Assignment 3

Assignment on Spark and Cloud Data Platform

Due by Oct 30, 2022

Note:

Submit a compressed archive (zip, tar, etc.) of your code, along with the output files and CLI screenshots (output/input commands with results). Please include a pdf document with answers to the below questions.

Contact your TA for any questions related to this assignment or post clarification questions to the Piazza platform.

Part A:

Input Data - kddcup.data_10_percent.gz 10% subset. (2.1M; 75M Uncompressed) from http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Please read the paper which is provided with your assignment in the Quercus and answer the following question.

1. [Marks: 15] What is an Intrusion Detection System? Is it possible to implement an Intrusion Detection System on this dataset? Explain the workflow as described in the paper for implementing Intrusion Detection System.

This part needs to be done by using PySpark or Spark-SQL in Databricks.

2. [Marks: 5] Use python urllib library to extract the KDD Cup 99 data from their web repository, store it in a temporary location and then move it to the Databricks filesystem which can enable easy access to this data for analysis. Use the following commands in Databricks to get your data.

   ```
   import urllib.request
   urllib.request.urlretrieve("http://kdd.ics.uci.edu/databases/kddcup99/kddcup.data_10_percent.gz", "/tmp/kddcup_data.gz")
   dbutils.fs.mv("file:/tmp/kddcup_data.gz", "dbfs:/kdd/kddcup_data.gz")
   display(dbutils.fs.ls("dbfs:/kdd"))
   ```

3. [Marks: 5] After storing the data into the Databricks filesystem. Load your data from the disk into Spark's RDD. Print 10 values of your RDD and verify the type of data structure of your data (RDD).

4. [Marks: 5] Split the data. (Each entry in your RDD is a comma-separated line of data, which you first need to split before you can parse and build your dataframe.) Show the total number of features (columns) and print results. See this link for more details. http://kdd.ics.uci.edu/databases/kddcup99/task.html

5. [Marks: 5] Now extract these 6 columns (*duration*, *protocol_type, service, src_bytes, dst_bytes, flag and label*) from your dataset. Build a new RDD and dataframe. Print schema and display 10 values.

6. [Marks: 5] Get the total number of connections based on the *protocol_type* and based on the *service*. Show result in an ascending order. Plot the bar graph for both.

7. [Marks: 15] Do a further exploratory data analysis, including other columns of this dataset and plot graphs. Plot at least 3 different charts and explain them.

8. [Marks: 20] Look at the label column where label == 'normal'. Now create a new label column where you have a label == 'normal' and everything else is considered as an 'attack'. Split your data (train/test) and based on your new label column now build a simple machine learning model for intrusion detection (you can use few selected columns for your model out of all). Explain which algorithm you have selected and why? Show the results with some success metrics.

Part B:

1. [Marks: 5] Read the below statements, choose the correct answer, and provide explanations. You can get more information by visiting this link. https://azure.microsoft.com/en-us/overview/what-is-paas/

Statements	**Yes**	**No**
1. A platform as a service (PaaS) solution that hosts web apps in Azure provides professional development services to continuously add features to custom applications.		
2. A platform as a service (PaaS) database offering in Azure provides built in high availability.		

2. [Marks: 5] Read the below statement, choose the correct answer, and provide explanations.

   A relational database must be used when:
   a. A dynamic schema is required
   b. Data will be stored as key/value pairs
   c. Storing large images and videos
   d. Strong consistency guarantees are required

3. [Marks: 5] Read the below statement, choose the correct answer, and provide explanations.

   When you are implementing a Software as a Service solution, you are responsible for:
   a. Configuring high availability
   b. Defining scalability rules
   c. Installing the SaaS solution
   d. Configuring the SaaS solution

4. [Marks: 5] Read the below statements, choose the correct answer, and provide explanations.

Statements	Yes	No
1. To achieve a hybrid cloud model, a company must always migrate from a private cloud model		
2. A company can extend the capacity of its internal network by using public cloud		
3. In a public cloud model, only guest users at your company can access the resources in the cloud		

5. [Marks: 5] Read the below statements, choose the correct answer, and provide explanations.

   a. A cloud service that remains available after a failure occurs _____
   b. A cloud service that can be recovered after a failure occurs _____
   c. A cloud service that performs quickly when demand increases _____
   d. A cloud service that can be accessed quickly from the internet _____

   Disaster recovery, Fault Tolerance, Low Latency, Dynamic Scalability