ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Evaluate BestPeer++ for Large-Scale Data Processing

Kodam Kamalakar
M.Tech Student
Department of CSE
Global Group Of Institutions
Batasingaram,
Ranga Reddy (Dist), India

Mr. S Dilli Babu, M. Tech
Assistant. Professor,
Department of CSE
Global Group Of Institutions
Batasingaram,
Ranga Reddy (Dist), India.

Mr. M V Narayana, M. Tech(Ph.D)
Associate Professor & HOD
Department of CSE
Global Group Of Institutions
Batasingaram,
Ranga Reddy (Dist), India.

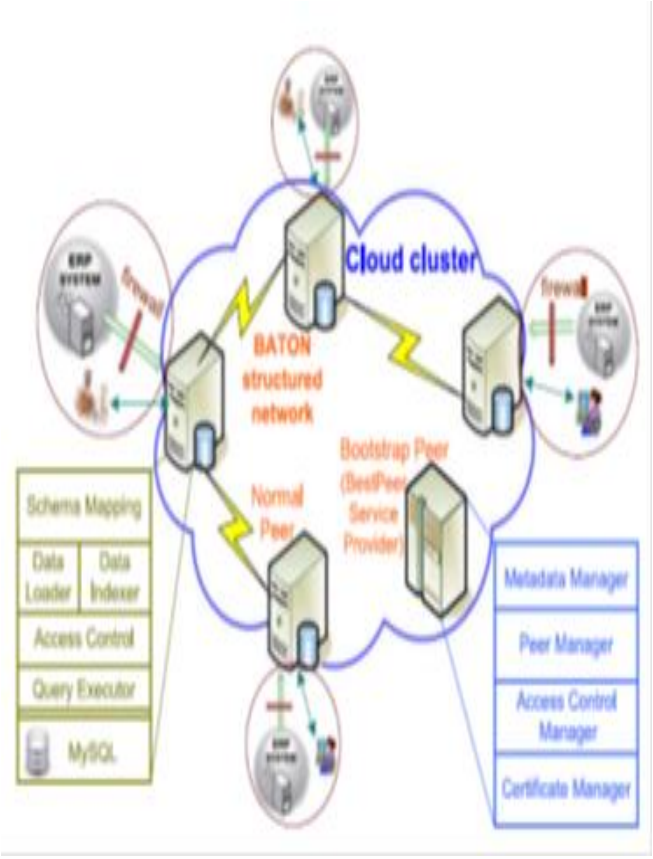
ABSTRACT-*Peer to peer network is a group of computers each of which acts as a node for sharing files within the group. To form a corporate network companies simply register their sites with the best peer++ service provider. The total cost of ownership is reduced and it leads to increases the revenue since companies don't require buying any hardware and software in advance. Best peer provides economical, flexible and scalable platform for corporate network applications by integrating cloud computing, p2p and data bas e technologies in to one system. The efficiency of best peer++ is demonstrated by benchmarking best peer against hadoop DB. At the end performance evaluation is done on Amazon EC2 cloud platform. The major contribution of this paper is design of a system in such way that it should delivers elastic data sharing services for the corporate network application in the cloud, based on the 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.*

Index Terms—*Peer-to-peer systems, cloud computing, MapReduce, query processing, index, BsetPeer, Cloud Platform, HadoopDB, Peer nodes.*

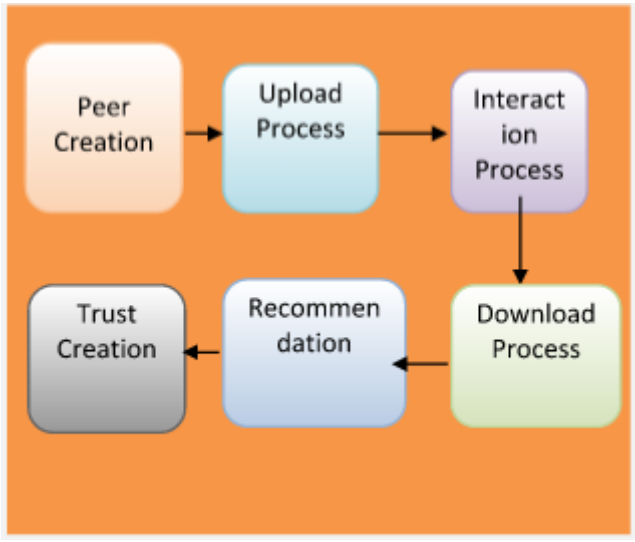
INTRODUCTION

Companies which belongs to same industry are often connected to a corporate network for association purpose each company maintains its own site and selectively shares a part of its business data with others include supply chain networks where organizations such as suppliers, manufacturers, and retailers cooperate with each others to accomplish their own business goals s uch as planning production-line, making achievement strategies and choose marketing solutions. As per technical perspective, selecting right data sharing platform for corporate network is very important is very important, a system which enables the pooled data supports capable logical queries over those data. Traditionally, data sharing was achieved by building a centralized warehousing, which regularly extracts data from the internal production system (e.g. ERP) of each company. Such warehousing solution has some deficiency in real consumptions. First the corporate network needs to scale up to support thousands of participants. In the real world, most of the companies are not ready to invest heavily on additional information system until they can clearly see the potential return on investment (ROI). Second, companies want fully customize the access control rule to determine which business partner can see which part of their shared data. Most of the data warehouse solutions fail to offer such flexibilities. Finally, to increase the revenues, companies often adjust their business process and may change their business partners. Therefore, the participants may join and leave the corporate network at resolve.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal



The data warehouse solution has not been designed to handle such dynamicity. For decrease such problem this paper design BestPeer ++ for corporate network. As an in-time response to the ever changing business demands and the appearance of the cloud computing techniques, best peer ++ has developed into its new stage of development the cloud enabled best peer ++ system. By integrating cloud computing, p2p and database technologies, BestPeer++ achieves its query processing competence in a pay-as-you-go cloud business model. This paper shows that design of BestPeer++ system that provides inexpensive, flexible solutions for corporate network. Performance of best peer++ will be demonstrated by benchmarking best peer++ against HadoopDB. In the below mentioned architecture the Bootstrap peer is run by the Best Peer++ s ervice provider, and its main functionality to manage the bets peer++ network. Bootstrap peer consist a peer manager, access control manager, metadata manager and certification manager.



The Best Peer++ system can communicate with many normal peers and each normal peer is managed and controlled by the bootstrap peer system. Each normal peer having three sub modules like query engine, Heartbeat (HB) and downloader. The results show that for simple quires, the performance of BestPeer++ is significantly better than HadoopDB.

Open nature of peer-to-peer systems exposes them to malicious activity. Building trust relationships among peers can mitigate attacks of malicious peers. This paper presents distributed algorithms that enable a peer to reason about trustworthiness of other peers based on past interactions and recommendations. Peers create their own trust network in their proximity by using local information available and do not try to learn global trust information. Two contexts of trust, service, and recommendation contexts are defined to measure trustworthiness in providing services and giving recommendations. Interactions and recommendations are evaluated based on importance, recentness, and peer satisfaction parameters. Additionally, recommender's trustworthiness and confidence about a recommendation are considered while evaluating recommendations. Simulation experiments on a file sharing application show that the proposed model can mitigate attacks on 16 different malicious behavior models. In the experiments, good peers were able to form trust relationships in their proximity and isolate malicious peers.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

LITERATURE SURVEY

1 "A Comparative Analysis of Methodologies for Database Schema Integration,"

One of the fundamental principles of the database approach is that a database allows a no redundant, unified representation of all data managed in an organization. This is achieved only when methodologies are available to support integration across organizational and application boundaries. Methodologies for database design usually perform the design activity by separately producing several schemas, representing parts of the application, which are subsequently merged. Database schema integration is the activity of integrating the schemas of existing or proposed databases into a global, unified schema. The aim of the paper is to provide first a unifying framework for the problem of schema integration, then a comparative review of the work done thus far in this area. Such a framework, with the associated analysis of the existing approaches, provides a basis for identifying strengths and weaknesses of individual methodologies, as well as general guidelines for future improvements and extensions.

2. "A Comparative Analysis of Methodologies for Database Schema Integration,"

One of the fundamental principles of the database approach is that a database allows a no redundant, unified representation of all data managed in an organization. This is achieved only when methodologies are available to support integration across organizational and application boundaries. Methodologies for database design usually perform the design activity by separately producing several schemas, representing parts of the application, which are subsequently merged. Database schema integration is the activity of integrating the schemas of existing or proposed databases into a global, unified schema. The aim of the paper is to provide first a unifying framework for the problem of schema integration, then a comparative review of the work done thus far in this area. Such a framework, with the associated analysis of the existing approaches, provides a basis for identifying strengths and weaknesses of individual methodologies, as well as general guidelines for future improvements and extensions.

3. "BATON: A Balanced Tree Structure for Peer-to-Peer Networks,"

We propose 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, we 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, we show that sideways routing tables maintained at each node provide sufficient fault tolerance to permit efficient repair. Specifically, in a network with N nodes, we guarantee that both exact queries and range queries can be answered in O(log N) steps and also that update operations (to both data and network) have an amortized cost of O(log N). An experimental assessment validates the practicality of our proposal.

4. "Data Sharing in the Hyperion Peer Database System,"

This demo presents Hyperion, a prototype system that supports data sharing for a network of independent Peer Relational Database Management Systems (PDBMSs). The nodes of such a network are assumed to be autonomous PDBMSs that form acquaintances at run-time, and manage mapping tables to define value correspondences among different databases. They also use distributed Event-Condition-Action (ECA) rules to enable and coordinate data sharing. Peers perform local querying and update processing, and also propagate queries and updates to their acquainted peers. The demo illustrates the following key functionalities of Hyperion: (1) the use of (data level) mapping tables to infer new metadata as peers dynamically join the network, (2) the ability to answer queries using data in acquaintances, and (3) the ability to coordinate peers through update propagation.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

5. "Adaptive Multi-Join Query Processing in PDBMS,"

Traditionally, distributed databases assume that the (small) set of nodes participating in a query is known apriori, the data is well placed, and the statistics are readily available. However, these assumptions are no longer valid in a peer-based database management system (PDBMS). As such, it is a challenge to process and optimize queries in a PDBMS. In this paper, we present our distributed solution to this problem for multi-way join queries. Our approach first processes a multi-way join query based on an initial query evaluation plan (generated using statistical data that may be obsolete or inaccurate); as the query is being processed, statistics obtained on-the-fly are used to (continuously) refine the current plan dynamically into a more effective one. We have conducted an extensive performance study which shows that our adaptive query processing strategy can reduce the network traffic significantly.

PROBLEM STATEMENT

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 (a.k.a total cost of ownership) and high maintenance cost (a.k.a total cost of operations). 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 (ROI). 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.

Drawbacks:-

- Its most of the data warehouse solutions fail to offer flexibilities.
- Its warehousing solution has some deficiencies in real deployment.
- It is expensive.

PROBLEM DEFINTION

BestPeer++ achieves its query processing efficiency and is a promising approach for corporate network applications, with the following distinguished features. BestPeer++ is deployed as 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++ 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. 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. Best-Peer++ is mainly optimized for these queries. For infrequent time-consuming analytical tasks, we provide an interface for exporting the data from Best-Peer++ to Hadoop and allow users to analyze those data using MapReduce.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Advantages:-

- It provides economical, flexible and scalable solutions for corporate network applications.
- It is more efficient.
- It prevent undesired business partners to access their shared data.

IMPLEMENTATION

Peer++ Processing Approach:-

BestPeer++ employs 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 submit-ted 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 (the list of these normal peers is determined by searching the indices stored in BATON). The subquery is then processed by each remote normal peer and the intermediate results are shuffled to the query submitting peer P. In the processing step, the normal peer P first collects all the required data from the other participating normal peers. To reduce I/O, the peer P creates a set of Mem Tables to hold the data retrieved from other peers and bulk inserts these data into the local MySQL when the Mem Table is full. After receiving all the necessary data, the peer P finally evaluates the submitted query.

Parallel P2P Processing:-

For each join, instead of forwarding all tuples into a single processing node, we disseminate them into a set of nodes, which will process the join in parallel. We adopt the conventional replicated join approach. Namely, the small table will be replicated to all processing nodes and joined with a partition of the large table.

Implementing MapReduce:-

The main difference between MapReduce method and native P2P method comes from the join processing. In MapReduce method, instead of doing replicate joins, the symmetric-hash join approach is adopted. Each mapper reads in its local data and shuffles the intermediate tuple according to the hash value of the join key. Therefore, each tuple only needs to be shuffled once on each level. Note that the configuration and launch of a MapReduce job also incurs certain overhead, which, can be measured in the runtime, is a constant value.

Adaptive Query Processing:-

For small jobs, the P2P engine performs better than the MapReduce engine, as it does not incur initialization cost and database join algorithms have been well optimized. However, for large-scale data analytic jobs, the MapReduce engine is more scalable, as it does not incur recursive data replications. Based on the above-mentioned cost models, we propose our adaptive query processing approach. When a query is submitted, the query planner retrieves related histogram and index information from the bootstrap node, analyzes the query and constructs a processing graph for the query. Then the costs of both the P2P engine and MapReduce engine are predicted based on the histograms and runtime parameters of the cost models. The query planner compares the costs between two methods and executes the one with lower cost.

RELATED WORK

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. These PDBMS can be classified as the unstructured systems such as PIAZZA, Hyperion and PeerDB, and the structured systems such as PIER . The work on unstructured PDBMS focus on the problem of mapping heterogeneous schemas among nodes in the systems. PIAZZA introduces two materialized view approaches, namely local as view (LAV) and global as view (GAV). PeerDB employs information retrieval technique to match columns of different tables. The main problem of unstructured

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

PDBMS is that there is no guarantee for the data retrieval performance and result quality.

The structured PDBMS can deliver search service with guaranteed performance. The main concern is the possibly high maintenance cost . To address this problem, partial indexing scheme is proposed to reduce the index size. Moreover, adaptive query processing and online aggregation techniques have also been introduced to improve query performance.

The techniques of PDBMS are also adopted in cloud systems. In Dynamo , Cassandra, and ecStore , a similar data dissemination and routing strategy is applied to manage the large-scale data. BestPeer++ is different from the systems based on the MapReduce/Hadoop framework (e.g., HadoopDB, Hive and Hadoop++). Hadoop-based systems are designed to process large-scale data sets in batch mode. They efficiently process aggregate queries by exploiting the parallelism. The SQL queries need to be translated into multiple MapReduce jobs, which are processed sequentially. BestPeer++, on the other hand, can handle both ad-hoc queries and costly analysis queries. It provides built-in MapReduce support and adaptively switches between its distributed processing strategy and MapReduce strategy based on the cost model. BestPeer++ shares a similar design philosophy with HadoopDB. In both systems, each processing instance maintains a local DBMS. The local DBMS helps manage the local data and
improve the query processing with the database techniques, such as index and optimizer.

CONCLUSION

This paper define exclusive challenges faced by contribution and open-handed out data in an interbusinesses environment and planned BestPeer++, a system which deliver elastic data sharing services, by Containing cloud computing, database, and peer-to-peer technologies. The standard conducted on Amazon EC2 cloud platform shows that our system can powerfully handle typical workloads in a corporate network. It can move near linear query throughput as the number of normal peers grows. Therefore, BestPeer++ is great solution for capable data sharing within corporate networks.

REFERENCES

[1] Gang Chen, Tianlei Hu, Dawei Jiang, Peng Lu, Kian- Lee Tan, Hoang Tam Vo, and Sai Wu, "Extended BestPeer: A Peer-to-Peer Based Large-Scale Data Processing Platform",VOL. 26,NO. 6, JUNE 2014.

[2] 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.

[3] 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.

[4] S. Wu, S. Jiang, B.C. Ooi, and K.-L. Tan, "Distributed Online Aggregation," Proc. VLDB Endowment, vol. 2, no. 1, pp. 443-454, 2009.

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

[6] 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.

[7] Beng Chin Ooi, Yanfeng Shu, "Relational Data Sharing in Peer-based Data Management Systems." KianLee Tan Sigmod Record special issue on P2P, 2003.

[8] B.C. Ooi, K.L. Tan, A.Y. Zhou, C.H. Goh, Y.G. Li, C.Y. Liau, B. Ling, W.S. Ng, Y.F. Shu, X.Y. Wang, M. Zhang " PeerDB: Peering into Personal Databases." The 2003 ACM SIGMOD Intl. Conf. on Management of Data (Demo). (SIGMOD 2003).

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

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

[10] Sai Wu, Dawei Jiang, Beng Chin Ooi, Kun Lun Wu" Efficient B+-tree Based Indexing for Cloud Data Processing VLDB 2010.

[11] Heng Tao Shen, Yanfeng Shu, and Bei Yu IEEE Trans. Knowl. "Efficient Semantic-Based Content Search in P2P Network." Data.