Research Challenges of Data Mining in Cloud Computing

M.L.Hari Prasad
Lecturer in Physics
The Adoni Arts and Science College
Adoni-518302, Kurnool (Dt), AP

Abstract- Cloud computing is a tool which provides resources to the users by means of the Internet. Cloud computing is characterized by its powerful computation capability and storage, and its policy of resource sharing accomplished by virtualization. In this paper, presented challenges in cloud computing, highlighting its main concepts, architectural principles, security and Data Mining models. In the present study, we have focused for different definitions of cloud computing and models of cloud computing, services and security issues in cloud computing. It was found that this is most useful to the researcher and others. The objective of this paper is to provide a good and better understanding for the design challenges of cloud computing and Data Mining identifying important research directions.

I. INTRODUCTION

Computing is being transformed into a model consisting of services that are commoditized and delivered in a manner similar to utilities such as water, electricity, gas, and telephony. In such a model, users access services based on their requirements, regardless of where the services are hosted. Cloud computing is the most recent emerging paradigm promising to turn the vision of "computing utilities" into a reality. Cloud computing refers to both the applications delivered as services over the Internet and the hardware and system software in the datacenters that provide those services [3].
Cloud computing is helping enterprises, governments, public and private institutions, and research organizations shape more effective and demand-driven computing systems. Access to, as well as integration of, cloud computing resources and Systems is now as easy as performing a credit card transaction over the Internet.
The main tasks of Data mining are generally divided in two categories: "Predictive and Descriptive". The objective of the predictive tasks is to predict the value of a particular attribute based on the values of other attributes, while for the descriptive ones, is to extract previously unknown and useful information such as patterns, associations, changes, anomalies and significant structures, from large databases. There are several techniques satisfying these objectives of Data Mining. Discovering association rules is at the heart of data mining. Mining for association rules between items in large database of sales transactions has been recognized as an important area of database research. These rules can be effectively used to uncover unknown relationships, producing results that can provide a basis for forecasting and decision making.
Cloud computing is a new computing paradigm that offers a huge amount of compute and storage resources to the masses. Individuals (e.g., scientists) and enterprises (e.g., startup companies) can have access to these resources by paying a small amount of money just for what is really needed.

1.1. Cloud Service Models

Three most common service models are [4]:
Software as a Service (SaaS): Service capability provided to the consumer to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user specific application configuration settings. Examples are Salesforce.com, CRM, Google Docs, and so on.
Platform as a Service (PaaS): platform service capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer may not manage or control the underlying cloud infrastructure

including networks, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment. Examples are Google's App engine and Microsoft's Azure, and so on.
Infrastructure as a Service (IaaS). This capability is provided to the consumer for processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer may not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications and possibly limited control of select networking components (e.g., host firewalls). Examples include Amazon's EC2 and S3, and so on.

1.2. Data mining in Cloud Computing

Data mining techniques and applications are very much needed in the cloud computing paradigm. As cloud computing is penetrating more and more in all ranges of business and scientific computing, it becomes a great area to be focused by data mining.
"Cloud computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users." [3]
As Cloud computing refers to software and hardware delivered as services over the Internet, in Cloud computing data mining software is also provided in this way.
The main effects of data mining tools being delivered by the Cloud are:

a) The customer only pays for the data mining tools that he needs – that reduces his costs since he doesn't have to pay for complex data mining suites that he is not using exhaustive;
b) The customer doesn't have to maintain a hardware infrastructure, as he can apply data mining through a browser – this means that he has to pay only the costs that are generated by using Cloud computing.

Using data mining through Cloud computing reduces the barriers that keep small companies from benefiting of the data mining instruments.

Cloud computing has become a popular buzzword; it has been widely used to refer to different technologies, services, and concepts. It is often associated with virtualized infrastructure or hardware on demand, utility computing, IT outsourcing, platform and software as a service, and many other things that now are the focus of the IT industry.

Figure 1: Architecture of Cloud computing services [8]

The NIST working definition [2] summarizes cloud computing is a model for enabling, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services). This can be rapidly provisioned and released with minimal management effort or service provider interaction.

II. CLOUD SECURITY

Cloud environment is widely used in industry, companies and researchers. Therefore security is an important aspect for organizations to run on these cloud environments. Since service providers normally mat not have to access the physical security system of data centers. They must rely on the infrastructure provider to achieve full data security.

Cloud computing comes with numerous security issues because it encompasses many technologies including networks, databases, operating systems, virtualization, and resource scheduling, transaction and memory management. Hence, security issues of these systems and technologies are applicable to cloud computing. It is very important for the network which interconnects the systems in a cloud to be secure. Virtualization paradigm in cloud computing results in several security concerns. For example, mapping of the virtual machines to the physical machines has to be performed very securely.

Data security not only involves the encryption of the data, but also ensures which appropriate policies are enforced for data sharing. In addition, resource allocation and memory management algorithms also have to be secure.

To secure data, most systems use a combination of techniques:

2.1. Encryption

A complex algorithm is used to encode information. To decode the encrypted files, a user needs the encryption key.

- Authentication processes

This requires a user to create a name and password.

- Authorization practices

The client lists the people who are authorized to access information stored on the cloud system. Many corporations have multiple levels of authorization. For example, a front-line employee might have limited access to data stored on the cloud and the head of the IT department might have complete and free access to everything.

III. ASSOCIATION RULES

Association rule mining is one of the most important techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among a large set of data items. A typical application is market basket analysis, which studies the buying habits of customers by searching for sets of items that are frequently purchased together [7]. Other application areas include customer segmentation, store layout, web usage mining, software defect detection, telecommunication alarm prediction, and bioinformatics.

Association rule is one of the important themes and essential of data mining which can find out the relationship between item sets in the database. We can use the interesting association relationships which are extracted among huge amounts of data. But, the discovery of association rule is a direct mass-oriented database system that often has hundreds of properties and millions of records, contains a complex relationship between data tables, and remains a time-consuming process. This will inevitably lead to a great surge in search of dimension and space. One of the important problems in data mining is discovering association rules from databases of transactions where each transaction consists of a set of items.

3.1 Problem statement

The problem of mining association rules over market basket analysis was introduced in [10] [12]. It consists of finding associations between items or itemsets in transactional data. As defined in [14], the problem can be formally stated as follows. Let I = {i1, i2 ,…,im} be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that T □ I. Each transaction has a unique identifier TID. A transaction T is said to contain X, a set of items in I, if X ⊆ T. An association rule is an implication of the form X ⇒Y, where X ⊆ I, Y ⊆ I and X ∩Y = ∅.

Each itemset has an associated measure of statistical significance called support. For An itemset X, we say its support is s if the fraction of transactions in D containing X equals s. The rule $X \Rightarrow Y$ has a support s in the transaction set D if s of the transactions in D contains $X \cup Y$. The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a support and confidence greater than given thresholds. These rules are called strong rules.
This association-mining task can be broken into two steps:
Step1. The large or frequent itemsets which have support above the user specified minimum support are generated.
Step2. Generate confident rules from the frequent itemsets.

IV. HADOOP

Hadoop is the parallel programming platform built on Hadoop Distributed File Systems (HDFS) for Map/Reduce computation that processes data as (key, value) pairs. Hadoop has been receiving highlights for the enterprise computing because business world always has the big data such as log files for web transactions. Hadoop is useful to process such big data for business intelligence so that it has been used in data mining for past few years.
Hadoop can compose hundreds of nodes that process and compute peta- or tera-bytes of data working together. Hadoop was inspired by Google's MapReduce and GFS as Google has had needs to process huge data set for information retrieval and analysis [2]. It is used by a global community of contributors such as Yahoo, Facebook, and Twitters. Hadoop's subprojects include Hadoop Common, HDFS, MapReduce, Avro, Chukwa, HBase, Hive, Mahout, Pig, and ZooKeeper etc [6].

4.1 MapReduce

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Map/Reduce programming platform is implemented in the Apache Hadoop project that develops open-source software for reliable, scalable, and distributed computing [8]. It is composed of two functions to specify, "Map" and "Reduce". They are both defined to process data structured in (key, value) pairs.

As the framework showed in Figure 2, Map Reduce specifies the computation in terms of a map and a reduce function, and the underlying runtime system automatically parallelizes the computation across large-scale clusters of machines, handles machine failures, and schedules inter-machine communication to make efficient use of the network and disks.

The map and reduce functions run on distributed nodes in parallel. Each map operation can be processed independently on each node and all the operations can be performed in parallel. But in practice, it is limited by the data source and/or the number of CPUs near that data. The reduce functions are in the similar situation because they are from all the output of the map operations. However, Map/Reduce can handle significantly huge data sets since data are distributed on HDFS and operations move close to data for better performance [3][4]].

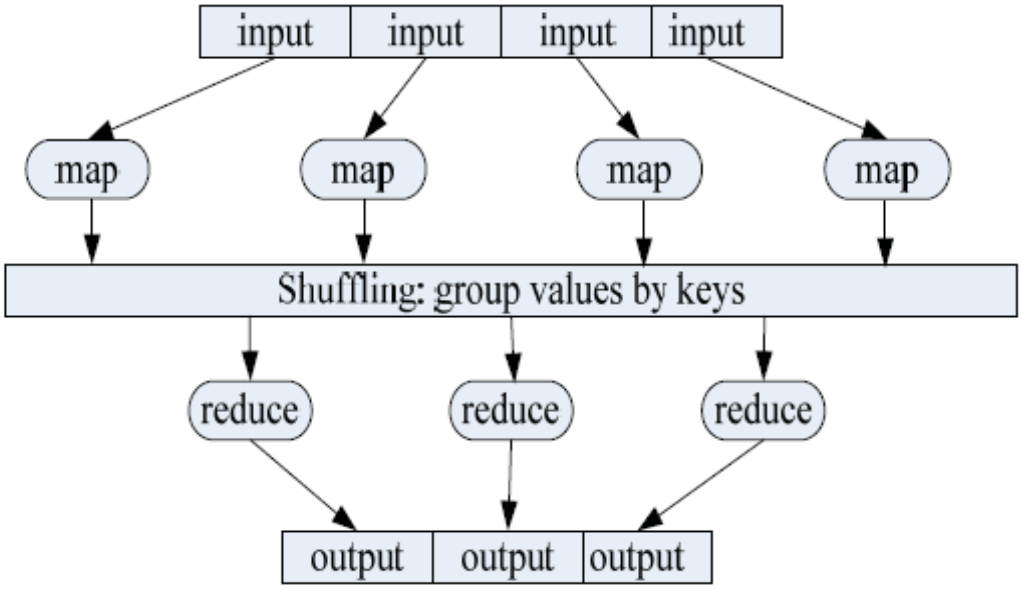

Figure 2: framework of Map Reduce

Hadoop is restricted or partial parallel programming platform because it needs to collect data of (key, value) pairs as input and parallely computes and generates the list of (key, value) as output on map/reduce functions. In map function, the master node parts the input into smaller subproblems, and distributes those to worker nodes. Those worker nodes process smaller problems, and pass the answers back to their master node. That is, map function takes inputs (k1, v1) and generates <k2, v2> where < > represents list or set. Between map and reduce, there is a combiner that resides on map node, which takes inputs (k2, <v2>) and generates <k2, v2>.

In reduce function, the master node takes the answers to all the sub-problems and combines them That is, reduce function takes inputs (k2, <v2>) and generates <k3,v3>..

4.2 Apriori Algorithm based on MapReduce

Figure 3 is the proposed Apriori-Map/Reduce Algorithm that runs on parallel Map/Reduce framework such as Apache Hadoop. prune(Ck+1) function is to remove the nonfrequent item set Ck+1 by eliminating non-frequent item sets Ck as non-frequent item sets cannot be a subset of frequent item sets.

MaPApriori. The steps are as follows.

Map transaction *t* in data source to all Map nodes;
Step1. Use MapReduce model to find the frequent 1-itemsets.
Step2. Set $k = 1$.
Step3.If the frequent (k+1)-itemsets cannot be generated, then goto Step6.
Step4. According to the frequent k -itemsets, use MapReduce model to generate the frequent (k+1)-itemsets.
Step5. If k is less than the max iteration times, then $k++$, goto Step3; Otherwise, continue to the next step.
Step6. According to the frequent itemsets L, generate the Association rules.

Figure 3: Apriori-Map/Reduce Algorithm

V. CONCLUSIONS

Cloud computing is characterized by its powerful capability of computation and storage, as well as its policy of resource sharing accomplished by virtualization. These features render cloud computing valuable merits favorable to data mining service in network environment. Searching for frequent patterns in transactional databases is considered one of the most important data mining problems. The task of finding all association rules requires a lot of computation power and memory.

We will work on bringing together ideas from MapReduce and data mining algorithms, also to combine the advantages of MapReduce-like software with the efficiency and shared work advantages that come with loading data and creating performance enhancing data structures. Therefore, the proposed framework can process large datasets on commodity hardware effectively. Big Data analysis tools like Map Reduce over Hadoop and HDFS, promises to help organizations better understand their customers and the marketplace, hopefully leading to better business decisions and competitive advantages.

REFERENCES

[1] Armbrust M, Fox A, Griffith R, Joseph A, Katz R, Konwinski A, et al. Technical Report No. UCB/ EECS-2009-28 Above the clouds: a berkeley view of cloud computing. USA: University of California at Berkeley; 2009

[2] Apache Hadoop Project, http://hadoop.apache.org/,

[3] Bhagyashree Ambulkar and Vaishali Borkar, "Data Mining in Cloud Computing", MPGI National Multi Conference 2012 (MPGINMC-2012), 7-8 April 2012, Link: http://research.ijcaonline.org/ncrtc/number6/mpginmc1047.pdf

[4] C.Jin, C.Vecchiola, R.Buyya. MRPGA: An Extension of MapReduce for Parallelizing Genetic Algorithms.Fourth IEEE International Conference on eScience 2008.

[5] Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters. In: Proc. of Operating Systems Design and Implementation, San Francisco, CA, pp. 137–150 , 2004

[6] Hadoop: Open source implementation of MapReduce, Available: http://hadoop.apache.org, June 24, 2010

[7] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005.

[8] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters, ACM Commun, vol. 51, Jan. 2008, pp. 107-113.

[9] Jeffrey Dean and Sanjay Ghemawa, "MapReduce: Simplified Data Processing on Large Clusters", Google Labs, pp. 137–150, OSDI 2004

[10] J. Han and M. Kamber, Data Mining—Concepts and Technique (The Morgan Kaufmann Series in Data Management Systems), 2nd ed. San Mateo, CA: Morgan Kaufmann, 2006.

[11] Lammel, R. Google's MapReduce Programming Model - Revisited. Science of Computer Programming 70, 1–30, 2008

[12] MELL, P. and GRANCE, T. 2009. Draft NIST Working Definition of Cloud Computing.

[13] NIST Definition of Cloud Computing v15, csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc

[14] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Database," Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Vol.22, Issue 2, pp. 207-216, 1993