# Question: 1

Answer the below questions:

1. What is Big Data?

Big Data basically refers to huge volume of data that cannot be stored or processed using the traditional approach with in the given timeframe.

Ex: Let us say we have 10TB of image files upon which certain processing needs to be done. For instance, we may want to resize and enhance these images with in a given timeframe. If we make use of traditional system to perform this task, we will not be able to accomplish this task in given timeframe as the computing resources of the traditional system would not be able to accomplish this task on time. So, this 10TB of images is referred to BIGDATA.

1. What is Hadoop?

As 90% of the worlds data was generated in just last two years. So, these quantities of data were simply to large to pump through in a series format and difficult to process.

So, Hadoop is an open source project which allows applications to run using the Map Reduce algorithm using the utilities available and a network of many computer. Simply to frame the data is getting processed in parallel by dividing data into small chunks of data using Hadoop rather than in traditional serial processing

1. What is OLTP?

Online Transaction Processing as the name says it is transactional data which has to be read optimized and write optimized. Typically, transactions get continuously generated. So OLTP is a class of systems that supports or facilitates high transaction-oriented applications, which has the ability to fully process or completely undo an order. OLTP’s primary system features are immediate client feedback and high individual transaction volume.

1. What is OLAP?

Online Analytical Processing*,* as name suggests it is an analytical tool that provides analysis of [data](https://www.webopedia.com/TERM/D/data.html) stored in a [database](https://www.webopedia.com/TERM/D/database.html). Before OLAP it was difficult to perform analytics on multidimensional data OLAP resolved that issue as the data is collected from multiple data sources and stored in [data warehouses](https://searchdatamanagement.techtarget.com/definition/data-warehouse) then cleansed and organized into data cubes and it also has special functions for analyzing the data.

1. What is RDBMS?

Relational Database Management System is an advanced version of a DBMS system. RDBMS system. It is based on relational model introduced by E.F. Codd. Data is managed and stored in rows and columns which is known as tuples and attributes. Due to a collection of organized set of tables, and relations between tables is created using keys data can be accessed easily.

1. What is HDFS?

Managing and storing huge data in one machine will take a lot of time to process which is resolved using HDFS (Hadoop Distributed File Management System) which is used to store and manage data across cluster data being divided into several blocks with default block size 128MB.

1. What is MapReduce?

MapReduce is one of the major components in Hadoop ecosystem. MapReduce is designed to process a large amount of data in parallel by dividing the work into some smaller and independent tasks and assign them to worker nodes. It takes input as a list and convert it to output as a list as well.

1. What is NameNode?

Name node is a Master Daemon which maintains and manages data nodes. Name node is the component which helps in recording the metadata and also receives heart beat and block report of performance from all data nodes.

1. What is DataNode?

Data Node is a Slave Daemon which is responsible for storing the data and serves read and write requests from clients.

1. What is SecondaryNameNode?

When there are cases of editlog file getting corrupted then the role of secondary name node is needed .The secondary name node instructs the name node to record the transactions to a new edit log file now the secondary name node copies the FS image and edit log file to its check point directory once this files are copied the secondary name node loads the FS image and applies all the transactions from the edit log file and stores this information onto a new compacted FS image file the secondary name node transfers this compacted FS image file to the name node .The name node adopts this new FS image file and also creates a new edit log file this process occurs every hour or when the edit log file reaches to 64MB

1. What is ResourceManager?

ResourceManager is the master that arbitrates all the available cluster resources and thus helps manage the distributed applications running on the YARN system. It works together with the per-node NodeManagers and the per-application ApplicationMaster.

1. What is NodeManager?

NodeManagers take instructions from the ResourceManager and manage resources available on a single node.It launches the container and runs services to determine the health of the node .

1. What is Distributed File System?

Distributed file system is a client server-based application. It allows clients to access and process data stored on the server as data is available on their own computer. DFS organize the file and directory services of individual servers into a global directory in such a way that remote data access is not location specific but is identical from any client

1. What is Parallel Processing?

Parallel processing consists of two major components: Parallel computing and parallel systems

Parallel computing is a computing where the jobs are broken into discrete parts that can be executed concurrently

Parallel systems deal with the simultaneous use of multiple computer resources that can include a single computer with multiple processors, a number of computers connected by a network to form a parallel processing cluster or a combination both

1. What is Commodity Hardware?

Hadoop does not require a very high-end server with large memory and processing power. Typically, it is a low-performance system that is IBM PC-compatible and is capable of running Microsoft Windows, Linux, or MS-DOS without requiring any special devices or equipment. Due to this we can use any inexpensive system with average RAM and processor. Such kind of system is called commodity hardware.

1. What is a process?

Process in Hadoop is referred to processing of data on most of the cases .When processing, Hadoop will indeed use the chunks of data on all nodes, to achieve a cluster and distributed computing.

1. What is a daemon?

Daemons in computing terms is a process that runs in the background. Hadoop has five such daemons. They are NameNode, Secondary NameNode, DataNode, JobTracker and TaskTracker. Each daemon runs separately in its own JVM.

1. What is a service?

A service is nothing but performing tasks Hadoop can also be said as Haas(Hadoop as a service ) is also known as Hadoop in the cloud, is a big data analytics framework that stores and analyzes data in the cloud using Hadoop. Users do not have to invest in or install additional infrastructure on premises when using the technology, as HaaS is provided and managed by a third-party vendor.

1. What is a Mapper?

Mapper is a function which process the input data. The mapper processes the data and creates several small chunks of data. The input to the mapper function is in the form of (key, value) pairs, even though the input to a MapReduce program is a file or directory (which is stored in the HDFS).

1. What is a Reducer?

Reducer is the 2nd phase of processing the data in Hadoop. The output from the Mapper (intermediate keys and their value lists) are passed to the Reducer in sorted key order.The Reducer outputs zero or more final key/value pairs. These are written to HDFS. The Reducer usually emits a single key/value pair for each input key

1. What is map?

Map is a Task in the Hadoop MapReduce framework. Hadoop creates a map task to process each input split. The map task: Uses the Input Format to fetch the input data locally and create input key-value pairs.

1. What is reduce?

Reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

1. What is Input Split?

An input to a MapReduce job is divided into fixed-size pieces called input splits. It is a chunk of the input that is consumed by a single map. Further split divides into records. Thus, during data processing in MapReduce program or other processing techniques use InputSplit. In MapReduce, important thing is that InputSplit does not contain the input data. Hence, it is just a reference to the data.

1. What are map intermediate results?

Intermediate data is a mapper output which is stored on the Local file system of each individual mapper nodes. This is typically a temporary directory location which can be setup in config by the hadoop administrator. The intermediate data is cleaned up after the Hadoop Job complete

1. What is shuffle and sort?

When mappers finish their tasks, their output is a series of key-value pairs. Shuffling is simply the act of transferring the mapper output to the reducers. Sorting is the process or sorting the mappers output by key. For example, unsorted output from the mappers might look like this:

(sam,1),(joe,1),(rajesh,1),(joe,1),(bill,1),(andy,1),(joe,1)

When it's sorted, it would look like this:

(andy,1), (bill,1), (joe,1), (joe,1), (joe,1), (rajesh,1), (sam,1)

Sorting helps the system determine when a new reducer should be started.

1. What are the components of Resource manager?

Scheduler and Application Manager

1. What is Application Master?

ApplicationMaster’s are responsible for negotiating resources with the Resource Manager and for working with the Node Managers to start the containers.

1. What is a container?

Container is a place where a unit of work occurs. For instance each MapReduce task(not the entire job) runs in one container. An application/job will run on one or more containers. Set of system resources are allocated for each container, currently CPU core and RAM are supported. Each node in a Hadoop cluster can run several containers.

1. What is a task?

MapReduce is the data processing layer of Hadoop. It is the framework for writing applications that process the vast amount of data stored in the HDFS.In Hadoop, Job is divided into multiple small parts known as Task.

1. What is a job in YARN?

In Yarn Job class is the most important class in the MapReduce API. It allows the user to configure the job, submit it, control its execution, and query the state. The set methods only work until the job is submitted, afterwards they will throw an IllegalStateException.

1. What is HQL?

Before HQL Map reduce program is written in java and it is a huge drawback because all the companies need to get their employees skilled in Java which is a huge drawback.

HiveQL is similar to SQL for querying on schema info on the Metastore. It is one of the replacements of traditional approach for MapReduce program. Instead of writing MapReduce program in Java, we can write a query for MapReduce job and process it.

1. What are managed and external tables in hive?

Managed Table: It is also know an internal table. When we create a table in Hive, it by default manages the data. This means that Hive moves the data into its warehouse directory.

External Table: We can also create an external table. It tells Hive to refer to the data that is at an existing location outside the warehouse directory.

The main difference is that when you drop an external table, the underlying data files stay intact. This is because the user is expected to manage the data files and directories. With a managed table, the underlying directories and data get wiped out when the table is dropped.

1. What are map side joins?

Map-side Joins allows a table to get loaded into memory ensuring a very fast join operation, performed entirely within a mapper and that too without having to use both map and reduce phases. In case your queries frequently run with small table joins , you might see a very substantial decrease in the time taken to compute the queries after usage of map-side joins.

1. What are the different modes of execution in hive?
2. Local mode
3. Map reduce mode
4. What is UDTF’s in hive

UDTF   – User Defined Tabular Function. UDTF is a User Defined Table Generating Function that operates on a single row and produces multiple rows a table as output.

1. Which command would you use to alter the hive configuration parameters?

SET is the command used to alter the hive configuration parameters

1. What are the different ways of creating tables and loading data?

Creating table can happen in two ways i.e.,

1. Table by declaring the column names and loading the data from a path.
2. Table using subquery.

Loading Data :

1. Loading data from text file or csv and other formats
2. Loading data using sub query.
3. What are the 2 types of partitioning in hive?

Static and Dynamic partitioning

1. How to extract hive data into a directory?

By using the Insert query, we can export the hive data into local directory or HDFS directory.

1. What is hive metastore?

Metastore is the central repository of Apache Hive metadata. It stores metadata for Hive tables (like their schema and location) and partitions in a relational database. It provides client access to this information by using metastore service API.

Hive metastore consists of two fundamental units:

1. A service that provides metastore access to other Apache Hive services.
2. Disk storage for the Hive metadata which is separate from HDFS storage.

# Question: 2

Case Study-1: Petrol Dataset: Data loaded in the path **/user/raghavendra/petrol** on the cluster

Run the below queries from the partition petrol\_part table

Initial Data Loading steps:

1. mkdir Assign\_Petrol – Creating a local directory
2. hadoop fs -get /user/raghavendra/petrol/petrol.txt /home/cdh7bmanipal6/Assign\_Petrol – Getting data from hdfs and loading the file in local directory
3. beeline -u jdbc:hive2:// --silent=true – Launching Hive
4. create database DB\_Petrol; - Creating the data base
5. use DB\_Petrol; - Using the DB which is created for working on this data
6. Creating the table :

create table Petrol\_DB

(ID String,

Name String,

Buy\_rate String,

Sell\_rate String,

Volume\_IN double,

Volume\_OUT double,

Year double)

row format delimited

fields terminated by ","

lines terminated by "\n"

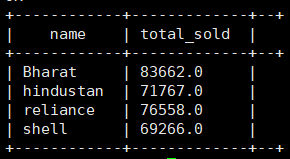
stored as textfile

tblproperties("skip.header.line.count"="1");

1. Loading the data
2. load data local inpath '/home/cdh7bmanipal6/Assign\_Petrol/petrol.txt' into table Petrol\_DB;

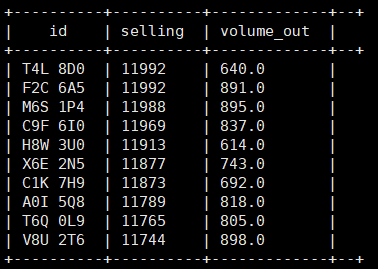
1)In real life what is the total amount of petrol in volume sold by every distributor?

Select NAME,SUM(Volume\_OUT) as Total\_Sold from Petrol\_DB group by Name;



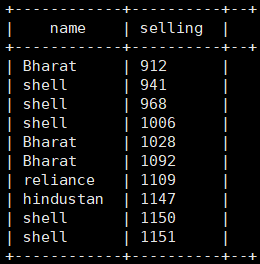
2)Which are the top 10 distributors ID’s for selling petrol and also display the amount of petrol sold in volume by them individually?

Select ID,int(REGEXP\_EXTRACT(sell\_rate,'[0-9]+',0)) as Selling,Volume\_OUT from Petrol\_DB order by Selling desc limit 10 ;



3)Find real life 10 distributor name who sold petrol in the least amount.

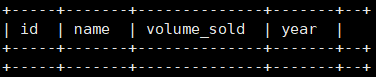
Select Name,int(REGEXP\_EXTRACT(sell\_rate,'[0-9]+',0)) as Selling from Petrol\_DB order by Selling limit 10 ;



4)The difference between volumeIN and volumeOuT is illegal in real life if greater than 500. As we see all distributors are receiving petrols on every next cycle.

List all distributors who have this difference, along with the year and the difference which they have in that year.

Select id,Name,SUM((Volume\_IN -Volume\_OUT)) as Volume\_Sold,year from Petrol\_DB group by id,Name,year having volume\_sold>500 order by Volume\_Sold desc;



Case Study-2: Olympic Dataset: **/user/raghavendra/olympicdata** on the cluster

Initial Data Loading steps:

1. mkdir Assign\_Olympicdata– Creating a local directory
2. hadoop fs -get /user/raghavendra/olympicdata/olympic\_data.csv /home/cdh7bmanipal6/Assign\_Olympicdata– Getting data from hdfs and loading the file in local directory
3. beeline -u jdbc:hive2:// --silent=true – Launching Hive
4. create database DB\_Olympic\_Amruth; - Creating the data base
5. use DB\_Olympic\_Amruth; - Using the DB which is created for working on this data
6. Creating the table :

create table Olympic\_Data

(Name String,

Age int,

Country String,

Year double,

Date date,

Sport String,

Gold double,

Silver double,

Bronze double,

Total\_medals double)

row format delimited

fields terminated by "\t"

lines terminated by "\n"

stored as textfile;

1. Loading the data

load data local inpath '/home/cdh7bmanipal6/ Assign\_Olympicdata /olympic\_data.csv ' into table Olympic\_Data;

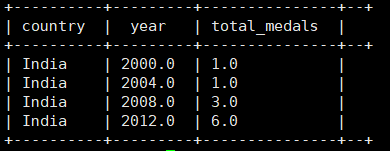
1.Using the dataset list the total number of medals won by each country in swimming.

Select country,SUM(total\_medals) as Total\_medals from Olympic\_Data where Sport='Swimming' group by country ;



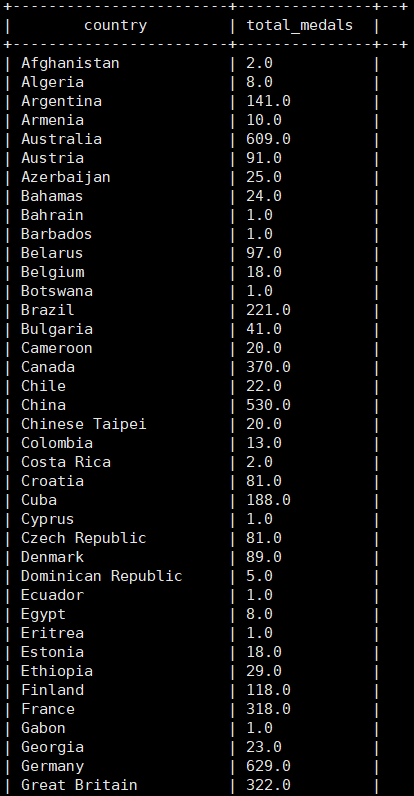
2.Display real life number of medals India won year wise.

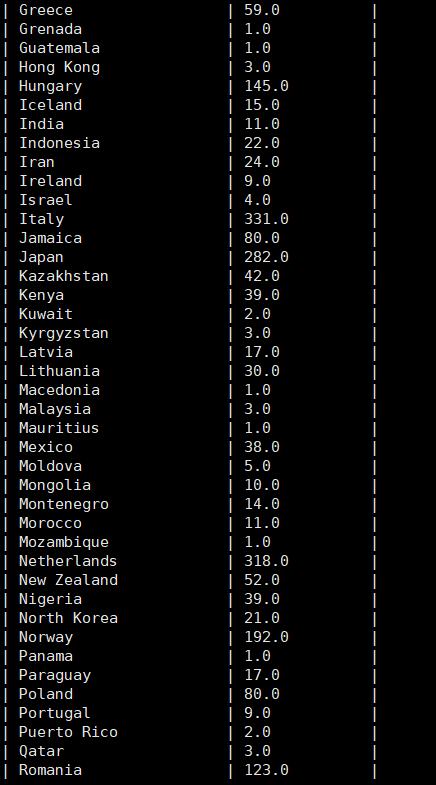
Select country,Year,SUM(total\_medals) as Total\_medals from Olympic\_Data where country='India' group by country,year;

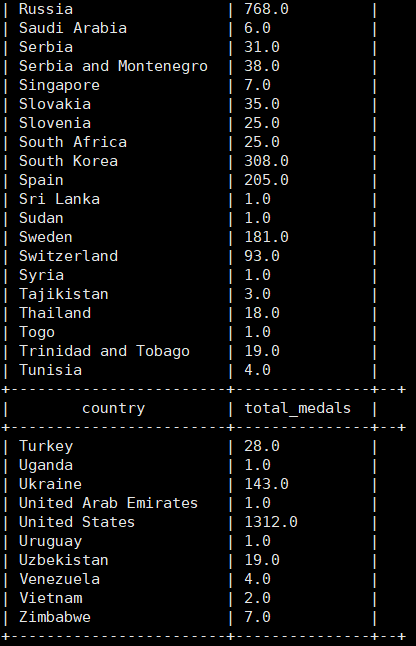


3.Find the total number of medals each country won display the name along with total medals.

Select country,SUM(total\_medals) as Total\_medals from Olympic\_Data group by country;

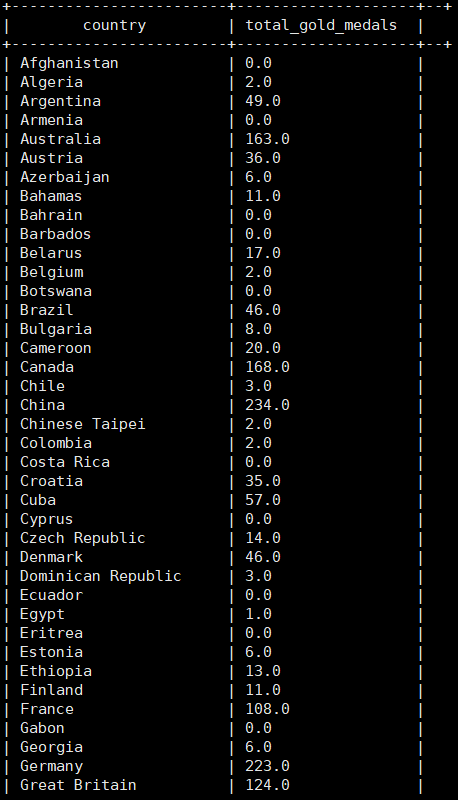




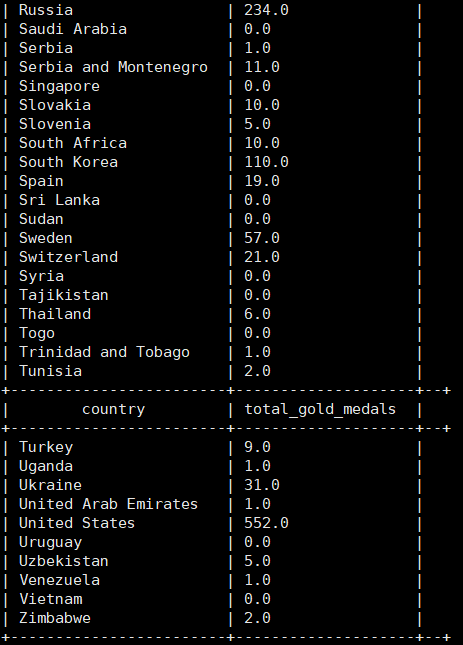


4.Find the real life number of gold medals each country won.

Select country,SUM(Gold) as Total\_Gold\_Medals from Olympic\_Data group by country;







5.Which country got medals for Shooting, year wise classification?

Select year,country,SUM(total\_medals) as Total\_medals from Olympic\_Data where Sport='Shooting' group by year,country ;

