Advances in Engineering Research (AER), volume 148
3rd Workshop on Advanced Research and Technology in Industry Applications (WARTIA 2017)

Research on Log Monitoring Technology Based on Flume

Wu Wenxuan[1], Wu Fei[1], Cai Yuxiang[2], Wang Qiulin[3], Lin Wei[3,a]

[1]State Grid Fujian Electric Power Company, Fuzhou 350003, China

[2]State Grid Fujian information&telecommunication company，Fuzhou,350003,China

[3]Fujian Yirong Information Technology Co.Ltd，Fuzhou,350003,China

[a]Linwei8868@163.com

Keywords: flume; log monitoring; data acquisition

Abstract. In the network environment, the significance of log monitoring lies in collecting resources and maintaining network security. This paper enumerates the log types in the network environment. Based on the illustration of flume structure, the design and implementation of log monitoring technology based on flume are further studied, hoping to provide a guarantee for the improvement of the network log monitoring level.

Introduction

Network log, namely, browsing and using records of network system, can be divided into various categories such as operating system log and application system log according to log source. Flume, also known as distributed log acquisition system, is applied to the network system to effectively realize the collection of logs generated during the operation of the network, which is of great value to analyze the characteristics of network browsing and improve the accuracy.

Category of weblog

The category of weblog is as shown in Table 1[1]:

Table 1 Category of weblog

Category	Category 1	Category 2	Category 3	Category 4
Content	Operating system log	Application system log	Security device log	Network device log

(1) Operating system log: including application program log, system logs, and security logs, etc. The record structures include event number, source, grade, and category. And data is generally represented in 16 hexadecimal. (2) System log (Table 2) consists of user name, request date, event, and page communication protocol, etc. (3) Security log: logs generated by firewalls and intrusion detection systems. (4) Network device log: the log generated by devices such as network switches and routers, consisting of time, messages, devices, and importance levels of the event.

Table 2 Application system log composition and format

Field	Format
User name	UserID
Request date and time	[DD/MM/YYYY.HH:MM:SS+TimeZone]
Page communication protocol	“Get xxx.host.subent.domain.net”

Flume structure

The composition of flume is as shown in Table 3[2]:

Table 3 Composition of flume

Number of plug-in	Plug-in 1	Plug-in 2	Plug-in 3
Type of plug-in	Interceptor	Pipeline selector	Sink thread

(1) Interceptors: The function is to implement data modification and check. (2) Pipeline selector: includes two categories of default pipeline selector and multiplexer channel selector. The former can manage different pipelines simultaneously based on the same event. The latter manages different channels according to the header address of each event. (3) Sink thread: The function is to activate the Sink and ensure that the system load is balanced.

Design and implementation of log monitoring system based on flume

Design of data acquisition based on Flume

Data acquisition process

Data acquisition process based on Flume is as shown in Table 4[3]:

Table 4 Data acquisition process based on Flume

	Agent1		
Data Generators	Agent2	Data Collector	Centralized Stores
	Agent3		

As shown in Table 3, in the operation of the software or network device, the data generator generates a large amount of data. After collecting the above data, Agent can store it in the HDFS HBase and analyze the data to complete the process of log monitoring.

Function of automatic data acquisition system

The functions of automatic data acquisition system based on flume are shown in Table 5.

Table 5 Functions of automatic data acquisition system based on flume

Constituent items	Item 1	Item 2	Item 3
Constituent content	Data storage and analysis	Data update and acquisition	Prevent “reading while writing”

The system is built on the basis of flume, and the database is mainly HBase.

Design of the system

Flume data acquisition system can create Java virtual machine according to the demand of data acquisition. Different virtual machines are named as event 1, event 2, and event 3. The function of the virtual machines is to collect data according to the system settings. Different events are independent, and data from different sources can be processed simultaneously.

System topology

In essence, the Flume data acquisition system is a distributed system, in which each machine has its own network topology. Therefore, the network topology of Flume data acquisition system mainly includes point-to-point topology, mesh topology, and ring topology. (1) Point-to-point topology: connects a information point to the other. There is only one logical and physical path between the two information points. Ethernet and frame relay networks are point-to-point topology. (2) A mesh topology connects multiple subnets together. In the same subnet, hubs, repeaters and other devices need to be connected to each other, and information transmission is mainly completed in the subnet. (3) Ring topology: compared with other topologies, ring topology does not require a terminator. Topologies have different characteristics. In order to reduce the network latency between the database and the Agent, the star structure is chosen as the main topology in the system.

System module design

(1) The front end calls flume client module design: flume client module can be designed based on the jar package of flume client. Firstly, the Server address is defined, and then the log information is defined, which lays the foundation for the module design. Secondly, flume Message is initialized and definition file is configured. Finally, the configured file can be directly sent to the Server to complete the module design process. (2) Flume Server module: flume client realizes the connection through thrift. When the information is returned, the connection has been implemented and the data is sent to flume Server. When the message is not returned, the connection is not implemented. Try connecting again until the connection is successful. (3) HDFS interaction: Data received by source can be transmitted to channel and stored in HDF. (4) SAcontroller module storage statistics: SAcontroller is composed of two modules, namely, HDFS and MapReduce. When the data is accepted and stored in the HDFS, the MapReduce can realize the statistics of the data, and ensure that the log data monitoring process can be implemented smoothly.

Implementation of log monitoring process based on flume

Implementation of flume client

The function of the data collection module is to collect the log data, while the log cannot be stored or monitored. The collected log is generally written to the server. The log information need to be actively sent to the server to implement the flume client. (1) Flume client framework is designed, and the data is sent by the queue method. The front-end APP stores the collected data into the emission queue and sends the data one by one in the order of storage. The back-end Thread pool tool can organize and access data in the queue. When the data is not empty, 100 messages need to be removed each time to realize communication.(2) Parameter validity judgment: Open cache queues, and create threads. The queue is stored in threads, and sent to server according to its sequence.(3) Parameter design: when the data is initialized, for example the front end has no API, and the information is empty, the default parameters need to be loaded into the information event to achieve log monitoring. When API exists and the information is not empty, events and running cycles can be set by defining the IP address and the Server port, thus enabling log monitoring. Client has the function to provide default configuration files that can be invoked by the front end. After the Message Params gets the configuration file, the parameter values are recognized. (4) Data protection: There are two types of flume client data protection mechanisms. One is retry and record policy mechanism, and the other is record policy mechanism. The two mechanisms can be used interchangeably to avoid data protection process vulnerabilities and improve the security of data monitoring.

Implementation of flume sever

（1）Flume sever consists of source and sink. The former refers to the data source, namely, the source of the log. The latter refers to log storage. The size of the log data is the main factor affecting the storage mode. If the logs are large enough, they need to be stored in the Hadoop. Otherwise, the logs can be stored directly in the file system. (2) Configuration files include flume, source, and flume sink. The attributes of the flume source (Table 6) include type, host, port, and channel. (3) Implementation process: the base class translates logs the log into event and sends it to channel.

Table 6 Attributes of flume source

Attribute	Default	Description
type	Null	Source type
host	Null	Binding domain name or IP address
port	1499	Binding port
channels	Null	Channel name

Implementation of SAcontroller

(1) Design framework: is composed of HDFS and MapResuce job. (2) Implementation process: the purpose of SAcontroller is to monitor the operation of an individual user in one hour. SAcontroller can run Report in a pseudo distributed environment, and realize function of map by integrating the

mapper class. In the main program, MapResuce can provide API to generate the number of independent users.

Monitoring test

Test environment

(1) The hardware of the system meets the standard of log monitoring test. (2) The operating system is windows7. (3) The test module is flume client and SAcontroller. (4) The unit test tool is Junit. (5) The test platform is DCSS. (6) The unit test method is white box test, and function and stability test method is black box test.

Test result

(1) Configuration file test results: configure the default file to initialize message successfully; Set the correct format for the Server ID address test successfully; Activate the data protection mechanism and obtain the parameters successfully. (2) Data protection mechanism test results: Data is sent to the message when the sever process is not started. The test results show that the client data transmission fails and the data record is located locally. The data is sent to the category when the sever process is not started. According to the test results, the client fails to send the message after three consecutive attempts, and the data record is located locally.

Conclusion

Log monitoring function test results show that the log monitoring system based on flume designed in this paper can effectively meet the demand of log monitoring. The system should be applied to network log monitoring process in relevant fields in order to achieve the collection and analysis of network data, and provide guarantee for improving the network operation security.

References

[1]Guo F, Zhang Z, Hu D, et al. Research on adaptive tide numerical simulation based on steering dynamic monitoring[J]. Environmental Earth Sciences, 2015, 74(10):1-11.

[2]Mao L, Carrillo R, Escauriaza C, et al. Flume and field-based calibration of surrogate sensors for monitoring bedload transport[J]. Geomorphology, 2016, 253:10-21.

[3]Dhont B E M, Rousseau G, Ancey C. Continuous Monitoring of Bed-Load Transport in a Laboratory Flume Using an Impact Sensor[J]. Journal of Hydraulic Engineering, 2017, 143(6).