Table of Contents

Abstract

Spatial Data Mining (SDM) technology has emerged as a new area for spatial data analysis. Geographical Information System (GIS) stores data collected from heterogeneous sources in varied formats in the form of geodatabases representing spatial features, with respect to latitude and longitudinal positions. GIS analysis is the process of modeling spatially, deriving results by computer processing, then examining and interpreting those model results. Spatial analysis is useful for evaluating suitability and capability, estimating and predicting, interpreting and understanding, and much more.

Acknowledgements

We would like to express our sincere gratitude to our supervisor **Mr. Abdul Qadeer** for providing their invaluable guidance, comments and suggestions through the course of the project. We would thank him for constantly motivating us to work harder.

I am immensely obliged to my seniors and friends for elevating inspiration, encouraging guidance and kind supervision in the completion of my project.

Mohammad Kamran

Saeem Ansari

26 November 2018

List of Figures

Chapter 1 : Introduction

The GIS tools for hadoop are a collection of GIS tools that leverage the Spatial Framework for Hadoop for spatial analysis of big data. The tools make use of the Geoprocessing Tools for Hadoop toolbox, to provide access to the Hadoop system from the ArcGIS Geoprocessing environment. Geographic Information System (GIS) has emerged as a new discipline due to the development of communication technologies. SDM techniques are widely used in GIS for inferring association among spatial attributes, clustering, and classifying information with respect to spatial attributes.

Analysis of data deposited in GIS has gained importance in domains related to knowledge management and data mining. The development of information and communication technologies in GIS Domain has generated huge volume of data representing spatial information of water bodies, forest reserves, urbanization, etc.,

1.1 Motivation

Spatial data are today needed in a wide range of application domains. Indeed, spatial properties are included in several application contexts requiring the management of very large data sets, such as, for instance, computer-aided design (CAD), very large scale integration (VLSI), robotics, and image processing. However, the primary target of systems dealing with spatial data remains geographical application,since they served as the first motivation for the development of such technology and still represent the most challenging application environment.

1.2 Objectives and Scope

Spatial Analysis is a module in the field of quantitative geography and geocomputation
At the same time this module provides the technical tools and skills to study the spatial dimension of various spatial phenomena from a geographic perspective. And have practical experience in applying spatial analysis methods in addressing geographical issues in the real world using special software. And also have an overview of modern methods of spatial analysis used in the industry and in research projects in which the science of geography plays an important role.such as finding places within study area having highest number of trips with common origin and destination using GPS.

1.3 Organisation

The report is organized as follows.Chapter 2 deals with Literature Review of the report in Which We discussed our progress done till now collecting the spatial data set for our project to learn the various GIS tools and use hadoop with technology map reduce.

Chapter 2 : Background Study

2.1 Literature Survey

Spatial data can be defined as pieces of information describing quantitative and/or qualitative properties that refer to space. Such properties can be represented as attributes of a set of objects
(like the path of a given highway or the technical drawing of the new version of a car engine). spatial data in geographical applications; thus it is focused on geographical data. This means that spatial data are used to describe objects or, more generally, natural phenomena and human activities that occur on the Earth's surface

Analysis of data deposited in GIS has gained importance in domains related to knowledge management and data mining. Recent widespread use of spatial databases has lead to the studies of Spatial Data Mining (SDM), Spatial Knowledge Discovery (SKD), and the development of SDM techniques.

2.2 Approach

Huge amount of data has been generated as a result of existing automated computer applications.

The MapReduce part of Hadoop abstracts all the details of parallel processing from the user and the user gets a very simplified framework for programming. The framework consists of mapper and reducer components that work on key-value pair concept.

The spatial data-set is input to the Hadoop cluster into the HDFS and queries are implemented in parallel on the distributed data. The output for the parallel query is taken and analyzed on the ArcGIS system.

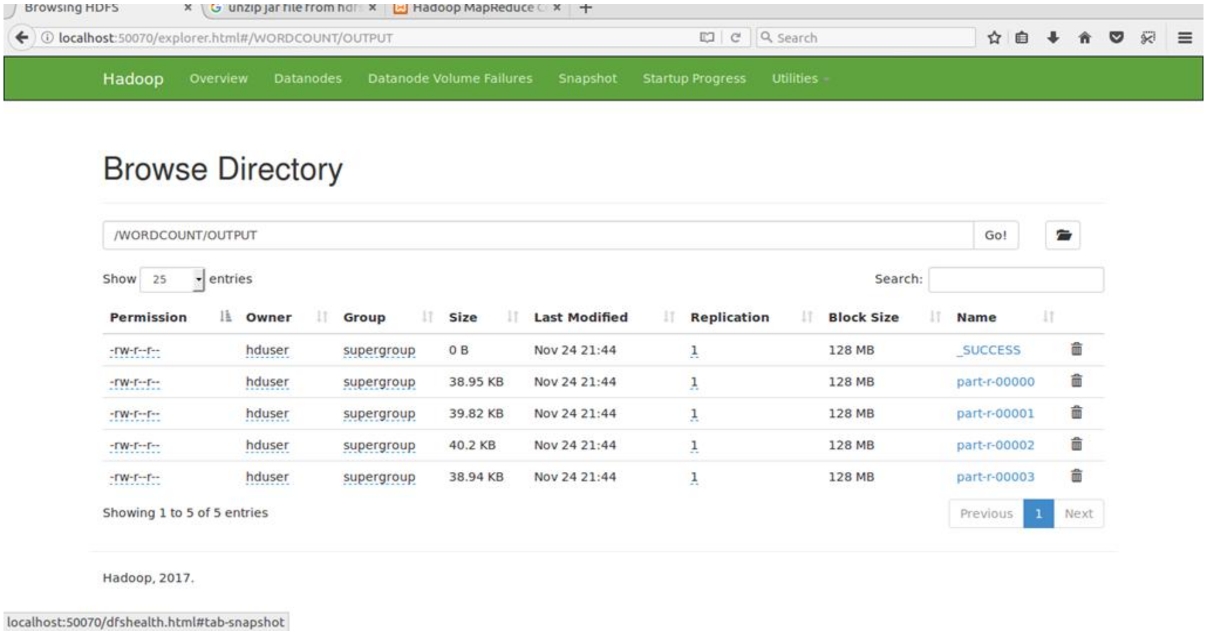
Chapter 3 : Overall Description

3.1 Overview

There has been a tremendous growth in the usage of Geographic Information System (GIS) in a variety of fields. It is due to the attractive pictorial and real time result display of the GIS software that provides an excellent decision support system for analysis. A large amount of data is available in a variety of formats, such as maps or images that can be integrated with other GIS data for GIS processing. The traditional sequential ways of spatial data processing lags in the efficiency of executing the queries. The modern parallel processing technique, the MapReduce, is used extensively for big data analysis. And analyzes the synthetic large spatial data-set on the MapReduce and ArcGIS to check the similarity of the outputs generated through the parallel framework and the specialized GIS software - ArcGIS. Firstly, the input data-set is processed in parallel for queries through Map and Reduce functions, and the query results are displayed on ArcGIS. Secondly, on the ArcGIS, the addresses are integrated with spatial data, through the geo-coding process that assigns addresses to locations on a map, and outputs a shapefile for display

3.2 Work Done

3.2.1 Installation of Hadoop



Browse Directory

/WORDCOUNT/OUTPUT Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hduser	supergroup	0 B	Nov 24 21:44	1	128 MB	_SUCCESS
-rw-r--r--	hduser	supergroup	38.95 KB	Nov 24 21:44	1	128 MB	part-r-00000
-rw-r--r--	hduser	supergroup	39.82 KB	Nov 24 21:44	1	128 MB	part-r-00001
-rw-r--r--	hduser	supergroup	40.2 KB	Nov 24 21:44	1	128 MB	part-r-00002
-rw-r--r--	hduser	supergroup	38.94 KB	Nov 24 21:44	1	128 MB	part-r-00003

Showing 1 to 5 of 5 entries

Hadoop, 2017.

localhost:50070/dfshealth.html#tab-snapshot

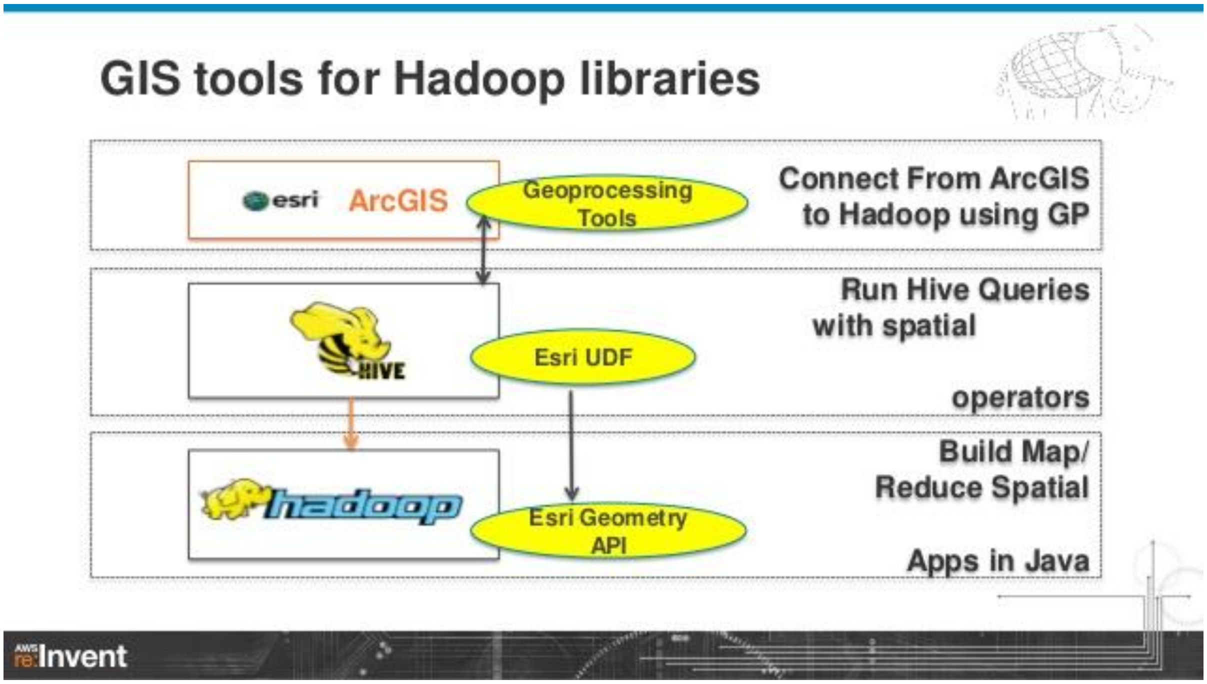
Fig 3.1

3.2.2 Working with ArcGis

ArcGIS is a geographic information system (**GIS**) for working with maps and geographic information. It is used for creating and using maps, compiling geographic data, analyzing mapped information, sharing and discovering geographic information, using maps and geographic information in a range of applications, and managing geographic information in a database.

The system provides an infrastructure for making maps and geographic information available throughout an organization, across a community, and openly on the Web.

3.3 System Architecture



GIS tools for Hadoop libraries

Connect From ArcGIS to Hadoop using GP

Run Hive Queries with spatial operators

Build Map/ Reduce Spatial Apps in Java

Fig 3.2

Chapter 4 : Tools and Technology

The requirement for this project include some software specifications which are: -

4.1 Spatial Framework (Hadoop based) :

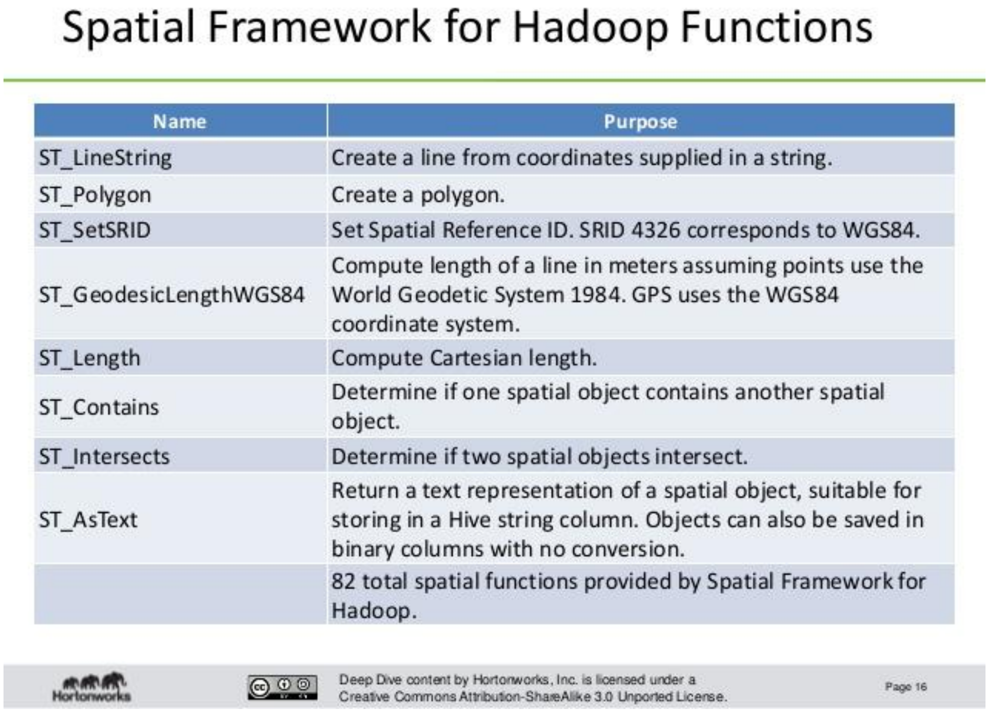
4.1.1 Spatial Data

Also known as geospatial data *or* geographic information it is the data or information that identifies the geographic location of features and boundaries on Earth, such as natural or constructed features, oceans, and more. Spatial data is usually stored as coordinates and topology, and is data that can be mapped. Spatial data is often accessed, manipulated or analyzed through Geographic Information Systems (GIS).

4.1.2 Spatial Framework used for hadoop

The Spatial Framework for Hadoop allows developers and data scientists to use the Hadoop data processing system for spatial data analysis.

The following are the hadoop functions used for Spatial Framework :-



Spatial Framework for Hadoop Functions

Name	Purpose
ST_LineString	Create a line from coordinates supplied in a string.
ST_Polygon	Create a polygon.
ST_SetSRID	Set Spatial Reference ID. SRID 4326 corresponds to WGS84.
ST_GeodesicLengthWGS84	Compute length of a line in meters assuming points use the World Geodetic System 1984. GPS uses the WGS84 coordinate system.
ST_Length	Compute Cartesian length.
ST_Contains	Determine if one spatial object contains another spatial object.
ST_Intersects	Determine if two spatial objects intersect.
ST_AsText	Return a text representation of a spatial object, suitable for storing in a Hive string column. Objects can also be saved in binary columns with no conversion.
	82 total spatial functions provided by Spatial Framework for Hadoop.

Hortonworks

Page 16

Fig 4.1

4.2 Geoprocessing Tools for Hadoop

GIS Tools for Hadoop is an open source toolkit that brings spatial analysis to your big data. The toolkit is composed of four projects:

4.2.1 The Building Blocks

Esri Geometry API for Java: This library includes geometry objects (e.g. points, lines, and polygons), spatial operations (e.g. intersects, buffer), and spatial indexing. By deploying the Esri geometry API library (as a jar) within Hadoop, you are able to build custom MapReduce applications using Java to complete analysis on your spatial data.

4.2.2 The Framework

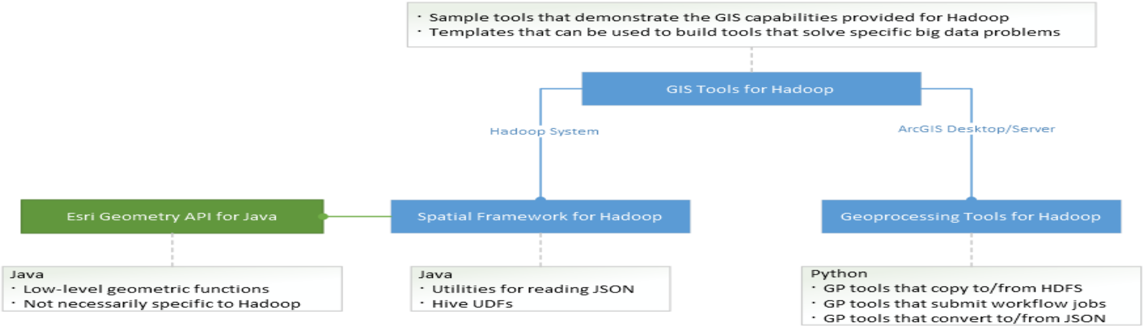
Spatial Framework for Hadoop: This library includes user defined functions (UDFs) that extend Hive and are built upon capabilities of the Esri Geometry API.This allows you to avoid complicated MapReduce algorithms and stick to a more familiar workflow.

4.2.3 The Connector

*Geoprocessing Tools for Hadoop***:** These tools are downloaded as a toolbox and applied in ArcMap – recreating a typical workflow for an ArcGIS user. Using these tools, you can connect data between Hadoop and ArcGIS, submit workflow jobs, and convert data to and from JSON. You can then transport your Hadoop results into ArcGIS for visualization.

4.2.4 The Toolkit

GIS Tools for Hadoop: This project synthesizes the above three projects into the toolkit. It includes samples and instructions that leverage the complete toolkit. The samples are available to help test your deployment of the spatial libraries.



· Sample tools that demonstrate the GIS capabilities provided for Hadoop
· Templates that can be used to build tools that solve specific big data problems
GIS Tools for Hadoop
Hadoop System
ArcGIS Desktop/Server
Esri Geometry API for Java
Spatial Framework for Hadoop
Geoprocessing Tools for Hadoop
Java
· Low-level geometric functions
· Not necessarily specific to Hadoop
Java
· Utilities for reading JSON
· Hive UDFs
Python
· GP tools that copy to/from HDFS
· GP tools that submit workflow jobs
· GP tools that convert to/from JSON

Fig 4.2

4.3 Access to a Hadoop cluster :

4.3.1 Introduction

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Hadoop's infinitely scalable flexible architecture (based on the HDFS filesystem) allows organizations to store and analyze unlimited amounts and types of data—all in a single, open source platform on industry-standard hardware.

4.3.2 Hadoop Distributed File System (HDFS)

HDFS supports the rapid transfer of data between nodes. At its outset, it was closely coupled with MapReduce, a programmatic framework for data processing.

When HDFS takes in data, it breaks the information down into separate blocks and distributes them to different nodes in a cluster, thus enabling highly efficient parallel processing.

4.3.3 Map-Reduce

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce.

Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

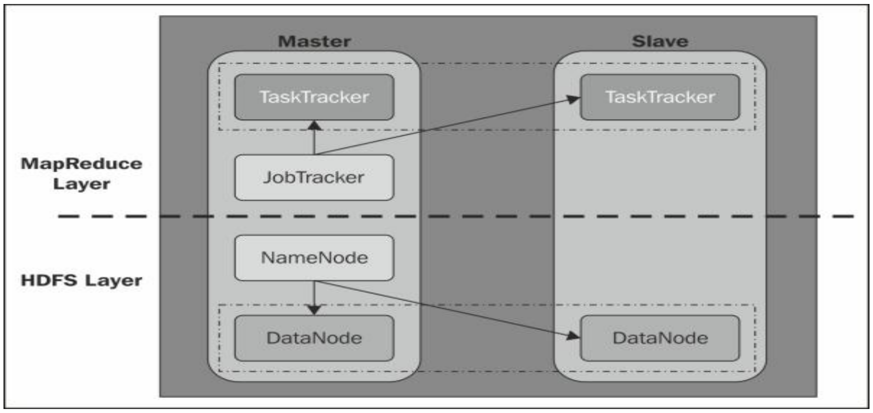


Fig 4.3

Chapter 5 : Conclusions and Future Work

5.1 Conclusions

Maps can be quite beautiful. They stimulate both sides of our brain: the right side that’s intuitive and aesthetic, and the left side that’s rational and analytical. Maps are this wonderful combination of both. It’s this neat marriage of utility and beauty that makes it alluring.

ArcGIS is rapidly emerging as the platform of choice for the creation and dissemination of authoritative geographic data content.

Spatial analysis is the most intriguing and remarkable aspect of GIS. Using spatial analysis, you can combine information from many independent sources and derive new sets of information (results) by applying a sophisticated set of spatial operators.

Spatial Framework for hadoop makes geo analytics simple with processing power of Hadoop.

This project integrates ArcGIS with Hadoop using a suitable toolkit.

More specifically, the tools provided allow ArcGIS users to run Hadoop workflow jobs and permits the exchange of data between ArcGIS geodatabase and a Hadoop system.

5.2 Future Work

The Future of ArcGIS coupled with the processing power of hadoop in the coming ages is challenging and bright when we would be having huge amounts of spatial data to deal with.

Through these tools, ArcGIS users can do more complex and sophisticated analysis that narrow their data to a specific subset. The following ideas would be implemented :

- Convert between Feature Classes in a Geo-database and JSON formatted files.
- Copy data files from ArcGIS to Hadoop, and copy files from Hadoop to ArcGIS.
- Run an Oozie workflow in Hadoop, and to check the status of a submitted workflow.
- Furthermore, users can leverage the ArcGIS platform capabilities to publish their maps to the web and mobile applications
- These new data would be in need of algorithms that could extract meaningful content from them and give rise to new ideas and solutions that could help change the society in a better way.

References

[1] Zhu Deng , Yuqi Bai " Floating car data processing model based on Hadoop - Gis Tools ", 2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)

[2] Yonggang Wang, Sheng Wang," Reasearch and implementation on spatial data storage and operations based on Hadoop Platform.

[3] Han Singh , Seema Bawa , "Spatial Data Analysis with ArcGIS and MapReduce".

[4] Manfred M.fischer , JInfeng Wang - "Spatial Data Analysis Models,Methods and Techniques"-Springer.

[5] Bhuvaneswari Velmani, Anantha Sadhasivam ," Spatial Data Mining Approaches for GIS - A Brief Review".

[6] Github Repository https://esri.github.io/gis-tools-for-hadoop/

**