Error Log Analytics using Hadoop

Ms. Akshaya A. Navaghane[1] Ms. Jyoti A. Patil[2] Ms. Nikita C. Patil[3] Ms. Ashwini J. Dive[4] Prof. R. S. Kamble[5]

[1,2,3,4]Student [5]Assistant Professor

[1,2,3,4,5]Department of Computer Science & Engineering

[1,2,3,4,5]Ahmedabad Institute of Technology, Ahmedabad, Gujarat, India

Abstract— Logs analysis is study and analysis of various logs generated by applications and various devices. Logs are integrated from various nodes or the servers. Whenever we access any application on system that time some data is record in background of system those records are termed as 'logs'. A meaningful information is stored in log file which is extract using hadoop tools and stored into hadoop framework. On the basis of log file data we can classify the logs data for a future use.

Key words: Data Mining, Predictive Analysis, Hadoop, Mapreduce, Flume, Sqoop

ABBREVIATION AND ACRONYMS

- Hadoop: It is a framework used for storing and processing heterogeneous data such as structured, semi-structured and unstructured data.
- HDFS: It is one of the main component of the Hadoop. It stands for Hadoop Distributed File System. It is used to store heterogeneous data in database.
- Mapreduce: It is another component of Hadoop. Mapreduce is an algorithm where mapping consist of java program used for processing, while reducing is used to compress the size of data.
- Flume: Flume is a tool used to extract data from various sources and store it in HDFS.
- Sqoop: Used for export and import of data from one database to other such as from HDFS to MySQL.

I. INTRODUCTION

Log files provide valuable information about the applications and devices. The log files are used by in many sectors like banking sector, online reservations, online shopping etc. These logs are useful for a developer to monitor application and find out any problems and to recover those. The structure of error log file varies from one application to another application. Traditional processing of logs takes huge amount of time and therefore it can be tedious work. Since to reduce the work use the big data hadoop technology for analysis the log files. Analytics involves to find out understandable format of log files.

Logs are in the form of semi-structured, unstructured or structured data. Using the Hadoop we can store those data into single database known as HDFS. Predictive Analytics is function which is used to predict the future status of application based on current status of application**.**

II. PURPOSE

Now-a-days log files plays very important role in all type of business strategy. Computer System involves large number of log files. Log files are files that contain lists of activity which are performed on application or the device. These log record are useful to administrator to monitor system like who has accessed the system and where from it was accessed, which type of websites have been accessed and also the health of applications and devices is stored in log files.

Here some log files can be found:

- Operating Systems
- Web Servers
- Web Browsers
- Applications

There are 2 types of log files- Access log and Error lo,

- Access Log: Access logs are generated by servers. It consist various parameter such as the client IP address, URL, response code etc.
- Error Log: Error logs are generated by application. It is used for analyzing the health of the application. These logs can be used to analyzing the errors and recovery from it. Error logs stores information such as Timestamp, application name, Error message ID, Error message details. When application is running it generate the error logs. It contains errors, warning and information. Error logs in heterogeneous format which looks something like this.

Fig. 1: Error Log

Fig. 2: Information Log

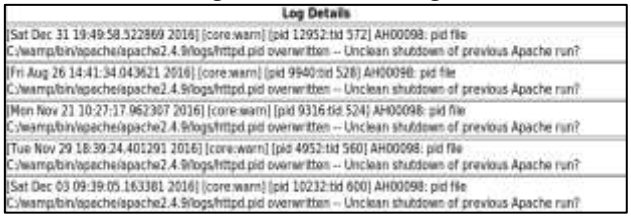

Fig. 3: Warning Log

III. ADVANTAGES

Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.

Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.

Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.

Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

IV. PROJECT PERSPECTIVE

Error Log Analytics using Hadoop is used to gather error log and perform analytics on gathered log data. Project is based on open source software such as Linux, Hadoop, Flume, Sqoop. The goal of error log analytics is to improve quality of application and reduce the future work.

V. PROJECT MODULES

- User Authentication**:** First module is user authentication where user can interact with the system. The User will login on the application and perform some functionality on error log and through that see the graphical view of project result.
- Data Extraction and Integration: In second module, we extract data from various sources and store heterogeneous data in hadoop distributed file system (HDFS).
- Pattern Matching: The keywords stored in a file are compared with the pattern in the source file. These are extracted from the source file and classified based on the keywords into various files.
- Predictive Analytics: The error logs are analyzed and this information is used to predict the performance of application.
- Visualization: The analysis result represent in the form of graph.

VI. OPERATING ENVIRONMENT

Hadoop consist two components HDFS and Mapreduce. HDFS is used to store heterogeneous data. Mapreduce is used to perform mapping and reducing process on data which is stored in HDFS.

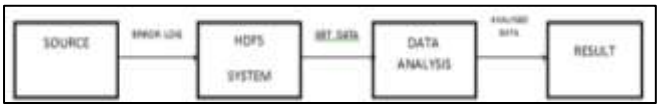

Fig. 4: Block Diagram

The logs are collected from the servers using tool such as flume which extracts the data from various sources and send it to sink though channel. Sink is an temporary storage such as buffer which in turn sends the data to HDFS.

The data from HDFS is used for analysis and we perform Mapreduce task on those log data. After performing analysis final result is created in form of graphical representation of data.

A. System Architechture

Streaming log data into HDFS is achieved using log collector tools such as Flume. Which efficiently collect and aggregate huge amounts of data. There exist 3 components in the Data Flow model of a log collector agent. They are Source, Channel and Sink. Server log data will be heterogeneous and semi-structured in nature. This data is collected and aggregated. ETL operations are to be performed and the missing values are filled by averaging the neighboring values present in the log entry.

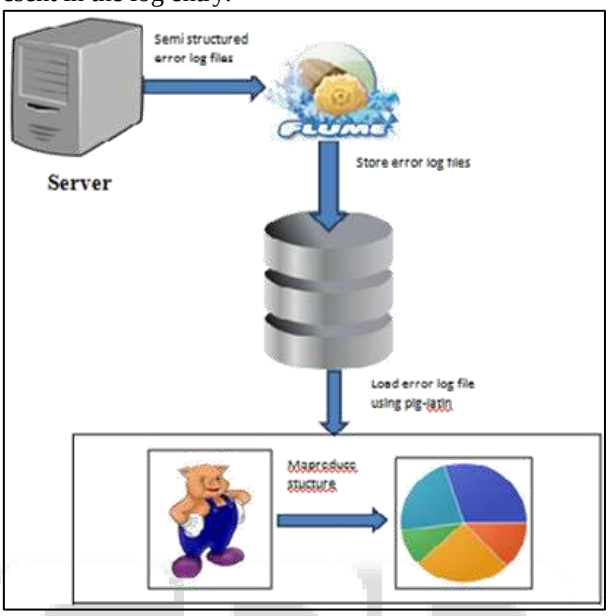

Fig. 5: System Architecture

The various fields include the timestamp, severity, the IP address of the machine that generated the error logs. Transformation of data is done to get the data into a proper structure for analytics and querying purpose. After the ETL operations are performed, the error log file is brought into a uniform homogeneous format. In order to perform analytics on this data, it is loaded into a warehouse. Getting the sum of errors and warnings on the severity attribute is achieved using the Grouping operation. Cubing operation generates aggregates for all combinations of values in the selected columns. Historical error log data is maintained in the warehouse, which is used in various Analytics and Business Intelligence techniques.

VII. RESULT

Sign up provides registration of new user in database. The sign up is required to keep secure each users file and to provide privileges to intended user only.

Fig. 4: Signup Page

Once the user is successfully logged in, the user will be provided with classified files based on the keywords provided. As shown in snapshot, the classification of various files based on keyword is done for better management.

Fig. 5: Home Page

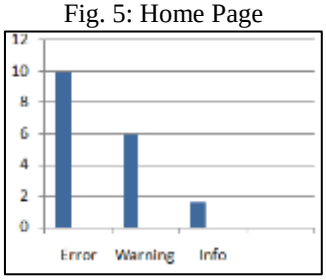

Fig. 6: Graph Presentation of Result

In the log file contain the error, warning and information which is generated by server or the application. The above figure shows the count of error, warning and information present in the error log file. That result used to predict the future values.

VIII. CONCLUSION

This report describes a detailed view of Hadoop framework used to process big data. It even gives a description of how the log file is processed for exceptions and errors. This report describes how a log file is processed using map reduce technique. Hadoop framework is used as it is beneficial for parallel computation of log files. As described in the report, the framework makes use of tableau tool for pictorial representation of log files accessed by the users.

REFERENCES

[1] Chuck Lam, "Hadoop in Action", Manning Publications.
[2] Charles Elkan, "Predictive Analytics and Data Mining", 2013.
[3] Hortonworks, "Chapter 2: Understanding the Hadoop Ecosystem".
[4] The Apache Software Foundation, "Flume User Guide". (IJU), Vol.4, No.3, July 2013.
[5] G.S.Katkar, A.D.Kasliwal, "Use of Log Data for Predictive Analytics through Data Mining", Current Trends in Technology and Science, ISSN: 2279-0535. Volume: 3, Issue: 3(Apr-May 2014).
[6] W.Peng, T. Li, S.Ma, "Mining Logs Files for Data-Driven System Management", Florida International University.
[7] A. Bruckman, "Chapter 58: Analysis of Log File Data to Understand User Behavior and Learning in an Online Community", Georgia Institute of Technology.

