MIT805 Assignment 1: Data Collection and Processing

Yolanda Nkalashe, u13193016@tuks.co.za

September 14, 2020

1 Introduction

Data and the use of big data to gain untapped information has been the buzz for the last 5 or so years. With the promise of data driven technologies and prospects of profit from the information the data provides. Although, the promise of big data is real, industries are still challenged about understanding what big data constitutes (Bekker, 2020). For a project to be considered a big data project in any industry it must satisfy the three fundamental Vs which define big data (Salzig, 2020).

- The first v is Volume: literature defines volume as the size of the data which will be processed and analysed (Salzig, 2020).
- Secondly, one needs to look at the Variety of data which refers to how divers the data is with respect to the source and data types within the data (Salzig, 2020).
- The last fundamental v is velocity which speaks to how fast the data is generated, analysed and processed (Salzig, 2020).

The list above is not exhaustive but consists of the fundamental Vs each data set should consists of to be considered big data. Further one could look at validity and veracity, which speaks to the quality and credibility of data respectively.
In this report we study and analyse the technical aspects of a big data set which we collected, and we further process the data.

2 The data

The data set collected is data set from a lending company, called the lending club. The data consists of loan stats data of borrowers that have been granted the loan and contains demographic details, financial details and post grant details of borrowers.
The purpose of the data is for the Lending club to be able to cluster the borrowers based on their, interest rate, grade and determine probability of default. Further, the data is crucial to determine when the peak periods are in which people request loans and the most common term of loans (36 months or 60months).
The data was extracted from the website and was collected via csv format with the following technical aspects:

- Volume: The data consists of 150 columns with just over a million records. With memory or storage size of around 1.2 GB.
- Variety: The data is of the period from 2016 Q1 to 2020 Q4 it consists of a mix of string, float, integer and date formats across the 150 columns.
- Velocity: The data is collecting through the Lending Clubs app platform when a borrower loads, their information and is updated as the borrower makes any payments or any new development forms based on their profile. The Lending club releases the data after each quarter. Thus, although the data is available every quarter in the background the velocity of data is frequent.
- Validity and veracity: The data is consistently verified by the lending club.

3 Processing and Exploratory Data Analysis

Once the data has been collected it needs to be processed. Processing is a step which is crucial for any big data use case, it involves transforming your raw data into more meaningful information for the use case (Scheer, 1985).
There are two common ways that data can be processed (Scheer, 1985):

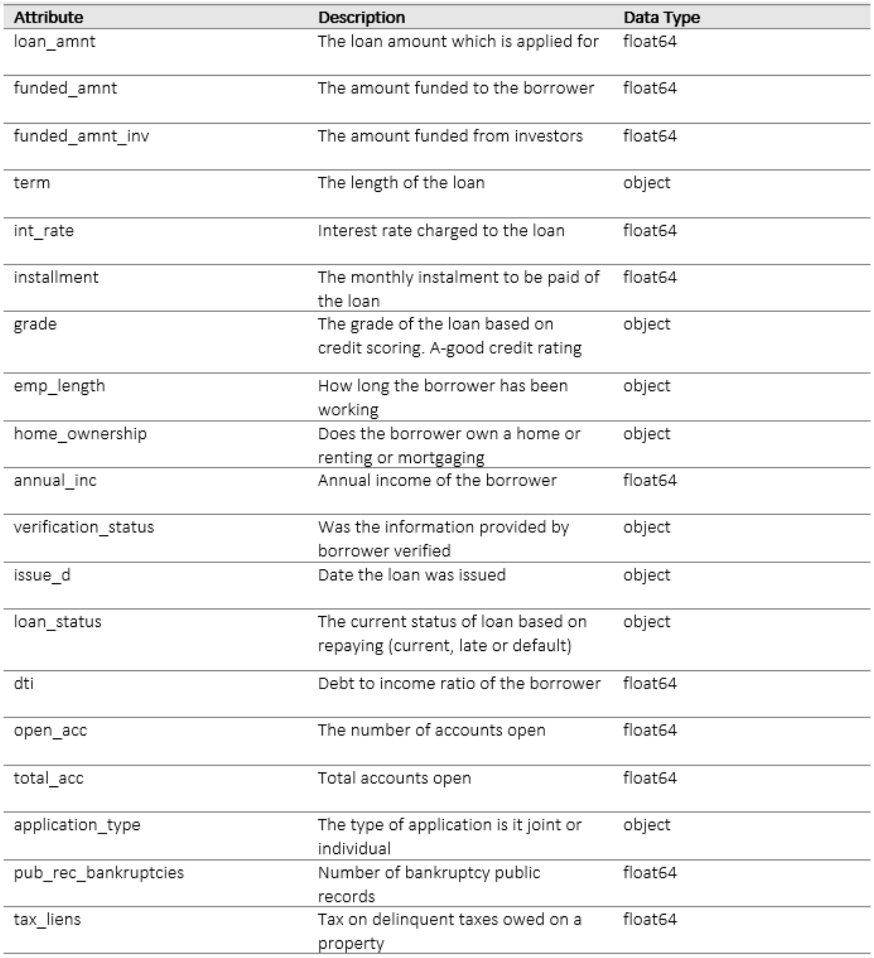
1. Batch processing: Data is collected and stored for a period before it is processed.
2. Real time: Data is processed as it is collected. In this section we describe how the Lending Club data was processed and the architecture of the system.

The first step taken before the data was processed was to identify the most suitable platform which would be able to host and house big data. Based on reviews and literature amazon web services chosen as the storage platform, whilst the processing platform used was Dask which integrates to aws at no fee compared to the services within aws. The data was processed via a jupyter notebook instance.

Once the platform was set up the following transformations and processing steps were taken:

- As described previously the data has 150 columns, however not all the columns will have predictive power and be useful to extract information. We thus first determine which columns have percentage of NA values which are greater than 30 %. Thereafter the columns are removed.
- Once the 'trivial' columns are removed we analyse each column in detail and determine and keep the attributes which would be critical to the use case described previously, clustering each borrower based on the information which is collected at origination of the loan. Through this exercise 19 columns where identified as crucial; these attributes are described in figure 1 below.

From the attributes identified the float data type variable where not transformed any further, however we do expect a correlation between employment length, debt-to income ratio, open accounts and the interest rate on the loan. We also expect a correlation between debt to income ratio and number of bankruptcy records. Categorical/ Object type variables where transformed as follows:



Attribute	Description	Data Type
loan_amnt	The loan amount which is applied for	float64
funded_amnt	The amount funded to the borrower	float64
funded_amnt_inv	The amount funded from investors	float64
term	The length of the loan	object
int_rate	Interest rate charged to the loan	float64
installment	The monthly instalment to be paid of the loan	float64
grade	The grade of the loan based on credit scoring. A-good credit rating	object
emp_length	How long the borrower has been working	object
home_ownership	Does the borrower own a home or renting or mortgaging	object
annual_inc	Annual income of the borrower	float64
verification_status	Was the information provided by borrower verified	object
issue_d	Date the loan was issued	object
loan_status	The current status of loan based on repaying (current, late or default)	object
dti	Debt to income ratio of the borrower	float64
open_acc	The number of accounts open	float64
total_acc	Total accounts open	float64
application_type	The type of application is it joint or individual	object
pub_rec_bankruptcies	Number of bankruptcy public records	float64
tax_liens	Tax on delinquent taxes owed on a property	float64

Figure 1: Final list of Attributes Chosen.

- Employment length was reduced from 10 categories to 4.
- Home ownership was reduced to 3 categories.
- Verification was transformed to Yes or No category and
- Loan status was reduced to 4 categories.

Once the data is processed based of the above transform at the data frame is converted to a new csv instance and stored in our S3 aws Bucket. The architecture of the data processing can be seen in the image below.

4 Conclusion

[h!] From collecting and processing of the Lending club big data these where the following observations where noted.

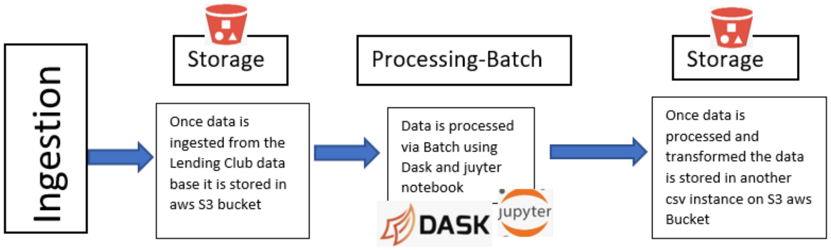

Ingestion
Storage
Once data is ingested from the Lending Club data base it is stored in aws S3 bucket
Processing-Batch
Data is processed via Batch using Dask and juyter notebook
DASK
jupyter
Storage
Once data is processed and transformed the data is stored in another csv instance on S3 aws Bucket

Figure 2: Overview of Processing architecture.

1. Storage and Processing of big data can take a lot up of ones time, therefore the right platform that can handle your data based on the use case must be found before processing and transformation.

2. Determining whether one will be processing in batch or real time is a decision that needs to be made prior as it affects the entire processing architecture.

3. Processing and collecting of big data is not an easy task one needs to plan and structure the process well and timely.

The next steps in any big data project would be to visualize the data and perform any algorithms or models necessary.

References

- Bekker, A., 2020. 7 Major Big Data Challenges And Ways To Solve Them. [online] Scnsoft.com. Available at: ¡https://www.scnsoft.com/blog/big-data-challenges-and-their-solutions¿
- Salzig, C., 2020. What Is Big Data? – A Definition With Five Vs. [online] Blog.unbelievable-machine.com. Available at: ¡https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs¿
- Scheer, A., 1985. The Importance of Data Processing for the Practical Realization of Theoretical Conclusions. Computer: A Challenge for Business Administration, pp.113-114.