Guidelines

You are to submit an 8-10 page PDF report using reasonable fonts and spacing. A suggested outline of the Group Final Report includes:

1. A description of the use case that you want to address. Discuss why the problem is useful and important.
2. Some basic facts about the data, including the following:
   1. Source of the dataset(s)
   2. Describe the dataset(s) - E.g., number of observations, number of variables, type of variables. (string, integer, etc.) You may present the variables in a tabular form.
   3. An explanation of what you think are the interesting aspects that you can look at in the dataset(s)
   4. A discussion of what you hope to mine from the dataset(s)/some hypotheses that you might have in mind that you hope to verify - E.g. spot increasing trends etc.
3. A detailed discussion of the following:
   1. Pre-processing steps to sanitize/manipulate/combine your dataset(s).
   2. Design of databases and data warehouses in the form of an ER diagram. Explicitly specify how you define the primary/foreign keys of the tables.
   3. Snapshots of tables with data for both databases and data warehouses.
   4. Describe the rationales that you have considered for your design of the pipeline
   5. Snapshot of the graph visualization of your pipeline.
   6. Snapshot of the tree view of your pipeline after triggering your DAG.
   7. Run time of each step in your pipeline.
   8. What visualizations (line chart/bar chart/histogram, etc.) and machine learning models you have used (E.g., Unsupervised? Supervised?)
   9. The insights that you have gathered from the downstream applications.
4. Discussions
   1. What are the performances of the pipeline (speed, accuracy if ML models are used)?
   2. How can the model be integrated into the respective business process?
   3. What insights can be obtained from the dashboards?
   4. Etc.
5. Conclusions

Grading Criteria

Your report will be evaluated based on the following:

1. Clarity and completeness of the report
2. Appropriateness of the models and methods applied for data processing and analysis
3. The usefulness of your analysis in other similar real-world problems
4. Concise summarization of your work
5. Reasonability of the discussion of advantages and limitations of applied methods to the defined problem

# 1. Introduction

# 2. Data Sources

# 3. Data Pipeline

* Workflow management (Airflow)
* Database/warehouse

# 4. Downstream Applications

* Dashboard
* ML model

# 5. Conclusions