EECS E6893 Big Data Analytics - Fall 2019

Homework Assignment 3: Twitter data analysis with Spark Streaming

Due Friday, November 1st, 2019 by 5:00pm

TL;DR

1. Do twitter streaming analysis using Spark Streaming
2. Count hashtags and special words
3. Submit your codes and report to Canvas

Abstract:

In this assignment, you will implement a streaming analysis process. The architecture is as follows. A socket request data from twitter API and send data to spark streaming process. Spark read real time data and do analysis. It also save temp streaming results to Google Storage. After the streaming process terminate, it reads the final data from Google Storage and save it to BigQuery, and then clean the data in Storage.

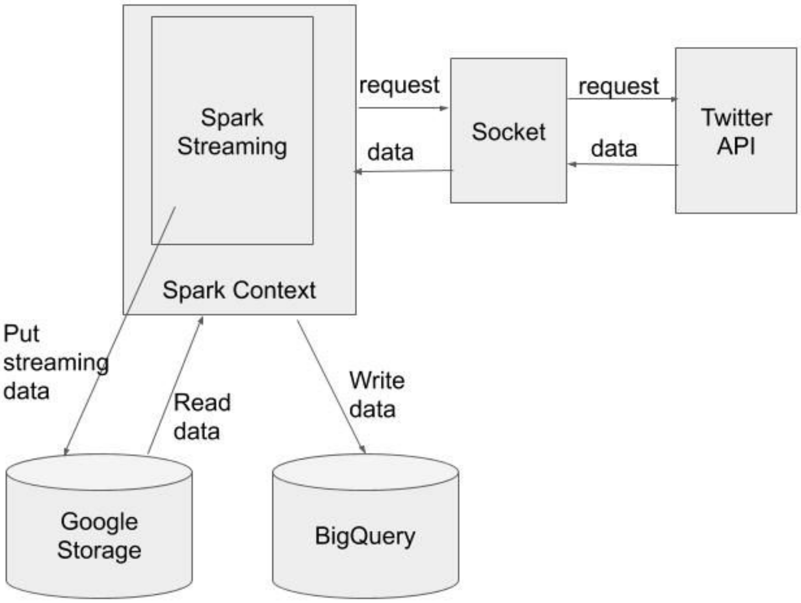


Fig 1. Twitter streaming architecture

You will be asked to do two analysis tasks based on the tweets you get. And you need to figure out how to save temp steaming results to google storage.

Tasks:

1. Calculate the accumulated hashtags count sum for 120 seconds and sort it by descending order of the count. Hashtag usually starts with "#" and followed by a series of alphanumeric. [40%]
2. Filter the chosen 5 words and calculate the appearance frequency of them every 20 seconds (no overlap). The words are: 'data', 'spark', 'ai', 'movie', 'good'. [40%]
3. Save temp steaming results to google storage. [20%]

Some important notes:

1. Remember to start the socket first and then the streaming process
2. If you want to terminate the socket, you have to go to job page of cluster and cancel it. Or you can use `gcloud dataproc jobs kill` **JOB**
3. Create a table first before saving results to it.
4. For tasks (2), it is ok if you don't see all of the words in your results every time.
5. You can stop your cluster instance in Computer Engine page and keep your cluster. If you stop your instance, you won't pay too much for having that cluster.

Steps:

Step1: Register on Twitter Apps

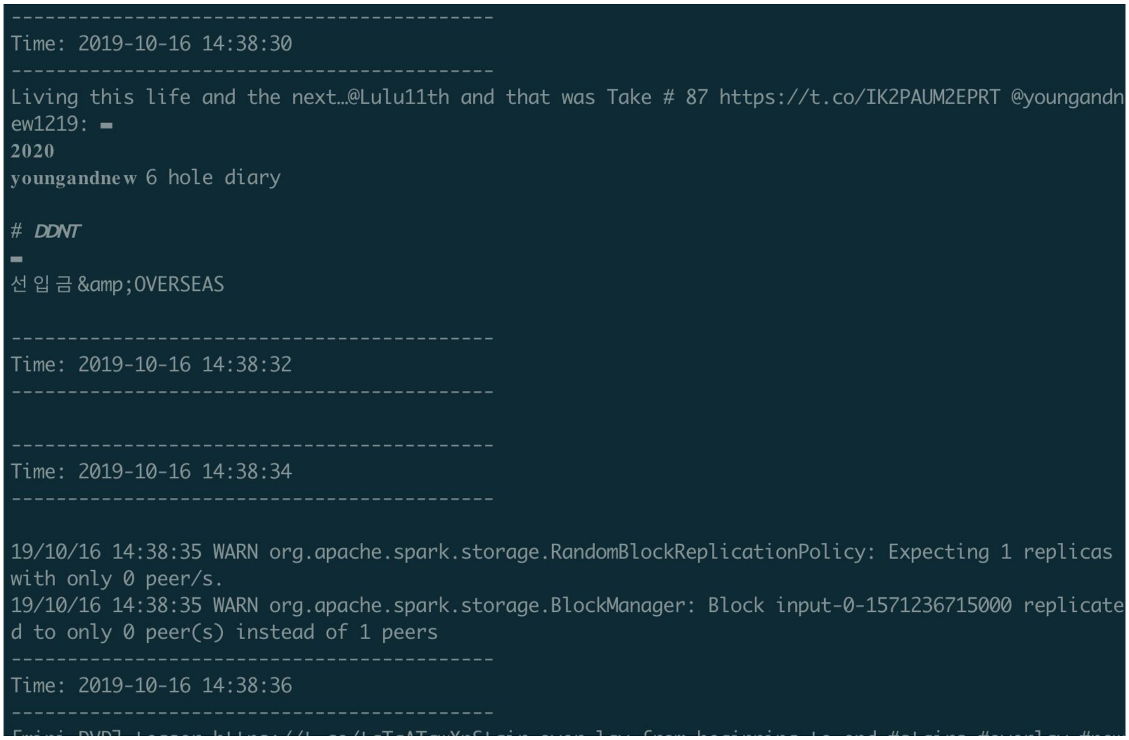
1. Go to https://developer.twitter.com/en/apply-for-access.html and apply for a twitter developer account.
2. Login at https://apps.twitter.com/
3. Click "Create New App", fill out the form, and click "Create your Twitter application"
4. In the next page, click on "API keys" tab, and copy your "API key" and "API secret".
5. Scroll down and click "Create my access token", copy your "Access token" and "Access token secret"

Step2: Create a cluster and run the get streaming data

1. Use the following command to create a cluster.

```
gcloud beta dataproc clusters create <cluster-name> \
--optional-components=ANACONDA,JUPYTER \
--image-version=preview --enable-component-gateway \
--metadata 'PIP_PACKAGES=requests_oauthlib google-cloud-bigquery
tweepy' \
--metadata gcs-connector-version=1.9.16 \
--metadata bigquery-connector-version=0.13.16 \
--project <ProjectID>
--bucket <Bucket> \
--initialization-actions=gs://dataproc-initialization-actions/pyth
on/pip-install.sh,gs://dataproc-initialization-actions/connectors/
connectors.sh \
--single-node
```

2. Download start code from canvas.
3. Open *twitterHTTPClient.py* Replace the Twitter API credential with your own and run on dataproc. Remember to stop the job after you finish streaming.
4. Open *sparkStreaming.py.* You can first comment code from line 140 (words=...) to line 148 (wordCount.pprint()) and from line 166(saveToBigQuery) to line 167saveToBigQuery). And then run it on dataproc. You are expected to see a tweet stream like this. Ignore all the warnings while running.



```
-------------------------------------------
Time: 2019-10-16 14:38:30
-------------------------------------------
Living this life and the next…@Lulu11th and that was Take # 87 https://t.co/IK2PAUM2EPRT @youngandn
ew1219: ▬
2020
youngandnew 6 hole diary

# DDNT
▬
선 입 금 &amp;OVERSEAS

-------------------------------------------
Time: 2019-10-16 14:38:32
-------------------------------------------

-------------------------------------------
Time: 2019-10-16 14:38:34
-------------------------------------------

19/10/16 14:38:35 WARN org.apache.spark.storage.RandomBlockReplicationPolicy: Expecting 1 replicas
with only 0 peer/s.
19/10/16 14:38:35 WARN org.apache.spark.storage.BlockManager: Block input-0-1571236715000 replicate
d to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2019-10-16 14:38:36
-------------------------------------------
```

Step3: Stream data analysis and store data in BigQuery

1. Once you can successfully get the stream, you can start writing your own code. The analytics tasks are listed before.
2. To save the data, remember to:
    a. Replace the bucket global variable in the code with your own bucket.
    b. Create a BigQuery dataset first using

        bq mk <your dataset name>

Homework Submissions

1. A report includes:
    a. Screenshot of your code to do all the tasks. Tasks
    b. Screenshot of the preview of your data stored in BigQuery. You have to include two tables: *hashtags* and *wordcount.*
2. Your code

Useful links:

https://spark.apache.org/docs/latest/streaming-programming-guide.html