Automation of Corporate ID Data Validation Process for a Global Financial Services Company

INTRODUCTION

A major credit card service company with a presence in over 140+ countries and annual revenue exceeding 15 billion USD faced challenges related to inaccurate financial statements resulting from improper hierarchies of accounts. Maintaining these hierarchies became increasingly difficult due to spin-offs, mergers, and acquisitions, posing the risk of missing transactions or data breaches that could lead to financial losses for clients and service providers.

To address this issue, the clients aimed to establish accurate records of all corporate IDs (CIDs), each tied to the respective Organization Unit ID. The existing hierarchy maintenance process heavily relied on manual tasks, demanding substantial time investments, with analysts spending 2-3 days on a single client's case.

The central objective of the project was to enhance efficiency by automating the intricate data validation process, relieving analysts from manual tasks, and accelerating hierarchy maintenance. Analysts would play a crucial role in verifying the final output's accuracy before forwarding the CID list to their Client Managers (CLMs), thereby ensuring data integrity and client satisfaction.

OBJECTIVE

The main objective is divided into the following tasks: -

1) Identifying various parameters for CID validation.
2) Collecting spending activity data of each client from the client’s internal systems.
3) Formulating rules for identifying CID status and its client name.
4) Developing a Python script for automation
5) Testing of results

METHODOLOGY

The data validation methodology consists of 4 stages, as shown in the figure below:

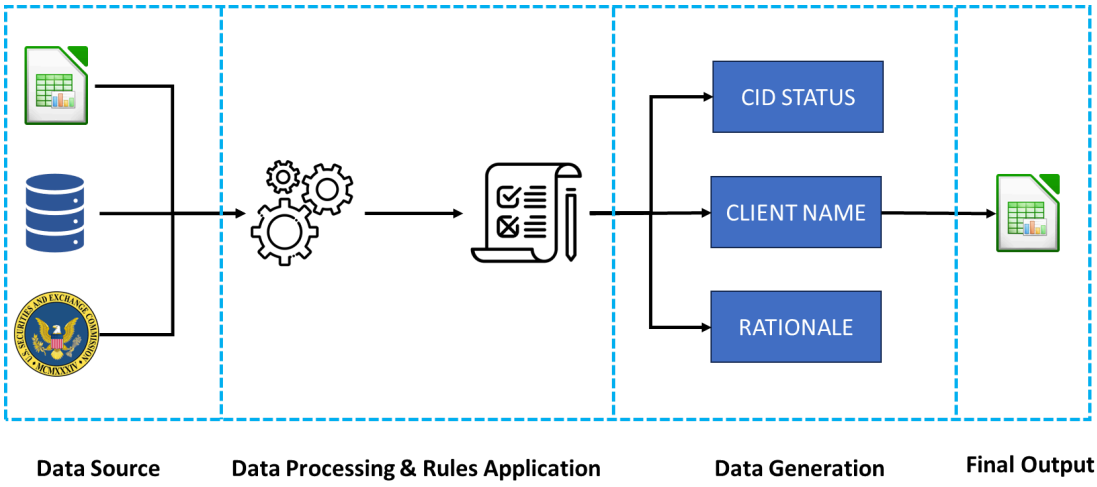

Figure 1: Flow chart showing the data validation process

1) **Data Source:** This involved collecting data from the Security and Exchange Commission website, internal repositories, and previously extracted CID data.
2) **Data Processing & Rules Application:** The collected data was then cleaned and processed to enhance the performance of string-matching algorithms. Data validation rules were subsequently applied to individual CID records.
3) **Data Generation:** This step includes the generation of CID Status such as “Active”, “Sunset,” or “Cancelled” along with CID’s organization name. A confidence score, ranging from 0 to 100, and rationales behind the status and organization name are also generated.
4) **Final Output:** Formatting the data and generating it in Excel sheets for review by analysts.

IMPACT

The automation of the CID Validation process significantly reduced the number of man-hours required. It enabled the successful data validation for all the major customers, accounting for 60% of the client's annual revenue. The automation greatly enhanced the overall efficiency of the department.

KEY CHALLENGES

There were many challenges involved in this project, including:

1) Spending activity records were not available for some clients.
2) Searching for the right data sources from repositories containing thousands of tables.
3) Determining the right combinations of rules for CID status.
4) String matching algorithms were not working correctly in some cases.