Formula 1

Evan de Guzman

Overview

Race Circuits

- 24 races
- 21 countries

Team/Constructors

- 10 Teams

Drivers

- 20 drivers

Practice

- Friday 60 min free practice (FP1 and FP2)
- Saturday final practice (FP3)

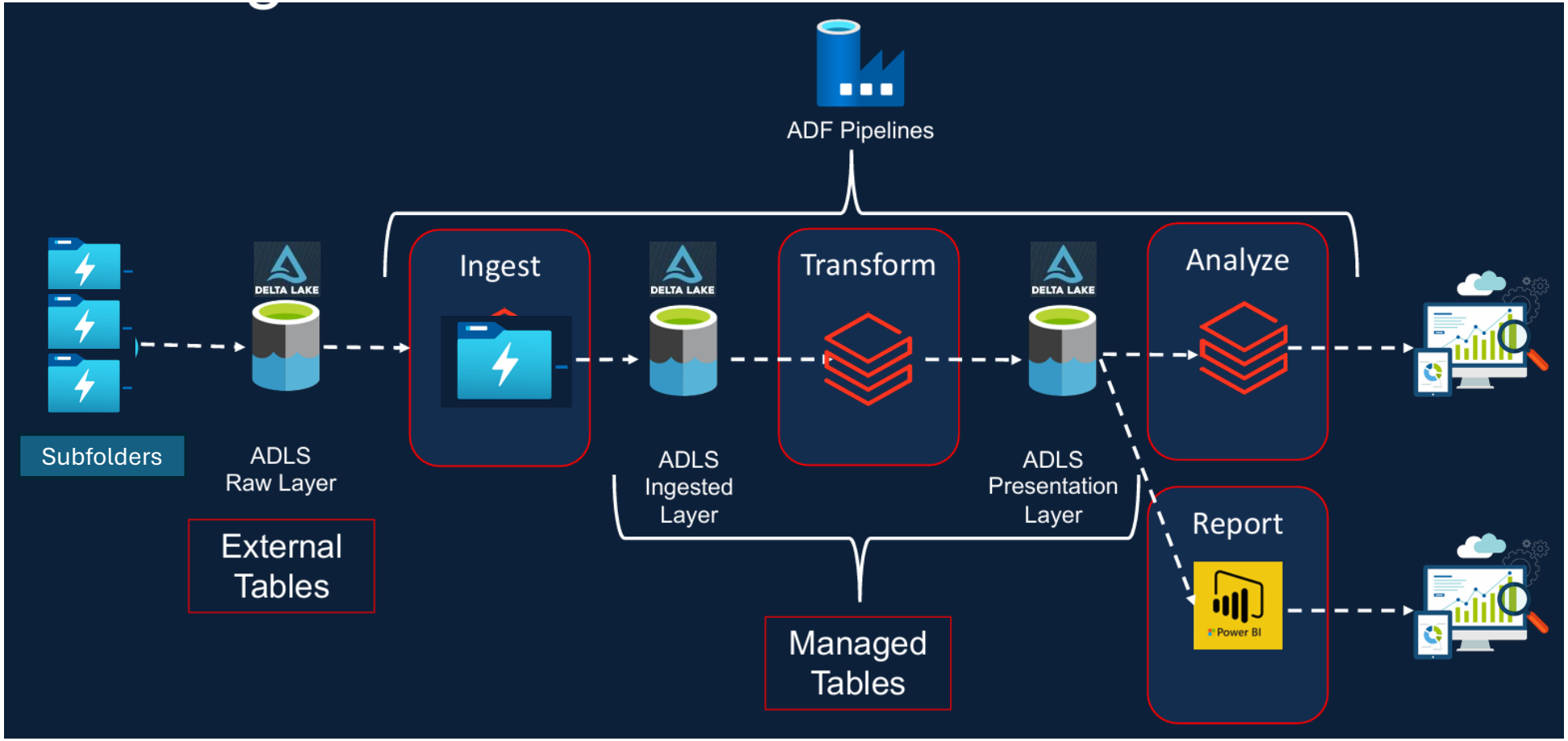
Qualifying

- 3 stage session (Q1, Q2, Q3)

Race

- Sunday the Grand Prix
- Laps
- Pit Stops

Solution Architecture



Data Files

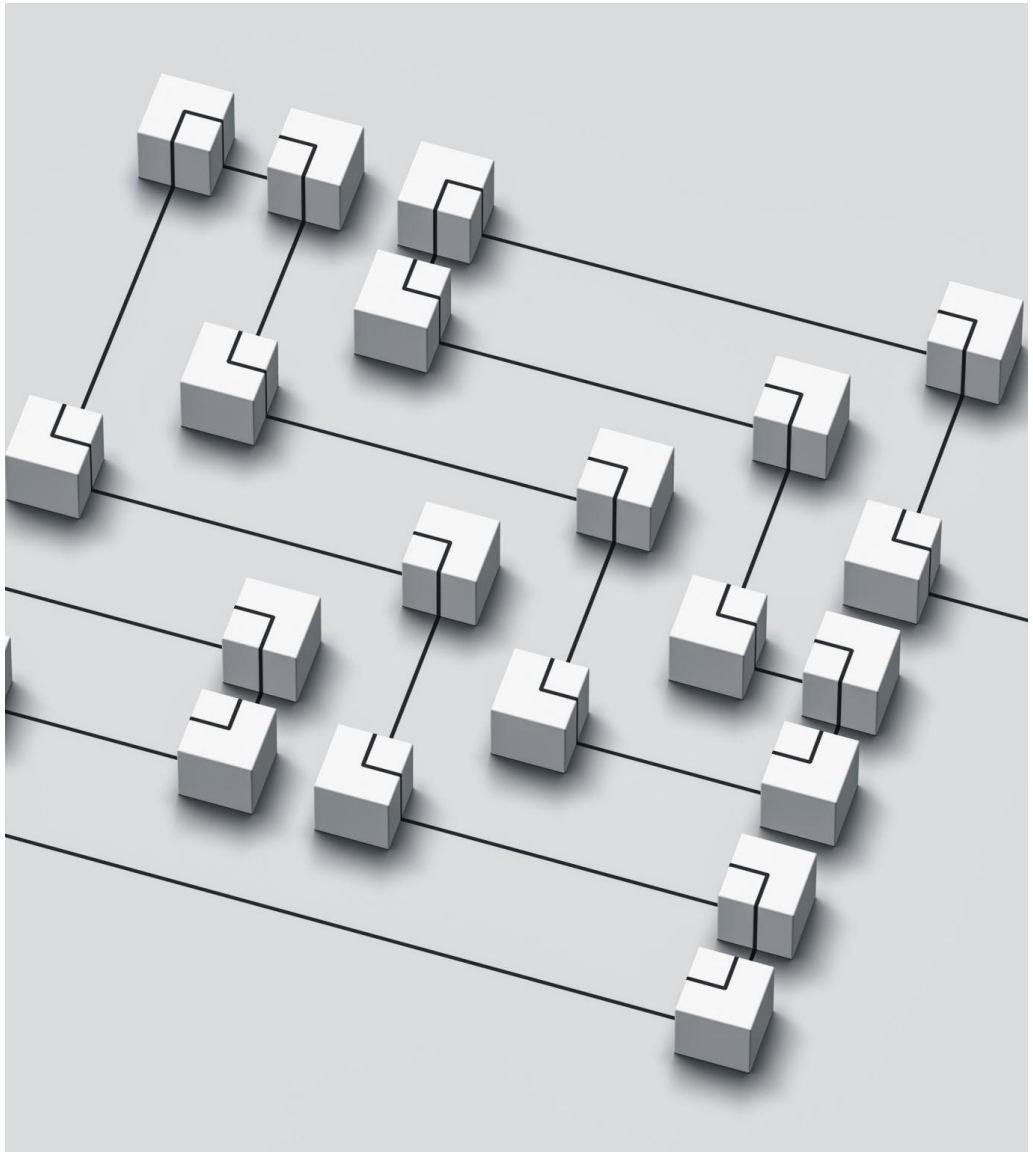
- Circuits - csv
- Races - csv
- Constructors – single line json
- Drivers – single line json
- Results – single line json
- PitStops – multi line json
- LapTimes – Folder/csv
- Qualifying - Folder/multi line json

Data Ingestion Requirements

- Ingest all 8 files into the data lake
- Ingested data must have the schema applied
- Ingested data must have audit columns
- Ingested data must be store in columnar format (i.e., Parquet)
- Must be able to analyze the ingest data via SQL
- Ingestion logic must be able to handle incremental load

Data Transformation Requirements

- Join the key information required for reporting to create a new table
- Join the key information required for analysis to create a new table
- Transformed tables must have audit columns
- Must be able to analyze the transformed data via SQL
- Transformed data must be stored in columnar format (i.e., Parquet)
- Transformation logic must be able to handle incremental load



Reporting Requirement

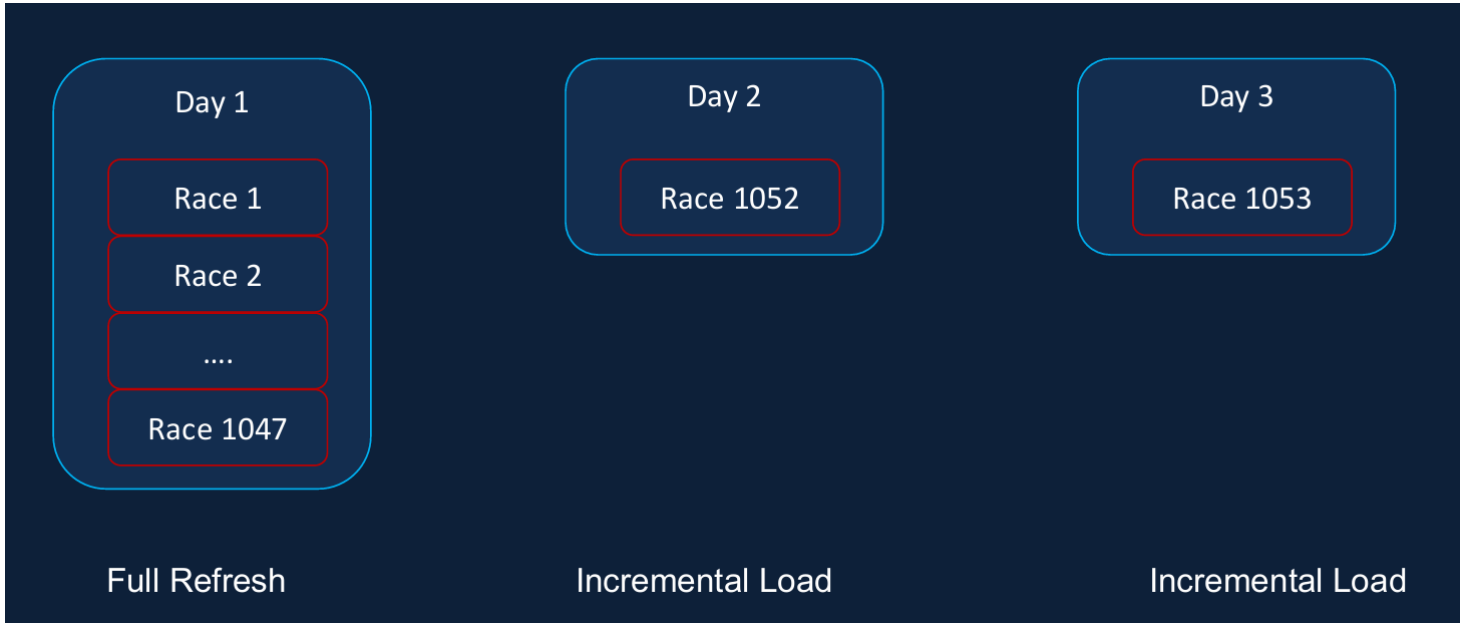
- Driver Standings
- Constructor Standings

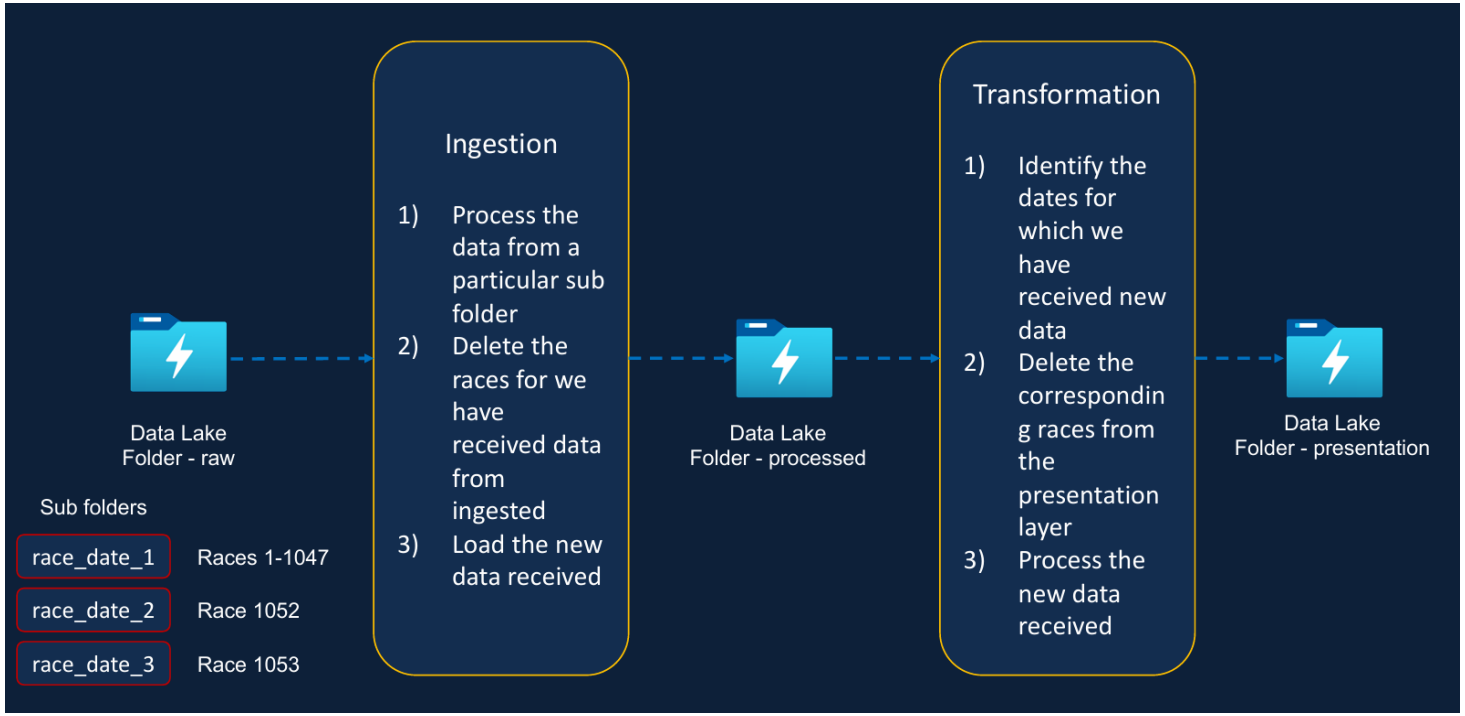
Analysis Requirement

- Dominant Drivers
- Dominant Teams/Constructors
- Visualize the outputs
- Create Databricks Dashboards

Data Load

- Full Load
  - Circuits – data from all races
  - Races – data from all races
  - Constructors – data from all races
  - Drivers - data from all races
- Incremental Load
  - Results – data only from that race
  - PitStops – data only from that race
  - Laptimes – data only from that race
  - Qualifying – data only from that race





Presentation Layer

Dominant Drivers / Dominant Teams

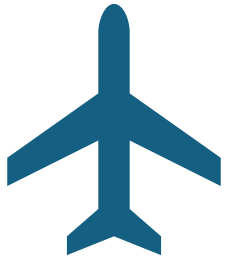
- Grouped Aggregations
- Granularity of the data – race_year, driver, team
- Rank the dominant drivers of all time/ last decade
- Rank the dominant teams of all time/ last decade

Scheduling Requirements

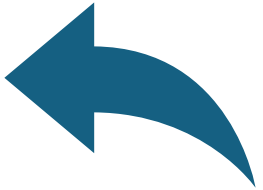
- Schedule to run every Sunday 10pm
- Monitor pipeline
- Ability to re-run failed pipelines
- Set up alerts on failures

Other Non-Functional Requirements

Ability to delete individual records



See history and time travel



Roll back to a previous version