2nd International Conference on Emerging Trends in Engineering, Technology, Science and Management
Lakshmi Narain College of Technology, Bhopal
20th May 2017, www.conferenceworld.in
(ICETETSM-17)
ISBN: 978-93-86171-45-0

LOG FILE ANALYSIS USING HADOOP AND ITS ECOSYSTEMS

Vandita Jain[1], Prof. Tripti Saxena[2], Dr. Vineet Richhariya[3]

[1]M.Tech(CSE) ,LNCT, Bhopal M.P.(India)*

[2]Prof. Dept. of CSE, LNCT, Bhopal M.P.(India)

[3]Prof. & Head, Dept. of CSE, LNCT, Bhopal M.P.(India)

ABSTRACT

In view of the fact that clusters used in large scale computing are on the rise, ensuring the wellbeing of these clusters is of paramount significance. This highlights the importance of supervising and monitoring the cluster. In this regard, many tools have been contributed that can efficiently monitor the Hadoop cluster. The majority of these tools congregates necessary information from each of the node in the cluster and takes it for processing. These diagnosis tools are mostly post execution analysis tools. This paper presents an exploratory assessment of the different log analyzers used for failure detection and monitoring in Hadoop. In this we proposed a various hadoop ecosystems through which we can analyse the complex hadoop logs.

Keywords-- Hadoop, HDFS, Mapreduce, Log analyzer, Hadoop ecosystems.

I. INTRODUCTION

Log files [3] provide valuable information about the functioning and performance of applications and devices. These files are used by the developer to monitor, debug, and troubleshoot the errors that may have occurred in the application. Manual processing of log data requires a huge amount of time, and hence it can be a tedious task. The structure of error logs vary from one application to another. Since Volume, Velocity and Variety are being dealt here, Big Data[5] using Hadoop is used. Analytics [7] involves the discovery of meaningful and understandable patterns from the various types of log files.

Error Log Analytics deals about the conversion of data from semi-structured to a uniform structured format, such that Analytics can be performed over it. Business Intelligence (BI) functions such as Predictive Analytics is used to predict and forecast the future status of the application based on the current scenario. Proactive measures can be taken rather than reactive measures in order to ensure efficient maintainability of the applications and the devices.

II. PURPOSE

A large number of log files [4] are generated by computers nowadays. A Log File is a file that lists actions that have taken place within the application or device. The computer is full of log files that provide evidence of what is going on within the system. Through these log files, a system administrator can determine what Web sites

(ICETETSM-17)
ISBN: 978-93-86171-45-0

have been accessed, who accessed and from where it was accessed. Also the health of the application and device is recorded in these files. Here are a few places where log files can be found:

- Operating systems
- Web browsers (in the form of a cache)
- Web servers (in the form of Access logs)
- Applications (in the form of error logs)
- E-mail

Log files are an example of semi-structured data. These files are used by the developer to monitor, debug, and troubleshoot the errors that may have occurred in an application. All the activities of web servers, application servers, database -servers, operating system, firewalls and networking devices are recorded in these log files.

There are 2 types of Log files - Access Log and Error Log. This paper discusses the Analytics of Error logs.

Access Log files contain the following parameters – IP Address, User name, visiting path, Path traversed, Time stamp, Page last visited, Success rate, User agent, URL, Request type.

1. Access Log records all requests that were made of this server including the client IP address, URL, response code, response size, etc.

2. Error Log records all the details such as Timestamp, Severity, Application name, Error message ID, Error message details.

Error Log is a file that is created during data processing to hold data known to contain errors and warnings. It is usually printed after completion of processing so that the errors can be rectified. Error logs contain the parameters such as:

- Timestamp (When the error got generated).
- Severity (Mentions if the message is a warning,

error, emergency, notice or debug).

- Name of application generating the error log.
- Error message ID.
- Error log message description

III. HADOOP

The Apache Hadoop[10] project develops open-source software for scalable, reliable, distributed computing. The Apache Hadoop library is a framework that allows for the distributed processing of large data sets beyond clusters of computers using a thousands of computational independent computers and large amount (terabytes, petabytes) of data. Hadoop was derived from Google File System (GFS) and Google's Map Reduce. Apache Hadoop is good choice for twitter analysis as it works for distributed huge data. Apache Hadoop is an open source framework for distributed storage and large scale distributed processing of data-sets on clusters. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different clusters nodes. In short, Hadoop framework is able enough to develop applications able of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data. Hadoop MapReduce

(ICETETSM-17)
ISBN: 978-93-86171-45-0

is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

IV. LITERATURE REVIEW

In [1] , With rapid innovations and growing Internet population, petabytes of information are being generated every second. Processing these enormous data and analysing is a tedious process now-a-days. The amount of data in real-time is growing tremendously. Nearly 80% of the data is in unstructured format. Analysis of unstructured data in real-time is a very challenging task. Existing traditional business intelligence (BI) tools perform best only in a pre-defined schema. Most of the real-time data are logs and dont have any defined schema. Doing queries over these large datasets takes long time. During streaming of realtime data, much unwanted information is extracted from the data source causing overhead in the system. This results in an increase in the cost of construction and maintenance. Each and every second, new data streams keeps accumulating in the system consistently about whats going on in the world. Gathering these data and processing is an essential skill to know, for preparing a vital report. In this paper, they propose a Piece of News (PoN) end-to-end solution where they used the appropriate Hadoop components for real-time data analytics. Our aim is to extract the health data from the normal news data so that we can predict any real-time breakouts immediately. Rather than collecting all the news, they filtered only the important news based on certain threshold, thus reducing the cost. We compared historical data with real-time data which leads to take prompt action as we already knew the outbreaks from the previous data. One step ahead we can even detect any dangerous outbreaks before anyone else in the world.

In [2], Web Page access prediction is a challenging task in the current scenario, which draws the attention of many researchers. Predictions need to keep track of history data to analyze the usage behavior of the users. Web Usage behavior of a user can be analyzed using the web log file of a specific website. User behavior can be analyzed by observing the navigation patterns. This approach requires user session identification, clustering the sessions into similar clusters and developing a model for prediction using the current and earlier accesses. Most of the previous works in this field have used K-Means clustering technique with Euclidean distance for computation. The drawbacks of K-Means is that deciding on the number of clusters, choosing the initial random center are difficult and the order of page visits are not considered. In these they proposed research work uses hierarchical clustering technique with modified Levenshtein distance, Page Rank using access time length, frequency and higher order Markov model for prediction.

V. OBSERVATION

Log files are commonly used at customer's installations for the purpose of permanent software monitoring and fine-tuning. Logs are essential in operating systems, computer networks , distributed systems and storage files. Error od access statistics will be useful for fine tuning the application functionality, based on frequency of an error messages in the past times, we can forecast its occurance in the future and before it's occurrence on customer's application, if we can provide a fix the error, then customer satisfaction will be improved which in turn business will increase.

(ICETETSM-17)
ISBN: 978-93-86171-45-0

VI. PROBLEM DEFINITION

As the log files are being continuously produced in various tiers with different type of information, the main problem is to store and process tis much data in an efficient manner to produce rich insights into the application. A server or application will generate logs of size in large amount. We cannot store this much of data into a relational database system. RDBMS systems can be very expensive and cheaper alternatives like MYSQL cannot scale to the volume of data that is continuously being added. A better solutn is to store all the log file in HDFS [8] which store data on commodity hardware, so it will be cost effective to store huge volumes of log files into HDFS.

VII. PROPOSED WORK

For analysing these large and complex data required a power tool, we are using hadoop[10] which is a open source implementation of mapreduce, a powerful tool designed for deep analysis and transformation of very large data.

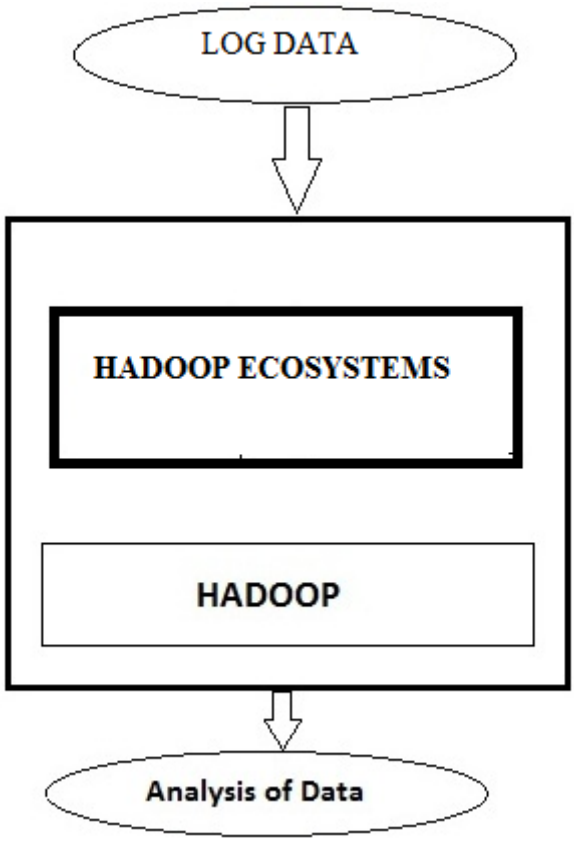

Figure1. Workflow Diagram

ISBN: 978-93-86171-45-0

This paper we design algorithm for handling the problems raised by the larger data volume and the dynamic data characteristics for finding and performing operation on log files. For analysing first we used standard platform as hadoop on single node ubuntu machine [9] to solve the challenges of big data through MapReduce framework where the complete data is mapped to frequent datasets and reduced to smaller sizable data to ease of handling ,After that we can use bigdata analytical tools [6] to refine such unstructured data and analyse the logs data using bigdata analytical tools.

VIII. PROPOSED METHODOLOGY

Our Steps or Algorithm Steps will follow:

1. First we collect the application error log files and store the lofs file in to HDFS.
2. The log files are collected from application. As the structure of the log files are unstructured, data cleansing plays a key role in bringing the collected log data into a uniform homogeneous structure format.
3. Patterns and useful information are extracted from the dataset and associations and relationships existing between the patterns are analyzed. Frequently occurring patterns are selected for Analysis.
4. Proactive measures can be taken using Predictive Analytics. The error log patterns are analyzed and the frequency of warning and error messages is used to predict the future performance of the overall system

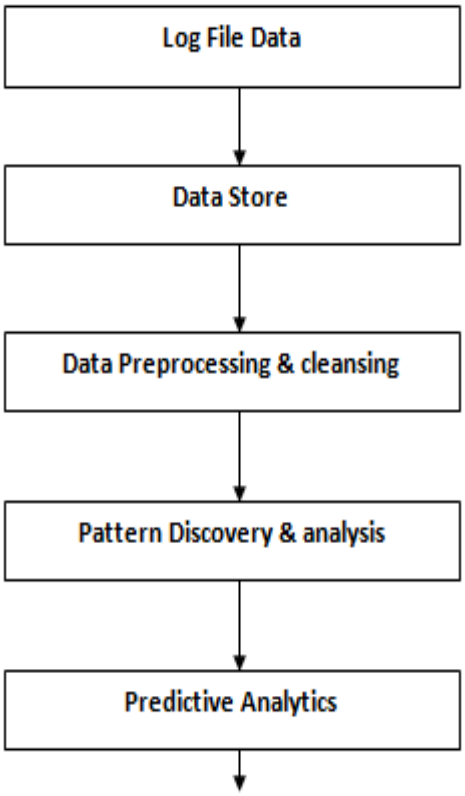

Figure 2. Analysis Steps

ISBN: 978-93-86171-45-0

IX. CONCLUSION

Hadoop being the most popularly used methodology for storage and processing of BigData, has several subprojects for failure monitoring and analysis. The majority of these tools seize the log files to capture the behavior of the cluster and the running application. They process the logs to retrieve the necessary information required for failure diagnosis, and some of the tools even support the failure recovery. In this paper we propsed hadoop and its ecosystem to analyze the complex and unstructured logs files.

REFERENCES

[1] Nikitha Johnsirani Venkatesan, Earl Kim, Dong Ryeol Shin, "PoN: Open Source solution for Real-time Data Analysis" in IEEE 2016,ISBN: 978-1-4673-9379-9

[2] Harish Kumar B T, Dr. Vibha L, Dr. Venugopal K R, "Web Page Access Prediction Using Hierarchical Clustering Based on Modified Levenshtein Distance and Higher Order Markov Model" in 2016 IEEE Region 10 Symposium (TENSYMP), Bali, Indonesia

[3] 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).

[4] Savitha K, Vijaya MS, "Mining of Web Server Logs in a Distributed Cluster Using Big Data Technologies", IJACSA, Vol. 5, 2014.

[5] McKinsey, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey & Company, 2011, http: //www.mckinsey.com/.

[6] White Paper Big Data Analytics Extract, Transform, and Load Big Data with Apache Hadoop-Intel corporation.

[7] Qureshi, S. R., & Gupta, A, "Towards efficient Big Data and data analytics: A review", IEEE International Conference on IT in Business, Industry and Government (CSIBIG),March 2014 pp-1-6.

[8] http://searchbusinessanalytics.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS.

[9] Michael G. Noll, Applied Research, Big Data, Distributed Systems, Open Source, "Running Hadoop on Ubuntu Linux (Single-Node Cluster)", [online], available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/

[10] Chuck Lam, "Hadoop in Action", Manning Publications.