Storage Database System in the Cloud Data Processing on the Base of Consolidation Technology

Alexander V. Bogdanov[1,2], Thurein Kyaw Lwin[1], and Elena Stankova[1,2(✉)]

[1] Saint-Petersburg State University, Peterhof, Universitetsky pr., 35, 198504, St.-Petersburg, Russia
{alex,trkl.mm}@mail.ru, lena@csa.ru
[2] Saint-Petersburg Electrotechnical University "LETI", ul.Professora Popova 5, 197376, St.-Petersburg, Russia

Abstract. In this article we were studying the types of architectures for cloud processing and storage of data, data consolidation and enterprise storage. Special attention is given to the use of large data sets in computational process. It was shown, that based on the methods of theoretical analysis and experimental study of computer systems architectures, including heterogeneous, special techniques of data processing, large volumes of information models relevant architectures, methods of optimization software for the heterogeneous systems, it is possible to ensure the integration of computer systems to provide computations with very large data sets.

Keywords: Database system · Cloud computing · Cosolidation technology · Hybrid cloud · Distributed system · Centralized databases · Federated databases · IBM DB2 · DBMS · Vmware · Virtual server · SMP · MPP · Virtual processing

1 Introduction

Cloud computing and cloud storage systems have gained popularity as the most convenient way of transferring information and providing functional tools on the Internet. Some cloud services offer a wide range of services and functions to individual consumers (online shopping and online multimedia technologies, social networks, environment for e-commerce and protect critical digital documents), and the other commercial structures, that support the work of small and medium-sized businesses, large corporations, government and other institutions[1].

Some cloud services provide consumers with space for the storage and use of data for free, others charge a particular fee for services provided by subscription. There are also private clouds, owned and operated by organizations. In fact, this secure network is used for storing and sharing critical data and relevant programs [1]. For example, hospitals can use the sharing services for archiving electronic medical records and images of patients or create your own backup storage network. Moreover, it is possible to combine budgets and resources of several hospitals and provide them with a separate private cloud, which the participants of the group will enjoy together[2]. To create a private cloud requires hardware, software, and other tools from

O. Gervasi et al. (Eds.): ICCSA 2015, Part IV, LNCS 9158, pp. 311–320, 2015.
DOI: 10.1007/978-3-319-21410-8_24

different vendors. Management of physical servers at the same time can be both external and internal. Hybrid clouds, as is clear from the title, pool resources of different public and private clouds into a single service or a decision [3]. The basis of all cloud services, products and solutions are software tools that functionality can be divided into three types, means for processing data and running applications (server computing) to move data (network) and for storage (NAS).

The problem of Cloud use for computation stands aside both due to the large overheads in parallel libraries in the Cloud and what is more important serious limitations due to file systems peculiarities in virtual cluster architectures. Thus the way of transferring the data to virtual processes is of a vital importance for large scale computations optimization.

2 Cloud Computing on the Based of Consolidation Technology

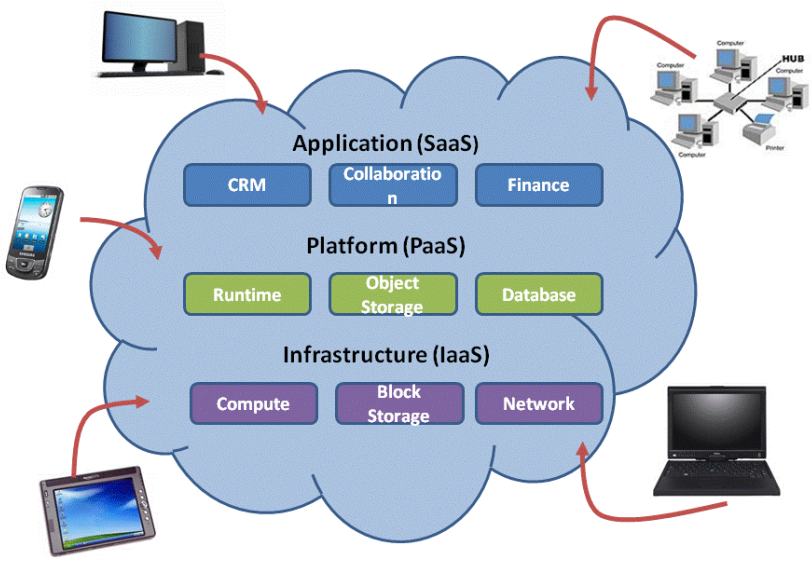

Fig. 1. Cloud Processing

The idea of "cloud computing" is to combine multiple computers and servers into a single environment designed to solve certain kinds of problems, such as scientific problems or complex calculations. Over time, this structure collects a lot of data, distributed computing and storage nodes. Typically, applications, running in a distributed computing environment, address only one of the data sources [4]. However, when the need arises to access simultaneously multiple sources, difficulties arise because these sources may contain different data and tools of heterogeneous access and also are

distributed at a distance from each other. In addition, for users performing an analysis of historical data, it is convenient to apply to a single source of information, forming a query and get results in the same format.

Thus, the main problem of the approach to the storage of information in distributed computing systems is the diversity and remote data sources. The solution is to create a centralized point of access, providing a single interface access to all data sources for cloud computing in real time. It is necessary to choose the most appropriate approach and the corresponding platform that provides a consolidation.

3 Data Consolidation Technology

All existing approaches to consolidate distributed data sources can be divided into two types:

1. The centralized approach

Data from all external sources are transferred to the central repository and are updated periodically. All users work directly with the central repository.

2. The federated approach

Data is stored directly in the sources, the central link provides transparent redirection of user requests and the formation of the results. In this case, all users can also refer to the central node only which translates requests more data sources.

Each of these approaches has its advantages, it is necessary to consider each and identify the most suitable for data consolidation in the cloud computing.

3.1 Architecture Centralized Databases

Centralized approach to consolidate distributed data sources is duplicating data from all sources in the central database. Such a database called a data store. Typically, the data warehouse using relational databases with advanced tools for integration with external sources. The availability of data combined in a single source, speeds up user access to data and simplifies the normalization and other similar processes compared with data scattered in different systems. However, the integration of information in a centralized source requires that data, which are often in different formats, were brought to a common format, and this process can lead to errors. Also for storage can be difficult to work with new data sources in unfamiliar formats. Moreover, the cost of treatment is often increased because of the need to duplicate the data processing and two sets of data[5].

3.2 Architecture Federated Databases

Federated databases access mechanism and management of heterogeneous data hides the features of the reference to a specific data source, but instead provides a single interface, similar to the classical relational databases [5].

Most applicable approach to creating a platform for the federated database approach is to develop the existing relational database management system and to ensure its interaction with external data sources. This database becomes central to a federal database that stores all the information about data sources, and redirects requests to it [5].

System database directory of the central node must contain all the necessary information about data sources in general and about each object in particular. Such information shall be used by the optimizer SQL-queries to build the most efficient query execution plan.

4 Comparison of Federal and Centralized Approaches

A feature of the federal database is a logical integration of data when the user has a single point of access to all the data, but the data itself physically remain in the original source. This feature is a key differentiator from the centralized federal approach that uses physical integration when data from disparate sources are duplicated at the general assembly that is accessed by all users. Federated approach involves storing data sources themselves, when the central node performs broadcasting requests, taking into account features of a particular source[5].

In the case of cloud computing, the federated database is a better choice for the following reasons:

1. Federated technology less prone to errors with the distortions and integrity because the data will remain in their original locations.
2. In a federated architecture easier to add new sources, this is especially important in dynamic systems.
3. The federated approach, as opposed to a centralized, always guarantees a real-time data from the original source, whereas the centralized approach transferring the data to a central site can become outdated.

It is worth noting that in complex cases that require large amounts of data intersection from different sources, federated database should provide the ability to store the information centrally, providing thus a hybrid approach [6].

5 Software Requirements for Federated Databases

Due to the heterogeneity and distribution of data sources in the cloud, unified information management environment is a challenge. Data sources can be relational databases, business applications, flat files, web services, etc. Each of them has its own storage format, challenges and way of delivery of results. Moreover, the sources may be located at a considerable distance from each other on different networks with different access protocols.

The software manages the federated database, must necessarily meet the following requirements:

1. Transparency
2. Heterogeneity

3. Scalability
4. Support for specific functionality
5. High performance
6. Separation of access rights

6 Existing Platforms of Federated Databases

1. IBM DB2 Information Integrator

This decision is based on DBMS IBM DB2 Universal Database and initially focused on the creation of distributed systems with federated access. Supported by a lot of variety of data sources, as well as standard SQL / MED, allowing you to create your own extensions.

Particular attention is paid to the performance and security platform, as well as ease of use and management.

2. Microsoft SQL Server

Integrating Microsoft SQL Server database with external sources is carried out through the use of Microsoft Integration Services - a platform to build integration solutions and data transformation at the enterprise level.

The Integration Services can extract and transform data from a number of sources, such as files XML, flat files, relational databases, etc. It is possible to use graphical tools of Integration Services to create a ready-made solutions or independent creation of the object model of Integration Services using the supplied software.

3. Oracle Streams

Integration of Oracle initially focused on the implementation of the approach with a centralized access, however, the technology Oracle Streams Transparent Gateways provides the means to implement the model with federated access. External data sources can be registered in the Oracle database in the form of links, called DB-links, and use the data from these sources in distributed queries. Supports access to flat files, XML-files, ODBC sources, etc[7].

7 Distributed Data Processing

Distributed data processing is an opportunity to integrate fragmented data resources. One approach to centralizing data is to decommission simply the existing database system and to build a new integrated database. An alternative approach is to build an integration layer on top of pre-existing systems. Building an integration layer on top of existing database systems is a challenge in complexity and performance, but this option sometimes gives the most effective results in business and engineering sense. In a data-sharing environment, there is no single best architecture that will solve all

problems. Large installations of database systems may be accessed by hundreds of thousands of times a minute. The irreducible latency present even in a fully optical network is not capable of supporting such a performance requirement. Indeed, local disks are also too slow, and most of this sort of information is cached off disk and into memory. In some organizations, if critical data is unavailable for even a matter of minutes, it could affect millions of dollars of revenue. This is why remote data access is not used in such large-scale situations where high availability is critical. There are many small and medium weight applications with modest performance requirements for data. Often, such applications are designed to work with a copy of data because getting a copy and loading it on a local database seems like the easiest solution. Such design does not factor in the cost of maintaining a separate copy of the data. When the applications are put into production and begin having problems keeping their data in sync, these costs become all too apparent. Such applications would probably do better to remotely reference their data. In such cases, it is a good architecture to remotely reference application databases for shared data. Such “distributed” databases need to incorporate some high availability design, depending on the weight of the applications served and their availability requirements. Each application should be analyzed to determine its performance and reliability requirements [1].

8 Distributed Databases are Working with Data on the Remote Server

Database management system (DBMS) has become universally recognized tool for creating application in software systems. These tools are constantly being improved, and the company database developers are closely monitoring the progress of their competitors, trying quickly to include in their packages the new features implemented in the competition. True internal architecture of the database is not always let to do this successfully. Distributed databases are implemented in a local or global computer network [7]. In this case, one of the logical database is located in different nodes of the network, possibly on different types of computers with different operating systems. Distributed DBMS provides users with access to information, regardless of what equipment and how the application software used in the network nodes. Members are not compelled to know where the data is physically located and how to perform physical access to them [8]. Distributed DBMS allows horizontal and vertical "splitting" of the tables and put the data in one table in different network nodes. Requests to the distributed database are formulated in the local database. Transaction processing operations and backup / restore distributed database integrity is ensured throughout the database.

We have been installing our database system on the server of linux platform in St.Petersburg State University GRID technoligies laboratory. The system includs a DB2 client and DB2 Administration Client, which implements the graphical tools that enable you to select the appropriate performance access to remote servers, to manage all servers from one location, to develop powerful applications and process requests. If the network is working properly, and protocols will function correctly on the workstation, the interaction of the "LAN - Local Area Network" between Database

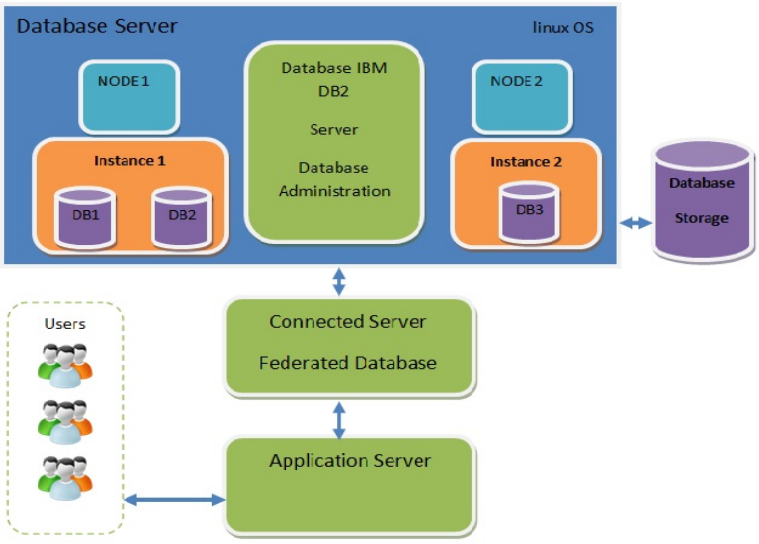

Fig. 2. Architecture, developed in Saint Petersburg State University

servers and clients require no additional software. As long as there is a connection between the local networks of any network client can access any server. Transactions provide access and update data in the databases of both servers while maintaining the integrity of data on both servers. Typically, such a mode of operation is called two-phase (two-phase commit) or access within a distributed unit of work. The first debit account and for the account of the second loan is very important that their update was carried out as a single transaction [8].

We can consider this solution as the DB2 database used for the consolidation of computing components and data storage in a virtual site. DB2 UDB is a completely parallel and support parallelized execution of most operations, including queries, insert, update, and delete data, create indexes, load and export data. Moreover, due to functionality of DB2, the transition of a standard, non-parallel execution environment to a parallel one is not limited by increasing efficiency [2]. DB2 UDB has been specifically designed to work successfully in a number of parallel media systems including MPP, SMP and MPP clusters of SMP nodes. DB2 provides computer data storage solutions for the target problem.

We choose DB2 UDB for our database system. We have installed IBM DB2 in the VMware Infrastructure environment. Then we installed VMware Tools in the guest operating system and created eight identical single-processor virtual machines. We have included only the virtual machines that are used in the particular test and made sure all unused virtual machines on the host ESX Server, were closed [8]. Results show, that one VCPU virtual machines advantages of working with IBM DB2 on a VMware ESX Server to become apparent when we run the test with multiple virtual machines.

We were modeling the virtual processor used in the test run with different load. In other words, the same number of concurrent virtual machines used, we have doubled the number of simulated users within two VCPU SMP virtual machines compared to one VCPU virtual machine. Figure-3 shows that the efficiency is almost doubled in two-VCPU SMP virtual machine. The results verify that the virtual environments can achieve the scale of SMP similar to that seen in the native environment.

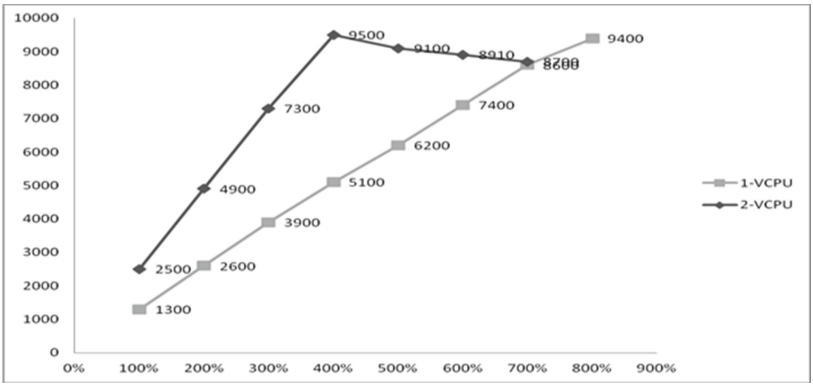

Fig. 3. Scalability compared for a fixed number of virtual machines

9 Analysis of the Database in the Distributed System

We were testing the distributed system in our university and have shown how to improve the performance and extend the range of applications of scientific methods and algorithms for parallel and distributed data processing by optimization of database applications from the point of view of search for promising architectural solutions[9]. The aim is to provide an operating environment for the database and its consolidation in a distributed computing environment, that is some general solution for relatively small networks and can be used in research institutions and commercial enterprises whose resources may be located in the same building and in geographically remote locations To achieve this goal it was necessary to solve fairly complex problem of choosing a prototype system architecture, algorithm development, as well as the problem of creating and adapting existing software products[10]. Such a system is implemented in the form of blocks that make up the distributed virtual computer system and we can call it virtual testbed.

10 Conclusions

Consolidation of data in distributed heterogeneous systems is an important and challenging task. Out of existing approaches to solving this problem, the most appropriate approach is that of federal databases. Creating and managing such a structure requires the use of specialized software, which in turn must meet a number of requirements for

transparency, heterogeneity, security, performance, etc. On market integration software there are a number of solutions from major manufacturers, build on industrial relational database, based on which you can organize a federal structure data access. To select a specific solution, the detailed examination for compliance with the requirements for systems of this type must be made. In this research we can point out the following results.

• There is a way to create an operating environment for database consolidation in a distributed computing environment. This environment is some common solution for relatively small networks and can be used in research institutes and commercial enterprises in which resources may be located in the same building as well as in geographically remote locations.

• A special synthetic method of data processing, which allows you to combine the power of SGE and DB2 DBMS for distributed heterogeneous computing resources, was an important step in the consolidation of the global pool of resources.

• The results of testing, which clearly showed that the databases in a distributed computing environment can be an effective means of consolidating software and the practical use of the developed techniques can significantly improve the efficiency of data processing and improve the scalability of distributed systems.

• The proposed methods and products have been tested on a variety of platforms and operating systems, and we hope, that they will be widely used not only in distributed computer systems, but also for cluster computing.

We listed the requirements for the cloud databases and compared the suitability of different database architectures to cloud computing. Based on our performance results, we believe that the Cloud Database vision can be made a reality, and we look forward to demonstrating an integrated prototype of next Big Data Solution. Whether we come to assembling, managing or developing of a cloud computing platform and need a cloud-compatible database is a challenge.

References

1. Bogdanov, A., Lwin, T.K.: (Myanmar)/ Storage database in cloud processing. In: The 6th International Conference “Distributed Computing and Grid-technologies in Science and Education” will be held at the Laboratory of Information Technologies (LIT) of the Joint Institute for Nuclear Research (JINR) on 30 June - 5 July 2014 in Dubna
2. Technical Details Architecture cloud data processing and storage . http://www.seagate.com/ru/ru/tech-insights/cloud-compute-and-cloud-storage-architecture-master-ti/
3. Bogdanov, A.: Private cloud vs Personal supercomputer, Distributed computing andGRID technologies in science and education, JINR, Dubna, 2012 Advanced high performance algorithms for data processing, - Computational science ICCS-2004, Poland
4. Bogdanov, A.: The methodology of Application development for hybrid architectures. Computer technologies and applications (4), 543 – 547 (2013)
5. Bogdanov, A.V., Stankova, E.N., Lin, T.K.: Distributed databases. SPb.: ``LETI'', pp. 39–43 (2013)

6. Bogdanov, A.V., Kyaw lwin, T., Stankova, E.: Using Distributed Database System and Consolidation Resources of Data Server Consolidation. In: Computer Science and Information Technology International Conference, pp. 290–294. Armenia, September 23–27, 2013
7. Bogdanov, A.V., Lwin, T.K., Naing, Y.M.: Database Consolidation used for Private Cloud. In: Proceedings of the 5th International Conference GRID 2012, Dubna, July 16–21, 2012
8. Bogdanov, A.V., Lin, T.K.: Database technology for system integration of heterogeneous systems and scientific computing. Proceedings of ETU ``LETI'' (4), 21–24 (2012)
9. Babu, S.: Automated control in cloud computing: challenges and opportunities. In: Babu, S., Chase, J., Parekh, S.(eds.) 1st workshop on Automated control for datacenters and clouds, pp. 13–18 (2009) (DOI:10.1145/1555271.1555275)
10. Armbrust, M., Fox, A., Griffith, R., et al.: A view of cloud computing. Communications of the ACM **53**(4), 50–58 (2010)