Project

Instruction:

- The project includes two parts:
  - (i) The first part is about individual work, including Problem 1 and 2. You need to complete the tasks all by yourself. This part accounts for 50%.
  - (ii) The second part is about team work, including Problem 3, 4 an 5. This part accounts for 50%. Each team may decide its own way of collaboration. All team members share the same score in this part.
- You should submit two reports:
  - (i) For the individual work part, each student should submit a report to moodle.
  - (ii) For the team work part, only one report is required for each team. This report shall include: the solutions to Problem 3 and 5 and the slides for Problem 4 and 5. Please send it to me by email.
- Important dates:
  - (i) Jan 9: Due date of informing me the choice of topic of Problem 4 for each team.
  - (ii) Jan 16: Due date of submitting a paragraph describing your dataset and proposed task of Problem 5 for each team.
  - (iii) Jan 29: The day of presentation for Problem 4 and 5.
  - (iv) Jan 30: Due date of submitting your reports (for both individual work and team work).

1 Individual work

1. Setup your hadoop in a pseudo-distributed mode (you may use either hadoop 2.x or hadoop 3.x).
   - (a) What is the Java environment used by your hadoop? Show the configuration files that you've modified to setup the pseudo-distributed mode.
   - (b) Create a folder named with your own name and generate a txt file named with your student ID number in this folder. The content of the txt file includes both your name and your ID number. Take screenshots of all the operations you've performed to show that you have done it successfully.
   - (c) By taking screenshots, show the content information of hadoop web UI for the namenode and datanodes in your report. (Be careful that hadoop 2.x and hadoop 3.x have different port numbers for web UI.)
2. In this problem, you are required to use the Monte Carlo Method to estimate the volume of the following objects:

   (1) The first object, denoted by O_A, is a sphere centered at $(0, 1, -2)$ of radius 4. As a set of points, we write $O_A = \{(x, y, z) \mid x^2 + (y-1)^2 + (z+2)^2 \leq 16\}$.

(2) The second object, denoted by O_B, is a cylinder defined by $O_B = \{(x, y, z) \mid x^2 + y^2 \leq 4,\ 0 \leq z \leq 4\}$.

(3) The third object, denoted by O_C, is the intersection of O_A and O_B, i.e., $O_C = O_A \bigcap O_B$.

(4) The fourth object, denoted by O_D, is the union of O_A and O_B, i.e., $O_D = O_A \bigcup O_B$.

For a brief description of the Monte Carlo method, you may check the following link

http://mathworld.wolfram.com/MonteCarloMethod.html

Please write pyspark codes to compute the volume of all the objects O_A, O_B, O_C and O_D using two different approaches:

(1) In the first approach, you should use the DataFrame API of pyspark. For help, a good reference is the pyspark example in "/spark/examples/src/main/python/pi.py" (in your folder of spark) to estimate the value of π using the Monte Carlo method.

(2) In the second approach, you should use Spark Streaming API or Structured Streaming API. The idea is that you need to create an input stream of random points in a fixed region while counting the cumulative number of points inside each of the objects. The requirement is that you should only use one input stream to compute the volumes of the objects all together instead of using one input stream for each object. On the other hand, the way to create such an input stream is up to you. For example, you may use stream sockets or RDD queues. Moreover, you may consider to use the methods "flatMap()" and "updateStateByKey()".

In your report, you should provide both your codes and your demonstration of the results. Take screenshots whenever necessary.

2 Team Work

3. In this problem, you are required to use spark.ml API. As in Problem 2, consider 3 objects:

(1) The first object, denoted by O_A, is a ball centered at $(0, 0, 0)$ of radius 1. As a set of points, we write $O_A = \{(x, y, z) \mid x^2 + y^2 + z^2 \leq 1\}$.

(2) The second object, denoted by O_B, is a cylinder defined by $O_B = \{(x, y, z) \mid x^2 + y^2 \leq 4,\ 2 \leq z \leq 4\}$.

(3) The third object, denoted by O_C, is an ellipsoid

$$O_C = \{(x, y, z) \mid \frac{(x-2)^2}{1.2} + y^2 + \frac{z^2}{4} \leq 1\}$$

Note that O_A overlaps with O_C a little bit.

Create a dataset in the following way:

(1) Each record in the dataset corresponds to a point contained in the union of O_A, O_B and O_C, which has a "features" part which is made of the xyz coordinates

of that point and a "label" part which tells which of O_A, O_B or O_C this point is contained in. Note that since $O_A \cap O_C$ is nonempty, if the point happens to locate in $O_A \cap O_C$, you still can only label it as O_A or O_C, but not both.

(2) The dataset you create should contain at least 500000 records. You should generate the records randomly in the following way:

i. Each time, choose O_A, O_B or O_C randomly. Suppose we choose O_X (X is A, B or C).

ii. Randomly create a point P contained in O_X (think of how to do it). Now the features of the newly created record is the coordinates of P and the corresponding label is "O_X".

iii. After creating all the records, you should load and transform the dataset to a spark Dataframe.

You are required to do the following work.

(1) Do classifications using both logistic regression and decision tree classifier. You should try several different training/test split ratio on your dataset and for each trained model, evaluate your model and show the accuracy of the test.

(2) Use K-means clustering to make cluster analysis on your data. Now only the "feature" part of your data matters. Set the number K of clusters to 2, 3 and 4 respectively and make a comparison. Show the location of the centroids for each case.

(3) Provide a visualization of the results of your classifications and cluster analysis.

In your report, you should provide both your codes and your demonstration of the results. Take screenshots whenever necessary.

4. Instruction:

- There are many techniques related to big data and big data analytics that we are not able to cover in class.
- Here some topics are listed in the table, including NoSQL databases, cloud services, and tools for big data analytics.
- For each team, please let me know your choice of the topic to present by Jan 9. If conflicts happen, the first one of notification takes it. So if possible, you should inform me the topic earlier than that.
- Each team is required to
  (a) make slides of about 8-12 pages for your topic, making a general introduction to the topic, and
  (b) give a 12 min presentation in a presentation session on Jan 29.
- For the slides, you can talk about a brief history, features and merits of the technique. You are not required to talk about technical details.
- To find resources for your slides, you can refer to the official websites provided below and also do some online search by yourself.
- You should submit your slides to me (by email) at least one day before the presentation.

Topic	Official Website	Presenter
Apache Storm	https://storm.apache.org/	
Apache Beam	https://beam.apache.org/	
Apache Cassandra	http://cassandra.apache.org/	
Apache CouchDB	https://couchdb.apache.org/	
Apache Flink	https://flink.apache.org/	
Apache Giraph	http://giraph.apache.org/	
Apache Hadoop YARN	https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html	
Apache Hive	https://hive.apache.org/	
Apache Ignite	https://ignite.apache.org/	
Apache Impala	http://impala.apache.org/	
Apache Kafka	https://kafka.apache.org/	
Apache Mahout	http://mahout.apache.org/	
Apache Mesos	http://mesos.apache.org/	
Apache Samza	http://samza.apache.org/	
Amazon DynamoDB	https://aws.amazon.com/dynamodb	
Amazon RDS	https://aws.amazon.com/rds/	
Couchbase	https://www.couchbase.com/	
Google BigQuery	https://cloud.google.com/bigquery/	
Microsoft Azure	https://azure.microsoft.com/	SWE group 2
MySQL	https://www.mysql.com/	SWE group 7
Neo4J	https://neo4j.com/	
PostgreSQL	https://www.postgresql.org/	
Redis	https://redis.io/	
TIBCO StreamBase	https://www.tibco.com/products/tibco-streaming	
VoltDB	https://www.voltdb.com/	

5. The last problem is of open tasks.

   - Go to the website https://www.kaggle.com/datasets.
   - Choose a public dataset from the above website and set up your own task of data analysis on your chosen dataset using Spark.
   - Send me by email a paragraph describing your dataset and proposed task to me by Jan 16. To avoid conflicts, you may want to send it to me as early as possible.
   - If approved, do your analysis using Spark.
   - Give a presentation based on the results you obtained in our presentation session on Jan 29.
   - You should submit your slides to me (by email) at least one day before the presentation.
   - In your report, you should provide both your codes and your demonstration of the results. Take screenshots whenever necessary.