MIT 805 Assignment 2: Map-Reduction and visualization

Yolanda Nkalashe, U13193016@tuks.co.za

11 November 2020

1. Introduction

In the previous project documentation, we presented a system of collecting and transforming our large data set. We used batch processing to transform the data by using Dask. This system/architecture we developed is a crucial step for any business as it ensures that the data which is used by the end users is clean and accurate and reduces the time for each user to clean the datasets themselves.

In this second portion of the project we use the cleaned processed dataset to process our big data using Hadoop tools such that we can derive the necessary insights from our data for business to make the necessary decisions and to ensure efficiency when dealing with their customers.

Hadoop is an architecture which allows us to store and process our data in a distributed manner (Sinha, 2020). The Hadoop ecosystem contains various tools that can be used based on the use case or project that you are undertaking (Sinha, 2020). For this project we use Apache spark for machine learning and MapReduce to a subset of our data for insights. Once we retrieve our data we visualize the data using a BI tool, the BI tool that we use is power BI. Power BI is a Microsoft tool which integrates with various source systems and allows for efficient data visualization.

In this report we will discuss the two systems 1. MapReduce 2. Spark for Machine learning we then discuss some of the visualization obtain through the BI tool.

2. The Systems:

2.1 MapReduce:

MapReduce is an algorithm which processes big data through distributed computing (Apache, 2020). The algorithms take individual records of a data and contains two components:

- Mapper: this portion of the algorithm splits the data into individual components and outputs key value pairs based on the output data needed to be seen based on the use case (Apache, 2020).
- Reducer: This stage contains a shuffle process where the output from the mapper is shuffled and grouped based on the key values and the second leg is the reducer which produces a new output from the shuffler (Apache, 2020).

In this project we use the Mrjob package which is easily integrated with AWS elastic map reduce. We create four map reduce jobs:

- Application Type: This map reduce class will output the split in volumes of the type of applications received in the company. This will ensure that as business we can get a sense of the type of customers we have and if we would like to attract a different customer segment the necessary decisions can be made.
- CustomerGrade: This map reduce class gives us a distribution of the grade split or the level of risk the customer base is throughout each quarter. This gives insight on whether we are catering for more riskier customers or low risk customers this will impact decision with respect to finance and residuals the company keeps based on the probability of default of the customers.
- MonthlyTrend: This output the monthly volumes and gives us a sense of what our seasonal trends are if we do have any.
- CustomerStatus: This mapper class will output the percentage customers that have defaulted and those that are late in payment.

2.2 Spark:

We would like to be able to cluster or segment our customers base on the data that we receive from customer once applied for the loan. This can assist with the decision on whether to accept a customer based on the cluster or segment that the customer falls in.

To create this cluster, we need to create kmeans machine learning model. Given the limitations of MapReduce not being able to iteratively transform parameters or datasets efficiently (Apache, 2020). To achieve this, we use the Hadoop apache spark which processes data using the distributed framework of processing data through various clusters.

3. Visualization:

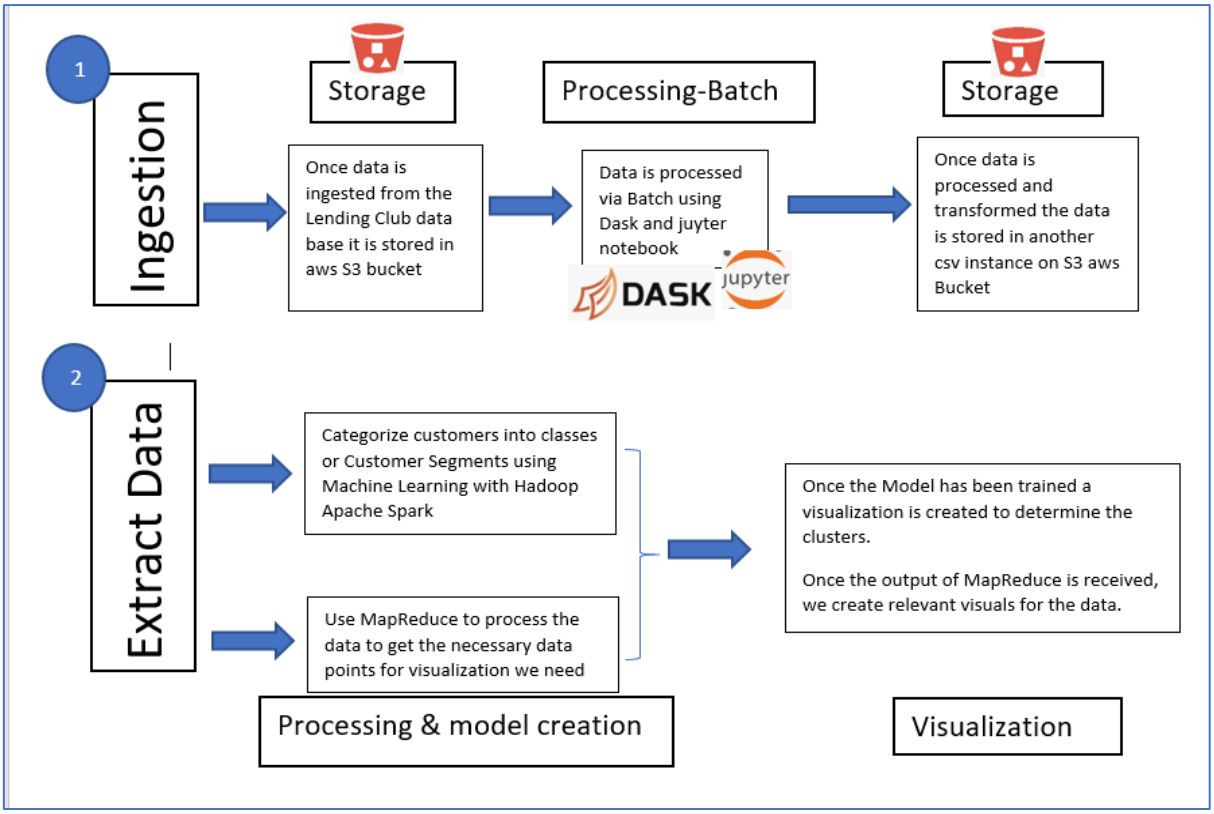
From the output of the model and MapReduce we visualize the summarized data to identify any interesting trends and patterns. Visualization ensures better communication and assists in faster decision making.

To visualize the data, we use a Microsoft BI platform power BI which seamlessly connects to various data sources either through connections strings, directly connecting to the data or using an API.

From the visualization one interesting note from the data is there’s not much of a seasonal trend and the average volumes per month are quite consistent which is not what we expected. We also not from the visualization that we take on customers which are most low risk customers. Further, our customer base is made largely on individual consumer applicants.

4. Architecture:

We extend the architecture proposed in part 1 to include the end user portion described in the report above.



References

Apache, H., 2020. *Hadoop - Mapreduce - Tutorialspoint*. [online] Tutorialspoint.com. Available at: <https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm#:~:text=MapReduce%20is%20a%20processing%20technique,(key%2Fvalue%20pairs).> [Accessed 1 November 2020].

Sinha, S., 2020. *What Is Hadoop | Introduction To Hadoop And It's Components | Edureka*. [online] Edureka. Available at: <https://www.edureka.co/blog/what-is-hadoop/> [Accessed 1 November 2020].