A Review on Cloud Enable Data Sharing Platform Using BESTPEER++

[1]Ms. Seema Shinde, [2]Ms.Dipali Khatwar

[1]M .tech Student, [2]Professor of M.Tech
Department of computer Science & Engineering

ABSTRACT
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. It can effectively help the companies to reduce their operational costs and increase the revenues. However, the inter-company data sharing and processing poses unique challenges to such a data management system including scalability, performance, throughput, and security. In this paper, we present BestPeer++, a system which delivers elastic data sharing services for corporate network applications in the cloud based on BestPeer – a peer-to-peer (P2P) based data management platform. By integrating cloud computing, database, and P2P technologies into one system, BestPeer++ provides an economical, flexible and scalable platform for corporate network applications and delivers data sharing services to participants based on the widely accepted pay-as-you-go business model. We evaluate BestPeer++ on Amazon EC2 Cloud platform. The benchmarking results show that BestPeer++ outperforms HadoopDB, a recently proposed large-scale data processing system, in performance when both systems are employed to handle typical corporate network workloads. The benchmarking results also demonstrate that BestPeer++ achieves near linear scalability for throughput with respect to the number of peer nodes.

I INTRODUCTION

Company of the same industry sector is 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 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. First, the corporate network needs to scale up to support thousands of participants, while the installation of a large-scale centralized data warehouse system entails nontrivial costs including huge hardware/software investments and high maintenance cost. In the real world, most companies are not keen to invest heavily on additional information systems until they can clearly see the potential return on investment. Second, companies want to fully customize the access control policy to determine which business partners can see which part of their shared data. Unfortunately, most of the data warehouse solutions fail to offer such flexibilities. 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 fore mentioned problems, this paper presents Best Peer++, a cloud enabled data sharing platform designed for corporate network applications. By integrating cloud computing, database, and peer-to-peer (P2P) technologies, Best Peer++ achieves its query processing efficiency and is a promising approach for corporate network applications.

By integrating cloud computing, database, and peer-to-peer (P2P) technologies, BestPeer++ achieves its query processing efficiency and is a promising approach for corporate network applications, with the following distinguished features. 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. BestPeer++ employs P2P technology to retrieve data between business partners. BestPeer++ instances are organized as a structured P2P overlay network named BATON . The data are indexed by the table name, column name and data range for efficient retrieval. BestPeer++ employs a hybrid design for achieving high performance query processing. The major workload of a corporate network is simple, lowoverhead queries. Such queries typically only involve querying a very small number of business partners and can be processed in short time. BestPeer++ is mainly optimized for these queries.

II.LITERATURE REVIEW

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 automatic call

Overlay networks based on distributed hash tables (DHTs), are based on the principle of peers maintaining routing tables which facilitate the forwarding of a query closer node to a responsible peer until the query can be answered. The ways to partition the search space and the routing methods may differ, but all DHT overlays depend on the consistent maintenance of routing tables at each peer, taking into account network dynamics, most importantly. This paper contains novel correctionon-failure (CoF) and correction-on-use (CoU) approaches that adaptively support network maintenance under various conditions of network.
A Comparative Analysis of Methodologies for Database Schema Integration. Different data models are used primary data models used for this implementation is hierarchical,network and relational data models.These models are compared with each other and used the best model.

The *Yahoo! Cloud Serving Benchmark* (YCSB) framework, and report performance results for four systems: Cassandra, HBase, PNUTS, and a simple sharded MySQL implementation. Their focus is on performance and elasticity, the framework is intended to serve as a tool for evaluating other aspects of cloud systems such as availability and replication.

To meet the reliability and scaling needs, Amazon has developed a number of storage technologies, of which the Amazon Simple Storage Service (also available outside of Amazon and known as Amazon S3), is probably the best known. This paper presents the design and implementation of Dynamo, another highly available and scalable distributed data store built for Amazon's platform. Dynamo is used to manage the state of services that have very high reliability requirements and need tight control over the tradeoffs between availability, consistency, cost-effectiveness and performance. Amazon's platform has a very diverse set of applications with different storage requirements. A select set of applications requires a storage technology that is flexible enough to let application designers configure their data store appropriately based on these tradeoffs to achieve high availability and guaranteed performance in the most cost effective manner..

Warehousing is a promising technique for retrieval and integration of data from distributed, autonomous and possibly heterogeneous information sources . A warehouse is a repository of integrated information that is available for queries. As relevant information sources are modified, the new information is extracted, and translated to the data model of the warehouse, and integrated with the existing warehouse data. In this paper, This focus on the detection and the extraction of the modifications to the information sources.

A balanced tree structure overlay on a peer-to-peer network capable of supporting both exact queries and range queries efficiently. In spite of the tree structure causing distinctions to be made between nodes at different levels in the tree, This show that the load at each node is approximately equal. In spite of the tree structure providing precisely one path between any pair of nodes, This show that sideways routing tables maintained at each node provide sufficient fault tolerance to permit efficient repair. In a network with N nodes, They guarantee that both exact queries and range queries can be answered in $O(logN)$ steps and also that update operations have an amortized cost of $O(logN)$.

III. PROPOSED APPROACH

We use the two-tier partial replication strategy to provide both data availability and load balancing, as proposed in our recent study. To enhance the usability of conventional P2P networks, database community have proposed a series of PDBMS (Peer-to-Peer Database Manage System) by integrating the state-of-art database techniques into the P2P systems. We have discussed the unique challenges posed by sharing and processing data in an inter-businesses environment and proposed Best Peer++, a system which delivers elastic data sharing services, by integrating cloud computing, database, and peer-to-peer technologies.

The existing system provides that 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 Best Peer++ instance's hours and storage capacity. The Best Peer++ service provider elastically scales up the running instances and makes them always available

Fig. 1. The BestPeer++ network deployed on Amazon Cloud offering

Fig.1. Block Diagram of Proposed Method

As shown in Fig. 1, there are 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 MySQL 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. The query processor first parses the query and then employs the BATON search algorithm to identify the peers that hold the data related to the query. Then, the query executor employs a pay-as you- go query processing strategy, the health of normal peers and scheduling fail-over and auto-scaling events. The bootstrap periodically collects performance metrics of each normal peer.

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 Best Peer++ instance's hours and storage capacity The Best Peer++ service provider elastically scales up the running instances and makes them always available.

IV.CONCLUSION

We have discussed 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 our 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

REFERENCES

[1] Gang Chen,Tianlei Hu,Dawei Jiang,Peng Lu,Kian-Lee Tan,Hoang Tam Vo and Sai Wu " BestPeer++: A Peer-to-Peer Based Large-Scale Data Processing Platform 2014"

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

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

[4] 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.

[5] D. Bermbach and S. Tai, "Eventual Consistency: How Soon is Eventual? An Evaluation of Amazon s3's Consistency Behavior," in Proc. 6th Workshop Middleware Serv. Oriented Comput. (MW4SOC '11), pp. 1:1-1:6, NY, USA, 2011.

[6] B. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears, "Benchmarking Cloud Serving Systems with YCSB," Proc. First ACM Symp. Cloud Computing, pp. 143-154, 2010.

[7] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: Amazon's Highly Available Key-Value Store," Proc. 21st ACM SIGOPS Symp. Operating Systems Principles (SOSP '07), pp. 205-220, 2007.

[8] 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.

[9] H. Garcia-Molina and W.J. Labio, "Efficient Snapshot Differential Algorithms for Data Warehousing," technical report, Stanford Univ., 1996.

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

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

[12] 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.
[13] H.V. Jagadish, B.C. Ooi, K.-L. Tan, C. Yu, and R. Zhang, "iDistance: An Adaptive B+-Tree Based Indexing Method for Nearest Neighbor Search," ACM Trans. Database Systems, vol. 30, pp. 364-397, June 2005.
[14] 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.
[15] A. Lakshman and P. Malik, "Cassandra: Structured Storage System on a P2P Network," Proc. 28th ACM Symp. Principles of Distributed Computing (PODC '09), p. 5, 2009.