Design and Implementation of Smart City Big Data Processing Platform Based On Distributed Architecture

Shuangmei Ma
School of Economics and Management
Communication University of China
Beijing, China
shuangmeima@126.com

Zhengli Liang
School of Economics and Management
Communication University of China
Beijing, China
liangzhengli@yeah.net

***Abstract*—For the problem about the low shared efficiency and the poor integrating degree of the massive multivariate data, in the paper, we propose a design scheme based on distributed architecture, mainly using Kafka, Storm and Spark clusters, and the real-time transmission among the structured data, unstructured data and database data are implemented. And the real-time analysis system of real-time data and statistical analysis system of offline data are developed. Finally, the test results show the performance of Big Data processing platform, and demonstrate fully the advantage of the distributed architecture.**

Keywords- Smart City, distributed architecture, Big Data processing platform

I. INTRODUCTION

With the development of the Internet of Things, Cloud Computing, Mobile Web, Big Data and other related technologies, the construction of Smart City in developed countries is in full swing, and the construction of Smart City in China also is developing rapidly. The Smart City is a new kind of urban form, which achieve the thorough perception of urban living environment, the comprehensive regulation of urban resources, the coordination of all parts of city, the convenient operation of the urban aspects, harmony between cities and human by means of dynamic monitoring, analyzing, integrating and utilizing all parts of urban data based on the new generational information technology.

The type and amount of urban data increase sharply, and the source of the data is extremely dispersible. In order interconnect these disparate data together, and achieve the sharing and integration of data, and improve the utilization of the urban data, it is necessary to build Smart City Big Data processing platform.

The current construction of the Smart City Big Data processing platform is relatively backward compared with the development of the urban infrastructure. The problems, such as, traffic congestion, the inefficient management of urban, the faultiness of environment monitoring system and emergency system, and so on, are all due to low utilization of urban data to a certain degree, therefore setting up an efficient Smart City Big Data processing platform is imperative.

Throughout the current research status about Smart City Information Platform, there are less theory researches in related areas in China, and most researches are concentrated in news reports and meetings, in addition, the deep systematic researches are less. The focus of reports and meetings is only on the blueprint of Smart City, lacking of the researches about technical system of the Smart City. In practice, the areas where economic and modern information technology are relatively developed, have begun construct the intelligent industry and information platform with the conceptual model of Smart City. However, the problems of the model of platform, the application of technology system, restricting access, the data transmission between terminals and the analysis of Big Data are discussed rarely. Overall, the construction of information platform is still in primary stage.

II. THE OUTLINE OF SMART CITY BIG DATA PROCESSING PLATFORM

A. *The type and characteristics of urban data*

During the construction of Informational City and Digital City, by the informational infrastructure providing information services, the city has accumulated vast amounts of dynamic data. The common types of Smart City's data are follows:

Map and POI data, GPS data, traffic data, video surveillance data, environment data, social activity data, and so on.

Thus we conclude characteristics of urban data:

The huge amount of data, the multidimensional in time-space, the multi-scale and multi-granularity, multiple and heterogeneous.

B. *Key issues of Smart City Big Data processing platform*

According to the characteristics of urban data and actual needs of management in Smart City, we can sum up key issues of the Smart City Big Data processing platform that are as follows:

- Transmission bottleneck of huge amounts of data
  Smart City Big Data processing platform can provide rapid feedback for every access request from each terminal, basing on the handling capacity of the server database. However, the rapid growth

978-1-4673-9323-2/15 $31.00 © 2015 IEEE
DOI 10.1109/ISKE.2015.7

of data scale and the expansion of the number of users put forward higher requirements for the Smart City Big Data processing platform. How to improve the transmission efficiency of massive data under the premise of ensuring the safety of data and transmission properly has been the key to the success of Smart City Big Data processing platform.

- Security of the transmission
  Certificate Authority:
  In order to ensure the security of urban data, we need to classify the departments and users and to conduct allocation of the functions, making sure that each user can obtain the data needed in his work and the data do not be leaked and abused and so on.
  Encryption:
  In order to prevent the disclosure of the urban data in transmission, we should conduct encryption. Sending data terminal should encrypt the data before which is transmitted, and the receiving data terminal should decrypt the data after which are received.
- The analysis of massive data
  In addition to achieving the data safe, efficient transmission, analyzing the huge amount of data implementing the data of intrinsic value has an important meaning to the urban manager's decision support system (DSS). Implementation of massive city data statistical analysis can help city managers to solve the traffic congestion, inefficiency of urban management, the faultiness of environment monitoring system and emergency system, and so on. Therefore, completing the analytical function of the big data processing platform, and raising the utilization of urban data better, can promote the development of smart city and improve the citizens' daily life.

III. THE KEY TECHNOLOGIES

This paper compares the mass, multiple and heterogeneous urban data which is constantly produced to the log data of IT companies, thus we can looking for solutions of constructing Smart City Big Data processing platform from the transmitting and analyzing system of the log data in IT companies.

A. The comparison of log transmission systems

First of all, there is a comparison of the excellent log transmission systems, as shown in TABLE 1

Overall, the design of Scribre is simple and easily to be used, but the fault tolerance and load balance is nevertheless unsatisfactory. Chukwa belonging to the Hadoop series, has pretty good scalability, however, it is not good in the real-time and load balance. Flume is designed ingeniously, having good scalability, and is also pretty in the fault tolerance and load balancing, but is not as good as Kafka in the real-time.

The former three systems adopt push/pull in architecture that the log producers push data to the controllers, then

TABLE 1 The comparison of log transmission systems

	Scribe	***Chukwa***	***Flume***	***Kafka***
Company	Facebook	Apache	Cloudera	Linkedin
Open time	2008/10	2009/11	2009/7	2010/12
Language	C/C++	Java	Java	Scala
Architecture	Push/push	Push/push	Push/push	Push/pull
Load Balance	--	--	Zookeeper	Zookeeper
Real-time	ordinary	ordinary	ordinary	excellent
Scalability	good	good	good	good

controllers push data to log consumers. There is delay in both two push. This designing concept matches offline massive data processing, but is not good at real-time.

Kafka is designed very ingeniously, bringing the push/pull of information system into its architecture, namely the log producers push data to the controllers but the log consumers pull data from controllers, thus achieving real-time consumption.

In summary, the Kafka that behaves pretty well in all aspects is used, as the log transmission system of Big Data processing platform in this paper.

B. Kafka

Kafka is a news subscription system, and is used as the basement of processing pipeline of activity stream and operational data.

Kafka is written by Scale, using many optimization mechanisms to raise efficiency, and its whole architecture is more new (push/pull), and is more suitable to the heterogeneous cluster, many different kinds of company use it as multiple types of data pipeline because of all these excellent characteristics.

As shown in Figure1, Kafka is a news subscription system essentially. Producer pushes data to the broker and consumer pulls data from broker. The transmission of data is organized by the topic. The producer sends information to one topic, while the corresponding consumer subscribe information of this topic. The broker will send all consumers who have subscribed topic at the same time as the topic has something new. Every topic can be divided into many partitions, thus it's easy to manage the data and balance the load, and meanwhile, zookeeper can also balance the load. Zookeeper is responsible for keeping the metadata of producer, consumer, broker, topic and partition, and matching them, while broker is responsible for transforming the data.

The advantages of Kafka:

- The cost of data access on the disk is O(1)
- High throughput
- A distributed architecture, which can fragment the message
- Support to load the data into the Hadoop in parallel
- Load balance and fault tolerance

C. Hadoop and Spark

Hadoop is distributed system architecture developed by Apache, and realizes Hadoop Distributed File System (HDFS). HDFS has high fault tolerance, generally deployed on inexpensive hardware, and can provide high throughput capacity to access the data of program, suiting to applied

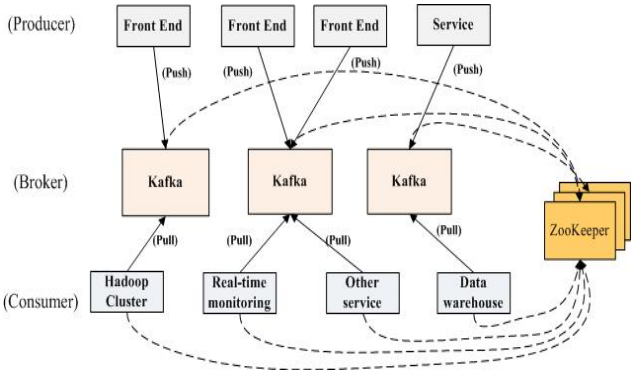

Figure 1 The architecture of Kafka

program with large data sets. HDFS can access the data in the file system by stream.

The core of the Hadoop is HDFS and MapReduce. HDFS provides storage for the vast amounts of data, and MapReduce provides computing for the vast amounts of data. The advantages of Hadoop are: reliability, scalability, efficient, fault tolerance, low cost.

Spark is a complete ecosystem of Big Data processing, except the HDFS of Hadoop used in the storage system on the basement, and it can replace Hadoop on other aspects, having functions more powerful than Hadoop, such as:

- The intermediate data of Spark is placed into memory, so Spark can calculate iteration more efficiently.
- Spark is more common than Hadoop. Spark offers a wide variety of dataset operations, providing convenient to the upper application, and its programming model is more flexible than Hadoop.
- Spark is associated with RDD by lineage, using minimal performance cost to ensure the robustness of the data.

D. Mllib

Mllib is algorithm library about Big Data machine learning based on Spark, concluding basic algorithm, such as classification, regression, clustering and collaborative filtering, and its advantages are stable and efficient.

Platform uses the storage capabilities of HDFS of Hadoop and the computational logic of Spark, and excavates the value of the massive data in Smart City by Mllib.

E. Storm

Storm is a distributed, fault-tolerant real-time computing systems, providing a set of common primitives for distributed real-time computing, and it can be used in "Stream processing", real-time processing messages and updating the database. It can also write and expand the complex real-time calculations in a computer cluster conveniently.

The advantages of Storm are as follows:

- Storm is a Simple programming model. It reduces the complexity of the real-time processing, just like the MapReduce reducing the complexity of the batch.
- A variety of programming languages can be used.

In addition, fault tolerance, scalability, reliability, fast are also Storm's superiority.

IV. Design and Implementation Of Smart City Big Data Processing Platform

A. Demand

Big Data Processing Platform integrates authentication and authorization, data exchange and data analysis, it not only can achieve accurate and efficient transmission of massive data, but also can help manager mine the value of log data. The demands of the platform are:

- Construct a data monitoring and managing system. This system is primarily responsible for monitoring the operating status of authentication and authorization, and the transmission system of log data and data analysis system. It can assign permissions to log data, then different departments and different users have different permissions.
- Construct a distributed log system that can achieve real-time transmission of data. And it should meet the following requirements: high concurrency, high load, scalable, and fault tolerance.
- Construct a data analysis system which integrates real-time and offline computing clusters. This system executes the real-time analysis of the real-time data by Storm and executes offline analysis of the offline data by Spark, and in both pathways, users can self-control the data filters and statistical analysis strategies to create the statistical analysis tasks.

B. Technical specifications and requirements of the platform

- Simple interface: Messaging middleware system supporting the processing platform should have simple interface, so that users can access to the platform conveniently and can easily conduct the implementation, maintenance and expansion of the system.
- The system is of high performance and reliability, and can quickly meet the requirements of massive data transmission, making sure that users can receive data correctly and the data will not be stolen and tampered.
- Support the definable data standards and data exchange standards.
- Support the new users joining the platform, and the joining of the new users should not affect the structure and the normal operation of the platform.
- Provide the correct results for the analysis request of massive data quickly.

C. The functions of the platform

The functions of the platform are: authentication and authorization, data exchange, data analysis and data mining.

- Authentication and authorization
  Different departments and different users have different permissions, and these information are

Stored in the corresponding database and can be maintained and modified.

- Data exchange
  Big Data processing platform can conduct unified construction and management for the services of data exchange, providing services for the trans-department or the cross- regional exchange of information.
- Data analysis
  The analysis function of the platform achieves the real-time analysis of the real-time data and statistical analysis of the offline data. The users can create their own data analysis task, then the platform import the relevant data to existing storage and run in the computing cluster according to the users' data filters and statistical analysis strategies, and finally return the results to the users.
- Data mining
  The platform can carry out the data mining, and extracting the value of massive urban data, and can improve the decision-making ability of the Smart City. For example, mining the portrait of every citizen or each uptown, the model of vehicle's trajectory, prediction of vehicle flowrate and accident rate, and so on, all of this can provide reliable support for the city manager.

D. *The process of platform*

This platform has the advantages of reliable, scalable, high fault tolerance and high efficiency. The technologies applied are Flume, Kafka, Storm, Spark, Mllib, MySQL, mongodb, and so on.

The overall architecture of the platform is shown in Figure 2 and Figure 3.

This platform achieves data collection system, data transmission system, real-time analysis system, offline analysis system and data mining system (there is no introduction about authentication and authorization any more here).

Data collection system: the data comes from the department of different functions, for example, consumption data, monitoring data of traffic, and so on. We collect data from different server by Flume. The platform can achieve collect data automatically only by deploying the directory of the data.

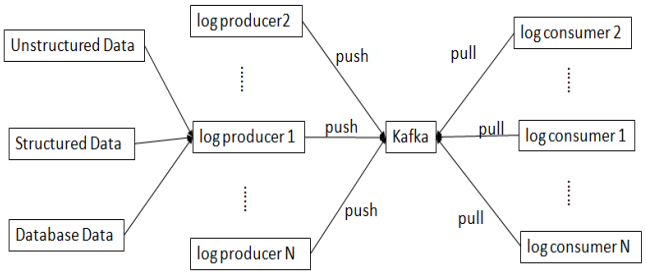

Figure 2 Architecture of the platform---data exchange

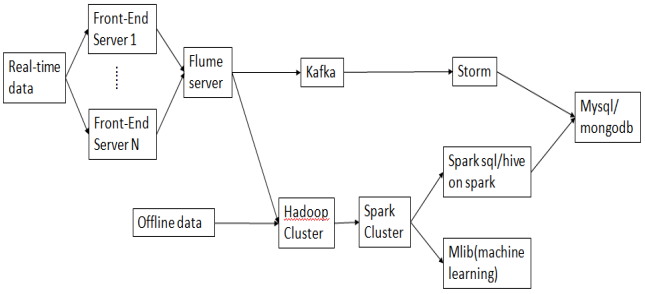

Figure 3 Architecture of the platform----data analysis

Data transmission system: the log producers and log consumers ------as the users of the platform only tell the designer of the broker of the Kafka about the directory of the data exchanged. When the platform is normally operating, the data will exchange in a particular time span automatically. For example, the information (Surveillance video of a corner) of the transport Sector is going to be shared with the police, the transport Sector only needs to tell the broker about the path of the directory of the video, the share is completed, then the same video that will be generated continuously will appear in the appropriate folder of the police in real-time.

Real-time analysis system: this platform achieves the real-time analysis of the real-time data and the statistical analysis of the offline data. The data of the example above can carry on real-time analysis through Storm, and the these analysis results not only can help to ease traffic congestion, but also can play a greater role in terms of social security with the technology of face recognition.

Offline analysis system: this system can provide analysis results of more dimensions and longer period, such as computing traffic situation under different dimensions of time, location and weather.

Data mining system: After the data are collected into Hadoop cluster, we can use Mllib to conduct analysis and big data mining, for example, providing portraits for population and residential areas by clustering, judging the categories of things by classification, and predicting abnormal users' behavior, and so on.

E. *Performance test and results*

1) *Performance test of the data transmission system*

a) *Performance testing of the log producer*

Test method:

It is set that the producer read the same data, and record the test results will be recorded according to the increase number of the producer progressively.

Test results:

As show in Figure 4, the abscissa is the number of producer, and the ordinate is the production of log data (the unit is MB/Sec). We can conclude that the production of the log data has been stable at around 40MB/Sec with the increasing number of producer.

Analysis and conclusions:

Broker meets the demand of high concurrency and high load by with the zookeeper, and the broker's performance is not affected by increasing of the concurrent number of the

producer. Besides, we can improve the performance of broker by increasing the number of servers.

b) Performance testing of the log consumer

Test method:

Create a consumer cluster, and gradually increase the number of consumer in the experiment.

Test results:

As show in Figure 5, the abscissa is the number of consumer, and the ordinate is the consumption of log data (the unit is MB/Sec). We can conclude that the consumption of the log data increases at the first and stables at a certain value later.

Analysis and conclusions:

The result also show that broker meets the demand of high concurrency and high load. In the beginning, the consumption of the log data go up with the increases of consumers, because consume data are in parallel in a same group. While the consume reaches a certain number of the topic's partition, the consumption is stable at a certain value.

2) Performance test of the data analysis system

This part mainly compares the Spark cluster with Hadoop cluster algorithm, which nowadays are used widely,the details are as follows:

Experimental scenario: the logistic regression algorithm is used, to judge whether patients have relative diseases.

Experimental data: data size is 50 G, the proportion of positive to negative samples is 1:6; Cluster for 10 nodes, one master, nine slave.

Server Configuration is 12core CPU, 192G RAM.

Hadoop edition is cloudera Hadoop 5.2.3, and spark is 1.2.0.

The data are tested by the logistic regression in the mahout and mllib.

The results of the execution are shown in Figure 6

Conclusion: the Spark's ability processing big data is more superior than the Hadoop's.

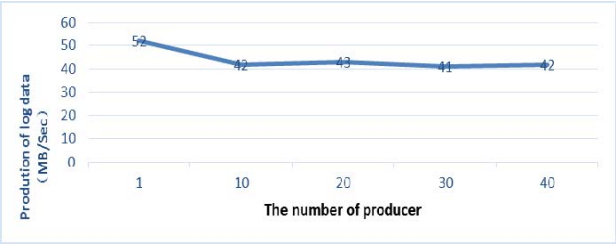

Figure 4 Test results of the producer

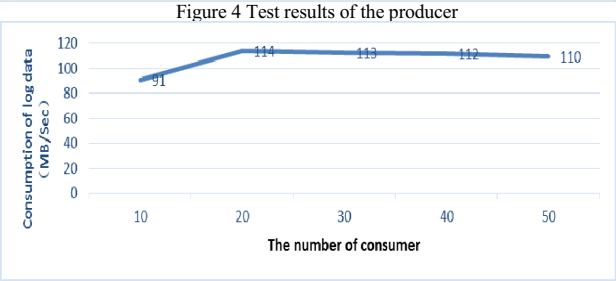

Figure 5 Test results of the consumer

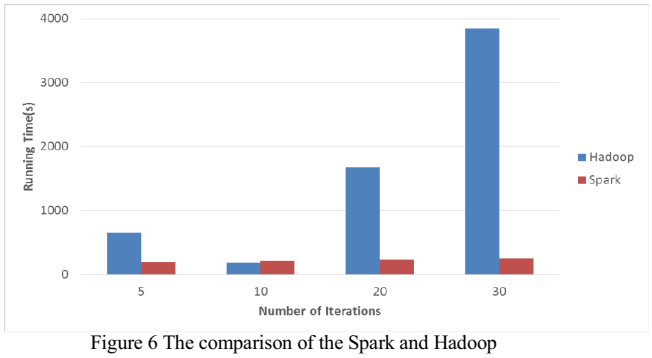

Figure 6 The comparison of the Spark and Hadoop

V. SUMMARY AND OUTLOOK

This Big Data processing platform can easily realize the structured and unstructured data transmission, and the users do not need any operation, log data consumer synchronizing the data automatically with the producer, and achieving the real-time analysis of the real-time data and statistical analysis of the offline data. However, this platform is merely an attempt constructing the data processing platform of the big Smart City basing on the distributed architecture. The implemented functions, the scale of users, magnitude of transmission and the functions of the data analysis are still in its infancy.

To apply the platform to smart city, we need to expand the number of producer and consumer. Because of the architectures we used are distributed, don't worry about that the speed of transmission after expansion will become slow provided to expand the platform transversely.

In order to excavate the value of the urban data, we ought to increase the statistical analysis of the data. In the future we must write different software packages according to different analysis, so that, the users just call the corresponding package, and change some parameters, and then can conduct various statistical analysis of data.

REFERENCES

[1] Haiyan Zhang. The foundation and principle of Smart City maturity assessment system[J].The introduction of economic.2012(3):33-34

[2] Bin Wei.A Design and Implementation of Data Cloud Service Platform Based on Scalable Log Svstem[D].Hangzhou:Zhejiang University,2013

[3] http://kafka.apache.org/documentation.html#introduction

[4] http://baike.baidu.com/link?url=s671-_1BrswqCSNrsXw6irrmG0JaH94uolavHD5hPENl_UC5_A2MTsRDTZBxdOkJRyKsjqOxzab4vbvn6SWeO_

[5] Zhengli Liang,Xiaofeng Jia, The theory and practice of decision support system[M]Beijing: Tsinghua University Press,2014

[6] Zhizhou Wu, Xiaoguang Yang, Zhizhou Li. The Research of some Key Technologies on the Application of WebGIS in ITS[J].IEEE,2003

[7] http://www.e-gov.org.cn/xinxihua/news008/201106/119275.html

[8] P.J.G.Teunissen. Theory of Carrier Phase Ambiguity Resolution. WUJNS,Vol.8 No.2B,2003:471—484

[9] Yanhu Kong.

[10] Lian Liu.Construction and empirical research on the Smart City information service system[D].Changchun:Jilin University2012

[11] Jingyuan Wang.Data-centric research review of Smart City[J]. Computer Research and Development.2014,51(2)

[12] Guowei Zou, Janbo Cheng.The application of Big Data in Smart City[J].Telecom Network Technologies.2013(04)

[13] Yuli Gong.The design and implementation of Smart City information platform based on open-source framework[D].Wuxi:Jiangnan University,2013

[14] Yuanming Yuan.Key technologies of Smart City information system[D].Wuhan:Wuhan University,2012

[15] Jay Kreps, Neha Narkhede, Jun Rao. Kafka: a Distributed Messaging System forLog Processing, NetDB, 2011

[16] Tang_Dali, Huang_Jixian. Structure and technology on Application WebGIS[J].IEEE,2001