A study in the technology implementation of the network massive Information Processing Platform

Yue LI
Information Communication Company
The Power Supply Company of He-nan Xin-yang
Henan, Xingyang, 464000, China
e-mail: 55609170@QQ.com

Ran LIU
Department of Foreign Language
The XinYang Normal University
Henan, Xinyang, 464000, China
e-mail: 305404571@QQ.com

***Abstract*—With the popularity and development of the network, the support of the high-performance computer technology becomes increasingly important as the huge information storage and the convenience of Information retrieval function of the internet that attracts more and more people join the netizens' team. Therefore, I proposed an Information Processing Platform based on the high performance data mining in order to improve the Internet mass information intelligence parallel processing functions and the integrated development of the system's information storage, management, integration, intelligence processing, data mining and utilization. The propose of this system is to provide certain references and guidance for the technology implementation and realization of the high performance and high efficiency network massive Information Processing Platform as on the one hand, I have analyzed the key technology of the implementation of the platform, on the other hand briefly introduced the implementation of the RDIDC.**

Keywords- the internet; Massive Information; information processing; technology implementation

I. FOREWORD

In this era with the information expansion,people are producing,spreading,retrievalling and applying any kind of information.Especially in recent 20 years，the total amount of the information produced by the whole society left the total amount of the information since human came into being far behind..From the aspect of internet medias,as the report by China Internet Network Information Center in 2009 shows,the webpages amount of our country grew 90% over previous year and reach the number of 16 billion.Among them,the bytes amount has passed 460TB. With the continuous improvement of the current acceleration of the speed of social information and network technology level，this number shows a ever-accelerating trend. In this context， the implementation of the network massive Information Processing Platform has become a core problem for network wokers and technicians.In fact, before we set about to study this problem systematically, it's necessary for us to summarize the information charateristics of network Medias briefly.

.

II. CHARACTERISTICS OVERVIEW OF INFORMATION IN THE INTERNET SPACE

A. The amount of information is huge, the kindsare various and the speed is fast

Currently, the information scale and kinds are expanding sharply.By the end of July 2008, the number of index pages in Google website has passed one trillion and kept growing.

B. Information generation is fast, and easy to change

As we all know, the data information in cycberspace mainly based on the high-speed network system and computer hardware with the characteristics of interactive, instant and integrative.The form of it possesses obviously characteristic of instant, we can just use our mouse to finish copying and changing.The information in the internet is using the form of data to store, so it's easy to change.

C. Information collection is messy, and difficult to find

The capacity of the information in the Internet space is extremely large and wide variety of, so the information in the cycberspace now is short of effective and rational orgazination.The messy characteristic brings down the efficiency for people to retrieval, obtain and use the information.So it's also one of the core problems of implementation of the network massive Information Processing Platform.

D. Information in the cycberspace is mixed

Presently, the information medias in our country show the develop mode of dispersion and liberalization.The threshold to get into the network is quite low, especially lack for effective supervision.And it leads to the fact that the related departments can't supervise the release and spread of the information.For example, there are full of various kinds of similar and fake information, even some violent reactionary and pornographic information.All these stuff disturb netizen's using.

E. Network information has obvious effectiveness of public opinion

With the popularize and develop of the internet,especially the consummation of instant talks、BBS forums、twitter and other applications,people prefer to post there daily affairs,things they heard and saw and their own opinion on

Published by Atlantis Press, Paris, France.

current affairs in cycberspace.Therefore, network information has obvious effectiveness of public opinion.Especially in some emergency; the massive network information will throw huge pressure on the government department so the problems will be sloved publicly timely and clearly.

This shows that the special information characteristics bring tougher challenge to the implementation of the network massive Information Processing Platform in information storage and management, data mining, real-time processing, intelligent processing, audio and video data, web text, and such other problems.

III. THE BASIC ARCHITECTURE OF THE NETWORK MASS OF INFORMATION PROCESSING PLATFORM

The network massive Information Processing Platform designed by me is on the basis of high-performance computer technology.It contains unified view of middleware, distributed parallel database, parallel data mining services, high-speed parallel computing environment, the cluster interconnect and so on.So that it can form a well-functional information process system to analysis and dig various kinds of information.What's more,the system can improve the ability of cyberspace in overall information processing, information retrieval, public opinion analysis, and predicting trends.As the figure 1 shows,the overall architecture of the platform, including data acquisition, storage and organization, business analysis layer, data integration layer, and user interface layer several.

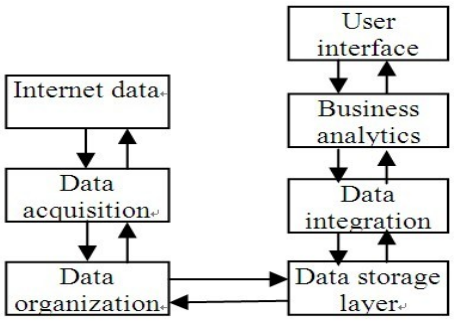

Figure 1. The basic architecture of network vast amounts of information processing platform.

In this system, data mining module is the core part of the whole mass of information processing platform,it provide a comprehensive analysis of the network mass information processing, retrieval and effective application platform including hotspot,tracing, statistics, hook-linked analysis, behavioral analysis and mining business analysis mode on the basis of open mining algorithm library type.Specifically, Data acquisition section is a data gateway access platform via the Internet, into the specific organization of data in the different categories of the data information through a series of data cleaning and formats unified after processing, so as to ensure that the data information in a subsequent processthe effectiveness and robustness. The main function of the data organization is to process online activities of data stream,which contains identification of the information, text extraction, fast scanning, feature extraction, data filtering, file deduplication, information classification, and many other functions role.Among them, Feature extraction mainly formed file feature data, text extraction mainly formed of text data information is derived.The data storage parallel and process three particle size and improve data processing capability of the system through a distributed and parallel database.On the basis of distributed middleware,it can achieve parallelism between the plurality of parallel data processing activities, but also in the parallel database internal parallel processing of multiple threads on the plurality of nodes and nodes, etc. After forming a particular data organization data information storage apply a unified view of the middleware to store and manage data.Eventually, the unified view of middleware enable the upper application to direct access to the underlying distributed parallel database transparently within the system.What is more, ata integration layer extract raw data from the distributed parallel database information through a specific unified view middleware.And by theme-oriented data integration and storage information，it can improve the efficiency and performance of the entire system of data mining and data analysis.Specifically,data integration layer contains data cleaning and loading, as well as integrated data modeling as the main target in order to analyze the various functions.Thereinto, Integrated data modeling to analyze the main target is based on a variety of mining applications,through a subject-oriented multidimensional data model based on the data information integration.This model can organize massive data information from Multi-angle and multi-level,and supportdifferent granularity materialized view of the effect to implement Real-time query data information from the macro to the micro.Furthermore, it can ensureDifferent particle size and a full range of data mining and data analysis.

Apart from this ， the main function of the businessanalysis layer is to dig parallelled data.It containsspecific data mining operations and the open data Mining algorithm library two parts.Among them, specific data mining business collusion include the analysis of the hot spot analysis,tracing analysis,behavioral analysis,statistical analysis,behavior mining.The open data Mining algorithm library includes clustering,classification,association,text mining,sequential patterns and content.At last ， user interface layer for network information is to retrieval and users to provide special automated background tasks mining task wizard, custom tasks, excavation analysis visualization, user screening and evaluation, and many other technical services.

IV. THE KEY TECHNICAL ANALYSIS OF THE NETWORK MASS INFORMATION PROCESSING PLATFORM

As the table1 shows, to form a network mass of information processing platform based on the high-performance data mining, we need to achieve the following key technologies:

Published by Atlantis Press, Paris, France.

TABLE I. THE CATEGORY AND LIST OF KEY TECHNOLOGIES OF MASS INFORMATION PROCESSING PLATFORM

Field	***Platform***	***Key Technology***
Storage and management of few internet information	Data storage layer	High-speed interconnect of parallel database based on InfiniBand RDS protocol
Real-time network data information processing activities	Data organization、Data storage layer、Data integration layer	Distributed parallel database、unified view of middleware
Internet data information mining	Business analytics layer	Parallel data mining based on database
Data index of internet media and text	Data integration layer、Business analytics layer	Web text data mining

A. High-speed interconnect based on InfiniBand RDS protocol parallel database

The system designed by me is to process the network mass information storage and management through distributed and parallel database. Each parallel database contains more than one database server node shared storage constitute a cluster database system.While each large distributed parallel database systems is formed by multiple parallel database.We can inplement the storage and management activities of hundreds of TB of huge amounts of data information in cycberspace.So distributed parallel database is also based on parallel database.From the aspect of technology implementation,the point that influences and even decides parallel database scalability with parallel processing efficiency or parallel processing performance is node interconnect bandwidth, delay, and processor overhead and so on…Nowadays, with the constant expansion of the computer database,and implementation of open interconnect technology such as gigabit Ethernet can not meet the current needs of the computer parallel database node interconnect.We must reconstruct large-scale nodes the computer parallel database system toi mprove information storage and management capabilities and efficiency of the entire platform

As we all know，InfiniBand is mainly defined by the InfiniBand as a Trade Association, an open, advanced interconnect standard.It is a channel-based, the use of the I/O system of the exchange structure. Reliable Datagram Sockets InfiniBand is the upper layer protocols, with characteristics of low-latency, low-load, high-bandwidth, the IB network provides reliable datagram service, in order to support the UDP protocol application.

The parallel database IB, the RDS network environment by the host channel adapter, the four parts of the IB switch, database applications support software, sub-network management. RDS improved malleability and performance of the application of parallel databases greatly.Compared to IPOIB,the CPU occupancy rate has dropped by about 50%.While compared to UDP protocols, the delay also reduces the half.The advantages of RDS over Gigabit Ethernet is in easy-to-use, low-latency and low processor utilization, high bandwidth and high availability, no the discarded or retransmission reliable packet transmission.Based on the InfiniBand RDS protocol, in the high-speed interconnect eight-node RAC parallel database experimental environment built by me，comparing Oracle RAC database IB RDS protocol and Gigabit Ethernet interconnect using TPC-H benchmark.In three typical TPC-H queries total running time, the latter increased by approximately 33% than the former. This shows obviously that， IB RDS protocol can greatly reduce the cluster database cluster latency and global cache coherence transfer time, its overall performance has been greatly enhanced.

B. Distributed and Parallel Database unified view of middleware

Through a unified view of middleware，the distributed parallel database system can synthesise the parallel database set into a relatively large number of parallel database set.So unified view of middleware plays an extremely important role in the whole system.Generally speaking, Unified view of middleware systems, distributed parallel database include client API, unified view of middleware services, the statistical backup module, system security, policy management services, database access, etc.In this database system, a unified view of the middleware can boost the upper applications transparent access to the underlying distributed parallel database to develop interface in response to the upper application for the network mass information processing platform.Besides,this application parallel query optimization, SQL parsing, multi-level load balancing and fault tolerance, and many other important technical,so that it can ensure the entire system reliability, availability, high-speed, concurrent and other properties greatly.

Parallel load and parallel query service are the core functionalities of the unified view of middleware.We can implement these funtionalities through data dictionary. Through the real practice, its overall performance to meet the network mass information is accurate and can meet the needs of the real-time processing..

C. Database-based parallel data mining

Data mining means that we can dig the valueable information that can meet the users' meet from the massive network information.And this is also one of the core technologies of the network massive Information Processing Platform.

Database-based parallel data mining underlying construct an open mining algorithms library，including clustering, classification, association, text mining, sequential patterns, anomaly identification, important attributes, feature extraction, etc.The upper-developed network information processing platform includes a variety of business analysis means hot spot analysis, statistical analysis, hook-linking analysis.Its advantage is that it can support for parallel

computing and parallel database mining, support for Windows / Linux platform, the performance of high-speed and concurrent better and other characteristics.

D. Web text data mining

Web text is the most popular way to present the network information; therefore, the data mining is a very basic function to the network of the mass information processing platform. It includes two aspects: text stored and managed and safety retrieval. Generally speaking，the former is based on the text system，organizing and storing the whole text by three layers model structure：logic store、physics store and user view. Among them, in physics store, the whole system supports multi-catalogue and multi-level store according to location、time and theme. It adopts multi-level text index techology, and uses all kinds of data structures to hierarchically store the location of the text physics store based on the text management that the computer system offers. Uesr view centers on the logic store and mavennages different text information via convenient and efficient text index techology.

As a whole, text index is a typical distributed text search tool, and its core function based on the web text information that the user needs includes getting web text information, dealing with text analysis and text collection, and wiping out repeated information; classifing the text to build up part index and form index storeroom finally; sending requests and index storerooms to other nodes.

V. THE END

At persent, the key techology of the network of the mass information processing platform, which is based on high performance data mining that the author designs, has gotten effective used. It has been found that the platform can deal with the intelligent parallel dispose of the information and analyse and dig out the characteristic information, and because of its high expansibility, it can meet the need of the internet information's rapid growth in some degree.

REFERENCES

[1] Fangyuan He. The basic research about the massive data processing technology [J]. Silicon Valley, 2009(8): 59-60(Ch).

[2] Lixin Qi. Content-oriented network to massive information processing platforms and systems research[J]. China's media technology, 2006(9): 68-72(Ch).

[3] Xiaoyu Li. The search about the using of Multi-Agent in network vast amounts of information[J]. Science and technology information, 2009(14): 24-27(Ch).

[4] Liangying Wang. The search about the technology of massive information resource storage and sharing[J]. Information systems engineering, 2011(11)：31-34(Ch).

[5] Yi Liao, Dong Liu. Asynchronous called to the SCADA system of massive information processing method， Automation of Electric Power Systems, 2006(20): 24-27.

Published by Atlantis Press, Paris, France.