Lake challenge analysis

Design:

The Uber Movement Dataset from Kaggle (<https://data.world/rmiller107/travel-time-uber-movement>) is ingested to AWS data lake and then processed in Snowflake data warehouse orchestrated using Airflow as below.

Timeline

Description automatically generated with medium confidence

The source data is ingested to AWS S3 Raw bucket as-is to maintain a copy of input Raw data using Airflow S3 connection.

The Raw data is processed with compression techniques and loaded to S3 Raw-plus bucket in parquet format using Pyspark/EMR. Parquet provides faster data retrievals with minimum latency and can be structed using Glue crawler and querying Athena tables.

The Raw-plus data of parquet is transformed with derived dataset and staged to s3 staging bucket for Snowflake data loads using Pyspark/EMR.

The Snowflake staged data is copied to stage table and curated into facts and dimensions tables using normalization techniques and dimensional modelling. The historical versions along with timestamps are captured which can be used for identifying day-over-day changes.

The Snowflake dimensional data is denormalized into a cube with data elements required for consumer reporting and analysis.