ISSN:0975-9646
Laurel Thejas Souza et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2364-2367

# Error Log Analytics using Big Data and MapReduce

Laurel Thejas Souza[1], Girish U R[2]

*SPAN INFOTECH (India) Pvt. Ltd. Bangalore, India*

**Abstract-Log Analytics is a technique for automatically understanding the meaningful patterns from heterogeneous log data. Every activity taking place in an application or device is recorded in a log file. Valuable information is stored on log data which can be extracted and stored into Big Data platforms. Prediction and Classification can be performed over the log data.**

**Keywords-*Data Mining, Machine Learning, Predictive Analytics, Hadoop, Pig, Flume, Hive, Oozie, R***

## I. INTRODUCTION

Log files 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 using Hadoop [1] is used. Analytics [5] 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 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 have been accessed [12], 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.[2]

1. Access Log records all requests that were made of this server including the client IP address, URL, response code, response size, etc. [3]

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

Access logs come in several different formats but they all look something like this:

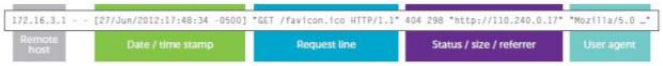


Figure 1: Access Log

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 are always found in a heterogeneous format which looks something like this.

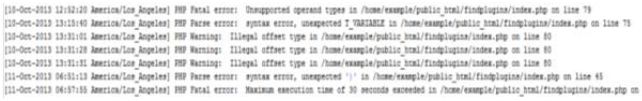

Figure 2: Error Log

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.

Analytics [8] often involves studying historical data to research potential trends, to analyze the effects of certain decisions or events, or to evaluate the performance of a given tool or scenario. [9] The goal of analytics is to improve the business by gaining knowledge which can be used to make improvements or changes. Log Analytics: Log analytics, is a study of Log Files to research about the records of a system for pattern discovery and analysis, through which the systems can be made Pro-active.

*1) Abbreviations and Acronyms*

- ELABD : Error Log Analytics using Big Data
- Log File: A log file is a recording of all the activities taking place in a particular server [2].
- Error Log: A log file which contains all the activities and error messages generated by an application. Every message has a unique Identifier.
- Log Analytics: A technique seeking to make sense out of computer-generated records (also called log or audit trail records). The process of creating such records is called data logging.
- Business Intelligence (BI): A set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purposes.
- HDFS [1]: Hadoop Distributed File System (HDFS) is a file system which provides scalable and reliable data storage that is designed to span large clusters of commodity servers.
- Pattern Recognition: A branch of Artificial Intelligence which aims to classify data (patterns) based on either a priori knowledge or on statistical information extracted from the patterns.
- Predictive Analytics: Practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends.
- Machine Learning: Study of algorithms that can learn and make predictions from the provided data.
- Server Log: A log file that is automatically created and maintained by a server consisting of a list of activities to be performed.
- Web log mining: Also called Web Usage Mining, is the type of Web mining [11] activity that involves the automatic discovery of user access patterns from one or more Web servers [3], [4].

## III. Project Perspective

ELABD is used for collecting error logs and perform analytics on those collected data. ELABD ensures scalability in terms of data storage and accumulates a large number of log files. Only open source software such as Linux, Hadoop, Pig, Hive, Flume, Oozie and R are used.

## IV. Project Functionality

**User Authentication** Front end is created so that the user can interact with the application in a convenient way. The user will login to the application and perform the functionalities on log data through the Graphical User Interface (GUI). There will be provision for user management.

**Data Preprocessing Cleansing and Integration.** The log files are collected from heterogeneous servers. As the structure of the log files are different, data cleansing plays a key role in bringing the collected log data into a uniform homogeneous format. Integration of data is achieved through statistical techniques [2].

**Pattern Discovery and Analysis** 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 [3], [4].

**Predictive Analytics** [5] 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 [3], [4].

**Visualization** The analysis carried out on the error log patterns are visualized using graphs and plots. The reports are also generated.

## V. Operating Environment

A Hadoop cluster is set up, where the different Linux distros are integrated for different software such as Apache Pig, Hive and Oozie [6] to perform data mining and analytics. Pig Latin and HiveQL are high level languages which generate Mapreduce. Flume[7] and Hadoop [1] are used for the collection of huge data or Big Data Collection. Machine Learning and Predictive Analytics is achieved using the R language.

## VI. Figures and Design

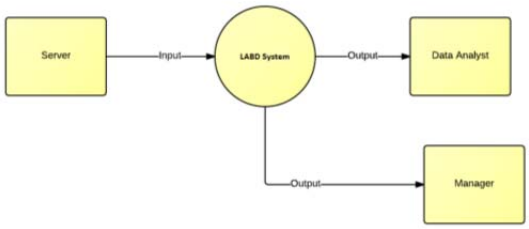


Figure 3: DFD - Context diagram


Figure 4: Use Case Diagram

Use Case Diagram describes about the roles and users interacting with the system. The various Use Cases and privileges are highlighted by this diagram. Here Administrator is having most of the control towards the data and specifying the roles. An IT Support is an actor who deals with the server logs, collects the log files and maintains access and error logs. Data analyst is the one who discovers the patterns, works with the patterns, analyzes the patterns and produces the visualization. Manager is an actor who views the generated report and visualization to take actions or make decisions.

## VII. System Architecture

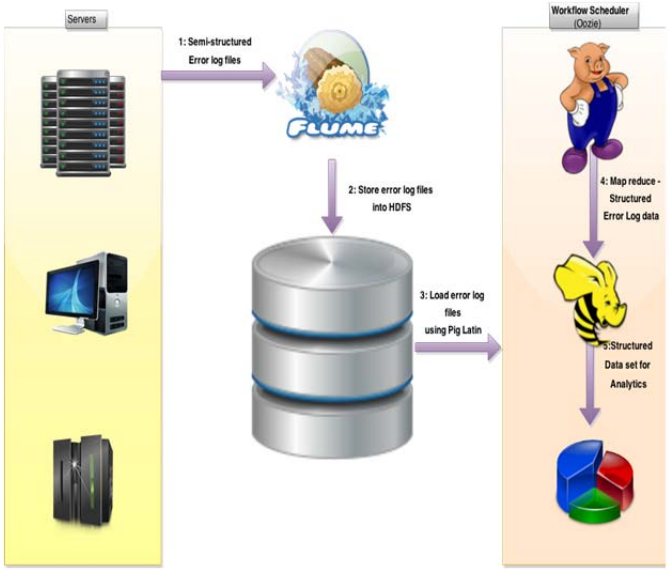


Figure 5: System Architecture Diagram

- Streaming log data into HDFS is achieved using log collector tools such as Flume[7] 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. 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.
- Data Analytics [9] is used to extract and represent meaningful information from the log data. It is used in many industries to allow companies and organization to make better business decisions and in the sciences to verify or disprove existing models or theories. Predictive Analytics [5] is a branch of Business Intelligence (BI) which is used to forecast and predict the future based on the current and historical data. [4] Regression analysis is used here to determine a Statistical model which will fit into the data pattern that is generated. Linear Regression and Decision Trees are some of the examples which are used to predict the future trend based on the available data.
- Scheduling of the tasks to be performed is necessary in order to achieve an effective output. Hence workflows are created and the various tasks are packaged inside a workflow. Similar to job scheduling, the workflow executes as a batch processing module and carries out the activities or tasks that need to be performed such as Extract-Transform-Load (ETL) operations, loading the cleansed and structured error log data from the Distributed File System into the Data Warehouse and performing Analytics over the structured data. At regular intervals of time, the workflow should be executed.

## VIII. Results Obtained

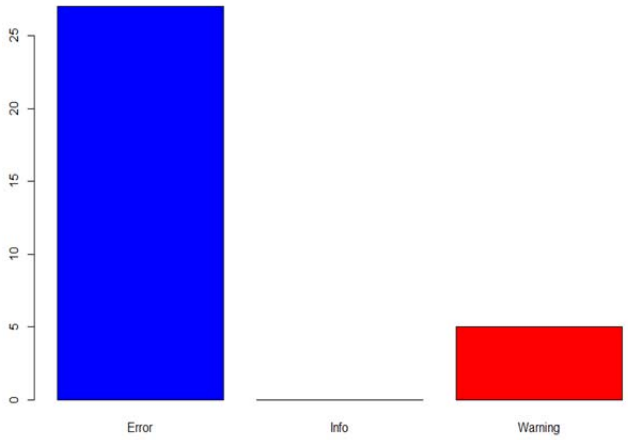


Figure 6: Severity count in the dataset

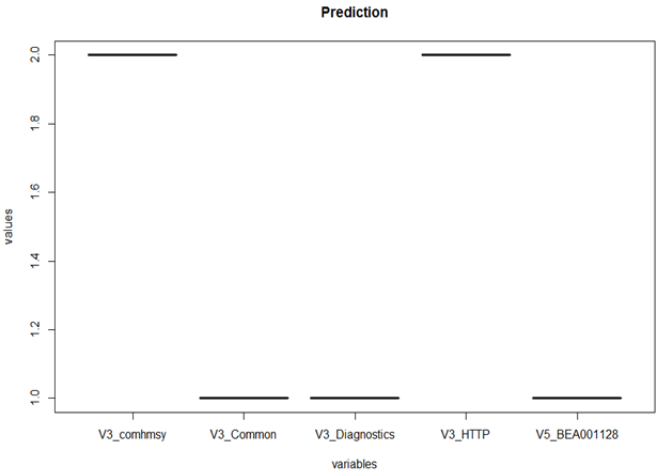


Figure 7: Prediction of errors based on subsystem

The various categories of severity are Error, Warning. And Info. Figure 2.4 shows the severity count present in the error log file. The straight line equation

$$Y = mx + c$$

Is used to predict the future severity value. The independent variable x consist of the influencing parameters for prediction, while y is the predicted value. In Figure 2.5, the final prediction of the severity values are shown. Values ranging from 1 to 2 denote Info and Error respectively.

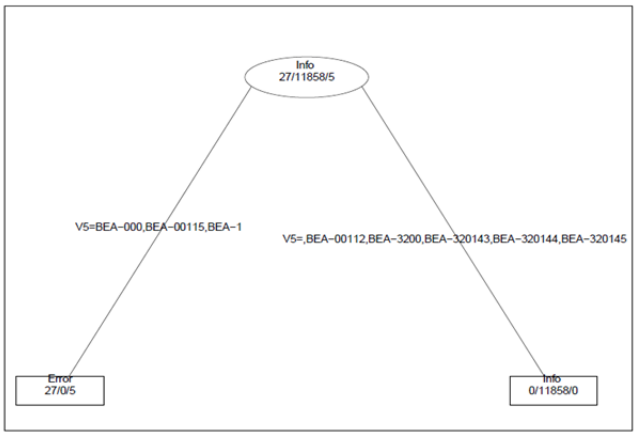


Figure 8: Decision tree for Severity

The Decision tree consists of a root node and 2 leaf nodes. The root node tells that out of the total rows in the dataset, 27 are classified as Errors, 11858 are Infos and the remaining 5 are Warnings. Clearly Infos are dominant in the dataset. Whenever message IDs are the values displayed in the left branch of the tree, then there is a high chance of the system going to an Error. When the message IDs are the values displayed in the right branch of the tree, the severity may consist of severity as Info.

## REFERENCES

[1] Chuck Lam, "Hadoop in Action", Manning Publications.
[2] Savitha K, Vijaya MS, "Mining of Web Server Logs in a Distributed Cluster Using Big Data Technologies", IJACSA, Vol. 5, 2014.
[3] Bamshad Mobasher, "Chapter 12: Web Usage Mining".
[4] Sasa Bosnjak, Mirjana Maric, Zita Bosnjak, "The Role of Web Usage Mining in Web Applications Evaluation", Management Information Systems, Vol 5.
[5] Charles Elkan, "Predictive Analytics and Data Mining", 2013.
[6] Hortonworks, "Chapter 2: Understanding the Hadoop Ecosystem".
[7] The Apache Software Foundation, "Flume User Guide".
[8] S. Narkhede, T. Baraskar, "HMR Log Analyzer: Web Application Logs over Hadoop Mapreduce", International Journal of UbiComp (IJU), Vol.4, No.3, July 2013.
[9] 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).
[10] W.Peng, T. Li, S.Ma, "Mining Logs Files for Data-Driven System Management ", Florida International Unversity.
[11] L.K.J. Grace, V. Maheswari, D. Nagamalai, "Analysis of web logs and web user in web mining", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.
[12] A. Bruckman, "Chapter 58: Analysis of Log File Data to Understand User Behavior and Learning in an Online Community", Georgia Institute of Technology.