A DATA WAREHOUSE SOLUTION FOR E-GOVERNMENT

Xiufeng Liu[1] & Xiaofeng Luo[2]

[1] Department of Computer Science Aalborg University, Selma Lagerlofs Vej 300, DK-9220 Aalborg , Denmark
[2] Telecommunication Engineering Co.Ltd., Meizhou City, Guangdong Province, China, 514000

ABSTRACT

The *eGovMon Data Warehouse* (eGovMon DW) is built as a data repository for eGovernment services benchmarking results. We propose a DW architecture with open source business intelligence technologies for eGovernment. This DW architecture uses PostgreSQL as the DBMS, eGovernment operational system as the data source, and a right-time ETL tool to populate the data. Through this proposal, we give the potential research interests and issues for our future work.

Keywords: *eGovMon, Data Warehouse, ETL, DW architecture*

1. INTRODUCTION

The term of *Business Intelligence* (BI) is defined as the technologies, applications and practices that collect, integrate, analyze and present large amounts of data from data warehouse and extract useful knowledge from it [4]. The objective of using BI is to improve the timeliness and quality of the input to the decision process. For example, the world's largest retailer - Wal-Mart use their business intelligence system to attain the products in stock, predict retail marketing trends and make marketing decisions. A Data Warehouse (DW), the information source and basic of business intelligence, is defined as "a subject-oriented, non-volatile, time-variant repository" [2]. It provides a single platform for an enterprise to make a plan and decision on the data analysis of integrated history data.

The technologies of business intelligence encompass a series of tools, architectures, information services and communication infrastructures that are useful for the data integration and analysis from heterogeneous data sources. Much often the business intelligence is characterized by the widely use of business intelligence tools. In the commercial world, business intelligence tools and technologies have come to their maturity after years' development. Many BI vendors not only give their own business intelligence tools but provide their business intelligence solutions as well, such as IBM, SAP, Business Objects, Cognos etc. However, in contrast with the commercial world, the use of business intelligence tools is still not very common in open source area, and not to mention complete business intelligence solutions. Although an increasing number of open source tools have been developed in recent years, most of them are still used in stand-alone in software development, other than in a full BI solution. This might be due to the reasons that, unlike the commercial tools having the full features required by the business intelligence and providing good technical support, many open source tools cannot be put into use immediately, instead, they might need to be intimately tailored and customized to the distinctive needs and requirements of the business. The main reason, however, is that open source business intelligence is still at the beginning phase of its development so that it still need several years to grow into maturity. Nevertheless, the benefit of open source business intelligence is obvious, such as its low-cost, manageable size, flexibility and reducing the dependence on software vendors, which make it to be a valuable alternative to the traditional business intelligence solutions. The investigation in section 2.1 evidences the open source tools are in bloom over the past few years. We can expect that in the next few years this trend will continue and will concentrate more research interests and efforts in this area [3].

In this paper, an open source business intelligence solution is proposed for the data warehouse of eGovMon project [1]. The construction of this data warehouse is to store the benchmarking eGovernment services results in four observatories: *Accessibility, Transparency, Efficiency and Impact* (ATEI). The ultimate aim of this system is to help governments deliver services to citizens in better, improve interactions with business and industry, and better government management. This project is also the continuous work of the other project - European Internet Accessibility Observatory (EIAO), which evaluates the accessibility of European websites.
The remainder of this paper is structured as follows. Section 2 reviews the open source business intelligence tools and latest research status in this area. Section 3 gives the overview of the proposed data warehouse architecture and details every component in the architecture. Section 4 analyzes potential research interests and issues for open source data warehouses. The final section summarizes this paper and outlines future work.

2. RELATED WORK

2.1. Investigation of Open Source BI Tools

For the aim of finding out the current available business intelligence tools and their development status in open source market, Thomsen and Pedersen conducted two investigations in the years of 2005 and 2008 [6, 8]. The reviewing of BI tools as below are based on their investigations. We consider the tools for making a complete data warehouse solution such that we will follow the order of ETL → DBMS → OLAP Server → OLAP Client.

- **ETL** There are many open source ETL tools found, such as OpenSrcETL, OpenETL, CloverETL, KETL, Kettle, Octopus and Talend etc. Most of them can meet the fundamental requirements of data processing, support extract data from heterogeneous data sources and load the data into ROLAP or MOLAP system. But in general, these open source ETL tools are still not very powerful. For example, most of the data cleaning requires to be coded by users, do not support the automatic incremental loading which is very important for the daily use and the documents are still not very good and comprehensive.
- **DBMS** There are several open source DBMS available, such as Ingres, LucidDB and Eigenbase, MonetDB, MaxDB, MySQL and PostgreSQL etc. We here only introduce the most popular open source DBMS: MySQL and PostgreSQL. Currently, the latest product release of MySQL is 5.0.67 and 6.0 is in its alpha stage. It provides consistent fast performance, high reliability and ease of use. But MySQL lacks of support of materialized views, bitmap indices and start joins which are all crucial for BI. PostgreSQL describes itself as the "*world's most advanced open-source database*". Its current product release version is 8.3.3. PostgreSQL can be compared favorably to other DBMS. It contains all the features that we can find in other commercial or open source databases. But the materialized views and start joins are not supported in PostgreSQL. The on-disk bitmaps are also net yet support, but they are planned for inclusion in a future version. On the whole, open-source DBMS have reached a high maturity, but still lacks of some features offered by leading commercial DBMSs [6].
- **OLAP Server** Not many open source OLAP servers are available. In the survey [6], only two open source OLAP servers were found which are Mondrian and Palo. Mondrian supports the multidimensional expressions query language and the XML for Analysis and JOLAP specifications and since Nov. 2005 it became part of the Pentaho business intelligence suite. While Palo is a memory resident multidimensional (OLAP or MOLAP) database server. Apart from multidimensional queries, data can also be written back and consolidated real-time. Both of the OLAP servers are used in BI industry.
- **OLAP Client** There are many choices on the OLAP clients. According to the surveys [6] in 2005 and 2008, within 3 years many products have been developed, such as FreeAnalysis, JMagallanes OLAP \& Reports, JRubik, Openl and REX. JPovit has been very actively developed since their survey in 2005. In general, all the OLAP clients are implemented in Java. The tools are run either on clients or web servers. But JPivot is the most widely used as it is included in different BI suites.

In summary, the ETL tools is the least mature compared with other BI tools since they are still lack many useful features required by BI projects. The open source DBMSs have the highest maturity, as the evidence that some can be compared with commercial DBMSs. With respect to the OLAP servers and clients, there is a great difference in their maturities. Therefore, the diversity of these tools makes it possible for us to propose a complete open source business intelligence solution.

2.2 Open Source BI Solutions

Not only have recent years witnessed a increasing number of open source softwares came into use in BI markets, but also the tendency of providing a complete open source BI solutions. For example, the company JasperSoft offers stand-alone BI tools: Jasper-Report, JsperAnalysis and JasperETL, as well as a complete BI solution: JasperIntelligence-Suite. Other vendors, like Pentaho, Jedox and Greenplum, also give their own BI tools or complete solutions. Research communities and universities also offer some open source BI solutions, such as the University of Waikato, Neuseeland, SpagoBI, and the Eclipse Foundation. Our previous EIAO project also gives an open source BI solution for its data warehouse - EIAO DW implementation [7].

3. THE PROPOSED SOLUTION OVERVIEW

In this section, we will propose an open source data warehouse architecture for eGovMon DW. This proposal is partially based on the literature review, but mainly on the findings of the completed open source web warehouse project EIAO [7].

3.1 Architecture Overview

The eGovMon project will develop large-scale web observatories for the benchmarking of eGovernment services. The *accessibility*, *transparency*, *efficiency* and *impact* (ATEI) of eGovernment services will be automatically evaluated, in which a large amount of evaluation data will be produced. The constructed data warehouse needs to be

met the increasing data storage demand while providing a high performance for the analysis of ATEI. Thus, we propose a distributed data warehouse architecture which is outlined in Figure 1.

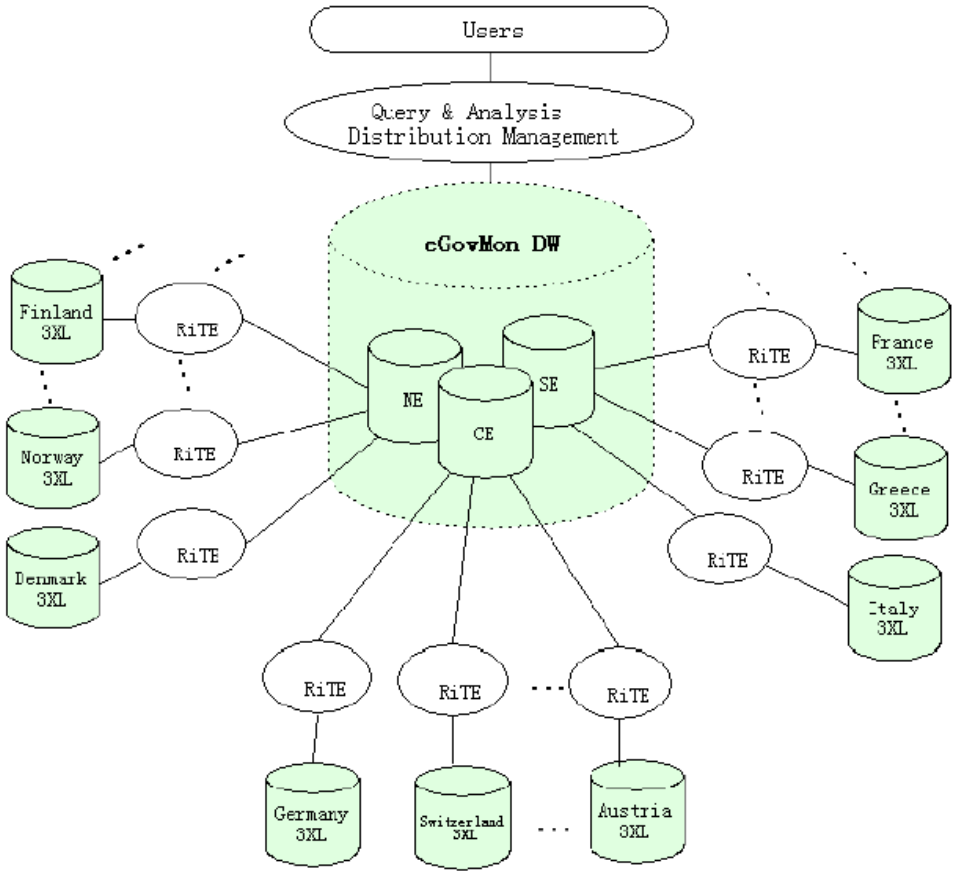

Figure 1: eGovMon Data Warehouse Architecture

The center of this architecture is the data warehouse, named eGovMon DW, which is composed of several physical data repositories, such as the repository for Northern Europe (NE), Central Europe (CE) and Southern Europe (SE) and so on, and each repository can be physically distributed on different servers and geographically in different places, but these separate data repositories function as one logical data warehouse by queries. Similarly, we can have separate data sources partitioned by regions or locations, such as by different countries Denmark, Norway, Finland and so on. For every data source, it will have its own ETL process to extract the source data to the corresponding target data repository. This partition by geographical regions is based on the consideration of the degree of difficulty, feasibility and extensibility, though there might exist one obvious problem - the load balance which might cause the different data loading overhead both in ETLs and data repositories. For example, some ETLs have big loading overhead while some only little; some repositories have huge data volume while some only small. However, the impact brought can be solved by distribution management, which will be further studied in the distribution DW design.

3.2 The Component Details

Data Model It includes conceptual, logical and physical data model. In the conceptual modeling phase, we need to capture and understand the user requirements so that the concepts revealed can be accommodated into the conceptual model. For the conceptual design of DW, we propose to use the multidimensional model rather than relational model as it provides us a more clear view of data structure - facts and dimensions, which is better to support analytical work. In the logical model design, the main issue is how to design the hierarchy levels of dimensions as the multi-levels of dimensions may cause a number of tables joins when do the query operations, such

as *roll up*. In database, join is an expensive operator, especially, for large tables. Hierarchy dimensions design can be less hierarchy levels but with a bigger dimension table, e.g., many fields populated in one table or more hierarchy levels but with a smaller dimension table. Thus, we need to considerate the impact of multi-joins, which the dimension hierarchy levels bring, to the query speed when we do the logical design of data warehouse. In the physical modeling, tables are declared for different dimensions and facts. Star or snowflake data schema is generated for the dimensions and facts physically, indexes are built and data partitions are used for the large amount of data populated in the fact tables. The partition technologies will be our main concern in the physical design. Usually, the data in the warehouse can be vertically partitioned by measures or horizontally partitioned by dimensions. In our design, we propose to horizontal partition according to the location dimension, which is also indicated in our proposed architecture: the repository for each European region.

Data Source 3XL storage system will be used for the eGovMon DW data source, which is proposed by [5]. The 3XL system can automatically generate a specialized schema for the data based on *Web Ontology Language* (OWL) descriptions of classes and their properties. It creates a table for each of the OWL classes. The data of instances is hold in the table of the class. In order to achieve high performance, the inserted data will be buffered in the main memory and only be flushed to the database when memory is needed or by committing. The experience from EIAO DW, in which 3store system was used, is that 90% - 99% of used time was spent on extracting data from the 3store [5]. However, by the use of data buffer and bulk loading technologies, 3XL storage system can be expected to gain a better performance than 3store storage system.

ETL The ETL is the way of extracting the data from different data sources, converting them into the uniform data format and loading into the data warehouse. In the proposed architecture, there are several number of *Right-Time ETLs* (RiTE) running in parallel to load the data from distributed 3XL data sources to the central data warehouse. RiTE is an ETL technique that can make the new inserted source data quickly available to data consumers, while still providing bulk-load insert speed. The trick is that *catalyst*, a middle-ware system, is used to achieve fast loading and concurrency control [9].

eGovMon DW The data warehouse is the central data repository that stores the materialized view of source data. It uses a multidimensional model where the data is stored as *facts* and *dimensions*. As the data volume will be very huge in our data warehouse, data partition is necessary to achieve a better query performance. In this proposal, the data partition by the location dimension, e.g., the different European regions, will be applied to our data warehouse but logically it still functions as a whole data repository. Open source PostgreSQL is proposed to use as the DBMS of eGovMon data warehouse as it provides several useful features that are crucial to data warehouse, such as good extensibility, table partitioning and bit-mapped indexes etc. Besides, PostgreSQL performs very well for complex queries on large databases.

OLAP A data warehouse stores and manages data. OLAP transforms data warehouse data into strategic information providing users multidimensional views for their analysis. Open source tools, like JasperAnalysis, Pentaho etc, can be chosen to fulfill the normal OLAP operations, such as slice and dice, pivot, filter, chart, drill-down, or roll-up a cube of data in real-time.

In summary, we gave an overview and the general components' information of eGovMon DW architecture above. The technologies used and architecture will evolve with the project maturity. More interesting open source technologies used in the business intelligence will be explored in the future.

4. RESEARCH ISSUES

The eGovMon DW project research will be based on the proposed architecture as shown in Figure 1. As we mentioned in section 3, this open source DW contains a collection of components, technologies and techniques. This project expects to focus its research effort on the issues described in the followings.

- **Resource Description Framework (RDF) 3XL Storage System** will be one of our research interests. As RDF data model is a directed graph, when a large number of data is persisted in the database, its insert and query speed should be highly considered. The research will focus on the 3XL storage system implementation and the optimization.
- **Data Model Design** includes conceptual, logical and physical design. In eGovMon project, there will be four indicators used to benchmark the eGovernment services such that the research interest will arise from scalability requirement of the data model design. At the same time, the data model design should also meet the efficiency requirement that is to achieve a good performance for the queries. Research interest will also focus on the design, implementation and maintenance of materialized views in the physical design.

- **Distributed Data Warehouse Architecture** in our proposal has distributed data repositories partitioned by geographical regions, such that the research interests include the scalability of this architecture, to find innovative technologies for the parallel data placement, history data handling, and load balancing problem etc.
- **Very Large Dimension Design** will be another research issue in our project as a dimension can be very big, like *subject* dimension in EIAO project, which all the dimension hierarchies populated in one dimension table with many fields. The implementation in this way is much easier, however, one problem might arise from the big table and the huge data volumes, which query performance will deteriorate. Therefore, it is of significance that some techniques are taken into account in the dimension design, such as vertical partitioning or hybrid design etc.
- **Right Time ETL (RiTE)** is responsible for the extraction of data from data sources, cleansing, customization and insertion into a data warehouse. Research issues can be concerning the optimization and further development of RiTE.

Other research interests or issues may be included, but depend on the extent of our project involved and their importance, such as meta-data management, query optimization and performance etc. More discoveries of research interests will proceed with the project development.

5. CONCLUSIONS AND FUTURE WORK

We have proposed an data warehouse architecture for the eGovMon project. Open source technologies are new introduced in the business intelligence area. It is of great significance of our attempt to propose data warehouse architecture with open source technologies. We expect the architecture to evolve as the project matures, which should help to fit open source technologies into business intelligence in better.

As it is a three-year project, we will have several milestones with the project progression. The next is a more concrete requirement will be acquired for the eGovMon DW so that a more concrete distributed data warehouse architecture can be designed and documented. After that, it comes to the implementation and optimization of 3XL storage system where we expect more interesting research issues can be found. Much effort will then go to the data model design. An ETL - RiTE will be implemented to achieve the INSERT-like data availability but with bulk-load speed for the data loading. At the later phase of this project, OLAP reports generation techniques, performance tuning and other issues will be into our research.

6. REFERENCES

[1] eGovMon - eGovernment Monitor. Available from http://www.egovmon.no/en as of 2010-07-15.
[2] W. H Inmon and R. D.Hackethorn. *Using the Data Warehouse*. John Wiley & Sons, Inc., New York, NY, USA, 1994.
[3] M. Podolecheva. *Open Source meets Business Intelligence An Introduction to Pentaho*, June, 2007.
[4] D.J. Power. *A Brief History of Decision Support Systems*. DSSResources.com, 2003.
[5] C. Thomsen. *Aspects of Data Warehouse Technologies for Complex Web Data*. PhD thesis, Aalborg University, Denmark, January 2008.
[6] C. Thomsen and T. B. Pedersen. *A Survey of Open Source Tools For Business Intelligence*. In Proc. of DaWaK, Berlin, Germany, 2005.
[7] C. Thomsen and T. B. Pedersen. *Building a Web Warehouse for Accessibility Data*. In Proc. Of DOLAP, New York, NY, USA, 2006.
[8] C. Thomsen and T. B. Pedersen. *A Survey of Open Source Tools for Business Intelligence*. In International Journal of Data Warehousing and Mining, 2009.
[9] C. Thomsen, T.B. Pedersen, and W. Lehner. RiTE: Providing On-demand Data for Right-time Data Warehousing. In Proc. of ICDE, 2008.