IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Exploratory Survey of Hadoop Log Analysis Tools

Sudhir Allam, *Sr. Data Scientist, Department of Information Technology, USA*

Abstract— This paper aims to explore more about Hadoop Log Analysis tools and how they can assist with technological breakthroughs. Since the usage of clusters in large-scale computation is increasing, maintaining the quality of these clusters is important. The significance of monitoring and controlling the network is illuminated by this. There have been several methods that can manage the Hadoop cluster effectively [1]. The bulk of these tools collect appropriate information in each cluster node and submit it to be processed. The majority of these diagnostic methods are post-processing research tools. An exploratory assessment of these tools is presented in this paper. A typical application for a first Hadoop program is log analysis. Indeed, one of Hadoop's first applications was for the large-scale study of clickstream logs — logs that document details regarding the web pages users visit and in what order they access them. The data logs created by the IT infrastructure are also called data exhaust. Like smoke emanating from a running engine's exhaust pipe, a log is a by-product of a running system. The term "data exhaust" connotes pollution or waste, as several businesses would certainly treat this type of data with that in mind [1]. The use of the internet is increasing every day in the modern world. Data are obtained from numerous sources including sensor data, web search page logs, social network pages, visual photographs and recordings, transaction details, and GPS signals on mobile phones. Managing this kind of data is a time-consuming process. This type of unstructured data is referred to as Big Data. Hadoop is the basis for conceptualizing Big Data and addressing the challenge of getting it useful for analytics.

Keywords: Hadoop, Log analyzer, MapReduce, Log files, Cloud computing, HDFS

I. RESEARCH PROBLEM

The problem that this research will try to solve is to show how challenges in log analysis can be addressed by utilizing the various Hadoop Log Analysis Tools. The most difficult aspects of log analysis are inconsistency, length, and distribution. The composition of log data varies with each source that produces this information. The scale of traffic logs is normally greater than the capability of a single IDS due to a large amount of data traffics. As a result, doing meaningful work of these data is virtually impossible [2]. Most analyzes are typically carried out after a computer forensics intrusion has occurred. Log data develops rapidly, and analyzing it can be time-consuming due to the large volumes produced.

Furthermore, the data's future meaning is often ambiguous. As a consequence, IT agencies are tempted to retain or archive these log data as quickly as possible. The importance of log analysis cannot be overstated, but it is much more critical to recognize that taking a cautious attitude is critical in predicting when bad stuff might arise and what might prove out to be a possibly fatal risk. Timestamps are particularly useful because they enable researchers to compare incidents dependent on time through a network or organization [2]. Timestamps can be difficult to work with. Internal clocks of machines are often inaccurate or have a time drift. This triggers clocks to differ, often dramatically, within a network or regulated device. The Network Time Protocol (NTP) solves the issue but contributes to the difficulty of machine and network administration. Another problem in log analysis is the sheer quantity of data [2]. The amount of machine data increases dramatically as the number of machines that generate logs increases. Identifying an answer to a particular problem necessitates the need to locate only valid data. This normally entails filtering out a significant amount of repetitive data.

II. INTRODUCTION

Log analysis is a valuable method for software or machine debugging. A programmer may use a smart logging build to recreate the logical sequence of execution of a program to verify that the machine is performing as expected. To track complicated applications written by several authors, it's necessary to combine logs from multiple sources. It's also important for keeping track of large networks of interconnected parts. A potential issue is that various developers assign different levels of priority to different log entries. Furthermore, an overly complex

logging scheme can cause more problems than it solves. It may also result in an unmanageably high log volume.

Hadoop's popularity is increasing largely due to its capacity to process massive volumes of data. Clusters for data storage and retrieval may range from a single computer to a collection of computers, most of which are generic machines [3]. Although the one-name node does have some shortcomings, this Hadoop is commonly used. The MapReduce programming model is supported by this software architecture for conducting decentralized processing of data processed on the cluster. Hadoop Distributed File System (HDFS) is the Hadoop framework's internal storage layer, and as a result, it is utilized for all Hadoop-based applications. The configuration of this layer was derived from Google File System (GFS) [4], and the entire architecture is reliant on HDFS's performance. The HDFS framework has a master-slave design which was developed for data storage durability. Hadoop's capacity to handle data at the same time is a big gain for analytics. A weblog analyzer is a method for analyzing weblogs created on the server-side and determining statistics for a particular site. Weblog analyzer is a simple, efficient log analysis tool [5]. It offers details on-site users and their behaviors, as well as information on accessible files, routes inside the domains, user interaction between websites, different browser statistics, and operating system statistics. It provides reports with pie charts, bar graphs, and text details that are simple to interpret.

Operating-systems reports have a distinct layout and provide different material than networking logs. The case logs of the Windows Operating System (OS) vary from those of the UNIX Operating System (OS). The former uses a binary format, whereas the latter uses a text format known as American Standard Code for Information Interchange (ASCII) [5]. Logs usually include timestamps added to the incident log entry that is registered by the computer as incidents arise. A difficult challenge is making use of the increasing amount of computer data. Machine data, in this case, machine logs, is knowledge provided by machines without the intervention of humans. Throughout this paper, a discussion will include a framework for successfully examining vast quantities of machine logs to enhance computer protection. "A log is a database of the activities happening within an organization's processes and networks," as per the National Institute of Standards and Technology (NIST). Operating-system incident records, network traffic logs, and program logs are all examples of this type of data. The reason for this research is addressed in this portion, as well as the possible advantages of log analysis. After that, we'll go over the analysis questions and study framework [5,6]. Computers are unlikely to document anything that exists on a device because this will result in more log data being generated than user data [6]. They do, nevertheless, need to keep a lot of data because complex security issues sometimes necessitate a statistical review of the available logs accompanied by a further inquiry, and all measures necessitate a lot of data to draw accurate conclusions. A consolidated log collection and processing framework facilitate network-wide correlation of incidents and an awareness of the dynamic challenges confronting security personnel and administrators. This study aims to figure out how to use a big-data framework to securely and easily handle a huge volume of log data. The emphasis of the topic is on how to use Hadoop log analysis tools for security, efficiency, and debugging in general.

III. LITERATURE REVIEW

A. Hadoop Frameworks

Hadoop is an open-source and distributed platform for extracting, handling, and storing massive data sets that operates on a cluster of commodity hardware. The architecture, which can scale from a single computer to thousands of machines and manage petabytes of data, is error resistant, meaning it was built to handle vulnerabilities [7]. The system provides 3 main frameworks in addition to the Hadoop Common module, which incorporates the key Hadoop utilities:

- **HDFS (Hadoop Distributed File System): HDFS** is a fault-tolerant distributed file system for storing data on commodity computers. HDFS is intended to operate on commodity machines and is extremely scalable and error. An HDFS cluster uses a master/slave architecture made of one NameNode and DataNodes [7]. The NameNode that is used to control the cluster, its metadata, and DatNodes (slaves) is a server that manages the file system like opening, shutting down, and changing the name files and folders on the file system.

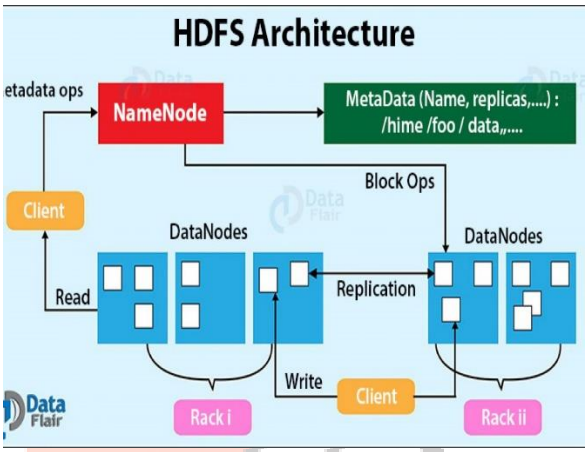

Fig I: HDFS architecture.

- **Hadoop YARN** is a big data processing framework. Its architectural unit oversees the cluster. It is the architectural core in control of handling Hadoop's various operations. It was developed in response to the need to address the bottleneck caused by the previous design, known as MRv1 or MapReduce v1 [7]. Under YARN, JobTracker tasks are divided into various daemons: NodeManager, ResourceManager, and ApplicationMaster.
- **MapReduce:** This is a programming model for handling vast amounts of data. MapReduce is a programming model and architecture for application development (jobs) that manage large amounts of data in synchronization through a cluster [7]. The map task and the reduce task are the two main components of MapReduce work.

B. The challenges of Using Hadoop

Not all challenges are well-suited to MapReduce technology. It works well with basic knowledge queries and issues that can be broken down into discrete components, but it is ineffective for iterative and dynamic analytic activities. MapReduce is a file-intensive algorithm [8]. Iterative algorithms take several map-shuffle/sort-reduce steps to complete since the nodes can interact through sorts and shuffles. This results in several files being created between the MapReduce phases, which is wasteful for sophisticated analytics computation. There is a well-known skills gap. It can be hard to find entry-level programmers with enough Java knowledge to be efficient

with MapReduce. One explanation for this is because delivery companies are rushing to implement relational (SQL) technologies on top of Hadoop [8]. SQL programmers are much more difficult to come by than MapReduce programmers. Furthermore, Hadoop administration seems to be a mixture of art and technology, necessitating a basic comprehension of operating systems, hardware, and Hadoop kernel configurations.

Data protection is essential. Another concern is the fragmented data protection challenges, which are being addressed through modern software and technology. Kerberos authentication is a significant move toward securing Hadoop settings. Data control and maintenance to the greatest degree possible. Hadoop lacks robust data storage, data cleansing, management, and metadata resources that are straightforward to use [9].

C. Hadoop Log analysis tools

Tools for data quality and standardization are particularly inadequate. Since the volume of big data is continuously growing (and shows no signs of slowing down), the clusters that deal with it must be carefully watched and managed. Because recognizing a system's success is associated with system resource use awareness, log analysis will assist in improving the overall performance of the system. The following are some example of Hadoop Log analysis tools:

1. Vaidya

Vaidya is a very useful method for performance review of Hadoop cluster implementations, which is one of the additions to the Hadoop platform. This method assists the cluster supervisor in finding positions that are doing poorly [10]. Vaidya evaluates the application's output using a series of diagnostic rules that the user may create based on the application. The value, threshold, and dosage may all be defined for each diagnostic test rule. The test's general value is determined by its importance. The recommendation outlines the recommendations that would be given to increase the job's results. The test report includes the following information: test name, the relevance of test, diagnostic test summary, severity, test outcome, and medication. In the default edition of Vaidya, five test rules are included [11]. One downside to this tool is the challenge of ensuring rule continuity amid an ever-growing Hadoop codebase.

2. Salsa and Mochi

Salsa analyses Hadoop logs (Data Node logs and Mission Tracking logs) to illustrate relaying process operation around each node and provides a logical data view of the activity on each node. Each operation releases log messages as a task reduction chart contains logging statements [15]. To model the control flow, each log message is viewed as an occurrence that can be maybe an execution beginning or finish [11]. The usage of pattern detection is used to identify activities from log files. Salsa also allows an effort to integrate fault identification and diagnosis for work MR job executions. This method would not necessitate any changes to the operating framework, middleware, or program code. Mochi, a log-based visualization framework for Hadoop, can also be used for Hadoop debugging. Mochi [16] connects the actions of execution in terms of time, space, and volume, as well as interconnections in a distributed system. It first collects the Task execution view per node by running the Map-Reduce node and then creates a Job Centric Data Flow (JCDF), which is a guided graph, by linking the collected execution experiences on each node as well as the HDFS layer [12]. The JCDF is spread across nodes as well. As a consequence, it attempts to produce and compare data efficiently, as well as monitor data flow between nodes, before evaluating Hadoop actions. Users will use Mochi's model to reason for and debug output problems. MIROS, REP, and Swimlanes are a handful of the visualizations. MIROS restricts and extracts the flow of data at each chart node. States that handle greater amounts of data should take longer to process, according to REP search [13]. Swimlanes are used to monitor work progress and display the length of each task.

3. Chukwa

Another Hadoop framework is Chukwa which resembles Ganglia in storage. This is a modular data storage and network management system developed on Hadoop clusters, i.e. the Hadoop distributed File System and Map Reduce architecture [13]. In contrast to other tools that hold logging data from the local storage, Hadoop's distributed file system, HDFS, stores logged data. For storing large data, HDFS has a high throughput and is extremely scalable, reliable, and dependable. Chukwa offers constant device tracking to help in fault detection. Chukwa [13,14]'s simple architecture is composed of an adaptor, an agent, a collector, and HDFS storage. Chukwa gathers device metrics, log archives, random log files, and X-trace [15] records, among other things. These data are obtained by adaptors and saved in HDFS as big files, as HDFS is more effective for large-file operations than small-file operations. Therefore, between the adapters and the HDFS plate, the collector and agents are installed [16]. As a consequence, this tool significantly enables the collection and retrieval of massive data chunks.

The above-listed log analysis tools can be contrasted. Vaidya performs a post-execution log review, focusing on device fault detection rather than system and network failure. By adding additional test rules, Vaidya's spectrum of failure prediction and diagnosis can be expanded [17]. However, this relies on the program to be tracked, so the accuracy of Vaidya's malfunction assessment lies mainly with the user. Both Salsa and the Mochi define the states through pattern detection. Although if a user has to adjust the log patterns (as part of every change in the source code), they can't distinguish patterns simply by using matching patterns. It will effectively capture task complexities since the change from one state to another reveals the interdependence of each task. As the time factor is taken into consideration, clock synchronization should be taken with caution [17]. Chukwa, on the other side, takes all available forms of knowledge from different sources as feedback and tracks the whole framework. It designs a Hadoop method of data collection and network monitoring.

IV. IT'S FUTURE IN THE UNITED STATES

The need for Hadoop big data analytics solutions is projected to grow in the U.S. due to the rising requirement for predictive asset maintenance, remote health management, sales, and consumer management, energy management, and inventory management. Hadoop is useful both for many organizations around the US in stored, analyzed, logged, and cached, and recovered financial data (logs from factory loads or other data processing). Most market players are considering ways to utilize it for archival purposes. Hadoop is utilized by various organizations for many departments. Today, a lot is spent a lot in historical data management and many companies are planning to switch it into Hadoop. There has been a much better performance in the handling of a significant volume of financial data in the Hadoop cluster as well. Financial institutions, such as JPMorgan, have millions of clients but are now able to run efficiently due to big data analytics applied to a growing array of unstructured and standardized data sets utilizing the open-

source Hadoop platform [18]. Big data analysis allows JPMorgan to determine the right range of solutions that its clients will produce. The key motive of JPMorgan is to get the proper service to the right consumer in a contextually appropriate manner at the right moment through the right platform.

The Hadoop Big Data Analytics industry is divided into promotion, sales, businesses, finance, and human resources through company functions. In the near future in the Hadoop, Big Data Analytics industry the sales and marketing segment is projected to expand rapidly. The increasing usage and improvement in consumer engagement of structured and unstructured data in business development are projected to encourage the implementation of Hadoop Big Data Analytics and services.

V. Benefits to the U.S economy

Businesses in different areas are now working to implement a diverse variety of Hadoop Big Data Analytics tools to turn their activities and consumer interactions digitally into mission-critical processes. Enterprises' core market and organizational goals are anticipated to accelerate the implementation of Hadoop big data analytics technologies worldwide, including lower technology and operational costs, improved user interactions, improved data protection, and safety, increased operational insight for different processes, and improved real-time business decision-making [18]. The US government has supported the creation of an optimal research and innovation ecosystem contributing to progress in different fields of science and technology. Investments are made in the incorporation of big data analytical solutions, IA and IoT services, and solutions across the consumer electronics segment, pushing growth in the United States.

JPMorgan is integrating nearly 30 million consumers' spending details with widely accessible US economic statistics. From October 2012 to December 2014, JPMorgan data analytics created a data collection of 2.5 million de-identified consumers to analyze the income and consumption habits of 2.5 million account holders. Throughout the development and review of these important data properties, JPMorgan implemented stringent privacy protocols to safeguard the privacy of its customers. The majority of homes already have smart meters that keep track of their energy use. Modern vehicles have a network of sensors that record their condition and operation. Whenever an item is purchased — even without having a credit card or debit card — systems report the business in files — and records [18]. One may view some of the most popular log data sources: IT servers, network click flows, sensors, and transaction systems. Each industry (and the log styles just described) have enormous potential for useful analyzes—especially when one can zero in on a particular form of operation and link the results to another collection of data.

VI. CONCLUSION

This paper offers a concise overview of the tools of Hadoop log analysis for massive data processing. The findings of this paper suggest that the Big Data period necessitates data debugging, output tracking, storage, as well as activity monitoring because data is increasing exponentially with no indications of slowing. Therefore, there is an appalling need for scalable and efficient solutions for this, and several resources to track and manage large clusters have been introduced. Hadoop is the most commonly deployed method for Big Data storage and processing and involves many performance intelligence gathering and research sub-projects. Most of these tools use log files to record the cluster activity and running applications. They process the logs and get the requisite data to identify a malfunction and even support failure repair by utilizing some of its tools. An analysis of some of these log analyzers indicates that several tools strive only to capture the diagnostic component of an application malfunction, without regard to hardware and system failures. Loss is not an anomaly in such clusters and the assessment of error must also be applied to all potential failure stages from node failure through system failures.

References

[1] M. Danthala and S. Ghosh, "Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using BigInsights", International Journal of Engineering Research and, vol. 4, no. 05, 2015.

[2] M. Fan and S. Xu, "Massive medical image retrieval system based on Hadoop", Journal of Computer Applications, vol. 33, no. 12, pp. 3345-3349, 2013.

[3] J. Kim, K. Park and D. Kim, "The Study on the Analysis System of Effective Attack Patterns in Kendo Match using Hadoop-based Hive", Indian Journal of Science and Technology, vol. 9, no. 1, 2016.

[4] Y. Lee and Y. Lee, "Yet Another BGP Archive Forensic Analysis Tool Using Hadoop and Hive", Journal of KIISE, vol. 42, no. 4, pp. 541-549, 2015.

[5] A. Chalkiopoulos, Programming MapReduce with Scalding. Birmingham: Packt Publishing, 2014.

[6] B. Luo, H. LI, Y. Wang and Z. Yu, "Design and Implementation of Web-based Census Cartography System", Geo-information Science, vol. 11, no. 6, pp. 826-833, 2010.

[7] R. Desai, "Real-Time Analysis using Hadoop", International Journal Of Engineering And Computer Science, 2016.

[8] P. Gupta, P. Kumar, and G. Gopal, "Sentiment Analysis on Hadoop with Hadoop Streaming", International Journal of Computer Applications, vol. 121, no. 11, pp. 4-8, 2015.

[9] P. R.Sahoo, "Performance Overhead on Relational Join in Hadoop using Hive/Pig/Streaming - A Comparative Analysis", International Journal of Applied Information Systems, vol. 4, no. 7, pp. 15-20, 2012.

[10] Y. Laxmi, "Customer Complaint Analysis Using Hadoop (Consumer Analysis)", International Journal Of Engineering And Computer Science, 2017.

[11] Y. Hu, X. Cai and B. DuPont, "Design of a web-based application of the coupled multi-agent system model and environmental model for watershed management analysis using Hadoop", Environmental Modelling & Software, vol. 70, pp. 149-162, 2015.

[12] K. Ahn, J. Lee, D. Yang, and B. Lee, "Design and Implementation of a Hadoop-based Efficient Security Log Analysis System", Journal of the Korea Institute of Information and Communication Engineering, vol. 19, no. 8, pp. 1797-1804, 2015.

[13] M. Mohandas and D. PM, "An Exploratory Survey of Hadoop Log Analysis Tools", International Journal of Computer Applications, vol. 75, no. 18, pp. 33-36, 2013.

[14] J. Hare, S. Samangooei and P. Lewis, "Practical scalable image analysis and indexing using Hadoop", Multimedia Tools and Applications, vol. 71, no. 3, pp. 1215-1248, 2012.

[15] M. Zhang, R. Zhang and C. Liu, "Design of Smart Healthcare Data Management System Based on Hadoop", Advanced Materials Research, vol. 998-999, pp. 1121-1124, 2014.

[16] B. Lee, J. Kwon, G. Go, and Y. Choi, "A Method for Analyzing Web Log of the Hadoop System for Analyzing an Effective Pattern of Web Users", Journal of the Korea society of IT services, vol. 13, no. 4, pp. 231-243, 2014.

[17] S. Pan, L. Zhu, B. Hu, and P. Yang, "Research and Design of a Massive Offline Data Analysis System Based on Hadoop", Applied Mechanics and Materials, vol. 631-632, pp. 1049-1052, 2014.

[18] Y. Wei, G. Zhou, D. Xu, and Y. Chen, "Design of the Web Log Analysis System Based on Hadoop", Advanced Materials Research, vol. 926-930, pp. 2474-2477, 2014.