Assignment 5 – Streams and Events

Tools & Techniques for Large-Scale Data Analytics (CT5105)
2019/2020 Semester 1

Submission deadline (strict): Friday, 29th November, 23:59

Instructions

- Please put your code into a single .zip archive with name "YourName_Assignment5.zip", submit via Blackboard
- Include a screenshot of the output of your application for each part of both questions.
- For Part A of this assignment use only Java 8 together with Apache Spark. Include all source code files (that is, files with name ending .java) required to compile and run your code.
- For Part B use the Esper Online tool http://esper-epl-tryout.appspot.com/epltryout/mainform.html Include the EPL statements in a text file together with a screenshot of the output of the Online tool.
- Please note that all submissions will be checked for plagiarism.
- Use comments to explain your source code. Insufficient comments can lead to mark deductions.

Part A – Streaming (60 Marks)

1. Write a standalone Spark Streaming program that connects to Twitter and prints a sample of the tweets it receives from Twitter every second.
    - **Hint:** You can use an existing Twitter account or create a new one. You will need to setup Twitter to allow your program to access your account. See: https://apps.twitter.com/
    - TwitterUtils has been moved to the Apache project 'Bahir' http://bahir.apache.org/ The TwitteUtils Jars should then be added separately from Bahir
    - Follow the tutorial on http://ampcamp.berkeley.edu/big-data-mini-course/realtime-processing-with-spark-streaming.html to help you connect to the twitter feed. It is also worthwhile to notice the difference between the code for Java and Scala in the tutorial.

    (10 marks)
2. Extend the code to count the number of characters, words and extract the hashtags in each tweet.

    (10 marks)

3. Extend the code to calculate:
    a. the average number of characters and words per tweet (10 marks)
    b. count the top 10 hashtags (15 marks)
    c. for the last 5 minutes of tweets, continuously repeat the computation every 30 seconds. (15 marks)

PTO for Part B

Part B – Complex Event Processing (40 Marks)

The Yellow Cab Company in NYC is using Event Processing to help it understand the movements and revenue generated by its fleet of cabs. The event processing engine is using two types of events:

Pickup(int taxi_id, int location_id) and

Dropoff(int taxi_id, int location_id, int amount).

You job is to write a query to find the ten least profitable routes in the last 40 minutes. The profitability of a route is the sum of the amounts of all the taxi trips for that route. A route is a pair of (pickup location, dropoff location). Consider routes that ended within the last 40 minutes.

Use for the following sequence of events and timings.

Pickup={taxiId=10, pickupLocation=1}
t=t.plus(5 minutes)

Dropoff={taxiId=10, dropoffLocation=2, amount=110}
t=t.plus(5 minutes)

Pickup={taxiId=10, pickupLocation=2}
t=t.plus(5 minutes)

Dropoff={taxiId=10, dropoffLocation=3, amount=90}
t=t.plus(5 minutes)

Pickup={taxiId=20, pickupLocation=1}
t=t.plus(5 minutes)

Dropoff={taxiId=20, dropoffLocation=3, amount= 80}
t=t.plus(5 minutes)

Pickup={taxiId=20, pickupLocation=3}
t=t.plus(5 minutes)

Dropoff={taxiId=20, dropoffLocation=1, amount=110}
t=t.plus(5 minutes)

Pickup={taxiId=20, pickupLocation=1}
t=t.plus(5 minutes)

Dropoff={taxiId=20, dropoffLocation=3, amount=100}
t=t.plus(5 minutes)

Pickup={taxiId=30, pickupLocation=1}
t=t.plus(5 minutes)

Dropoff={taxiId=30, dropoffLocation=3, amount=110}
t=t.plus(5 minutes)

Pickup={taxiId=40, pickupLocation=6}
t=t.plus(5 minutes)

Dropoff={taxiId=40, dropoffLocation=7, amount=140}
t=t.plus(5 minutes)