Case Study Analysis: Data Science Lifecycle

Prepared by: Georges Assaf

Task 1

The company integrates multiple data sources to consolidate different data source to unify customer insights, optimize operations, and improve decision-making. This integration will help the data science team address key business questions like predicting customer churn, optimizing inventory, and identifying effective pricing and marketing strategies.

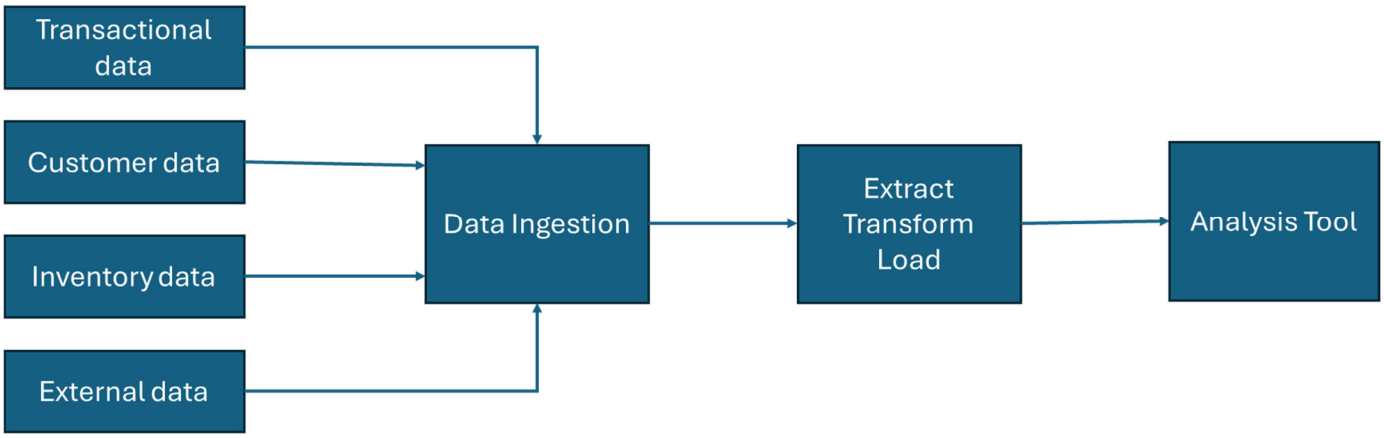
Data engineering supports these efforts by building scalable pipelines to aggregate, clean, and process data from various sources. It ensures data quality, consistency, and availability of data for analysis. This collaboration between data engineering and data science transforms raw data into actionable insights, driving growth and competitive advantage.

Task 2

Data Sources: The pipeline will handle data from different sources:

- Transactional Data: Point-of-sale (POS) systems and online transactions normally from SQL/Oracle databases
- Customer Data: CRM systems, loyalty programs, and customer feedback
- Inventory Data: Warehouse management systems /structured databases
- External Data: Market trends through API

The pipeline will the read the different data from different datasource, extract, transform and load into a centralized data warehouse for analysis.



Task 3

To clean and transform the data, start by assessing its structure and identifying issues such as missing values, duplicates, and inconsistent formats. Handle missing values through imputation, removal, and eliminate duplicates to maintain unique records. Correct errors and outliers by applying appropriate statistical or logical methods. Standardize field names, data types, and formats across all sources to ensure uniformity. Align units of measurement and encode categorical variables consistently. Validate the cleaned data by performing quality checks on ranges, relationships, and critical fields. Finally, document all transformations and prepare the dataset for seamless integration and analysis.

Task 4

The ETL process involves extracting data from various sources using tools like SQL for relational databases, API calls for third-party data, and Python libraries such as Pandas for file extracts. During the transformation phase, data is cleaned and preprocessed using tools like Pandas, ensuring consistency through techniques like missing value handling, data type conversion, and aggregation. After transformation, the data is loaded into target systems such as data warehouses like Google BigQuery, or data lakes like AWS S3, depending on the data structure. Considerations for efficient loading include using partitioning, incremental loads, and compression to optimize performance. Workflow automation with Apache Airflow ensures smooth execution and scheduling of the ETL tasks.