ISSN: 0974-1011
Vol. 13, No.1, Jan-Mar 2020

# A Novel Approach to Develop Big Data Analytics Architecture for Spatial Data

**Sneha V. Hadole**[1]
[1] Research Scholar,
Post Graduate Dept. of
Computer Science,

**Dr. D. P. Kale**[2]
[2]Asso. Prof. and Head
Dept of CSE,
STC's School of
Engineering, Shegaon.
dpkale2015@gmail.com

**Dr. S. S. Sherekar**[3]
[3]Professor,
Post Graduate Dept. of
Computer Science,
SGBAU, Amravati.

**Dr. V. M. Thakare**[4]
[4] Professor and Head
Post Graduate Dept. of
Computer Science,
SGBAU, Amravati.
vilthakare@yahoo.co.in

***Abstract*-The ubiquity of complex data generated from different data sources has evolved into a new technology called Big Data Analytics. These are associated with some combination of data volume, data velocity, and data variety that may include complex analytics and complex data types. Big Data Analytics has shown its potential in many application domains. Big data is characterized by Volume, Variety, Velocity, Veracity, and Value. Huge quantity of data is generated and collected such as real-world applications, scientific applications, Global Positioning System (GPS) enabled devices, scientific applications, observations from satellites, government agencies, etc from multiple data sources. These massively distributed data are equipped with spatial attributes and characteristics. Spatial data are complex and multi-dimensional. This research Paper focuses on to study an open source and scalable big-data analytics architecture for massive scale data management together with spatial data. The hundreds of millions of users demand efficient analytics on massively distributed data with low latency. Also focuses on relevant analytics of real-life datasets in the agriculture.**

***Keywords: Big Data Analytics, Spatial data, real-life datasets***

## I. INTRODUCTION

Now a day's user demands sophisticated, scalable, high speed, accessible and economic solutions to perform relevant analytics on difficult and scattered data together with spatial data. To meet the existing user's demand the conservative systems for data management are becoming less capable to scale out extensively due to limited computational power and storage. The Big-Data Analytics has shown its potential in many application domains with its different goals like discover cost-effective, design data products, find the insight based on historical data, locate valuable patterns and information, and reliable methods to extract social and economic values from terabytes (TB) and petabytes (PT) of data carrying out by research organizations and industries. It has accomplished success to fulfill the modern user's demands up to some extent.

## II. BIG-DATA AND BIG-DATA ANALYTICS

Big-data is a term associated with the new types of workloads and underlying technologies needed to solve business problems that we could not previously support due to technology limitations, prohibitive cost, or both [1]. Big data is characterized by 5 V's viz. Volume, Variety, Velocity, Veracity, and Value. Data volume can be measured by the number of transactions, events, and the amount of historical data. Data variety is the assortment of data. There can be a wider variety of data sources such as machine-generated data, social media data, weblogs, Internet data, geospatial data, and image data. Also huge amount of data generate a Data-driven applications from web and Internet. Data velocity is the speed at which data is accumulated, created, processed and ingested. Veracity points to the trustworthiness of the data. The quality and accuracy are less controllable. The value signifies the quantifiable outcome from the data.

Big-Data tools and technologies are developed to store and process massively distributed, batch or real-time data with a high degree of resource utilization. Regardless of conventional systems, big data tools and technologies have gained more popularity and success for massive scale data management as they are high-speed, highly scalable, and fault tolerant.

## III. BIG SPATIAL DATA

Spatial data are complex and multi-dimensional. Compared to non-spatial data, spatial data are characterized by special data types, complex operations and methods, and index structures. The specializing systems and architectures are required to manage the exponential growth of spatial data. Spatial data is mainly characterized by 3 V's: Volume, Variety, and Velocity. Big spatial data have acquired a major portion of big-data. Big spatial analytics has achieved significant attention in many application domains such as finance, retail and e-commerce, agriculture, and banking. The hundreds of millions of users demand efficient analytics on massively distributed data with low latency. The implementation and design of low latency query processing on the exponential growth of geospatial data have made geospatial analytics more challenging and complex for traditional systems. The state-of-the-art big data storage and processing frameworks such as Hadoop [2], Not only SQL (NoSQL) databases and Spark [3] are highly scalable, highly available, fault tolerant, and efficient to hold immense range geospatial data. However, these frameworks are providing very limited geo-functionality and Open Geospatial Consortium (OGC) [4] standard methods compared to traditional systems. Hence, it is highly desirable to take advantage of such frameworks to store and process geospatial data at scale.

## IV. Big Data Analytics in Agriculture

In developed countries and developing countries, the challenges related to big data application development is different. In the world of digitalization, Indian farmers are still considering fellow farmers, agro-retailers, various digital media channels, and agriculture experts as a preferred source of information and make decisions based on such informal sources. The information provided by extension services is perceived to be either biased (e.g. agro-marketing companies) or less actionable due to lack of consistency, accuracy, and personalization. Hence, farmers never reach near 100% production potential. Farmers have a dearth of information on weather, soil nutrients, optimal market price, availability of agro-products, and timely advice on pest or disease management.

### *A. 3.1 Spatial Analytics in Agriculture*

In the agriculture domain, voluminous and variety of disparate information is consumed and generated at a higher velocity. Information is available in the form of reports on weather, GPS mapping, soil conditions, fertilizer, water resources or field characteristics, pesticide usage, and commodity market conditions in agriculture. Geospatial processing, remote sensing, advanced analytics algorithms, cloud resources, and advanced storage systems has a huge potential to refer to this information and produce comprehensive insight via Big-data analytics.

### *B. 3.2 Spatial analytics applications in agriculture*

For developing multidisciplinary applications such as helping in planning, managing, and utilizing natural resources efficiently using spatial analysis to develop flexible and versatile functions and applications, Geospatial data is very important. Spatial data in agriculture can help farmers to enhance knowledge of farmland, increase value of farmland, and market potentials. The major contributors are location-based and geospatial data applications to big data.

There is an urgent need to manage spatial data at scale with specialized systems, techniques, and algorithms. A variety of MapReduce systems and cloud infrastructures (e.g., Hadoop, Hive [5], HBase [6], Impala [7], Dremel [8], Vertica [9], and Spark) are available for big data management. They do not provide any special support for spatial data. A number of comprehensive systems that support spatial data management efficiently are Map Reduce systems such as Hadoop-GIS [10], ESRI Tools for Hadoop [11], SpatialHadoop [12], parallel DB systems such as Parallel Secondo, and systems that use resilient distributed datasets (RDD) [13] such as SpatialSpark [14], GeoTrellis[15], and GeoSpark [16]. Effective use of such systems and infrastructures is always crucial in application development.

## V. Existing Agricultural Information Systems and Applications

It has reviewed the existing ICT based agricultural systems and applications developed using modern technologies such as web, mobile, GIS, Big Data, and Big Spatial Data. The Mobile-based applications [17] developed for the agriculture domains are depicted in Table 1.

TABLE I. Description of Mobile applications in agriculture

| Mobile Applications | Describe |
|---|---|
| CropInfo | India-specific android based application provides making practices of cultivable crops in Kannada & English Language. |
| KisanYogana | It provides useful information about Government's schemes to the farmers from their perspective. |
| Mkishan | It is an advisory and information system. |
| Shetkarimasik | It is a most accepted review magazine in the agriculture sector, under publication since 1965. It has been published by the Department of Agriculture, Maharashtra. |
| Farm-o-pedia | A mobile application developed for the rural community of Gujarat, INDIA |

### *A. 4.1 Web GIS Based Systems*

In agriculture the web-GIS based systems are reviewed [18,19,20,21,22] have demonstrated and implemented web-GIS based information systems for agriculture domain using traditional GIS tools and technologies. These technologies are often insufficient to give a complete representation of analytics in a geographic context. Table 2 shows existing web-GIS based applications in the agriculture domain.

TABLE II. Table 2. Web GIS based applications in agriculture

| Web GIS based Applications | Description |
|---|---|
| Agrifootprint | Web GIS based information system which facilitate small agriculture enterprises by delivering customized, comprehensive, and user friendly solutions using open source technology. |
| Kumar SK et al. | An open source web GIS-based decision support system to monitor and map sugarcane crop at farm level by using remote sensing and GIS at Medak district of Andhra Pradesh, India. |
| CropScape | An interactive web Cropland Data Layer (CDL) exploring system which is developed to analyze, query, disseminate, and visualize CDL data geospatially through standard geospatial web services in a publicly accessible online environment. |
| Z. Zhu et al. | A study of GIS-based agriculture expert-system which can provide decision-making services for soil fertilizer at Hebei farmland. |

| | |
|---|---|
| Zhang, H. et al. | Web based GIS system for generating online fertilization recommendation at village scale of Hua country. |

## B. *4.2 Big Data Applications*

It has reviewed the big data applications and systems that have been merged in the agriculture domain to manage complex data at scale. Table 3 shows the existing big-data systems and applications [23, 24, 25, 26, 27] developed for agriculture.

TABLE III. TABLE 3. BIG DATA SYSTEMS AND APPLICATIONS IN AGRICULTURE

| **Big Data Systems and Applications** | **Objective** | **Big Data Tools and Technology** |
|---|---|---|
| Garg, R. et al. | A structure for generating approval solutions for the paddy leaf blast in Punjab state, India. | Hadoop and Hive |
| S. Lamrhari et al. | A profile based architecture for precision agriculture to improve decision making in real time. | Dynamic big data service selection and composition methods are introduced. |
| Claudia Vitolo et al. | An overview of the current state-of-art for processing large and heterogeneous data sets of web based tools and technologies. | Most relevant web based environment data processing tools in the big-data period are investigated. |
| Peisker, A. et al. | A conceptual framework for an integrated analytics approach towards rural development. | No tools and technologies specified |
| Xie, N. F. et al. | A conceptual system for an agricultural information system hierarchy based on big data technology. | Scoop, Hadoop, Hive, Mahout, IBM Infosphere, Biginsight |
| Chalh, R. et al. | Conceptual architecture of big data open platform used for supporting water resource management. | No tools and technologies specified |

## VI. CONCLUSIONS

This paper has studied an open source and scalable architecture to manage massively distributed data as well as spatial data. In comparison with the proposed architecture is in-memory, the existing platforms and architectures, open source and cost-effective. The architecture intends to build and implement distributed and scalable APIs for spatial data organization on top of analytics and integrated infrastructure. The architecture is realized to develop analytical services for agriculture and provide customized solutions in the form of interactive maps and Restful ad-hoc services to the end users.

## References:

[1] Ferguson, Mike. "Architecting a big data platform for analytics."*A Whitepaper prepared for IBM* 30 (2012).

[2] Hadoop, Apache. "Hadoop."*2009-03-06. http://hadoop.apache.org*(2009).

[3] Zaharia, Matei, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. "Spark: Cluster computing with working sets." *HotCloud*10, no. 10-10 (2010): 95.

[4] Website of Open Geospatial Consortium, http://www.opengeospatial.org/.

[5] Thusoo, Ashish, JoydeepSen Sharma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. "Hive: a warehousing solution over a map-reduce framework." *Proceedings of the VLDB Endowment* 2, no. 2 (2009): 1626-1629.

[6] Dimiduk, Nick, AmandeepKhurana, and Mark Henry Ryan. *HBase in action*. Shelter Island: Manning, 2013.

[7] Bittorf, M. K. A. B. V., TarasBobrovytsky, C. C. A. C. J. Erickson, Martin Grund Daniel Hecht, M. J. I. J. L. Kuff, Dileep Kumar Alex Leblang, N. L. I. P. H. Robinson, David RorkeSilviusRus, John Russell DimitrisTsirogiannis Skye Wanderman, and Milne Michael Yoder. "Impala: A modern, open-source SQL engine for Hadoop." In *Proceedings of the 7th Biennial Conference on Innovative Data Systems Research*. 2015.

[8] Melnik, Sergey, AndreyGubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. "Dremel: interactive analysis of web-scale datasets." *Proceedings of the VLDB Endowment* 3, no. 1-2 (2010): 330-339.

[9] Lamb, Andrew, Matt Fuller, Ramakrishna Varadarajan, Nga Tran, Ben Vandiver, Lyric Doshi, and Chuck Bear. "The vertica analytic database: C-store 7 years later." *Proceedings of the VLDB Endowment* 5, no. 12 (2012): 1790-1801.

[10] Aji, Ablimit, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang, and Joel Saltz. "Hadoopgis: a high performance spatial data warehousing system over MapReduce." *Proceedings of the VLDB Endowment* 6, no. 11 (2013): 1009-1020.

[11] Esri, G. I. S. "Tools for Hadoop." (2015).

[12] Eldawy, Ahmed, and Mohamed F. Mokbel. "Spatialhadoop: A MapReduce framework for spatial data." In Data Engineering (ICDE), 2015 IEEE 31st International Conference on, pp. 1352-1363. IEEE, 2015.

[13] Zaharia, Matei, MosharafChowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, and Ion Stoica. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing." In Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, pp. 2-2.USENIX Association, 2012.

[14] Web site of Spatialspark, http://simin.me/projects/spatialspark/

[15] Kini, Ameet, and Rob Emanuele. "Geotrellis: Adding geospatial capabilities to spark." Spark Summit (2014).

[16] Yu, Jia, Jinxuan Wu, and Mohamed Sarwat. "Geospark: A cluster computing framework for processing large-scale spatial data." In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 70.ACM, 2015.

[17] Website of mobile applications, http://mkisan.gov.in/downloadmobileapps .aspx.

[18] De Oliveira, Tiago H. Moreira, Marco Painho, Vitor Santos, Otávio Sian, and André Barriguinha. "Development of an agricultural management information system based on Open-Source solutions." *Procedia Technology* 16 (2014): 342-354.

[19] Kumar SK, Babu SDB (2016) A Web GIS Based Decision Support System for Agriculture Crop Monitoring System-A Case Study from Part of Medak District. J Remote Sensing & GIS 5:177. doi: 10.4172/2469-4134.1000177.

[20] Han, Weiguo, Zhengwei Yang, Liping Di, and Richard Mueller. "CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support." *Computers and Electronics in Agriculture* 84 (2012): 111-123.

[21] Zhu, Zhiqing, Rongmei Zhang, and Jieli Sun. "Research on GIS-based agriculture expert system." In *Software Engineering, 2009.WCSE'09. WRI World Congress on*, vol. 3, pp. 252-255. IEEE, 2009.

[22] Zhang, Hao, Li Zhang, YannaRen, Juan Zhang, XinXu, Xinming Ma, and Zhongmin Lu. "Design and implementation of crop recommendation fertilization decision system based on WEBGIS at village scale." In *International Conference on Computer and Computing*

*Technologies in Agriculture*, pp. 357-364.Springer, Berlin, Heidelberg, 2010.

[23] 23] Garg, Raghu, and HimanshuAggarwal. "Big data analytics recommendation solutions for crop disease using Hive and HadoopPlatform."*Indian Journal of Science and Technology*9, no. 32 (2016).

[24] Lamrhari, Soumaya, Hamid Elghazi, TayebSadiki, and Abdellatif El Faker. "A profile-based Big data architecture for agricultural context." In *Electrical and Information Technologies (ICEIT), 2016 International Conference on*, pp. 22-27. IEEE, 2016.

[25] Vitolo, Claudia, YehiaElkhatib, DominikReusser, Christopher JA Macleod, and WouterBuytaert. "Web technologies for environmental Big Data."*EnvironmentalModelling& Software* 63 (2015): 185-198.

[26] Peisker, Anu, and Soumya Dalai. "Data analytics for rural development."*Indian Journal of Science and Technology* 8, no. S4 (2015): 50-60.

[27] Xie, N. F., X. F. Zhang, W. Sun, and X. N. Hao. "Research on Big Data Technology-Based Agricultural Information System."In*International Conference on Computer Information Systems and Industrial Applications.Atlantis Press*. 2015.