NASA MARS DATA ANALYSIS

Github: nasamarsdata

Introduction:

- This document outlines the technical design for a data pipeline that retrieves data from the NASA API, stores it in Amazon S3, loads it in Snowflake's RAWZ layer, and then cleans and loads it into the final destination tables and schema. The pipeline leverages AWS Lambda functions for serverless data processing and Snowflake for data storage and transformation.

Objective of the ETL Pipeline:

- Streamlining data retrieval from NASA API
- Ensuring data consistency and reliability through transformation
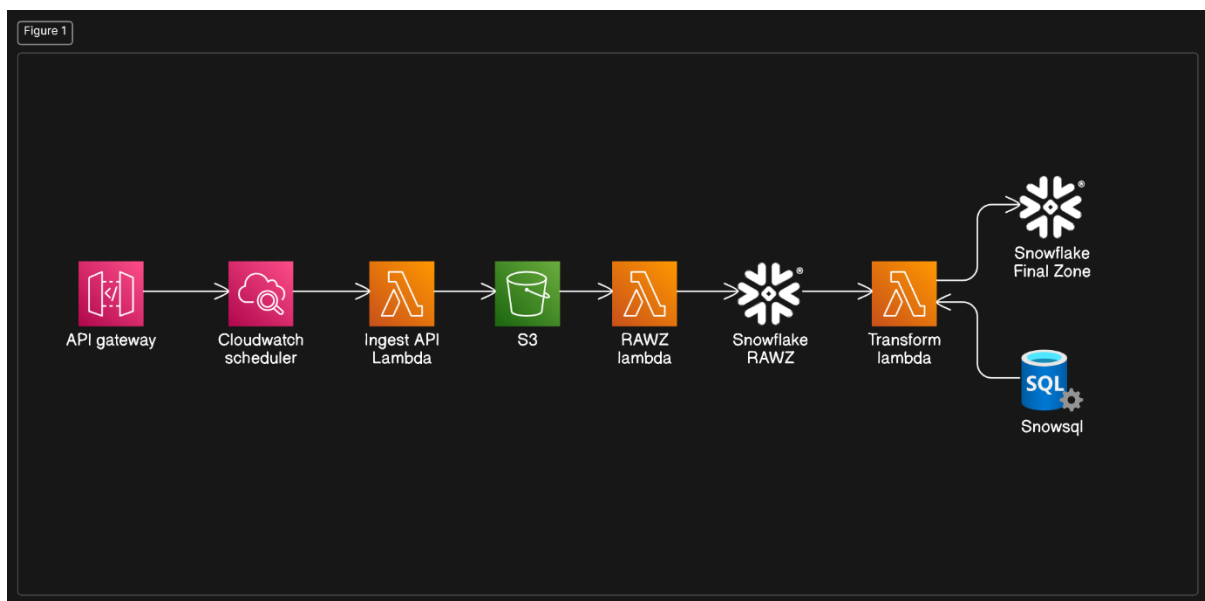- Loading transformed data into Snowflake for further analysis

Fig: ETL FLOW

NASA API:

- The Space Weather Database Of Notifications, Knowledge, Information (DONKI) is a comprehensive on-line tool for space weather forecasters, scientists, and the general space science community. DONKI chronicles the daily interpretations of space weather observations, analysis, models,

forecasts, and notifications provided by the Space Weather Research Center (SWRC), comprehensive knowledge-base search functionality to support anomaly resolution and space science research, intelligent linkages, relationships, cause-and-effects between space weather activities and comprehensive webservice API access to information stored in DONKI.

System Architecture:

The system consists of the following components:

- **Trigger**: An external trigger (e.g., scheduled event, API call) initiates the data pipeline.

- **Lambda Function 1 (NASA API Lambda)**:
1. Fetches data from the NASA API.
2. Transforms the data to JSON format (if necessary).
3. Uploads the JSON data to an S3 bucket.

- **S3 Bucket**: Temporary storage for the downloaded NASA data in JSON format.

- **Lambda Function 2 (Snowflake Ingestion Lambda):**
1. Reads the JSON data from the S3 bucket.
2. Stages the data in Snowflake's RAWZ layer (potentially using external stages).
3. Performs basic data validation (optional).

- **Snowflake RAWZ Layer**: Temporary storage for the staged data in Snowflake.
1. Lambda Function 3 (SnowSQL Execution Lambda):
2. Executes a SnowSQL script containing transformations and loading logic.
3. Transforms the data in the RAWZ layer according to the defined SnowSQL script.
4. Loads the transformed data into the final Snowflake tables.

Technology Stack:

- Cloud Platform: AWS
- Serverless Functions: AWS Lambda
- Scheduler: AWS Cloudwatch

- Data Storage:
1. Temporary: Amazon S3
2. Permanent: Snowflake
- SQL Scripting: SnowSQL