THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Building the Central Regional Education Data Warehouse

Henry Techie-Menson

Lecturer, Department of Mathematics and ICT,
OLA College of Education, Cape Coast, Ghana

Paul Nyagorme

Senior Lecturer, Department of Mathematics and Science,
University of Cape Coast, Cape Coast, Ghana

Abstract:

This paper seeks to build a data warehouse for the central region directorate of the Ghana Education Service (GES). The project deployed the top-down approach by relying on the hub and spoke architecture based on a Joint Application Development (JAD) session organized to elicit user requirements. Again, dependent data marts, as well as the star schema, were chosen due to their uniqueness and appropriateness for the project. There were, however, some shortcomings involving data quality, efficiency, user acceptance, and cost. The data warehouse project, however, helped to improve management decision-making in terms of tracking, quality and efficiency in resource and logistics use, as well as predicting growth in the near future.

Keywords: Datawarehouse, data mart, data staging, online analytical processing

1. Introduction

The data warehouse supports online analytical processing (OLAP), this allows high-level end individuals to experience insight into the company's activities through immersive and iterative access to the information stored (Matei, Chao, Godwin, 2015). Through the querying of the data warehouse to analyze business processes, results and patterns, organizational executives will optimize corporate strategy and tactical decision-making (Gonzalez, 2006).

The data warehouse can be used to perform the following tasks at the Ghana Education Service regional directorate:

- Manage and Track performance.
- Track and change the in-and-out activities of the region's teaching and non-teaching staff.
- Optimize, review logistics and operations
- Increase inventory management, production performance and effectiveness.
- Query, enter and view fragmented data obtained from various sources.
- Managing and strengthening partnerships with the key stakeholders.
- Projected future development, specifications and deliverables.
- Cleanse and maximize the accuracy of data from the company.

There are several advantages of operating and using a data warehouse efficiently. Separating the theoretical processes in a data warehouse from the operating processes and transactions will boost the efficiency of all fields from an IT perspective (Wang, Luo & Yuan, 2018). A data warehouse platform can provide a realistic way to interpret the history from a corporate viewpoint without disrupting the company's everyday operations (Ghrab, Romero, Skhiri, Vaisman & Zimányi. 2015). It enhances operations and facilitates more productive market processes by querying and reviewing data in the data warehouse, thus increasing efficiency and achieving operational goals.

2. Building Central Regional Education Data Warehouse

The GES regional directorate has some core functions, which include strategic decision-making with strategic knowledge as a requirement. The regional directorate core functions including but not limited to the following:

The Cape Coast regional educational directorate is responsible for:

Oversight responsibilities of all Pre-tertiary schools; second cycle educational institutions (both public and private) which include 56 (Asare & Nti, 2014) Secondary/Senior High Schools, Senior High/Technical Schools, Technical/Vocational Schools; and Basic Schools (JHS, Primary schools, Kindergarten, Nursery, and/or Crèche) in the region. These include;

- The number of Schools, both public and Private
- Staff strength (both teaching and Non-teaching) of all schools
- Professional and/or academic qualifications of each staff
- Infrastructure quality and quantity of the schools
- The 21 (Asare & Nti, 2014) district directorates of education

- Staff movement, in-and-out transfers of staff
- Staff recruitment for the various schools
- Availability and conveyance of logistics
- Monitoring and evaluation of teaching and learning in schools

2.1. Analysis of Data Sources

The existing data sources at the regional education office are described as follows:

- Record of teaching and non-teaching staff across all metropolitan, municipal and district education directorates.
- Records of academic and professional qualification of teaching and non-teaching staff.
- Records of student's placement or admissions in various schools as well as their gender distribution
- Information on students' performance under the metropolitan, municipal and district directorates.
- Information on quizzes and competitions with their respective winners.
- Data on scholarship and sponsorship packages, their requirements and beneficiaries
- Data on inter-schools' athletics and games competitions and their winners
- Information on the organization and distinguished sports personalities discovered
- Data on criteria for selection into the district and regional teams
- Punctuality and regularity of teaching and non-teaching staff
- Data on student absenteeism and dropout rates with a gender distribution
- Records on lesson note preparation and delivery
- Records on Teaching Learning Materials (TLM) available and quantities needed
- Records on visits by supervisors and the adherence to code of ethics
- Records on appointments, transfers, promotion, demotion, reinstatements of teaching and non-teaching staff
- Information on study leave applicants and those under study leave
- Data on the workshops, in-service training, counseling programs undertaken within the various districts
- Records on expenditure and the disbursement of funds
- Audit report on departments and schools
- Data on physical structures, those under construction and those that require renovation
- Information on various facilities like laboratories, vehicles, busses etc.

2.2. Data Warehouse Requirements

The next step in our approach to building a regional education warehouse is to allow users to put in their requirements. Joint Application Development (JAD) discussion sessions were held to extract the requirements from the anticipated stakeholders and users. Key stakeholders included the Regional Director of Education, the various District Directors of Education, teacher unions, teaching and non-teaching staff within the Central Region. It is divided into functional and non-functional specifications in order to illustrate their development goals.

2.3. Functional Requirements

Functional requirements describe what the proposed Regional Education Warehouse should do or the behaviors of the system and below are the requirements it should satisfy.

- Data visualization: Through the sketch of dynamic charts that transform according to the queries and the values resulting from the change and most importantly in order to represent the results of end user requests, various types of charts will be used.
- Data browsing and dynamic queries: The targeted data warehouse should promote a quick and seamless response to complex changes in user queries. Using OLAP tools/ instruments will fulfill this.
- Data integration: Integrating data from many sources has become one of the key functions of data warehousing.
- Data validation: It is important to verify and validate the loaded data in the data warehouse, so the values are considered to be logical and accurate.
- Remote access: To allow the end user (top management) to access the data warehouse through the internet through the upload of the data warehousing cubes on a website.
- Security: By using various authentication levels, the protection aspects can be accomplished. The data warehouse will be accessed and browsed by each user according to his or her level of authority.

2.4. Non-Functional Requirements

A non-functional requirement places constraint on how the proposed Regional Education Warehouse should work or elaboration of the performance characteristics of the system. Below are the requirements it must satisfy:

- Simplicity in the sense that usage and learning of the data warehouse can be easy and straightforward and does not require any specific ability.
- Secondly, owing to the improvements in the customer specifications, the data warehouse should support the upcoming changes and the planned upgrades.

3. Regional Education Data Warehouse Architecture

The main steps for building the data warehouse prototype are shown in Figure 1. The data warehouse would be coherent and well aligned with the top-down strategy. The Hub and spoke architectures will ensure that a broad view of atomic level data is retained in the warehouse which can help to create subject by subject and an educational institution-wide view (hub) (Beheshti, Benatallah & Motahari-Nezhad, 2016). The overwhelming majority of people are going to access the data from data marts with dependent dimensionally modeled data (spokes). There are three main areas or stages within this architecture. Until it is incorporated and deposited in the production data warehouse for further analysis, data is imported from multiple sources and transformed inside a staging area. The reasons for utilizing the Huband Spoke architecture include;

- Its ability to addresses the need for dependent data marts
- Marts receive data from the central source, that is the warehouse
- It offers medium and large contexts
- It is scalable a most often enterprise-wide
- Sometimes serves as a corporate information factory

Figure 1: Data Warehouse Architecture

3.1. Data Acquisition

This covers the whole method of collecting data in an access database from data sources, hardcopy files on enrollment, academics, sports, monitoring and supervision, human resource, finances and budgeting then transferring all extracted information to the staging area and organizing the extracted data for loading into the repository of the data warehouse. Apart from moving data from the above OLTP systems, it can also be used to move data from one data warehouse to another (Challal, Boussaid & Boukhalfa, 2017). The three distinct functional elements of Extraction, Transformation and Loading for the Central Region Educational warehouse is explained below.

3.1.1. Extraction

A replica of the initial data was made during the extraction process, preserving the original data in the source system. In order not to place processing burden on the restricted legacy systems at the different educational offices, this choice was chosen. For the necessary manipulation, extracted data is then transferred into the staging area via ODBC, OLE DB and other APIs. In data extraction procedures, SQL stored processes, Data Transformation Services (DTS) as well as other programming and scripting languages were used.

3.1.2. Transformation

Data transformation was conducted to verify, ensure consistency, conversion of data type and the implementation of business rules for all educational directorate data sources. (Winter, Winter & Strauch, 2003). In addition to the complex application that is often needed when dealing with sophisticated business rules, simple tools containing SQL stored procedures were used (Lujan-Mora & Trujillo, 2006). Data validation ensures that the rows in the tables of facts and dimensions adhere to maintain data integrity. Conversion of the data form provides the means to ensure that values are processed in the data warehouse precisely regardless of how they were stored in the data source. Application of business

rules helps to comply with rules set by organizations that are imposed on records stored in the data warehouse. For instance, ensuring that all teaching and non-teaching staff have **StaffId** and **Date of First Appointment** fields.

3.1.3. Loading

After collecting and transforming data from various sources, data loading puts data into the data warehouse database dimension and fact tables. For the Regional Education Data Warehouse, Bulk Insert, as well as Bulk Copy API, was used to load data into the warehouse

3.1.4. Metadata

With the support of a personalized application which automatically updates it, the metadata for the Regional Education warehouse is extracted from the extraction, transformation and loading process. The approach requires the identification of data through the ETL structures and also the efficacy of the activities of the ETL. Accurate data mapping records, pattern prediction, automatic administrative management and reusable code are provided through metadata logging.

Data Lineage, Schema Changes, Transition Statistics and User Statistics were the metadata used to analyze the operation and output of the Regional Education warehouse. Data Lineage metadata contains data about the time the data warehouse was filled with a specific collection of documents. Metadata on Schema Changes contains details on changes made to tables. Metadata from Transformation Statistics provides details on the time taken for operation at the transformation points, as well as the executed rows and the last time for a transformation task. Metadata from User Statistics offers data on the query periods for report generation.

3.1.5. Data Storage

The phase of loading the transformed data from the staging area into the data warehouse repository is covered by the data storage stage. In the data staging area, all processes for transformation and integration between the access database and data source files are performed.

At this stage, there are a set of functions and services such as:

- Load Regional education data into data warehouse tables.
- Optimize the loading process.
- Perform incremental loads at regularly prescribed intervals.
- Support loading integrated data into multiple tables.

3.2. Data Marts

Data mart is a subject-oriented data in data warehouse which comprise of relevant data for a particular business function (marketing, sales, and finance) (Kämpgen, O'Riain & Harth, 2015). The Regional Education Warehouse is based on a secure data-mart infrastructure in which data is collected from a central data warehouse that has already been developed. It is a small data warehouse that serves the needs of a smaller community of users (Jukic, 2006). The data marts are, as shown by this method, viewed as sub-sets of a data warehouse (Jakawat, Favre & Loudcher, 2016). The explanation for this option is that the method of extraction simplifies the transition due to the centralized loading of summarized clean data into the data warehouse. This will help achieve improved performance and availability of lower communication costs. Scalability in terms of warehouse expansion and the management of massive databases is the main downside. The data marts for the warehouse are;

- Enrollment data mart
- Sports data mart
- Supervision data mart
- Human Resource data mart
- Finance data mart

3.2.1. Schema

A The Regional Education Warehouse chose a star schema in which a single object fact table is related to other tables described as dimensions (Kimball & Ross, 2013). The primary key in each of the dimensions is related to a foreign key in the fact table. The fact itself includes a collection of measurements that reflect various metrics of numerical analysis. The reasons for the star schema are;

- It's trouble free to design, set up, execute and understand
- Many other dimensions may be adapted later.
- Adequate space to tackle data redundancy
- Supports faster performance for lookups and ETL processes.
- Quick access to numerous aggregation levels

Figure 2: The Star Schema for Central Region Education Data Warehouse

3.2.2. Fact

The data in a fact table is always known and is numerically represented (Jakawat, Favre & Loudcher, 2016). In the Regional Education Warehouse, the fact tables apply to the dimension tables in which the aggregation of data is carried out using counts, summaries, minimum values and maximum values. Without returning to the base, transaction-level data, the lower level of granularity was enforced to influence how much users can drill down.

3.2.3. Dimension

Dimensions are text-based attributes that are related. Dimensions are the viewpoint entities that institutions would like to hold information with respect to (Abello, Samos & Salto, 2001). In the Regional Education Warehouse, the dimensions include student dimension, staff dimension, Workshop dimension, funds disbursement dimension, study leave dimension, payments dimension, competition dimension, staff eligibility dimension, district schools' dimension, facilities dimension and monitoring dimension. The staff dimension, for instance, had attributes such as Staff Id, Staff Name, Date Of Birth, Qualification, Appointment Date, School Id and Rank.

3.2.4. Measure

A measure is a numerical feature of a fact and determines an element of quantitative fact that is important to the analysis (Denis, Ghrab & Skhiri, 2013). The Regional Education data warehouse is based on a measure such as a school performance, salaries, staff promotion, school enrollment, scholarships budgets and audits

4. Information Delivery

The information delivery aspect enables users to access the data directly from the data warehouse for strategic decision-making. In this Regional Educational data warehouse, the star schema model was adopted, based on a fact table and multiple dimensions tables. Using the data cubes, decision makers conduct analysis in the MDDBs.

At this stage, there are a set of functions and services such as:

- Allow directors and decision makers to browse data warehouse content.
- Enable queries of aggregate tables for faster results and analysis.
- Provide multiple levels of data granularity.
- Monitor directors and decision-making access to improve service and for future enhancements.

4.1. Challenges in Constructing the Central Regional Education Directorate Data Warehouse

The process involved in constructing the Regional Education data warehouse had several shortfalls which were categorized under data quality, performance, user acceptance and cost. Since data comes from several different sources from all of the institution's divisions, it is important to integrate them. It experiences bugs, duplicates, logic clashes, and incomplete data while attempting to merge conflicting data from disparate sources, all resulting in problems with data quality. The low quality of data results in incorrect reporting and analytics needed for proper decision-making.

Secondly, because the data warehouse was constructed or built to meet overall performance criteria, there were performance problems. Whereas the final product may be tailored to meet the education unit's performance requirements, due to scalability problems, the desired results cannot be reached.

Again, because people are not interested in altering their daily schedule, there have been some problems with user acceptance, particularly if the new process is not intuitive. This posed a challenge at the onset due to the lower-level requirement elicitation inputs which may affect or influence the rate of adoption by the institution.

Finally, Building the Central Educational Data Warehouse was costly, refuting to some degree the common belief among educational units that they could build in-house data warehouses to save money. There are many hidden problems in building a data warehouse in which cost is a factor. Even if an education unit incorporates a data warehouse specialist to its workforce size, with only a few qualified practitioners, the level of knowledge and skills necessary to achieve a successful outcome are simply not feasible. In terms of cost, the hiring of experts coupled with the required logistics is enormous for an educational warehouse.

5. Conclusion

In this paper, an overview of the design and implementation of a data warehouse for the central region directorate of education has been presented. Because deployment is really the first step, comprehensive maintenance progress is needed to provide a long-term solution to strategic decision-making. In order to support information delivery, future works should increase functionality accompanied with more analytical resources. Due to time constraints, we decided to build the data warehouse only one region, Central Regional Education Directorate, instead of the Ministry of Education or Ghana Education Service, which will need strategic information for strategic decision making.

6. References

- i. Abello, A., Samos, J., Saltor, F. (2001). Understanding facts in a multidimensional Object Oriented Model. In Proc. of the 4th Int. Workshop on Data warehousing and OLAP (DOLAP), pages 32 -39, ACM press, 2006.
- ii. Asare K. B. and Nti, S. K. (2014). 'Teacher education in Ghana: a contemporary synopsis and matters arising,' SAGE Open, vol. 4, no. 2, 2014
- iii. Beheshti, S., Benatallah, B., & Motahari-Nezhad, H. (2016). Scalable graph-based OLAP analytics over process execution data. Distributed and Parallel Databases 34, 3 (2016), 379--423. DOI: <https://doi.org/10.1007/s10619-014-7171-9>
- iv. Challal, Z., Boussaid, B., Boukhalfa. K. (2017). Minimizing negative influence in social networks: A graph OLAP based approach. In Proceedings of the Database and Expert Systems Applications. 378--386. DOI: https://doi.org/10.1007/978-3-319-64471-4_30
- v. Denis, B., Ghrab, A., & Skhiri. S. (2013). A distributed approach for graph-oriented multidimensional analysis. In Proceedings of the 2013 IEEE International Conference on Big Data. IEEE, 9--16. DOI: <https://doi.org/10.1109/BigData.2013.6691777>
- vi. Ghrab, A., Romero, O., Skhiri, S., Vaisman, A., & Zimányi, E., (2015). A framework for building OLAP cubes on graphs. In Proceedings of the Advances in Databases and Information Systems. 92--105. DOI: https://doi.org/10.1007/978-3-319-23135-8_7
- vii. Gonzales, M. L., (2006). Hands-On OLAP, Seattle, WA: The Data Warehousing Institute.
- viii. Jakawat, W., Favre, C., & Loudcher, S., (2016). Graphs enriched by cubes for OLAP on bibliographic networks. International Journal of Business Intelligence and Data Mining 11, 1 (2016), 85. DOI: <https://doi.org/10.1504/IJBIDM.2016.076435>
- ix. Jukic, N., (2006). 'Data Modeling Strategies and Alternatives for Data Warehousing Projects,' Communications of the ACM, 49, 4, pp. 83-88
- x. Kämpgen, B., O'Riain, S. & Harth. A. (2015). Interacting with statistical linked data via OLAP operations. In Proceedings of the Semantic Web: ESWC 2012 Satellite Events. 87-101. DOI: https://doi.org/10.1007/978-3-662-46641-4_7

- xi. Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. John Wiley & Sons.
- xii. Larson, D., Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *Int. J. Inf. Manag.* 2016, 36, 700–710.
- xiii. Lujan-Mora, S.; Trujillo, J. Applying the UML and the Unified Process to the Design of Data Warehouses. *J. Comput. Inf. Syst.* 2006, 46, 30–58.
- xiv. Matei, A., Chao, K., & Godwin, N. (2015). OLAP for multidimensional semantic web databases. In *Proceedings of the International Workshop on Business Intelligence for the Real-Time Enterprise*. 81--96.
DOI:https://doi.org/10.1007/978-3-662-46839-5_6 Google Scholar Cross Ref
- xv. Winter, R., Winter, R.; Strauch, B. A Method for Demand-driven Information Requirements Analysis in Data Warehousing Projects. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, Big Island, HI, USA, 6–9 January 2003.
- xvi. Wang, J., Luo, S. & Yuan. J., (2018). Analysis of computer network and communication system. *Journal of Networking and Telecommunications* 1, 1 (February 2018), 507--550. Retrieved from <http://systems.enpress-publisher.com/index.php/JNT/article/view/228/217>