Volume 3, Issue 2 (2015) 359-361 ISSN 2347 - 3258
International Journal of Advance Research and Innovation

Bootstrap Based Large Scale Data Processing Using Cluster

M. Premalatha, G. Baskaran
Department of Computer Science Engineering, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamil Nadu, India

Article Info
Article history:
Received 4 April 2015
Received in revised form
30 April 2015
Accepted 20 May 2015
Available online 15 June 2015

Keywords
Cluster,
BestPeer++,
Data Mining,
Cloud Computing,
BATON

Abstract

Cloud computing is model in which large groups of remote servers are networked to allow centralized data storage and online access to computer services or resources. Clouds can be classified as public, private. A corporate network is a group of computers, connected together in a building or in a particular area, which are all owned by the same. The corporate network is often used for sharing information among the participating companies and facilitating collaboration in a certain industry sector where companies share a common interest, there is so challenges and some security issues appeared .So in the proposed System it implement the Best Peer++, a system which delivers elastic data sharing services for corporate network applications in the cloud based on Best Peer—a peer-to-peer (P2P) based data management platform. A data management platform is the backbone of data-driven marketing, and serves as a unifying platform to collect, organize, and activate your first- and third-party audience data from any source, including online, offline, or mobile. A true Data Management Platform should have the ability to collect unstructured audience data from any source, including mobile web and app, web analytic tools, CRM, point of sale, social, online video, and other available offline data sources.

1. Introduction

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 behaviours, 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. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They clean databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?".

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. Data mining consists of five major elements. Those are extract, transform, and load transaction data onto the data warehouse system. Store and manage the data in a multidimensional database system.

Provide data access to business analysts and information technology professionals. Analyze the data by application software. Present the data in a useful format, such as a graph or table.

Companies of the same industry sector are often connected into a corporate network for collaboration purposes. Each company maintains its own site and selectively shares a portion of its business data with the others. Examples of such corporate networks include supply chain networks where organizations such as suppliers, manufacturers, and retailers collaborate with each other to achieve their very own business goals including planning production-line, making acquisition strategies and choosing marketing solutions. From a technical perspective, the key for the success of a corporate network is choosing the right data sharing platform, a system which enables the shared data (stored and maintained by different companies) network-wide visible and supports efficient analytical queries over those data.

Traditionally, data sharing is achieved by building a centralized data warehouse, which periodically extracts data from the internal production systems (e.g., ERP) of each company for subsequent querying. Unfortunately, such a warehousing solution has some deficiencies in real deployment in the real world, most of the data warehouse solutions fail to offer such flexibilities.

BestPeer data management platform, while traditional P2P network has not been designed for enterprise applications, the ultimate goal of BestPeer is to bring the state-of art database techniques into P2P systems. In its early stage, BestPeer employs unstructured network and information retrieval technique to match columns of different tables automatically. After defining the mapping functions, queries can be sent to different nodes for processing. Existing system are having different

Corresponding Author,
E-mail address: prema89latha@gmail.com

disadvantages, those are the basic approach is the inefficiency of query processing. The main problem of existing system is unstructured PDBMS. There is no guarantee for the data retrieval performance.

2. Related Work

Finally, to maximize the revenues, companies often dynamically adjust their business process and may change their business partners. Therefore, the participants may join and leave the corporate networks at will. The data warehouse solution has not been designed to handle such dynamicity. To address the aforementioned problems, this paper presents BestPeer++, a cloud enabled data sharing platform designed for corporate network applications. BestPeer++, a cloud enabled evolution of BestPeer. BestPeer++ is enhanced with distributed access control, multiple types of indexes, and pay-as-you-go query processing for delivering elastic data sharing services in the cloud.

The key idea of BestPeer++ is to use dedicated database servers to store data for each business and organize those database servers through P2P network for data sharing. The Amazon Cloud Adapter provides an elastic hardware infrastructure for BestPeer++ to operate on by using Amazon Cloud services. The infrastructure service that Amazon Cloud Adapter delivers includes launching/terminating dedicated MySQL database servers and monitoring/backup/auto-scaling those servers.

BestPeer++ achieves its query processing efficiency and is a promising approach for corporate network applications, with the following distinguished features. BestPeer++ employs P2P technology to retrieve data between business partners. BestPeer++ instances are organized as a structured P2P overlay network named BATON. BestPeer++ employs a hybrid design for achieving high performance query processing. The major workload of a corporate network is simple, low overhead queries. Such queries typically only involve querying a very small number of business partners and can be processed in short time.

To maximize the revenues, companies often dynamically adjust their business process and may change their business partners. Therefore, the participants may join and leave the corporate networks at will. In large scale data processing system, a set of queries designed for data sharing applications.

3. System Model

BestPeer introduces a series of techniques for improving query performance and result quality to enhance its suitability for corporate network applications. In particular, BestPeer provides efficient distributed search services with a balanced tree structured overlay network (BATON) and partial indexing scheme for reducing the index size.

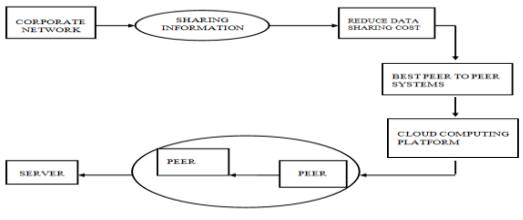

Figure: Architecture Diagram

Proposed systems are having different advantages. It can efficiently handle typical workloads in a corporate network. It can deliver near linear query throughput as the number of normal peers grows. In the BestPeer++ is a promising solution for efficient data sharing within corporate networks.

3.1 Normal Peer-To-Peer Systems

Peer-to-peer file sharing is the distribution and sharing of digital media using peer-to-peer (P2P) networking technology. P2P file sharing allows users to access files such as books, using a P2P software program that searches for other connected computers on a P2P network to locate the desired contents. The nodes (peers) of such networks are end-user computer systems that are interconnected via the Internet.

Two data flows inside the normal peer: an offline data flow and an online data flow. In the offline data flow, the data are extracted periodically by a data loader from the business production system to the normal peer instance. In particular, the data loader extracts the data from the business production system, transforms the data format from its local schema to the shared global schema of the corporate network according to the schema mapping, and finally stores the results in the databases hosted in the normal peer. In the online data flow, user queries are submitted to the normal peer and then processed by the query processor. The query processor performs user queries using a fetch and process strategy.

3.2 Data Loader

Data Loader is a component that extracts data from production systems to normal peer instances according to the result of schema mapping. While the process of extracting and transforming data is straightforward, the main challenge comes from maintaining consistency between raw data stored in the production systems and extracted data stored in the normal peer instance (and subsequently data indices created from these extracted data) while the raw data being updated inside the production systems. There is no adjacent node available for load balancing.

3.3 Access Control

BestPeer++ extends the role-based access control for the inherent distributed environment of corporate networks. Through a web console interface, companies can easily configure their access control policies and prevent undesired business partners to access their shared data.

The challenge is for BestPeer++ to provide a flexible and easy-to-use access control scheme for the whole system; at the same time, it should enable each business to decide the users that can access its shared data in the inherent distributed environment of corporate networks. BestPeer++ develops a distributed role-based access control scheme. The basic idea is to use roles as templates to capture common data access privileges and allow businesses to override these privileges to meet their specific needs.

3.4 Query Processing

In query processing, two query processing approaches, basic processing and adaptive processing. The basic query processing strategy is similar to the one adopted in the

distributed databases domain. Overall, the query submitted to a normal peer P is evaluated in two steps: fetching and processing. In the fetching step, the query is decomposed into a set of sub queries which are then sent to the remote normal peers that host the data involved in the query.

4. Best Peer to Peer Networks

BestPeer++ is deployed as a service in the cloud. To form a corporate network, companies simply register their sites with the BestPeer++ service provider, launch BestPeer++ instances in the cloud and finally export data to those instances for sharing. BestPeer++ adopts the pay-as-you-go business model popularized by cloud computing. The total cost of ownership is therefore substantially reduced since companies do not have to buy any hardware/software in advance. Instead, they pay for what they use in terms of BestPeer++ instance's hours and storage capacity.

5. Conclusion

The unique challenges posed by sharing and processing data in an inter-businesses environment and proposed BestPeer++, a system which delivers elastic data sharing services, by integrating cloud computing, database, and peer-to-peer technologies. The benchmark conducted on Amazon EC2 cloud platform shows that the system can efficiently handle typical workloads in a corporate network and can deliver near linear query throughput as the number of normal peers grows. Therefore, BestPeer++ is a promising solution for efficient data sharing within corporate networks. In the future data sharing Cloud storage is a model of data storage where the digital data is stored in logical pools, the physical storage spans multiple servers (and often locations), and the physical environment is typically owned and managed by a hosting company. These cloud storage providers are responsible for keeping the data available and accessible, and the physical environment protected and running. People and organizations buy or lease storage capacity from the providers to store user, organization, or application data. And also implementing the Virtual Machine Virtualization helps enterprises make more efficient use of hardware resources. It facilitates a greater degree of abstraction of the software environment from its hardware. Servers now exist as a single file. You can easily move them from one piece of hardware to another, duplicate them at will, and create a more scalable and flexible infrastructure. Cloud computing has taken that degree of efficiency and agility realized from virtualization and magnified it further.

References

[1] A. Abouzeid, K. Bajda-Pawlikowski, D. J. Abadi, A. Rasin, A. Silberschatz, Hadoop DB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads, VLDB Endowment, 2(1), 2009, 922-933.

[2] K. Aberer, A. Datta, M. Hauswirth, Route Maintenance Overheads in DHT Overlays, 6th Workshop Distrib. Data Struct., 2004

[3] B. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, R. Sears, Benchmarking Cloud Serving Systems with YCSB

[4] Google Inc, Cloud Computing-What is its Potential Value for Your Company? White Paper, 2010

[5] R. Huebsch, J. M. Hellerstein, N. Lanham, B. T. Loo, S. Shenker, I. Stoica, Querying the Internet with PIER, Proc. 29th Int'l Conf. Very Large Data Bases, 2003 321-332

[6] H. V. Jagadish, B.C. Ooi, K.-L. Tan, Q.H. Vu, R. Zhang, Speeding up Search in Peer-to-Peer Networks with a Multi-Way Tree Structure, ACM SIGMOD Int'l Conf. Management of Data

[7] H. V. Jagadish, B. C. Ooi, K.-L. Tan, C. Yu, R. Zhang, iDistance: An Adaptive B+-Tree Based Indexing Method for Nearest Neighbor Search, ACM Trans. Database Systems, 30, 2005, 364-397

[8] H. V. Jagadish, B. C. Ooi, Q. H. Vu, BATON: A Balanced Tree Structure for Peer-to-Peer Networks, Proc. 31st Int'l Conf. Very Large Data Bases (VLDB '05), 2005, 661-672

[9] A. Lakshman, P. Malik, Cassandra: Structured Storage System on a P2P Network, 28th ACM Symp. Principles of Distributed Computing (PODC '09), 2009, 5

[10] W. S. Ng, B. C. Ooi, K.-L. Tan, A. Zhou, Peer DB: A P2P-Based System for Distributed Data Sharing, Proc. 19th Int'l Conf. Data Eng., 2003, 633-644.

[11] I. Tatarinov, Z. G. Ives, J. Madhavan, A. Y. Halevy, D. Suciu, N. N. Dalvi, X. Dong, Y. Kadiyska, G. Miklau, P. Mork, The PiazzaPeer Data Management Project, SIGMOD Record, 32(3), 2003, 47-52

[12] S. Wu, J. Li, B.C. Ooi, K.-L.Tan, Just-in-Time Query Retrieval over Partially Indexed Data on Structured P2P Overlays, ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '08), 2008, 279-290

[13] S. Wu, S. Jiang, B.C. Ooi, K.-L.Tan, Distributed Online Aggregation," Proc. VLDB Endowment, 2(1), 2009, 443-454

[14] A. Thusoo, J. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, R. Murthy, HIVE: A Warehousing Solution over a Map-Reduce Framework, VLDB Endowment, 2(2), 2009, 1626-1629

[15] H. T. Vo, C. Chen, B. C. Ooi, Towards Elastic Transactional Cloud Storage with Range Query Support, VLDB Endowment, 3(1), 2010, 506-517