CS 6350- ASSIGNMENT 6

Please read the instructions below before starting the assignment.

- This assignment can be done using Cloudera Docker or on another cluster resource such as AWS. You have to mention your environment clearly in the README file. The TA should be able to run your code.

- You should use a cover sheet, which can be downloaded at: http://www.utdallas.edu/~axn112530/cs6350/CS6350_CoverPage.docx

- You are allowed to work in pairs i.e. a group of two students is allowed. Please write the names of the group members on the cover page. Only one submission per team is required.

- You have a total of 4 free late days for the entire semester. You can use at most 2 days for any one assignment. After that, there will be a penalty of 10% for each late day. The submission for this assignment will be closed 2 days after the due date.

- Please ask all questions on Piazza, and not through email to the instructor or TA.

ASSIGNMENT 6

Analytics Using Spark and HBase

In this assignment, you will use Spark to connect to data stored in HBase tables and run analytical queries. Since HBase is not available on UTD cluster, you would have to use either Cloudera's Docker container or another installation (such as AWS). The assignment consists of the following steps:

Step I:

1. Download the bike sharing dataset from:
http://www.utdallas.edu/~axn112530/cs6350/data/bikeShare/201508_trip_data.csv

Hint: On the UNIX shell, you can run the following
curl –o 201508_trip_data.csv
http://www.utdallas.edu/~axn112530/cs6350/data/bikeShare/201508_trip_data.csv

2. Analyze the data and look at the fields. Check if it has a header. Create table and at least one column family in HBase so that this data can be imported. You can do this using the command line or using the Hue GUI.

3. Import the data into the table that you created in step 2. You can do this using any of the Hadoop technologies, such as Pig or Spark. An example of this was shown in the class.

4. Make sure that the data has been imported correctly by looking at it on the Hue GUI.

Step II:

1. In this step, you will use Spark to connect to the HBase table that you created in step I. Below are some hints:

- Download the Spark HBase connector jar file from:
  https://github.com/nerdammer/spark-hbase-connector
  The above page also contains helpful hints and code snippets.
  You can directly download the jar file as:
  curl -o spark-hbase-connector.jar
  http://central.maven.org/maven2/it/nerdammer/bigdata/spark-hbase-connector_2.10/0.9.2/spark-hbase-connector_2.10-0.9.2.jar
  (No space in the above lines)
- When starting Spark shell use the following command:
  spark-shell --jars spark-hbase-connector.jar

- On the first line of the Spark shell, import the library as:
  import it.nerdammer.spark.hbase._

2. Study the examples available on the connector page
https://github.com/nerdammer/spark-hbase-connector
and learn how to connect to the table you created in step I.

3. Connect to the table and answer the following queries:

- List the top 10 most popular start stations i.e. those start stations that have the highest count in the dataset

- List the top 10 most popular end stations i.e. those end stations that have the highest count in the dataset

- List the top 10 start stations that have the highest average trip duration

- Which zip code has the highest number of stations (you can take either start or end stations)

- What is the average duration of the trips that start from any station that contains 'San Francisco' in their name

- Give the breakdown of subscriber type of users and the count of their occurrence.
Something like:

```
Subscriber Type        Count
--------------------   -------------
Subscriber             100
Customer               50
```

- Give summary statistics for the duration column e.g. count, min, max, mean, stddev

Hint: Dataframe has a describe command that you might find useful
df.describe().show()

What To Submit:

- Source code for all the steps
- Output for the queries in step II
- README file indicating which environment you used, and other relevant details