Constructing National Geospatial Big Data Platform: Current Status and Future Direction

Junghee Jo, In-Hak Joo, and Kang-Woo Lee
IoT Research Division
Electronics and Telecommunications Research Institute
Daejeon, South Korea
{dreamer, ihjoo, kwlee}@etri.re.kr

***Abstract*—With the increasing predominance of Internet of Things (IoT) applications, a considerable amount of geospatial information has been accumulating from various sources. In addition to the common features of big data, the unique characteristics of spatial data make the treatment of big spatial data even more complicated. To facilitate developers creating big spatial data applications, it is imperative to develop new technologies to efficiently handle the massive amount of big spatial data. Given this impetus, the Korean government launched a five-year national project involving businesses, government, and the research community. The goal is to develop a platform for efficiently storing, extracting, processing, and analyzing geospatial big data. This paper explains the expected outcome from the project including the overall architecture of the platform, along with its current status and future direction.**

Keywords—big spatial data, IoT, Hadoop, MapReduce, geospatial information.

I. Introduction

In the settings of Internet of Things (IoT), numerous sensors in various domains create enormous volumes of data at a rapid pace, a considerable portion of which is big spatial data. It is widely agreed that 80% of data in the world has a geospatial component [1-3]. According to an estimate by the McKinsey Global Institute, the amount of location data was about 1 PB in 2009 and is growing at an annual rate of 20% [4]. Google creates about 25 PB of data per day, a large amount of which consists of spatiotemporal features [5]. The United Nations Initiative on Global Geospatial Information Management estimates that about 2.5 quintillion bytes of data are being generated daily and a considerable portion includes location features [6]. Due to this rapid increase of big spatial data, use of the data is increasingly drawing the attention not only of industry but also government.

In the US, the Federal Geographic Data Committee (FGDC) develops national geospatial standards to promote coordinated development, use, and sharing of big spatial data on a national basis [7]. To provide one-stop access to a variety of federal geospatial datasets and associated services to government agencies, the private sector, and academia. The FGDC created Geospatial Platform that embodies the principles and spirit of open government [8]. Similarly, the European Global Earth Observation System of Systems (GEOSS) is a platform where all data providers are connected to a single infrastructure, accessible via the GEOSS Portal. GEOSS Portal provides single access to earth observation data and supports intuitive interfaces enabling easy search of the resources. The system is continually upgraded, recently addressing big data challenges [9].

With the introduction of the fourth industrial revolution, the Korean government has initiated big data projects to enhance citizens' lives and address national challenges. Recently, the government announced plans to launch a national big data initiative, aiming to manage the national data sets kept at various government offices and offer data analysis and consulting services for the private sector. To accomplish this, the government plans to launch a task force team playing the lead role to construct necessary infrastructure technologies to converge knowledge and administrative analytics via big data. Several ministries and agencies have proposed associated action plans – currently there are about 30 independently managed government big data teams.

The Ministry of Land, Infrastructure, and Transport of the Korean Government has focused especially on big spatial data as the key issue of the fourth industrial revolution and initiated a new research project on big spatial data by involving businesses, government, and the research community [10]. This is a five-year project to research and develop three core big spatial data technologies: geospatial big data management, geospatial big data analytics, and geospatial big data service. In this paper, we present the architecture of the national geospatial big data platform, focusing particularly on the geospatial big data management system. This paper also explains the outcome from the research project, along with its current status and future direction.

II. National Geospatial Big Spatial Platform

The overall architecture of the national geospatial big data platform currently developed by the Ministry of Land, Infrastructure, and Transport is shown in Fig. 1. The platform is divided functionally into six layers: geospatial big data, geospatial ETL (i.e., extract, transform, load), geospatial big data management, analytical engine, presentation layer, and service layer.

978-1-5386-4980-0/19/$31.00 ©2019

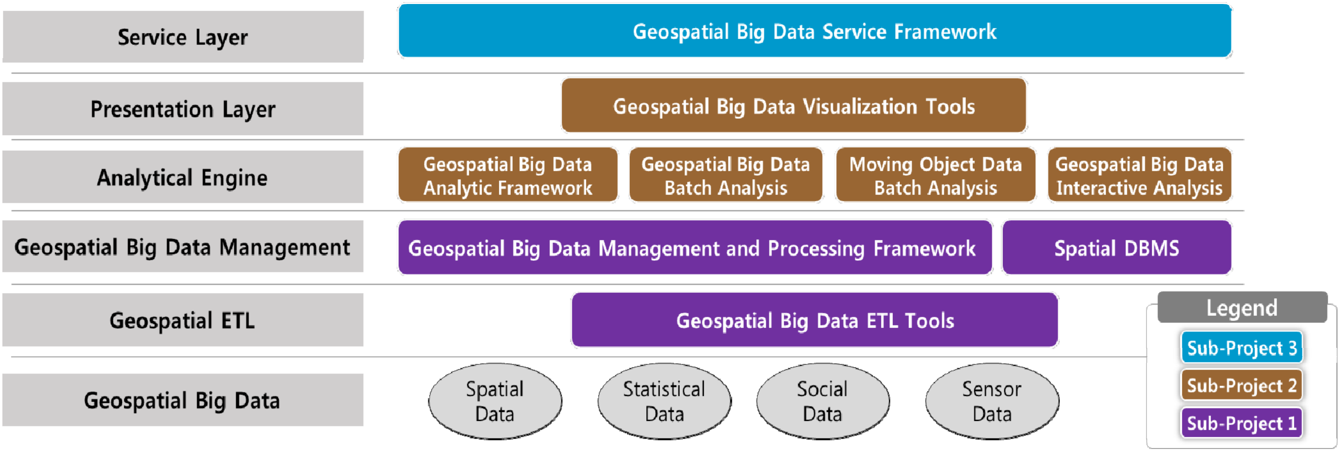

Service Layer
Geospatial Big Data Service Framework
Presentation Layer
Geospatial Big Data Visualization Tools
Analytical Engine
Geospatial Big Data Analytic Framework
Geospatial Big Data Batch Analysis
Moving Object Data Batch Analysis
Geospatial Big Data Interactive Analysis
Geospatial Big Data Management
Geospatial Big Data Management and Processing Framework
Spatial DBMS
Geospatial ETL
Geospatial Big Data ETL Tools
Geospatial Big Data
Spatial Data
Statistical Data
Social Data
Sensor Data
Legend
Sub-Project 3
Sub-Project 2
Sub-Project 1

Fig. 1. Overall architecture of the national geospatial big data platform.

This project is composed of three sub-projects, each of which is responsible for researching and developing specific layers. The first sub-project is responsible for geospatial ETL and geospatial big data management; the second, the analytical engine and presentation layer; the third, the service layer in the platform.

Geospatial big data is collected from various sources including spatial data, statistical data, social data, or sensor data. The Ministry of Land, Infrastructure, and Transport, for example, has been collecting considerable amount of geospatial data, e.g., 2D/3D digital maps, cadastral maps, topographic maps, satellite images, aerial photographs, and digital elevation model information. We currently utilize those data sets along with outside data sources (e.g., Movebank [11]) in order to test the geospatial big data platform.

Software components in the layers of geospatial ETL and the geospatial big data management are responsible for storing, extracting, and processing geospatial big data. The framework of geospatial big data management plays a significant role in the entire platform. Fig. 2 shows the system architecture of geospatial big data management. The system is composed of two components: the processing engine and spatial query engine. A geospatial big data processing engine has been developed by extending the original Hadoop to support spatial functions. Hadoop is an open source MapReduce implementation which is recognized as the most representative big data framework. During the initial phase of developing the engine, we surveyed most of the existing geospatial big data processing frameworks on Hadoop and found that the majority had been built as plugins for NoSQLs, such as Pig Latin and Hive.

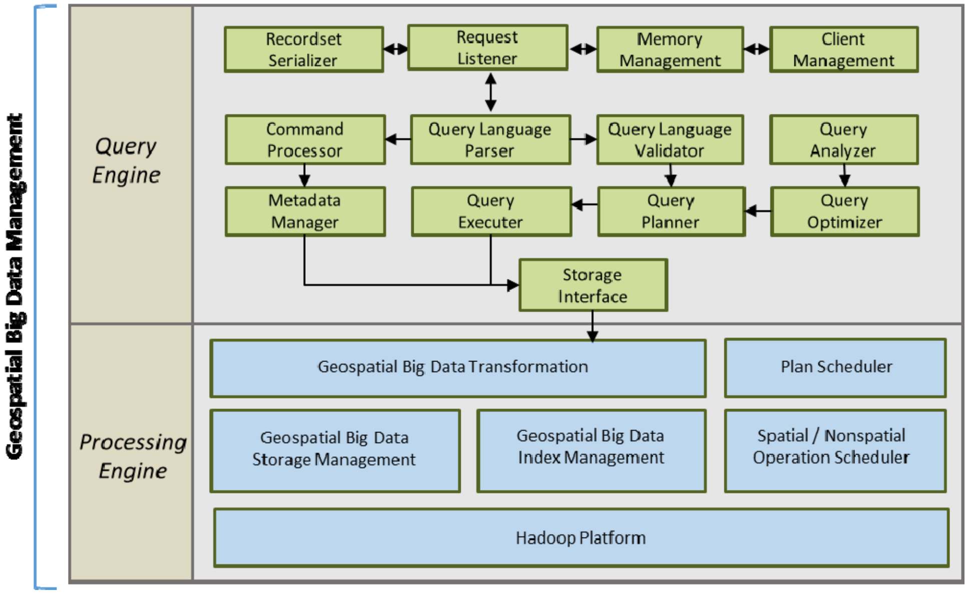


Fig. 2. System architecture of geospatial big data management.

Due to the inherent limitations of plugin-based approaches have created, they suffer from performance problems especially when an application is built as a combination of spatial operations and non-spatial operations. Doing so causes massive volumes of data transfer between NoSQL and spatial plugins.

Therefore, we concluded that if we decreased the overhead associated with massive data transfer between NoSQL and spatial plugins, we could improve data processing performance. Hence, we proposed and implemented a new method that maps a sequence of operations (both a spatial one and a non-spatial one) into a sequence of MapReduce jobs. In addition, developers who use MapReduce frequently conduct spatial analysis that is too complex to be transformed to MapReduce jobs. To address these issues, the processing engine automatically constructs one or more MapReduce jobs from a given spatial analysis task. When the task is transformed to MapReduce jobs, the processing engine controls the number of jobs in such a way to achieve better performance by decreasing the overhead of conducting mapping and reducing [12].

From the perspective of big spatial data query language, most big data systems are already furnished with high-level languages such as HiveQL [13] or Pig Latin [14]. Similarly, a geospatial big data platform needs to be equipped with high-level languages that reduce all the complexities of the platform. The query engine in our platform supports Open Geospatial Consortium (OGC)'s standards compliant spatial data types and spatial operations. This is because these standards are already being used by existing spatial systems making it easier for application developers to use. The engine supports user-familiar SQL-based query language that enables developers who utilize spatial data to easily transform from the existing spatial RDBMs to a Hadoop-based big data system. This engine involves core functions for query processing such as query language parsing, query validation, query planning, and connection with the processing engine.

The analytical engine, presentation layer, and service layer are designed to conduct spatial analysis using either batch data or real-time data and provide the results via various visualization tools including digital maps.

III. Current Status and Future Direction

It has been four years since this project was initiated and the development of each individual layer is now nearing completion. For the next two years, each developed layer will be integrated into one complete platform for final release. The platform will then be stabilized by incorporating outside feedback.

Fig. 3 shows the current test environment of the national geospatial big data platform. The test was conducted on the four nodes of a Hadoop cluster. Each node is a desktop PC with a 4.0 Ghz 4 core CPU, 32GB main memory, 4TB disk. The operating system is CentOS 6.9 and the Hadoop version is Hortonworks HDP 2.6.1.0 with Ambari 2.5.0.3. The specific information about the test environment are shown in Table 1.

TABLE I. Specifications of the Test Environment

Items	Specifications
CPU	Intel Core i7, 4 cores
RAM	32GB
Storage	4TB
OS	CentOS 6.9 (64bit)
JDK	Oracle JDK 1.8
DB	PostgreSQL 9.5

Hadoop Echosystem	Specifications
HDFS	2.7.3
YARN	2.7.3
MapReduce2	2.7.3
ZooKeeper	3.4.6
Storm	1.1.0
Flume	1.5.2
Kafka	0.10.1
Spark	1.6.3
Spark2	2.1.1

The national geospatial big data platform could be used to provide various services including personal, regional, and nationwide. For example, information on civil affairs is currently represented as plain text and is manually processed by the person in charge. By analyzing cumulative information on civil affairs, the big data platform can extract where locations of interest occur and display the results on digital maps in connection with spatial information. It can supplement the intuitive judgment of the person in charge and provide effective policy support, as customized responses are possible.

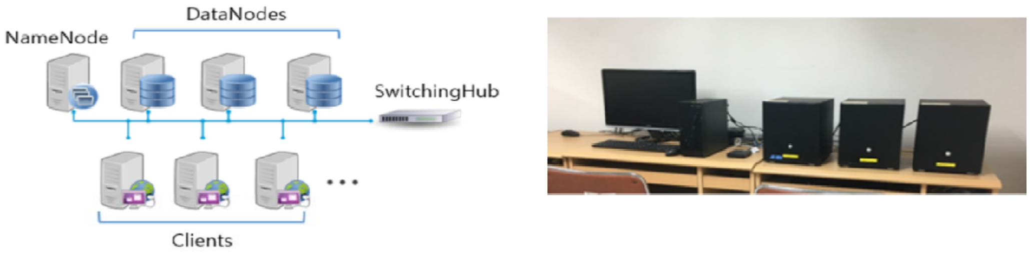
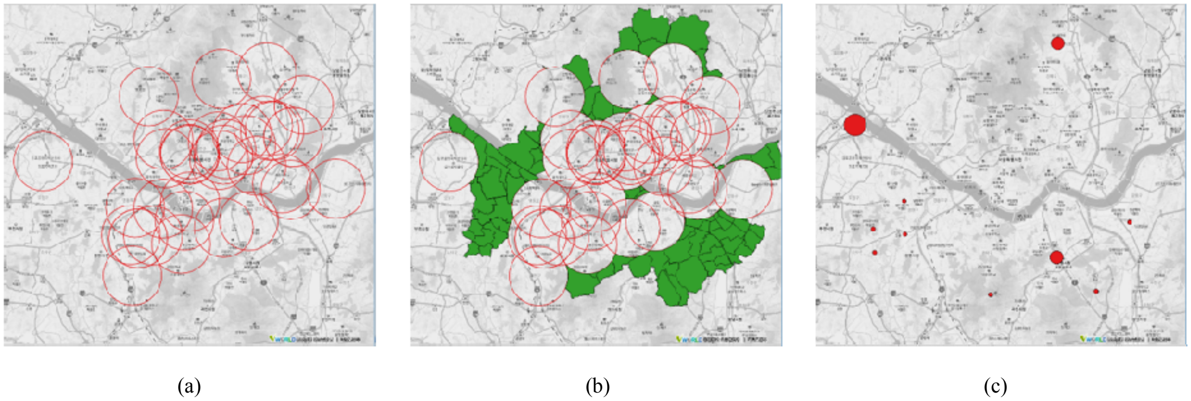


Fig. 3. Test environment of the national geospatial big data platform.



(a)

(b)

(c)

Fig. 4. A case analysis: extracting specific areas requiring construction of new hospitals with advanced burn units.

As another example, it is possible to derive unusual land areas by integrating spatial information, official land price, and floating population data in a specific area and analyzing the unusual value considering the spatial distribution of the official land price. Fig. 4 shows an example application of using the national geospatial big data platform. The goal of this analysis is to extract specific areas in Seoul city requiring construction of additional hospitals to treat patients with thermal injuries. This analysis is conducted via the following three steps. First, a certain buffer distance (e.g., 3km) is created around each of the hospitals in Seoul city, Fig. 4(a). Second, based on the generated buffers, non-adjacent areas from hospitals are extracted, Fig. 4(b). Finally, key areas are extracted which are not adjacent to any hospital but in which burn injury accidents have occurred, Fig. 4(c).

IV. Conclusion

This paper presents the architecture of a national geospatial big data platform for efficiently storing, extracting, processing, and analyzing geospatial big data. It is a five-year project undertaken by the Ministry of Land, Infrastructure and Transport of the Korean government to research and develop core big spatial data technologies. We expect that the platform will play a key role in helping citizens and address major national challenges that cannot be solved with current technology. In the coming two years, we will continue to develop the platform and conduct experiments in anticipation of its final release.

Acknowledgment

This research, 'Geospatial Big Data Management, Analysis and Service Platform 499 Technology Development', was supported by the MOLIT (The Ministry of Land, Infrastructure and Transport), 500 Korea, under the national spatial information research program supervised by the KAIA (Korea Agency for 501 Infrastructure Technology Advancement)(18NSIP-B081011-05).

References

[1] C.D. Morais, "Where is the Phrase "80% of Data is Geographic" From?", Available online: http://www.gislounge.com/80-percent-data-is-geographic (accessed on 10 October 2018).

[2] R. Jeansoulin, "Review of forty years of technological changes in geomatics toward the big data paradigm", ISPRS Int. J. Geo-Inf., 2016, 5, 155.

[3] Z. He, Q. Liu, M. Deng, and F. Xu, "Handling multiple testing in local statistics of spatial association by controlling the false discovery rate: A comparative analysis", In Proceedings of the 2017 IEEE International Conference Big Data Analysis (ICBDA), Beijing, China, 2017, pp. 684-687.

[4] A. Dasgupta, "Big Data: The Future is in Analytics", Available online: https://www.geospatialworld.net/article/big-data-the-future-is-in-analytics (accessed on 10 October 2018).

[5] R.R.Vatsavai, A. Ganguly, V. Chandola, A. Stefanidis, S. Klasky, S. Shekhar, "Spatiotemporal data mining in the era of big spatial data: Algorithms and applications", In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, Redondo Beach, CA, USA, November 2012, pp. 1-10.

[6] J. Carpenter and J. Snell, "Future trends in geospatial information management: The five to ten year vision", United Nations Initiative on Global Geospatial Information Management.

[7] D.J. Maguire and P.A. Longley, 2005, "The emergence of geoportals and their role in spatial data infrastructures", Computers, environment and urban systems, 29(1), pp.3-14.

[8] Geospatial Platform, Available online: https://www.geoplatform.gov (accessed on 10 October 2018).

[9] S. Nativi, P. Mazzetti, M. Santoro, F. Papeschi, M. Craglia, and O. Ochiai, 2015, "Big data challenges in building the global earth observation system of systems", Environmental Modelling & Software, 68, pp.1-26.

[10] J. Lee, M. Kang, "Geospatial big data: challenges and opportunities", Big Data Research. 2015. 2, pp.74-81.

[11] Movebank, Available online: http://www.movebank.org (accessed on 10 October 2018).

[12] J. Jo and K.Lee, "High-Performance Geospatial Big Data Processing System Based on MapReduce", ISPRS Int. J. Geo-Inf., 2018, 7, 399.

[13] A. Thusoo, J. S. Sen, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy, "Hive: A Warehousing Solution over a Map-Reduce Framework," PVLDB, 2009, pp. 1626–1629.

[14] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, "Pig Latin: A Not-so-foreign Language for Data Processing," in SIGMOD, 2008, pp. 1099–1110.