BIG DATA

Hands On Session - 2

HIVE

K V Subramaniam

Usha Devi B G

Dept of Computer Science and Engineering

OVERVIEW

- HIVE is an open-source system for **querying and managing structured data** built on top of Hadoop.
- Hive supports queries expressed in a SQL-like declarative language.
- HiveQL, which are compiled into mapreduce jobs are executed using Hadoop.
- Metastore – A system catalog that contains schemas and statistics, which are useful in data exploration, query optimization and query compilation.

Objective

- HIVE queries on a real world dataset.

Problem Statement

- Find the frequency of books published each year from the data set.

```
"ISBN";"Book-Title";"Book-Author";"Year-Of-Publication";"Publisher";"Image-URL-S";"Image-URL-M";"Image-URL-L"

"0195153448";"Classical Mythology";"Mark P. O. Morford";"2002";"Oxford University Press";"http://images.amazon.com/images/P/0195153448.01.THUMBZZZ.jpg";"http://images.amazon.com/images/P/0195153448.01.MZZZZZZZ.jpg";"http://images.amazon.com/images/P/0195153448.01.LZZZZZZZ.jpg"

"0002005018";"Clara Callan";"Richard Bruce Wright";"2001";"HarperFlamingo Canada";"http://images.amazon.com/images/P/0002005018.01.THUMBZZZ.jpg";"http://images.amazon.com/images/P/0002005018.01.MZZZZZZZ.jpg";"http://images.amazon.com/images/P/0002005018.01.LZZZZZZZ.jpg"
```

SPECIFICATIONS

1. Hadoop: 3.2
2. Java: 1.8
3. Hive : apache-hive-2.1.0
4. Dataset: Please download the dataset from the forum.

Step 1. To start the Hive Terminal:

a) Run,

- *$ start-dfs.sh*
- *$ start-yarn.sh* (Start hadoop)

b) $ *cd $HIVE_HOME*

c) Run Hive.

```
$ sudo bin/hive
```

OUTPUT Shell will look like

```
Logging initialized using configuration in jar:file:/usr/lib
/hive/apache-hive-0.13.0-bin/lib/hive-    common-0.13.0.jar!/
hive-log4j.properties
hive>
```

d) If hive command gives an error, try removing metastore_db
$ *rm -rf metastore_db* (It is present in the $HIVE_HOME))
directory or $HIVE_HOME/bin directory)

e) $ cd bin/

f) $ schematool -dbType derby -initSchema

g) Run hive again.

Step 2: To create a database

Syntax: create database <database name>;

Example: create database sample_database;

HIVE Hands-on

Step 3: To create a table

Syntax: create table <table name>(attribute_name_1 datatype, attribute_name_2 datatype) row format delimited fields terminated by '<delimiter type>';

Example: *create table sample_table(id INT, name string) row format delimited fields terminated by ' ';*

Step 4: To load data into table

Syntax: load data local inpath '<local absolute path to data.txt>' overwrite into table <table name>;

Example: *load data local inpath '/home/xyz/data.txt' overwrite into table sample_table;*

Step 5: Query the Hive Database.

Syntax*:* SELECT <attribute_name_1>, <attribute_name_2> FROM <table_name > GROUP BY <attribute_name_2> ;

Problem Statement

- Find the number of cars in every city which use gas as a mode of fuel using Hive.

HIVE Hands-on

- **Columns of the Dataset :** The columns are indexed from [0-25] (Ex. Transmission is the 11th index)

- **Sample output :**

City	Number of Cars that use Gas
Bangalore	10
Chennai	12

- Actual output to be displayed as two columns on the terminal inside HIVE shell with each line of the answer having the pair <cityname> <number> .

PES UNIVERSITY ONLINE

THANK YOU

K V Subramaniam
Usha Devi B G

Department of Computer Science and Engineering