Stage 1
1 Pass all **sales files** from local to **S3** bucket using python
2 once data is available in S3 , do data cleaning and merge all files in S3 in another bucket (new bucket) using Glue with schedule
3 If file is available in S3 New Bucket then lambda function should trigger then pass all data to **AWS RDS**

Stage 2
1.Pass all the **features file** and Store the file in **HDFS** (HADOOP FILE SYSTEM)
2.Push the feature data from HDFS to **Hive** Table database via **PYSpark job (optional: Do data cleaning if required)**

Stage 3
1 Pass all **store data** to **AWS RDS (optional: Do data cleaning if required)**
2 Push all rds data to **Hive** via **Sqoop**

Stage 4
This environment should be capable to merge **stage 1, 2, 3**

Stage 5
Based on requirements stage 4 should be capable to store in rds/file using pyspark job

Stage 6
From stage 1 to stage 5, based on possible scenario these stage should be connected in airflow pipeline

Stage 7
Stage 6 should be monitored via prometheus

Note : Participants choice to decide Docker or Cloud environment

Requirements :

- Identify average customer visit in the **type B** store in **April Months**
- Identify best average sales in holiday week for **all store** types
- Which store had a worst sales in **leap year**
- What is the expected sales of each department when unemployment factor is greater **>8**
- Aggregate the net(total) sales of each department on month wise
- Which store performs high sales in week wise
- Identify better department performance based on the store on all the week
- Identify the store which has minimum fuel price based on the week
- Identify overall performance of the store based on year wise
- Identify the performance of the store on week wise with without offers