

Designing a new Applications of Data Warehousing Using 3-tier Architecture for Petroleum Industry

Dr. D.V.S.S.Subrahmanyam, Professor,
Dept of CSE,
Sphoorthy Engineering College.

Abstract-In this paper we present three tier architecture of a middleware that addresses challenges facing accessibility, availability, and consistency of data in mobile environments. The architecture supports the automatic hoarding of data from multiple, heterogeneous sources into possibly a variety of different mobile devices. Organizations be it industry or business or even educational institutes, need to improve their information inventory system so as to survive in the competitive environment. The Organizations have to increase their efficiency and effectiveness in maintaining the cycle of activities, in their planning, decision-making processes, and analytical needs. There are several ways to acquire this goal; one of it is with data mining which is able to make a prediction using existing data in their database in order to forecast future demand. In addition, with data mining they would be able to determine which activity is more important and what trend is prevailing. An information system, which is based on both World Wide Web technology and a 3-tiered architecture, is proposed herein to meet the above requirements. This paper is an attempt to provide the initial concept about data mining model that most likely will be used in various department including libraries of the teaching institutes. The initial concepts covered by the paper are the appropriate data warehouse schema; data mining tasks and techniques that are best suited, and applications.

And also In this paper we investigated the various data capturing systems which they currently in use at the Ghana National Petroleum Cooperation (GNPC) – a state agency responsible for the overall management and supervision of the hydrocarbon endowment of the sub region and recommending a tree tier level petroleum data warehouse architecture for cutting- edge decision making on petroleum resources.

Keywords-Data warehouse architecture, Petroleum data warehousing, Ghanaian petroleum industry.

INTRODUCTION

Ghana came into the spotlight in 2007 for its first ever discovery of hydrocarbon in commercial quantities. According to reports, the country is set to produce 1.4million gallons of crude oil per year, shooting the country's wealth up by a warping 4 billion dollars a year. Economic growth is also estimated to increase more than 13% per annum [1]. All these economic developments in the country raised the issue of how all these plans have been made without using a reliable data. As a matter of fact that for any country to achieve international reputation should depends greatly on the pro-activeness of its government and its citizens, especially the research reports to be able to anticipate the ramifications of their decisions and to put in place appropriate contingency plans to

mitigate them. However this cannot be done without using the right information carefully formatted, faultlessly triangulated, craftily and rightly positioned at the right place relative to each other for cutting edge decision making. Quality data and vibrant is the key to success of this fast growing industry. The importance of data is once again captured by the fact that the organization makes approximately about US\$2 million from data licensing per annum and presently provisions are being made to embark on a US\$10 million project that would see to the establishment of an online four levels architectures data store (the so termed data farm) supported by a meta data for faster and easier access to information. In this paper we looked at how data warehousing can be integrated into this data farm for efficient and effective extraction of necessary data for tactical decision making.

This paper presents a new data warehouse framework for the Ghanaian petroleum industry, Ghana is a relatively small but powerful country in sub-Saharan Africa – the republic of Ghana which recently idiomatically struck gold fortunes on its sea beds in the form of hydrocarbon deposits. In this context, some of the key questions that can be asked; what are the government's plan for continual discovery, exploration and preservation of this resource for the citizens of Ghana, and if there are any, are these plans made based on readily available, factual and highly summarized data or they are the decisions of the pre-computer age. Data warehouses are said to provide artificial support for management decision making but can this be of any use to the Ghanaian petroleum exploration industry. This paper seeks to provide answers to these questions by investigating the various data capturing systems currently in use at the Ghana National Petroleum Cooperation – a state agency responsible for the overall management and supervision of the hydrocarbon endowment of the sub-region and recommending petroleum data warehouse for cutting- edge decision making on petroleum by government. Interviews supported by some level of participant observation methods are adopted for gathering information in this paper. Transcription was the main analysis tool utilized in identifying themes, trends and building interpretations around them.

The term data warehouse just like any other term in the professional domain has received many scholarly attentions over the years with criticisms and counter criticisms by many database technologists. However, Inman in 1993 has it that “a data warehouse is a subject oriented, integrated, time variant and non-volatile collection of data in support

of management's decision making" [2]. This is not a one stop definition since many scholars have varying views about what should be the embodiment of the definition for the term and what should not. However from the above definition of a data warehouse it is lucid clear that the definition of the term data warehouse should rest on four main building blocks which I would comfortably want to refer to as the four "defi-pillars" of data warehousing – that is its non-volatility, subject-orientation, integrated and it's time-variant.

The concept of petroleum data warehousing has been written about by many professionals. The following are some of the academic works done in the domain and for an exhaustive digestion of the topic, it is recommended that this paper be read in conjunction with some of these academic papers as listed below,

- Mapping and modeling of oil and gas relational data objects for warehouse development and efficient data mining & Onissues of data warehouse architectures – managing Australian resources data [3].
- Building the data warehouse [2].
- Research on the application of data-mining for quality analysis in petroleum refining industry [4].
- The application of spatial data warehouse technology in the national petroleum resources database [5].

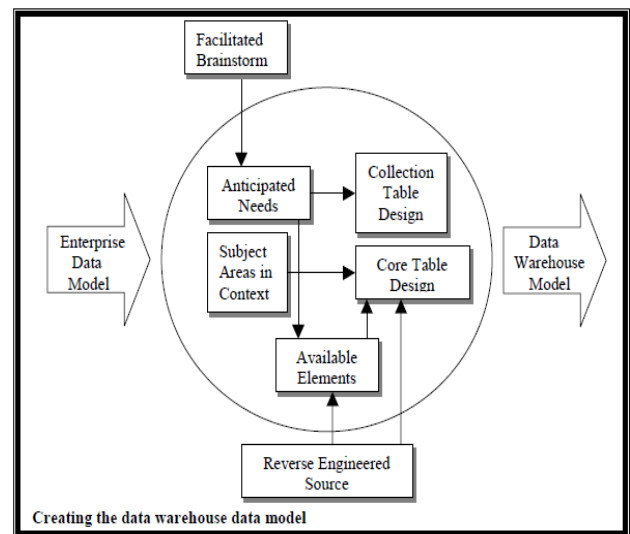
Data warehousing involves the polling together of data from various data islands and operational data bases and storing it in a persistent, integrated, subject specific and time variant format for an informed strategic management decision making [2]. On the business front, It is always promoted by developers and other major stakeholders as having the capability to revolutionise businesses by transforming management decision making process in terms of its speed, precision, reliability and real-timeliness. This task is virtually unachievable by relational database management system built for online transaction processing due to its rigidity and inability to offer the required multiple dimensions that would intend bring out aspects for OLAP analysis [7]. In an investigative approach to model the Australian resource data, Nimmagadda and Dreher [3, 6] provided a mapping and modeling technique for extracting and preparing data and offered an exhaustive methodology on how this is then integrated into a framework to form the Australian petroleum resource data warehouse. In a related work by Ying, Yang, Ya-fu and Li-xhou [5], generic data warehouse approaches were utilized to achieve a spatial data warehouse design by systematically illustrating how the various enterprise-wide data is coordinated into a single repository and multi- dimensionally modelled into a high level decision making tool. However despite the enormity of research work in the subject domain, many of these works have not been able to provide a logical approach and methodology on how the data warehouse can be integrated into the global information system for rapid, economical and cost saving development while ensuring that the existing information flow channel is not compromised. In this paper we provide an approach to achieve this by introducing the three tier level data warehouse architecture in the Ghanaian petroleum exploration information management system.

Data Warehouse Services:

- Acquisition – bringing data into the warehouse from the operational environment.
- Preparation – packaging the data for a variety of uses by a diverse set of information consumers
- Navigation – assisting consumers in finding and using warehouse resources
- Access – supporting direct interaction with data warehouse data by consumers using a variety of tools
- Delivery – providing agent-driven collections and subscription information delivery to consumers.

Data Warehouse Design

A data warehouse model must be comprehensive, current and dynamic, and provide a complete picture of the physical reality of the warehouse as it evolves. The data warehouse database schema should be generated and maintained directly from the model. The model not just a design aid, but an ongoing communication and management tool. Managing the dynamic nature of a data warehouse is one of the most daunting challenges. New feeds must be supported from new and existing sources. New configurations of data must be packaged with tools to meet continually changing needs. Under-utilized datasets must be retired or consolidated with others. Staying on top of these changes is easier when the model is religiously used to maintain the database structure.



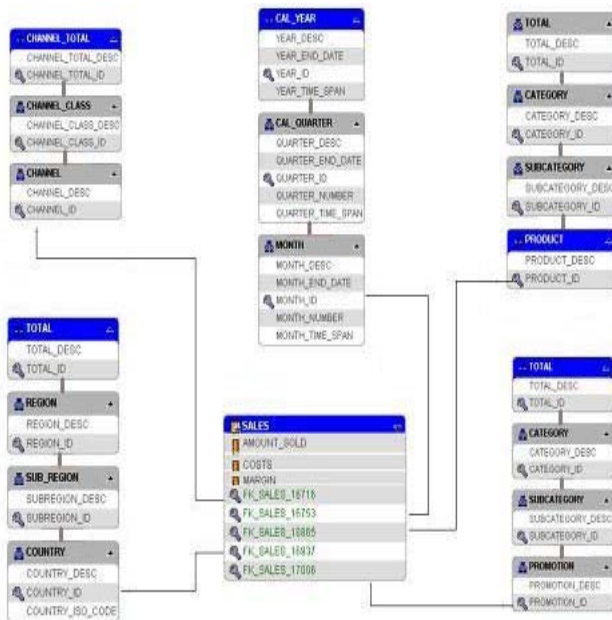
Multidimensional Structures

The need for a deep understanding of query technology is not an ideal solution for all users within an organization. In relational analysis there is one basic storage structure, a flat table. This object takes on many different roles depending on how it relates, or joins, to other tables. In a star schema, fact tables and dimension tables describe the relationship between the various tables within the schema. The fact table is usually the main driver within the query. In multi-dimensional analysis the same basic structures also exist. The fact table is referred to as a cube, and the columns within the table are referred to as measures. Each cube has additional structures over and beyond a simple table. The cube has edges, which are referred to as

dimensions. Each dimension is a grouping of common or related columns from one or more tables into a single entity. For example a Product dimension could be based on a single or collection of tables containing columns that refer to categories, sub-categories, and products. All these structures (cubes, measures and dimensions) interact with each other to provide an extremely powerful reporting environment. Each object adds new levels of interactivity when it is fully exploited. The use of a multi-dimensional model does not exclude the use of SQL as the query language. It merely requires the supporting query products to understand and exploit the additional structural metadata, provided by the multi-dimensional environment. In fact, Oracle's OLAP BI reporting products all use SQL to execute multi-dimensional queries although this is of no interest to end users who want to be insulated from any database semantics. Users want to interact with their data using business terms and objects without having to comprehend the underlying storage model. The abstraction of the physical storage model from the logical presentation model is key to the success of multi-dimensional models.

Modeling the Business

Using a relational data model and mastering the concept of table joins is an easy task. Most ad-hoc query tools do a good job of protecting the user community from the basic relational concepts. Tables and views are usually mapped to concepts such as folders. An administrator creates all the required joins and adds additional metadata to enhance the relational objects by changing the names of columns and folder names.



The Ghanaian Petroleum Data Management System

The Data Management Centre of the Department (GNPC) is a modern 9100sq. ft. building complex, containing laboratories and office space, and data storage facilities (approximately 4000 sq. ft) for the storage of hydrocarbon data. The centre is located on the Accra-Tema bypass road

(Spintex Road), approximately 5 miles east of the airport in Accra separated from the main office located in the industrial city of Ghana -Tema. The mission of the unit is to provide a secure and managed storage environment for Ghana's heritage of geological and geophysical data. These data have been obtained over a period of many years during the search for hydrocarbon resources. The data types include technical data and reports, rock cores and geological samples, seismic sections, and digital tapes. The Centre has developed a catalogue of all exploration and production data and related information. The catalogue is computer-based, to provide the exploration with complete and comprehensive listing of all available material and information to assist in the on-going search for hydrocarbon resources. Each catalogued item is assigned a unique code and storage location, and catalogues are managed by a computerized data management system. It is envisaged to provide on-line global data access and geological services. It must be noted that the existence of a computerized data gives a positive indication for a possible warehouse to be constructed. Also the reference to unique keys (primary keys) allocated to each data entity further gives the process additional boost. These keys usually serve as indexes to the data cubes and are positioned in the fact table to facilitate easy retrieval of data.

Data Types

Data stored is catalogued under the following major categories: Archival includes any corporate information that GNPC have deemed to be inactive, but necessary to keep in an organized manner for possible future reference. Displays include well information, cores, samples, reports, logs and surface samples. Geophysical includes seismic and gravity/magnetic line detail, reports, sections, and digital tapes. Library includes books, magazines, and technical reports.

METHODOLOGY

Since the research followed the consultancy path which was aimed at producing a consultancy form of report with recommendations to the client organization, we emphasized that it was not meant to find out how many people does A or B or how many more women in the organization carry out an activity or set of activities than men. Quantitative research which usually embraces the positivist ideology follows the scientific investigation path and objectively investigates issues and deals normally with statistical data. The objective of this research therefore differs from the positivist ideology of tabulating figures and establishing the relationship between them. This research therefore embraced the interpretive ideology and trailed on the path of the qualitative research. In addition, the idea of the iterative research value was introduced to depict the flexibility that must be employed in investigating IT issues as illustrated in the diagram below.

In support of the above, Wisker in 2001 hammered that researchers employing interviews under qualitative research have the competitive advantage of unintentionally forcing much details from their respondents by giving them the free will to express their deeper feeling about the concept while the researcher on the other hand is given the opportunity to

explore the idea behind their way of thinking unlike the restrictive quantitative approach [8].

To this end, the investigation employed both observation and interviews while relying on some amount of secondary data for critical analysis of existing trends and methodologies. Weaver in 2004 established that the use of interviews in probing and demystifying computing issues helps build a long lasting bond between the interviewer (the project team) and the interviewee [9].

THE SURVEY

Two main companies in the oil industry in Ghana - the Ghana National Petroleum Corporation (GNPC) and Tullow Oil Ghana (TOG) were interviewed to gather mainly qualitative data to enable the researcher to develop a conceptual view into their activities (background information) and their data storage workflow and requirements [10]. In addition to this, secondary data on one company- the Kosmos Energy Ghana limited (KEG) was also obtained in order to have a more complete picture of the investigation [11]. Among the people interviewed are the information systems managers, administrators, members of management decision making body and some selected non-managerial staff to help in obtaining details about operational data held by the department?

Transcription was the analysis tool employed to reveal the core meaning in the array of obtained set of responses. However this is not enough to bring out trends and aspect of the data. To do this, the power of thematic analysis was employed to segregate the data and bring out aspects which then helped shaped the development of the architecture and as well guided the recommendations suggested to management.

The researcher prior to the start of the investigation planned to interview as many as five companies involved in the oil exploration activity in Ghana. However at the preliminary stage of the actual fieldwork, it came to light that many of the companies including the Saltpond offshore production company limited (SOPCL), Sabre oil, Anardako Ghana limited and the E.O group Ghana Limited have not started any serious exploratory work in their respective fields.

This is attributed to the infant nature of the industry and hence goes a long way to buttress the researcher's ambitiousness and pro-activeness to research into the area. The only alternative the situation presented to the researcher is to focus on the two fields currently under exploration. In this direction therefore, the researcher was strained to rely on data obtained from interviewing the active companies notably the Tullow Oil Ghana Limited (TOG) and the Ghana National Petroleum Corporation (GNPC) while relying on secondary data obtained on Kosmos Energy Ghana Limited (KEG) to build a more complete picture of layout [10, 11].

However upon several participant observations with these two companies and as contained in the responses to the interviews, it came out that there is indeed a logical layout of flow of information between these three companies thus fulfilling the requirement of the research. Consequently, the government of Ghana as stipulated by the legal provisions requires that all data relating to Ghanaian petroleum

industry generated by the various stakeholders in the industry be made available to the GNPC who then persistently house these data for the consumption of the general public, the government, and other state holders both abroad and localized thus serving as a candidate for possible implementation of data warehouse. The GNPC works closely with its partners namely the KEG and TOG which in conjunction with the other dormant companies categorized above serve as the primary generators or sources of these data.

Recommendations to The Corporation

The benefits presented by data warehousing in businesses and more especially petroleum domain are wide and unlimited. Kimball et al in 2008 wrote that these benefits can be demarcated into two main categories – the tangible and the intangible [12].

The tangible benefits they said ranges from:

- Increment in business turnover
- Elimination of the cost of producing legacy reports
- Reduction in time in pulling together data for analysis.

While the intangible among other things embrace those benefits that direct value cannot be put on and includes the following:

- Timely and easy access to managerial decision data
- The discovery of relationships or knowledge that would have been embedded in data without being noticed through data mining technologies.
- Reduction in work burden on employee
- Precise rapid and cutting edge decision making.

Based on these numerous benefits and by the proclamation of the corporation that some managerial decisions are currently and constantly being withheld or hindered due to the sheer fact that data is currently available but does not exist as a common pool to facilitate rapid decision making, the researcher having thoroughly analyzed the data and with personal communications held with management during his observation session recommends that a data warehouse could be the most probable solution to this problem. The data islands in this situation are notably Tullow oil Ghana (TOG) and Kosmos Energy Ghana (KEG) the operators of the Jubilee fields who form the prime focus of this research and the other operators both in the process of securing a licence and those already with licence but currently doing the ground work for official exploratory work to begin [10, 12].

Apart from TOG and TEG the cooperation also houses various databases both in its head office location in the industrial city Tema and the data management unit off the Spintex road in Accra. As indicated by the department data in the data management unit are catalogued under various categories. The research thus established that all these entries be scanned if possible or captured and integrated into their respective operational databases (first level data store). This is necessary to allow for easy and rapid upload into the oracle data store (second level data store). We have for some time now been considering the establishment of a common repository of data (data farm) supported by online tools for accessing the data and currently the necessary

propo sals are being approved for the establishment of one, a manager proclaimed. Under this same Project a super server is to be purchased and this would form the backbone for pulling together b locks of data from the various islands or silo s including those currently being held by TOG an TEG due to unavailability of adequate storage facility at GNPC. However as a common practice in data warehousing, this online data farm cannot be deemed to provide answers to management questions due to the fact data warehouses and Online Transaction Processing (O LTP) systems have very different requirements.

It is therefore recommended that double server architecture be formed rather than the single server architecture. The super server as it was proposed would still go ahead to pool together heterogeneous operational data from the various data islands into a single homogeny ous repository of data farm and would serve as the main feeder of data to the data warehouse.

In effect a three level architecture re has been recommended to the department and his would be discussed by the sub-chapter below.

Building The Conceptual Architecture Model

Just like any software development project, data warehouse also follow a set of milestones to ensure that a working system is delivered on time and to ensure that all user requirements has been fully captured by the design.

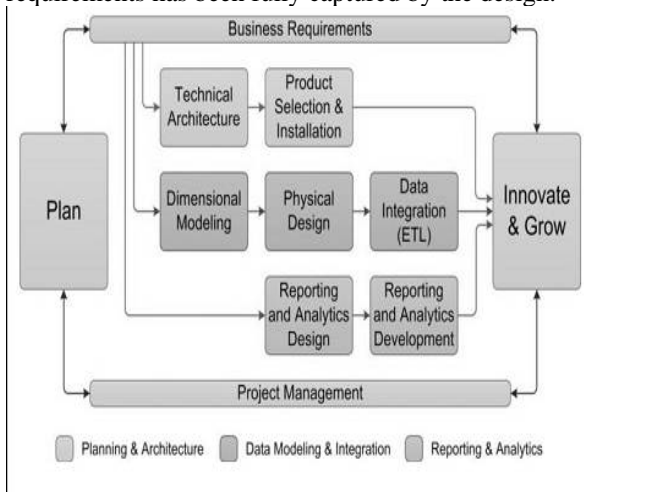


Figure :. The Kimball data warehouse development lifecycle.

Sour ce: www.teknonusa.com (2011) [13]

The successful completion of the activities within a systems lifecycle reassures the project team that the information system developed is in line with organizational needs [1]. Thus the architecture shall employ the Kimball’s lifecycle methodology. Kimball in 2008 outlined that for a data warehouse project to be successful does not really depend on the quantity and quality of data or the best combination of sophisticated modern technologies but rather how the task and components in the development project are blended and coordinated for the best delivery of product [11]. Thus the lifecycle ensures that the project constituents are positioned at the right place within the project and that this is done at the right time. The diagram below is a depiction of the proposed project lifecycle by Kimball.

However, a lifecycle cannot operate in a vacuum. It must have an approach or methodology within which it operates or shape thus the approach recommended to management is the top-down approach as outlined by Paulraj Ponniah. “The top-down approach is to start at the enterprise-wide data warehouse, although possibly build it iteratively and then data from the overall enterprise-wide data warehouse flow into departmental and subject data marts” [14].

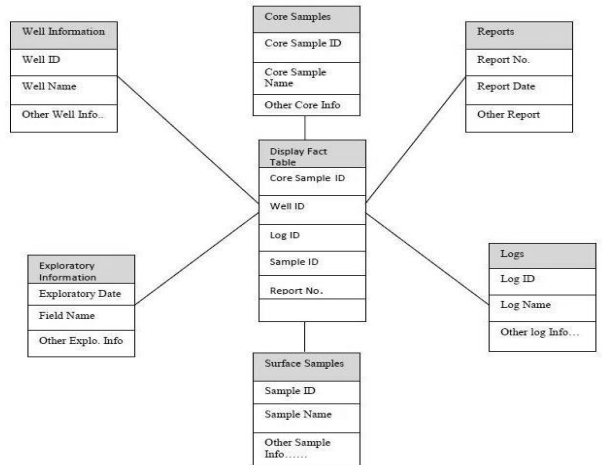


Figure : A multi-dimensional model of the display category

This approach is a strategic step following the nature and layout of the architecture. From the architecture, the corporation has the vision of integrating data from the various operational databases into one big store house (the super server). From here it would be a waste of time and resources to re- disintegrate the data into data marts before re-integrating into the enterprise-wide data warehouse. Thus a more economical way to do this is to follow the top-down approach for a faster and timely delivery.

Model of the schema: For the purpose of the schema, the multidimensional database design shall be the main focus to help us capture the various dimensions accurately. However emphasis is placed on st a portion of the holding that are involved in active decision making by the department notably those that fall under the display and the geophysical data. The various dimensions are captured diagrammatical by the schema below:

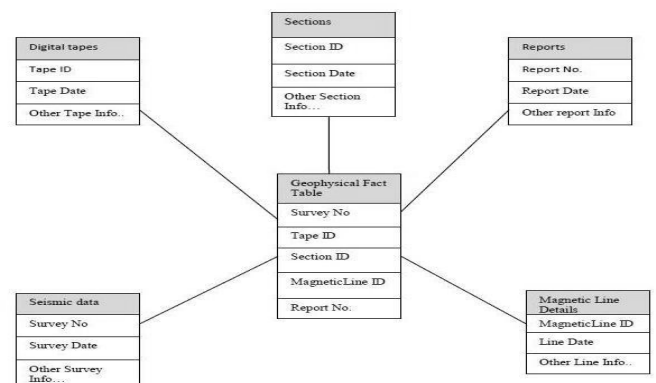


Figure :. A multi-dimensional model of the geophysical category

Data warehouse architecture is a way of representing the overall structure of data, communication, processing and presentation that exists for end-user computing within the enterprise [6] thus the model in Figure 6 below represent the three tier level data warehouse architecture recommended to the department.

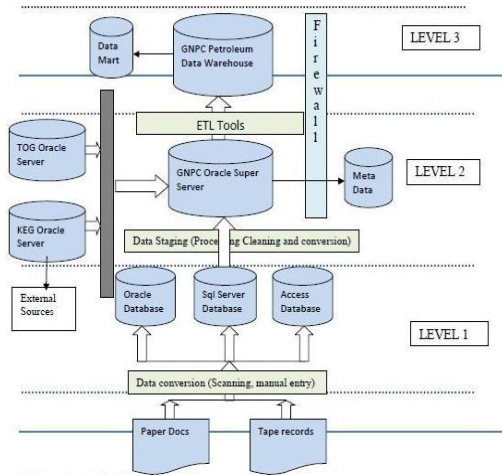


Figure 6. The three tier level data warehouse architecture

CONCLUSION

This paper has accomplished the mission of constructing a data warehouse model for the Ghanaian petroleum industry. In all three companies has been investigated in the petroleum exploratory sector and in effect it was revealed that a producer-ownership relationship in terms of data exist between these three companies. However GNPC as the custodians of the information lacks the appropriate facilities to well house and manage these data and in effect has been facing constant managerial decision making due to information dispersion across the organization and its partners. This leads to the formulation of various goals and objectives and in effect a number of investigative approaches, methodologies and strategies have been brought together resulting in the construction of the model

in Figure 6. This if implemented is believed to eliminate the current obscurities encountered in managerial cum government decision making on petroleum resources.

REFERENCES

- [1] BENNET, S., MCROBB, S. & FARMER, R. (2010), Object-oriented systems analysis and design using UML, 4th ed. London: McGraw-Hill.
- [2] INMON, W.H. (2005), Building the data warehouse, 4th ed. Indiana: Wiley Publications Inc.
- [3] NIMMAGADDA, S.L & DREHER, H. (2009), Mapping and Modeling of Oil and Gas Relational Data Objects for Warehouse Development and Efficient Data Mining, Institute of Electrical and Electronic Engineers, (August), pp.1201-1205.
- [4] CHEN, J. et al. (2004), Research On the Application of Data-mining For Quality Analysis In Petroleum Refining Industry, Institute of Electrical and Electronic Engineers, (June), pp.4214-4318.
- [5] WANG, Y. et al (2010), The Application of Spatial Data Warehouse Technology in The National Petroleum Resources Database, Institute of Electrical and Electronic Engineers, (July), pp. 257-260
- [6] MOHANTY, S. (2006), Data warehousing design, development and best practice, New Delhi: Tata McGraw-Hill.
- [7] CONNOLLY T. & BEGG, C. (2010), Database systems: a practical approach to design, implementation and management, 5th ed. London: Pearson.
- [8] WISKER, G. 2001. The postgraduate research handbook. New York: Palgrave.
- [9] MÜCI-KUCHLER, K., AND WEAVER, J. 2004. Learning how to identify customer requirements: a key component of product development courses. In: Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition. Salt Lake City, UT: American Society for Engineering Education.
- [10] TULLOW OIL GHANA. (n.d), About Tullow Ghana, [WWW]. Tullow oil Ghana. Available from: <http://www.tullowoil.com/ghana/index.asp?pageid=61> [Accessed 04/08/11].
- [11] KIMBALL, R. et al. (2008), The data warehouse lifecycle toolkit, Indiana: Wiley publishing inc.
- [12] KOSMOS ENERGY GHANA. (n.d), Kosmos Energy Genesis, [WWW]. Kosmos energy Ghana. Available from: <http://www.kosmosenergy.com/aboutus.html> [Accessed 04/08/11].
- [13] TEKNION (n.d), Business Intelligence, [WWW]. Teknion. Available from: <http://www.teknionusa.com/Solutions/Business-Intelligence/Pages/BI-Methodology.aspx> [Accessed 09/09/11].
- [14] PONNIAH, P. (2001), Data warehousing fundamentals: a comprehensive guide for IT professionals, New York: John Wiley and Sons