A Network Traffic Analysis System based on Big Data

Changhua Yi [1, a], Lixin Ke [2, *]

[1] Modern information and Educational Technology Center, Shanghai Ocean University, Shanghai, China

[2] Education Committee Team, Shanghai Municipal Big Data Center, Shanghai, China

[a] chyi@shou.edu.cn, *Corresponding author e-mail: klx@shec.edu.cn

Abstract. With the rapid development of network technology, network applications are becoming more and more popular. In order to ensure the availability of network and the stable operation of key business, the methodology of network traffic analysis appeared accordingly. Compared with the normal non-big data traffic analysis, using big data for traffic analysis and monitoring can solve the problems of single data source, difficulty in chart expansion horizontally and chart customization, etc. Based on big data technology to build a unified platform on traffic analysis and monitoring, it is able to integrate with the network analysis of data and other dimensions of the associated monitoring information to achieve efficient service on fault warning and navigating location.

Keywords: Big data; Cloud platform; Network management; Traffic analysis; Traffic mirroring.

1. The Value of Big Data Applied in Network Traffic Analysis

1.1 The characteristics of large network data

Big data generally refers to huge amount of unstructured data that cannot be processed in real time by conventional analysis methods, and its characteristics can be decomposed into five “V” to describe:(1) Velocity: the speed of collecting and analyzing new data; (2) Volume: new data Volume is large; (3) Value: the Value of extracting data;(4) Variety: Variety of data types;(5) Veracity: quality or credibility of data collection [2].

Network traffic monitoring data and Web applications monitoring data in high-speed network infrastructure, are in full compliance with the main characteristic of big data, so we can call the network traffic and service monitoring data network big data for short.

1.2 Challenges of big data applied in the network management

The basic process of big data analysis is as the following steps: data acquisition, data storage, business analysis and intelligent exhibition. Big data analysis is a process to examine a large number of different data collection, the main purpose to find out the correlation of hidden information is unknown and useful information, such as business development trend so as to help users make more wise business decisions. Compared with the traditional business intelligence analysis, big data analysis is different, with user analysis as a case, the traditional business intelligence analysis requires the following process: sampling information events, artificial question &answer, data analysis, summing up reason, summary event rules and responsive for future events. And big data analysis process is completely different, it is a form of advanced analysis, involving complex applications, including prediction model, statistical algorithms, and supported by the high performance analysis system hypothesis analysis element, its analysis process mainly includes: collect all of the information event data (after cleaning), machine learning, build multi-factor model events, forecast warning events, etc., by comparing the comprehensiveness of traditional business intelligence analysis and big data analysis processing, it is clearly that big data has more advantage to deal with diverse large data analysis.

The main trend of the industry's big data analysis method is using large server clusters, it adopts the method of machine learning to holographic large-scale continuously to all data analysis, so as to support the prediction and early warning to of business goals. Such big data analysis has a serious

drawback, that is, the data contains amount of impurities, it need to consume a large amount of computing and storage resources, at the same time, if handled improperly, it will seriously affect the accuracy of the analysis, resulting in failure. This is a big problem that big data analysis has to face.

If there is no efficient tool to collect and analyze full traffic data in the network in real time, there are only two possibilities.

1). The real-time analysis of data after collecting the flow samples, the obvious disadvantage of this method is that it is easy to cause the analysis results to be biased and the prediction results are distorted;

2). The original data of the network was saved firstly, then the offline analysis was carried out, which required a large amount of computing resources for analysis, and the analysis results were not timely.

Thus, without the support of good big data analysis application, the analysis results are not conformed to the purpose and significance of big data analysis. Network big data analysis urgently requires innovative acquisition storage analysis and display technologies.

2. Design of a Network Traffic Analysis System Based on Big Data

The network data analysis system mainly consists of five modules: The data acquisition module can implement data collection for network traffic data packages, network equipment log, equipment information, SNMP information and server probe information; The data receiving module implements operations such as verification, cleaning, conversion and reduplication of the collected data, and converts it into the data format required by the system., and converts it into the data format required by the system.; The data storage module is in charge of persistent storage of data; The data processing module mainly realizes the implementation of machine learning algorithms to predict and analyze results of data collection. The data display module mainly displays the data results processed by various analysis algorithms, and its display method matches WEB front-end technology mainstream.

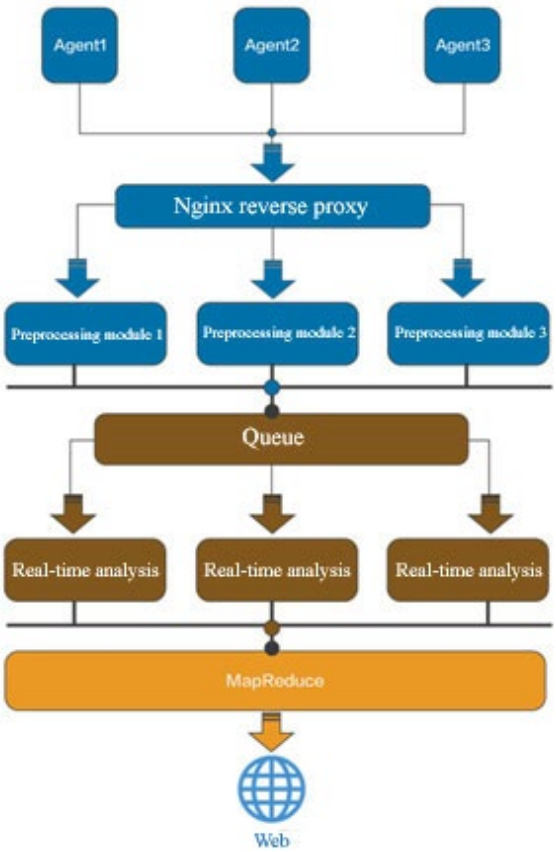


Fig. 1 System architecture based on virtual cloud

The functional design of each part is as follows:

a. Let Nginx as a unified data interface, to receive collected data information through the HTTP protocol, and convert to JSON format, then distribute data to the preprocessing module component group, to play a role in load balancing;

b. The preprocessing module performs formatting and preprocessing on the received source data, and can modify field names, increase data types, time stamps, etc., and forward them to the message queue;

c. Data buffering the message queue, enabling batch processing for real-time analysis module component group reading;

d. The real-time analysis module takes the data out of the cache, parses the fields in it, and stores the parsed results in the file system.

e. MapReduce framework is used for data storage and query;

f. Web provides data visualization and human-computer interaction, turning application requirements into complex query syntax to query and combine data.

2.1 System Framework

The overall framework of the system is shown in Figure 2:

2.2 Data Acquisition Module

The data acquisition module mainly includes network traffic hardware probes, software probes, SNMP probes, and log collection services. Through the bypass mirror (Tap or Span), local monitoring (Local mode), probes and other methods, the packets can be parsed into more meaningful flow information.

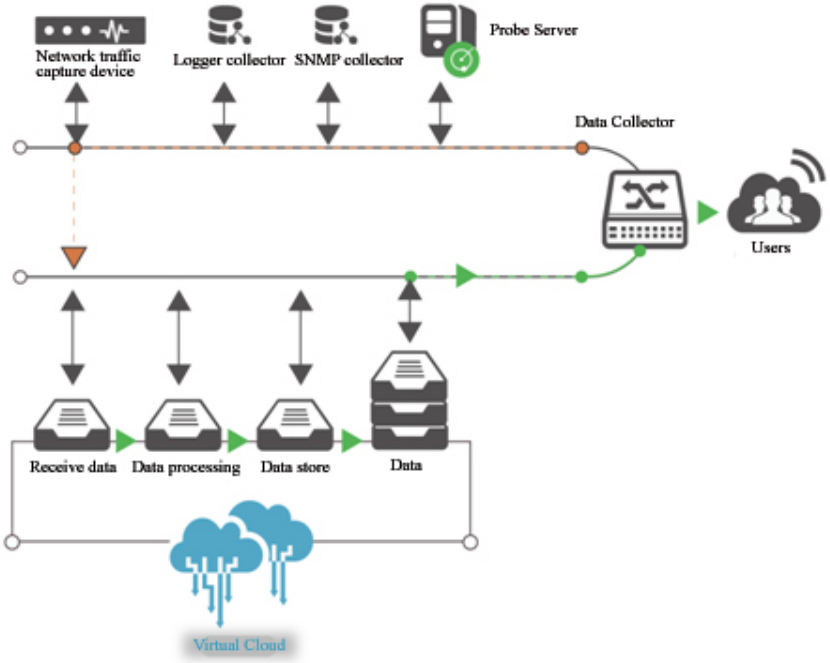


Fig. 2 Overall System Framework

For example: IP 61.140.38.91 visits 110.75.196.70. The website produces the message on the network, through message depth analysis, parses into a structured record:

{
“dst_ip”: 1850459206,
“dst_port”: 80,
“host”: “wpstatic.china.alibaba.com”,
“http_method”: 1,

“https_flag”: 0,
“in_bytes”: 2139,
“in_pkts”: 2,
“l4_protocol”: “tcp”,
“latency_nsec”: 27902,
“latency_sec”: 0,
“out_bytes”: 449,
“out_pkts”: 1,
“refer”: “http://cjgdj556.cn.alibaba.com/”,
“src_ip”: 1032595035,
“src_port”: 55366,
“status_code”: 304,
“url”: “/js/module/part/album.js”,
“url_query”: “”,
“user_agent”: “Mozilla/5.0 (compatible; MSIE 9.0; Qdesk 2.3.1185.202; Windows NT 6.1; Trident/5.0)”
}

2.3 Data Receiving Module

The data receiving module is mainly composed of multiple preprocessing and message queue components. The queue implements load balancing through the Nginx reverse proxy collection service. The log collection program uses the Rsyslog protocol to collect the system logs and service logs of different devices. Through the regular analysis of these text-type data, the system finally generates the structured Json data required by the system.

For example: System log from a server:

<134>Jun 27 14:50:29 localhost fwlog: log type: service control or application control, user: null), source IP: 106.14.19.151, source port: 46809, destination IP: 10.1.10.17, destination port: 25, Application Type: Other, Application Name: null), System Action: Logged\n

Convert to:

{
"facility": "local0",
"hostname": "localhost",
"log_msg": "fwlog: log type: service control or application control, user: null), source IP: 106.14.19.151, source port: 46809, destination IP: 10.1.10.17, destination port: 25, application type: Other, Application name: null), System Action: Recorded,"
"log_prio": "134",
"log_ts": "Jun 27 14:50:29",
"severity": "info"
}

2.4 Data Processing Module

The network data that needs to be analyzed mainly includes HTTP, DNS, FTP, and DHCP types. The module can perform address database parsing on all IP type fields and expand the field on the original structured data. Extended information includes: province, city, and street. Geographic information, operator information, etc.

E.g:

"dst_ip": 1944526390

Expands to:

"dst_ip": {
"decimal": 1944526390,
"dotted": "115.231.30.54",

"isp": "Telecommunications",
"region": "Huzhou"
}

For HTTP type data, the module parses the User-Agent attribute in the HTTP header field to obtain client browser-related information, including: browser and browser type, operating system and operating system version, browser engine, and version, as well as whether it is a script trigger, etc.

E.g:

Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.1 (KHTML, like Gecko) Maxthon/4.1.3.5000 Chrome/26.0.1410.43 Safari/537.1

Expands to:

{
"browser": "Safari",
"browser_version": "4.1.3.5000",
"os": "Windows 7",
"platform": "Windows",
"engine": "AppleWebKit",
"engine_version": "537.1",
"bot": false,
}

2.5 Big Data Processing Based on MapReduce

MapReduce is a programming model based on the principle of divide and conquer algorithm. It mainly encapsulates details such as parallel computing, fault tolerance, data distribution, load balancing, etc. The big data processing module uses MapReduce technology to easily realize the horizontal expansion of the cluster, enabling the cluster to have efficient parallel computing ability.

MapReduce-based big data processing system supports automatic allocation and execution of tasks on cluster nodes, collecting calculation results, and completing the whole process of data distribution and storage, data communication, fault-tolerant processing and other parallel computing [4]. The processing process is shown in Figure 3.

2.6 Data Storage Module

The data storage module divides multiple data collected by different probes according to the time stamp. Different types of data which have similar time series are stored in different tables under the same sub-database.

In order to improve the efficiency of data query and the utilization rate of disk, the system provides a series of script tools to periodically optimize, close and age the library.

The system supports uniform settings for a variety of data fields, including field types, whether strings open participles, and so on.

Here's an example of an inline associative array:

“comments”: [
{
“name”: “Zhang San”,
“comment”: “Undergraduate Student”,
“age”: 30,
“stars”: 5,
“date”: “2017-01-11”
},
{
“name”: “Li Si”,
“comment”: “Staff”,
“age”: 45,

"stars": 2,
"date": "2017-01-22"
}
]

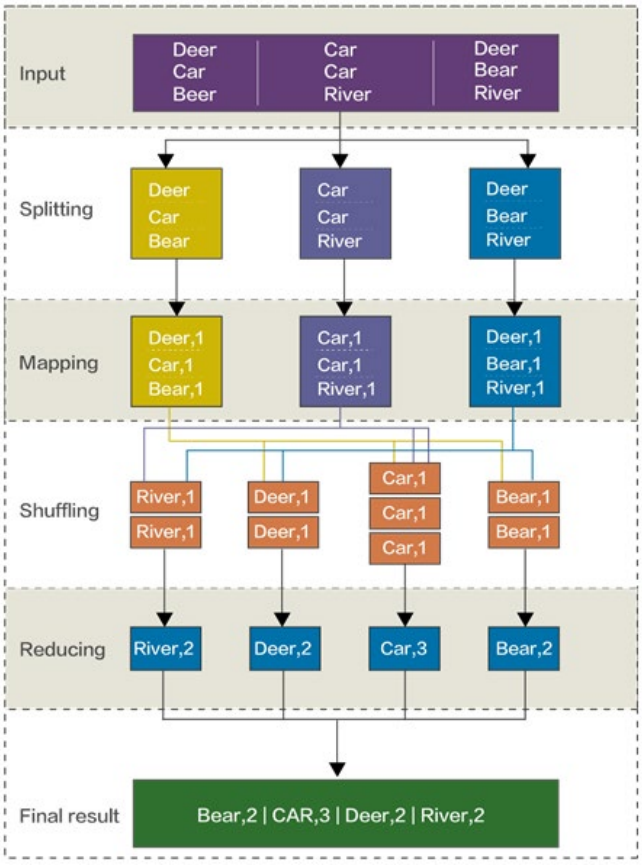


Fig. 3 Big Data Processing Based on MapReduce

3. System Implementation

We set up an experimental platform on the VMware virtual data center. The platform consists of 18 virtual machines, they are, 4 data processing and analysis nodes, 9 data storage nodes, and 2 data display integrated nodes. The machine is configured with 16 vCPUs, 16 G bytes of memory, and the operating system is Ubuntu 16.04. Nginx 1.10 is used to achieve load balancing of key data nodes. The platform can monitor the interface traffic of the device in real time through the SNMP protocol (as shown in Figure 4). When the traffic exceeds the threshold fluctuation range, the alarm information is sent to the operation and maintenance personnel in real time.

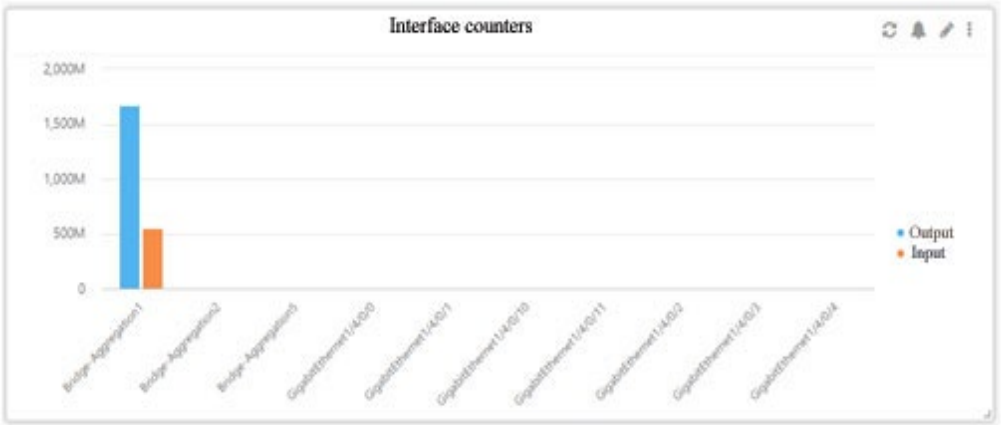


Fig. 4 Interface traffic monitoring

In the application layer traffic analysis module, we can analyze the common applications such as Http, DNS and other common applications.

4. Conclusion

Building a comprehensive analytics platform based on big data can point the way for all aspects of business development, including new development opportunities, more effective business promotion, better user services, and higher operational efficiency. It also can improve the status of IT operation and maintenance from the perspective of business development. This paper designs and simply implements a big data network traffic analysis system based on a virtual cloud environment, which can be applied to regional networks of different sizes to analyze traffic flow, flow direction and find abnormal traffic. These methods would have great practical value for reducing operational risks and improving the quality of service. In the later period, we will use the platform's big data analysis capabilities to expand the intelligent functions, integrate the operation and maintenance expert experience, so as to provide more accurate problem positioning for IT operation and maintenance.

Reference

[1] Chen Hao. The principle of NetFlow and its application in the analysis of Internet traffic [D]. East China Normal University, 2006.

[2] Xie Pingjun. Parallel Frequent Itemset Mining Base On MapReduce [D]. Nanjing University, 2012.

[3] YANG Houyun, WANG Jungang, GONG Hanming. Design And Implementation of Dataflow Monitoring in Campus Network[J]. Journal of Beijing Information Science and Technology University, 2009,24(4): pp87-91.

[4] Hadoop Learning Notes: Detailed Explanation Of The Mapreduce Framework. http://blog. jobbole.com/ 84089 / [EB/OL].

[5] Xia Guangfeng. Network Analysis Technology And Its Application In Network Management [J]. Computer Knowledge and Technology, 2009 (03).

[6] He Changlin, Zhang Lei. Analysis on NetFlow data of worm virus based on PATRICIA tree [J]. Automation and instrumentation, 2015(1):pp135-137.

[7] Liu Danfeng. Research on Network Data Flow Analysis and Network Optimization for Enterprises [D]. Harbin Engineering University, 2012.

[8] Yang Qing. Mobile Internet Traffic Analyzing Of Cell based on Cloud Computing [D]. Beijing University of Posts and Telecommunications, 2014.