Joining Data Sets with Hive

|  |  |
| --- | --- |
| **Exercise Dir** | ~/material/data |
| **Eclipse Proj** | hive |

**Joining Datasets and Contents Analysis**

**PriceData load into Hive**

1. Go to your ~/workspace/hive directory

2. Check contents of the files using the head command on local file system. Note these two local files

$ head NYSE\_daily\_prices\_F.csv

Do the same on the dividends file:

$ head NYSE\_dividends\_C.csv

3. Verify the two files you have put into HDFS in previous lab

$ hdfs dfs -ls hive\_data/\*.csv

Results:

hive\_data/dividends\_C.csv

hive\_data/prices\_F.csv

Question: How do you quickly check NYSE\_dividends\_C.csv under ~/workspace/hive and prices\_F.csv file in HDFS under ~/hive\_data folder have the same data?

file size is the same: 9537527 bytes

4. Check that the prices\_f table is still in Hive:

$ hdfs dfs -ls /apps/hive/warehouse

**Dividend Data load into Hive**

1. Let's create a new hive file called aggregate.hive

$ touch aggregate.hive

2. Use a text editor to add below lines to create dividend\_c table:

CREATE TABLE dividend\_c (market string, symbol string, trade\_date string, dividend float) row format delimited fields terminated by ',';

3. To load the data into the table, add below command to the file:

LOAD DATA INPATH 'hive\_data/dividends\_C.csv' OVERWRITE INTO TABLE dividend\_c;

4. Now we need to create a table to have the data sets joined. Let's add below command to the file to build the table named stock\_aggregate

create table stock\_aggregate (symbol string, year string, high float, low float,

average\_close float, total\_dividends float);

5. View the file:

$ cat aggregate.hive

Make sure it has the correct commands as above.

6. Execute the hive file

$ hive -f aggregate.hive

7. Verify that data exists in the dividend\_c table in hive:

hive>

select \* from dividend\_c limit 10;

Results:  
 exchange stock\_symbol date NULL

NYSE CY 2008-09-30 16.418

NYSE CY 2008-09-29 0.0

NYSE CPO 2009-12-30 0.14

NYSE CPO 2009-09-28 0.14

NYSE CPO 2009-06-26 0.14

NYSE CPO 2009-03-27 0.14

NYSE CPO 2009-01-06 0.14

NYSE CPO 2008-09-30 0.14

NYSE CPO 2008-06-24 0.14

Time taken: 0.266 seconds, Fetched: 10 row(s)

8. Verify the stock\_aggregate table:

hive>

describe stock\_aggregate;

OK  
 symbol string

year string

high float

low float

average\_close float

total\_dividends float

Question: do you see any data in the stock\_aggregate table?

Answer: No

**Now join the price and dividend data in Hive**

1. Create a new hive file

$ touch price\_dividend\_join.hive

2. Edit the price\_dividend\_join.hive file with a text editor, first adding the data insert statement:

insert overwrite table stock\_aggregate

3. In the same line add below command to the above statement:

select p.symbol, year(p.dates), max(.price\_high), min(p.price\_low), avg(p.price\_close), sum(d.dividend) from prices\_fp

4. Now we need to do a join to the dividend\_c table

left outer join dividend\_c d

5. Add the join:

on(p.symbol=d.symbol and p.dates=d.trade\_date)

6. the below will need to be added to get the aggregate result:

group by p.symbol,year(p.dates);

7. Verify the file:

$ more price\_dividend\_join.hive

insert overwrite table stock\_aggregate select p.symbol,year(p.dates), max(p.price\_high), min(p.price\_low), avg(p.price\_close), sum(d.dividend) from prices\_fp left outer join dividend\_c d on (p.symbol=d.symbol and p.dates=d.trade\_date) group by p.symbol, year(p.dates);

**Execute the hive script**

1.  Now execute the script:

$ hive -f price\_dividend\_join.hive

...

Total MapReduce CPU Time Spent: 7 seconds 910 msec

OK

2. How long does this job run?

Time taken: 43.654 seconds

Note the time taken might be different in your environment.

3. How many Map and Reduce jobs executed?

1 Map and 1 Reduce job

**View the aggregate table contents**

1. Verify a file is generated:

$ hdfs dfs -ls

000000\_0

2. Now use the text command to check the data again and get a row counts:

$ hdfs dfs -text /apps/hive/warehouse/stock\_aggregate/ 000000\_0 | wc -l

1748

**View the aggregate table contents in Hive**

1. Retrieve data from the table:

$ hive

hive>

select \* from stock\_aggregate;

....

FXEN 200 68.53 3.84 5.27 79284 NULL

FXEN 200 710.6 4.6 47.69 3825 NULL

FXEN 200 88.6 81.9 55.24 664 NULL

FXEN 200 94.9 72.1 23.27 0635 NULL

FXEN 201 03.2 32.8 43.03 16 NULL

symbol NULL NULL NULL NULL NULL

2. How many rows were returned from the above query?

Fetched: 1748 row(s)

Does it match the count from Step 4.4?

*Answer: Yes*

3. Use count command to get the total rows:

hive>

select count(\*) from stock\_aggregate;

What is the result?

Answer: 1748

**END**