Big Data Research 2 (2015) 74–81

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Geospatial Big Data: Challenges and Opportunities ☆

Jae-Gil Lee *, Minseo Kang

Department of Knowledge Service Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea

ARTICLE INFO

Article history:
Received 19 September 2014
Accepted 4 January 2015
Available online 18 February 2015

Keywords:
Geospatial big data
Spatial big data
Complex event processing
Spatial online analytical processing

ABSTRACT

Geospatial big data refers to spatial data sets exceeding capacity of current computing systems. A significant portion of big data is actually geospatial data, and the size of such data is growing rapidly at least by 20% every year. In this paper, we explore the challenges and opportunities which geospatial big data brought us. Several case studies are introduced to show the importance and benefits of the analytics of geospatial big data, including fuel and time saving, revenue increase, urban planning, and health care. Then, we introduce new emerging platforms for sharing the collected geospatial big data and for tracking human mobility via mobile devices. The researchers in academia and industry have spent a lot of efforts to improve the value of geospatial big data as well as take advantage of its value. Along the same line, we present our current research activities toward the analytics of geospatial big data, especially on interactive analytics of real-time or dynamic data.

1. Introduction

Geospatial data has always been big data. In these days, big data analytics for geospatial data is receiving considerable attention to allow users to analyze huge amounts of geospatial data. *Geospatial big data* typically refers to spatial data sets exceeding capacity of current computing systems. McKinsey Global Institute says that the pool of personal location data was in the level of 1 PB in 2009 and is growing at a rate of 20% per year [1]. This estimation did not include the data from RFID sensors and those stored in private archives. According to the estimation by United Nations Initiative on Global Geospatial Information Management (UN-GGIM), 2.5 quintillion bytes of data is being generated every day, and a large portion of the data is location-aware. Also, in Google, about 25 PB of data is being generated per day, and a significant portion of the data falls into the realm of spatio-temporal data [2]. This trend will be even accelerated since the world becomes more and more mobile in these days. As in Fig. 1, in India, the internet traffic from mobile devices already exceeded that from desktop computers [3].

Along with this exponential increase of geospatial big data, the capability of high performance computing is being required greatly than ever, for modeling and simulation of geospatially enabled contents. However, because of limited processing power, it has been hard to fully exploit high-volume or high-velocity collection of geospatial data in many applications. Recently, distributed, parallel processing on a cluster of commodity computers or a cloud such as Amazon EC2[1] has been becoming widely available for use, breaking the existing limitations on processing power. In addition, big data platforms such as Hadoop [4], Hive [5], and MongoDB [6] have been developed such that users can implement big data analytics software very easily on a distributed, parallel computing platform. It is obvious that these recent improvements are providing us with a lot of opportunities for advanced analytics for geospatial big data [7–9]. According to Garner's hype cycle in Fig. 2, geospatial big data analytics belongs to the stage of *peak of inflated expectations* as of July 2012 [10].

Geospatial big data or simply spatial big data are societal opportunities [11,12]. The Millennium Project identified 15 global challenges that the human kind is facing as in Fig. 3 [13]. Many of them can benefit from geospatial big data. Shashi Shekhar [14], a renowned computer scientist, says that the seven challenges are related to geospatial big data, as indicated by boxes in the figure. For example, as for energy, eco-routing is one example that can save energy using geospatial big data. This technology minimizes fuel consumption rather than travel time or travel distance. For this purpose, eco-routing tries to find a route that avoids congestion, idling at red lights, turns and elevation changes, and so on. Compared to using the "Fastest Route" option, Ford researchers told that using the "Eco Route" option offered as much as 15% reduced

☆ This article belongs to Visions on Big Data.

* Corresponding author.
E-mail addresses: jaegil@kaist.ac.kr (J.-G. Lee), minseo@kaist.ac.kr (M. Kang).

[1] http://aws.amazon.com/ec2/.

http://dx.doi.org/10.1016/j.bdr.2015.01.003
2214-5796/

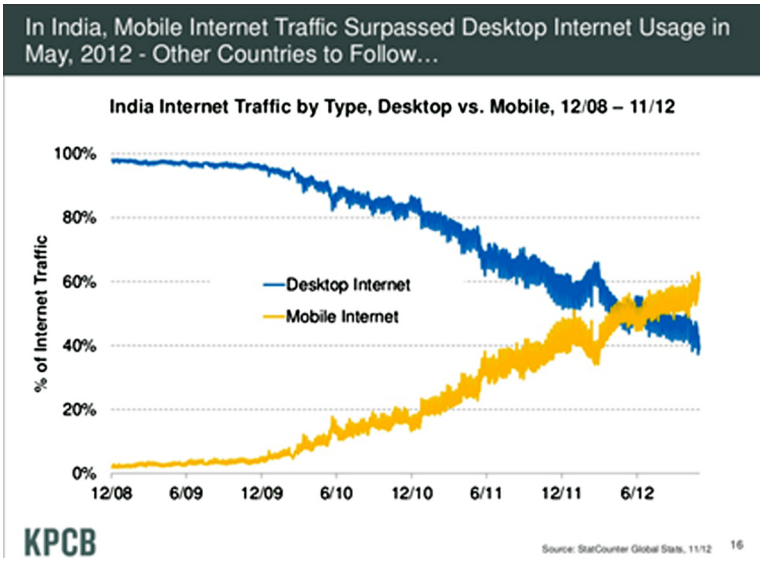


Fig. 1. Mobile internet traffic in India [3].

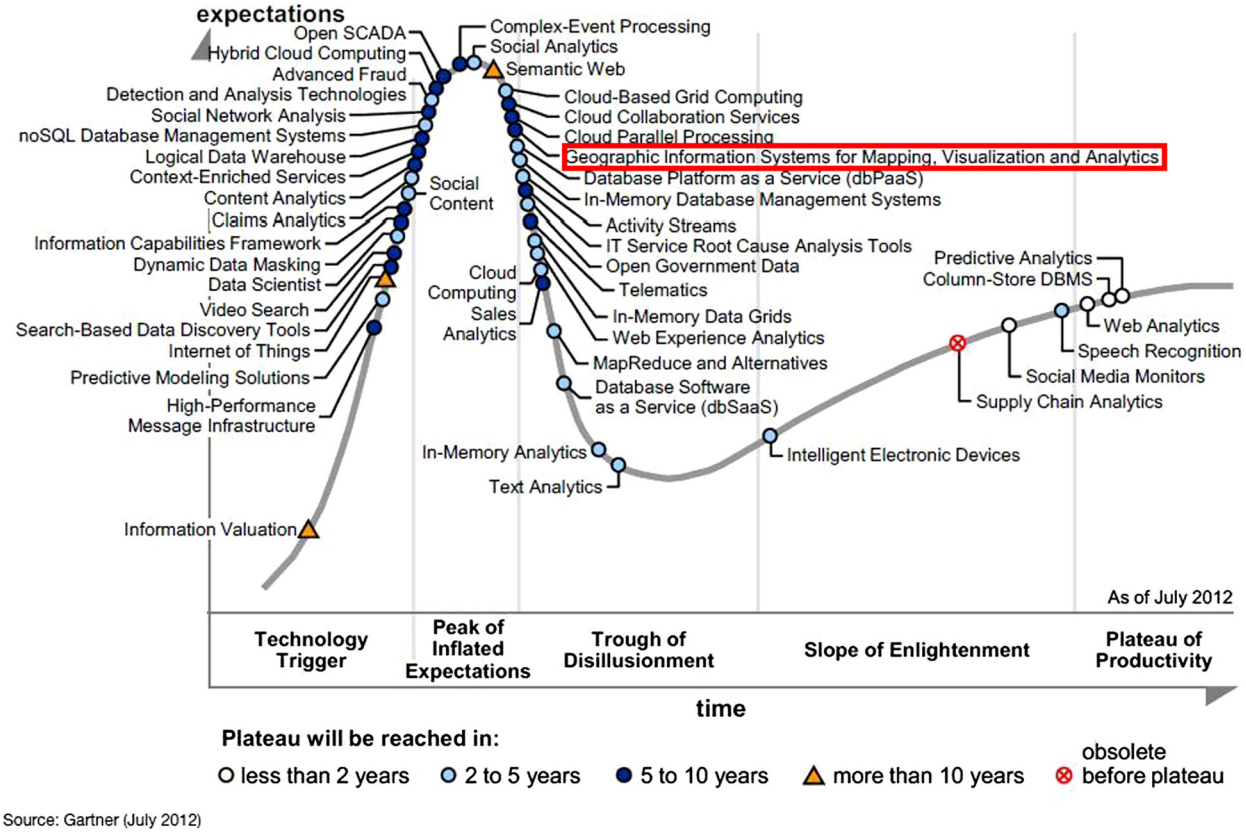


Fig. 2. Gartner's hype cycle (source: Gartner) [10].

fuel consumption in some of their vehicles [15]. Now, in many Ford cars, we can find the eco-route option, as in Fig. 4.

McKinsey Global Institute conducted a study on how big data can innovate our world [16]. As for geospatial big data, the study says "the use of personal location data could save consumers worldwide more than $600 billion annually by 2020." One can find out users' current locations by tracking their mobile devices such as smart phones. The study mentioned geosocial networking services such as Foursquare used for locating friends and for finding nearby stores and restaurants, where many users check-in at various places and reveal their current location [17]. On the other hand, according to the study, the biggest consumer benefit will be obtained from time and fuel saving thanks to location-based services that, by taking account of real-time traffic and weather data, help driver avoid traffic congestion and recommend alternative routes. Location tracking can be done by using a driver's smart phone or a global positioning system (GPS) equipped with a car.

2. Power of location

Sir Martin Sorrell [18], the CEO of WPP Group, says "Location targeting is holy grail for marketers." Big data analytics is an effec-

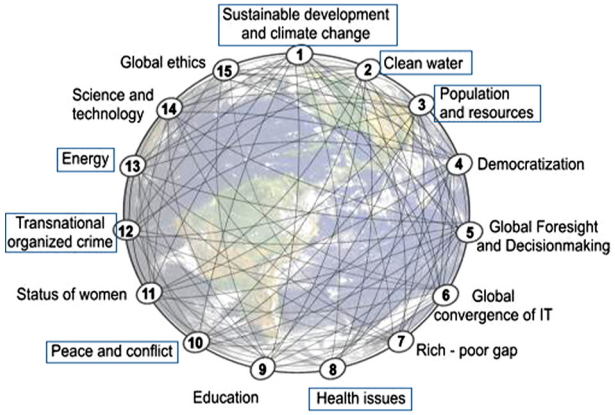


Fig. 3. 15 global challenges facing humanity [13].

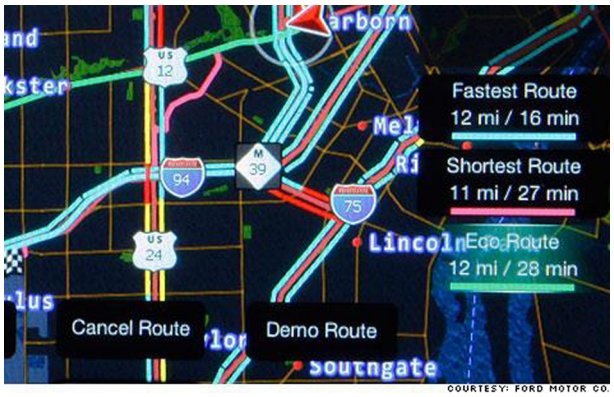


Fig. 4. Eco-route option in navigation systems.

tive way to enhance the power of location [19]. For example, video rental services of Netflix can benefit from analyzing rental patterns of the regions designated by zip codes [20]. In Fig. 5, which is the result of this study, Netflix generated the data on the top-50 rentals in 2009 in each zip code. Titles were listed in the approximate order of popularity across selected metropolitan areas. The most rented movies in a specific region are significantly different from those in another region. Such patterns are very useful for recommending movies to their users. In addition, the future location of a human can be predicted by analyzing his/her records of previous traces. Song et al. [21,22], using the mobile-phone traces of three million persons, showed that human traces are barely random. That is, human behavior is unlikely to significantly deviate from his/her daily routine behavior in most days, thus being highly predictable. Analysis of human mobility can boost many applications ranging from epidemic modeling to traffic prediction and urban planning.

Another example that can benefit from analysis of human mobility is direct marketing. This field is supported by Tobler's first law of geography [23]: "Everything is related to everything else, but near things are more related than distant things." This law was also validated by the analysis of Wikipedia [24]. That is, suggesting the services or stores close to the current location of a user should be more effective than suggesting those far-away. Location functionality has been combined with social networking, news, information, search, and entertainment services, and it is estimated that almost 800 million location-based service users exist worldwide as of 2012 [25]. Thus, businesses are willing to spend their money on the power of location, expecting larger revenue. Gartner predicts that consumer location-based services will generate revenue of $13.5 billion in 2015, where advertising becomes the main contributor [19]. Location targeting is known to improve the performance of mobile advertising by over 200%. Nowadays, many geosocial networking services such as Foursquare, when a user connects to the services, are providing those places close to his/her current location, as shown in Fig. 6.

Let's think about an interesting story reported on New York Times in 2004 [26], which shows the benefit of the analysis of geospatial data. Hurricane Frances was approaching Florida's Atlantic coast across the Caribbean. The executives at Walmart decided to adopt one of big data technologies—predictive analytics [27]. Linda M. Dillman, Walmart's chief information officer, asked her staff to predict what would happen soon based on what had happened when Hurricane Charley landed several weeks ago. By analyzing the transaction records stored in Walmart's data warehouse, the company could predict which items were bought just before or after an event (*i.e.*, a hurricane) at a *specific region*. People who had lived in Florida's Atlantic coast did not increasingly buy some products directly related to hurricanes, *e.g.*, water and flash lights. Surprisingly, strawberry PopTarts increased in sales, by seven times compared with their usual sale rate, just before a hurricane. In addition, the top-selling item immediately before the hurricane was beer. This kind of predictive analytics can be used for reducing the cost of maintaining the inventory and shipping items between warehouses [28].

3. Data collection

There are several forms of geospatial big data. Traditionally, geospatial data can be categorized into three forms: raster data, vector data, and graph data [14]. First, the raster data include geo-images typically obtained by unmanned aerial vehicles, security cameras, and satellites. Recently, the military is collecting huge amounts of raster data by utilizing drones, and the satellites keep providing us with the remote sensing data of the Earth. Table 1 shows some of the climate and earth system data stored at the Earth System Grid (ESG) portal. The raster data is being provided by digital map services, *e.g.*, Google Earth. Data analysts extract the tracks of moving objects or useful features from these raster data. Representative use cases include life pattern mining and change detection. Second, the vector data consists of points, lines, and polygons. The map data belongs to this form, and there are various data sources. For example, points can be collected through check-in's on Foursquare, and lines and polygons correspond to roads in OpenStreetMap,[2] which is a collaborative project to create a free editable map of the world. Representative use cases include detection of hot spots and spatial correlation patterns. Third, the graph data mainly appears in the form of road networks. Here, an edge represents a road segment, and a node represents an intersection or a landmark. The trajectories of vehicles on the road network are represented by sequences of road segments (edges) [29].

With the advancements of sensor and communication technologies, new sources of geospatial big data are emerging [11,30]. First, sensors (or sensor networks) become more prevalent in these days. The examples are loop detectors for detecting traffic in roads, electrical grids, environmental sensors for measuring air quality, and so on. These sensors are usually connected through wired or wireless communications and form sensor networks. Second, mobile devices become almost ubiquitous in these days. Smart phones can be used for tracking the trajectories of persons very easily. Especially, since the capacity of batteries and the efficiency of application processors have improved significantly, it becomes possible to record the location of a person very frequently and to

[2] http://www.openstreetmap.org/.

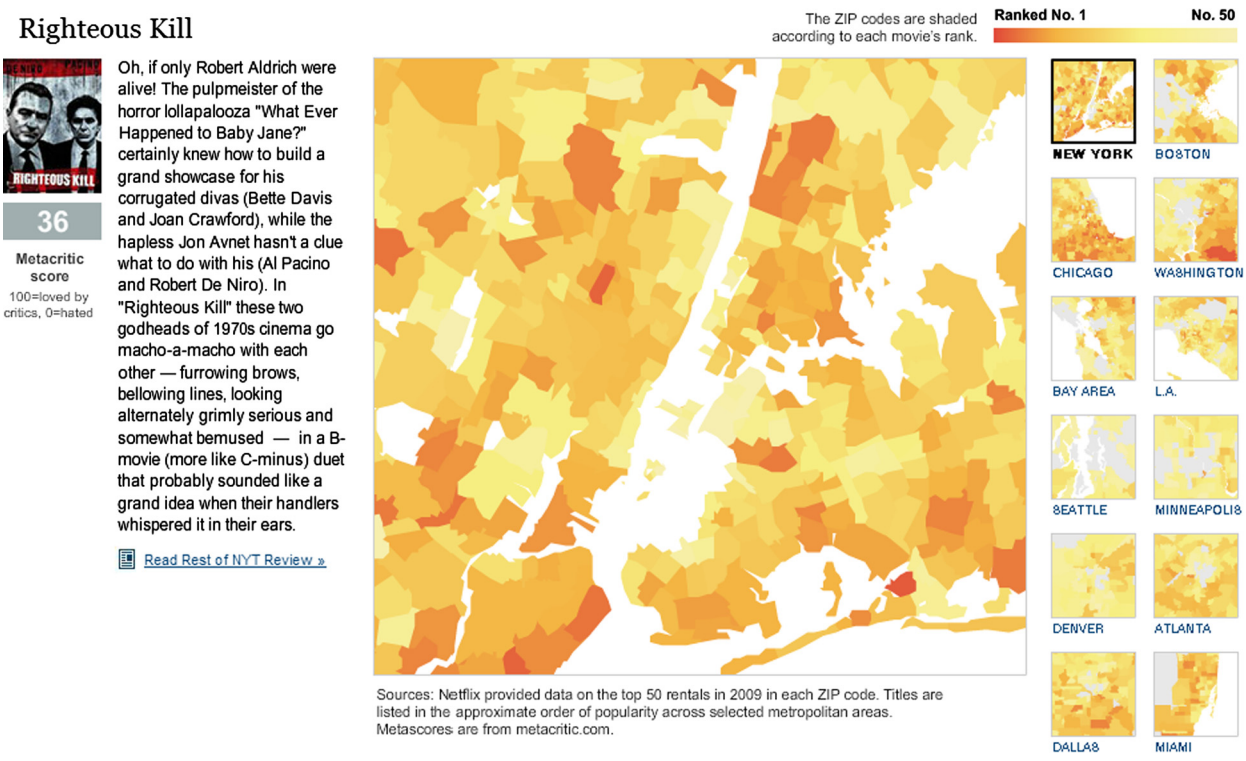


Fig. 5. Analysis of rental patterns based on regions by Netflix [20].

Table 1
ESG integrated data archive [2].

	CMIP5	ARM	DACC
Sponsor	SciDAC	DOE/BER	NASA
Description of data	40+ models	Atmospheric processes and cloud dynamics	Biogeochemical dynamics, FLUXNET
Archive size	~ 6 PB	~ 200 PB	~ 1 TB
Year started	2010	1991	1993

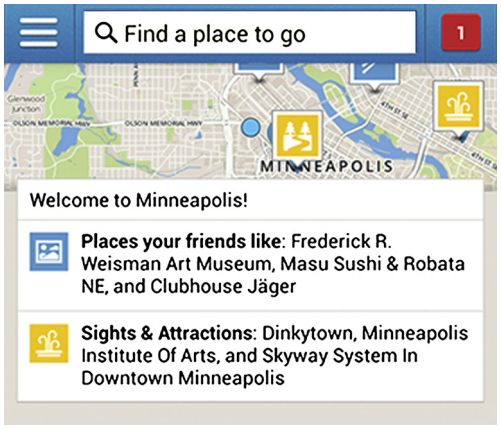


Fig. 6. Foursquare Android app with recommendations.

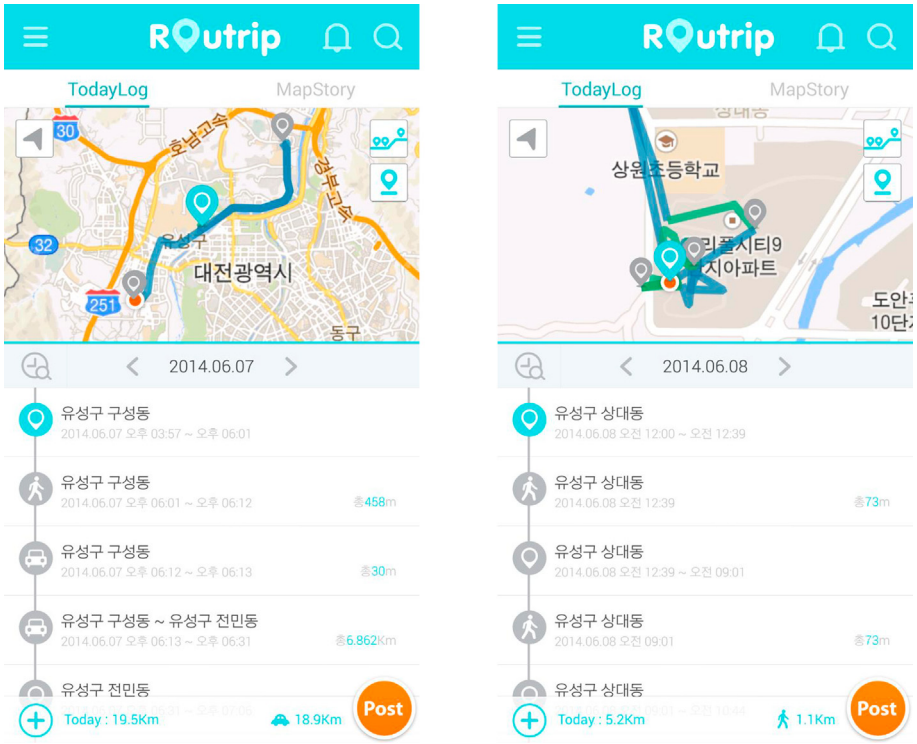
record his/her entire daily life on his/her smart phone. For example, Routrip,[3] an Android application which is being developed by Samsung Electronics, records the track of each person and shows his/her history in the form of the timeline, as shown in Fig. 7. Interestingly, by calculating movement speed, the transportation mode (*e.g.*, by car and by foot) is displayed as well. In Fig. 7b, the blue lines represent the movements by car, and the green lines represent those by foot.

As we discussed, the progress and innovation is no longer hindered by the ability to collect data. The most important issue is how we exploit these geospatial big data [7–9]. Toward this direction, there have been some efforts to share the collected geospatial data such that the data can be used by many other researchers. Movebank[4] is a free online infrastructure initiated to help researchers manage, share, analyze, and archive animal movement data, which has been hosted by the Max Planck Institute for Ornithology. The database of Movebank has been designed primarily for data sets that include consecutive locations of individual animals, which are generally called *tracking data*. The current status of MoveBank, as of November 2013, is as follows: 970 studies, over 250 contributors, 335 taxa, 41,170 tracks, and 61 million locations. Fig. 8a shows the home page of Movebank, and a small map in the upper-right corner shows the locations where data were collected so far. By selecting a circle on the map, as in Fig. 8b, one can investigate and download the data set.

The geospatial data collected from new emerging sources have the 3V's properties—volume, velocity, and variety—as the traditional big data have. Especially, these new types of geospatial data are being received continuously at a very high speed. Thus, the property "velocity" should be considered more importantly. For these data, instead of storing them in a data warehouse and analyzing them later in a batch, we need to look at the incoming data on the fly and make decisions in time. That is, we need to

[3] http://routrip.co, under beta testing as of September 2014.

[4] http://www.movebank.org/.



(a) A daily route. (b) Part of a daily route.

Fig. 7. The Routrip service by Samsung Electronics. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

pay more attention on *interactive* or *dynamic* analysis on geospatial big data, in order to better support this new trend.

4. On-going efforts

We recently initiated a new research project on spatial big data, supported by the Ministry of Land, Infrastructure, and Transport of the Korean Government. This project is planned for five years, and the outcomes will be integrated into the public services for Korean citizens. Fig. 9 shows the entire system architecture we are planning. The system consists of three layers: geospatial big data integration & management, geospatial big data analytics, and geospatial big data service platform. The first layer is responsible for quickly storing, retrieving, indexing, and searching geospatial big data. The second layer is responsible for performing data analytics on the data. This layer is further decomposed into the module of interactive analytics for real-time or dynamic data and the module of batch analytics for static or archived data. For interactive analytics, we have identified two main components: complex event processing (CEP) and spatial online analytical processing (SOLAP).

Fig. 10 shows how the interactive analytics module works in more detail. Geospatial big data are coming from various sources including satellites, drones, vehicles, geosocial networking services, mobile devices, and cameras. A cluster of commodity computers receive these data and perform filtering and preprocessing in parallel. Then, data streams generated by the cluster are propagated to the spatial CEP engine. The complex events to detect are defined using the continuous query language (CQL), and the continuous queries are stored in a rule database. The spatial CEP engine processes multiple continuous queries on multiple data streams on the fly. If a complex event is detected, the result is notified to the user. On the other hand, the incoming data are stored in a geospatial data warehouse after extract, transform, and load (ETL). The spatial OLAP engine can perform further analysis by conducting several OLAP operations such as roll-up, drill-down, and aggregation on the geospatial data warehouse. Especially, the result is visualized on various kinds of maps, such that a user can easily grasp the result of each operation. The user is likely to continue with other operations until he/she obtains the intended knowledge.

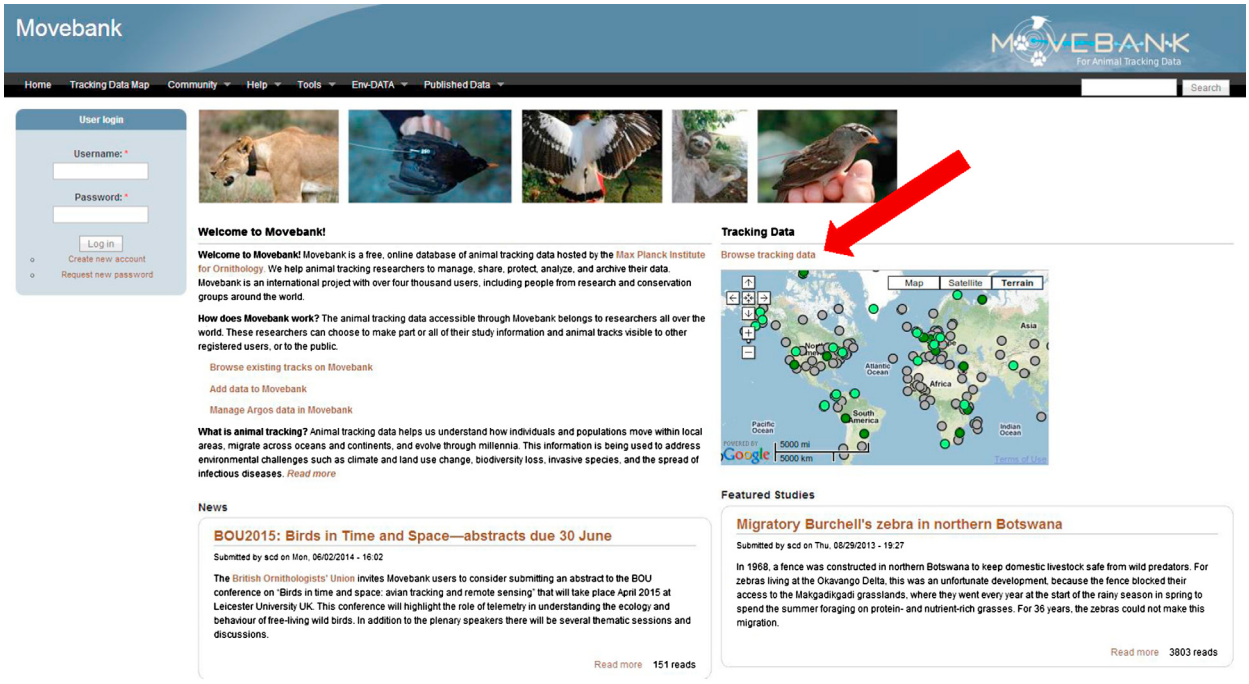
4.1. Complex event processing

Event processing is a method of tracking and analyzing data streams about events (*i.e.*, what is happening) in order to draw a conclusion from them [32]. *Complex event processing (CEP)* is event processing that involves multiple data streams to infer events or patterns that imply more complex situations [33]. The most representative products include Oracle CEP [34] and Esper.[5] Since the CEP engines are inherently designed to support data streams coming at a high speed, they can support big data to some extent. For example, Esper processes over 500,000 events per second on a dual CPU 2 GHz Intel-based hardware, with engine latency below 3 microseconds on average [35]. However, most of the existing engines do not support distributed, parallel processing, which is required to improve scalability and lower latency. To the best of our knowledge, the only product that supports distributed, parallel processing is Interstage Big Data CEP Server by FUJITSU. In addition, the existing engines do not support geospatial features. That is, the CQL for defining complex events is not capable of specifying spatio-temporal conditions such as nearest neighbor and range conditions. Thus, we plan to extend these existing CEP engines for both incorporating geospatial functionality and further improving scalability as well as latency.

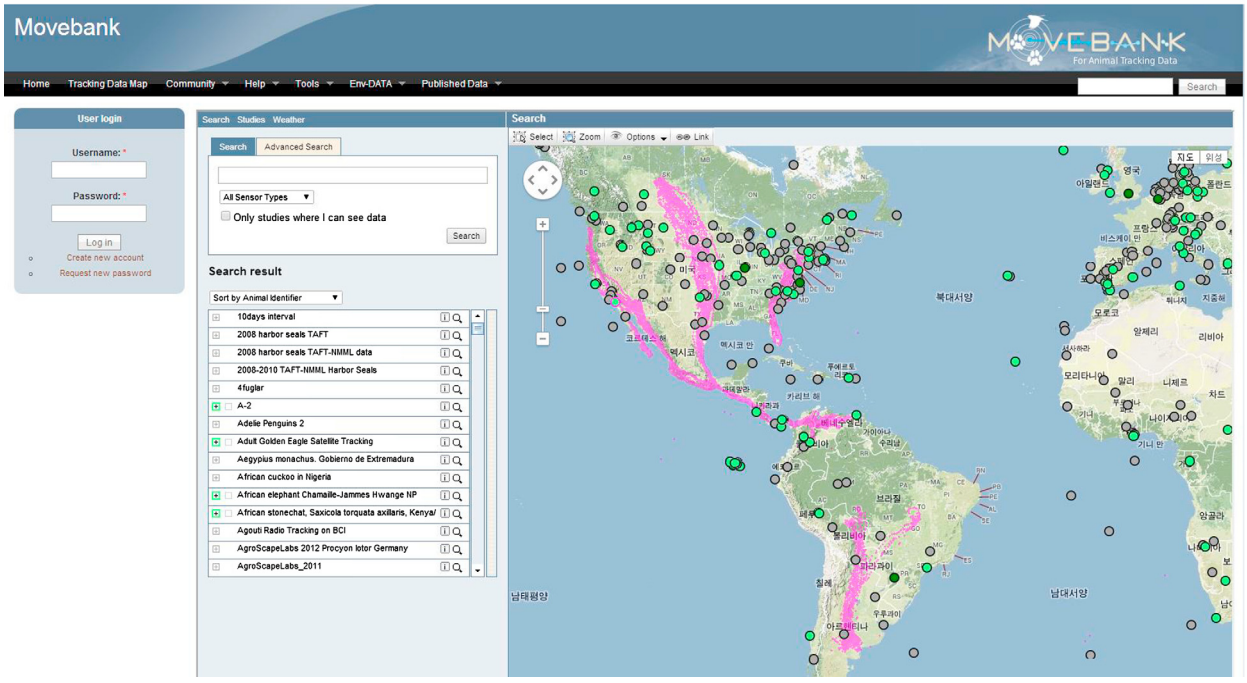
4.2. Spatial online analytical processing

Spatial OLAP is a visual platform for analyzing spatio-temporal data quickly and easily as well as for exploring the data in a

[5] http://esper.codehaus.org/.



(a) Home page.



(b) Data set page.

Fig. 8. The Movebank service.

multi-dimensional way with aggregation levels being available on cartographic displays and in tabular and diagram displays. The concept of Spatial OLAP, which tries to combine geographic information systems (GIS) and OLAP, was invented in late 1990's. However, even though a few research prototypes were released, no commercial product has never become widely available. One notable research prototype is JMap,[6] but it is no longer maintained since November 2009. As far as we know, the only product which is currently active is GeoMondrian.[7] It is an open-source SOLAP engine, *i.e.*, a spatially-enabled version of the Mondrian OLAP engine. Since GeoMondrian basically uses PostGIS as its data warehouse,[8] it has limitations in supporting geospatial "big" data. Our approach is first to extend Hadoop for geospatial data and then to utilize the Spatial Hadoop as our base platform. There have been some research efforts to extend the original Hadoop for geospa-

[6] http://www.spatialbi.com/.

[7] http://www.spatialytics.org/projects/geomondrian/.

[8] A commercial version can support Oracle Spatial, Microsoft SQL Server, and MySQL.

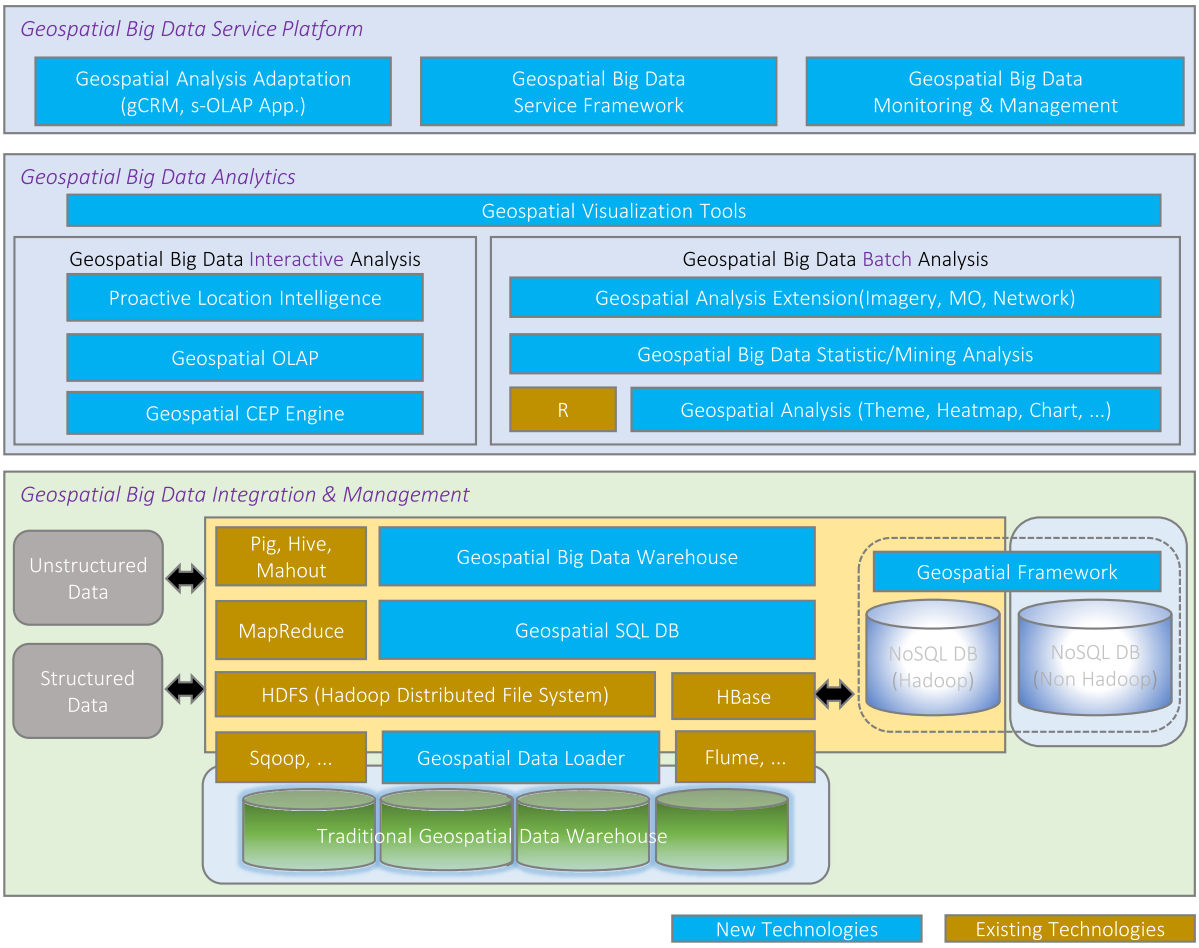


Fig. 9. The entire system architecture of our project [31].

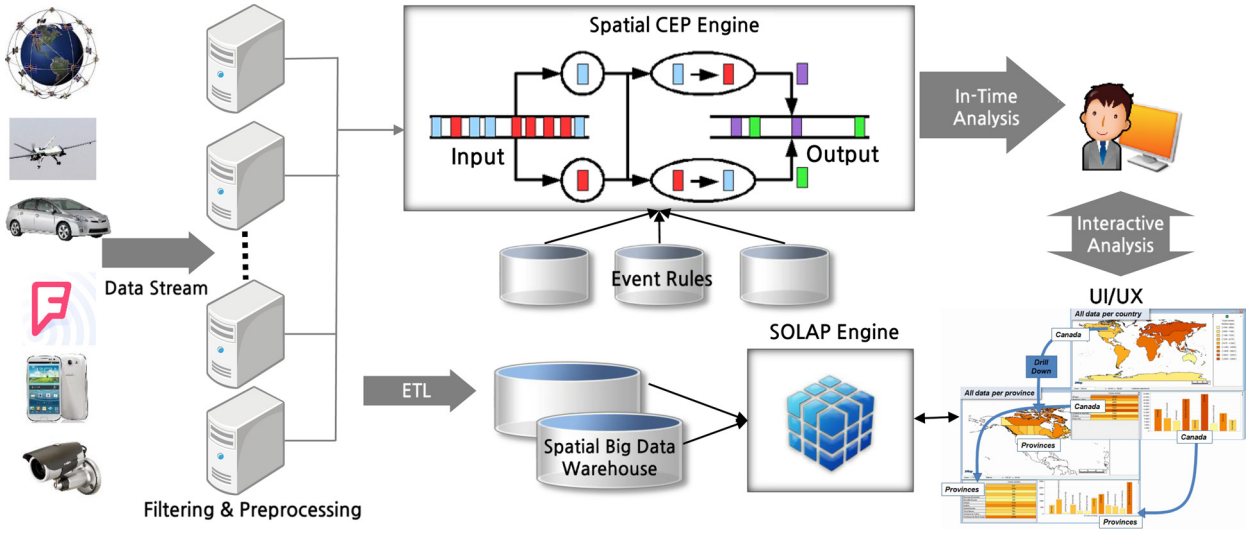


Fig. 10. The interactive analytics module.

tial data, including SpatialHadoop[9] in University of Minnesota [36]. To the best of our knowledge, there is no SOLAP engine that supports truly big data, and we need to work toward this direction.

[9] http://spatialhadoop.cs.umn.edu/.

5. Conclusion

In this paper, we have discussed the challenges and opportunities which geospatial big data brought us. Many evidences have witnessed that a significant portion of big data is, in fact, geospatial big data. We can innovate our daily life and business by exploiting the power of location embedded in geospatial big data.

A few cases are introduced to show the real benefits of geospatial big data. The collection of geospatial big data becomes pretty easy thanks to the advancements of sensor and communication technologies. A few agencies are devoting their efforts to the development of the platforms for sharing the collected data. The recent geospatial big data is usually being generated at a very high speed. This, in our new research project, we are concentrating on interactive analytics for real-time or dynamic data. In more detail, the two components, the spatial CEP and OLAP engines, are being developed by our KAIST team for upcoming five years.

Acknowledgements

This research, "Geospatial Big Data Management, Analysis and Service Platform Technology Development," was supported by the MOLIT (The Ministry of Land, Infrastructure and Transport), Korea, under the national spatial information research program supervised by the KAIA (Korea Agency for Infrastructure Technology Advancement) (14NSIP-B091011-01).

References

[1] A. Dasgupta, Big data: the future is in analytics, http://www.geospatialworld.net/Magazine/MArticleView.aspx?aid=30512, Apr. 2013, Geospatial World.
[2] 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 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, Redondo Beach, CA, 2012, pp. 1–10.
[3] M. Meeker, 2012 KPCB internet trends year-end update, http://www.slideshare.net/kleinerperkins/2012-kpcb-internet-trends-yearend-update, Dec. 2012.
[4] T. White, Hadoop: The Definitive Guide, 3rd edition, Yahoo Press, 2012.
[5] A. Thusoo, J.S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, R. Murthyin, Hive: a warehousing solution over a map-reduce framework, Proc. VLDB Endow. 2 (2) (2009) 1626–1629.
[6] E. Plugge, P. Membrey, T. Hawkins, The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing, 1st edition, Apress, 2010.
[7] J.-G. Lee, J. Han, K.-Y. Whang, Trajectory clustering: a partition-and-group framework, in: Proceedings of 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, 2007, pp. 593–604.
[8] J.-G. Lee, J. Han, X. Li, Trajectory outlier detection: a partition-and-detect framework, in: Proceedings of 24th International Conference on Data Engineering, Cancun, Mexico, 2008, pp. 140–149.
[9] J.-G. Lee, J. Han, X. Li, H. Gonzalez, TraClass: trajectory classification using hierarchical region-based and trajectory-based clustering, Proc. VLDB Endow. 1 (1) (2008) 1081–1094.
[10] A. Lapkin, Hype cycle for big data, 2012, http://www.gartner.com/document/2100215, Jul. 2012.
[11] N. Eagle, K. Greene, Reality Mining: Using Big Data to Engineer a Better World, 1st edition, The MIT Press, 2014.
[12] V. Mayer-Schonberger, K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Eamon Dolan/Houghton Mifflin Harcourt, 2014.
[13] M. Marien, Global challenges for humanity, http://www.millennium-project.org/millennium/challenges.html, accessed: 2014-08-30.
[14] S. Shekhar, Spatial big data challenges, in: Keynote at ARO/NSF Workshop on Big Data at Large: Applications and Algorithms, Durham, NC, 2012.
[15] P. Valdes-Dapena, GPS systems that save gas, http://money.cnn.com/2011/03/03/autos/navigation_gps_fuel_economy/, Mar. 2011, CNN Money.
[16] S. Lohr, New ways to exploit raw data may bring surge of innovation, a study says, http://www.nytimes.com/2011/05/13/technology/13data.html?_r=0, May 2011, The New York Times.
[17] M. Choy, J.-G. Lee, G. Gweon, D. Kim, Glaucus: exploiting the wisdom of crowds for location-based queries in mobile environments, in: Proceedings of 8th International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI, 2014, pp. 61–70.
[18] S.M. Sorrell, The power of apps, in: The 2011 GSMA Mobile World Congress, Feb. 2011, http://www.youtube.com/watch?v=5gfTQUq0mHw.
[19] G. Percivall, The power of location, http://www.opengeospatial.org/blog/1817, Apr. 2013, Open Geospatial Consortium.
[20] M. Bloch, A. Cox, J.C. McGinty, K. Quealy, A peek into Netflix queues, http://www.nytimes.com/interactive/2010/01/10/nyregion/20100110-netflix-map.html, Jan. 2010, The New York Times.
[21] C. Song, Z. Qu, N. Blumm, A.-L. Barabasi, Limits of predictability in human mobility, Science 327 (5968) (2010) 1018–1021.
[22] C. Song, T. Koren, P. Wang, A.-L. Barabasi, Modelling the scaling properties of human mobility, Nat. Phys. 6 (10) (2010) 818–823.
[23] W. Tobler, A computer movie simulating urban growth in the Detroit region, Econ. Geogr. 46 (2) (1970) 234–240.
[24] B. Hecht, E. Moxley, Terabytes of Tobler: evaluating the first law in a massive, domain-neutral representation of world knowledge, in: Proceedings of 9th International Conference on Spatial Information Theory, Aber Wrac'h, France, 2009, pp. 88–105.
[25] N. Brandweiner, Gartner outlines mobile services to reach mainstream by 2014, http://www.mycustomer.com/topic/technology/gartner-outlines-mobile-services-reach-mainstream-2014/158276, Oct. 2012, My Customer.
[26] C.L. Hays, What Wal-Mart knows about customers' habits, http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html?_r=0, Nov. 2004, The New York Times.
[27] E. Siegel, T.H. Davenport, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, 1st edition, Wiley, 2013.
[28] F. Provost, T. Fawcett, Data Science for Business: What you Need to Know About Data Mining and Data-Analytic Thinking, 1st edition, O'Reilly Media, 2013.
[29] J.-G. Lee, J. Han, X. Li, H. Cheng, Mining discriminative patterns for classifying trajectories on road networks, IEEE Trans. Knowl. Data Eng. 23 (5) (2011) 713–726.
[30] S. Madden, Going big on spatial data: a mobile systems perspective, in: Keynote at 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, 2012.
[31] Y. Kim, et al., Report on technology development for big data analysis and use based on spatial information, Tech. rep., Korea Agency for Infrastructure Technology Advancement, Nov. 2013.
[32] D.C. Luckham, Event Processing for Business: Organizing the Real-Time Enterprise, 1st edition, Wiley, 2011.
[33] I. Schmerken, Deciphering the myths around complex event processing, http://www.wallstreetandtech.com/latency/deciphering-the-myths-around-complex-event-processing/d/d-id/1259489, May 2008, Information Week.
[34] Oracle Product Management and Development Teams, Oracle complex event processing: lightweight modular application event stream processing in the real world, Tech. rep., Oracle Corporation, Jun. 2009, http://www.oracle.com/technetwork/middleware/complex-event-processing/overview/oracle-37.pdf.
[35] Esper Team, EsperTech Inc., Performance in Esper reference, Tech. rep., EsperTech Inc., Apr. 2012, http://esper.codehaus.org/esper-4.6.0/doc/reference/en-US/html/performance.html.
[36] A. Eldawy, M.F. Mokbel, A demonstration of SpatailHadoop: an efficient mapreduce framework for spatial data, Proc. VLDB Endow. 6 (12) (2013) 1230–1233.