Challenges in Developing a Data Warehouse to Manage the Rollout of Antiretroviral Therapy in a Developing Country

J. E. Kotzé, University of the Free State, South Africa; E-mail: eduan@xsinet.co.za
T. McDonald, University of the Free State, South Africa; E-mail: Theo.SCI@mail.uovs.ac.za

ABSTRACT

With a global HIV/AIDS epidemic, developing countries are facing an enormous challenge in combating the disease. Public health will be placed under severe pressure in providing treatment such as highly active antiretroviral therapy to all its HIV infected patients. This paper will describe the challenges involved in establishing a data warehouse to provide strategic information during the rollout of antiretroviral therapy (ART). The construction of a Human Resources Data Mart, which is critical to the successful rollout of antiretroviral therapy in South Africa, will be discussed in detail. Special attention will be given to extraction, transformation and loading, slowly changing dimensions type 2 and materialized views.

Keywords: Data Warehousing, Healthcare, Human Resource Data Mart, Antiretroviral Treatment, Developing Countries.

1. INTRODUCTION

The HIV/AIDS epidemic is a global crisis which threatens development gains, economies and societies. At the end of 2004, the total number of people worldwide living with HIV/AIDS was estimated to be just under 40 million. In South Africa the estimated number of AIDS related deaths in 2003 ranged anywhere between 270 000 and 520 000 according to the UNAIDS Global Report (UNAIDS, 2004).

In response to this epidemic, the South African Government created the HIV/AIDS and STD Strategic Plan. This plan includes the provision of antiretroviral therapy in the public health sector in an attempt to reduce AIDS mortalities. Antiretroviral treatment (ART) for HIV infection consists of drugs that slow down the reproduction of the HIV virus in the body.

The Free State Department of Health (FSDOH) launched its provincial antiretroviral treatment program during May 2004. By the end of June 2006 a total of 31 public health facilities were empowered to provide antiretroviral drugs for 6200 patients in the Free State. The Actuarial Society of South Africa (ASSA) has developed an AIDS Demographic Model that can be used to project the impact of this disease on each province in South Africa. According to Chapman (2003) using the ASSA 2000 Model, it is estimated that in the Free State

- Approximately 480 000 people are HIV positive (based on 30.1% HIV positive mothers in the 2003 HIV Antenatal Survey);
- Seven percent (7%) of all HIV infected patients are in a World Health Organization (WHO) Stage 4 AIDS defining illness, which is approximately 31 111 patients;
- Annually, 28 290 patients will develop a WHO Stage 4 AIDS defining illness.

The WHO recommends that all people in a WHO stage 4 AIDS defining illness should commence with antiretroviral treatment immediately. This recommendation will pose serious challenges in managing the resources required for treating all these patients by the FSDOH. Mechanisms have to be developed to effectively monitor the antiretroviral treatment programme but at the same time provide the necessary **strategic information** in managing and evaluating the programme as well. It is clear that a number of factors are forcing the FSDOH in the direction of a data warehouse (DW).

This paper will indicate how a Health Department in South Africa, the FSDOH, tackled and successfully managed the challenges of creating a data warehouse. A background section will provide the history of the current operational system and the shortcomings of the system. That will be followed by a detailed discussion of the challenges involved in constructing the human resource data mart (HRDM) which is critical to the successful rollout of ART in South Africa. The Extraction, Transformation and Loading (ETL) process will be examined and slowly changing dimensions will be addressed. The paper then concludes with a discussion of a modified staging area that uses a materialized views approach to provide the platform for developing the human resource (HR) online analytical processing (OLAP) cube.

2. BACKGROUND

2.1 General

A data warehouse differs significantly from a conventional operational or transactional database in several aspects. First of all, a complex data structure must be maintained in order to offer flexible and dynamic retrieval of rich decision-support knowledge (Shin, 2003). For this, it maintains data that is more integrated, subject-oriented, non-volatile and time-variant in comparison with transactional or operational databases (Dodge & Gorman, 2000; Hristovski *et. al.,* 2000; Shin, 2003). Data structures of a data warehouse should also be more cross-functional (Shin, 2003) and support management decisions (Hristovski *et. al.,* 2000).

According to Saraceni *et. al.,* (2005), the linkage of several databases can assist with studying the distribution of diseases and for analysis of AIDS-related mortalities in Brazil. Although the linkage of databases is in essence not a data warehouse, it demonstrates the importance of analyzing information and using it to provide strategic information for the decision-making process.

Data warehouses have previously been used in the areas of health and public health (Davis *et. al.*, 2002; Lau & Catchpole, 2001; Prather *et. al.,* 1997). However, most data warehouses in the health areas are used for ***clinical treatment outcome*** or for ***biomedical studies*** and limited research has been done on the usage of data warehouses in public health for holistic decision-making.

2.2 Lack of Strategic Information

The Personnel and Salary (PERSAL) system is an online transaction processing (OLTP) based payroll system and is used by all the National and Provincial governments in South Africa. The system has been in a production environment since 1990 and was developed in Natural Adabas. At present, the system is being maintained by a private company.

Because the system is OLTP based, it proved inadequate in providing the necessary human resource statistics needed by antiretroviral programme managers.

Furthermore, every new change or new report must be submitted to a central *System Change Control* system. From there it will be prioritised and once accepted, handed over to the private company for development.

This process was cumbersome, inflexible and time consuming which led to overall frustration. In 2004, the FSDOH received approval from National Treasury to extract all relevant human resource data from PERSAL, thus allowing them the freedom of incorporating the data into a data warehouse.

3. CHALLENGES

The following sections will provide details on the challenges that were faced during the development of a data warehouse.

3.1 Limited Budget

According to Schubart & Einbinder (2000), research has showed that the key factors for successful data warehouse implementation are organizational in nature. Management support and adequate resources are most important because these address political resistance. Gatziu & Vavouras (1999) stated that data warehouse development is a demanding and costly activity of which the establishment thereof could be in excess of $1m. This can be a major obstacle in a developing country.

Taking these factors into consideration, top management was approached to direct the development of the data warehouse in early 2005. Because of a limited IT budget (0.68%), a decision was taken to break the project down into several data marts and to develop the data warehouse over a longer period of time. To cut back on costs, in-house existing staff was used to construct the data warehouse in lieu of making use of expensive outside consultancy firms.

Oracle is the current worldwide leader in the data warehouse tools marketplace (Vesset, 2006). Furthermore, Oracle 10.2g also offers all the functionalities required in both OLTP and DW based databases. Both these reasons guided the FSDOH decision to upgrade their existing Oracle9i infrastructure to Oracle 10.2g and make use of it for the data warehouse. The upgrade process was covered in an existing maintenance contract, resulting in no additional expenditure.

Human resources and pharmaceutical (ART drugs) costs were identified as the main cost drivers, and more strategic information was required in order to obtain sufficient funding for the ART programme.

Listed below are the identified data marts:

- human resources
- clinical patient ART treatment
- pharmaceuticals (ART drugs)
- patient mortalities
- tuberculosis

The Human Resource Data Mart (HRDM) was chosen as the first data mart to be constructed and will be the focus of the rest of this paper.

3.2 Extraction, Transformation and Loading Challenges

3.2.1 Data Extraction

The data in the OLTP system (PERSAL) is *transient data* of nature. According to Bruckner & Tjoa (2002), the key characteristic of *transient data* is that alterations and deletions of existing records physically destroy the previous data content. In order to keep the history of the data in tact, all modifications to the data had to be considered.

The ETL processes commenced with a data extraction process which are performed **twice** a month by Treasury. At the beginning of every month, the FSDOH will receive two sets of data. The reason for this approach is entrenched in the manner in which government officials receive their salaries in South Africa. The salary of permanent staff is paid on the 15th of each month. All the information related to this event constitutes data set one. However, additional or supplementary payments (i.e. S&T claims, overtime, fuel allowance) and workforce operations (promotions, staff re-allocations) can be made to government officials from the 16th until the end of the month. All these additional information constitutes data set two.

In essence, the first extraction consists of a full *data snapshot* taken on the 15th of the month from the *transient data set*. This includes **staff, posts** and the **hierarchical organization structure**. According to Bruckner & Tjoa (2002), a *data snapshot* is a stable view of data as it exists at some point in time. It is a special kind of periodic data. Snapshots usually represent the data at some time in the past, and a series of snapshots can provide a view of the history of an organization.

The second extraction consists of the supplementary changes to the *data snapshot* picture of the first extraction set in terms of **staff** and **posts** but **exclude** any changes to the hierarchical organization structure. This extraction process was preformed at the end of the month. It can be regarded as *semi-periodic data*. According to Bruckner & Tjoa (2002), almost all operational systems retain only a small history of data changes due to performance and/or storage constraints.

The challenge pertaining to more than one set of extraction data in the update window is the issue of *late-arriving data.* According to Bruckner & Tjoa (2002), *late-arriving data* is bothersome because it is difficult to integrate with existing fact and dimension tables, especially when surrogate keys are used in order to cope with slowly changing dimensions. Aggregates have to be updated, because the newly integrated data sets will change counts and totals of the prior history. Late-arriving data can therefore possibly change analysis results unexpectedly from the analyst's perspective.

In order to deal with the problem of late-arriving data, it was agreed that the data warehouse will be updated during the **first week** of the following month, reflecting the *transient data picture* and supplementary changes (*semi-periodic data)* that was made to it.

3.2.2 Time Stamping

The standard approach for storing periodic data (typically found in Data warehouses) is to use time stamped status fields for each record. For the HRDM the *load timestamp* method was used.

Slow changing dimensions (SCD) Type 2 will be used as far as possible. According to Berndt & Fisher (2001), this type of change adds rows to maintain an arbitrarily long history. The keys must be "generalized" in this approach by using a version number or some other mechanism, so that related rows can be retrieved as a coherent history.

Each table in the staging area had a column added called EXTRACT_DATE which translated to the record *load timestamp.*

3.2.3 Dealing with Slowly Changing Dimensions

One of the biggest challenges with the HRDM was the monthly changes to the organizational structure. Changes occur when new components (organizational units) are created, moved or become obsolete during the month. Components contain the posts for that particular unit and the links of the child components directly reporting to it.

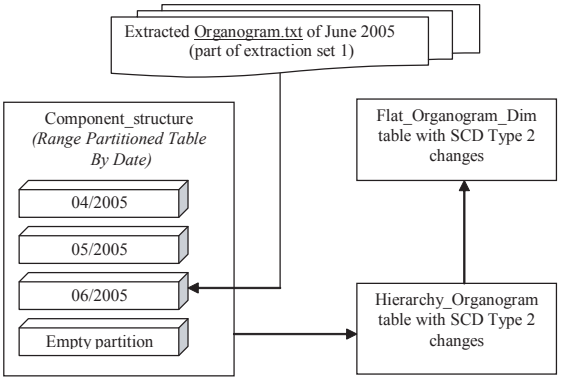
Changes in the organizational structure were not directly reflected in each month's download and had to be identified with specially developed algorithms in order to perform SCD Type 2. This was because the organizational structure was only included in extraction set 1 as a *data snapshot picture* called **Organogram.txt** and not a list of changes. See Figure 1 for the organizational structure data flow of June 2005 as an example.

The data for the organizational structure was imported into the COMPONENT_STRUCTURE partitioned table from the Organogram.txt file. This partitioned table then contained the organizational hierarchy for each month. The organizational hierarchy in turn, consisted of component details and linkages between child and parent components.

A table called HIERARCHY_ORGANOGRAM was constructed and populated with the hierarchy on the date the HRDM project commenced (April 2005). For each following month, COMPONENT_STRUCTURE was algorithmically compared to HIERARCHY_ORGANOGRAM using complex SQL statements and SET operators to help identify the following changes:

- New Component
- Component name change
- Parent component position change
- Parent component name change
- Deleted Component

Figure 1. Data flow for extracted organogram.txt (June 2005)



Each time a change was detected, a new record was inserted into HIERARCHY_ORGANOGRAM, with a new surrogate key. The superseding record was changed to the last date of the previous month. A surrogate key called ORGANOGRAM_KEY was created with the extract date (or *load timestamp*) concatenated with the component number to stay within the bounds of SCD Type 2.

Kimball & Margy (2002) pointed out that hierarchical structures of variable depth presents several problems in the relational environment. Some examples are the difficulty of navigation or rolling up of facts within these hierarchies using non-procedural SQL. This posed a problem for the FSDOH when using Oracle 'CONNECT BY' SQL extension in the same statement as a join. While 'CONNECT BY' is very useful when navigating recursive points in a dimension table, it can not be used by an ad hoc query tool. If the tool could generate this syntax to explore the recursive relationship, it cannot in the same statement be joined to a fact table. Even if Oracle was to remove this somewhat arbitrary limitation, the performance at query time would probably be not too good (Corr, 2001).

To overcome this problem, a bridge table or often called helper tables are inserted between the hierarchical dimension table and the fact table (Kimball & Margy, 2002). The problem the FSDOH experienced with this approach was entrenched in the manner the multidimensional online analytical processing (MOLAP) tool used the dimensional model for its analytical model. The MOLAP tool required a flat organizational view which in theory meant a totally denormalized view of the hierarchical organizational structure and relationships in HIERARCHY_ORGANOGRAM.

Kimball & Margy (2002) also pointed out that when navigating the bridge table via the standard SQL code, it is not for the faint of heart. In order to overcome the prerequisite of the MOLAP tool together with minimizing the SQL complexity for the FSDOH users, a ***modified version*** of a bridge table was introduced. The table FLAT_ORGANOGRAM_DIM was created and used as one of the dimensions in the dimensional model (See Figure 2).

This ***modified version*** of a bridge table might not be the perfect solution should the organizational structure consist of more than 10 levels. To overcome this, the table is re-created every month from all the SCD Type 2 changes captured within HIERARCHY_ORGANOGRAM. In this manner the algorithm will allow an extra level (meaning an extra table column) when it detects it, thus avoiding the possibility of missing data. However, the only manual action to be taken is to insert this additional level (table column) within the MOLAP tool.

3.2.4 Example of a SCD Type 2 on a Parent Component

The following example (See Figure 3) will illustrate a SCD Type 2 on a parent Organizational Unit (Component) between April 2005 and June 2005 and the domino effect it will have on its child components.

- Step 1: Algorithm detects a change in *parent name* in component 011200
- Step 2: Perform SCD Type 2 and load changes into HIERARCHY_ORGANOGRAM

HIERARCHY_ORGANOGRAM

Organogram Key	Extract date	Component Name	Parent Component	Date From	Date To
01-APR-2005-011200	01-APR-2005	Hospital A	011000	01/04/2005	30/06/2006
01-JUL-2006-011200	01-JUL-2006	Pelonomi Hospital	011000	01/06/2006	

- Step 3: Force the change in all child components of component 011200. Only ***Component 011333*** will be illustrated below.

HIERARCHY_ORGANOGRAM

Organogram Key	Extract date	Component Name	Parent Component	Date From	Date To
01-APR-2005-011333	01-APR-2005	AIDS (Unit)	011200	01/04/2005	30/06/2006
01-JUL-2006-011333	01-JUL-2006	AIDS (Unit)	011200	01/06/2006	

- Step 4: Convert the hierarchical organizational structure into a flat organizational structure

Figure 2. Dimensional model

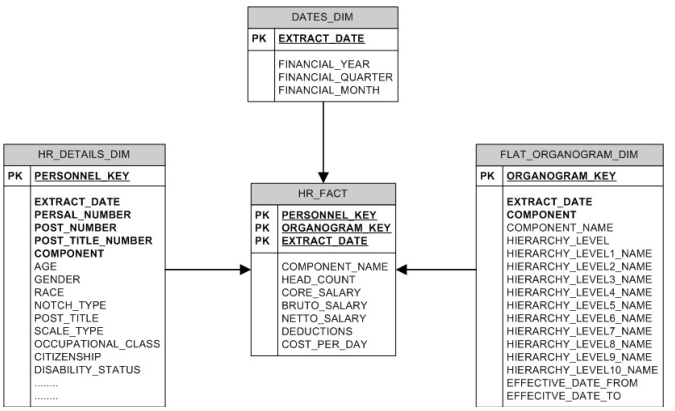
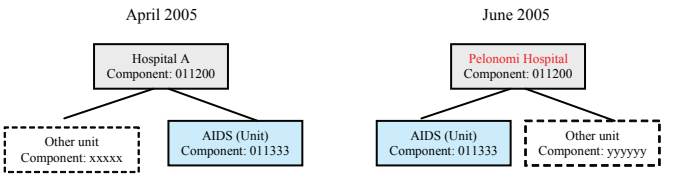


Figure 3. Parent component name change



FLAT_ORGANOGRAM_DIM							
Organo-gram Key	Extract date	Component Name	Effective Date From	Effective Date To	Level 1	…	Level x
01-APR-2005-011200	01-APR-2005	Hospital A	01/04/2005	30/06/2006	HEALTH		Region A
01-JUL-2006-011200	01-JUL-2006	Pelonomi Hospital	01/06/2006		HEALTH		Region A
01-APR-2005-011333	01-APR-2005	AIDS (Unit)	01/04/2005	30/06/2006	HEALTH	….	Hospital A
01-JUL-2006-011333	01-JUL-2006	AIDS (Unit)	01/06/2006		HEALTH	….	Pelonomi Hospital

3.2.5 Using Materialized Views and SCD Type 2

According to Becker (2004), one of the problems of the SCD Type 2 technique is the large number of additional rows required to support all the changes. Barbusinski *et. al.*, (2003) pointed out that joining the fact and associated dimensions would also require complex temporal joins at analysis time. Furthermore the SQL statement must include time reference for both the fact and associated dimensions. All these factors will lead to an undesired environment for non-sophisticated users such as in the case of the FSDOH.

One possible way of overcoming these obstacles, is by using a materialized view (mview) to hide the complexity. A materialized view also physically stores the data that corresponds to the view's defined query (Dodge & Gorman, 2000). According to Goldstein & Larson (2001) query processing time can be improved through the use of materialized views.

For these reasons it was decided to make use of Oracle's materialized views. HR_DETAILS_DIM (mview) was created by joining all the posts with the matching staff member details. A staff member could also belong to more than one post. In order to uniquely identify a staff member with a particular post, a surrogate key called PERSONNEL_KEY was constructed for this purpose.

The PERSONNEL_KEY was constructed using a concatenated combination of the following fields from the posts and staff tables:

- EXTRACT_DATE (staff details)
- PERSAL_NUMBER (staff details)
- POST_NUMBER (post details)
- POST_TITLE (post details)
- COMPONENT (post details)

Thereafter, HR_FACT (mview) was created by joining FLAT_ORGANOGRAM_DIM (table) and HR_DETAILS_DIM (mview) to ensure consistency with all the SCD Type 2 changes in FLAT_ORGANOGRAM_DIM.

3.2.6 Building OLAP Cubes

The HRDM OLAP cube was constructed from the dimensional model (See Figure 4). This was all done using Cognos Framework Manager and Cognos Transformer Series 7.

Research done by Gorla (2003) to evaluate OLAP tools in ease of use and usefulness, suggested that MOLAP be used for non-sophisticated computer users and relational online analytical processing (ROLAP) for sophisticated users.

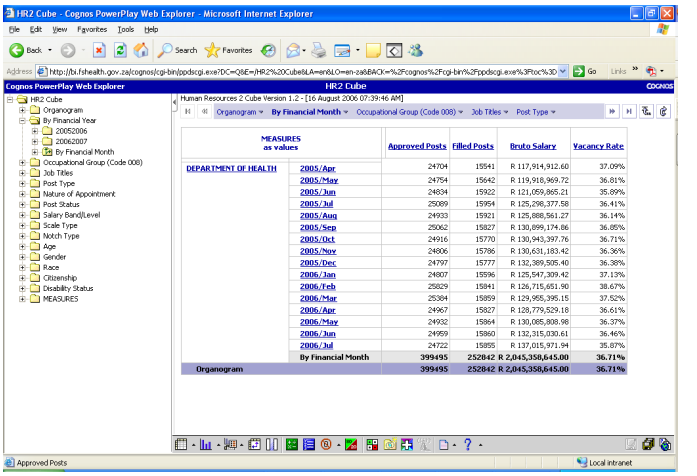
Since most of the users at FSDOH can be categorized as non-sophisticated computer users, the **MOLAP** architecture was the choice of platform. The cube was deployed using Cognos Enterprise Server Series 7 which delivers **Web-based OLAP (WOLAP)** content, but using an underlying architecture that is still MOLAP. According to De Beer (2006), WOLAP is also seen as the next generation BI tool providing "thin-client" viewing tools for analyzing information.

The users were able to generate pivot tables (See Figure 4) from the WOLAP cube to assist them in obtaining strategic human resource information.

4. CONCLUSION

Efficient resource management is critical for the success of the rollout of antiretroviral therapy in South Africa. Human resources management and ART drugs

Figure 4. Pivot table from human resource data mart



management was identified as the key factors but also the main cost drivers. A HRDM was built to provide strategic information for the ART programme. FSDOH management was now able to perform efficient staff allocations, monitor absenteeism and identify overworked personnel in time. Problematic ART clinics and hospitals in terms of staff turnaround could now also be easily identified by using trends, providing the FSDOH management team enough time to address the problem.

Future work and research could be done to link the HRDM to the ART clinical data set to identify health workers infected with HIV and AIDS. With this information, FSDOH management can obtain a better picture on the infection rate of HIV and AIDS on its health workers.

In conclusion, this paper demonstrated that it is possible to overcome the challenges of building a large-scale data warehouse, by starting small, using in-house knowledge and skills and to build data mart by data mart. The ETL process was modified to overcome the challenge of using SCD Type 2 within a hierarchical dimension. Materialized views were used to assist with the construction of the OLAP cube by camouflaging the complexities created by SCD Type 2. The end result was a MOLAP cube which provided an environment, conducive for analytical HR operations.

5. REFERENCES

1. Barbusinski L, Howard S, Jennings M, Kelley C, Oates J. (2003). The relationship between a fact and dimension table, DMReview.com http://www.dmreview.com/article_sub.cfm?articleId=6349
2. Becker B. (2004). Kimball Design Tip #53: Dimensions Embellishments. Retrieved September 3, 2006, from http://www.rkimball.com/html/designtipsPDF/KimballDT53Dimension.pdf
3. Berndt DJ, Fisher JW. (2001). Understanding Dimension Volatility in Data Warehouses (or Bin There Done That). *Sixth INFORMS Conference on Information Systems and Technology* (INFORMS/CIST-2001)
4. Bruckner RM, Tjoa AM. (2002). Capturing Delays and Valid Times in Data Warehouses – Towards Timely Consistent Analyses. *Journal of Intelligent Information* Systems, 19(2), p169-190. Kluwer Academic Publishers.
5. Chapman RD. (2003). Plan for Implementation of ARV's in the Free State Province. Unpublished.
6. Corr L (2001). Kimball Design Tip #17: Populating Hierarchy Helper Tables. Retrieved September 22, 2006, from http://www.kimballgroup.com/html/designtipsPDF/DesignTips2001/KimballDT17Populating.pdf
7. Davis X, Wan C, Ross L, Wen X & Thomas B. (2002). A data warehouse concept for HIV prevention program evaluation. *AIDS Education and Prevention: Official Publication Of The International Society For AIDS Education*, 14(3 Suppl A), p120-122
8. De Beer, E (2006). Mobilising BI with data quality. Computing SA, July 2006, p35

9. Dodge G, Gorman T. (2000). Essential Oracle8i Data Warehousing: Designing, Building and Managing Oracle Data Warehouses. Wiley. New York.
10. Department of Health (Free State) (2006). Corporate Strategic Plan 2006/2007 to 2014/2015
11. Gatziu S, Vavouras A. (1999). Data warehousing: Concepts and mechanisms. INFORMATIK Volume 1 of 1999
12. Goldstein J, Larson P. (2001). Optimizing Queries using Materialized Views: A Practical Scalable Solution, ACM SIGMOD, 2001
13. Gorla N. (2003). Features to consider in a data warehousing system. *Communications of the ACM,* 46(11), p111-115
14. Hristovski D, Rogac M, Markota M. (2000). Using Data Warehousing and OLAP in Public Health Care. *Journal – American Medical Informatics Association*, 7, p369-373.
15. Kimball R, Margy R. (2002). The Data Warehouse Toolkit. Wiley Computer Publishing. Second Edition.
16. Lau RKW, Catchpole M. (2001). Improving data collection and information retrieval for monitoring sexual health. *International Journal of STD and AIDS*, 12(1), p8-13
17. Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. (1997). Medical data mining: Knowledge discovery in a clinical data warehouse. *Proceedings of the 1997 AMIA Annual Fall Symposium, Bethesda: American Medical Informatics Association,* p101-105
18. Saraceni V, Da Cruz MM, Lauria LM, Durovni B. (2005). Trends and Characteristics of AIDS Mortality in the Rio de Janeiro City after the Introduction of Highly Active Antiretroviral Therapy. The Brazilian Journal of Infectious Diseases 2005, p209-215
19. Schubart J, Einbinder J. (2000). Evaluation of a data warehouse in an academic health sciences center. *International Journal of Medical Informatics*, p319-333
20. Shin B. (2003). An Exploratory Investigation of System Success Factors in Data Warehousing. *Journal of the Association for Information Systems,* Volume 4, p141-170.
21. UNAIDS.ORG. (2004). Report on Global AIDS epidemic, July 2004. Retrieved February 2, 2006, from http://www.unaids.org/bangkok2004/report_pdf.html
22. Vesset, D (2006). Worldwide Data Warehousing Tools 2005 Vendor Shares, IDC #203229, Volume 1. Retrieved December 1, 2006, from http://www.oracle.com/corporate/analyst/reports/infrastructure/bi_dw/idc-dw-tools-2005-1.pdf