**Assignment 1**

**Definitions:**

1. What is Big Data?

Big Data is the terminology used to refer the data which is very huge available in various formats.

1. What is Hadoop?

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

1. What is OLTP?

OLTP (Online Transaction Processing) class of software program capable of supporting transaction-oriented applications on the internet.

1. What is OLAP?

OLAP (Online Analytical Processing) is said to be more of an offline data store. It handles all historical data. It provides analysis of data stored in a dataset.

1. What is RDBMS?

RDBMS (Relational Data Base Management System). It is a DBMS designed specifically for relational databases. It refers to a database that stores data in a structured format, using rows and columns. It is relational because the values within each table are related to each other.

1. What is HDFS?

HDFS stands for Hadoop Distributed File System. It distributes the data blocks in to various nodes. Optionally replicates the blocks for high availability.

1. What is MapReduce?

It is a software framework for easily writing applications which process vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner. It has two phases MAP phase and REDUCE phase.

1. What is NameNode?

It is the master component of HDFS which does several tasks, few of them are monitoring the disk storage of all Data nodes, Stores file system Meta Data, check pointing etc.

1. What is DataNode?

Data Node is the slave component of HDFS. It is software demon which runs on one or more dedicated machines. It stores the actual data and default block size is 128MB with replication factor of 3.

1. What is SecondaryNameNode?

It is a software daemon which runs on a different machine other than a name node. It helps to restart name node when it fails. Creates check points of the file system

1. What is Resource Manager?

The Resource Manager is responsible for resource management and scheduling. Manages resources required for execution like (CPU+RAM+HD). It allocates containers for execution

1. What is Node Manager?

It runs on the same node where data node is running. Responsible for launching and managing containers on node.

1. What is Distributed File System?

Distributed File System is a system that can handle accessing data across multiple clusters (nodes).

1. What is Parallel Processing?

Parallel processing is a method in computing of running two or more processors (CPUs) to handle separate parts of an overall task. ... Parallel processing is commonly used to perform complex tasks and computations.

1. What is Commodity Hardware?

Computer hardware that is affordable and easy to obtain. 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.

1. What is a process?

In computing, a process is the instance of a computer program that is being executed by one or many threads. It contains the program code and its activity.

1. What is a daemon?

In multitasking computer operating systems, a daemon is a computer program that runs as a background process, rather than being under the direct control of an interactive user.

**In Linux:**

A daemon (also known as background processes) is a Linux or UNIX program that runs in the background. Almost all daemons have names that end with the letter "d".

1. What is a service?

In reference to computer software, a service is software that performs automated tasks, responds to hardware events, or listens for data requests from other software. In a user's operating system, these services are often loaded automatically at startup, and run in the background, without user interaction.

1. What is a Mapper?

Mapper in Map Reduce maps input key/value pairs to a set of intermediate key/value pairs.

1. What is a Reducer?

Reducer in Map Reduce reduces a set of intermediate values which share a key to a smaller set of values.

1. What is map?

Maps are the individual tasks that transform input records into intermediate records. The transformed intermediate records do not need to be of the same type as the input records.

1. What is reduce?

The Reduce can iterate through the values that are associated with that key and produce zero or more outputs. In other words, worker nodes now process each group of output data, per key, in parallel.

1. What is Input Split?

It will read the data blocks from the disk and give every line of that block to the mapper.

1. What are map intermediate results?

Mapper output are the map intermediate results and are stored on the local file system of each individual mapper nodes.

1. What is shuffle and sort?

Map Reduce makes the guarantee that the input to every reducer is sorted by key. The

Process by which the system performs the sort and transfers the map outputs to the reducers as inputs is known as the shuffle.

1. What are the components of Resource manager?

It has two components. Scheduler and Application Manager.

1. What is Application Master?

It has the responsibility of negotiating appropriate resource containers from the scheduler.

Tracking their status and monitoring for progress.

1. What is a container?

The container refers to a collection of resources such as memory, CPU, disk and network IO.

1. What is a task?

A job which is divided in to smaller parks is known as task.

In Hadoop, “MapReduce Job” splits the input dataset into independent chunks which are processed by the “Map Tasks” in a completely parallel manner.

1. What is a job in YARN?

Yarn takes care of the execution of the program by allocating necessary resources.

It manages the resources required for execution of the program.

Job Scheduling and monitoring.

1. What is HQL?

HQL stands for Hive Query Language. It is a dialect of SQL which hive uses.

It has a rich set of built in functions to be applied. It is not completely ANSI complaint. It supports most of the SQL commands for data querying.

1. What are managed and external tables in hive?

Managed Tables:

Hive controls the life cycle of a table along with data, when the table is dropped, data is deleted from HDFS

External Tables:

Hive just control the structure not the data, when the table id dropped, the data remains on HDFS and only the structure is removed.

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.

1. What are the different modes of execution in hive?

Modes of execution of hive

* Interaction Mode
* Batch Mode
* Inline Mode

1. What is UTDF’s in hive?

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” command used to alter the hive configuration parameters

“Hive-site.xml” file is used to override any of the values of hive configuration parameters from the java class file.

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

Creating Tables:

* Creating tables through a sub query-CTAS(Create Table As Select)
* Conditional creation of databases and tables
* Cloning of tables(using only its structure)

Loading data:

* Populating data to a table through subquery
* Multi table insert
* Exporting output of the queries in to a file

1. What are the 2 types of partitioning in hive?

Static Partitioning and Dynamic Partition

1. How to extract hive data into a directory?

INSERT OVERWRITE DIRECTORY ‘path of the directory’

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘,’

LINES TERMINATED BY ‘\n’

SELECT \* FROM TABLE\_NAME;

1. What is hive metastore?

Provides information about the data layout or structure of data in HDFS.

**QUERIES**

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

select distributer\_name, sum(volumeOUT) total\_volume\_OUT from petrol\_partition group by distributer\_name;

1. **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 district\_id, sell\_rate\_million, volumeout petrol\_amount\_sold from petrol\_partition order by sell\_rate\_million desc limit 10;

1. **Find real life 10 distributor name who sold petrol in the least amount.**

select distributer\_name,sell\_rate\_million from petrol\_partition

order by sell\_rate\_million limit 10;

1. **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 distributer\_name, (volumein-volumeout) Vol\_Diff from petrol\_partition where (volumein-volumeout)>500;

**OLYMPICS DATA SET**

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

select country,sum( total\_medals)from Olympics where type\_of\_game='Swimming' group by country;

1. **Display real life number of medals India won year wise.**

select year,sum(total\_medals) as Total\_medals from olympics

where country='India' group by year;

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

select country,sum(gold\_medal) as Total\_gold\_medals from olympics

group by country;

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

select country,sum(gold\_medal) as Total\_gold\_medals from olympics

group by country;

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

select year,country,type\_of\_game from olympics

where type\_of\_game='Shooting' group by year;