Big Data Analytics Peiran Zhou
Hive Installation pz2210@columbia.edu

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. This guide teaches you how to install, configure and use basic functions of Hive.

1. Installation:

```
$ brew install hive
$ export HIVE_HOME=/usr/local/Cellar/hive/1.2.1/libexec
$ export JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_60.jdk/Contents/Home
```

2. Start Hadoop:

```
$ hstart
```

```
$ jps
```

```
9906 ResourceManager
11465 Jps
9594 NameNode
9789 SecondaryNameNode
9679 DataNode
9999 NodeManager
```

3. Hive

```
$ cd /usr/local/Cellar/hive/1.2.1/libexec/bin
$ hive
```

1> Using "SHOW TABLES" to see if any tables exist:

```
hive> SHOW TABLES;
```

2> Create a table:

```
hive> CREATE TABLE test (name STRING, gender STRING, year INT, month INT);
hive> SHOW TABLES;
hive> SELECT * FROM test;
hive> quit;
```

3> Practical Example:

Ok, let's create a new table and add real data. The dataset I used is a simple version of 2007.csv airline data. And the data located at /usr/local/Cellar/Data/2007.csv on my laptop. To save more time, I just kept first five columns of original dataset. However, you should keep integrate data.

```
hive> CREATE TABLE airline(
> Year INT,
> Month INT,
> DayofMonth INT,
> DayOfWeek INT,
> DepTime INT)
> row format delimited fields terminated by ',' stored as textfile;

hive> SHOW TABLES;
hive> DESCRIBE airline;
hive> quit;
```

Next step, we need to upload our dataset to HDFS, and so that we can import it to Hive.

```
$ cd /usr/local/Cellar/hadoop/2.7.2/libexec
$ bin/hadoop fs -copyFromLocal /usr/local/Cellar/Data/2007.csv /user
$ bin/hadoop fs -ls /user
$ cd /usr/local/Cellar/hive/1.2.1/libexec/bin
$ hive

hive> LOAD DATA INPATH '/user/2007.csv' INTO TABLE airline;
hive> SELECT * FROM airline LIMIT 100;
```

Do some queries:

```
hive> SELECT * FROM airline WHERE DepTime=802;
hive> SELECT avg(DepTime) FROM airline WHERE Month=2;
```

Now you know how to use Hive and HiveQL to analyze your data that is stored on HDFS. Please try some other examples by yourself~