**Data Collection and Management Map**

Srilasya Garigipaty

Student ID 100822953

School of Business, IT and Management

DATA 1201: Data Collection

Dr. Mina Labib

February 1st, 2021

**Table of Contents:**

ASSIGNMENT INTRODUCTION…………………………………………………………1

CONTENTS

1.What is the name of your organization?.............................................................................. 2

2. What business problem is your organization trying to solve?............................................2-5

3.Data Management Map Steps

* Data Generation…………………………………………………………………….. 5-6
* Data Collection &Transmission……………………………………………………. 6-7
* Data Pre-Processing………………………………………………………………… 7-8
* Data Storage………………………………………………………………………… 8-10
* Data Analysis……………………………………………………………………….. 10
* Decision Making……………………………………………………………………. 10-11

4.References…………………………………………………………………………………12

**Assignment Introduction:**

The purpose of the assignment is to create a Data Collection and Management Map for an organization of choice. Data Management consists of 6 stages. **Data Management** is defined as the end-to-end tools and processes used to integrate, govern, secure, and administer the transformation of source data into data that is in compliance with corporate and regulatory policies**.** The **6 stages of Data Management consist of the Data Value Chain. The stages are Data Generation, Data Collection, Data Transmission, Data Preprocessing, Data Storage, and Data Analytics.** **Data Generation** is to locate the source the data is coming from a trustful source and ensuring accuracy**. Data Collection** analyzes the 5 V’s of the data. The **velocity** of the data is the speed at which data is emitting data and creating changes. The **volume** of the data refers to amount of data being generated every second**. Variety** of data includes the various forms of data that is being created and consumed. **Veracity** is the quality and credibility of the collected data. **Value** is the business insights or solutions that can generated from the data. ((Sedkaoui, n.d.) **Data Transmission** ensures the communication of the data being is received and sent through a secure source. **Data Preprocessing** involves cleaning the data and dealing with errors**. Data Storage** is the most important step. Data storage is the persistent management of data in a scalable way that satisfies the needs of applications that require fast access to data. Lastly, **Data Analytics** determines the metrics being extracted, and what visualizations are needed to address the business problem. (Sedkaoui, n.d)

**Keywords:** Data Management, Data Value Chain, Data Generation, Data Collection, Data Transmission, Data Preprocessing, Data Storage, Data Analytics, Velocity, Volume, Veracity, Variety, Value



* **The organization name**: **Tangerine Bank**
* **Tangerine Bank** is a subsidiary of Scotia Bank, and it is an online bank.
* Tangerine, launched in Canada in April 1997, and introduced the country’s first no fee, high interest savings account.
* It was the first bank in Canada to offer the direct banking business model, and the **motivation** for Tangerine Bank was to offer more favorable rate compared to competing banks by avoiding the costs associated with running a network of branches.
* The **reason** I have chosen this organization is because it is new banking business model. With many industries turning virtual, the online banking business model has a large potential for growth. There is access to data that can be used to determine the customer satisfaction, and through the data available an analysis can be done to determine how customer growth can be achieved within the organization.

**Business Problem Organization is trying to solve:**

* Tangerine Bank has a growing customer base. It offers services such as savings accounts, chequing accounts, credit card services, mutual funds and mortgages. Majority of the customers at the bank are liability customers with deposits of varying sizes.

**Problem Definition:** The **number of customers who are borrowers (asset-based customers) is small**, and the bank is **interested in expanding** **its asset-based customers** to bring in **more loan business, and in the process, earn more revenue for the bank through interest on loans.**

* The retail marketing departmenthas decided to devise campaigns to better target their marketing and **increase the success ratio of loan borrowers and retain them as bank customers.**
* The department wants to build a model that will help identify the potential customers who have a higher probability of purchasing a loan. This will increase success ratio and **reduce the costs of the marketing campaign by only targeting those with more chance of success.**

**Details of the problem:**

* Tangerine Bank is not a big brand bank name amongst Canadians. Canadians trust the big bank names more such as Scotia Bank, Royal Bank, Bank of Canada, Toronto-Dominion (TD) Bank, HSBC Bank of Canada,etc.
* Since it is an online bank, there is a lack of face-to-face meetings with a bank agent, or manager (for example, if there are fraud issues or to speak to a loan officer)
* Slow Deposits as compared to ATM or big brand bank.
* Concerns about internet security.
* It is a longer process to get notarization, or bank signatures for loan approvals.
* The demographic is mostly younger, as most of the 40+ age prefer to use in-person branches.
* Due to these issues, the bank is making **less revenue** and not **reaching its potential, especially in pandemic times where virtual businesses have increased.**

**What the company has done to solve the problem:**

* **Business-Driven approach to big data:** business requirements are identified, and existing resources are utilized to increase marketing campaign.
* **Changes in the delivery of financial products and services.** With the bank services online, the company has tried to make the banking and financial transaction process handling **more convenient** for customers.
* **Compare current business model** with the big-name banks and implement areas of success into own development plan.
* Tangerine Bank collects large volumes of data, but the current model has considered a very small data sample. The **small sample used has created inaccuracies, incorrect future projections, and many assumptions which has led to poorer market performance**.

**Conclusion:** **Service quality** is the top reason customers switch banks. Banks have a lot of customer data, but effective and personalized service is lacking because there is a lack of big data analysis to **learn about customers and their expectations**. Tangerine Bank needs **to employ data analysts** and professionals in the marketing and finance field to **gain insights** from the large amount of data the bank collects. Tangerine Bank plans to **achieve greater customer loyalty** through **personalized offers.** The bank will **utilize the power of big data** to get an understanding of each customer. Through analytics process, Tangerine Bank **will reduce business risk by using predictive analytics and saving costs**.

**Process Mapping:** Using the process map located in the text (Figure 2 Big Data Value Chain - Chapter 1) as a guide, **please provide information on each** of the following steps. If a particular step does not apply, please explain why.

**Data Generation: What data do you need? Where are you going to collect the data from?**

Tangerine Bank will collect the data from the following information about the customer:

**Structured Data:** Data with **a high degree of organization**. It is located in a relational database and easily searchable.

|  |  |
| --- | --- |
| Customer ID | Account Number |
| Age | Numerical number |
| Job | (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown') |
| Marital | marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed) |
| Education | (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown') |
| Default | has credit in default? (categorical: 'no', 'yes', 'unknown') |
| Balance | Average yearly balance |
| Housing | (categorical: 'no', 'yes', 'unknown') |
| Contact | contact communication type (categorical: 'cellular','telephone') |
| Day | last contact day of the week (categorical: 'mon','tue','wed','thu','fri') |
| Month | last contact month of year (categorical: 'jan', 'feb', 'mar', 'nov', 'dec') |
| Duration | last contact duration, in seconds (numeric). |
| Campaign | number of contacts performed during this campaign and for this client (numeric, includes last contact) |
| pdays | number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) |
| previous | number of contacts performed before this campaign and for this client (numeric) |
| poutcome  Deposit | outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')  Has the client subscribed to a term deposit? “Yes” or “No” |

|  |  |
| --- | --- |
| Loan | Has a personal loan? “Yes” or “No” |
| Account Transactions | Number of Banking Transactions |

**Unstructured Data:** Data that **does not come in specific shape or form**. It can be analysed for **content and sentiment**, rather than slotted in a database table and analysed.

|  |
| --- |
| Customers’ Feedback/experiences |
| Article/Blog Engagement |
| Phone Call Audio Recordings with Customers |

During **Data Generation**, data is **gathered, and selected based on business goal.** The data should be **verified for accuracy**, and the **source** should be examined to make sure it is **reliable. High transaction volumes** of information take place, and the generated data can support either one or many data types to get as many insights as possible **without bias.**

**Data Collection & Transmission: What overall process will you use to collect the data? What format will the data come in? Are there any special considerations in transmitting the data to the storage center?**

In **Data Collection and Transmission**, data is **collected from various formats and communicated in secured way to the data storage.**

Tangerine Bank will collect **the structured data in an SQL (relational) database server**. The **Unstructured Data is collected in NoSQL(non-relational) databases and MongoDB.**

Tangerine Bank’s **internal financial records** can be used to get data on:

-Customer ID

-Customer Defaulted Credit

-Customer Yearly Balance

-Customer Deposit

-Customer Loan

-Account Transactions

Tangerine Bank’s **internal marketing and customer department records** can be used to get data on:

-Customer Job Type

-Customer Marital Status

-Customer Education

-Customer Housing

-Customer Contact Information

-Marketing Campaign number of contacts with customer, duration of last contact, outcome, and days passed since last contact.

**Customer Service Feedback/ Experiences** can be collected **through a survey published on the Tangerine Bank website, or Online App.**

**Article/Blog Engagement** from **Bank Website, and online bank related blogs such as Tangerine Bank blog, Reddit, Quora, etc.** can be used to learn about the feedback, comments, bank- related topics of interest, and customer satisfaction in different areas of service the bank office.

**Phone call Audio Recordings and Transcripts** with customers can be accessed through the **organization’s customer service records.**

**The file format for Audio files** will be **WAVE, AIFF, MP3, and MXF**. The **database comes in various formats in CSV, XML and Tab formats.** The **transcript format which will be in HTML and TXT.**

Tangerine Bank uses **256 bit AES advanced encryption standard** for storing the databases. **Encryption** is done to **protect the customer’s privacy** and **prevent the leak of confidential information of the customer**. The data collected that is sent to the data storage will be worked on by only **authorized personal** who are responsible for maintaining the databases by the organization.

The bank data **transmitted has large volume and very high velocity.**

**Data Pre-Processing: What pre-processing will you need to do on the data? Integration? Cleaning? Elimination of Redundant Data? Error Correction? Try to be as specific as possible.**

In **Data Pre-processing**, the data is transformed from data that is inconsistent, contains errors, or is incomplete into quality data that is consistent and reliable.

The **Major Tasks of Data Pre-Processing** includes **Data Cleaning, Data Integration, and Data Transformation.**

**Data Cleaning:**

* Data is identified for missing values, and noisy data. This can be done through combining human and computer inspection to detect any suspicious values.
* Inconsistencies such as repeated, or incomplete records are corrected. Inconsistencies can be due to the data entry problems, data transmission problems, or technology limitations due to improper software, slow processing.
* Creating a unified date format.
* Metadata is reviewed according to business problem and target feature.
* Converting nominal data to numeric data. Nominal data is “labeled” data which can be divided into various groups.

**Data Integration:**

* Combining the bank data from the various sources into coherent storage.
* Schema integration: bank metadata will be integrated from different sources.
* Redundant data occurs from the integration of multiple databases. The same attribute can have different names in different databases ex. Cust ID and Cust #. Redundant data can be detected through correlational analysis.
* Careful integration of the bank data from multiple sources may help reduce or avoid the redundancies and inconsistencies and improve the mining speed and quality before building model.

**Data Transformation:**

* Transform the data from their file format into one format, all in SQL database or NoSQL/MongoDB database.
* Convert the column values into same currency, date, character format, numeric categories, etc.
* In SQL Server, based on their storage characteristics, some data types are designated as belonging to the following groups:

-Large value data types: **varchar(max)**, and **nvarchar(max),**

**-**Large object data types: **text**, **ntext**, **image**, **varbinary(max)**, and **xml**

**Data Storage: What kind of storage system do you need? Which parts of ACID are required? How will you handle the CAP theorem?**

-**The data will be stored in a complex database or set of databases known as a banking core. Banking core** is a **centralized system** established by banks which allows its customers to conduct their business irrespective of location. Thus, it removes the barriers of branch specific transactions.

-**Banking Cores** have a very high degree of security, multiple verifications, and can be backed up to external backup servers or hard drives as well.

**-Banking Cores** are usually **located on in the cloud** and retain client data, perform transactions, and allow the bank to close each day/month/ week/year and automatically calculate balances and statements on transaction information.

**ACID property** is an essential part of an organization. **ACID stands for atomicity, consistency, isolation, and durability.**

**Atomicity:**

All changes to the data are performed as if they are a single unit.

-When a bank customer is inserting their deposit into both their checking and saving account, both accounts will have the money credited.

**Consistency:**

Data is in a consistent state when the transaction starts and also when it ends.

-When a transfer of funds from a savings to a checking account occurs, the consistency ensures that the same amount of transfer is credited and debited in the account.

**Isolation:**

The intermediate state of a Transaction is invisible to other transactions.

-When transfer funds are sent from one account to the other, isolation ensures that the other transaction sees the funds in one account or the other, but not both accounts concurrently.

-When a customer is filling out a satisfaction survey on mobile app, the individual must fill out all details and complete the form before bank can receive the results.

**Durability:**

After a transaction is successful, changes to the data are permanent and unable to change.

-When the transaction is completed, the transaction processing system ensures that the changes made to the data are permanent and visible to other transactions.

**CAP Theorem:** CAP theorem comprises of three components. The three components are related to distributed bank data storage. The three components are Consistency, Availability and Partition tolerance.

In **normal banking operations the data storage provides all the three components.**

The **banking transactions should be fully consistent**. In case of network failure, or server maintenance access can be disrupted. However, **availability to customers and to the databases is very important** and availability should be there.

It is very important to maintain **high consistency and availability**. Financial Transactions are **crucial services.**

. **Partition tolerance** is needed so that even if the server is down in one cloud location, it is still available on backup or in multiple locations on the server. Customer bank account balances and messages are consistent database.

**Data Analysis: What kind of analysis do you plan to do with the data? What kind of tools might you need? What kind of visualization will you need?**

**Data Analysis:** An approach to de-synthesize data, informational and or factual elements to answer business related problem.

Tangerine Bank needs **the following software tools** to complete the data analysis:

**SQL Database**- For creating, deleting, modifying, and sorting queries on the structured data to create relational databases.

**R**- used to determine descriptive statistics (mean, median, standard dev, covariation, quartiles, etc.) on numerical data and gain insights.

**Tableau**-Connecting the databases to Tableau and create meaningful visualizations such as Age group and Balances, Age Group and House Ownership, Gender statistics, Customer Job and Loan counts etc. with key insights that can be shown to those involved in business marketing campaign goal.

**Python**- Create a machine learning model. Upon selecting from various machine learning algorithms such as Linear Regression, Decision Tree, Naïve Bayes, and Random Forest, proceed to generate the test design. The test model is necessary to test the model’s quality and validity. The test design will split the dataset into a training and testing set. The model will be built on the training set, and the quality of the model will be estimated on a separate test dataset.

**Online Social Blog Analysis-** To see the opinion of the customers on various bank services.

**Decision Making: How will decision makers want to receive the data?**

**Decision Making:** The process of selecting a course of action from different alternatives to achieve a desired goal or solve a problem.

* After the model is completed, the model must be **assessed and interpreted on success criteria.**
* In the banking situation, **the success criteria would be which customers should be targeted? What factors should be considered? What to consider in future marketing campaigns? How to determine if a customer is willing to take a loan?**
* The data analyst or engineer should contact the marketing department of the bank and other experts **to discuss the results in a business context**. A **review process** should occur to determine if resulting models satisfy business needs. A **quality assurance** on the models should be done to make sure nothing gets overlooked.
* The Stakeholders will see the visualizations through **a PowerPoint presentation**:

-**Scatter plots** can be used to visualize trends between Balance vs. Loans, Balance vs. Deposit.

-**Tableau tables** can be used to show trends in opinions on blogs.

**-Bar Charts** can be used to categorize customers by age group, education, marital status and job.

**-For machine learning model** results, and the rest of the team findings various presentation techniques will be used to give a clear understanding of results and prepare solutions.

* After the model evaluation results have been reviewed, **the model will be deployed within the bank organization’s system and a monitoring plan will be put in place for future changes.**

**References:**

DATA 1201- Data Collection Class Lectures

Tangerine Bank Website& Blog Posts. Quora and Reddit posts on customer opinions of Tangerine Bank.

José María Cavanillas, Curry, E., & Wolfgang Wahlster. (2016). *New Horizons for a Data-Driven Economy*. Cham Springer International Publishing.

Miller, M. (2013). From Data to Decisions: A Value Chain for Big Data. IT Professional, 15(1), 57–59. <https://doi.org/10.1109/MITP.2013.11>

Moro, Sérgio & Cortez, Paulo & Laureano, Raul. (2011). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. Proceedings of the European Simulation and Modelling Conference.

Parlar, T., & ACARAVCI, S. K. Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data. International Journal of Economics and Financial Issues, 7(2), 692-696, 2017.

Sedkaoui, S., n.d. *Data analytics and big data*.

Tuga Mauritsius et al., International Journal of Advanced Trends in Computer Science and Engineering, 8(5), September - October 2019, 2322- 2329