Institute : Great Lakes Institute of Management, Great Learning,

Hyderabad Batch : Oct 2017, PGP - Big Data Analytics

Assignment : Lab | Big Data on Hadoop

Date : 30th Dec 2017

Author : Utsav Singhi

Data Set : NYSE_daily_File.csv

About Data set-

The NYSE stock Data set which can also be downloaded from the HDFS file browser from Hue in Big Data Lab. The path of file is : "/data/NYSE_daily_File"

NYSE dataset Daily stock data of each company is available live on yahoo finance for each stock exchange worldwide. We have taken the NYSE stock exchange data for this study. The data set is composed of: stock exchange,company symbol, date, open price of the day, high of the day, low of the day, close of the day , volume and adjusted close price.

Problem Statement :

Please complete following tasks based on the data set mentioned above:

1. Create NYSE_Partition table based on the date field. 2. This table needs to store the data as ORC file. 3. Load data of only those records

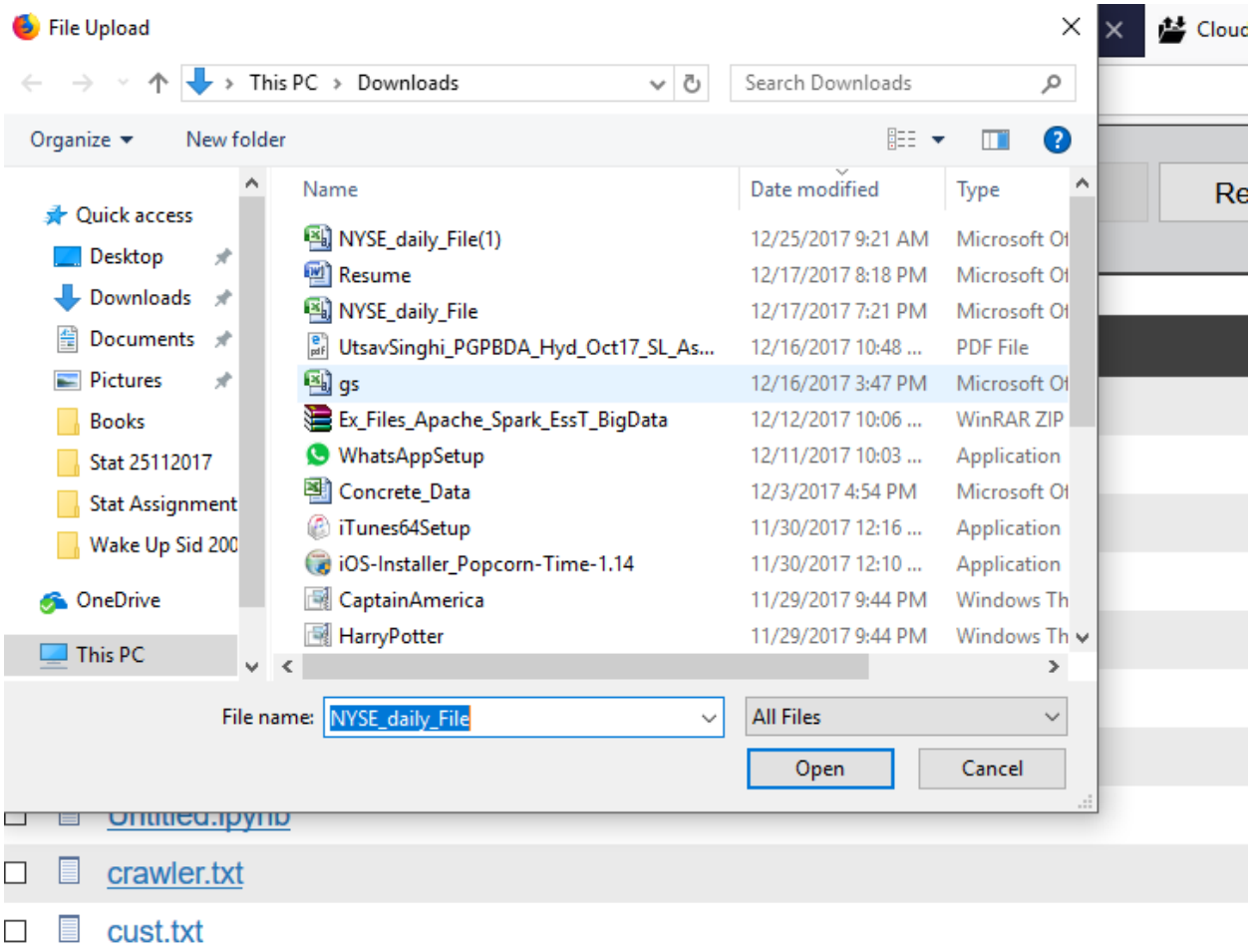
where open price of the day is greater than 68 and high price of the day is less than 70.

Tools available (any of the below) :

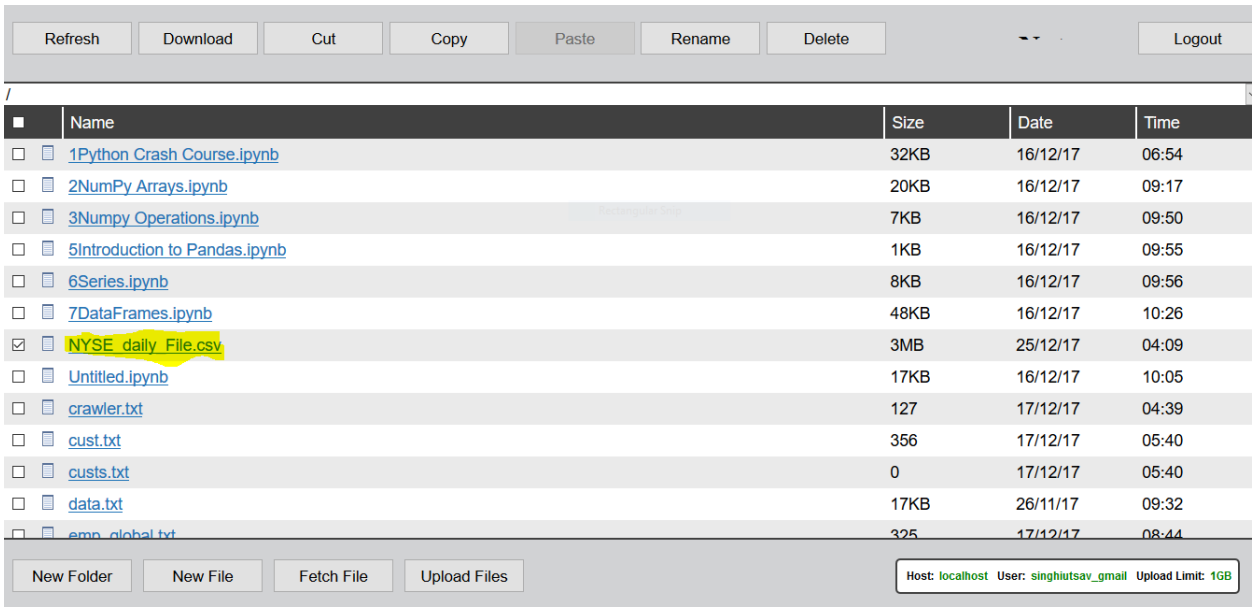
FTP, Webconsole, Hue, Cloudera Manager

Solution

Using FTP Tool to Upload file in local machine



File Uploaded to Local Machine



Now login to Cloudera Manager and creating new directory as ‘Big Data Assignment’

- hdfs dfs –mkdir bigdataassignment
- Using hdfs dfs –ls to see all the directories

```
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-18 12:00 .Trash
drwx------   - singhiutsav_gmail hadoop          0 2017-12-17 07:46 .staging
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-17 05:20 CRM
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-17 05:47 ERP
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-25 03:56 bigdataassignment
-rw-r--r--   3 singhiutsav_gmail hadoop          0 2017-12-17 04:23 crawler.txt
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-17 04:42 crawler_data
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-17 08:54 emp_global
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-11-26 09:37 input
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-17 04:39 inverted_out
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-17 04:41 inverted_out1
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-17 04:42 inverted_out2
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-17 04:49 inverted_out4
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-12-17 05:48 reduce_join_out
drwxr-xr-x   - singhiutsav_gmail hadoop          0 2017-11-26 10:03 wc_out
```

Putting the CSV in the bigdataassignment directory

```
[singhiutsav_gmail@ip-172-166-41-247 ~]$ hdfs dfs -put NYSE_daily_File.csv /user/singhiutsav_gmail/bigdataassignment/
```

Login in Hive

hive> use utsav_test;

```
hive> use utsav_test;
OK
Time taken: 0.528 seconds
```

CREATING external table for reading the CSV and taking the unstructured in structured form.

- CREATE EXTERNAL TABLE IF NOT EXISTS nyse_input_stg2(se STRING, cs STRING, ip_date STRING, op DOUBLE, hd DOUBLE, ld DOUBLE, cd DOUBLE, vol BIGINT, acp DOUBLE) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/singhiutsav_gmail/bigdataassignment'

```
hive>  CREATE EXTERNAL TABLE IF NOT EXISTS nyse_input_stg2(se STRING, cs STRING, ip_date STRING, op DOUBLE, hd DOUBLE, ld DOUBLE, cd DOUBLE,vol BIGIN
T,acp DOUBLE)ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION'/user/singhiutsav_gmail/bigdataassignment';
OK
Time taken: 0.087 seconds
```

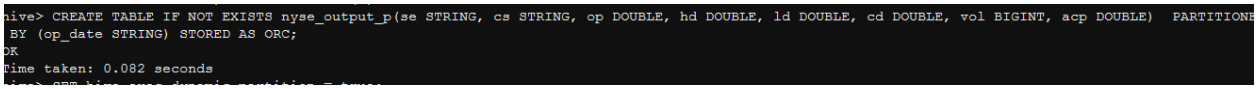
Checking that the data is been loaded properly

- Select * from nyse_input_stg2 limit 5;

```
hive> select * from nyse_input_stg2 limit 5;
OK
NYSE    CLI     31-12-2009      35.39   35.7    34.5    34.57   890100  34.12
NYSE    CLI     30-12-2009      35.22   35.46   34.96   35.4    516900  34.94
NYSE    CLI     29-12-2009      35.69   35.95   35.21   35.34   556500  34.88
NYSE    CLI     28-12-2009      35.67   36.23   35.49   35.69   565000  35.23
NYSE    CLI     24-12-2009      35.38   35.6    35.19   35.47   230200  35.01
Time taken: 0.062 seconds, Fetched: 5 row(s)
```

Create a partitioned table based on date and store it in ORC format

CREATE TABLE IF NOT EXISTS nyse_output_p(se STRING, cs STRING, op DOUBLE, hd DOUBLE, ld DOUBLE, cd DOUBLE, vol BIGINT, acp DOUBLE) PARTITIONED BY (op_date STRING) STORED AS ORC;

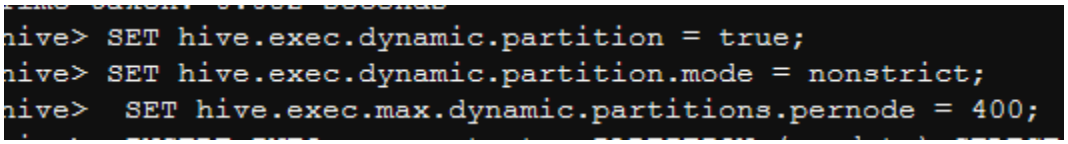


Set some parameters in Hive for dynamic partitioning

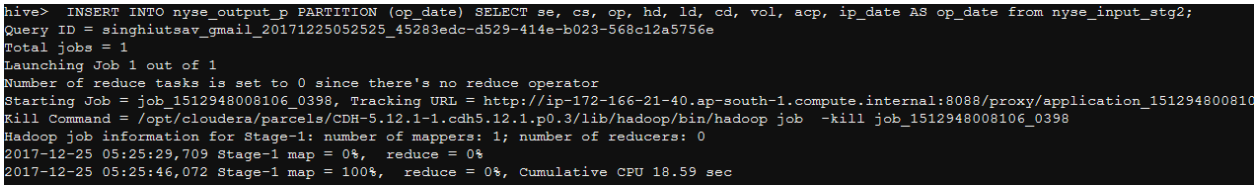
hive> SET hive.exec.dynamic.partition = true;<enter>

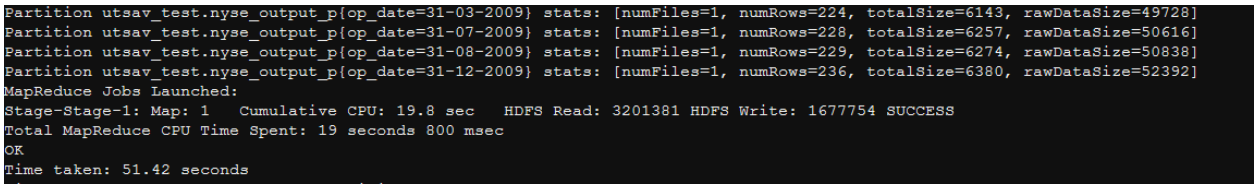
hive> SET hive.exec.dynamic.partition.mode = nonstrict;

hive> SET hive.exec.max.dynamic.partitions.pernode = 400;



hive> INSERT INTO nyse_output_p PARTITION (op_date) SELECT se, cs, op, hd, ld, cd, vol, acp, ip_date AS op_date from nyse_input_stg2;





Verifying data present in Partition table

Hive> select * from nyse_output_p limit 5;

```
hive>  select * from nyse_output_p limit 5;
OK
NYSE    CAJ     53.22   53.22   52.71   53.02   200900  52.15   01-02-2007
NYSE    CVC     25.87   26.48   25.81   26.38   2009000 26.38   01-02-2010
NYSE    CR      25.01   25.36   24.31   25.01   200900  19.35   01-03-1999
NYSE    CLI     18.99   19.26   18.3    18.6    2497900 17.75   01-04-2009
NYSE    CSL     19.31   20.13   18.88   20.07   633500  19.74   01-04-2009
Time taken: 0.633 seconds, Fetched: 5 row(s)
```

As per requirement it value should be between less than 70 and greater than 68

Hive> select * from nyse_output_p where op>68 and hd<70

```
hive>  select * from nyse_output_p where op>68 and hd<70;
Query ID = singhiutsav_gmail_20171225052626_872784aa-1b11-4054-8403-161ab7f6af37
Total jobs = 1
Launching Job 1 out of 1
```