ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

A Peer-to-Peer Based System Used for Sharing Large Scale Data

Bhavsar Harshada V.[1], Dr. S. V. Gumaste[2], Prof. Deokate Gajanan S.[3]

ME Student, Department of Computer Engineering, Sharadchandra Pawar College of Engineering, Dumbarwadi, Otur, Pune, India[1]

Professor, Department of Computer Engineering, Sharadchandra Pawar College of Engineering, Dumbarwadi, Otur, Pune, India[2]

Assistant Professor, Department of Computer Engineering, Sharadchandra Pawar College of Engineering, Dumbarwadi, Otur, Pune, India[3]

ABSTRACT: The Companies used sharing data where they need to contribute or they share common interest. As per increasing business trends and maximum used of cloud computing, the new system evolved in new stage of growth towards cloud enabled system. In this system based on peer to peer system develop data sharing service in shared network. This system is the combination of cloud computing, databases and peer to peer based technologies. This system gives the efficiency as pay as you go manner. This paper used benchmarking method for analysis and evolution by comparing this method with Hadoop DB which is Large Scale data processing platform. This system having security by providing private key and admin authorized to provide access to other user. Using Cloud Computing they allowed to user to store their data into cloud and access when required.

KEYWORDS: Cloud computing, Map Reduce, Network Security, Peer-to-peer systems, Query processing

I. INTRODUCTION

Sharing Companies having common interest are always connected to a corporate network for sharing purposes [2]. A company creates its own website and shares a part of its business data with others which include supply chain networks such as supplier, manufacturer, and retailer who co-operate with each other to achieve their goals such as business planning, reducing production cost, developing business strategies and marketing solutions. Selecting right data sharing platform is very important task for sharing network. Usually, centralized data such as Data warehouse is used for data sharing, which extracts data from the internal production systems (e.g., ERP) of each company for following querying. Actually this data warehouse having some deficiency Such as [4], First, The share data network wants to scope up to support thousands of participants. Second, companies want to fully modify the access control rule to determine which business partners can see which part of their shared data. Most of them failed to overcome such problem. At last to increase the revenue; companies may change their business partners. Therefore, the participants may join and leave the share networks at resolve [1]. This situation cannot be handled by physical data warehouse, to overcome such problem this designs the system for Shared Network for data sharing. This system is the combination of cloud computing, databases and peer to peer based technologies. This system gives the efficiency as pay as you go manner.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

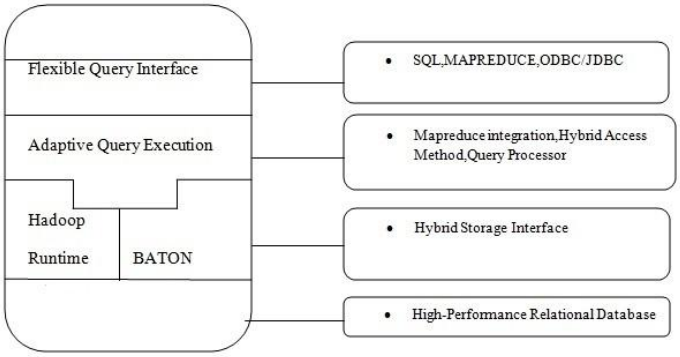

Fig.1: The Structure of System

This system has developed into its new stage of development as a cloud-enable System [3]. The structure of the system shows in Fig 1. In summary, this paper shows that designed system provides inexpensive, Flexible solutions for shared network. They evaluate the system against Hadoop DB [2], an approach for data sharing applications which shows that the proposed system is significantly better than Hadoop DB.

II. LITERATURE SURVEY

This section explains the existing System and some basic concept used in this system.

A. Existing System

The original system attempt to develop peer-to-peer (P2P) technologies for shared Network. Existing system was designed to work as a scalable, sharable, and secure P2P-based Data Management system 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. Supply chain management and national healthcare network are consider as shared network application. Existing system provides an effective and efficient way to share data among to different association and also provide enterprise quality query facility, without the requirement to set up a big centralized server [5]. Data warehouse used in existing system have some problem. First, the shared network environment, most companies are not dedicated to invest deeply on additional information systems until they can clearly see the potential return on investment (ROI) [11]. Second, companies want to completely modify the access control policy to determine which business partners can see which part of their shared data.

B. What is Peer to Peer Distributed Data Management :

Practically all existing P2P systems are designed to support sharing of data at a coarse granularity (e.g., files, documents) [3]. In this section, it first distinguishes between P2P systems and distributed database systems. Then define P2P distributed data management by looking at three examples (due to space constraints) of how P2P technology can be employed for distributed database applications. This will also serve to motivate the need for database technology in P2P systems [6].

C. P2P Distributed Data Management Systems Applications

Instead of formalizing the concept of P2P distributed data management systems, It show with sample applications on what such systems may be like [4].

1) Health Care

In a hospital, each specialist has a group of patients that are solely under his care. While some patient data are stored in a centralized server of the hospital (e.g., name, address, etc), other data (e.g., X-rays, prescription, allergy to drugs, history, Reaction to drugs, etc) are typically managed by the specialist on his personal PC [5]. For most of these patients, the specialist is willing to share their data, but there are always some cases that he is unwilling to share for different reasons (e.g., part of his research program on a new drug, etc). By making the shareholder patient data

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

available to other specialists, it allows them to look for other patients who may have similar symptoms as their own patients, and hence can help them in making better decisions on the treatment (e.g., drugs to prescribe, reactions to look out for, etc).

2) Genomic Data

The discovery of new proteins necessitates complex analysis in order to determine their functions and classifications [5,7]. The main technique that scientists use in determining this information has two phases. The first phase involves searching Known protein databases for proteins that ”match” the unknown protein. The second phase involves analysing the functions and classifications of the similar proteins in an attempt to infer commonalities with the new protein. While there are several known servers on genomic data (e.g., GenBank, SWISS-PROT and EMBL) [8], there are many more data that are produced each day in the many lab oratories all over the world. These scientists create their own local databases of their newly discovered proteins and results, and are willing to share their findings to the world! Clearly, this is an application for P2P distributed data management systems for the same reasons as the health care application.

3) Data Caching

In the above two examples, each participant is actively involved in the process of consuming and supplying data. P2P distributed data management can also be deployed in passive nodes: nodes that are used to share resources (storage or computational power) on data that they may or may not be interested. Caching results from earlier queries is one such example - a node may have issued a query to some server (e.g., a data warehouse), the results of the query can be cached on the node (or some other neighbouring nodes). In this way, another node that requests for data that overlap the query result can potentially obtain partial answers quickly from this node, and the remainder from the original server. This also lightens the load on the original server. Indeed, Kalnis et. al. [5,8,9] have shown how distributed caching can be deployed in P2P environments to speed up OLAP queries.

D. Query Processing

This system gives the two mode processing approach i.e parallel processing approach and adaptive processing approach [1]. Query is submitted over normal peer using fetching and processing. The normal peer remote the subquery and the results are shuffled to the query submitting peer P [11]**.**

III. SYSTEM ARCHITECTURE

A. System Architecture

This system uses the pay-as-you-go business model popularized by cloud computing. By combining cloud computing, database, and peer-to-peer (P2P) technologies [6] containing a cloud development of Best Peer. At the last stage of its development, this System is improved with distributed access control, multiple types of indexes and pay-as-you-go query processing for deliver elastic data sharing services in the cloud [5]. The software components of the system are separated into two parts: Core and Adapter. The Architecture is shown in Fig. 2. The core having all function of data sharing and it shows that it is platform independent.

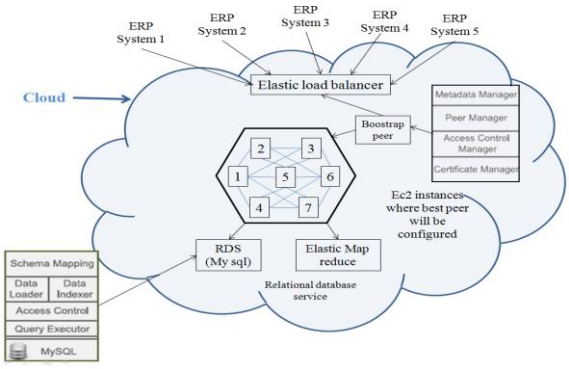

Fig. 2. The System Architecture

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

The adapter contains one abstract adapter for getting portable service edge and a set of real adapter components which by specific cloud service providers (e.g., Amazon). To achieve portability it build two level design, With appropriate adapters, this System can be portable to any cloud environments (public and private) or even non-cloud environment (e.g., on-premise data centre) [7]. The System can be implemented an adapter for Amazon cloud platform.

B. Components of the System

1) *Amazon Cloud Adapter:* The main approach of the system is to use committed database servers to store data for each business and arrange those database servers through P2P network for data sharing [3, 10]. The Amazon Cloud Adapter provides portable hardware transportation for system by using Amazon Cloud services.

2) *The Core:* This system having core contains all portable logic, containing query processing and P2P overlay [3]. The core consists of two parts of software i.e Bootstrap Peer, Normal Peer. In Bootstrap peer which is containing only single instances as service provider and having set of normal peer instances. It is considered as entry point of whole network [10]. In normal peer works under two mode i.e on-line mode and off-line mode. The data flow using normal peer in the system that id off line data flow and on-line data flow shown as Fig. 3: (a) and Fig.3: (b)

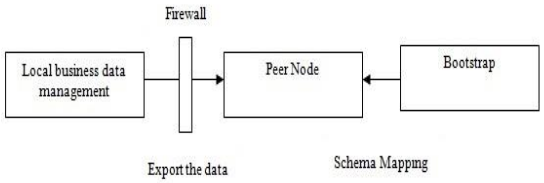

Fig.3: (a) Off-line Data Flow

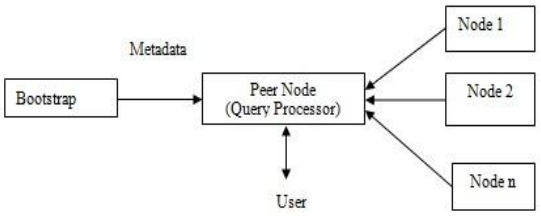

Fig.3: (b) On-line Data Flow

3) *Query Processor:* This system is design for achieve high query processing. These queries occupy querying a very small number of business partners and can be processed in small moment [3]. This system is mainly developed for these queries. For less time-consuming analytical tasks, they provide an interface for exporting the data from the system to Hadoop and allow users to analyse those data using Map Reduce [10].

C. Future Scope

Most of the enterprises are migrating their physical infrastructure to cloud based platform to reduce operational cost and achieve best performance of enterprise .This system will not be just a data sharing system in cloud but it will play an important role of plug and play adaptor to easily migrate the whole physical infrastructure in cloud.

IV. ADVANTAGES

1. This system can powerfully handle characteristic workloads in a shared network .
2. This system provide pay-as-you go model for business due to using of cloud computing, such as, what they use of system instance's hours and storage capacity they pay for it.
3. This system prolongs the role-based access control for the natural distributed environment of shared networks.
4. This system accepted P2P technology to retrieve data between employees in business model.
5. This system gives important solution to shared data in shared network of business model.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

V. EVOLUTION

This section shows evolution of the performance and throughput of existing system on Amazon cloud platform. For the performance benchmark, it evaluate the query latency of system with Hadoop DB using queries from typical shared network applications workloads as shown in Fig .4. For the throughput benchmark, it produces a simple supply-chain network consisting of suppliers and retailers and studies the query throughput of the system [8].

In Performance Benchmarking compares the performance of access system with HadoopDB. It chooses HadoopDB as benchmark target since it is an alternative promising solution for this problem and adopts architecture similar to it. Comparing the two systems (i.e., HadoopDB and Access System) reveal the performance gap between a general data warehousing system and a data sharing system specially designed to shared network applications [11].

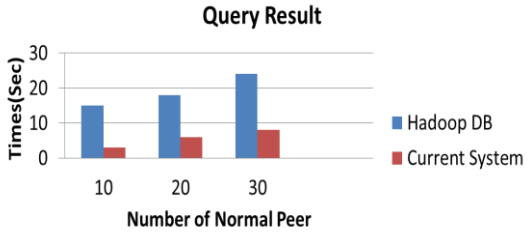

Fig 4. Comparison Between Access System with Existing System

In Throughput benchmarking the query throughput of Access system, HadoopDB is not designed for high query throughput, therefore, it intentionally skip the results of HadoopDB and only present the results of Access system [11]. It conducted two tiers of benchmark evaluation for the performance and scalability of Access system, respectively.

VI. CONCLUSION

This paper define limited challenges pose by contribution and generous out data in an inter-businesses environment and planned System, a system which developed portable data sharing services, by Combining cloud computing, database, and peer-to-peer technologies. This is created on Amazon EC2 cloud platform shows that our system can powerfully handle typical workloads in a corporate network and can move near linear query throughput as the number of normal peers grows. Therefore, This System having great capacity for sharing data within shared networks.

ACKNOWLEDGEMENT

We would like to thank the publishers for making their resources available and teachers for their guidance. We are also thankful to all authorities of Savitribai Phule Pune University for their constant guidelines and support. We are also thanks the college authorities for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

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. 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", VOL. 26,NO. 6, JUNE 2014.
3. 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.
4. 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.
5. 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

6. 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.
7. 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.
8. Beng Chin Ooi, Yanfeng Shu, "Relational Data Sharing in Peer-based Data Management Systems." *Kian-Lee Tan Sigmod Record special issue on P2P*, 2003.
9. 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).
10. G. Chen, H. T. Vo, S. Wu, B. C. Ooi, T. "A Framework for Supporting DBMS-like Indexes in the Cloud." *Ozsu* VLDB 2011.
11. Heng Tao Shen, Yanfeng Shu, and Bei Yu IEEE Trans. Knowl. "Efficient Semantic-Based Content Search in P2P Network." *DataEng.*16(7):813-826(2004)

BIOGRAPHY

Miss. Bhavsar Harshada V completed her B.E. in Information Technology from Babasahe b Ambedkakar MaratWada University, Pune and doing M.E. from Sharadachandra Pawar College of Engineering, Dumberwadi.

Prof. G. D Deokate currently working as an assistant professor in SPCOE, Dumbarwadi.

Dr. S.V.Gumaste, currently working as Professor and Head, Department of Computer Engineering, SPCOE-Dumberwadi, Otur. Graduated from BLDE Association's College of Engineering, Bijapur, Karnataka University, Dharwar in 1992 and completed Post- graduation in CSE from SGBAU, Amravati in 2007. Completed Ph.D (CSE) in Engineerng & Faculty at SGBAU, Amravati. Has around 22 years of Teaching Experience.