Peer-to-Peer Data Sharing and Deduplication using Genetic Algorithm

Prof . J. R. Waykole, Ms. S. P. Band*, Ms. V. D. Amritkar*, Ms. P. R. Adsul*, Ms. S. P. Agawane*.**

**(Associate Professor,Department of Computer Engineering, Pune University)

*(UG Student, Pune University)

ABSTRACT:

To form corporate network organization simply join using register their sites with the peer-to-peer(P2P) service provider, and share their information among the participating organization. It can effectively help the organization to reduce their operational costs and increase the revenues. However, the inter- organization data sharing and processing posses unique challenges to such a data management system including scalability, performance, throughput, and security , a system which delivers elastic data sharing services for corporate network applications in the cloud based on a peer-to-peer based data management platform. By integrating cloud computing, database, and P2P technologies and genetic algorithm for deduplication into one system. P2P 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.

***Keywords*: Cloud computing, Deduplication, Genetic algorithm.**

1. INTRODUCTION

Different companies which have common interest for sharing data are always connected to corporate network[1].

The era of cloud computing technology provides various services to the human which is need. Cloud computing provides a platform for other advanced technology like big data, mobile computing to inculate its service and provides QOS to the customers[1]. The cloud has grown to a vast extend over the period of years. All the services that are provided to the customer are done using cloud as their backbone, it give vast amount of resources and infrastructure and consumer to act as vendors to small scale business and cloud could provide services to fully fledged organization less cost. Cloud provides space for extending the services as service provider and also it can provide infrastructure service to small scale service vendors[2].

Deduplication is key operation in integrating data from heterogeneous sources. The main challenge in this task is designing a function that can be resolve when a pair of records refers to same entity inspite of various data inconsistencies. Deduplication reduce amount of storing data by eliminating redundant copy of data. Problems in sharing and processing data in corporate network and proposed a new system peer to peer, which is used to deliver data sharing facilities by including P2P technology[4]. To configure a corporate network, organization simply register their sites provider; launch peer to peer instances in the network and exports the data to those instances for sharing purpose[3].

2. LITERATURE SURVEY

PeerDB: A P2P-based System for Distributed Data Sharing Peer-to-peer (P2P) technology is an emerging paradigm that is now viewed as a potential technology that could distributed architectures (e.g., the Internet). In a P2P distributed system, a large number of nodes(e.g., PCs connected to the Internet) can potentially be pooled together to share their resources, information and services. These nodes, which can both consume as well as provide data and/or services, may join and leave the P2P network at any time, resulting in a truly dynamic and ad-hoc environment. The distributed nature of such a design provides exciting opportunities for new killer applications to be developed[4].

Detection of Duplicate Record using Genetic Algorithm: Genetic algorithms are ideal for these types of problems where the search space is large and the number of feasible solutions is small. To apply a genetic algorithm to a scheduling problem we must first represent it as a

genome. One way to represent a scheduling genome is to define a sequence of tasks and the start times of those tasks relative to one another. Each task and its corresponding start time represent a gene. A specific sequence of tasks and start times (genes) represents one genome in our population. To make sure that our genome is a feasible solution we must take care that it obeys our precedence constraints. We generate an initial population using random start times within the precedence constraints. With genetic algorithms we then take this initial population and cross it, combining genomes along with a small amount of randomness (mutation). We let this process continue either for a pre allotted time or until we find a solution that fits our minimum criteria.Several systems such as digital libraries another database system likes organization databases are affected by the duplicates[3].

Efficient data processing in peer network using cloud computing: A cloud called intensive technique with p2p, in which different companies(peers) will stored data. This cloud will be web based so it will be available any time any where online. Company needs to login into the cloud system to upload their data. Data stored on cloud securely[2].

Amazon Cloud Adapter:The Amazon Cloud Adapter provides an elastic hardware infrastructure for P2P 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. We use Amazon EC2 service to provision the database server. Each time a new business joins the P2P corporate network, a dedicated EC2 virtual server is launched for that business. The newly launched virtual server (called Peer-to-Peer instance) runs a dedicated MySQL database software and the P2Psoftware[2].

2.1 GENETIC ALGORITHM

Genetic Algorithm is one of the evolutionary technique based on natural selection. For solving optimization problems a genetic algorithm (GA) is an evolutionary algorithm used. The algorithm repeatedly refines an initial population of possible solutions until a solution is found. An initial population of solutions is created randomly. These solutions are then evaluated using a fitness function. A selection method is applied in order to choose a parent. Genetic operators are applied to the chosen parents to create offspring. This process of evaluation, selection and recreation is continued until either a solution has been found or a number of iterations/generations have been reached. It is well known for its best performance in searching large spaces and as well as its capability to operate over the population of individuals. It not only creates new solutions but also allows new combination of features[5]. The basic flow of genetic algorithm is shown in figure below

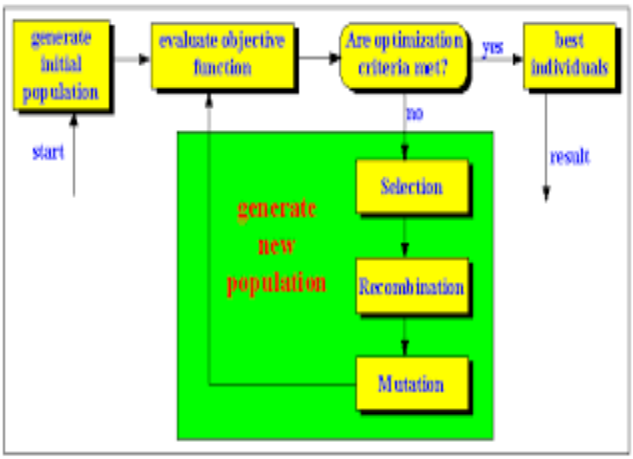

Fig1.1: GA Flow

3. Proposed System

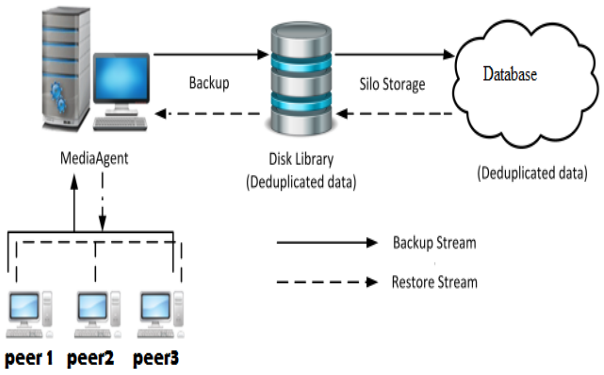

Fig1.2. Proposed System

3.1Peer++ Processing Approach:

Peer to Peer 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 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.

3.2 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.

3.3 Deduplication using Genetic algorithm:

Deduplication is the operation of integrating data from different data sources(i.e industrial sector, medical and social sector). The main task of deduplication is eliminate the duplicate data from the database storage like cloud. It also checks attributes of the records present in database. Genetic algorithm is an evolutionary algorithm. It is used for solving the optimization problems.

3.4 Auto failover Condition

The bootstrap periodically collects performance metrics of
each normal peer. If some peers are malfunctioned, the bootstrap peer will trigger an automatic fail-over event. The automatic fail-over is performed by first launching a new instance from bootstrap peer. Then, the bootstrap peer asks the newly launched instance to perform data-base recovery from
the latest database backup stored in bootstrap peer. Finally, the failed peer is put into the blacklist.

3.5 Auto Scaling-Up Condition

Similarly, if any normal peer is overloaded (e. g., CPU is overutilized
or free storage space is low), the bootstrap peer triggers an auto-scaling event to either promote the normal peer to larger instance or allocate more storage spaces.

4. CONCLUSION

Problem in sharing and processing data in corporate network are solved by including P2P (Peer-to-Peer) technology, query processing and access control which is used to delivered data effectively. To configure a corporate network, organization simply register their sites with P2P service provider , launch P2P instances in the network and finally exports the data to those instances for sharing purpose. Genetic algorithm is used to reduce the duplicate records from the cloud. P2P accepts the pay-as-you-go bussiness model popularized by cloud computing. The benchmark conducted on

cloud platform shows that our system can efficiently handle typical workloads in corporate network and can deliver near a linear query throughput as the number of normal peers grows. Therefore, P2P is a promising solution for efficient data sharing within corporate networks

REFERENCES

[1] B.Cooper.A. SilbersteinE.Tam. R. Ramakrishnan, and R. Sears, “Benchmarking cloud serving system with YCSB,proc.” First ACM Symp. Cloud computing, 143-154, 2010.

[2] Shilpa V. Paralkar, GayatriKabra, “Efficient data processing in peer network using cloud computing”, 2015.

[3] ShitalGujar, AvinashShrivas, “Detection of Duplicate records using Genetic Algorithm”, 2014.

[4] W.S. Ng, B.C. Ooi, K.-L. Tan, and A. Zhoui,”PeerDB: A P2P-Based System for Dis-tributed Data Sharing,Proc”. 19th International Conf. Data Eng., pp. 633-644, 2003.

[5] J. Stender, Brainware GmbH, "Introduction to Genetic Algorithm", Berlin, London, 1997.

[6] J. R. Waykole, S. M. Shinde "A survey paper on Deduplication using Genetic Algorithm alongwith Hash Based Algorithm", 2014.

[7] Oracle Incl,"Achieving the Cloud Computing Vision", White Paper, 2010.

[8] 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 Platforms", 2015.

[9] Prof.S.A.Agrawal 1, Kalyani Pathak2 , Yogesh Barhe3 , Chetan Chavan4 , ShrishailyaBhinge, "Large scale Data Sharing using BestPeer++ Technique", 10 oct 2015.

[10]N. Vijayalakshmi, B. Ramesh, "BestPeer++:- A Peer-to-Peer Based Large Scale Data Processing Platforms", 2015.