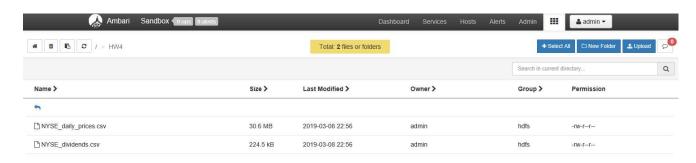# STSCI 5065 HW4

Franklin Zhao (qz297)

05/03/2019

Problem Set 1

A. `hadoop fs -mkdir /HW4`

B.

stocks_prices = LOAD '/HW4/NYSE_daily_prices.csv'
            USING PigStorage(',')
            AS (exchng: chararray, symbol: chararray, ymd: chararray,
                price_open: double, price_high: double, price_low: double,
                price_close: double, volume: int, price_adj_close: double);

DUMP stocks_prices;

Job ID	job_1551915748870_0033
Started	2019-03-08 23:23

⌄ Results

```
(NYSE,BGY,2010-02-08,10.25,10.39,9.94,10.28,600900,10.28)
(NYSE,BGY,2010-02-05,10.53,10.53,9.86,10.17,965800,10.17)
(NYSE,BGY,2010-02-04,10.64,10.69,10.58,10.6,497300,10.6)
(NYSE,BGY,2010-02-03,10.61,10.77,10.61,10.75,521200,10.75)
(NYSE,BGY,2010-02-02,10.58,10.71,10.52,10.68,661400,10.68)
(NYSE,BGY,2010-02-01,10.67,10.73,10.5,10.57,446500,10.57)
(NYSE,BGY,2010-01-29,10.66,10.74,10.53,10.63,477100,10.63)
(NYSE,BGY,2010-01-28,10.62,10.69,10.43,10.64,501000,10.64)
(NYSE,BGY,2010-01-27,10.75,10.79,10.41,10.61,713900,10.61)
(NYSE,BGY,2010-01-26,10.74,10.95,10.73,10.8,650300,10.8)
```

```
stock_dividends = LOAD '/HW4/NYSE_dividends.csv'
                  USING PigStorage(',')
                  AS (exchng: chararray, symbol: chararray, ymd: chararray, dividend: double);

DUMP stock_dividends;
```

Job ID	job_1551915748870_0035
Started	2019-03-08 23:32

✔ Results

```
(NYSE,BCE,2009-12-11,0.386)
(NYSE,BCE,2009-09-11,0.373)
(NYSE,BCE,2009-06-11,0.35)
(NYSE,BCE,2009-03-12,0.309)
(NYSE,BCE,2008-12-19,0.296)
(NYSE,BCE,2008-03-12,0.363)
(NYSE,BCE,2007-12-12,0.363)
(NYSE,BCE,2007-09-12,0.346)
(NYSE,BCE,2007-06-13,0.344)
(NYSE,BCE,2007-03-13,0.311)
```

C.

```
grpd = GROUP stocks_prices BY symbol;
cnt = FOREACH grpd GENERATE group, COUNT(stocks_prices);
sorted = ORDER cnt BY group;
DUMP sorted;
```

Job ID	job_1551915748870_0045
Started	2019-03-08 23:54

✔ Results

```
(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)
```

D.

grpd = GROUP stocks_prices BY symbol;
avgopen = FOREACH grpd GENERATE group, ROUND_TO(AVG(stocks_prices.price_open), 4) AS ap;
DUMP avgopen;

```
(BXS,23.523)
(BYD,17.449)
(BYI,11.3204)
(BYM,13.8966)
(BZA,14.7472)
(BZH,37.9346)
(BBVA,23.2422)
(BRFS,37.8637)
(BSBR,13.1079)
(BZMD,24.5267)
```

E.

grpd = GROUP avgopen ALL;
highest = FOREACH grpd GENERATE MAX(avgopen.ap) AS maxopen;
joined = JOIN avgopen BY ap, highest BY maxopen;
result = FOREACH joined GENERATE $0, $1;
DUMP result;

✔ Results

```
(BLK,97.0208)
```

```
grpd = GROUP avgopen ALL;
lowest = FOREACH grpd GENERATE MIN(avgopen.ap) AS minopen;
joined = JOIN avgopen BY ap, lowest BY minopen;
result = FOREACH joined GENERATE $0, $1;
DUMP result;
```

Job ID	job_1551915748870_0072
Started	2019-03-09 01:00

∨ Results

(BZ,4.6387)

F.

```
stock_div = LOAD '/HW4/NYSE_dividends.csv'
            USING PigStorage(',')
            AS (exchng: chararray, symbol: chararray, ymd: chararray, dividend: double);
```

```
grpd = GROUP stock_div ALL;
highest = FOREACH grpd GENERATE MAX(stock_div.dividend) AS maxdiv;
joindiv = JOIN stock_div BY dividend, highest BY maxdiv;
joinprice = JOIN joindiv BY (exchng, symbol, ymd), stocks_prices BY (exchng, symbol, ymd);
result = FOREACH joinprice GENERATE stocks_prices::symbol, stocks_prices::ymd,
                                    stocks_prices::price_open;
DUMP result;
```

Job ID	job_1551915748870_0084
Started	2019-03-09 03:07

∨ Results

(BCE,2000-05-09,26.62)

Problem Set 2

A.

```
alldata = LOAD '/HW4/flight12.csv'
        USING org.apache.pig.piggybank.storage.CSVExcelStorage(',')
        AS (YEAR, FL_DATE, UNIQUE_CARRIER, CARRIER, FL_NUM, ORIGIN_AIRPORT_ID, ORIGIN,
            ORIGIN_CITY_NAME, ORIGIN_STATE_ABR, DEST_AIRPORT_ID, DEST,
            DEST_CITY_NAME, DEST_STATE_ABR, DEP_DELPAY_NEW: FLOAT, ARR_DELAY: FLOAT,
            ARR_DELAY_NEW: FLOAT, CARRIER_DELAY: FLOAT, WEATHER_DELAY: FLOAT,
            NAS_DELAY: FLOAT, SECURITY_DELAY: FLOAT, LATE_AIRCRAFT_DELAY: FLOAT);

DUMP alldata;
```

Job ID   job_1551915748870_0087

Started   2019-03-09 03:43

❯ Results

```
(2013,2013-12-01,9E,9E,2900,12478,JFK,New York, NY,NY,10693,BNA,Nashville, TN,TN,0.0,-1.0,0.0,,,,,)
(2013,2013-12-02,9E,9E,2900,12478,JFK,New York, NY,NY,10693,BNA,Nashville, TN,TN,78.0,36.0,36.0,0.0,0.0,0.0,0.0,36.0)
(2013,2013-12-03,9E,9E,2900,12478,JFK,New York, NY,NY,10693,BNA,Nashville, TN,TN,0.0,-16.0,0.0,,,,,)
(2013,2013-12-04,9E,9E,2900,12478,JFK,New York, NY,NY,10693,BNA,Nashville, TN,TN,0.0,-36.0,0.0,,,,,)
(2013,2013-12-05,9E,9E,2900,12478,JFK,New York, NY,NY,10693,BNA,Nashville, TN,TN,0.0,-8.0,0.0,,,,,)
(2013,2013-12-06,9E,9E,2900,12478,JFK,New York, NY,NY,10693,BNA,Nashville, TN,TN,29.0,21.0,21.0,10.0,0.0,0.0,0.0,11.0)
(2013,2013-12-07,9E,9E,2900,12478,JFK,New York, NY,NY,10693,BNA,Nashville, TN,TN,0.0,-12.0,0.0,,,,,)
(2013,2013-12-08,9E,9E,2900,12478,JFK,New York, NY,NY,10693,BNA,Nashville, TN,TN,1.0,-18.0,0.0,,,,,)
(2013,2013-12-09,9E,9E,2900,12478,JFK,New York, NY,NY,10693,BNA,Nashville, TN,TN,13.0,14.0,14.0,,,,,)
(2013,2013-12-10,9E,9E,2900,12478,JFK,New York, NY,NY,10693,BNA,Nashville, TN,TN,83.0,87.0,87.0,0.0,0.0,4.0,0.0,83.0)
```

B.

```
wanted_data = FOREACH alldata GENERATE FL_DATE, FL_NUM, CARRIER_DELAY, WEATHER_DELAY,
                                        NAS_DELAY, SECURITY_DELAY, LATE_AIRCRAFT_DELAY;
DUMP wanted_data;
```

Job ID   job_1551915748870_0090

Started   2019-03-09 03:51

❯ Results

```
(2013-12-01,2900,,,,,)
(2013-12-02,2900,0.0,0.0,0.0,0.0,36.0)
(2013-12-03,2900,,,,,)
(2013-12-04,2900,,,,,)
(2013-12-05,2900,,,,,)
(2013-12-06,2900,10.0,0.0,0.0,0.0,11.0)
(2013-12-07,2900,,,,,)
(2013-12-08,2900,,,,,)
(2013-12-09,2900,,,,,)
(2013-12-10,2900,0.0,0.0,4.0,0.0,83.0)
```

C.

```
grpd = GROUP wanted_data ALL;
longest = FOREACH grpd GENERATE MAX(wanted_data.CARRIER_DELAY),
                                 MAX(wanted_data.WEATHER_DELAY),
                                 MAX(wanted_data.NAS_DELAY),
                                 MAX(wanted_data.SECURITY_DELAY),
                                 MAX(wanted_data.LATE_AIRCRAFT_DELAY);
DUMP longest;
```

Job ID   job_1551915748870_0092

Started   2019-03-09 04:03

❯ Results

(1975.0,1451.0,1174.0,175.0,892.0)

D.

```
grpd = GROUP wanted_data ALL;
avg_result = FOREACH grpd GENERATE ROUND_TO(AVG(wanted_data.CARRIER_DELAY), 2),
                                   ROUND_TO(AVG(wanted_data.WEATHER_DELAY), 2),
                                   ROUND_TO(AVG(wanted_data.NAS_DELAY), 2),
                                   ROUND_TO(AVG(wanted_data.SECURITY_DELAY), 2),
                                   ROUND_TO(AVG(wanted_data.LATE_AIRCRAFT_DELAY), 2);
DUMP avg_result;
```

Job ID   job_1551915748870_0094

Started   2019-03-09 04:10

❯ Results

(16.49,2.68,11.69,0.06,23.8)

E.

```
REGISTER '/HW4/delay_udf.py' USING jython AS myudf;
out = FOREACH avg_result GENERATE myudf.delay_udf(*);
DUMP out;
```

```
@outputSchema("res:chararray")
def delay_udf(carrier, weather, nas, security, late_aircraft):
        dic = {carrier: 'CARRIER_', weather: 'WEATHER_', nas: 'NAS_',
                security: 'SECURITY_', late_aircraft: 'LATE_AIRCRAFT_'}
        values = [carrier, weather, nas, security, late_aircraft]
        max_value, min_value = max(values), min(values)
        res = 'The most common delay category is ' + dic[max_value] + \
                'DELAY, which has an average delay of ' + str(max_value) + \
                ' minutes, and the most uncommon delay category is ' + \
                dic[min_value] + 'DELAY, which has an average delay of ' + \
                str(min_value) + ' minutes.'
        return res
```

Job ID	job_1551915748870_0142
Started	2019-03-09 07:44

⌄ Results   ⬇ Download

```
(The most common delay category is LATE_AIRCRAFT_DELAY, which has an average delay of 23.8 minut
es, and the most uncommon delay category is SECURITY_DELAY, which has an average delay of 0.06 m
inutes.)
```