IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 1, Feb-Mar, 2014
ISSN: 2320 - 8791
www.ijreat.org

Design of Event Log Analyser Algorithm Using Hadoop Framework

Banupriya P[1], Mohandas Ragupathi[2]
PG Scholar, Department of Computer Science and Engineering, Hindustan University, Chennai
Assistant Professor, Department of Computer Science and Engineering, Hindustan University, Chennai

Abstract:-Event logs in windows desktop and servers in an enterprise are voluminous. However there is a need for analysing logs related to security. System administrator will find it difficult to collate these distributed logs. Never – theless in the current security environment it is imperative to put an automated system to collect, process and analyse the disparate log files. We cannot use on-hand database tools for this problem. This intends to use BIG DATA infra-structure for aforementioned problem. It will be developing Hadoop based log processing application. The main objective is to showcase the qualities of proposed solution. This application will contain interfaces to submit jobs in Hadoop along with the log files and map-reduce programming modules to process the logs with high performance.

1. Introduction

In this paper we discuss the how to use Event Viewer as a troubleshooting tool. Today, event logs contain vast amounts of data. Event Viewer displays detailed information about system events. Event logging and log files are playing an important role in system and network administration. With the growth of communication network, event logs are increasing in size rate.

The "Big Data" phrase refers to some new challenges, known as the "Three V", and the most important is the Volume. It refers to the necessity to deal with large amounts of data and the most popular solution is provided by Hadoop. Indeed, this software handles wide datasets by splitting them into chunks that are scattered among a cluster of computers. This distributed solution eases the data process since it is possible to split the working load without overcharging a single computer. As regards the hardware, the Hadoop advantage is that it does not have any particular requirement about that. As a matter of fact, it may use a cluster of cheap computers and in that way it avoids overloading any node of the cluster. In addition, larger volumes may be handled through adding some new nodes.

The Map-Reduce structure is capable of handling Big Data problems. Indeed, a mapping consists in performing the same chosen operation on each datum, so the computational cost linearly grows as the volume increases. As regards reducing, things are more complicated, but if projected properly it is scalable too. In conclusion, the efficiency is generally low but it depends on the algorithm.

Hadoop solution is Map-Reduce that is a specific programming model for writing algorithms. As its name suggests, the algorithms are divided into two steps that are Map and Reduce. Mapping consists in extracting from each chunks of data all information that is necessary. Once the information is extracted, the Reduce step aggregates it and computes the output. Every Hadoop procedure follows the Map-Reduce logic and there are different high-level tools that present this structure, although it is hidden. However, in order to develop new algorithms, it is necessary to follow the Map-Reduce logic.

Roles of event log.

Event logs play an important role in modern IT systems, many system components like applications, servers, and network devices have a built-in support for event logging (with the BSD syslog protocol being a widely accepted standard) In most cases event messages are appended to event logs in real-time, event logs are an excellent source of information for monitoring the system. Information that is stored to event logs can be useful for analysis at a later time, e.g., for audit procedures.

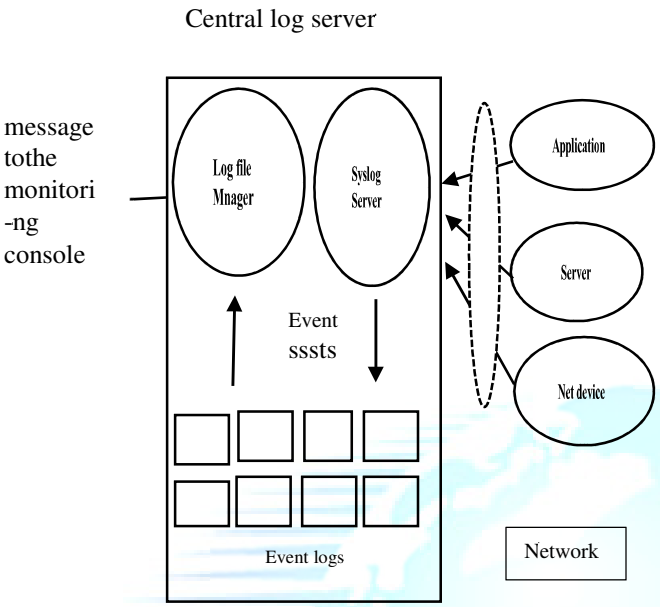
Event viewer Descriptions

Windows machine, an event is any significant occurrence in the system or in a program that requires users to be notified, or an entry added to a log.

The Event Log Service records application, security, and system events in Event Viewer.

With the event logs in Event Viewer, you can obtain information about your hardware, software, and system components, and monitor security events on a local or remote computer.

Event logs can help you to identify and diagnose the source of current system problems, or help you to predict potential system problems.



Event Log types:-Microsoft windows server2003, Windows XP, windows 2000 server and window NT record event in three kinds of logs:

Application log: - The application log contains events logged by programs. For example, a database program may record a file error in the application log. Events that are written to the application log are determined by the developers of the software program.

Security log:-The security log records events such as valid and invalid logon attempts, as well as events related to resource use, such as the creating, opening, or deleting of files. For example, when logon auditing is enabled, an event is recorded in the security log each time a user attempts to log on to the computer. You must be logged on as administrator or as a member of the administrators group in order to turn on, use, and specify which events are recorded in the security log.

System log:-The system log contains events logged by Windows system components. For example, if a driver fails to load during start up, an event is recorded in the system log

Event Types:-The description of each event that is logged depends on the type of event. Each event in a log can be classified into one of the following types :

Information:-An event that describes the successful operation of a task, such as an application, driver, or service. For example, an Information event is logged when a network driver loads successfully.

Warning:-An event that is not necessarily significant, however, may indicate the possible occurrence of a future problem. For example, a Warning message is logged when disk space starts to run low.

Error:-An event that describes a significant problem, such as the failure of a critical task. Error events may involve data loss or loss of functionality. For example, an Error event is logged if a service fails to load during start up.

*Success Audit (Security log):-*An event that describes the successful completion of an audited security event. For example, a Success Audit event is logged when- a user logs on to the computer.

Failure Audit (Security log):-An event that describes an audited security event that did not complete successfully. For example, a Failure Audit may be logged when a user cannot access a network drive.

2. System Architecture

Event Viewer is a tool that displays detailed information about significant events (for example, programs that don't start as expected or updates that are downloaded automatically) on your computer. Event Viewer can be helpful when troubleshooting problems and errors with Windows and other programs.

Open Event Viewer by clicking the **Start** Button, clicking **Control Panel**, clicking **System and Security**, clicking **Administrative Tools**, and then double clicking **Event Viewer**. If you're prompted for an administrator password or confirmation, type the password or provide confirmation.

This module deals with the collecting windows security event logs by pushing the logs from all systems to the hadoop system.by using certain commands push log will perform. The process of acquiring log files is to enhance the security of the enterprise.

HDFS is the primary distributed storage used by Hadoop applications. A HDFS cluster primarily consists of a Name Node that manages the file system metadata and Data Nodes. The HDFS architecture depicts basic interactions

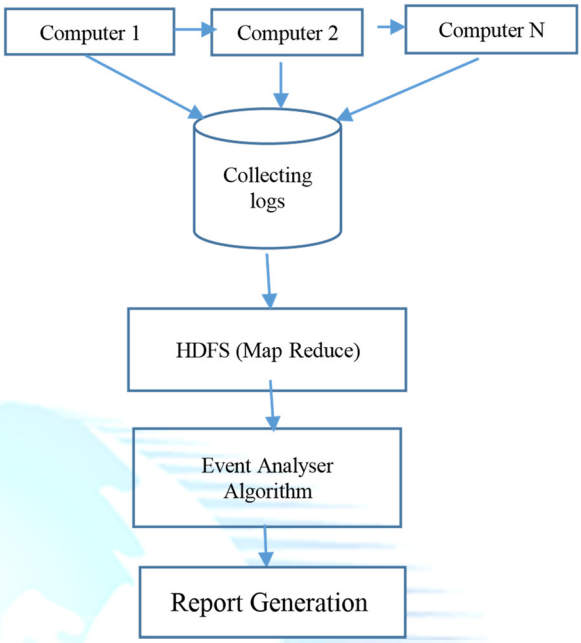
among Name Node, the Data Nodes, and the clients. Clients contact Name Node for file metadata or file modifications and perform actual file I/O directly with the Data Nodes.

The **Hadoop distributed file system** (**HDFS**) is a distributed, scalable, and portable file-system written in JAVA for the Hadoop framework. Each node in a Hadoop instance typically has a single namenode; a cluster of datanodes form the HDFS cluster. The situation is typical because each node does not require a datanode to be present. Each datanode serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses the TCP/IP layer for communication. Clients use Remote Procedure Call (RPC) to communicate between each other.

HDFS is the primary distributed storage used by Hadoop applications. A HDFS cluster primarily consists of a Name Node that manages the file system metadata and Data Nodes. The HDFS architecture depicts basic interactions among Name Node, the Data Nodes, and the clients. Clients contact Name Node for file metadata or file modifications and perform actual file I/O directly with the Data Nodes.

The **Hadoop distributed file system** (**HDFS**) is a distributed, scalable, and portable file-system written in JAVA for the Hadoop framework. Each node in a Hadoop instance typically has a single namenode; a cluster of datanodes form the HDFS cluster. The situation is typical because each node does not require a datanode to be present. Each datanode serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses the TCP/IP layer for communication. Clients use Remote Procedure Call (RPC) to communicate between each other.

Map-Reduce is the heart of Hadoop. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster.The term Map-Reduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map-Reduce implies, the reduce job is always performed after the map job.



3. Analysing the Events

These are some sample security related events which should not be committed inside the office premises. The numbers denoting the event Ids for each events it possesses distinct event Id. For different machines like Windows, Linux etc., has separate event id list.

Event ID	*Description*
41	The system has rebooted without proper shut down.
592	New process has been created.
1102	Logs cleared in windows events.
4726	User account was deleted.
4782	Password hash an account was accessed.
4616	System time was changed.
7306	USB stick inserted.

Event log Descriptions

keywords	Date& Time	Source	Event ID	Task Category	Event Source Name

Audit success	16012014 23:53: 05	Ms Win Security Audit	4797	User Acct Management	An attempt was made to query the exist of blank passwd for an account

Fields in Security Log

This sample security log depicts that the current event attempted to query the existing blank password. Likewise many events occur with its unique event ID.

This paper aims about in analysing and generating a report of the windows security event log analysis carried out by hadoop infrastructure. The project is to help the administrator of the large enterprise to analyse the security related events occurred in that Organisation faster and generate report of huge volume log file in short period of time. This aims about making the work simpler. The complexity of the project is that it is typical to collect such voluminous log file say some Terabytes (TB) or in Petabytes (PB).

The main objective to use event Ids in this paper is to find the employee who violates the policies of the organization. They may inserted the USB stick or downloaded any movies, or deleted any software etc , to know all these the administrator wanted to make ease the analysing of such security related events among Big Data he intends to apply in Hadoop infrastructure.
The aim of this approach is to simplify complex programming tasks.

Here are the some sample security event logs with some audit failure and success.

Audit Success 15-Jan-14 11:58:59AM Microsoft-Windows Security-Auditing 4672 Special Logon "Special privileges assigned to new logon.

Subject
Security ID: SYSTEM Account Name: SYSTEM
Account Domain: NT AUTHORITY
Logon ID: 0x3e7

Privileges: SeAssignPrimaryTokenPrivilege
SeTcbPrivilege
SeSecurityPrivilege
SeTakeOwnershipPrivilege
SeLoadDriverPrivilege
SeBackupPrivilege
SeRestorePrivilege
SeDebugPrivilege
SeAuditPrivilege
SeSystemEnvironmentPrivilege
SeImpersonatePrivilege"

5. Hadoop Algorithm Significance

Hadoop Run jobs processing 100's of terabytes of data, it Need cheap computers. Fixes speed problem (15 minutes on 1000 computers), but, Will be efficient and reliable, content optimization. Fault tolerance by detecting faults and applying quick and automatic recovery. The Data can be accessed via Map Reduce streaming. It is simple and robust coherency model .The Processing logic close to the data, rather than the data close to the processing logic. Portability across heterogeneous commodity hardware and operating systems. Scalability to reliably store and process large amounts of data. Economy by distributing data and processing across clusters of commodity personal computers. Efficiency by distributing data and logic to process it in parallel on nodes where data is located. Reliability by automatically maintaining multiple copies of data and automatically redeploying processing logic in the event of failures.

JOB TRACKER & TASK TRACKER

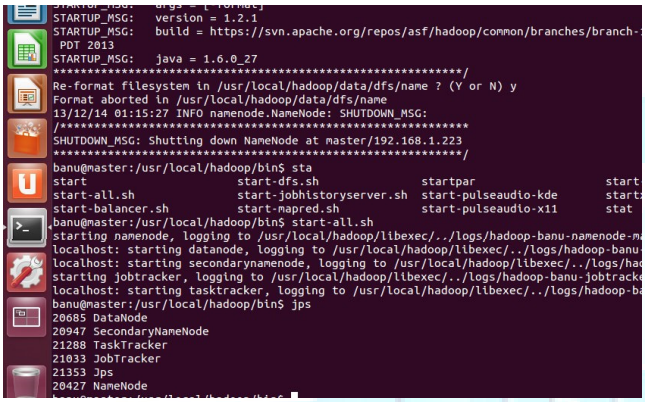
Above HDFS comes the map reduce engine. This has one Job tracker, to which the client application used to submit its map reduce jobs. Then this job tracker splits work out to available task tracker. With rack-aware file system Job tracker knows which node contains what data. If the work cannot be hosted on the actual node where data is present, then the job is assigned to some other node in the same rack. This reduces network traffic in the backbone network. If the Task tracker on each node has its own Java Virtual Machine process to prove the Task Tracker itself from failing itself from running job crashes the JVM.

A heart beat is sent from the Task tracker every time ti the Job tracker every single minute to update its status of presence. The allocation of the work to the Task tracker is almost single. Every task tracker has some specific slots. Every map or reduce task has its own slot. The Job tracker allocates work to track nearest to the data with available slot. If the Task tracker is very slow it can delay the entire

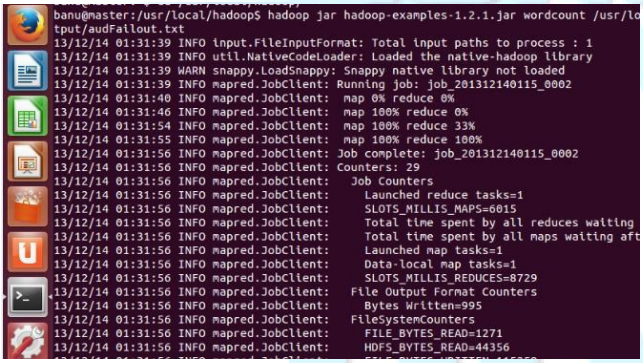
map-reduce job where everything can be end up with *Contents of the Directory* slowest task. However single task executes on multiple slave nodes

Screen Shots

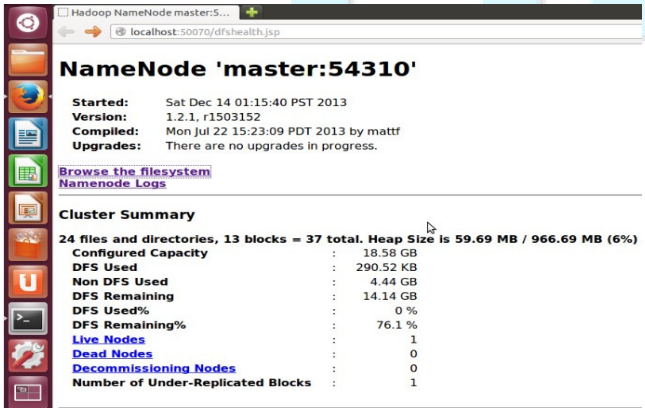
After starting up Hadoop

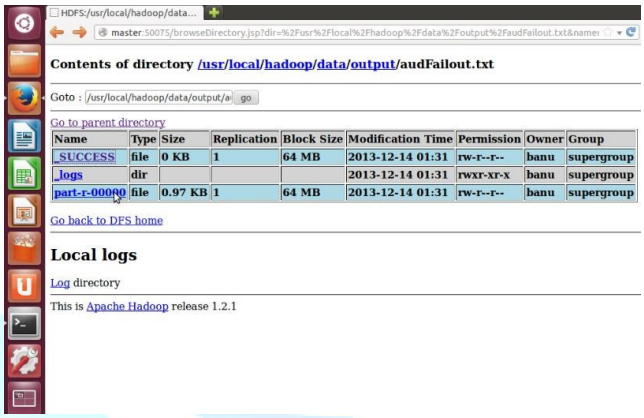


Running of Map-reduce task

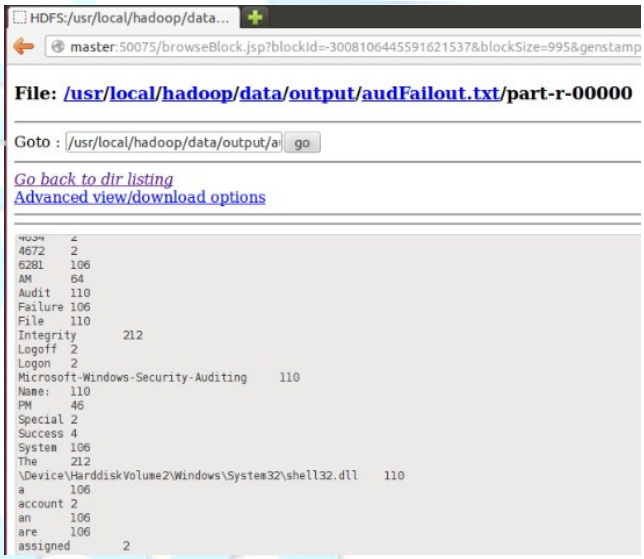


Hadoop Namenode





Output of Word Count Program of audit failure record



CONCLUSION

We tried to give detail information to industry that worked on operating systems to remove errors, so we do firstly and its Applications thorough study on event viewer of windows operating systems (client, server). Do analysis on every error which is occurred in runtime or time in installation of any application. Also analysis the type of error which is frequently occurred when operating system is run. We also save the log files from different type of operating systems for analysis and making approach for appropriate result, and graph. Our Project analysis can be useful for an administrator in an enterprise to collate and manage log files.

References

[1]. A. I. Verkamo ,H. Mannila, & H. Toivonen, "Discovery of frequent episodes in event sequences", Data Mining and Knowledge Discovery Vol. 1(3), 1997

[2]. G. Jiang, H. Chen, C. Ungureanu, and K. Yoshihira.Multiresolution Abnormal trace detection using varied-length ngramsand automata. In Proc. of ICAC, pages 111 –122, 2005.

[3]. J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. In Proc. of OSDI, 2004.

[4]. J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. In Proc. of OSDI, 2004.

[5]. R.Agrawal,T.lmielinski and A.Swami," Mining Association Rules between sets of items in Large Database", in proceeding of the ACM SIGMOD International Conference on Management of data ,1993,pp.207-206.

[6]. RistoVaarandi, A Breadth-first algorithm for mining frequent patterns from event logs, IEEE international conference, 2003-04.

[7]. RistoVaarandi "A Data Clustering Algorithm for Mining Patterns From Event Logs", IEEE international conference, 2003.

[8]. RistoVaarandi, "SEC - a Lightweight Event Correlation Tool", Proceedings of the 2nd IEEE Workshop on IP Operations and Management, 2002.

[9]. RistoVaarandi, "SEC - a Lightweight Event Correlation Tool",Proceedings of the 2nd IEEE Workshop on IP Operations and Management, 2002.

[10]. H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of frequent episodes in event sequences", Data Mining and Knowledge