Developing Big Data Analytics Architecture for Spatial Data

Purnima Shah
Supervised by Sanjay Chaudhary
School of Engineering and Applied Science,
Ahmedabad University, India

purnima.shah@iet.ahduni.edu.in

ABSTRACT

The main goal of the research work is to solve a wide range of data problems by offering batch, iterative, and interactive computations in a unified architecture. The dissertation proposes an integrated architecture to manage a large amount of massively distributed data including spatial data. The implementation architecture has three major components: data preparation, data analytics, and data visualization. As a part of the implementation, a novel big spatial data analytics framework is developed to load, store, process, and query spatial and non-spatial data at scale. As proof of concept, spatial analytics applications are developed using agricultural real-life datasets.

1. INTRODUCTION

In the mobile and Internet era, massive scale data is generated from disparate sources with spatial components. Today's users demand high speed, scalable, sophisticated, economic, and accessible solutions to perform relevant analytics on complex and distributed data including spatial data. The conventional systems for data management are becoming less capable to scale out extensively to meet the current users' demand due to limited computational power and storage. The modern technologies like Big Data and Big Data Analytics (BDA) have a huge potential to handle massive scale data with high scalability and low latency. Though the modern big data management tools such as Not only SQL (NoSQL) databases, Hadoop [1], and Spark [2] are highly efficient, they offer limited functions and methods for spatial data management. In addition, in modern application development, only one specific big data tool would not be able to manage big data efficiently and effectively. Hence, it is highly enviable to exploit the potential features of big data tools and technologies and propose integrated frameworks and architectures built on top of more than one technology to develop robust and powerful applications including geospatial data.

Proceedings of the VLDB 2019 PhD Workshop, August 26th, 2019. Los Angeles, California.

2. REVIEW STATUS

The research reviewed existing databases, frameworks, and architectures for spatial data management.

2.1 Database technologies for spatial data

NoSQL databases such as Cassandra [3] do not offer native support for spatial data. As an exception, MongoDB [4] offers query operations on geospatial data with index support. Though MongoDB is the best suitable NoSQL database for geospatial data, it does not offer complex spatial operations like KNN search, spatial join, and KNN join. It also does not provide support for aggregated queries. Ben Brahim et al. [5] have developed a spatial extension for the Cassandra database to solve spatial range queries.

2.2 Big data computational frameworks for spatial data

The Big data computational frameworks such as Spark and Hadoop don't offer native support for spatial data. A number of extended systems have made important contributions to extend the functionality of Hadoop/Spark engine for spatial data management. These extension systems include parallel DB systems such as Parallel Secondo, MapReduce systems such as ESRI Tools for Hadoop [6], SpatialHadoop [7], Hadoop-GIS [8], and systems that use Resilient Distributed Datasets (RDD) [9] such as GeoTrellis [10], SpatialSpark [11], GeoSpark [12], Magellan [13], LocationSpark [14], and Spatial In-Memory Big Data Analytics (SIMBA) [15]. However, these frameworks are only able to execute spatial operations on datasets that are available in text-based file formats (CSV/GeoJSON/shapefiles and WKT), and stored in HDFS or local disk. There is no big data analytics framework available which reads data from the NoSQL database and performs spatial analytics on those data.

2.3 Big data architectures for spatial data

Generally, big data architectures are designed and developed to achieve a specific goal. Many big data architectures such as Lambda [16], Kappa [17], Liquid [18], BDAS [19], SMACK [20], and HPCC [21] have been developed on top of integrated infrastructures. However, there have been insufficient discussions about how these architectures perform spatial data management.

The research reviewed the existing platforms and architectures such as IBM PARIS [22], SMASH [23], and ORANGE [24] for spatial data management. In comparison with the existing big spatial data

architectures, the proposed architecture is in-memory and open source.

There are many big spatial data frameworks have been developed on top of big data stack composed of Spark and Cassandra. The Cassandra-Solr-Spark framework has been developed by Datastax to enable spatial query processing on top of the big data stack. The framework provides SQL like query interface to perform spatial operations. However, it does not support join operations. It has also not been evaluated based on the performance metric. P. Shah et al. [25, 26] have developed a big data analytics framework including geospatial data. The spatial analytics applications in the agriculture domain have been developed using third-party GeoSpark libraries. The major drawback of the framework is data duplication. GeoSpark is a spatial extension for spark which can only access data available in HDFS or local disk.

3. BIG DATA ANALYTICS ARCHITECTURE FOR SPATIAL DATA

The big data analytics architecture [26] is built and implemented on big data open source technologies for the enrichment of massive scale data including spatial data. The architecture is designed to provide scalable, flexible, extendible, and cost-effective solutions with available infrastructures and tools for agriculture. Architecture implementation has three components: data preparation, big spatial analytics, and data visualization. There are four types of user interaction with the architecture: 1) System developer, 2) Data scientist, 3) Domain expert, and 4) End users. The proposed architecture is shown in figure 1.

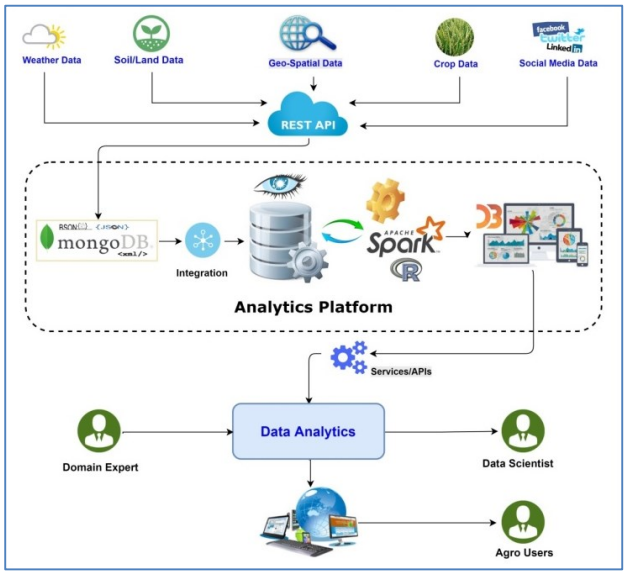

Figure 1 Big data analytics architecture

3.1 Data Preparation Framework

The data preparation framework is implemented to fetch consistent and clean data from disparate sources and store into a persistent database. It provides two layers of data abstraction. First, it hides all physical data sources from the data repository. Second, it further unifies the data available in a data repository using various complex tools and techniques such as data fusion algorithms, schema mapping tools, and record linkage algorithms. The implementation architecture for data preparation framework is shown in figure 2.

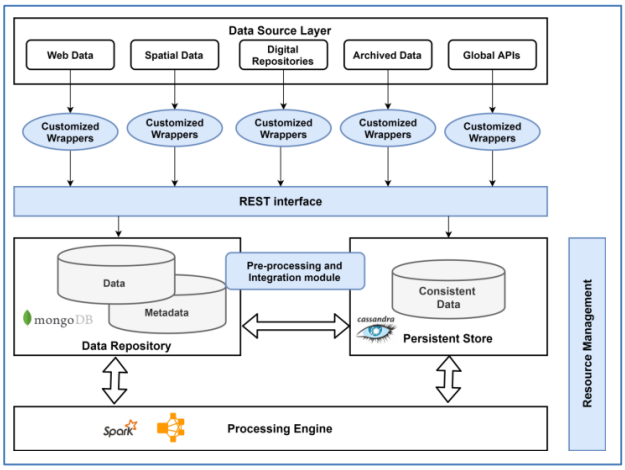

Figure 2 Implementation architecture for data preparation Framework

3.2 Big Spatial Data Analytics Framework

The main purpose of a big spatial data analytics framework [27] is to enable spatial data management on large scale data. The consistent datasets generated by data preparation services are processed and analyzed using data analytics framework. It is an integrated infrastructure designed to manage spatial data efficiently and effectively by exploiting the potential features provided by the standard storage and processing big data frameworks. It is realized on top of big data stack with Spark as a core processing engine and Cassandra as a data storage. The big spatial data analytics framework is shown in figure 3.

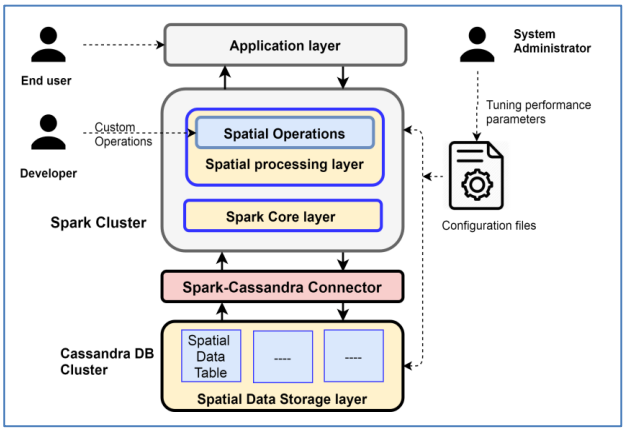

Figure 3 Big spatial data analytics framework

A NoSQL i.e. Cassandra based spatial data storage framework is built and implemented. The framework offers distributed and scalable APIs for spatial operations such as location search, proximity search, and KNN search.

The framework provides a convenient web-based REST interface to the end user. Cassandra performs fast data retrieval based on partition key and clustering key compared to Spark. The application architecture facilitates end users to execute ad-hoc queries on a suitable framework either Spark or Cassandra via a common user interface. The low latency queries are executed on Cassandra,

whereas complex queries (e.g. aggregated and spatial queries) are executed on the Spark framework. The analytical results are explored to end users through visualization and REST interface.

The performance of the framework is evaluated in terms of latency against the variable size of data. The performance of the framework is compared with the baseline technology, i.e. Cassandra for low latency queries.

3.3 Data Visualization Framework

Data visualization makes complex data more accessible, understandable, and usable. The implementation architecture provides a web-based user interface by developing analytical and visualization services through Restful ad-hoc APIs and interactive maps.

The Data visualization framework [26] is implemented to showcase the analytical results with dynamic layouts. A dashboard application is designed and implemented to depict the analytical results in the agriculture domain.

4. REALIZATION OF ARCHITECTURE IN APPLICATION DOMAIN

The challenges related to big data application development in agriculture is different in developed countries and developing countries. In developing countries, the major barriers for big data application development in agriculture are lack of tools, infrastructures, data standards, semantics, integrated data models, developers APIs, unified access points for public and private data, technical expertise, and finally the data.

The prototype applications in the agriculture domain are developed on top of the big data analytics architecture. Spatial and non-spatial data on weather, crop, and market are collected from different sources like meteorological departments, agriculture universities, and web portals. The summary of data collection is given in Table 1. The snapshots of the dashboard results are shown in figure 4 and 5.

Table 1 Data Collection

Datasets	Data source	Data format
Weather data for seven districts of Gujarat (1992 – 2007)	Archived data, www.Indiastat.com	Spreadsheet/ document
Crop data for cotton crop for eighteen districts of Gujarat (1960 – 2007)	www.Indiastat.com, http://apy.dacnet.nic.in/crop_fryr_toyr.aspx, Archived data	Spreadsheet/ document
Market data for 429 agro-markets in Gujarat	http://agmarknet.gov.in/	PDF document
Spatial data for Gujarat	www.diva-gis.org	Shapefile

5. Implementation Status – Present and Future

The big data analytics architecture is designed and developed for massive scale data management including spatial data. The architecture is developed to address two big data challenges: Variety and Volume. The data preparation framework is designed with two levels of data abstraction. As a part of implementation, the REST interface is designed and implemented to fetch and collect data from different data sources with different formats such as PDF, spreadsheets, documents, web pages, and online services. The integration of data collected through the REST interface is the most critical module in data preparation framework. An algorithmic solution is to be devised to link a variety of data from diverse sources in aid of the unified search, query, and analysis.

The core component of big data analytics architecture, i.e. big data analytics framework is implemented for spatial data management. The framework is to be extended by developing complex spatial operations like spatial join and kNN join. The spatial applications like spatial aggregation and spatial auto-correlation are to be developed on top of the framework. The complex applications in the agriculture domain are to be developed by identifying new data sources, formats, and data types. The real-life datasets including real-time and streaming data are to be collected and stored in a data repository to perform further analytics. The near real-time data analytics and visualization algorithms are to be devised to process real-time data like weather, disaster, etc. The analytical services like rainfall prediction, crop recommendation, crop price prediction, agro-inputs procurement, supply chain management, crop disease alert, fertilizer recommendations, etc. are to be implemented. These services can be used to generate customized and multilingual solutions in the form of weather-based crop calendar and alerts based on adverse events.

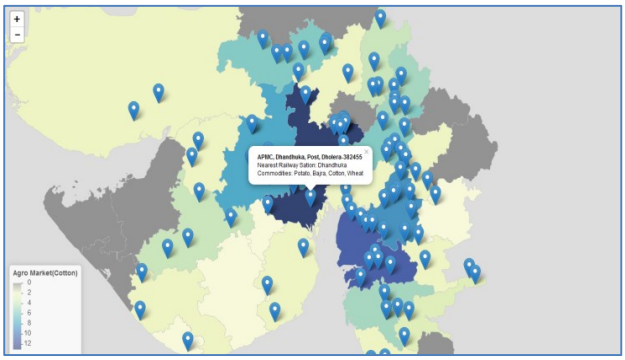

Figure 4 Snapshot of agro-market search analysis

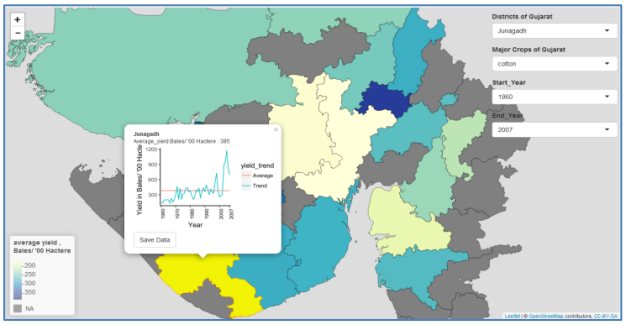

Figure 5 Snapshot of crop yield data aggregation

6. ACKNOWLEDGMENTS

I would like to acknowledge and thank my Ph.D. thesis supervisor Dr. Sanjay Chaudhary, for his excellent guidance, constant encouragement, patience, care, and support. This work is a part of a research project on 'Developing Data Analytics Architecture,

Applications in Agriculture', funded by NRDMS and NSDI, Department of Science and Technology, Government of India.

7. REFERENCES

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

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

[3] Lakshman, Avinash, and Prashant Malik. "Cassandra: a decentralized structured storage system." ACM SIGOPS Operating Systems Review 44, no. 2 (2010): 35-40.

[4] Website of MongoDB, http://www.mongodb.org.

[5] Brahim, Mohamed Ben, Wassim Drira, Fethi Filali, and Noureddine Hamdi. "Spatial data extension for Cassandra NoSQL database." Journal of Big Data 3, no. 1 (2016): 11.

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

[7] 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.

[8] 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. \

[9] Zaharia, Matei, Mosharaf Chowdhury, 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.

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

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

[12] 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.

[13] Web site of Magellan. https://github.com/harsha2010/magellan. Magellan - https://hortonworks.com/blog/magellan-geospatial-analytics-in-spark/; https://github.com/harsha2010/magellan.

[14] Tang, Mingjie, Yongyang Yu, Qutaibah M. Malluhi, Mourad Ouzzani, and Walid G. Aref. "Locationspark: a distributed in-memory data management system for big spatial data." Proceedings of the VLDB Endowment 9, no. 13 (2016): 1565-1568.

[15] Xie, Dong, Feifei Li, Bin Yao, Gefei Li, Liang Zhou, and Minyi Guo. "Simba: Efficient in-memory spatial analytics." In Proceedings of the 2016 International Conference on Management of Data, pp. 1071-1085. ACM, 2016.

[16] "Lambda Architecture," http://lambda-architecture.net/, 2014.

[17] "Kappa Architecture," http://radar.oreilly.com/2014/07/ questioning-the-lambda-architecture.html, 2014.

[18] Fernandez, Raul Castro, Peter R. Pietzuch, Jay Kreps, NehaNarkhede, Jun Rao, Joel Koshy, Dong Lin, Chris Riccomini, and Guozhang Wang. "Liquid: Unifying Nearline and Offline Big Data Integration." In CIDR. 2015.

[19] Website of Berkeley Data Analysis Stack. , [https://amplab.cs.berkeley.edu/software/]

[20] Website of HPCC, https://hpccsystems.com/.

[21] Estrada, Raul, and Isaac Ruiz. "Big Data SMACK." Apress, Berkeley, CA (2016).

[22] Klein, Levente J., Fernando J. Marianno, Conrad M. Albrecht, Marcus Freitag, Siyuan Lu, Nigel Hinds, Xiaoyan Shao, Sergio Bermudez Rodriguez, and Hendrik F. Hamann. "PAIRS: A scalable geo-spatial data analytics platform." In Big Data (Big Data), 2015 IEEE International Conference on, pp. 1290-1298. IEEE, 2015.

[23] Sinnott, Richard O., Luca Morandini, and Siqi Wu. "SMASH: A cloud-based architecture for big data processing and visualization of traffic data." In Data Science and Data Intensive Systems (DSDIS), 2015 IEEE International Conference on, pp. 53-60. IEEE, 2015.

[24] S. Cho, S. Hong, and C. Lee, "ORANGE: Spatial big data analysis platform," In Big Data (Big Data), 2016 IEEE International Conference on, pp. 3963-3965. IEEE, 2016.

[25] Shah, Purnima, Deepak Hiremath, and Sanjay Chaudhary. "Big data analytics architecture for agro advisory system." In High Performance Computing Workshops (HiPCW), 2016 IEEE 23rd International Conference on, pp. 43-49. IEEE, 2016.

[26] Shah, Purnima, Deepak Hiremath, and Sanjay Chaudhary. "Towards development of spark based agricultural information system including geo-spatial data." In Big Data (Big Data), 2017 IEEE International Conference on, pp. 3476-3481. IEEE, 2017.

[27] Shah, Purnima, and Sanjay Chaudhary. "Big Data Analytics Framework for Spatial Data." In International Conference on Big Data Analytics, pp. 250-265. Springer, Cham, 2018.