An optimized Geospatial Data Storage and Retrieval Using SpatialHadoop in Cloud Environment

Karthi S
School of Computer Science and Engineering
VIT University
Vellore, TN, India
s.karthi2014@vit.ac.in

Prabu S
School of Computer Science and Engineering,
VIT University
Vellore, TN, India
sprabu@vit.ac.in

ABSTRACT
Technology advancement increase the volume of remote sensing data growth exponentially increase processing capability. Objective of this research proposes a secured and efficient way of storing and retrieval the geospatial data in public cloud using Spatial Hadoop mechanism in cloud computing environment. To secure the geospatial data retrieval in the public cloud Attribute based authentication protocol. The spatialHadoop-GIS is a high performance and secure accessing spatial data for running gauge spatial queries on spatialHadoop.

Keyword: Remote sensing data, Geospatial Information Systems, Cloud Environment, Attribute based Authentication

INTRODUCTION
GIS applications play a major role in different sections of modern cities[13][14]. Computerized cities are the sources of the required data. As an example, vector data (focuses, polygon and line) , raster data (aeronautical and satellite images) use spatial analysis on a large scale. Satellites, GPS gadgets and unprecedented sensors deliver this kind of information once in a while. Because of the voluminous amount of facilities given for spatial storage and also because of the geospatial model’s complexity, it requires an upper time multifaceted nature and also needs large resources for registering. Since spatial information sharing and organization undergo considerable changes from time to time, unborn synthetically GIS platform requires displaying a lot of copious information’s, responding to a large number of parallel requests and finishing taking care of shrewd data. For these kinds of data organization, going to Internet is not preferred as it is not exactly the same as preventing exploratory registering and exchange preparing. It exhibits appeal to expansibility and usability of systems supporting it and normal prevalent preparing and database structures cannot meet those requirements [4]. Considering the colossal flexible figuring resources, circulated processing gives organizations on the Internet to the customers. Considering the latest examples in the IT industry, it can be considered as a standout amongst all of them. With the advancements in geographic information, disseminated registering can possibly give a relatively good making environment. It finds out the present problematic issues in the organization of geospatial information by passing on arrangements. A strong IT base to the GIS is supplied by appropriated figuring. Considering the amassment of significant scale servers, circulated figuring helps in upgrade of structure execution, enrolling and limiting the capacity to significantly reduce gear and programming costs. Cloud stages have excellent system adaptability and can supply mass geospatial handling needs because they realize virtualization organization with physical resources.

Geographic Information System (GIS) focuses on identifying natural changes and find out the laws of economic or social development all by processing huge attributed and spatial data of surface of the earth[1]. With development of performing computations in clouds, it is possible to give a solution to voluminous information storing, processing information and analyzing spatial data. With the coming of clouds, the huge information that knowledge acquisition and decision support information Requires could be arranged and processed in parallel entirely inside cloud. For centralized storage and backup, data security and protection is guaranteed. Enterprises apart from always renting hardware and data resources, could also employ the service inside cloud itself thus giving unique services for processing data.

CLOUD COMPUTING
Cloud computing can be considered as a relatively recent Internet dependent concept and also as a subset within supercomputing, that gives on-demand and scalable services to computing, hardware, software and storing with great security for the users in the cloud has many definitions for cloud terms and the definition given by the pioneers, are marked on reference. Though it may seem that cloud computing is a recent web service, it relies on old already existing technologies like virtualization, utility computing and distributed computing. But, the core features that separate cloud computing is the service dependent attribute of cloud.
CLOUD COMPUTING IN GEOGRAPHICAL INFORMATION SYSTEM (GIS)
GIS encompasses many types of high level features related to data processing, and its rudimentary function [2] is to collect information, managing, processing, analysing and outputting. Depending on such rudimentary functions, GIS provides lots of application by using spatial analysis, database, modeling, data integration technology, network technology and more development environment to meet the voluminous user needs. With the coming of cloud computing, opportunities for GIS have been increased.

There are lot many variants of spatial data: geometric data, attribute data and relational data that evaluates mutual relationship amongst map elements, and auxiliary data which helps in map processing. As the data varies with time and the size and of data is huge and voluminous, maintenance and management of spatial data storage involves high costs of human, financial and hardware resources. The concept of storing data in cloud that brings together many types of storage devices to work in unison by using software’s with the ability of clustering, distributed or grid file systems, is a complex system consisting of many devices like storage equipment’s, network device, servers, the access network, public access interface and client programs. Storage equipment’s from the core along with other parts render services of business access and data storage taking the help of application software’s. For voluminous data residing on cloud, the cloud platform renders services needed for storing information, managing, backup and maintaining by using its cluster technology and distributed file systems and also builds the user interface for recognized users which can be accessed by a Unified API. Authenticated users can hire servers on cloud to store their own data and pay to use the data of other service providers stored in the cloud that results in greatly reducing the cost of hardware and software maintenance and upgradation.

In this paper, we import conveyed processing advancement including Spatial Hadoop as the primary unquestionable MapReduce framework. Spatial Hadoop works similarly to Hadoop where undertakings are formed similarly as guide and diminish limits, and subsequently existing Hadoop activities can continue running as if it is on Spatial Hadoop. In all cases, if the activities oversee spatial data, Spatial Hadoop will have request of degree best execution over Hadoop. Spatial Hadoop keeps pushing its spatial forms in all layers of Hadoop, particularly for stockpiling, vernacular, operations and MapReduce layers. In keeping aside the advantages of the development, affiliations now go up against rudimentary safety problems, for example, security, data spillage and character protection while transporting applications and tricky data to and from the cloud. Out of these challenges baffling correspondence still needs basic thought and tending. This paper discusses some open benchmarks related to validation and approval by unknown verification conventions associated with the proposed tradition.

SPATIAL HADOOP ARCHITECTURE
A Spatial Hadoop [5][6][8][9][11] collection has 1 expert hub that splits a MapReduce work into tinier sub activities, completed with the slave nodes. It contains 3 types of clients which interface with Spatial Hadoop. The usual clients manage the framework through the accessible applications. Designers shall do novel processes by using MapReduce. Framework admins have profound comprehension of architectural issues which shall tune up the framework via the arrangement files. The center of Spatial Hadoop comprises of four layers [6] as shown below. (1) The first layer has a language consisting of high level data types of spatial and methods. (2) The second layer possesses three indexes of spatial, R-tree, grid file and R-tree, wherein everything is done internally, the Hadoop Distributed File System (HDFS) that allows the functions of spatial to execute quickly as opposed to loading files in Hadoop. The organization of it is done in 2 layers, 1 global index which divides data into nodes and also several indexes of limited nodes to consolidate records with every node. (3) The third layer consists of two novel modules namely SpatialFileSplitter and SpatialRecordReader, which permits functions of spatial to get the built indexes, the spatial indexes and new MapReduce modules. (4) The fourth layer summarizes the spatial processes maintained by Spatial Hadoop. A simple procedure encloses 3 standard spatial processes, range query, kNN and spatial link.

INPUT LAYER
SpatialHadoop [5][6] gives a straightforward abnormal state dialect that revamps the participation with the structure for non-specific customers. This dialect gives a verifiable sponsorship to spatial data sorts, spatial primitive limits, and spatial operations. Spatial data sorts (polygon, rectangle, point) are used to describe the outline of an information record upon its stacking strategy. The spatial primitive limits MBR, distance and overlaps are related with spatial credits to learn the distinction between two shapes’ centroids, check if two shapes cover or not and figure the unimportant skipping rectangle of a polygon separately. To feature queries with a declarative query language, the paper has used Hive, that gives SQL like language atop of Map Reduce. We can use Hive with spatial query support by extending HiveQL with spatial constructs, spatial query translation and execution, with addition of the spatial query engine into the Hive query engine. The spatial operations go address, k-nearest neighbor, and spatial join are associated with data archives of spatial qualities and deliver the results in another yield record.
MAPREDUCE LAYER
The Hadoop’s MapReduce [5][6] layer is anticipated to work with non-recorded stack archives. In all cases, spatial operations in Spatial Hadoop take spatially requested records as data, which needs specific dealing with some spatial operations. Spatial Hadoop provides 2 new modules in the MapReduce layer, the SpatialFileSplitter and the SpatialRecordReader, which attempts the worldwide and adjacent records, separately, for capable data to get in. The SpatialFileSplitter takes as input a couple spatially recorded reports despite the fact that client gave channel work. Both Spatial-RecordReader and SpatialFileSplitter together help engineers design a lot of spatial operations as MapReduce projects.

EXECUTION LAYER
The MapReduce layer [5][6] gives the opportunity of identifying many spatial operations in Spatial Hadoop. In this paper, we present the utilization of achieve question, spatial and kNN join as three relevant examples of how to mishandle the new stockpiling and MapReduce layers in Spatial Hadoop. Other spatial operations, for example, kNN join can be in similar manner be completed by taking a similar procedure. In achieve questions, only the bundles that cover the request run are picked by the SpatialFileSplitter. The picked bundles then individually encounter the SpatialRecordReader that separates the close-by rundown in that apportioning and runs a standard degree address on that rundown to find out the planning records. Since couple records are imitated in the middle of requesting process, the reference point duplicate avoiding technique [2] is uapplied on planning records to make sure that each answer record is represented one time only. The k-nearest neighbor operation is executed as a combination of two cycles. The SpatialFileSplitter makes use of the overall record to find a match that has the question point in the essential cycle. The adjacent rundown in the fragment is separated by the SpatialRecordReader and is also used to find the kNN in the match.

WAREHOUSE LAYER
SpatialHadoop [5][6] incorporates many new spatial files that are undoubtedly suitable for the MapReduce operations. These records defeat a confinement in Hadoop that underpins just non-listed pile documents. Spatial Hadoop sorts out files in two levels namely worldwide ordering and nearby ordering. The worldwide record segments data over nodes belonging to group.
and the nearby file sorts out data efficiently inside each and every node. The detachment of worldwide and nearby lists fits the MapReduce programming worldview where the worldwide list is used to set up the MapReduce work and the neighborhood files are used to prepare map errands. Splitting each record into many small packets allows independent ordering of every segment in the memory and composing it to a document in a successive manner, but this is a difficult task. The principle memory of the expert hub holds the worldwide file while every neighborhood record is put together as a single document piece in a slave hub. Spatial Hadoop bolsters lattice document, R-tree and R+-tree files. By issuing the new record framework order write SpatialFile presented in Spatial Hadoop, a record is built for the current document, where the client gives the information record, section to list, and file sort to be build. A file is created in Spatial Hadoop with the help of a MapReduce work that keeps running in 3 stages, namely apportioning, nearby ordering, and worldwide ordering. In the apportioning stage, a document is parcelled spatially with the final goal that every segment is contained within a rectangle.

SPATIAL INDEXING
To adjust to the MapReduce condition SpatialHadoop [12] includes new spatial indexes, indexes files succeed an impediment in Hadoop which bolsters just non-filed store documents. There are 2 confinements when we utilize the Hadoop spatial indexes. Initially these records are outlined only for the procedural programming approach yet the SpatialHadoop utilizes the MapReduce programming method. Second, customary files are only intended for nearby record frameworks while then again SpatialHadoop utilizes the Hadoop File System (HDFS), which is default restricted as documents can be composed in add mode just, and once it is composed, the substance can’t be changed. To succeed such obstructions, SpatialHadoop orchestrates its indexes in two layers, to be specific local and global indexes. The global indexes parts information crosswise over node in the group and the local index then again organized information dwelling inside every node. The qualification of global and local indexes makes the MapReduce programming strategy less demanding in light of the fact that the global index is utilized for arrangement of the MapReduce work and the neighborhood records are utilized to handle the guide undertakings. Breaking the document into littler segments permits to file each parcel independently in memory and composing it to a record in a consecutively. The records are made accessible to MapReduce programs by presenting two moderately new modules in the MapReduce layer, to be specific, SpatialFileSplitter and SpatialRecordReader. SpatialFileSplitter takes as info an information record which is spatially filed and a channel work characterized by client and it misuses the worldwide to remove parcels immaterial to the appropriate response. SpatialRecordReader acknowledges a parcel that is privately ordered and which is returned by the channel capacity and endeavors its local index to get the records coordinating with client question.

QUERY PROCESSING
The new MapReduce [12] components and the efficient indexes from in the indexing layer gives the core of SpatialHadoop that opens the door to efficiently realize many spatial operations. Here, we present some case studies of 3 different types of operations, join operations, computational geometry and, basic operations. If the developers implement a similar techniques, they can include many such similar operations like reverse nearest neighbor operations or kNN join.

BASIC OPERATIONS
A large number of basic spatial operations [11] like k-nearest neighbor query and range query are supported by SpatialHadoop. A range query takes as input a query area A and a collection of spatial records R, and returns as output the overlapping records with A. It makes use of the global index with the SpatialFileSplitter to choose only those partitions which overlap with the query range A. Then, it makes use of the SpatialRecordReader for processing the local indexes in matching partitions and finding the matching records. At the end, a duplicate avoidance step removes the duplicate results formed from replication in the index. Even though this algorithm is relatively effective since it fast removes non-relevant partitions, it takes quite a lot of time for very small ranges due to the costs caused by Hadoop for beginning the MapReduce job. So, if the range of the query is too small and only few patterns are matching, this algorithm can be executed on a single machine without starting a MapReduce job, which responds with an interactive response.

CP-Attribute-based encryption
In the CP-ABE [3] is describe in each user is concomitant with a set of attributes and its private key is generated based on these attributes. When encrypting a spatial data SD, the user particularly an access structure which is expressed in terms of a set of selected attributes for SD. The spatial data is then encrypted based on the access structure such that only those whose attributes satisfy this access structure can decrypt the spatial data. Unauthorized persons are not able to access and hack the spatial data. In the access structure is sent in plaintext. CP-ABE [10] has the four algorithms
THE CP-ABE POLICY HAVING FOUR ENTITIES

**Key Generator:** The key era focus, which is completely confidential by every single other member of the framework. It creates open constraints and the master key, which are the essential key materials for the whole procedure of the proposed conspire. It additionally produces client particular private keys which compare to the arrangement of attributes for spatial data get to, ciphertext converted to plain text for user, and unknown keys for spatial data owner.

**Cloud Service provider:** The cloud service provider gives spatial data storage and retrieval to clients. The spatial data owner outsourced in the storage spatial data so the user easily access and retrieve the spatial data in effectively. We are accept that the cloud Service Provider is trustworthy, which implies that it takes after the convention indicated in the framework.

**Spatial Data owner:** The volume endorser who needs to transfer its spatial data satisfied secretly to the dispersed framework storage after encryption. The encrypted data can be imparted to expected user who has adequate qualifications as indicated by the spatial data owner.

**user:** which inquiries the Service provider for encrypted spatial data in the capacity framework by utilizing an alias the spatial data owner. Just users who have legitimate rights fulfilling the access policy indicated by the spatial data owner can get to the encrypted substance and reestablish the original data from it.

THE CP-ABE POLICY ALGORITHM

1) *Operation:* This algorithm that yields a security factor as input, and outputs the common factors $P_k$ and a key (master) $K_M$. $P_k$ is used for encryption and $K_M$ is used to produce user keys and is recognized only to the administrator.

2) *Encryption:* The input spatial data $SD$, an access structure $A$, and the public factors $P_k$. It outputs the encrypted text $ET$.

3) *Key Generator:* This algorithm yields as input the set of a user (say $X$)’s attributes $X_s$, the key (master) $K_M$ and outputs a secret key $K_S$ that finds with $X_S$.

4) *Decryption:* The input the encrypted text $ET$, a secret key $SK$ for an attribute set $X_s$. If $X_s$ satisfies the access structure inserted in $ET$, it will coming back the original spatial data $SD$. 

![Figure 2: Attribute-based Spatial Data retrieval in Cloud Environment](image)
SECURITY CONCERNS

1. Data Confidentiality: whether the spatial data stored on cloud storage should be protected against unintended or unauthorized access which does not have legal rights to them. Since cloud has used to manage sensitive spatial data it’s often a measure of the ability of the cloud to protect its spatial data.

2. Access control: service protects against unintended or unauthorized access to spatial data. Access is further qualified in terms of read, write and execute rights. Access control service works in tandem with authentication service since conceding entitled access right to an entity requires prior authentication of the entity.

3. Data availability: the procedure of certify that spatial data is available to end users and its applications, when and where they need it. It is important to confirm that the spatial data concerned is access to the authorized viewer at any times. Selected types of security attack attempt to deny access to the suitable user.

4. Data protection: the process of protect the spatial data from the malware, any attacks and nature or man-made disaster.

CONCLUSION

The proposed work provides a secured and efficient way of storing and retrieving spatial data in public cloud environment. This demonstrates the scalability of execution of substantial scale queries on spatialHadoop. Similarly the framework gives moderateness and simplicity to share the data which furnishes decision successfully with high confidence in the cloud. The improvement in the capacity of storage, integration and correlation of information in a productive way has additionally been managed.

REFERENCE


[4]. Wang, Y., & Wang, S.”Research and implementation on spatial data storage and operation based on Hadoop platform” Geoscience and Remote Sensing (IITA-GRS), Second IITA International Conference on (Vol. 2, pp. 275-278), 2015 IEEE.

[5]. Eldawy, A., Alharthi, S., Atzaidy, A., Daghistani, A., Ghanai, S., Basalamah, S., & Mokbel, M. F. A “demonstration of Shahed: A MapReduce-based system for querying and visualizing satellite data”. In 2015 IEEE 31st International Conference on Data Engineering (pp. 1444-1447), 2015,IEEE.


[7]. Quan, Q., Tian-hong, W., Rui, Z., & Ming-jun, X. “A model of cloud data secure storage based on HDFS”. In Computer and Information Science (ICIS), 2013 IEEE/ACS 12th International Conference on (pp. 173-178), 2013, IEEE.


