Big Data Service Engine (BISE): Integration of Big Data Technologies for Human Centric Wellness Data

Muhammad Idris *, Shujaat Hussain *, Mahmood Ahmad *, Sungyoung Lee *
*Department of Computer Engineering, Kyung Hee University,South Korea
Email: {idris, shujaat.hussain, rayemahmood, sylee}@oslab.khu.ac.kr

***Abstract*—The advancement in new technologies and their data generation at substantial rate gave birth to the Big Data and requires a robust platform to capture, retrieve, store, and process it. Data generated by Human centric services and applications such as sensors, healthcare applications, social networks, and smart-phones need to be collected and processed to provide in-depth knowledge. In this paper, we propose Hadoop Distributed File System (HDFS) as convergence platform where all these multi-structured data is stored and use Hadoop No-SQL solutions to build warehouse for applications real time access to the data. We manage users clinical, personalized, and feedback data to provide clinical, physical, social, and mental health monitoring platform. We implement a Big Data service engine which provides storage services to health monitoring systems and analytics services to visualize and monitor clinical information, physical activities and emotions performed by the users. Our prototype system successfully integrates various technology platforms and provide centralized health monitoring system.**

***Keywords*-BigData; Hadoop; MapReduce; Cloud; Healthcare;**

I. Introduction

With the sound and fast development of Information Technology (IT), the data growth has increased substantially. Ubiquitous devices/sensors, social networking sites (SNS), healthcare applications, business transactions/logs, and other devices generate data in peta bytes per second in different formats and structures [1]. Data generation at exponential rate gives birth to volumetric data, in variety of formats including unstructured (text, CSV), semi-structured (XML, CDA), and structured (business transaction, RDBMS). Key parameters of generated data include data growth rate, volume, variety, and uncertainty/noise leads to the generation of Big Data [1]. To handle data at such large scale, cloud computing is a point of interest in the Big Data research community.

Cloud vision was first given by McCarthy at MITs centennial celebration in 1961 [2] as computing may be some time organized as public utility. It is a group of commodity hardware sharing computation resources to provide software, infrastructure, and platform as a service. Many cloud based solution for Big Data has been proposed to provide compute and data intensive [3] computation platforms such as GFS, HDFS, BSP, Twitter Storm, and GPGPU. Such systems provide scalable, distributed, efficient, and fault tolerant mechanisms on commodity machines either as open-source implementation or proprietary services. Hadoop Distributed File System [4] has emerged as an efficient distributed storage and processing framework supporting MapReduce [5] computational model on the cloud. We use Hadoop as convergence platform for healthcare related data storage, processing, and analytics.

Healthcare applications such as clinical information processing [6], physical activities recognition [4], physical emotions recognition, and recommendations [7] are widely spread and individually available to the users. However, there is lack of a convergence platform where all these related information are collected and provided to the users for their health monitoring, activities analytics, and recommendation based on their data. In this paper, we propose Big Data Service Engine (BISE) using Hadoop where all healthcare related data are stored, processed, used by other applications and perform analytics on it.

Big Data Service Engine (BISE) is a platform where HDFS is used as storage system for multi-structured data and MapReduce computation model for processing it. BISE transforms multi-structure data stored in HDFS into an intermediate relational database which is used by other applications for activity recognition [4], emotion recognition [8], recommendation, and sentiment analysis [9]. It uses MapReduce for legacy and new data transformation and visualization. Clinical data such as Electronic Health Record (EHR), Patient Health Record (PHR), and Electronic Medical Record (EMR) is used as service from Smart CDSS [10] to provide the physicians with clinical recommendations for their patients. We have built a prototype system for ATHENA [11] and PULSE [12] with all its functions. This paper covers the introduction of Hadoop for integration of various multi-structure data and its processing to provide healthcare services. However, functional statistics and algorithms of each service components are out of the scope of this paper. The objectives of the paper include

1) Exploiting Big Data technology Hadoop-MapReduce to provide health-care services
2) To present a service engine that provides health-care services for maintaining physical, social, and mental health
3) A platform to handle variety of data at a single place to perform analytics and present services. It will help the researchers to guide themselves towards the future trends in the research area of personalized Big Data.

The rest of the paper is organized as follows. Section II discusses the related work aligned close to our work. In section III, we describe the proposed architecture and its components. Section IV shows the implementation and results of the system. Section V concludes the paper with discussion.

II. Related Work

Hadoop is a framework designed to process Big Data on commodity cluster with its computational paradigm as MapReduce. It was first proposed by Google as Google File System (GFS) [13] and Apache Hadoop is an open source implementation of GFS. Hadoops Distributed file System (HDFS) is adopted by many large organizations like Facebook [14], Oracle [15], and Yahoo [16]. MapReduce has been adopted for large scale indexing strategies and recommended for indexing at terabyte scale data. Due to its fault-tolerance, efficiency, reliability, and cluster of commodity machines framework; it has been widely adopted by machine learning, healthcare, and data analytics community.

Hadoop and MapReduce has been widely used in healthcare. Rose et al. [17] from Teradata have proposed the Big Data in healthcare ecosystem. To investigate the potential value in the data and know significant challenges, Hadoop and MapReduce have been adopted for Big Data processing such as data exploration and analytics.

To support large-scale medical image analysis, Markonis et al. [18] have proposed three use cases including: 1) parametric optimization for lung texture classification, 2) content-based medical image indexing, and 3) Three dimensional directional wavelet analysis. Their results show considerable decrease in processing time by reducing a task performing for 130 hours to just 6 hours. Hadoop MapReduce has also been adopted for large scale peptide identification [19] for searching in spectral library database. A search application called MR-MSPolygraph presents a hybrid approach to match experimental spectrum against some protein sequence in a spectral library. Their results on 400 cores cluster of commodity machines compared to a single state-of-the-art desktop takes less time by a factor of 333 approximately. However, these systems lack support for multi-structured data and daily healthcare services convergence.

In today's fast business environment, StreamCentral[20] is an end user application/product that allows its users to quickly process new stream of data, add new data from a source,and analyzing and uncovering the business performance. StreamCentral provides the features including operational intelligence, on-demand business intelligence, knowledge discovery, and social collaboration. However, our BISE framework is intended for health-care services integrating mental, physical, and social health information. BISE framework is different from StreamCentral in its nature(BISE uses Hadoop), purpose (BISE is focused on active life-style), and services.

Large data generated by Social Networking Sites(SNS) including Twitter, Facebook, and Myspace could be used to predict and analyze the trend analysis in the affairs of the world, healthcare, and other known areas. Hyeokju et al. [21] have proposed an implementation for large-scale social data analysis on MapReduce. Their system consists of two parts: Data gathering agent and data analysis module. Data analysis module is composed of improved versions of TF-IDF and Weighted-MINMAX analysis to extract high priority trending words. System is evaluated based on efficiency and performance and shows 3-5 times improved performance comparing to the existing systems and it can even more be decreased by additional nodes in the cluster. To provide smart-phone users the advancement of distributed cheap, available, and reliable technologies; Adam et al. [22] have proposed a framework form mobile applications based on MapReduce to provide strong connectivity model of computation in open networks.

In existing works, the proposed frameworks and developed models mostly work and use specific structured (relational) or unstructured (textual, sensory, XML) data. Other Bulk Synchronous Parallel[23] processing and GraphLab[24] are useful when the data is posses high density of relationships. These frameworks are mostly efficient for high compute intensive processing on single format of the data. However, Hadoop exhibit the property of dealing with variety data including structured, semi-structured, and unstructured. Twitter Storm can be configured to process high velocity of data and works on streaming data. However, it also does not posses the ability of processing and managing variety of data. Therefore, depending on the data structures and the problem we are addressing, Hadoop is suitable candidate to handle the data at such large scale in various formats.

III. BISE Architecture

In this section, we present our Big Data Service Engine (BISE). Our system stores and processes multi-structured data from various sources to provide health-care services. The proposed system is composed of three layers and an adaptive intermediate database. The three layers include data acquisition layer, data storage and processing layer, services layer. Data acquisition layer consists of multiple sources of collecting/generating data for the system; storage and processing layer consists of data processing modules; and services layer consists of multiple service engines facilitating healthcare. The overall architecture of the proposed system is shown in Figure 1. Adaptive intermediate database is used by services layer sub-components to perform computation and calculations on relational data for clinical, physical, mental, and social health services. Each sub-component in the services layer perform statistical and analytical algorithms to provide healthcare services. However, these algorithms are out of the scope of this paper. Layers and its sub-component are explained are discussed in the subsections in detail.

A. Data Acquisition Layer

Big Data Service Engine (BISE) process variety of data collected from various sources in different formats including XML, CSV, relational, and images. Data sources include smart-phone sensors, wearable sensors, embedded sensors, location sensors, 2D/3D cameras, Social Network Sites (SNS),

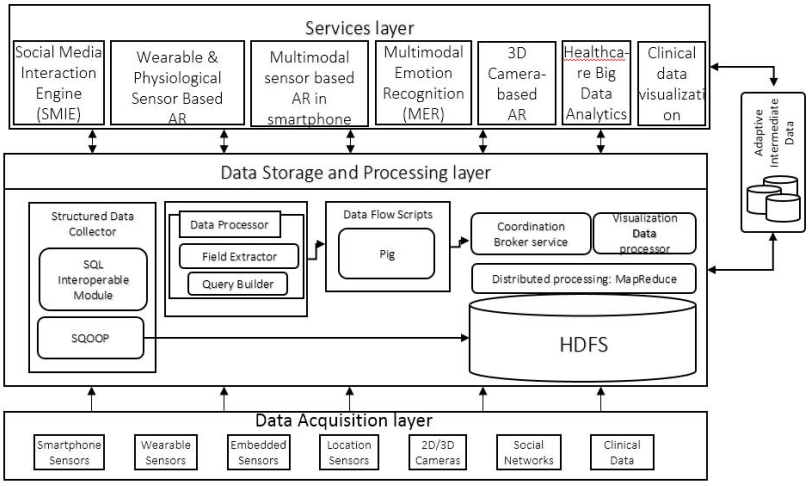

Fig. 1. BigData Service Engine Detailed Architecture

and health-records. Sensors generate data in different formats and it is used for activities recognition [4] in the services layer. 2D/3D cameras generate data in the form of images is used for emotion recognition using facial expression [8] techniques. SNS data is in textual and XML formats, and used by Social Media Interaction engine (SMIE) in services layer to monitor users daily activities and interests. Clinical data is in Clinical Document Architecture(CDA)/relational formats and used to generate healthcare analytics. These analytics are based on sugar, blood pressure diseases trends based on gender, location, age, and various other parameters. Currently data acquisition is performed in batch mode to store in HDFS. Techniques such as timestamps, hierarchy, and directory structures are used to differentiate the data from various input sources and maintain the versions.

B. Data Storage and Processing Layer

In this layer, the data stored in HDFS is processed according to the needs of health-care services in services layer. Apache Sqoop [25] and Pig [26] are used to transfer bulk data to populate adaptive intermediate database and process large data present in Hadoop for analytics respectively. Data storage and processing layer has the following sub-components.

1) Structured Data Collector: The objective is to generate and collect structured data from variety data in HDFS to be processed by services modules in services layer. The structured data is stored in the intermediate database for quick access and processing. Each module in the services layer such as SMIE, Activity Recognition (AR), and Emotion recognition (ER) provide schema to the structured data collector based on requirements of the service and data collected in the acquisition layer. Sqoop is used to transfer the data from large unstructured formats (HDFS, Pig) to structured format (intermediate database).

2) Data Processor: In BigData Service Engine, data comes from diverse input sources in variety of formats. Various data formats include structures such as relational data (clinical data i.e. EHR), semi-structured such as XML (Twitter streams), unstructured such as sensory data. Sub-components in services layer need to process data of a specific format such as activity recognition(AR) needs to process data using machine learning techniques. To provide required data to the sub-components, data processor module generates and executes Apache Pig script to perform operations on these large data sets. Pig scripts are generated based on the visualization requirement from the services layer. Visualization and analytics could be performed on any data including physical activities, facial expressions and emotions, visited places, clinical findings, and clinical disease analysis. Data processor is meant to process HDFS data to identify abnormal behavior in a users all data including clinical, social, emotional, and activities. It has two sub components: Field extractor, and Query builder. Field extractor extracts the parameters passed from the upper layer for data processing. Query builder generates queries based on the parameters extracted by field extractor and executes on HDFS through Data Flow Scripts (Apache Pig). The results returned are stored back in HDFS and consumed by visualization module in services layer to produce interactive visual images as shown in Figure 2.

3) Coordination Broker: Data in HDFS is multi-structured and multi-purpose, therefore coordination broker serves as decision point to select appropriate data required by each services layer module. Data stored in HDFS is integrated from diverse input sources, coordination broker works to parse and process the required data set requested by services layer sub-components instead of parsing the whole repository. It works as switch, where based on the request, the required data in HDFS is processed and provided back as response. BISE is a cloud based multi-user, multi-service framework, therefore broker schedules the requests to for efficient processing. Some sub-components such as Activity Recognition needs to process data in structured format, for this purpose, the data needs to be converted from semi-structured or unstructured format to a structured format. Coordination broker needs to differen-

tiate between data transformation requests and data retrieval requests. By introduction of coordination broker, running and response time of data transfer can be reduced significantly. It also serves as selection point for the data required for visualization.

C. Services Layer

In BISE, we provide users health monitoring in terms of their clinical, social, mental, and physical health services and recommendations. Services layer in BISE includes sub-components to provide individual services such as social media interaction, activity recognition based on wearable sensor logs, activity recognition in smart-phones, emotion recognition in 2D/3D video and images, clinical data visualization, and recommendations based on these different data-sets. From all these different services, a user/patient can monitor and analyze current physical activities, emotions, and clinical conditions. All the services in this layer access and use the data in the adaptive intermediate database while visualization and analytics consume the data directly generated using pig scripts.

IV. Implementation and Results

In this section, we describe the experimental framework for the prototype development of BISE and present some results based on a scenario which covers the details of which data is used and what health information and visualization are generated. The goal of this paper is to provide the framework for heterogeneous huge data sources to integrate and provide an easily/cheaply available system on the cloud. The scenario presented only presents about how do we collect and use the data from heterogeneous sources for the better life style of user.

A. Experimental framework and implementation

We have implemented Big Data Service Engine (BISE) as sub-part of two healthcare projects including Activity awareness for Human Engaged wellness Applications (ATHENA) [11] and Personalized Ubiquitous Life care Support Engine (PULSE) [12]. Hadoop Distributed File System (HDFS) is used as variety data storage mechanism and MapReduce as its processing framework. Apache Pig is configured on top of Hadoop to provide quick data processing for activities and clinical data visualization. Apache Sqoop is used to transfer and convert data from HDFS (various formats) to adaptive intermediate database which is used by services layer for service provisioning. Data upload to HDFS is performed as batch upload and service layer access the data in HDFS and intermediate database in real time.

Hadoop cluster consists of 10 nodes with five physical machines and one virtual machine on each physical machine. All physical machines have 4GB RAM, 500GB HDD, Intel core i5 processor, and Ubuntu 12.10 Operating System. Data is collected from Social network Sites like Facebook and Twitter, Smartphone activity recognition sensors such as accelerometer, video 2D/3D cameras, and clinical observation data from hospital such as electronic Health record (EHR). Services layer sub-components are developed in Oracle Jdeveloper[27] as Web application along with analytics and activities monitoring. The prototype system successfully integrates various technologies and provides centralized health monitoring to the users. Some of data visualization results are shown in the **??** illustrating activities and clinical observations.

B. Case Study Scenario and Discussion

Considering a user who wants to monitor his physical, mental activities and his/her food intake. The activities data is generated by various activity recognition sensors and wearable devices, daily routines, and mental health data is collected from his social activities on Facebook[28], twitter[29], KakaoStory[30] and images. Food information along with their calorie intake is collected from his daily routine check-in and check-outs at restaurants. Based on these information of the user, the system visualizes the user data and shows a dashboard as shown in Figure 2. In this Figure, user activities, calorie consumption/intake, daily mood, and overall mental health are visualized. Calorie consumption information are provided by the user and calorie intake information are collected using the check-in and food intake that the user have done in the restaurants. Based on these information, users can monitor their mental, physical, and social health and compare with normal human life style. We have proposed an integrated solution to use variety of data in a single place as Big Data Service engine (BISE).

V. Conclusion and Future Work

A health-care system which offer services in low cost by managing information from diverse modalities is a challenging issue. In this paper we describe design and development of Big Data Service engine (BISE) that provide storage and processing services for large and variety of data sets from diverse sources. To achieve the objective, we use multiple Big Data technologies for storage and processing, and collect activities data from sensors, user experience and social information from social media data, and medical history records. The framework and data is used by various services such as activity/emotion recognition, social media interaction, and clinical visualization. Clinical data visualizations are performed based on medical expert knowledge to infer the trends in the user medical history. The current system provides an integrated, easy, and anytime availability to users to monitor their healthy life style.

In future research, we plan to enable and extend the system for online streaming of data acquisition as well as data used by different services. We are also developing real-time streaming engine to provide an easy, direct, and real-time access to the data so that the users can monitor their data easily.

Acknowledgment

This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2013-(H0301-13-2001))

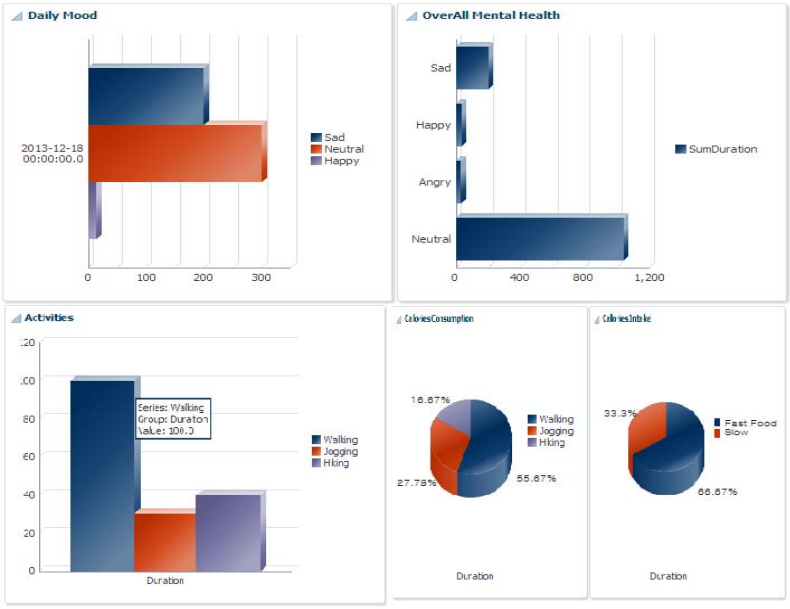

Fig. 2. visualization of daily mood(top-left), overall mood(top-right), daily activities(bottom-left), and daily calorie consumption/intake(bottom-right).

REFERENCES

[1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," 2011.

[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica *et al.*, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[3] J. Shamsi, M. A. Khojaye, and M. A. Qasmi, "Data-intensive cloud computing: requirements, expectations, challenges, and solutions," *Journal of grid computing*, vol. 11, no. 2, pp. 281–310, 2013.

[4] Y. Han, M. Han, S. Lee, A. Sarkar, and Y.-K. Lee, "A framework for supervising lifestyle diseases using long-term activity monitoring," *Sensors*, vol. 12, no. 5, pp. 5363–5379, 2012.

[5] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[6] M. Hussain, M. Afzal, W. A. Khan, and S. Lee, "Clinical decision support service for elderly people in smart home environment," in *Control Automation Robotics & Vision (ICARCV), 2012 12th International Conference on*. IEEE, 2012, pp. 678–683.

[7] W. Yuan, D. Guan, Y.-K. Lee, S. Lee, and S. J. Hur, "Improved trust-aware recommender system using small-worldness of trust networks," *Knowledge-Based Systems*, vol. 23, no. 3, pp. 232–238, 2010.

[8] M. H. Siddiqi, S. Lee, Y.-K. Lee, A. M. Khan, and P. T. H. Truc, "Hierarchical recognition scheme for human facial expression recognition systems," *Sensors*, vol. 13, no. 12, pp. 16 682–16 713, 2013.

[9] R. Batool, A. M. Khattak, J. Maqbool, and S. Lee, "Precise tweet classification and sentiment analysis," in *Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on*. IEEE, 2013, pp. 461–466.

[10] M. Hussain, A. Khattak, W. Khan, I. Fatima, M. Amin, Z. Pervez, R. Batool, M. Saleem, M. Afzal, M. Faheem *et al.*, "Cloud-based smart cdss for chronic diseases," *Health and Technology*, vol. 3, no. 2, pp. 153–175, 2013.

[11] M. Fahim, M. Idris, R. Ali, C. Nugent, B. Kang, E.-N. Huh, and S. Lee, "Athena: A personalized platform to promote an active lifestyle and wellbeing based on physical, mental and social health primitives," *Sensors*, vol. 14, no. 5, pp. 9313–9329, 2014.

[12] W. A. e. a. Khan, "Personalized ubiquitous life-care decision support system (pulse): http://uclab.khu.ac.kr/pulse," 2013.

[13] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," in *ACM SIGOPS Operating Systems Review*, vol. 37, no. 5. ACM, 2003, pp. 29–43.

[14] D. Borthakur, "Facebook has the worlds largest hadoop cluster," *Retrieved April*, vol. 20, p. 2012, 2010.

[15] S. Shankar, A. Choi, and J.-P. Dijcks, "Integrating hadoop data with oracle parallel processing–an oracle white paper," *Oracle Corporation*, 2010.

[16] R. T. Kaushik and M. Bhandarkar, "Greenhdfs: Towards an energy-conserving storage-efficient, hybrid hadoop compute cluster," in *Proceedings of the USENIX Annual Technical Conference*, 2010.

[17] T. Data, "Asterdata-aster: http://www.asterdata.com/solutions/big-data-analytics-discovery.php," 2013.

[18] D. Markonis, R. Schaer, I. Eggel, H. Müller, and A. Depeursinge, "Using mapreduce for large-scale medical image analysis." in *HISB*, 2012, p. 1.

[19] A. Kalyanaraman, W. R. Cannon, B. Latt, and D. J. Baxter, "Mapreduce implementation of a hybrid spectral library-database search method for large-scale peptide identification," *Bioinformatics*, vol. 27, no. 21, pp. 3072–3073, 2011.

[20] S. solutions, "Streamcentral," 2013.

[21] H. Lee, J. Her, and S.-R. Kim, "Implementation of a large-scalable social data analysis system based on mapreduce," in *Computers, Networks, Systems and Industrial Engineering (CNSI), 2011 First ACIS/JNU International Conference on*. IEEE, 2011, pp. 228–233.

[22] A. Dou, V. Kalogeraki, D. Gunopulos, T. Mielikainen, and V. H. Tuulos, "Using mapreduce framework for mobile applications," *Multimedia Services and Streaming for Mobile Devices: Challenges and Innovation*, p. 181, 2011.

[23] T. Cheatham, A. Fahmy, D. Stefanescu, and L. Valiant, "Bulk synchronous parallel computinga paradigm for transportable software," in *Tools and Environments for Parallel and Distributed Systems*. Springer, 1996, pp. 61–76.

[24] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein, "Graphlab: A new framework for parallel machine learning," *arXiv preprint arXiv:1006.4990*, 2010.

[25] Apache, "Sqoop: http://sqoop.apache.org/," 2014.

[26] ——, "Pig: https://pig.apache.org/," 2014.

[27] Oracle, "Oracle jdeveloper: http://www.oracle.com/technetwork/developer-tools/jdev/overview/index.html(last accessed 05-05-2014)," 2014.

[28] "Facebook: http://www.facebook.com(last accessed 05-05-2014)," 2014.

[29] "Twitter: http://www.twitter.com(last accessed 05-05-2014)," 2014.

[30] "Kakao story, social network: https://play.google.com/store/apps/ details?id=com.kakao.story (last accessed 05-05-2014)," 2014.