Cloud Based Advanced Bestpeer for Data Sharing in Corporate Network

Divya B[1] Vatsala B R[2]
[2]Assistant Professor
[1]Department of PGSCEA [2]Department of Computer Science & Engineering
[1,2]The National Institute of Engineering, Mysore

Abstract— Many computer networks and business organisation of the same industry sector share some of theselective data and collaborate with other companies who share a common interest. This data sharing helps to achieve their business goals and to reduce the operational cost which in turn increases the revenue of the corporate network. There are many unique challenges faced by inter-company data sharing and data processing processes such as scalability, performance and security. In this paper we present a cloud based system, Advanced BestPeer which delivers elastic data sharing services with better performance for corporate network and business organisation. Advanced BestPeer system is developed by combining cloud computing, peer to peer technology and database to provide economical, flexible, scalable platform and also based on pay-as-you-go business model. Advanced BestPeer is evaluated on Amazon EC2 cloud platform. The benchmarking result shows that Advanced BestPeer performs better than HadoopDB technology. It also demonstrates that Advanced BestPeer achieves near linear scalability for throughput with respect to the number of peer nodes.

Keywords: HadoopDB, BestPeer, Amazon EC2 cloud Platform.

I. INTRODUCTION

Cloud computing is the delivery of computing services over the Internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications. The services that business use is provided by Advanced BestPeer and managed on behalf of organisation. Cloud services are available on-demand and often bought on a "pay-as-you go" or subscription basis [4]. Cloud computing can significantly reduce the cost and complexity of owning and operating computers and networks. If an organization uses a cloud provider, it does not need to spend money on information technology infrastructure, or buy hardware or software licenses. Cloud services can often be customized and flexible to use, and providers can offer advanced services that an individual company might not have the money or expertise to develop.

Peer to peer technology is networking method of delivering computer network services in which the participants share a portion of their own resources, such as processing power, disk storage, network bandwidth etc. In this paper we are presenting hybrid decentralized peer architecture [1], it is a combination of peer-to-peer and client-server models. A common hybrid model is to have a central server that helps peers find each other. Currently, hybrid models have better performance than either pure unstructured networks or pure structured networks because certain functions, such as searching, do require a centralized functionality but benefit from the decentralized aggregation of nodes provided by unstructured networks.

In this paper we integrate both cloud computing and peer-to-peer technology along with the database to support data sharing in companies. Companies of the same industry are often connected to a corporate network for association purposes. Each company maintains its own public and private data in the site and selectively shares a part of its business data with the other companies include supply chain networks where organizations such as supplier, manufacturer, and retailer cooperate with each accomplish their own business goals such as planning production-line, making achievement strategies and choose marketing solutions. As per technical perspective, selecting the right data sharing platform is very important to support network-wide visibility of data and to support queries over those shared data

II. PROBLEM STATEMENT

Traditional data sharing is based on centralized datawarehouse which extracts the data from internal production system. This method has several deficiencies. First the corporate or the business network needs to scale up to support millions of participants without huge investments on hardware/software and also the cost of maintenance. These costs are also known as total cost of ownership and total cost of operations. But most of the companies are not ready to invest more until they see potential return on invest (ROI). Second, most of the companies want to customize the access control policy to determine which business partners can see which portion of the shared data. Finally, to maximize their revenue, and also to achieve dynamicity where the business partners change their business processes and also change their business partners. Therefore, the business partners may join and leave the corporate network. The data warehouse method is not designed to handle these problems.

III. EVOLUTION OF ADVANCED BESTPEER

This section gives the brief explanation of existing System and then overview of the propose System.

A. Existing System

The initial BestPeer system attempted to develop a peer-to-peer (P2P) technology for corporate network. BestPeer was designed to work as a scalable, sharable, and secure P2P-based Data Management system with full functionalities for building corporate networks in which a part of association controlled by different executive domains work together in order to reduce operation cost and pick up efficiency(3). The goal of BestPeer is to involve the database technique into P2P systems. Earlier BestPeer was unstructured network

andinformation retrieval was done automatically by matching the columns of different tables (15). Once the mapping functions are defined queries can be sent to different nodes for processing. BestPeer also introduces many techniques for improving query performance and also to enhance its suitability for corporate network applications.

B. Prosed System

The main contribution of this paper is the design of Advanced BestPeer system that well-structured, scalable and elastic solution for corporate or the business network. The unique challenges faced are the sharing of data and processing it in a business environment is resolved by Advanced BestPeer, a system which gives economical, flexible and elastic data sharing services for corporate applications. Advanced BestPeer is based on research on P2P database system, and offers an accelerate data processing engine and more portability via MapReduce framework (2) and Software-as-a-Service (SaaS) paradigm.

To form a corporate network, companies and the business organisationregisters their site with the Advanced BestPeer service provider, which initiates the Advanced BestPeer instances in the cloud and exports the company's data for sharing. Advanced BestPeer is based on pay-as-you-go model popularized by cloud computing (4). Since the companies do not have to purchase any hardware/software in advance, they reduce the total cost of ownership. In exchange they have to pay for Advanced BestPeer instance in terms instance hours and storage capacity. The Advanced BestPeer service provider scales up the instances and makes them available always. Advanced BestPeer employs a hybrid p2p design for achieving high performance query processing. Advanced BestPeer is mainly optimized for simple queries. Majority of the workload in a corporate network is simple, light-weight queries. Such queries involve querying only a small number of business partners and require a short time for processing. For rare time-consuming analytical tasks, an interface is provided for exporting the data from Advanced BestPeer to Hadoop (2)

And allows the user to analyse these data using MapReduce technology(2) .Advanced BestPeer also inherits the features of existing system such as efficient distributed query processing, support for semi-automatic schema mapping and data mapping (6), effective system load balancing and other functionalities that a corporate network requires.

IV. COMPONENT OF ADVANCED BESTPEER

Advanced BestPeer, a cloud enabled evolution of BestPeer is improved with distributed access control, pay-as-you-go query processing for deliver elastic data sharing services in the cloud and multiple types of indexes. The software components of Advanced BestPeer are separated into two parts: core and adapter. The Architecture is shown in fig. 1.

A. Core

The core contains all the data sharing functionalities and is designed to be platform independent.

B. Adapter

The adapter contains one abstract adapter which defines the elastic infrastructure service interface and a set of concrete adapter components which implement such an interface through APIs provided by specific cloud service providers (e.g., Amazon).

Core and Adapter forms the "two-level" design to achieve portability. With proper adapters, Advanced BestPeer can be ported to any cloud environments such as public and private clouds or even for non-cloud environment.

V. SYSTEM ARCHITECTURE

Advanced BestPeer core contains all the platform independent logic like query processing and peer-to-peer overlay. It runs on top of cloud adapter. Cloud adapter consists of software components bootstrap peer and normal peer. Advanced BestPeer network can have only one bootstrap peer instance launched and maintained by Advanced BestPeer service provider and many normal peer instances.

The architecture is depicted in Fig.1. The bootstrappeer instance isthe only entry point for the completenetwork. Bootstrap peer has many responsibilities.

- Bootstrap peer serves for various administration purposes such as monitoring and managing normal peers.
- Scheduling various network management events.
- It acts as central repository for metadata of corporate network application.
- It also stores shared global schema, participant normal peer list and role definitions

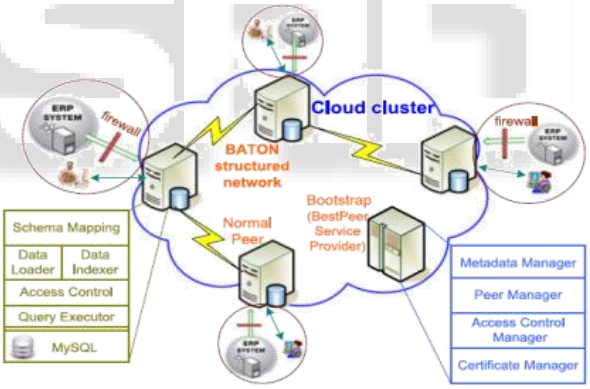

Fig. 1: Advanced Bestpeer Network Architecture

Advanced BestPeer also employs PKI encryption scheme to encrypt/decrypt data transmitted between normal peers to increase the security of the data in the network. Each normal peer owned and managed by separate business peer and processes the data retrieval requests given by the users owning the business.Normal peers are organized as balanced tree peer-to-peer overlay based on BATON which meets the required throughput (5)(9). Query processing is performed in a distributed manner.

VI. ADVANTAGES OF PROPOSED SYSTEM

- This system powerfully handles the workloads of a corporate network.
- Advanced BestPeer is based on pay-as-you-go business model. Therefore the company has to pay for what they use in terms of Advanced BestPeer instance's hours and storage capacity.

- Advanced BestPeer extend the role-based access control for distributed corporate network.
- Advanced BestPeer integrated P2P technology to retrieve data between the normal peers.
- Advanced BestPeer is a great solution for efficient data sharing system within the corporate network.

VII. BOOTSTRAP PEER

There is only one bootstrap peer for the whole network and it is run by the bootstrap service provider. The main function of bootstrap peer is to manage the normal peers.

A. Managing and Monitoring of Normal peers

In a corporate network normal peers may join the network and leave the network at any time. Bootstrap is the first connection point for every normal peer. The bootstrap peer will put the joined new peer into the peer list of corporate network only if the join request is permitted by the service provider.Simultaneously as soon as the peer is joined it will receive information such as global schema, role definition, issued certificates and current participants. Whenever a normal peer wants to leave the network, it should first notify the bootstrap peer. Then the bootstrap peer will move the departure peer into the black list and make its respective certificate as invalid. Bootstrap peer will get back all the resources assigned to the departing peer and then removes it from the normal peer list.

Bootstrap peer also monitors the health of the normal peers and takes cares of fail-over and auto-scaling events. Bootstrap peer collects performance data at some interval of time. If some nodes are crashed or if performance is very low, it will trigger an automatic fail-over event which will launch the new instance from the cloud. The new instance does the database recovery from the latest backup stored. Whenever normal peers are overloaded, it triggers auto-scaling event to increase the storage space or to increase the CPU speed. Within the recovery time the data of the crashed nodes can be recovered back forming a reliable network. Node failure problem leads to another issue i.e consistency of the system which affects the order of execution of the query. The most used eventual consistency model or the weakened consistency models doesn't suit this case. A more powerful consistency is applied by blocking all the affected queries when a node is crashed until the fail-over event is completed. This provides both the correctness and consistency of the data with the expense of some latency.

VIII. NORMAL PEER

Normal peer is instances created whenever a business organization registers with this system. It consists of five software components: schema mapping (6), data loader, data indexer, query executor and access control.

Schema Mapping: It defines the mapping between the local schema of the production system and the global shared schema which is used by the corporate network. Advanced BestPeer supports relational schema mapping which consists of metadata mappings and value mappings. Metadata mapping is mapping of table definitions of local table to global table and Value mapping is mapping of local terms to global terms. Schema mapping involves human to map and it is also time consuming but it is done only once. Further to reduce the time templates can be used which can be modified by the business organization according to their business needs.

Data Loader: It is a software component which extracts the data from individual production system to normal peer instances according to the schema mapping results. Extracting and transforming the data is easy but the main challenge is maintaining the consistency of the data between the production system and the normal peer instance.

Data Indexer: In this system the data is stored in the local MySQL database present in the normal peers. In order to process a query we need to locate the tables in the normal peers. P2P technology is used to resolve the data locating in the peers for this it uses a BATON tree structure to organize the normal peers (5). Indexes are to speed up the data retrieval (8); we use three types of the index: table index, column index and range index.

IX. BENCHMARKING

Benchmarking shows the evolution of throughput and performance of Advanced BestPeer on Amazon cloud platform.

A. PerformanceBenchmark

In performance benchmark the query latency of Advanced BestPeer is compared with HadoopDB using some queries selected from common workloads of corporate network (10). Hadoop technology is chosen as a target because it is an alternative best solution for this problem. Comparison of these two systems reveals the performance gap between a more general data warehousing systems and specially designed data sharing system for business or corporate network applications.

B. Throughput Benchmark

Throughput benchmark studies the query throughput of Advanced BestPeer. In this, a simple supply-chain network consisting of suppliers and retailers is produced and the query throughput of the system is studied. HadoopDB is not built for high query throughput (10) .Advanced BestPeer achieves linear scalability in both heavy-weight workload and light-weight workload.

X. CONCLUSION

This paper defines the exclusive challenges faces by sharing of data and also data processing in inter-business environment. Proposed Advanced BestPeer system delivers elastic data sharing services by combining cloud computing, database and p2p technology. This system reduces the operational and maintenance costs by increasing the ROI. The experiment conducted on Amazon EC2 cloud shows that this can successfully handle all workloads in a corporate network with linear query throughput. Therefore Advanced BestPeer is a best solution to handle data sharing between the corporate networks.

REFERENCE

[1] Yang, Beverly; Garcia-Molina, Hector (2001). "Comparing Hybrid Peer-to-Peer

Systems". Very Large Data Bases. Retrieved 8 October 2013.

[1] A. Abouzeid, K. Bajda-Pawlikowski, D.J. Abadi, A. Rasin, and A.Silberschatz, "HadoopDB: An Architectural Hybrid of MapReduceand DBMS Technologies for AnalyticalWorkloads," Proc.VLDB Endowment, vol. 2, no. 1, pp. 922-933, 2009.

[2] S. Wu, Q.H. Vu, J. Li, and K.-L. Tan, "Adaptive Multi-Join Query Processing in PDBMS," Proc. IEEE Int'l Conf. Data Eng. (ICDE '09), pp. 1239-1242, 2009.

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

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

[5] C. Batini, M. Lenzerini, and S. Navathe, "A Comparative Analysis of Methodologies for Database Schema Integration," ACM Computing Surveys, vol. 18, no. 4, pp. 323-364, 1986.

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

[7] G. Chen, H. T. Vo, S. Wu, B. C. Ooi, T. "A Framework for Supporting DBMS-like Indexes in the Cloud." Ozsu VLDB 2011.

[8] H.V. Jagadish, B.C. Ooi, K.-L. Tan, Q.H. Vu, and R. Zhang, "Speeding up Search in Peer-to-Peer Networks with a Multi-Way Tree Structure," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2006.

[9] J. Dittrich, J. Quian_e-Ruiz, A. Jindal, Y. Kargin, V. Setty, and J.Schad, "Hadoop++: Making a Yellow Elephant Run Like a Cheetah (without it Even Noticing)," Proc. VLDB Endowment, vol. 3, no. 1/2, pp. 515-529, 2010.

