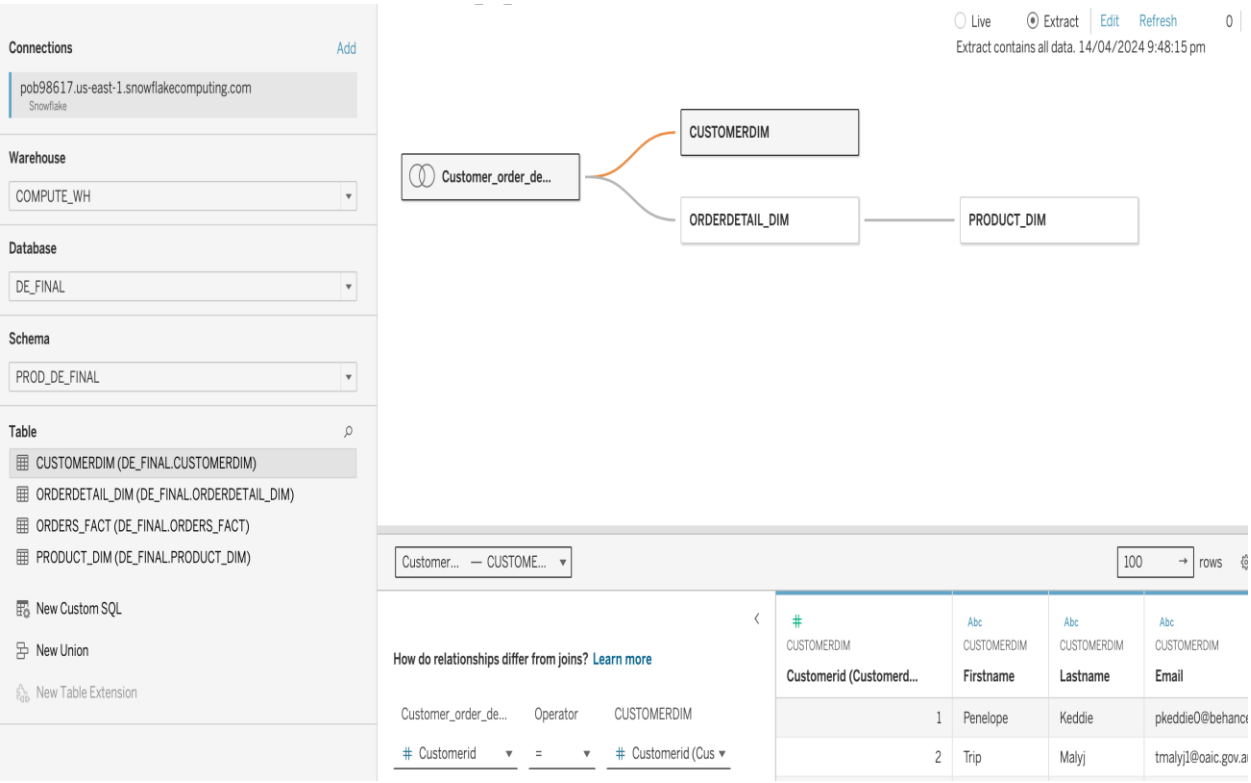
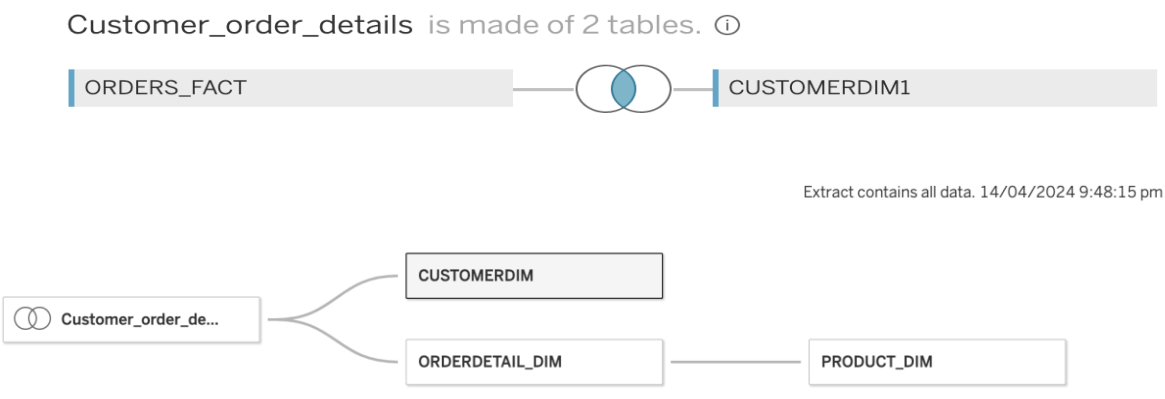
DE Final Deliverable Project Reflection

Final Deliverable

- Properly Connecting Data from **Snowflake** into **Tableau** environment.



- **Proper Joins in Tableau**



- **Dashboards-**

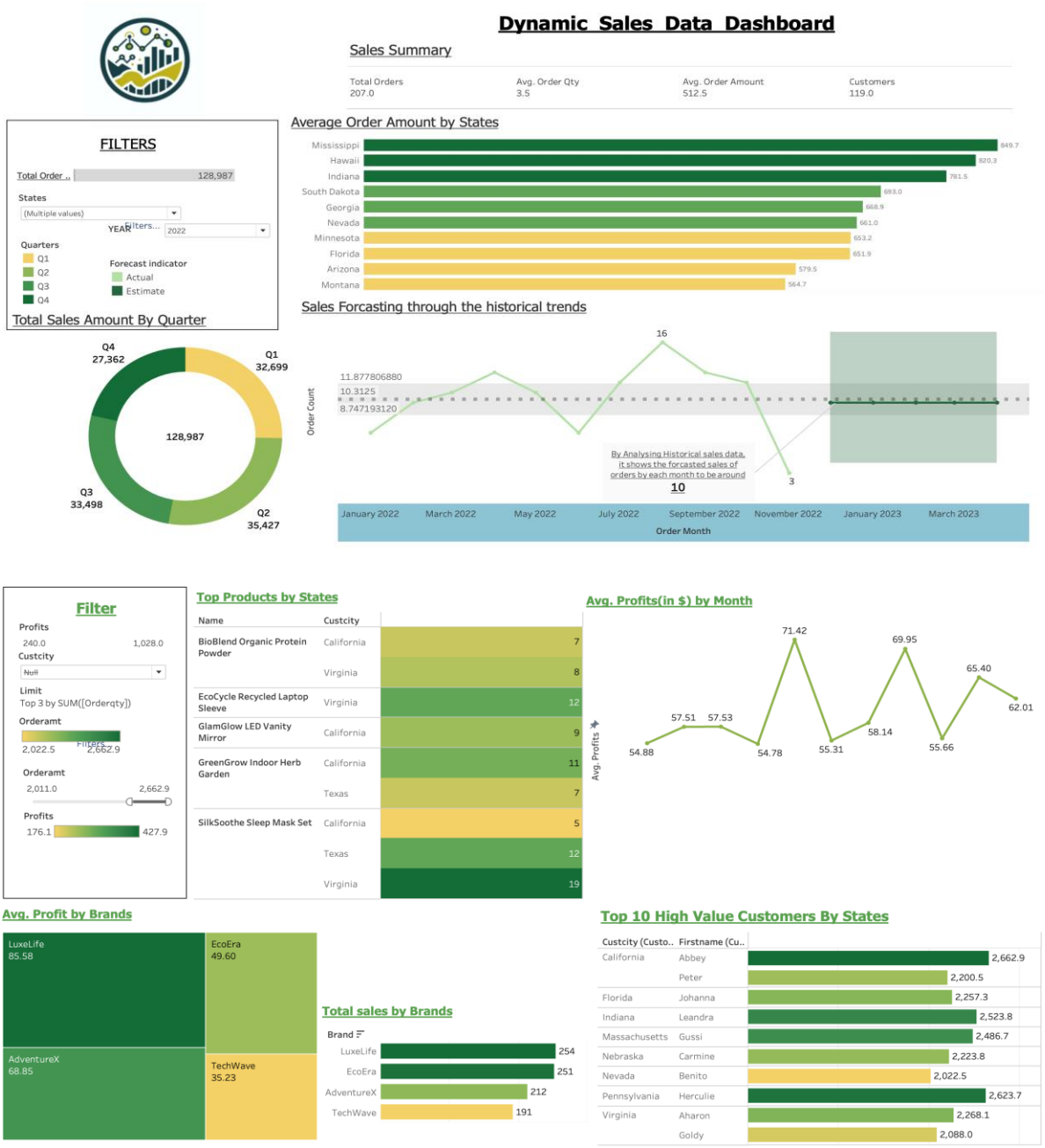


Tableau Project Link: Click Here

Project Reflection-

The project reflects the culmination of knowledge and skills acquired over my Data Engineering studies, integrated into a comprehensive project. The project involved several critical phases: data creation, ingestion, transformation, and visualization. By utilizing a variety of tools and platforms—including mockaroo, Salesforce, Airbyte, Snowflake, and Tableau—I was able to simulate a real-world data pipeline that would be applicable in a professional data engineering context. This reflection details each step of the project, the challenges faced, the solutions implemented, and the key takeaways.

Data Creation and Ingestion-

The initial phase of the project involved creating a robust dataset that could simulate real-world business operations. Using mockaroo, I generated structured data that represented business entities such as customers, products, and orders. This dataset was designed to mimic the complexities found in typical business environments, including varying data formats and volumes.

Once the data was created, the next challenge was to ingest it into a Salesforce environment. Salesforce served as the initial receiving platform for our data, acting as a CRM system might in a typical business setting. The ingestion process required careful mapping of data fields to Salesforce objects to ensure data integrity and usability.

Data Integration and Transformation-

After successfully ingesting data into Salesforce, the next step involved extracting this data and loading it into a Snowflake environment for further processing. This was achieved using Airbyte, an open-source ETL tool that facilitated the data movement from Salesforce to Snowflake. The choice of Airbyte was driven by its flexibility and broad compatibility with various data sources and destinations.

Within Snowflake, the data was organized into two schemas: Stage(STAGE_DATA) and Prod(PROD_DE_FINAL). The Stage schema was used for initial data reception and included raw data that might still contain inconsistencies or errors. The Prod schema, on the other hand, was intended for data that had been cleaned and transformed, ready for analysis and visualization.

The transformation process in Snowflake involved several tasks:

- Data Cleaning: Removing duplicates, correcting erroneous entries, and filling missing values.

- Data Formatting: Ensuring that all data adhered to a consistent format suitable for analysis.

- Data Transformation:Aggregating, summarizing, and restructuring data to support efficient querying and analysis.

The transformation logic was applied to four main tables: Customer Dim, Product Dim, Order Fact, and Order Dim. These tables were designed to support a dimensional modeling approach, facilitating complex analytical queries.

Data Visualization with Tableau-

The final step of the project was to connect the transformed data in the Snowflake Prod schema to Tableau, a data visualization platform. In Tableau, I developed an interactive dashboard that provided dynamic insights into the data through various visualizations.

The dashboard included different visualizations, each designed to highlight specific aspects of the data:

1. Customer demographic distributions.

2. Sales performance overtime.

3. Product popularity and profitability.

4. Order volume and patterns.

5. Geographic distribution of sales.

6. Interactive trend analysis using parameters and dynamic visuals.

Two actions were implemented to allow users to interact with the data dynamically: filtering options and data highlighters. The use of container usage in Tableau ensured that the dashboard was not only functional but also aesthetically pleasing and easy to navigate.

Reflections and Learnings-

Throughout this project, I gained invaluable experience in handling real-world data scenarios from creation to visualization. I was surprised by the complexity and intricacy involved in ensuring data integrity throughout the data pipeline. The project also highlighted the importance of choosing the right tools for each stage of the data lifecycle.

Looking forward, I plan to further explore advanced features of ETL tools like Airbyte and delve deeper into Snowflake's capabilities for data warehousing. The skills and insights gained from this project will undoubtedly enhance my future endeavors in data engineering.

Conclusion-

This project not only reinforced my data engineering skills but also provided a comprehensive understanding of the data lifecycle in a business context. It demonstrated the critical importance of each stage in the data pipeline and the interconnectedness of data creation, ingestion, transformation, and visualization. This project stands as a testament to my growth as a data engineer and my readiness to tackle complex data challenges in the professional world.