**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | *Advanced Data Analytics*  *Big Data Storage and Processing* |
| **Assessment Title:** | *MSC\_DA\_BD\_ADAv5* |
| **Lecturer Name:** | *David McQuaid*  *Muhammad Iqbal* |
| **Student Full Name:** | Jose Luis Alvarado Carranza |
| **Student Number:** | 2020092 |
| **Assessment Due Date:** | 08/11/2023 |
| **Date of Submission:** | 08/11/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Contents**

[ABSTRACT 3](#_Toc151155846)

[INTRODUCTION 3](#_Toc151155847)

[KEY TAKEAWAYS 4](#_Toc151155848)

[COMPARATIVE ANALYSIS OF DATABASES 5](#_Toc151155849)

[TIME-SERIES AND SENTIMENT ANALYSES 6](#_Toc151155850)

[FURTHER STUDIES 7](#_Toc151155851)

[CONCLUSION 7](#_Toc151155852)

[Bibliography 8](#_Toc151155853)

[References 9](#_Toc151155854)

# **ABSTRACT**

The following report details the many techniques, methods, and models used to store and process a dataset which contains over a million-and-a-half tweets. It is comprised of two main parts, one of processing and storing data and the other using the processed data to get a sentiment analysis and a time-series analysis in addition to their respective results and insights. Keywords: Sentiment, Time-Series, HBase, SparkSQL, Hadoop.

# **INTRODUCTION**

In this day and age, it has become of the utmost importance being able to identify and carry out analyses of large datasets. In this continuous assessment we have used the knowledge gathered in our course to complete said analysis. We begin by giving some information regarding the dataset to then be able to describe the processes and methods used with the technologies we have been required to using.

We have been given a dataset that contains more than a million tweets. As we know, Twitter (now known as ‘X’) is a social media website where users can post short text messages, videos, images, animated GIFs, polls, and more. Each post on Twitter is called a “tweet”, and each tweet consists of a maximum of 280 characters. These tweets are a great way to get a short bit of information out to followers, whether it’s an announcement about a product, a link to a blog post, or a poll you’re using to gather information. (Chimp, 2022)

As of 2022, Twitter has [229 million monetizable daily active users](https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/) across the globe. At the very least, your business should have a Twitter account to help boost your web presence. Even if you’re not especially active on Twitter, it gives your customers another way to find you and connect with your brand online. (Chimp, 2022).

Brands use Twitter to promote content related to their brand. For example, you might tweet out a link to a sale you are currently having, or maybe your tweets will include links to events you are holding in the near future. For brands, Twitter can be a great way to reach a larger audience and promote sales, products, and events in a more organic way. Individuals use Twitter for many different reasons. Some people have a Twitter account to follow their favorite athletes while others use Twitter to post one-liners and opinions. (Chimp, 2022)

As part of our assessment, our goal is to Critically assess the data storage and management requirements of a given data project from a modern perspective and evaluate limitations of legacy approaches to Big Data. We also have to assess the design concepts and architectural patterns of distributed Big Data systems and analyse the components that form their technology stack, and finally, critically evaluate and select a big data environment suitable for retrieving and processing a given Big Data set, perform data management and select appropriate analytic algorithms for the required scale and speed.

Firstly, we have uploaded the Project Tweets file into three technologies used in the course. These are Hbase, SparkSQL, and Hadoop.

We begin by uploading our dataset onto Hadoop. We open a terminal on Ubuntu as a virtualized environment and entered the commands shown in the figure below.

A close up of numbers

Description automatically generated

Figure 1: File uploaded to Hadoop.

The figure above shows that a new folder has been created in Hadoop called ‘ca2’ and there is where our dataset will be stored. This allows us to manage the large number of tweets in the dataset and from there, we can call it from a different technology so as to be processed.

For the purposes of developing this continuous assessment, we have carried on processing the data as it was more efficient for us rather than starting, stopping and switching between technologies every time.

We use Pyspark to be able to manipulate the data prior to processing so later it can be populated into the NoSQL database. To do so, we establish a connection between Hadoop and Spark.

A screenshot of a computer

Description automatically generated

Figure 2: File read from Hadoop.

And so, after creating a connection with Hadoop through pyspark, we defined a schema like so.

A screen shot of a computer code

Description automatically generated

Figure 3: Defining Schema.

We obtained the type of data in the dataset and then proceeded to creating a dataset which could be stored as a CSV file, as shown in the figure below.

A screenshot of a computer

Description automatically generated

Figure 4: Output from SparkSQL.

Now we proceed to store the dataset onto Hbase. To do this we have to begin by accessing Hbase from our virtual machine environment using a new terminal and changing the directory. Then, before importing the file from Hadoop, we must create a table in Hbase, as it is the way in which we can later import our tweets to Hbase.

A screenshot of a computer screen

Description automatically generated

Figure 5: Table creation.

Later, we input the code as shown below to import the dataset from Hadoop to Hbase.

A computer screen shot of a computer code

Description automatically generated

Figure 6: Upload files to Hbase.

The next step is to process the dataset. To do so we use the Hbase operational commands. There are 5 atomic commands which carry out different operations by Hbase. Get, Put, Delete, Scan and Increment. For our assessment we will use ‘Get’ to get the text of the tweets.

The data written is forwarded to MemStore which is actually the RAM of the data node. Afterward, all the data is dumped in HFile and, as we know from the classes of the course, the actual data is stored in HDFS. We were able to corroborate this as the project was being developed.

We can later use Hadoop to download the output file.



Figure 7: downloading processed data.

# **KEY TAKEAWAYS**

Through our assessment development we were able to test and try technologies such as Hbase and SparkSQL. It has become blatantly apparent that they are extremely useful as well as powerful in processing large datasets and more importantly, that they can be used with very little knowledge of coding, even though they are based on higher coding languages like Java.

We could appreciate the value of their use not only in a practical sense but also in its cost. This type of computational power could easily bankrupt a small company as Tech Giants like Microsoft, Amazon and Google, to name a few, charge large fees for their storage and virtualized environments.

Nevertheless, it is worth mentioning that we found many differences between these technologies in our approach to the assessment.

|  |  |
| --- | --- |
| SparkSQL | HBase |
| Spark provides a simple way to analyze data interactively. | Using a terminal to see the results can be confusing, especially because we are still not used to its interface |
| Uses SQL- like statements | It uses its own statements. |
| It is not schema free | It is schema free |
| It is newer technology and therefore limited in programming languages | It is older and it can be used with more programming languages. |
| It does not have map-reduce capabilities | It has map-reduce capabilities. |
| It becomes slower as the dataset gets bigger | It is slow regardless of the dataset |

Table 1: Comparison between used technologies

A computer screen shot of a program

Description automatically generated

Figure 9:Data read from Hbase Shell

# **COMPARATIVE ANALYSIS OF DATABASES**

As previously discussed, the location of the dataset and the way it can be processed by using mainstream methods can make or break a company. It is especially useful in cases when the resources are limited and so, there needs to be a way speedy way in which companies can determine the performance and capabilities of these technologies. This is where the Yahoo Cloud Service Benchmark come in handy. YCSB can be used to benchmark multiple database systems and can be used to compare them.

In our assessment, we have used YCSB to compare HBase and MongoDB as shown below.

A computer screen shot of white text

Description automatically generated

Figure 8: YCSB for HBase

The YCSB framework contains a core set of workloads to evaluate different aspects of a system’s performance, called the YCSB Core Package. In YCSB, a package is a collection of related workloads. The workload defines the data that will be loaded into the database during the loading phase, and the operations that will be executed against the data set during the transaction phase and can be used to evaluate systems at one particular point in the performance space. While the core package examines several interesting performance axes. (Researchgate, 2015)

Our approach towards implementing YCSB in our project was to try different workloads on the databases and then determine some of the most interesting points to mention. However, it is worth noting that we did some research online before benchmarking the databases and we based our approach on our findings and so, this part of our assessment is mostly corroborating the results. The figures included are from a different research which will be referenced further.

A table with numbers and letters

Description automatically generated

Figure 9: Read operation against Average latency for workload A.

A screenshot of a spreadsheet

Description automatically generated

Figure 5: Update operation against average latency for Workload A

A table with numbers and letters

Description automatically generated

Figure 6:Operational counts against Throughput workload A

A table with numbers and letters

Description automatically generated

Figure 7: Read Operation workload D

A table with numbers and text

Description automatically generated

Figure 8:Insert Operation Workload D

A table with numbers and letters

Description automatically generated

Figure 9: Operational counts workload A

# **TIME-SERIES AND SENTIMENT ANALYSES**

Performing time-series analysis involves examining historical data to uncover patterns, trends, and other valuable insights. It is a crucial step in understanding the behavior of time-dependent data and making predictions for the future. (Timescale, 2023). In our assessment this is not different.

We have gotten outputs from handling different databases. We obtained from previous processes two dataset which we have called ‘tweet.csv’ and ‘tweets\_sparksql.csv’. The former obtained from our HBase processes and the latter from using SparkSQL. We have used two different jupyter notebooks one for the sentiment analysis and the other for time-series analysis.

For our sentiment analysis, we followed the classes to the best of our understanding and we got some results. First we began by using the different libraries and models and so, we calculated the sentiment scores by classifying and counting negative and posives words given in the tweets while taking the ratio of the difference of the word count and total word count. Later, we creaed an instance of wordNetLemmatizer. Next, we defined a function to preprocess the text. Then, we extracted letters from the strings. To accomplish this, we used “Regular Expressions”. Later, we tokenized the string so that we could then filter stop words. Finally, we lemmatized the remaining words and returned them.

We applied the function to every text in the ‘text’ column. By doing this we have calculated the number of resultant words. The idea here was to compare the resultant words with what is know as Opinion Lexicon, which are two different .txt files named positive-text.txt and negative-text.txt. These files are commonly used in sentiment analysis and have to be stored in the same folder as the jupyter notebook to be read by the algorithm. The words are counted and stored. Finally, we calculated thee sentiment. All of these steps allowed us to assign number to the words and therefore, we have effectively turned the words to numerical values.

Our reasoning behind this was that if we followed the logic behind what has been taught to us in class, we would get the desired results which are taking words, assign them values to use the models determine the sentiment of the tweets dataset. Additionally, we did some research and it showed that the procedures followed in class are a norm in order to perform a sentiment analysis.

For our time-series analysis, again we followed the classes and the logic behind them as there needed to be a lot of modifications for the process to yield results.

Firstly, we begin by adding the necessary libraries and algorithms. Then, we determined the type of data which resulted in an object. We used pandas to change the time to a format that could be used to use the time-series analysis. This then allowed us to generate some visualizations to try and observe correlations, stationarity, or some key feature to then be able and build our model. Then, we tried to validate thhe model and finally use some model to make predictions.

RESULTS

Both the Sentiment and Time-series Analyses yielded results. The caveat being that they were not the expected or desired results. We were limited by either our computational power or the amount of comprehension on the necessary topics to achieve our objectives.

In terms of the computational power, we were limited by the processing power of our equipment. It seems that the dataset was to big to handle. At some point, we even decided to wrangle the data so as to create a smaller dataset, half of the amount of tweets. Even then, the computer kept on crashing. It was particularly evident that the virtualized environment on top the operating system was too much for it, let alone be able to wrangle that many tweets. In some instances, the equipment responded, however, it took over thirty minutes to come up with some sort of result. It hindered our results as well as the overall outcome of this continuous assessment.

ALTERNATIVE

In our assessment we were expected to analyse a set of requirements to determine the type of advanced data analysis for any problem posed. So, to accomplish this we would like to give some guidelines that hopefully can be followed and then get desired results, because it seem that even though the situation at hand maybe different, the logic behind it remains and so do the steps to follow.

Assuming we have a proper equipment, which is at least capable of handling a high computational load, we would begin by collecting the data and cleaning it. This will allow us to prepare visualizations with respect of time. Additionally, we would need a key feature which for our case is the number assigned to the sentiment. From there we could observe if there our series has stationarity. If there is Stationarity, which is very likely due to the nature of our data set, we would proceed develop charts to understand its nature. This would become a precursor to building our model, this being AR, MA, ARIMA, Etc. Finally we could extract insights from the predictions.

# **FURTHER STUDIES**

In our Big data storage and processing part of the assessment, we used convenience technologies because they are open source and therefore free. In class we saw that the similar results can be achieved with paid technologies, like the ones Google offers its customers. What is really the main advantage of paying for Storage technology. There could be a comparison between the capabilities and the time it takes for processing data.

In our attempt of coming up with predictions, we have proposed a methodology to follow in order to get results. So, there needs to be further studies in terms of time series analyses as they possess inheritent limitations like the data points need to be linear in their relationships or the act that they mostly work on univariate data. In our case, we used time and value of the sentiment. So, perhaps These tweets can be analyzed as mulltivariable data.

# **CONCLUSION**

In the real world companies are shifting their storage needs to cloud services and the demand to manage big data increases with the passing of time because as technology evolves, so do the requisites of the companies. However, Hadoop is still used by companies in many different sectors. Thus, its true worth was shown through our project because we demonstrated that it can handle large volumes of data at low costs.

The performance of any database can be tested using YCSB. This is an extremely valuable tool because it is reliable and more importantly scalable. It allows us to find out if the database that we want to use is the appropriate for our interest. So, if there is a company which needs to pay for databases regularly, this tool can be used to detemine if said service is worth paying for or no. In our case, this was not evidence because we inly worked with convenince technologies.

The main downside of working with HBase is that it does not support some of the more traditional model features, and so, it is not an ideal replacement of database. Thus migrating data from a different RDBMS to HBase servers would require a new design which in turn things more complicated when querying needs to be done.

# Bibliography

Chimp, M., 2022. *https://mailchimp.com/resources.* [Online]   
Available at: https://mailchimp.com/resources/twitter-guide-for-beginners/#:~:text=Each%20post%20on%20Twitter%20is,re%20using%20to%20gather%20information.  
[Accessed 10 11 2023].

Dataversity, 2022. *https://www.dataversity.net.* [Online]   
Available at: https://www.dataversity.net/brief-history-deep-learning/  
[Accessed 30 09 2023].

Flow, T., 2023. *https://www.tensorflow.org/.* [Online]   
Available at: https://www.tensorflow.org/  
[Accessed 06 10 2023].

Hassaballah, M. a. A. A. e., 2020. *Deep learning in Computer VIsion: principles and applications..* Boca Raton, FL: Taylor & Francis Group.

IBM, 2023. *https://www.ibm.com.* [Online]   
Available at: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/  
[Accessed 30 09 2023].

Keras, 2023. *https://keras.io/.* [Online]   
Available at: https://keras.io/why\_keras/  
[Accessed 06 10 2023].

Linkedin, 2023. *https://www.linkedin.com/advice.* [Online]   
Available at: https://www.linkedin.com/advice/0/what-can-you-learn-from-using-open-source-machine  
[Accessed 30 09 2023].

Microsoft, 2023. *https://azure.microsoft.com/.* [Online]   
Available at: https://azure.microsoft.com/en-us/solutions/ai/data-scientist-resources/  
[Accessed 30 09 2023].

Observatory, E., 2022. *https://www.economicsobservatory.com/.* [Online]   
Available at: https://www.economicsobservatory.com/minimum-wages-and-living-wages-what-happens-in-times-of-inflation  
[Accessed 06 10 2023].

Oracle, 2022. *The Evolution of Big Data and the Future of the Data Platform.* [Online]   
Available at: https://www.oracle.com/ie/a/ocom/docs/big-data/big-data-evolution.pdf  
[Accessed 05 10 2023].

Researchgate, 2015. *https://www.researchgate.net/publication.* [Online]   
Available at: https://www.researchgate.net/publication/349239668\_A\_Study\_of\_YCSB\_-tool\_for\_measuring\_a\_performance\_of\_NOSQL\_databases  
[Accessed 7 11 2023].

Target, T., 2020. *https://www.techtarget.com/.* [Online]   
Available at: https://www.techtarget.com/searchdatamanagement/definition/big-data  
[Accessed 05 10 2023].

Timescale, 2023. *https://www.timescale.com.* [Online]   
Available at: https://www.timescale.com/blog/how-to-work-with-time-series-in-python/#:~:text=Python%20provides%20the%20Matplotlib%20library,%2C%20scatter%20plots%2C%20and%20histograms.&text=Performing%20time%2Dseries%20analysis%20involves,trends%2C%20and%20other%20valu  
[Accessed 7 11 2023].

# References

Chimp, M., 2022. *https://mailchimp.com/resources.* [Online]   
Available at: https://mailchimp.com/resources/twitter-guide-for-beginners/#:~:text=Each%20post%20on%20Twitter%20is,re%20using%20to%20gather%20information.  
[Accessed 10 11 2023].

Dataversity, 2022. *https://www.dataversity.net.* [Online]   
Available at: https://www.dataversity.net/brief-history-deep-learning/  
[Accessed 30 09 2023].

Flow, T., 2023. *https://www.tensorflow.org/.* [Online]   
Available at: https://www.tensorflow.org/  
[Accessed 06 10 2023].

Hassaballah, M. a. A. A. e., 2020. *Deep learning in Computer VIsion: principles and applications..* Boca Raton, FL: Taylor & Francis Group.

IBM, 2023. *https://www.ibm.com.* [Online]   
Available at: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/  
[Accessed 30 09 2023].

Keras, 2023. *https://keras.io/.* [Online]   
Available at: https://keras.io/why\_keras/  
[Accessed 06 10 2023].

Linkedin, 2023. *https://www.linkedin.com/advice.* [Online]   
Available at: https://www.linkedin.com/advice/0/what-can-you-learn-from-using-open-source-machine  
[Accessed 30 09 2023].

Microsoft, 2023. *https://azure.microsoft.com/.* [Online]   
Available at: https://azure.microsoft.com/en-us/solutions/ai/data-scientist-resources/  
[Accessed 30 09 2023].

Observatory, E., 2022. *https://www.economicsobservatory.com/.* [Online]   
Available at: https://www.economicsobservatory.com/minimum-wages-and-living-wages-what-happens-in-times-of-inflation  
[Accessed 06 10 2023].

Oracle, 2022. *The Evolution of Big Data and the Future of the Data Platform.* [Online]   
Available at: https://www.oracle.com/ie/a/ocom/docs/big-data/big-data-evolution.pdf  
[Accessed 05 10 2023].

Researchgate, 2015. *https://www.researchgate.net/publication.* [Online]   
Available at: https://www.researchgate.net/publication/349239668\_A\_Study\_of\_YCSB\_-tool\_for\_measuring\_a\_performance\_of\_NOSQL\_databases  
[Accessed 7 11 2023].

Target, T., 2020. *https://www.techtarget.com/.* [Online]   
Available at: https://www.techtarget.com/searchdatamanagement/definition/big-data  
[Accessed 05 10 2023].

Timescale, 2023. *https://www.timescale.com.* [Online]   
Available at: https://www.timescale.com/blog/how-to-work-with-time-series-in-python/#:~:text=Python%20provides%20the%20Matplotlib%20library,%2C%20scatter%20plots%2C%20and%20histograms.&text=Performing%20time%2Dseries%20analysis%20involves,trends%2C%20and%20other%20valu  
[Accessed 7 11 2023].