LOG BASED ANALYTICAL ENGINE USING HADOOP AND SPARK

Prof. Avinash Taskar¹, Roshan Patil², Dwineeth Shetty³, Jagruti JadHAV⁴ and Vrushali Ahirrao⁵

¹,²,³,⁴,⁵ Department of Computer Engineering
Sandip Institute of Engineering and Management, Nashik (India)

Abstract—Due to digitalization in today's modern era log file analysis has become a necessary task to track user behavior and get important knowledge based on it. Log files are generated at massive rate and to analyze them is tedious task. In order to analyze large dataset user need effective solution for integrating the data and also parallel processing for that spark's machine learning is used which will give power to run machine learning algorithm in distributed environment on large volume dataset (Big Data).

Keywords—Analytical Engine, Hadoop, HDFS, Machine learning, Spark.

I. INTRODUCTION

Log contains a large amount of information. There are many different types of logs available like surfing log, error log, etc. These logs are stored on server. The log production rate can lead up to several Terabytes (TB) or PetaBytes (PB). Log based analytical engine is an engine which analyzes the gathered logs in such a way that we can understand the number of access-users, system operation status, user-behavior, etc. There have been some log analysis tools but either they are standalone or have the limitation of data scale. With today's increasing speed of development of internet technology, how to deal with large scale log data has become a challenge.

[1]Hadoop is a popular open source computing distributed framework which has been mainly used to store and analyze data. It also provides a distributed file system called as HDFS. Hadoop provides features like horizontal scalability and fault-tolerance. [3] Hive is a analytical tool provided by Hadoop which provides facilities like writing, reading and managing large size of datasets that are stored in HDFS using SQL statements. Hadoop breaks up the log files into number of blocks and distribute them to the nodes which are present in Hadoop cluster. [4] The analytical engine needs stable and massive data processing ability as well as efficiency required to work under variety of scenarios. Hadoop does not provide fast performance and also not suitable for real time applications because each time it performs operations from hard disk.

[2]Spark is an open source in-memory distributed cluster system. Using Spark the speed required for data analysis is increased as it uses main memory to perform operations. It is used for interactive data mining as well as iterative algorithms like graph mining algorithm and machine learning algorithm.

Unlike spark, for working with the structured-data Apache Spark SQL module is used which lets you query structured data inside the Spark programs, using SQL. Spark SQL reuses the Hive front-end and meta-store giving us a full compatibility with the existing Hive data, queries etc. It also includes column storage, cost based optimizer and code-generation to make the queries execute faster.

II. MACHINE LEARNING

Machine learning is a computer science’s subfield that has emerged from the study of computational learning theory in artificial intelligence and pattern recognition. In other words it means that it is the field of study that helps the computers to get the ability of learning without being explicitly programmed. This is mainly used to understand our analytical engine in a proper way, as it is co-related to computational statistics. Machine learning sometimes is also closely related with
Data Mining in which the latter sub field is mainly focused on data which is explanatorily analyzed and in other words it is called as Unsupervised Learning.

2.1 Supervised learning

Supervised learning is a task of machine learning that follows inferring a function from labeled training data. A supervised algorithm is an algorithm which analyzes the training data and then after all the execution generates an inferred function. Hadoop and Spark are the two parallelized technologies which are basically used for testing the system whether it is horizontal scalable or vertical scalable or both by running two supervised learning algorithms i.e. KNN and SVM.

2.2 Unsupervised Learning

Unsupervised learning is also a task of machine learning that infers a function describing hidden structure from the data which is unlabeled. There are many approaches to unsupervised learning that includes clustering, hierarchical clustering, k-means, etc. and from these we are going to use clustering and k-means in our system. Clustering performs the in such a way that it groups a set of objects in the same group i.e., a cluster becomes more similar to each other as compared to other groups(cluster).

III. LITERATURE SURVEY

Log Analysis becomes an important field to study the behavior of user. Researches on log files has increased significantly, there are deficiencies in the comprehensive literature reviews of that research

Matei Zaharia ET. Al[5]: In this, a first-class primitive in-memory data sharing, Spark a powerful tool for Interactive data mining is used, as well as efficient runtime for complex machine learning & graph algorithms. By the same time, Spark’s which has the capability to call into existing Java libraries (through the Scala language) and to access any Hadoop-supported storage system i.e. HDFS (by reusing Hadoop’s input/output APIs) made it a pragmatic and great choice to complement Hadoop for large-scale data analysis.

Sayalee Narkhede ET. Al[6]: To summarize the data results for a particular web application, in this paper they gave a way of log analysis which will help us to improve the business strategies and also to generate statistical re-ports. HMR log analysis tool is used which provides the reports graphical showing hits for a web-page, user's behaviour and activity, also in which part of website users are interested, etc. From these generated reports business organisations and communities can decide which parts of the website need to be improved, which are the valuable and potential customers, from which geographical region or place website is getting maximum number of hits, etc, which helps us in designing future marketing plans.

Xiaokui Shu Et. Al[7]: In this paper, they intended to design and implement a light-weight distributed Framework, that consisted of a minimized set of components. The framework was different from general ones and was specially designed for Security based Log-Analysis. It also featured the streaming of log analysis very efficiently in the design.

N. Ramasubramanian Et. Al[8]: Here in this paper, stream logging application was designed which computed over a distributed file system and was used to facilitate applications that involved transaction based computing. Since, the computation was done using normal desktop systems and the transactions performed were very large in number which was considerably massive. The application helps us to reduce the number of special hardware's (server systems) required for computation. Hence the overall cost of computation is minimize and made considerably low.

Muhammad Ali Gulzar Et. Al[9]: In this, BIGDEBUG which is an interactive debugging primitives for the in-memory Data-Intensive Scalable Computing (DISC) framework. It enables the user to determine crash culprits and resolve the problem at runtime, it helps avoid program re-run from the scratch. It also scales with massive large data in the order of terabytes (TB), and improves fault localizability by order of millions than baseline Spark, and provides near up to 100% time-saving with respect to a post hoc instrumentation replay debugger.
Jorge L. Reyes-Ortiz ET. Al[10]: In this paper, Spark on Hadoop with in-memory data-processing which reduces the gap between Hadoop Map Reduce and HPC for Machine Learning. Here, MPI/Open MP on Beowulf that is high-performance oriented and exploits multi-machine and multicore infrastructures, and the Apache Spark on Hadoop which targets an iterative algorithm through in-memory computing. Results which are obtained from experiments with a particle data set show MPI/OpenMP outperforms Spark is better in terms of its processing speed and consistent performance.

Ilias Mavridis ET. Al[11]: To investigate the analysis of log files with the two most widespread computing frameworks in cloud computing, the well-established Hadoop and rising Spark. The two frameworks have developed, executed and evaluated realistic programs for analyzing logs. The two frameworks have the common goal of parallel processes execution on distributed files or other input files. The differences and similarities between Hadoop Map Reduce and Apache Spark and evaluates the performance of them.

IV. PROPOSED SYSTEM

To make Analytical Engine for Log files very generic and allow user to see the statistical analysis in the form of charts for their huge data set. Apache Hadoop used to store this unstructured log files in database using Hive in structured manner and Apache Spark is used to analyze that structured data set very efficiently and effectively. Let M be a Analytical Engine give by

\[ M = \{ D, N, F, A, L, R \} \]

D is input data set from database

\[ D = i_1, i_2, i_3 \ldots \text{in} \]

\[ i_1 = \text{host} \]

\[ i_2 = \text{URL} \]

\[ i_3 = \text{date and time} \]

N is normalized form of data set

\[ I \in D \]

Figure 1
N=j_1, j_2, j_3... j_m where m<=n

Figure 2

F is number of feature required for analysis
F=k_1, k_2, k_3...k_p, where p<m

Figure 3

A is log association matrix with respect to F
L is Machine Learning model

Figure 4

R is resultant data set.

Figure 5: Mathematical model of proposed System

V. FUTURE SCOPE

In addition to this proposed system it can be extend for supporting more complex log files and also build a very flexible system to analyze the various kinds of log data.

For security of web server it can be modify to check the malicious attacks and user behavior based on log.

VI. CONCLUSION

Analytical system has become ubiquitous. People use this system for analyzing the log data very efficiently to get aware of user behavior. In this paper we proposes very efficient and flexible system for log analysis which can take any unstructured or structured log that user provides.

REFERENCES


