DIFFERENT TYPE OF LOG FILE USING HADOOP MAPREDUCE TECHNOLOGY

Rahul Pawar [1], Rajkumar Bhosale[2], Archana Panhalkar[3]
Information Technology AVCOE Sangamner
Sangamner, India
[1]rahulpwr023@gmail.com

Abstract— **There are various applications which have a huge amount of database in different format. All databases maintain log files that keep records of database changes in a system formation. This can include tracking various user events and activity. Apache Hadoop can be used for log processing at scale of a system. Log files have become a standard part of large applications and large scale industry , computer networks and distributed systems. Log files are often the only way to identify and locate an error in software as well as system, because log file analysis is not affected by any time-based issues known as probe effect or a system effect. This is opposite to analysis of a running program, when the analytical process can interfere with time-critical or resource critical conditions within the analyzed program in a system. Log files are often very large and can have complex structure of database. Although the process of generating log files is quite simple and straightforward, log file analysis could be a tremendous task that requires enormous computational resources, long time and sophisticated procedures and time consuming . This often leads to a common situation, when log files are continuously generated and occupy valuable space on storage devices, but nobody uses them and utilizes enclosed information. The overall goal of this project is to design a generic log analyzer using hadoop map-reduce framework. This generic log analyzer can analyze different kinds of log files such as- Email logs, Web logs, Firewall logs Server logs, or System produces.**

Index Terms— **Hadoop, Map-reduce framework, Log files, log analyzer, Heterogeneous database, different Kind of log database.**

I. INTRODUCTION

Current software applications often produce (or can be configured to produce) some auxiliary text files known as log files or activity log (1). Such files are used during various stages of software development or software installation, mainly for debugging and profiling purposes of logs. Use of log files (2) helps testing by making debugging and developing easier. It allows you to follow the logic of the program, at high level, without having to run it in debug mode and runing mode. Nowadays, log files are commonly used at customer's installations for the purpose of permanent software monitoring and/or software installation. Log files became a standard part of large application and large scale of database. Log files are often the only way how to identify and find an error in software, because log file analysis is not affected by any time-based issues known as probe effect or system effect. This is an opposite to an analysis of a running program or software, when the analytical process can interfere with time-critical or resource-critical conditions within the analysed software of an system. Log files (2) are often very large and can have large structure. Although the process of generating log files is quite simple and easy to understand, log file analysis could be a tremendous task that requires enormous computational area, long time and easy understand procedures. This often leads to a common situation, when log files are continuously generated and occupy valuable space on storage devices or storage dtabase, but nobody uses them and utilizes enclosed information. The overall goal of this project is to design a generic log analyser using hadoop map-reduce (4) framework of analyzing importance things. This generic log analyser can analyse different kinds of log files such as- Email logs, Web logs, Firewall logs Server logs, Call data logs different kind of log database (3).

There are various applications or program(known as log file analysers for log files (1)to produce easily human readable summary reports. Such tools are undoubtedly useful, but their usage is limited only to log files or any activity of certain structure. Although such products visualization tools of database) that can digest a log file of specific vendor and have configuration options, they can answer only built-in questions and create built-in reports to design an open, very flexible modular tool, that would be capable to analyse almost log file present database. There is also a belief that it is useful to research in the field of log files analysis and any log file and answers any questions, including very complex ones and not understands. Such analyser should be programmable, extendable, efficient (because of the volume of log files) and easy to use for end users of analysis. It should not be limited to analyse just log files of specific structure or type and also the type of question should not be restricted in a log file. Big Data is a high volume, high velocity and high variety information assets that demand cost-effective, innovative forums of information processing for enhanced insight and decision making.

II. LITERATURE SURVEY

In the past decades there was surprisingly low attention paid to problem of getting useful information from log files (1) of an software development process. That seems there are two main streams of research and implementation. First one concentrates on validating program runs by checking conformity of log files (3) to a state system. The Records in a log file are interpreted as transitions of given state machine records or files. Some illegal transitions occur, then there is certainly a problem, created by software under test or debugging or in the testing software itself of a system. Second branch of research is represented by articles that just de-scribe various ways of production graphical output of a system. Following item summarizes current possible usage of log files of analysis (3):

- The generic program debugging.
- Whether program conforms to a given state machine various usage statistics, top fifteen, etc.
- Security monitoring and testing.

According to the available scientific papers it seems that the most evolving and developed area of log file analysis is the www industry or an organization. The Log files of HTTP servers are now-a-days used not only for system load statistic but they offer a very valuable and cheap source of feedback of log files. Log file Providers of web content were the first one who lack more detailed and sophisticated reports based on server logs or HTTP logs . They should require detecting behavioral patterns, paths, trends or type of log files etc. The Simple statistical methods do not satisfy these needs so an advanced approach must be used of an log file analysis. There are over 30 commercially available applications for web log analysis and many more free available on the internet. Regardless of their price, they are disliked by their user and considered too low, inflexible and difficult to maintain. Some log files, especially small and simple, can be also analyzed using common spreadsheet or database programs or a system program. In such case, the logs are imported into a worksheet or database and then analyzed using available functions and tools of an database.

III. VIRTUAL DATABASE SYSTEM

A virtual database (or VDB) is a container for components used to integrate data from multiple data sources in a network, so that they can be accessed in an integrated manner through a single or multi, uniform API. The standard VDB structure is shown in the Fig. 1 It consists of four components. The Mapper, Publisher, Executor and Wrapper of an log files.

1) Publisher: The publisher provides a query language for the users to access the system.
2) Mapper: Mapper needs the Metadata information. The query given by the users are processed and decomposed into sub queries according to the actual data retrieval instructions defined in the Metadata of an database.
3) Executor: The Executor provides an abstraction layer between Query Engine and physical data source that knows how to convert issued user query commands into source specific commands and execute them using the wrapper. It also has intelligence logic to convert the result data that came from the physical source into a form that Query engine is expecting of an system.

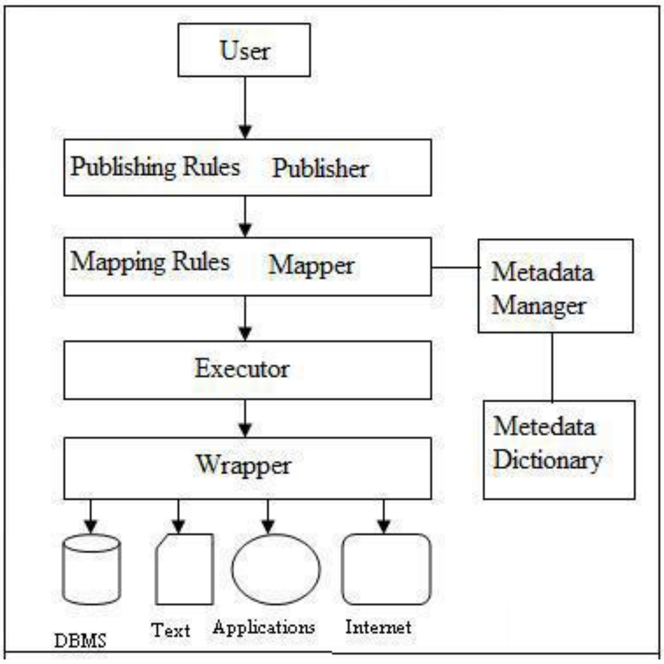

Fig. 1. Virtual Database System-General structure

4) Wrapper: A wrapper provides the connectivity to the physical data source. This also provides way to natively issue commands and gather results. A wrapper can be a RDBMS data source, Web Service, text file, connection to main frame etc.

5) Metadata: Metadata is data that describes a specific item of content and where it is located. Metadata is capturing important information about the enterprise environment, data, and business logic to accelerate development, drive integration procedures, and improve integration efficiency.

Metadata captures all technical, operational, and business metadata in real time in a single open repository. This repository ensures that metadata is always up to date, accurate, complete, and available. Fig. 1 shows the extraction of data from heterogeneous data sources using VDB.

IV. PRAPOSED SYSTEM

A. Problem Statement

To build a system for generic log analysis using Hadoop Map Reduce Framework by providing user to analyze different type of large scale of log file and malware present that log files.

B. Feature

- Increased efficiency do to use of Hadoop-Map Reduce framework.
- Ability to analyze the different kinds of log files.

C. Scope

Generic log analyzer can be used to analyze various kind of logs such as:

- Email logs.
- Web logs.
- Firewall logs.
- Serever logs.

This system build Generic Log Analyzer for different type of large scale log files.By taking advantage of Hadoop Map Reduce framework and polymorphism for log analysis and will increase efficiency and reliability of log analysis.

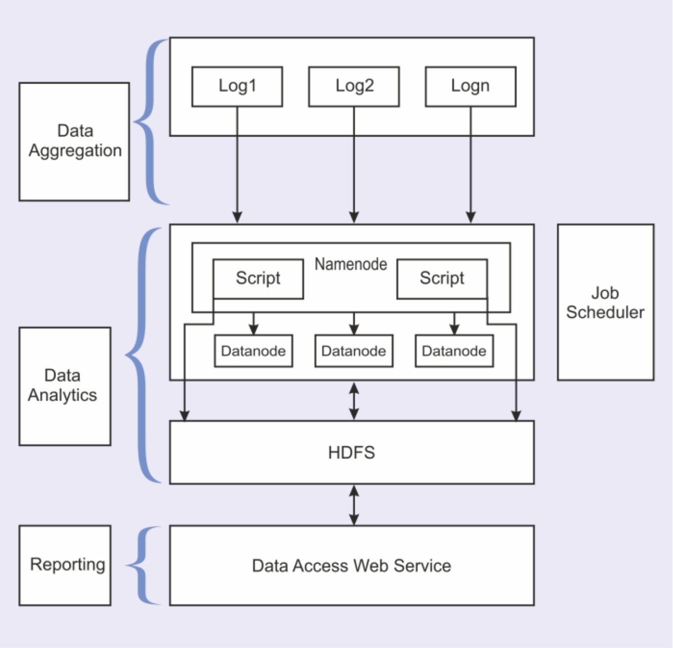

Fig. 2. Architecture Diagram

The above Fig. 2 shows that the architecture of proposed system. It consists of different type of log files as a input i.e. mail log, Firewall log, Server log. In first process this log files goes to name node and data node where name node is primary and data node is secondary there is writing the script related to that log files. After the name node and data node that log files goes to an HDFS i.e. Hadoop Distributed File System. Finally that file is represented in to the graphical represented format.

V. WORKING METHOD OF MAPREDUCE

Map Reduce a java based distributed programming model consists of two phases: a massively parallel "Map" phase, followed by an aggregating "Reduce" phase. MapReduce is a programming model and an associated implementation for processing and generating large data sets . A map function processes a key/value pair (k1,v1,k2,v2) to generate a set of intermediate key/value pairs, and a reduce function merges all intermediate values [v2] associated with the same intermediate key (k2) (1).

Maps are the individual tasks that transform the input records into intermediate records. A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks. The framework sorts the output of the map, which are then input to the reduce tasks. Both the input and the output of the processed job are stored in a file-system. Typically just zero or one output value is produced by the reducer. In MapReduce, a mapper and reducer is identified by the following signature,

$$\text{Map } (k1, v1) \rightarrow [(k2, v2)] \quad ` \qquad (1.1)$$
$$\text{Reduce } (k2, [v2]) \rightarrow [(k3, v3)]\ (1) \qquad (1.2)$$

Mapreduce suits applications where data is written once and read many times. The data stored in a file system namespace contributes to HDFS (2) which allows master-slave architecture.

VI. RESULT

The preprocessing work executed in parallel results in 2.15 minutes and 3.04 minutes for session identification of 550MB NASA web log data file. The same work when performed in single node produces 2.48 minutes for preprocessing and 3.02 minutes for session identification. The performance evaluation of Non hadoop approach, Pseudo distributed mode and fully distributed mode is shown in Table-I and Fig.3, which proves that performing the work in distributed mode improves the time in few milliseconds. Expanding the cluster and using terabytes of data would result in better time efficiency. The preprocessed data is analyzed using R, a free statistical programming tool. The log files don't contain any specific format, due to which fields could not be separated easily. Data frames are used in the analysis of log files to read the content of the file using read.

TABLE I. PERFORMANCE OF DIFFERENT APPROACHES

NASA Server logs	**Milliseconds**	**Minutes**
Non hadoop approach	369391	6.15
Pseudo distributed mode	330001	5.50
Fully distributed mode	311400	5.21

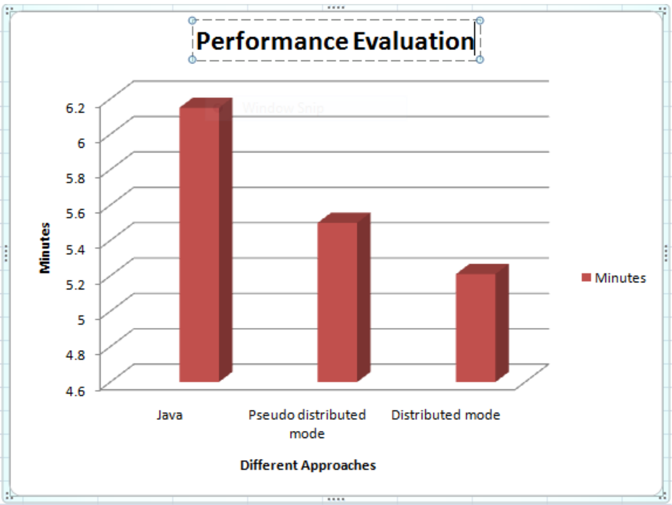

Fig. 3. Performance Evaluation of different approaches in hadoop

The data in R is used to filter, analyze, or manipulate to make it more usable .Using data editor the data frame is created from the preprocessed log file with row ranging till 50, 00,000 as shown in Fig.3.

VII. CONCLUSION

The main concept of data integration is to combine data from different resources and provide users with a unified view of these data. So this System will be able to analyze different types of log files. Due to use of Hadoop framework, Efficiency of log analysis has improved. If any new standard format log file is created then it will be easy to extend our project to analyze that log file. Our project can also be implemented on windows so that novice users find it easy to use.

REFERENCE

[1] Sayalee Narkhede and Tripti Baraskar, "HMR Log Analyzer: Analyze Web Application Logs over Hadoop MapReduce," International Journal of UbiComp (IJU) vol.4, No.3, July 2013.

[2] Konstantin Shvachko, et al., "The Hadoop Distributed File System," Mass Storage Systems and Technologies (MSST), IEEE 26th Symposium on IEEE, 2010.

[3] Milind Bhandare, Vikas Nagare et al., "Generic Log Analyzer Using Hadoop Mapreduce Framework,"International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol.3, issue 9, September 2013.

[4] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data" Processing on Large Clusters," Google, Inc.

[5] Savitha K and Vijaya MS "Mining of Web Server Logs in a Distributed Cluster Using Big Data Technologies" International Journal of Advanced Computer Science and Applications (IJACSA). Vol. 5, No. 1, 2014.