Practice Selecting Data from Hive

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

**Preparation: Load the data**

1. Go to the ~/workspace/hive directory:

2. Move the raw sample data into HDFS:

$ hdfs dfs -mkdir hive\_data/

\*If you already created “hive\_data” in a previous lab, continue on to the next command.

$ hdfs dfs -put NYSE\_daily\_*prices\_F*.csv hive\_data/ prices\_F.csv

$ hdfs dfs -put NYSE\_dividends\_C.csv hive\_data/ dividends\_C.csv

3. What command do you use to confirm the files are in their expected location in HDFS?

$ hdfs dfs -ls -R hive\_data

**Create a Hive script to create table**

1. Create a file called prices.hive either in the file explorer or vi a terminal editor:

$ touch prices.hive

2. Use terminal to edit the file be issuing [ nano prices.hive ]

create table prices\_F (

market string,

symbol string,

dates date,

price\_open float,

price\_high float,

price\_low float,

price\_close float,

volume int,

price\_adj\_close float)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ',' ;

Tip: You can use cat command to view the contents of the file

**Create a Hive table**

1. Let's execute the hive file to create a Hive table

$ hive -f prices.hive

....

OK

Time taken: 5.598 seconds

2. [OPTIONAL] Verify the table was created on your own, by guessing where it would be

3. Start Hive shell

$ hive

hive>

4. Load data into the Hive table

hive>

LOAD DATA INPATH 'hive\_data/prices\_F.csv' OVERWRITE INTO TABLE prices\_f;

Loading data to table default.prices\_f

...

OK

Time taken: 0.755 seconds

5. Verify the table structure:

hive>

DESCRIBE prices\_f;

OK

market string

symbol string

dates date

price\_open float

price\_high float

price\_low float

price\_close float

volume int

price\_adj\_close float

Time taken: 0.549 seconds, Fetched: 9 row(s)

6. What's the command to get a total count of rows in this table?

select count(\*) from prices\_f;

You should see similar results:

....

OK

386435

Which means the prices table has 386,435 rows.

Question: For this job, how many Mappers were used? How many Reducers?

1 mapper, and 1 reducer

**Analyze the data contents**

1. Before we start checking on the data in the prices table, we could use below command to find out more details of the table in interest

hive>

DESCRIBE EXTENDED prices\_f;

Which display detailed information on the table "prices\_f", such as CreateTime, table location, whether it is compressed, as well as tableType as Managed\_Table.

2. Now select the first 10 rows of data:

hive>

select \* from prices\_f limit 10;

...

NYSE FCFS 2010-02-08 20.59 20.92 0.3920.6413260020.64

NYSE FCFS 2010-02-05 21.09 21.12 0.3720.8926640020.89

NYSE FCFS 2010-02-04 21.4 21.42 1.021.119780021.11

NYSE FCFS 2010-02-03 21.85 22.02 1.3921.477120021.47

NYSE FCFS 2010-02-02 22.46 22.51 21.7121.9116020021.91

NYSE FCFS 2010-02-01 23.02 23.39 22.4522.5120200022.51

NYSE FCFS 2010-01-29 22.81 22.92 22.5122.8353390022.83

NYSE FCFS 2010-01-28 23.34 23.42 22.6222.7638680022.76

NYSE FCFS 2010-01-27 21.8 23.43 21.5823.2340150023.23

3. Now take a look of how many trades for the most active stocks over the years in the table with ORDER (assuming only one trade on a given day):

hive>

select symbol, count(dates) as count\_date from prices\_f group by symbol order by count\_date desc limit 20;

.....

OK

FWLT 6766

FEIM 6227

FSYS 5655

FMER 5245

FAST 5010

FSBI 5010

FFBC 5010

FARM 5010

FELE 5010

FISV 5004

FITB 5004

FLOW 5003

FULT 4987

FMBI 4968

FFCH 4955

FSTR 4924

FSII 4924

FCNCA 4525

FRME 4525

FLXS 4525

Time taken: 42.142 seconds, Fetched: 20 row(s)

Mark down your elapsed time of execution.

Question: Which stock has the most trades over the years?

4. Now let us take a look what is the biggest volumes of trade

hive>

select max(volume) from prices\_f;

Does it invoke a MapReduce job?

Now how many Map jobs?\_\_\_    How mnay Reduce jobs? \_\_\_

result:

...

OK

198530200

Time taken:19.266 seconds, Fetched: 1 row(s)

Mark down the time it takes.

5. Let's see fora given stock which has big trades, the dates and its highest price, using below conditions:

hive>

select symbol, dates, price\_high from prices\_f where symbol='FACT' and volume>10000 order by price\_high desc;

....

OK

FACT 2009-10-01 18.35

FACT 2009-12-11 18.34

FACT 2009-12-21 18.3

FACT 2009-12-10 18.26

FACT 2009-10-16 18.25

FACT 2009-12-18 18.15

FACT 2009-10-14 18.14

FACT 2009-10-15 18.05

FACT 2009-12-29 18.02

FACT 2008-12-18 18.0

FACT 2009-10-12 18.0

FACT 2010-01-11 17.98

FACT 2009-11-17 17.96

FACT 2009-12-22 17.95

FACT 2009-10-02 17.93

FACT 2009-12-24 17.92

FACT 2009-12-23 17.92

FACT 2009-12-28 17.92

FACT 2010-01-04 17.89

FACT 2009-10-19 17.84

FACT 2009-11-18 17.81

FACT 2010-01-08 17.81

....

Time taken:20.131 seconds, Fetched: 286 row(s)

Question: what is the highest price for the stock FACT over the years? \_\_\_\_

Answer: 18.35

6. Use this query to find out how many different stocks in the table:

hive>

select distinct symbol from prices\_f order by 1;

....

symbol

FXEN

FXCB

FWRD

FWLT

FUQI

FUND

FUNC

FULT

FUEL

....

Time taken:49.592 seconds, Fetched: 138 row(s)

7. Select these columns on DISTINCT:

hive>

select distinct symbol, price\_adj\_close from prices\_f order by 1;

FABK 9.12

FABK 9.11

FABK 9.08

FABK 9.07

FABK 9.06

FABK 9.02

FABK 9.0

FABK 8.99

FABK 8.97

FABK 8.86

FABK 8.85

FABK 8.84

Time taken: 40.555 seconds, Fetched: 132937 row(s)

8. You can also run the command below to get the highest price for each stock during the years

hive>

select symbol, max(price\_high) from prices\_f group by symbol ;

....

OK

FABK 12.97

FACE 23.25

FACT 50.0

FALC 15.5

FARM 393.94

FARO 50.27

FAST 86.71

...

Time taken:18.801 seconds, Fetched: 138 row(s)

**Use EXPLAIN for execution plan review**

1. Practice the EXPLAIN command

hive>

explain select symbol,dates,price\_high from prices\_f where symbol='FACT' and volume>10000 order by price\_highdesc;

STAGE DEPENDENCIES:

Stage-1 is a root stage

Stage-0 is a root stage

STAGE PLANS:

Stage: Stage-1

Map Reduce

Map Operator Tree:

TableScan

alias: prices\_f

filterExpr: ((symbol = 'FACT') and (volume> 10000)) (type: boolean)

Statistics: Num rows: 134580 Data size:22071232 Basic stats: COMPLETE Column stats: NONE

Filter Operator

predicate: ((symbol = 'FACT') and (volume> 10000)) (type: boolean)

Statistics: Num rows: 22430 Data size:3678538 Basic stats: COMPLETE Column stats: NONE

Select Operator

expressions: symbol (type: string),dates (type: date), price\_high (type: float)

outputColumnNames: \_col0, \_col1, \_col2

Statistics: Num rows: 22430 Data size:3678538 Basic stats: COMPLETE Column stats: NONE

Reduce Output Operator

key expressions: \_col2 (type: float)

sort order: -

Statistics: Num rows: 22430 Data size: 3678538 Basic stats: COMPLETE Column stats: NONE

value expressions: \_col0 (type:string), \_col1 (type: date), \_col2 (type: float)

Reduce Operator Tree:

Extract

Statistics: Num rows: 22430 Data size:3678538 Basic stats: COMPLETE Column stats: NONE

File Output Operator

compressed: false

Statistics: Num rows: 22430 Data size:3678538 Basic stats: COMPLETE Column stats: NONE

table:

input format:org.apache.hadoop.mapred.TextInputFormat

output format:org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat

serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

Stage: Stage-0

Fetch Operator

limit: -1

Time taken: 5.119 seconds, Fetched: 39 row(s)

How many stages in the plan?

**Now store the query result to files**

1. Now we store the results to a file, either in HDFS or to the local drive.

hive>

insert overwrite directory '/user/[username]/hive\_data/ stock\_f/' select symbol,dates, volume from prices\_f where price\_open >20;

....

Moving data to: /user/[username]/hive\_data/stock\_f

MapReduce Jobs Launched:

Job 0: Map: 1 Cumulative CPU: 2.96 sec HDFS Read: 22071462 HDFS Write: 3113455 SUCCESS

Total Map Reduce CPU Time Spent: 2 seconds 960 msec

OK

Time taken:24.846 seconds

2. We can verify the file created in the expected location

hive>

dfs -ls hive\_data/stock\_f;

Result:

hive\_data/stock\_f/000000\_0

Question: what command do you use to store a query result to a local file system?

insert overwrite LOCAL directory '/tmp/temp\_dir' select \* from prices\_f order by symbol;

**Create a table using SELECT command**

1. Create a table based on select results

hive>

create table stock\_price\_FN as select symbol, dates, price\_open, price\_close,price\_high, volume from prices\_f where symbol like '%FN%';  
 ...

OK

Time taken: 25.136 seconds

2. You can use DESCRIBE to see the table structure:

hive>

DESCRIBE stock\_price\_fn;

symbol string

dates date

price\_open float

price\_close float

price\_high float

volume int

**END**