Documentation Data Engineer

This page is documentation for the level of DE expertise.

- Incoming
- Airflow
- MySQL
  - Landing
  - Aggregations
- Metabase Reporting system

Incoming

Incoming data appears in S3 bucket 'eastrockvalley-test' (root folder) as standart CSV files with header line

- games.csv (GameID,UserID,GameType,PlayDate,Duration)
- users.csv (UserID,UserName,SignUpDate,Country)
- payments.csv (TransactionID,UserID,Amount,TransactionDate,Type)

The examples of the files can be found in GIT.

If the file is already present in the bucket, it means it has not been processed yet.

Airflow

The AirFlow is in charge of processing files. In the next step, the data lands in MySQL database, and the source files are deleted.

The AirFlow has 3 DAGs for each file. Scheduled for every 3 min. Should be monitored in UI.

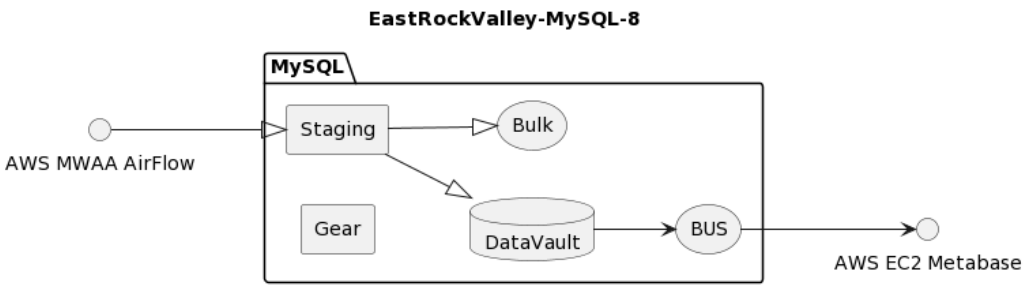
The code of DAGs (Python 3.10) in GIT (please update in case of changes).

MySQL

Instance run in Amazon RDS

Datawarehouse is split into a few schemas:

- Staging - all incoming data
- Bulk - all successfully processed data
- Vault - datavault storage
- Bus - presentation layer
- Gear - storage for routines and temporary tables



The last code of routines also must be always put in GIT. UML link files are also presented.

Landing

The Staging has 3 tables for incoming files to reflect data 1:1. The event gear.ten_minutes_routine_staging (every 10 minutes) is calling gear.exec_process_staging which processes all the lines in staging tables. After processing, data is copied into the bulk schema, and staging is truncated.

While processing, it populates the tables with datavault methodology in the vault schema.

Aggregations

The aggregation is triggered by the event gear.ten_minutes_routine_aggregations, which calls gear.calc_aggregations.

This routine is checking what data was imported or updated using table gear.recent_changes. In case the changes are registered, the serial of aggregations is running. Recalculation takes place only for new/updated period and a list of elements.

The result of aggregations is several tables in the schema bus. This is the storage for tables and views requested for the reporting system or external outcome.

Metabase Reporting system

The Metabase is placed in AWS EC2 instance.

Only the BUS schema is opened.

Metabase can call directly certain tables or views, but also SQL selections can be written in a reporting system.
Need to try to keep the SQL code in GIT to be able to monitor what and how is used,