International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 14938-14943

Design and Implement of Bigdata Analysis Systems

Jeong-Joon Kim

**Department of Computer Science & Engineering, Korea Polytechnic University, Gyeonggi-do Siheung-si 15073, Korea.*

Abstract

The development of networks and the Internet has resulted in large amounts of data on the Web. There was a new paradigm called big data to handle this. Big data uses unstructured data including existing structured data and social data such as SNS. It is creating new value by analyzing it in various ways. However, we can not blindly trust social data created by human subjective thinking. Therefore, in this paper, we design and implement a system that can verify social data and a big data analysis system.

Keywords: Big Data, System Architecture, Trust System, Social Data, Analytics, Ranking System, Text Mining

INTRODUCTION

With the development of networks and the Internet, data has become popular throughout the network. With the spread of various media and convenient web, many data have been transferred to the web. With the increase in web data, searching companies has become more active in large volume data research.

In particular, in the mid-2000s, Google's Big Table[1] and Amazon's Dynamo[2] presented a paradigm for big data research by presenting modeling for parallel processing. On the other hand, many open source applications based on Apache's Hadoop[3] have been developed, and companies will also use the bigdata system, which incorporates a variety of software such as storage, search, analysis, and visualization in a large parallel processing system based on Hadoop. As a result, social network services such as Facebook, Twitter, and Instagram have developed, and these companies are creating new value through social network analysis.

However, social data can not be considered to have expertise and objectivity. Political, cultural, and moral, personal views are entered, often manipulating public opinion. On the other hand, quantitative aspects of social data can not be ignored. Therefore, if the expertise and objectivity are verified, it will be useful information. In order to judge the expertise and objectivity of a document, it is possible to quantify it through the history of the author and the human network relation.[4]

Therefore, in this paper, we designed and implemented a system that can provide better results by estimating a measure that can guarantee professionalism and objectivity and reflecting it on big data system.

RELATE WORKS

Hadoop ECO System

Hadoop is a software framework that allows the Apache Foundation to process large amounts of data, and various big data business tools are available through Hadoop subprojects. The Hadoop framework and the Hadoop subprojects help collecte, store, process, analyze and visualize the Hadoop ECO system[5]. The following Fig. 1. shows the structure of the Hadoop ECO system.

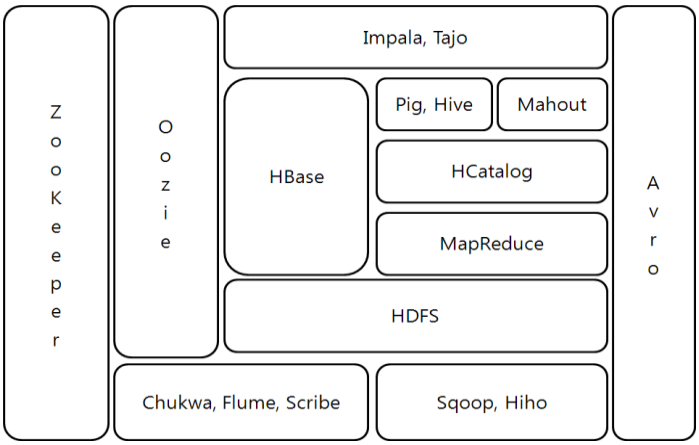

Figure 1: Hadoop ECO System.

ZooKeeper is a distributed administrator to ensure traffic load balancing and availability, and Oozie is a workflow manager that manages the work in Hadoop. HBase is a column based database based on HDFS, Pig and Hive are similar to SQL but control data stored in HDFS. Mahout provides a variety of analysis functions such as classification, recommendation, filtering, clustering, regression analysis, pattern mining, and so on. HCatalog is a data manager that allows data generated by specific modules to be used in other modules. Avro supports remote procedure calls and data serialization, and Chukwa and Flume allow stable storage of generated data in HDFS. Unlike Chukwa and Flume, Scribe transfers data to the main server instead of HDFS, and it can be stored in HDFS using JNI (Java Native Interface). Sqoop is responsible for transferring

data to various repositories such as HDFS, RDBMS, and DW, while Impala and Tajo are responsible for retrieving and storing data through SQL, such as RDBMS.

The Strasphere Platform for Big Data Analytics

The Stratosphere platform [6] is a very large scale analytical application and an extension platform for parallelization and optimization iterative programming. A platform that includes data warehousing, information extraction and integration, data organization, graph analysis and statistical analysis.

Fig. 2. is the architecture of the Stratosphere Platform and consists of repository, cloud platform, job manager, parallel program manager, and script manager from below.

As shown in Fig. 2. data warehouses including HDFS, Amazon S3, and other databases are connected to cloud platforms such as Amazon EC2 and Apache YARN, and the priority of parallel processing is determined using a job planner named Nephele. PACT (PArallelization ConTract) is a process for parallel processing and is based on MapReduce. Sopremo analyzes the data based on the requested script and derives the data through processes such as deduplication and data sorting.

The Stratosphere Platform creates a script that specifies the input and output values of a Meteor script and creates a specification of the data and results to be found and passes it to Sopremo.

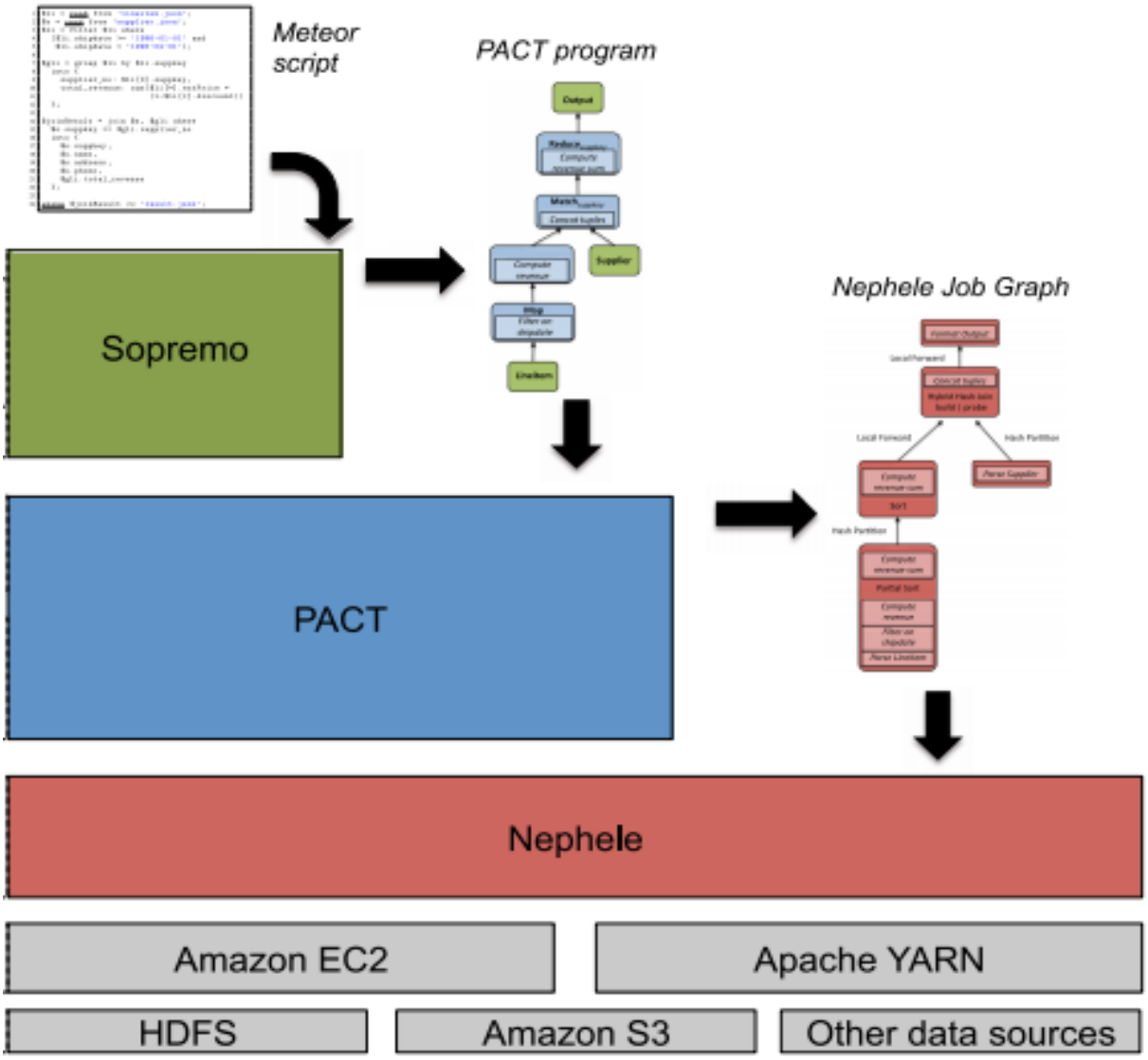

Figure 2: Stratosphere Platform System Architecture

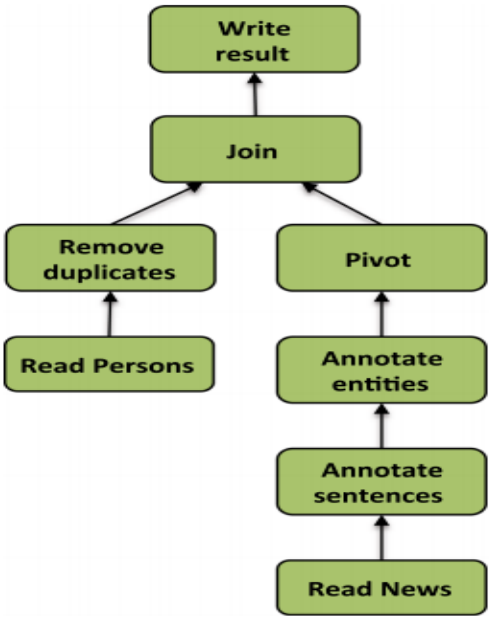

Figure 3: Sopremo

Sopermo in Fig. 3. requests data in accordance with the results requested by the Meteor script, extracts duplicates, and extracts the main content from the sentence.

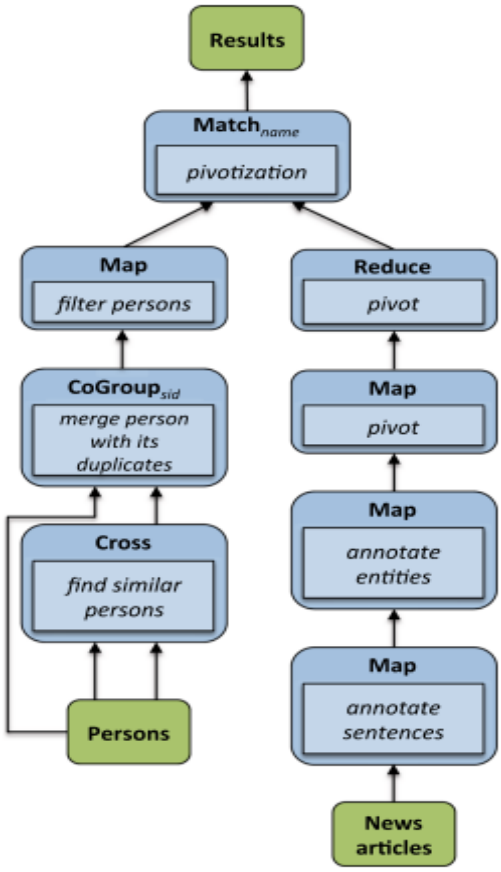

Figure 4: PACT

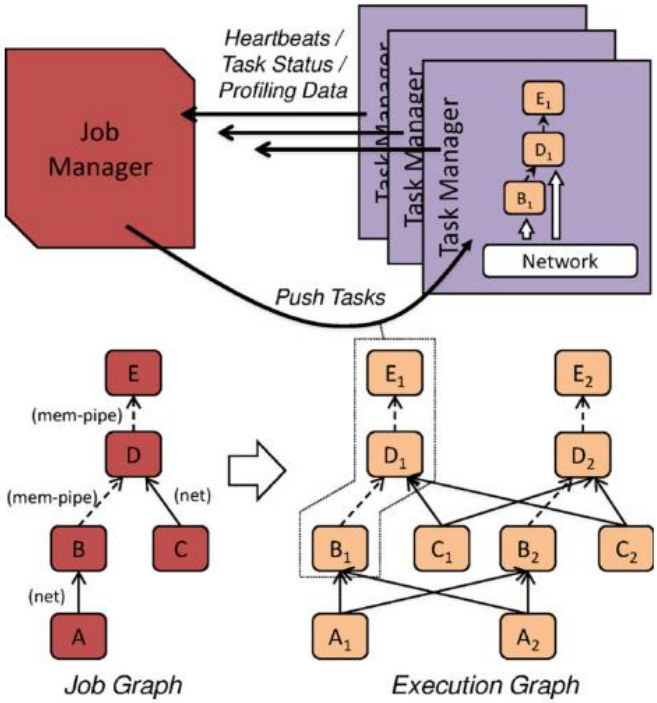

Figure 5: Nephele

PACT in Fig. 4. refines the necessary data through Map, Reduce step, and mutual group operation. Nephele, the job manager of Fig. 5. received the tasks to be executed from the task manager, changed them into a graph, created an execution graph based on the graph, and reflected the same on the task manager, thereby improving the efficiency of the parallel operation.

The Anatomy of a Large-Scale Hypertextual Web Search Engine

It is the basis of the ranking system of Google search engine[7,8,9,10]. The arguments in this paper are summarized as follows.

$$PR(A) = (1-d) + d\left(\frac{PR(T1)}{C(T1)} + \cdots \frac{PR(Tn)}{C(Tn)}\right), \quad (2.1)$$

PR is abbreviation of PageRank, and PR (A) is a page rank of a web page named 'A'. T1, T2, ... Tn refers to other pages that point to the page. And PR (T1) is the page rank value of the page T1. D is called 'Damping Factor', which means the probability that a visitor clicks on a link on another page. C (T1) means the total number of links that the page T1 has.

If the above equation is modified to fit the social data environment, the relevance can be used for the measurement. This is described in the implementation of Chapter 3.

DESIGN AND IMPLEMENTS TRUST BIGDATA SYSTEMS

System Architecture

The reliability large data system proposed in this paper has the system architecture as shown in Fig. 7.

As shown in Fig. 6, the Trust Big Data system System Architecture performs the big data analysis procedure from data collection to refinement, analysis and visualization.

Trust Big data system data collection can be seen in two major ways. Sample data for analytical purposes includes data based on logs generated by devices, data warehouses, and systems, as well as social networking data such as Twitter, Facebook, blogs, newspaper articles, and community writings to gain confidence in this data analysis. . In this process, logs of devices, data warehouses, and systems are not artificially generated data, so they can be viewed as trustworthy data, but social networks and social data are subjective information, so a filter is needed. Therefore, it is necessary to separately process the two types of data.

Trust Data Measurement

Trust points can be obtained by modifying the ranking system equation as described in Section 2.3. However, the documents created on the Web and the documents created on social data are not the same. The same can be cited, so we can use the equation in 2.3, but we do not cite much in actual social data. The same can be hyperlink, so we can use the equation in 2.3, but we do not hyperlink much in actual social data. Therefore, additional factors must be added human network map.

$$TP(A) = (1-d) + d\left(\frac{TP(T1)}{C(T1)} + \cdots \frac{TP(Tn)}{C(Tn)}\right), \qquad (3.1)$$

Equation 3.1 is shown in Equation 2.1. However, it can be applied to new environments by changing the corresponding elements.

D is the document that the author created mainly for the document of the query among the entire documents of the author, unlike the one used in the information retrieval with df (document frequency). TP(T1) is the TP value for the people who press the 'favorite' button, and C(T1) is the hyperlink contained in the body and comments. However, there is a contradiction in this equation. If there is no hyperlink, the denominator becomes zero, so the equation does not hold. Therefore, when C(T1) is 0, it is replaced with 1.

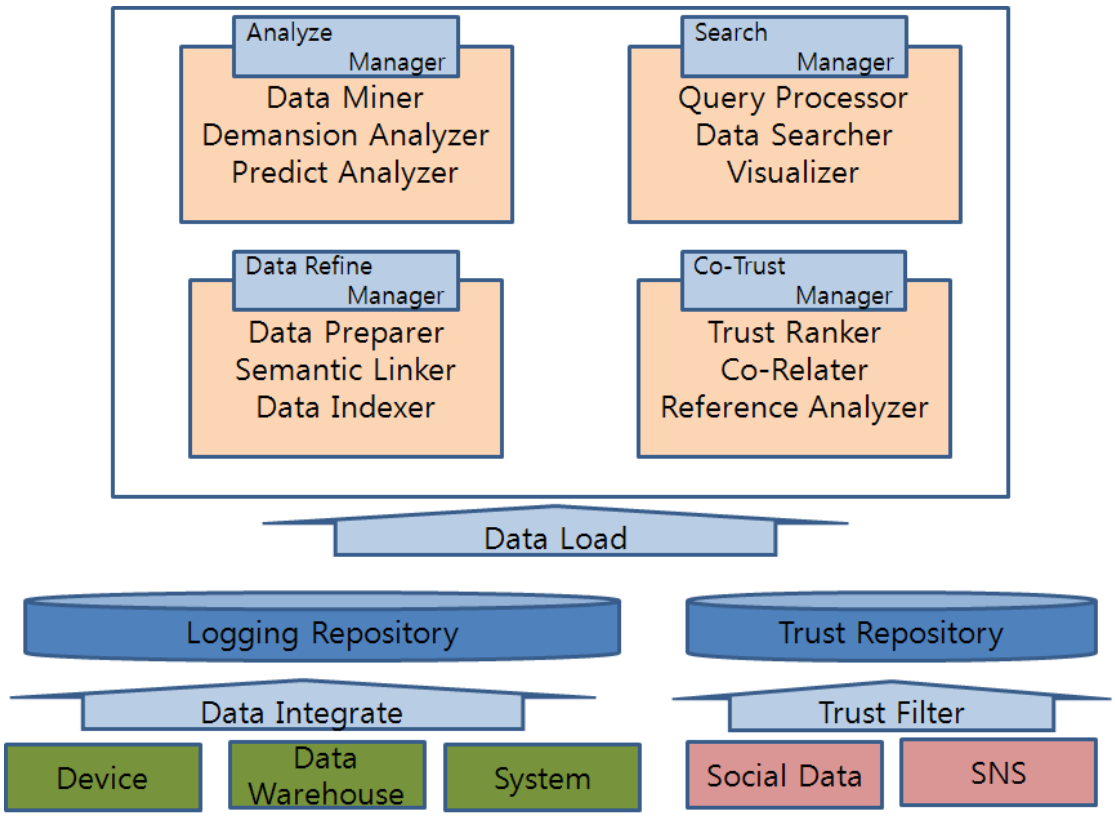

Figure 6: Trust BigData system System Architecture

Using this formula, the expertise of social data and the interest of surrounding network characters can also be evaluated, and the degree of influence can be evaluated.

SCENARIO AND TEST

Recent issues include political issues, Avian Influenza and earthquakes. Political issues can be a sensitive issue, creating a scenario for Avian Influenza.

First, we looked at the recent trend of the Korea Broiler Council to manage the prices of chicken for influencing and spreading power of avian influenza. Fig. 8. is a graph of recent prices.

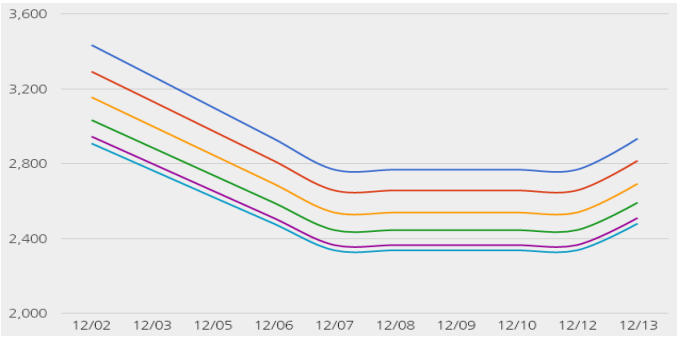

Figure 7: Prices of Before Avian Influenza(KR Won, Korea Broiler Council, 2016)

As shown in Fig. 8, the market began to plunge in December. This is because Avian Influenza has been reported and demand has decreased. Although not visible in the graph, the November data remained at a constant level.

From December 1 to December 13, 2016, we collected articles containing chicken and avian influenza in data collection conditions. We randomly collected 2,500 data including chicken and 2,500 data including avian influenza.

We evaluated these articles through Trust Document Measure and analyzed the results through analysis manager and search manager.

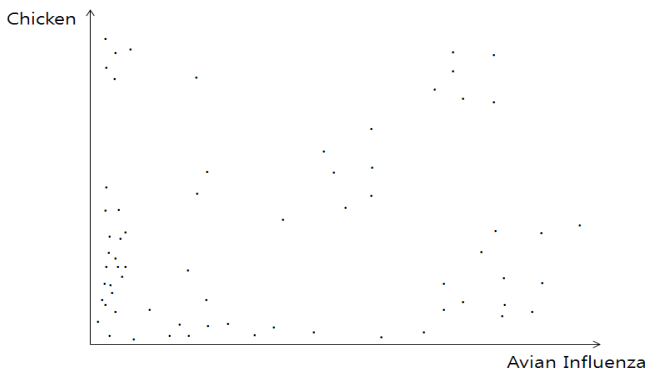

Figure 8: Dimension Analyze

Fig. 8. is a visualization of dimensional analysis. In large measure, they can be divided into four groups. First, poor data at a distance of zero can be viewed as having little or no impact, or in daily life or in personal emotions. The second group with a high score for Chicken is a famous chef or restaurant. For the third time, the group that scored high on Avian Influenza is the writer of the scholar, the government office, and the media. Finally, Chicken and Avain Influenza were two of the most influential source groups from the writings of Hwasung City Hall.

CONCLUSION

In this paper, we study the value of social data to build a Trust big data system. We assessed and tested the influence of social data through quotations, comments, and relationships with surrounding people. However, our Big Data system is currently in its infancy and is not fully implemented. This need to be supplemented. In order to satisfy the system architecture proposed in this paper, additional implementation and researches are needed on the analysis model and the collection model.

REFERENCES

[1] Chang, F., Dean, J., Ghemawat, S., Hsieh, W.-C., Wallach, D.-A., Burrows, M., Chandra, T., Fikes, A. and Gruber, R.-E., Bigtable: A Distributed Storage System for Strutured Data. *ACM Transactions on Computer Systems*, 26.2 (2008), 205-218.

[2] DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman A.,Pilchin, A., Sivasubrmamanian, S., Vosshall, P. and Vogels, W., Dynamo: Amazon's Highly Available Key Value Store. *ACM SIGOPS Operating System Review.*, 41.6 (2007), 205-220.

[3] Shvachko, K., Kuang, H., Radia, S., and Chansler, R., The Hadoop Distributed File System. *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*, IEEE (2010), 1-10.

[4] Son, J.-E., Jang, Y.-J., Lee, S.-H. and Kim, H.-W., A Systems Thinking Approach to the Enhancement of Social Capital: In Case of Social Media Users. *Information Systems Review*, 15.2 (2013), 21-40.

[5] Landset, S., Khoshgoftaar, T.-M., Richter, A.-N., and Hasanin, T., A Survey of Open Tools for Machine Learning with Big Data in the Hadoop Ecosystem. *Journal of Big Data.*, 2.24 (2015), DOI 10 1186/s40537-015-0032-1.

[6] Alexandrov, A., Bergmann, R., Ewen, S., Freytag, J.-C., Hueske, F., Heise, A. and Naumann, F., The

Stratosphere Platform for Big Data Analytics. *The VLDB Journal*, 23.6 (2014), 939-964.

[7] Brin, S., and Page, L., Reprint of: The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer network,* 56.18 (2012), 3825-3833.

[8] Kim, J., J.,, Kwak, K., J.,, Lee, D., H.,, and Lee, Y., S., Study of Trust Bigdata Platform. Journal of IIBC, Vol. 16, No 6, (2016)

[9] Julie Kim, Hyokyung Bahn, An Efficient Log Data Processing Architecture for Internet Cloud Environments. Journal of IIBC, Vol. 8, No 1, (2016)

[10] Shin, J., S., SNS using Big Data Utilization Research. Journal of IIBC, Vol. 12, No. 3, (2012), 257-265