Capstone Project Summary

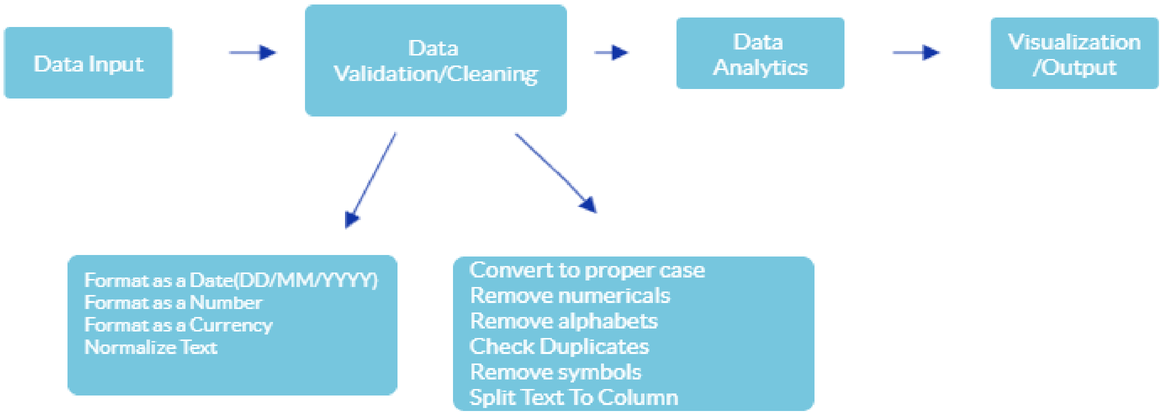
Development of a Data Flow Validation Process

This project focuses on the introduction of data cleaning algorithms or bots, coded within the client's infrastructure to clean and transform data in the manner the user desires. In most instances, data is structured in columns and rows.

a) For columns that contain numerical data, it is recommended to the client develop quick data fixes for the incorrect date, number and currency formats: Format as a Date, Format as a Number, Format as a Currency.

b) For columns that contain both alphabets and numerical, where either category is desired, it is recommended the creation of quick algorithms that work to implement these outcomes: convert to proper case, remove duplicates, remove numerals, remove alphabets, remove symbols and split texts into columns.

The Figure below exhibits the student's process mapping of the intended data flow validation process and the data cleaning algorithms proposed.



Source: Student's own

Outcomes

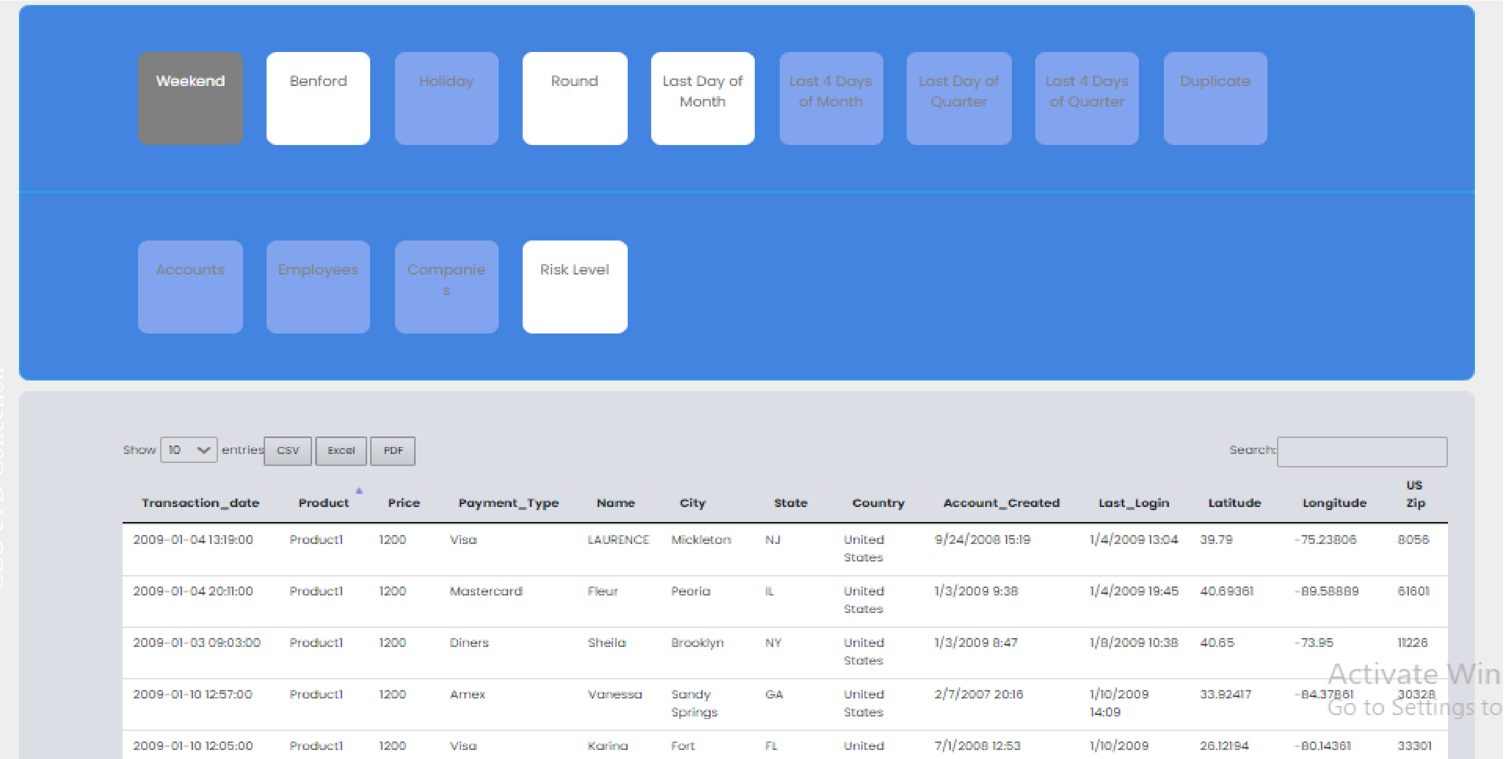
This proposal was pitched and recommended to the client's management and the software developers began working on it at the beginning of January 2021. As of the date of submission, these are the tools that have been implemented and tested.

a) The creation of a data upload module that recognizes csv, pdf and xls data format. The module also includes a search tool where the user may use to confirm certain entries before analysis is implemented.

b) The Client has developed the 'remove duplicates' algorithm. Numerous tests with different datasets confirm that it works to recognize duplicate numerical groups that include all amounts, bank accounts, personnel numbers, invoice numbers and cheque numbers.

c) The algorithms capture and flag transactions completed during unofficial business hours. Business hours are recognized the world over to comprise of weekdays between 8:00 am to 5:00 pm. Therefore, transactions done during weekends are suspect, unless there is official permission for a resource to do so. In particular countries, procurement laws oblige accounting resources to suspend operations at the end of the month or the last day of the fiscal year. The system can capture these anomalies.

The test below recognizes transactions made on the weekend. 3rd, 4th and 10th January 2009 were dates on Saturday, Sunday and Saturday respectively.



Transaction_date	Product	Price	Payment_Type	Name	City	State	Country	Account_Created	Last_Login	Latitude	Longitude	US Zip
2009-01-04 13:19:00	Product1	1200	Visa	LAURENCE	Mickleton	NJ	United States	9/24/2008 15:19	1/4/2009 13:04	39.79	-75.23806	8056
2009-01-04 20:11:00	Product1	1200	Mastercard	Fleur	Peoria	IL	United States	1/3/2009 9:38	1/4/2009 19:45	40.69361	-89.58889	61601
2009-01-03 09:03:00	Product1	1200	Diners	Sheila	Brooklyn	NY	United States	1/3/2009 8:47	1/8/2009 10:38	40.65	-73.95	11226
2009-01-10 12:57:00	Product1	1200	Amex	Vanessa	Sandy Springs	GA	United States	2/7/2007 20:16	1/10/2009 14:09	33.92417	-84.37861	30328
2009-01-10 12:05:00	Product1	1200	Visa	Karina	Fort	FL	United	7/1/2008 12:53	1/10/2009	26.12194	-80.14361	33301

Source: Screengrab from TraceRiser.com

Benefits to the client

This research on Data Flow Validation is integrated under the client's adoption of Artificial Intelligence and Machine Learning in its fundamental code. It seeks to complement the Risk Analysis technology that Client infrastructure focuses on. It packages the client as an All-In-One platform where the user is provided with a menu of auditing tools to clean, manage and transform data according to their criteria.

By implementation of this process, it puts the client's versatility above the existing auditing systems. In this way, the client uniquely packages its technology and brand by leveraging the power of innovating ahead of rivals and peers in the industry. The Data Flow Validation Tool stands to market and attracts demand for the software more than if it was absent.

Lessons Acquired from This Capstone Project

1. Product Development: I have gained substantial experience working with software developers of a startup. My role involved product development of audit software from the start to launching a demo version with no funding whatsoever.
2. The Danger of Moral Hazard: I have learned that the costs of using dirty data to make decisions in companies are very high, however, most firms don't realize the source of the error until it's too late. Correcting the error once a decision has been made is time-consuming and costly, adding to the already burden of financial loss. Therefore, I have learned that it is imperative for management to hire qualified data scientists and analysts to certify data before it's used for decision-making.
3. Corporate Consulting & Communication: This research has trained me on the approaches of consulting and communication with founders of startups and taught me to exercise patience with available scarce resources. A sense of collaboration and brainstorming as a team creates a synergy that greases working relationships and ensures a smooth flow of shared tasks.