**Project Reflection**

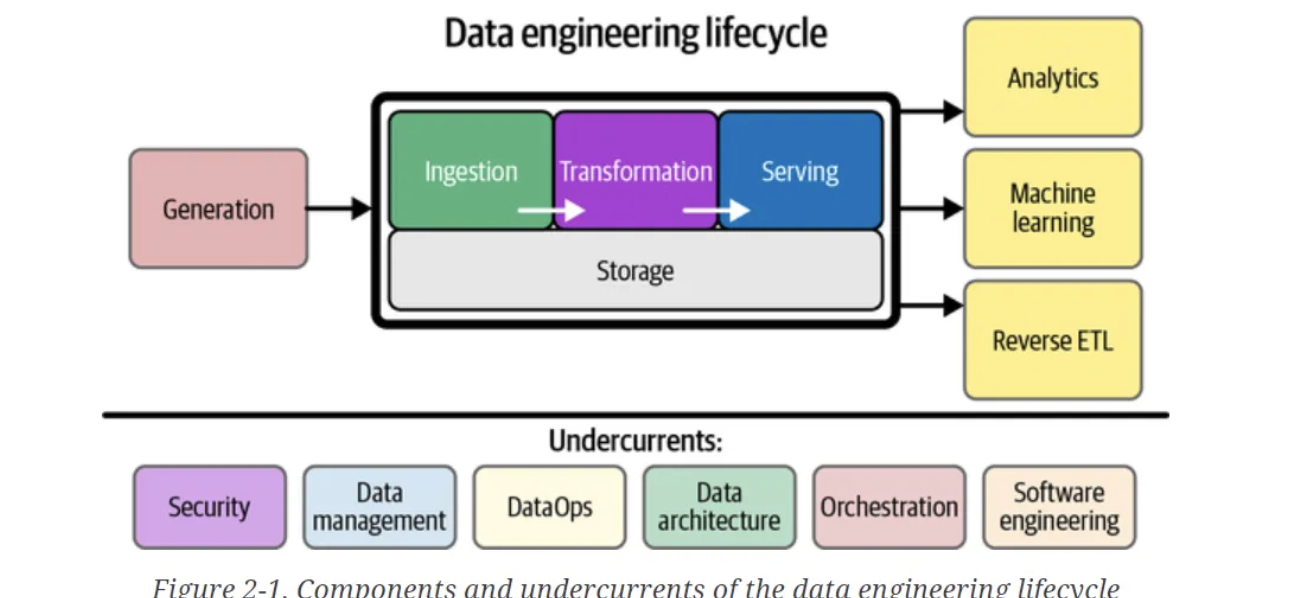
With a growing influx of data previously managed in Excel over the past two years, the emerging startup *Stadium Essentials* has sought assistance from Clark interns to develop a solution for efficiently storing, collecting, and reviewing data to support business operations.

Stadium Essentials aims to scale its sales data analysis on a yearly and quarterly basis. The primary objective of the project is to identify the top-performing brands and establish key performance indicators (KPIs).

As interns, our role in this project involves exploring various processes in data engineering, including ETL workflows, creating data pipelines, and preparing the data for analysis.

Below are the tools we would require for the Project

|  |  |  |  |
| --- | --- | --- | --- |
| Technology | Component | About the tools | Purpose |
| Tableau | Analytics tool | Data digitalization tool used to create interactive dashboards, charts, and graphs from various data sources, helping leaders derive insights and make informed decisions. | Self-explanatory dashboards to display metrics |
| Salesforce | Datasource | A leading cloud-based CRM tool that helps businesses manage customer interactions, offering customized applications for sales, service | Help customers manage customer data effectively, track sales leads |
| Snowflake | Data lake | A leading cloud-based CRM tool that helps businesses manage customer interactions, offering customized applications for sales, service, e-commerce, and various aspects of customer information, all integrated into a single platform. | A centralized repository as well as mechanism through which high volume of queries can be processed and analyzed at scale making the process smoother. |
| Airbyte | Connecter | It’s an open-source data integration tool that enables admins to replicate data from multiple sources to a data warehouse or other destinations, simplifying the ETL process and creating a centralized data source. | It is designed to simplify data processes by allowing users to gather data from various sources, transform it according to specific needs, and load it into a centralized data warehouse. This enables organizations to consolidate the data for analysis, reporting, and informed business decision-making. ns |



**DATA GENERATION**

In Salesforce, data generation involves creating or updating records such as leads, opportunities, and customer data to test workflows, automate processes, or validate system configurations. This helps ensure data pipelines are optimized for accurate analysis and seamless integration.

Sales, order, and customer data are typically recorded in PSF Excel sheets, which are then integrated into a master sheet. However, as the company continues to grow and expand its sales operations, it faces significant challenges in centralizing data due to the high volume of information handled in Excel. This has prompted the need for a solution that enhances the accuracy and speed of managing new sales data.

As data scales, the entire data engineering lifecycle is impacted, often leading to issues such as unstructured data, data loss, or leakage by third parties, competitors, or hackers. To address these security concerns, implementing robust software like Salesforce or Microsoft Dynamics is advisable. Given that the company is still expanding and lacks an existing office suite, Salesforce is recommended due to its ease of integration and flexibility.

Once customer sales data is uploaded into Salesforce, it will be securely stored in the Salesforce cloud. However, to support future business decisions, a complementary tool is necessary for cleaning and organizing the sales data. Additionally, SQL-based systems can be employed to store and analyze large datasets from various sources effectively.

**Why snowflake?**

Snowflake offers a pay-per-use model based on compute and storage consumption, providing a straightforward and flexible pricing structure. Unlike Microsoft Azure Synapse, which may have more complex pricing options depending on usage and resource allocation, Snowflake operates as a standalone cloud service.

For organizations heavily reliant on Microsoft products and services, Azure Synapse might be a more suitable choice. However, Snowflake is ideal for companies seeking multi-cloud flexibility and a simpler pricing model.

In the future, Snowflake allows for the creation of data marts, enabling different teams to access data tailored to their specific business units (BUs).

Once customer data is uploaded to Salesforce, a tool is required to store the data in a database. Options include SQL, MongoDB, or Snowflake, with Snowflake standing out as an efficient cloud-based storage solution.

**DATA INGESTION**:

Data ingestion is the process of extracting data from various sources and transferring it into a centralized system or database for storage and analysis. Tools like Airbyte automate this process, ensuring seamless and efficient data flow between systems such as Salesforce and Snowflake.

During the ingestion and transformation stage, data from Salesforce is transferred to Snowflake using a low-code, plug-and-play connector called Airbyte.

Given the large volume of data, one might wonder how long it takes for the data to move from Salesforce to Snowflake. This process can be explained using data orchestration principles, which include streaming or batch processing. Since Stadium requires real-time data to strategize for the Paris Olympics sales, they have opted for batch processing. By scheduling data transfers at a specific time each day, they ensure the data is effectively moved from Salesforce to Snowflake in alignment with their operational needs.

**DATA TRANSFORMATION**

Data transformation in Snowflake involves modifying, cleaning, and structuring raw data into a usable format for analysis and reporting. Using SQL queries and Snowflake's powerful processing capabilities, businesses can automate transformations to ensure data is accurate and optimized for decision-making.

To load data into Snowflake, we need to create tables with rows and columns, granting admin access based on the structure of the tables in Salesforce. Two environments have been set up: **staging** and **production**.

The staging environment is essential for performing any necessary transformations or customizations. This ensures that the admin has clean and well-structured data to work with before it moves to production.

During this process, we reviewed and transformed the customer table, addressing numerous null values and applying required standardizations. After conducting a series of quality checks, the data is loaded into the production environment, making it ready for use by end users.

**SERVING DATA:**

Serving data in Tableau involves connecting to data sources like Snowflake and creating visualizations that enable interactive analysis. Tableau provides a user-friendly interface to explore, filter, and present data insights for business decision-making.

After copying the data to the production environment, the next step is to establish a connection with Tableau.

Once connected, the necessary tables for analysis are made available in Tableau. The **extract feature** in Tableau allows users to create a local copy of the data from the original source, enabling offline access and improving performance. This feature is particularly useful for faster analytics.

In Tableau, relationships can be created between two or more physical tables. For example, a left join can be established between the order table and the customer table to obtain order details at the customer level, as well as between the order detail table and the product table for further analysis.

In this project, I learned about Airbyte and its application in data engineering, application of ETL and data engineering lifecycle in real life projects .

**Analysis and Dashboard** can be found here

Public URL <https://public.tableau.com/app/profile/archana.sahu/viz/TableauFinalProject12/Dashboard1?publish=yes>

Insights from Analysis

