PROJECT REFLECTIONS

Presented by Keerthana Goka

Introduction

This project aims to gain practical experience and knowledge of the role of a Data Engineer along with the tools used in the current industry.

Data Engineering:

Data engineering is essential in the present industry due to the increasing importance of data-driven decision-making and the massive growth of data. It involves designing and building data systems to collect, store, and analyze data at scale. Data engineers ensure data quality, scalability, security, and compliance, forming the backbone of effective data management and analysis.

Tools used for the Project:

1. Salesforces:
    a. Salesforce is a powerful CRM platform that helps businesses manage their customer relationships, automate sales and marketing processes, and provide personalized customer experiences.
    b. As a cloud-based CRM platform, Salesforce stores vast amounts of customer data, including contact information, sales data, and customer interactions. This data can be used as a source for data engineering projects, such as data integration, data warehousing, and data analytics.
    c. Using Salesforce as a source system can provide several benefits, including access to rich customer data, integration with other systems, scalability, and security.
2. Snowflake:
    a. Snowflake is a cloud-based data warehousing platform that provides a fully managed service for data storage, processing, and analytics.
    b. It is designed to handle large volumes of structured and semi-structured data, providing fast query performance, scalability, and ease of use.
    c. Snowflake's architecture separates compute and storage resources, allowing them to scale independently and providing on-the-fly scalable compute. It also offers features such as data sharing, cloning, and third-party tools support, making it a popular choice for many organizations.
    d. Snowflake is often used as a destination system for data engineers, who use it to store and process data that has been ingested from various source systems.
3. Airbyte:
    a. Airbyte is an open-source data integration platform that extracts, transforms, and loads data from various sources to a target database or data warehouse.
    b. It is known for its ease of use, fast query performance, scalability, and extensive library of pre-built connectors.
4. Tableau:
    a. Tableau Desktop is a data visualization tool for creating interactive dashboards and reports. It allows users to connect to various data sources, clean and transform data, and create visualizations.

b. Tableau Public is a free version of Tableau Desktop that allows users to publish their visualizations on the web for public viewing. It has some limitations compared to Tableau Desktop, such as the inability to connect to certain data sources and the requirement to publish visualizations publicly.

Methodology:

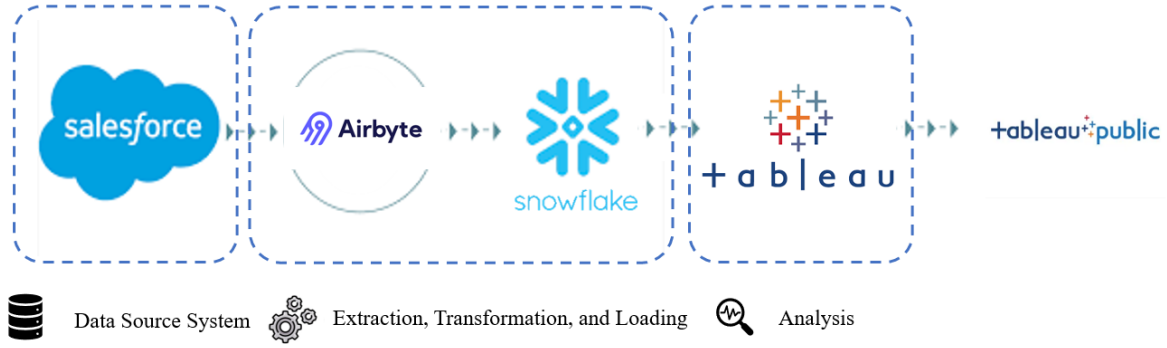
Data source:

CRM data of an e-commerce company for the year 2022 and 2023.

Data consists of 4 CSV files :

1. Sales – Fact table
2. OrderDetail – Dimension table
3. Customer – Dimension table
4. Product – Dimension table

Overview:



The project was implemented in three stages:

Project Part 1:

In Part 1 of the Project, by configuring the automated ETL tool: Airbyte, the CRM Data from the Source System: Salesforce was loaded into the Destination system: Snowflake.

- The CRM Data was loaded externally into Salesforce.
- Airbyte Connection was setup in Salesforce.
- In the Snowflake account, Database named MSDA3040_TERMPROJECT was created and schema STAGE_DATA was created in it.
- In Airbyte account, the required attributes of the salesforce CRM data was setup as a source system. The destination system was connected to the schema created in the Snowflake account.

Thus the CRM data was extracted from Salesforce by Airbyte and load into Snowflake.

Project Part 2:

In Part 2 of the project, the loaded data by Airbyte, into the staging schema in Snowflake, was cleaned, formatted, and transformed to a usable state. This data is later inserted in the Production environment schema.

Data Cleaning:

- In snowflake, using SQL, missing values in the customer dimension table were handled by replacing with NA.

Data Transformation:

- The abbreviated State names were transformed to full forms.
- Further, State names with 'NA' values were populated with appropriate state names based on the available zipcodes.

The cleaned and transformed data was pushed to the production schema with additional timestamp of the insertion.

Later a connection was established between Snowflake and Tableau Desktop for further analysis on the data.

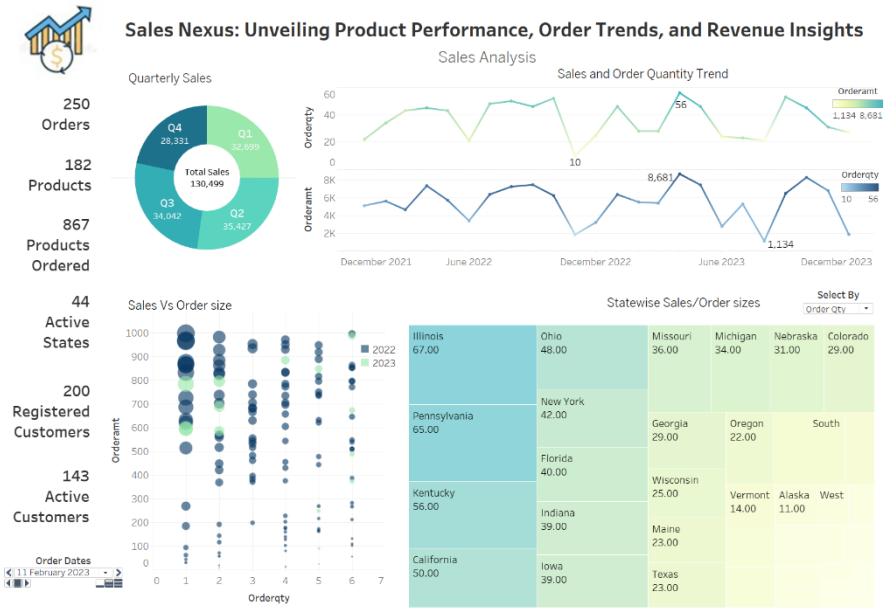
Project Part 3:

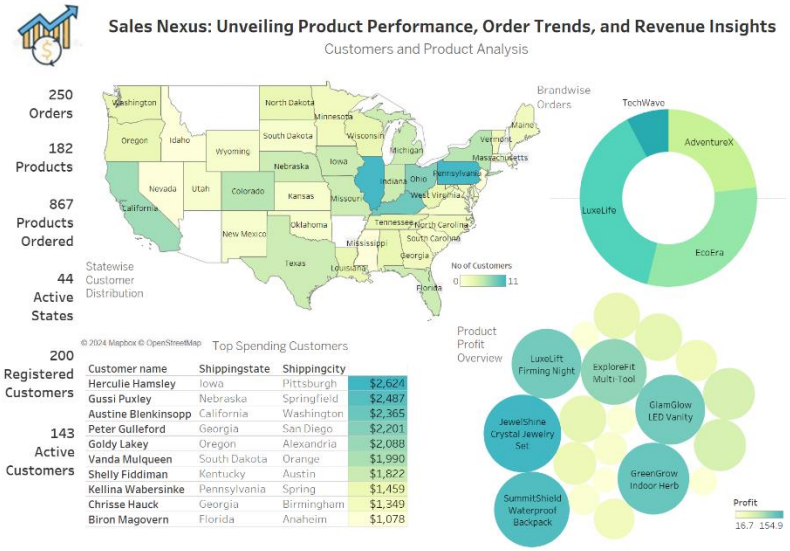
In the final phase, deep dive analysis was conducted using Tableau Desktop, and visualizations were published on Tableau Public. The link for the same:

https://public.tableau.com/views/DataEngineeringSales/Dash2?:language=en-US&publish=yes&:sid=&:display_count=n&:origin=viz_share_link

The analysis revealed insights such as the absence of seasonality in sales, the relationship between order quantity and order amount, and the sales/order sizes state-wise. The analysis also identified top-spending customers and profitable products.

Insights from the visualizations:





1. From the quarterly sales donut diagram, no seasonality can be seen as all the quarters have almost the same sales.
2. From the animated scatter plot of Order quantity vs Order Amount, it can be inferred that in general order amounts are higher when the quantity ordered is lower.
3. The trends of Order quantity and Order amount may be understood from the trend chart and further forecasting of sales may be done as required.
4. To understand the sales/order sizes state-wise, the treemap of Order amount/Order quantity may be utilized. While Pennsylvania customers gave the highest sales, customers in Illinois have placed the highest number of orders.
5. As part of customer and product analysis, for the customer distribution map, the concentration of customers may be understood, which may be utilized to select the locations for advertisement and such.
6. From the color formatted text table, top spending customers are identified, who may be targeted for further recommendations on discounts and such.
7. From the Brand-wise Orders donut chart, the distribution of sales of products in each brand may be analysed. This would be vital to understand brand performance.
8. The packed bubble chart of the profits prices of each product may help identify profitable products to focus on to increase sales.

Challenges:

The project faced some challenges.

- Including the presence of numerous null values in the data affected the visualization results.
- Additionally, since the data used was mock data, the insights derived from it did not reflect any real-world patterns or have any practical implications.

Conclusion

Overall, this project provided a comprehensive understanding of the data engineering lifecycle and hands-on experience with tools such as Snowflake and Tableau. The experience has given me the confidence to use any ETL and analytical tool in the industry.