Hive Based Geospatial Analysis for Tracking and Envisioning of Geospatial Data in Hadoop Environment

Asha Kiran M, M. Sreedevi

***ABSTRACT---** Nowadays, the spatial data has gained prevalence as it has become an emerging subject in the technological world. It deals with the geographical location, boundaries, and features on the earth. To handle this spatial data, varied technologies and tools are available but are limited to some constraints. Different GIS tools are getting used to create and handle the spatial data for visualization. However, the outcomes have been unsatisfactory when it comes to handling huge data and analyzing the data by far. In this paper, as an improvement, a Hive-based spatial analysis has been proposed in Hadoop ecosystem to handle the spatial data, in fact, it can be called as spatial big data, as Hadoop can process a huge amount of data. Both GIS and Hadoop are integrated here to produce efficient outcomes, i.e.., in a pliable manner.*

***KEYWORDS:** GIS (Geographic Information System), Urbanization, Geospatial data, Hadoop, Hive, ArcMap, ArcPy.*

1. INTRODUCTION

The geospatial data is widely being used by different developers, analysts to develop any place or an area. But, according to the previous technologies, these analyses are not sufficient to utilize the resources to the great extent. proper data gathering, designing is not made to ensure the precise analysis and visualization of the data. Different developers are using different technologies for data visualization and analyzation. but each technology has their own disadvantages that cannot overcome easily. "In recent years, Geospatial data is being evolved to identify the geolocation, land of a particular area and the resources available in that area. It reflects several aspects such as economic development, regional development, and social change. This ultimately makes an area into urbanization. But there is a problem that affects the quality of analysis and visualization. Gathering the spatial data, cleaning and filtering the data. These are difficult tasks among the whole process.

There is another glitch that can make the whole process more complicated. for example, the spatial data with less amount can be handled easily. But the data is always being generated day by day [1].so it is very difficult to handle such a huge amount of data. That data is known as “Big Data” [6].

There are numerous technologies available to handle the big data. one such technology is HADOOP [9]. It allows the data to store and process to produce the desired results [2]. Because it segregates the main task into several subtasks for efficient results [3].

The geospatial data is often now called geospatial big data because of its size and structure of the data.

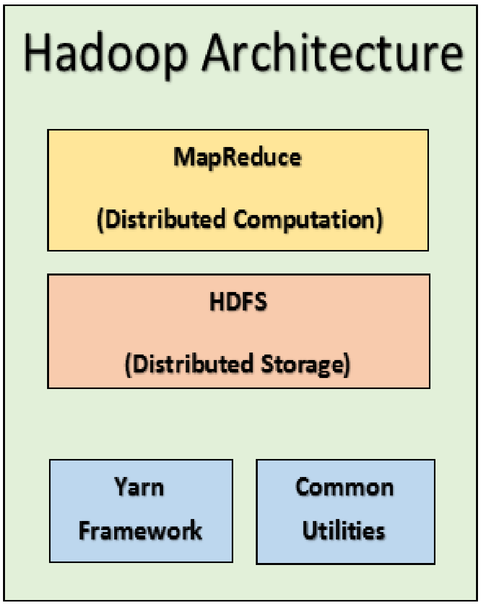

Fig 1: HADOOP Architecture

2. LITERATURE SURVEY

A survey has been conducted a geospatial database has been developed for assessing the maps in an area to monitor that area based on some map visualizations for urban planning. This development has been made based on GIS and remote sensing satellites. For easy accessing a GUI has been developed to incorporate several components of the geospatial database and to implement under GIS environment [8].

For the development of geospatial database software called ArcGIS 8.3, GIS software has been used. The land cover map has been acquired from IRS-1D LISS-III data of 2nd Dec. 2000 using digital image processing techniques. A GUI has been implemented using Visual Basic to develop a menu-driven interface. But there are some disadvantages in visual basic that is visual basic is a platform dependent.

]

Revised Manuscript Received on March 08, 2019.

Asha Kiran M, M. Tech Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India (Email: asha3577@gmail.com)

Dr. M. Sreedevi, Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India (Email: msreedevi_27@kluniversity.in)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

3. PROPOSED METHODOLOGY & RESULTS

In the proposed methodology, a system has been developed based on some techniques, which are helpful in such a way that it provides desired results in the form of representation of maps and geospatial analysis for better future planning.

3.1 OVERVIEW OF THE SYSTEM:

The principal of the below system is a two-part subsystem that mainly consists of two technologies: one such technology is a query engine, here we call it as spatial query engine. This technology provides diverse queries of spatial data with peerless access methods which are performed in map reduce framework. The spatial queries are interpreted into string of MapReduce which is a default processing code run by java.

3.1.1 Query Language:

Hadoop query languages such as HIVE, PIG both are acquiring ample momentum in the recent times. From naive based developers to experts such as application developers as well as data scientists feel more comfortable with the sql rather than other complex programming languages [11].so here HQL (HIVE Query Language) is used that yields easy to use interface. The following queries show that HQL queries are same as SQL to create a table.

```
1: CREATE TABLE
2: ROADNETWORK (OSMID 3: BIGINT, NAME STRING, AREA BIGINT)
4: ROW FORMAT DELIMITED FILEDS
5: TERMINATED BY '/t';
```

Fig2: schema of a table in HQL

3.1.2 LOAD THE DATA:

In HIVE,data can be loaded in two ways.i.e.., users can load the data either from local path or from HDFS path.

```
1: LOAD DATA [LOCAL] INPATH '/data/road.csv
2: [OVERWRITE] INTO TABLE roadnetwork;
```

Fig 3: command for data loading

3.1.3 QUERYING THE DATA:

For querying the data,users can write sql like commands effortlesly by using select statements,functions etc, into the HIVE shell.

```
1: SELECT osm_id, name,area from
2: roadnetwork where osm_id=567398;
```

Fig 4: commands for querying the data

3.2 INTEGRATING HIVE AND ARCMAP

Another technology is a GIS which deals with the queried spatial data. These queried spatial data are loaded into the GIS to view and analyse the data.ArcPy is an inheritor to ArcGIS.In Arcpy ,python is used as a scripting language which access all functions and modules.These helps us automate the GIS jobs.Different functions and classes are used in Arcpy to create the objects.

3.2.1 Classes in Arcpy

out of all the available classes,spatialReference and Extent classes are often used in geoprocessing tools.After the classes are instantiated,the methods(constructors) and properties of classes are used .These constructors are helpful in initializing the new instances of a class.Below code shows that creating an object "SpatialRef" for a SpatialReference(prj file) constructor.

```
import arcpy
prjFile = "c:/data/America Equidistant Conic.prj"
spatialRef = arcpy.SpatialReference(prj)
```

Fig 5: creating an object using SpatialReference class.

Some other important classes used in Arcpy are:

category	*Class name*
cursor	cursor
Environments	env
General	Array
charts	chart
Environments	EnvManager
Exceptions	ExecuteError
Exceptions	ExecuteWarning
General	Extent
FutureSet	FutureSet

Fig 6: Table with list of classes

3.2.2 Functions in Arcpy

A function in Arcpy consists of a functionality that perpetrates a specific task. These functions can be incorporated into large codes and are useful for validating a table name, retrieving the properties of a dataset. As other functions, it takes arguments and returns some value.

```
import arcpy
input= arcpy.GetparameterAsText(0)
if arcpy.Exists(input):
   Print ("data is existed")
else:
Print ("data is not existed")
```

Fig 7 : Using functions in Arcpy

Other functions used in Arcpy :

Category	*Function name*
Datastore	AddDataStoreItem
Fields	AddFieldDelimiters
Messages and Error Handling	AddMessage
Tools and toolboxes	AddToolBox
General	command

Fig 8: Table with list of functions

From the use of above classes and functions data can be manipulated and customized based on user requirements.These manipulated data is then loaded into ArcMap for viewing and analysing the data.

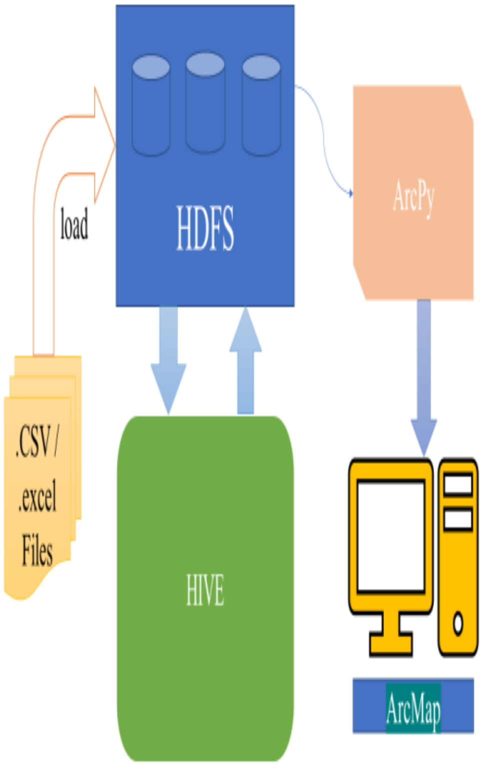

Fig 9: workflow of the proposed system.

Major building blocks of the proposed system:

HADOOP: In our model, we impart a software framework which stores both structural and unstructured data [4]. It has immense processing power by map-reduce that handles limitless tasks [5]. It helps us to process any amount of data (.csv files) for yielding the desired analysis.

HDFS: HDFS is a dominant chunk in all Hadoop ecosystems. It stores the data like other databases that provides well-grounded means to manage the big data [10]. Here, a pool of CSV files is loaded into the HDFS for storing.

HIVE: HIVE is a Datawarehouse tool for processing the CSV files (structured data). Each CSV file is stored as tables. It furnishes SQL like interface which is very simple and trouble-free software for querying the tables. HIVE helps the users to write as many queries as the user wants for analyzing the data to get achieved results.

ArcPy: ArcPy is a successor to the ArcGIS scripting module. It is a python site package that provides numerous functions and classes used for data conversion, geographical data analysis. It assists the users in converting .csv files (from Hadoop, after cleaning and filtering the data) to the .shp files which are most suitable for map representation. It acts as an integration tool for Hadoop and ArcMap.

ArcMap: ArcMap is a crucial component in ArcGIS for processing Geospatial data. Here, ArcMap used to take the data from ArcPy for creating, editing and viewing the data.

Steps to follow and obtain result for the proposed system:

1) Initially, all the required files (shapefiles) needs to be taken.Each shapefile holds a number of files including .SHP, SHX, PRJ, .DBF files . out of all the files .dbf files with .csv extension are loaded into the HDFS for further processing.
2) Now, tables are to be created for each CSV file for cleaning and filtering the data by writing the queries.
3) In the next step, the filtered data is then loaded into the ArcPy which is an integration tool for Hadoop and ArcMap. It provides the necessary classes and methods to convert the CSV files to the respective shp files. It provides an easy to view the data.
4) In the final step, the converted data is then loaded into the ArcMap to create the data, edit the data and view the data for map representation and analysis.

Advantages of the system:

1) This model helps us in analyzing any kind of data by following the above steps. Likewise,
2) It helps the users in vehicle tracking analysis with GPS data.
3) It is used for planning and analyzing the data of an area by taking the shapefiles. It helps the decision makers to identify the advantages and disadvantages by observing the road, railway, buildings, water supply etc.,
4) Each tool has its own benefits like HDFS stores any amount of data as it can handle. HIVE provides an easy and simple interface for querying the data to get the desired results.

CONCLUSION AND FUTURE SCOPE

A simple and efficient work has been done for planning, developing and analyzing by taking the data in any area either a city or a village or a state. For this, the number of input files has been taken for analyzing the data. Moreover, A Hadoop framework has been used for analyzing. Querying and filtering the data based on user requirements. And GIS tool has been used for maps visualization for better understanding. This research can be extended based on the user requirements in future for further planning, developing and analysing on Hadoop.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

REFERENCES

1. Krish Krishnan, “Data Warehousing in the Age of Big Data”
2. Raghavendra Kune, Pramod Kumar Konugurthi, Arun Agarwal, Raghavendra Rao Chillarige, and Rajkumar Buyya, “The anatomy of big data computing,” Software: Practice and Experience, vol. 46.
3. Ahmed Eldawy, M Mokbel, and Christopher Jonathan, “Hadoop viz: A MapReduce framework for extensible visualization of spatial data,” in IEEE Intl. Conf. on Data Engineering (ICDE), 2016.
4. Ahmed Eldawy and Mohamed F Mokbel, “Spatial Hadoop: A MapReduce framework for spatial data”.
5. Rakesh K. Lenka1, Rabindra K. Barik2, Noopur Gupta1, Syed Mohd Ali1, Amiya Rath3, Harishchandra Dubey4” Comparative Analysis of Spatial Hadoop and GeoSpark for Geospatial Big Data Analytics”.
6. Hervais Simo Fraunhofer-Institut für Sichere Information technology, Darmstadt, Germany” Big Data: Opportunities and Privacy Challenges”
7. S. Kazemi , S. Lim, C. Rizos School of Surveying & Spatial Information Systems, the University of New South Wales, Sydney, NSW 2052, Australia" A review of a map and spatial database generalization for developing a generalization framework"
8. Anuj Bariar1 *, R.D. Gupta2, S.C. Prasad3 Department of Civil Engineering, Motilal Nehru National Institute of Technology (MNNIT), Allahabad-211004, U.P.,” Geospatial database development for urban planning using Satellite data under GIS environment”.
9. Sunil Pratap Singh and Preetvanti Singh Department of Physics and Computer Science.” Modelling a geospatial database for managing travelers demand”.
10. G. Vijay Kumar[1], M. Sreedevi[2,] K. Bhargav[3], Mohan Krishna[4] “Incremental Mining of Popular patterns from Transactional Databases.
11. Ablimit Aji Xiling Sun# Hoang Vo Oioaling Liu Rubao Lee Xiaodong Zhang Joel Saltz Fu sheng Wang MathCS and BMI, Emory University.” Demonstration of HADOOP GIS: A Spatial Data Warehousing System Over MapReduce”.