Application of Data Warehouse Technology in Data Center Design

Hu Xuanzi
Computer Engineering Department
Dongguan Polytechnic College
Dongguan City, China
huxuanzi@126.com

Wang Kuanfu
Computer Engineering Department
City College
Dongguan University of Technology
Dongguan City, China

Abstract

Chinese electronic government (E-Government) has achieved great success in several years, to quicken the pace of building E-Government, the central government of China has programmed to establish four governance information resource databases in next five years, data center construction in the developed cities is becoming an important project for present China's E-Government. Through investigation and analysis requirements of Nanhai city for data and business, this paper presents the architecture of data center, which is composed of six main components: data share and exchange platform, kernel database, support application platform, application database, data center management platform, and data center security platform. Extensible markup language (XML) and data warehouse technology are adopted for data center. Proposed method has already been successfully used in data center of Nanhai city, which is listed as the pilot cities for Model Project of China's E-government Application.

1. Introduction

E-government allows governments to service citizens in a more timely, effective, and cost-efficient method. The development of an E-government system has already been very popular all over the world, because E-government helps to disseminate information. Further, it aids in the collection of information that helps decision makers serve citizens more effectively. E-Government allows government agencies to centralize decision making.

China have started first E-Government program in the late 1980s, in which the governments both at central and local levels built up office automation (OA) systems and established an intranet, subsequently the Central Government of China had formally launched five Golden Projects (Golden Bridge Project, Golden Customs Project, Golden Card Project, Golden Tax Project and Government online Project) aimed at building E-Government in China ever since 1990s. After realizing five Golden Projects, Chinese government has set ambitious visions in the implementation of E-Government: quicken the pace of change in government functions to suit the requirement of reform, opening up and modernization policies, improve the performance of government operation, introduce new government measures in a scientific manner and more effective mechanisms to monitor the economic activities, place a greater emphasis on central co-ordination and transparency of government work, carry out administrative functions in accordance with law and provide better service for the public. To meet above ambitious visions, the Central Government of China have programmed to establish four governance information resource databases in next five years, governance information resource databases consist of four databases: population basic information database, judicial entity basic information database, natural resource and geography basic information database, and macroscopically economy database [1][2].

A data warehouse has been defined as a collection of data in support of management decisions which is: subject oriented, integrated, nonvolatile, time variant. Data warehouse, as a collection of database or data management technologies, emerged in the early 1990s. The data warehouse has now been more generally seen as a strategy to bring heterogeneous data together under a common conceptual and technical umbrella and to make the data available for new operation or decision support application. Three intrinsic features of data warehouse are data integration, data completeness and decision-making support [3][4].

Management of government is from top to down, but collecting data of government is from down to top. To support construction of four databases, building data center is becoming an important project for E-government. Nanhai city of Guangdong province is a leading the way of E-government in China, this paper put forward the design of data center based on data

978-0-7695-3508-1/08 $25.00 © 2008 IEEE
DOI 10.1109/CIS.2008.124

warehouse technology for Nanhai city, this solution not only consider to provide data for high level four databases, but also solve integration, share and exchange of data in various departments of Nanhai city, especially devised application database based on data warehouse technology for better utilizing accumulated data.

The remainder of this paper is organized as follows. In Section 2, we briefly describe requirement analysis of data center. In Section 3, we give the architecture of data center based on data warehouse technology. Section 4 gives some pivotal techniques of realizing data center. Finally, Section 5 gives the conclusion.

2. Requirement Analysis

Nanhai city of Guangdong province is the state-level pilot city for information-based. Consequently Nanhai city is listed as the pilot cities for Model Project of China's E-government Application. Nanhai city has successfully programmed and developed a series of E-government projects, including village management, finance decision-making, education, irrigation management, soil management, police management, etc, which vigorously promoted information-based construction in such fields as governments, rural areas, education, culture. With many system applications in every department, data exchange among departments, data share among departments and data integration application are becoming a big demand for improving greatly management efficiency and service standards. The relations of data provider and user for population data and for judicial entity data are depicted in Figure 1 and Figure 2, respectively.

To solve data share and exchange among departments, establishing data center is a prime method, main requirements of data center are summarized as following:

Realizing data share and exchange in the various departments, supplying data collecting, processing and loading from data sources to data center, achieving centralized storage of government information resource, in favor of the higher level construction of four databases, offering integration application service for government, enterprises and citizens, Offering integration management for population information.

3. The Architecture of Data Center Based on Data Warehouse Technology

The architecture of data center based on data warehouse is presented in Figure 3. Data center is composed of six main components: data share and exchange platform, kernel database, support application platform, application database, data center management platform, and data center security platform.

The very essence of the data warehouse is the flexible and unpredictable access of data [5][6]. Thus, required is the ability to access data quickly and easily. If data is not efficiently indexed and users cannot access data rapidly, the data warehouse will not succeed. In addition, the data in the data warehouse needs to be able to be monitored at will. The cost of monitoring data cannot be so high and the complexity of monitoring data cannot be so hard that a monitoring program cannot be run whenever necessary. The data warehouse also needs to be able both to receive data from and pass data to the various departments of Nanhai city.

3.1. Data Share and Exchange Platform

Data share and exchange platform is made up of ETL (extraction, transformation and load) and data share agent. ETL is data provider for kernel data, which collects data from data source and load to kernel data. Data share agent deal with dispensing share data stored in kernel database into various departments that need data.

3.2. Kernel Database

Kernel database of data center is composed of four parts: population basic information database, judicial entity basic information database, natural resource and geography basic information database, and macroscopical economy database, which is organized according to user requirements and is maintained by administrators of data center.

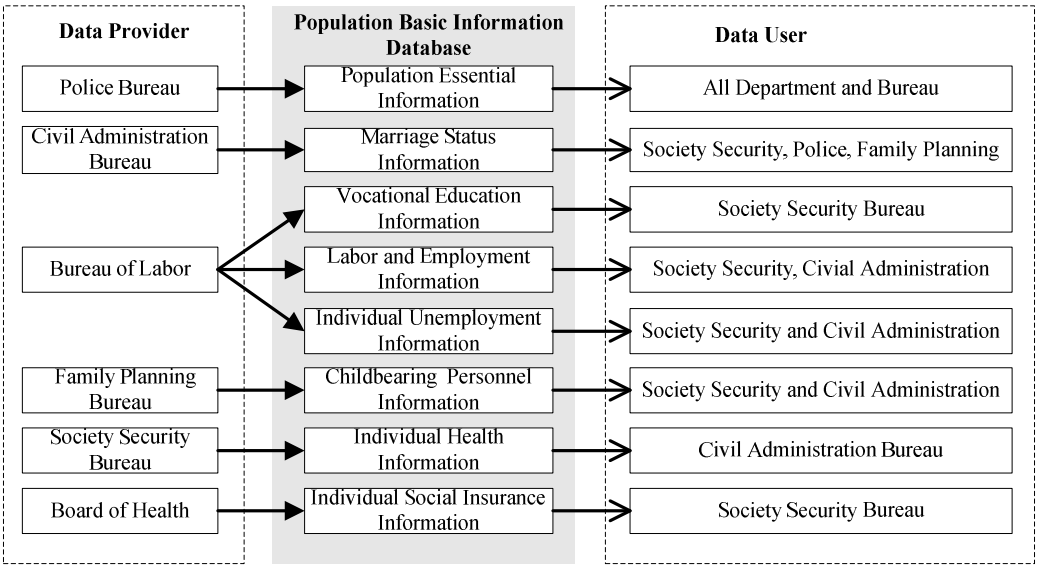

Figure 1. Relation of Provider and User for Population Data

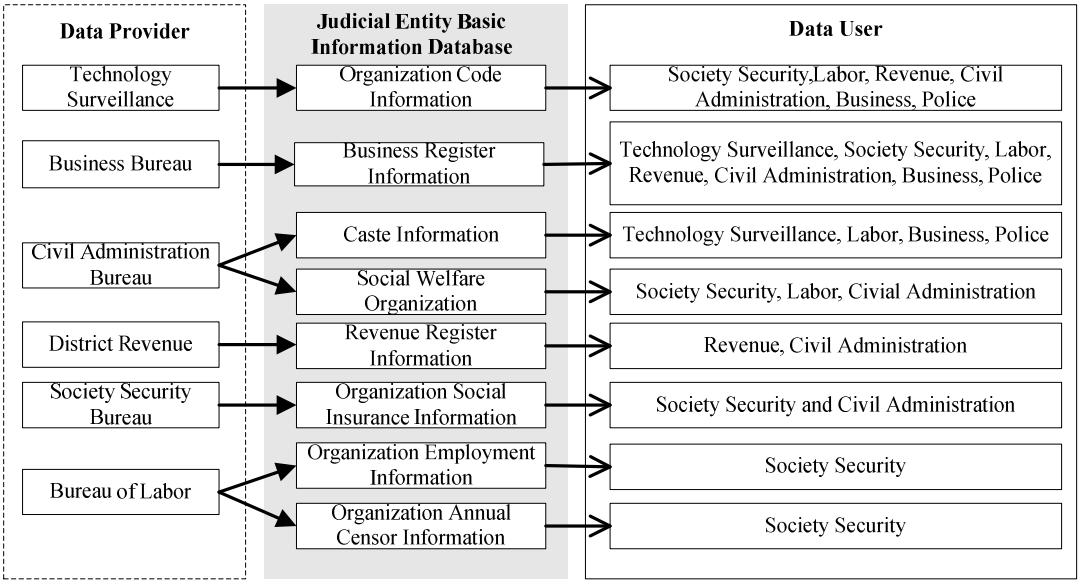

Figure 2. Relation of Provider and User for Judicial Entity Data

3.3. Support Application Platform

Support application platform is a secondary development tool possessed by data center, main function of which is establish special database by extracting data directly from kernel database according requirements of application.

3.4. Application Database

Application database may be treated as data mart, which is composed of three parts: public service database, decision support database and special application database. Data Special application database is provided by kernel database and source database according to application requirement. Application database is organized according to requirements of decision-making, which is maintained by both administrators of data center and department.

3.5. Data Center Management Platform

Its main function is to manage and control data center, including share management, exchange management, run management, log management, authorization management, backup management and recovery management.

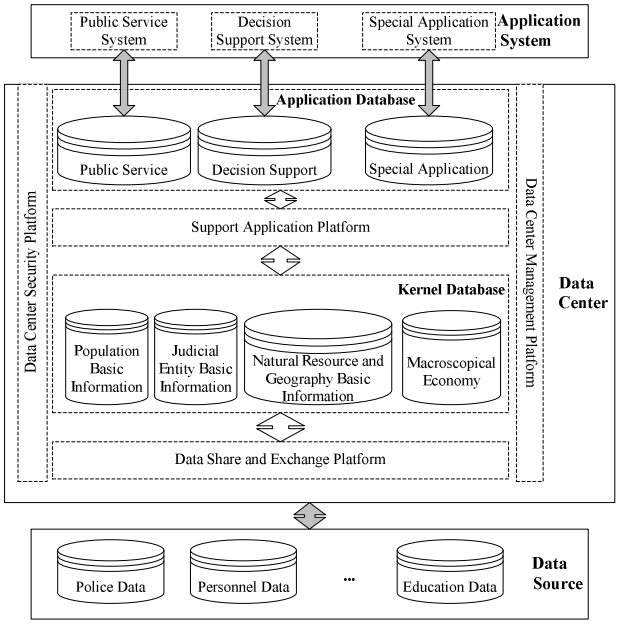

Figure 3. Architecture of Data Center

3.6. Data Center Security Platform

To deal with alarming and unpredictable security threats, data center must consider security. Data center Security Platform is a base of other platform that monitors and protects data center.

4. Pivotal Techniques

4.1. ETL Technique

The process of extracting data from data sources and bringing it into the kernel database is commonly called ETL, which stands for extraction, transformation, and loading. During Extraction, the desired data has to be identified and extracted from data sources. Very often, it is not possible to identify the specific subset of interest; therefore more data than necessary has to be extracted, since the identification of the relevant data will be done at a later point in time. The size of the extracted data varies from hundreds of kilobytes to hundreds of gigabytes, depending on the source system and the organization situation. Just as the size of the data extraction may vary widely, the frequency at which the data is extracted may also vary widely: the time span may vary between hours and minutes to near real-time [3][4][5]. After extracting (and transporting) the data, the most challenging and time consuming parts of ETL follow: Transformation and Loading into the target system. This may include applying complex filters; validating the incoming data against information which already existing in target database tables; comparing new data to existing data in the data warehouse, to determine whether the new data needs to be inserted or updated; computing aggregations and other derived data based on the new data. Generally there are three kind approaches of ETL: transformation-then-load, load-then-transformation and transformation-while-load.

4.2. Data Storage Technique

Data storage techniques is very important and complicated for realizing goal of data center, to support future decision support system, star schema for the warehouse is adopted to build application databases. Every application database is composed of several fact tables and a set of dimensional tables, the fact table contains a list of all measures and points to the key value of the lowest level of each dimension. Each of these measurements is taken at the intersection of all dimensions. Dimensions are qualifiers that give meaning to measures. They organize the data based on the what, when, and where components of a organization question. Dimensions are stored in dimension tables made up of dimensional elements and attributes. Each dimension is composed of related items or elements. Dimensions are hierarchies of related elements. Each element represents a different level of summarization.

Choosing the appropriate fact measures for the grain in the fact table depends on the organization and analysis purposes. For example, the star schema for the contract data is constructed as shown in Figure 4.

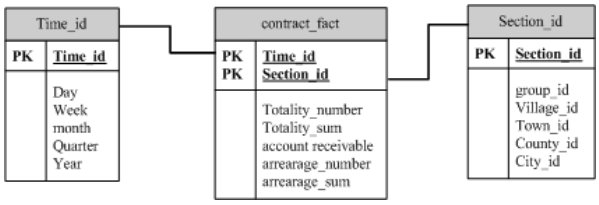

Figure 4. Star Schema of Contract Data

4.3. Data Share and Exchange Technique

Data exchange is the problem of finding an instance of a target schema, given an instance of a source schema and a specification of the relationship between the source and the target. Such a target instance should correctly represent information from the source instance under the constraints imposed by the target schema, and should allow one to evaluate queries on the target instance in a way that is semantically consistent with the source data. XML stands for extensible markup language. XML was released in the late 90's and received a great amount of application. The XML standard was created by World Wide Web Consortium to provide an easy to use and standardized way to store self-describing data. The main benefit of XML is that you can take data from a platform, convert it into XML, and then share that XML with other platforms. Each of these receiving platforms can then convert the XML into a structure the platform uses normally and you have just communicated between two potentially very different platforms! So XML is adopted to realize data share and exchange. Data stored in the data source is converted into XML file, and then send it to data share and exchange platform, when receive XML file, data share and exchange platform again convert XML to structure data of kernel database, data stored in the kernel database is sent to application department in the same way, therefore, function of data share and exchange is realized easily.

5. Conclusion

In this paper, relation of provider and user for data in various departments of government is given in detail. On the base of data analysis, the architecture of data center based on data warehouse technology has been presented. In our proposed method, XML technology is used for data share and exchange. The proposed method has already been successfully used in development of data center system for Nanhai city.

6. References

[1] Donna Evansa, David C. Yenb. E-Government: Evolving relationship of citizens and government, domestic, and international development Government Information Quarterly 23, Pp. 207–235, 2006.

[2] State Information Construction Promotion Office (SICPO) (2001). Report on China's Internet Resources. July 2001. http://www.cei.gov.cn.

[3] Lijuan Zhou, Chi Liu, Chunying Wang. The Design and Application of Data Warehouse during Modern Enterprises Environment. *Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security*. 2006.

[4] Xiaohua Hu, Nick Cercone. A Data Warehouse/Online Analytic Processing Framework for Web Usage Mining and Business Intelligence Reporting. *International Journal of Intelligent Systems*. Vol. 19, Pp. 585 – 606, 2004.

[5] Dr. Katherine Jones. An Introduction to Data Warehousing: What Are the Implications for the Network? *International Journal of Network Management*. Vol. 8, Pp. 42–56, 1998.

[6] Keqin Wang, Shurong Tong and et al. Fourth. Review on Application of Data Mining in Product Design and Manufacturing. *International Conference on Fuzzy Systems and Knowledge Discovery* (FSKD 2007).

[7] Alex A. Freitas, Jon Timmis. Revisiting the Foundations of Artificial Immune Systems for Data Mining. *IEEE Transactions on Evolutionary Computation*, Vol.11, No.4, August, 2007.