DATA CENTER SCALE COMPUTING - EXAM 2

Hi! I believe you had a great time performing labs and now it's time to connect the dots! You will be solving each task to make a complete Data Engineering pipeline.

Instructions

- You can collaborate with your other peers from the course, but have to submit the solution individually
- There will be a grading Interview for this exam
- I will not be assisting you with coding errors etc. That said, you can contact me if you are facing issues with the platform.
- I have provided hints so that you can execute this smoothly without having to look for code online.

Platform

You will be performing this exam on Azure DataBricks. In case you find issues with Azure Databricks you can use GCP but please intimate before you switch with a valid reason. Find the tutorial to set up a Azure Databricks account and its usage to get your exam started at https://drive.google.com/drive/folders/1bWtt9sty6bPcQw5UixA6Lr4YWGftVBW5?usp=share_link

Task 1 - Creating a Data Source - Databricks [Component 1]

1. Upload the StudentData.csv and create a table to train your model (as explained in the video)
2. Check and verify the datatypes of the column
3. Code to read the table - *sdf = spark.read.table("default.<table_name>")*

Task 2 - Create a Kafka Producer - Google Colab [Component 2]

1. Create a Kafka Producer that can read the json file(unseenData.json) and publish it to a topic
2. Create a topic with the required number of partitions. You are an architect now and should be able to decide on the number of partitions
3. Publish the data as JSON String.

Task 3 - Train a model in PySpark and Save (Refer Lab 4) [Component 3]

1. Train a classification Model in Spark using 'AtRisk_academic' as the target variable
2. You can use the same model that was developed as part of Lab 4

3. Persist the model i.e save the model and load the model to predict on unseen data
4. You must use ML pipeline for all the operations (StringIndexer, VectorAssembler etc)

Task 4 - Create a Kafka Consumer [Component 4]

1. Create a Consumer that consumes the data from the topic you created. Write the code in your pyspark coding environment
2. Consume the data as spark DataFrame. You will face issues while predicting for a single instance. Hence append to a dataframe.
3. Predict for a batch size of 10

Task 5 - Save the data onto a database [Component 5]

1. Save the predicted data in a database of your choice.(Select AtRisk_academic and prediction columns)
2. You may use Atlas MongoDB service
3. Use pymongo to connect to the database

Task 5 - Display the data in Tableau [Component 6]

1. Query the database and display the table on your Tableau Dashboard
2. Congratulations! You have successfully completed building a basic Data Engineering pipeline.

Datasets

The datasets are available in the repository. You will use the same dataset that you had used for Lab-4(PySpark). You will use the unseenData.json for predictions.

Schema - Data Set Description

Attribute	**Description**
ID	The identification number
Major	Computer Science, Electric Engineering, Mathematics, Information Science, Liberal Arts, and others
Gender	M (for Male), F (for Female)
C01-C10	The score of a course C01, C02, …, C10.

Academic	The academic performance (0 to 100)
Campus	Campus evolvement attribute record how many student-organizations joined, how many events participated/organized, etc.
Internship	Internship record how many internship the student have taken. To quantify it, we use the # of months for this record. Full time internship contributes 100% month and parr-time contributes 50% of time. For example, a student have a part-time internship for 2 months, and a full-time internship for 3 months, the total month will be 2/2 + 3 = 4
AtRisk_{academic,campus,internship}:	Student may need help in {}, 0 indicates no help needed, 1 indicates help needed
AtRisk	We may identify students at risk and reach out to help them (0 as no risk, 3 as high risk)
Graduate_program	The likelihood of the student will continue in a graduate program
Government	The likelihood of the student will take a government position
Industry	The likelihood of the student will take an industry position
Placement	The possible placement includes graduate program, industry positions,government. It will be measured by 0 to 3 where 0 is no placement, and 3 is highest.
Annual	The annual salary of the student.