# DESIGN AND IMPLEMENTATION OF REAL-TIME LOG ANALYSIS SYSTEM OF MAP WORLD PLATFORM

Hongping Zhang [1,2,*], Wei Huang [1] , Jing Yang [1]

[1] National Geomatics Center of China, Beijing, China- (zhanghongping, huangwei, yangjing)@ngcc.cn
[2] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China

**Commission IV, WG IV/8**

**KEY WORDS:** Real-time, Log analysis, World Map

**ABSTRACT:**

In the big data era, real-time log data analysis is becoming the important demands for Internet enterprises, behind these data hiding big value. Map Word, the National Platform for Common GeoSpatial Information Services in China, is a one-stop website providing geospatial information services to personal users, enterprises, professional agencies and governments. After 7 years’ development of the platform, the traffic increased significantly, reaching 200 million service requests per day. But due to the lack of effective analysis and processing technology, the log data did not play its value, which lead to disconnection between the top-level design of national common geospatial information services and the actual demands of the users in a way. Now, the geospatial information service in China is trying to shift from the data production driven to the demand driven actively, and how to understand the demands of users became one imperious issue for research. In addition, the access behaviour of group users to common geospatial information services has a social nature and there is a certain group access behaviour pattern. This mode has high intensity of access aggregation and spontaneity, and determines the demand of common geospatial information services for cloud computing resources. Parts of the above demands can be analysed from the log data. Therefore, how to develop a log analysis system for unified real-time collection, real-time analysis, centralized storage, and graphical display is the key to support the demands. Flume、Kafka、Storm、Redis and HBase have been integrated to design and implement a distributed real-time log analysis system supporting online and offline log analysis. The system is composed of log collection module, asynchronous communication module, real time analysis and calculation module, data cashing and storage module, and visualization module. The system was release and integrated with Map World in June 2017 successfully, and the implementation of the system indicates that it can efficiently solve the problems of real-time log data collection, real-time analysis, real-time storage, real-time query, massive data storage, offline analysis, etc. It played an important role in map data update, policy making, product decisions, online server load prediction, resource allocation optimization, Internet security improvement and operation funs evaluation.

## 1. INTRODUCTION

As the unstructured records produced during the operation of the software systems, logs are a management tool to record the behaviours of systems and network users, which describe the behaviours on application services and user interaction (Qu G., 2016). More important, it can be used to monitor the abnormality of websites and application services. Through log data mining, statistic and analysis, it can help product managers and decision makers to find the potential production problems so as to promote replace of products, improve user experience and optimize product operation. At present, massive amounts of logs are highly valuable to all Internet firms, for example, Taobao and Baidu. According to the logs, users behavior patterns and personas are minded and this is the basement for particularly recommendations (Bai Y., 2013; Cheng M. & Chen H., 2011). To deal with logs with the characteristics of huge data volume, high complexity and high demand of real time, it puts forward higher requests to the capability of real-time computing and massive storage in the whole process. The study on the real-time streaming data processing technology to the massive log data has been applied widely in the Internet, such as real-time monitoring, real-time recommendation, and real-time statistics (Liu F., 2017). The Real-time collection and analysis of massive logs has become the important big data service in the Internet firms, which is the main method to understand user behaviour deeply, evaluate marketing effectiveness, optimize the product experience, and improve operation efficiency. Map World, the National Platform for Common GeoSpatial Information Services in China, established in 2011, has become an important component of digital China to provide authoritative, standard and unified geospatial information services to personal users, enterprises, professional agencies and governments (Jiang J., Wu H.& Huang W., 2017). With the deep and wide application of Map World, the users of the platform increase rapidly, it has over 200 million log records of page and service requests everyday, which presents challenges to the existing log data collection and analysis system. Meanwhile, in the era of cloud computing and big data, the geospecial information service is gradually transforming from data production driven to user demands driven, and the core issue for this transformation is how to understand the users’ demands by extracting useful information form the log data promptly and effectively (Wu H., Li R., Zhou Z., Jiang J., & Gui Z.,2015.). Neither the storage mode nor the calculation efficiency, the traditional log collection and analysis technology

* Corresponding author

cannot meet the above requirements. Therefore, there is an urgent need to use the existing open source technology framework to build a reliable and efficient real-time collection, analysis and processing system for massive log data of Map World.

It has proved that Flume, Kafka, Storm, Redis and HBase are the mainstream technologies of open source frameworks and databases for solving massive logs now (Chen Q., & Zhou L.,2012.). This paper aimed at the requirements of real-time statistical analysis, deep mining and visualization of the Map World logs, a distributed log real-time processing system based on the above technologies was designed and implemented, the system can be used to solve the log data collection and storage, real time log stream data processing and visualization, which can provide reliable evidences for monitoring and optimizing services of the platform.

## 2. RELATED TECHNOLOGIES

### 2.1 Flume

Flume (Apache Flume[OL]) is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application. Flume can store log information collected from multiple website servers into HDFS/HBase efficiently. A Flume event(JVM) is defined as a unit of data flow having a byte payload and an optional set of string attributes, consisting of three components: Source, Channel and Sink. Source is responsible for log production or collection, and encapsulating the data source into a Flume event. The event is stored into one or more channels, and the channel is a passive store that keeps the event until it’s consumed by a Flume sink. Sink is responsible for deliver the collected log data to the storage and analysis modules (Chen F., 2016).

### 2.2 Kafka

The logs collected include page views, search, map roam, API and other user interactive behaviours and the amount of log records each day is over 60GB. The traditional solution for log processing and analysis is offline, and it has a great time delay. Kafka (Apache Kafka[OL]) is a distributed publish-subscribe messaging system, which can handle all the action flow data in a consumer-scale website. As a lightweight messaging system, it contains three components: Producer, Broker and Consumer. Data is transmitted to Broker from Producer, and Broker’s role is mid-tier caching and distribution.

### 2.3 Storm

Storm (Apache Storm[OL]) is a free and open source distributed real-time computation system. Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing. The processing speed of Storm is fast, and according to the statistical data from Twitter, over a million tuples can be processed per second per node. Asynchronous communication message queues, relational databases and non-relational databases can be integrated with Storm. Zookeeper is used to coordinate the Storm nodes and save some public information, such as system heartbeat. The most important components of the cluster are Spout and Bolt. The former is responsible for reading data from the external data source and transmit it to the calculation task, and the latter is responsible for streaming data calculation received from Spout.

### 2.4 ECharts

ECharts (ECharts [OL]) is a pure JavaScript library for data visualization based on the lightweight Canvas library ZRender, supporting mainstream browsers such as IE8/9/10/11, chrome, Firefox. It is comprehensive charting library offering an easy way of adding intuitive, interactive, and highly customizable charts to your commercial products. With original features like Drag-Recalculate, Data View and Scale Roaming, ECharts improved the user’s ability for data mining and integration. More important, it supports data visualization with online maps easily.

## 3. GENERAL DESIGN OF THE SYSTEM

As a basic core service for Map World, the real-time log analysis system is a distributed cluster for real-time log data collection, processing, storage and data presentation. It also provides log analysis services to the authority management system and monitoring system. The design of the system bases on the modular method and standard interfaces to reduce the dependence among different modules as much as possible, and increase the stability and reliability. The system is composed of log collection module, asynchronous communication module, real time analysis and calculation module, data cashing and storage module, and visualization module, as shown in Figure 1.

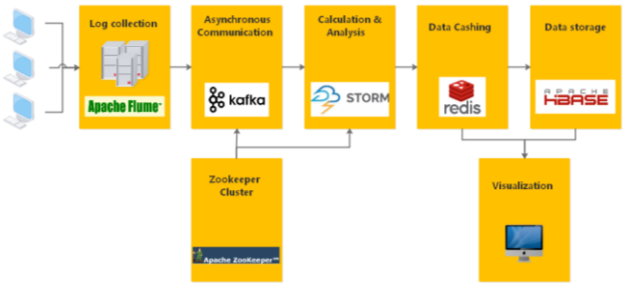


Figure 1. Overall architecture of the system

### 3.1 Distributed Log Collection Module

All the external service interfaces of Map World platform are proxied to the users by Nginx, which is used for a reverse-proxy server onto the web-servers. Nginx records every request from the users and store it into the log file for analysis. Flume is used as the collection framework to collect massive distributed logs on the different servers. The Agents are installed on different servers for pushing the real time log data in the specified directory to the Collector. The Collector is deployed on a central server and can management the flows intelligently. After data pre-processing, they are delivered to the publish/subscribe asynchronous messaging module, and as the endless data flow to the Storm. The whole collection process is shown in Figure 2.

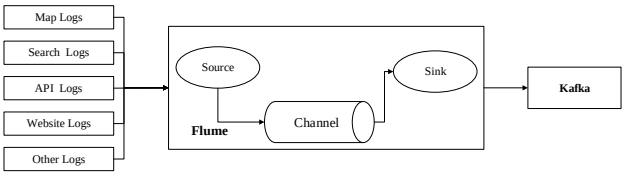


Figure 2. Distributed log collection architecture

### 3.2 Asynchronous Communication Module

The asynchronous communication module is used to reduce coupling degree between log collection and calculation, and ensures the stability of log calculation. The log records to be processed is nearly 10-15 thousand in each second, and the calculation models are different because of different service interfaces. It is a bottleneck that keeping the data to be processed transmitting between the modules efficiently. As a middleware, Kafka is used in the system as the asynchronous communication message queue to receive any kind of log data source, and Producer writes the logs on the server into Kafka continually. The real-time calculation module pulls the log data in the messaging queues to consume.

### 3.3 Real-Time Analysis and Calculation module

This module is responsible for processing and analysing the data flow in the messaging queues of Kafka, and calculating the data with different calculation model based on different business requirements, and the results are stored in the database. The real-time analysis and calculation service is developed based on Storm cluster, including data acquisition and logical analysis engine. Spout, the component of Kafka is used to pull data, and after internal transformation, the data is assigned to the logical analysis engine for calculation. In this way, the problem of data congestion while real-time calculation in Storm directly can be avoided. The logical analysis engine is responsible for data cleaning, formatting, statistic, space conversion, IP conversion, etc. Parts of the processed data is stored in the Redis cluster, and others is stored in the HBase for the periodic offline analysis.

### 3.4 Data Cashing Module and Storage Module

This module has two parts: Redis and HBase. The hot data for the day from 0:00:00-23:59:59 is stored temporarily in the Redis cluster to increase data reading efficiency for visualization module. After 0:00, the data will be stored in the HBase and deleted from Redis.

### 3.5 Visualization Module,

Data visualization has become one of the most important means of man-machine interaction. This module mainly attempts to demonstrate the operation and visiting situation from the logs of the platform more intuitively, provides auxiliary decision making support and improve the management level of Map World. A big screen system for the monitor center and a log management system for operations staff are separately developed, supporting real-time data, many chart types, such as line (area), column (bar), scatter (bubble), pie (doughnut), word cloud. In addition to the regular visualization methods, the special visualization methods are integrated, for example, hot map, administrative diagram, map layer overlay.

## 4. DATASHEET DESIGN AND CLEANING RULES

### 4.1 Log source Datasheet Design

Since Map World platform has hundreds of servers, it is important to access the logs according to the log source datasheet, avoiding unnecessary resource usage. The configuration information such as log types, routes, server name is designed in the sheet. The log source rout must be the same as the Agent of the collection module. When the log data is processed in Storm, the existence of the log’s source route will be checked in the database. If there is no confirmation about the log in the storm, the log can be abandoned. Only the logs with confirmation information, they can be processed in Storm. The main columns of log source datasheet are shown in Table 1.

| Fileds | Types | Description |
|---|---|---|
| log _source _id | int | Log source ID |
| log _source _name | varchar(100) | Log source name |
| Server_ip | varchar(100) | Server IP |
| Parth | varchar(255) | Path name |
| File _pattern | varchar(255) | File regular expressions |
| Record_start_tag | varchar(255) | log start tag |
| Record_filter _keyword | varchar(255) | Log filter keyword |
| Status | int | Monitoring status |

Table 1. Log source datasheet design

### 4.2 Statistical Index Datasheet Design

The statistical index datasheet is designed for statistical purposes to understand the system's real responding situation and operation situation for some time. The datasheet includes specific user behaviours and system’s response indexes, such as the request URL, source IP system’s response status, system’s response time, and returned data. Through aggregated analysis of the logs, two parts of the platform can be understood: the first is the service quality level, for example, whether the response time is normal or the response status is true. And the second is the user level, some statistics and recommendation can be mined according to the maps where users visits mostly.

Because the statistical index datasheets are different according to different services, an example of Map service (the most-visited service) is shown in Table 2 (only main indexes).

| Indexes | Types | Description |
|---|---|---|
| nid | int | Unique ID |
| log_time | unix time | Record time |
| log_ip | varchar(100) | Request IP |
| layer | varchar(100) | Map layer |
| projection | varchar(100) | Map projection |
| x | int | Row |
| y | int | Column |
| z | int | Level |
| latitude | float | Latitude |
| longitude | float | Longitude |
| Code | int | Administrative code |
| response_time | unix time | Response time |
| response_status | int | Response status |

Table 2. Statistical index datasheet design for map service

## 5. SYSTEM IMPLEMENTATION

### 5.1 Functional Design

The main functions of design and realization consist of four big parts in the system: Statistics, Comparison Analysis, Export and Visualization, as shown in Figure 3.

**5.1.1 Statistics**: The function module include statistics on PV/UV, interface requests, response status, average response error rate, average response time, key words and API requests. The operator can get an overall sense of the platform how it is going on at any time based on the statistics.

**5.1.2 Comparison Analysis**: It contains the functions of regional user comparison, access arrival rate comparison, service time comparison, word cloud comparison and layers comparison. The comparison analysis is mainly based on administrative regions and time. Comparison analysis is the key to make decisions about data update plan.

**5.1.3 Export:** The statistics and analysis results of a specified duration can be exported in the format of Excel.

**5.1.4 Visualization:** Two kinds of user interfaces are designed in the system with the same data interfaces, one is a big scene website for display the main running indexes of the platform, and the other includes the whole functions of log analysis for the operator.

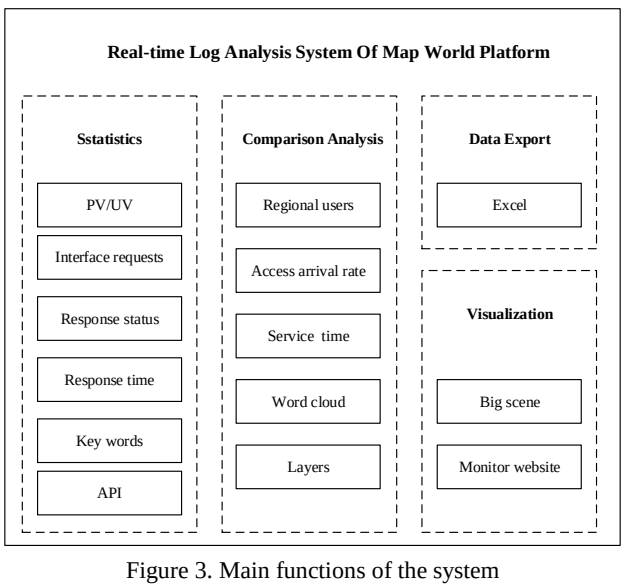


Figure 3. Main functions of the system

### 5.2 System Development

Take the above key technologies and designs as the foundation, the system is developed based on the model with B/S architecture to ensure the needs of system integration, stability and easy maintainability, and the server-side development language is Java, while the client language is JavaScript. The detailed development environment is listed in table 3.

| Software | Version | Function |
|---|---|---|
| Eclipse | 4.5 | Development platform |
| Jdk | 1.8.0_151 | Development tools |
| Nginx | 1.1.3 | Proxy server |
| Flume | 1.8.0 | Log collection |
| Kafka | 0.11.0.1 | Messaging system |
| Zookeeper | 3.4.9 | Coordination service |
| Storm | 1.1.0 | Flow data analysis |
| Redis | 4.0.0 | Hot log data storage |
| HBase | 1.2.5 | Log data storage |
| MySQL | 5.7 | Basic confirmation |
| Echarts | 3.0 | Data visualization |
| Boostrap | 3.0 | The front frame |

Table 2. Development environment

### 5.3 System Application

The system was started in 2014 and deployed in June, 2017 in the Map World platform successfully, and now it has become one of the most important operation tool for the platform. As shown in Figure 4 and 5, the overall running situation can be seen on the big scene conveniently for the leaders, and the operator can also query the historical log indexes to analyse the problem in time when necessary. The amount of log data processed in the system is more than 200 million per day. The system has high performance, high availability and high reliability.

More important, according to the analysis results that where users viewed most, a new update program was made including fast update area, normal update area, and annual update area. The new update plan improved the availability and quality of the map data, leading a significant increase in traffic.

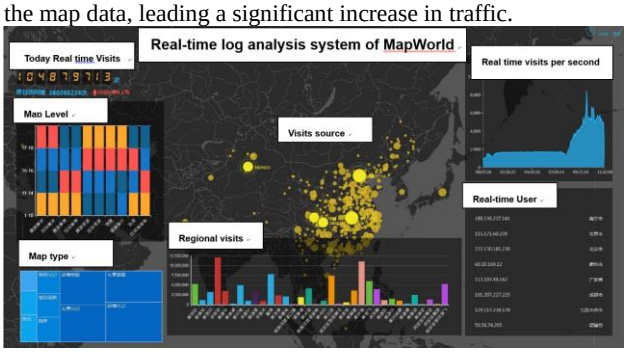


Figure 4. Big scene of the system

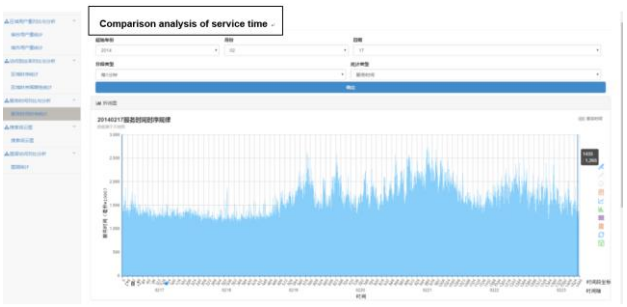


Figure 5. Monitor website of the system

## 6. CONCLUSION

According to the characteristics of many products, large amount of log data, real-time analysis requirements and multi-dimensional data in the Map World platform, a real-time log analysis system is designed and development based on the mature open source framework technology, including a set of solutions and software systems, such as real-time data collection,

data cleaning, real-time streaming calculation, real-time query and analysis, massive elastic storage, offline multi-dimensional analysis, and data monitoring and display. The system has been successfully applied to the operation and management, product decisions, and asset allocation improvement. Results of application indicates that it has achieved all its prospective constructive goals.

By making full use of data visualization method, the system realized the presentation of all the platform running indexes more intuitively, and the builder can understand the requirements for geospecial information service more directly and faster, and the top-level design of national Common geospecial information Services can be made more reasonable to meet the actual demands of the users. It can also provide important references of online server load prediction, resource allocation optimization, Internet security improvement and operation funs evaluation.

## ACKNOWLEDGEMENTS

Financial supports from the National Key R&D Program of China (No.2017YFB0503700) are gratefully acknowledged.

I also would like to express my thanks to Prof. Li Rui (Wuhan university), who provided many log analysis models to the system, which made the analysis method richer and more intelligent.

## REFERENCES

Apache Flume[OL]. http://flume.apache.org/FlumeUserGuide.html

Apache Kafka[OL]. http://kafka.apache.org/

Apache Storm[OL]. http://storm.apache.org/

Bai Y., 2013. Cloud computing environment large-scale data processing research. *Computer engineering & Software*, pp. 128-129

Chen F., 2016. Design and implementation of distributed log collection and analysis system based on Flume. *Computer engineering & Software*, Vol. 37, No. 12, pp.82-88

Chen Q., & Zhou L.,2012. HBase-based storage system for large-scale data in wireless sensor network. *Journal of Computer Applications*, Vol 32, No.7, pp. 1920-1923,1977

Cheng M. & Chen H., 2011. Weblog Mining Based on Hadoop. *COMPUTER ENGINEERING*, Vol. 11, pp.37-39

ECharts[OL]. http://echarts.baidu.com/

Jiang J., Wu H.& Huang W., 2017. Key Techniques and Project Practice for Establishing National Geo-information Service Platform "Tianditu". *Acta Geodaetica et Cartographica Sinica*, Vol. 10, pp. 1665-1671

Liu F., 2017.Design and implementation of real-time stream processing system based on massive network log data. *Beijing University of Posts and Telecommunications*, Master

Li Y. &Lv J., 2017. Research on log data real-time processing system based on Hadoop and Storm. *Journal of Southwest China Normal University (Natural Science Edition)*, Vol. 42, No. 4, pp.119-126

Qu G., 2016. The design and implementation of log analysis based on Storm. *Nanjing University*, Master

Wu H., Li R., Zhou Z., Jiang J., & Gui Z.,2015. Research and Prediction on Time-Sequence Characteristics of Group-User Access Behavior in Public Map Service. *Geomatics and Information Science of Wuhan University*, Vol. 40, No. 10, pp. 1279-1286,1316