**Title: Data Engineering Capstone**

This project is supposed to be done in the groups of three to four.

There are four roles that a person can assume:

1.Data architect

2.Data pipeline builder

3. Apache spark developer

4.cloud engineer for data leak and data security

**project objective** is to collect and process the data so that other data analysts and data scientists can utilize your data for their business use cases

As a data engineer you would first build the bronze layer where you would collect the data as is. Then you will build a silver layer where you would clean the data and remove all impurities such as missing values or datetime correction. Finally, you will build a gold layer where you will design the distribution of the data sets and save the graphs back into the data lake.

The bronze layer should be built using data lake and data factory and after that use Azure data bricks to move the data from bronze to silver finally load this data into synapse analytics and generate the charts there. Alternately you can also use data factory to build silver layer and use data bricks finally in the gold layer.

As a part of this submission, you will be required to upload your data factory pipelines and Py spark notebooks into a GitHub account and share the repository link, additionally also submit a diagram showing the flow of your data and applications.