University of Dundee

Real time data warehousing with Airflow (Formula1 data)

Hamsa Ravikumar
SUPERVISOR: Andrew Cobley

Introduction

F1 is one of most technologically advanced and competitive sports in the world. Behind the world of fast cars and exotic locations is a sport driven by data. The ability to collect and analyse huge amounts of data is critical to teams and drivers allowing them to develop and evolve race strategies and adapt them in real-time as they attempt to gain a split-second advantage over the opposition.

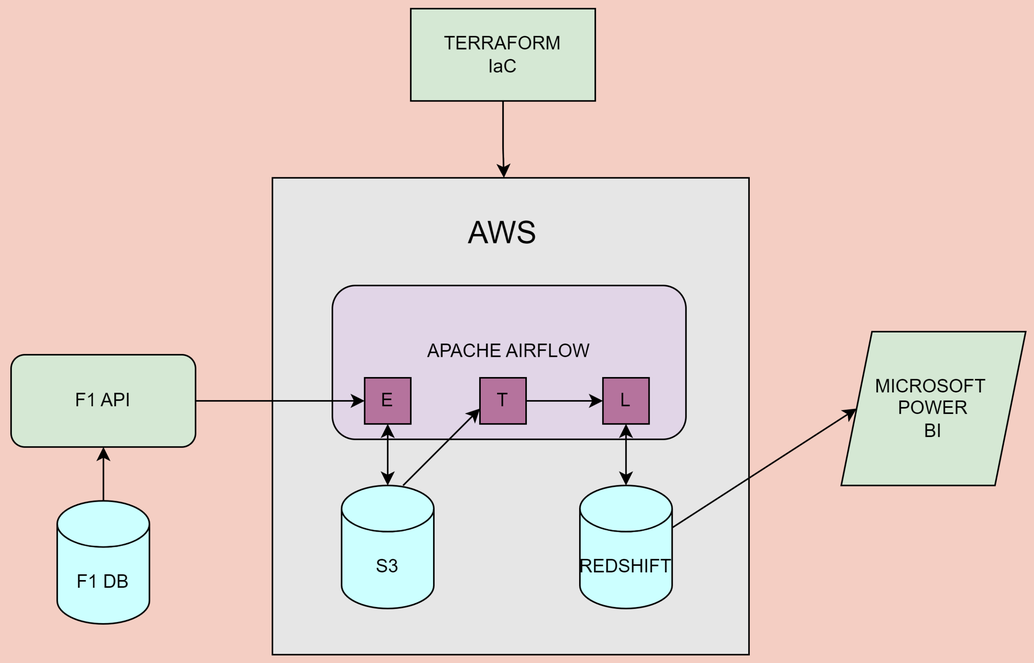
The project aims to demonstrate a key technique used in F1 data analysis creating a real time data pipeline to extract, transform and load (ETL) Formula1(F1) data from a dynamic data source to a cloud-based data warehouse where it can be used to provide real-time analysis allowing the user to make live decisions based on latest continuously changing data.

All the resources used in this project are programmatically configured using Terraform, an open-source infrastructure as code (IaC) software tool. Terraform defines and provisions the infrastructure resources, such as virtual machines, network, storage and services using declarative configuration files in AWS cloud.

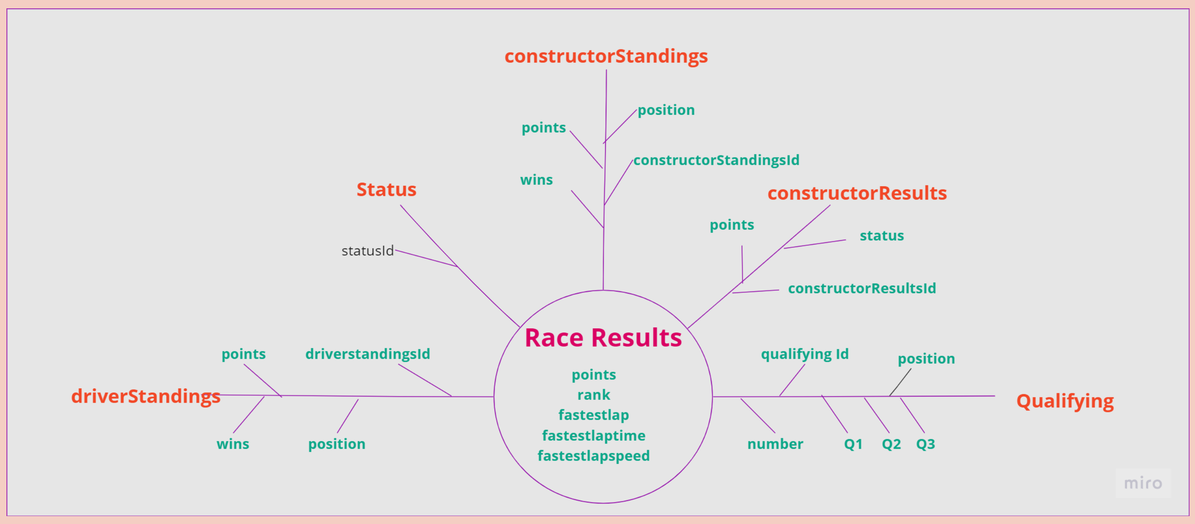
Apache Airflow is used to develop python workflows to query a fastf1 library in real time and extract F1 data, transform it for analysis and load into a cloud-based warehouse. Airflow is also configured to schedule periodic ETL for a data refresh thereby establishing a real-time automated ETL data pipeline.

AWS Redshift, a cloud-based data warehouse is used to store large volumes of data through columnar storage format and high compression ratios. The parallel processing architecture allows faster query performance and supports complex analytical queries on the large datasets.

System Architecture



Realtime Datawarehouse with Airflow for F1 data



Sun model - Race table measures and dimensions

Objective

- Implement a streamlined cloud infrastructure provisioning process to support the real-time data pipeline.
- Design and deploy a containerized deployment architecture to enhance scalability and resource utilization.
- Develop and maintain a dynamic real-time ETL data pipeline leveraging Apache Airflow.
- Utilize Python frameworks for efficient Extract, Transform, and Load (ETL) processing in the real-time data pipeline.
- Optimize data loading and querying operations to handle large-scale data volumes efficiently.
- Establish a cloud-based data warehouse infrastructure to serve as the central repository for real-time data.
- Perform comprehensive data analysis to derive actionable insights from the real-time data streams.
- Implement advanced data visualization techniques to present analytical findings effectively for decision-making.

Technology

The tools enumerated below, along with their respective technologies, are employed across various phases of the project

Discussion

The inability to process streaming data limits the agility and flexibility of organizations to adapt to data changes and trends. It introduces a costly latency between the time data is generated and its analysis.

A real-time data warehouse facilitates instantaneous collection and analysis of vast amounts of data at an unprecedented pace. For instance, during a race where the top speed of car has direct correlation with the lap-times a driver is achieving. Realtime analysis of this F1 data allows a team to adjust car performance during the race, redirecting resource away or toward top speed to improve the driver's overall race performance. Furthermore, introducing additional data such as tire degradation and weather, F1 teams can start making decisions that collectively give them a significant competitive edge real-time.

Cloud-based real-time data warehouses operate through a blend of scalable infrastructure, distributed computing and optimization. This allows the scale of the analysis to grow substantially and dynamically, without the constraints associated with an on-premise data warehouse which can struggle with unpredictable spikes in memory and compute, crucial for effective big data processing.

Maximising the benefits of cloud-based real-time analysis allows F1 teams to win the race.