Project Documentation and reflection

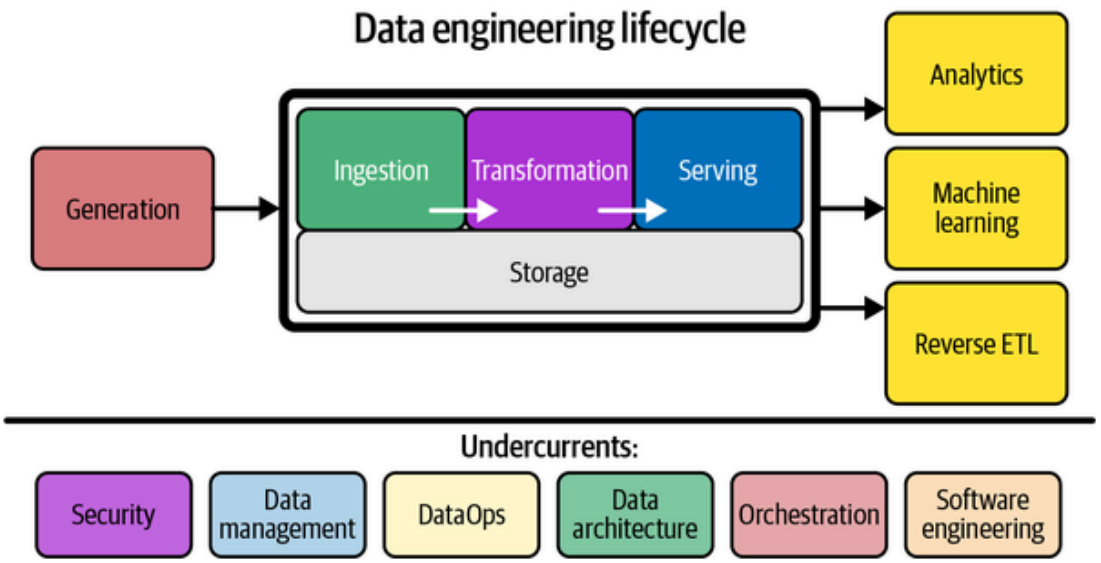
A new emerging start-up "**Sales-sparta Ltd**" has reached out to Clark students to develop a data pipeline so that the company can collect, store, and utilize data for essential business understanding. Their data engineering team is yet to be built, and they need a simple solution to scale for their sales team to collect and nurture leads, in this process, the management also wants to understand key business metrics to gauge which brands are doing well in real-time.

In this project, we will explore the different processes involved in creating a data engineering pipeline and serving the data for an analytics tool.

Understanding requirements and tools used in this project.

App Components	Technology	What is it?	Why it is used?
Analytics Solution	Tableau	Tableau is a powerful data visualization tool that allows users to create interactive and shareable dashboards, charts, and graphs from various data sources, helping businesses gain insights and make informed decisions.	Provides an easy way to create a dashboard to display metrics
Data Source	Salesforce	Salesforce is a leading cloud-based customer relationship management (CRM) platform that helps businesses manage their interactions with customers and prospects. It offers a suite of applications for sales, service, marketing, commerce, and more, all integrated into a single platform	It helps organizations manage customer data effectively, automate repetitive tasks, track sales leads.
Backend	Snowflake DB	Snowflake is a cloud-based data warehousing platform that allows users to store, manage, and analyze large volumes of data. It offers scalability, performance, and concurrency, enabling organizations to efficiently query and analyze their data for actionable insights.	Snowflake allows companies to establish both a centralized data repository as well as a mechanism through which queries can be processed and analyses can be conducted at speed and scale
Deployment	Airbyte	Airbyte is an open-source data integration platform that allows users to replicate data from various sources to a data warehouse or destination of their choice. It simplifies the process of data extraction, transformation, and loading (ETL), enabling organizations to centralize and analyze their data effectively.	Airbyte is used to streamline data by enabling users to extract data from different sources, transform it as needed, and load it into a centralized data warehouse or destination. This allows organizations to consolidate their data for analytics, reporting, and other business purposes.

Understanding the components and lifecycle involved



The data engineering lifecycle mainly involves

Generation
Data generation refers to the process of creating or collecting raw data from various sources, such as sensors, user interactions, or external databases. This step lays the foundation for the entire data project, as the quality and relevance of the generated data will directly impact the downstream processes.

Project scenario: Data generated are normally uploaded on an Excel sheet, this Excel sheet will be added to the master sheet. In this phase, the company faces a lot of challenges in channeling data, and most importantly speed and accuracy are reduced. Even though spreadsheets are quick and easy to use, when data scales all the undercurrents involved in the data engineering lifecycle are severely affected leading to un-localized data, the threat of stealing by other competitors, and loss of data and information always looms over.

To address this problem the company is recommended to use external software called Salesforce.

Storage
Data storage involves the secure and efficient storage of the generated data. This can include using databases, data lakes, or cloud-based storage solutions to ensure the data is readily available for further processing and analysis.

<u>Project Scenario:</u> Even if an employee uses Salesforce to upload leads or a sale, the data gets stored in the Salesforce cloud. It is recommended to use an on-premise or cloud to store the data so that, other data sources can be connected in the future to build a company's data lake. To store the data we have decided to go with "Snowflake".Through Snowflake security concerns can addressed by giving data access to certain users and restricting usage to some. In the near future data marts can be created in Snowflake enabling different teams to access data that is specific to the business unit.

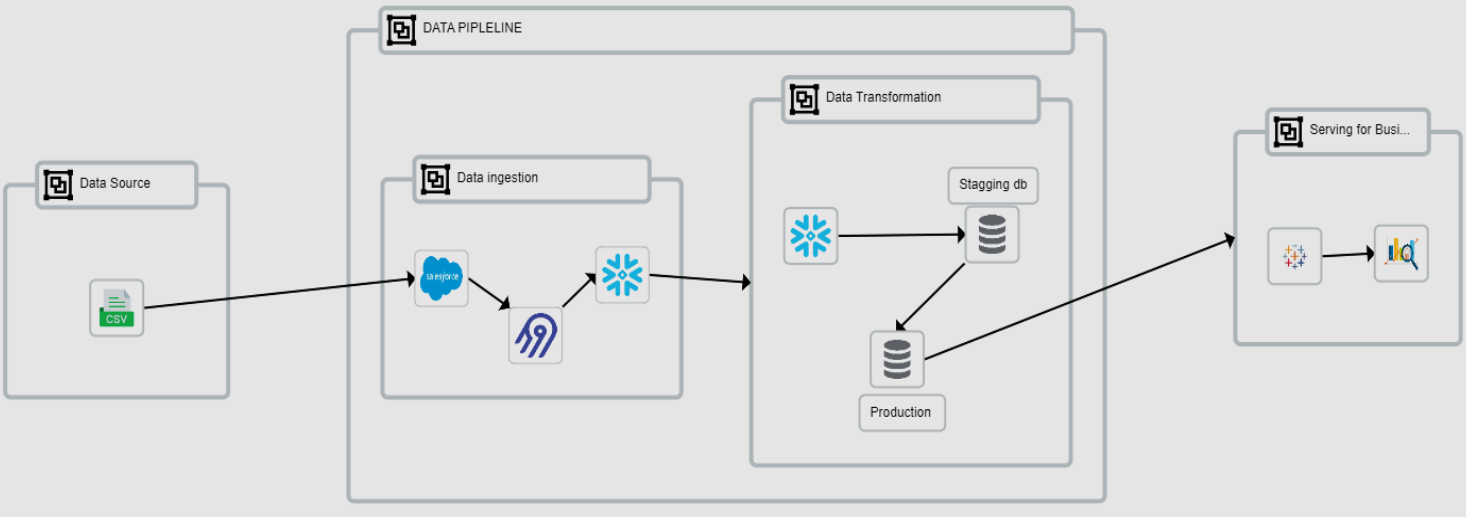
Ingestion
Data ingestion is the process of transferring data from its source into the storage system. This step ensures the data is properly formatted, validated, and integrated into the data infrastructure, preparing it for subsequent transformation and analysis.

Transformation
Data transformation involves the manipulation and processing of the ingested data to extract meaningful insights. This may include cleaning, filtering, aggregating, or enriching the data to align it with the project's objectives and requirements.

<u>Project Scenario:</u> In the Ingestion and Transformation phase the data from salesforce is sent to Snowflake. This is done through the low-code plug-and-play tool called Airbyte Integration. During this phase, an important question needs to be answered ie How long will it take for the data to reach Snowflake? We understood the data orchestration principles to be applied ie Streaming or Batch processing of data, the company requirement says they need data in real-time, so we have approached this problem by choosing batch processing/manual at a particular given time of the day, data is moved from salesforce to snowflake db.

<u>Transformation:</u> For the data to be loaded in Snowflake db, the DB has to be configured based on tables sent from salesforce. We have added to environments stagging and production environment. A staging environment is required so that any transformation or featurization needed can be done in stagging so that the end users will have clean and structured data to work on. We transformed customer tables as they had a lot of null values and few columns of information had to be standardized. After certain quality checks, the data is inserted in the production environment and is ready to be used by the end user.

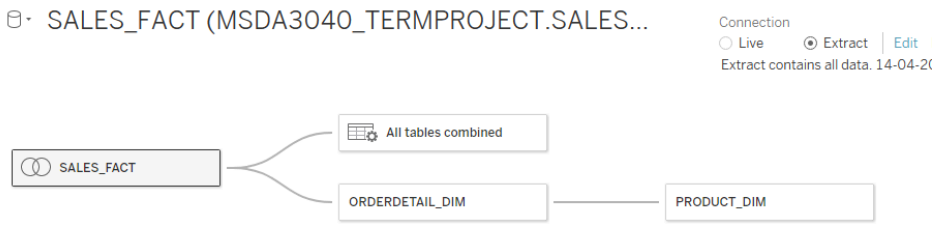
Here is the data pipeline architecture used for the project.



Serving Data
The final step is serving the transformed data to end-users, whether through dashboards, reports, or APIs. This ensures the data is accessible, understandable, and actionable for decision-makers and stakeholders.

<u>Project Scenario</u>: Post-copying data to a production environment, a connection is set up from Snowflake to Tableau. Post authentication, all the tables needed for creating analytics is served in tableau. In Tableau, the "extract" feature allows users to create a local copy of data from the original source, enhancing performance and enabling offline access. The "live" feature connects directly to the data source, providing real-time updates but potentially slower performance due to frequent queries to the source. So for faster analytics extract feature is used.

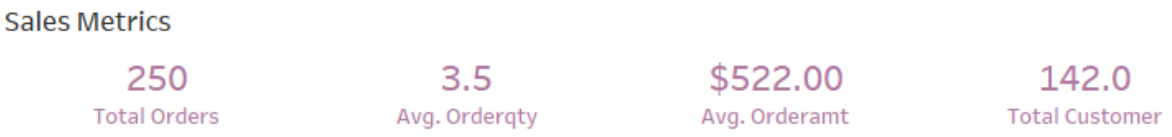
Tableau allows the user to create a logical layer on top of physical tables through this relationships between all the tables can be created. We have created tables that combines info or product level, a left joint table for orders and a customer table to get details orders at a customer level, order detail table, and product table respectively.



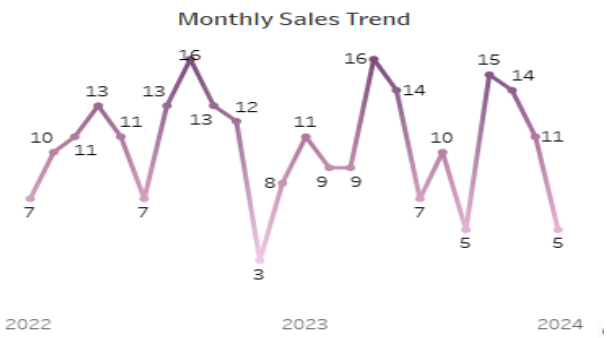
Analysis and Dashboard can be found here
https://public.tableau.com/app/profile/santosh.govardhan.kyathsandra.badarinath/viz/Sales-spartametrics/Dashboard1

Insights from Analysis.

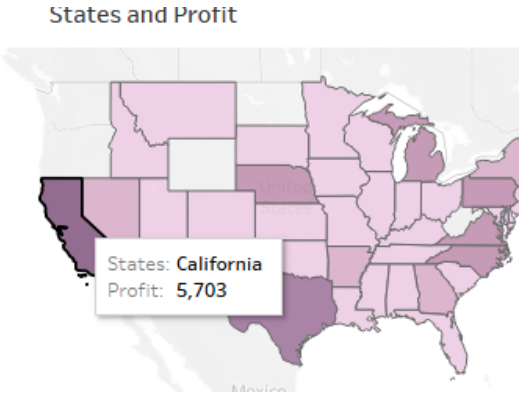
1. **A total for 250 orders have been created, with $522 being the avg order value.**



2. **Sales are highest in august 2022 and september 2023 are the highest with 16 orders each.**

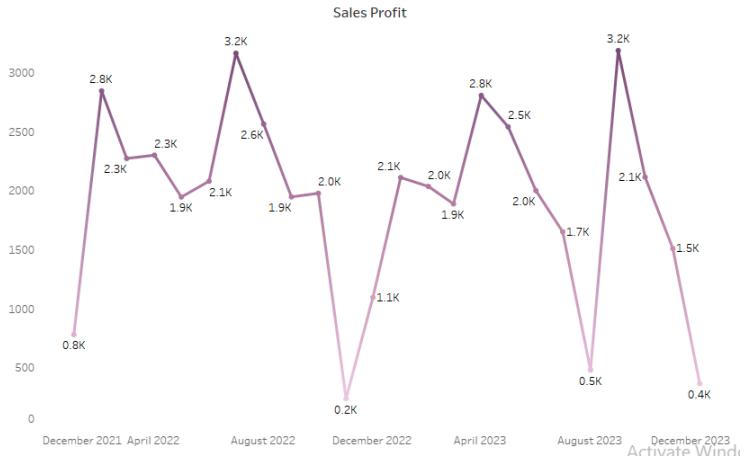


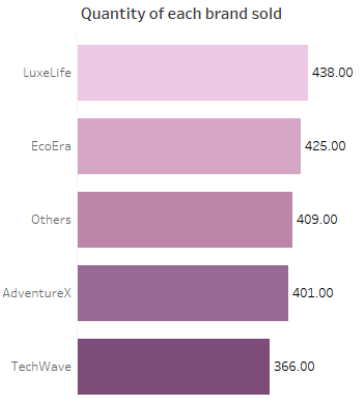
3. **Most of the sales happens in California and texas, majority of the profit comes from these states.**



4. **The profit margin is always positive as the cost involved in manufacturing is really low. A total profit of $28k is achieved by the sales team and company.**

5. **Luxelife brand is the most sold brand followed by Ecoera**





Challenges

The challenges faced during this process include:

1.Data Quality: The CSV files containing the data have a significant number of null values, indicating potential data loss and poor data management practices. Addressing this issue requires thorough data cleaning and validation processes.

2.Data Model Integrity: The predefined dimension table keys violate primary key rules, specifically regarding the foreign key relationships between fact and dimension tables. Ensuring data model integrity is crucial for accurate analysis and reporting, necessitating adjustments to maintain referential integrity.

Addressing these challenges involves implementing robust data cleaning, validation, and transformation procedures, along with refining the data model to adhere to primary key constraints and ensure data integrity throughout the pipeline

Tableau connection from snowflake