iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** III **Month of publication:** March 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 3 Issue III, March 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Data mining in cloud computing

Snehal Govind Tagalpallewar[1], Prof.prajakta chapke[2]
Computer Science and Engineering, H.V.P.M College of engineering Amravti, India

Abstract: ***Data Mining is a process of extracting potentially useful information from raw Data, so as to improve the quality of the information service. With the rapid development of the Internet, the size of the data has increased from KB level to TB even PB level; The object of data mining is also more and more complicated, so the data mining algorithm need to be more efficient. Cloud computing can provide infrastructure to massive and complex data of data mining, as well as new Challenging issues for data mining of cloud computing research are emerged. This paper introduces the basic concept of cloud computing and data mining firstly, and sketches out how data mining is used in cloud computing; Then summarizes 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 base parallel programining service based on the cloud computing.***
Keywords: Data Mining, Cloud Computing, Map-Reduce, Hadoop

I. INTRODUCTION

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. The term "cloud" is used as a metaphor for the Internet, based on the cloud drawing used in the past to represent the telephone network, The actual term "cloud" borrows from telephony in that Telecommunications companies, who until the 1990s offered primarily dedicated point-to-point data Circuits, began offering Virtual Private Network (VPN) services with comparable quality of service But at a much lower cost. In early 2008, Eucalyptus became the first open-source, AWS API-compatible platform for deploying private clouds. In early 2008, Open Nebula, enhanced in the RESERVOIR European Commission-funded project, became the first open-source software for deploying private and hybrid clouds, and for the federation of clouds. Cloud computing is becoming one of the buzz words of next industry. It joins the ranks of terms including: grid Computing, utility computing, virtualization, clustering, etc.Cloud computing overlaps some of the concepts of distributed, grid and utility computing, however it does have its own meaning if contextually used correctly. The conceptual overlap is partly due to technology changes, usages and implementations over the years. the computing paradigm shift on the last half century through six distinct phases:

Phase 1: people used terminals to connect to powerful mainframes shared by many users.
Phase 2: stand-alone personal computers became powerful enough to satisfy users' daily work.
Phase 3: computer networks allowed multiple computers to connect to each other.
Phase 4: local networks could connect to other local networks to establish a more global network.
Phase 5: the electronic grid facilitated shared computing power and storage resources.
Phase 6: Cloud Computing allows the exploitation of all available resources on the Internet in a scalable and simple way.

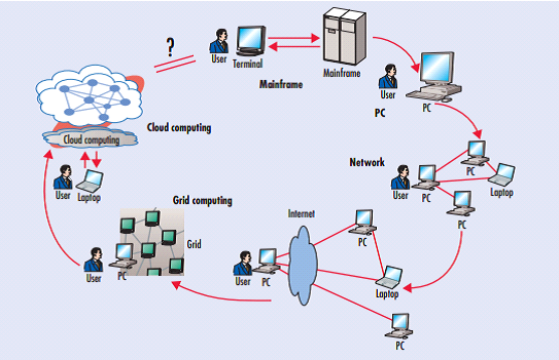


Figure 1. Computing paradigm

©IJRASET 2015: All Rights are Reserved

A. Data Mining

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. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees (1960s) and support vector machines (1990s). 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:

1) Association - Looking for patterns where one event is connected to another event.
2) Sequence or path analysis - Looking for patterns where one event leads to another later event
3) Classification - Looking for new patterns
4) Clustering - Finding and visually documenting groups of facts not previously known
5) Forecasting - Discovering patterns in data that can lead to reasonable predictions about the future This area of data mining is known as predictive analytics. So there are many applications of Data mining in real world As, Hospital, Student Management,Airline Reservation, Forecasting, Biometrics,Mathematics, Geographical, Web Mining, Parallel

Processing, Space Organization, Data Integrity, etc. And in which the data mining term is very useful. But how to efficiently implement data mining in the platform of cloud computing.

B. Parallel programming model

In order to make the users achieve parallel computing results through a simple development, a series of parallel computing models have been proposed by researchers. 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. Commonly used parallel programming interfaces and models include:

pThread: pThread is a common multithreaded programming API on Unix systems, it provides users with a series of function to created and manage the threads, and enables users to easily write multithreaded programs.

MPI: MPI(Message Passing Interface) which provides users with a range of interfaces.in this model, the users establish inter-process communication mechanism by messages, so the algorithms can be parallel implemented easily.

Prege: Google's Pregel is a programming model for graph algorithms , it provides parallel
algorithm support of large - scale graph computing. A typical Pregel calculation process will be carried out on graph by a series of Super Steps , in each super- step , all the vertices of calculations perform in parallel function of the user-defined , and the process is stopped by a vote mechanism.

CUDA:CUDA is a GPU-based parallel computing model proposed by NVIDIA .

C. Data mining techniques and technique

1) Clustering: Useful for exploring data and finding natural groupings. Members of a cluster are more like each other than they are like members of a different cluster. Common examples include finding new customer segments and life sciences discovery.

2) Classification: Most commonly used technique for predicting a specific outcome such as response / no-response, high / medium / low value customer.

3) Association: Find rules associated with frequently co-occurring items, used for market basket analysis, cross-sell, root cause analysis. Useful for product bundling, in-store placement, and defect analysis.

4) Regression: Technique for predicting a continuous numerical outcome such a customer lifetime value, house value, process yield and rates.

5) Attribute Importance: Ranks attributes according to strength of relationship with target attribute. Use cases include finding factors most associated with customers who respond to an offer, factors most associated with healthy patient.

©IJRASET 2015: All Rights are Reserved

II. APPLICATION

A. Hospital
B. Student management
C. Airline Reservation
D. Forecasting
E. Biometrics

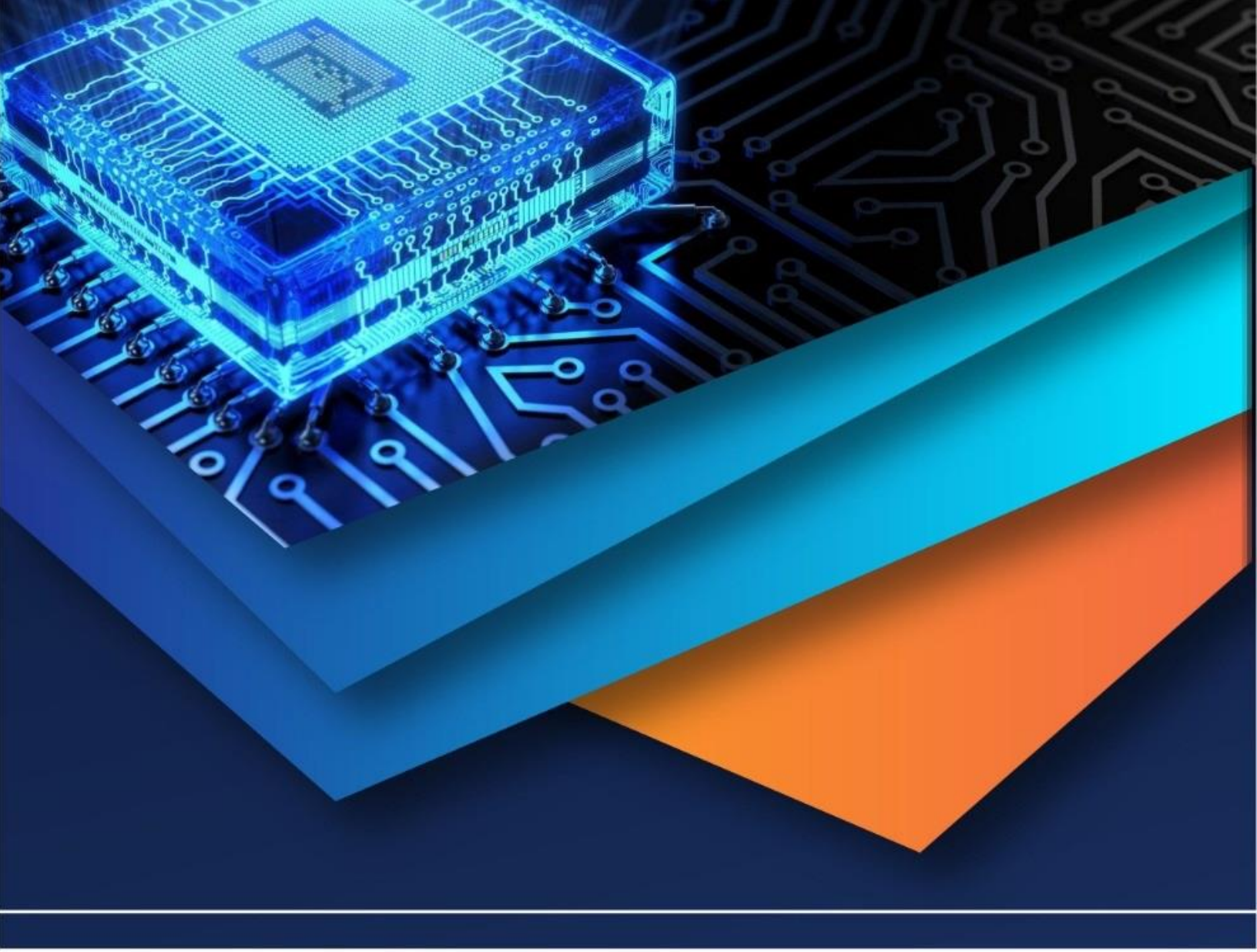
III. CONCLUSIONS

Data mining technologies provided through Cloud computing is an absolutely necessary characteristic for today's businesses to make proactive, knowledge driven decisions, as it helps them have future trends and behaviors predicted. This paper provides an overview of the necessity and utility of data mining in cloud computing. As the need for data mining tools is growing every day, the ability of integrating them in cloud computing becomes more and more stringent.

REFERENCES

[1] Ruxandra-stefania PETRE Database system jornal vol.no.3/2012
[2] IEEE standard for information technology.
[3] www.technetwork,com

©IJRASET 2015: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)