Low Level Design

Amazon Records Data Analysis

Written By	Dipankar Modak, Akash Sahu
Document Version	0.1
Last Revised Date	

DOCUMENT CONTROL

Change Record:

VERSION	DATE	AUTHOR	COMMENTS

Contents

1. Introduction

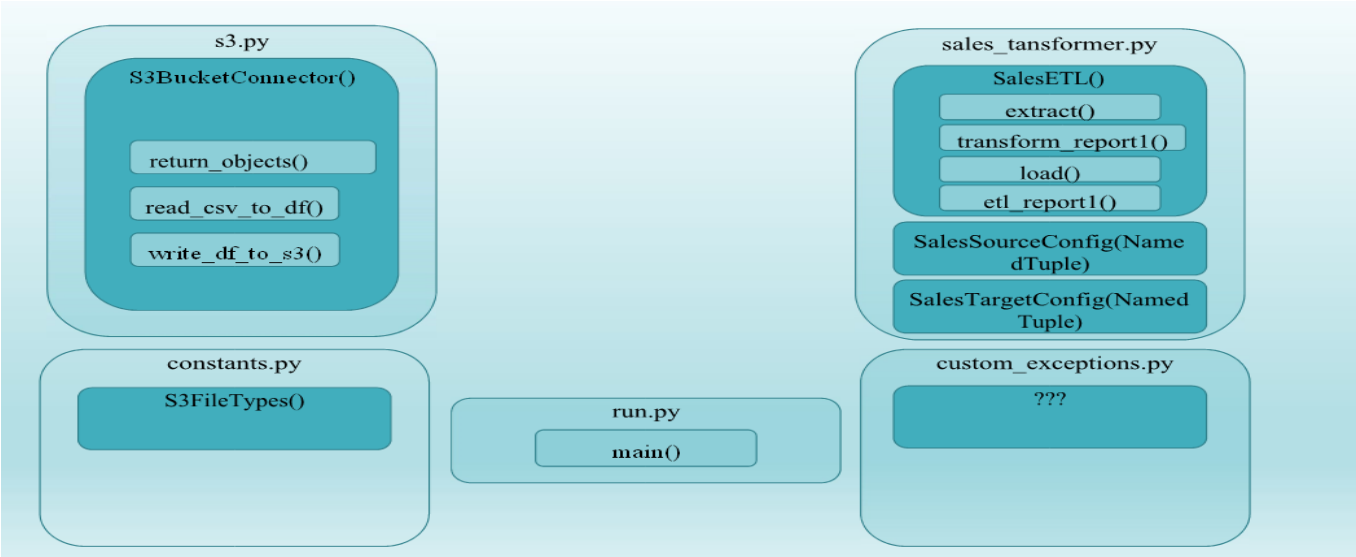
1.1 What is Low-Level design document?

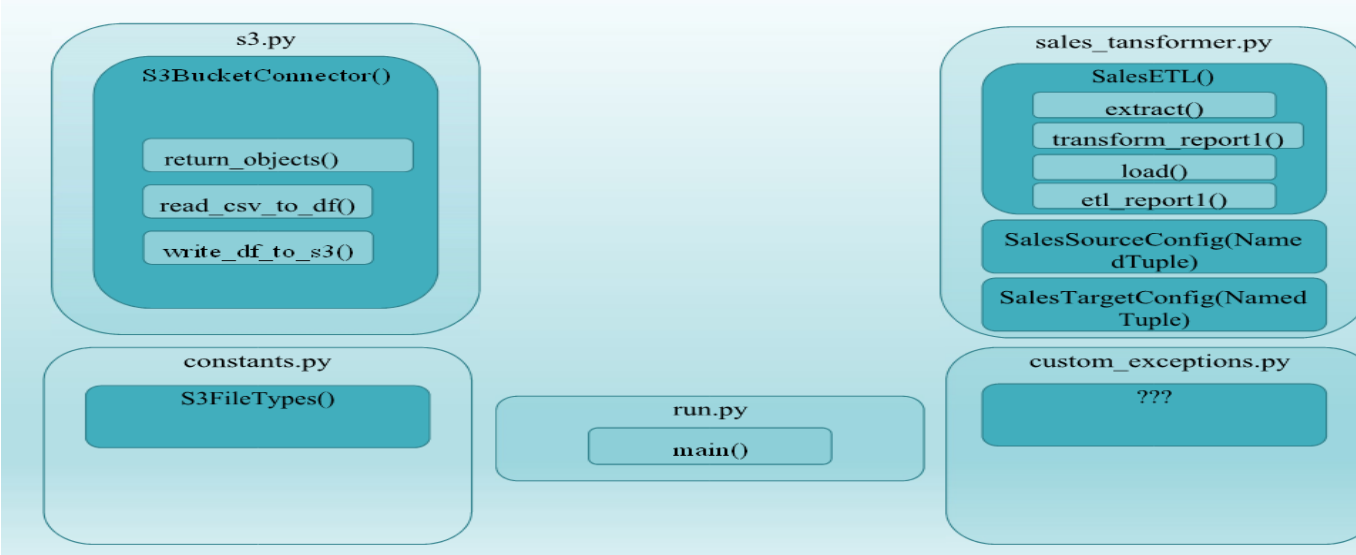
The goal of the LLD or Low-level design document (LLDD) is to give the internal logic design of the actual program code for the Amazon Sales Analysis Records dashboard. LLD describes the class diagrams with the methods and relations between classes and programs specs. It describes the modules so that the programmer can directly code the program from the document.

1.2 Scope

Low-level design (LLD) is a component-level design process that follows a step-by-step refinement process. The process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work.

2. Architecture





The ETL pipeline created in python

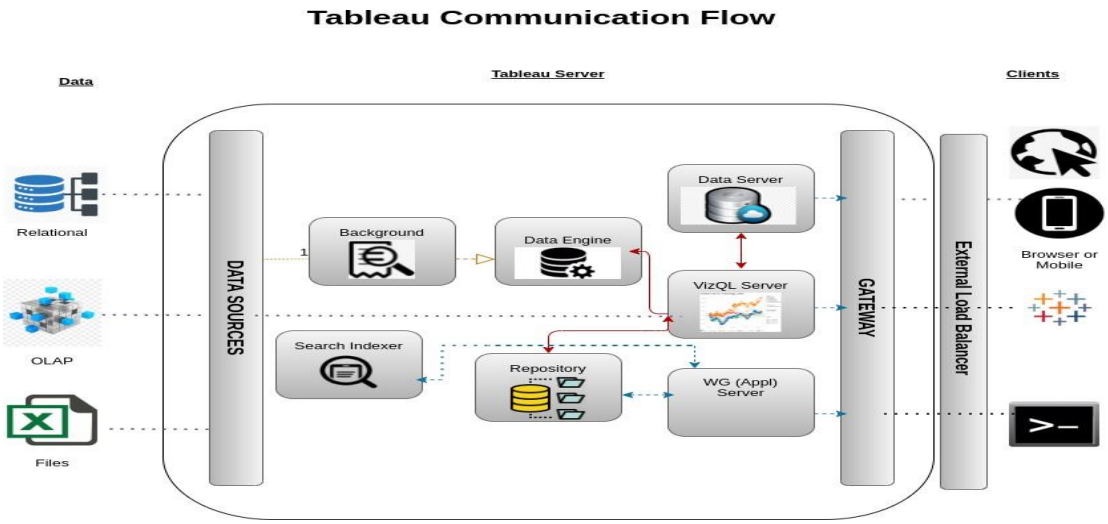


Tableau Server is internally managed by the multiple server processes.

1. S3 Bucket

A public cloud storage resource available in the Amazon web services.

2. ETL pipeline written in Python: -

Python script written in editor to perform Extract Transform and Load operations with python- AWS s3 bucket interaction tool like boto3.Unittest on each of the operations were performed using module like pytest. AWS Access and Secret key are stored as environment variables. Configuration file contains all the variable names.

3.AWS crawler

A crawler that is used to populate the Amazon Glue catalogue with tables. It can crawl multiple data stores in a single run.

4.AWS Athena

It is a service that enables data analysts to perform interactive queries in the web-based cloud storage service. It is used with large scale datasets.

5.Tableau online

Tableau is a visual analytics platform used in the business intelligence industry. Data analytics and creating interactive dashboards become faster-using tableau. Tableau connects to all kinds of resources like excel, JSON, databases. Tableau services can be accessed through tableau online, tableau server, tableau public as per the difficulty of the task.

Gateway/Load Balancer

- It acts as an Entry gate to the Power bi Server and also balances the load to the Server if multiple Processes are configured.

Application Server

- Application Server processes (wgserver.exe) handle browsing and permissions for the Tableau Server web and mobile interfaces. When a user opens a view in a client device, that user starts a session on Tableau Server. This means that an Application Server thread starts and checks the permissions for that user and that view.

Repository

- Tableau Server Repository is a PostgreSQL database that stores server data. This data includes information about Tableau Server users, groups and group assignments, permissions, projects, data sources, and extract metadata and refresh information.

VIZQL Server

- Once a view is opened, the client sends a request to the VizQL process (vizqlserver.exe). The VizQL process then sends queries directly to the data source, returning a result set that is rendered as images and presented to the user. Each VizQL Server has its own cache that can be shared across multiple users

Data Engine

- It Stores data extracts and answers queries.

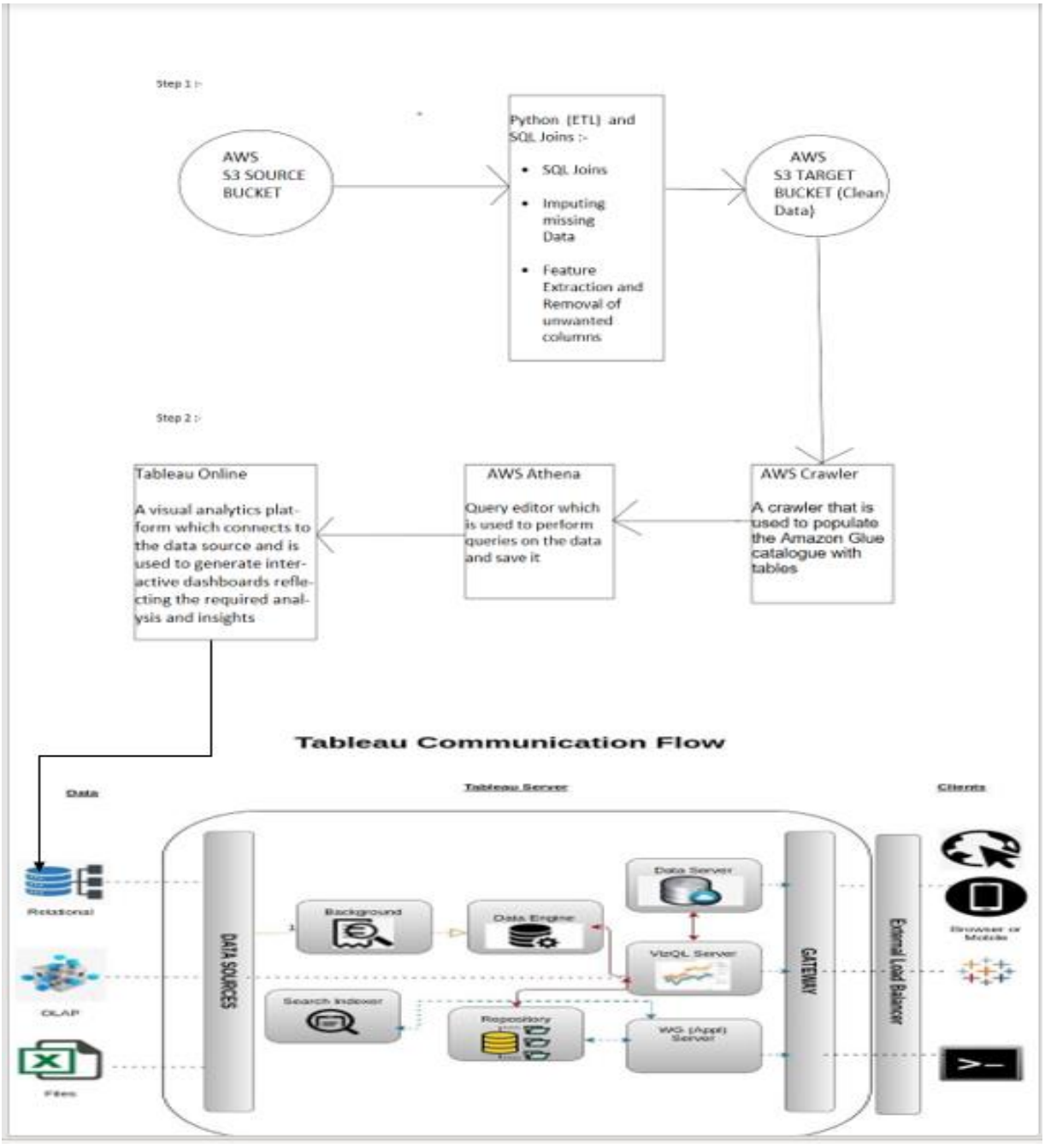
Backgrounder

- The backgrounder Executes server tasks which includes refreshes scheduled extracts, tasks initiated from tabcmd and manages other background tasks.

Data Server

- Data Server Manages connections to Tableau Server data sources
- It also maintains metadata from Tableau Desktop, such as calculations, definitions, and groups.

Total System Architecture



3. Architecture Description

3.1. Data Description

The S3 bucket contains 5 datasets. Each of them has certain attributes.

1. Region: Region contains regional division (Canada, Western, Southern,Northeast,Central,Northwest).
2. Sales Data: Sales Data information for different Product of the company.
3. Division: Division contains International, Domestic section.
4. Customer: contains Customer information data.
5. Customer address: contains information regarding Customer address.
6. GPS: Contains latitude longitude of Cities. This was done using geocode library.

3.2. Data Insertion into Database

a. Database Creation and connection – Create two S3 bucket database with name passed. One is the source bucket, and another is the target bucket.

b. Insertion of necessary files in the source bucket.

3.3. Extract Transform and Load operation

Step 1: Extract the datasets from the database to the ETL pipeline.

Datasets from the source S3 bucket were extracted as Byte objects. This is done using the Boto3. Boto3 is the name of the Python SDK for AWS. It allows you to directly create, update, and delete AWS resources from your Python scripts. The objects are then converted to Excel files.

Step 2: Transformation of the extracted excel files to a clean dataset

Here the excel files are joined based on SQL data table joining principle (left, right, outer, inner) to combine every attribute of each table into a consolidated data frame. Duplicated columns and rows were removed from the consolidated data frame. Missing Values were imputed using RandomSampleImputer, MeanMedianImputer and Categorical Imputer using Feature engine tool. Finally, all the locations in the dataset were converted to latitude and longitude features.

Step 3: Writing the data frame to S3 target bucket.

The cleaned dataset is then upload into the target bucket in the form of a comma separated file. Completion of the ETL pipeline is done.

3.4 Export Data from Database to Tableau

Step 1: Extract the dataset from the target bucket to the Tableau.

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run.

Athena is easy to use. Simply point to your data in Amazon S3, define the schema, and start querying using standard SQL. Most results are delivered within seconds. With Athena, there's no need for complex ETL jobs to prepare your data for analysis. This makes it easy for anyone with SQL skills to quickly analyze large-scale datasets.

Athena is out-of-the-box integrated with AWS Glue Data Catalog, allowing you to create a unified metadata repository across various services, crawl data sources to discover schemas and populate your Catalog with new and modified table and partition definitions, and maintain schema versioning.

Athena is connected to Tableau online or Tableau server via a JDBC Driver.

Step 2: Create Dashboard using Tableau.

Dashboard is created using Tableau.

3.5. Deployment.

Once you've completed your dashboard, follow these steps: - Publish, Save to Tableau Server As

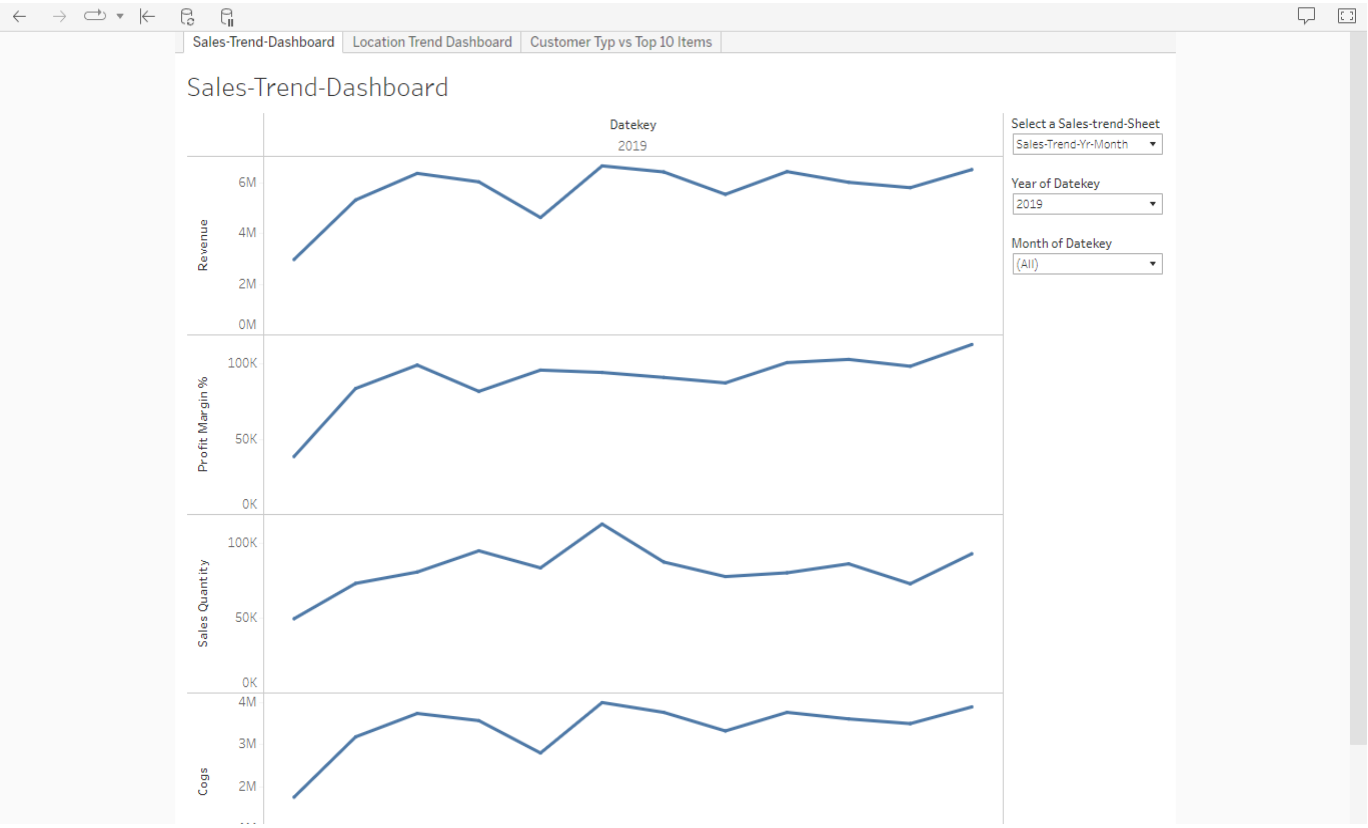
You may be prompted to log into your Tableau online profile first if this is your first-time publishing.

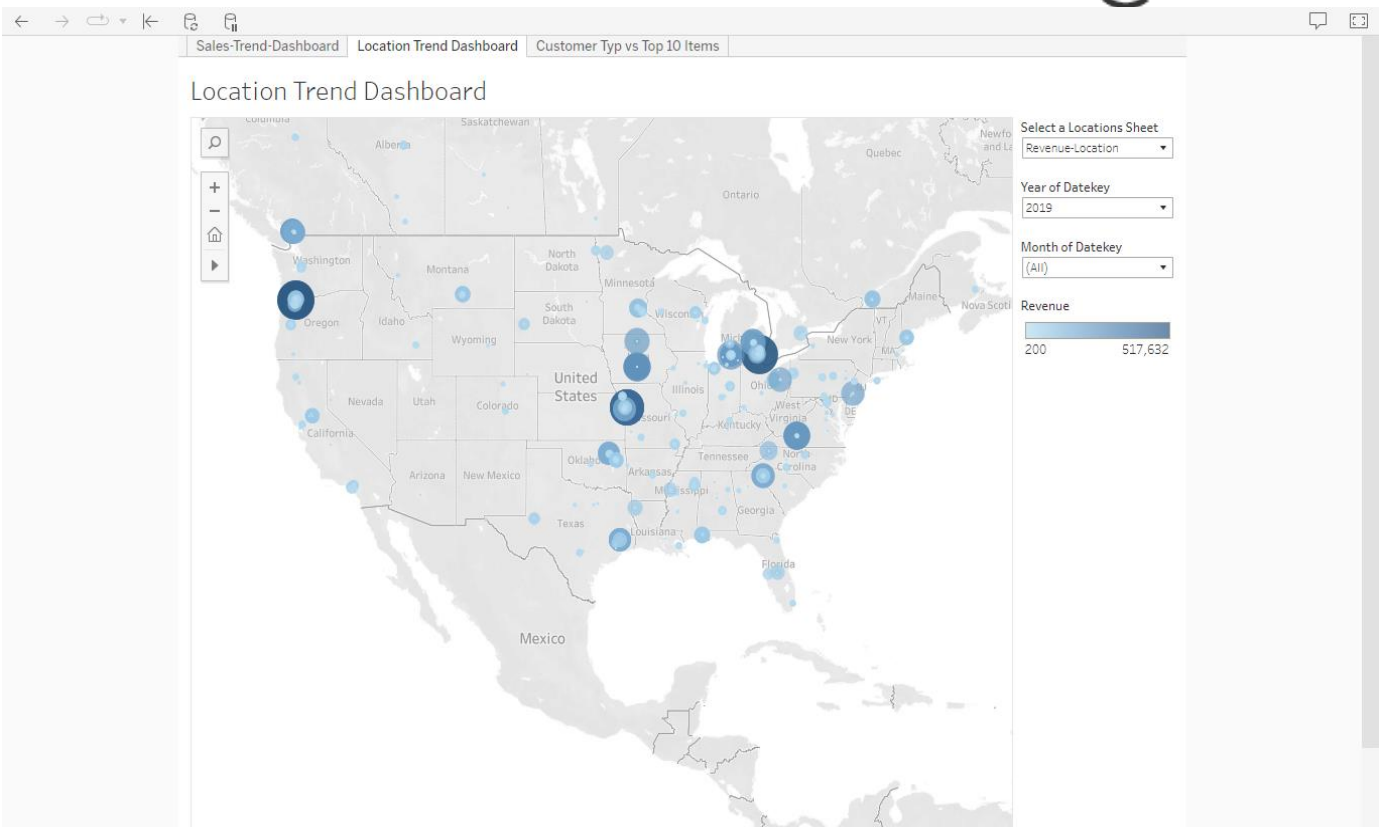
Next, fill out the title you want your viz to have and click.

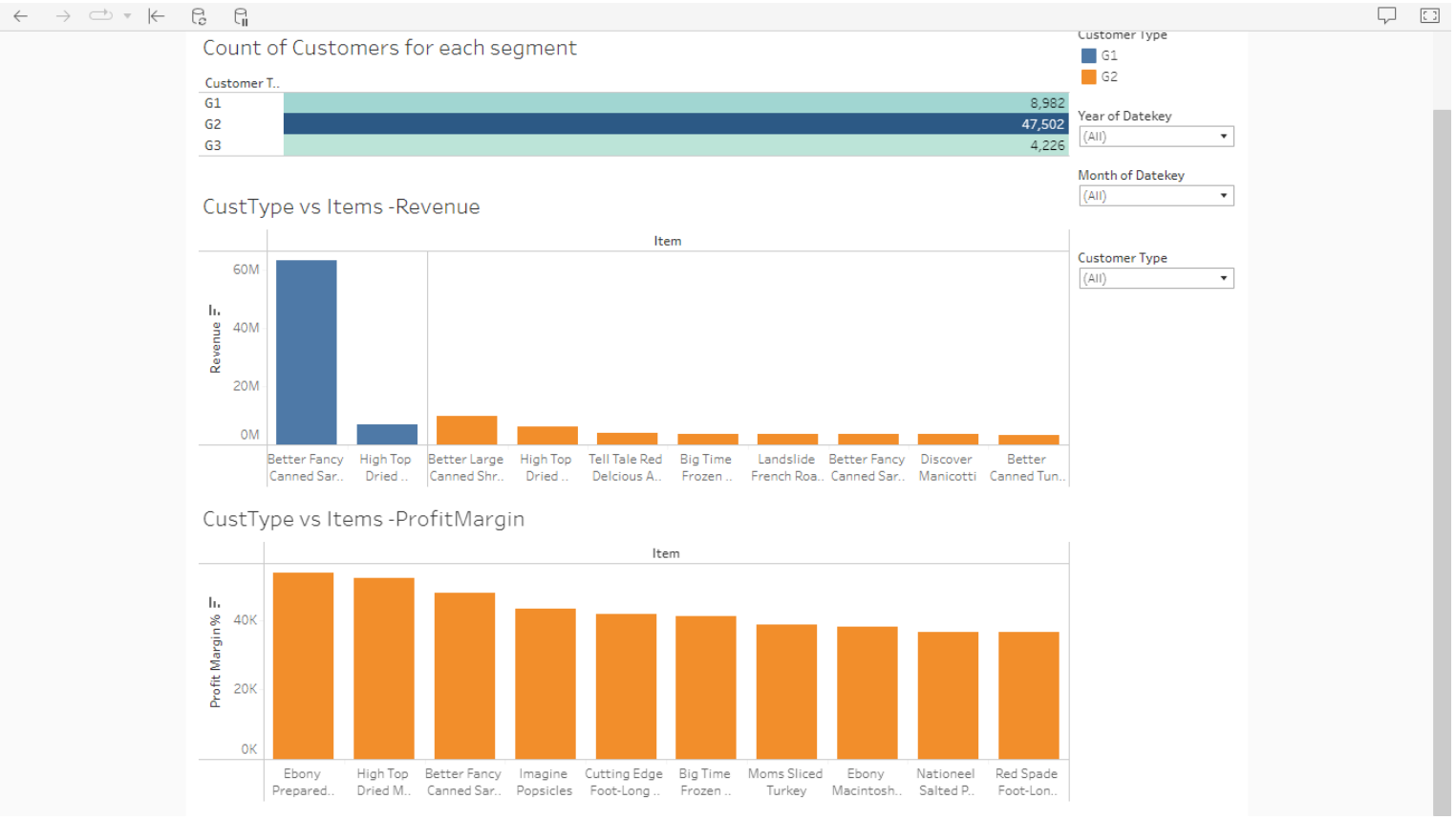
This message means that your connection to the dataset set is a live connection.

Here in the below screenshot, we can see that out workbook has been published to tableau online.

Tableau Dashboards




Sales-Trend-Dashboard
Location Trend Dashboard
Customer Typ vs Top 10 Items
Location Trend Dashboard
Select a Locations Sheet
Revenue-Location
Year of Datekey
2019
Month of Datekey
(All)
Revenue
200
517,632
United States
Mexico


Count of Customers for each segment
Customer T..
G1
8,982
G2
47,502
G3
4,226
Customer Type
G1
G2
Year of Datekey
(All)
Month of Datekey
(All)
CustType vs Items -Revenue
Item
Revenue
60M
40M
20M
0M
Better Fancy Canned Sar..
High Top Dried ..
Better Large Canned Shr..
High Top Dried ..
Tell Tale Red Delcious A..
Big Time Frozen ..
Landslide French Roa..
Better Fancy Canned Sar..
Discover Manicotti
Better Canned Tun..
Customer Type
(All)
CustType vs Items -ProfitMargin
Item
Profit Margin %
40K
20K
0K
Ebony Prepared..
High Top Dried M..
Better Fancy Canned Sar..
Imagine Popsicles
Cutting Edge Foot-Long ..
Big Time Frozen ..
Moms Sliced Turkey
Ebony Macintosh..
Nationeel Salted P..
Red Spade Foot-Lon..

4. Unit Test Cases

test_return_objects	returns all the objects present in the defined bucket
test_read_excel_to_df_ok	reads the objects from the source s3 bucket and returns a data frame
test_write_df_to_s3_csv	Writes the data frame to the target s3 bucket as a csv file.
test_extract_xls	Read the source data and converts them to one Pandas Data Frame
test_transform_report1_ok	Joining, Cleaning of the data frames happens
test_load_ok	Saves a Pandas data frame to the target bucket
test_etl_report	Extract, transform and load to and from S3 bucket