MIE1628
Assignment 3: *Spark and Cloud Data Platform*

Due Date: Jul 12, 2025

- *No submission is accepted via email. We have no exception.*
- *Students are responsible for submitting the correct files on time.*
- *Assignments submitted up to 48 hours late will incur a 20% penalty.*
- *Your grade will be zero if you submit your answer after 48 hours.*

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

Submission Requirements:

- This assignment must be completed using PySpark or Spark SQL.
- Submit your code scripts (Python files or Jupyter notebooks).
  - A PDF document that details your methodologies, results, and theoretical insights and result screenshots as needed.

Part A: [Marks 60]

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 provided with your assignment in the Quercus and answer the following question.

1. [Marks: 10] What is an Intrusion Detection System? Is it possible to implement an Intrusion Detection System on this dataset? Explain the workflow described in the paper for implementing the Intrusion Detection System.

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

2. [Marks: 4] Use the 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. {Hint: You can 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: 4] After storing the data in 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: 4] 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 data frame.) 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: 4] Now extract these 6 columns (*duration*, *protocol_type, service, src_bytes, dst_bytes, flag and label*) from your dataset. Build a new RDD and data frame. Print schema and display 10 values.

6. [Marks: 4] Get the total number of connections based on the *protocol_type* and based on the *service*. Show results in an ascending order. Plot the bar graph for both.

7. [Marks: 10] 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: [Marks: 20]

1. [Marks: 4] 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 provide 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: 4] 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: 4] 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: 4] 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 a public cloud		
3. In a public cloud model, only guest users at your company can access the resources in the cloud		

5. [Marks: 4] 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

Part C: [Marks: 20]

1. [Marks: 10] **Hybrid Cloud Strategy:** A large financial institution is considering migrating its legacy systems to the cloud. They have stringent regulatory requirements for data security and latency. They also need to maintain control over sensitive data and ensure compliance.

   a. **Design a hybrid cloud strategy** outlining which components should reside on-premises versus in the cloud, considering security, regulatory compliance, latency, and cost-effectiveness. Justify your decisions.
   b. **Discuss the challenges** in implementing and managing this hybrid cloud environment. What specific technical and operational considerations must be addressed?
   c. **Propose a plan for monitoring and managing** the hybrid cloud infrastructure to ensure security, performance, and compliance.

2. [Marks: 10] **Database Selection:** A rapidly growing e-commerce company needs a database solution to handle increasing volumes of transactional data and customer information. They require high availability, scalability, and strong consistency guarantees. The data model involves relational and non-relational elements.

   a. **Recommend a database solution** (or a combination of solutions) suitable for this scenario, explaining your choice based on the specific needs. Consider both relational and NoSQL options.
   b. **Outline a data migration strategy** for moving their existing data to the chosen solution, including considerations for downtime, data integrity, and testing.
   c. **Discuss potential scalability challenges** and how your chosen solution(s) can address them.

Additional Notes:

- Ensure clear documentation of your code and methodologies throughout the assignment.
- Include visualizations (e.g., graphs, charts) where appropriate to reinforce your findings and enhance result clarity.
- The PDF report should detail your methodologies, findings, and insights, citing any references for external sources utilized during your research and code development.