Context

Formula 1 (a.k.a. F1 or Formula One) is the highest class of single-seater auto racing sanctioned by the Fédération Internationale de l'Automobile (FIA) and owned by the Formula One Group. The FIA Formula One World Championship has been one of the premier forms of racing around the world since its inaugural season in 1950. The word "formula" in the name refers to the set of rules to which all participants' cars must conform. A Formula One season consists of a series of races, known as Grands Prix, which take place worldwide on purpose-built circuits and on public roads.

Our upcoming project revolves around Azure Data Engineering, with a primary focus on Databricks. This initiative is set to span 15 Days, equivalent to two weeks. Regular updates are essential and should be performed every Saturday, commencing on the 5th of Nov.

The project is centered around the Formula 1 dataset, comprising eight raw files. Kindly locate the input file paths for reference. "adlssonydatabricks/raw/project2/file\_names."

Throughout the project, you will gain expertise in various technologies, including Azure Databricks, Delta Lake, PySpark, Spark SQL, and Data Lake, Databricks Workflow, Databricks Repos(CI/CD), Databricks SQL. Get ready to embark on a journey of learning and innovation!

Find raw files here

Storage Account Name: adlssonydatabricks

Container= raw

Directory= project2

## **1. Project Description**

This project focuses on developing a scalable, efficient data pipeline in Databricks to process and analyze Formula 1 historical data. Utilizing the medallion architecture (bronze, silver, gold layers) on Delta Lake, we will perform data ingestion, transformation, and enrichment, ensuring data quality, scalability, and efficient querying. The project includes advanced PySpark transformations, SCD Type 2 implementation, Delta Lake optimization, and automation of workflows for robust end-to-end data processing.

### **Objectives**

1. **Ingest and Clean Data**: Ingest Formula 1 data from multiple sources and ensure data quality through validation and cleaning.
2. **Medallion Architecture**: Transform data through bronze, silver, and gold layers, structuring it for various levels of analysis.
3. **Optimize Data and Implement SCD**: Use Delta Lake for data versioning and implement SCD Type 2 for historical tracking.
4. **Data Governance and Security**: Ensure data governance using Unity Catalog, with role-based access control and data masking.
5. **Automate the Workflow**: Automate the entire process using Databricks Workflows to ensure timely data refreshes.

**2. Dataset Schema** (Details are mentioned at last of project document)

## **3. Solution Architecture**

### **Phases of the Project**

#### **Phase 1: Data Ingestion**

* **Source Files**: Ingest CSV and JSON files from a source directory into the Databricks environment.
* **Storage in Bronze Layer**: Load raw data into Delta tables in the bronze layer of the medallion architecture. Each table mirrors the source structure, serving as a raw data repository.
* **Quality Checks**: Implement basic validations (e.g., data type checks, null handling) to ensure initial data consistency.

#### **Phase 2: Data Transformation and Enrichment**

* **Silver Layer Processing**:
  + Standardize and clean data.
  + Apply advanced PySpark transformations (joins, aggregations, window functions) to create enriched views:
    - Join race results with driver and constructor details.
    - Aggregate pit stops, lap times, and qualifying results to create consolidated race data.
  + **SCD Type 2**: Implement SCD Type 2 on dimensions (e.g., Constructors, Drivers) to maintain historical records.
* **Data Quality and Validation**: Apply additional transformations and quality checks, including duplicate handling, logical validation (e.g., lap times must be positive), and integrity checks between tables.

#### **Phase 3: Data Aggregation and Analysis in the Gold Layer**

* **Aggregated Tables**: Create final tables for reporting and analysis:
  + **Driver Performance**: Aggregate total points, wins, podiums, and average position per season.
  + **Constructor Standings**: Total points and wins by constructor per season.
  + **Circuit Insights**: Analyze pit stop frequencies, average lap times, and finishing positions per circuit.
* **Delta Lake Optimization**: Optimize gold tables by applying partitioning, z-order indexing on fields like race\_id and driver\_id, and optimizing for fast retrieval.

#### **Phase 4: Workflow Automation**

* **Databricks Workflow**: Use Databricks Workflows to schedule and automate the pipeline, ensuring periodic updates (e.g., after each race season).
* **Error Handling**: Implement error logging to capture and report ingestion, transformation, and load errors.

#### **Phase 5: Data Governance and Security**

* **Unity Catalog and Role-Based Access**: Implement role-based access control to secure data.
* **Data Masking**: Apply masking on sensitive information, if needed.
* **Audit Logging**: Track and log access and modifications to Delta tables for compliance.

## **4. Project Deliverables**

1. **Data Pipeline**: A fully automated data pipeline in Databricks for ingesting, processing, and storing F1 data.
2. **Medallion Architecture Implementation**: Bronze, silver, and gold tables created and optimized in Delta Lake.
3. **SCD Type 2 Implementation**: Historical tracking for dimensions such as Drivers and Constructors.
4. **Optimized Delta Tables**: Partitioned and z-ordered gold tables for efficient querying.
5. **Dashboards**: Visual dashboards to display key metrics (can be developed in Databricks or integrated with Power BI).
6. **Data Governance Controls**: Access control and data masking configurations in Unity Catalog.
7. **Documentation**: Complete documentation covering the pipeline design, architecture, schema, and automation setup.

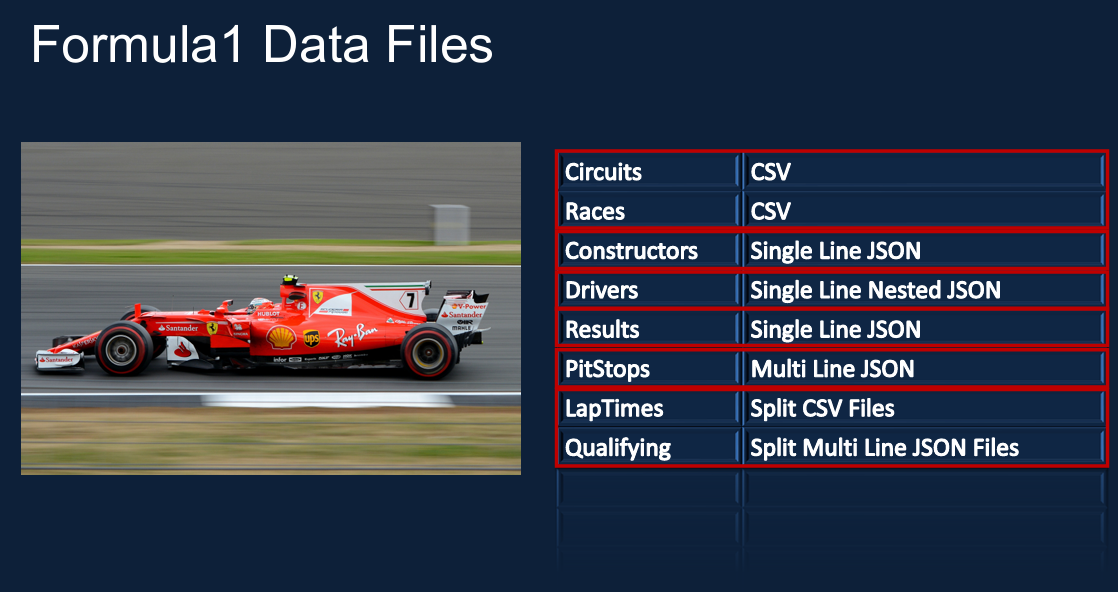
## **5. Key Performance Indicators (KPIs)**

1. **Data Freshness**: Time taken for the pipeline to update after new data ingestion.
2. **Query Performance**: Average time for retrieving data from gold tables.
3. **Data Accuracy**: Percentage of records passing validation checks at each stage.
4. **Processing Time**: Time taken to process data from raw ingestion to gold layer.
5. **SCD Implementation Accuracy**: Correctness of historical records for dimension tables.

## **6. Project Expectations**

1. **Scalability**: The pipeline should handle future F1 data without significant changes.
2. **Accuracy**: Ensure data transformations preserve data integrity and meet quality standards.
3. **Automation**: The pipeline should be fully automated and require minimal manual intervention.
4. **Compliance**: Data access should be restricted and logged per governance requirements.
5. **High-Performance Retrieval**: Gold tables should be optimized for analytical queries with minimal latency.

**Dataset Schema**



A diagram of a diagram of a structure

Description automatically generated with medium confidence

A diagram of a data architecture

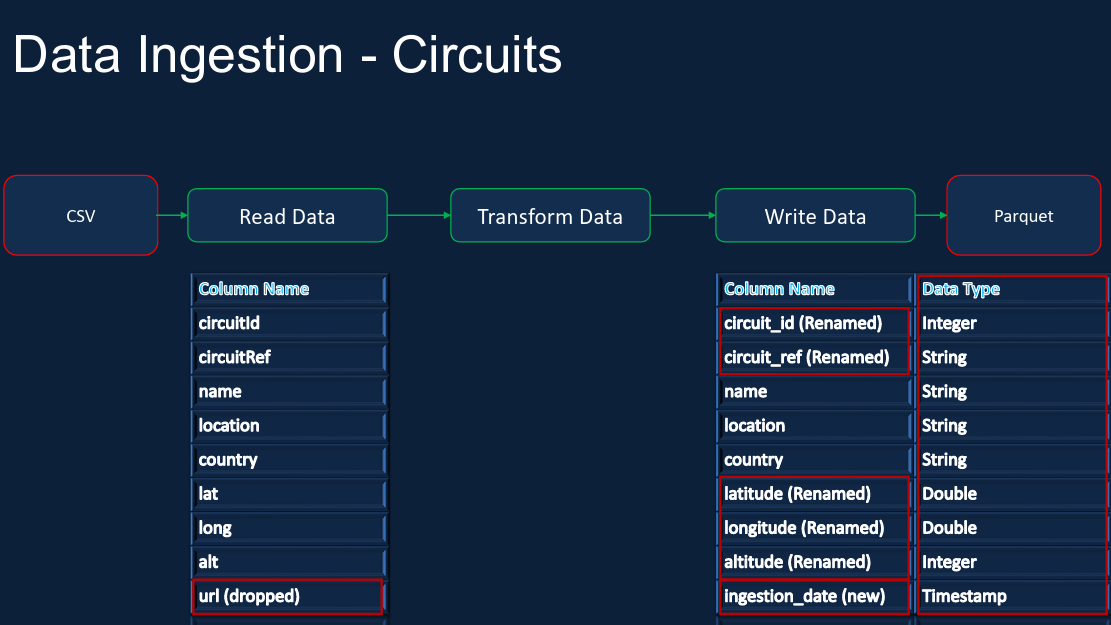
Description automatically generated

A red race car on a track

Description automatically generated

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated