**Suresh Gyan Vihar University Journal of Engineering & Technology**
*(An International Bi-Annual Journal)*
**Vol . 3, Issue 1, 2017, pp 13-18**
**ISSN: 2395-0196**

# Cloud Data Mining Using Sparks & Mapreduce Programming

Pankaj Dadheech[1], Dinesh Goyal[2], Sumit Srivastava[3]

[1]Research Scholar, Department of Computer Science & Engineering, SGVU, Jaipur
[2]Principal, Suresh Gyan Vihar University, Jaipur
[3]Professor, Department of Information & Communication Technology, Manipal University, Jaipur

***Abstract:*** Cloud computing can provide infrastructure to complex and massive data of data mining, as well as new challenging issues for data mining of cloud computing, Big Data Analytics are emerged. The research of parallel programming mode especially analyses the Map-reduce programming model and it's development platform-Hadoop; finally, overviews efficient mass data mining algorithm based on parallel programming model and mass data mining service based on the cloud computing. This paper gives the basic concept of cloud computing and data mining and explains how data mining is used in cloud computing. It describes the research of parallel programming mode especially analyses the Map-reduce programming model and it's development platform-Hadoop. It introduces the efficient mass data mining algorithm based on parallel programming model and mass data mining service based on the cloud computing. It will help to Make in India campaign of Government of India by overcome the complexity of Data Mining during peak hours while all users are accessing simultaneously the same data for many times on a cloud platform or to process the large amount of data.

***Keywords****: Cluster, MapReduce, Hadoop, HSim, Spark.*

## INTRODUCTION

Cloud computing provides people the way to share distributed resources and services that belong to different organizations or sites. A cloud based application is based on network appliance software, with its operating system, running in a virtual machine in a virtualized environment. A virtual appliance relieve some of the notable management issues in enterprises because most of the maintenance, software updates, configuration and other management tasks that they are done by cloud provider which responsible for them. Data Mining is a process of extracting potentially useful information from raw Data, so as to improve the quality of the information service. MapReduce is a distributed programming model for data intensive tasks which has become an enabling technology in support of Cloud Computing. Popular implementations of the MapReduce model include Mars, Phoenix, Hadoop and Google's implementation. Among them, Hadoop has become the most popular one due to its open source feature. Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. Hadoop MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

## PRESENT STATUS

Cloud computing is a general term for anything that involves delivering hosted services over the Internet. These services are broadly divided into three categories: Infrastructure as a

Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). The name cloud computing was inspired by the cloud symbol that's often used to represent the Internet in flowcharts and diagrams. Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery, from data. Data mining Techniques involve sophisticated algorithms, including Decision Tree Classifications, Association detection, and Clustering. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns in large data sets.

Data mining parameters include:

- Association - Looking for patterns where one
- event is connected to another event.
- Sequence or path analysis - Looking for patterns
- where one event leads to another later event
- Classification - Looking for new patterns
- Clustering - Finding and visually documenting
- groups of facts not previously known.
- Forecasting - Discovering patterns in data that can
- lead to reasonable predictions about the future.

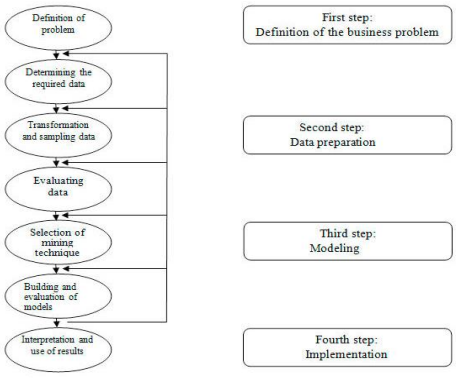


Figure 1: Phases of the Data Mining Process

The key contributions of H Sim lie in its high accuracy in simulating the dynamic behaviour of Hadoop environments and the large number of Hadoop parameters that can be modelled in the simulator. The main focus in H Sim is to accurately simulate the behaviour of Hadoop framework. Using H Sim, the performances of Hadoop applications can be studied from a number of angles, including the impacts of the parameters on the performance of a Hadoop cluster, the scalability of a Hadoop application in terms of the number of nodes used, and the impact of using heterogeneous environments. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage. As an emerging framework, Spark, which is designed to have a global cache mechanism, can achieve better performance in response time since the in-memory access over the distributed machines of cluster will proceed during the entire iterative process.

**OBJECTIVES & SCOPE**

Big Data is (relatively) new term for large and complex data sets that cannot be processed and maintained by using traditional tools for managing databases. The Hadoop File System was developed using the distributed file system design. It runs on commodity hardware. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware. HDFS holds a very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in a redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing. HDFS is the Hadoop file system and comprises two major components: namespaces and block storage service. The namespace service manages operations on files and directories, such as creating and modifying files and directories. The block storage service implements data node cluster management, block operations and replication. YARN is often called the operating system of

Hadoop because it is responsible for managing and monitoring workloads, maintaining a multi-tenant environment, implementing security controls, and managing high availability features of Hadoop. Hadoop has an architecture consisting of a master node with many client workers and uses a global queue for task scheduling, thus achieving natural load balancing among the tasks. The Map Reduce model reduces the data transfer overheads by overlapping data communication with computations when reduce steps are involved. Hadoop performs duplicate executions of slower tasks and handles failures by rerunning the failed tasks using different workers. Spark is a top-level Apache project that leverages memory to improve the speed of large-scale data analysis programs. Programs can be written in Java, Scala or Python.

**HADOOP WORKING CONFIGURATION**

Hadoop is an open source software framework that supports data-intensive distributed applications. Hadoop's most touted benefit is its ability to store data much more cheaply than can be done with RDBMS software. Hadoop YARN is a framework for job scheduling and cluster resource management. Hadoop is a large-scale distributed batch processing infrastructure. Hadoop is also designed to efficiently distribute large amounts of work across a set of machines. Hadoop Distributed File System (HDFS) provides high-throughput access to application data. HDFS, the Hadoop Distributed File System, is a distributed file system designed to hold very large amounts of data (terabytes or even petabytes), and provide high-throughput access to this information. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. HDFS uses a master/slave architecture where master consists of a single NameNode that manages the file system metadata and one or more slave DataNodes that store the actual data. Spark has designed to run on top of Hadoop and it is an alternative to the traditional batch map/reduce model that can be used for real-time stream data processing and fast interactive queries that finish within seconds. So, Hadoop supports both traditional map/reduce and Spark. Data mining in Cloud is a very tedious process that requires a special infrastructure based on application of new storage technologies, handling and processing. Big Data/Hadoop is the latest hype in the field of data processing. Based on the algorithms and technologies developed by large Internet companies, there is a quite widespread ecosystem of solutions for processing and analysis of huge amounts of data.

**EXPERIMENTS & ANALYSIS**

MapReduce is a module that is used for highly distributed processing of large data sets using thousands of computers. Introduced in 2004 by Google, MapReduce can be seen as a framework or system for the execution of a query in the background. Regardless of the amount of data, the system processes the entire data set for each query.

Processing is defined by two functions:

1. Map: transparently reading “raw data” from a distributed file system, filtering and generating pairs of key-value.
2. Reduce: processing of associated and sorted pairs generated Map functions and generating output in the key-value format.

MapReduce is a fundamental concept of processing in Hadoop environment. Subsystem for performing MapReduce programs in Hadoop makes a major node, which is called „job tracker“, and a set of node’s workers is called „task tracker“. MapReduce program sent to perform an action is called "job". Hadoop divides the job into a set of tasks. Entrance to the MapReduce program is a set of data stored within the distributed file system.

Performing of tasks is completely under the control of the main node. Before the performance of specific tasks, "job tracker" must choose to which job task it belongs, that will run. Anticipated job scheduler selects the first job that comes into the job queue. After selecting the job, job tracker assigns

tasks that make him free worker. Task tracker periodically reports its state to head node, where the situation represents information on the number of available slots for Map/Reduce tasks. After Map / Reduce tasks are granted, significant optimization is accomplishing. Specifically, the Map tasks are assigned to nodes' workers that contain their own data that handles just the assigned task. This is extremely important because in this way we avoid the (expensive) network communication. The job ends when a node worker that performs the last task is presented to the head node as the one that has completed the assigned task. Parallel computing model is a bridge between user needs and the underlying hardware system, it makes the parallel algorithm become more intuitive and more convenient for processing the large-scale data. According to the user the hardware environment, parallel programming model can be divided into multi-core machines, GPU computing, mainframe computers and computer clusters. One of the solutions, surely, can offer the integration of in-depth analysis of data (data mining) and Cloud Computing. Huge storage and processing potential of Cloud Computing, and well known techniques and methods of data mining, which have "moved to the Cloud," create a powerful platform for analyzing vast amounts of data that is produced daily and in itself it hides much (useful) information, which is the basis for new knowledge and better business decisions, which, in return, is ultimately the main goal. By developing cloud based data mining solutions accessing data mining services every time and everywhere and from various platforms and devices will be made possible. Ultimately, the application of Cloud Data mining solutions can provide a sort of knowledge discovery ecosystem built of a large numbers of decentralized data analysis services.

HSim follows a master–slave mode. The simulated Map instances (MapperSim), Reduce instances (ReducerSim), JobTracker and TaskTrackers are located on these nodes. The Master node is the Namenode of Hadoop framework which contains JobTracker to correspond and schedule the tasks. The Slave nodes are the Datanodes of Hadoop framework which contains TaskTrackers. On Slave nodes Map instances and Reduce instances perform data processing tasks. when a job is submitted to a simulated Hadoop cluster, the JobTracker splits the job into several tasks. Then TaskTracker and JobTracker will communicate with each other via messaging based on heartbeats. If the Job-Tracker finds that all the Map tasks have been finished, then the Reduce instances will be notified to be ready for merging phase. Moreover if the JobTracker finds all Reduce tasks have been finished, then the job will be considered as finished. If the Map tasks have not been finished yet, the TaskTrackers will be notified to choose a Map task or a Reduce Task based on their availabilities.

**RESULTS & DISCUSSION**

Huge amounts of data daily produce and in themselves hide potentially useful information. The data that is processed does not originate only from multiple information system of companies, giant amount of it comes from "on-line" environment, with a variety of services that users use for both commercial and private purposes. The Hadoop framework is a complex system involving a number of components. HSim is designed and implemented to simulate such components and interactions. It works similarly like the way of the Hadoop framework works. However we cannot simply conclude that HSim can accurately simulate Hadoop without any limitations. The accuracy of HSim can be affected by a number of factors such as the time of job propagations, cold starts of Map instances, key distributions, system communications, shared hardware resources and dynamic IO loads. These dynamic factors may affect the performance of both experimental and simulated results depending on user applications. Enabling the Combiner feature of Hadoop also can affect the accuracy of HSim. However, the combiner instance has not been fully implemented in HSim. A combiner can be considered as an

in-memory sort process. The output of mappers will be combined and written into an intermediate file by a combiner. And then the file will be sent to a reducer. So when the number of mappers is small, the benefits gained from using combiners are not significant. HSim was validated with established benchmark results and also with experimental environments which have shown that HSim can accurately simulate the dynamic behaviors of Hadoop clusters. HSim can be used to investigate the impacts of the large number of Hadoop parameters by tuning their values. It can also be used to study the scalability of MapReduce applications which might involve hundreds of nodes. A remarkable characteristic of the Hadoop framework is its support for heterogeneous computing environments. Therefore computing nodes with varied processing capabilities can be utilized to run MapReduce applications in parallel.

## CONCLUSION

Through big data storage and distribution of computing in cloud computing, the new ways to effectively solve the distributed storage of massive data mining and efficient computing has been found. To carry out the research of the data mining based on cloud computing can provide the new theory and support tools for data mining in more complex and more mass data. As extension of traditional data mining, mass data mining based on cloud computing will drive the Internet advanced technological achievements in the public service, is a new method to share and use information resources efficiently.

## REFERENCES

[1]. B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha, P. Dhavachelvan, "Big Data and Hadoop-A Study in Security Perspective", *Elsevier: 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)*, Procedia Computer Science 50 ( 2015 ) 596 – 601

[2]. Jiangtao Yin, Yong Liao, Mario Baldi, Lixin Gao, Antonio Nucci, "GOM-Hadoop: A distributed framework for efficient analytics on ordered datasets", *Elsevier:* J. Parallel Distrib. Comput. 83 (2015) 58–69

[3]. Ivanilton Polato, Reginaldo Ré, Alfredo Goldman, Fabio Kon, "A comprehensive view of Hadoop research-A systematic literature review", *Elsevier: Journal of Network and Computer Applications* 46(2014)1–25

[4]. Mohd Rehan Ghazi, Durgaprasad Gangodkar, "Hadoop, MapReduce and HDFS: A Developers Perspective", *Elsevier: International Conference on Intelligent Computing, Communication & Convergence (ICCC-2014)*, Procedia Computer Science 48 (2015 ) 45 – 50

[5]. Hamza Zafar, Farrukh Aftab Khan, Bryan Carpenter, Aamir Shafi and Asad Waqar Malik, "MPJ Express Meets YARN: Towards Java HPC on Hadoop Systems", *Elsevier: ICCS 2015 International Conference on Computational Science*, Volume 51, 2015, Pages 2678–2682

[6]. Xuanhua Shi, Ming Chen, Ligang He, "Mammoth: Gearing Hadoop Towards Memory-Intensive MapReduce Applications" *IEEE Transactions on Parallel and Distributed Systems*, Vol. 26, No. 8, August 2015

[7]. Tom White, "Hadoop: The Definitive Guide", *ISBN: 978-1-449-31152-0*, O'Reilly, 3rdEdition May 2012 (2012-01-27)

[8]. Alex Holmes, "Hadoop in Practice", *ISBN 9781617290237*, Manning Publications Co. 2012 (NY 11964)

[9]. Srinath Perera, Thilina Gunarathne, "Hadoop MapReduce Cookbook", *ISBN 978-1-84951-728-7*, Packt Publishing Ltd. February 2013 (2250113)

[10]. Jonathan R. Owens, Jon Lentz, Brian Femiano, "Hadoop Real-World Solutions Cookbook", *ISBN 978-1-84951-912-0*, Packt Publishing Ltd. February 2013 (1280113)

[11]. Xia Geng, Zhi Yang, "Data Mining in Cloud Computing", *2013*, Published by Atlantis Press

[12]. Robert Vrbic, "Data Mining and Cloud Computing*", Journal of Information Technology and Applications* (*JITA), December 2012, 2:75-87* DOI: 10.7251/JIT1202075V

[13]. Yang Liu, Bin Wu, Hongxu Wang, and Pengjiang Ma, "BPGM: A Big Graph Mining Tool", *Tsinghua Science and Technology*, ISSN:1007-0214, 04/10, pp33-38, Volume 19, Number 1, February 2014