IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

Data Mining with Big Data under Cloud Environment – Opportunities, Issues and Challenges

R. Kabilan[1], Dr. N. Jayaveeran[2]

Research Scholar, P.G and Research Department of Computer Science, Khadir Mohideen College, Adirampattinam[1]

Professor and Head, P.G and Research Department of Computer Science, Khadir Mohideen College, Adirampattinam[2]

Abstract: We are awash in a flood of data today, the day-by-day generation of data becomes unprecedented scale. The true of it lies in collecting these data, explore and carry out some computations to get some significant information. This information can be derived using some data mining techniques. In short we can call big data as an asset and data mining is a technique that is used to provide data visualization. To perform these study data mining techniques can be used and also the big data methods under cloud environment.

Keywords: Data Mining, Big Data, Cloud Computing.

I. INTRODUCTION

Data Mining over Big Data and Cloud Computing are considered as major technologies that can support suitable resource sharing. Data mining is considered as an important method as it is used for finding fresh, suitable, valuable and clear forms of data. Big Data is a new term used to recognize the datasets that due to their huge size and complexity, we cannot manage them with our current data mining tools.

Data Mining over Big Data is the potential of extracting useful information from these large datasets or streams of data, that due to its quantity, inconsistency and speed, it was not possible before to do it. Cloud computing is a ingenious technology that can support a wide range of applications. Data mining tasks and applications can be effectively used in cloud computing model. The data mining tasks in cloud computing provides an elastic and scalable structural design which can reduce the cost of infrastructure and storage and used for efficient mining of vast amount of data from virtually incorporated data sources with the aim of producing useful information which is supportive in decision making to predict the future trends and behavior. But it has the risk of privacy of data user and security.

II. DATA MINING TECHNIQUES

Data mining is the method of finding interesting patterns from huge amounts of data, where the data can be stored in various data sources. Various data Mining techniques are [2]:

A. Popular Data mining tools

WEKA – It is java based tool. Free to use. It includes association, classification, clustering, visualization and modeling techniques.

Techniques	Algorithm
Association Rule mining	Apriori Algorithm,
Classification	ID3,C4.5, SLIQ, Nearest-neighbor, Naïve Bayes, Oblivous read-Once, Lazy decision trees, Decision Table
Clustering	K-Means, K-Medoids, CLARANS, HAC, Self-Organizing Feature Map
Regression	Multivariate Linear Regression Generalized Linear Models, Support Vector Machines
Time Series	Randomized Algorithms, Las Vegas algorithm

Rapid miner – Open source, no-coding required. Java based. It integrates data mining techniques data preprocessing, prediction and visualization

R-Programming Tool – FORTRAN and C based tool. It supports data miners for classification, analysis, clustering and time-series data analysis.

ORANGE – It is python based tool. It supports machine learning, data analytic, and text analysis.

KNIME – GUI supported tool. Popular tool for Data Analysis.

SAS– It provides prediction and Visualization of models dimension.

III. BIG DATA

Big data means huge amount of data shaped very fast by variety of sources.[1] Data can either be created by public or generated by technology such as sensors, satellite images, digital pictures and videos, GPS signals, etc.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

A. Big data Characteristics

Volume - The amount of stowed and made data. The size of the data determines the worth and possible insight- and whether it can actually be considered big data or not.

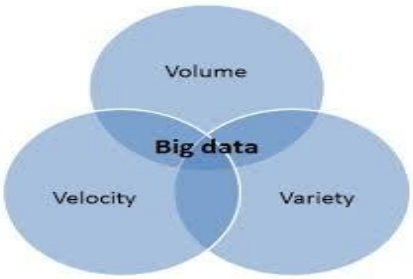

Figure : 3Vs of Big Data

Variety - The type and environment of the data. It assists vision into the individuality of dissimilar classes of big data.

Velocity - The speed at which the data is processed and to meet the load and challenges that lie in the path of expansion and development.

B. Trendy BIG Data Tools

Apache Hadoop - Hadoop is an open source, Java-based. It supports the processing and storage of large data sets in a distributed computing environment.

Apache Mahout - Open source software based mostly in Hadoop. It has implementations of machine learning and data mining algorithms: classification, clustering, collaborative filtering and frequent pattern mining.

R - Open source. Designed for statistical computing and visualization.

Pegasus – Open source. Graph mining system built on top of Map Reduce. It permits discover patterns and anomalies in enormous real-world graphs.

SPARK- Spark provides an interface for programming entire clusters with hidden data parallelism and fault-tolerance.

C. Challenging issues with Big data

When managing big data challenges take place in the following areas.

1. Data Capture and Storage.
2. Data Searching and Indexing
3. Data Sharing
4. Data Integration
5. Data Transmission.
6. Data Curation.
7. Data Analysis.
8. Data Visualization.

D. Comparison of Data Mining with Big Data

Data Mining	Big Data
Data mining is the old	Big data is the whole thing in the world now
Data size is less significant	Data size is bigger
Extracting or "mining" knowledge from large amounts of data.	Big data is described as high in Volume, Velocity and Variety - data
All data mining tasks are not big data	All big data tasks are data mining
Details and close look for data	Overall look among hug data and relationship
Data mining relates to the process of going through large sets of data to identify relevant or pertinent information	It can be clear simply as huge data sets that outgrow simple databases and data management architectures

Figure: Data Mining Techniques with Big Data under Cloud Environment

IV. CLOUD COMPUTING

Cloud computing is a type of Internet- based computing. It delivers common resources and data to PCs and other devices on request.

Private cloud – This model especially operated for an organization. More control and Security.

Public cloud – This model made available to the universal public. Free or pay-per - usage. Simple and reasonably priced model.

Hybrid cloud - It is an arrangement of two or more clouds-private, community or public. It can join comparison, achieved and/or devoted services with cloud resources.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

A. Types of cloud services
Cloud services can be classified into three categories:
Infrastructure as a service
Platform as a service
Software as a service

Infrastructure-as-a-service (IaaS)
Provide virtualized server and software. It will be hosted inside the Cloud data center. Beneficiaries pay only the agreed on cost. No need to purchase hardware / software and license.

Platform as a service (PaaS)
In addition to the IaaS application layer, which the customer intends to use.

- Runtime Environment for Applications
- Development and Data Processing Platforms

Software as a service (SaaS)
SaaS is offering software to the others remotely as a Web-based service. It allows organizations to access business functionality at a price usually below paying for certified applications and using resources for that application.

B. Cloud challenges
Cloud computing moved away from PCs and an enterprise application server to cloud services. The outward show of cloud computing has made a spectacular impact on the industry. Right now industry desires Cloud services for best opportunities to the society.

C. Cloud computing issues
We get any new technology, the implementation of cloud computing is not free from issues. Some of the most important issues are follows.

Technical issues	Legal issues
Security and Privacy	Data protection and Privacy
Service quality	Lack proper Contract and Service Level Agreement between Cloud user and Provider
Interoperability, Portability, Performance and Bandwidth Cost	Conflict of Law
Reliability and Availability	Multi jurisdictional issues
Problem with internet bandwidth	Lack of Transparency and Accountability
Data recovery and backup	Data process and data storage
Loss or bankruptcy of service provider	E-discovery and digital investigation
Improper Cloud Service Exit policy	Data deletion, alteration and leak
Vendors Lock-in	IPR protection Issues
Multi-tenancy , Virtualization	

V. DATA MINING WITH BIG DATA IN CLOUDS

It is a new paradigm[7] for next generation analytics development, enabling large scale data organization, distribution, knowledge discovery, decision making and penetrating of large volumes rapidly growing diversity forms of data using Cloud computing as a back end large scale service-oriented computational set-up facility.
This paradigm combines huge scale compute, new data exhaustive techniques and precise models to build data analytics for built-in information extraction.
Organizations uphold to store more and more data in cloud environments, which signifies enormous, expensive source of information to mine and clouds offer commercial users scalable resources on request.

Cloud computing is emerged as service oriented computing model, to distribute infrastructure, platform and applications as services from the providers to the consumers meeting the Quality of Service parameters by enormous volumes of data at faster scale based on market models.
Data mining is a technique that is used to provide data visualization. Big Data demands enormous computing data resources and Clouds present huge scale set-up, hence both these technologies could be combined.

A. Data Mining, Big data across public and private clouds
Data Mining with big data can be used to paddling through log files, internal strife click streams, transaction analysis, dealing with the social media, to avoid fraud and trying to manage protein series. Cloud computing is a normal fit for Data Mining and Big data analytics. Flexible compute capacity and on-demand provisioning make analytics accessible to more teams within society, while Apache Hadoop has reduced the time to complete analysis. Here design, deployment, operation and Mining big data cloud applications across public and private clouds maintained by Cloud Management.
Cloud architectures includes an arrays of virtual machines that are model for the processing of very huge data sets, to the extent that processing can be segmented into several parallel processes.
Mining of data from various data sources is monotonous. And also the big data is stored at cloud environment. Applying data mining techniques with big data and the data is processed using parallel computational methods. Finally, the data is combined, predicted using visualization methods.

VI. OPEN AREAS FOR FURTHER RESEARCH

Data visualization is becoming more and more significant component of analytics in the age of big data under cloud environment.

Big data tools and operation platforms -Conservative tools are incompetent to handle Big data, Various research is needed in these areas.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

Managing Consistency for Big Data Applications on Clouds Analytics-as-a-service models for cloud-based Data Mining with big data analytics Use Natural Language processing techniques on Data Mining and Big Data under cloud to find out the current sentimental trend.
Development of high performance clustering algorithms that can run in parallel on distributed architectures

VII. CONCLUSION

In this paper, we have identified the challenges, opportunities and issues. We have provided an insight into the possible solutions to these problems even though lot of work is needed to be done in this regard. With the maturity of Cloud computing technologies, Data Mining and Big Data technologies are accelerating in several areas of business, science and engineering to solve data intensive problems.

REFERENCES

[1] www.rightscale.com/solutions/cloud-computing-uses/big-data
[2] Shobana.V, Maheshwari. S, Savithri.M - Study on Big data with Data Mining International Journal of Advanced Research in Computer and Communication Engineering. Vol. 4, Issue 4, April 2015
[3] Wikipedia.org/wiki/Big_data
[4] Xindong Wu, Xingquan Zhu, Gong Qing Wu, WeiDing, Data mining with Big data, IEEE, Volume 26, Issue 1,
[5] January 2014.
[6] Global Journal of Computer Science and Technology, Volume 11, Issue 11 Version 1.0,July 2011
[7] Publisher: Global Journals Inc. (USA)
[8] IbrahimAbakerTargioHashem A.N IbrarYaqoob.A BadrulAnuar .A Salimah Mokhtar .A, AbdullahGani. A, SameeUllahKhan.B - The rise of "big data" on cloud computing: Review and open research issues NDSU-CIIT
[9] Raghavendra Kune - Big Data Computing in Clouds– Data Aware Scheduling and Extended MapReduce for Scientific Analytics Thesis , April, 2016.
[10] Intel IT Center, Solution Brief Big Data in the Cloud: Converging Technologies,
[11] How to Create Competitive Advantage Using Cloud-Based Big Data Analytics April- 2015
[12] https://selecthub.com/business- intelligence/bi-vs-big-data-vs-data-mining/
[13] https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/
[14] www.sas.com/resources/asset/five-big-data-challenges-article.pdf
[15] www.qubole.com/resources/article/big-data-cloud-database-computing/
[16] wikipedia.org/wiki/Cloud_computing