MRDS Data Processing and Mining using Hadoop in Cloud

Ravindra P. Bachate
Department of Computer Engineering,
JSPM's JSCOE, Hadapsar
Pune, Maharashtra, India

H. A. Hingoliwala
Department of Computer Engineering,
JSPM's JSCOE, Hadapsar
Pune, Maharashtra, India

ABSTRACT

This project explores the use of Hadoop framework for MRDS (Mineral Resources data system) data processing and mining in cloud. Cloud computing provides efficient computation and analysis for large data. To improve the performance of system for massive data, Hadoop provides Map Reduce technique. Hadoop has a distributed file system (HDFS) that stores data on the cluster nodes. This project focuses on to provide real time information of mineral resources stored in cloud environment with minimum data processing time. Storing MRDS data in to the cloud ensures the availability and reliability of it.

Keywords

Hadoop, cloud computing, data processing, data mining

1. INTRODUCTION

Due to the drastic development in various sectors, size of data increases day by day. One computer can read 30-35 MB data per second. For example if data size is 100 TB, approximately it will take 1 month to process it [4]. So obviously lots of data mining and data processing required for getting important information from the available data. Our world is data driven. For example Science has databases from astronomy, genomics, transportation data, environment data etc. Likewise medicine, entertainment, commerce, humanities and social sciences [4]. The data which challenges to current technologies to store, process and use called as a big data. Big Data are high-volume and high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. Now a day's more than 80% businesses are relying on cloud because of the services and features provided with cloud environment. Cloud is a repository of such huge data and challenge of cloud provider is to manage and process the data into it.

Mineral Resources Data System is a collection of data describing metallic and nonmetallic mineral resources throughout the world [8]. It includes resource name, location, commodity, geologic characteristics, resource description, production, reserves, and references. As MRDS contains mineral resources data around the world, it is large and complex. If data size goes beyond the Tera Byte, it is difficult to manage and process the data by using RDBMS. The performance of RDBMS decreases as data size increases. To make MRDS data available to all the time, we need to keep it in the cloud environment. Traditional approach to deal with such massive data is ETL i.e. extract, transform and load it into RDBMS. But spatial data is available in the unstructured format and it is not easy for RDBMS is to cope with it.

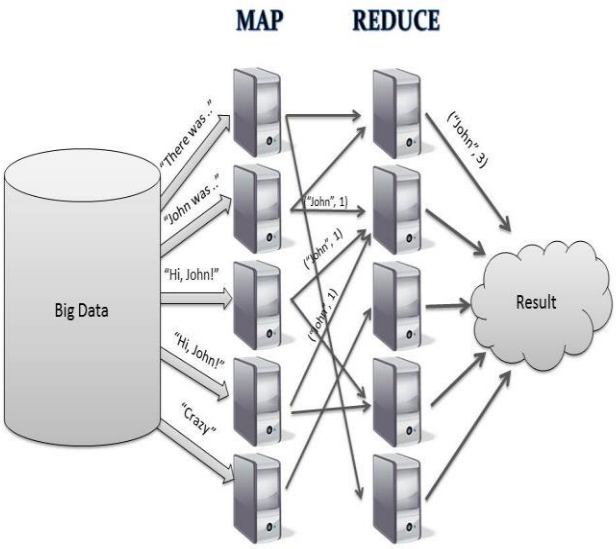

Fig.1 Map Reduce Architecture

To deal with big data like MRDS in the cloud, we need a best technology which can cope with it. There are two options available, parallel DBMS and Hadoop Map Reduce technology. But Hadoop Map Reduce gives a better data processing performance with minimum cost and time as compare to parallel DBMS because it works with commodity hardware. Hadoop has HDFS file system for storing a big data into it. The Hadoop framework provides a solution for problems of massive data processing; because it runs applications on large cluster built of commodity hardware with failure tolerance [5].Unstructured data can be processed with Hadoop Map Reduce technique which is not possible with RDBMS. Map Reduce provides flexibility and fault tolerance which is not with parallel DBMS. Map Reduce provides automatic parallelization, data partitioning, task scheduling, handling machine failures and manages inter-machine communication. Hadoop is totally transparent from the end user. The rate of growing an unstructured data is much more as compare to the structured data. The unstructured data includes media files, heavy text files etc.

2. LITERATURE SURVEY

Hongyong Yu, Deshuai Wang [1] proposed a system for data processing and mining log data of SaaS cloud using Hadoop. We focused on Hadoop's Map Reduce technique and the algorithm used for data mining by Hongyong Yu, Deshuai

Wang. The results given in this paper proved that Hadoop data processing performance is very high as compare to RDBMS i.e. 28% improvement in the data processing [1].Apriori algorithm is used for data mining in the cloud which is the best to find association rules from big data. It uses tree structure and bottom up approach to count item sets efficiently from data. Parallel computing approach is used in adaptive Apriori algorithm to improve the performance of the system having large data size.

Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years [4].The commonly used software technology cannot cope with massive data and big challenge is to extract a important information from it. Big data has large volume, heterogeneous format and decentralized data control. The example of big data applications are Facebook, Twitter and Google. It is a big challenge to manage and mining a massive data because of its volume, different file formats and growing rate of the data in the world. There are many challenges with big data such as storage, processing, variety and cost.

MRDS is a spatial data about the mineral resources available across the world. As GIS (Geographic Information System) is growing, massive data is generated from the different geographic locations of the world [2].There are several algorithms available to cope with data mining such as K-means, AdaBoost, kNN (k-nearest neighbor), Apriori algorithm.

3. IMPLEMENTATION

3.1 Proposed framework

The proposed system framework is an enhancement to techniques introduced in [1].The main motive of proposed system is to improve mineral resources data system(MRDS) performance for finding the demand, supply, flow of mineral and data mining. This helps user to make efficient data processing and mining as compare to previous system which has been implemented by using RDBMS. The main advantage of proposed system is its user-friendly interface design and quick response time.

This proposed framework makes use of Apriori algorithm for finding association rules and Hadoop data processing framework with less processing time. The proposed system uses Hadoop’s Map Reduce technique for data processing and mining. The mineral resources data system database is available in various formats like .dbf, .mdf and .txt. We have more than 2 lack world records of mineral resources having attributes such as state, country name, product name, product size, work type like underground. Use of Hadoop provides a capability to a system to deal with any kind of data which makes the system strong. Because mineral data can be available in different data format and system must be able to cope with such heterogeneous and massive data. Mineral resources data system gives the information about resources like metallic and non-metallic across the world.

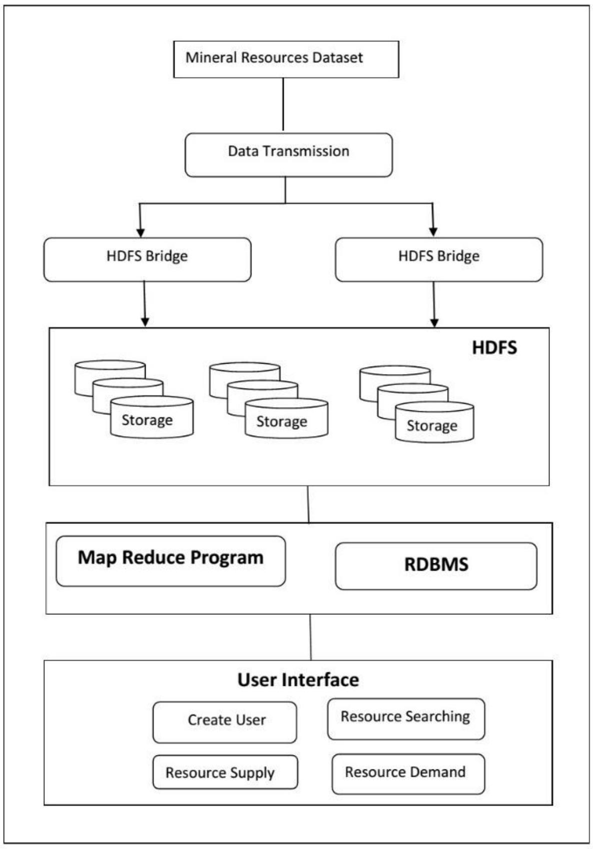

Fig.2 Proposed System

3.2 Details of System Model

In this section we will study details of proposed framework for MRDS data processing and mining. Major problem associated with RDBMS is it takes more processing time if data is big. Hadoop based data processing and mining using Apriori algorithm [1] is implemented in proposed architecture for MRDS. MRDS data is available in various formats, so first we have to convert that data into HDFS format to work with Hadoop. Data transmission program mentioned in the diagram converts that available data in to HDFS. This converted data is stored on HDFS storage. Users can access the data through portals and HDFS data is processed by using Map Reduce technology of Hadoop. Proposed system provides three options for user- finding general information about resources, resource demand and resource supply.

1) *Create User*: To register a user for mineral resources data system, this module is used. We can keep all the log information of the user by using this module.

2) *Resource Searching:* This is the first module which describes all the available information about the resources. We can give any resource name for which we are looking in the world. This module will gives the details of the resource such as location, resource availability information, name of site. To find the resource location, mapper and reducer is used as described in algorithm. It takes id, data set as a input and do a mapping between them.

For example: user gave “*gold*” as input query then application programming interface will pass that value to Hadoop’s Map Reduce processing tool. It will look for the data available in HDFS and will give back the result.

3) *Resource Demand:* This model illustrates the demand of resources from various region across the world. Supplier and can view the demand of resources across the world. As this is the data collected from all the regions of world, volume of data is huge. So it can be easily processed and mined by applying the proposed system to it. In fraction of second, result for which user is searching will be provided.

4) *Resource Supply:* Such as demand, the proposed system has another module i.e. Supply. This module is also for both supplier and buyer. If supplier is interested to check region in the world where their product can be sale and also buyer can search for the supplier over here.

3.3 Algorithm

The Apriori algorithm is used for data processing and mining for MRDS using Hadoop in cloud. Data mining technology is used to extract and find knowledge from big data.Apriori algorithm is a nnovative way to find association rules on large scale, allowing implication outcomes that consist of more than one item.There are two stages to find the result – mapping and reducing. We assume that D is mineral resources data set, FSet is hash set, key means ID. Reporter is used to report progress or just indicate that mapper and reducer are alive. In scenarios where the application takes significant amount of time to process individual key/value pairs, this is crucial since the framework might assume that the task has timed-out and kill that task.

Mapper:

```
public void init ()
    {
     Read all records from F and store them  into HashSet FSet
    }
    public void map<key, value, reporter>
    {
    for each f in FSet{
             c=f.attr1#f.attr2#...;
             //property of relation extension
             D=key ;}
    output.Collect (new Text(c), new Text (d));
    }
```

Reducer:

```
    public void reduce<key,
    Iterator<Text>values, output, reporter>
    {
    While (values.hasNext ())
          {
             f=values.next ();
            if (!f Є values)
             values = values#1;
    }
    Output.collect (key, new Text (values));
    //combine and output the results      }
```

Mapper and Reducer algorithms are used for implementing the data mining and data processing for MRDS. Mapper maps input key/value pairs to a set of intermediate key/value pairs [5].Mapper generates intermediate records. Reducer reduces a set of intermediate values which share a key to a smaller set of values [5].

4. CONCLUSION

Dealing with a massive data is a big challenge and requires a best technology to cope with it. In this we proposed a system for processing and mining a big mineral resources data system's data. To enhance the performance of data processing Hadoop's Map Reduce architecture is used. For better improvement in data mining for MRDS, Apriori algorithm is used. The proposed system can improve the performance of the existing system more than 30%.By grouping a good and open source technology such as Hadoop and aproiri algorithm together, we can achieve a better data processing and mining for the MRDS system

5. ACKNOWLEDGMENTS

I express true sense of gratitude towards my project guide Prof. H.A. Hingoliwala, Associate Professor Computer Department for his invaluable co-operation and guidance that he gave me throughout my project. I specially thank our P.G coordinator Prof. M. D. Ingle for inspiring me and providing me all the lab facilities. I would also like to express my appreciation and thanks to HOD Prof. S.M. Shinde & JSCOE Principal Dr. M.G. Jadhav and all my friends who knowingly or unknowingly have assisted me throughout my hard work.

6. REFERENCES

[1] Hongyong Yu, Deshuai Wang, "Mass Log Data Processing and Mining Based on Hadoop and Cloud Computing" .The 7th International Conference on Computer Science & Education (ICCSE 2012)July 14-17, 2012. Melbourne, Australia.

[2] Duck-Ho Bae Coll. of Inf. & Commun., Hanyang Univ., Seoul, South Korea Ji-Haeng Baek ; Hyun-Kyo Oh ; Ju-Won Song ; Sang-Wook Kim, "SD-Miner: A SPATIAL DATA MINING SYSTEM" Network Infrastructure and Digital Content, 2009.

[3] Weikuan Yu, Member, IEEE, Yandong Wang, and Xinyu Que, " Design and Evaluation of Network-Levitated Merge for Hadoop Acceleration", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS.

[4] Xindong Wu,Fellow, IEEE,Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding,Senior Member, IEEE, "Data mining with big data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014.

[5] Hadoop: The definitive Guide, 3rd ed., O'Reilly, Tom White, 2012

[6] Hadoop in Action, Manning, Chuck Lam, 2011

[7] Hadoop, http://hadoop.apache.org/

[8] MRDS, http://tin.er.usgs.gov/mrds/