World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:4, No:3, 2010

Dimensional Modeling of HIV Data Using Open Source

Charles D. Otine, Samuel B. Kucel, Lena Trojer

***Abstract*—**Selecting the data modeling technique for an information system is determined by the objective of the resultant data model. Dimensional modeling is the preferred modeling technique for data destined for data warehouses and data mining, presenting data models that ease analysis and queries which are in contrast with entity relationship modeling. The establishment of data warehouses as components of information system landscapes in many organizations has subsequently led to the development of dimensional modeling. This has been significantly more developed and reported for the commercial database management systems as compared to the open sources thereby making it less affordable for those in resource constrained settings. This paper presents dimensional modeling of HIV patient information using open source modeling tools. It aims to take advantage of the fact that the most affected regions by the HIV virus are also heavily resource constrained (sub-Saharan Africa) whereas having large quantities of HIV data. Two HIV data source systems were studied to identify appropriate dimensions and facts these were then modeled using two open source dimensional modeling tools. Use of open source would reduce the software costs for dimensional modeling and in turn make data warehousing and data mining more feasible even for those in resource constrained settings but with data available.

***Keywords*—**About Database, Data Mining, Data warehouse, Dimensional Modeling, Open Source.

I. Introduction

DATA models form the foundation of data warehousing and data mining systems since they help to describe how data is to be represented and acccessed. It is critical that the underlying data model correctly represent the data that is being studied [6], with accurate identification and representation of the required measures and variables. [2] notes that increasing development in the concept of information systems has resulted in interest in data models, since in essence data models form the blue print for the development of databases which is at the backbone of information systems. Database data models such as the flat model, hierarchical model, network model and the relational model have been suggested. The most common of these is the relational model with specific types such as the Entity relational model [1], the concept oriented model and the star and snow flake schemas for data warehouses. [17] refer to other variations of the Entity relationship (E-R) model such as the Multidimensional Entity relationship (ME/R) model, the EVER model and the StarER that combines the star model and the ER model.

A data warehouse is an all inclusive system that enables the extraction of data from different and often heterogeneous source systems [15] and their management in the 'warehouse' to provide user access and analysis. The accessed data can then be data mined for new information. [16] reports that multi-dimensional data model have proved to be most suitable for data warehouse applications. Muti-dimensional models for data warehouses are generated by using the dimensional modeling technique which is in contrast with entity relationship modeling which aims to generate models that ensure efficiency of record insertion and upates not retrievals like in the case of data warehouses. This fundamental difference in the architecture renders the retrieval of large number of records from E-R model based systems resource intensive and therefore not suitable for data warehousing and data mining that deals with the retrieval of large volumes of data at a time.

[8] notes that detailed guidance for dimensional modeling during the complex data warehousing information systems projects is lacking. Also, [8] indicate that the large and complex nature of data warehousing projects result in difficulties during the design stage. The design stage is made more complicated by the little guidance available for dimensional modeling, with literature available suggesting instead the models suitable for particular situations. Furthermore the dimensional modeling and data warehousing tools are more common, more developed and more documented and reported for the case of 'over the counter[1]' commercial softwares [15]. These softwares prove to be prohibitively expensive for a majority of information system developers who may wish to engage in data warehousing and data mining. This leads to a loss of opportunity for establishements who may have abundant and continuously growing data from taking advantage of data warehousing due to the high costs inolved, especially software costs.

Take the case of sub-saharan Africa, a region most affected by the HIV-virus [20]. This culminates into large quantities of

Charles D. Otine is an Assistant Lecturer at Makerere Universityy Faculty of Technology. (Phone: +256772-923585; fax: +256-414-; e-mail: hautine@tech.mak.ac.ug).

Samuel B. Kucel is a senior Lecturer at Makerere University Faculty of Technology Uganda. (email: sbkucel@tech.mak.ac.ug)

Lena Trojer is a professor at Blekinge Institute of Technology in Sweden (e-mail: lena.trojer@bth.se)

[4] Vendors such as Oracle, IBM and Microsoft have developed data warehousing and data mining in their Database Management system commercial tools

data on HIV infection but little is done to take advantage of this information with a bid to generate new knowledge using data warehousing and data mining. This is further hindered by the high cost of data warehousing and data mining tools availbale in the market and the little information on the cheap and free open source tools. This research paper looks at using open source data modeling tools in developing dimensional models for use in HIV patient data warehousing.

The use of open source is championed because of the high cost of 'off- the- shelf' data modeling and data mining tools and the limited literature on open source modeling tools.

II. DIMENSIONAL MODELING

Dimensional modeling is used to conceptualize data warehouses which are then implemented using star schemas or snow-flak schemas. It differs from Entity relationship (E-R) modeling that is used for ordinary transaction databases in that it aims to implement a database that eases user navigation [10], enhances performance [4] and interaction thereby improving analysis. Analysis of data in a data warehouse is key to data mining [6]; this is facilitated by the underlying warehouse data model. E-R modeling on the other hand aims to improve ease of understanding by users, enforce consistency and reduce redundancies in the data. With this architecture in mind E-R models are normalized to a large extent and therefore not suited for extensive and complex analysis of data.

Dimensional modeling helps to generate the star schemas. Star schemas are constituted by a fact-table in the centre surrounded by a range of dimensions (Figure 1). The fact table represents a concept of primary interest to the decision maker [5]. The fact table contains attributes known as measures that can be analyzed along different perspectives or dimensions. This assists in giving the data a multidimensional view [13]. Each of the dimensions that connect to the fact table in the centre of the model adds a primary key that acts as a foreign key and forms part of the composite primary key for a row of the fact table. One of the core dimensions of the star schema is the time dimension; this is used to give the information in the data warehouse a lifeline.

The data represented by star-schemas are extensively de-normalized with significant number of redundancies; this architecture improves analysis of the data. This is related to the fewer number of joins required to obtain the results of a query. The snow-flake schema may be interpreted as an extension of the star schema [14]. The reason for this is that the snow flake schema attempts to reduce on redundancies in its architecture by introducing a degree of normalization. Star schemas may be extended to snow-flake (star-flake) schemas when there is a significant increase in the number of rows for a dimension that would impede the performance of the data warehouse. It is due to this that during dimensional modeling, the dimension in question could be normalized to reduce the size of the resultant table in the data warehouse. [14] notes a third schema, the 3NF schema, but contends that it is possible to present the schematics of any application in either of the schemas.

The star schema is considered to be the most efficient design and is suited for modeling data marts. The snow-flake schema may suffer from potential performance issues from the relatively higher number of query joins needed as opposed to the star- schema. Star schemas with more than 1 fact table are commonly referred to as constellations.

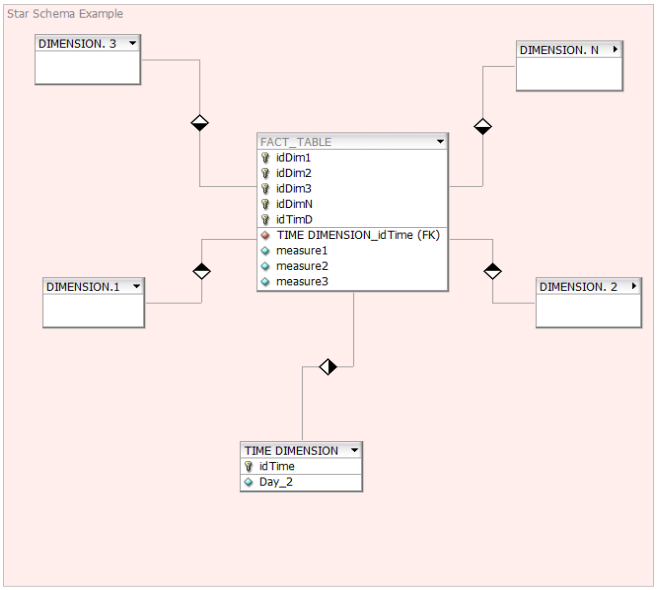

Fig. 1 Star Schema

III. OPEN SOURCE

The cost of commercial software is at times a stumbling block for information systems. This is a lot evident in the moderately young field of data warehousing and data mining. Commercial vendors of different database management systems have developed data warehousing and data mining capabilities for instance Oracle, IBM, and Microsoft. The costs of such software, especially license fees, render the acquisition process prohibitively expensive for the resource constrained settings.

The response to the above has been the development of open source software [3]. Areas that are resource constrained can take advantage of this to acquire information systems that would otherwise be considerably expensive for them. For the case of data warehousing and data mining the use of open source has been scarce and literature on the above limited. However several open source database management systems (DBMS) have come out to compete against the commercial versions. These, to mention a few, include MySQL, PostgreSQL, Firebird, Ingres and Berkeley DB; of this MySQL is by far the most successful.

They (open source DBMS) have however lagged behind in terms of dimensional modeling tools suited to these database management systems. In sections to come, the paper shall draw attention to two open source dimensional modeling tools that were sampled. A key factor for their consideration was the flexibility in terms of the database management system where the dimensional models developed by these tools could be implemented.

IV. HIV/AIDS

The HIV epidemic has affected the countries of sub-Saharan Africa both socially and economically. The HIV virus results in the destruction of the body's immune system rendering it unable to fight off opportunistic infections and therefore resulting in the condition AIDS. Although this has resulted in a large number of deaths it has also offered an opportunity in terms of the data available on HIV. The numerous advances in ICT (data warehousing and data mining) mentioned above can be used to put this data to use. Due to the economic situation in some of these countries like Uganda the use of commercial software, consequential maintenance and sustainability of these data warehousing endeavors may outweigh the benefits of the resultant system. It is important to research and identify alternatives to the high cost of commercial software in the data warehousing process, and why not at the early stages of the data warehousing process, that is, at the dimensional modeling stage.

This paper highlights the development of a dimensional model to support HIV data warehousing using open source. This was done using information from the Ugandan HIV scenario and assistance from different health care partners in Uganda dealing with HIV cases. The government of Uganda in a bid to improve access to antiretroviral therapy (ART) by the infected people has championed the provision of free antiretroviral drugs to patients at Health care centers. Some private non-governmental organizations have also taken the lead in supporting HIV patients. These organizations and government health centers provide support in the form of ART, voluntary counseling and testing, prevention of mother to child transmission of the virus, medical checkups and adherence monitoring for the patients. It is the information generated by these activities that form the basis of the data to be used for analysis.

V. MODELING PROCESS

A. Methodology

[12] and [11] recommend that collecting the objectives and requirements should be done by involving the end users. This is the case since organizations have a large spectrum of users with distinct needs to be addressed. Selected government and non government HIV health centers were visited and the professionals interviewed. The views of some prominent health care givers in HIV were sought.

The dimensional modeling process was articulated in the following phases after collecting the objectives and the requirements.

- Selection of the appropriate open source dimensional modeling tool(s).
- Analysis with selected sample of stakeholders to identify the HIV cares process to be modeled.
- Identification of the dimensions, hierarchies for each fact table.
- Identification of measures for the fact table.
- Verification of the technical system.

B. Selection of the Modeling Tool

Two open source modeling tools were identified. The emphasis was placed on modeling tools that allows for connectivity with several database management systems as well as enabling capabilities for database synchronization and reverse engineering. Synchronization allows the tool to generate the corresponding data warehouse dimensions and fact tables from the model directly into the database management system it has connected to. Once the dimensions have been generated in the data warehouse, it is not uncommon for changes to be made directly on the data warehouse. Changes such as definition of new dimensions; attribute additions or removal, new measures in the fact tables can then be directly generated onto the dimensional model; this is known as reverse engineering. The functionality of reverse engineering allows for modifications to the dimensional model from changes to the physical data warehouse dimensions and schemas in the database management system.

The two open source dimensional modeling tools studied were SQL Power Architect and DB designer. Both these tools allows for working with open source database backend as well as commercial database backend including commercial versions search as Oracle, SQL, DB2, and IBM. This would provide a huge flexibility for the dimensional modelers to choose whatever platform was more suitable to the data warehouse design problem in question.

Both tools allow for the definition of the appropriate dimensions with their respective attributes. The relationship and interactions between the different dimensions can then be defined with the appropriate cardinalities. Figure 2 and Figure 3 indicate the screens for the Power Architect and DB designer tools respectively with sample models.

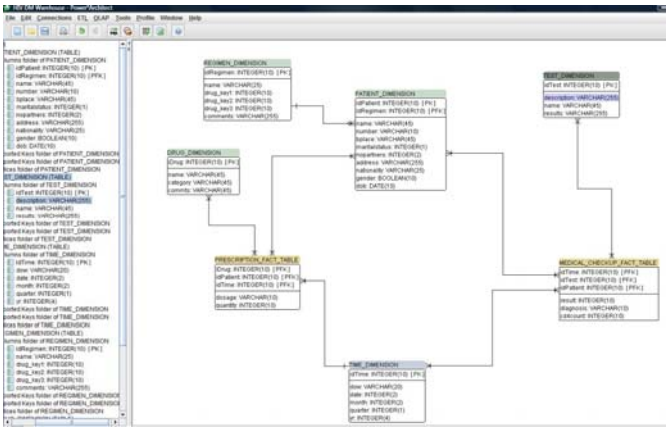

Fig. 2 Power Architect Modeling Screen

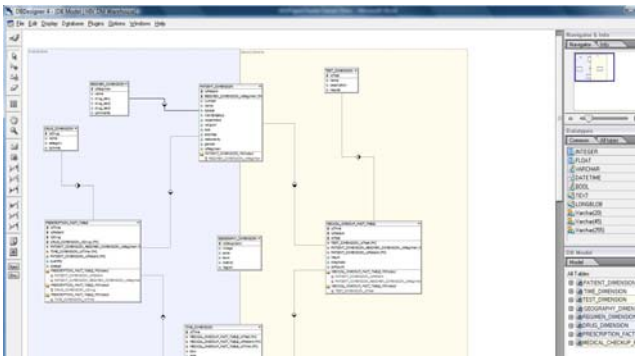

Fig. 3 DB Designer Modeling Screen

C. Identification of process to be modeled, dimensions, Hierarchies and Measures

[7] emphasize the importance of selecting the dimensions or features to be used by a data mining algorithm correctly. [11] argues that correctly identifying dimensions and interrelationships with facts is crucial to coming up with a model that correctly represents users data requirements and the analysis intended on the data. Features that are either irrelevant or unreliable may render the data mining process difficult and make the results complicated to analyze. The objective of dimensional modeling is to represent a set of measurements in a standard frame work. The idea behind this phase is to identify the key process of interest for the HIV health care providers and this was done after repeated consultative sessions with this target group. Analysis of some selected source system[2] from selected HIV health care providers was also done.

Two main processes of interest were noted the *i)* periodic medical check-ups and *ii)* Prescriptions to patients. Patients access antiretroviral therapy (ART) from different ART government distribution centers or other NGO assisted ART centers. ART is the treatment of HIV patients with pharmacological agents (antiretroviral drugs) that slow down the progression of the HIV virus in the body. In either of these centers; patients (from here referred to as clients) are given medical checkups at the onset of their ART treatment and periodically when replenishing their ARV drug supplies. Medical checkups may also be conducted in an ad hoc manner whenever the client experiences a relapse of any kind or at the discretion of the care giver. The second process is the prescription; this is given to a client by a physician or medical person in response to a medical checkup or diagnosis that has been done. It may involve the regimen (ART treatment option) that the client is on, or supplementary drugs to assist with opportunistic infections.

The warehouse would then be modeled to monitor the two processes or in this case 'facts' identified above. [8] describes dimensions as entities that are used for analyzing the measurements in the fact table. The dimensions identified include the patient, drug, regimen, test and the time dimensions. In summary the dimensions identified assist in keeping track of what patient underwent what medical checkup and the prescriptions that they were given at what point in time. Items that are being monitored include medical tests that have been done on the patient, the drugs given out as prescriptions and those that make up the patient's ART treatment regimen. The time dimension is necessary to keep track when each of the two processes of interest have been carried out for each patient.

A number of measures were identified for the facts represented by the two processes. The process medical checkup monitors the patients CD4[3] count, weight gain or loss, the tests for opportunistic infections, blood pressure, and pregnancy. The prescription process would monitor measures such as drugs given, the quantity and the dosage dispensed. Dimensional models are extensible because they allow for the addition of new data elements; new facts, dimensions and attributes can be added so long as they are consistent with the present facts. New measures for the facts can be added for increased analytical capabilities.

Figure 4 indicates the dimensional model generated. This is a constellation with two fact tables and conformed dimensions to enable comparison between the two main processes identified during this stage. This model was generated using the open source modeling tool DB designer. The geography dimension has been normalized from the patient dimension to form a hierarchy along which role up and or aggregation can be done during analysis. Aggregation or role up is done to provide summarized views of the data. It would be important to view the aggregated analysis of each of the two processes, for instance the average CD4 count value for patients in a geographical region on a treatment regimen and an alternate prescription. There are other interesting dimensions that can be analyzed against the time line and tests like the effects of administering ARVs over a period of time, based on the attribute *doPTest* (date of patient testing positive) in the patient dimension, attributes of time dimension and the different facts in the medical_check_up fact table.

The test dimension enables monitoring of not only the different opportunistic infections but also gives information about pregnancies that could assist with the prevention of mother to child transmission (PMTCT) of the virus. The PMTCT program reduces the risk of mother to child transmission of the virus. It is reported that in the absence of any intervention 15-30% of mothers with HIV will transmit the infection through pregnancy and delivery and others during breast feeding. The response to HIV studies has highlighted ways of reducing this risk one of which is the provision of ART for the HIV positive mothers and the new born babies. The dimensional model offers the opportunity for comparison and optimization on what regimen dimensions for expecting mothers would result in the significant reductions in the HIV virus basing on the test dimension and the largest increments in CD4 count indicated in the medical_check_up

[2] The source systems analyzed include the Adherence monitoring system at reach out Mbuya (an organization that specializes in ART for patients in a Kampala Suburb in Uganda) and Infectious disease institute information system (IDI) in Kampala Uganda.

[3] CD4 count is a measure of the strength of the human immune system. HIV continually kills CD4 cells; overtime the body may not be able to replace these lost cells.

fact table. The time dimension would also indicate the most opportune moment to begin the intervention and the progress that is being made during the intervention.

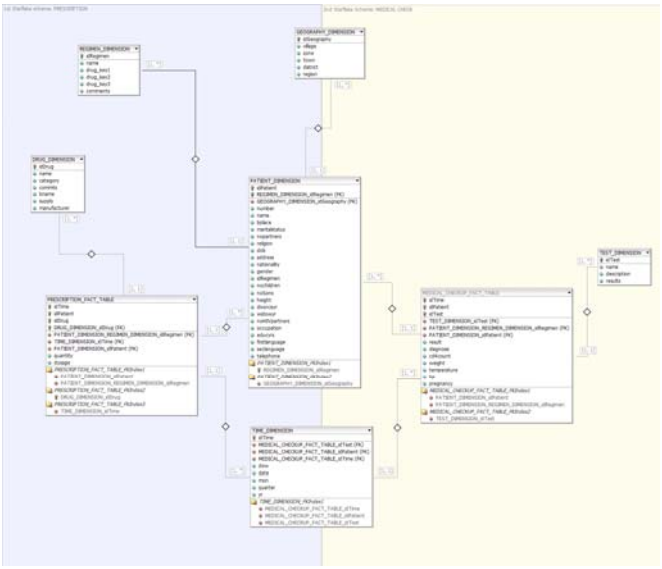

Fig. 4 Data warehouse Dimensional Model

D. Verification Technical System

The performance of a model is determined during verification. [9] highlights the different techniques of verification including; general good modeling and programming practices, verification of intermediate simulation outputs, comparison of final simulation outputs with analytical results and animation. The verification process involves checking that the schema is a correct model of the data warehouse. The attributes of the dimensions identified are meticulously cross checked for conformity to requirements, and conformity to the two source systems that were selected for analysis during modeling.

The open source tools selected were flexible in that they offer connectivity with different database management systems both commercial and open source. Part of the verification process was done by connecting to two database management systems MySQL and PostgreSQL as well as trial version of a commercial database Microsoft SQL (MSSQL). The tool allows for the defined models to be generated in the corresponding database management system (synchronization), this was successfully done in all the three database management systems selected.

The reverse engineering aspect was also tested altering the generated data warehouse tables in the different database management system. The changes were successfully reflected in the respective models in the modeling tools.

VI. Conclusion

This paper reported on the use of open source tools in building dimensional models for HIV patient information, with the long term result of implementing an HIV patient data warehouse. This is in a bid to reduce on the impact of the high cost of commercial dimensional modeling tools and database management systems in the market and to take advantage of the cheaper open source tools available and the data available on HIV in regions of sub-Saharan Africa such as Uganda.

Star schemas generated using dimensional models are flexible in that they allow for modifications as the data warehouse grows from different data marts organized around the key processes. This is a good property as the dimensional model for HIV patient data would be envisioned to grow as new processes of interest are identified and added to the schema with allowance for new dimensions and hierarchies. This can be done through additions of other new data marts analyzing new processes that are identified with time.

The open source dimensional tools have a weakness in handling complex data types as compared to various new tools that have been researched on and incorporated into some commercial database management systems. These are capabilities to handle complex data types as indicated in [18] and [19]. This would improve on analysis of complex data in open source systems such as patient x-ray screens, brain scans and heart scans. This is still lacking in the open source domain of data warehousing

VII. Conclusion

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

Acknowledgment

We would like to acknowledge the assistance of Sida/SAREC, the Swedish research cooperation, Makerere University (Faculty of Technology), Uganda and Blekinge Institute of Technology, Sweden for all the assistance rendered.

References

[1] Chen, P. (1976). The Entity Relationship model-Towards a unified view of data, ACM Transactions on Database Systems, 1, 1, 9-36.

[2] Chilton, M.A. (2006). Data Modeling Education: The changing technology, Journal of Information Systems Educaion, 17,1, 17-20.

[3] Coar, K. (2006). The Open source Definition , Retrieved on 18th Nov 2008 from opensource.org: http://www.opensource.org/docs/osd

[4] Dash, A.K and Agarwal, R. (2001). Dimensional modeling for Data warehouse, ACM SIGSOFT software engineering notes, 26, 1, 83-84.

[5] Golfarelli, M., Maio, D. and Rizzi, S. (1998). Conceptual Design of Data warehouses from E-R schemes, Proceedings of the Hawaii International Conference On System Sciences, January 6-9, Hawaii

[6] Gui, Y., Tang, S., Tong, Y. and Yang,D. (2006). Tripple Driven Data Modeling Methodology in Data warehousing: A case study, ACM workshop on Data warehousing and OLAP, 59-66

[7] Ilczuk, G. and Wakulicz-Deja, A. (2007). Selection of Important attributes for Medical Diagnosis Systems. *Transactions on Rough Sets* , *7*,1, 70-84.

[8] Jones, M. E. and Song, I.Y. (2008). Dimensional modeling: Identification, classification and evaluation of patterns. *Decision Support Systems* , 59-76.

[9] Kleijen, J. P. (1995). Verification and validation of simulation models. *European Journal of Operations Research* , *82*,1, 145-162.

[10] Kortinik, M. A. and Moody, D. L. (2003). From ER Models to Dimensional Models: Bridging the Gap between OLTP and OLAP Design. *Business Intelligence Journal* , 8,3, 1-17.
[11] Laender H. F., Freitas, G.M., and Campos, M.L. (2002). MD2- Getting Users Involved in the Development of Data Warehouse Applications. *4th International Conference Workshop Design and Management of Data warehouses.* May 27, Toronto, University of British Columbia, 3-12.
[12] Lambert, B. (1995). Break Old Habits To Define Data Warehousing Requirements. *Data Management Review* .
[13] Malinowski, E. and Zimanyi, E. (2007). A conceptual model for temporal data warehouses and its transformation to the the ER and object-relational model. *Data and Knowledge Engineering* ,64, 101-133.
[14] Martyn, T. (2004). Reconsidering Multi-Dimensional Schemas. *ACMs Special Interest Group On Management of Data , 33*,1, 83-88.
[15] Nguyen, T. M., Tjoa, A. M., and Trujillo, J. (2005). Data Warehousing and Knowledge Discovery: A Chronological View of Research Challenges. *Springer* , 530-535.
[16] Pearson, W. (2008, 1 24). Dimensional Model components: Dimensions part 1. Retrieved 11 19, 2008, from Database Journal: http://www.databasejournal.com/features/mssql/article.php/3723311/Dimensional-Model-Components--Dimensions-Part-I.htm
[17] Phipps, C. and Davis, K.C. (2003). Automating Data warehouse conceptual Schema Design and Evaluation. Proceedings of the 4th international conference on Design and Management of Data warehouses. May 27, Toronto Canada, 23-32
[18] Pokorny, J. (2003). Modeling stars using XML.
[19] Riadh, B. M., Omar, B., & Sabine, R. (2004). A new OLAP Aggregation Based on the AHC Technique. DOLAP (pp. 65-71). Washington,DC: ACM.
[20] UNAIDS. (2008). 2008 Report on the Global AIDS epidemic. Geneva: *WHO Library Cataloguing-in-Publication Data.*