November/December 2020
ISSN:0193-4120 Page No. 129 – 138

Design and Implementation of Cloud Computing Based Mixed Big Data Mining

1Dr. BVS Varma, Professor, Dept. of CSE, DNR College of Engineering and Technology, Bhimavaram, A.P, India
2Dr. A Ramamurthy, Professor, Dept. of CSE, DNR College of Engineering and Technology, Bhimavaram, A.P, India

Article Info
Page Number: 129 – 138
Publication Issue:
November/December 2020

Article History
Article Received: *25 October 2020*
Revised: *22 November 2020*
Accepted: *10 December 2020*
Publication: *31 December 2020*

ABSTRACT: With the application and popularization of Internet, the data of network resources are becoming more and more massive and diversified. The amount of digital data is increasing beyond any previous estimation and data stores and sources are more and more pervasive and distributed. So much data will undoubtedly bring people a vast amount of information, but the difficulty of finding useful knowledge for the enterprise or individual from the vast amount of data has increased. This paper elaborates the design and implementation of Cloud Computing based mixed Big Data Mining. Cloud computing platform can perform dynamic resource scheduling and allocation, with the characteristics of highly virtualization and high availability, which meets the needs of efficient data mining. Taking the multifunctional Hadoop big data mining platform as an example, this article analyses the internal workflow of big data mining. The performance of described model is evaluated through the execution of workflow-based data analysis applications on a pool of virtual servers hosted by a Microsoft Cloud data center. The experimental results demonstrated the effectiveness of the framework.

KEYWORDS: Cloud Computing, Data Mining, Cloud data center, Big Data.

I. INTRODUCTION

In recent years, with the continuous development of information technology, communication technology and network technology, and related network derivative

Published by: The Mattingley Publishing Co., Inc.

November/December 2020
ISSN:0193-4120 Page No. 129 – 138

services, such as broadcasting network, mobile network and Internet have expanded rapidly [1]. These data formed a large amount of distributed data based on cyberspace. There is tremendous value in these data that can provide the basis for decisionmaking.

Data centres and Web servers supporting Internet applications are uninterruptedly more and more pervasive and distributed [2]. New ways to efficiently compose different distributed models and paradigms are needed and relationships between hardware resources and programming levels must be addressed. Cloud computing platforms offer a real and scalable support for addressing both the computational and data storage needs of big data mining and parallel knowledge discovery applications [3]. Complex data mining tasks involve data-intensive and compute-bound algorithms that require large and efficient storage facilities together with high performance processing units to get results in adequate times.

Data mining is a process to extract implicit potentially useful information and knowledge from large, incomplete, noisy, fuzzy and random of the actual data [4]. Data mining can be seen as an important technology in the field of knowledge discovery from the definition of data mining. It involves artificial intelligence, machine learning, pattern recognition, statistics such as technology; specific techniques include characteristic, association, clustering, prediction analysis and etc [5]. Data mining has been widely applied in the Internet, mobile Internet, telecommunication, finance, scientific research and other fields. The traditional data mining technology is to make data compute based on relational database and data warehouse, and find out the relations hidden in the data model [6]. And it makes data access in large-scale data and statistical calculation, the entire mining process needs to consume large amounts of computing resources and storage resources.

Data mining algorithms which are employed currently are only appropriate for small-scale data discovery. Mining efficiency gets lower for generous I/O data and nodes working differently. To overcome shortcomings mentioned above and to mine super-large scale of data of information in online examination systems, it's necessary to introduce a new data mining algorithm which can run in the cloud computing environment, using fully computing resources in clusters in the cloud computing environment to support parallel execution of mining algorithms, make up disadvantages of old data mining techniques, and find out useful data of information from massive examination data resources [7].

Cloud computing is an emerging technology for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or

Published by: The Mattingley Publishing Co., Inc.

service provider interaction [8]. It is composed of three service models: Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). The Cloud SaaS provides the users with applications that they can run and get results. The Cloud PaaS provides the users with the possibility to deploy applications onto the cloud. The Cloud IaaS provides the users with the capability to provision processing, storage, and networks for running their applications.

The Cloud computing is promising for high performance computing of many scientific and engineering application. Cloud computing systems implement a computing model in which virtualized resources dynamically scalable are provided to users and developers as a service over the Internet. In fact, clouds implement scalable computing and storage delivery platforms that can be adapted to the needs of different classes of people and organizations by exploiting the Service Oriented Approach (SOA). The advent of clouds offered large facilities to many users that were unable to own their high-performance computing systems to run applications and services. In particular, big data analysis applications requiring access and manipulate very large datasets with complex mining algorithms will significantly benefit from the use of cloud platforms.

II. LITERATURE SURVEY

Rayan Dasoriya et. al. [9] presents a review of big data analytics over cloud. The existing techniques are insufficient to analyze the Big Data and identify the frequent services accessed by the cloud users. Results can be analyzed in a better way by visuals like graphs, charts etc. and it helps in faster decision making. It also includes MapReduce Algorithm which will help in maintaining a log of user's activities in the cloud and show the frequently used services. This paper proposes a scheme for making Big Data Analytics more accurate, efficient and beneficial.

Jianguo Chen et. al. [10] presents a Parallel Random Forest (PRF) algorithm for big data on the Apache Spark platform. The PRF algorithm is optimized based on a hybrid approach combining data-parallel and task-parallel optimization. From the perspective of data-parallel optimization, a vertical data-partitioning method is performed to reduce the data communication cost effectively. From the perspective of task-parallel optimization, a dual parallel approach is carried out in the training process of RF, and a task Directed Acyclic Graph (DAG) is created according to the parallel training process of PRF and the dependence of the Resilient Distributed Datasets (RDD) objects. Extensive experimental results indicate the superiority and notable advantages of the PRF algorithm over the relevant algorithms implemented by Spark MLlib and other studies in terms of the classification accuracy, performance, and scalability.

Published by: The Mattingley Publishing Co., Inc.

Yong Wang et. al. [11] introduces a short review of Cloud Computing and Big Data, and discussed the portability of general data mining algorithms to Cloud Computing platform. It revealed the Cloud Computing platform based on Map-Reduce cannot solve all the Big Data and data mining problems. Transplanting the general data mining algorithms to the real-time Cloud Computing platform will be one of the research focuses in Cloud Computing and Big Data. Chun-Chieh Chen et. al. [12] presents a new approach, the cloud-based SpiderMine (c-SpiderMine), which exploits cloud computing to process the mining of large patterns on big graph data. Although cloud computing is effective at solving traditional algorithm problems, mining frequent patterns of a massive graph with cloud computing still faces the three challenges: 1) the graph partition problem, 2) asymmetry of information, and 3) pattern-preservation merging. The proposed method addresses the above issues for implementing a big graph data mining algorithm in the cloud. We conduct the experiments with three real data sets, and the experimental results demonstrate that c-SpiderMine can significantly reduce execution time with high scalability in dealing with big data in the cloud.

Tao Chen et. al. [13] presents a cloud-based data mining platform which demonstrates the solution of data mining as a service (DMaaS). In the backend, the data processing engine is based on hadoop, an open-source implementation of Google MapReduce. The user can access DMaaS from his browser for analyzing general purpose data mining problems. In this paper, we give an overview of DMaaS, present the system architecture and implementation techniques, and elaborate on a demonstration scenario. Yang Lai et. al. [14] describes a data mining framework on Hadoop using the Java Persistence API (JPA) and MySQL (My Structured Query Language) Cluster. Hadoop, a cloud computing project using the MapReduce framework in Java, has become of significant interest in distributed data mining. The framework is elaborated in the implementation of a decision tree algorithm on Hadoop. The results show the algorithm is more efficient than naïve MapFile indexing. We compare the Java Database Connectivity (JDBC) and JPA (Java Persistence API) implementations of the data mining framework. The performance shows the framework is efficient for data mining on

Hadoop. Yu Hua Zhang et. al. [15] discussed the intelligent cloud computing system (ICCS) is presented and discussed. Some tactics of the performance optimization of the CCS can be found and the management problems such as resource allocation policies, infrastructure development plan, and capabilities management and so on can be quantified through analyzing all this data. Based on these, building a cloud computing management information system (CCMIS) which is different from the general

Published by: The Mattingley Publishing Co., Inc.

November/December 2020
ISSN:0193-4120 Page No. 129 – 138

management information systems (MIS) is the major task in this paper.

III. DESIGN OF CLOUD COMPUTING BASED MIXED BIG DATA MINING

Design and implementation of Cloud Computing based mixed big Data Mining architecture is represented in below Fig. 1. The structure is divided into supporting platform layer, functional layer and cloud computing service layer, as shown in fig. 1.

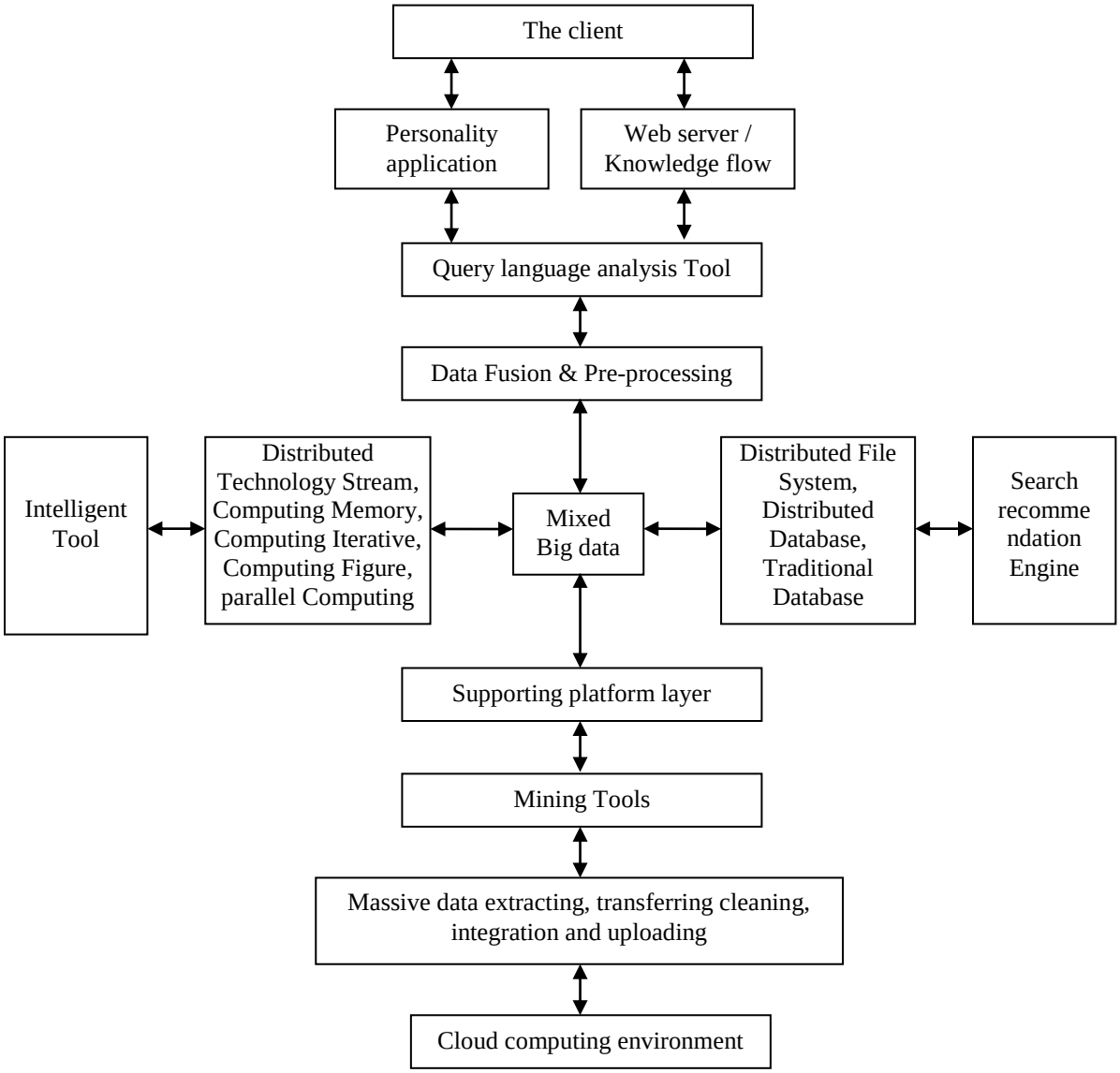

Fig. 1: ARCHITECTURE OF CLOUD COMPUTING BASED MIXED BIG DATA MINING

Main goal of huge amounts of data mining based on cloud computing service is to use cloud computing ability of parallel processing and mass storage to solve the problem of huge amounts of data processing in front of data mining. The client cognizes

Published by: The Mattingley Publishing Co., Inc.

November/December 2020
ISSN:0193-4120 Page No. 129 – 138

system and receives service by interaction. Supporting Platform Layer As the resource

and power support of big data mining, the platform will build a cloud environment with strong and abundant resources by combining a mixed big data with many kinds of support processing technologies based on cloud computing. This cloud environment can not only provide resources such as data, hardware, software to the world, but also calculate move data.

Functional Layer can automatically analyze and excavate according to the user's demand and preference. The analysis, mining and other tools rely on the high efficient ability of storage and computing, which high scalability and expandability have presented to users in the form of visualization, data source and other technologies.

The cloud computing service layer is located in the bottom, which provides a distributed parallel data processing and massive data storage. In the cloud computing environment, huge amounts of data storage need to consider the high availability of data and to ensure the safety. Cloud computing uses distributed data storage to store multiple copies of redundant storage to ensure that when the data has problems, it does not affect the normal use of the user.

Cloud computing environment or layer is mainly used to provide distributed file storage, database storage and computing power. The architecture can be based on the enterprise independent research and development of cloud computing platform, can also be based on cloud computing platform provided by the third party. Data mining ability layer mainly provides the foundation of mining capacity, contains algorithm service management, scheduling, data parallel processing framework, and provides the ability to support the data mining of cloud services layer. Cloud services layer mainly provides data mining cloud services. Service ability to encapsulate interface can be varied in form, including SOAP based Web service, Restful, Hypertext Transfer Protocol (HTTP), XML (Extensible Markup Language), or local Application Programming Interface (API) and other forms.

In order to specifically analyze the big data mining, this article shows each part of processing by constructing a multifunction Hadoop big data mining platform. Based on Hadoop platform, the big data mining is divided into 3 layers, which are the data source, the big data mining platform and the user layer. The data source is a complex processing object, which is composed of structure, semistructure and unstructured data; The big data mining platform combines various training-calculating modes, analyses, mining and other functions based on Hadoop to deal with the characteristics of real-time data; The user layer cognizes system and receives service by interaction.

Published by: The Mattingley Publishing Co., Inc.

November/December 2020
ISSN:0193-4120 Page No. 129 – 138

IV. RESULTS ANALYSIS

The Cloud environment used for the experimental evaluation was composed by 128 virtual servers, each one equipped with one dual-core 1.66 GHz CPU, 3.5 GB of memory, and 135 GB of disk space. To summarize, we present in Table 1 the speedups achieved by executing some data analysis applications with cloud computing based data mining.

Table 1: SPEEDUP VALUES OF SOME CLOUD COMPUTING BASED DATA MINING

Application	Tasks	Server	Speedup
Parallel clustering	17	17	9.2
Association rules analysis	56	16	15.3
Trajectory mining	135	64	50.4
Ensemble learning	118	20	15.7

The table reports the speedups achieved by the following applications: i) Parallel clustering, where multiple instances of a clustering algorithm are executed concurrently on a large census dataset to find the most accurate data grouping; ii) Association rule analysis, which is a workflow for association rule analysis between genome variations and clinical conditions of a group of patients; iii) Trajectory mining, a data analysis workflow for discovering patterns and rules from trajectory data of vehicles in a wide urban scenario. The best speedup is achieved in applications where many tasks can be run in parallel, and the concurrent tasks are homogeneous in terms of execution times. This is the case of the association rule analysis, where, after a short sequential task, several data mining tasks of similar duration are executed in parallel.

The scalability achieved using cloud computing based data mining can be further evaluated through Fig. 2, which illustrates the relative speedup obtained by using up to 128 servers. For the 5 GB dataset, the speedup passes from 6.8 using 8 servers to 84.7 using 128 servers. For the 10 GB dataset, the speedup ranges from 7.4 to 91.9. Finally, with the 20 GB dataset, we obtained a speedup ranging from 7.5 to 95.7. This is a very positive result, taking into account that some sequential parts of the implemented application (namely, partitioning and voting) cannot run in parallel.

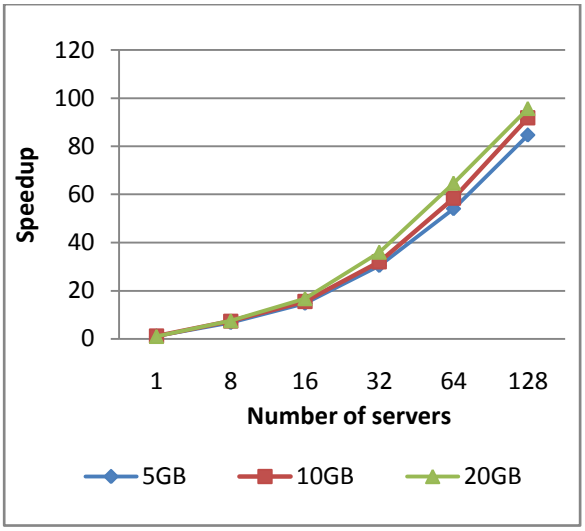

Fig. 2: SPEEDUP VERSUS SERVERS

An example in which task heterogeneity limits scalability is the parallel clustering application, where the tasks in charge of grouping data into a high number of clusters

Published by: The Mattingley Publishing Co., Inc.

November/December 2020
ISSN:0193-4120 Page No. 129 – 138

are much slower than those looking for a lower number of clusters. Therefore, in this case the speedup does not increase linearly with the number of servers used, because the turnaround time is bound to the execution time of the slowest task instances.

In big data mining platform, the core of Hadoop is HDFS, HBase and MapReduc;e. It has high reliability, high extensibility, high fault tolerance and high efficiency. Its calculation modes are composed of batch processing and stream processing. MapReduc;e is suitable for batch processing of static data with huge volume and low update rate, while Flume, Pig has scalability for the processing of dynamic data stream. However, Hadoop is not suitable for small amounts of low latency data and graph data with complex relation, which is also hard to support memory computing. Therefore, during building this system, traditional database, processing tools and memory calculation Spark are integrated into Hadoop platform.

In this way, traditional data structure can improve its rate of query processing by distributed storage and computing technology, while semi-structured and unstructured data can be processed quickly in real-time by memory computing and graphic computing. It has been confirmed in the theory and practice. Distributed computing contains two aspects of content that are distributed storage and parallel computing. And the cloud platform provides a distributed file storage and parallel computing ability, so it's a good solution to the content on both the two levels. The experimental results demonstrate the good scalability achieved using cloud computing based data mining to execute different types of data analysis workflows on a Cloud platform.

V. CONCLUSION

The emergence of cloud computing makes data mining platform have new developing direction, makes a new generation of data mining platform possible. Design and implementation of Cloud Computing based mixed big Data Mining architecture is described in this paper. As for the massive, complex uncertain and dynamic data, traditional data processing methods are facing serious challenges with the capacity of computing and storage. Its extensibility, flexibility and other abilities cannot meet the requirements of big data real-time processing. On the contrary, cloud computing provides a powerful impetus of computing and storage for big data processing. Cloud computing platform can be used to develop high-performance applications. We evaluated the performance of cloud computing based data mining through the speedup and number of server's workflow-based data analysis applications. Experimental results demonstrated the effectiveness of the framework, as well as the scalability that can be achieved through the execution of data analysis applications on the Cloud.

Published by: The Mattingley Publishing Co., Inc.

V. REFERENCES

[1] Ahmed El-mekkawi, Xavier Hesselbach, Jose Ramon Piney, "Evaluating the impact of delay constraints in network services for intelligent networks slicing based on SKM model", Journal of Communications and networks, Volume: 23, Issue: 4, Year: 2021

[2] Heru Nurwarsito, Verio Brika Sejahtera, "Implementation of Dynamic web server Based on Operating System-Level Virtualization using Docker Stack", 2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE), Year: 2020

[3] Lin Lin, "Research and Analysis on Key Technologies of cloud computing platform Based on IPv6", 2020 IEEE International Conference on Power, Intelligent computing and Systems (ICPICS), Year: 2020

[4] Mingda Li;Hongzhi Wang;Jianzhong Li, "Mining conditional functional dependency rules on big data", Big data mining and Analytics, Volume: 3, Issue: 1, Year: 2020

[5] Harsh Jain, Misha Kakkar, "Job Recommendation System based on machine learning and data mining Techniques using RESTful API and Android IDE", 2019 9th International Conference on Cloud Computing, data Science & Engineering (Confluence), Year: 2019

[6] Ján Cigánek, "Design and Implementation of Open-data-data- Warehouse", 2019 6th International Conference on Advanced Control Circuits and Systems (ACCS) & 2019 5th International Conference on New Paradigms in Electronics & information Technology (PEIT), Year: 2019

[7] Zhirui Huang, Pengfei Liu, Xiaxu He, Yanhua Chen, Weifeng Zhang, "Mobile data mining System based-on cloud Computing", 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP), Year: 2018

[8] Cody Balos, David De La Vega, Zachariah Abuelhaj, Chadi Kari, David Mueller, Vivek K. Pallipuram, "A2Cloud: An Analytical Model for Application-to- cloud Matching to Empower Scientific computing", 2018 IEEE 11th International Conference on cloud computing, Year: 2018

[9] Rayan Dasoriya, "A review of big data analytics over cloud", 2017 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Year: 2017

[10] Jianguo Chen, Kenli Li, Zhuo Tang, Kashif Bilal, Shui Yu, Chuliang Weng, Keqin Li, "A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment", IEEE Transactions on Parallel and Distributed Systems, Volume: 28, Issue: 4, Year: 2017

[11] Yong Wang, Ya-Wei Zhao, "Transplantation of Data Mining Algorithms to Cloud Computing Platform When Dealing Big Data", 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Year: 2014

[12] Chun-Chieh Chen, Kuan-Wei Lee, Chih-Chieh Chang, De-Nian Yang, Ming-Syan Chen, "Efficient large graph pattern mining for big data in the cloud", 2013 IEEE International Conference on Big Data, Year: 2013

November/December 2020
ISSN:0193-4120 Page No. 129 – 138

[13] Tao Chen, Jidong Chen, Baoyao Zhou, "A System for Parallel Data Mining Service on Cloud", 2012 Second International Conference on Cloud and Green Computing, Year: 2012

[14] Yang Lai, and Shi Zhongzhi, "An Efficient Data Mining Framework on Hadoop using Java Persistence API", IEEE 10th International Conference on Computer and Information Technology (CIT), Bradford, UK, Pages 203-209, 2010

[15] Yu Hua Zhang, Jian Zhang, Wei Hua Zhang, "Discussion of Intelligent Cloud Computing System", 2010 International Conference on Web Information Systems and Mining, Volume: 2, Year: 2010

Published by: The Mattingley Publishing Co., Inc.