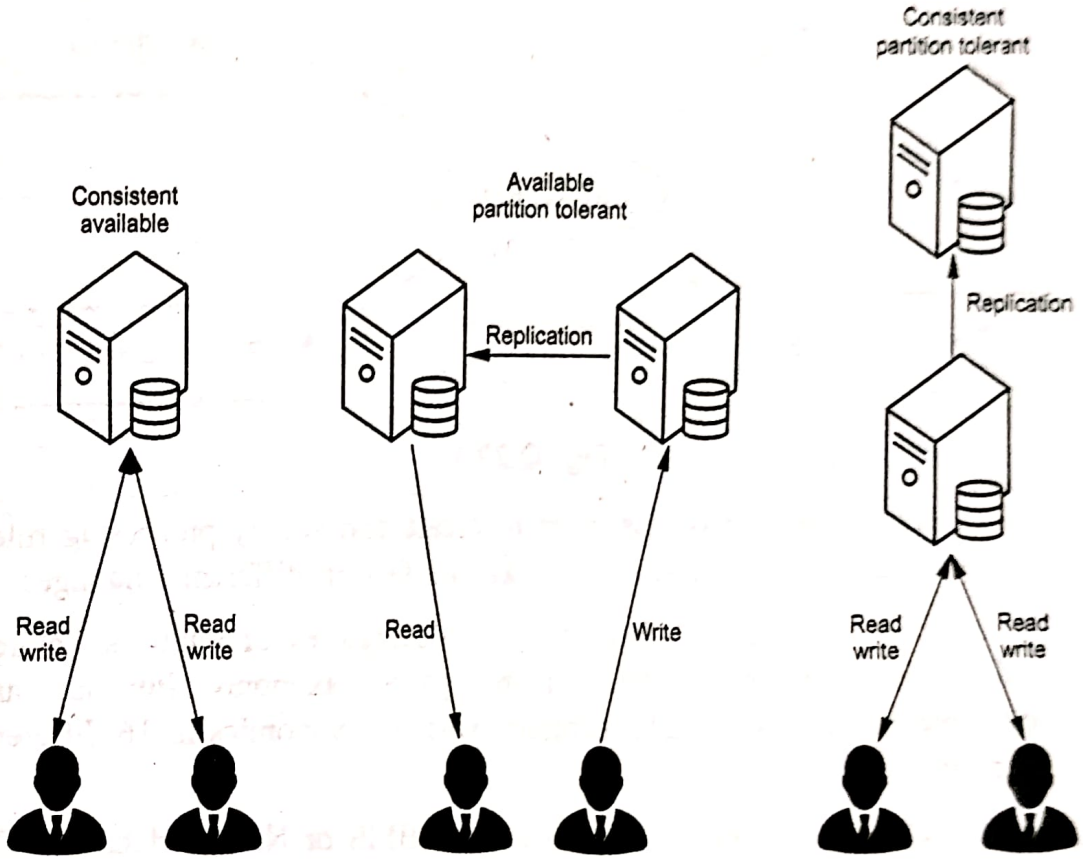


Fig. Q.21.1

- Availability : In this case, user should always access his shopping cart eventually regardless of its consistency.
- Partition tolerance : What partition tolerance forces our system is that we should be accessing our system even some nodes are partitioned.

Q.22 Explain ETL processing. ☞ **[SPPU : June-22, End Sem, Marks 4]**

OR Explain ETL in Big data. ☞ **[SPPU : Dec.-22, End Sem, Marks 5]**

Ans. : • The components of textual ETL (Extract, Transform, Load) processing are Textual ETL rules engine. User Interface, Taxonomies, output database. Fig. Q.22.1 shows ETL processing.

- Textual ETL rules Engine takes large unstructured data and parse it to extract value for integration. Rules engines contains series of data processing steps and algorithms such as classification, clustering affinity, proximity.
- Taxonomy integration, master data integration, metadata integration are some integration processing techniques are available in rules engine.

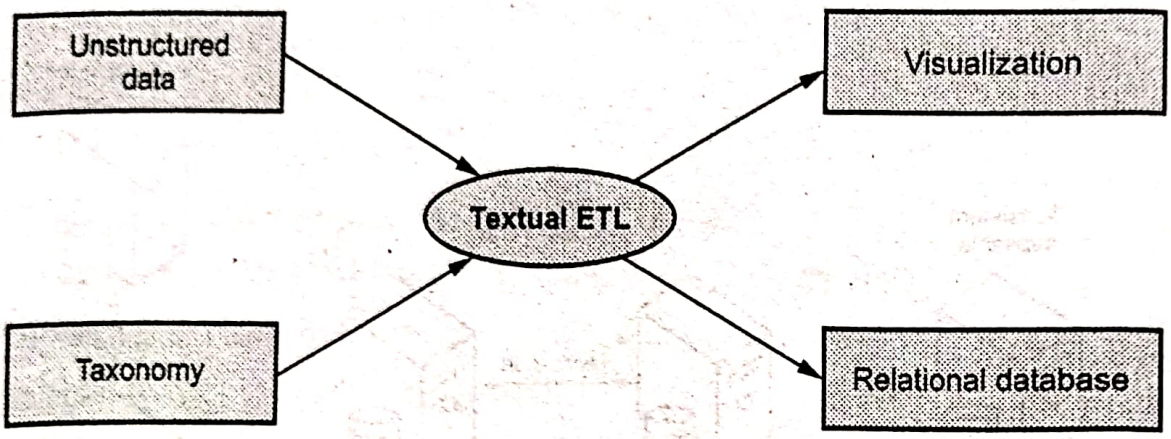


Fig. Q.22.1

- User Interface helps business users to create and supply processing rules through drag drop and free - form text interface in different languages.
- Taxonomies are required that of several categories of multi - structures and multi - hierarchical data. Third party taxonomy libraries are available, where textual ETL supports variety taxonomies in 16 different languages.
- Output Database in textual ETL is any RDBMS or NoSQL database. To integrate structured and unstructured database uses result sets which are key value pairs.

Q.23 Differentiate between SQL and NOSQL.

☞ **[SPPU : April-19, In Sem, Marks 4]**

Ans. :

Sr. No.	SQL	NoSQL
1.	These are called RDBMS	These are called not only SQL database.
2.	Based on ACID properties i.e. Atomicity, Consistency, Isolation and Durability.	Based on CAP properties i.e. (Consistency, Availability and Partition tolerance)
3.	These are table based database i.e. the data are stored in a table with rows and columns.	These databases are documented based, key - value pairs or graph based etc.

4.	These are scaled vertically. Load can be managed by increasing CPU, RAM etc in the same server.	These are scaled horizontally. A few servers can be added to manage large traffic.
5.	Preferred for complex, query execution.	Not preferred for complex query execution.
6.	Examples : DB2, MySQL, Oracle, Postgress, SQL server.	Examples : ConchDb, MongoDB, RavenDb, Redis, Cassandra, Hbase, Neo-4j, BigTable.

END...

Unit IV

4 Big Data Analytics

4.1 : Big Data Analytics - Architecture and Life Cycle

Q.1 Explain different steps in data analytics project life cycle.

☞ [SPPU : June-18, 22, End Sem, Marks 7]

Ans. : • The data analytic lifecycle is designed for Big Data problems and data science projects. With six phases the project work can occur in several phases simultaneously. Fig. Q.1.1 shows data analytic life cycle model.

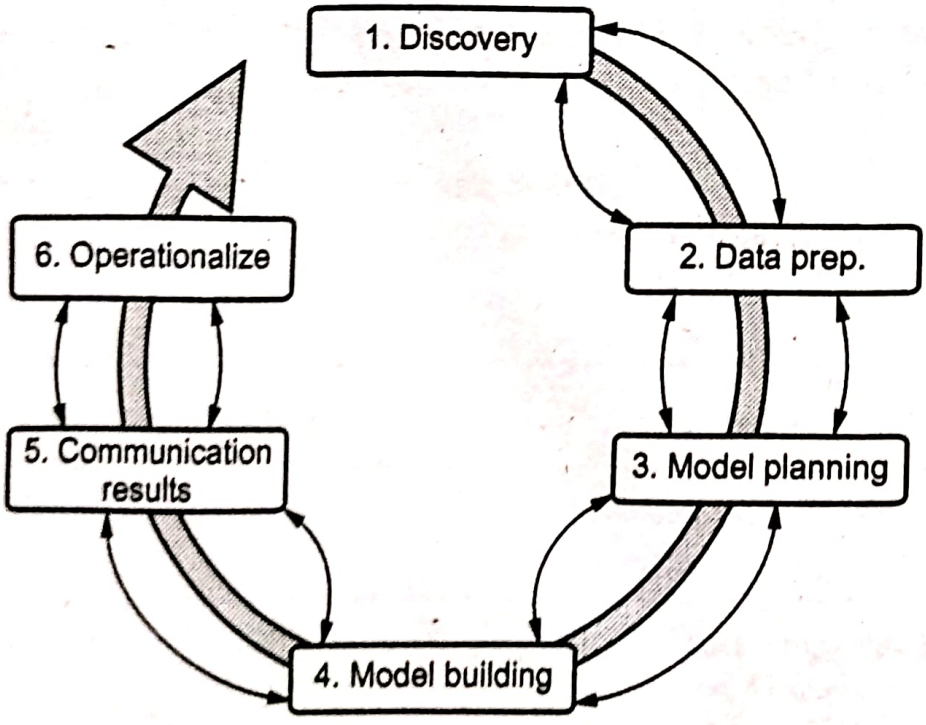


Fig. Q.1.1

1. **Discovery :** In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time, and data.

2. **Data preparation :** Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox.
3. **Model planning :** The team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase.
4. **Model building :** In Phase 4, the team develops datasets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase.
5. **Communicate results :** In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1.
6. **Operationalize :** In Phase 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.

Q.2 Explain in detail how the model building phase is built by team in data analytics life cycle ?

Ans. : Phase 3 : Model planning

- The team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase.
- The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.
- **Activities to consider :**

a) Assess the structure of the data -this dictates the tools and analytic techniques for the next phase

b) Ensure the analytic techniques enable the team to meet the business objectives and accept or reject the working hypotheses

c) Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow

d) Research and understand how other analysts have approached this kind or similar kind of problem.

a) Data exploration and variable selection

- Explore the data to understand the relationships among the variables to inform selection of the variables and methods. A common way to do this is to use data visualization tools.
- Often, stakeholders and subject matter experts may have ideas. For example, some hypothesis that led to the project.
- Aim for capturing the most essential predictors and variables. This often requires iterations and testing to identify key variables.
- If the team plans to run regression analysis, identify the candidate predictors and outcome variables of the model.

b) Model selection

- The main goal is to choose an analytical technique, or several candidates, based on the end goal of the project. We observe events in the real world and attempt to construct models that emulate this behavior with a set of rules and conditions.
- A model is simply an abstraction from reality. Determine whether to use techniques best suited for structured data, unstructured data, or a hybrid approach.
- Teams often create initial models using statistical software packages such as R, SAS, or Matlab. Which may have limitations when applied to very large datasets.
- The team moves to the model building phase once it has a good idea about the type of model to try.

c) Common tools for the model planning phase

- R programming language has a complete set of modeling capabilities. It contains about 5000 packages for data analysis and graphical presentation.
- SQL Analysis services can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models.
- SAS/ACCESS provides integration between SAS and the analytics sandbox via multiple data connections.

Phase 4 : Model building

- The team develops datasets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase.
- Building a model involves two phases :
  - a) **Design the model :** identify a suitable model. This step can involve a number of different modeling techniques to identify a suitable model. These may include decision trees, regression techniques and neural networks.
  - b) **Execute the model :** The model is run against the data to ensure that the model fits the data.
- Common Commercia Tools for the Model Building Phase
  - a. SAS Enterprise Miner used for building enterprise-level computing and analytics.
  - b. SPSS Modeler (IBM) provides enterprise-level computing and analytics.
  - c. Matlab is a high-level language for data analytics, algorithms, data exploration
  - d. Alpine Miner provides GUI frontend for backend analytics tools.
  - e. STATISTICA and MATHEMATICA is popular data mining and analytics tools

Q.3 List and explain the steps in data preparation phase of data analytics life cycle.

Ans. : • This stage involves in collecting, processing and cleaning data. Here the focus shift from business requirement to data requirement. In this early phase, data is collected but not analyzed.

- The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often.

a) Preparing the analytic sandbox :

- Create the analytic sandbox. It also called workspace. It allows team to explore data without interfering with live production data.
- Sandbox collects all kinds of data. The sandbox allows organizations to undertake ambitious projects beyond traditional data analysis and BI to perform advanced predictive analytics.

b) Performing ETLT (Extract, Transform, Load, Transform) :

- The team needs to execute Extract, Load, and Transform (ELT) to get data into the sandbox.
- Extract, Transform, Load (ETL) : It transforms the data based on a set of business rules before loading it into the sandbox.
- Extract, Load, Transform (ELT) : It loads the data into the sandbox and then transforms it based on a set of business rules.
- Extract, Transform, Load, Transform (ETLT) : It's the combination of ETL and ELT and has two transformation levels.

c) Learning about the data :

- Data is captured through three main ways :
  - i. **Data acquisition :** Obtaining existing data from outside sources.
  - ii. **Data entry :** Creating new data values from data inputted within the organization.
  - iii. **Signal reception :** Capturing data created by devices.

d) Data conditioning :

- Data conditioning includes cleaning data, normalizing datasets, and performing transformations. It often viewed as a preprocessing step prior to data analysis, it might be performed by data owner, IT department, DBA, etc.
- Best to have data scientists involved and data science teams prefer more data than too little.

e) Common tools for data preparation :

- Hadoop can perform parallel ingest and analysis.
- Alpine Miner provides a graphical user interface for creating analytic workflows.
- OpenRefine is a free, open source tool for working with messy data.
- Similar to OpenRefine, Data Wrangler is an interactive tool for data cleansing an transformation.

4.2 : Data Analytics Architecture

Q.4 Explain big data analytics architecture with diagram.

Ans. : • Analytics architecture refers to the systems, protocols, and technology used to collect, store, and analyze data. ... Analytics architecture also focuses on multiple layers, starting with data warehouse architecture, which defines how users in an organization can access and interact with data.

• Fig. Q.4.1 shows data analytical architecture.

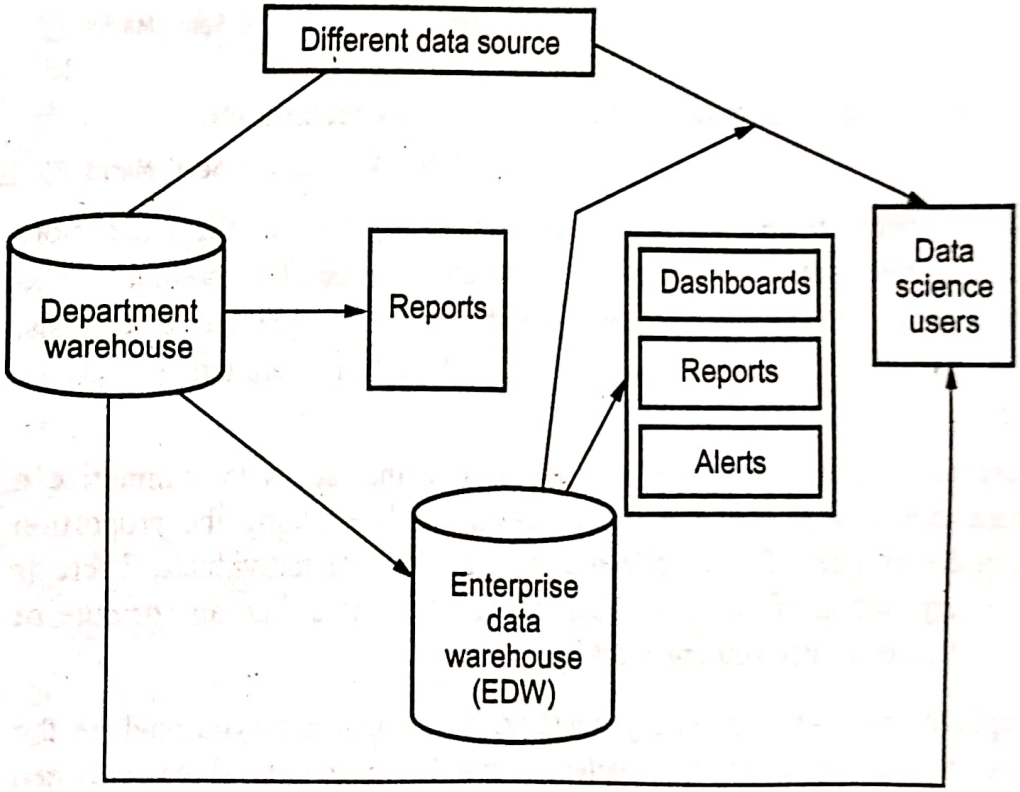


Fig. Q.4.1 Data analytical architecture

- Data to be loaded into the data warehouse. It must be well understood structured and normalized with the appropriate data type. Centralization provides security, backup facility. Also provides significant pre-processing and checkpoints facility before storing data.
- Required level of control on the EDM with additional local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis. Sometime local data marts allow users to do some level of more in-depth analysis.

- Once in the data warehouse, data is read by additional applications across the enterprise for BI and reporting purposes. These are high-priority operational processes getting critical data feeds from the data warehouses and repositories.
- At last, analysts get data provisioned for their downstream analytics. Many times, these tools are limited to in-memory analytics on desktops analyzing samples of data, rather than the entire population of a dataset.

4.3 : Types of Analysis

Q.5 Explain different kinds of big data analysis.

☞ [SPPU : June-22, End Sem, Marks 7]

OR Explain different types of big data analysis techniques.

☞ [SPPU : Dec.-22, End Sem, Marks 8]

Ans. : • There are many types of data analysis. Some of them are more basic in nature, such as descriptive, exploratory, inferential, predictive and causal. Some, however, are more specific, such as qualitative analysis, which looks for things like patterns and colors and quantitative analysis, which focuses on numbers.

- **Descriptive :** A descriptive question is one that seeks to summarize a characteristic of a set of data. For example, determining the proportion of males in a set of data collected from a group of individuals. There is no interpretation of the result itself as the result is a fact, an attribute of the set of data that you are working with.
- **Exploratory :** An exploratory question is one in which you analyze the data to see if there are patterns, trends or relationships between variables. These types of analyses are also called "hypothesis-generating" analyses because rather than testing a hypothesis as would be done with an inferential, causal or mechanistic question, you are looking for patterns that would support proposing a hypothesis.
- **Inferential :** An inferential question would be a restatement of this proposed hypothesis as a question and would be answered by analyzing a different set of data. By analyzing it you are both determining if the association you observed in you exploratory analysis holds in a different sample and whether it holds in a sample that is representative.

- **Predictive :** A predictive question would be one where you ask what types of people will eat a diet high in fresh fruits and vegetables during the next year. In this type of question you are less interested in what causes someone to eat a certain diet, just what predicts whether someone will eat this certain diet.
- **Causal :** A causal question asks about whether changing one factor will change another factor, on average, in a population. Sometimes the underlying design of the data collection, by default, allows for the question that you ask to be causal.
- **Mechanistic :** Finally, mechanistic questions tell us how something happens. For instance, a question that asks how a diet high in fresh fruits and vegetables leads to a reduction in the number of viral illnesses would be a mechanistic question.

Q.6 Explain predictive and descriptive task.

☞ [SPPU : Dec.-18 (End Sem), Marks 5]

Ans. : Predictive task :

- To make prediction, predictive mining tasks performs inference on the current data. Predictive analysis provides answers of the future queries that move across using historical data as the chief principle for decisions.
- It involves the supervised learning functions used for the prediction of the target value. The methods fall under this mining category are the classification, time-series analysis and regression.
- Data modeling is the necessity of the predictive analysis, which works by utilizing some variables to anticipate the unknown future data values for other variables.
- It provides organizations with actionable insights based on data. It provides an estimation regarding the likelihood of a future outcome.
- To do this, a variety of techniques are used, such as machine learning, data mining, modeling and game theory.
- Predictive modeling can, for example, help to identify any risks or opportunities in the future.
- Predictive analytics can be used in all departments, from predicting customer behaviour in sales and marketing, to forecasting demand for operations or determining risk profiles for finance.

Descriptive task :

- Descriptive analytics is the conventional form of business intelligence and data analysis, seeks to provide a depiction or "summary view" of facts and figures in an understandable format, to either inform or prepare data for further analysis.
- Two primary techniques are used for reporting past events : data aggregation and data mining.
- It presents past data in an easily digestible format for the benefit of a wide business audience.
- A set of techniques for reviewing and examining the data set to understand the data and analyze business performance.
- Descriptive analytics helps organisations to understand what happened in the past. It helps to understand the relationship between product and customers.
- The objective of this analysis is to understanding, what approach to take in the future. If we learn from past behaviour , it helps us to influence future outcomes.
- Company reports is an example of descriptive analytics which simply provides a historic reviw of company operations, stakeholders, customers and financials.
- It also helps to describe and present data in such format, which can be easily understood by a wide variety of business readers.

Q.7. Explain the difference between descriptive and predictive model.

Ans. :

Sr. No.	Descriptive model	Predictive model
1.	It use data aggregation and data mining to provide insight into the past and answer.	Use statistical models and forecasts techniques to understand the future and answer.
2.	What has happened ?	What could happened ?
3.	Descriptive analytics is the analysis of past or historical data to understand trends and evaluate metrics over time.	Predictive analytics predicts future trends.

4.	Examples of tools used : Data aggregation and data mining.	Examples of tools used : Machine learning, statistical models and simulation.
5.	Used when user want to sumarize results for all or part of your business.	Used when user want to make an educated guess at likely results.
6.	Limitation : Snapshot of the past, often with limited ability to help guide decisions.	Limitation : Guess at the future, helps inform low complexity decisions.

Q.8 Explain difference between descriptive, predictive and prescriptive analytics.

Ans. :

Descriptive model	Predictive model	Prescriptive model
It use data aggregation and data mining to provide insight into the past and asnwer.	Use statistical models and forecasts techniques to understand the future and answer.	Use optimization and simulation algorithms to advice on possible outcomes and answer.
What has happned ?"	What could happened ?	What should we do ?
Descriptive analytics is the analysis of past or historical data to understand trends and evaluate metrics over time.	Predictive analytics predicts future trends.	Prescriptive analytics showcases viable solutions to a problem and the impaact of considering a solution on future trend.
Examples of tools used : Data aggregation and data mining.	Examples of tools used : Machine learning, statistical models and simulation.	Examples of tools used : Optimization and heuristics.
Used when user want to sumarize results for all or part of your business.	Used when user want to make an educated guess at likely results.	Used when user have importance iterdependent, complex or time-sensitive decisitions to make.

Limitation : Snapshot of the past, often with limited ability to help guide decisions.	Limitation : Guess at the future, helps inform low complexity decisions.	Limitation : Most effective where user have some control over what is being modeled.

4.4 : Data Ingestion from Different Sources

Q.9 Write short note on CVS and JSON.

Ans. : CVS

- A CSV (comma-separated values) file is a simple text file in which information is separated by commas. CSV files are most commonly encountered in spreadsheets and databases.
- Each line in a CSV file is a new record from the set of records contained in the file. Such files are generated when data transfer is intended from one storage system to another. Since all applications can recognize records separated by comma, import of such data files to database is done very conveniently.
- CSV files can be converted to several different file formats using the applications that can open these files. For example, Microsoft Excel can import data from CSV file format and save it to XLS, XLSX, PDF, TXT, XML and HTML file formats.
- CSV file format is known to be specified under RFC4180. It defines any file to be CSV compliant if :

a) Each record is located on a separate line, delimited by a line break (CRLF). For example :

aaa,bbb,ccc CRLF

zzz,yyy,xxx CRLF

b) The last record in the file may or may not have an ending line break. For example :

aaa,bbb,ccc CRLF

zzz,yyy,xxx

JSON

- JavaScript Object Notation (JSON) is used to format data. It is commonly used in Web as a vehicle to describe data being sent between systems.

- JSON is much easier to use with JavaScript than XML. When it comes to Ajax and JavaScript, JSON Web Services are replacing XML Web Services.
- The JSON format is often used for serializing and transmitting structured data over a network connection. It is often used to transmit data between a server and web application, serving as an alternative to XML.
- JSON is based on a subset of JavaScript, containing object and array. Objects contain pairs of property and value. Arrays contain values. A value could be a string, number, object array, true, false or null.
- On average, JSON requires less characters and so less bytes, than the same data in XML. Because it uses JavaScript syntax, it requires less parsing than XML when used in Ajax Applications.

Q.10 How data can be ingested in python ? Write syntax in python for the same. ☞ **[SPPU : June-22, End Sem, Marks 7]**

Ans. : • Data ingestion is the process of obtaining and importing data for immediate use or storage in a database.

- Data can be ingested from many sources like Kafka, Kinesis, or TCP sockets, and can be processed using complex algorithms expressed with high-level functions like map, reduce, join and window. Finally, processed data can be pushed out to file systems, databases, and live dashboards.
- Data Ingestion with Pandas, is the process, of shifting data, from a variety of sources, into the Pandas DataFrame structure. The source of data can be varying file formats such as Comma Separated Data, JSON, HTML webpage table, Excel. Also refer Q.11.

Q.11 What is dataset ? Explain with python syntax of 2 different types of dataset used in big data. ☞ **[SPPU : Dec.-22, End Sem, Marks 6]**

Ans. : • The term data set refers to a file that contains one or more records. The record is the basic unit of information used by a program running on z/OS. Any named group of records is called a data set.

- A data set is a structured collection of data points related to a particular subject. A collection of related data sets is called a database.
- Data sets can be tabular or non-tabular. Tabular data sets contain structured data that is organized by rows and columns. Non-tabular data sets contain unstructured data contained by brackets.

- Set is an unordered collection of simple objects in Python.
- You can use curly braces to give an expression whose value is a set. Python prints sets using curly braces.

```
>>> {1+2, 3, "a"}
{'a', 3}
>>> {2, 1, 3}
{1, 2, 3}
```

- Note that duplicates are eliminated and that the order in which the elements of the output are printed does not necessarily match the order of the input elements.
- Python provides collections, called dictionaries, that are suitable for representing such functions. Conceptually, a dictionary is a set of key-value pairs.
- The keys defined for a dictionary need to be unique. Though values in a dictionary can be mutable or immutable objects, only immutable objects are allowed for keys.
- A dictionary in Python is defined using key-value pairs, separated by commas, enclosed within curly braces. The key and value are separated using a colon. the general syntax of a dictionary is :

  d = {"Key1":"Value1","Key2":"Value2"}

  a. Key and values are separated by a colon
  b. Paris of entries are separated by commas
  c. Dictionary is enclosed within curly braces
- Key and values in the dictionaries can be of any type, but the keys should be unique to that particular dictionary. The empty dictionary is denoted {}.

4.5 : Data Cleaning

Q.12 Explain task of data cleaning. How to handle missing data?

Ans. : Tasks of data cleaning are as follows :

1. Deal with missing values
2. Identify outliers and smooth out noisy data
3. Correct inconsistent data

Missing value : Data is not always available. E.g., many tuples have no recorded value for several attributes, such as customer income in sales data.

- Missing data may be due to
  1. Equipment malfunction
  2. Inconsistent with other recorded data and thus deleted
  3. Data not entered due to misunderstanding
  4. Certain data may not be considered important at the time of entry
  5. Not register history or changes of the data
  6. Missing data may need to be inferred. Missing values may carry some information content: e.g. a credit application may carry information by noting which field the applicant did not complete.

Handling of Missing Data :

- Ignore records(use only cases with all values).
- Usually done when class label is missing as most prediction methods do not handle missing data well.
- Not effective when the percentage of missing values per attribute varies considerably as it can lead to insufficient and/or biased sample sizes.
- Ignore attributes with missing values.
- Use only features (attributes) with all values.
- Fill in the missing value manually.

Q.13 What is data cleaning ? What kind of issues affect the quality of data ?

Ans. : • Data cleaning is the process of detecting and correcting or removing corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

- All data sources potentially include errors and missing values, data cleaning addresses these anomalies.
- Not cleaning data can lead to a range of problems, including linking errors, model mis-specification, errors in parameter estimation and incorrect analysis leading users to draw false conclusions.

- The main data cleaning processes are editing, validation and imputation. Editing and validation are sometimes used synonymously. Editing describing the identification of errors, and validation their correction. The remaining process, imputation, is the replacement of missing values.
- Issues affect the quality of data are as follows :
  1. Invalid values : Some datasets have well-known values, e.g. gender must only have "F" (Female) and "M" (Male). In this case it's easy to detect wrong values.
  2. Formats : The most common issue. It's possible to get values in different formats like a name written as "Name, Surname" or "Surname, Name".
  3. Attribute dependencies : When the value of a feature depends on the value of another feature. For example, if we have some school data, the "number of students" is related to whether the person "is teacher?". If someone is not a teacher he/she can't have any students.
  4. Missing values : Some features in the dataset may have blank or null values.
  5. Misspellings : Incorrectly written values.
  6. Misfiled values : When a feature contains the values of another.

Q.14 Explain data standardization. ☞ **[SPPU : Dec.-22, End Sem, Marks 3]**

Ans. : • Data standardization is a data processing workflow that converts the structure of different datasets into one common format of data.

- Data standardization converts data into a standard format that computers can read and understand. This is important because it allows different systems to share and efficiently use data. Without data standardization, it would not be effortless for different approaches to communicate and exchange information.
- Data standardization is also essential for preserving data quality. When data is standardized, it is much easier to detect errors and ensure that it is accurate. This is essential for making sure that decision-makers have access to accurate and reliable information

4.6 : Data Transformation

Q.15 What is data integration ? Discuss issues of data integration.

Ans. : • Data integration combines data from multiple sources to form a coherent data store. Metadata, correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute toward smooth data integration

- With the increasing volume of data collected through a variety of sources and at a much faster velocity every day, it is very much clear that Data is and has been the most valuable possession.
- Data integration is important as it provides a unified view of the scattered data not only this it also maintains the accuracy of data.
- Issues in Data Integration : While integrating the data we have to deal with several issues.

1. Entity Identification Problem

- As we know the data is unified from the heterogeneous sources then how can we 'match the real-world entities from the data'. For example, we have customer data from two different data source.
- An entity from one data source has **customer_id** and the entity from the other data source has **customer_number**. Now how does the data analyst or the system would understand that these two entities refer to the same attribute? The schema integration can be achieved using metadata of each attribute.

2. Redundancy

- Redundancy is one of the big issues during data integration. Redundant data is an unimportant data or the data that is no longer needed. It can also arise due to attributes that could be derived using another attribute in the data set.
- For example, one data set has the customer age and other data set has the customers date of birth then age would be a redundant attribute as it could be derived using the date of birth.
- The redundancy can be discovered using correlation analysis. The attributes are analyzed to detect their interdependency on each other thereby detecting the correlation between them.

Q.16 Explain different data transformation techniques.

☞ **[SPPU : June-22, Dec.-22, End Sem, Marks 3]**

Ans. : • In data transformation, the data are transformed or consolidated into forms appropriate for mining.

- Data transformation can involve the following :

1. **Smoothing** : It removes noise from the data. Such techniques include binning, regression, and clustering.
2. **Aggregation** : An aggregation or summary operation is applied to the data.
3. **Generalization** of the data, where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.
4. **Normalization** : The attribute data are scaled so as to fall within a small specified range.
5. **Attribute construction** : New attributes are constructed and added from the given set of attributes to help the mining process.

Q.17 Explain mean, mode and variance and standard deviation with suitable example. ☞[SPPU : Dec.-22, End Sem, Marks 8]

Ans. : Mode :

It is the value of maximum frequency. It occurs most frequency.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} h$$

L = Lower limit of the class containing mode

Δ_1 = Excess of modal frequency over frequency of preceding class

Δ_2 = Excess of modal frequency over following class

h = Size of modal class

Mean :

Let $x_1, x_2, x_3, \ldots, x_n$ be the set 'n' values of the variate, then arithmetic mean or mean is given as,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

- In the frequency distribution if $x_1, x_2, \ldots, x_n$ are the midvalues of the class intervals with frequencies $f_1, f_2, \ldots, f_n$ respectively, then we have,

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum f_i x_i}{\sum f_i}$$

Variance :

- The variance is a measure of variability that utilizes all the data. It is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$) for a sample, μ for a population.
- The variance is the average of the squared between each data value and the mean.

$$\text{Sample variance : } S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\text{Population variance : } \sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Standard Deviation :

- The standard deviation of a data set is the positive square root of the variance. It is measured in the same in the same units as the data, making it more easily interpreted than the variance.
- The standard deviation is computed as follows :

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

$$\text{Sample standard deviation} = S = \sqrt{S^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

Q.18 Minimum salary is ₹ 20, 000 and maximum salary is ₹ 1,70,000 Map the salary ₹ 1,00,000 in new range of ₹ (60, 000, 2,60,000) using min-max normalization method. 2) If mean salary is ₹ 54,000 and standard deviation is ₹ 16,000 then find z score value of ₹ 73,600 salary.

Ans. :

Solution 1 :

Old range = (20000, 1,70,000)

max = 1,70,000

min = 20000

New range = (60000, 260000)

new_max = 260000

$$\text{new\_min} = 60000$$

$$V_i = 100000$$

$$V_i' = [\{V_i - \min)(\max - \min)\} \times \{\text{new\_max} - \text{new\_min})\}] + \text{new\_min}$$

$$= [\{(80000/150000) \times 200000\}] + 60000$$

$$= [106666] + 60000 = 166666$$

Salary ₹ 100000 in old range is equal to salary ₹ 166666 in the new range.

Solution 2 :

$$\text{mean} = ₹\ 54,000$$

$$\text{Standard deviation} = ₹\ 16,000$$

$$\text{Z-score value of } 76,300 = \frac{(76,300 - \text{mean})}{\text{Standard deviation}} = \frac{(54000)}{16000}$$

$$= \frac{54000}{16000} = 3.375$$

Z-score value of ₹ 73,600 salary is 3.375

Q.19 Use min-max normalization method to normalize the following group of data by setting min = 0 and max = 1, 200, 300, 400, 600, 1000.

Ans. : i) Min-max normalization by setting min = 0 and max = 1.

Original data	200	300	400	600	1000
0, 1 normalized	0	0.125	0.25	0.5	1

ii) Z-score normalization

Original data	200	300	400	600	1000
0, 1 normalized	– 1.06	– 0.7	– 0.35	0.35	1.78

4.7 : Handling Categorical Data

Q.20 Explain various methods used for generation of concept hierarchies for categorical data.

Ans. : • Categorical data are discrete data. Categorical attributes have a finite number of distinct values, with no ordering among the values.

- Example : geographic location, job category, and item type.
- Various methods are used for the generation of concept hierarchies for categorical data :

a) Specification of a partial ordering of attributes explicitly at the schema level by users or experts

- Example : A relational database or a dimension location of a data warehouse may contain the following group of attributes : street, city, province or state, and country.
- A user or expert can easily define a concept hierarchy by specifying ordering of the attributes at the schema level.
- A hierarchy can be defined by specifying the total ordering among these attributes at the schema level, such as : *street < city < province or state < country*

b) Specification of a portion of a hierarchy by explicit data grouping

- We can easily specify explicit groupings for a small portion of intermediate-level data.
- For example, after specifying that area and country form a hierarchy at the schema level, a user could define some intermediate levels manually, such as: {India, Maharashtra, Pune} < SPPU.

c) Specification of a set of attributes, but not of their partial ordering

- A user may specify a set of attributes forming a concept hierarchy, but omit to explicitly state their partial ordering.
- The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept.
- Example : Suppose a user selects a set of location-oriented attributes, street, country, state, and city, from the database, but does not specify the hierarchical ordering among the attributes.
- Fig. Q.20.1 shows automatic generation of a schema concept hierarchy based on the number of distinct attribute values.

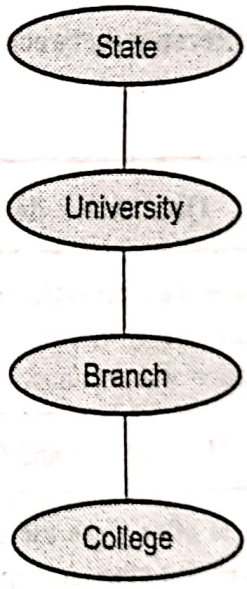


Fig. Q.20.1 Automatic generation of a schema concept hierarchy

Q.21 What is qualitative data ?

Ans. : • **Qualitative data** provides information about the quality of an object or information which cannot be measured. Qualitative data cannot be expressed as a number. Data that represent nominal scales such as gender, economic status, religious preference are usually considered to be qualitative data.

- **Qualitative data** is data concerned with descriptions, which can be observed but cannot be computed. Qualitative data is also called categorical data. Qualitative data can be further subdivided into two types as follows :

  1. Nominal data
  2. Ordinal data

Q.22 What is quantitative data ?

Ans. : • **Quantitative data** is the one that focuses on numbers and mathematical calculations and can be calculated and computed.

- **Quantitative data** are anything that can be expressed as a number or quantified. Examples of quantitative data are scores on achievement tests, number of hours of study or weight of a subject. These data may be represented by ordinal, interval or ratio scales and lend themselves to most statistical manipulation.
- There are two types of quantitative data : Interval data and Ratio data.

Q.23 Difference between qualitative data and quantitative data.

Ans. :

Qualitative data	Quantitative data
Qualitative data provides information about the quality of an object or information which cannot be measured	Quantitative data relates to information about the quantity of an object; hence it can be measured
Types : Nominal data and Ordinal data	Types : Interval data and Ratio data
Narratives often make use of adjectives and other descriptive words to refer to data on appearance, color, texture, and other qualities	Measure's quantities such as length, size, amount, price, and even duration.

They are descriptive rather than numerical in nature	Expressed in numerical form.
For example : • The team is well prepared. • The leaf feels waxy. • The river is peaceful.	For example : • The team has 7 players. • The leaf weighs 2 ounces. • The river is 25 miles long.

4.8 : Hive Data Analytics

Q.24 Draw and explain architecture of HIVE.

☞ [SPPU : June-22, Dec.-22, End Sem, Marks 7]

Ans. : • Hive allows users to simultaneously access data and, at the same time, increases the response time, i.e., the time a system or a functional unit takes to react to a given input. In fact, Hive typically has a much faster response time than most other types of queries.

• Fig Q.24.1 shows architecture of Hive.

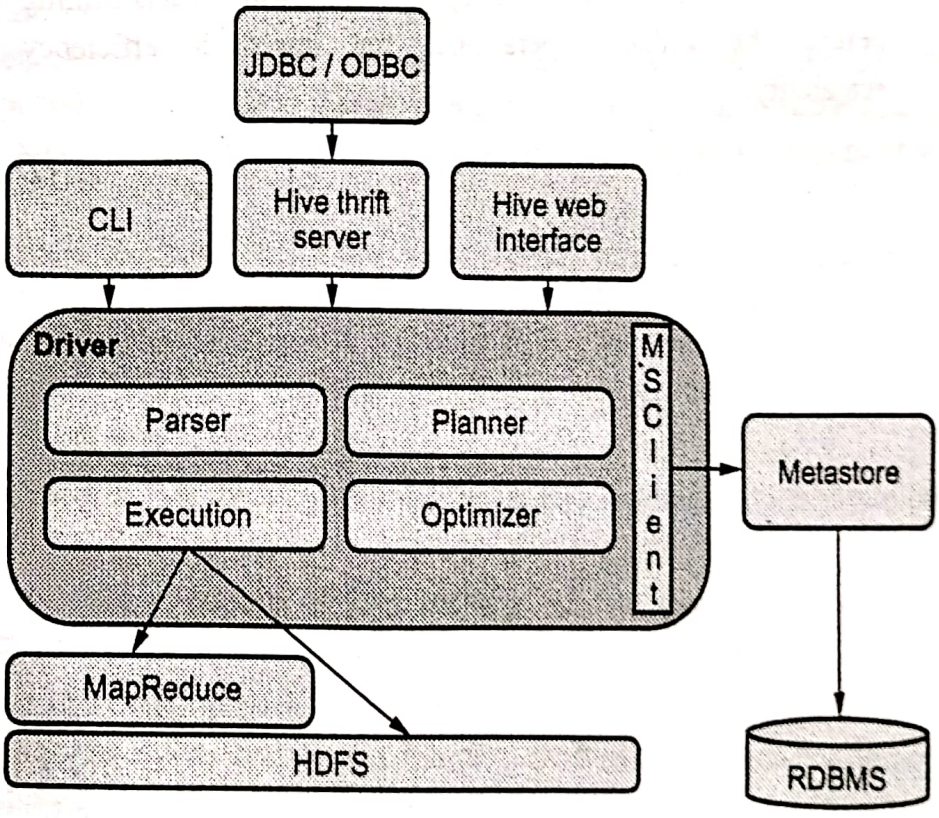


Fig. Q.24.1 Architecture of hive

- Major Components of Hive Architecture

1. **Metastore :** It is the repository of metadata. This metadata consists of data for each table like its location and schema. It also holds the information for partition metadata. The metadata keeps track of the data, replicates it, and provides a backup in the case of data loss.
2. **Driver :** The driver receives HiveQL statements and works like a controller. It monitors the progress and life cycle of various executions by creating sessions. The driver stores the metadata that is generated while executing the HiveQL statement. When the reducing operation is completed by the MapReduce job, the driver collects the data points and query results.
3. **Compiler :** The compiler is assigned with the task of converting a HiveQL query into a MapReduce input. It includes a method to execute the steps and tasks needed to let the HiveQL output as needed by MapReduce.
4. **Optimizer :** This performs various transformation steps for aggregation and pipeline conversion by a single join for multiple joins. It also is assigned to split a task while transforming data, before the reduce operations, for improved efficiency and scalability.

END...

Unit V

5 Big Data Visualization

5.1 : Introduction to Data Visualization

Q.1 Explain data visualization with the help of example? What are the advantages of data visualization ☞ **[SPPU : Dec.-22, End Sem, Marks 8]**

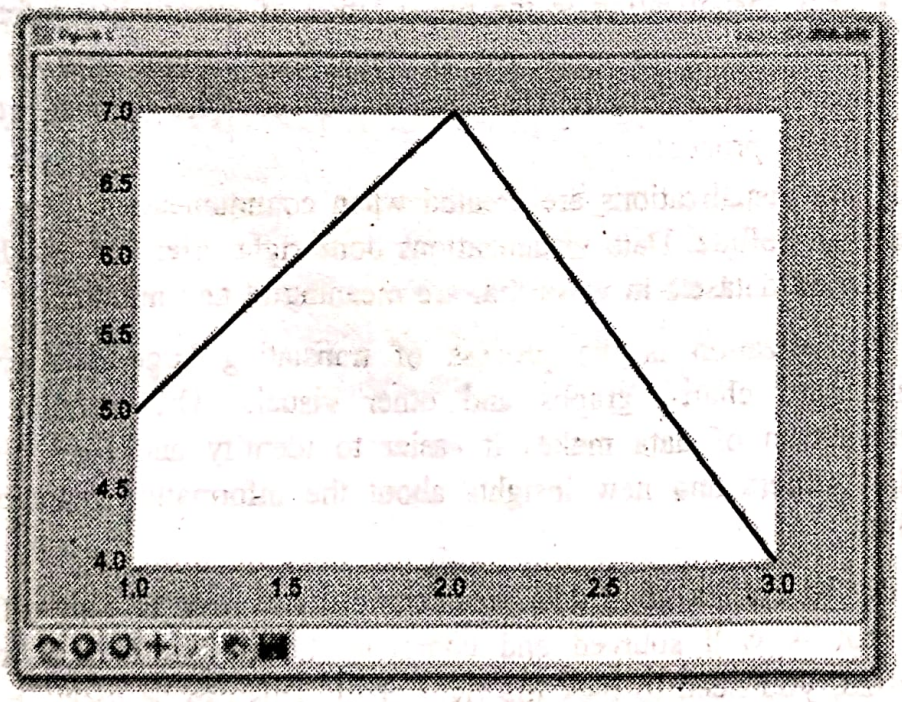
Ans. : • Data visualization is the presentation of quantitative information in a graphical form. In other words, data visualizations turn large and small datasets into visuals that are easier for the human brain to understand and process.

- Good data visualizations are created when communication, data science and design collide. Data visualizations done right offer key insights into complicated datasets in ways that are meaningful and intuitive.
- Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers and new insights about the information represented in the data.
- In order to craft a good data visualization, you need to start with clean data that is well sourced and complete. Once your data is ready to visualize, you need to pick the right chart. This can be tricky, but there are many resources available to help you choose the right type of chart for your data.
- A graph is simply a visual representation of numeric data. MatplotLib supports a large number of graph and chart types.
- MatplotLib is popular Python package used to build plots. MatplotLib can also be used to make 3D plots plots and animations.
- Line plots can be created in Python with MatplotLib's pyplot library. To build a line plot, first import MatplotLib. It is a standard convention to import MatplotLib's pyplot library as plt.

- To define a plot, you need some values, the matplotLib. pyplot module and an idea of what you want to display.

```
import MatplotLib.pyplot as plt
plt.plot([1, 2,3], [5,7,4])
plt.show( )
```

- The plt.plot will "draw" this plot in the background, but we need to bring it to the screen when we're ready, after graphing everything we intend to.
- plt.show() : With that, the graph should pop up. If not, sometimes it can pop under or you may have gotten an error.
- Your graph should look like :



- This window is matplotlib window, which allows us to see our graph, as well as interact with it and navigate it.
- **Three principal drivers of this technology :**
  1. **Visual :** Data are represented in a graphic/visual format.
  2. **Insight :** Data visualization, helps manager to understand data immediately and provides advice and suggestions on the possible actions to take.
  3. **Sharing :** Advice and suggestions on the possible actions can be easily shared across the company which will lead to a consequent.

- To fully document graph, user usually have to resort to labels, annotations and legends. Each of these elements has a different purpose, as follows :
  1. **Label :** Make it easy for the viewer to know the name or kind of data illustrated.
  2. **Annotation :** Help extend the viewer's knowledge of the data, rather than simply identify it.
  3. **Legend :** Provides cues to make identification of the data group easier.
- Benefits of data visualization :
  1. Constructing ways in absorbing information. Data visualization enables users to receive vast amounts of information regarding operational and business conditions.
  2. Visualize relationships and patterns in businesses.
  3. More collaboration and sharing of information
  4. More self-service functions for the end users.
- Big data visualization is important because :
  1. It provides clear knowledge about patterns of data.
  2. Detects hidden structures in data.
  3. Identify areas that need to be improved.
  4. It helps us to understand which products to place where.
  5. Clarify factors which influence human behaviour.

Q.2 Explain role of visualization in big data analytics.

☞ **[SPPU : June-22, End Sem, Marks 3]**

Ans. : Refer Q.1.

Q.3 Explain challenges in big data visualization.

☞ **[SPPU : June-22, End Sem, Marks 7]**

OR What are the major challenges in visualizing the big data and how to overcome these challenges. ☞ **[SPPU : May-18, End Sem, Marks 8]**

OR Explain various challenges in big data visualization and explain the mechanism to overcome the challenges.

☞ **[SPPU : Dec.-18, End Sem, Marks 8]**

Ans. : • Big data analytics plays a key role through reducing the data size and complexity in big data applications. Visualization is an important approach to helping big data get a complete view of data and discover data values.

- Scalability and dynamics are two major challenges in visual analytics.
- Volume : The methods are developed to work with an immense number of datasets and enable to derive meaning from large volumes of data.
- Variety : The methods are developed to combine as many data sources as needed.
- Velocity : With the methods, businesses can replace batch processing with real - time stream processing.
- Value : The methods not only enable users to create attractive info graphics and heat maps, but also create business value by gaining insights from big data.
- Visualization of big data with diversity and heterogeneity (structured, semi - structured and unstructured) is a big problem. Speed is the desired factor for big data analysis.
- There are also following problems for big data visualization :
  1. **Visual noise :** Most of the objects in the dataset are too relative to each other. Users cannot divide them as separate objects on the screen.
  2. **Information loss :** Reduction of visible data sets can be used, but leads to information loss.
  3. **Large image perception :** Data visualization methods are not only limited by aspect ratio and resolution of device, but also by physical perception limits.
  4. **High rate of image change :** Users observe data and cannot react to the number of data changes or its intensity on display.
  5. **High performance requirements :** It can be hardly noticed in static visualization because of lower visualization speed requirements - high performance requirements.
- Potential solutions to some challenges and problems of data visualization are as follows :
  a) **Meeting the need for speed :** One solution is to deal with hardware, by increasing memory and massive parallel processing can be used.

b) **Understanding the data :** select proper domain and Expertise is solution.

c) **Displaying meaningful results :** solution is to cluster data into smaller groups that are visible effectively.

d) **Dealing with outliers :** solution is to remove outliers from data or create separate chart for outliers

Q.4 Write two data visualization functions from matplotlib.

☞ [SPPU : June-22, End Sem, Marks 3]

Ans. : Refer Q.1.

Q.5 How data visualization help big data analytics ?

☞ [SPPU : Dec.-22, End Sem, Marks 4]

Ans. : • Big data Visualization is a visual representation of big data. Visualization techniques vary depending on the goal of the illustration. It could be as simple as line charts, histograms and pie charts or a bit complex like scatter plot, heat maps, tree maps, etc. Visualization of big data can also be done in 3-Dimensional graphs, based on the use case.

- Big data analytics makes it possible for organizations to sift through captured data in order to produce viable, actionable conclusions related to causes, processes, and trends.
- Generally, when big data analytics and algorithms are applied to data sets, the results are meant for the decision makers. The best part of big data visualization tools is that they are capable of capturing data sets in the visual format without loss of accuracy. One can control the factors like accuracy, precision, level of aggregation that is required to serve the purpose.
- Big Data visualization techniques helps in following ways :

a) Enable decision - Makers to understand what the amount of data means very quickly;

b) Capture trends - The use of appropriate techniques can make it easy to recognize this information;

c) Reveal patterns - identify correlations and unexpected connections that could not be found with specific questions;

d) Provide a highly effective way to communicate any insights that surfaces to others.

5.2 : Types of Data Visualization

Q.6 Explain any four types of data visualization with example.

☞ [SPPU : Dec.-22, End Sem, Marks 8]

OR Describe different types of data visualization.

☞ [SPPU : Dec.-19, End Sem, Marks 8]

Ans. : Various types of data visualization are as follows :

1. Multidimensional : 2D Area
2. Temporal
3. Hierarchical
4. Network

Sr. No.	Types	Descriptions
1.	Multidimensional : 2D Area	1. **Cartogram :** It distorts map space to express information such as travel time or population of the alternate variable. It mainly consists of two main types : Area based and distance - based cartograms. 2. **Choropleth :** It is used to represent the statistical measurement such as population density rate or website visitors count per city. 3. **Dot distortion map :** It uses a dot symbol to represent a feature on the map, depending on the visual scatter for displaying spatial patterns.
2.	Temporal	1. **Pie chart :** The circle is divided into sectors to represent numeric proportions. The length of the arc and angle length of the sector is proportional to the particular quantity it represents. 2. **Histogram :** In a histogram, the data are grouped into ranges (e.g. 10 - 19, 20 - 29) and then plotted as connected bars. Each bar represents a range of data. The width of each bar is proportional to the width of each category and the height is proportional to the frequency or percentage of that category. 3. **Scatter plot :** It displays collection of all the points for the set of data limited only for two values.

3.	Hierarchical	1. **Dendrogram :** It is nothing but a tree diagram used to represent clusters generated by hierarchical clustering. 2. **Ring chart :** It is a multi-level pie chart which is represented by the nested circles. 3. **Tree diagram :** It represents the data or the hierarchy in the graph form, which can be visualized from left to right or top to bottom.
4.	Network	1. **Alluvial diagram :** It is a flow diagram which visualizes over time changes in network structure. 2. **Node link diagram :** In this representation, nodes are visualized as dots whereas links are represented as line segments to display the data connection. 3. **Matrix :** It shows relation between two to four groups of information and gives information regarding the same.

5.3 : Data Visualization Techniques

Q.7 Explain different techniques of big data visualization.

☞ **[SPPU : June-22, End Sem, Marks 7]**

Ans. : Different data visualization techniques are line graph, Box plots, Histograms, Heat maps, scatter diagram etc. Also refer Q.6.

Q.8 Explain scatter plot, histogram and heat map with example.

☞ **[SPPU : June-22, End Sem, Marks 7]**

Ans. : Scatter plot :

- Scatter diagram is also called scatter plot, X-Y graph. The scatter plot is the model of data visualization depicting two sets of unconnected dots as parameter values.
- Scatter plots which use horizontal and vertical axes to plot data points and display how much one variable is affected by another. The position of each dot on the horizontal and vertical axis indicates values for an individual data point.

- Fig. Q.8.1 shows scatter plots of two variables.

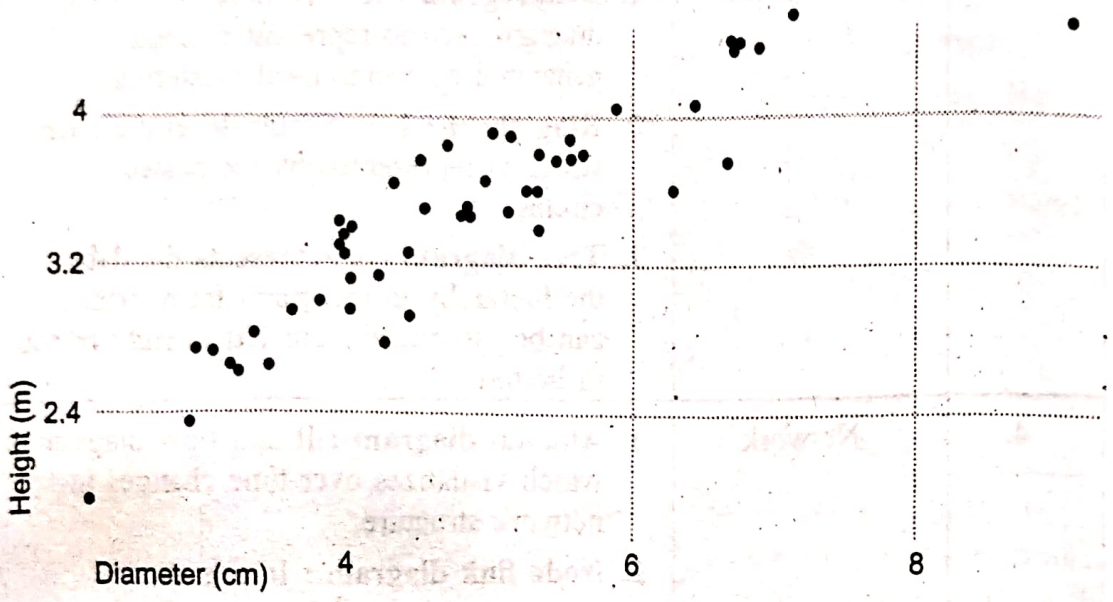


Fig. Q.8.1 Scatter plot

- The example scatter plot above shows the diameters and heights for a sample of fictional trees. Each dot represents a single tree; each point's horizontal position indicates that tree's diameter (in centimeters) and the vertical position indicates that tree's height (in meters).

Histogram :

- In a histogram, the data are grouped into ranges (e.g. 10-19, 20-29) and them plotted as connected bars. Each bar represents a range of data. The width of each bar is proportional to the width of each category and the height is proportional to the frequency or percentage of that category.
- It provides a visual interpretation of numerical data by showing the number of data points that fall within a specified ranges of values called "bins".
- Fig. Q.8.2 shows histogram. (See Fig. Q.8.2 on next page.)
- Histograms can display a large amount of data and the frequency of the data values. The median and distribution of the data can be determined by a histogram. In addition, it can show any outliers or gaps in the data.
- Matplotlib provides a dedicated function to compute and display histogram : plt.hist()

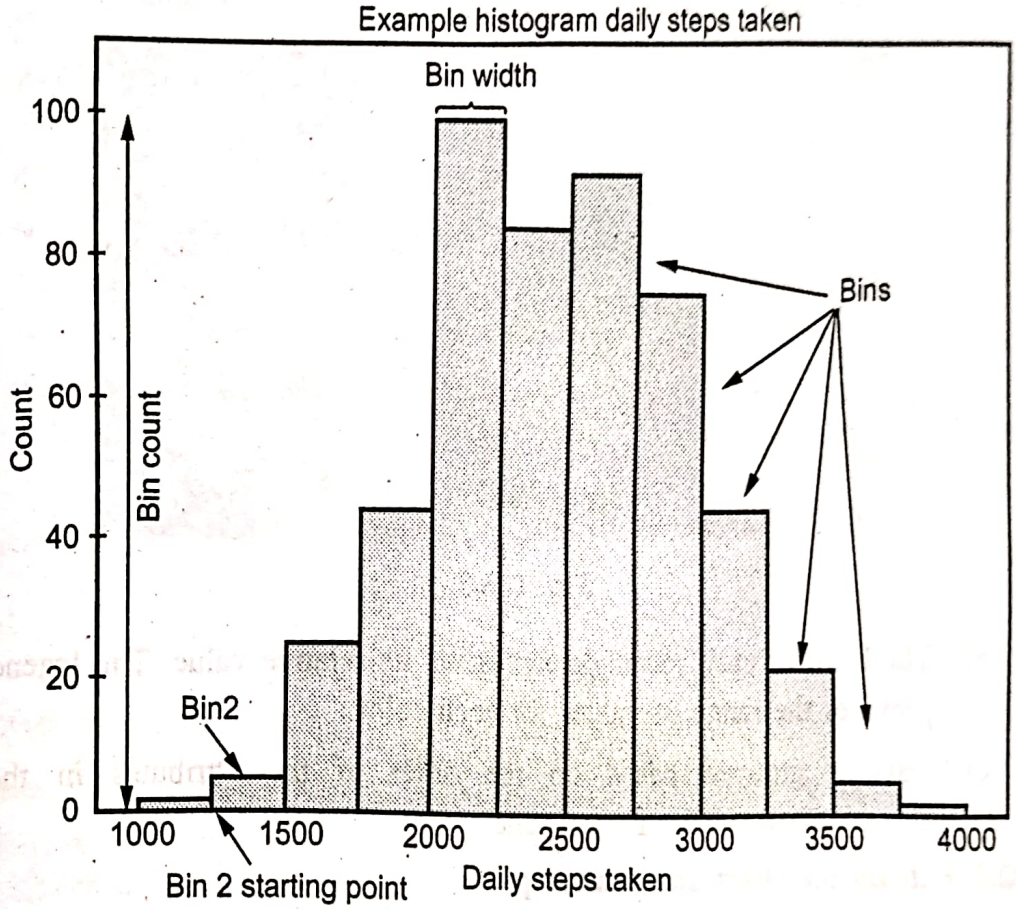


Fig. Q.8.2 Histogram

Heat map :

- Use a heat map visualization to visualize the relationship between columns, represented in a matrix type view. A heat map visualization uses color and intensity of the color to show the relationship between two columns.
- A heat map visualization is a combination of nested, colored rectangles, each representing an attribute element. Heat Maps are often used in the financial services industry to review the status of a portfolio.
- For example, this heat map visualization shows the average customer lifetime value by gender and education.
- The rectangles contain a wide variety and many shadings of colors, which emphasize the weight of the various components. In a heat map visualization :

a) The size of each rectangle represents its relative weight. The legend provides information about the minimum and maximum values.

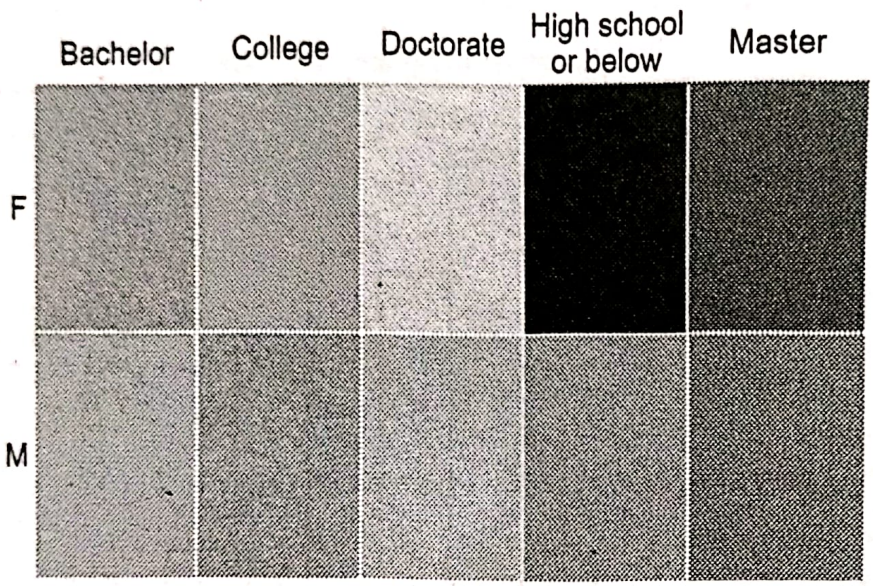


Fig. Q.8.3

b) The color of each rectangle represents its relative value. The legend provides the range of values for each color.

c) Data is grouped based on the order of the attributes in the Grouping area of the Editor panel.

Q.9 Explain pie chart and scatter plot.

☞ **[SPPU : Dec.-19, End Sem, Marks 8]**

Ans. : Pie chart :

- A type of graph in which a circle is divided into sectors that each represent a proportion of the whole. Each sector shows the relative size of each value.
- A pie chart displays data, information and statistics in an easy to read "pie slice" format with varying slice sizes telling how much of one data element exists.
- Pie chart is also known as circle graph. The bigger the slice, the more of that particular data was gathered. The main use of a pie chart is to show comparisons. Fig. Q.9.1 shows pie chart.
- Various applications of pie charts can be found in business, school and at home.

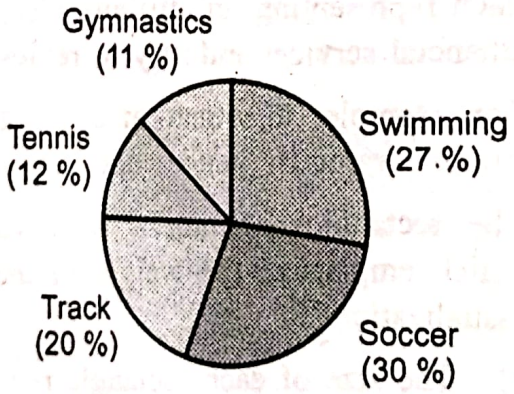


Fig. Q.9.1 Pie chart

For business pie charts can be used to show the success or failure of certain products or services.

- At school, pie chart applications include showing how much time is allotted to each subject. At home pie charts can be useful to see expenditure of monthly income in different needs.
- Reading of pie chart is as easy as figuring out which slice of an actual pie is the biggest.
- Pie charts can be drawn using the function pie() in the pyplot module. The below python code example draws a pie chart using the pie() function. By default the pie() fucntion of pyplot arranges the pies or wedges in a pie chart in counter clockwise direction.

Scatter Plot : Refer Q.8.

5.4 : Visualizing Big Data

Q.10 Write short note on : Visualizing Big Data.

Ans. : • Big data visualization is the process of displaying data in charts, graphs, maps and other visual forms.

- There are various analytical techniques used in big data processing in order to extract, collect, store, process and analyze the huge amount of data coming very fast with the different variety.

1. Machine Learning :

- A machine learning algorithm then takes these examples and produces a program that does the job. The program produced by the learning algorithm may look very different from a typical hand-written program. It may contain millions of numbers. If we do it right, the program works for new cases as well as the ones we trained it on.
- Main goal of machine learning is to devise learning algorithms that do the learning automatically without human intervention or assistance. The machine learning paradigm can be viewed as "programming by example." Another goal is to develop computational models of the human learning process and perform computer simulations.
- The goal of machine learning is to build computer systems that can adapt and learn from their experience.

- Algorithm is used to solve a problem on computer. An algorithm is a sequence of instructions. It should carry out to transform the input to output.
- For example, The addition of four numbers is carried out by giving four number as input to the algorithm and output is the sum of all four numbers. For the same task, there may be various algorithms. It is interesting to find the most efficient one, requiring the least number of instructions or memory or both.
- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.
- Machine learning provides business insight and intelligence. Decision makers are provided with greater insights into their organizations. This adaptive technology is being used by global enterprises to gain a competitive edge.
- Machine learning algorithms discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system.
- Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. The task of the supervised learner is to predict the output behaviour of a system for any set of input values, after an initial training phase.
- In contrast to supervised learning, unsupervised or self-organized learning does not require an external teacher. During the training session, the neural network receives a number of different input patterns, discovers significant features in these patterns and learns how to classify input data into appropriate categories.
- Unsupervised learning algorithms aim to learn rapidly and can be used in real - time. Unsupervised learning is frequently employed for data clustering, feature extraction etc.
- Reinforcement learning : This is an advanced machine learning technique. This is based on probability theory where mapping can be done based on input received and changes based on the environment around it.
- Deep learning : This is also advanced machine learning technique which has multiple processing layers so as to produce non-linear response

based on input data. There are so many small processors called as **neuron working** parallel in data processing.

- Predictive analytics : This technique refers to prediction based on past experience and it uses both data mining and machine learning.
- Association rule learning : This is used to identify interesting relations between different attributes from large datasets.

5.5 : Tools used in Data Visualization

Q.11 Explain different tools for data visualization.

☞ [SPPU : June-22, End Sem, Marks 7]

OR Explain different data visualization tools.

☞ [SPPU : Dec.-19, 22, End Sem, Marks 6]

Ans. : Data visualization tools include Google Charts, Tableau, Grafana, Chartist.js, FusionCharts, Datawrapper, Infogram, ChartBlocks, and D3.js.

Pentato :

- Pentaho tightly couples data integration with full business analytics to solve data integration challenges while providing business analytics in a single, seamless platform.
- Pentaho's Java-based data integration engine integrates with the MapRHadoop cache for automatic deployment as a MapReduce task across every data node in a Hadoop cluster, making use of the massively parallel processing and high availability of Hadoop.
- Pentaho's open source heritage drives our continued innovation in a modern, integrated, embeddable platform built for the future of analytics, including diverse and big data requirements.
- Within a single platform it provides visual tools to extract and prepare our data plus the visualizations and analytics that will change the way we run our business.
- Pentaho's modern, simplified and interactive approach empowers business users to access, discover and blend all types and sizes of data. With a spectrum of increasingly advanced analytics, from basic reports to predictive modeling, users can analyze and visualize data across multiple dimensions, all while minimizing dependence on IT.

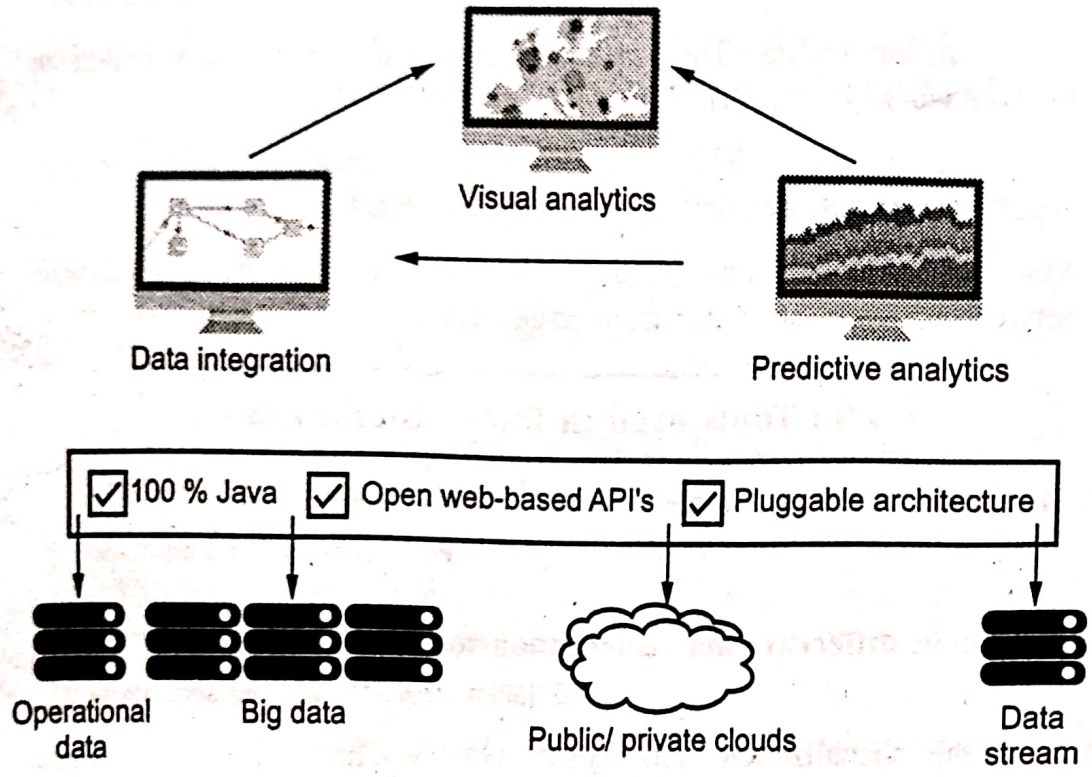


Fig. Q.11.1

- The Business analytics platform is a web application that allows users to publish and manage reports within an enterprise business intelligence system.
- The Business analytics platform offers many capabilities, including the management and execution of Pentaho reports. By combining Pentaho reporting and pentaho's business analytics platform, information technologists may utilize Pentaho reporting in their environment without writing any code.
- In addition to the publishing and execution of reports, the open source Business analytics platform allows for scheduling, background execution, security, and much more.

Datameer :

- Datameer's Flipside provides simple, highly accessible, visual data profiling that lets users easily spot outliers in data, quickly and early in the analytics process.
- Datameer runs natively on Hadoop. Fig. Q.11.2 shows all Datameer functionality occurs across three major components.

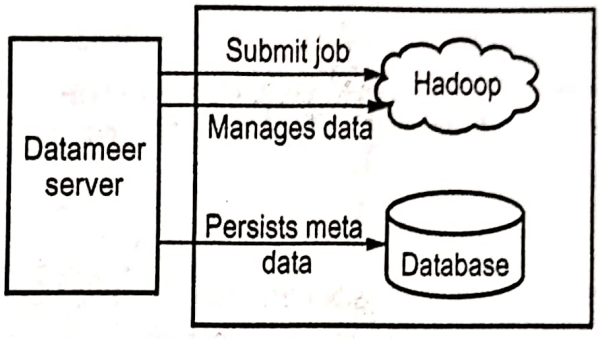


Fig. Q.11.2

- The Datameer server : Server is also called Conductor. This server orchestrates all work and manages the configuration of all jobs performed on the Hadoop cluster. It also hosts the web app that lets users interact via the software's web UI. All processing done during the design of a workbook in real time on the Datameer server. Datameer provides real-time feedback during the design phase using intelligent previews generated by our Smart Sampling technology.
- Database for metadata storage : Datameer uses a database to store all metadata.
- Hadoop cluster : The Hadoop cluster provides persistent storage for all data, pre-views and other job artifacts, as well as a big data processing framework for executing long-running operations. Fundamental to the design of Datameer software is the fact that all resource-intensive processes are submitted to Hadoop clusters. This approach allows Datameer to scale up and scale out easily by distributing work across the entire Hadoop cluster.

Q.12 List the conventional data visualization tools. Explain any two.

☞ **[SPPU : Dec.-19, 22, End Sem, Marks 6]**

Ans. : • Various tools in data visualizations are Plotly, DataHero, Chart.js., Tableau, Raw, Dygraphs, ZingChart and InstantAtlas.

1. DataHero

- DataHero is the world's first truly self-service data visualization and data dashboard platform.
- Turn siloed data into deep insights with powerful data visualization tools that everyone on your team can understand.

- With DataHero, you can connect easily to the cloud services you rely on most, create stunning visualizations, and build automated KPI dashboards so your team can get deeper insights and make better decisions.

2. Dygraphs

- Dygraphs is an open source JavaScript library that produces interactive, zoomable charts of time series. It is designed to display dense data sets and enable users to explore and interpret them.
- It can handle large data sets with millions of plot points. It works in all browsers and zooms down for mobile devices.
- Some of the features of dygraphs :
  1. Plots time series without using an external server or flash
  2. Works in Internet Explorer (using excanvas)
  3. Lightweight (69 kb) and responsive
  4. Displays values on mouseover, making interaction easily discoverable
  5. Supports error bands around data series
  6. Interactive zoom
  7. Displays annotations on the chart
  8. Adjustable averaging period
  9. Can intelligently chart fractions
  10. Customizable click-through actions
  11. Compatible with the Google Visualization API
  12. Intelligent defaults make it easy to use

Q.13 Discuss about JasperReport.

Ans. : • JasperReports is a powerful open source reporting package, but generating reports with data from multiple sources is hard and often impossible without the enterprise version.

- Fig Q.13.1 shows JasperReport.
- JasperReports is an open source java reporting engine. JasperReports is a Java class library, and it is meant for those Java developers who need to add reporting capabilities to their applications.
- The main purpose of JasperReports is to create page oriented, ready to print documents in a simple and flexible manner.
- JasperReports Server is a stand-alone and embeddable reporting server.

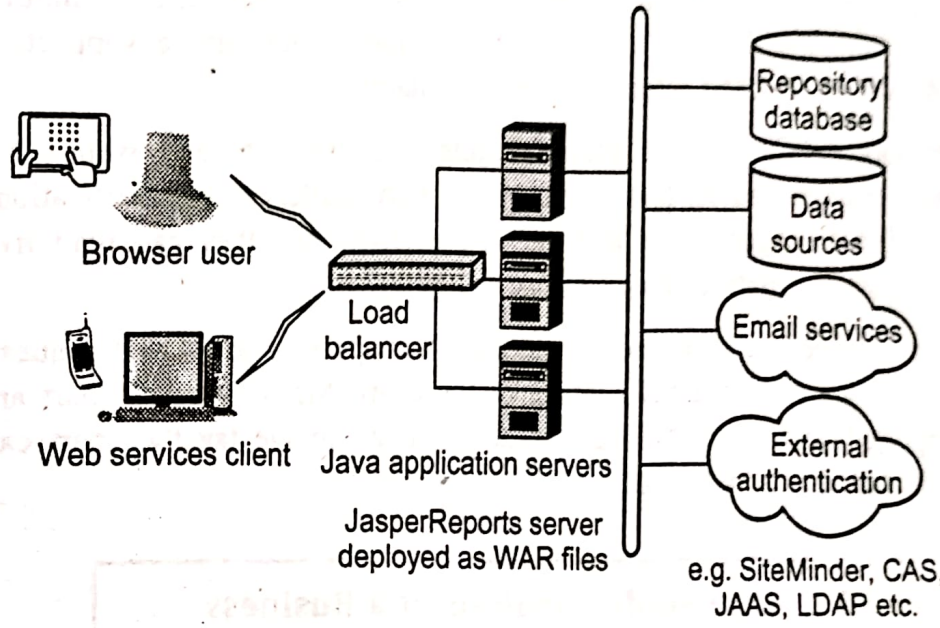


Fig. Q.13.1

- It provides reporting and analytics that can be embedded into a web or mobile application as well as operate as a central information hub for the enterprise by delivering mission critical information on a real-time or scheduled basis to the browser, mobile device, or email inbox in a variety of file formats.
- JasperReports Server is optimized to share, secure, and centrally manage your Jaspersoft reports and analytic views.
- Datasources are structured data container. While generating the report, JasperReports engine obtains data from the datasources. Data can be obtained from the databases, XML files, arrays of objects, and collection of objects.
- JasperReports has a feature <style> which helps to control text properties in a report template. This element is a collection of style settings declared at the report level.
- Properties like foreground color, background color, whether the font is bold, italic, or normal, the font size, a border for the font, and many other attributes are controlled by <style> element.

Q.14 What is Google Chart API ?

Ans. : • The Google Chart API is an interactive Web service that creates graphical charts from user-supplied data.

- Google servers create a PNG image of a chart from data and formatting parameters specified by a user's HTTP request. The service supports a wide variety of chart information and formatting.
- The Google Chart Tools enable adding live charts to any web page. They provide advantages such as a rich gallery of visualizations provided as image charts and interactive charts and they can read live data from a variety of data sources.
- Users embed the data and formatting parameters in an HTTP request, and Google returns a PNG image of the chart. Many types of chart are supported, and by making the request into an image tag the chart can be included in a web page.

5.6 : Case Study : Analysis of a Business Problem of Zomato using Visualization

Q.15 Discuss analysis of a business problem of Zomato using visualization.

Ans. : • Founded in 2008 Zomato is a major food delivery aggregator with a markdown cap of 1 Trillion INR. It started as Foodiebay, a restaurant recommendation product, at its peak, it has 35000 menus and ₹ 60 Lakh monthly revenue. Foodiebay.com reroutes to zomato.com now.

- Zomatlo offered customer : A solution to have access to all the restaurants through a database, the type of food menu, location of the eateries and most importantly their feedback and the reviews.
- Zomato Kitchens under the banner of Zomato Infrastructure Services provides cloud kitchens to the best and reliable restaurants only. It provides kitchen equipment, tech stack, POS, and delivery, and tracking systems. Zomato earns a share of restaurants profit, thus making sure it's a win-win situation.
- Zomato is dynamic on Instagram, Facebook and Twitter. Beginning at July 2019, it has 154 K followers on Instagram, 1,899,405 supporters and 1.42 Million lovers on Twitter.
- The dataset of restaurant was carried out by the researchers based on Zomato registered restaurant through Zomato API and it is publicly availabe on "www.kaggle.com". The dataset has multiple different variety of columns which are used to analyze and identify which city

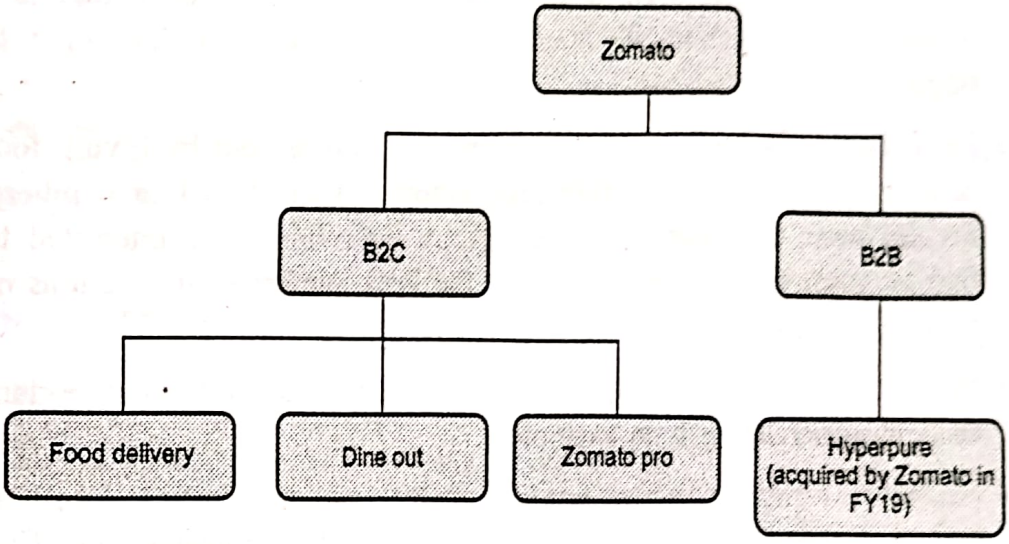


Fig. Q.15.1

has highest number of good restaurants based on ratings, votes and analyzing pattern of expensive restaurant with quality of food.

5.7 : Analytical Techniques used in Big Data Visualization

Q.16 Explain analytical techniques used in big data visualization.

Ans. : • There are various analytical techniques used in big data processing in order to extract, collect, store, process and analyze the huge amount of data coming very fast with the different variety.

Machine Learning :

- A machine learning algorithm then takes these examples and produces a program that does the job. The program produced by the learning algorithm may look very different from a typical hand-written program. It may contain millions of numbers. If we do it right, the program works for new cases as well as the ones we trained it on.
- Main goal of machine learning is to devise learning algorithms that do the learning automatically without human intervention or assistance. The machine learning paradigm can be viewed as "programming by example." Another goal is to develop computational models of human learning process and perform computer simulations.
- The goal of machine learning is to build computer systems that can adapt and learn from their experience.

- Algorithm is used to solve a problem on computer. An algorithm is a sequence of instruction. It should carry out to transform the input to output.
- For example, for addition of four numbers is carried out by giving four number as input to the algorithm and output is sum of all four numbers. For the same task, there may be various algorithms. It is interested to find the most efficient one, requiring the least number of instructions or memory or both.
- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.
- Machine Learning provides business insight and intelligence. Decision makers are provided with greater insights into their organizations. This adaptive technology is being used by global enterprises to gain a competitive edge.
- Machine learning algorithms discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system.
- Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. The task of the supervised learner is to predict the output behavior of a system for any set of input values, after an initial training phase.
- In contrast to supervised learning, unsupervised or self-organized learning does not require an external teacher. During the training session, the neural network receives a number of different input patterns, discovers significant features in these patterns and learns how to classify input data into appropriate categories.
- Unsupervised learning algorithms aim to learn rapidly and can be used in real-time. Unsupervised learning is frequently employed for data clustering, feature extraction etc.
- Reinforcement learning : This is advanced machine learning technique. This is based on probability theory where mapping can be done based on input received and changes based on environment around it.
- Deep learning : This is also advanced machine learning technique which has multiple processing layers so as to produce non-linear response based on input data. There are so many small processors called as neuron working parallel in data processing.

- Predictive analytics : This technique refers to prediction based on past experience and it uses both data mining and machine learning.
- Association rule learning : This is used to identify interesting relations between different attributes from large datasets.

5.8 : Data Visualization using Tableau

Q.17 Explain data visualization with Tableau.

☞ [SPPU : Dec.-22, End Sem, Marks 4]

OR Explain data visualization with Tableau.

☞ [SPPU : May-18, End Sem, Dec.-19, End Sem, Marks 8]

Ans. : • Tableau is one of the fastest evolving Business Intelligence (BI) and data visualization tool. Tableau Server is a business intelligence application that provides browser-based analytics anyone can use. It's a rapid-fire alternative to the slow pace of traditional business intelligence software.

- A business intelligence and data visualization tool allowing users to make sense of their data through interactive charts, graphs, and diagrams.
- Why use Tableau ?
  1. Traditional BI tools require complex installations
  2. Rapid results to useful information
  3. Easy to use for all skill levels
  4. Excellent migration path for Excel users
  5. It can use many different sources of data.
- Tableau uses a visual query language. The tableau data engine is a breakthrough in-memory analytics database designed to overcome the limitations of existing databases and data silos.
- Capable of being run on ordinary computers, it leverages the complete memory hierarchy from disk to L1 cache. It shifts the curve between big data and fast analysis.
- Tableau allows the users to directly connect to databases, cubes, and data warehouses etc. After analyzing the data, the results can be shared live with just a few clicks . The dashboard can be published to share it live on web and mobile devices.

- Tableau is relatively new in the Business Intelligence market but its market share is growing on a daily basis. It is being nearly all industries, from transportation to healthcare.
- Tableau Software does not support expanded analytics such as Box plots, network graphs, tree - maps, heat-maps, 3D-scatter plots, Profile Charts or data relationships tool which allow users to mine data for relationships like another data visualization software does.

5.9 : Introduction to : Candela, D3.js, Google Chart API

Q.18 Write two data visualization functions from seaborn.

☞ [SPPU : June-22, End Sem, Marks 3]

Ans. : Seaborn is another Python data visualization library built on top of Matplotlib. Creating histograms in Seaborn:

```
sns.histplot(x=sample["price"])
```

- Histograms only work on numeric variables. They divide the data into an arbitrary number of equal-sized bins and display how many diamonds go into each bin.
- **Create a scatter plot :** Scatter plots are useful to show relationships between numeric variables.

```
import seaborn as sns
import matplotlib.pyplot as plt
age = [10,12,15,16,17,17,20,25,30,35,37,39,40,42,45,50]
height = [120,130,145,143,182,186,170,172,172,182,178,168,182,187,
          160,166]
sns.scatterplot(x=age, y=height)
plt.xlabel('age')
plt.ylabel('height')
plt.title('Height vs age')
plt.show()
```

Q.19 What is Google Chart API ?

Ans. : • The Google Chart API is an interactive Web service that creates graphical charts from user-supplied data.

- Google servers create a PNG image of a chart from data and formatting parameters specified by a user's HTTP request. The service supports a wide variety of chart information and formatting.

- The Google Chart Tools enable adding live charts to any web page. They provide advantages such as a rich gallery of visualizations provided as image charts and interactive charts and they can read live data from a variety of data sources.
- Users embed the data and formatting parameters in an HTTP request, and Google returns a PNG image of the chart. Many types of chart are supported, and by making the request into an image tag the chart can be included in a web page.

Q.20 Write short note on Candela.

Ans. : • As an open-source suite of web visualization components that make use of the Python language, Candela emphasizes scalable, rich visualizations created with a normalized API for use in real-world data science situations.

- Candela is an open-source suite of inter-operable web visualization components. It provides library for JavaScript, package for Python and R.
- The tool works on Kitware's resonant platform and offers a range of elements for data visualization. It allows user to make super-rich visualizations that are scalable and available within a normalized API.
- Candela is a JavaScript library that provides reusable visualization components for the web. Let's consider following points :

1. **Reusable :** Candela provides a general API not tied to any particular framework or library, so the components user create with it can be ported from application to application very easily.
2. **Visualization :** That general API does not provide many constraints; the main one is that user must implement a function called render() which will carry out visualization semantics, whatever they may be.
3. **Components :** Use object-oriented concepts to implement the notion of a visualization component.
4. **For the web :** Candela uses modern JavaScript, taking advantage of features and modern tooling to produce a library that can be used to do visualization almost anywhere on the web.

Q.21 Explain : i) Google chat API ii) Candela.

☞ **[SPPU : May-18, End Sem, Marks 8]**

Ans. : Refer Q.19 and Q.20.

END...

Unit VI

6 Big Data Technologies Application and Impact

6.1 : Social Media Analytics

Q.1 What is social network analysis ?

Ans. : • Social Network Analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs and other connected information / knowledge entities. The term "social network" has been introduced by Barnes in 1954.

- SNA is the study of social relations among a set of actors. The methods of data collection in network analysis are aimed at collecting relational data in a reliable manner. Data collection is typically carried out using standard questionnaires and observation techniques that aim to ensure the correctness and completeness of network data.
- Social network analysis is based on an assumption of the importance of relationships among interacting units. The social network perspective encompasses theories, models and applications that are expressed in terms of relational concepts or processes.
- The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships.
- The advantage of social network analysis is that, unlike many other methods, it focuses on interaction. Network analysis allows us to examine how the configuration of network influences how individuals and groups, organizations or system function.

Q.2 What do you mean social media analytics ?

Ans. : • Social Media Analytics deals with development and evaluation of tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data

- Buyer's perspective about various brands and businesses can be statistically analyzed to extract various insights required for decision

making from a very large amount of unstructured and semi structured social media data.

- In social media, the two sources of information are the content (images, audios, customer feedbacks, product reviews, videos, bookmarks, sentiments etc.) generated by users and the relationships between the entities of network (people, organizations, products etc.).
- The social media analytics can be categorized into two parts : Content-based analytics and Structure-based analytics.
- In content-based analytics, analytics is performed on the content posted by the users on the social media platforms. Such content is of high volume, unstructured, noisy and dynamic nature.
- To extract insights from such data, the text, audio and video analytics techniques can be applied. The data processing challenges are addressed by the big data technologies.
- In structure-based analytics, the focus is on the structural attributes of the social network. Insights are extracted from the relationships of the entities

Q.3 Describe process of social media data analytics.

Ans. : • Fig. Q.3.1 shows process of social media data analytics.

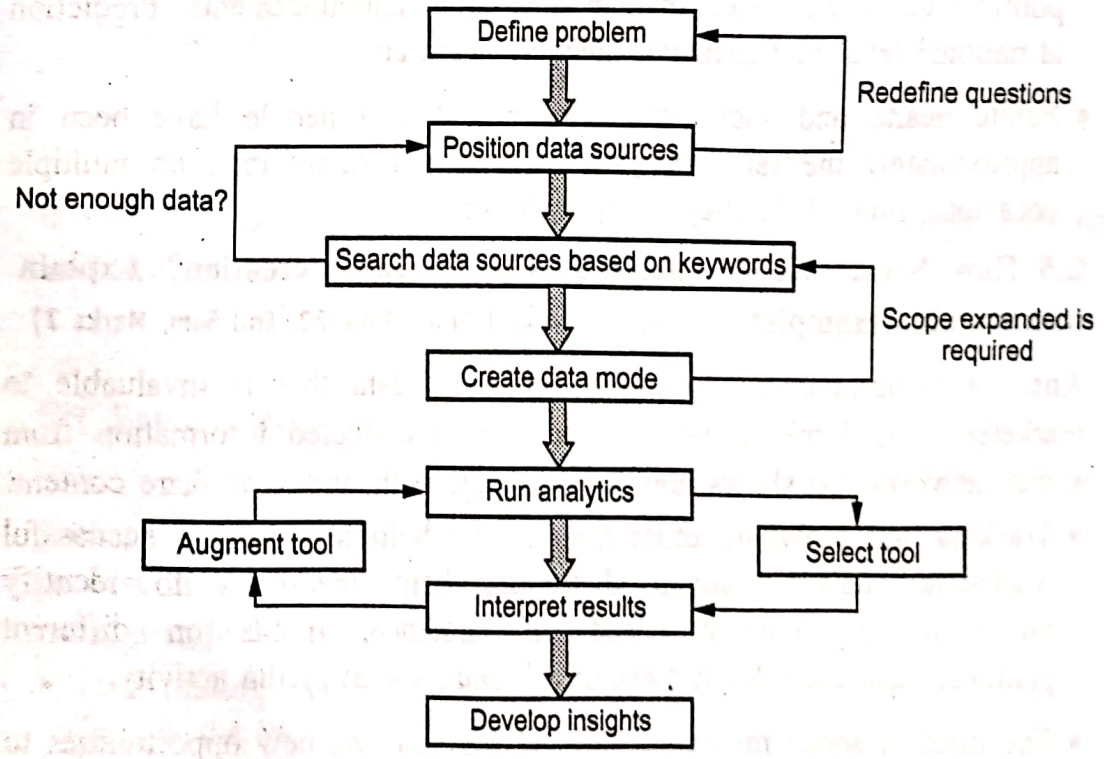


Fig. Q.3.1

- Social Media Analytics as a part of social analytics is the process of gathering data from stakeholder conversations on digital media and processing into structured insights leading to more information-driven business decisions and increased customer centrality for brands and businesses.
- Data analysis is the set of activities that assist in transforming raw data into insight, which in turn leads to a new base of knowledge and business value.
- In other words, data analysis is the phase that takes filtered data as input and transforms that into information of value to the analysts.
- Many different types of analysis can be performed with social media data. The data analysis step begins once we know what problem we want to solve and know that we have sufficient data that is enough to generate a meaningful result.

Q.4 List and explain applications of social media.

Ans. : • Retail companies - To harness their brand awareness, service improvement, advertising/marketing strategies, identifying influencers.

- Finance : To determine market sentiment, news data for trading.
- Government and public officials : Monitoring public perception on political candidates, election campaigns and announcements. Prediction at national level of happiness, unemployment etc
- Public health and sociology : Given that two people have been in approximately the same geographic locale, at same time, on multiple occasions, how likely they know each other ?

Q.5 How Social Media analytics helps in value creation? Explain with suitable examples. ☞ **[SPPU : June-22, End Sem, Marks 7]**

Ans. : • Social media contains a wealth of data that is invaluable to marketers. Social media data is made up of collected information from social networks that shows how users engage with, view, or share content.

- Tracking and analyzing those metrics can help to inform a successful marketing strategy. Social data can help marketers to identify high-performing content based on audience, trends on different platforms, and the effectiveness of a brand's social media activity.
- The value of social media expands beyond just the new opportunities to connect with consumers. It brings value to other aspects of your company as well, including learning more about your company's target

audience, building brand awareness and retaining existing customers. Social media can also be used as an effective customer service tool.

- Social media delivers measurable results in sales, leads and branding. It also enables a company to reach a large number of people at a low cost. The number of followers, shares and the reach can impact the value of your brand. With an effective social media plan in place, a company may be able to secure a higher valuation at the time of sale.
- Another social media strategy is to create a hashtag personal to your brand or one that encourages followers to create their own content as a way to increase engagement and ideally sales too.
- For example, Oreo's #PlayWithOreo Instagram campaign resulted in followers taking their own pictures featuring the ways they play with Oreo cookies. Followers that love Oreo had a chance to interact with the brand, and followers purchased the product in order to participate.
- Finding a way to adapt your company's product or services into a visually appealing, fun, and entertaining social media campaign could be a good way to gain extra exposure and more deeply connect with your followers, as well as increase sales.
- Showing a potential acquirer that your company can evolve and use new forms of social media to consistently expand sales may make your company more appealing and lead to a higher valuation.

6.2 : Text Mining

Q.6 What is text mining ? Draw and explain text mining architecture and its use.

☞ [SPPU : May-18, Dec.-19, June-22, End Sem, Marks 7]

OR Explain text mining with example.

☞ [SPPU : Dec.-22, End Sem, Marks 8]

Ans. : • Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools.

- Text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns.
- Fig. Q.6.1 shows Simple input-output model for text mining.

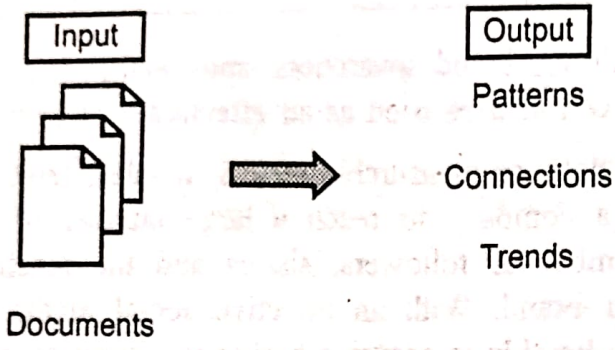


Fig. Q.6.1

- General model roughly divisible into four main areas : (a) preprocessing tasks, (b) Core mining operations, (c) Presentation layer components and browsing functionality, and (d) Refinement techniques.
- Fig. Q.6.2 shows high-level text mining functional architecture.

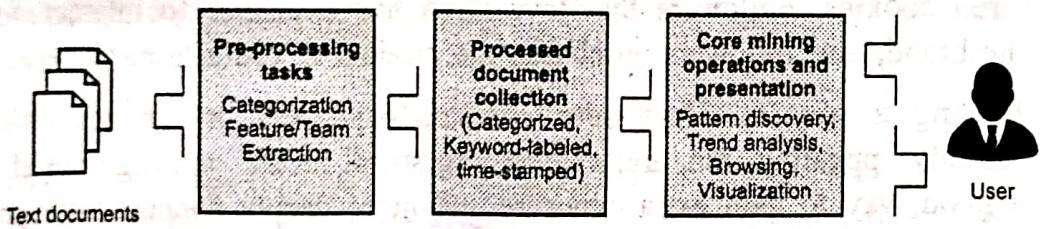


Fig. Q.6.2

- Preprocessing tasks : It include all those routines, processes, and methods required to prepare data for a text mining system's core knowledge discovery operations. Preprocessing tasks generally convert the information from each original data source into a canonical format before applying various types of feature extraction methods against these documents to create a new collection of documents fully represented by concepts.
- Core Mining Operations are the heart of a text mining system and include pat-tern discovery, trend analysis, and incremental knowledge discovery algorithms.
- Presentation Layer Components include GUI and pattern browsing functionality as well as access to the query language. Visualization tools and user-facing query editors and optimizers also fall under this architectural category.
- Refinement Techniques include methods that filter redundant information and cluster closely related data but may grow, in a given text mining system, to represent a full, comprehensive suite of suppression, ordering, pruning, generalization.

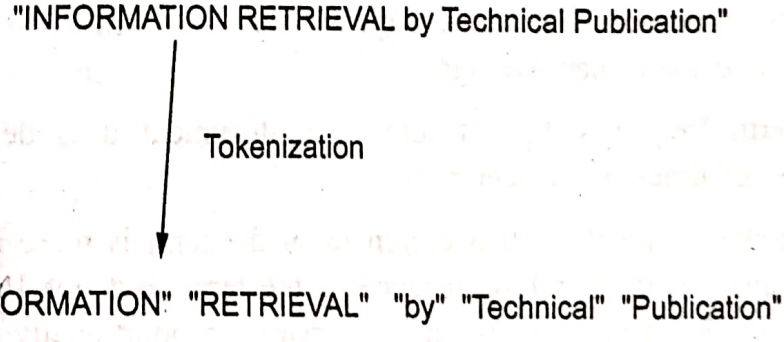
Q.7 What is text pre-processing ? Explain tokenization with example.

Ans. : Text pre-processing :

- Text pre-processing is required to transform the text into an understandable format so that machine learning algorithms can be applied to it.
- As we know Machine Learning needs data in the numeric form. We basically used encoding technique to encode text into numeric vector. But before encoding we first need to clean the text data and this process to prepare or clean text data before encoding is called text pre-processing.
- The various text pre-processing steps are : Tokenization, Lower casing, Stop words removal, Stemming and Lemmatization.

Tokenization

- Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. The simplest form of analysis is to reduce different word forms into tokens.
- Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords.
- Tokenization can be broadly classified into 3 types : Word, character, and subword (n-gram characters) tokenization.
- These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.



- Tokenization can be done to either separate words or sentences. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.
- Tokens are the building blocks of Natural Language, the most common way of processing the raw text happens at the token level.

- The major question of the tokenization phase is what are the correct tokens to use ? You chop on whitespace and throw away punctuation characters.
- Types of tokenization are white space, dictionary based, rule based, penn tree, spacy, subword etc.

Q.8 Write short note on : TF - IDF.

Ans. : • Term Frequency (TF) : Frequency of occurrence of query keyword in document

- Inverse Document Frequency (IDF) : How many documents the query keyword occurs in.
- Inverse Document Frequency (IDF) is a popular measure of a word's importance. It's defined as the logarithm of the ratio of number of documents in a collection to the number of documents containing the given word. This means rare words have high IDF and common words have low IDF.
- Term frequency is a measure of the importance of terms *i* in document *j*.
- Inverse document frequency is a measure of the *general* importance of the term.
- High term frequency for "apple" means that apple is an important word in a specific document. But high document frequency (low inverse document frequency) for "apple", given a particular set of documents, means that apple is not all that important overall, since it is in all of the documents.
- The weight increases as the number of documents in which the term appears decreases. High value indicates that the word occurs more often in this document than average.
- The term frequency $tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d.
- A document with tf = 10 occurrences of the term is more relevant than a document with tf = 1 occurrence of the term. But not 10 times more relevant. Relevance does not increase proportionally with term frequency.
- The document frequency is the number of documents in the collection that the term occurs in. We define the idf weight of term t as follows :

$$\text{idf weight } (idf_t) = \log 10 \frac{N}{df_t}$$

here N is the number of documents in the collection

- The tf-idf weight of a term is the product of its tf weight and its idf weight

$$W_{t,d} = (1 + \log tf_{t,d}) \cdot \log \frac{N}{df_t}$$

Stop lists and stemming :

- **Stoplists :** This is a list of words that we should ignore when processing documents, since they give no useful information about content.
- **Stemming :** This is the process of treating a set of words like "fights, fighting, fighter, ..." as all instances of the same term - in this case the stem is "fight".

6.3 : Mobile Analytics

Q.9 Explain types of mobile analytics.

☞ [SPPU : Dec-22, End Sem, Marks 9]

Ans. :

1. **Advertising/Marketing Analytics :** In today's marketplace, even if we develop an incredible app, the probability of it organically standing out among a million other apps is very low. The success of an app often hinges on whether marketing campaigns are able to attract the right types of users - those that install, remain engaged, and contribute to the financial components of the app.
   - Partnering with one or more ad networks is one of the most common ways to market an app. In the olden days, this practice consisted of establishing a budget and permitting an ad network to arrange for our ad to be displayed on a variety of publisher websites and apps. If the campaign was successful, we would see an increase in installs, engagement, and financial metrics.
   - Examples of common marketing analytics data that can be collected includes : Installs, Opens, Clicks, Purchases, Registrations, Content viewed, Level achieved, Shares and Custom events.

2. **In-App Analytics :** Regardless of whether an app delivers content, sells products, or offers a gaming experience, in order to be successful, the app must satisfy the expectations of its users.

   - Every app has one or more goals or objectives. In theory, apps are designed to enable users to achieve these objectives in the simplest manner possible.
   - We may follow intuitive hunches or make marginally-educated guesses regarding user demographics and in-app behaviors, but with no user or in-app behavior data, knowing where to make improvements may as well be determined on the roulette wheel.
   - In-app analytics is essentially "in-session" analytics - what users are actually doing inside the app and how they are interacting with the app. This is where conversion funnel, pathway, and feature optimization is the primary focus. Although marketers sometimes get looped into this data, it is primarily used by product managers.

Q.10 How mobile analytics is different than social media analytics ? Explain with suitable example. ☞ **[SPPU : May-18, End Sem, Marks 9]**

Ans. : • Mobile analytics involves measuring and analyzing data generated by mobile platforms and properties, such as mobile sites and mobile applications.

- With mobile analytics data, you can improve your cross-channel marketing initiatives, optimize the mobile experience for your customers, and grow mobile user engagement and retention.
- Social media analytics encompasses the collection, measurement, analysis, visualization and interpretation of digital data illustrating user behavior.
- Mobile analytics can provide more significant data and understanding than traditional web analytics.
- Mobile analytics do not only track the use of mobile apps, but also mobile web traffic. This combination of tracking mobile browsing and also offering a deeper understanding into user engagement with an app provide valuable insights how users react, interact and engage with different mobile features, pages and advertising.
- This approach of tracking provides feedback to developers, designers, advertisers and marketers to help them understand why users are or are not registering, buying or returning.

- Real time analytics are also key to understanding and improving user experience. These analytics focus on understanding user behaviour, instead of just providing a narrow set of metric data such as the amount of downloads.
- Web analytics offer great services, tracking the number of visits, recording how long they remained on a site and also providing information how they arrived at the site.
- However this information is starting to be diluted by the fact that even if some browser tabs are open, it doesn't mean that they are being engaged with.
- Or if a movie or TV show is being watched online, a web analytics can't know if a second screen is being engaged with at the same time. This doesn't make the web analytics irrelevant, but it is a thought that needs to be added to the equation.
- Social media analytics track IP addresses and the user agent, however with users working from different locations such as home, work, cafes, airports, etc., switching browsers for different reasons or clearing cookies suddenly a user can become anonymous.
- Mobile analytics face similar issues, a user might own several mobile devices, including tablets, alongside a PC and another PC at their workplace.
- Although there are a few advantages to keeping mobile users as they can connect through social authentication across several traditional web devices with their mobile devices. Mobile users also can clear their cookies, but it is not as common to reset their mobile identifiers.

Q.11 What is mobile analytics ? How it helps to organization ?

Ans. : • Analytics is the practice of measuring and analyzing data of users in order to create an understanding of user behavior as well as website or application's performance. If this practice is done on mobile apps and app users, it is called "mobile analytics".

- Mobile analytics is the practice of collecting user behavior data, determining intent from those metrics and taking action to drive retention, engagement, and conversion.
- Mobile analytics is similar to web analytics where identification of the unique customer and recording their usages.

- With mobile analytics data, you can improve your cross-channel marketing initiatives, optimize the mobile experience for your customers, and grow mobile user engagement and retention.
- Analytics usually comes in the form of a software that integrates into companie's existing websites and apps to capture, store, and analyze the data.
- It is always very important for businesses to measure their critical KPIs (Key Performance Indicators), as the old rule is always valid: "If you can't measure it, you can't improve it."
- To be more specific, if a business finds out 75% of their users exit in the shipment screen of their sales funnel, probably there is something wrong with that screen in terms of its design, user interface (UI) or user experience (UX) or there is a technical problem preventing users from completing the process.

Q.12 Explain working of mobile analytics.

Ans. : • Most of the analytics tools need a library (an SDK) to be embedded into the mobile app's project code and at minimum an initialization code in order to track the users and screens.

- SDKs differ by platform so a different SDK is required for each platform such as iOS, Android, Windows Phone etc. On top of that, additional code is required for custom event tracking.
- With the help of this code, analytics tools track and count each user, app launch, tap, event, app crash or any additional information that the user has, such as device, operating system, version, IP address (and probable location).
- Unlike web analytics, mobile analytics tools don't depend on cookies to identify unique users since mobile analytics SDKs can generate a persistent and unique identifier for each device.
- The tracking technology varies between websites, which use either JavaScript or cookies and apps, which use a software development kit (SDK).
- Each time a website or app visitor takes an action, the application fires off data which is recorded in the mobile analytics platform.

Q.13 How are mobile analytics different from web analytics ?

Ans. :

Mobile analytics	Web analytics
When web site is using, then mobile user called as USER.	When web site is using, then user called as VISITER.
Interaction with site is called as SESSIONS	Interaction with site is called as VISITS
On mobile, users have less screen real estate (4 to 7 inches) and interact by touching, swiping, and holding	On a desktop, users have larger screens (10 to 17 inches) and interact by clicking, double-clicking, and using key commands
Session timeout may be as short as 30 seconds	Session will end after 30 minutes of inactivity for websites
Unique users are identified via user IDs	Cookies are used to identifies user.

6.4 : Data Analytics Life Cycle of Case Studies

Q.14 Explain in brief data analytics life cycle.

☞ **[SPPU : June-22, End Sem, Marks 7]**

Ans. : Data analytics lifecycle is listed below :

Business user (BI analyst)	• Main responsibility is to define the business process through identifying key performance indicator and metrics to measure these processes. • The business users must be clear with the set of questions and answers for taking business decisions.
Data warehouse manager	• Main responsibility is to define, develop and manage the data platform.

	• New innovative technologies help data warehouse manager to broaden responsibility by using new technologies like Hadoop, data federation and in-memory computing. • Processing of new technologies structured and unstructured data
Data Scientist	• Responsibility is to mine the unstructured and structured data from inside as well as outside the organization to discover new business insights. • They are continuously in search of the new data sources to fulfill the requirement of analytical insights for the improvement of key business processes. • The data scientist needs the working environment so that they can collect; transforms; combine; cross-examine and visualize data so as to find the hidden relationship and new insights across available data items.
Business Intelligent analyst	• Responsibility is to identify, manage, present and publish key performance indicator and metrics against which all the business users will calculate and watch business success. • Various dashboards and reports are developed by BI analyst so that business users can run the business and discover new business insights based on real time data.
Business user	• Finally the analytical process comes back to business users who can use these reports and dashboard to the business. • At the end, the effectiveness of various decisions made by the business users is totally depending on the effective work done by data scientist, BI analyst and data warehouse manager.

6.5 : Organizational Impact

Q.15 Explain big data impact on organizations.

☞ [SPPU : June-22, End Sem, Marks 4]

Ans. : • Business Intelligence and data science including wide statistics-helps to describe a possible future event information, data engineering, programming, and data seeing have very varied aspects and require varied skills and approaches.

- Big data has reverted the details for defining and putting into numbers terms such as valuable, important, and successful. It is these details that fuel the that are the source of competitive wholesomeness.
- New big data sources and new wide capabilities of deep learning, supports higher loyalty answers to these questions.
- Most organizations now understand that if they capture all the data that streams into their businesses, they can deploy big data analytics to get significant leverage in understanding their customers, forecasting business trends, reducing operational costs, and realizing more profits. Regardless of industry, data analysis and visualization become accessible and impact business in critical ways.
- Big data analytics helps to improve quality in industries where inconsistencies are hard to reduce.
- Big Data Analytics is beneficial for organizations like Travel and hospitality, Healthcare, Government, Retail and more, that relies on agile and quick decisions to stay competitive.
- The high-performance analytics retrieved from Big Data analytics resources helps organization do things they have never thought before.
- They can get quick results in few seconds rather than days which can in turn accelerate quick reactions to key business challenges and questions.
- Big Data and Analytics are becoming closely intertwined and work together on the unstructured data to get precise answers for hard-to-solve problems and uncover new growth opportunities.
- Optimize the allotment of sufficient sales resources in view of best sales opportunities by the sales department. Identification of great potential and important business account is an equally important task done by the sales team.

- To identify and confirm suppliers who are cost-effective and supply good quality products in a timely manner.
- To measure the device performance and process variance are the main indicators of manufacturing, processing or quality problems.

6.6 : Understanding Decision Theory

Q.16 What is decision theory ? How is decision theory related to data analysis ? Explain Business Intelligence Challenge.

Ans. : • Decision theory is the analysis of the behavior of an individual facing non-strategic uncertainty, that is, uncertainty that is due to what we term "Nature" or, if other individuals are involved, their behavior is treated as a statistical distribution known to the decision maker.

- Decision theory depends on probability theory. Decision theory is based on various decisions taken at different stages of analytical life cycle.
- Statistical problems can be interpreted using decision theory. In this view, problems are considered solved when an optimal "decision rule" is chosen from a set of allowed rules.
- The optimal choice is usually given in terms of an optimization problem involving a risk/loss/objective function to be minimized.
- Most decisions are not made in short time due to complicated processes. So it is therefore natural to divide them into phases or stages.
- One interesting specialty of big data is it is challenging to the ordinary thinking. The reason is the suitable large amount of data is generated which need to model by sampling datasets.
- Business intelligence (BI) is the broad category of skills sets, processes, technologies, applications and practices used to support better decision making.
- BI can also be defined as the ability to access the right data needed at the right time to make informed strategic decisions.
- Significant training and handcrafting were needed or demanded that help to merchant users.
- Business users, who are by their nature are not experts and have struggled to learn and need to combine varied things together so they work as one unit. Also should be able to convert into expertness for their daily merchantry processes.

- Merchantry intelligence tools did not help merchantry users to make the transformation from one thing to another. This comes from deep thinking by understandings of deep things and optimization considering that tools were not unbearable in helping users understanding why something happened.

Q.17 What is business process ? Explain analytics chasm.

Ans. : • Decision processes : Most decisions are not made in short time due to complicated processes. So it is therefore natural to divide them into phases or stages. One interesting specialty of big data is it is challenging to the ordinary thinking. The reason is the suitable large amount of data is generated which need to model by sampling datasets.

- Big data professionals don't necessarily have to exercise on huge data in diverse ways. As doing of these things to huge data sets may be the pioneer of some obstacles to understanding why un-revokable behaviors happen.
- Specific area of data analysis where users are trying to apply the statistical algorithm to their data in order to measure cause and effects. This helps to identify the correlation between certain activities and results of it. Users hopes are if they can quantify cause
- The BI vendors added statistical and analytic capabilities to their products, all in the hope of moving business users beyond retrospective reporting into the area of predictive analytics.

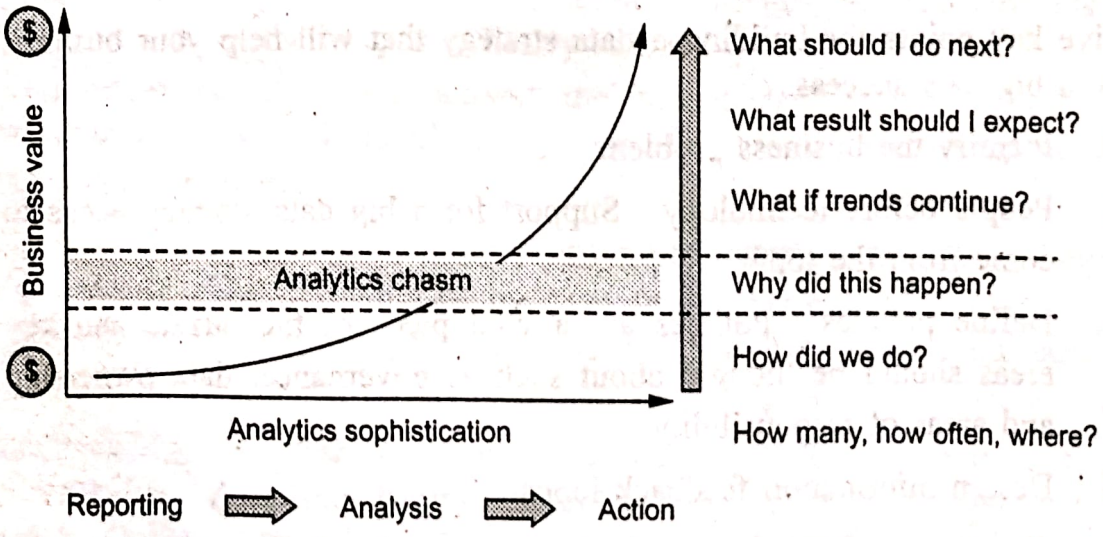


Fig. Q.17.1 Analytics Chasm

- Unfortunately, the BI tools failed to help business users make the transition because the tools were totally inadequate in helping users understand.

- Something happened these tools required business users to quantify cause-and-effect in order to build the models necessary to predict what to do next, and that was beyond their training and interest. As a result, the users' transition to a forward-looking view of their business fell into the "analytics chasm."
- Trying to turn the average business user into a statistical specialist failed, in the early 2000's and it continues to fail today.
- The average business user's career aspiration is not to become a statistical expert. They are in the retail or medical or telecommunications or banking industries because they like that industry, not because they want to master statistics or manipulate large data sets. The tools today are way too hard to make that process trivial.

6.7 : Creating Big Data Strategy

Q.18 How to create strategy for big data ?

Ans. : • Big data is like traditional data in many ways : It must be captured, stored, organized, and analyzed, and the results of the analysis need to be integrated into established processes and influence how the business operates.

- The importance of big data and analytics has seen it rise to the top of the decision making tree, becoming a C-level decision to address what should be done.
- Five key points for building a data strategy that will help your business be a big data success. :

1. Identify the business problem
2. People before technology : Support for a big data strategy needs to come from the top
3. Define policies : Policies are a vital piece of the puzzle and key areas should be thought about such as governance, data ownership and areas of responsibility.
4. Design information feedback loops
5. Plan for the future

- Document for Big Data Strategy :

1. Business Strategy : The targeted business strategy always clearly defines the scopes where the big data initiative will be the focus. This includes the title of the document.
2. Business Initiatives : This section breaks business plan into various new supporting subsets. This subset is allotted time duration from total approximately 9 months to one-year project duration.
3. Outcomes and Critical Success Factors (CSF) : This section helps to generate the result and through successful execution of the organizations merchantry approach to finding something new.
4. Tasks : This is the next level which has various writing task that needs to be done for successful completion of merchantry attempts.
5. Data Source : Data documents focus on the key data source required and support merchantry plan to achieve their goals through various merchantry attempts

6.8 : Big Data Value Creation Drivers

Q.19 Explain big data drives value creation processes.

Ans. : • Understanding value creation process : As some organization has their fixed understanding or imagination about how solutions on big data help to achieve their key merchantry attempt.

- Visualization exercise help among the merchantry user to identify particular areas so can have an idea wherever big data affect their organization business.
- All business users recognize a different type of questions they are trying to answer in view of their key business processes. Fig. Q.19.1 shows Big data drives value creation processes. (See Fig. Q.19.1 on next page.)
- Understanding big data value creation process : It is important to understand big data "merchantry" drivers and the values of "megacosm" process. Different "merchantry" drivers are shown below :

1. Structured Data : Mostly business transactional data. It helps to get Enable more coarse as well as exhaustive decisions.

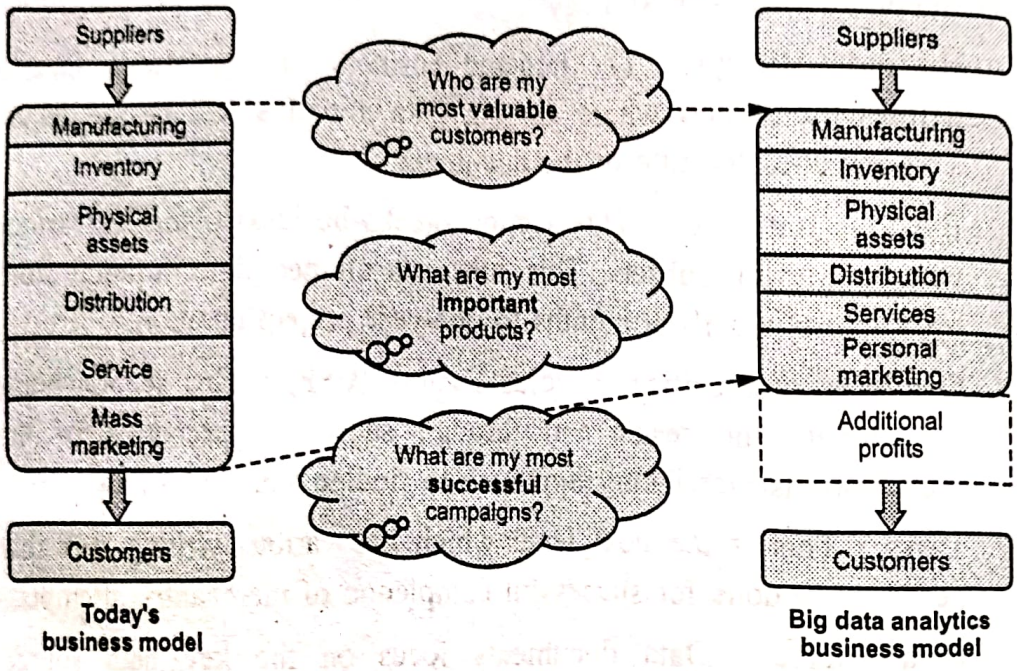


Fig. Q.19.1 Big data drives value creation processes

2. Unstructured Data : Data from variety of sources mostly in text documents. It helps to get complete and accurate business decisions.
3. Velocity : Real time data with high speed and very low data latency. It helps in real time business decisions frequently
4. Predictions : Predictor, Experiment, cause, instrumentations. It help to predict actionable business decisions.

Q.20 Explain big data value terminology.

☞ **[SPPU : June-22, End Sem, Marks 4]**

Ans. : • **Big data :** Big data is an umbrella term for datasets that cannot reasonably be handled by traditional computers or tools due to their volume, velocity, and variety. This term is also typically applied to technologies and strategies to work with this type of data.

• **Batch processing :** Batch processing is a computing strategy that involves processing data in large sets. This is typically ideal for non-time sensitive work that operates on very large sets of data. The process is started and at a later time, the results are returned by the system.

- Cluster computing : Clustered computing is the practice of pooling the resources of multiple machines and managing their collective capabilities to complete tasks. Computer clusters require a cluster management layer which handles communication between the individual nodes and coordinates work assignment.
- **Data warehouse :** Data warehouses are large, ordered repositories of data that can be used for analysis and reporting. In contrast to a data lake, a data warehouse is composed of data that has been cleaned, integrated with other sources, and is generally well-ordered. Data warehouses are often spoken about in relation to big data, but typically are components of more conventional systems.
- **ETL :** ETL stands for extract, transform, and load. It refers to the process of taking raw data and preparing it for the system's use. This is traditionally a process associated with data warehouses, but characteristics of this process are also found in the ingestion pipelines of big data systems.

6.9 : Michael Porter's Valuation Creation Models

Q.21 Explain Michael Porter's Analysis model.

Ans. : • Five Forces Analysis provides inputs, perspective on an organization's competitive as industry wide and outside-in data as a driver. Following are the five forces :

1. Competitive jealousy : This mainly focuses on the total number and volume of other organizations competing to the organization. It also include by and large organizational size and its directions.
2. Power of supplier : This includes a factor those are related to the supplier as the reputation of the brand, an area covered by the supplier, various products and provided services. Also gives an idea about the capacity to bid on different products and services available
3. Power of buyer : This force helps to "merchantry" as it include buyer choice and preferences, a volume of a buyer and switching frequency and tendency.
4. Development of Products and Technologies : This is about quality and price of various products and services offered. Also, exposure to marketplace allocation, market trends, and compliance risk, legislative and government actions is also included under product and technology development.

5. New market entrants : This gives an idea about the barriers to the newcomer or the new entry. Different geographical and cultural factors those affects on "merchantry" are also discovered under this title.

Q.22 Explain Porter's value chain analysis.

Ans. : • The Porter's value chain concept says that there is a chain of events which occur in a company right from the procurement of raw materials to the delivery of goods as well as the post sales service

• Value chain analysis is an analytical framework that assists in identifying business activities that can create value and competitive advantage to the business. Fig. Q.22.1 shows the essence of apple value chain analysis.

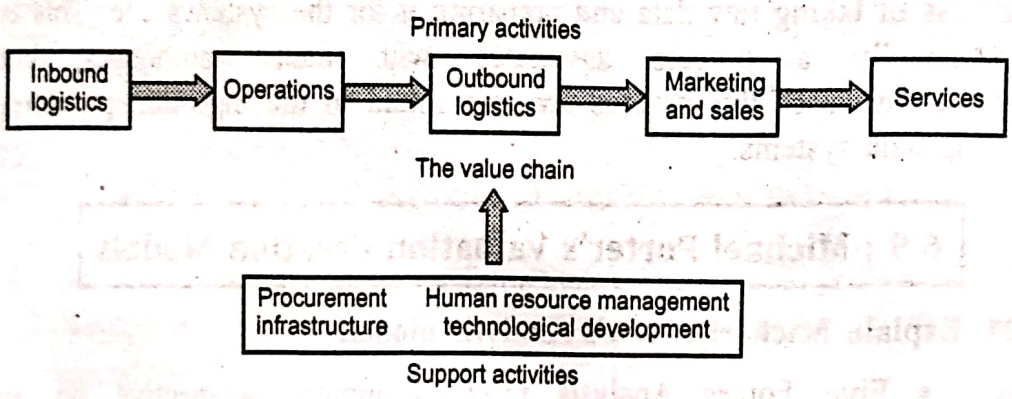


Fig. Q.22.1 Primary and secondary activities

• Most organizations engage in hundreds, even thousands, of activities in the process of converting inputs to outputs. These activities can be classified generally as either primary or support activities that all businesses must undertake in some form.

• According to Porter, the primary activities are :

1. Inbound Logistics - Involve relationships with suppliers and include all the activities required to receive, store, and disseminate inputs.
2. Operations - Are all the activities required to transform inputs into outputs (products and services).
3. Outbound Logistics - Include all the activities required to collect, store, and distribute the output.
4. Marketing and Sales - Activities inform buyers about products and services, induce buyers to purchase them, and facilitate their purchase.
5. Service - Includes all the activities required to keep the product or service working effectively for the buyer after it is sold and delivered.

- Secondary activities are :

1. Procurement - Is the acquisition of inputs, or resources, for the firm.
2. Human Resource management - Consists of all activities involved in recruiting, hiring, training, developing, compensating and (if necessary) dismissing or laying off personnel.
3. Technological Development - Pertains to the equipment, hardware, software, procedures and technical knowledge brought to bear in the firm's transformation of inputs into outputs.
4. Infrastructure - Serves the company's needs and ties its various parts together, it consists of functions or departments such as accounting, legal, finance, planning, public affairs, government relations, quality assurance and general management.

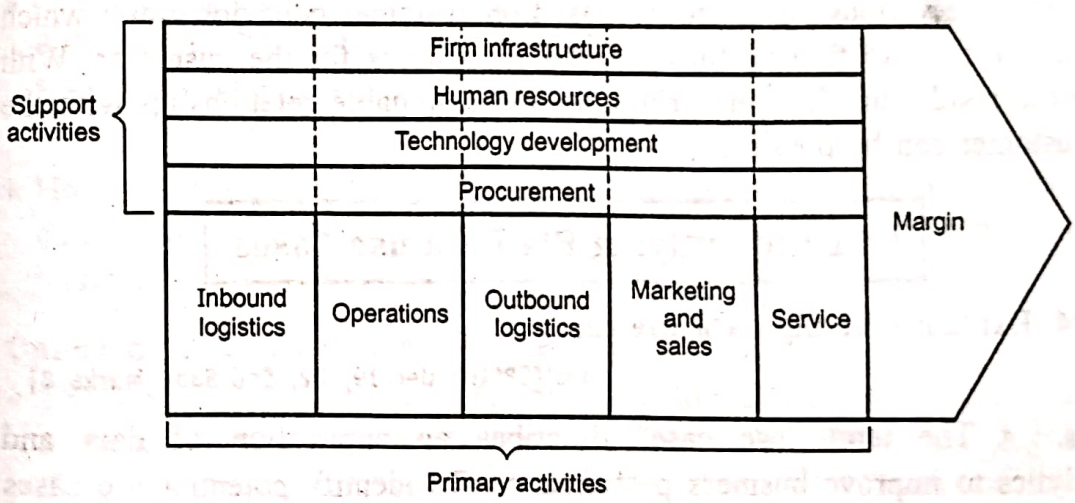


Fig. Q.22.2 Porter's value chain analysis

6.10 : Big Data user Experience Ramifications

Q.23 How big data user experience ramifications helps to organization?

Ans. : • This use case provides a good example of the process that one can employ in order to identify the most relevant questions that need to be answered in order to support an organization's key business decisions.

- And it all starts by understanding your organization's key business initiatives.

Step 1 : Understand your organization's key business initiatives.

Step 2 : Capture the decisions that an organization needs to make in order to support the organization's key business initiatives.

Step 3 : Identify those questions that need to be answered in order to facilitate making the decisions.

- Understanding the relationship between your customer's objectives, decisions, and questions that need to be answered is key to creating a user experience that provides the right information to the right customer to make the right decisions at the right time.
- It defines Big Data Model Maturity Index so that products insights extracted from big data have a very powerful impact on consumer user experience.
- There are various opportunities based on customer behavior graph which can be used to find useful and relevant insights for the customer. With these results building the profitable and actionable relationship with the customer can be possible.

6.11 : Identifying Big Data use Cases

Q.24 Explain four big data use cases.

☞ **[SPPU : Dec-19, 22, End Sem, Marks 8]**

Ans. : • The term "use case" describes an application of data and analytics to improve business performance. To identify potential use cases in a business, the first step is to understand what data is available and what modeling approaches fit the business challenge to solve with the information at hand.

- An effective use case is one in which the application of analytics to a business challenge provides benefits sufficient to justify the investments required to acquire, prepare, analyze, and act on the data. Even as the cost of data capture and storage and analytical processing power fall, this ROI threshold is a high bar for potential use cases to exceed.
- What are the characteristics of potential use cases likely to meet the required ROI threshold? They often solve problems in which even small performance improvements can yield large returns. An example of such a use case is customer price segmentation, often called yield management. Reducing the loss associated with pricing actions, even if by relatively small amounts, can yield large revenue gains.

- Other use cases with high ROIs are those that avoid substantial costs, such as unexpected down time in a production operation, for example, predicting the failure of important parts where unexpected failure can cause substantial delays.
- A key attribute for successful use case outcomes is that the incremental revenue or reduced cost is substantial and measurable.
- An approach to operationalizing data and analytics that maximizes the potential ROI is to define the application and possible approaches before starting down the path of data capture. Most data will come from the systems in place, so finding out what data they keep and how to access the data will help frame the available modelling options.
- How do merchantry and IT work together to identify the right merchantry opportunity upon which to focus the big data effort to doing something, and then diamond the right use of big data money-making opportunities ?
- How do you make sure of the successful use service of these new big data skills given the upper rate of failure for the adoption of new technologies ?

Cases of use of Big Data in Factories 4.0

- The amount of information produced by IoT and today's manufacturing systems must be translated into actionable ideas. That's why Big Data classifies the information collected and draws relevant conclusions that help improve companies' operations in the following ways :
  a) Improving warehouse processes : Using sensors and portable devices, companies can improve operational efficiency by detecting human errors, performing quality controls and showing optimal production or assembly routes.
  b) Elimination of bottlenecks : Big Data identifies variables that can affect performance, at no extra cost, guiding manufacturers in identifying the problem.
  c) Predictive demand : More accurate and meaningful predictions thanks to the visualization of activity through internal analysis (customer preferences) and external analysis beyond historical data. This allows the company to modify/optimise its product portfolio.

d) Predictive maintenance : Data fed sensors identify possible failures in the operation of machinery before it becomes a breakdown, by identifying breakdowns in patterns. The system sends an alert to the equipment so that it can react in time.

- Finding the right use case is essential to the success of data science project because it enables :

a) Understand your problem from an end-user perspective.

b) Find the right data-driven solution.

c) Define how to measure the project's success which ensures that your solution adds business value.

d) Go beyond data science in a technical sense and view your use case from a business perspective.

6.12 : Big Data Analytics Challenges and Research Directions

Q.25 Explain Big data analytics in research.

☞ [SPPU : June-22, End Sem, Marks 7]

Ans. : • Big data analytics and data science are becoming the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction from data.

- Applications of big data and data science include information science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing.
- Effective integration of technologies and analysis will result in predicting the future drift of events. The research issues pertaining to big data analysis are classified into three broad categories namely internet of things (IoT), cloud computing, bio inspired computing, and quantum computing.

IoT for Big Data Analytics :

- Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Currently, machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things.

- Thus, appliances are becoming the user of the internet, just like humans with the web browsers. IoT is attracting the attention of recent researchers for its most promising opportunities and challenges. It has an imperative economic and societal impact for the future construction of information, network and communication technology.

Cloud Computing for Big Data Analytics :

- The development of virtualization technologies have made supercomputing more accessible and affordable. Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processors, disk space, memory, and operating system.
- The use of these virtual computers is known as cloud computing which has been one of the most robust big data technique. Big Data and cloud computing technologies are developed with the importance of developing a scalable and on demand availability of resources and data.
- Cloud computing harmonize massive data by on demand access to configurable computing resources through virtualization techniques.

Q.26 Explain big data analytics challenges in brief.

☞ **[SPPU : Dec.-22, End Sem, Marks 9]**

Ans. : • In recent years, applications of big data and AI in education have made significant headways. This highlights a novel trend in leading-edge educational research. The convenience and embeddedness of data collection within educational technologies, paired with computational techniques have made the analyses of big data a reality.

- We are moving beyond proof-of-concept demonstrations and applications of techniques, and are beginning to see substantial adoption in many areas of education. The key research trends in the domains of big data and AI are associated with assessment, individualized learning, and precision education.
- Model-driven data analytics approaches will grow quickly to guide the development, interpretation, and validation of the algorithms. However, conclusions from educational analytics should, of course, be applied with caution.
- At the education policy level, the government should be devoted to supporting lifelong learning, offering teacher education programs, and protecting personal data.

- With regard to the education industry, reciprocal and mutually beneficial relationships should be developed in order to enhance academia-industry collaboration.
- Furthermore, it is important to make sure that technologies are guided by relevant theoretical frameworks and are empirically tested.
- Intelligent educational systems employing big data techniques are capable of collecting accurate and rich personal data. Data analytics can reveal students' learning patterns and identify their specific needs. Hence, big data and AI have the potential to realize individualized learning to achieve precision education.

END...

JUNE - 2022 [5870] - 1149 | Solved Paper

Course 2019

Time : $2\frac{1}{2}$ Hours] **[Maximum Marks : 70**

Instructions to the candidates :

1) Answer Q.1, or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8.

2) Neat diagrams must be drawn wherever necessary.

3) Figures to the right side indicate full marks.

4) Assume the suitable data, if necessary.

Q.1 a) *Explain big data ecosystem with suitable diagram.* **(Refer Q.1 of Chapter - 3)** **[7]**

b) *Explain anatomy of file read and write in HDFS.* **(Refer Q.6 of Chapter - 3)** **[7]**

c) *Write and explain any two Hadoop shell commands.* **(Refer Q.7 of Chapter - 3)** **[4]**

OR

Q.2 a) *Explain map reduce with proper diagram for word count example.* **(Refer Q.8 of Chapter - 3)** **[7]**

b) *Explain Google file system.* **(Refer Q.2 of Chapter - 3)** **[7]**

c) *Explain ETL processing.* **(Refer Q.22 of Chapter - 3)** **[4]**

Q.3 a) *Explain different steps in data analytics project life cycle.* **(Refer Q.1 of Chapter - 4)** **[7]**

b) *Draw and explain architecture of HIVE.* **(Refer Q.24 of Chapter - 4)** **[7]**

c) *Explain different data transformation techniques.* **(Refer Q.16 of Chapter - 4)** **[3]**

OR

Q.4 a) *Explain different kinds of big data analysis.* **(Refer Q.5 of Chapter - 4)** **[7]**