Manish. B. Shriyan
4AL16CS131
8th Sem 'B'

D - Apache hive with minimum 5 of hive query language commands.

Hive is data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big data & makes querying & analyzing easy. Initially Hive was developed by Facebook, later the Apache software foundation took it up and developed it further as an open source, under the name apache Hive.

- To work in Hive with Hadoop user with access the HDFS can run the Hive queries
- Simply enter the hive command. If Hive start correctly to get a hive > prompt

```
$hive
(some message may show up here)
hive>
```

- Hive command to create and drop the table that hive commands must end with a semicolon (;)

```
hive> CREATE TABLE pokes (foo INT, bar STRING);
ok
Time taken: 1.705 seconds
```

- To see the table is created

```
hive> SHOW TABLES;
ok
Pokes
Time taken: 0.17 seconds, Fetched: 1 row(s)
```

- To drop the table.

```
hive> DROP TABLE pokes;
ok
Time taken: 4.038 seconds
```

- The first step is to create the table & can be developed using a web server log file

```
hive> CREATE TABLE log (t1 string, t2 string, t3 string, t4 string, t5 string, t6 string, t7 string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ' ';
```

- NEXT to load the data from the sample log file, the file is found in the local directory & not in HDFS.

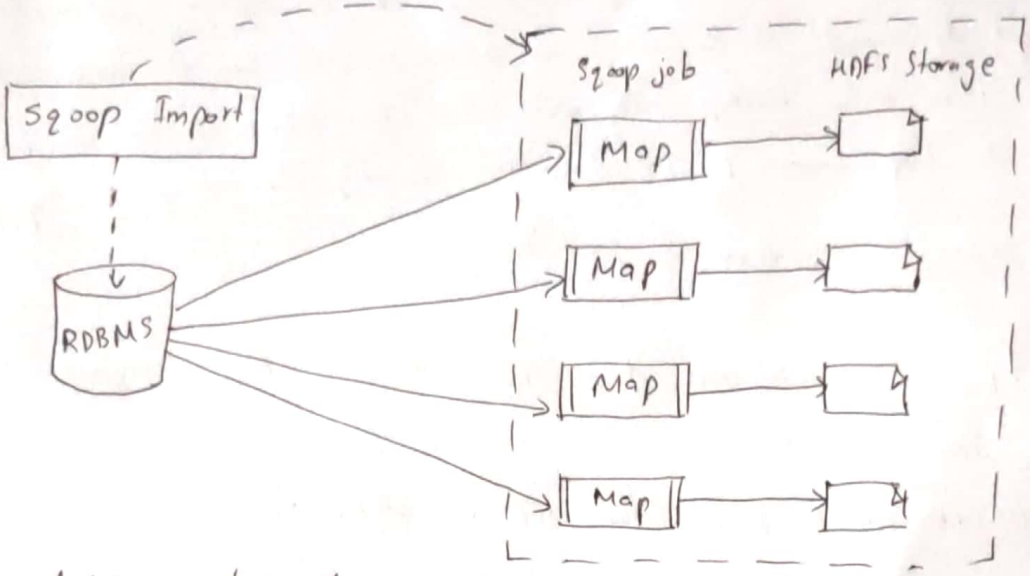
```
hive> LOAD DATA LOCAL INPATH 'SAMPLE.log' OVERWRITE INTO TABLE logs;
```

- Finally the select step. that this invokes a Hadoop MapReduce operation. The results appear at the end of output.
- To exit Hive, simply type exit.

```
hive> exit;
```

2) Explain Apache Sqoop import & export methods with neat diagram.

- Sqoop is used to:
  - → import data from a relational database management System.
  - → export the data back into an RDBMS.



The data import is done in 2 steps:

① Sqoop examines the database to gather the necessary metadata for the data to be imported.

② Map-only Hadoop job: Transfers the actual data using the metadata.

- The imported data are saved in an HDFS directory.
- Sqoop will use the data name for the directory, or the user can specify any alternative directory where the file should be populated.

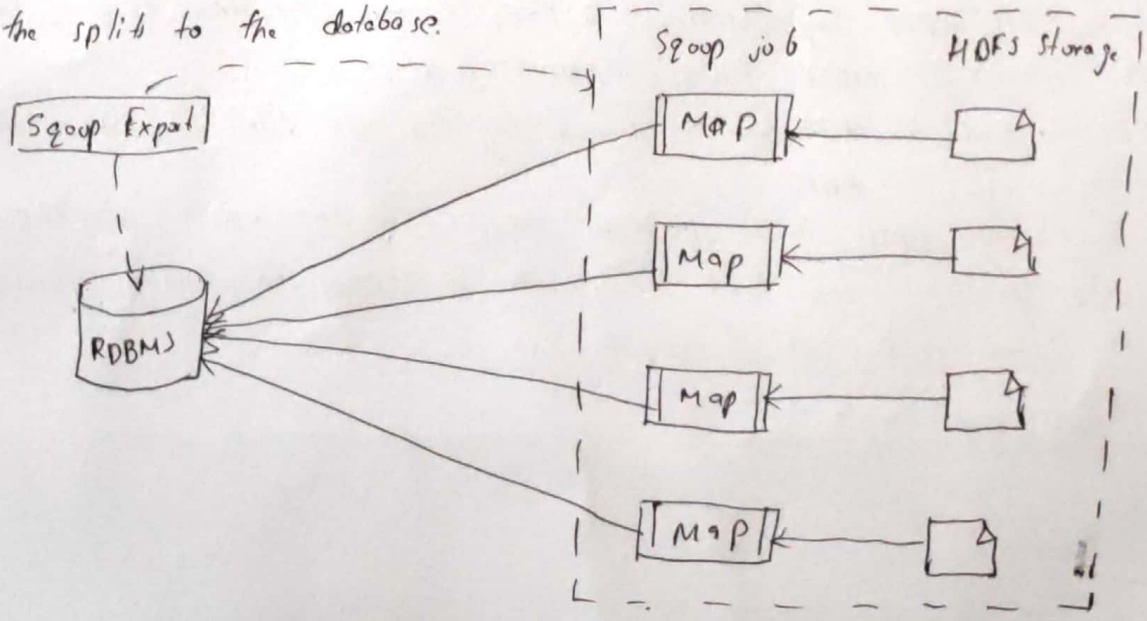
⇒ Sqoop Export method:

Done export from the cluster works in a similar fashion the export is done in:

① Examine the database for metadata

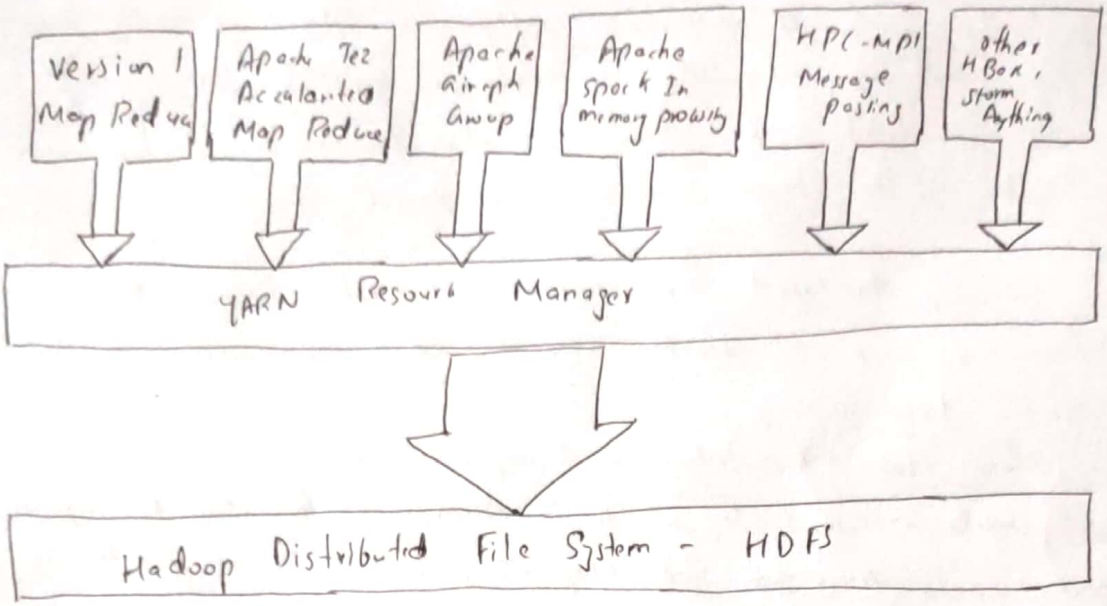
② Map-only Hadoop job in to write the data to the database.

Sqoop divides the input data set into splits, then uses individual map tasks to push the splits to the database.



3) Explain YARN application framework.

YARN presents a resource management platform which provides Servis such as scheduling, fault monitoring, data locality & more to map reduce & other frameworks. Below figure illustrates some of the frameworks that will run under YARN



Distributed - shell:

- Distributed-shell is an exampl application included with the Hadoop core components that demonstrates how to write application on top of YARN
- It provides a simple method for running shell commands & scripts in containers in parallel on Hadoop YARN cluster.

Hadoop Map Reduce:

- MapReduce was the first YARN framework and drove many of YARN's requirements. It is integrated tightly with the rest of Hadoop ecosystem projects, such as Apache Pig, Apache Hive & Apache Oozie.

→ Apache Tez:

- Many Hadoop jobs involve the ~~exp~~ execution of a complex directed acyclic graph (DAG) of tasks using separate MapReduce stages. Apache Tez generalize their process & enables these tasks to be spread across stages so that they can be run as a single all encompassing job.
- Tez can be used as a MapReduce replacement for project such as Apache Hive & Apache Pig. No changes are needed to the Hive or Pig application.

→ Apache Giraph:

- Apache Giraph is an iterative graph processing system built for high scalability. Facebook, Twitter & LinkedIn use it to create social graphs of users
- Giraph was originally written to run on standard Hadoop V1 using the MapReduce framework but that approach proved inefficient and totally unnatural for various reasons.
- The native Giraph implementation under YARN provides the user with an iterative processing model that is not directly available with MapReduce.

- In addition, using the flexibility of YARN the Giraph developers plan on implementing their own web interface to monitor job progress.

→ Hoya : HBase on YARN

- The Hoya project creates dynamic & elastic Apache HBase clusters on top of YARN
- A client application creates the persistent configuration files, sets up the HBase cluster XML files, and then asks YARN to create an application master.
- YARN copies all files listed in the client's application-launch request from HDFS into the local file system of the chosen server, and then executes the command to start.

→ Apache Spark:

- Spark was initially developed for application in which keeping data in memory improves performance, such as iterative algorithm which are commonly in machine learning & Interactive data mining.
- Spark differs from classic MapReduce in 2 important ways.
  - First, Spark holds intermediate results in memory, rather than writing on disk.
  - Second, Spark supports more than just MapReduce functions is, it greatly expands the set of possible analysis that can be executed over HDFS data stores.

→ Apache Storm:

- This frame work is designed to process unbounded streams of data in real time.

4) Explain Apache Spark & Apache REEF.

- Apache Spark:

- Spark was initially developed for application in which keeping data in memory improves performance such as iterative algorithm.
- Since 2013 Spark has been running on production YARN clusters at Yahoo.
- The advantage of porting and running Spark on top of YARN is the common resource management & single underlying File System.

Apache REEF:

- The REEF project by Microsoft recognizes the task of writing a custom application on YARN & factors out several components that are caching, fault detection & checkpoints.
- Designers can build their apps on top of REEF more easily than they can directly on YARN and reuse these common services/libraries.

5) Write short note on Apache Ambari:

→ Apache Ambari is an open-source administration tool deployed on top of Hadoop clusters and it is responsible for keeping track of the running applications & their states. Apache Ambari can be referred to as a web-based management tool that manages, monitors &

provision the health of Hadoop clusters.
It provides a highly interactive dashboard that allows administrators to visualize the progress and states of every application running over the Hadoop clusters.
- Instantaneous insight into the health of the Hadoop cluster using preconfigured operational metrics.
- User-friendly configuration providing an easy step-by-step guide for installation.
- Installation of Apache Ambari is possible through Hortonworks Data Platform (HDP)
- Monitoring dependencies & performance by visualizing & analyzing jobs & tasks.
- Authentication, authorization & auditing by installing kerberos-based Hadoop clusters.
- Flexible & adaptive technology fitting perfectly in the enterprise environment.

6) Define data warehouse & write the design consideration for DW?

→ - A data warehouse (DW) is an organized collection of integrated subject oriented database designed to support decision support function.
- DW is organized at the right level of granularity to provide clean integrated data in a standardized format for reports, queries & analysis
- DW is physically & functionally seperate from an operational & transactional database.

DW supports business reporting & data mining activities.
- The objective of DW is to provide business knowledge to support decision making.

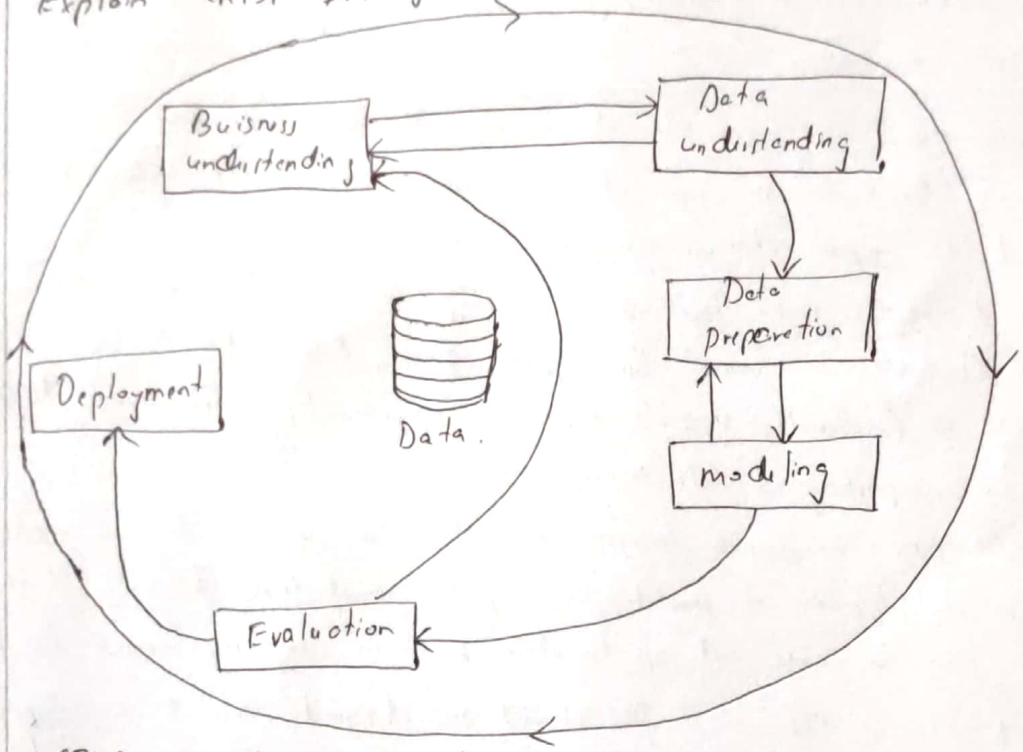
For DW to serve its objectives, it should be aligned around those decisions. It should be comprehensive easy to access and up-to date. Here are some requirements for DW.

① Subject oriented: To be effective a DW should be designed around a subject domain, i.e. to help solve a certain category of problems.
② Integrated: DW should include data from many function that can shed light on a particular subject area. Thus the organization can be benefited from a comprehensive view of the subject area.
③ Time variant (time series): The data in DW should be grown daily or other chosen intervals. That allows latest comparison over time.
④ Non volatile: DW should be persistent, that is, it should not be created on the fly from the operational database, thus, DW is consistently available for analysis across the organization & overtime.
⑤ Summarized: DW contains rolled-up data at the right level for queries & analysis. The process of rolling up time data helps create consistent granularity for effective comparisons. It also helps reduce the number of variables or dimensions. Data to make them more meaningful for the decision makers.
⑥ Not Normalized: DW often uses a store schema, which is rectangular central table

Surrounded by some look-up tables. The sigle table view significantly enhances speed of queires.

- Metadata: Many of the variables in the database are compund from other variables in the operational database. Ex:- total daily sales may be a computed field. The method of calculation for each variable should be effectively documented. Every element in the DW should be sufficiently well-defined.
- Near Real-time &/or right-time: DW's should be updated in real-time in many high transaction volume industries such as airlines.

7) Explain CRISP-DM cycle with neat Diggram.



- CRISP-DM has six essential steps.

① Business understanding
② Data preparation
③ Modelling
④ ~~data~~ model evaluation
⑤ dat understanding
⑥ Dissemination and

There are the best practies of data mining.

8) Explain 5 important data mining techniques

- The most important class of problems solved using data mining are classification problems classification techniques are called supervised learning as there is a way to supervise whether the model is providing the right or wrong answers. These are problems where data from past decision is mined.

There are most popular data mining technique for many reasons.

→ Decision tree:

① Decision trees are easy to understand and easy to use, by analysts as well as executives. They also show a high predictive accuracy.

② Decision trees select the most relevant variables automatically out of all the available variable for decision making.

③ Decision trees are tolerated of data quality issues and do not require much data preparation from the users.

→ Regression:

① This is a most popular statistical data mining technique the goal of regression is to derived smooth well-defined over to best the data.

② Regression analysis techniques for example can be used to model and predict the energy consumption as a function of daily temperature.

→ Artificial Neural Network:

① ANN is a sophisticated data mining technique from the stream in computer Science. It mimics the behavior of human neural structure. Neurons receive stimuli process and communicate their results to other neurons successively & eventually a neuron outputs a decision.

② The neural network can be trained by making a decision over and over again with many data points. It will continue to learn by adjusting its internal computation & communication parameter based on their feedback received on its previous decisions.

→ Cluster Analysis:

- It is an exploratory learning technique that helps in identifying a set of similar groups in the data. It is a technique used for automatic identification of natural grouping of things. Data instances that are similar to each other categorized into one clusters, while data instances that are very different from each other are categorized into separate clusters. There can be any number of clusters that could produced by the data. The k-means technique is a popular technique and allow the user guidance in selecting the right number (k) of clusters from the data.

→ Association rules:

- These are a popular data mining method in business especially where selling is involved. Also known as market basket analysis it helps in answering question about cross-selling opportunities
- This is the heart of the personalization engine used by ecommerce sites like Amazon.com & streaming sites like netflix.com. The technique helps find inter relationship between variable.

9) Define data visualization & explain the types of charts.

- Data visualization is the art & science of making data easy to understand and consume, for the end user. Ideal visulization shows the right amount of data in the right order, in the right visual form to convey the high priority information

Types of charts:

① line graph:
This is a basic and most popular type of displaying information. It shows data as a series of points connected by straight line segments. It mainly with time-series data, time is usually shown on the x-axis. Multiply variables can be represented on the same scale on y-axis to compare of line graph of variables.

② Scatter plot:
This is another basic & useful graphic form. It helps reveal the relationship blw two variables. In the easiest it shows two dimensions. unlike in a line graph there are no line segments connecting the points.

③ Bar graph:
A bar graph shows thin colorful rectangular bars with their lengths being proportional to the values represented the bars can be plotted vertically or horizontally The bar graphs use a lot of more Ink than the line graph & should be used when line graphs are inadequate.

④ Stacked Bar graph:
These are a particular method of doing bar graph. values of multiple variable are stacked one on top of other to tell an interesting story.

⑤ Histogram:
These are like bar graphs, except that they are useful showing data frequencies or data values on classes (or range) of a numerical variables.

⑥ Pie charts: These are very popular to show the distribution of a variable such as sales by region. The size of a slice is representative of relative strength of each value.

⑦ Box charts: These are special form of charts to show the distribution of variable. The box show the middle half of the values. which ~~not~~ on both sides extend to the extreme values in either direction.

⑧ Bubble graph: This is an interesting way of displaying multiple of in scatter plot with many data point marked on the dimensions.

⑨ ~~These are~~ Dials: These are charts like the speed dial in the shows whether the variables value is the low range, medium range, or high range. These range can be coloured.

(10) Geographical: Data maps are particularly useful maps to denote statistics.

7th question Continuation:

(1) Business understanding:
- The first & important step in data mining is asking the right business questions. A question is a good one if answering it would lead to large payoffs for the organization, financially and otherwise
- In other words, selecting a data mining projects is like any other project. in that it should strong payoffs its the project is successful.

(2) Data understanding:
A related step in to understand the data available for mining. One needs to be imaginative in sourcing for many elements of data through many sources in helping address the hypothesis to solve a problem without relevant data, the hypothesis cannot be tested.

(3) Data preparation: The data should be relevant, clean & of high quality. It's important to assemble a team that has a mix of technical & business skills, who understand the domain & the data. Data cleaning can take 60-70% of the time in data mining project.

(4) Modelling: This is the actual task of running many algorithms using the available data to discover if the hypothesis are supported. patience is required in continously engaging with the data until the data yields some good insights. A host of modeling tools and algorithms should be used.

(5) Model Evaluation: One should not accept what the data says at first. It is better to triangulate the analysis by applying multiple data mining techniques & conducting many what-if scenarios to build confidence in the solution. One should evaluate & improve the models predictive accuracy with more test data.