WQD7007 Big Data Management

Introduction to Hive

Introduction

- In this lab, we are going to practice how to install, load and access data using Hive.
- Hive is a data warehouse infrastructure tool to process **structured data** in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

Installation

- Online reference: https://www.dezyre.com/hadoop-tutorial/install-hive

Installation

```
• wget https://www.apache.org/dist/hive/hive-1.2.2/apache-hive-1.2.2-bin.tar.gz

• tar -xzf apache-hive-1.2.2-bin.tar.gz

• mv apache-hive-1.2.2-bin /home/{yourname}/hive/
```

- In ~/.bashrc:

```
• export PATH=$PATH:/home/{yourname}/hive/bin
• source ~/.bashrc
```

Installation

- In hive bin folder:
  - In hive-config.sh
    - Add

      ```
      export HADOOP_HOME=/home/wlhoo/hadoop
      ```

      at the end of the file
- Create Hive warehouse

  ```
  – hadoop fs -mkdir /user/hive/warehouse
  – hadoop fs -chmod 765 /user/hive/warehouse
  ```

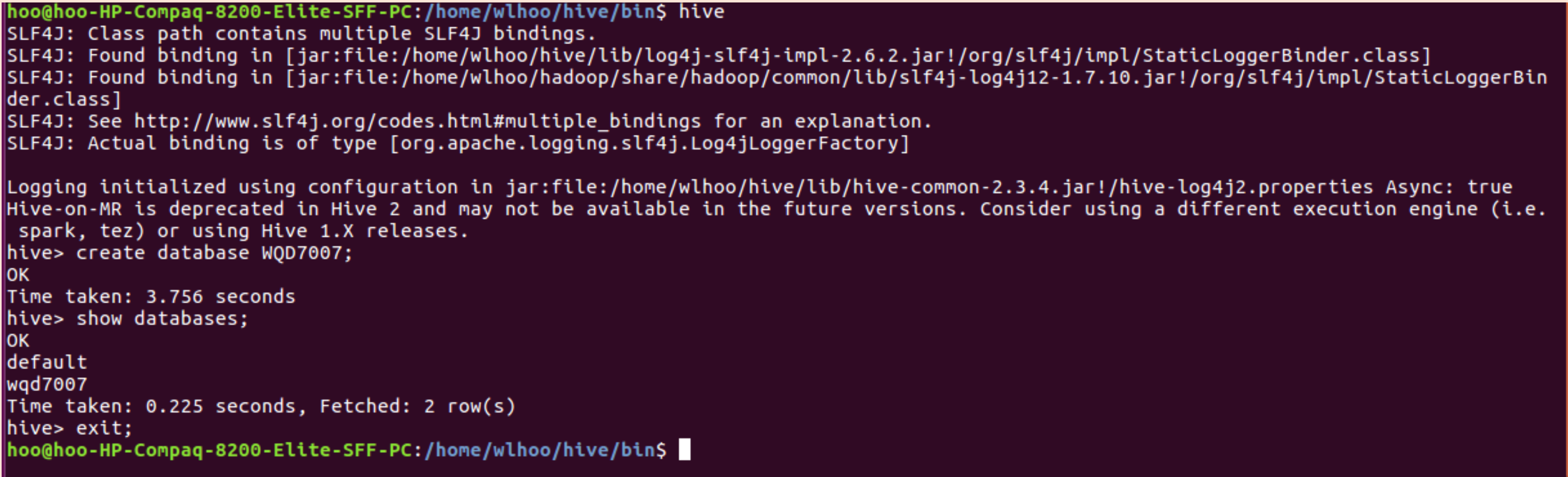
- In bin folder:

  ```
  – ./schematool -initSchema -dbType derby
  ```

- Run ‘hive’ in terminal

Installation

- In hive:
  - create database wqd7007;
  - show databases;
  - exit;



```
hoo@hoo-HP-Compaq-8200-Elite-SFF-PC:/home/wlhoo/hive/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/wlhoo/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/wlhoo/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBin
der.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/home/wlhoo/hive/lib/hive-common-2.3.4.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e.
 spark, tez) or using Hive 1.X releases.
hive> create database WQD7007;
OK
Time taken: 3.756 seconds
hive> show databases;
OK
default
wqd7007
Time taken: 0.225 seconds, Fetched: 2 row(s)
hive> exit;
hoo@hoo-HP-Compaq-8200-Elite-SFF-PC:/home/wlhoo/hive/bin$ █
```

DIY

- try to upload a csv file (batting.csv, download from Spectrum) to hive table.

Queries

- Let’s group the records by year, and select the highest run each year

```
•  SELECT yearid, max(r) FROM batting GROUP BY yearid;
```

- We also need to know the player_id in order to know who the player are:

```
•  SELECT a.yearid, a.player.id, a.r from batting a
      JOIN (SELECT yearid, max(r) FROM batting
      GROUP BY yearid) b
      ON (a.yearid = b.yearid AND a.r=b.r)
```

Other building functions

- Standard deviation of ”run”
  - `SELECT stddev(r) from batting`

- `Min? Max?`

Try another database

1. Download geolocation.csv and trucks.csv
2. Create tables geolocation and trucks from these two CSV files
3. Find out the truck, driver, truck model and their effective miles per gas in Jun13.
4. Find the average miles in May13 by the truck model
5. Find the drivers who “overspeed” using geolocation table
6. Find the city where the drivers overspeed