INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 2)

Available online at: www.ijariit.com

Specialized data Analysis of stack overflow

Akash Tyagi
tyagiaakash001@gmail.com
IMS Engineering College, Ghaziabad, Uttar Pradesh

Agam Tiwari
agamtiwari.1994@gmail.com
IMS Engineering College, Ghaziabad, Uttar Pradesh

Abhisheke
abhishekkumar3727@gmail.com
IMS Engineering College, Ghaziabad, Uttar Pradesh

Vivek Jain
vivek.jain@imsec.ac.in
IMS Engineering College, Ghaziabad, Uttar Pradesh

ABSTRACT

This research paper analyzes and recommends a Question and Answer site for programmers, Stack Overflow, that dramatically improves on the utility and performance of Q&A systems for technical domains. It would be very difficult to analyze the real-time data collected from the users by using traditional techniques. Using a mixed methods approach that combines statistical data analysis with user interviews, we seek to understand this success. Our main focus is to develop a system that provides recommendations and statistical analysis on the basis of data of thousands of users.

Keywords: *Big Data, Hadoop, Stack Overflow, Map Reduce, Pig, Apache Spark, Scala, Apache Sqoop, Apache Hive, Apache Oozie, Data, Clustering.*

1. INTRODUCTION

Big Data encloses everything from structured data (tables in rows and columns like DBMS tables) to unstructured data like email attachments, images, PDF documents etc. Big Data can be termed as data which is beyond the ability of databases to capture, store, manage and analyze. Hadoop is used in organizations to make decisions based on comprehensive analysis of multiple variables and data sets, rather than a small size of data. The ability to process large sets of different kind of data gives Hadoop users a complete view of their customers, operations, opportunities, risks, etc.

Stack Overflow is a Q & A site answering questions related to programming. Over 92% of Stack Overflow questions about expert topics are answered — in a median time of 11 minutes. The recommendations we will be providing will be based on the basis of the statistical analysis so that the website can improve its quality of content and provide suggestions according to their relevant field. By evaluating the data over a long period of time we can study the various patterns related to it and various customizations can be made to the website according to it on a timely basis.

2. TECHNOLOGIES USED

2.1 Apache Hadoop

Hadoop is an open source, a Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.

2.2 Map Reduce

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks namely Map and Map Reduce. The map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

2.3 Apache Pig

Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called PigLatin. Pig can execute its Hadoop jobs in MapReduce, ApacheTez, or Apache Spark.

2.4 Apache Oozie

Apache Oozie is a server-based workflow scheduling system to manage Hadoop jobs. Workflows in Oozie are defined as a collection of control flow and action nodes in a directed acyclic graph.

2.5 Apache Spark

Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data workers to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets. With Spark running on Apache Hadoop YARN, developers.

3. LITERATURE REVIEW

Kiran kumara Reddi & Dnvsl Indira et.al. Enhanced us with the knowledge that Big Data is a combination of structured, semi-structured, unstructured, homogenous and heterogeneous data [1]. Under this model, these transfers are relegated to low demand periods where there is a large amount of idle bandwidth available.
Bhandarkar, M.[2] discussed how Apache Spark can be used for developing large-scale intensive applications. He also describes several frameworks and utilities developed using Hadoop that increase programmer-productivity and application performance.
Jimmy Lin et.al. used Hadoop which is currently the large-scale data analysis " hammer" of choice, but there exist classes of algorithms that aren't " nails" in the sense that they are not particularly amenable to the MapReduce programming model.[3]. He focused on finding simple non-iterative algorithms to find solutions to the same problem.
Wei Fan & Albert Bifet et.al. [4] Introduced Big Data as the ability to reduce the redundancy and extract useful information from large-sized datasets which were otherwise not possible to do. They also raised certain challenges such as compression, visualization etc.
Albert Bifet et.al. [5] Stated in his paper that streaming data analysis in real time is becoming the fastest way to attain knowledge thereby allowing organizations to react quickly in case of a problem or to improve their performance. He also instructed that our handling of the huge amount of data is dependent on software framework like Apache Spark
Wang et.al [6] suggested that Map Reduce provides a simplified model for data-intensive business as well as scientific applications
Borthakur et.al [7] Apache Hadoop platform is the first user-facing application of Facebook which is built on Apache HBase is a database-like layer built on Hadoop designed to support billions of messages per day.
Das, S. et.al [8] in his paper told that enterprises are maintaining petabytes of data in their data repositories the ability to apply sophisticated statistical analysis methods to this data is becoming essential for marketplace competitiveness.
Liu et.al [9] Hadoop framework has been widely used in various organizations to build large-scale, high-performance systems. However, Hadoop distributed file system (HDFS) is designed to manage large files but due to the presence of a large number of small files, it suffers a performance penalty.
Luca Cagliero & Alessandro Fiori [10] The increasing availability of user-generated content coming from online communities allows the analysis of common user behaviors and trends in social network usage which can be easily maintained in data repositories and analyzed for a better recommendation

4. DATASETS

Dataset has been downloaded from official StackOverflow blog insight.stackoverflow. The command contains data of around 1,59,000 users. StackOverflow conducts a survey every year to find out interests of their users and on basis of those answers, various analysis can be made using the results.

5. METHODOLOGY

Our dataset contains answers of users to around 20 Questions asked during the survey. First, the dataset is feed to HDFS for making calculations on it. As the dataset contains n no. of answers and we want to work with answers to just some specific questions the dataset will be converted into a set of files that contain answers only of those questions that are needed by us and rest of them is neglected. This work is done using Apache Pig. Now as we have the result to our questions various combinations of the different result set can be combined and analysis is made on those result to find out certain fact and figures by using Apache Pig and Spark. Pig helps us is collection and combination of our data according to our needs and the resultant calculation on this dataset is done with the help of Apache Spark.

6. RESULT

Various result analysis and recommendation can be made from the analysis of dataset such as countries with maximum no of site users, salary of person with comparison to languages known, highest paying job in a country, occupation of people with respect to their country, preferred os type by different type of developers, comparison of age to job satisfaction. This result can be used as a recommendation for StackOverflow site so that respective steps can be taken according to the result to improve site quality, content, and marketing. Results from this analysis can also be used to estimate the trends occurring in a specific country in terms of technology. our result also focuses on the fast processing and manipulation speed of the apache spark, if the same process has to be done using map reduce this could have taken much longer time the spark thus increasing latency.

7. CONCLUSION AND FUTURE WORK

The need to process enormous quantities of data has never been greater. Not only are terabyte- and petabyte-scale datasets rapidly becoming commonplace, but there is a consensus that great value lies buried in them, waiting to be unlocked by the right computational tools. In the commercial sphere, business intelligence, driven by the ability to gather data from a dizzying array of Sources. Big Data analysis tools like Map-Reduce over Hadoop and HDFS, promises to help users of Stack Overflow get a better user experience by showing them only the relevant results which will attract new users. By doing the users do not get any unnecessary suggestions that are not relevant to them.

8. ACKNOWLEDGMENTS

Our thanks to the teachers who have contributed towards the development of this paper.

9. REFERENCES

[1] Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data: survey" IEEE Transactions on 52(8) (Aug.2013) 2348
[2] Bhandarkar, M. (2010) MapReduce Programming with Apache Hadoop. Parallel &Distributed Processing (IPDPS) IEEE, 19-23 April 2010.
[3] Jimmy Lin "MapReduce Is Good Enough?" The control project. IEEE Computer 32 (2013).
[4] Wei Fan, Albert Bifet. Mining big data: current status, and forecast for the future, ACM SIGKDD Explorations Newsletter, Volume 14 Issue 2, December 2012
[5] Albert Bifet "Mining Big Data In Real Time" Informatica 37 (2013) 15–20 DEC 2012
[6] Graph classification based on pattern co-occurrence. N Jin, C Young, W Wang.Proceedings of the 18th ACM conference on Information and knowledge … (2009)
[7] Dhruba Borthakur, Apache hadoop goes realtime at Facebook, Proceedings of the 2011 ACM SIGMOD International Conference on Management of data.
[8] Das et al., 2010, Das S., Sismanis Y., Beyer K.S., Gemulla R., Haas P.J., McPherson J., Ricardo: Integrating R and Hadoop, In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD '10)*, 2010
[9] Huang Z, Shen H, Liu J, Zhou X (2011) Effective data co-reduction for multimedia similarity search. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM
[10] Cagliero, L., & Fiori, A. (2013). Dynamic social network mining: Issues and prospects. In *Data Mining in Dynamic Social Networks and Fuzzy Systems* (pp. 122-144). IGI Global.

