Data Warehouse for Interactive Decision Support for Addis Ababa City Administration

Yilma Desta
Ethiopian Airlines, Addis Ababa, Ethiopia
yilmadessta@yahoo.com

Mesfin Kifle
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

Information Technology plays a vital role for Ethiopia in its effort for development. Effective information provision is necessary for agricultural and industrial development, improvement of health services, education, family planning and the protection of the environment. From the Addis Ababa City administration point of view, to increase the quality of decision making, there is a need to process information, scattered across databases under sector bureaus or sub cities.

As a technology solution for information management and provision, a data warehouse is a collection of data to support decision making which is subject oriented, integrated, non-volatile, and time variant. Implementing a data warehouse can help decision makers at city, sub city or wereda level to make the right decisions at the right time. In this paper, related works have been reviewed to get a clear understanding of the data warehouse concepts and implementation strategies. The Ethiopian e-government framework is also reviewed to see the relationship with data warehouse. In the review, it is noticed that a data warehouse can be used as a source database for e-government.

Furthermore, current data warehouse technologies are also explored and analyzed to select an appropriate data warehousing platform considering cost, scalability, performance, and security as decisive factors to compare the available technologies. Based on these factors, Microsoft SQL Server outweighed for the target data warehouse. In effect, SQL Server best practice is reviewed to support the implementation. Additionally, source databases of Acts & Civil Status and Registration Document Office, Finance and economy development and Education bureaus of the City are identified and analyzed. User's analytical requirements are also gathered.

By taking the SQL Server best practice, source database design and user analytical requirements as input, a data warehouse is designed and a prototype is developed. The lesson drawn from the prototype is that decision makers can access the integrated data online from their office, meeting hall or even at home according to their access privilege and analyze it to get information which supports their decision.

Keywords: Data Warehousing; SQL Server; Database

1. Introduction

Information Technology plays a vital role for Ethiopia in its effort for development. Effective information provision is necessary for agricultural and industrial development, improvement of health services, education, family planning, and the protection of the environment.

Information is a major development resource. It allows governments to serve citizens in a timely, effective, and cost effective manner. Information helps government agencies to make informed decision.

In order to get valuable information from historical data, data must be extracted, organized and analyzed. Currently the City Administration has a set of databases across Education bureaus, Health bureaus, etc. that are running independently. Large amount of data is stored in the databases and are not integrated and suitable to draw information that can support the City Administration's decision process.

Data warehousing is increasingly being used by businesses to store data for the purpose of decision support [1]. A data warehouse is populated with data from preexisting systems and/or other external sources. Data from these sources are transformed and

integrated to provide a more complete picture of an institution [1, 2].

Currently many data warehouse software are available in the market. IBM, SAP, Teradata, Microsoft, Oracle, Sybase, and Netezza are among the leading data warehouse vendors.

Microsoft is among the data warehouse vendors making most significant headways. SQL Server's ease of use, availability and tight Windows operating system integration makes it an ideal choice for firms that choose Microsoft products for their enterprises [10]. Oracle Database 11g and Oracle Warehouse Builder (OWB) are the tools that have let Oracle reach its great position with significant market share.

2. Problem Statement

A data warehouse is an enabled relational database system designed to support very large databases at a significantly higher level of performance and manageability. Any organization or a system in general is faced with a wealth of data that is maintained and stored, but the inability to discover valuable, often previously unknown information hidden in the data, prevents it from transferring these data into knowledge or wisdom [4].

Addis Ababa City Administration is part of the government body which collects a large amount of data from different sectors. But the valuable information within the collected data is hidden and the City Administration is not able to convert these data into knowledge. This makes it difficult to the City Administration to make informed decisions.

Currently the City administration bureaus exchange reports using hard copies. Each bureau generates its own report and sends a print out to the concerned office. This requires days to reach to the office. In addition to this, the report might not show exactly what the decision maker wants to see.

On the other hand, even though a couple of data warehouse technologies are available, the internal capacity to explore, analyze, and select an optimal technology to the context of the City Administration is limited. The available data warehouse technologies are varied in terms of cost, scalability, performance, and security. To choose which platform to invest in, first identifying the pros and cons of the available data warehousing platforms is important. Proper comparison enables the City Administration to evaluate the currently available data warehouse technologies and select the one that suits its need.

Establishing a data warehouse will help to discover valuable information out of the data collected by the Addis Ababa City Administration which is vital for the administration to make strategic decisions. The user can interact with the remote data and make analysis according to his/her need. This helps the decision maker to get the appropriate information being in his office or any where according to his/her access privilege to the data warehouse.

3. Literature Review

An organization can manage information in two dominant forms: operational systems of record and data warehouses. Operational systems are designed to support online transaction processing (OLTP) whereas data warehousing systems are designed to support online analytical processing (OLAP) [3]. This section discusses review of data warehousing and related topics.

3.1 Data Warehouse

Operational systems concentrate on high-volume transaction processing on a day-to-day basis using real-time data. They are optimized for simplicity and speed of modification, allowing for efficient and effortless data entry and retrieval. While the operational systems primarily focus on current data management, a data warehouse updates and stores historical data.

The fundamental goal of a data warehouse is to support strategic planning, modeling, and forecasting at the organizational level. It must fulfill the need for knowledge for an area of uncertainty or growth in the organization. In order to accomplish this task, it must provide a single, comprehensive, and consistent view of the organization. A data warehouse must present the information consistently and securely to its users.

As shown in Figure 1, a data warehouse has four main components: operational systems of record, the data staging area, the data presentation area, and data access tools [2]. Each component of the data

warehouse serves a unique function in preparing data for manipulation and examination.

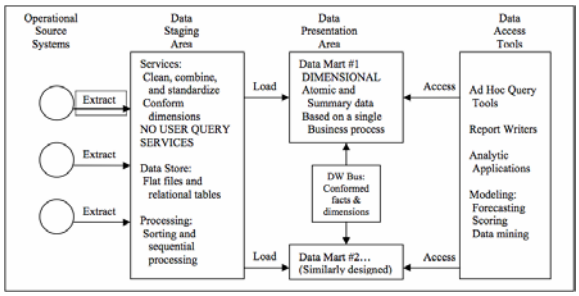

Figure 1: The basic elements of data warehouse

First, the data is extracted from various source systems and copied into the data staging area. There, the data is combined, cleansed and transformed into a standard format and structure. Missing elements, incorrect labels, duplicate data, misspellings, and other errors are manipulated and corrected in this phase. Once the data is standardized, it is loaded into the data presentation area [3].

The formatted data is organized, located and available for user queries in the data presentation area. Once the data presentation area contains the formatted data, users can utilize various data access tools to perform queries. Some data access tools include ad hoc query tools and data mining search specific segments of the data presentation area [3].

3.2 E-Government

Figure 2 [8] shows the architecture framework of e-government which is divided into four layers: access layer, e-government layer, e-business layer, and infrastructure layer.

Access layer involves the channels that government users can access the various government services. Government users can be citizens, business firms, employees, other governments, and other community members.

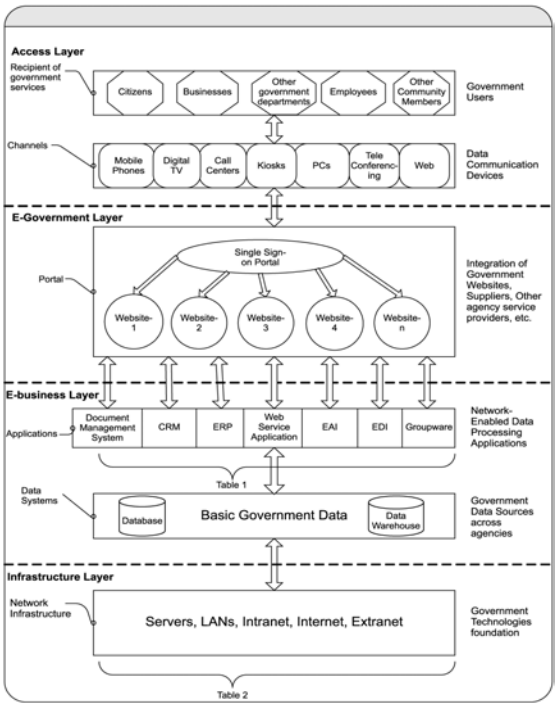

Figure 2: Framework of e-government architecture

E-government layer is about integrating digital data of various organizations into a web-portal of government services, in the form of a one-stop e-government portal. This may result in improved access to government resources, reduces service-processing costs, and enables organizations to provide a higher quality of service [8].

E-business layer is focused on using ICT applications and tools to harness a network of trust, knowledge sharing and information processing that takes place both within and between organizations. Infrastructure layer focuses on technologies that should be in place before e-government services can be offered reliably and effectively to the public.

3.3 Ethiopian E-Government

The Ethiopian Government Information System Development and Data Centre Administration team is managing projects targeted towards the development of integrated information systems to avail government services online [7].

The e-Government strategy for Ethiopia has identified several strategic priority projects. These projects, when implemented, would deliver electronic services. In Figure 3 [7], the Ethiopia e-government core projects are shown.

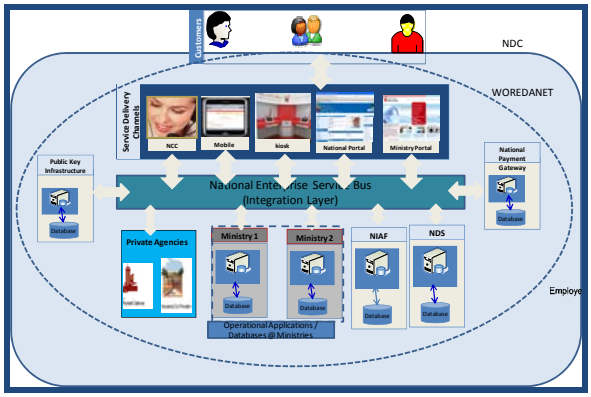

Figure 3: Ethiopia e-government Core Projects

4. Exploring and Analyzing Data Warehouse Technologies

The available data warehouse technologies are evaluated against the factors identified: cost, scalability, performance, and security. The analysis is presented in this section.

Oracle licenses much of its premium database functionality through options, which can only be added on to an Enterprise Edition license (Oracle Database options are not available with its Standard Edition). For these options, payment is required in addition to the base Enterprise Edition license fee. Considering that the base cost of an Oracle DB 11gR2 Enterprise Edition processor license is $47,500, these options can add significantly to the acquisition and maintenance costs of an Oracle database. Considering these additional licenses required to achieve comparable functionality to that offered by SQL Server, the total aggregate least price for an Oracle configuration would be $2,024,0007 [11, 12].

Microsoft does not require the purchase of any add-on options in order to have full rights to use all available database capabilities. Figure 4 shows license price comparison including Oracle options (data warehouse scenario, 4 x 8-core x86 processors).

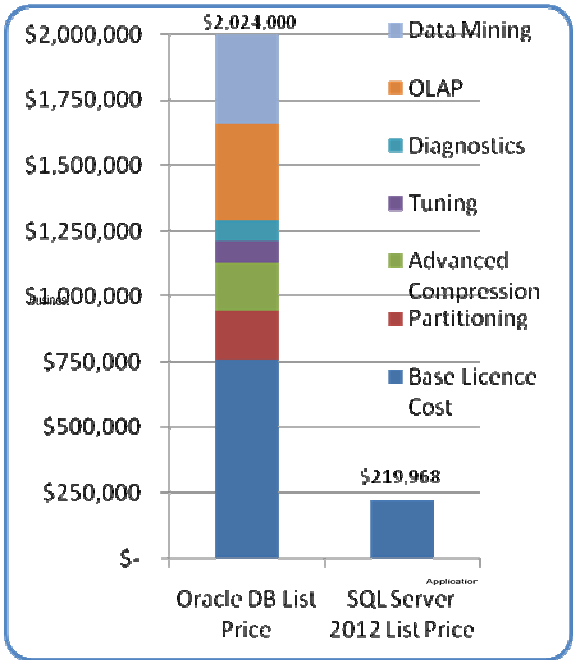

Figure 4: License price comparison including Oracle options (data-warehouse scenario, 4 x 8-core x86processors)

The SQL Server product group has made sizable investments in improving scalability and performance associated with the SQL Server Database Engine. Some of the main enhancements that allow organizations to improve their SQL Server workloads are Column store Indexes, Partition Support, Online Index, Achieve Maximum Scalability with Windows Server 2008 R2 [6].

Strong security is a must for every organization irrespective of size or vertical industry. Databases are among the most crucial applications in the entire network infrastructure. Information in databases is the organization's intellectual property and life blood [5].

Since 2002, Microsoft's SQL Server has compiled an enviable record. It is the most secure of any of the major database platforms [5]. SQL Server has recorded the fewest number of reported vulnerabilities, just 49 from 2002 through June 2010, of any database [5]. These statistics were compiled independently by the National Institute of Standards and Technology (NIST), the US government agency that monitors security vulnerabilities by technology, vendor, and product [5]. So far in 2010, through June, SQL Server has a perfect record, no security bugs have been recorded by NIST.

During the same eight-and-a-half year period spanning 2002 through June 2010, NIST recorded 321 security vulnerabilities associated with the Oracle database platform, the highest total of any major vendor [5]. Oracle had more than six times as many reported security flaws as SQL Server during the same time span. This is shown on Figure 5 [5].

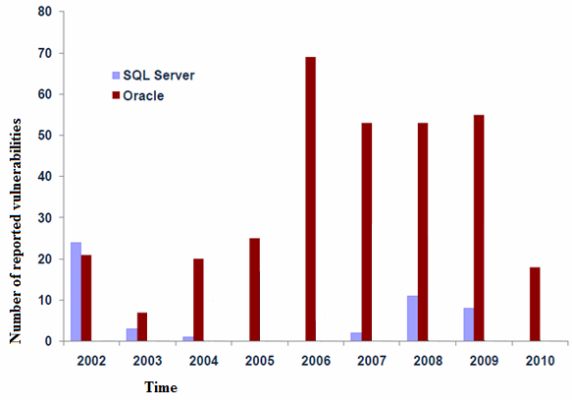

Figure 5: Reported Vulnerabilities of Database Platforms from 2002 to June 2010

To conclude the comparison, SQL Server 2012 outweighs Oracle 11g in many aspects. Microsoft provides full database capabilities for data warehousing with the basic license cost. Whereas most of the data warehouse tools in Oracle 11g are separately licensed. This makes the cost of Oracle 11g with options for data warehousing more than 9 times of the SQL server 2012.

SQL Server has recorded the fewest number of reported vulnerabilities, from 2002 through June 2010, of any database [5]. Oracle had more than six times as many reported security flaws as SQL Server during the same time span. From these we can see that SQL Server is more secure than Oracle.

Both SQL Server 2012 and Oracle 11g have high availability. The AlwaysOn new feature of SQL Server 2012 is more than the data guard in Oracle 11g. In SQL Server 2012, the fail substitute server is an active, already in-use instance. But Oracle 11g uses a separate standby server which takes over when the primary server fails.

The SQL Server has made improvement in scalability and performance associated with the SQL Server Database Engine through Columnstore Indexes, Partition Support, Online Index Create, Rebuild & Drop and Achieved Maximum Scalability with Windows Server 2008 R2.

This shows that choosing SQL Server data warehouse platform is useful from cost, security, availability, performance, and scalability aspects. Based on these facts, Addis Ababa City Administration is recommended to use SQL Server for the data warehouse implementation.

5. Source Database Design Analysis

To build a data warehouse, performing analysis of the existing source databases that will potentially contribute to the data warehouse is mandatory. It helps to know relationships between data. This provides the necessary information to design the extraction, transformation and loading (ETL). Operational databases from Acts & Civil Status and Registration Document Office, Education Bureau and Finance & Economy Development Bureau are identified and analyzed.

Analytical requirements of Acts and Civil Status and Registration Document Office, Education Bureau, Finance and Economy Development Bureau, and Addis Ababa City Administration are collected.

6. The Data Warehouse Design

As discussed in the previous sections, the source databases are identified and analyzed. The user analytical requirements are also collected. This section discusses the logical and physical design of the data warehouse, transformation, and architecture of the data warehouse.

The design approach is first the logical model is designed based on the user requirement and source databases of the Acts and Civil Status and Documents Record Agency, Education Bureau, Finance and Economy Bureau, and the Addis Ababa City Administration logical model. Then the physical model is designed for each based on the logical model. Finally the architecture of the data warehouse is presented.

6.1 Data Warehouse Design

Using the soured databases and the user analytical requirements as input, the logical and physical model of the data warehouse is presented as follows.

The data warehouse design follows the analytical requirements of users. The users tend to split into many groups, each with different analytical requirements. Users in each group often articulate the analysis they need (the User model of analysis) in terms of graphs, grids of data (worksheets), and printed reports. These are visually very different but all essentially present numerical measures filtered and grouped by the members of one or more dimensions.

So we can capture the analytical requirements of a group simply by formalizing the measures and dimensions that they use. These are captured together with the relevant hierarchical information in a sun model, which is the Logical model of the analytical requirements.

Once these analytical requirements have been successfully captured with logical model, technical detail is added. This transforms the Logical model into a Physical model.

6.2 Data Warehouse Architecture

Figure 6 shows the Architecture of the data warehouse. The architecture of the data warehouse has three layers: source database layer, data warehouse layer, and the analysis layer. The data source database layer consists of the databases which the data warehouse uses as source of data. The data source databases can be as many as the number of independent databases under the City Administration.

Currently three data sources are indicated. SQL Server agent will be scheduled to refresh the data warehouse periodically. The schedule time will be when the source databases are at minimum load which can be during the night. After the first load to the data warehouse, the schedule to refresh the data warehouse will be based on the nature of the source databases change. If we compare the IBEX, RIDRS, and StatEdu2 source databases, IBEX and RIDRS change more frequently than StatEdu2.

After loading the data warehouse, facts and dimensions will be created based on the user's analytical requirement. Triggers on the data warehouse tables trigger stored procedures whenever there is insertion, update or delete on tables.

Using the dimensions and fact tables, cubes will be created. Cubes will be presented to the user for analysis.

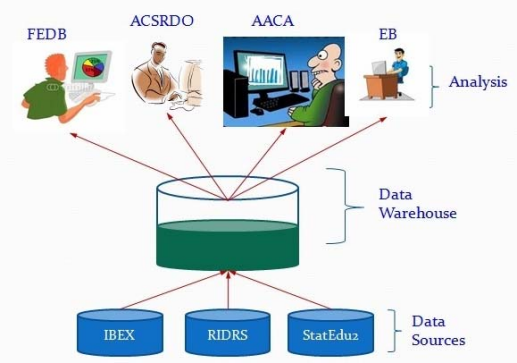

Figure 6: Data Warehouse Architecture

7. Implementation of Data Warehouse

The implementation of a data warehouse is resource intensive and time taking. Data warehouses are not built all at once. Instead, they are designed and populated a step at a time, and as such are evolutionary, not revolutionary. Building a data warehouse all at once, the resources required, and the disruption to the environment all dictate that the data warehouse be built in an orderly, iterative, step-at-a-time fashion [10]. The "big bang" approach to data warehouse development is simply an invitation to disaster and is never an appropriate alternative [9].

For the purpose of this study, the prototype is limited to part of the whole. The same approach will be applied to implement the rest of the data warehouse iteratively.

SQL Server 2008 R2 platform is used to implement the data warehouse. Two instances of the SQL Server (one for the source database and the other for the data warehouse database) are used.

After making the cubes ready for analysis, the next is making it available to decision makers. Based on the SQL Server best practice, Microsoft Office Excel is used to analyze the data.

The user can analyze the data by charts, tables, etc. Microsoft Excel 2010 has an additional feature which is called slicer. The slicer makes the analysis more interactive for the user. Figure 7 shows the different options to analyze the data.

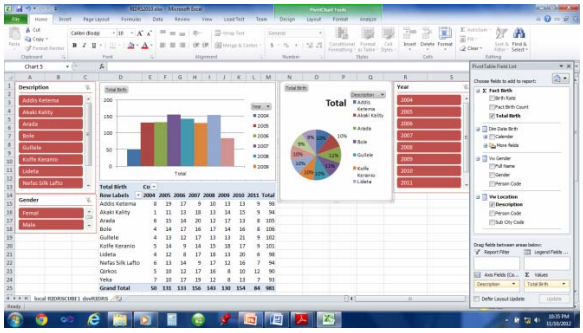

Figure 7: Analysis

8. Conclusions and Future Work

Information is a major development resource. It allows governments to serve citizens in a timely, effective, and cost effective manner. Addis Ababa City Administration is a government body which collects a large amount of data from different sectors. Establishing a data warehouse will help to discover valuable information out of the data collected by the Addis Ababa City Administration databases which is vital for the administration to make strategic decisions.

In this study, literature is reviewed to get a clear understanding of data warehousing. Current data warehouse technologies are also explored and analyzed to select data a warehousing platform. Cost, scalability, performance, and security are taken as factors to compare the technologies. Based on the results Microsoft SQL Server is suggested for the Addis Ababa City Administration data warehouse implementation.

By taking best practices of the SQL Server, source database design and user analytical requirements as an input, the data warehouse is designed and the prototype is developed. This enables the decision makers to analyze data, generate reports, and make informed decisions.

To enrich the result, additional works are expected in the future. One major work for the data warehouse could be to consider disaster recovery. Making a data warehouse available is not easy. A data warehouse can take months to set up, yet can fail in seconds. Therefore, disaster recovery is a challenge for data warehouse that must be tackled.

The other is security. The data in the data warehouse must be available to the right users at the right time, and the system must keep a record of activities performed by its users.

References

[1] W. H. Inmon, “Building the Data Warehouse”, 4th ed., Wiley Publishing Inc., Indiana, 2005.

[2] Kimball, Ralph, Ross, and Margy, “The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling”, 2nd edition, John Wiley and Sons Inc., 2002.

[3] Priya Chaplot, “An Introduction to Data Warehousing “, Mt. San Antonio College, January 9, 2007.

[4] S. Sumathi and S. N. Sivanandam, “Introduction to Data Mining and its Applications”, Springer-Verlag, Berlin Heidelberg, 2006.

[5] Laura Di Dio, ”SQL Server Delivers Industry-Leading Security”, Information Technology Intelligence Corp. (ITIC), 2010.

[6] Ross Mistry and Stacia Misner, “Introducing Microsoft SQL Server 2012”, Microsoft Press, 2012.

[7] “E-Government Strategy and Implementation Plan - Report”, Ministry of Communications and Information Technology, January 2011.

[8] Zakareya Ebrahim and Zahir Irani, “E-government Adoption: Architecture and Barriers”, Brunel University Business Process Management Journal, Vol. 11, No. 5, 2005, pp. 589-611.

[9] Mark Whitehorn, Keith Burns, and Eric N Hanson, ”Best Practices for Data Warehousing with SQL Server 2008 R2”, Microsoft, 2010.

[10] http://www.serverwatch.com/trends/article.php/3883441/Top-10-Enterprise-Database-Systems-to-Consider.htm, last accessed on December 04, 2012.

[11] http://www.oracle.com/us/corporate/pricing/price-lists/index.html, last accessed on August 11, 2012.

[12] http://download.microsoft.com/download/5/9/5/59527629-ABD3-4C12-8117-DFABB86E2CFA/SQL2012_Licensing_datasheet_USA_Dec2011.pdf, last accessed on August 11, 2012.