Data Mining in Cloud Computing

Xia Geng[1,a],Zhi Yang[2,b]

[1]School of Computer Science and Telecommunication Engineering , Jiangsu University, Jiangsu Zhenjiang, P.R. China

[2]School of Management, Jiangsu University, Jiangsu Zhenjiang, P.R. China

[a]gengxia@ujs.edu.cn,[b]yangzhi@ujs.edu.cn

Keywords: Data Mining, Cloud Computing, Map-Reduce, Hadoop

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 based on parallel programming model and mass data mining service based on the cloud computing.

Introduction

A) Cloud computing

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, OpenNebula, 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[1]. 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 cloud is a virtualization of resources that maintains and manages itself. There are of course people resources to keep hardware, operation systems and networking in proper order. But from the perspective of a user or application developer only the cloud is referenced.

Cloud computing really is accessing resources and services needed to perform functions with

dynamically changing needs. An application or service developer requests access from the cloud rather than a specific endpoint or named resource.

B) 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, we'll discuss it in more detail.

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[2]: 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[3]: 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[4]: 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[5]:CUDA is a GPU-based parallel computing model proposed by NVIDIA . Since the

design requirements of GPU is different to general CPU , so GPU is usually designed to be slower perform multiple concurrency threads , rather than faster execute continuous threads , and GPU has inherent advantages on the parallel computing . CUDA GPU computing provide users with a variety of interfaces to enable programmers to program GPU like ordinary computer CPU.

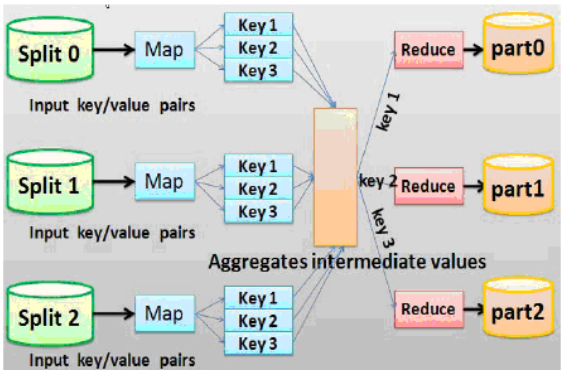

Fig. 1. Map-Reduce process architecture

Map-Reduce[6]:Map-Reduce model is a parallel programming framework proposed by Google., it provide users with a distributed file system firstly, which allow users to easily handle large-scale data; then all the procedures for computing are abstracted into two basic operations of Map and Reduce. in Map stage, data will be decomposed into smaller scale, and executed on different nodes of the cluster, and the results are integrated summary in the Reduce phase. Map-Reduce model is a simple but very effective parallel programming model. Figure 1 shows the overall flow of a Map-Reduce operation in Google's implementation.

As mentioned in the above intruduction, MapReduce is used most widely. And Apache Hadoop[7] is an open-source MapReduce style distributed data processing framework, which is an implementation similar to the Google MapReduce. Apache Hadoop MapReduce uses the HDFS[8] distributed parallel file system for data storage, which stores the data across the local disks of the computing nodes while presenting a single file system view through the HDFS API. HDFS is targeted for deployment on unreliable commodity clusters and achieves reliability through the replication of file data. When executing Map Reduce programs, Hadoop optimizes data communication by scheduling computations near the data by using the data locality information provided by the HDFS file system. Figure 2 shows the comparison between Hadoop and Google core technology.

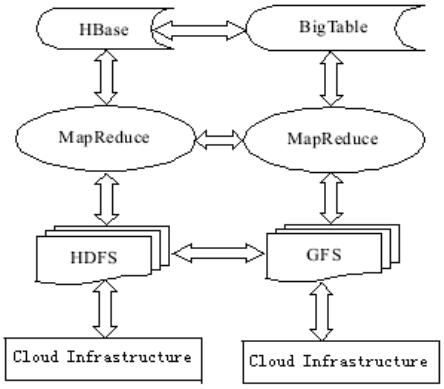

Fig. 2. Comparison of core technology between Hadoop and Google

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.

Data mining algorithm based on parallel programming model

In order to achieve the data mining on mass data, a large number of distributed and parallel data mining algorithms have been proposed. Bhaduri et al[9] put a very detailed parallel data mining algorithms bibliography, which not only include four major categories of distributed data mining algorithms in association rule learning, classification, clustering, streaming data mining, but also include related research works such as distributed systems, and privacy protection.

Map-Reduce parallel programming model has powerful ability to handle large-scale data, and

thus is an ideal programming platform for mass data mining. Data mining algorithms often need to traverse the training data to obtain the relevant statistical information for solving or optimizing the parameters of model. But frequent access on large-scale data requires a lot of compute time.

In order to improve the efficiency of the algorithm, Chu et al[10] propose a general parallel programming method for traditional machine learning algorithms. Analysis of the classical machine learning algorithms, they find the process of learning algorithm can be transformed to a number of the summation operation on the training data set, and the summation operation can be preformed independently on the subset of data. So it is easy to achieve parallel execution on Map-Reduce platform.

Firstly big data set is divided into a number of a subset and those subsets are assigned to corresponding Mapper nodes, then the Mapper node perform a variety of summation operation to emerge the intermediate results, Reduce node will sum the results finally, so the learning algorithm is executed parallelly. Under this framework, they implement ten classic data mining algorithms, including Locally Weighted Linear Regression (LWLR), Naive Bayes (NB), Gaussian Discriminative Analysis (GDA), k-means, Logistic Regression (LR), Neural Network (NN), Principal Components Analysis (PCA), Independent Component Analysis (ICA), Expectation Maximization (EM), Support Vector Machine (SVM).

Ranger et al[11] proposed a application programming interface called Phoenix which based on Map-Reduce and support parallel programming under the environment of multicore or multiprocessor systems. Phoenix can perform cache management, error recovery and concurrent management. They realized K-Means, PCA(principal component analysis) and linear regression by Phoenix.

Mahout[12] is a new open-source project which is developed by Apache Software Foundation (ASF).It is based on Hadoop liberary and its goal is to build scalable machine learning libraries. With scalable mean:

Scalable to reasonably large data sets. Mahout's core algorithms for clustering, classification and batch based collaborative filtering are implemented on top of Apache Hadoop using the Map-Reduce paradigm. However it is not restrict contributions to Hadoop based implementations: Contributions that run on a single node or on a non-Hadoop cluster are welcome as well. The core libraries are highly optimized to allow for good performance also for non-distributed algorithms.

Scalable to support user's business case. Mahout is distributed under a commercially friendly Apache Software license.

Scalable community. The goal of Mahout is to build a vibrant, responsive, diverse community to facilitate discussions not only on the project itself but also on potential use cases. Come to the mailing lists to find out more.

Currently Mahout supports mainly four use cases:

Recommendation mining takes users' behavior and from that tries to find items users might like.

Clustering takes e.g. text documents and groups them into groups of topically related documents.

Classification learns from existing categorized documents what documents of a specific category look like and is able to assign unlabelled documents to the (hopefully) correct category.

Frequent itemset mining takes a set of item groups (terms in a query session, shopping cart content) and identifies, which individual items usually appear together.

Data mining service Based on cloud computing

Cloud computing not only provide users with a common parallel programming model and big data processing capacity, but also provide users with an open computing services platform.

nowadays, a series of cloud computing service platforms have been developed to provide data mining services for the public.

Talia et al[13] summarize four levels of data mining services in cloud computing .(see Figure 3)

Single KDD steps: the underlying composition data mining algorithms.

Single data mining tasks: a separate data mining services, such as classification, clustering, etc.

Distributed data mining patterns: distributed data mining models, such as parallel classification, aggregation, and machine learning.

Data mining applications or KDD processes: complete data mining application based on the elements of all above.

On the basis of this design, they designed a Data Mining open service framework based on cloud computing, and developed a series of data mining services, such as Weka4WS, Knowledge Grid, Mobile Data Mining Services etc.

A) Weka4WS

Weka[14] is a widely used open source data mining toolkit that runs on a single machine. Weka4WS[15] extends the Weka toolkit by implementing a distributed framework that supports data mining in WSRFenabled Grids. Weka4WS integrates Weka and the WSRF technology for running remote data mining algorithms and managing distributed computations as workflows. The Weka4WS user interface supports the execution of both local and remote data mining tasks. On a Grid computing node, a WSRF-compliant Web service is used to expose all the data mining algorithms provided by the Weka library.

B) BC - PDM

China Mobile Institute begin cloud computing research and development from 2007 , it is the one of the earliest enterprises in cloud computing research and practice. In 2009, it officially announced his developing and testing cloud computing platform "BigCloud". Including the parallel data mining tools (BC-PDM).

BC-PDM[16] is a set of mass data processing analysis and mining system, it has high performance low cost high reliability high scalability characteristics .This system provides the mass data parallel ETL and parallel mining algorithm, supports enterprise BI application and accurate marketing; Provides business logic complex SQL ability, supports mass data cleaning conversion associated summary and operation, supports generation enterprise statements such as mining applications. Provides the SaaS service mode based on Web, and reduce the IT system investment of enterprise.

BC-PDM is a SaaS tools, and is based on the MapReduce implementation of cloud computing. Users can use the data from big cloud by BC-PDM only need to register rather then to buy or deployment, Because it is based on cloud computing, so BC-PDM overcome the traditional tools, and can deal with TB level mass data mining.

C) PDMiner

PDMiner[17] is a b parallel distributed data mining platform ased on Hadoop, which developed by the Institute of Computing Technology, PDMiner provide the vast majority of a series of parallel mining algorithms and ETL operations components, development of ETL algorithm to achieve a linear speedup, meanwhile has good fault tolerance. PDMiner has open architecture that allows the user to pack and loaded algorithm components into the

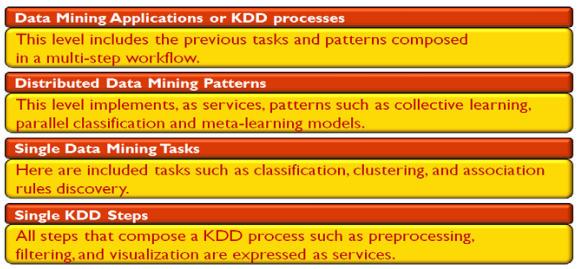

Fig. 3. Four levels of data mining services

system through a simple configuration.

The system can provide overall data mining solution for business decisions and intelligent information processing The system provides a variety of parallel data conversion rules and parallel data mining algorithms, the full support of the production, sales, marketing, financial management, corporate decision-making activities in the field, has broad application prospects.

In addition, major companies in the field of Business Intelligence provides business-oriented large-scale data mining services, such as micro-strategy, IBM, Oracle and other companies own the data mining services based on cloud computing platform.

Summary

Through big data storage and distribution of computing in cloud computing. we find a new ways to effectively solve the distributed storage of massive data mining and efficient computing. 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] Information on http://en.wikipedia.org/wiki/Cloud_computing

[2] IEEE standard for information technology - portable operating system interface (POSIX) - part 1: System Application program interface (API) - amendment 2: Threads Extension, 1995.

[3] W. Gropp, E. Lusk, A. Skjellum, Using MPI: Portable Parallel Programming with the Message-Passing Interface, seconde ed., the MIT Press, 1999.

[4] G. Malewicz, M. H. Austern, A. J. C. Bik, J. C. Dehnert, I. Horn, N. Leiser, Pregel: a system for large-scale graph processing, Proceedings of the 2010 international conference on Management of data. (2010)135--146.

[5] Information on http://www.nvidia.com/object/cuda_home_new.html

[6] J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters, Commun. ACM. 51 (2008) 107-113,

[7] Information on http://hadoop.apache.org/

[8] Information on http://hadoop.apache.org/docs/r1.0.4/hdfs_design.html

[9] K. Bhaduri, K. Das, K. Liu, H. Kargupta, and J. Ryan, Distributed Data Mining Bibliography, Distributed Data Mining Bibliography. 2011.

[10] C. T. Chu, S. K. Kim, Y. A. Lin, Y. Y. Yu, G. Bradski, A. Y. Ng, and K. Olukotun, Map-reduce for machine learning on multicore, Advances in neural information processing systems. 19 (2007) 281-287

[11] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, Evaluating mapreduce for multi-core and multiprocessor systems, IEEE 13th International Symposium on High Performance Computer Architecture. (2007) 13--24.

[12] Information on http://mahout.apache.org/

[13] D. Talia and P. Trunfio, How distributed data mining tasks can thrive as knowledge services Communications of the ACM. 53(2010) 132-137

[14] Information on http://researchcommons.waikato.ac.nz/handle/10289/1040

[15] D. Talia, P. Trunfio, O. Verta, The Weka4WS framework for distributed data mining in service-Oriented Grids, Concurrency and Computation: Practice and Experience. 20(2008) 1933-1951

[16] L. Yu, J. Zheng, W. C. Shen, B. Wu, B. Wang, L. Qian, B. R. Zhang, BC-PDM: data mining, social network analysis and text mining system based on cloud computing, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. (2012) 1496-1499.

[17] Information on http://www.chniot.cn/news/JSQY/2010/526/1052617494366_3.html