**INTRODUCTION**

**Background**

Social networking sites connect people in the world, allowing them to share pictures, content, videos and share their first-hand opinions on various issues. Big data analytics techniques are highly applied in the social networks because they are characterized by the 5V (Velocity, Volume, Veracity, Value AND Variety) s of big data.(Bazzaz Abkenar *et al.*, 2021). Some examples of social networking sites include Twitter, Facebook etc. Due to social media providing a big source of data, there has been an increase in application of machine learning, deep learning and time series techniques to better understand various problems. A lot of these analysis has been done mostly on twitter data. Twitter which has over 313 million monthly active users and which in a day over 500 million tweets are made, is one of the most utilised social media platforms when it comes to data.(Jianqiang and Xiaolin, 2017)

Some key techniques employed for understanding social media data include, sentiment analysis, times

**Problem statement**

**DATA**

The dataset used for this project was project Tweets dataset, a csv dataset with 1,600,000 observations and five features extracted using the twitter api. It contains the following 5 fields:

* ids: The id of the tweet (eg. 4587)
* date: the date of the tweet (eg. Sat May 16 23:58:44 UTC 2009)
* flag: The query (eg. lyx). If there is no query, then this value is NO\_QUERY.
* user: the user that tweeted (eg. bobthebuilder)
* text: the text of the tweet (eg. Lyx is cool)

**Big data Processing and Storage**

Big data processing is techniques utilised to access large scale data and extract meaningful information from them for decision making.(Mehdipour, Noori and Javadi, 2016), while big data storage are storage technologies that are not relational database systems that can be able to address the Volume, variety and velocity challenges of data.(Strohbach *et al.*, 2016). There are different big data storage and processing technologies available. Processing technologies include Hadoop Map-Reduce or Apache Spark, etc. While storage include either SQL or NoSQL databases such as HBase, HIVE, Spark SQL, Cassandra, MongoDB. Etc

For this project Apache Spark was used for Processing the Project Tweets Data, while MongoDB and spark SQL were used to Populate, Store and save Processed Data.

**Spark SQL**

This is a spark module for structured data processing.

**Big Data**

1. Details of the data storage and processing activities carried out, including preparation of the data and processing the data in a MapReduce/ Spark environment;**[0-30]**

● Source dataset(s) can be stored into an appropriate SQL/ NoSQL database(s) prior to processing by MapReduce / Spark (HBase / HIVE / Spark SQL /Cassandra / MongoDB / etc.) The data can be populated into the NoSQL database using an appropriate tool (Hadoop/ Spark etc.)

● Post Map-reduce processing dataset(s) can be stored into an appropriate NoSQL database(s) (Follow a similar choice as in the previous step)

● Store the data and then follow-up analysis on the output data. It can be extracted from the NoSQL database into another format, using an appropriate tool, if necessary (e.g. extract to CSV to import into R/ Python etc.).

* Source dataset(s) can be stored into an appropriate SQL/ NoSQL database(s) prior to processing by MapReduce / Spark (HBase / HIVE / Spark SQL /Cassandra / MongoDB / etc.) The data can be populated into the NoSQL database using an appropriate tool (Hadoop/ Spark etc.)
* Post Map-reduce processing dataset(s) can be stored into an appropriate NoSQL database(s) (Follow a similar choice as in the previous step)
* Store the data and then follow-up analysis on the output data. It can be extracted from the NoSQL database into another format, using an appropriate tool, if necessary (e.g. extract to CSV to import into R/ Python etc.).

**3. YCSB FOR Comparative analysis OF MYSQL AND MONGODB**

There are various test strategies that can be implemented in order to perform comparative analysis of the capabilities of various big databases. One of the most common is Yahoo cloud service benchmark client (YCSB). YCSB is an open-source license tool used to benchmark new cloud database systems. Through YCSB one can be able to benchmark multiple systems and compare them by creating “workloads”. (Gaikwad and Goje, 2015). A YCSB Comparative analysis was conducted to compare Capabilities of MySQL and MongoDB.

The comparison involved the following areas: -

1. *Workloads a, b, c*
2. *Read and update options*
3. *Distribution (Zipfian and uniform)*

YCSB has various distributions such as: -

Uniform distribution allows all records in the database to be equally chosen.

Zipfian distribution allows one to choose records in accordance to popularity. The most popular become the heads of the distribution and the least popular the tail.

The metrics considered for quantitative analyses were: -

1. *Average Latency*
2. *Throughput*
3. *Runtime*

The comparison involved comparing the two databases performance based on record counts 1,000, 10,000 and 100,000. Further the read and update proportions were also compared for different proportions and lastly the request distributions compared were Zipfian and uniform distribution, as shown in table below

|  |  |  |
| --- | --- | --- |
| **Comparison Parameters** | **MYSQL** | **MONGODB** |
| Record Counts | 1,000, 10,000 and 100,000 | |
| Request distribution | Zipfian and Uniform | |
| Workload A – Update heavy | * Read: 50% and Update: 50% * Read: 70% and Update: 30% | |
| Workload B- Read heavy | * Read: 95% and Update: 5% * Read: 70% and Update: 30/5 | |
| Workload C- Read only | * Read: 100% and Update:0% * Read: 70% and Update: 30% | |

**WORKLOAD A**

**Runtime**

**Throughput**

**Average Latency**

**WORKLOAD B**

**Runtime**

**Throughput**

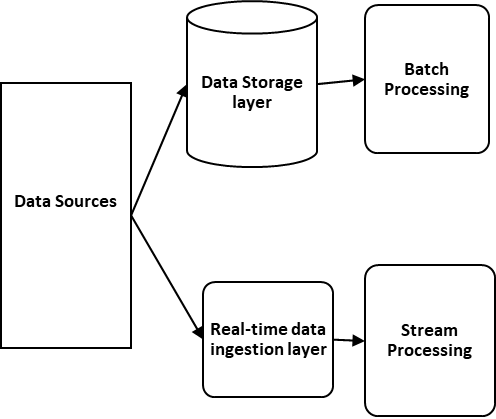
**Average Latency**

**WORKLOAD C**

* The runtime(ms) when read and update is 50/50 and the distribution is Zipfian for both MySQL and MongoDB are increasing as the records count increases. MongoDB has a lower runtime compared to MySQL across the three record counts. The runtime(ms) when read and update is 70/30 and the distribution is Zipfian for both MySQL and MongoDB are increasing as the records count increases. MongoDB has a lower runtime compared to MySQL across the three record counts. When you increase the read and update from 50/50 to 70/30, MongoDB still performs well compared to MySQL, but MongoDB with 50/50 has a lower runtime compared mongo with 70/30, while MySQL with 50/50 has higher runtime compared to MySQL with 70/30.
* The runtime(ms) when read and update is 50/50 and the distribution is uniform for both MySQL and MongoDB are increasing as the records count increases. MongoDB has a lower runtime compared to MySQL across the three record counts. The runtime(ms) when read and update is 70/30 and the distribution is uniform for both MySQL and MongoDB are increasing as the records count increases. MongoDB has a lower runtime compared to MySQL across the three record counts. When you increase the read and update from 50/50 to 70/30, MongoDB still performs well compared to MySQL, but MongoDB with 50/50 has a higher runtime compared mongo with 70/30, while MySQL with 50/50 has lower runtime compared to MySQL with 70/30.
* When the read and update is either 70/30 or 50/50 and the distribution is Zipfian for MySQL, the throughput increases, decreases then increases, while for MongoDB it increases only. The throughput is higher in MongoDB than in MySQL for both 50/50 and 70/30. Increasing the read to 70 and reducing update to 30 for both databases, causes an increase in throughput for both databases.
* When the read and update is 70/30 and the distribution is uniform for MySQL, the throughput decreases, increases then decreases, while when it is 50/50 the throughput decreases as the record counts increase, while for MongoDB it only increases in both proportions. The throughput is higher in MongoDB than in MySQL for both 50/50 and 70/30. Increasing the read to 70 and reducing update to 30 for both databases, causes an increase in throughput for both databases.

1. A discussion of the rationale and justification for the choices you have made in terms of data processing and storage, programming language choice, that you have implemented.**[0-20]**
2. Design the architecture for the processing of big data using all the necessary technologies (HADOOP/SPARK,NOSQL/SQL databases and programming). Present your Design in the form of a diagram and discussion in your report **.[0-20]**

Big data architecture handles the ingestion, storage, processing and analysis of data that is big data.



**Data Analytics and Machine Learning**

**Data Visualization and Reporting**

**Advanced Data Analytics**

1. A discussion of the rationale, evaluation, and justification for the choices you have made in terms of EDA, data wrangling, machine learning models and algorithms that you have implemented**.[0-40]**
2. Provide evidence and justification of your choice of sentiment extraction techniques.
3. Explore at least 2 methods of time-series forecasting including at least 1 Neural Network and 1 autoregressive model (ARIMA, SARIMA etc…) . (Hint: that this is a Short time series, How are you going to handle this?)
4. Evidence and justify your choices for your final analysis and include your forecasts at 1 day, 3 days and 7 days going forward.
5. Your dashboard must be dynamic and interactive. Include your design rationale expressing Tufts principles.
6. **E**valuation and justification of the hyperparameter tuning techniques that you have used **[0-20]**
7. Your analysis of any change sentiment that occurs and your forecast of the sentiment at 1 day, 3 days and 7 days going forward**[0-20]**
8. Presentation of results by making appropriate use of figures along with caption, tables, etc and your dashboard for your forecast, Discuss Tufts Principles in relation to your Dashboard **.[0-20]**