A Synopsis on

"Streaming Analytics"

Submitted in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

(ARTIFICIAL INTELLIGENCE)

Submitted by

A Rishab Vanigotha
19BTRCR018

M R Naveen Kumar
19BTRCR005

Shraddha Hiremath
19BTRCR037

Sujay Sukumaran Adityan
19BTRCR051

Under the guidance of

Dr. John Basha

Designation

Assistant professor

Faculty of Engineering & Technology

Jain (Deemed-To-Be University)

Department of Computer Science & Engineering

Jain Global Campus, KanakapuraTaluk - 562112

Ramanagara District, Karnataka, India

2020-2021

1. About the Problem.

Every day, social media networks such as Twitter, Facebook, Instagram, and YouTube generate billions of bytes of data. This information is commonly referred to as big data. The majority of the information is in text format. These text data are then used to generate profits for any business organisation by developing recommendation systems, correctly targeting online advertisements, sentiment analysis, customer segmentation, and many other strategies. To analyse such enormous amounts of data on a single system is highly computationally expensive and takes a long time. When it comes to streaming data, the rate at which it is generated is quite quick. A single system is incapable of processing large amounts of data that enter the system at high rates. To run a big data process efficiently we need current distributed parallel computing systems.

2. The primary reason to choose this topic (can also include the current technology and improvements being made with this project).

Organizations may use social media to assess their customers' reactions to material and events in real time. Furthermore, the initial stage of sentiment analysis is the pre-processing of data gathered from social media.

This project covers the use of Twitter in a variety of recommended themes, as it is the largest social networking website, and Twitter data is rising at an increasing pace every day, making it a Big Data Source.

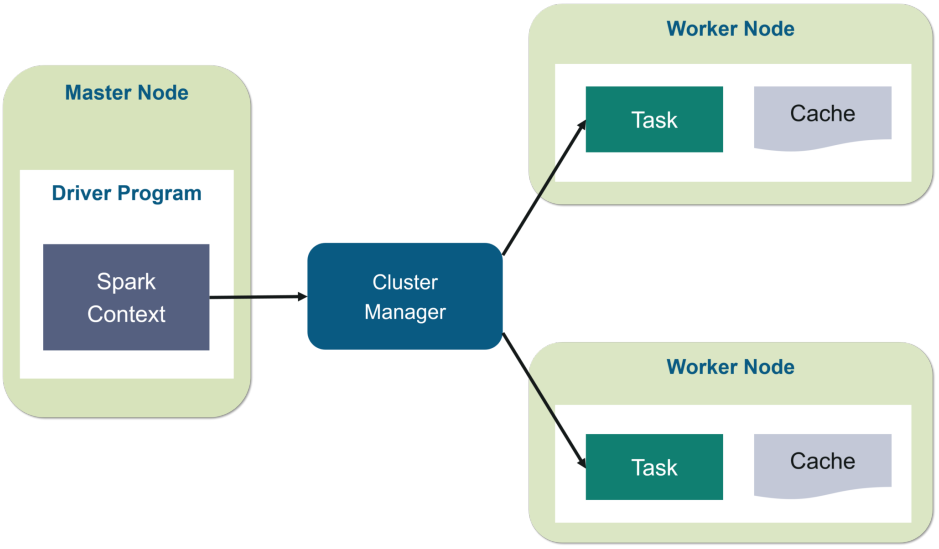
It is becoming more and more clear that the solution to artificial intelligence issues lies in efficient large data analysis. As the amount of data created on a daily basis reaches quintillions of bytes, it is becoming more crucial than ever to have a robust platform for effective big data analytics.

However, using machine learning techniques on large and complex datasets is computationally expensive and uses a lot of logical and physical resources, including CPU, memory, and data file space.

One of the most well-known platforms for big data analysis is Apache Spark MLlib, which provides a number of excellent functionalities.

3. The main objective of the project (a clear picture of the project).

To reduce computation time and distribute resources throughout the available memory for large-scale data processing, the project's primary goal is to handle big data using distributed parallel computing platforms.



4. Scope of the Project.

Ideal goal of this project is to analyse the real time streaming data using web scraping tool, snscrape for data ingestion, Apache Spark to analyse the streaming data using Spark MLlib and Power BI for creating dashboard

Once the data preparation on the fetched data is completed, we will work on sentiment analysis and cyber bullying in this project. In this segment, we will use python and spark frameworks to retrieve and analyse data.

Once sentiment analysis and cyberbullying are completed, we will create a graphical user interface (GUI)-based dashboard for better data visualisation, allowing us to gain multiple perspectives on a single piece of data.

5. Details about the Software or technologies that will be used.

Our project uses Spark as the most computationally efficient tool for performing analysis on streaming data.

In particular, Spark MLlib will be used to create models for streaming analytics in order to take advantage of in-memory cluster computation, which is quicker than conventional machine learning algorithms.

In this project, we will analyze Twitter data that has been collected using the Python module snscrape, which has no restrictions whatsoever on retrieving streaming data from Twitter.

6. Limitations of the system proposed (if applicable).

Cost ineffective:

While we talk about cost-effective large data processing, maintaining data in memory is difficult. When we operate with Spark, the memory consumption is really significant. Spark takes a lot of RAM to run in memory. The additional memory required to run Spark is quite expensive, so in-memory computing can be rather costly

Limited support to Model Ensembles:

Model ensembles such as stacking, boosting, and bagging are not supported by Spark ML. It offers limited boosting support in Random Forest training and Gradient Boosted Trees.

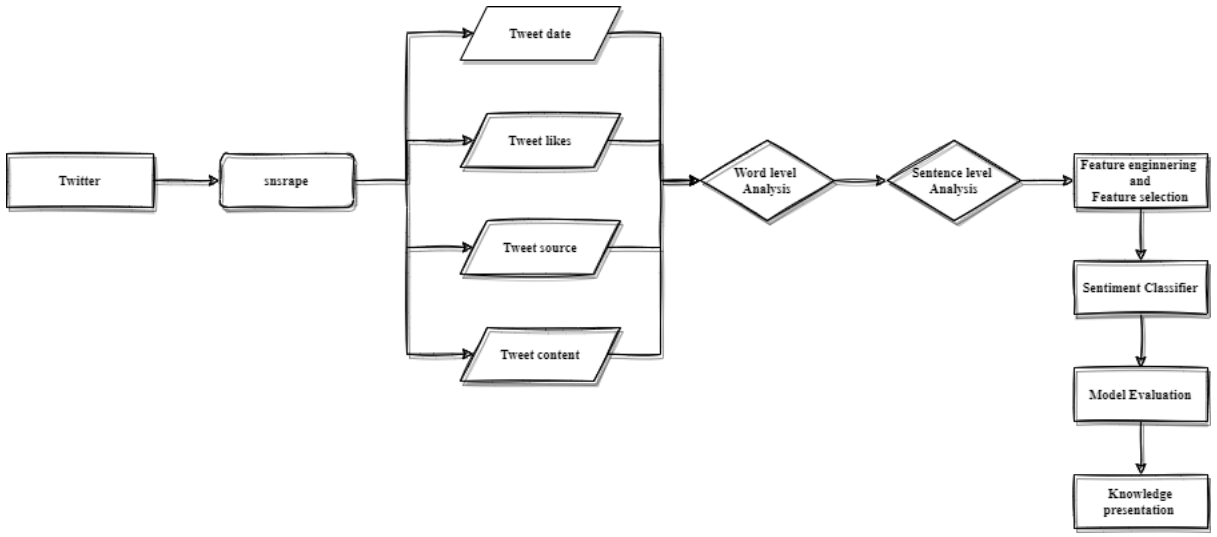
Spark ML has limited support of statistic functions:

Although you may compute the relationship between two columns and test hypotheses, Spark ML does not support advanced statistical methods like ANOVA, other descriptive statistics, and so on.

There is no file management system:

Spark does not have its own file management system. It does not include a file management system. It is usually dependent on other file system systems. As a result, it must integrate with any of the file management systems like HDFS. It's one of Spark's fundamental problems.

7. Flow chart to complete this project



REFERENCES

[1]. https://www.tandfonline.com/doi/pdf/10.1080/23311916.2018.1534519?needAccess=true

[2]. http://203.201.63.46:8080/jspui/bitstream/123456789/6117/1/PR3112.pdf

[3]. https://www.researchgate.net/publication/322577160_Detecting_Offensive_Language_in_Tweets_Using_Deep_Learning

[4]. Ha, I. , Back, B. , & Ahn, B. (2015). MapReduce functions to analyze sentiment information from social big data. International Journal of Distributed Sensor Networks , 11, 417502. doi:10.1155/2015/417502

[5]. Sheela, L. J. (2016). A review of sentiment analysis in twitter data using Hadoop. International Journal of Database Theory and Application , 9(1), 77–86. doi:10.14257/ijdta

[6]. Tare, M. , Gohokar, I. , Sable, J. , Paratwar, D. , & Wajgi, R. (2014). Multi-class tweet categorization using map reduce paradigm. International Journal of Computer Trends and Technology (IJCTT)

[7]. Barskar, A. , & Phulre, A. (2017). Opinion mining of twitter data using Hadoop and Apache Pig. International Journal of Computer Applications , 158, 9. doi:10.5120/ijca2017912854

[8]. Shang, S. , Shi, M. , Shang, W. , & Hong, Z. (2015, June). Research on public opinion based on big data. In Computer and Information Science (ICIS), 2015 IEEE/ACIS 14th International Conference on (pp. 559–562). Las Vegas, NV, USA: IEEE.

[9]. Jain, A. , & Bhatnagar, V. (2016). Crime data analysis using pig with Hadoop. Procedia Computer Science , 78, 571–578. doi:10.1016/j.procs.2016.02.104