# Cloud-based Data Analytics
# MIE1628
# Assignment 5

**MIE1628**
***Assignment on Azure Cloud Platform***

**Due Date: Aug 15, 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.*

## *** Instructions ***

1. **Note:**

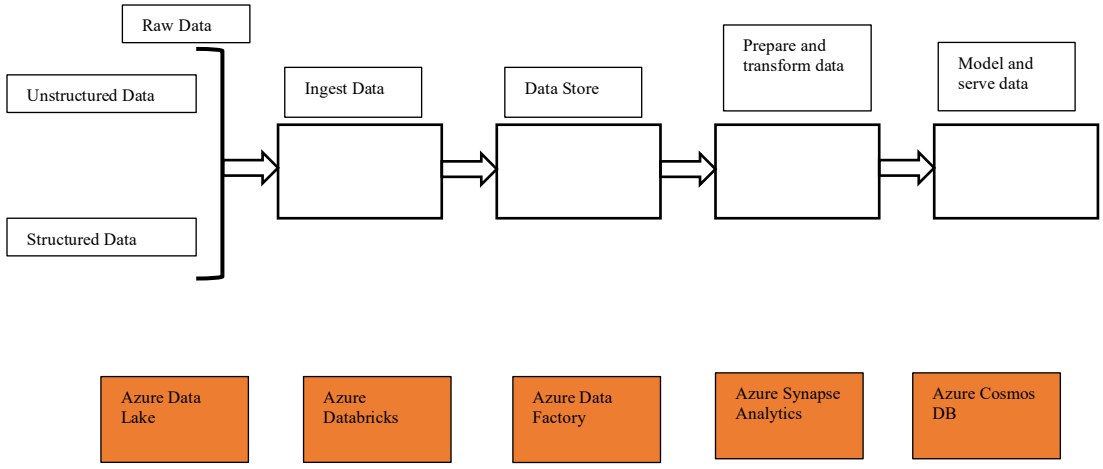
   **Part B of this assignment can be done in a group of two students or individually. Both students need to submit the assignment for both parts and provide both names, email, and student IDs at the top of the assignment.**

   Submit your complete project, including your Python Notebook with markdown explanations, and a comprehensive PDF document. The PDF should contain clear screenshots of input/output commands with results, images of your deployed Azure portal resources, detailed step-by-step explanations for each process, and final output screenshots. Additionally, include a section in the PDF answering the provided questions (to be specified separately).

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

**PART A:**

1. [Marks: 5] Explain below the 5 components shown in orange boxes. Explain which Azure components you will use where in this big data architecture and why.

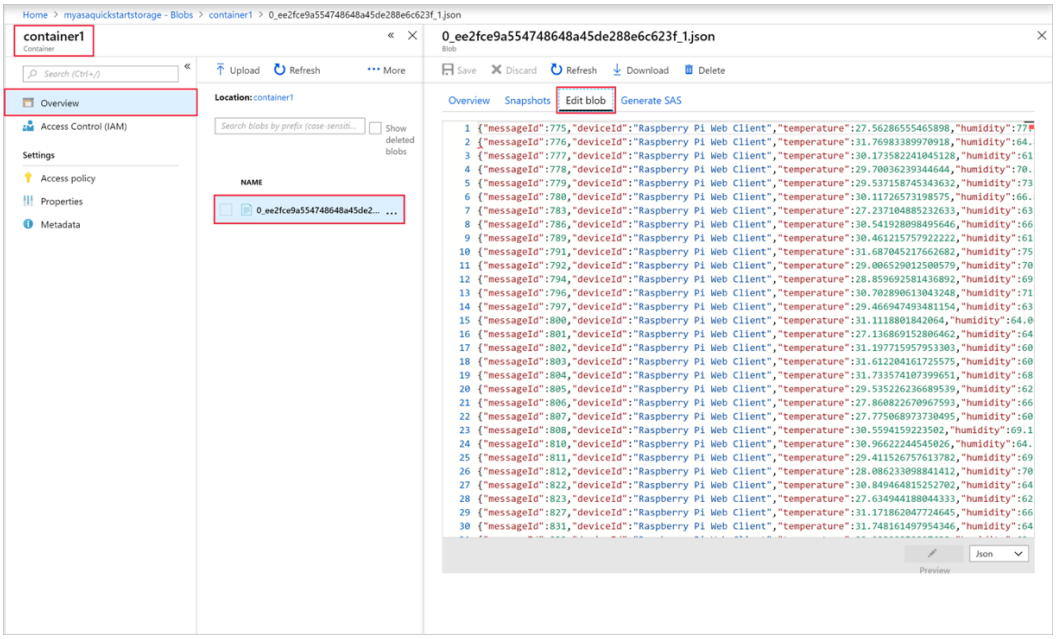


# Cloud-based Data Analytics
# MIE1628
# Assignment 5

2. [Marks: 5] Explain how Stream Analytics works in Azure. Mention at least two common use cases or applications for this service.

3. [Marks: 10] Deploy all the resources in Azure Portal. Implement a Stream Analytics job by using the Azure portal. See this for reference - https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-quick-create-portal

   For query use below:
   SELECT *
   INTO BlobOutput
   FROM IoTHubInput
   HAVING Temperature > 25
   See the below screenshot and show the top 30 results for your output.




**Part B:**

Data Input: Claim a dataset from Piazza - link. If the dataset is too large, you can take a subset of the data as well. No two groups can have the same dataset.

Your selected dataset should meet the following criteria:

1. It must contain a minimum of 1,000 instances (rows or data points).
2. It should include at least six features (columns or attributes).

Using this dataset, you are required to address a substantial and meaningful problem. Your analysis should demonstrate:

1. A clear understanding of the dataset's context and potential applications.
2. The ability to formulate relevant questions or hypotheses based on the data.
3. Appropriate use of data analysis techniques to extract insights.

# Cloud-based Data Analytics
# MIE1628
# Assignment 5

4. The capacity to draw meaningful conclusions that could inform decision-making or further research.

Some problems to consider:

1. Fraud Detection System
2. Customer Churn Rate Prediction
3. Segmentation using Clustering
4. Recommendations with your Dataset
5. Sales Forecasting
6. Stock Price Predictions
7. Human Activity Recognition with Smartphones
8. Wine Quality Predictions
9. Breast Cancer Prediction
10. Sorting of Specific Tweets on Twitter etc.

**Implement this part in Azure Machine learning using Azure Notebook**

1. [Marks: 15] Clearly define the problem you intend to address using this dataset. Present a comprehensive problem statement that includes:
    a. A detailed description of the meaningful issue you're tackling
    b. An outline of all necessary steps, including:
        i. Data preprocessing
        ii. Data cleaning
        iii. Modeling approach

    Your problem statement should be thorough, spanning approximately half to one full page. If you determine that data cleaning is unnecessary, please provide a justification for why this dataset doesn't require cleaning. In such a case, allocate more attention to other crucial aspects such as EDA and the modeling process.

    Ensure your problem statement is well-structured, coherent, and provides a clear roadmap for your data analysis project.

2. [Marks: 10] Explore your dataset and provide at least 5 meaningful charts/graphs with an explanation.

3. [Marks: 10] Do data cleaning/pre-processing as required and explain what you have done for your dataset and why?

4. [Marks: 15] Implement 2 machine learning models and explain which algorithms you have selected and why. Compare them and show success metrics (Accuracy/RMSE/Confusion Matrix) as per your problem. Explain results.

5. [Marks: 15] Deploy a run-time pipeline for your dataset using Azure Designer Studio.
    Or
    Do hyperparameter tuning for your algorithms. Explain your results.
    Or
    Use Automated ML for your data set. Explain the best model results.

# Cloud-based Data Analytics
# MIE1628
# Assignment 5

6. [Marks: 15] Summarize your project's key findings and overall conclusions in a brief paragraph. Ensure your summary is firmly grounded in the data and analysis you've presented throughout your project. Offer meaningful insights that not only encapsulate your work but also lay a foundation for potential future research in this area. Your conclusions should be well-reasoned and directly supported by your results.