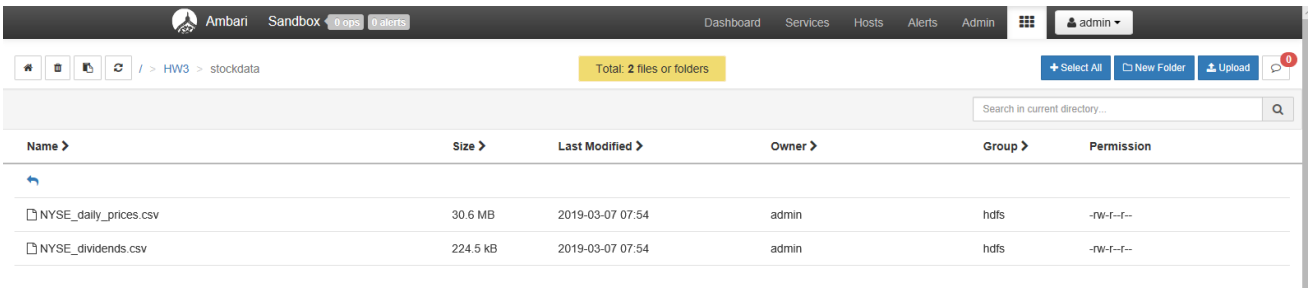
STSCI 5065 HW3

Franklin Zhao (qz297)

03/29/2019

Problem Set 1

A.
```
hadoop fs -mkdir /HW3
hadoop fs -mkdir /HW3/stockdata
```



B.
```
hive
```

```
CREATE DATABASE stocksdb;

CREATE TABLE IF NOT EXISTS stocksdb.stock_prices (
        exchng STRING,
        symbol STRING,
        ymd STRING,
        price_open DOUBLE,
        price_high DOUBLE,
        price_low DOUBLE,
        price_close DOUBLE,
        price_volume DOUBLE,
        price_adj_close DOUBLE)
        ROW FORMAT DELIMITED
        FIELDS TERMINATED BY ','
        LINES TERMINATED BY '\n';

CREATE TABLE IF NOT EXISTS stocksdb.stock_dividends (
        exchng STRING,
        symbol STRING,
        ymd STRING,
        dividend DOUBLE)
        ROW FORMAT DELIMITED
        FIELDS TERMINATED BY ','
        LINES TERMINATED BY '\n';

DESCRIBE stocksdb.stock_prices;

DESCRIBE stocksdb.stock_dividends;
```

```
hive> DESCRIBE stocksdb.stock_prices;
OK
exchng                  string
symbol                  string
ymd                     string
price_open              double
price_high              double
price_low               double
price_close             double
price_volumn            double
price_adj_close         double
Time taken: 0.534 seconds, Fetched: 9 row(s)
hive> DESCRIBE stocksdb.stock_dividends;
OK
exchng                  string
symbol                  string
ymd                     string
dividend                double
Time taken: 0.566 seconds, Fetched: 4 row(s)
hive>
```

C.

```
LOAD DATA INPATH '/HW3/stockdata/NYSE_daily_prices.csv'
OVERWRITE INTO TABLE stocksdb.stock_prices;

LOAD DATA INPATH '/HW3/stockdata/NYSE_dividends.csv'
OVERWRITE INTO TABLE stocksdb.stock_dividends;
```

D.

```
SELECT symbol, COUNT(*)
FROM stocksdb.stock_prices
GROUP BY symbol;
```

```
BA      12109
BAC     5977
BAF     1830
BAK     2785
BAM     6495
BAP     3539
BAS     1047
BAX     7115
BBD     1891
BBF     2142
```

```
BXG     5562
BXP     3175
BXS     6132
BYD     4103
BYI     6354
BYM     1830
BZ      658
BZA     1953
BZH     4018
BZMD    18
Time taken: 18.296 seconds, Fetched: 168 row(s)
```

E.

```
SELECT symbol, ROUND(AVG(price_open), 4) ap
FROM stocksdb.stock_prices
GROUP BY symbol;
```

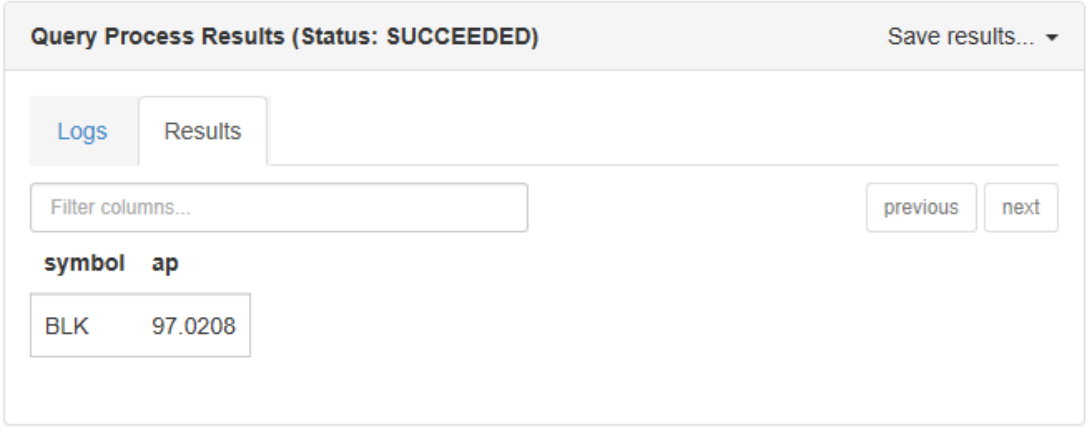
```
BXG     5.3577
BXP     55.4753
BXS     23.523
BYD     17.449
BYI     11.3204
BYM     13.8966
BZ      4.6387
BZA     14.7472
BZH     37.9346
BZMD    24.5267
Time taken: 7.004 seconds, Fetched: 168 row(s)
```

F.

```
CREATE VIEW average_price_v AS
SELECT symbol, ROUND(AVG(price_open), 4) ap
FROM stocksdb.stock_prices
GROUP BY symbol;
```

G.

```
SELECT symbol, ap
FROM average_price_v
ORDER BY ap DESC
LIMIT 1;
```




```
SELECT symbol, ap
FROM average_price_v
ORDER BY ap
LIMIT 1;
```

Query Process Results (Status: SUCCEEDED)

Save results...

Logs | Results

Filter columns...

previous next

symbol	ap
BZ	4.6387

H.

```
SELECT d.symbol symbol, d.ymd ymd, p.price_open price_open, d.dividend dividend
FROM stocksdb.stock_dividends d
JOIN stocksdb.stock_prices p
ON d.exchng = p.exchng
AND d.symbol = p.symbol
AND d.ymd = p.ymd
JOIN (SELECT MAX(dividend) d FROM stocksdb.stock_dividends) m
ON d.dividend = m.d;
```

Query Process Results (Status: SUCCEEDED)

Save results...

Logs | Results

Filter columns...

previous next

symbol	ymd	price_open	dividend
BCE	2000-05-09	26.62	87.057999

Problem Set 2

A.

```
CREATE DATABASE flightsdb;

CREATE TABLE IF NOT EXISTS flightsdb.flight_delays_hw3 (
        ymd STRING,
        flight_num STRING,
        carrier_delay DOUBLE,
        weather_delay DOUBLE,
        nas_delay DOUBLE,
        security_delay DOUBLE,
        late_aircraft_delay DOUBLE);
```

B.

```
CREATE TABLE IF NOT EXISTS temp_flight (tmp_flight STRING)
TBLPROPERTIES("skip.header.line.count" = "1");

LOAD DATA INPATH '/HW3/flight12.csv'
OVERWRITE INTO TABLE temp_flight;
```

C.

```
INSERT OVERWRITE TABLE flightsdb.flight_delays_hw3
SELECT REGEXP_EXTRACT(tmp_flight, '.*?,(.*?),', 1) AS ymd,
REGEXP_EXTRACT(tmp_flight, '(.*?,){4}(.*?),', 2) AS flight_num,
REGEXP_EXTRACT(tmp_flight, '(.*?,){18}(.*?),', 2) AS carrier_delay,
REGEXP_EXTRACT(tmp_flight, '(.*?,){19}(.*?),', 2) AS weather_delay,
REGEXP_EXTRACT(tmp_flight, '(.*?,){20}(.*?),', 2) AS nas_delay,
REGEXP_EXTRACT(tmp_flight, '(.*?,){21}(.*?),', 2) AS security_delay,
REGEXP_EXTRACT(tmp_flight, '(.*?,){22}(.*?),', 2) AS late_aircraft_delay
FROM temp_flight;
```

Comment: I just don't understand why should we use regexp_extract() in this particular question. Apparently split() using regex will be a better option. Regex groups cannot be implemented in a dynamic way (i.e., you cannot do something like (foo)* to match those groups – "groups" rather than "matches" since the 3rd parameter of regexp_extract() refers to group index), so I have to implement it in a "hard coding" way (match exactly all those fields in brutal force so that they can be matched in groups), which took longer time than I thought (34 seconds on a high-performance computer in my lab) in total.

D.

SELECT * FROM flightsdb.flight_delays_hw3 LIMIT 10;

```
hive> SELECT * FROM flightsdb.flight_delays_hw3 LIMIT 10;
OK
2013-12-01      "2900"  NULL    NULL    NULL    NULL    NULL
2013-12-02      "2900"  0.0     0.0     0.0     0.0     36.0
2013-12-03      "2900"  NULL    NULL    NULL    NULL    NULL
2013-12-04      "2900"  NULL    NULL    NULL    NULL    NULL
2013-12-05      "2900"  NULL    NULL    NULL    NULL    NULL
2013-12-06      "2900"  10.0    0.0     0.0     0.0     11.0
2013-12-07      "2900"  NULL    NULL    NULL    NULL    NULL
2013-12-08      "2900"  NULL    NULL    NULL    NULL    NULL
2013-12-09      "2900"  NULL    NULL    NULL    NULL    NULL
2013-12-10      "2900"  0.0     0.0     4.0     0.0     83.0
Time taken: 0.173 seconds, Fetched: 10 row(s)
```

E.

SELECT MAX(carrier_delay) max_carrier_delay, MAX(weather_delay) max_weather_delay, MAX(nas_delay) max_nas_delay, MAX(security_delay) max_security_delay, MAX(late_aircraft_delay) max_late_aircraft_delay FROM flightsdb.flight_delays_hw3;

Query Process Results (Status: SUCCEEDED) Save results...

Logs | Results

Filter columns... previous next

max_carrier_delay	max_weather_delay	max_nas_delay	max_security_delay	max_late_aircraft_delay
1975.0	1451.0	1174.0	175.0	892.0