CprE 419 Lab 7: Pig User Defined Functions for analyzing stock data

Department of Electrical and Computer Engineering
Iowa State University
Spring 2017

Purpose

In this lab you will use Pig to analyze stock prices from 440 companies in the S&P 500. The data consists of a stock ticker symbol, a date, the open price, the high, the low, the closing price, and the volume traded that day. You will write some pig scripts to analyze the dataset and the analysis will require the use of creating your own function, a UDF (User Defined Function). More details about UDF is given in the last section of this lab manual.

Submission

Create a zip archive with the following and hand it in through blackboard**. Failure to include all elements will result in a loss of points.**

- Snapshots of results from your terminal and summarize when necessary.
- Commented Code for your program. Include all source files needed for compilation.

Dataset

We have collected a dataset of stock prices from the S&P 500. The dataset covers every day from Jan 1st 1980 – Jan 31st 2014. The data is in the following format:
Ticker, Date, open, high, low, close, volume

The tickers can be obtained from:
https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

For the purposes of price analysis we are only concerned with the **opening price**. You will not need high, low, or close prices.
The dataset is located at: **"/class/s17419/lab7/historicaldata.csv"**.

Pig Job Submission

Due to the large number of submissions, when submitting your pig script, please use the PBS like submitting Hadoop jobs, example job file like following:

```
#!/bin/sh
#PBS -lnodes=1:ppn=1:studenthadoop,walltime=0:20:00
hdfs dfs -rm -r <files on HDFS you want to delete>
pig <your pig script>
```

Here is the command to run your job file, **run on cystorm not n0**:

```
qsub -q studenthadoop -j oe -o output_file.txt -N pigjob_YourID -v job=pigjob_YourID job.file
```

Experiment 1 (30 pts)

Find the companies with highest stock price growth for the following time periods.

Time Periods:
Jan 1st, 1990 – Jan 3rd, 2000 (both days inclusive)
Jan 2nd, 2005 – Jan 31st, 2014 (both days inclusive)

Growth means the largest growth factor, which can be calculated by taking the starting stock price, divided by the ending price (ie, $1 at start to $2 at end indicates a growth factor of 2). If a company doesn't have a price for the start date (for example Google, didn't exist in 1982) then pick the next closest day **after** that date. If a company isn't listed on the end date pick the next closest day **before** that date. If a company doesn't exist for more than 1 day in a time period then ignore it.

Include your results in your write up.

Experiment 2 (40 pts)

Calculate the 20-day moving average **for every company** and **for each day** between Oct 1st, 2013 and Oct 31st, 2013 (both days inclusive). Moving average for a particular company and for a particular date is defined as the average of the last available 20 trading days before that date for which stock prices are available for that company. For each company, you will need to find what day you need to start on to get a result for Oct 1st (20 trading days before). Output your answer in the following format:
Ticker, date, open price, moving average price

Include your results for General Electric, IBM, Intel, Microsoft, Google and Apple's moving average in your write up. You don't need to submit the other companies in your write up (answer would be very large), but your code should still generate them.

User Defined Functions (UDFs)

Pig allows the creation of custom java methods as a way to do custom processing. Some of these are included with Pig such as, UPPER, and COUNT. UDFs can be used to extend Pig functionality where the required logic cannot be easily expressed as a series of built in Pig statements. Thus, UDFs are useful when more complicated logic is needed in a Pig script and enables the user to write that custom logic.

All UDFs must extend the **EvalFunc** class. This class has a single required method called exec. To write custom code, you must override the exec method. When the UDF is called, this exec method is invoked and the method receives a tuple. Thus, using your custom code, the tuple can be processed and then you can output a data type of your choice.

There is a good documentation of UDF on the apache pig website:
http://pig.apache.org/docs/r0.14.0/udf.html

In order to use a UDF you must compile the java file and packaging in a JAR file. In addition to the **hadoop jars**, another essential external library jar file is "**pig-0.14.0-SNAPSHOT-core-h2.jar**". These jar files can be downloaded from Piazza-Software to Download. You should add them as external jar files in your project.

After you compile the program you need to include a reference to it in your Pig script. In Pig this is really easy; you just need to use the **REGISTER** command at the top of your pig script:

```
REGISTER ‘<Your_UDF.jar>’;
```

Now you should be able to use your User Defined Function in your Pig script.