Ampera

Building the DEI edge

Problem Statement

A global Pharmaceutical giant wanted to create a Customer Data Platform (CDP) focused on Marketing data as a 1st milestone, encompassing all their sources of data (in-house, 3rd party and Market Research) across different markets to get a holistic view of their Campaigns' effectiveness

Scope of Work

It's a combination of Data Engineering and Data Sciences. There are 50+ data sources that need to be ingested, curated together to arrive at a single-source-of-truth. All these data points need to be transformed into a normalized and standardized data structure for the Marketing Business team's consumption. There are a set of dashboards to be created in MicroStrategy and Tableau for the Business stakeholders to start consuming these data points easily

Solution Approach

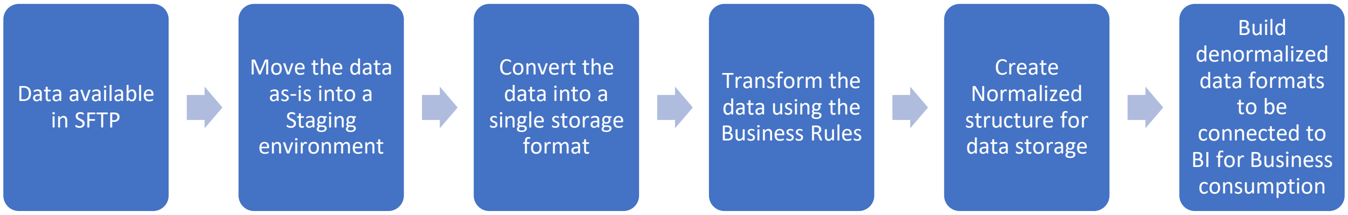
There are 2 kinds of data inputs –

1. Batch processed
2. Streaming

For Batch processing data sources, the data would be placed in the SFTP folder location. The data exchange framework set in place allows for specific data stewards to have access to place data sets and retrieve them from the NAS (Network Access Storage) drives. The data pipelines need to extract these data points/files and ingest them for further data transformations and storage

For Streaming data sources, the data would be extracted from the source in time intervals, loaded into Staging areas and transformed, pushing it for storage later. There were specific Service Accounts created that would be leveraged for data access via automated pipelines for data extraction. This would fall into the ELT protocol of data extraction, load and transformation.

High-level architecture –



Data available in SFTP

Move the data as-is into a Staging environment

Convert the data into a single storage format

Transform the data using the Business Rules

Create Normalized structure for data storage

Build denormalized data formats to be connected to BI for Business consumption

Technology Stack

- Amazon Web Services
    - S3 → Storage
    - Athena → Data Validation checks to be done on S3 Storage
    - Glue → Python script for file format & data standardization
    - Lambda Function → To check all the relevant parameters are in place and trigger the orchestration workflow
    - Step Function → Orchestration system for end-to-end pipeline visibility
    - Kinesis → Streaming data service for Extraction
    - Others like IAM, KMS, SNS, Glacier, CloudWatch etc.,
- Denodo
    - Data Virtualization Layer
- Snowflake
    - Data Warehouse
- MicroStrategy and Tableau
    - BI tools for Business consumption
- Liquibase
    - Change Management & CI/CD automation
- Collibra
    - Data Governance System
- Git & Confluence
    - Code Repo & documentation