Using distributed database system and consolidation resources for Data server consolidation

Alexander Bogdanov [1]
[1] High-performance computing Institute and the integrated systems, e-mail: bogdanov@csa.ru, Saint-Petersburg, Russia

Thurein Kyaw Lwin[2], **Elena Stankova**[3]
[2,3]St.Petersburg State University, e-mail: trkl.mm@mail.ru, lena@csa.ru, Saint-Petersburg, Russia

ABSTRACT

This paper describes the consolidation of large distributed computational complexes on the basis of database. The organization of access to the distributed computing resources, especially in the solution of the large complex problems, is a challenge for any general purpose computer centre. Consolidating data in distributed heterogeneous systems is an important and challenging task. The existing approach to solving this problem is the most suitable approach to the organization of federal databases.

1. INTRODUCTION

Data consolidation is a strategically important task for effective management of corporate and scientific information within the distributed computing environment. Due to the continuous increase complexity of applications and volumes of data generated, the company spends on such tasks more money, time and effort. Analysts at Gartner estimated the market size of integration at the end of 2007 to be 1.44 billion dollars, with annual growth of more than 17%.

The main reason for consolidation is the geographical distribution of data sources, as well as their syntactic, systemic, semantic and structural heterogeneity. Funds Consolidation used in corporate enterprise systems, with integration of heterogeneous information sources to support applications, business analysis, forecasting and management, as well as in distributed scientific systems for sharing access for knowledge bases and research findings. Data Consolidation covers practices and architectural approaches and software tools to ensure having consistent access and delivery of data for all range of applications and business processes. Therefore, the strategy and program tools used for consolidation fully depend on the characteristics of each particular system. The purpose of this study is to examine the software and approaches to overcome the problems associated with geographic remote data sources, as well as their syntactic and systemic heterogeneity. The problems of semantic and structural heterogeneity can also be solved through the creation of ontology and patterns of compliance using the methods of data cleaning that goes beyond the scope of this article.

The cloud computing is bringing together multiple computers and servers in a single environment designed to address certain types of tasks, such as scientific problems or complex calculations. This structure builds up a lot of data, distributed computing nodes and storage. Typically, applications executed in a distributed computing environment, apply to only one data source. However, when simultaneous access to multiple sources is required, difficulties arise because these sources may contain heterogeneous data and methods of access, and located at a distance from each other. In addition, for users engaged in an analysis of accumulated data, it is convenient to apply to a single source of information, creating queries and get results in the same format[1]. Thus, the main approaches to the problem of storage of information in distributed computing systems are heterogeneous and remote data sources. The solution is to create a centralized access point, providing a single interface to access all sources of data computing clouds in real time. You must select the most appropriate approach and relevant platform that provide consolidation.

2. Consolidation Technology

All existing approaches to the consolidation of distributed data sources can be divided into two types of Centralized approach and Federated approach.

2.1. Architecture of centralized databases

The centralized approach to the consolidation of distributed data sources is duplication of data from all sources in the central database. Such databases are called data stores. Usually the data warehouse used by a relational database with advanced tools for integration with external sources[2]. Availability of data combined in a single source speeds user access to data and facilitates normalization and other similar processes compared to

the case of data scattered in different systems. However, integration of information in a centralized source requires the data that are often in different formats, are reduced to a single format, a process that can lead to errors[2]. Also for the repository it can be difficult to work with new sources of data in unfamiliar formats. Moreover, the processing costs are often increased because of the need to duplicate data and process the two sets of data.

2.2. Architecture federated databases

Federated database - a mechanism to access and manage heterogeneous data, hiding from user a particular data source, but providing a uniform interface instead, similar to the classical relational database. The most applicable approach to creating a platform for federated database is an approach to develop the existing relational database management system to ensure its interaction with external data sources. This database is a central node of a federal database that keeps all the necessary information on the sources of data, and forwards requests to the sources of their parts[2]. System database directory of the central node should contain all necessary information about data sources in general and on each of the objects in particular[2]. Such information should be used by the optimizer of SQL-queries to build the most efficient query execution plan.

2.3.Comparison of federal and centralized approaches

Feature of federal databases is a logical integration of data when the user has a single point of access to the totality of data, but the data itself physically remain in the original source[2]. This feature is a key difference from the centralized approach that uses physical integration, where data from disparate sources are duplicated on a common node that is accessed by all users. The federated approach involves storing data in their respective sources, while the central node performs the query translation, taking into account the characteristics of the source[2].

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

1. Federated technology is less prone to distortions and integrity errors, because the data remain in their original locations.

2. In federated architecture it is easier to add new sources, which is especially important in dynamic systems.

3. The federated approach, in contrast to a centralized, always guarantees the receipt of actual data from the primary source, whereas in the centralized approach, a copy of the data in the central site may become outdated.

It should be noted that in complex cases that require the intersection of large data sets from different sources, federated database must provide the ability to store the information centrally, thus ensuring a hybrid approach.

3. Forms of Consolidation

Database consolidations can take many different forms. A physical consolidation focuses on reducing the number of physical servers, disk storage, database instances and databases. Geographic consolidation involves centralizing servers in one location. Logical consolidation entails centralizing applications or data by their business functionality. And vendor consolidation trims the number of database suppliers. Database consolidation projects can be triggered by several different factors. Often, an enterprise will have acquired new applications and databases through mergers and acquisitions. Or it may have multiple versions of a database purchased over time. Finally, when a company opts to re-architect its underlying infrastructure, perhaps moving to a grid infrastructure, for example, database consolidation may be appropriate[3].

A database consolidation project is not a trivial task. Like other major IT projects, a database consolidation project has six phases--analysis, design, development, test, implement, and monitor. The goal of the evaluation is to determine the performance of the existing infrastructure, assess which parts of the infrastructure should be retained, and develop a blueprint for the new architecture. Close cooperation will insure that the consolidation project will achieve its goals. Once the blueprint is developed, the hardware infrastructure must be configured and the databases migrated to the new platform. Finally, the applications must be moved.

In a database consolidation project involving heterogeneous databases, maintaining existing applications can represent a potential hidden cost. Each major database vendor uses its own version of SQL, and reworking existing applications can represent as much as 30 percent of a database consolidation project[3]. Clearly, consolidating on a universal database platform with multiple language compatibility offers significant advantages. It can radically decrease the cost of application maintenance and cut the time needed to complete the consolidation project.

4. Consolidating Servers

Server consolidation is a big topic for data center managers. Server consolidation is an approach to the efficient use of computer server resources in order to reduce the total number of servers or server

locations that an organization requires. The practice was developed in response to the problem of server sprawl, a situation in which multiple, under-utilized servers take up more space and consume more resources than can be justified by their workload[4]. Servers are still the primary focal point for consolidation because they are so obvious. Whether you have 100 servers or 5000 servers, you probably have too many to manage effectively. Today's distributed computing environment lends itself to a proliferation of servers. Reducing and controlling the number of devices to manage and simplifying ways to manage them is the goal of most IT groups[5].

4.1. Identifying Patterns in an End-to-End Architecture

The end-to-end architectures that are prevalent today, tiers of servers are specialized for particular tasks. When you look at consolidating servers, you need to look for patterns in your server population. When you identify these patterns within tiers, you can start to devise a consolidation strategy[5]. Scalability is the key, here. Because you are expected to deliver predictable service levels in response to unpredictable workloads, it is important that you use the right type of scalability for each part of a consolidated architecture. The following sections describe common patterns in an end-to-end architecture. For consolidation discussions, we generally assume that there are three server types, or tiers:

- The presentation tier is the closest tier to the end user.
- The business, or middleware, tier is where applications or middleware run in conjunction with the other tiers.
- The resource tier is where large, scalable servers run mission-critical applications and databases.

Although architectures with these characteristics have been around for a while, most corporations still have many servers running monolithic applications. In many cases, these are older servers running mature applications[5]. These servers are generally excellent candidates for server and application consolidation.

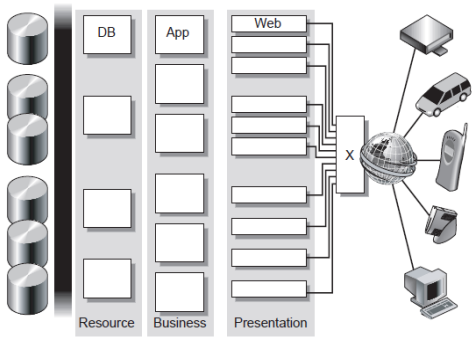

Figure -1. End- to- End Architecture

5. Data Consolidation

Data consolidation is the main approach, which uses data warehousing applications for building and maintaining operational data warehouses and enterprise storage[5]. Consolidation of data can also be used to create the dependent data marts, but in this case, the process of consolidation is only one data source (e.g., enterprise storage). In the data warehouse environment one of the most common technology is ETL (extract, transform, and load - extract, transform, and load)[5]. Another common technique of data consolidation is content management (enterprise content management, abbr. ECM). Most ECM solutions aimed at consolidating and managing unstructured data such as documents, reports and web-page.

6. Consolidating Data Centers

Many organizations are looking to consolidate multiple data centers into one site. These consolidations range from simple city-wide consolidations to complex region wide consolidations[5]. Shutting a data center is a huge task, and before you even start down the path, it is vital that you can articulate and defend your reasons for doing it. Further, once a data center is shut down, the costs of reopening it can be enormous. From there, data center consolidations are similar to other types of consolidation, except that assessment (especially application, networking, and physical planning) and implementation become much more complex[6].

7. Optimization

As networks, applications, and services grow more complex and users expect to conduct unified communications without a compromise in functionality or performance, a company’s distributed legacy infrastructure is hard-pressed to withstand the strain. Toss in the occasional corporate merger or acquisition that expands the enterprise and ratchets up network and application disparity and the situation borders on untenable. Consolidation promotes several

avenues to optimization. One of them is the aforementioned transport. With a more centralized approach, there are fewer “pipes” to monitor, the architecture is more straightforward and easier to control, and traffic patterns and volumes are more visible and clearly defined. This environment offers the option to implement more advanced protocols and management strategies that maximize bandwidth utilization and performance of the overarching network and its applications. Data center consolidation also goes hand-in-hand with application virtualization. The objective of application virtualization is to segregate applications from servers. Instead of running on a physical server with which it is co-located, an application executes on a virtual server which can reside anywhere in the enterprise, such as in the consolidated data center[7]. As a result, fewer physical servers are needed, because each is multi-tasked to handle many applications, each of which performs as if the server were dedicated to it. When properly planned and maintained, the adoption of shared services is transparent to the end users of the applications, yet delivers a more manageable quality of service. Automation solutions in the datacenter, for example, can restart failed applications, dynamically allocate new servers, conduct scheduled backups, and perform configuration management of the operating environment. Automation brings a number of advantages, including process consistency and enforcement of corporate rules and regulations, accelerated process execution, and minimization of human error. It also allows for more efficient adaptation to changing conditions, and it increases the productivity of the IT and operations teams whose manual input and support for the automated processes and systems is no longer required.

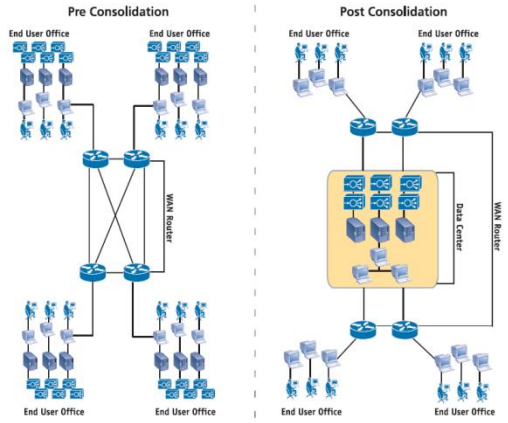

Figure 2 – Datacenter consolidation is necessary not only to simplify the infrastructure, but to optimize it so quality of service can be maintained and ultimately improved.

Conclusion

Consolidating data in distributed heterogeneous systems is an important and challenging task. The existing approach to solve this problem is the most suitable approach to the organization of federal databases. Creating and managing such a structure requires the use of specialized software, which in turn must meet several requirements for transparency, heterogeneity, security, performance, etc. In the market integration software there are a number of solutions from major manufacturers, based on industrial relational DBMS, you can use to organize a federal structure of data access. From a technical point of view, data integration has traditionally submitted to a centralized repository and tools Extract, Transform and Load (ETL). However, the main disadvantages of this approach is the large overhead storage information and delays in receipt of information. With the increasing number consolidated sources deny overhead grows proportionally. This study shows that within distributed computing systems, huge amount of applicable federal approach does not require integration of all data in a single source. The author also explains why this approach is more flexible and provides the fastest way to connect new sources of data, which is particularly important in dynamic changing systems. Also this study describes a hybrid approach that combines benefits of the approaches to the repository and federated access to data. In this case, part of the data is replicated to a central database, and part of it is still stored in the original sources, depending on the type and strategy used. The leaders in this area are IBM and Informatics, providing comprehensive support for solving problems of data consolidation. Finally, it is worth noting that despite of rapid growth, it remains many unsolved problems that require thorough study and new solutions. The process of consolidating data in this phase of development technology requires a large amount of manual work overcome semantic structural discordant and setting performance. Therefore, in the near future efforts of companies, producing many of consolidating data will be used to increase the level of automation and self-management of their products.

References

1. Alexander Bogdanov, Thurein Kyaw lwin// System integration of heterogeneous complexes for scientific computing, based on the use of DB2 technology //Proceedings of International Conference «Computer Science & Information Technologies», 28 September - 2 October, 2009, Yerevan, Armenia, pp.397-399

2. A.B.Vogdanov, Thurein Kyaw Lwin, Anatoly Shuvalov// Консолидация данных в системах распределенных вычислений (бэта-версия)// Consolidating data in distributed computing systems

3. Database Trends and Applications, Solution for the information Project Team, Volume21, Number 5//www.dbta.com.

4. Высокопроизводительные вычислительные алгоритмы (учебное пособие)/ А.В.Богданов,

М.И.Павлова, Е.Н.Станкова, Л.С.Юденич/ http://www.csa.ru/old/analitik/distant/q_start.html

5. Consolidation in the datacenter, David Hornby, Global Sales Organization Ken Pepple, Enterprise Services Sun BluePrints™ OnLine—September.

6. Ken Milberg Planning Your Data-Center Consolidation Strategies for a hassle-free deployment//www.ibmsystemsmag.com/aix/enewsletterexclusive/25075p1.aspx

7. Ensuring the Success of Datacenter Consolidation over the Long Haul / Fluke white paper //www.flukenetworks.com /ensuring the success of datacenter Consolidation