ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 3, Issue 11, DOI: 10.29027/IJIRASE.v3.i11.2020, 560-563, May 2020

Comparative study of Spatial Hadoop and Geospark for Geospatial Big Data Analysis

Jayati Gandhi[1], Nekita Chavhan[2], Girish Kumar[3]
Department of Computer Science and Engineering[1,2]
G H Raisoni College of Engineering,Nagpur,India
Scientist/Engineer[3]
RRSC-ISR0(Central),Nagpur
gandhi_jayati.ghrcemtechcse@raisoni.net[1], nekita.chavan@raisoni.net[2], girish.isro@gmail.com[3]

***Abstract*— Earth Observation (EO) is constantly producing large amount of Geospatial data over the last few years which is used in resource monitoring, protection of environment and disaster predictions. The applications like Ground surveying, remote sensing and mobile mapping produces geo-spatial data. The growth of EO data has been a challenge in recent approaches for data management and processing. For Big data scenario Geospatial data are the major contributors. There are various tools for analysis of big data that can support large amount of geospatial big data. The main aim of this paper is to do the comparative analysis of Spatial Hadoop and GeoSpark which are the most popular open source geospatial big data analytical tools and they can be used for observing and processing the geospatial big data in an accurate way. The architectural view of Spatial Hadoop and GeoSpark are also compared in this paper.**

Keywords*— *Earth Observation, Big Data, Geospatial Data, Spatial Hadoop , Geospark

I. INTRODUCTION

Earth Observation is done for the following activities such as gathering, managing, processing, observing and representing the physical, chemical and biological information related to earth system by using various techniques of remote sensing. It has tremendous applications for environmental monitoring [1]. Earth Observation satellite produces regular stream of multi-spectral, multi-resolution, multi temporal remote sensing images due to the development of sensor technologies. Many platforms like planes, satellites and vehicles have been used as sensor carriers to collect different data of earth and generates large amount of data types [2]. In the past decade the Geospatial technologies have developed in different ways and such modifications have possibly developed from the enhanced difficulty of data acquisition. In today's technoholic world the volunteered Geographic Information (VGI), regular use of satellite images of very high resolution , aerial imagery and sensor streams etc are provided by social media which consist different data of Internet of Things (IoT).Hencit is a proof that we manage and process the geospatial big data on a regular basis. Where as "big data" is a developing term and the very popular definitions usually indicates a situation in which managing the huge amount of data is a very complicated task [3]. For understanding and observing geospatial big data platforms there are some characteristics of Big data which are studied first.

A. Characteristics Of Big Data

1. Volume-Data volume is the information generated by the records of science and training, human and business connection records. Data volume plays an important role in storage and handling [4].

2. Value- In decision finalizing the value of data is used. Value of data is very important characteristics in Big Data. Value of data provides large amount of benefits in a business.

3. Velocity-Velocity of data means the velocity at which the data is delivered. For picking up the data Stream handling is used because the methods which are going on constantly is time manifesting and requests faster and closer results of investigations. In a group Hadoop efficiently process recorded information but Apache Spark is more efficient for regular and repetitive job results of investigation [5][6].

4. Veracity- Veracity describes the Data quality and accuracy.When there is a need to take decision about the gathered data, Veracity plays an important role. Its main aim is how to make data reliable. Because of data inconsistency and incompleteness data is categorized as good, bad and undefined data [7].

5. Viscosity- Data Viscosity is very hard to manage for huge datasets if they come from different sources, because correlating, similarity and conversion are very important work[8]. Complications of Big Data mainly analyse the association degree and some dependencies in big data which means that minute modifications can cause a large effect or no change in the behavior of system [9][10].

6. Variability- Variability in data occurs because of inconsistent and discontinous flow of data. Because of Inconsistency there is difficulty in accessing data. This property is very challenging as it has increased the usage of digital media [11].

7. Volatility- Volume, Variety, and Velocity of Big Data are the terms which is defined in Volatility. Volatility is defined as upto how much long time the data is stored and up to how much long time the data is considered as valid. That is why there is a need to check the validity of data. So there is a need to set up protocols for efficient running of work processes in Big Data research environment.

8. Viability- Due to Viability the Big data should have the potential to be live and active constantly, and should be capable enough to create, and to give more information when required.

9. Validity- Authenticity of information is same as that of accuracy of information. When the information status changes from exploratory to significant then information is considerable. Index of information might not have any accuracy issues but rather may not be likewise authenticate if they are not correctly valid. Validity is the foundation to explore the closeness of hidden connections among various components inside the huge Big Data age sources .

10. Variety- The level of association of information is called as Data assortment. The satisfactory level of association is required by Unstructured information and on the other hand the structure association has a huge degree [12].Organized database along with information design can be effortlessly deal with instruments of database. A Framework for the Examination of Conventional Information is used by the Relational Data Base Management System (RDBMS). But this traditional RDBMSs needed costly tools and then applied to organized information [13].

B. Spatial Hadoop

Inorder to compute huge amount of dataset , Spatial Hadoop is used to perform distributed storage. It is an open source network which is written in Java . Powerful computers were used in the traditional methods to process these large amount of data but they were not efficient and could not cope up with the large and regular growth of data. By the help of Distributed computing Hadoop overcome this complexity because it is considered as scalable. In order to achieve the distributed storage Hadoop divides the huge data into smaller parts. The processing is also divided into chunks and subtasking of each single node is also performed in Spatial Hadoop. In the end all nodes results are merged to produce the final result. Thus parallel and distributed processing is achieved in this way.

C. GeoSpark

To observe and work on Big Data spark is considered as an efficient tool. Apache Spark provides both batch processing and real time processing that is why maximum companies are using it in today's world. Driver Program and Worker Program are the two types of programs in spark. Driver Program is run on Master and on Slave, Worker Program is run. Spark is very flexible because along with its own cluster manager it also gets connected to another sources such as Hadoop and Apache Memos. Geospark is merged with Hadoop because both are considered to have the same family for example we are running Hadoop and and we want to shift to Spark then we can run Spark on Hadoop's top thus all the transferring of data can be saved to spark from HDFS . Geospark can also be merged with Open Stack Swift, Amazon S3 and Apache Cassandra.

II. LITERATURE SURVEY

Geospatial data is tremendously used on the platforms of big data. Because of different data acquisition methods there is a quick increase in the handling of data and computing abilities of different geospatial platforms. Hence it is a proof that methods to deal with the geospatial big data have attracted special attention. In the past years different platforms have developed like Spatial Hadoop and GeoSpark. Developed at the Minnesota University , Spatial Hadoop uses powerful mapping method of Hadoop by replacing their spatial counterparts with record reader, file splitter and indices. The performance of Spatial Hadoop with huge raster data was given by Eldawy and Mokbel [14]. In order to build a spatial index of two level for the environment of Map Reduce, Spatial Hadoop uses spatial index structures, Grid, R-tree and R+- tree . For implementing spatial operation Spatial Hadoop has three basic operations and they are range query, kNN, and spatial join. For spatial joins it is observed that the triple performance spatial Hadoop range queries and queries of k nearest-neighbor has larger throughput than Hadoop for datasets up to 4.6TB . GeoSpark implements the spatial resilient distributed datasets (SRDD) in the Spark distributed memory to allow the spatial query processing. There are three layers in Geospark such as Apache Spark , Spatial RDD and Spatial Query Processing Layer. Basic functionalities of Spark are given by Apache Spark layer which include loading and storing of data into the disk and also provides continous operations of RDD. The three unique Spatial Resilient Distributed Datasets (SRDDs) is present in Spatial RDD layer that increased regular Apache For supporting spatial and geometrical objects, Spark RDDs is used. For performing basic operations of Geometry like intersection and overlapping GeoSpark provides the library of geometrical operations which can accesses Spatial RDDs. After loading from the storage system Geo Spark along with SRDDs automatically stores these datasets into memory, thus it has an advantage where it can refer to these SRDDs from memory various times without doing any modification of data and loading of data in repititive jobs like spatial co-location. Geo Spark has better run time performance as compared to Map Reduce based counterparts according to Yu et al.

Hagedorn et al has studied the characteristics and performance estimation between the frameworks of big spatial data and found some queries were not present in GeoSpark. Similarly, there are some limitations on the platform of GeoSpark [15]. The above mentioned studies T were some kinds of spatial implementations. For processing geospatial data on Hadoop there are various options like oracle spatial and ESRI spatial framework but they are not so flexible.

III. Architecture View of Spatial Hadoop and Geospark

Spatial

A. Architecture View of Spatial Hadoop

Hadoop is completely developed Map reduce framework which provide help for huge geospatial data and it removes the shortcomings of Hadoop-GIS. Spatial Hadoop architecture has four layers which are as follows- [16].

1. Language Layer-Spatial Hadoop uses Pigeon language. An addition of pig is pigeon and due to this language only Spatial Hadoop can add geospatial data types, functions and operations.
2. Operation Layer- The Range query, KNN and geospatial join etc are some kinds of geospatial operations which is performed by Spatial Hadoop. Some more operations like KNN join, RNN etc are also added in this layer
3. Map Reduce Layer-Map reduce program are run in this layer. The input files that are supported by Spatial Hadoop are usually a geospatially indexed. As compared to traditional Hadoop, Map reduce layer is better because the two main components are added in it which are as follows -
   3.1 Geospatial File Splitter: File Splitter is an extension of it and is used in the system of Hadoop.
   3.2 Geospatial Record Reader: The slitted file which is originating from input files can be read by it. So local index are used for efficient processing of files.

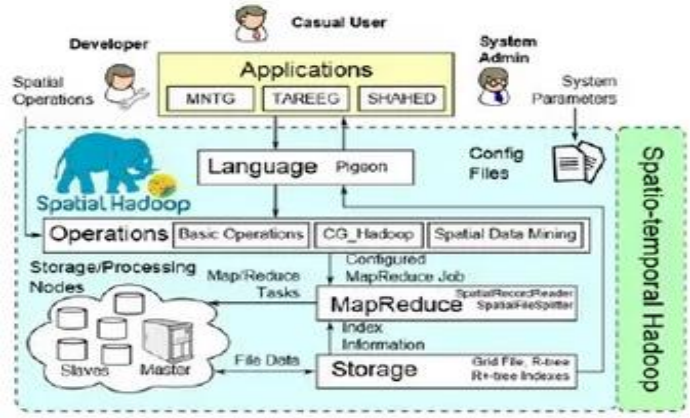

Fig1 Architecture of Spatial Hadoop

B. Architecture View of Geospark

Geospark is called as an in-memory cluster computing system for handling large amount of Geospatial data. Apache Spark is an extension of Geospark because it supports geospatial data types, indexes and operations.

There are three layers in the architecture of Geospark and which are as follows:

1. Apache Spark Layer- The loading and querying of the data is performed in this layer and it also contains all the components that are present in this layer.
2. Geospatial Resilient Distributed Dataset Layer- Spark is extended by this layer and in this layer there are three types of RDD and they are Point, Rectangle and Polygon RDD. For every RDD , it contains the library for various geometrical operations.
3. Geospatial Query Processing Layer- The different queries of Geospatial are performed in it.

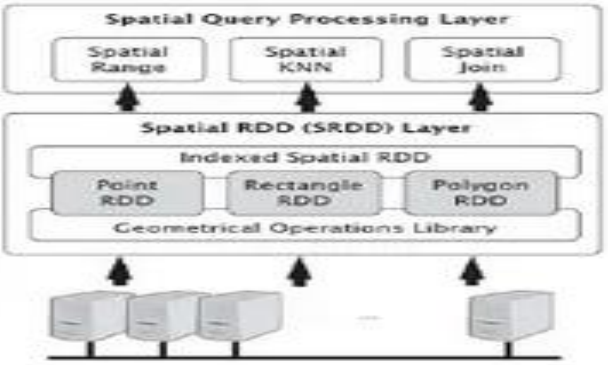

Fig 2:Architecture View Of Geospark

IV. Conclusion

In this paper, the two crucial geospatial big data analytical tools is discussed i.e. Spatial Hadoop and GeoSpark for processing of Geospatial data. The architectural comparative analysis is also carried out for these two crucial open source tools which manage the geospatial big data efficiently. Both GeoSpark and Spatial Hadoop are very flexible to handle the geospatial data but GeoSpark is much more efficient than Spatial Hadoop because it is very fast for real time processing of geospatial big data. For better predictive results there is a plan to implement it in future for disaster management and mitigation and also for the management of geospatial health information and infrastructure.

References

[1] Guo, H.; Liu, Z.; Jiang, H.; Wang, C.; Liu, J.; Liang, D. "Earth big data: A new challenge and opportunity for Digital Earth's development". Int. J. Digit. Earth 2017, 10, 1–12.

[2] Di, L.; Moe, K.; van Zyl, T.L." An overview of Earth observation sensor web". IEEE J. Sel. Top. Appl. Earth Observation Remote Sens. 2010, 3, 415–417.

[3] Mike, Loukides. What is data science. [Online] June 2, 2010.[Cited: September 5, 2018.] www.oreilly.com/ideas/what-is-data-science.

[4] Nada Elgendy and Ahmed Elragal. "Big data analytics: a literature review paper" In Industrial Conference on Data Mining, pages 214–227. Springer, 2014.

[5] Soulmaz Salehian and Yong hong Yan "Comparison of spark resource managers and distributed file systems. In Big Data and Cloud Computing

(BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) , 2016 IEEE International Conferences on, pages 567–572. IEEE, 2016.

[6] N.Khan, M. Alsaqer "The 10 Vs, Issues and Challenges of Big Data", ICBDE '18, March 9–11, 2018, Honolulu, HI, USA © 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-6358-7/18/03…$15.00 https://doi.org/10.1145/3206157.3206166

[7] S Mills, S Lucas, L Irakliotis, M Rappa, T Carlson, and B, Perlowitz. Demystifying big data: a practical guide to transforming the business of government. Tech America Foundation, Washington, 2012

[8] Abdullah Gani, Aisha Siddiqa, Shahaboddin Shamshirband, and Fariza Hanum. "A survey on indexing techniques for big data: Taxonomy and performance evaluation", Knowledge and Information Systems, 46(2):241–284, 2016.

[9] Mokbel, Ahmed Eldawy and Mohamed F.SpatialHadoop: A MapReduce Framework for Spatial DataSeoul. IEEE Conference on Data Engineering ICDE South Korea, 2016.

[10] Stephen Kaisler, Frank Armour, J Alberto Espinosa, and William Money. "Big data: Issues and challenges moving forward", In System Sciences (HICSS), 2013 46th Hawaii International Conference on, pages 995–1004.IEEE, 2013.

[11] AvitaKatal, Mohammad Wazid, and RH Goudar. "Big data: issues, challenges, tools and good practices." In Contemporary Computing (IC3), 2013 Sixth International Conference on, pages 404–409.IEEE, 2013.

[12] Peter Geczy. Big data characteristics. The Macrotheme Review, 3(6):94– 104, 2014.

[13] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. Mobile Networks and Applications, 19(2):171–209, 2014.

[14] Basel Kayyali, David Knott, and Steve Van Kuiken" The big-data revolution in us health care: Accelerating value and innovation" Mc Kinsey & Company, 2(8):1–13, 2013.

[15] Soulmaz Salehian and Yong hong Yan "Comparison of spark resource managers and distributed file systems. 2016 IEEE International Conferences on, pages 567–572. IEEE, 2016.

[16] Eldawy, Ahmed, "Spatial Hadoop: towards flexible and scalable spatial processing using mapreduce, " Proceedings of the 2014 SIGMOD PhD symposium, 2014, pp. 46-50