Web-based Collaborative Big Data Analytics on Big Data as a Service Platform

Kyounghyun Park, Minh Chau Nguyen, Heesun Won

Big Data SW Research Department, Electronics and Telecommunication Research Institute, South Korea

hareton@etri.re.kr, chau@etri.re.kr, hswon@etri.re.kr

***Abstract*— As data has been increasing explosively due to development of social networks and cloud computing, there has been a new challenge for storing, processing, and analyzing a large volume of data. The traditional technologies do not become a proper solution to process big data so that a big data platform has begun to emerge. It is certain that big data platform helps users develop analysis service effectively. However, it still takes a long time to collect data, develop algorithms and analytics services.**

We present a collaborative big data analytics platform for big data as a service. Developers can collaborate with each other on the platform by sharing data, algorithms, and services. Therefore, this paper describes big data analytics platform that effectively supports to manage big data and develop analytics algorithms and services, collaborating with data owners, data scientists, and service developers on the Web. Finally, we introduce a CCTV metadata analytics service developed on the platform.

***Keywords*— BDaaS (Big Data as a Service), Big data analytics, Collaborative platform, Big data platform, CCTV MVS, CCTV video analysis**

I. INTRODUCTION

Today there has been an enormous data explosion due to the new trend and paradigm such as social networks and cloud computing. Moreover, data has been getting more diverse, more complex, and less structured and it also needs to be processed rapidly. This situation has caused a new challenge for the traditional technologies such as relational databases and scale-up infrastructures[1][2].

Literally, big data is a large volume of data so that it is difficult to collect, store, and analyze. However, big data is not just about a large volume of data. It is a concept that provides an opportunity to find useful knowledge into existing data [3].

In this context, big data is characterized by 3Vs: Volume, Velocity, and Variety[4]. First of all, there has been the exponential growth in the data storage as data format is diversifying. We can find various data formats such as video, music, and large image files on the social network service channels. It is common to have the storage system of which size is Terabytes or Petabytes.

Secondly, velocity is often identified with real-time analytics. However, velocity is also about the rate of changes. There was a time when data had been hardly changed or the rate of changes is very low. Yet, today people reply on SNS (Social Network Service) to update their messages. Sometimes, a few seconds old messages is discarded or updated because users are not interested in any more. They often remove old messages and pay attention to recent updates. This high velocity data illustrates big data.

Lastly, data can be stored in multiple formats. For example, the data would be stored in a simple text file or stored in the form of video, SMS, pdf or user defined format. In this case, we should consider how to process a variety formats of data. This variety of data describes big data.

On the other hand, big data also refers to massively parallel and distributed processing architectures in the technology perspective of data processing.

Big data analytics platforms enable users to collect, organize, and analyze large sets of data to discover patterns and other useful information. Big data platforms focus on processing large data and do not support collaboration among users so that it takes enormous time for users to develop services including data collection, data pre-processing, data analysis, and algorithm development.

To support more efficient service development environment, we introduce a new collaborative big data analytics platform that helps users concentrate on developing their own services efficiently and rapidly by sharing data, algorithms, and services among them.

This paper is organized as follows. In the next section, we introduce Big Data as a Service (BDaaS) and show the collaborative big data analytics platform for BDaaS. In section 3, we describe CCTV metadata analytics service implemented on the platform. The final section contains concluding remarks and future work.

II. BIG DATA AS A SERVICE PLATFORM

A. Collaborative Analytics Platform for BDaaS

There are plenty of use cases using big data analytics such as after-sales service, searching for missing people, smart traffic control system, customer behavior analytics, and crisis management system. In order to provide these analytics services, we traditionally conduct many steps separately such as data collection, data pre-processing, information extraction, and visualization. Therefore, developing separate systems to analyze big data is getting to require high cost and professional knowledge on big data technologies.

Big Data as a Service is the delivery of statistical analysis tools or information by an outside provider that helps organizations understand and use insights gained from large data sets in order to gain a competitive advantages.

Big Data as a Service typically consists of three layers, Big Data Infrastructure as a Service, Big Data Platform as a Service, and Big Data Analytics Software as a Service[5].

The purpose of our platform is to provide Big Data as a Service and to enable users to develop cloud service more efficiently. Especially, we have designed collaborative analytics platform with focusing on Big Data Analytics Software as a Service in BDaaS layers

Figure 1 shows the general concept of collaborative big data platform for Big Data as a Service. For these services, the platform provides various web-based service environments and enables resource sharing.

For data owners, the platform offers data management environment. It also provides data scientists and service developers with algorithm and service development environment.

The platform supports role-based access control on data owner, data scientist, service developer, and platform manager. Data owners access to the platform to post their data information and share their own data. Data owners collect various data and register information of data via the web portal.

Data scientists develop and optimize analytics algorithms on the platform. Data scientists can explore and request data registered by data owners and measure the performance of algorithm. Finally, service developers implement analytics services using available components in the workflow-based tool. Service developers can explore and request workflow components registered by other data scientists.

The platform supports collaborative environments to users so that they can reuse other's data and algorithms and concentrate on their own work such as developing algorithms or services.

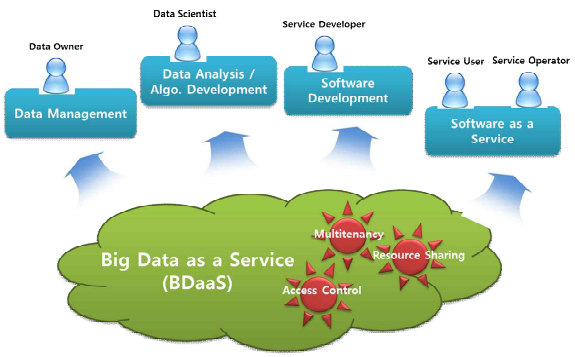

Figure 1. Collaborative Big Data Platform Concept for Big Data as a Service

To support collaboration on the platform, the platform provides users with the multi-tenant environment. Multi-tenancy refers to an architecture in which multiple users or tenants are served in a single instance with the guarantee of isolation and on the other hand, they can share computing resources with each other.

For multi-tenancy in the platform, we have extended access control management and resource management in hadoop. We have integrated kerberos into the platform for access control and extended YARN for efficient resource management. Figure 2 shows the platform architecture for multi-tenancy. Especially, we focused on metaDB synchronization and multi-tenant scheduler for performance improvement.

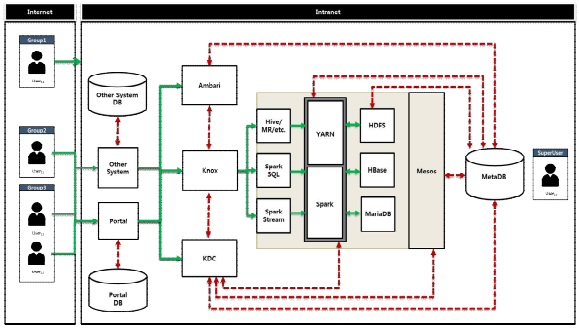

Figure 2. Platform Architecture for Multitenancy

B. Web Service Portal

One of the web user interfaces for the BDaaS platform is the web service portal illustrated in Figure 3. The main purpose of the web service portal is to facilitate the users' communication and information sharing on the platform for efficient and rapid service development. Therefore, the web service portal provides a web board on which all users can share their information. They can explore data, algorithms, and services using catalogue services and post their requirements on the board. This collaboration among users is possible due to the multi-tenancy architecture.

Figure 3. Web Service Portal Pages

Web service portal provides each user with different web pages according to its role as follows.

- Data owners access data list page, data registration page, data modification page, data catalogue page and data monitoring page to manage their own data. Data catalogue enables users to search data easily and rapidly by categorizing data and managing metadata.
- Web service portal has a link for analytics portal so data scientists can access a sandbox page which is running in

the analytics portal to develop algorithms. The sandbox page provides a text-based editor tool so developers can implement algorithms using editor tool and register them. They also access algorithm list page, algorithm catalogue page, algorithm registration/modification page.

- Service developers access an IDE tool page that is another sandbox page via web service portal. The IDE tool page supplies two kinds of development environments, named text-based and workflow-based tools. They also access service list page, service registration/modification page, service catalogue page, and service monitoring page.

C. *Analytics Portal*

Many big data platforms that mostly consist of various open software do not give useful development tools for users.

Analytics portal is a web-based development tool to improve development productivity under big data environment. Analytics portal is illustrated in Figure 4 and it is accessed via web service portal and supports various tools as follows.

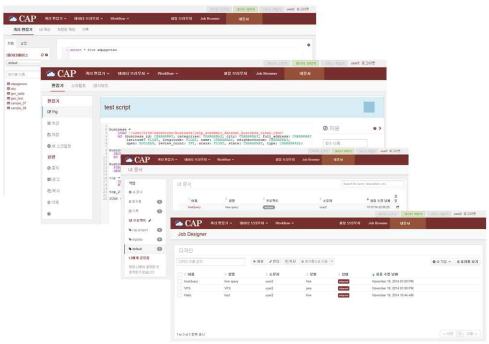

Figure 4. Analytics Portal Pages

- Query Editor: Analytics portal provides a web interface for hive queries and relational database queries. The hive query editor consists of query list, text window, and result window. The result window shows log and recent query history. The window also shows query results and reports them in the form of chart. Data scientists can test their hive queries and register them as the algorithms.
- Data Browser: Analytics portal offers two kinds of data browsers. One is for the metastore and the other is for HBase. Metastore keeps database schema information that is used in the platform cluster. A platform manager can manage metadata stored in the metastore using a data browser. In addition, developers can manage data stored in HBase and can be used as input or output data of each process through the data browser.
- Workflow Designer: Analytics portal provides users who have little experiences in developing services with a useful IDE tool, a web-based service modeling tool that enables users to develop services rapidly using algorithm component. Workflow consists of independent processes that are the unit of job. We can define the job as a process and monitor the status of the workflow using the workflow designer.
- Job Browser: Analytics portal supports a job browser for map-reduce job monitoring. Job browser enables users to check job information and the status of the jobs.
- Monitoring Tool: Analytics portal provides a monitoring tool to monitor cluster resources in the platform. The monitoring tool includes node monitoring that checks the cluster nodes, host monitoring that provides system information in the form of tables, and resource monitoring that checks the usage of cluster resources.

III. CCTV METADATA ANALYTICS SERVICE

For the demonstration on the platform, we have developed a prototype of CCTV metadata analytics service.

In the analytics perspective on CCTV, many useful use cases could be developed if we can handle a huge volume of CCTV video data.

The traditional CCTV VMS (Video Management System) provides real-time video monitoring and event detection by collecting video stream from a number of CCTV cameras, extracting the features from each video data, and analyzing them. Because the system focuses on real-time event detection, it extracts the features from the video in real-time and does not store them after processing. To analyze CCTV metadata, we modified CCTV VMS and stored the features in XML format in the HDFS (Hadoop Distributed File System). The Figure 5 illustrates advanced CCTV VMS architecture that we extended for CCTV metadata analytics. CCTV metadata includes information on moving objects and user defined events as illustrated in Figure 6.

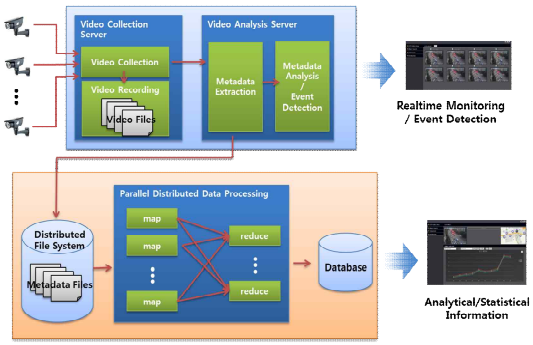

Figure 5. CCTV VMS Architecture for Metadata Analaysis

Figure 7 shows CCTV web service pages in which we can search the objects under the specific condition and extract statistical information from a large volume of CCTV video data. On the basis of this CCTV metadata analytic service, we expect to develop various CCTV related services such as searching for missing people and smart traffic control service.

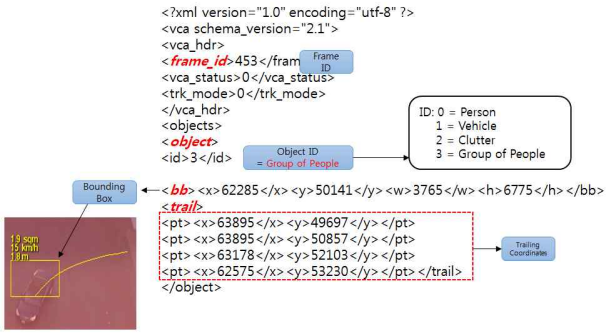

Figure 6. CCTV Metadata in XML

Figure 7. CCTV Web-based User Interface for CCTV Analysis

IV. CONCLUSIONS

In this paper, we introduced collaborative big data analytics platform. The big data platform provides two kinds of web portals: web service portal for collaboration and analytics portal for developing BDaaS services. We have enhanced access control and YARN for multi-tenancy so that the platform can support collaboration among actors.

Web service portal is a common web interface for communication. Analytics portal is linked to web service portal and supports various big data management and development tools. Finally, we demonstrated CCTV metadata analytics as an analytics service.

Currently, we have been extending the streaming processing system and plan to integrate it into the platform for real-time analytics service.

ACKNOWLEDGMENT

This research was financially supported by the Ministry of Trade, Industry and Energy(MOTIE) and Korea Institute for Advancement of Technology(KIAT) through the International Cooperative R&D program

REFERENCES

[1] M. A. Beyer and D. Laney, "The importance of 'big data': A definition," Gartner, Tech. Rep., 2012
[2] S. Lohr, "The age of big data, " New York Times, vol. 11, 2012
[3] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, "Interactions with big data analytics," Interactions, 2012
[4] J.Manyika, M. Chui, B. Brown, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, pp.15-17, 2011
[5] Z. Zheng, J. Zhu, and M. R. Lyu, "Service-generated big data and big data-as-A-service: An overview," pp. 403-410, 2013.
[6] S. Lohr, "The age of big data," New York Times, vol. 11, 2012
[7] "Challenges and opportunities with big data," leading researchers across the United States, Tech. Rep., 2011.
[8] E. Slack, "Storage infrastructures for big data workflows," Storage Switchland, LLC, Tech. Rep., 2012.
[9] C. Lynch, "Big data: How do your data grow?" Nature, vol. 455, no. 7209, pp. 28–29, 2008.
[10] "Big data-as-a-service: A market and technology perspective," EMC Solution Group, Tech. Rep., 2012.
[11] J. Horey, E. Begoli, R. Gunasekaran, S.-H. Lim, and J. Nutaro, "Big data platforms as a service: challenges and approach," in Proceedings of the 4th USENIX conference on Hot Topics in Cloud Ccomputing, ser. HotCloud'12, 2012, pp. 16–16.

Kyounghyun Park Kyoung Hyun Park is a senior research staff of the Electronics and Telecommunications Research Institute(ETRI). He received her M.S. degree in Computer Science from Chungbuk National University, South Korea, in 2001. He has been involved in many projects related to continuous speech recognition systems and database systems. His current research interests include big data management systems and cloud computing.

Minh Chau Nguyen Minh Chau Nguyen is a researcher of the Big Data Software Research Department, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. He received his BS degree in computer science from the University of Sciences, Ho Chi Minh, Vietnam, in 2009. He then went on to receive his MS degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Rep. of Korea, in 2013. His research interests include big data management, software architecture and distributed systems.

Heesun Won Heesun Won is a principal researcher of the Electronics and Telecommunications Research Institute(ETRI). She received her M.S. degree in Computer Science from KAIST, South Korea in 1992. Her current research interests include BDaaS (Big Data as a Service) and cloud computing platform.