MIS581 Capstone Project: The Dataset

Nicholas Laeder

Colorado State University - Global Campus

MIS581 - Data Analytics Capstone

Steve Chung

6/1//2025

The Dataset

The Capstone Project leverages a rich, real-world dataset derived from two prominent email marketing platforms: Constant Contact and Mailchimp. This dataset is critical to addressing the project's core objective of providing actionable insights to small business owners who currently lack sophisticated analytics capabilities for their marketing efforts. The selection of these specific data sources is rooted in their widespread use by small to medium-sized enterprises, making the project's findings and developed solution highly relevant and scalable.

The dataset is comprised of raw, granular data extracted directly from the Constant Contact API (V3) for Peer 1 and the Mailchimp Marketing API (v3.0) for Peer 2. Crucially, to ensure data integrity, privacy, and manage the inherent structural differences between these two distinct source APIs, separate dedicated tables are maintained for each peer's raw data. This approach means that Peer 1's Constant Contact data resides in its own set of tables, distinct from Peer 2's Mailchimp data.

For both platforms, the dataset includes several key entities and their associated attributes:

* Contact/Member Data: This includes subscriber details such as email addresses (nominal), first and last names (nominal), subscription status (e.g., 'subscribed', 'unsubscribed', 'active', 'bounced' - binary/nominal), and critical timestamps like creation or signup dates (interval/ratio). This data allows for audience segmentation and understanding subscriber lifecycle.
* Campaign Data: This encompasses details about individual email campaigns, including campaign IDs (nominal), titles or names (nominal), subject lines (nominal), creation dates (interval/ratio), and send times (interval/ratio). This provides context for analyzing campaign performance.
* Engagement/Report Data: This is perhaps the most valuable component, detailing how contacts interact with campaigns. It includes metrics such as opens (ratio), clicks (ratio), bounces (ratio), and unsubscribes (ratio), often linked to specific contacts and campaigns. Timestamps for these events (interval/ratio) are also crucial for temporal analysis. While some metrics like open and click rates are often summarized by the APIs, the goal is to obtain the underlying event data where possible to allow for more granular analysis.

The raw nature of this data, before extensive transformation, allows for maximum flexibility in analysis, preserving the "purity" of the information as it was captured by the source systems.

The analytical framework for this Capstone Project relies on a robust combination of cloud-native tools and established data analysis techniques:

1. Google BigQuery (Data Warehousing & SQL Analytics): BigQuery serves as the central data warehouse and the primary tool for data transformation and analytical querying. Following an ELT (Extract, Load, Transform) paradigm, raw data is first loaded into BigQuery. Subsequent data cleaning, joining, aggregation, and feature engineering are performed using BigQuery's powerful SQL capabilities. This allows for complex analytical queries to test hypotheses, identify trends, and prepare data for the next stage of analysis. Its serverless architecture ensures scalability for handling varying data volumes.
2. Python (ELT Scripting & API Interaction): Python is the core language for developing the Extract and Load components of the ELT pipeline. Libraries such as requests are utilized for interacting with the Constant Contact and Mailchimp APIs, handling authentication (OAuth 2.0), managing pagination, and gracefully handling API rate limits. The google-cloud-bigquery library is used to efficiently load the extracted raw data into BigQuery. Python will also be instrumental in preparing the data for the LLM and orchestrating the overall data flow.
3. Google Cloud Vertex AI (Large Language Model Analysis): Vertex AI, specifically leveraging its Large Language Model capabilities, will be employed for advanced qualitative and quantitative analysis. After data is processed in BigQuery, relevant insights and summaries will be fed into the LLM via carefully crafted prompts. The LLM will then synthesize complex data patterns, identify correlations, and generate human-readable narratives, including the "podcast-style" delivery, to make the insights accessible to non-technical business owners. This represents a novel technique for delivering business intelligence.

A comprehensive data dictionary and metadata strategy are integral to understanding and managing the dataset. While a literal visual model (like an ERD) is not presented here, the conceptual model for the raw data is directly reflected in the BigQuery table schemas. Each table within the raw\_data dataset (e.g., raw\_data.peer1\_constant\_contact\_contacts\_raw, raw\_data.peer2\_mailchimp\_members\_raw) is meticulously defined with explicit column names, data types (STRING, TIMESTAMP, INT64), and specific naming conventions that clearly indicate the source API and the associated peer. BigQuery's built-in table and column description features are utilized to embed metadata directly within the database, providing immediate context on data origin and purpose for future maintenance and analysis. This internal documentation is supplemented by external project documentation in Obsidian, detailing API endpoints, extraction logic, and transformation rules.

The selection of Constant Contact and Mailchimp email marketing data is driven by a clear problem statement and the potential for significant organizational benefits. Many small businesses, including those in my peer group, actively use these platforms but perform minimal to no analytics on the rich data they generate. This leads to missed opportunities for optimizing marketing spend, improving campaign effectiveness, and understanding customer behavior.

By focusing on this dataset, the project directly addresses this gap. It aims to demonstrate that even with limited resources, small businesses can leverage cloud analytics to gain a competitive edge. The data is sufficiently granular to allow for advanced analysis (e.g., correlating campaign types with enrollment trends, identifying characteristics of high-performing emails), moving beyond simple dashboard metrics. The project's success hinges on proving that valuable, actionable insights can be extracted from this readily available data, thereby providing a compelling proof of concept for a cost-effective, scalable analytics solution that can empower small business decision-making.