

Performing operations on GIS by using Hadoop for making multi node cluster and ARCGIS for visualizing

Prepared at



ISO 9001:2008
ISO 27001:20013
CMMI LEVEL-5

Bhaskaracharya Institute for Space Applications & Geo-informatics

Science & Technology Department, Govt. of Gujarat.

Gandhinagar

Prepared By

Anant Sharma
2014A7PS051G

Dhruv Passey
2014A7PS020G

Mahesh Hada
2014B3A7963G

Guided By

Prashant Chauhan
BISAG Gandhinagar
Project Coordinator

Mr. Gavax Joshi
PS 1, Instructor
BITS Pilani, Goa Campus

Submitted To

Birla Institute of Technology and Science



DEPARTMENT OF COMPUTER SCIENCE
BITS, PILANI – 333 031

Bhaskaracharya Institute for Space Applications and Geo-informatics



ISO 9001:2008
ISO 27001:20013
CMMI LEVEL-5

Department of Science & Technology

Government of Gujarat

Phone: 079 - 23213081 Fax: 079-23213091

E-mail: info@bisag.gujarat.gov.in, website: www.bisag.gujarat.gov.in

CERTIFICATE

*This is to certify that the project report compiled by **Mr. Anant Sharma, Mr. Mahesh Hada and Mr. Dhruv Passey** students of 4th Semester **B.E. (Hons.)** from **Department Of Computer Science, Birla Institute of Technology and Science**, have completed their Practice School-I project satisfactorily. To the best of our knowledge this is an original and bonafide work done by them. They have worked on a software application for **“Performing operations on GIS by using Hadoop for making multi node cluster and ArcGIS for visualizing”**, starting from May 23rd, 2016 to July 16th, 2016.*

During their tenure at this Institute, they were found to be sincere and meticulous in their work. We appreciate their enthusiasm & dedication towards the work assigned to them.

We wish them every success.

Prashant Chauhan

Project Scientist

BISAG, Gandhinagar

T. P. Singh

Director

BISAG, Gandhinagar

Organization Profile

1. BACKGROUND

The applications of space technologies and geo-informatics contribute significantly towards socio-economic development of the society.

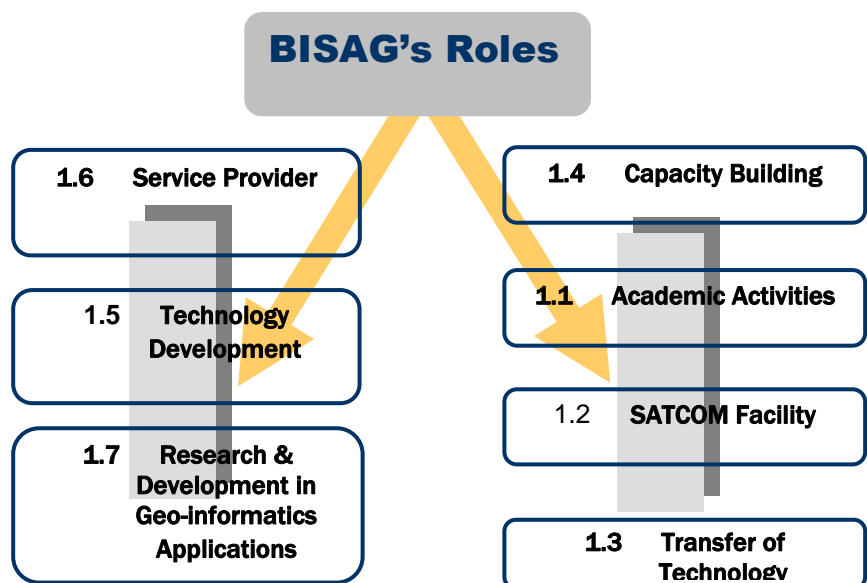


Recognizing the importance and need of Space technology and geo-informatics for developmental planning purposes, the Government of Gujarat established the Bhaskaracharya Institute for Space Applications and Geo-informatics (BISAG) in the year 1997, as the State nodal agency to utilize space technology and geo-informatics for various developmental activities of the State.

Since its foundation, the Institute has experienced extensive growth in the spheres of space technology and geo-informatics. The objective with which BISAG was established is manifested in the extent of services its renders to almost all departments of the State. Year after year the institute has been endeavoring to increase its outreach to disseminate the use of geo-informatics up to grassroots level. In this span of eleven years, BISAG has assumed multi-dimensional roles and achieved several milestones to become an integral part of the development process of the Gujarat State.

2. PROFILE

BISAG's has strengthened its role as a facility provider, a technology developer and as a facilitator for transferring technology to the grass root level.



Further reinforcing its functions, BISAG has achieved ISO 9001:2008 and ISO 27001:2005 certifications for quality management and security management services respectively. This has led to an organized and systematic development of its services and outputs.

3. ACTIVITIES OF BISAG

BISAG's activities are multi-fold and have expanded in a big way and focused on the following:

- ❑ **Satellite Communication** Promoting and facilitating the use of satellite broadcasting networks for distant interactive training, education and extensions
- ❑ **Remote Sensing** Inventory mapping, developmental planning and monitoring of natural and man-made resources
- ❑ **Geo-informatics System** Conceptualizing, creating and organizing multi-purpose common geo-spatial database for sectoral and thematic applications for various users
- ❑ **Photogrammetry** Creation of Digital Elevation Model, Terrain characteristics, Resource planning, etc.
- ❑ **Global Navigation Satellite System** Location based services, geo-referencing, engineering applications and research
- ❑ **Software Development** For providing low-cost Decision Support Systems, desktop as well as web-based geo-informatics applications to users for wider usage.
- ❑ **Disaster Management** For preparing geo-spatial information to provide necessary inputs to the Government to assess and mitigate extent of damage in the event of a disaster

❑ **Education, Research and Training**

For providing education, research and training facilities to promote number of end users through the Academy for Geo-informatics.

❑ **Value Added Services**

For providing services which can be customized as per the needs of the users.

❑ **Technology Transfer**

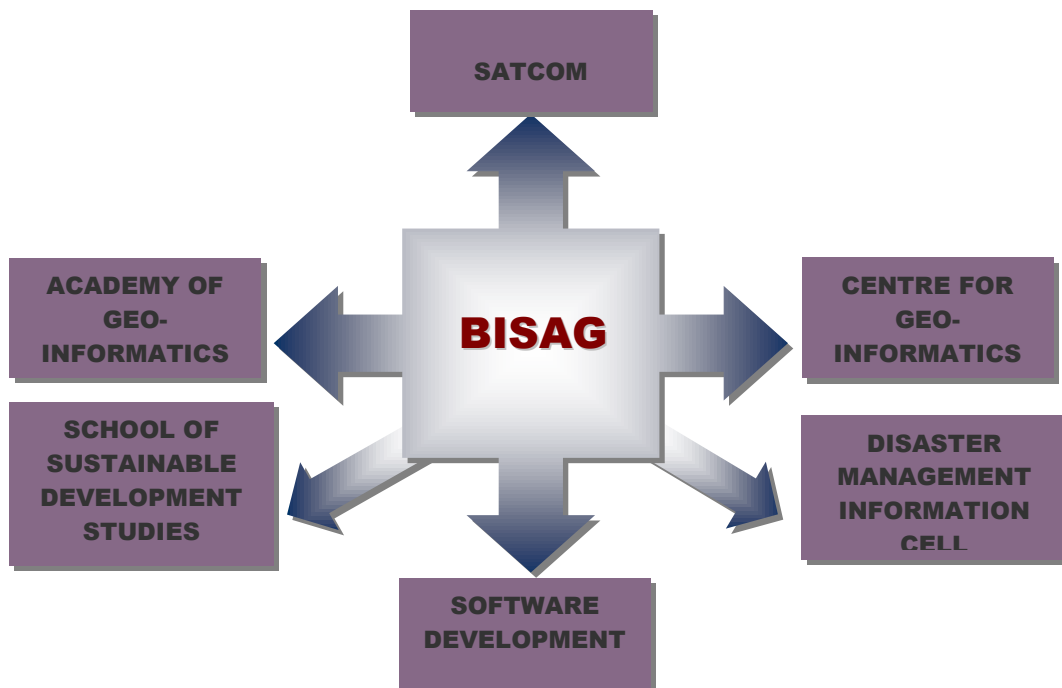
Transferring technology to a large number of end users.

4. UNITS OF BISAG

BISAG initially set up to carry out Space Technology applications, has evolved into *an Academic Institute, a Centre for Research and Technology Innovations, a Facility Provider, a Technology Developer and a Facilitator* for transferring technology to the grass root level. BISAG is the first such State Centre having such multifarious activities with ISO certification. BISAG has gradually progressed over the years and has grown into several units. Each unit focuses on specific functions and objectives to ensure efficiency in over all activities of the institute.

- **Gujarat Satellite Communication Network (GUJSAT):** SATCOM facilitates the promotion and facilitation of the use of broadcast and teleconferencing networks for distant interactive training, education and extension.
- **Centre for Geo-Informatics Applications:** The Centre for Geo-informatics provides services for the developmental and planning activities pertaining to Agriculture, Land and Water Resources Management, Wasteland/ Watershed development, Forestry, Disaster Management, and Infrastructure etc.
- **Software Development:** For wider usage of geo-spatial applications, customised software are developed by the Software Development Team. The institute has provided many indigenous software solutions in the field of Geographic Information Systems, Decision Support Systems and Image Processing.
- **Academy of Geo-informatics:** The Academy for Geo-informatics carries out Education, Research and Training activities.
- **Disaster Management Information cell:** BISAG works closely with the Gujarat State Disaster Management Authority (GSDMA), for assessment of existing situation through integrated analysis and for planning appropriate preventive and

preparatory measures, providing necessary support through data generation and analysis.



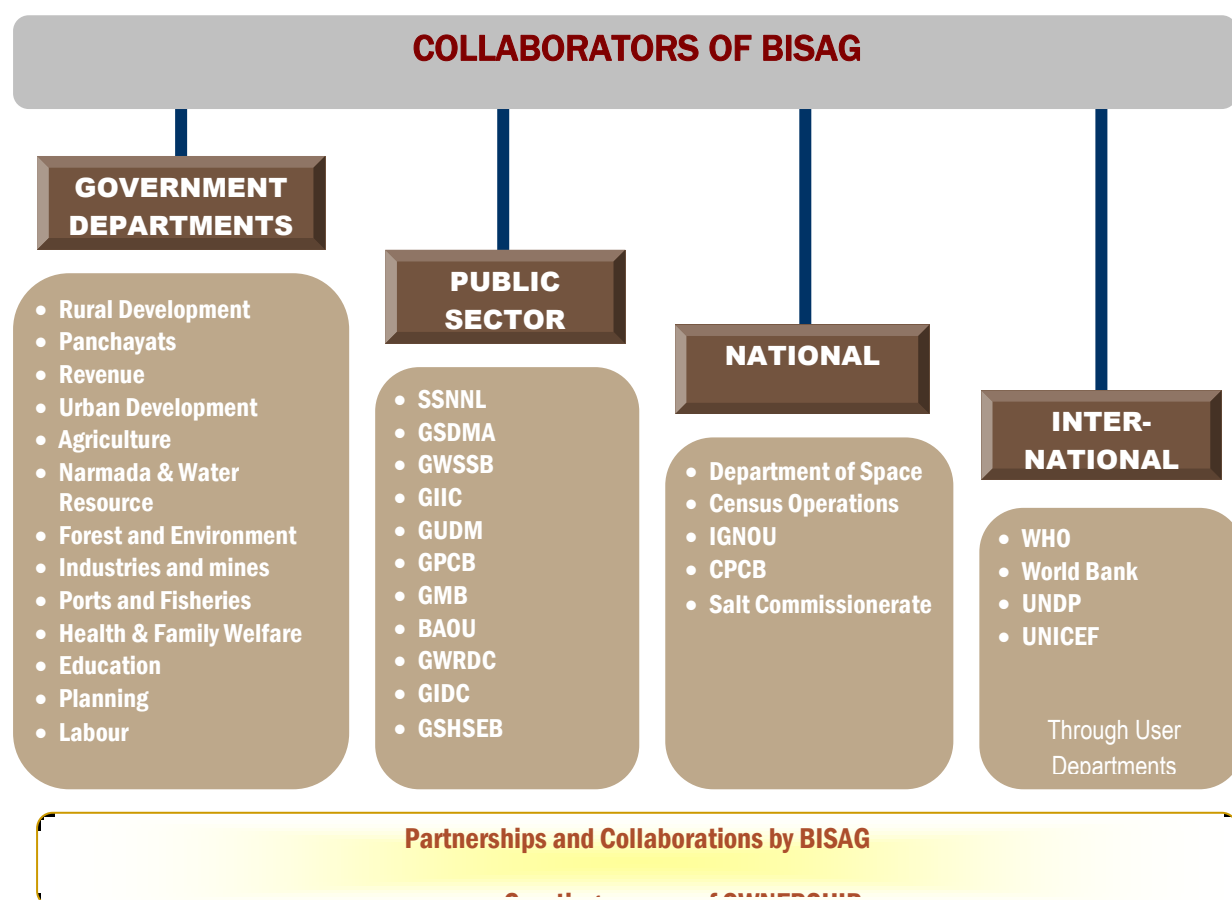
5. INFRASTRUCTURE DEVELOPMENT

The growth and progress of any institute is gauged by the infrastructure it develops and possesses. BISAG has a sound infrastructure setup that has developed in tandem with the growth of the institute. Having started with one building, there are now dedicated facilities for different units.

- 1.1** The laboratories are equipped with state-of the art technology with latest Hardware and Software required for executing its activities. BISAG also has a rich satellite data archive, which includes Satellite data of different spatial, spectral and temporal resolutions.

6. COLLABORATIONS OF BISAG.....Creating A Sense Of Ownership

BISAG works with almost all Government Departments and Organizations. Each of these Departments/Organization contributes in preparation of the respective projects. With strong Government support and proactive efforts on part of the staff of BISAG, the list of Collaborators is expanding and increasing.



7. INSTITUTIONAL STRENGTHENING

BISAG has achieved institutional strengthening through:

- **Reinforcement of Decision Support Systems**

Developing customized solutions as per user requirements through partnerships and collaborations, which are affordable and easy to use. Areas of natural and manmade resources, socio-economic parameters, are being effectively addressed with the help of Geo-informatics.

- **Establishing Linkage between Government and People through GUJSAT**

GUJSAT facility is being constantly employed for the promotion and facilitation of the use of teleconferencing networks for distant interactive training, education and extension. Experts, leaders, specialists and professionals can conduct their programs from a central location reaching out to remote areas through two-way audio-video channel making them interactive and meaningful.

- **Developing Innovative Education Programmes**

Innovative educational programmes are conducted regularly through GUJSAT, allowing people residing in remote areas to have an access to good quality educational and awareness programmes.

- **Solving real life problems through Human Resource Development**

The institute has a young multi-disciplinary team of professionals and a continuing induction programme. Multi-nationals and IT agencies pick up the trained staff that in turn is replaced by new people. This results in availability of more and more trained manpower in the realm of space applications. Every year BISAG provides training to about 300 students in the field of Geo-informatics.

- **Creation of the multipurpose sectoral comprehensive databases for the entire state of Gujarat**

The institute has made efforts towards conceptualization, creation and organization of multi-purpose common digital database for sectoral / integrated decision support

systems. This has provided impetus to planning and developmental activities at grass root level as well as monitoring and management potential in various disciplines like water resources, land resources, disaster management, infrastructure, urban management.

CANDIDATE'S DECLARATION

We declare that Practice School-1 report entitled “**Performing operations on GIS by using Hadoop for making multi node cluster and ARCGIS for visualizing**” is our own work conducted under the supervision of our guide **Mr. Prashant Chauhan** from BISAG (Bhaskaracharya Institute for Space Applications & Geo-informatics). We further declare that to the best of our knowledge the report for B.E. (Hons.) Practice School 1 does not contain part of the work which has been submitted for the award of Bachelor's Degree either in this or any other university without proper citation.

Candidate 1's Signature

Dhruv Passey

Student ID: 2014A7PS020G

Candidate 2's Signature

Mahesh Hada

Student ID: 2014B3A7963G

Candidate 3's Signature

Anant Sharma

Student ID: 2014A7PS051G

Submitted To:

Department Of Computer Science

BITS, Pilani

ACKNOWLEDGMENT

We are grateful to **Mr. T.P. Singh, Director (BISAG)** for giving us this opportunity to work the guidance of renowned people of the field of GIS Based Portal and also providing us with the required resources in the institution.

We would like to express our endless thanks to our external guide **Mr. Prashant Chauhan**, Project Scientist at Bhaskaracharya Institute of Space Applications and Geo-informatics for their sincere and dedicated guidance throughout the project development.

At this juncture I feel deeply honored in expressing my sincere thanks to **Mr. Viraj Choksi** of Bhaskaracharya Institute of Space Applications and Geo-informatics for making the resources available at right time and providing valuable insights leading to the successful completion of my project.

We would also like to express our hearty gratitude to our Head of Department **Mr. Bharat Deshpande** and our PS faculty **Mr. Gavax Joshi** for giving us encouragement and technical support on the project.

Anant Sharma

Student ID: 2014A7PS051G

Dhruv Passey

Student ID: 2014A7PS020G

Mahesh Hada

Student ID: 2014B3A7963G

PROJECT INDEX

1. Introduction.....	1
2. System Requirement Study.....	3
1. Hadoop.....	3
2. GIS tools for Hadoop.....	4
3. Hive.....	5
4. Geo processing tools for Hadoop.....	6
5. ArcGIS	6
6. QGIS	7
7. Virtualbox	8
8. NoSQL	9
9. JAVA	10
3. Methodology and Implemetation.....	11
4. Result and Conclusion.....	20
5. Bibliography.....	21

Chapter 1

Introduction

A **geographic information system (GIS)** is a computer system for capturing, storing, checking, and displaying data related to positions on Earth's surface. **GIS** can show many different kinds of data on one map. This enables people to more easily see, analyze, and understand patterns and relationships.

GIS is any information system that integrates, stores, edits, analyzes, shares, and displays geographic information. GIS applications are tools that allow users to create interactive queries (user-created searches), analyze spatial information, edit data in maps, and present the results of all these operations. Geographic information science is the science underlying geographic concepts, applications, and systems.

GIS is a broad term that can refer to a number of different technologies, processes, and methods. It is attached to many operations and has many applications related to engineering, planning, management, transport/logistics, insurance, telecommunications, and business. For that reason, GIS and location intelligence applications can be the foundation for many location-enabled services that rely on analysis and visualization.

However storing and processing large amount of geospatial data is not an easy task. So it requires some platform to process such data such as Apache Hadoop which is an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Thus, Hadoop multi node cluster is used for processing such large data sets and also it avoids the risk of failure of the operation if some datanode fails then some other datanode would be working.

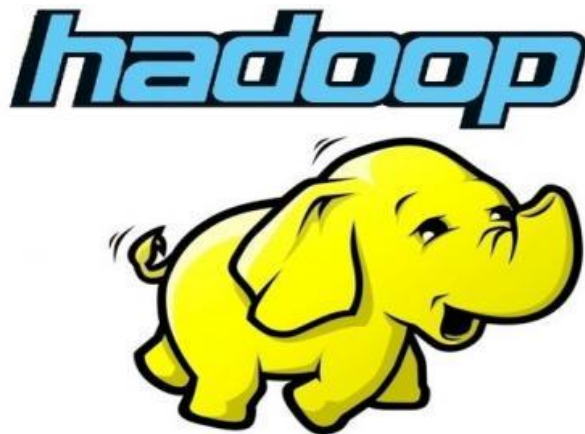
ArcGIS is a geographic information system (GIS) for working with maps and geographic information. It is used for: creating and using maps; compiling geographic data; analyzing mapped information; sharing and discovering geographic information; using maps and geographic information in a range of applications; and managing geographic information in a database. We have used ArcGIS for processing our geospatial data. **Apache Hive** is a data warehouse infrastructure built on top of Hadoop for providing data

summarization, query, and analysis. We used Apache Hive for sending queries to Hadoop which Hadoop processed using HDFS i.e **Hadoop Distributed File System** is a Java-based file system that provides scalable and reliable data storage that is designed to span large clusters of commodity servers and it completes operations using MapReduce programming model. **MapReduce** is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduce and scatter operations.

Chapter 2

System Requirement Study

2.1 Hadoop :-



Apache Hadoop is an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common and should be automatically handled by the framework.

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce. Hadoop splits files into large blocks and distributes them across nodes in a cluster. To process data, Hadoop transfers packaged code for nodes to process in parallel based on the data that needs to be processed. This approach takes advantage of data locality— nodes manipulating the data they have access to— to allow the dataset to be processed faster and more efficiently than it would be in a more conventional super computer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

2.2 GIS Tools for Hadoop



For many big datasets, location is a crucial component to truly understand underlying patterns and trends. Without location, datasets are less valuable, or in extreme circumstances - meaningless. GIS Tools for Hadoop works with big spatial data (big data with location) and allows you to complete spatial analysis using the power of distributed processing in Hadoop.

The GIS Tools for Hadoop toolkit allows us to leverage the Hadoop framework to complete spatial analysis on spatial data; for example:

- 1] Run a filter and aggregate operations on billions of spatial data records based on location.
- 2] Define new areas represented as polygons, and run a point in polygon analysis on billions of spatial data records inside Hadoop.
- 3] Visualize analysis results on a map and apply informative symbology.
- 4] Integrate your maps in reports, or publish them as map applications online.

2.3. Hive



Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. While developed by Facebook, Apache Hive is now used and developed by other companies such as Netflix and the Financial Industry Regulatory Authority. Amazon maintains a software fork of Apache Hive that is included in Amazon Elastic MapReduce on Amazon Web Services.

Apache Hive supports analysis of large datasets stored in Hadoop's HDFS and compatible file systems such as Amazon S3 filesystem. It provides an SQL-like language called HiveQL with schema on read and transparently converts queries to MapReduce, Apache Tez and Spark jobs. All three execution engines can run in Hadoop YARN. To accelerate queries, it provides indexes, including bitmap indexes. Other features of Hive include:

Indexing to provide acceleration, index type including compaction and Bitmap index as of 0.10, more index types are planned.

Different storage types such as plain text, RCFile, HBase, ORC, and others.

Metadata storage in an RDBMS, significantly reducing the time to perform semantic checks during query execution.

Operating on compressed data stored into the Hadoop ecosystem using algorithms including DEFLATE, BWT, snappy, etc.

Built-in user defined functions (UDFs) to manipulate dates, strings, and other data-mining tools. Hive supports extending the UDF set to handle use-cases not supported by built-in functions.

SQL-like queries (HiveQL), which are implicitly converted into MapReduce or Tez, or Spark jobs.

By default, Hive stores metadata in an embedded Apache Derby database, and other client/server databases like MySQL can optionally be used.

Four file formats are supported in Hive, which are TEXTFILE, SEQUENCEFILE, ORC and RCFILE. Apache Parquet can be read via plugin in versions later than 0.10 and natively starting at 0.13. Additional Hive plugins support querying of the Bitcoin Blockchain.

2.4 Geoprocessing-tools-for-hadoop

The Geoprocessing Tools for Hadoop provides tools to help integrate ArcGIS with Hadoop. More specifically, tools are provided that:

Enable the exchange of data between an ArcGIS Geodatabase and a Hadoop system, and

Allow ArcGIS users to run Hadoop workflow jobs.

See these tools in action as part of the samples in GIS Tools for Hadoop.

Features -

- Tools to convert between Feature Classes in a Geodatabase and JSON formatted files.
- Tools that copy data files from ArcGIS to Hadoop, and copy files from Hadoop to ArcGIS.
- Tools to run an Oozie workflow in Hadoop, and to check the status of a submitted workflow.

2.5. 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; analysing 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.

ArcGIS includes the following Windows desktop software:

- ArcReader, which allows one to view and query maps created with the other ArcGIS products;
- ArcGIS for Desktop, which is licensed under three functionality levels:
 - ArcGIS for Desktop Basic (formerly known as ArcView), which allows one to view spatial data, create layered maps, and perform basic spatial analysis;
 - ArcGIS for Desktop Standard (formerly known as ArcEditor), which in addition to the functionality of ArcView, includes more advanced tools for manipulation of shapefiles and geodatabases;
 - ArcGIS for Desktop Advanced (formerly known as ArcInfo), which includes capabilities for data manipulation, editing, and analysis.

2.6. QGIS



QGIS (Quantum GIS) is a cross-platform free and open-source desktop geographic information system (GIS) application that provides data viewing, editing, and analysis.

It is a user friendly Open Source Geographic Information System (GIS) licensed under the GNU General Public License. QGIS is an official project of the Open Source Geospatial Foundation (OSGeo).

Similar to other software GIS systems, QGIS allows users to create maps with many layers using different map projections. Maps can be assembled in different formats and for different uses. QGIS allows maps to be composed of raster or vector layers. Typical for this kind of software, the vector data is stored as either point, line or polygon-feature. Different kinds of raster images are supported and the software can Geo-reference images.

QGIS integrates with other open-source GIS packages, including PostGIS, GRASS, and MapServer to give users extensive functionality. Plugins written in Python or C++ extend QGIS's capabilities.

2.7. Virtualbox



Oracle VM VirtualBox (formerly Sun VirtualBox, Sun xVM VirtualBox and Innotek VirtualBox) is a free and open-source hypervisor for x86 computers from Oracle Corporation. Developed initially by Innotek GmbH, it was acquired by Sun Microsystems in 2008 which was in turn acquired by Oracle in 2010.

VirtualBox may be installed on a number of host operating systems, including: Linux, OS X, Windows, Solaris, and OpenSolaris. There are also ports to FreeBSD and Genode.

It supports the creation and management of guest virtual machines running versions and derivations of Windows, Linux, BSD, OS/2, Solaris, Haiku, OSx86 and others, and limited virtualization of OS X guests on Apple hardware.

For some guest operating systems, a "Guest Additions" package of device drivers and system applications is available which typically improves performance, especially of graphics.

2.8. NoSQL

A NoSQL (originally referring to "non SQL" or "non relational") database provides a mechanism for storage and retrieval of data which is modeled in means other than the tabular relations used in relational databases. Such databases have existed since the late 1960s, but did not obtain the "NoSQL" moniker until a surge of popularity in the early twenty-first century, triggered by the needs of Web 2.0 companies such as Facebook, Google and Amazon.com. NoSQL databases are increasingly used in big data and real-time web applications. NoSQL systems are also sometimes called "Not only SQL" to emphasize that they may support SQL-like query languages.

Motivations for this approach include: simplicity of design, simpler "horizontal" scaling to clusters of machines (which is a problem for relational databases), and finer control over availability. The data structures used by NoSQL databases (e.g. key-value, wide column, graph, or document) are different from those used by default in relational databases, making some operations faster in NoSQL. The particular suitability of a given NoSQL database depends on the problem it must solve. Sometimes the data structures used by NoSQL databases are also viewed as "more flexible" than relational database tables.

Many NoSQL stores compromise consistency (in the sense of the CAP theorem) in favor of availability, partition tolerance, and speed. Barriers to the greater adoption of NoSQL stores include the use of low-level query languages (instead of SQL, for instance the lack of ability to perform ad-hoc JOINS across tables), lack of standardized interfaces, and huge previous investments in existing relational databases. Most NoSQL stores lack true ACID transactions, although a few databases, such as MarkLogic, Aerospike, FairCom c-treeACE, Google Spanner (though technically a NewSQL database), Symas LMDB and OrientDB have made them central to their designs. (See ACID and JOIN Support.)

Instead, most NoSQL databases offer a concept of "eventual consistency" in which database changes are propagated to all nodes "eventually" (typically within milliseconds) so queries for data might not return updated data immediately or might result in reading data that is not accurate, a problem known as stale reads. Additionally, some NoSQL systems may exhibit lost

writes and other forms of data loss. Fortunately, some NoSQL systems provide concepts such as write-ahead logging to avoid data loss. For distributed transaction processing across multiple databases, data consistency is an even bigger challenge that is difficult for both NoSQL and relational databases. Even current relational databases "do not allow referential integrity constraints to span databases." There are few systems that maintain both ACID transactions and X/Open XA standards for distributed transaction processing.

2.9. JAVA



Java is a general-purpose computer programming language that is concurrent, class-based, object-oriented, and specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere" (WORA), meaning that compiled Java code can run on all platforms that support Java without the need for recompilation. Java applications are typically compiled to bytecode that can run on any Java virtual machine (JVM) regardless of computer architecture. As of 2016, Java is one of the most popular programming languages in use, particularly for client-server web applications, with a reported 9 million developers. Java was originally developed by James Gosling at Sun Microsystems (which has since been acquired by Oracle Corporation) and released in 1995 as a core component of Sun Microsystems' Java platform. The language derives much of its syntax from C and C++, but it has fewer low-level facilities than either of them.

The original and reference implementation Java compilers, virtual machines, and class libraries were originally released by Sun under proprietary licences. As of May 2007, in compliance with the specifications of the Java Community Process, Sun relicensed most of its Java technologies under the GNU General Public License. Others have also developed alternative implementations of these Sun technologies, such as the GNU Compiler for Java (bytecode

compiler), GNU Classpath (standard libraries), and IcedTea-Web (browser plugin for applets).

The latest version is Java 8, which is the only version currently supported for free by Oracle, although earlier versions are supported both by Oracle and other companies on a commercial basis.



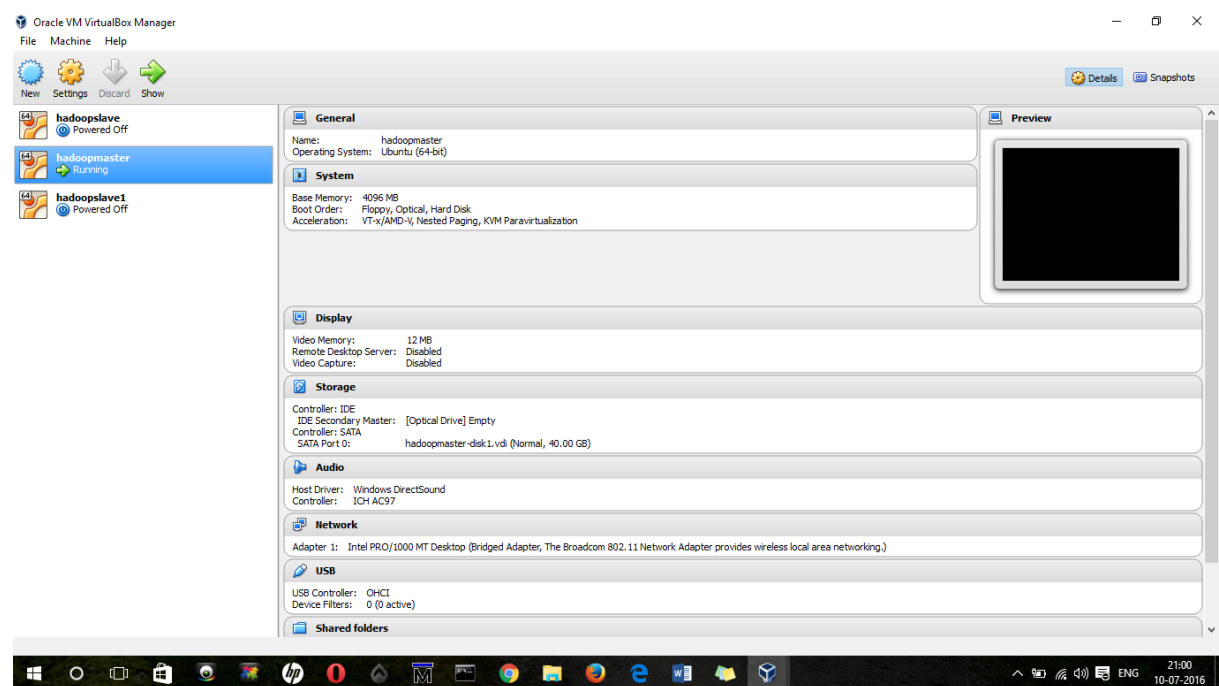
Chapter 3

Methodology and Implementation

On our VM VirtualBox we installed Ubuntu as our operating system. We installed java on our operating system and set the environment variables. After that Hadoop was installed further we created a Hadoop user and configured Hadoop. After setting this single node Hadoop cluster we cloned this VM and made two other VMs namely HadoopMaster and HadoopSlave1. These two VMs together formed the multi node cluster where HadoopMaster acting as the namenode and HadoopSlave1 acting as the datanode. The two clusters were connected through the Bridged Network. We installed GIS tools for Hadoop so that we can complete our spatial analysis of spatial data. Further we installed hive on our system for passing queries to Hadoop. Further we installed ArcGis which allows one to view spatial data, create layered maps, and perform basic spatial analysis.

Implementation -

VM VirtualBox: VM VirtualBox is a free and open-source hypervisor for x86 computers from Oracle Corporation. Virtualization is the process of converting from a purely physical implementation to one using a hypervisor which abstract the underlying physical hardware and provide an idealized, or virtual, implementation upon which some higher-level services and/or implementations can be designed and built. Once a physical cluster is virtualized, then higher level services, such as cloning a data node, or providing high-availability to a specific node, or providing user controlled provisioning, can be built. We installed VM Virtual Box on our machine so that we can create Hadoop clusters by cloning the single Hadoop node on LINUX Ubuntu operating system.



Ubuntu: Ubuntu is a Debian based Linux operating system and distribution for personal computers, smartphones and network servers. It uses Unity as its default user interface. We used Ubuntu as the Desktop operating system in our VM VirtualBox on which we installed the java based programming framework Hadoop.

Java: Java is a programming language and computing platform first released by Sun Microsystems. Java is the main prerequisite for Hadoop as Hadoop is a java based programming framework.

We installed Java 7 in our system, set the environment variables.

Hadoop Single node cluster: Apache Hadoop is an open source framework for storing and distributed batch processing of huge datasets on clusters of commodity hardware. Hadoop can be used on a single machine (Standalone Mode) as well as on a cluster of machines (Distributed Mode – Pseudo & Fully). One of the striking features of Hadoop is that it efficiently distributes large amounts of work across a cluster of machines/commodity hardware. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce. Hadoop splits files into large blocks and distributes them across nodes in a cluster. To process data, Hadoop transfers packaged code for nodes to process in parallel based on the data that needs to be processed. This approach takes advantage of data locality nodes manipulating the data they have access to allow the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

The base Apache Hadoop framework is composed of the following modules:

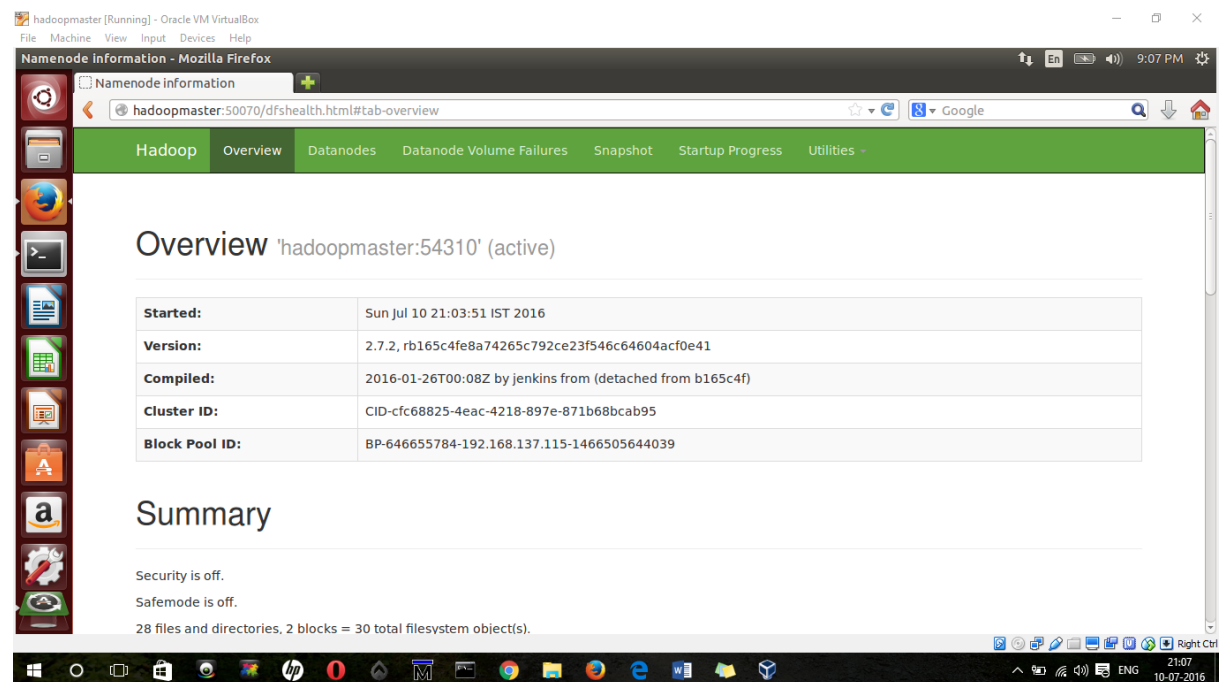
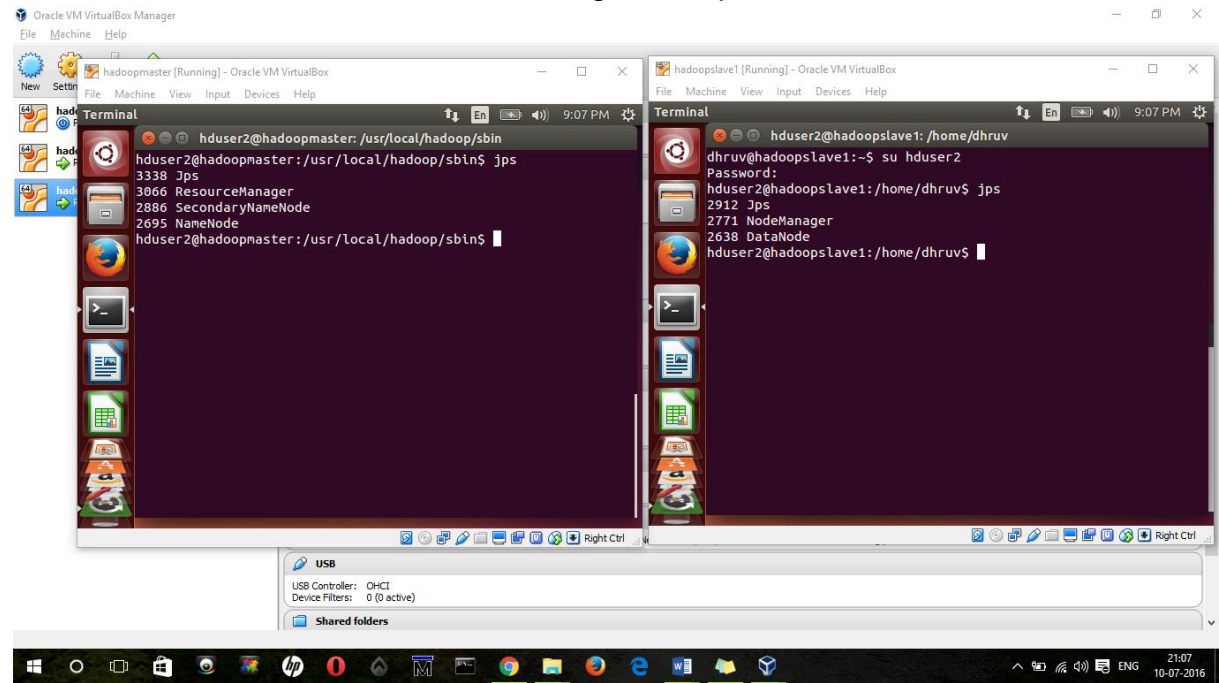
- *Hadoop Common* – contains libraries and utilities needed by other Hadoop modules;
- *Hadoop Distributed File System (HDFS)* – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;
- *Hadoop YARN* – a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications, and
- *Hadoop MapReduce* – an implementation of the MapReduce programming model for large scale data processing.

The Hadoop framework itself is mostly written in the Java Programming Language, with some native code in C and command line utilities written as shell scripts. Though MapReduce Java code is common, any programming language can be used with "Hadoop Streaming" to implement the "map" and "reduce" parts of the user's program. Other projects in the Hadoop ecosystem expose richer user interfaces.

We created Hadoop user in our system and configured Hadoop environment variables. Configured yarn-site.xml, mapre-site.xml, core-site.xml and hdfs-site.xml.

Hadoop multi node cluster:

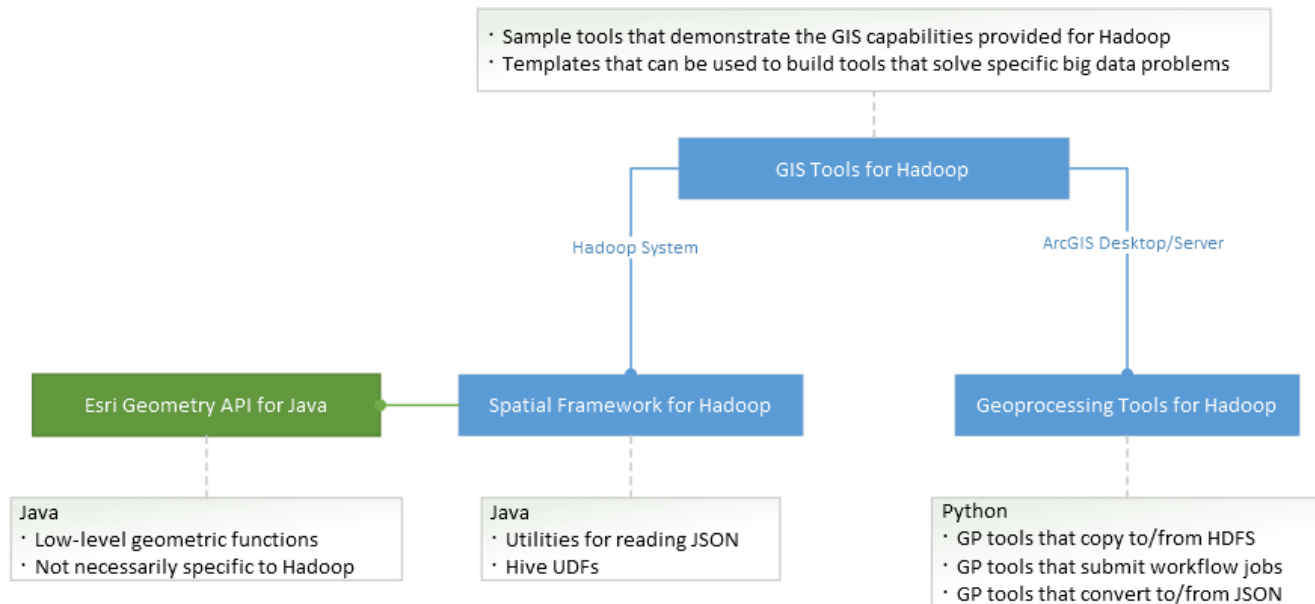
Following the tutorial of Michael Noll “Running Hadoop on Ubuntu Linux” from his website <http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/> we created the multi node cluster. We cloned the single node Hadoop cluster to two VMs HadoopMaster and HadoopSlave1 connected them through Bridged networking. Thus we created the multi node cluster using Hadoop.



[illegible]

- Features of GIS tools for Hadoop:

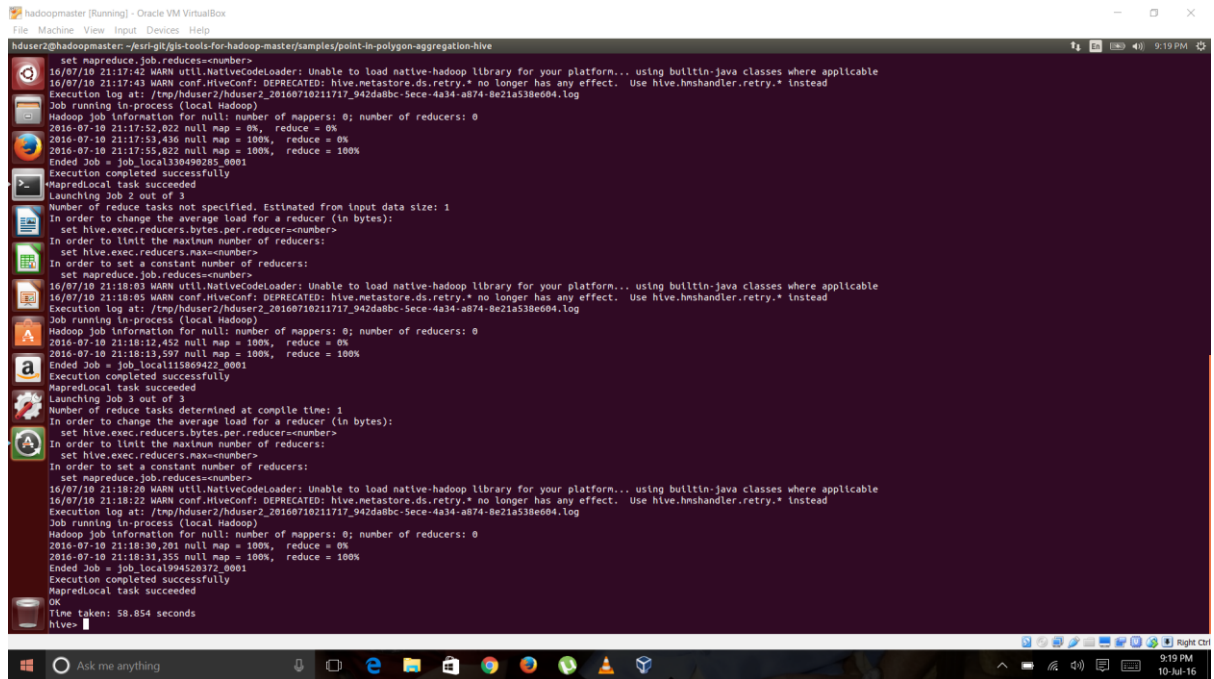
Sample tools that demonstrate full stack implementations of all the resources provided to solve GIS problems using Hadoop



We installed the GIS tools for Hadoop on the HadoopMaster. We executed the sample example from <https://github.com/Esri/gis-tools-for-hadoop/tree/master/samples> the aggregate sample <https://github.com/Esri/gis-tools-for-hadoop/tree/master/samples/point-in-polygon-aggregation-hive>.

```

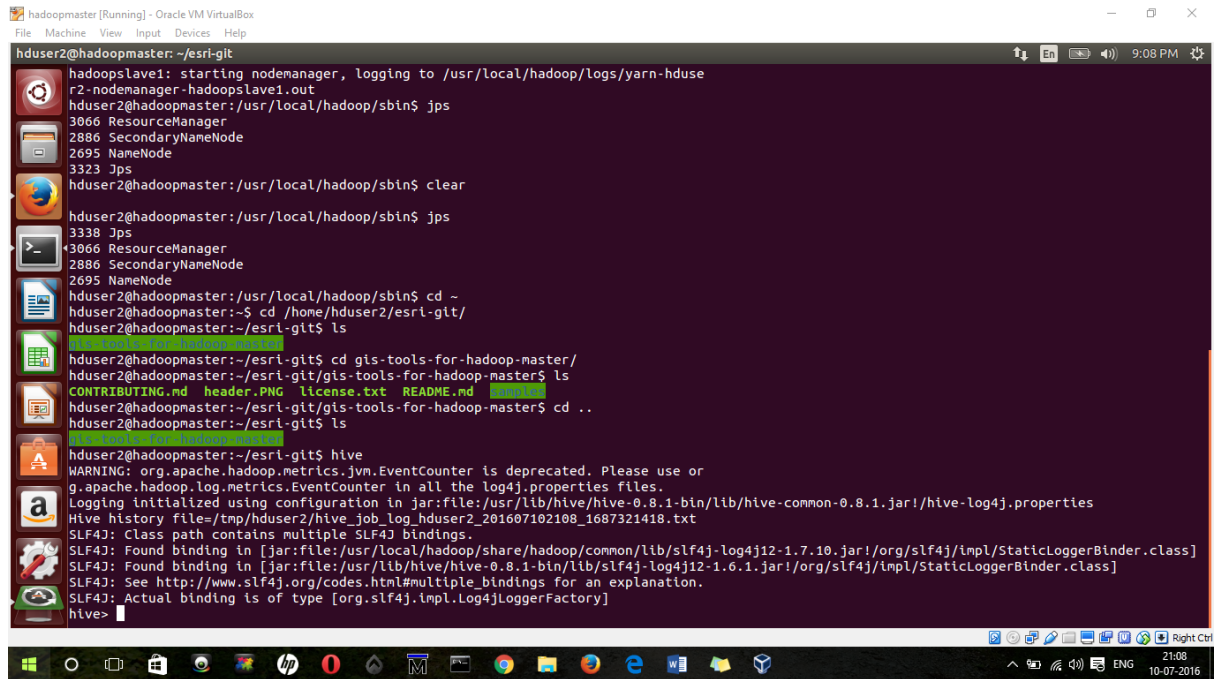
hadoopmaster [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
hduser2@hadoopmaster:~/esri-gis-tools-for-hadoop-master/samples/point-in-polygon-aggregation-hive$
hduser2@hadoopmaster:~/esri-gis-tools-for-hadoop-master/samples$ ls
derby.log  metastore.db
hduser2@hadoopmaster:~/esri-gis-tools-for-hadoop-master/samples$ cd point-in-polygon-aggregation-hive/
hduser2@hadoopmaster:~/esri-gis-tools-for-hadoop-master/samples/point-in-polygon-aggregation-hive$ hive
10/07/10 21:17:02 WARN conf.HiveConf: DEPRECATED: hive.metastore.ds.retry.* no longer has any effect. Use hive.hms.handler.retry.* instead
Logging initialized using configuration in jar:file:/usr/lib/hive/apache-hive-0.13.0-bin/lib/hive-common-0.13.0.jar!/hive-log4j.properties
hive> source run-sample.sql
> ?
Added ./lib/esri-geometry-api.jar to class path
Added resource: ./lib/esri-geometry-api.jar
Added ./lib/spatial-sdk-hadoop.jar to class path
Added resource: ./lib/spatial-sdk-hadoop.jar
OK
Time taken: 0.613 seconds
OK
Time taken: 0.021 seconds
OK
Time taken: 0.876 seconds
OK
Time taken: 0.285 seconds
Warning: Shuffle Join JOIN[4][tables = [counties, earthquakes]] in Stage 'Stage-1:MAPRED' is a cross product
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
10/07/10 21:17:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Execution log at: /tmp/hduser2/hduser2_20160710211717_942da8bc-Sece-4a34-a874-8e21a538e604.log
Job running in-process (local Hadoop)
Hadoop job information for null: number of mappers: 0; number of reducers: 0
2016-07-10 21:17:52,022 null map = 0%, reduce = 0%
2016-07-10 21:17:53,430 null map = 100%, reduce = 0%
2016-07-10 21:17:55,022 null map = 100%, reduce = 100%
Ended Job = job_local330490285_0001
Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 3
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Trash
ice.job.reduces=<number>
10/07/10 21:18:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
  
```



```
hduser2@hadoopmaster:~/src-git/gis-tools-for-hadoop-master/samples/point-in-polygon-aggregation-hive
File Machine View Input Devices Help
16/07/10 21:17:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/07/10 21:17:43 WARN conf.HiveConf: DEPRECATED: hive.metastore.ds.retry.* no longer has any effect. Use hive.hmsHandler.retry.* instead
Execution log at: /tmp/hduser2/hduser2_20160710211717_942da8bc-Sece-4a34-a874-8e21a538e604.log
Job running in-process (local Hadoop)
Hadoop job information for null: number of mappers: 0; number of reducers: 0
2016-07-10 21:17:52,022 null map = 0%, reduce = 0%
2016-07-10 21:17:53,436 null map = 100%, reduce = 0%
2016-07-10 21:17:55,022 null map = 100%, reduce = 100%
Ended Job = job_local1330450285_0001
Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 3
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=number
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=number
In order to set a constant number of reducers:
  set mapreduce.job.reduces=number
16/07/10 21:18:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/07/10 21:18:05 WARN conf.HiveConf: DEPRECATED: hive.metastore.ds.retry.* no longer has any effect. Use hive.hmsHandler.retry.* instead
Execution log at: /tmp/hduser2/hduser2_20160710211717_942da8bc-Sece-4a34-a874-8e21a538e604.log
Job running in-process (local Hadoop)
Hadoop job information for null: number of mappers: 0; number of reducers: 0
2016-07-10 21:18:13,597 null map = 100%, reduce = 100%
Ended Job = job_local115809422_0001
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=number
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=number
In order to set a constant number of reducers:
  set mapreduce.job.reduces=number
16/07/10 21:18:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/07/10 21:18:22 WARN conf.HiveConf: DEPRECATED: hive.metastore.ds.retry.* no longer has any effect. Use hive.hmsHandler.retry.* instead
Execution log at: /tmp/hduser2/hduser2_20160710211717_942da8bc-Sece-4a34-a874-8e21a538e604.log
Job running in-process (local Hadoop)
Hadoop job information for null: number of mappers: 0; number of reducers: 0
2016-07-10 21:18:30,201 null map = 100%, reduce = 0%
2016-07-10 21:18:31,355 null map = 100%, reduce = 100%
Ended Job = job_local094520372_0001
Execution completed successfully
MapredLocal task succeeded
OK
Time taken: 50.854 seconds
hive>
```

Hive: Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. Apache Hive supports analysis of large datasets stored in Hadoop's [HDFS](#) and compatible file systems such as Amazon S3 filesystem. It provides an SQL -like language called HiveQL with schema on read and transparently converts queries to MapReduce, Apache Tez and Spark jobs. All three execution engines can run in HadoopYARN. To accelerate queries, it provides indexes, including bitmap indexes.

We installed Hive on the HadoopMaster for passing the queries.



```
hadoopmaster [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help

hduser2@hadoopmaster: ~/esri-glt
hduser2@hadoopmaster:~/esri-glt$ jps
3066 ResourceManager
2886 SecondaryNameNode
2695 NameNode
3323 Jps
hduser2@hadoopmaster:~/esri-glt$ clear
hduser2@hadoopmaster:~/esri-glt$ jps
3338 Jps
3066 ResourceManager
2886 SecondaryNameNode
2695 NameNode
hduser2@hadoopmaster:~/esri-glt$ cd /home/hduser2/esri-glt/
hduser2@hadoopmaster:~/esri-glt$ ls
CONTRIBUTING.md  header.PNG  license.txt  README.md  esri-glt
hduser2@hadoopmaster:~/esri-glt$ cd gis-tools-for-hadoop-master/
hduser2@hadoopmaster:~/esri-glt/gis-tools-for-hadoop-master$ ls
CONTRIBUTING.md  header.PNG  license.txt  README.md  esri-glt
hduser2@hadoopmaster:~/esri-glt/gis-tools-for-hadoop-master$ cd ..
hduser2@hadoopmaster:~/esri-glt$ ls
CONTRIBUTING.md  header.PNG  license.txt  README.md  esri-glt
hduser2@hadoopmaster:~/esri-glt$ hive
WARNING: org.apache.hadoop.metrics.jvm.EventCounter is deprecated. Please use or
g.apache.hadoop.log.metrics.EventCounter in all the log4j.properties files.
Logging initialized using configuration in jar:file:/usr/lib/hive/hive-0.8.1-bin/lib/hive-common-0.8.1.jar!/hive-log4j.properties
Hive history file=/tmp/hduser2/hive_job_log/hduser2_201607102108_1687321418.txt
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/hive-0.8.1-bin/lib/slf4j-log4j12-1.6.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
hive>
```

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. We installed ArcGis platform on our HadoopMaster. We visualized our results of Hive query in ArcGis and thus got a proper interpretation and clearer results. We used the following link to execute it.

<https://github.com/Esri/gis-tools-for-hadoop/wiki/Getting-the-results-of-a-Hive-query-into-ArcGIS>

ArcGis Desktop is only available for windows so if BISAG implements on their server they need to run it in windows after installing VM there.

RESULT AND CONCLUSION

We created a multi node Hadoop cluster so that we can concurrent high queries efficiently and effectively. Also it reduces the possibility of failure of data loss as we have our data distributed across various nodes in our cluster. The MapReduce algorithm which Hadoop uses to process queries is quiet effective in processing large queries. Analyzing the geospatial data was our main aim, using GIS tools on Hadoop. We analyzed hive queries results in ArcGis platform. ArcGis platform helped us to clearly visualize and interpret our results of analysis of geospatial data.

Chapter 4

Bibliography

http://www.tutorialspoint.com/hadoop/hadoop_multi_node_cluster.htm

<http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/>

<http://doctuts.readthedocs.io/en/latest/hadoop.html>

<http://www.quuxlabs.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/>

<https://www.youtube.com/watch?v=MzdyM3N5SIE>

<https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Prerequisites>

<http://www.slideshare.net/darugar/cloud-computing-hadoop-presentation>

http://cs.smith.edu/dftwiki/index.php/Setup_Virtual_Hadoop_Cluster_under_Ubuntu_with_VirtualBox

http://cs.smith.edu/dftwiki/index.php/Setup_Virtual_Hadoop_Cluster_under_Ubuntu_with_VirtualBox

<http://archive.oreilly.com/pub/a/other-programming/excerpts/hadoop-tdg/installing-apache-hadoop.html>

<http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html>

https://www.youtube.com/watch?v=CRLq_aG_r6s

<http://trimc-hdfs.blogspot.in/2014/11/enabling-ssh-for-hadoop-cluster.html>

<https://www.anonproxy.org/index.php?q=aHR0cDovL2NoYWZscHJpdGFtLmJsbn2dzcG90LmNvbS8yMDE1LzAxL2hhZG9vcC0yNjAtbXVsdGktbm9kZS1jbHVzdGVyLXNldHVwLW9uLmh0bWw%3D&hl=2ed>

<https://github.com/Esri/spatial-framework-for-hadoop>

<https://github.com/Esri/gis-tools-for-hadoop>

<http://stackoverflow.com/questions/16004172/missing-hive-execution-jar-usr-local-hadoop-hive-lib-hive-exec-jar>

<https://github.com/Esri/gis-tools-for-hadoop/tree/master/samples/point-in-polygon-aggregation-hive>

<http://gis.stackexchange.com/questions/144454/how-to-use-hadoop-tools-in-arcmap-10-3>

<https://blogs.esri.com/esri/arcgis/2013/03/25/gis-tools-for-hadoop/>

<https://blogs.esri.com/esri/arcgis/2015/03/25/new-spatial-aggregation-tutorial-for-gis-tools-for-hadoop/>

<https://github.com/Esri/spatial-framework-for-hadoop/tree/master/hive>

<https://www.youtube.com/watch?v=1rMaeBd0JWQ>

<http://www.tutorialspoint.com/hive/>

<https://github.com/Esri/gis-tools-for-hadoop/wiki/Getting-the-results-of-a-Hive-query-into-ArcGIS>

<http://stackoverflow.com/questions/9995694/json-output-format-for-hive-query-results>

<http://stackoverflow.com/questions/11185528/what-is-hive-return-code-2-from-org-apache-hadoop-hive-ql-exec-mapredtask>

http://cs.smith.edu/dftwiki/index.php/Setup_Virtual_Hadoop_Cluster_under_Ubuntu_with_VirtualBox

<http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6194600&newsearch=true&queryText=geoserver>

<https://sites.google.com/site/hadoopgis/>

<http://webapps.esri.com/s3.amazonaws.com/esri-proceedings/devsummit14/papers/dev-062.pdf>