**RETRACED - DATA ANALYTICAL PIPELINE**

The goal of this project is to construct an analytic pipeline utilizing AWS services to transform un-optimized CSV files into data structures that can be quickly queried for insights. The pipeline must specifically:

* Using Python lambda functions, standardize the input CSV files
* The standardized CSV files should be converted to Parquet files and kept on S3.
* To read the Parquet files from S3 and save them in Athena tables, use an AWS Glue Crawler job.
* To facilitate simple analysis and insights, visualize the data with QuickSight.

****

**Step 1** : Data Ingestion

Place two CSV files into an S3 bucket using the Boto3 package.

**Step 2 :** Data Transformation

Convert data from CSV to Parquet format for effective querying

Using Amazon Glue Python shell.

**Step 3 :** Data Storage

Organize converted data in an S3 bucket for storage.

The stored data can be directly queried with SQL.

The data can also utilize Amazon Glue to create an Athena table.

**Step 4 :** Data analysis

Create a star schema based on date, company, and the number of logins.

Create fact table with login information.

Create dimension tables with date and company information.

**Step 5 :** Data cubing

Construct external tables that "cube" the data for simple aggregate querying,

by using AWS EMR with Apache Hive.

**Step 6 :** Server-less Querying

Perform serverless SQL queries on the Athena table using Amazon Lambda.

To retrieve data for the relevant duration's, Lambda functions can be scheduled to run at predetermined intervals.

**Step 7 :** Data Visualization

Date can be visualized with Amazon QuickSight, such as a bar chart showing the number of active organizations over time.

