>> from Apache Spark import Technology, Fun

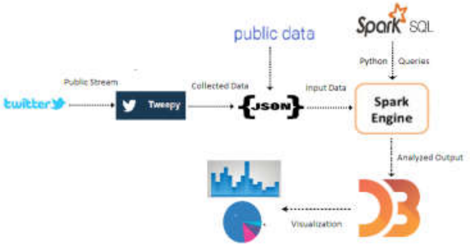
(Analysis on Payment Technologies, Batman v Superman movie)

Sai Venkatesh Gatiganti, Meghsai Reddy Bodimani, Sri Chaitanya Patluri

Abstract

The Project involves collection of Twitter Data using Tweepy, A Python based API used to collect Tweets from public streams & storing it in json format. The Analysis of collected data was done using Apache Spark, an Open Source Software used to analyze large amounts of data by running it through custom queries. The visualization of analyzed data was done through D3.js, A JavaScript library for manipulating documents based on data.

Architecture



Datasets

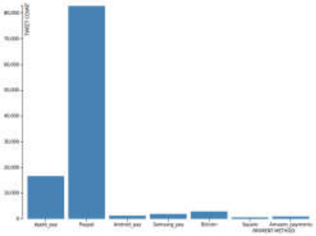
Dataset I: Tweets collected over a period of week on Different Payment Technologies like Apple Pay, Samsung Pay, Android Pay, PayPal etc.

Dataset II: Tweets collected about the movie "Batman v Superman: Dawn of Justice"

Dataset III: Public Dataset from U.S. Department of Education on all Accredited Universities in U.S.A.

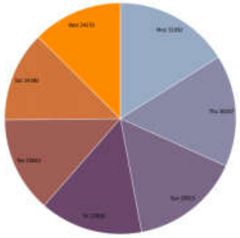
Queries & Visualization

Query I: Number of Tweets for each Payment Technology over a week in Dataset I.



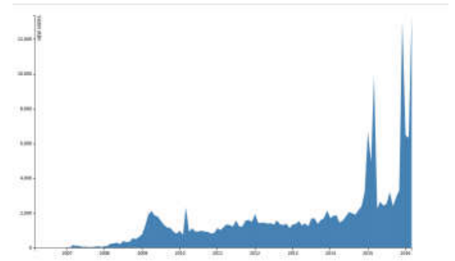
Query Time: 116 sec

Query II: Tweet count on each day over a period of week for all Payment Technologies in Dataset I.



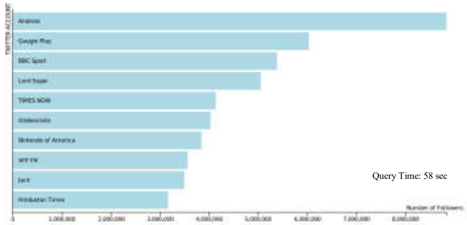
Query Time: 59 sec

Query III: Number of Twitter accounts created according to month & year from Dataset I.



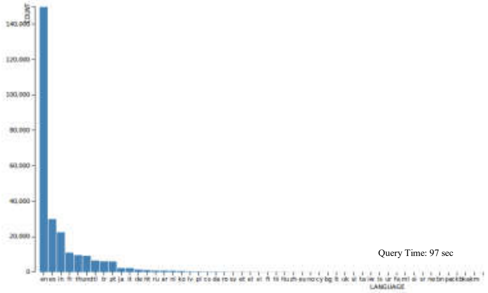
Query Time: 62 sec

Query IV: Top 10 Verified Accounts with Highest Follower Count in Dataset I.



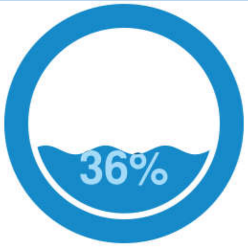
Query Time: 58 sec

Query V: List of all the Languages that were used to tweet and its count on Dataset - II.



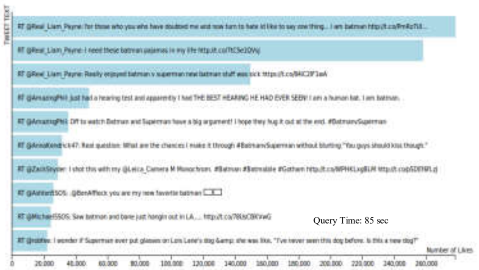
Query Time: 97 sec

Query VI: Percentage of Tweets with External Links in Tweet Status for Dataset II.



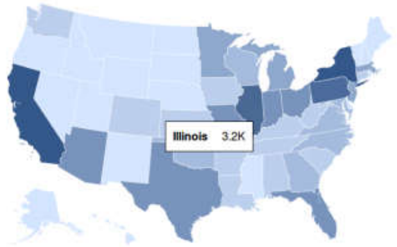
Query Time: 100 sec

Query VII: Top ten liked tweets with like count for Dataset II.



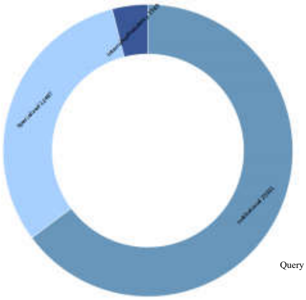
Query Time: 85 sec

Query VIII: Number of Colleges in each state for Dataset III.



Query Time: 39 sec

Query IX: Ratio of Accreditation Types of all Colleges in Dataset III.



Query Time: 15 sec

Tools

Environment: PySpark(Apache Spark)

Programming Language: Python

Visualization: D3.js

Conclusion

- Use of Apache Spark has significantly reduced Query processing time.
- Analytical results used in this project can be used by companies to review their products.

References

1. https://d3js.org
2. https://www.tweepy.org
3. https://databricks.com
4. https://spark.apache.org/docs/1.6.0/api/python/pyspark.sql.html