HW3 Spark

Deadline: Nov. 7th 5:59 P.M. (before class)

Part I. PairRDD

Description

PairRDD is the main component in Spark for performing map-reduce algorithm. In this assignment, you will need to translate the part I of last homework in Spark way. Specifically, you need to

1. Read the pg100.txt input file.
2. Write transformations to count the 2nd letter of each word
   - Requirements remain same with HW1
3. Write transformations to sort values in descending order
   - Use .sortByKey() after your mapper to mimic the default sorting process in Hadoop
4. Print out the count on your screen with the same format in HW1, which is
   letter1<\tab>count1
   letter2<\tab>count2
   …

Submission

Submit source code only (.py, .scala, or .java file), no output file required.

Note: your program can be in Python/Scala/Java, but the language must be consistent for all the Spark homework.

Part II. SparkSQL and DataFrames

Description

The tick_data.csv file is tick data of stocks. The file includes 5 columns:

- DATE: (int) date when transaction made
- TIME_M: (str) time in minute when transaction made
- SYM_ROOT: (str) stock name of transaction
- SIZE: (int) transaction volume
- TRADE: (double) price of transaction

In this part, you need to

1. Using Spark, report
   a. How many days this file contains
   b. How many stocks this file contains

2. Assuming TRADE reflects the stock price at the time, calculate
   a. total trading volume in a certain hour
   b. hourly return of each stock with the following formula,

$$r_t = \frac{p_t^n - p_t^1}{p_t^1}$$

   where in a certain hour t, p_t^n represents the price of last trade and p_t^1 represents the price of first trade.

   The output should include

   DATE,TIME_H,SYM_ROOT,SIZE_H,RETURN

   where TIME_H represents hour, SIZE_H represents total trading volume which calculated from 2a, and RETURN represents the hourly return which calculated from 2b.

Submission

1. Attach your spark code (1a & 1b) along with the answer for your report in a PDF file.
2. Submit the spark code (2a & 2b) with the output file.