ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

Data Mining for Data Cloud and Compute Cloud

Prof. Uzma Ali[1], Prof. Punam Khandar[2]

Assistant Professor, Dept. Of Computer Application, SRCOEM, Nagpur, India[1]
Assistant Professor, Dept. Of Computer Application, SRCOEM, Nagpur, India[2]

ABSTRACT: Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation. A storage cloud provides storage services, while a compute cloud provides compute services. We describe the design of the Sector storage cloud and how it provides the storage services required by the Sphere compute cloud. A cloud can be a storage cloud that provides block or file based storage service or it can be a compute cloud that provides computational services. Moreover in this paper we have reviewed the design and implementation of sector storage cloud and sphere compute cloud. Sector is the distributed file system, while sphere is the parallel in-storage data processing framework that can be used to process data stored in sector. Sector and Sphere are designed for analyzing large data sets using computer clusters connected with wide area high performance networks (for example, 10+ Gb/s). We describe a distributed data mining application that we have developed using Sector and Sphere. Mining association rules is one of the most important aspects in data mining. Association rules are dependency rules which predict occurrence of an item based on occurrences of other items. Sector is the distributed file system, while sphere is the parallel in-storage data processing framework that can be used to process data stored in sector.

Keywords**:** Cloud Computing, Sector, Sphere, Association Rule, Data Mining

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information- information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Mining Association rule is a way to find interesting associations among large sets of data items. Data mining system is designed for taking advantage of powerful and shared pools of processors. In that data is distributed over the processors and the computation is done using message passing paradigm. Then all the computation results are gathered and this process is repeated on the new data on the processor. By a cloud, we mean an infrastructure that provides on-demand resources or services over the Internet, usually at the scale and reliability of a data centre. A storage cloud provides storage services (block or file-based services); a data cloud provides data management services (record-based, column-based or object-based services); and a compute cloud provides computational services. Often these are stacked together to serve as a computing platform for developing cloud-based applications.

Sector is designed to provide long term persistent storage to large datasets that are managed as distributed indexed files. Different segments of the file are scattered throughout the distributed storage managed by Sector. Sector generally replicates the data to ensure its longevity, to decrease the latency. When retrieving it, and to provide opportunities for parallelism. Sector is designed to take advantage of wide area high performance networks when available.

Sphere is designed to execute user defined functions in parallel using a stream processing pattern for data managed by Sector. We mean by this that the same user defined function is applied to every data record in a data set man-aged by Sector. This is done to each segment of the dataset independently (assuming that sufficient processors are available), providing a natural parallelism.

The design of Sector/Sphere results in data frequently being processed in place without moving it. In this paper we have described the design of data and compute cloud. We have also describe a data mining application developed using sector and sphere that searches for evolving behavior in distributed network data.

II. BACKGROUND RELATED WORK

By a cloud, we mean an infrastructure that provides resources and/or services over the Internet. A data cloud provides data management services (record-based, column based or object-based services); and a compute cloud provides computational services. These all types of clouds are set up as a stack of cloud services that provides computing platform to develop cloud based applications.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

Application1	--	Application N
Compute Cloud Services		
Data Cloud Services		
Storage Cloud Services		

Figure 1: A data stack for a cloud consist of layered services

Examples include Google's Google File System (GFS), BigTable and MapReduce infrastructure Amazon's S3 storage cloud, SimpleDB data cloud, and EC2 compute cloud and the open source Hadoop system.
In this paper we have described a Sector storage cloud and a sphere data cloud. Sector/Sphere is a software platform that supports very large distributed data storage and simplified distributed data processing. The system consists of Sector, a distributed storage system, and Sphere, a runtime middleware to support simplified development of distributed data processing.

III. SECTOR

Sector is a storage cloud as defined above. Specifically, Sector provides storage services over the Internet with the scalability and reliability of a data centre.
The characteristics of sector are
1. Sector is designed to support a community of users, not all of whom may have write access to
the Sector infrastructure.
2. Sector provides long term archival storage and access for large distributed datasets.
3. Sector is designed to utilize the bandwidth available on wide area high performance networks.
4. Sector supports a variety of different routing and network protocols. [
Sector makes three assumptions.
(i) Sector assumes that it has access to a large number of commodity computers (which we sometimes call nodes). The nodes may be located either within or across data centres.
(ii) Sector assumes that high-speed networks connect the various nodes in the system. For example, in the experimental studies described below, the nodes within a rack are connected by 1 Gb sK1 networks, two racks within a data centre are connected by 10 Gb sK1 networks and two different data centres are connected by 10 Gb sK1 networks.
(iii) Sector assumes that the datasets it stores are divided into one or more separate files, which are called Sector slices. The different files comprising a dataset are replicated and distributed over the various nodes managed by Sector. For example, one of the datasets managed by Sector in the experimental studies described below is a 1.3 TB dataset consisting of 64 files, each approximately 20.3 GB in size.

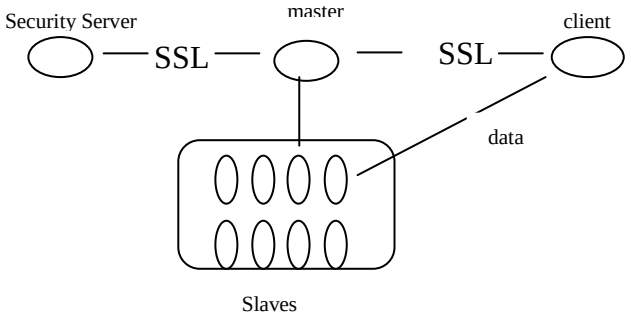

Figure 2: The Sector system architecture.

Figure 2 shows the overall architecture of the Sector system. The security server maintains user accounts, user passwords and file access information. It also maintains lists of internetwork protocol (IP) addresses of the authorized slave nodes, so that illicit computers cannot join the system or send messages to interrupt the system. The master server maintains the metadata of the files stored in the system, controls the running of all slave nodes and responds to users' requests. The master server communicates with the security server to verify the slaves, the clients and the users. The slaves are the nodes that store the files managed by the system and process the data upon the request of a Sector client. The slaves are usually running on racks of computers that are located in one or more data centres.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

IV. SPHERE

The Sphere Compute Cloud is designed to be used with Sector Storage Cloud. Sphere is designed so that certain specialized, but commonly occurring, distributed computing operations can be done very simply. Specially, if a user defines a function **p** on a distributed data set **a** managed by Sector, then invoking the command
Sphere. Run(a,p);

Applies the user defined function **p** to each data record in the dataset **a**. In other words, if the dataset a contains 100,000,000 records **a[i],** then the sphere command above replaces all the code required to read and write the array a[i] from disk , as well as loop:

```
for(int i=0; i<100000000;++i)
p(a[i]);
```

The Sphere programming model is a simple example of what is commonly called a stream programming model. Although this model has been used for some time, it has recently received renewed attention due to its use by the general purpose GPU (Graphics Processing Units) community. Large data sets processed by Sphere are assumed to be broken up into several files.
For example, the Sloan Digital Sky Survey dataset [12] is divided up into 64 separate files, each about 15.6 GB in size. The files are named sdss1.dat, . . ., sdss64.dat. Assume that the user has a written a function called *find-Brown Dwarf* that given a record in the SDSS dataset, extracts candidate Brown Dwarfs. Then to find brown dwarfs in the Sloan dataset, one uses the following

Sphere code:

```
Stream sdss;
sdss.init(...); //init with 64 sdss files
Process* myproc = Sector::createJob();
myproc->run(sdss, "findBrownDwarf");
myproc->read(result);
```

With this code, Sphere uses Sector to access the required SDSS files, uses an index to extract the relevant records, and for each record invokes the user defined function find- BrownDwarf. Parallelism is achieved in two ways. First, the individual files can be processed in parallel. Second, Sector is typically configured to create replicas of files for archival purposes. These replicas can also be processed in parallel. An important advantage provided by a system such as Sphere is that often data can be processed in place, without moving it. In contrast, a grid system generally transfers the data to the processes prior to processing.

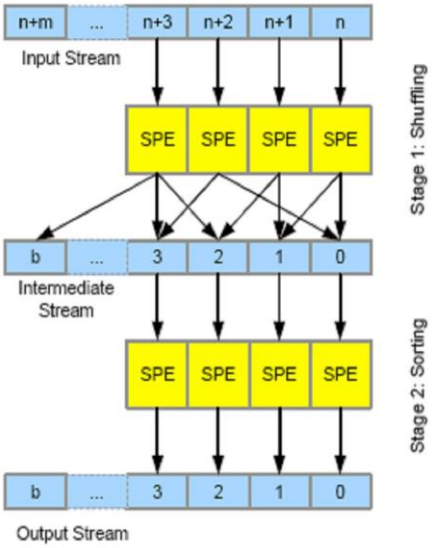

Figure 3**:** Sphere operators process Sphere streams over distributed Sphere Processing Elements (SPE)

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

V. ASSOCIATION RULE

A. DEFINITION 1 CONFIDENCE

Set up I={i1,i2,im}for items of collection, for item in ij(1≤j≤m), (1≤j≤m) for lasting item, D={T1,TN} it is a trade collection, Ti⊑I (1≤i≤N) here T is the trade. Rule r → q is probability that concentrates on including in the trade.

The association rule here is an implication of the form r → q where X is the conjunction of conditions, and Y is the type of classification. The rule r →q has to satisfy specified minimum support and minimum confidence measure .The support of Rule r → q is the measure of frequency both r and q in D S(r) = |r|/|D|

The confidence measure of Rule r → q is for the premise that includes r in the bargain descend, in the meantime includes q C(r → q) =S (rq) /S(r)

B. DEFINITION 2 WEIGHTING SUPPORT

Designated ones project to collect I = {i1, i2, im}, each project ij is composed with the value wj of right(0≤ j ≤1, 1≤j ≤m). If the rule is r → q, the weighting support is

$$S_w(r) = \frac{1}{k} \sum_{i \in r} w_j S(r)$$

And, the K is the size of the Set rq of the project. When the right value wj is the same as ij, we are calculating the weighting including rule to have the same support.

VI. STEPS FOR DATA MINING USING SECTOR, SPHERE AND ASSOCIATION RULE

1. Select the minimum support threshold (Ts) and minimum confidence threshold (Tc), minimum data size (Sizemin) and maximum data size (Sizemax).
2. We now input the data stream to the sphere processing elements. The stream is divided into data segments. The number of data segments per SPE is calculated on the basis of number of SPE and the entire stream size. Data segments from the same file are not processed at the same time until other SPE become idle.
3. The SPE accepts a new data segment from the client, which contains the file name, offset, number of rows to be processed, and additional parameters.
4. The SPE reads the data segment and its record index from local disk or from a remote disk managed by sector.
5. For each data segment find out the frequent term set with length of 1 which is recorded as L1, L1 is used to find the aggregate L2 of frequent 2-term sets, L2 is used to find the aggregate L3 of frequent 3-term sets, and so the cycle continues, until no new frequent k - term sets can be found.
6. We generate strong association rules on the basis of the found frequent term sets i.e. we generate those association rules whose support and confidence respectively greater than or equal to the pre-given support threshold (Ts) and confidence threshold (Tc).
7. For each data segment (single data record, group of data records, or entire data file), the Sphere operator processes the data segment using the association rules and writes the result to a temporary buffer. In addition, the SPE periodically sends acknowledgments and feedback to the client about the progress of the processing.
8. When the data segment is completely processed, the SPE sends an acknowledgment to the client and writes the results to the appropriate destinations, as specified in the output stream. If there are no more data segments to be processed, the client closes the connection to the SPE, and the SPE is released.

VII. CONCLUSION

In this paper, we have described a cloud-based infrastructure designed for data mining large distributed data sets over clusters connected with high performance wide area networks. Sector/Sphere is open source and available through Source Forge. We have used it as a basis for several distributed data mining applications.

The discovery of the association rule is a most successful and most vital duty in the data mining, is a very active research area in current data mining, its goal is to discover all frequent modes in the data set, and the current research work carrying on are mostly focused on the development of effective algorithm. On the basis of in-depth study of the existing data mining algorithms, in this paper a new data mining algorithm based on association rules is presented.

In this paper we have also discussed about the integration of Sector/Sphere framework and Association rule.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

REFERENCES

[1]. en.wikipedia.org/wiki/Cloud_computing
[2]. Yunhong Gu and Robert L. Grossman. "UDT: UDPbased data transfer for high-speed wide area networks". *Computer Networks*, 51(7):1777—1799, 2007.
[3]. Han J , Kamber M. "Data Mining: Concepts and Techniques". 2/e San Francisco: CA Morgan Kaufmann Publishers, an imprint of Elsevier. pp-259-261, 628-640 (2006)
[4]. Robert L. Grossman and Yunhong Gu "Data Mining using high performance data clouds: Experimental Studies using Sector and Sphere" Retrieved from http://sector.sourceforge.net/pub/grossman-gu-ncdm-tr- 08-04.pdf.
[5]Kanhaiya lal and N. C. Mahanti, "A Novel Data Mining Algorithm for Semantic Web Based Data Cloud".
[6]The Sector Project. Sector, a distributed storage andcomputing infrastructure, version 1.4.
[7] Yunhong Gu and Robert L. Grossman "Sector and Sphere: The Design and Implementation of a High Performance Data Cloud".
[8] Ramakrishnan Srikant and Rakesh Agrawal "Mining Generalized Association Rules".

BIOGRAPHY

MISS UZMA ALI working as Assistant Professor in Department of Computer Application at Shri Ramdeobaba College Of Engineering and Management (SRCOEM), Nagpur. She completed her MCA in year 2009. Her research area includes Cloud Computing and networks.

Mrs. Punam Khandar working as Assistant professor in Department of Computer Application at Shri Ramdeobaba College of Engineering and Management (SRCOEM), Nagpur. She completed her MCA in year 2002 and her research area includes Data Mining.