A Cloud Enable Data Sharing Platform using BESTPEER++

Seema Shinde
Department of Computer Science & Engineering, INDIA

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.

Keywords---- BestPeer, ERP, Data warehouse

I. INTRODUCTION
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. 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.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then next steps is to determine which operating system and language can be used for developing the tool. Once
the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from books or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

A. Survey On Networking

In the world of computers, networking is the practice of linking two or more computing devices together for the purpose of sharing data. Networks are built with a mix of computer hardware and computer software.

Networks consist of the computers, wiring, and other devices, such as hubs, switches and routers that make up the network infrastructure. Some devices, such as network interface cards, serve as the computer’s connection to the network. Devices such as switches and routers provide traffic control strategies for the network. All sorts of different technologies can actually be employed to move data from one place to another, including wires, radio waves, and even microwave technology.

Network architecture

Asynchronous Transfer Mode

Asynchronous Transfer Mode (ATM) is a switching technique for telecommunication networks. It uses asynchronous time-division multiplexing and encodes data into small, fixed-sized cells. This differs from other protocols such as the Internet Protocol Suite or Ethernet that use variable sized packets or frames. ATM has similarity with both circuit and packet switched networking. This makes it a good choice for a network that must handle both traditional high-throughput data traffic, and real-time, low-latency content such as voice and video. ATM uses a connection-oriented model in which a virtual circuit must be established between two endpoints before the actual data exchange begins.

Network topology

Common layouts

A network topology is the layout of the interconnections of the nodes of a computer network. Common layouts are:

A bus network: all nodes are connected to a common medium along this medium. This was the layout used in the original Ethernet, called 10BASE5 and 10BASE2.

A star network: all nodes are connected to a special central node. This is the typical layout found in in a Wireless LAN, where each wireless client connects to the central Wireless access point.

A ring network: each node is connected to its left and right neighbor node, such that all nodes are connected and that each node can reach each other node by traversing nodes left- or rightwards. The Fiber Distributed Data Interface (FDDI) made use of such a topology.

A mesh network: each node is connected to an arbitrary number of neighbors in such a way that there is at least one traversal from any node to any other.

A fully connected network: each node is connected to every other node in the network.

Overlay networks based on distributed hash table [2]

An overlay network is a virtual computer network that is built on top of another network. Nodes in the overlay are connected by virtual or logical links, each of which corresponds to a path, perhaps through many physical links, in the underlying network. The topology of the overlay network may (and often does) differ from that of the underlying one.

A sample overlay network: IP over SONET over Optical

For example, many peer-to-peer networks are overlay networks because they are organized as nodes of a virtual system of links run on top of the Internet. The Internet was initially built as an overlay on the telephone network [11].

The most striking example of an overlay network, however, is the Internet itself: At the IP layer, each node can reach any other by a direct connection to the desired IP address, thereby creating a fully connected network; the underlying network, however, is composed of a mesh-like interconnect of subnetworks of varying topologies (and, in fact, technologies). Address resolution and routing are the means which allows the mapping of the fully-connected IP overlay network to the underlying ones.

Overlay networks have been around since the invention of networking when computer systems were...
Another example of an overlay network is a distributed hash table, which maps keys to nodes in the network. In this case, the underlying network is an IP network, and the overlay network is a table (actually map) indexed by keys.

Overlay networks have also been proposed as a way to improve Internet routing, such as through quality of service guarantees to achieve higher-quality streaming media. Previous proposals such as IntServ, DiffServ, and IP Multicast have not been seen wide acceptance largely because they require modification of all routers in the network. On the other hand, an overlay network can be incrementally deployed on end-hosts running the overlay protocol software, without cooperation from Internet service providers. The overlay has no control over how packets are routed in the underlying network between two overlay nodes, but it can control, for example, the sequence of overlay nodes a message traverses before reaching its destination.

Routers

A router is an internetworking device that forwards packets between networks by processing information found in the datagram or packet (Internet protocol information from Layer 3 of the OSI Model). In many situations, this information is processed in conjunction with the routing table (also known as forwarding table). Routers use routing tables to determine what interface to forward packets (this can include the "null" also known as the "black hole" interface because data can go into it, however, no further processing is done for said data).

Network security

In the field of networking, the area of network security consists of the provisions and policies adopted by the network administrator to prevent and monitor unauthorized access, misuse, modification, or denial of the computer network and network-accessible resources. Network Security is the authorization of access to data in a network, which is controlled by the network administrator. Users are assigned an ID and password that allows them access to information and programs within their authority. Network Security covers a variety of computer networks, both public and private that are used in everyday jobs conducting transactions and communications among businesses, government agencies and individuals. Networks can be private, such as within a company, and others which might be open to public access. Network Security is involved in organization, enterprises, and all other type of institutions. It does as its titles explains, secures the network. Protects and oversees operations being done.

III. EXISTING METHODOLOGY

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. 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 non trivial 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.

The local administrator at each normal peer can assign the new user with an existing role if the access privilege of that role is applicable to the new user. If none of the existing roles satisfies the new user, the local administrator can create new roles by three operators. Note that Best Peer++ does not collect the information of existing users in the collaborating ERP databases, since it will lead to potential security issues. Instead, the user management module of Best Peer++ provides interfaces for the local administrator at each participating organization to create new accounts for users who desire to access Best Peer++ service.

The Existing system attempt to develop peer-to-peer (P2P) technologies for Corporate Network. Corporate network applications such as supply chain management and national healthcare network. A warehousing solution has some disadvantages in real consumption. First, the
corporate network needs to extent support thousands of participants, while the fitting of a large-scale centralized data warehouse system entails nontrivial costs including big hardware/software investments and high preservation cost. In the environment, most companies are not dedicated to invest deeply on additional information systems until they can clearly see the potential return on investment. Second, companies want to completely modify the access control policy to determine which business partners can see which part of their shared data. Disadvantages of Existing System is most of the data warehouse solutions fail to present flexibilities and Solution has not been designed to grip dynamicity.

If we look from the technical point of view, the key for the success of any 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 shown in figure 2) of each company for subsequent querying. Unfortunately, such a warehousing solution has some deficiencies in real deployment. From figure 2 we can clearly get an idea that the following departments which come under ERP like purchasing, production, accounting, HR, Engineering and so on

IV. PROPOSED METHODOLOGY

MODULE DESCRIPTION

THE BEST PEER++ SYSTEM

BestPeer data management platform.
Best Peer++, a cloud enabled evolution of Best Peer
Amazon Cloud Adapter

The BestPeer++ Core
Managing Normal Peer Join/Departure
Auto Fail-over and Auto-Scaling

Schema Mapping

Data Loader

MODULE DESCRIPTION

BestPeer++ core contains all platform-independent logic, including query processing and P2P overlay. It runs on top of adapter and consists of two software components: bootstrap peer and normal peer. A BestPeer++ network can only have a single bootstrap peer instance which is always launched and maintained by the BestPeer++ service provider and a set of normal peer instances. The architecture is depicted in Figure 1. This section briefly describes the functionalities of these two peers. Individual components and data flows inside these peers are presented in the subsequent sections.

Managing Normal Peer Join/Departure

Each normal peer which wants to join an existing corporate network must first connect to the bootstrap peer. If the join request is permitted by the service provider, the bootstrap peer will put the newly joined peer into the peer list of the corporate network. At the same time, the joined peer will receive the corporate network information including the current participants, global schema, role definitions, and an issued certificate. When the normal
peer needs to leave the network, it will also notify the bootstrap peer. The bootstrap peer will put the departure peer on the black list and mark the certificate of the departing peer invalid. Then, the bootstrap peer will release all resources allocated for the departing peer back to the cloud and finally remove the departing peer from the peer list.

- **Auto Fail-over and Auto-Scaling**

  In addition to managing peer join and peer departure, the bootstrap peer is also responsible for monitoring the health of normal peers and scheduling fail-over and auto-scaling events. The bootstrap periodically collects performance metrics of each normal peer. If some peers are malfunctioned or crashed, the bootstrap peer will trigger an automatic fail-over event for each failed normal peer. The automatic fail-over is performed by first launching a new instance from the cloud. Then, the bootstrap peer asks the newly launched instance to perform database recovery from the latest database backup stored in Amazon EBS. Finally, the failed peer is put into the blacklist. On the other hand, if any normal peer is overloaded (e.g., CPU is over-utilized or free storage space is low), the bootstrap peer triggers an auto-scaling event to either upgrade the normal peer to a larger instance or allocate more storage spaces. At the end of each maintenance epoch, the bootstrap releases the resources in the blacklist and notifies the changes to all participants.

- **Schema Mapping**

  Schema mapping is a component that defines the mapping between the local schema employed by the production system of each business and the global shared schema employed by the corporate network. Currently, BestPeer++ only supports relational schema mapping, namely both local schema and the global schema are relational. The mapping consists of metadata mappings (i.e., mapping local table definitions to global table definitions) and value mappings (i.e., mapping local terms to global terms). In general, the schema mapping process requires human to be involved and is time consuming. However, it only needs to perform once. Furthermore, Best- Peer++ adopts templates to facilitate the mapping process. For each popular production system (i.e., SAP or PeopleSoft), we provide a mapping template which defines the transformation of local schema of those systems to the global schema. The business only needs to modify the mapping template to meet its own needs. We found that this mapping template approach works well in practice and significantly reduces the service setup efforts.

- **Data Loader**

  Data Loader is a component that extracts data from production systems to normal peer instances according to the schema mappings. While the process of extracting and transforming data is straightforward, the main challenge is in maintaining the 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) when the raw data are updated inside the production systems.

V. IMPLEMENTATION OF PROPOSED METHODOLOGY

**PROPOSED SYSTEM**

Specifically, 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 BestPeer++, a system which delivers elastic data sharing services, by integrating cloud computing, database, and peer-to-peer technologies.

**ADVANTAGE**

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.

**ALGORITHMS**

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. Given two consecutive data snapshots, we employ a similar algorithm as the one proposed. The algorithm executes the sort merge algorithm on the tables in both snapshots.

The system employs bloom join algorithm to reduce the volume of data transmitted through the network.

VI. 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.


