BIG DATA ERA IN GIS

A MINI PROJECT REPORT SUBMITTED TO BHARATHIAR UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

MASTER OF SCIENCE IN DATA ANALYTICS

Submitted by

NANDHINI K

20CSEG20

Under the guidance of

Dr.J RAMSINGH, MCA., M.Phil., Ph.D.,

Department of Computer Applications

DEPARTMENT OF COMPUTER APPLICATIONS

BHARATHIAR UNIVERSITY

COIMBATORE – 641 046

February-2022

DECLARATION

I hereby declare that this Mini project Report titled **“BIG DATA ERA IN GIS”** submitted to the Department of Computer Applications, Bharathiar University is the record of original work done by **NANDHINI K (20CSEG20)** under the guidance of **Dr. J. RAMSINGH , MCA., M.Phil,. Ph.D.,** Department of Computer Application, Bharathiar University and this project work has not formed the basis for the award of any Degree/Diploma/Associate ship/ Fellowship or similar title to any candidate of any University.

Place: Coimbatore

Date: Signature of the Candidate

Countersigned by

Dr.J RAMSINGH, MCA., M.Phil., Ph.D.,

Department of Computer Applications

ACKNOWLEDGEMENT

I thank the Almighty God who showered his grace to make this project a grand success. The success and final outcome of this study required a lot of guidance and assistant from many people and I am extremely fortunate to have got those people, all along the completion of my project work.

I respect and thank our **Dr. Prof. T.Devi M.C.A, M.Phil.,Ph.D.(UK)**,Professor & Head, Department of Computer Applications, Bharathiar University, Coimbatore for giving me an opportunity to do the project work.

I really deem a special privilege **Dr. J RAMSINGH, MCA., M.Phil., Ph.D**., Department of Computer Applications, Bharathiar University, for his valuable guidance and interest on my project work, till the completion of the work by providing the necessary information for developing a good report.

My heartfelt thanks to my dear Parents and my dear friends for their support and encouragement for the successful completion of this project. I am highly obliged to those who have helped me directly and indirectly in making this project a successful one.

ABSTRACT

The project is entitled as **“BIG DATA ERA IN GIS”**. Geographic information system (GIS) is designed to generate maps, manage spatial datasets, perform sophisticated “what if” spatial analyses, visualize multiple spatial datasets simultaneously, and solve location-based queries. The impact of big data is in every industry, including the GIS. spatial data volume, being generated from a wide variety of sources, the need for efficient storage, retrieval, processing and analyzing of spatial data is ever more important. Hence, spatial data processing system has become an important field of research. In recent times a number of Big Spatial Data systems have been proposed by researchers around the world. Spatial Data or Geo spatial data are increasing day by day generating huge volume of data from satellite images providing details related to orbit and from other sources for representing natural resources like water bodies, forest covers, soil quality monitoring etc. This paper gives the idea to understand GIS data representation, data sets, data sources to provide better understanding of GIS for analyzing the data using Big Spatial Data.

TABLE OF CONTENTS

1. INTRODUCTION:

GIS has advanced as a new emerging field as enhancement of communication technologies. The rapid growth of data, information, and communication has generated voluminous data of earth surface. Just of increasing powerful remote sensors, more computing power and enhancement in GIS technologies themselves, GIS is developing platform of selection for integrating and analyzing massive amount of earth data. Today’s GIS is become indispensable and used in multidisciplinary areas to access information with respect to position. Enormous amount of GIS data is collected in numerical, text, graphics and analogues forms from satellite imagery sensors and other devices which represent the spatial and temporal situation. Spatial data mining has been emerging as innovative research area for data analysis with respect to spatial relations. SDM techniques has strong relationship with GIS and widely used in GIS for inferring association among spatial attributes, clustering, and classifying information with respect to spatial attributes. This paper gives the idea to understand GIS data models, data sets, data sources to provide better understanding of GIS for analyzing the data using spatial data mining techniques.

2. BIG DATA:

According to some authors there have been big-bang explosion of data in recent years. The data acquisitions at the rates of terabytes per day are quite common. The spatial data on mobile phone users, rail and air travelers, marketing, consumers, goods production and wide range of daily activities are producing large volume of data which are of interest for Data Mining. The remote sensing data are no exception to these. A large number of meteorological, land and ocean observations sensors on board satellites are continuously down pouring data from space. The analyses of such large data present their own challenges, even though with highly powerful processors and high speed data access possible presently.

One of the characteristics of this volume of data generated is their formats. The data may be structured, semi-structured and un-structured formats, which further complicate their analysis. The Big Data data are sets whose size (volume), complexity (variability), and rate of growth (velocity) make them complex to be collected, managed, processed or analyzed by current technologies and tools .The Big Data analysis require ingenious steps for data analysis, characterized by its three main components: variety, velocity and volume**.**

3. GIS:

Geographical Information System-GIS acts as information systems for importing, storing, analyzing, managing, exporting and presenting spatial referenced data(linked to location) . Because of the technological advancement in automated data acquisition in GIS domain has generated voluminous geographic data to represent spatial nomenclature of earth surface. GIS data received from heterogeneous discipline such as high resolution remote sensors, global positioning systems (GPS), location aware services and surveys etc. GIS database stored geo-reference spatially related dataset received from aforesaid heterogeneous components connected to each other and provide the spatial information about location, links with other, and description of nonspatial (attribute) features. Spatially located dataset gives the information about to what, when and where. GIS has powered significance in analysis related to knowledge management and data mining, It is a collection of components with respect to position namely Data, Software, Hardware, Procedures and Methods for analysis and decision making as shown in Fig.1.

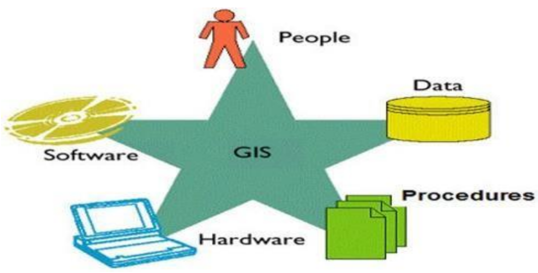

Fig.1. Component of GIS

The Big Data approach to GIS allows analysis and decision making from huge datasets, by using algorithms, query processing and spatiotemporal data mining. In simple words, this means extracting information from maximum possible sources using established procedures and computational technique. This section throws the light on GIS data source, data formats, trends and applications in GIS.

3.1 DATA SOURCE:

GIS dataset are of two types-spatial dataset and attribute dataset. Spatial data is about to where while attribute data is about to what. An entity contains the information about both spatial and attributes data. For example a spatial point entity is represented in GIS by georeferenced location using latitude and longitude and related to what type of information contain about that point, is described by nonspatial attribute data. The geodatabase is used to store a collection of geographical datasets having three elements-space, theme, and time. The object is characterized in the form of line, point and polygon, pertaining to object in the database. GIS objects are uniquely identified by latitude and longitude. Spatial data can store, retrieve and manipulate with GIS. The various available dataset of GIS are listed in Table1 to summarize details.

Table 1 GIS Data Source

Source of Data	Description
Esri Open Data Hub	Download formats are in spreadsheet, KML,shapefile. API's are OGC WMS, GeoJSON and GeoService.
Natural Earth Data	Cultural, Physical and raster(basemap)data.Quick start kit(MXD and QGS files)with all the essential stylized layers.
USGS Earth Explorer	Landsat,Sentinel-2 and land cover. Digital Elevation Models such as NASA's ASTER and SRTM
OpenStreetData	High Spatial resolution cultural vector data such as buildings, roads, waterways.
Geocoder.us	Provides latitude & longitude of any US address , Geocoding for incomplete address ,Bulk Geocoding and Calculates distances
CityGrid	Incorporates local content into web and mobile applications.
Yahoo! BOSS	BOSS (Build your Own Search Service). Provides a facility of Place finder & Place Spotter to make location aware.
GeoNames	It contains geographical names, populated places and alternate names. All categorized into one out of nine feature classes.
ArcGis,MapGis	Helps to organize and analyze geographic data
MaxMind	GeoIP - IP Intelligence databases and web services minFraud-transaction fraud detection database
US Census	Offers several file types for mapping geographic data based on data found in our MAF/TIGER database

3.2 DATA REPRESENTATION:

GIS Systems deposits data from different varied data sources from wide range of communicating devices in different representation and file format. Representation of data can be done in two main categories as raster and vector data types, shown in Fig. 2.

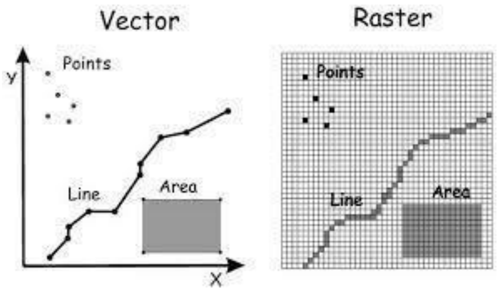

Fig.2. GIS Data Model

Raster data type is a two dimensional grid of rows and columns and stores the information as the value of pixel colors in a cell and the attributes values are contiguous in nature. Raster data type is used to represents information about to real world objects precisely such as aerial photograph, scanned maps, remote sensing data. It is best suitable representation technique for two dimensional spatial entities such as line, area and network. Vector data types are used to represent discrete features in nature. It is better representation method for surface representation in GIS and has a layered architecture representing point, line, and polygon. Vector data types are used to represents information from sources such as roads, rivers, cities, lakes, park boundaries with a layered hierarchy. TIN, contour, interpolation, elevation etc features can be easily extracted using vector data model. Raster to vector conversion shown in Fig.3.

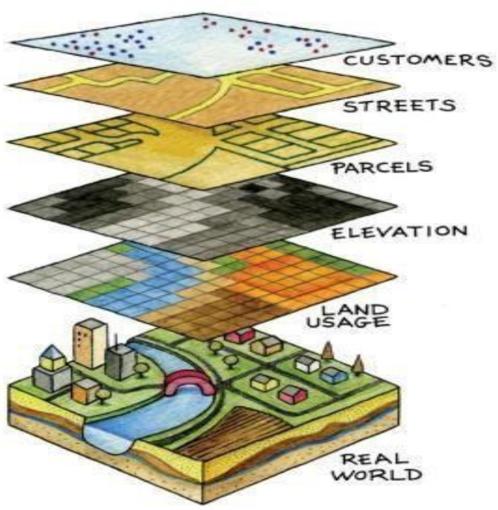

Fig.3. Layers-Raster to Vector

4. SPACIAL DATA MINING FOR GIS

A large amount of spatial dataset is increasing and accumulated from remote satellite imagery, sensors, aerial photography, etc. and this voluminous data have exceeded beyond to human capability to fully modeling, interpreting, analyzing and using. Therefore some advanced and efficient technique is required to discover knowledge from huge spatial database . Spatial Data Mining is the technique to find out the knowledge from huge geospatial dataset for extracting unknown, necessary spatial relationship, trends or patterns, not stored explicitly in spatial database . SDM techniques is nothing new, just an expansion of data mining techniques applied on Spatial data, GIS data and satellite dataset associated in various domain. Spatial means the data associated with the geographic location of the earth. Spatial data mining has been often used in many applications, like remote sensing, Marine Ecology, space exploration, environmental science, resource management, agriculture, geology, climatic change, NASA Earth Observing System (EOS), traffic analysis, Census Bureau, National Inst. of Health, National Inst. of Justice, etc. there is a need of mining the frequent trajectory patterns in a spatial–temporal database, extracting the spatial association rules from a remotely sensed database, generating polygon data from heterogeneous spatial information, and analyzing the change of land use . Spatial data mining is the extraction of spatial rules and has been used in many real time applications. proposed a model to that selects the locations of land-use by using the decision rules generated by nearest neighbors. It enables to recognize spatial dataset, find out relationship between spatial and nonspatial dataset, building of spatial knowledgebase, query optimizations, reorganizations of the spatial data, capture the basic features in easy and summarize way, etc. Spatial data using traditional data mining methods such as association, classification, clustering, trend detection, outliers generates interesting facts in associated domain. Spatial data mining techniques deals with spatial and nonspatial objects, attributes of neighboring objects and their spatial relations to find class, making clustering rules to detect outliers and deviations of trends, find association to extract multilevel topological relations.

4.1 SPACIAL DATA MINING TOOLS

DBMiner (DBlearn), GeoMiner are open source tools, used for data mining and query language, geomining and query language respectively and developed by Data Mining Research Group, Simon Fraser University, Canada. GeoDA(ESDA,STARS) is also open source Python language based tool which supports spatial autocorrelation statistics, spatial regression. Another open source Java language based tool is Weka-GDPM, developed by University of Waikato, NZ, which supports several standard data mining tasks. R language (sp, rgdal, rgeos) is also open source tool, compatible with C, FORTRAN, and Python language and support to analysis statistical and graphical techniques. Descrates is open source tool for Python language to visualize and analysis source data and display the outcome of classification on the plot. ArcGIS (ArcView, ArcInfo, ArcEditor) is paid tool, developed by Environmental Systems Research Institute (ESRI) used for Web API, Python, .NET language and used for spatial analysis and modeling features. It includes surface, network analysis, overlay, interpolation analysis and geo-statistical modeling techniques.

5. BIG DATA IN GIS:

O'Reilly Media coined the term big data in 2005. O'Reilly Media also created the term Web 2.0 one year previously. Big data referred to large datasets that cannot be managed and processed using traditional business intelligence tools. Yahoo also created Hadoop on top of Google's MapReduce in year 2005. The goal of Hadoop was to index the world wide web. Nowadays, the impact of big data is in every industry, including GIS. The datasets with massive amount of data in GIS are called big spatial data.

Big data is defined using volume, velocity and variety. GIS is also facing the same problem as other industries with an increase in variety, size and frequency of updates. These problems are exceeding the capacity of traditional GIS architecture, methodologies and spatial computing techniques. Spatial and Spatio-temporal data is not directly supported by the state-of-the-art tools and concepts used in big data like Hadoop, map-reduce, HBase, Hive and Spark. Hence, most of the big spatial data is processed as non-spatial data or by using some wrapper, which has a high processing cost. The limitations in existing big data tools and concepts inspired researchers to come up with extensions and architectures to handle big spatial data efficiently.[6] These include Hadoop-GIS, Spatial Hadoop, GeoSpark and integration of NoSQL with spatial data. ESRI, the world leader in the GIS software, also released a suite of GIS tools on Hadoop that work with their flagship ArcGIS product. GIS tools for Hadoop suffer from limitations like dealing with spatial data as non-spatial data and using Hive, a layer on top of Hadoop among others. In the coming sections, we will look into the work done by researchers on these extensions and architectures that enable processing of big spatial data.

5.1 BIG DATA COMPUTATIONAL FRAMEWORKS

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, SpatialHadoop ,Hadoop-GIS, and systems that use Resilient Distributed Datasets (RDD) such as GeoTrellis SpatialSpark , GeoSpark and Spatial In-Memory Big Data Analytics (SIMBA). 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.

5.1.1 HADOOP-GIS

Hadoop-GIS, basically integrating the Geographic information System (GIS) with HDFS file system for distributed data processing.It underpins various sorts of spatial queries on Hadoop-MapReduce system through spatial partitioning, adjustable Real-time Spatial Query Engine **(**RESQUE). It uses global partition and adaptable local spatial indexing to accomplish the efficient query processing.[8] It incorporates Hive to boost declarative spatial inquiries with an integrated construction modeling and accessible as an arrangement of library for handling spatial inquiries.
Some of the salient features of the Hadoop-GIS system are given here. There are five major categories of queries which can be summarized as follows:

- **Feature aggregation queries** (non-spatial queries), for example, queries for finding mean values of attributes or distribution of attributes,
- **Fundamental spatial queries,** such as point based queries, containment queries and spatial joins,
- **Complex spatial queries,** such as spatial cross-matching or overlay (large scale spatial join) and nearest neighbor queries,

- **Integrated spatial and feature queries,** for example, feature aggregation queries in a selected spatial regions, and

- **Global spatial pattern queries,** include queries on finding high density regions, or queries to find directional patterns of spatial objects.

Traditional Methods for Spatial Queries have the following limitations:

- Managing and querying spatial data on a massive scale,
- Reducing the I/O bottleneck by partitioning of data on multiple parallel SDBMSs disks,
- Optimizing for computationally intensive operations such as geometric computations,
- Locking of effective spatial partitioning mechanism(s) to balance data loads and task loads across database partitions, and
- High data loading overheads.

5.1.1.1 REQUE

It stands for **RE**al-time **S**patial **QU**ery **E**ngine. It takes advantage of global tile indexes and local indexes which exist to support more efficient spatial queries.It is completely advanced engine. It underpins the data compression and low overheads on data loading. It is compiled as a shared library which can be effectively conveyed in a cluster environment. To handle boundary objects in a parallel query processing situation two methodologies are proposed, viz. (i) Assignment and (ii) Multiple Matching. The multiple task methodology is easy to implement and fits well in the MapReduce programming model.

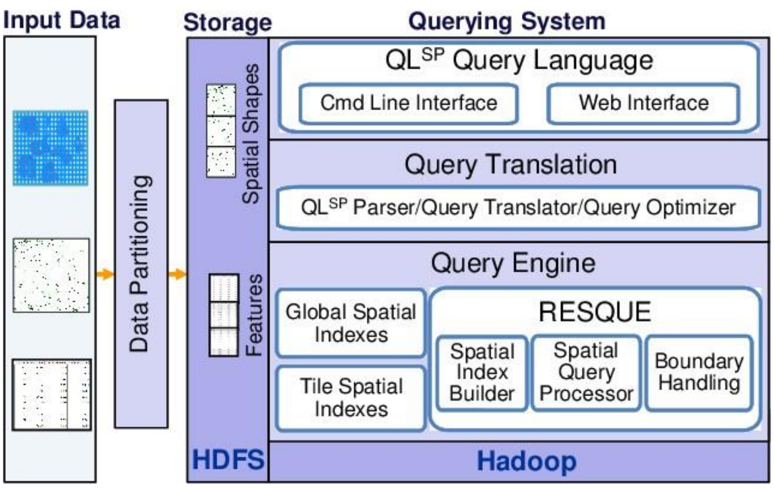

Fig.4.Architecture of Hadoop-GIS Hive

5.1.1.2 DATA PARTITIONING

It is a starting step to define, produce and show partitioned data. In spatial data partitioning, focus is on dividing high density parts into smaller density ones. For boundary intersecting objects, taking multiple assignments based approach, in which objects are replicated and assigned to each intersecting tile, a model is proposed as follows:

- First Count the number of objects in each tile, and sort them based on the counts.
- Define a a threshold Count Max (Cmax) as the maximal count of objects allowed in a tile.
- Pick Pick all tiles with object counts larger than Cmax.
- Split each of them into two equal half-sized tiles.
- This process is repeated until all tiles have counts lower than Cmax

5.1.1.3 INTEGRATION WITH HIVE

To provide an integrated query language and unified system on MapReduce, the Hive is extended to HiveQL, with spatial query support for spatial constructs, spatial query translation and their execution. The HiveQL is an integrated version of Hadoop-GIS. There are several core components in HiveQL to provide spatial query processing capabilities, such as:

- Spatial Query Translator, which parses and translates SQL queries into an abstract syntax tree,
- Spatial Query Optimizer, which takes an operator tree as an input and applies rule based optimizations, and
- Query Engine to support the following infrastructure operations: spatial relationship comparison, spatial measurements, and spatial access methods for efficient query processing.

5.1.1.4 QUERY PROCESSING

Hive uses the traditional –plan-first-execute-next‖ approach for query processing. This consists of three steps: (1) query translation, (2) logical plan generation, and (3) physical plan generation. The main differences between Hive and Hive^{SP} are in the logical plan generation. If the query contains spatial operations, the logical plan is regenerated with special handling of spatial operators. Specifically, following two additional steps are performed to rewrite such a query.

- First step: The operators involving spatial operations are replaced with internal spatial query engine operators at tile level query processing.
- Second step: The serialization/de-serialization operations are added before and after the spatial operators, respectively, to prepare Hive query for communicating with the spatial query engine.

5.1.2 SPACIAL HADOOP

SpatialHadoop is designed to handle big spatial data on Apache Hadoop. SpatialHadoop is extension of Hadoop which supports spatial data natively. SpatialHadoop is spatially aware. SpatialHadoop adapts traditional spatial indexing structures, Grid, R-tree and R+-tree to form a two-level spatial index. SpatialHadoop has built-in Hadoop base code, which makes Hadoop aware of spatial constructs and spatial data inside the core functionality of Hadoop. This is key for increasing power and efficiency. SpatialHadoop supports a set of spatial index structures including R tree-like indexing, which is built-in feature in Hadoop Distributed File System (HDFS). SpatialHadoop has functionalities, which users can directly modify.SpatialHadoop has four components :

- Language : Language provided by Pigeon, a high-level SQL language.
- Operations : Operation Layer Have spatial operation.
- MapReduce : MapReduce has 2 new components i.e. SpatialFileSplitter and SpatialRecordReader
- Storage : Storage applies 2 level indexing, global and local. Global indexing partition data at node level. Local indexing works inside node. Grid, R-Tree and R+-Tree are supported index.

Fig.5. Architecture of SpatialHadoop

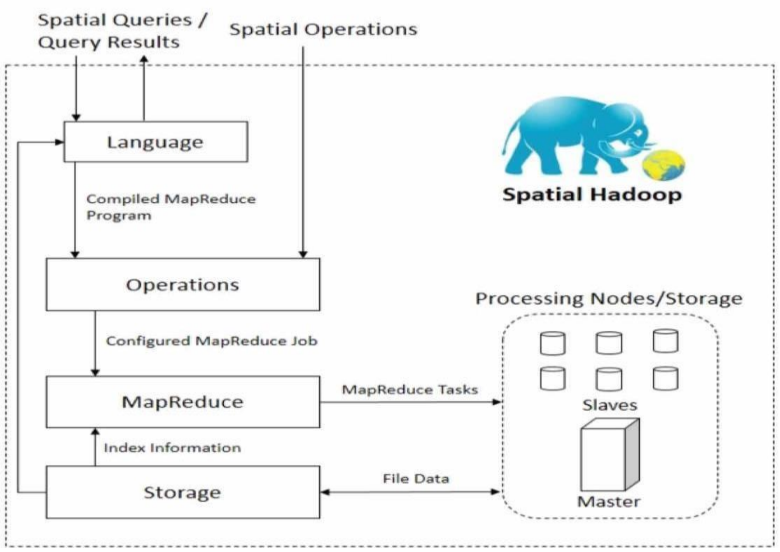

Fig.4.shows SpatialHadoop architecture at high level. Developers will be able to interact at operation level and could be able to perform operations like SpatialJoin, RangeQuery etc. Casual user would be able to do spatial query and see the result via visualization. And system admin is responsible for tuning up the configuration files and system parameters. MapReduce, which is layer, will interact with storage layer, where data would be stored using spatial indexing.

5.1.3 GEOSPARK

Apache Spark is an open-source engine for big data processing within Hadoop and runs workloads 100 times faster than MapReduce-the standard Hadoop engine. Apache Spark can program entire clusters with implicit fault tolerance and data parallelism. GeoSpark is built on top of Apache Spark to cater to the challenges of System Scalability and interactive performance with big spatial data. GeoSpark is also used to solve challenges like co-location pattern mining GeoSpark is designed to process big spatial data using in-memory cluster computing framework GeoSpark consists of Apache Spark layer that provides basic Spark functionalities,Spatial Resilient Distributed Dataset (SRDD) layer that provides support to geometrical and spatial objects and Spatial Query Processing layer that efficiently executes spatial query processing algorithms. The overview of GeoSpark layers is shown in Fig.4.

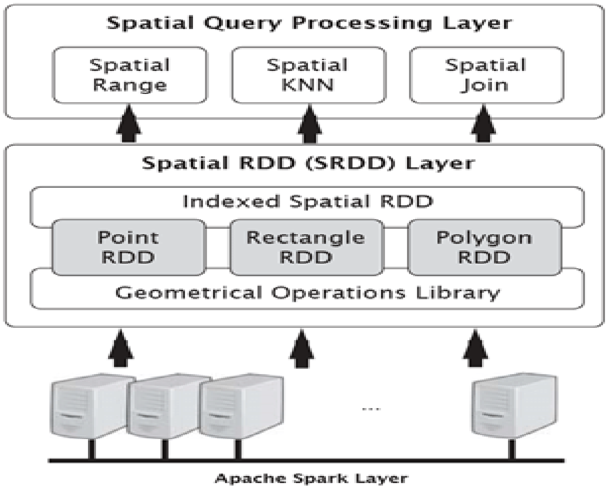

Fig.4. Architecture of GeoSpark

The run time performance of GeoSpark is two orders of magnitude faster than existing Hadoop-based systems and up to an order of magnitude faster than Spark-based systems.GeoSpark also comes with some limitations, it does not include integrated support for clustering methods, it does not cater to specific application requirements via customizable modules, and it did not consider critical concerns like data load balancing in particular application scenarios.

5.1.4 NoSQL

NoSQL databases provide an alternative data storage and retrieval to the relational databases. NoSQL is built for large-scale database clustering in web and cloud applications, as the performance and scalability outweigh the need for rigid data consistency in these applications. Distinctly, NoSQL databases are not required to follow the established relational schema and specialize in storing unstructured data .Hence, they are particularly useful for large sets of distributed data. The trade-off for not using relational schema is the level of consistency.

The rise in big data has coincided with the increase in popularity of NoSQL databases. The challenges of working with big data are mirrored in big spatial data as well. The big spatial data also suffers from less or no support for spatial data in big data technologies. The relational databases are not considered an ideal platform for spatial data because it is expensive and hard to maintain, especially when working with large systems. Hence, the researchers have worked with various NoSQL databases to provide support for big spatial data. There are efficiency problems associated with GIS data in relational databases.

Notable NoSQL systems that provide support for big spatial data include CouchDB, MongoDB, Neo4j.worked with MongoDB and Python to store and access big spatial data in NoSQL system. pointed at the challenges of data storage and processing big spatial data in a web service and proposed a cache-based GIS web service using Memcache and MongoDB.created a coherent healthcare system by exploiting the benefits of combining MySQL, MongoDB and GIS databases The authors argued that this system would help applications to comply with EHR (Electronic Health Record) and HIPAA (Health Insurance Portability and Accountability Act of 1996) requirements without compromising the performance and scalability.

6. ISSUES AND CHALLENGES:

With the explosive growth in GIS data is viewing in respect of data integration and mining huge spatial data volume. To overcome this issue, architecture is proposed to deal the problems of huge volume spatial data integration on the basis of analysis of spatial data. Data warehouse technique is providing the facility for data integration and deposits summarized data. The fast growing data is the origin of big data. In recent the big data approach has coined for mining spatial data using parallel algorithm on Hadoop and MapReduce architecture in a distributed environment to deal with such huge amount of datasets. We proposed a MRPrePost, hadoop architecture based, Pre-Post algorithm, used for mining big data in this paper. MRPrePost is hybrid of the Dis-Eclat and used as association rule mining to find frequent item set from huge spatial data sets.

7. CONCLUSION:

Big data refers to the high volume, high variety and high veracity of data. The increasing volume of storage and variety in the geospatial data collected from different sources, pose new investigations in data storage, data management and processing including quality verification of data, analysis and visualization of the data.This will provides with a detailed knowledge on Geographic information system (GIS) in Big Data Technology. A detailed analysis and description of GIS data sources, data formats and data representation is presented from various literatures. The Big Data computational framework used for different applications in GIS are also detailed.. We also provided summary tables for GIS data sources, tools. Furthermore, research challenges are discussed with focus on multiple dimensions, complexity, time-consumption, expense, storage size, integration, validation, and ground reality. Important open issues, i.e. privacy, security, ethics, cost, and generalization, are also summarized for future studies.

8. REFERENCES:

[1] Mousi Perumal, Bhuvaneswari Velumani, Ananthi Sadhasivam, and Kalpana Ramaswamy : Spatial Data Mining approaches for GIS , DOI: 10.1007/978-3-319-13731-5_63,(2015)

[2] Heywood, I., Cornelius, S., Carver, S., Raju, and S. An Introduction to Geographical Information Systems, Second Edition, Pearson Education, pp.10.

[3] Lingjun Zhao1,2 · Lajiao Chen1 · Rajiv Ranjan3 · Kim-Kwang Raymond Choo4 · Jijun He5 : Geographical information system parallelization for spatial big data processing: a review , Cluster Comput DOI 10.1007/s10586-015-0512-2,(2015)

[4] Eldawy, A. :A demonstration of shahed: A mapreduce-based system for querying and visualizing satellite data. 2015 IEEE 31st International Conference on Data Engineering, 1444–1447. doi:10.1109/ ICDE.2015.7113397(2015)

[5] Purnima Shah(&) and Sanjay Chaudhary : Big Data Analytics Framework for Spatial Data , A. Mondal et al. (Eds.): BDA 2018, LNCS 11297, pp. 250–265, 2018.

[6] Purnima Shah,: Developing Big Data Analytics Architecture for Spatial Data , ceur-ws.org/Vol-2399/paper03.pdf,(2018)

[7] P. Praveen, Ch.Jayanth Babu, B. Rama, : Big Data Environment for Geospatial Data Analysis, DOI: 10.1109/CESYS.2016.7889816,(2016)

[8] Raja Sher Afgun Usmani, Taylor's University, Malaysia : Geographic Information System and Big Spatial Data: A Review and Challenges , International Journal of Enterprise Information Systems, Volume 16 • Issue 4 • October-December 2020

[9] Ablimit Aji : Hadoop-GIS : A High Performance Spatial Data Warehousing System Over MapReduce, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3814183/

[10] https://www.semanticscholar.org/paper/Hadoop-GIS%3A-A-High-Performance-Spatial-Data-System-Aji-Wang/4cb4af8e5bb1c4d9569a7920b2a419873540092d/figure/0

[11] Md Mahbub Alam, Suprio Ray, Virendra C. Bhavsar : A Performance Study of Big Spatial Data Systems, DOI: 10.1145/3282834.3282841,(2018)

[12] Ahmed Eldawy Mohamed F. Mokbel: The Era of Big Spatial Data: Challenges and Opportunities, DOI 10.1109/MDM.2015.82,(2018)

[13] Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., & Saltz, J. (2013). Hadoop GIS: a high performance spatial data warehousing system over mapreduce. Proceedings of the VLDB Endowment, 6(11), 1009-1020

[14] Eldawy, A., & Mokbel, M. F. (2014). Pigeon: A spatial mapreduce language. Paper presented at the Data Engineering (ICDE), 2014 IEEE

[15] Mazin Alkathiri, Jhummarwala Abdul, M.B. Potdar, PhD : Geo-spatial Big Data Mining Techniques , International Journal of Computer Applications (0975 – 8887), Volume 135 – No.11, February 2016