Big Data Analytics

Manoj. L
4AL16CS125

Assignment - 2

1) Apache Hive with Minimum 5 of Hive query language Command.

⇒ Apache Hive is a data Warehouse Infrastructure build on top of Hadoop for providing data summarization and hoc queries & Analysis of large data sits using SQL language Called HiveQL. SQL Queries Over Petabytes of data using Hadoop & others The following features

(i) Tools to enable easy data extraction, Transformation and loading.

(ii) A Mechanism to Impose Structure On variety of data formate

(iii) Query Execution via Map Reduce.

⇒ Hive Query Language Command:-

(i) To start Hive, Simply Enter The hive Command

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

(ii) As a simplest test, Create and drop table: Hive Must End with (;)

```
hive > CREATE TABLE Pokes (foo INT, bar STRING);
OK
Time taken : 1.075 Seconds.
```

(iii) hive > SHOW TABLES;

```
OK
Pokes
Time taken : 0.174 Seconds, fetched : 1 row(s)
hive > DROP TABLE Pookes;
OK
Time taken : 4.08 Seconds
```

(iv) A More detailed Example Can be developed using Web Server log file to summarize text Message types

```
hive > CREATE TABLE log (t1 String, t2 String, t3 String, t4 String, t5 String
ROW FORMAT DELIMITED FIELDS TERMINATED BY ' ';
OK
Time Taken: 0.129 Seconds.
```

(v) Load The Data in this case from sample log file.

This file is Available from Example code download.

```
hive> LOADDATA LOCAL INPATH 'sample.log' OVERWRITE INTO
TABLE LOG;
Loading data to table defaults. logs
Table default. logs Stats : [numfiles = 1, numRows = 0, totalSize =
99271, rawDataSize = 0]
OK
TIME taken : 0.953 seconds.
```

(2) Explain Apache sqoop import and export Methods With neat diagram

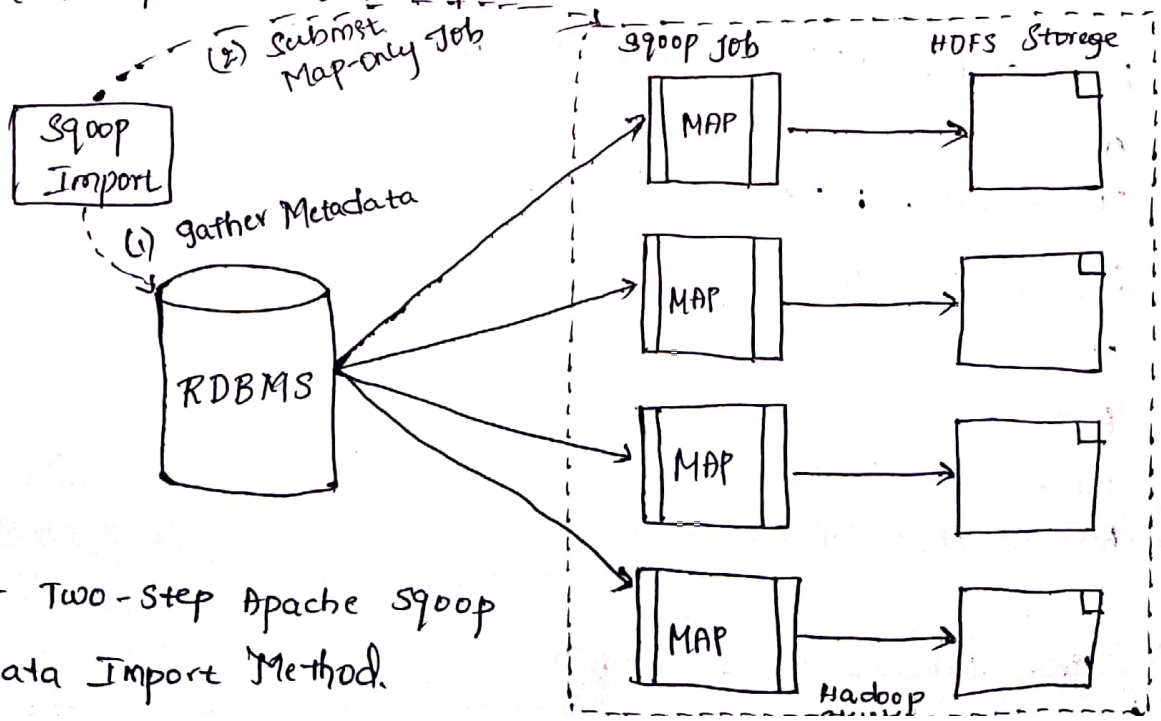


fig 1:- Two-step Apache Sqoop Data Import Method.

→ fig 1 describes Sqoop data Import (HDFS) process. The data Import is done in 2 steps

→ Step I :- Sqoop Examins the database together the Necessary Metadata for data to be Imposted.

→ Step II :- It is Map-only hadoop job that sqoop Submits to the cluster. This job does actual data transfer using Metadata captured in previous steps.

→ Note Each Node doing Import must have access to Database.

→ The Imported data are Saved in HDFS directory

→ Sqoop will use database name for the directory where the files should be populated.

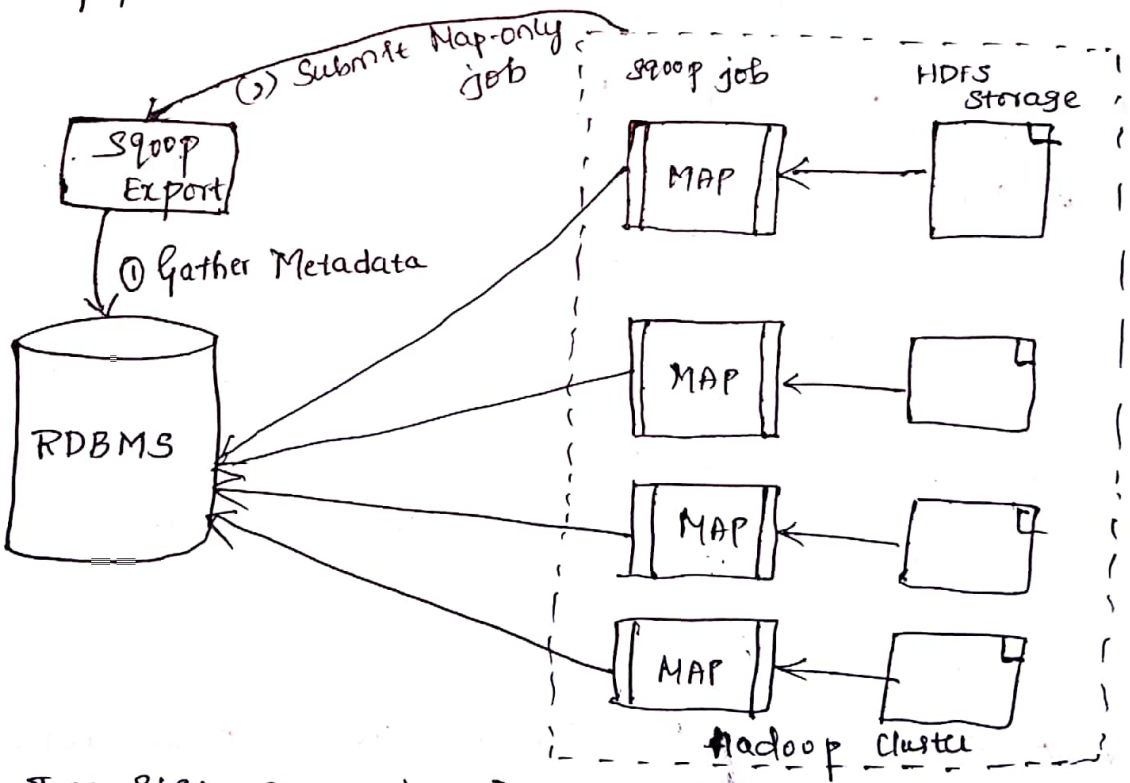


Fig. 2 :- Two Step Sqoop data Export Method.

(i) The Export is done in 2 steps as shown in fig 2

(ii) Step-1 :- is to examine database for Metadata

(iii) The Export Step again use Map-only Hadoop job to write data to database

(iv) Sqoop divides I/p data set into splits, Then uses Individual Map takes to push the splits to databases.

(v) Again, This process assumes The MAP tasks have Access to database.

(3) Explain YARN Application framework.

⇒. The yarn farework exists to Manage Application.

- A yarn application Implements a specific function that runs on Hadoop.
- A yarn Application Involves 3 Components.
  - Client
  - Application Master (AM)
  - Container.

(i) YARN Client :-

* Launching a new yarn application start with yarn client Communicated with Resource Manager to Create New yarn
* Part of this process Involves yarn client informing the Resource Manager of Application Master's physical resources requirments.

(ii) YARN APPLICATION MASTER :-

* Application Master is Master Process of a YARN application
* It doesn't perform any application specific work, as there function are delegated to container.
* Once the application master is started it will periodically send heartbeats to the Resource Manager to affirm its health & to update the record of its resource demand.

(iii) YARN Container

* A Container is an Application - Specific process that's Created by Node-Manager on behalf of an Application Master
* At the fundamental level, a Container is Collection of Physical resource such as RAM, CPU wires & disks on the single Node.

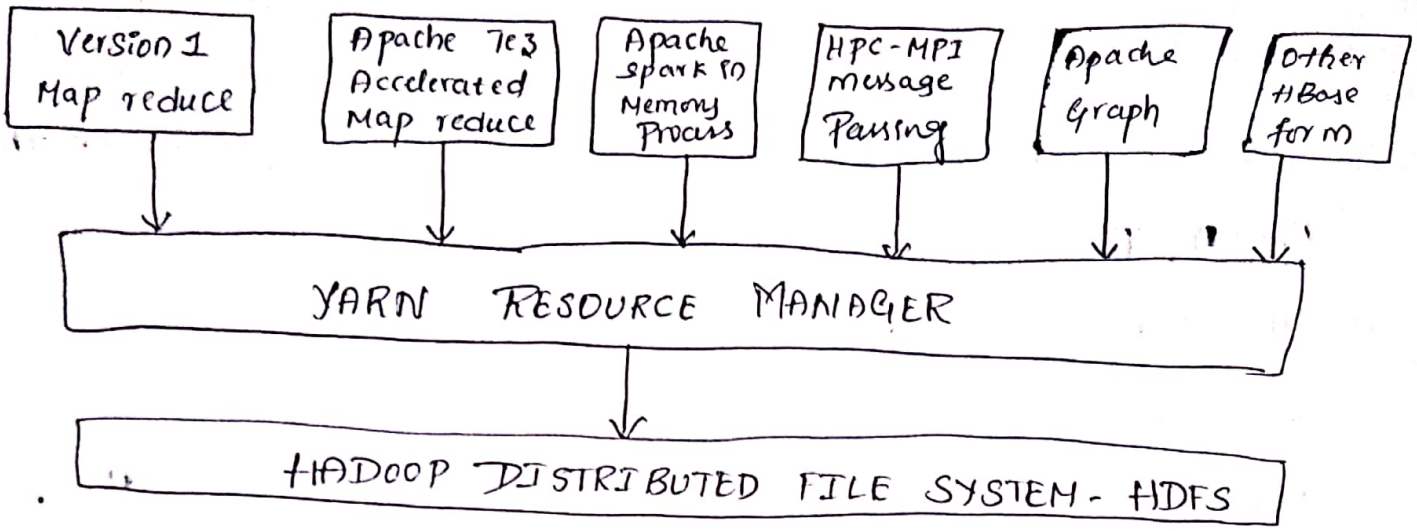


Fig :- Example of Hadoop Version 2 Eco System.

④ Explain Apache Spark and Apache REEP

(*) Apache Spark :-

(i) Spark was initially developed for an application in which keeping data in Memory Improve Performance Such as Algorithm, which are Common in Machine learning.

(ii) Spark differ from classic MapReduce in 2 Important ways:

(1) Spark holds Intermediate results in Memory rather than Writing Them to disk.

(2) Spark Supports more Than just MapReduce function.

(iii) It greatly Expands the set of Possible Analyses That can be Executed Over HDFS data Stores.

(iv) It also Provides API's in Scala, Java & Python.

(v) The advantage of Porting & running Spark on Top of YARN is common resource Management and a single underlying file sy.

Apache REEF:

(i) YARN's flexibility sometimes requires significant efforts on the part of application implements.

(ii) The steps involved in writing a custom application on YARN include building your own application master, performing client and container management and handling aspects of faults, execution flow.

(iii) It greatly expands the set of possible analyses that can be executed over HDFS data stores.

(iv) It also provides API's in Scala, Java & Python.

(v) REEF's design makes it suitable for both MapReduce & DAG-like execution as well as interactive and iterative computation.

(5) Write a short note on Apache Ambari.

⇒ (i) Managing a Hadoop installation by hand can be tedious and time consuming. In addition to keep configuration files synchronized across a cluster, starting, stopping and restarting Hadoop services & dependent services in right order is not a simple task.

(ii) The Apache Ambari's graphical management tool is designed to help you easily manage these and other Hadoop administration issue.

(iii) Along with being an installation tool Ambari can be used as a centralized point of admination for Hadoop cluster.

(iv) Using Ambari, The user can Configure cluster services, monitor the status of cluster host or services, visualize hotspots by service Metric, start or stop Service and add news hosts to the cluster.

(v) All of these features infuse a high level of agility into the Process of Managing and Monitoring a distributed Computing Environment.

(vi) Ambari also attempts to Provide real-time reporting of Important Metrics.

(6) Define Data Warehouse and write design Consideration for DW.

⇒ . A datawarehouse (DW) is an organized collection of integrated, subject-Oriented databases design to supposed decision support function.

. DW is Organized at the right level of granularity to provide clean

. Enterprise-wide data in standarlized formate for reports quaries and analyses.

* Design Consideration for DW

The Objective of DW is to provide business knowledge to support decision Making. Three are some requirment for good 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 Problem.

(2) Integrated : The DW should include data from many function that can shed light on particular subject area. Thus the Organization can benifit from a Comprehensive view of subject area.

③ Time-Variant:- The data in DW should grow at daily or other choosen Interval. That allows latest Comparisons Over times.

④ Non-Volatile:- DW should be persistent, i.e it should not be Created on the fly from the Operational databases

⑤ Summarized:- DW Contains rolled up data at right level of queries and analysis. The process of rolling up the data helps to Create Consistent for effective Comparisons.

⑥ Not Normalized :- DW often uses a star schema, which is rectangular central table, surrounded by some look-up tables. The single table view significantly enhances speed of queries.

⑦ Metadata :- Many of variables in databases are Computed from other variables in Operational databases. The Method of Calculation for each variable should be effectively documented Every element in DW should be well-defined.

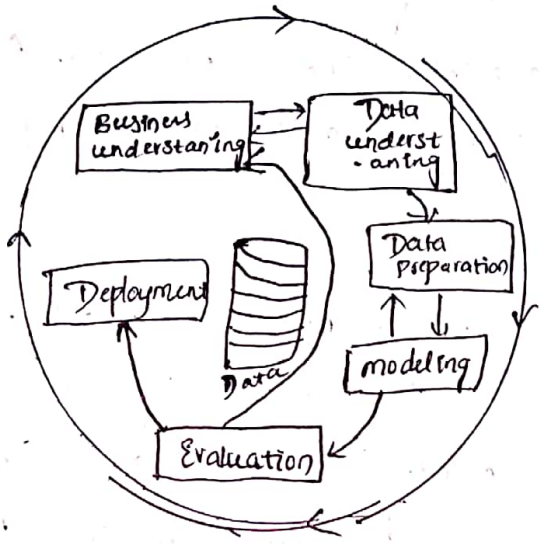
⑧ Near Real-time or Right-time :- DW should be updated in near real-time in many high transaction volume industries. The cost of updating DW in real-time could be disclouing Through. Another Clownside of real-time DW is possibilities of Inconsisten cies in reports Drawn just a few Minutes apart.

Q7) Explain CRISP-DM cycle with Neat Diagram.

⇒ ⊙ Effective and Succesful use of data Mining activity reaires both business and technology skills.

⊙ It also helps one imagine possible relationship in data and Create hypotheses to test it.

- There are several best pratices learned from the use of data Mining techniques over a long Period of time.
- The data Mining Industry has proposed a cross-Industry standard-ard proces for Data Mining.
- It has six Essential steps.



(1) Bussiness Understanding :-

The Most important step in data mining is asking right business question. A question is good one if answering it would be lead to large Payoffs for Organization, financially and otherwise. There should be strong support for DM Project, which means that Project aligns well with business strategy Thinking Outside the box is important, both in terms of a Proposed model as well as data sets available are required.

(2) Data Understanding :-

A related Important step is to Understand the data avaiable for mining. One need to be Imaginative in slowing for many elements of data through many sources in helping address of hypotheses to solve a Problem. Without relevant data, the hypothesis cannot be tested.

(iii) Data Preparation :-

The data should be clean, relevant & hight quality. It's important to assemble a team that has mixture of technical and business skills Data chaning can take 60-70% of time in data Mining Project.

(iv) Modeling :- This is actual task of running many Algorithms using the available data to discover if hypotheses are supported. a hosts of Modeling tools and algorithm should be used. A tool could be tried with different Opertions such as running different decision tree Algorithm.

(v) Model Evaluation :- One should not accept what the data says at first. It is better to triangulate the analysis of by applying multiple data mining Techniques and Conducting Many what if scenarios, to build confidence in the Sol". One should Evaluate and Improve the Model's predictive accuracy with More test-data.

(vi) Dissemination and rollout :- It is Important that data Mining solution is presented to key stakeholders, and is deployed in the Organization. Otherwise the project will be wasted of time and will be setback for establishing and supporting a data-based decision process culture in the Organization. The model should be eventually embedded in the Organization's business processes.

Q8] Explain 5 Important data Mining Techniques.

(1) Decision Trees :- This are most important DM techniques for many reasons :

(a) Decision tree are easy to understand and use by analyst, as well as Executives. They also show high predictive accuracy.

(b) Decision tree select most relevant Variable automatically out of all the available variables for decision making.

(c) Decision tree are tolerant of data quality issues & do not require much data preparation from users.

(d) Even Non-linear relationship can be handled well by decision tree.

(ii) Regression :-

(a) Most popular statistical data Mining Technique

(b) The goal is to derive smooth well-defined curve to best the data.

(c) It can be used to Model & predict the energy Consumption as function of daily temperature.

(d) Applying Non-linear regression Equation will fit data very well with high Accuracy.

(e) The Accuracy of regression Model depends entirely upon the data-sets used and Not at all on the Algorithm or tools used.

(iii) Artifical Neural Network :-

(a) ANN is sophisticated DM technique from Artifical Intelligence in Computer Science.

(b) A decision task may be processed by just one neuron & result may be communicated soon.

(c) The Neural Nw can be trained by making a decision over again with many data points.

(d) The Neural Nw can be learned Enough & begin to match the predictive Accuracy of human Expert or Alternative classification Tech

(e) ANNs require a lot of data to train it to develop from Predictive Ability.

(9) Define data visulization and Explain types of charts.

⇒ Data Visulization is the art and science of Making data easy to understand and Consme for the end User.

* Types of charts

(i) line graph:-

It shows data as a series of Point connected by straight line segment. If Mining with time-series data, time is usually shown on x-axis. Multiple Variable shows on y-axis to compare the graph lines of all variables.

(ii) Scatter plot:-

It Helps to reveal the relationship blw 2 Variables. Unlike in a line graph, there are no line segments Connecting the Points.

(iii) Bar graph:-

A Bar graph shows thin colorful rectangular bars with their length being Proportional to values represented the bars can be Plotted vertically or Horizontally.

(iv) Stacked Bar graph:-

Values of Multiple Variable are stacked up one on top of other to tell an Intersing story. Bar can be classified such as total height of every bar is equal can be relative composition of each bar.

⑤ Histograms :-

These are like bar graphs, Except that they are useful in showing data frequencies or data values on classes Numerical values.

⑥ Pie chart :-

Pie chart shows distribution of variable such as sales by region.

⑦ Box chart :-

Shows distribution of variables. The box shows Middle half of values while whiskers on both sides extend to Extreme value.

⑧ Bubble graph :-

This is Interesting way of displaying multiple dimension in one chart. It is variant of scatter plot with many data Points Marked on 2 dimensions.

⑨ Dials :-

Charts are like speed dial in Car, that shows whether the variable value is in low range, Medium or high range.

⑩ geographical Data Maps :-

Useful Map to denote statistices

⑪ Pictographs :-

One can use picture represent data.