Development of Real Time Analytics of Movies Review Data using PySpark

Prakash K. Aithal, Dinesh Acharya U., Geetha M.

Abstract— The data play the vital role in every organization. The data can be divided into structured, semi-structured and unstructured. One can not process the unstructured data in real-time using RDBMS or Hadoop. Spark is an extension of Hadoop architecture which clubs the goodness of both Hadoop and Storm. Spark supports languages such as Scala, Java, Python, and R. The proposed method uses PySpark to analyze the movies review dataset of 50000 reviews by 36409 people for 1539 movies in real-time. Since movie reviews are written by many users in real-time, it is necessary for real-time data analysis. This method finds all the users who are very active in writing the reviews of the movies. This analytics may be used for giving incentives to the active reviewers. Further, the information about more popular movies based on reviews can be gained through analytics. To achieve these tasks basic map, reduce and filter functionalities have been applied. It is found from the analytics that the Movie code B002VL2PTU has been reviewed by the maximum number of people and also it is determined that maximum of 112 reviews were written by the single user with code A3LZGLA88K0LA0. The frequency count of words in the movie review is accomplished, and sentiment of the user can be analyzed using unigrams.

Keywords-Real-time Analytics; BigData; PySpark

I. INTRODUCTION

Amazon.com is the most successful e-commerce site and is also one of the fortune500 companies. Amazon.com was found by Jeff Bezos in 1994. It is successful because of the review system, an integral part of the company. Amazon.com provides a review for every product it sells. The review sys- tem has benefitted all the stakeholders. The recommendation system is another integral part of the amazon.com which cross sells and up sells the products based on user interests. The proposed method analyzes the movie reviews data found on the website of Stanford University[1]. The advantage of analyses are as follows

1) One can tell which movie has highest reviews and which is the most popular movie?
2) Which is the best and the worst movie?
3) Average rating of each movie.
4) Helpfulness of the reviews.
5) Interest of a particular user.
6) Recommending movies.
7) Sentiment analysis.

The proposed method has used PySpark to analyze the movie reviews data in real-time. Spark is a framework which has positives of both Hadoop and Storm. As processing is in-memory in Spark it is 100 times faster than Hadoop. The architecture of Spark is shown in Fig. 1. Spark uses Resilient Distributed Datasets (RDD) to achieve the efficiency. Instead of remembering the state of the data it remembers which operations have resulted in current RDD. By remembering the operations performed to get the RDD, it achieves fault tolerance. Spark is a single framework which supports both batch processing and real-time processing. Spark supports machine learning and interactive algorithms. Unlike Hadoop, Spark supports Scala, Java, Python, and R as first citizen languages. Python is one of the fast developing languages with vast library support.

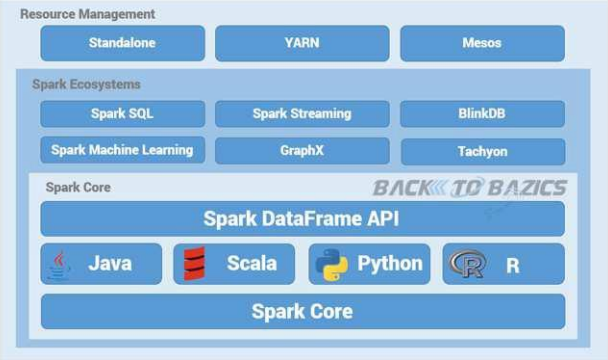

Figure 1: Spark Architecture [2]

II. LITERATURE SURVEY

Building a recommender system requires knowledge about customer preferences and emotions. The dataset is rare for building such a system. A personalized recommendation system is developed based on past feedback. The author also finds the changing fashion trend over a period of time [3]. A cross-sell opportunity is explored by analyzing the human buying pattern of clothes and shoes together. The task is achieved by crunching large dataset which is visual in nature[4]. Amazon reviews the book. Whether the earlier reviews of the book gets a more favorable vote than the later reviews are analyzed [5]. The quality of review reveals the opinion of the customer with the product that has been purchased by them. The new customer can purchase based on the quality review. A quality review is the one which is helpful and which is not

Revised Manuscript Received on March 10, 2019.

Prakash K. Aithal, Department of Computer Science and Engineering Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. (prakash.aithal@manipal.edu)

Dinesh Acharya U., Department of Computer Science and Engineering Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India.

Geetha M. Department of Computer Science and Engineering Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India.

populated maliciously. Feedback on Amazon products is distilled, and helpful reviews are statistically analyzed[6]. The reviews are useful for the companies, reviewers, and readers of the review. The companies can get the pulse of the customers, reviewers are given incentives and customers are reading the review will come to know about the product by comparing it with the other product reviews. An automatic review mining and summarizing of the lengthy reviews for the benefits of all stakeholders is designed[7]. Whether the movie is good or bad is decided by applying the naive base classifier along with neural network. The unigram feature is used for the classification[8]. Exactly 50000 reviews of 15 products of Amazon are considered to classify whether the product is liked by the customer or not. The naive base classifier is used for classification. A product such as Kindle are the subject of the study[9]. A support vector machine based classifier is used to classify the sentiments of the people into positive or negative on movie reviews data[10]. To analyze semistructured and unstructured data BigData technologies need to be used.

BigData is the data which has high Volume which do not fit into the storage media, High Velocity, Variety, Veracity and Value. Veracity determines how valid is the data. The five V's of BigData and their meaning is depicted in Fig. 2. BigData analytics aid informed decision making[11].

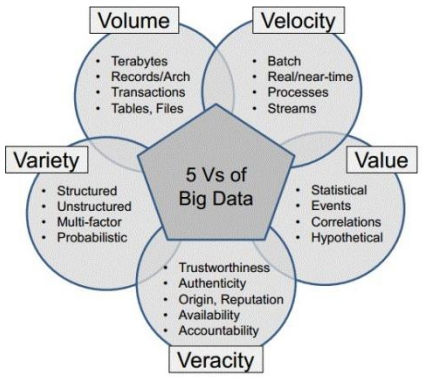

Figure 2: V's of BigData [12]

BigData analytics has following steps

1) Problem Formulation
2) Data Collection
3) Data Storage and Transformation
4) Data Analysis
5) Visualization
6) Evaluation

Apache Spark is an opensource BigData analytics framework. It supports processing of streaming data, graphs and SQL queries[13]. The BigData will transform the way governments, corporate houses and industries work. The economic value of BigData is explored[14]. The user feedback which is unstructured can be exploited to recommend using machine learning techniques[15]. The real-time Twitter data has been analyzed in real-time[16].

III. RESEARCH METHODOLOGY

A summary of research design is pictorially represented in Fig. 3.

- **Approach:** Quantitative approach.
- **Strategy:** An empirical study is conducted.
- **Methodological paradigm:** Positivism
- **Data Collection:** Secondary Data from Stanford University data repository is chosen. The data contains reviews of different movies

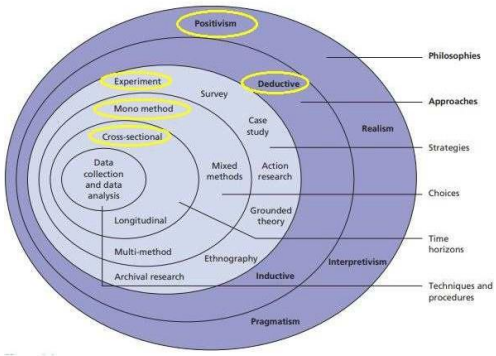

Figure 3: Research Framework[17] The dataset consists of following fields

1) **User id:** Unique id of the reviewer
2) **Product id:** Unique id of the movie
3) **Review:** Review written by the reviewer
4) **Profile name:** Name of the reviewer
5) **Helpfulness:** The helpfulness of the review to the user
6) **Time:** Timestamp of the movie review
7) **Score:** The rating given by the reviewer

Fig. 4. gives the block diagram of movie review analytics system.

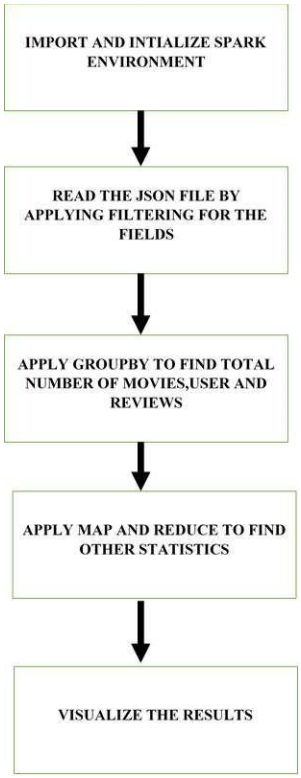

Figure 4: Block Diagram of the Movie Review Analytics System

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The spark environment is first imported using findspark. After findspark is imported spark environment is initialized. The second step is to create SparkContext. Once the Spark-Context is created one is ready to use spark under Python. Later all the fields in the Java Script Object Notation(JSON) file has been read by applying filter and a map function. To get the number of users groupBy clause to user_id has been applied and to get the number of movies groupBy clause to product_id has been applied. Total reviews are obtained by applying count member function. The number of reviews for a particular movie is obtained by applying map,mapvalues and reduceByKey operations. A similar approach is followed to find the average rating and number of movies reviewed by the particular user. Pandas, math, numPy, datetime, and matplotlib libraries are imported for visualization and converting RDD to data frame. Frequencies of words in the review are computed by applying map, flatmap, and reduceByKey functionalities. The words with frequency more than ten are obtained by applying a filter. Visualization is done through line and histogram plots. Fig. 5. Depicts real-time algorithm for updating of movie statistics. The proposed method has divided the 50000 records into six parts. The divided dataset of 25000 records is used for training and testing the model and getting the statistics. Once the model is built, it is incrementally tested on divided data set of 5000 records each to get the best rank and number of iterations for the model. The Alternating Least Squares(ALS) model is a collaborative filtering model. In collaborative filtering, one will predict the interests of a user based on interests of the similar user. In movie dataset if user-A has written similar reviews as user-B then both users have the same taste and user-A movie can be recommended to user-B and vice-versa.

Figure 5 Real-Time Updation of Movie Statistics

Input Incremental Movies Data
Output Movie Statistics
1: **procedure** MOVIE STATISTICS
2: Historical movies review records is read into an RDD
3: Find out the best movie by counting highest number of reviewers
4: Find out the best reviewer based on the highest number of reviews written
5: Find out average rating of a movie
6: Find out the best rank and number of iteration to train the recommendation model
7: Repeat the steps 3 to 6 on incremental movie dataset
8: **end procedure**

IV. RESULTS

The number of reviewers have increased over the years as depicted in Fig. 6. The blue line on top represents yearly count of reviews, the green line in the middle represents quarterly count of number of reviews and saffron line at the bottom represents the monthly count of number of reviews. The average rating of four bottom most least popular movies is depicted in Fig. 7. Average rating is helpful as it will give insight into quality and popularity of movie. The user A3LZGLA88K0LA0 has reviewed the highest 112 movies as depicted in Fig. 8. The people who review large number of movies can be given incentives to review further. Total number of users reviewing the movie is depicted in Fig. 9.

It shows the popularity of the movies. Fig. 10. depicts the frequency of words in the reviews which can be utilized to analyze the sentiment of the user. Table1 gives Statistics about the movie data. Time taken to testing of the incrementally updated data for the given model is depicted in Table 2. Time taken for training and testing of model on incrementally updated historical data is given in Table 3. The recommendation model is best trained with rank 5 and number of iteration 5. Root Mean Square error of the model decreases with increase in the available dataset.

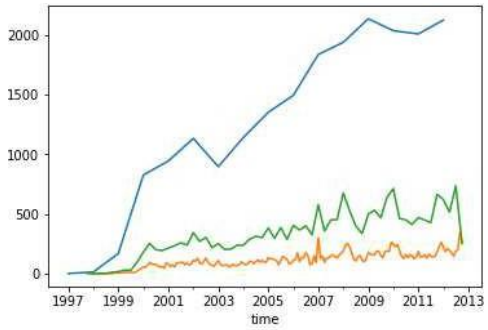

Figure 6: Monthly-Quarterly-Yearly-Count of Reviews

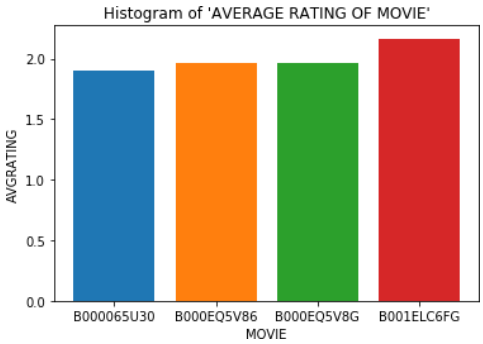

Figure 7: Average Rating of Movies

V. CONCLUSION AND FUTURE WORK

A system to analyze the movie review data is implemented using PySpark. The statistics of best movie, worst movie, average movie rating, and best reviewer is useful for all the stake holders. The recommendation system for different users is developed based on their reviews for earlier movies. The proposed work can be extended to user sentiment analysis.

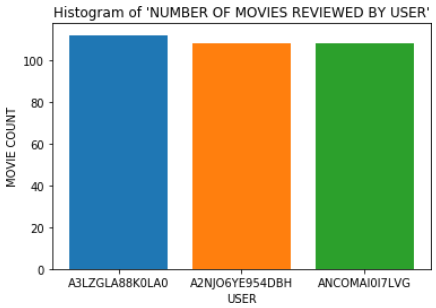

Figure 8: Number of Movies Reviewed by User

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

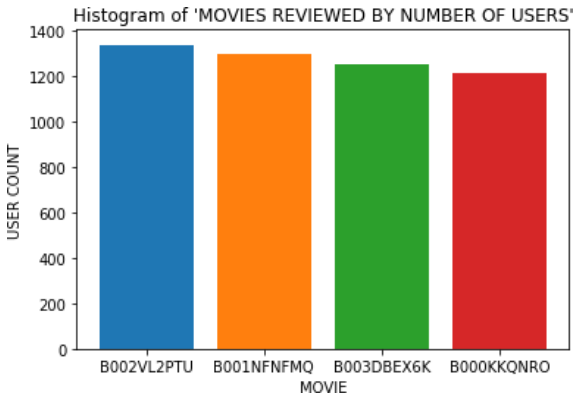

Figure 9: Total Number of Users Reviewing the Movie

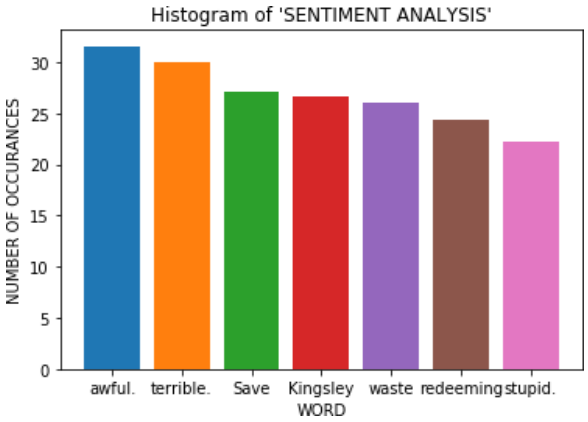

Figure 10: Frequency of Words in the Review

Table I: Statistics about Processed Data

	Count
Total Number of Reviews	50000
Number of Movies	1539
Number of Users	36409
Number of users reviewing the most popular movie	1338
The highest number of reviews written by single person	112

Table II: Time taken for incremental testing of data in real- time

Number of Records	Time Taken in Seconds
4096	1.96
5120	1.10
5120	1.05
5120	0.96
5120	0.95
5120	0.94

Table III: Time taken for training and testing of incremen- tally increasing data in a real-time recommendation system

Number of Records	Time Taken in Seconds
25000	36.07
30000	39.83
35000	42.56
40000	45.61
45000	48.32
50000	51.41

REFERENCES

1. https://snap.stanford.edu/data/web Amazon.html.
2. http://backtobazics.com/big-data/spark/understanding-apache-spark architecture/.
3. Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collab- orative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517. International World Wide Web Conferences Steering Committee, 2016.
4. Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2015.
5. Timothy Wong. *Exploratory Data Analysis of Amazon. com Book Reviews*. PhD thesis, 2009.
6. Sumit Kawate and Kailas Patil. An approach for reviewing and ranking the customers'reviews through quality of review (qor). *ICTACT Journal on Soft Computing*, 7(2), 2017.
7. Li Zhuang, Feng Jing, and Xiao-Yan Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50. ACM, 2006.
8. Lina L Dhande and Girish K Patnaik. Analyzing sentiment of movie review data using naive bayes neural classifier. 2014.
9. Callen Rain. Sentiment analysis in amazon reviews using probabilistic machine learning. 2013.
10. Neelu Rani, Nishant Singh, Sujay Pawar, et al. Sentiment analysis by data mining of past movie reviews/ratings. *Impe- rial Journal of Interdisciplinary Research*, 3(6), 2017.
11. Xiaomeng Su. Introduction to bigdata. NTNU.
12. *http://iihtofficialblog.blogspot.com/2014/07/5-vs-of-hadoop- big-data.html.*
13. Debi Prasanna Acharjya. A survey on big data analytics: challenges, open research issues and tools. 2016.
14. Liran Einav and Jonathan Levin. The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1):1–24, 2014.
15. Christoph Lofi and Philipp Wille. Exploiting social judge- ments in big data analytics. 2015.
16. B. Yadranjiaghdam, S. Yasrobi, and N. Tabrizi. Developing a real-time data analytics framework for twitter streaming data. In *2017 IEEE International Congress on Big Data (BigData Congress)*, pages 329–336, June 2017.
17. http://theservicemanagement.blogspot.com/2016/09/research onion.html.