INTERMEDIATE PROGRESS REPORT

Introduction: Performing analysis and visualization on Formula1 race results using Databricks and Azure. We are getting data for all the F1 races from an open-source API called Ergest developer API. It has different tables of data like drivers, constructors, lap times, pit stops, races, results, seasons, status and the circuits etc. Tables are represented as single-line or multi-line CSV/JSon folder.

Technologies used:

- Databricks
- Azure
- PySpark
- Python

Team Information:

Deepthi Reddy Kallam – 16341887
Chandra Sekhar Akkandra – 16343150
Bhavana Navari – 16341885
We formed a group on Canvas called ***'Cloud Project'*** and we collaborated with each other as we live nearby.

Things that have been tried/accomplished:

→We collected raw data from the API and stored it in containers on the Azure.
→We have joined the key information required for reporting to create a new table.
→We processed this data using Databricks notebook to ingest into ingested raw layer.
→We have ingested data which has requirements like - the data should be in columnar format and the schema must have been applied to the ingested data and the status must be analyzed via SQL.

→We have created a project structure and was able to categorize different kinds of data in the project.
→We have setup notebooks to mount ADLS storages in Databricks.

Things to do:

→We need to transform and analyze the data.
→We must use Azure data factory to schedule and monitor the data with different requirements like scheduling should be done on a particular day of a week and ability to re-run the failed pipelines and ability to setup alerts on failures.
→We are trying to add a feature like – the ability to roll back to a previous version.
→We are yet to finish the programming part of the project.
→We still need to work on the visualization of the results.

Challenges/Comments:

→It was challenging to import data from the API to the cloud containers.
→We learned how to use the SQL to transform the raw data to required logical data.
→We got to know how to use the Databricks notebook to ingest data.
→It was difficult for us to transform the data by applying the schema.
→It took long time to add additional information to the data, which was transformed via a Databricks notebook for presentation layer.
→ We learned how to operate on the data and how to use multiple cloud services on the data.