A survey on spatial data analysis tools

Smt. Veena G
Assistant Professor
Department of Computer Science and Engineering
Vemana Institute of Technology

Smt. Shilpa Reddy K
Assistant Professor
Department of Computer Science and Engineering
Vemana Institute of Technology

Abstract *- In recent days the amount of spatial data collected from various sources such as remote sensors and satellite systems are really huge. Different data analysis techniques are used to extract useful information or knowledge from large voluminous data. Geographical or geo-spatial data collected from various sources are not necessarily in the same format, data may be structured or unstructured. From last few decades, Research and Development is going on to efficiently extract useful information from geo-spatial data. Number of techniques has evolved to analyse spatial data in an efficient manner. This paper reviews different tools available for geo-spatial data analysis.*

Index Terms - Spatial database, OGC, GeoMiner, SpatialHadoop, GIS, GeoSpark

1. INTRODUCTION

Spatial data represents the location, shape and size of an object such as lakes, buildings, parks, mountains, forests etc. Spatial data may also include other attributes which provide more information about an object. Spatial data is usually stored as co-ordinates and topology, the data can be mapped i.e. data can be represented on a map for better viewing. The process of discovering useful, non-trivial patterns from large spatial data sets is known as **spatial data analysis.** Spatial data can be collected from different sources; the important sources are data from **sensors attached to platforms**. Sensors consist of **cameras**, **lidar** (it is a surveying method that measures distance to a target by illuminating the target with laser light) and **digital scanner.** Platforms consist of satellites and aircraft. The amount of spatial data collected is very huge; NASA's EOSDIS (**E**arth **O**bserving **S**ystem **D**ata and **I**nformation **S**ystem) archives about 6.4 TB of data every day [1].

Spatial data analysis finds its applications in many areas, historically in the mid of 19th century in Soho district of London the number of human death was increased to 578 per day because of deadly disease cholera. Dr. John snow published a map of an area where the number of deaths was high, map included locations of 13 public wells. He also showed that the cause for cholera is drinking infected water from a particular well out of 13 wells [2]. In modern times spatial data analysis is used in vast areas, cancer clusters [3] to investigate environmental health hazards, the cancer cluster can be defined as a greater than expected number of cancer cases that occurs within a group of people in a geographic area over a period of time, using spatial data analysis the environmental hazards responsible for cancer causing agent can be determined. Identifying Crime hotspots [4] (areas of concentrated crime) is one area where spatial data analysis is used widely, which helps police department to plan for police patrol routes.

The rest of the paper is organized as follows. Section 2 describes characteristics of spatial data analysis. Section 3 surveys tools available for spatial data analysis.

2. SPATIAL DATA

Spatial data identifies the geographic location with details of boundaries and features of the location. Usually spatial data will be stored as co-ordinate and topology. Extracting non-trivial, interesting patterns from spatial data is known as spatial data mining. Important characteristics of spatial data mining are as follows.

2.1 AUTO CORRELATION

First law of geography states that "Everything is related to everything but nearby things is more related than distant things". Spatial auto correlation helps to understand the degree to which one object is similar to other nearby object. Moran's I [5] is used to measure spatial auto-correlation. Positive spatial auto correlation occurs when similar spatial objects cluster together. It occurs when Moran's I is close to +1. Fig.1 shows an image with positive auto correlation, white and black blocks are uniformly clustered.

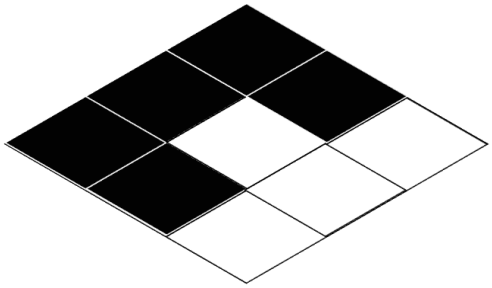

Fig.1 positive spatial auto correlation.

Negative spatial auto correlation occurs when dissimilar spatial objects cluster together. It occurs when Moran's I is close to -1. Fig.2 shows an image with negative auto correlation, white and black blocks are dispersed.

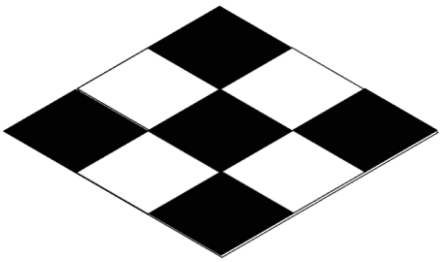

Fig.2 negative spatial auto correlation.

2.2 REPRESENTATION

Pictorial representation explains better than textual representation, so spatial data mining results have to be summarized on top of maps.

2.3 STORING

Longitudes, latitudes and topological information of the spatial object along with non-spatial attributes are stored in spatial databases. The stored data will be used as an input to spatial data analysis tools.

2.4 TYPES OF SPATIAL RELATIONS

Spatial databases do not share spatial relations explicitly additional functionalities are required to compute them. **OGC** (Open Geospatial Consortium) specified the following types of spatial relations.

- Distance relations: Euclidean distance between two spatial features.
- Direction relations: Ordering of spatial features in space
- Topological relations: Characterize the type of intersection between spatial features.

3 AVAILABLE TOOLS FOR SPATIAL DATA ANALYSIS

3.1 GEOMINER

GeoMiner [6] is a spatial data analysis tool which mines characteristic rules, comparison rules and association rules in geo spatial database. GMQL (Geo Mining Query Language) is designed and implemented as an extension to spatial SQL. GeoMiner system includes spatial data cube construction module, spatial OLAP (Online Analytical Processing) module, spatial data mining module.

Fig.3 shows architecture of GeoMiner. Geo-characterizer modules mines set of characteristic rules at multiple levels of abstraction in a spatial data base. Geo-characterizer answers questions like given spatial hierarchies of Western Canada Geo-characterizer describe general weather patterns according to region partitions. Given non spatial hierarchies such as temperature, precipitation etc. geo-characterizer describe the regions in Western Canada based on their generalized weather patterns.

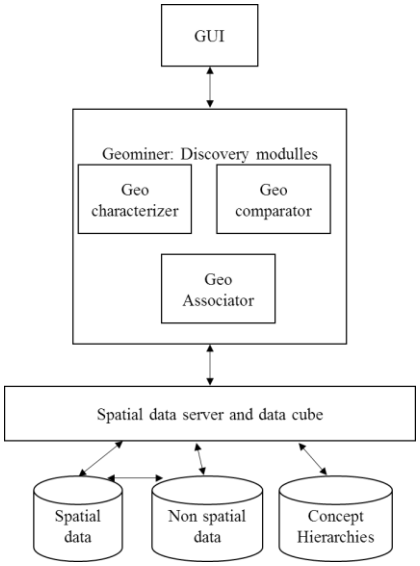

Fig.3 Architecture of GeoMiner

Geo-comparator- This module mines a set of comparison rules. It compares one set of data known as target class to other set of data known as contrasting class. Ex. The comparator module show difference in weather patterns between Columbia and Alberia.

Geo-associator - This module finds a set of strong spatial association rules from the relevant set of data in spatial database. It shows frequently occurring patterns of a set of data items in spatial database. For example while finding the relationships among Canadian towns, closeness to water, population and closeness to the national border, many association rules were found, one among them is if a Canadian town is large and is adjacent to large water body it is close to the U.S. border, with the possibility of 78%.

3.2 SPATIALHADOOP

Huge volumes of spatial data collected from various different sources like satellites, cell phones and medical data need to be efficiently managed and analysed to extract useful, interesting patterns. Hadoop MapReduce framework is the best choice to analyze huge volumes of data, but Hadoop fails to support spatial data since it was designed to analyse non-spatial data. To efficiently analyse huge volume of spatial data **SpatialHadoop [7]** was introduced. SpatialHadoop is accessible through a high level language named Pigeon. SpatialHadoop has better performance when compared to Hadoop. For example, consider range queries in Hadoop and SpatialHadoop.

Obj = LOAD ‘points’ AS (id:int, a:int, b:int);
Result = FILTER Obj BY a < a2 AND a > a1 AND b < b2 and b > b1;

RANGE QUERY IN HADOOP

```
Obj = LOAD 'points' AS (id:int, Location:POINT);
Result = FILTER Obj BY
Overlaps(Location,RECTANGLE(a1,b1,a2,b2));
```

RANGE QUERY IN SPATIALHADOOP

For 70M spatial objects on a 20-nodes cluster the above range query execution in Hadoop takes 200 seconds. Whereas execution of above range query in SpatialHadoop takes only 2 seconds.

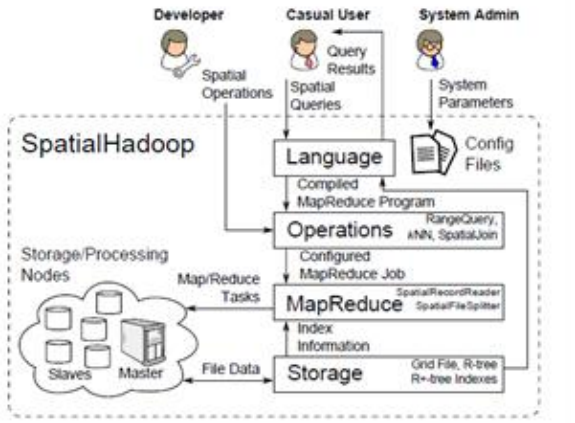

Fig.4 Architecture of SpatialHadoop

Fig.4 shows architecture of SpatialHadoop. It follows master slave architecture, a cluster of it contains one master node which accepts a user query and divides it into smaller tasks and each smaller task is executed in many slave nodes. Developers, casual users and administrators are the three types of users who interact with SpatialHadoop. Developers understand the system better they can implement new spatial operations required for their application. Casual users can process their datasets using GUI. System administrators work on fine tuning system by adjusting system parameters in the configuration files. SpatialHadoop has storage, mapreduce, operations and language layers.

Storage layer: Hadoop uses input files which are non-indexed heap files stored in HDFS (Hadoop Distributed File System), whereas in SpatialHadoop index structure is used within HDFS. There are two levels of indexes are used, global index and local index.

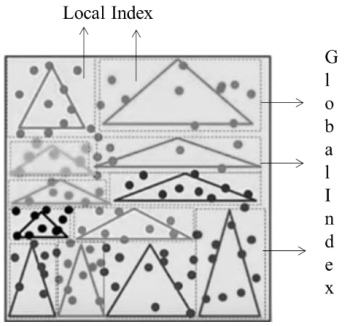

Fig.5 Indexing in SpatialHadoop

Fig.5 shows the indexing in SpatialHadoop. Huge data set is divided into blocks so that closer objects are placed in a single block, each block can be processed in a separate machine. The data within a block is indexed using local index which can be either grid index or R tree index or R+ tree index.

MapReduce layer: MapReduce [8] is a programming paradigm which allows massive scalability across thousands of servers in a cluster. It refers to two distinct and separate tasks that Hadoop programs perform. First is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken into key/value pair. The reduce process takes the output from a map as input and combines those data tuples into a smaller set of tuples. SpatialHadoop has two new components SpatialFileSplitter and SpatialRecordReader that allow spatial operations to access the indexes.

Operations layers: The spatial operations supported by SpatialHadoop are range query, kNN, spatial join.

Language Layer: SpatialHadoop uses Pigeon as high level language which has functions and data types which are OGC complaint.

3.3 GEOSPARK

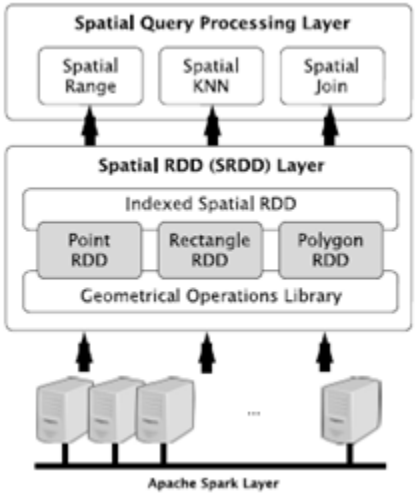

Fig.6 Architecture of GeoSpark

Fig.6 shows architecture of Geospark [9]. It is an extension of Apache Spark. Apache Spark does not support spatial operations. GeoSpark was built on Apache Spark to analyze spatial data. Geospark contains three layers, the Apache spark layer's main functionality is to load the data from persistent storage and also save data to persistent storage, the persistent storage includes hard disk or HDFS (Hadoop Distributed File System). SRDD (Spatial Resilient Distributed Dataset) layer extends Apache Spark with Spatial RDD. Spatial indexes like Quad-tree, R-tree are provided. Spatial Query Processing layer provides spatial range query, spatial join query spatial KNN query. GeoSpark is based on map-reduce paradigm. It supports only java and SCALA.

CONCLUSION

In this paper an introduction to spatial data and spatial data analysis is provided. To analyze spatial data and to extract non-trivial and interesting patterns many tools exist. In this paper three different tools namely GeoMinier, SpatialHadoop, GeoSpark is considered. Each tool has its own pros and cons. GeoMiner is not efficient to process huge volumes of spatial data. SpatialHadoop uses map-reduce framework, it provides high reliability, availability and scalability. Compared to GeoSaprk it is much slower. GeoSpark also uses mao-reduce framework, it is built on top of Apache Spark. It is much faster than SpatialHadoop. Its main limitation is it supports only java and SCALA.

REFERENCES

[1] https://earthdata.nasa.gov/getting-petabytes-to-people-how-the-eosdis-facilitates-earth-observing-data-discovery-and-use
[2] https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak
[3] https://www.cdc.gov/nceh/clusters/factsheet.htm
[4] John E. Eck, Spencer Chainey, James G. Cameron, Michael Leitner, and Ronald E. Wilson, "Mapping Crime: Understanding Hot Spots", Aug. 2005, National Institute of justice
[5] https://en.wikipedia.org/wiki/Moran's_I
[6] Han, J., Koperski, K., & Stefanovic, N. (1997). GeoMiner: a system prototype for spatial data mining. Paper presented at the AcM sIGMoD Record.
[7] Ahmed Eldawy and Mohamed F. Mokbel, "A Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial data", August 26th 30th 2013, Proceedings of the VLDB Endowment, Vol. 6, No. 12
[8] https://www- 01.ibm.com/ software/ data/ infosphere/ hadoop/mapreduce/
[9] Jia Yu, 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. ACM, 2015, p. 70.