Information and Knowledge Management
ISSN 2224-5758 (Paper) ISSN 2224-896X (Online)
Vol.4, No.5, 2014

Leveraging Data Mining and Data Warehouse to Improve Prison Services and Operations in Nigeria

Georgina N. Obunadike[1&2*],Dr Arthur Umeh[2]
1. Department of Mathematical Sciences and IT, Federal University, Dutsin-ma, Katsina State
2. Department of Information and Media Technology, Federal University of Technology, Minna, Niger State
*Email: nkoliobunadike@yahoo.com

Abstract
Crimes are social nuisance and cost our society dearly in several ways. In Nigeria, any research geared towards helping to solve crimes faster will be beneficial to the society at large. It has been observed that the major challenge facing all law-enforcement and intelligence-gathering organizations in Nigeria is how to accurately and efficiently analyze the growing volume of crime data. As the volume of this crime data becomes enormously large, new techniques have to be used to turn this data into valuable information and actionable knowledge so that appropriate actions can be taken accordingly. Sometimes it is usual to find that the data needed to be analyzed to produce report are scattered throughout different operational States and jurisdictions of Nigeria and must first be carefully integrated. Moreover, observations show that the process required to extract the existing data from each operational system demand so much of the system resources such that the IT professional must wait until nonoperational hours before running targeted queries required for producing operational reports. These delays are not only time-consuming and frustrating for both the IT professionals and the decision-makers they are dangerous for the sector whose primary task is to control crime spread and explosion. It should be noted that when such operational reports are finally produced, they may not be relied upon, because the data use in producing them many a times are inconsistent, inaccurate, or obsolete. This paper therefore highlights the increasing growing need for Data integration, Data warehouse and Data mining as ways to improve the operations of principal actors within the prisons sector of Nigeria. The paper explains what these Data management techniques mean and entail, and furthermore suggests ways to effectively leverage the techniques to help detect existing crime patterns and speed up the process of solving crimes.
Keywords: Crime-data, data mining, data mining techniques, data warehouse, data integration

Introduction

Data mining is an emerging field gaining acceptance in research and industry. It incorporates techniques from a number of fields including statistics, machine learning, database management, artificial intelligence, pattern recognition, and data visualization. A number of definitions for data mining are presented in literature (Julio and Adem, 2009). The convergence of computing and communication has resulted in a society that highly relies on information (Witten and Frank, 2000). Technological developments that aid to collect and store vast quantities of data have enabled organizations to capture and accumulate huge amount of data in their databases, within which, large amount of valuable information is buried (Fayyad, Piatesky-Shapiro and Smyth, 1996). Most organizations gather data because it is needed for some operational purpose and ones it has served its purpose, it is left to languish on disk or tape or is discarded. Such data can be mined to discover hidden patterns or trends which can help improve the organization and even the society at large. Much crucial information is usually buried in such operational data and only data mining techniques has the capabilities to bring such information to bear. John Naisbitt (1986) in his book megatrends observed that we are drowning in information but are starved for knowledge.

Database systems have had great success during the past two decades. More and more data are being collected and saved in databases—a database with a pedabyte of data is no longer uncommon. Such a large volume of data can be hard to store, but it is a valuable source of data for understanding trends and the way the enterprise functions. This data can be very helpful for making projections that lead to successful strategic decisions, as well as for improving everyday decision making. Finding useful information in these databases has become an important focus of many enterprises; and more and more attention has turned to data mining as a key component to such information discovery. Data-mining algorithms and visualization tools are being used to find important patterns in data and to create useful forecasts. This technology is being applied in virtually all business sections including banking, telecommunication, manufacturing, marketing, and e-commerce (ZhaoHui Tang and Jamie MacLennan, 2005).

Data mining make the most sense when there are large volumes of data. Most data mining algorithms require large amount of data in order to build and train the model. The large amounts of data being produced are being extracted from the operational data and are fed into a data warehouse to become part of the corporate memory. Data warehouse brings together data from many different sources in common format with consistent definitions for keys and fields. Operational system store data in format designed to optimize performance of operational task.

Thus, the format is generally not well suited for decision support activities like data mining. Data warehouse on the other hand is designed exclusively for decision support which simplify job of the data miner (Berry and Linoff, 2004) .

Data warehousing is a natural ally of data mining. Data mining seeks to find actionable patterns in data and therefore has a firm requirement for clean and consistent data. Much of the effort behind data mining endeavors is in the steps of identifying, acquiring, and cleansing the data. A well-designed corporate data warehouse is a valuable ally. Data warehousing has become increasingly important for any kind of decision support or information analysis (Berry and Linoff, 2004).

It has been observed that the major challenge facing all law-enforcement and intelligence-gathering organizations especially in developing countries like Nigeria is how to accurately and efficiently analyze the growing volume of crime data (Chen et al., 2004). As the volume of this crime data becomes enormously large, new techniques have to be used to turn this data into usable information and knowledge so that appropriate actions can be taken accordingly. Data mining holds the promise of making it easy, convenient, and practical to explore very large databases for the Nigerian Prisons Service. Data mining is a powerful tool that uses complex algorithms to look for patterns in very large sets of data. Criminal investigators who may lack extensive training as data analysis can easily explore large databases quickly and efficiently through the implementation of data mining techniques in the form of commercial and other applications (Chen et al., 2004; Fayyad and Uthurusamy, 2002).

According to Megaputer Intelligence (2002), the analysis of crime patterns and trends is very important for Security agencies and analysts to learn from historical crime patterns and enhance crime resolution rate. It also helps to prevent future incidents by putting in place preventive mechanisms based on the observed patterns. Another possible advantage is that it can reduce the training time for officers assigned to a new location and having prior knowledge of site specific patterns to assist them in investigations. In line with the crime patterns extracted from previous records, the right scarce resources can be deployed to the right place at the right time.

Law enforcement agencies like that of Nigerian prisons today are faced with large volume of data that must be integrated, processed and transformed into useful information and hence data mining can greatly improve crime analysis and aid in reducing and preventing crime.

The rest of this paper proposes a framework for the integration of the Nigerian prison databases and explores the applicability of data mining techniques in crime patterns extraction.

2. The Data Warehouse Concept

Database computing has shifted its focus from operational to decisional concerns. The differences in operational and decisional information requirements presented new challenges that old computing practices could not meet. This change in computing focus has become the impetus for the development of data warehousing technologies. A Data Warehouse (DW) is defined as "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process" (Batini et al, 1992). Data warehouses store huge amount of information from multiple data sources which is used for query and analysis. The data warehouse holds transformed and integrated enterprise-wide operational data appropriate for strategic decision-making, as shown in Figure 1. A data warehouse also contains data obtained from external sources, whenever this data is relevant to decision-making. Decision support applications usually used to analyze and make data warehouse information available in formats that are readily understandable to decision-makers.

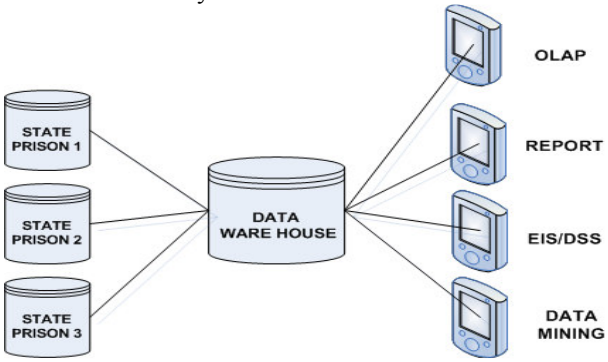

Figure 1: Data Warehouse as an integrated enterprise-wide operational data appropriate for strategic decision-making

It is worth noting that since cohesive sector-level operational information requirements cannot be anticipated with disparate data pools and repositories, operational system (which currently focus on recording different types of crimes in different State prisons) are unable to provide decision-makers with the information they need. As a result, executives and operational managers fall back on the time-consuming, and often frustrating, process of

going through operational inquiries or reports already supported by operational systems in an attempt to find or derive the information they really need. Oftentimes, IT professionals are pressurized to produce ad hoc reports from the operational systems as quickly as possible. It is common for IT professionals to realize that the data needed to produce the crucial reports are scattered throughout different operational systems and platforms; thus existing pockets of data pools must first be carefully integrated.

Moreover, IT expert usually discover that the process required to extract the data from each operational system demands so much of the system resources, in most cases they are forced to wait until nonoperational hours before running targeted and diced queries required to produce decision making report. These delays are not only time-consuming and frustrating both for the IT personnels and the decision-makers but portend gross ineffectiveness and inaccuracy of decision. In this prevalent circumstance, there is a predominant tendency to produce reports which in actual sense trigger request for more ad hoc reporting. A data warehouse maintains a copy of information from the source operational systems; this provides the opportunity to:

- Maintain data history, even if the source transaction systems do not.
- Integrate data from multiple source systems, enabling a central view across the enterprise. This benefit is always valuable, but particularly so when the organization has grown by merger.
- Improve data quality, by providing consistent codes and descriptions, flagging or even fixing bad data.
- Present the organization's information consistently.
- Provide a single common data model for all data of interest regardless of the data's source.
- Restructure the data so that it makes sense to the business users.
- Restructure the data so that it delivers excellent query performance, even for complex analytic queries, without impacting the operational systems.
- Add value to operational business applications, notably customer relationship management (CRM) systems

A data warehouse platform (which is shown in figure 2 below) consists of one or more hardware servers, an operating system, a database management system (DBMS), and data storage. These communicate via a LAN or WAN, although a multi-node data warehouse platform may have its own specialized network. Note that a data warehouse platform manages a data warehouse, defined as a collection of metadata, data model, and data content, designed for the purposes of reporting, analyzing information, and making decisions. But the data warehouse is not part of the platform per se (Russom, 2009).

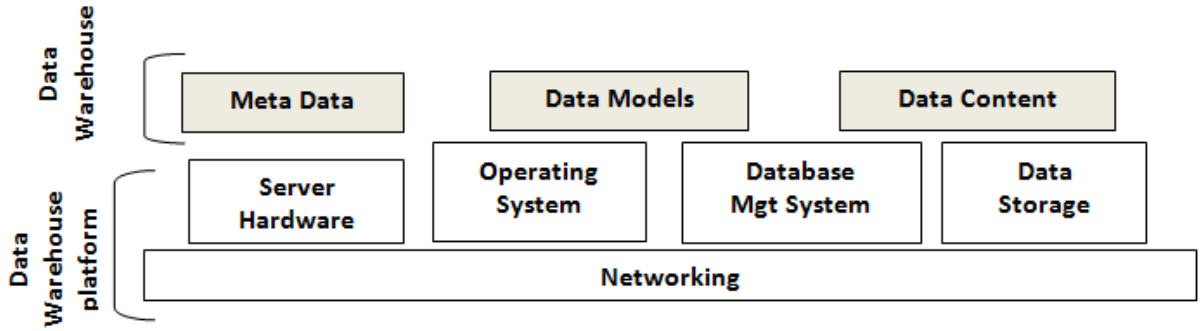

Figure 2: A data warehouse platform manages a data warehouse, but the two are separate
Source: (Russom, 2009).

3. Related Works

There are many data mining techniques that can be applied to Nigerian prisons crime data. Some examples of the specific applications of these techniques are given below as derived from current scholarly literature on the data mining of crime data.

Lin and Brown (2006) presented an outlier-based data association method for linking criminal incidents. In this technique, according to them, "an outlier score function is defined to measure the extremeness of an observation, and a data association method is developed based upon the outlier score function." They applied this method to the robbery data in Richmond, Virginia, and compared the result with a similarity-based association method. Their results show that the outlier-based data association method is promising.

Estivill-Castro and Lee (2001) incorporated two knowledge discovery techniques, clustering and association-rule mining, into a fruitful exploratory tool for the discovery of spatial-temporal patterns. They presented two methods for this exploratory analysis and the detail algorithms to effectively explore geo-referenced data. They illustrated the algorithms with real crime data. They demonstrated their approach to a new type of analysis of the spatial-temporal dimensions of records of criminal events.

Brown (1998) describes a software framework for building and applying data mining algorithms to crime analysis problems. This framework provides specific focus on spatial data mining. The author provides several reasons to justify this focus: 1) spatial queries are more time consuming, 2) spatial analysis is harder to do than analyses based on attribute matching, 3) spatial data mining has the potential to yield important immediate benefits for crime analysis as crimes have
an inherently spatial component (Brantingham and Brantingham, 1984), and 4) spatial analysis is a key to law enforcement resource allocation.
Chen et al. (2004) present a general framework that shows the relationship between data mining techniques applied in criminal and intelligence analysis and the crime types. They identify and arrange eight crime types (traffic violations, sex crime, theft, fraud, arson, gang/drug offences, violent crime, and cyber-crime) in increasing order of public harm on the horizontal axis. On the vertical axis, they arrange the techniques in increasing order of analysis capability. They identified four major categories of crime data mining techniques: entity extraction, association, prediction, and pattern visualization. Each category represents a set of techniques for use in certain types of crime analysis. They then identified the intersection of the techniques with the crime types denoting where each technique could be effectively used for each crime type, completing the framework.

4. Data Mining Tasks Which May Be Applied To Manage Prison Data

4.1 Association Rules

Mining association rules searches for interesting relationships among items in a given data set. It allows finding rules of the form If antecedent then (likely) consequent where antecedent and consequent are item sets (El-Halees and M. Abu Tair, 2012). Item set is set of one or more items. In crime analysis association rule can be used to detect things that go together with a particular crime

4.2 Classification

Classification is a data mining task that is used to predict values for some variable (El-Halees and M. Abu Tair, 2012). For example, given all the records of criminal activities in an area, one may want to predict whether the area is prone to a particular criminal activity. Thus, classification technique can be used to predict the crime occurrence.
It is important to know that classification rules are different from the rules generated from association rules. Association rules are characteristic rules (it describes current situation), but classification rules are prediction rules (it describes future situation) (El-Halees, 2008).

4.3 Clustering

Clustering is a data mining task that aims at finding homogeneous groups in data. The objective of clustering is to find high-quality clusters such that the Inter-cluster distances are maximized and the Intra-cluster distances are minimized (El-Halees and M. Abu Tair, 2012). The clustering method can be used to detect areas with high criminal activities or criminal activities that are common to a particular region to enable necessary planning and actions.

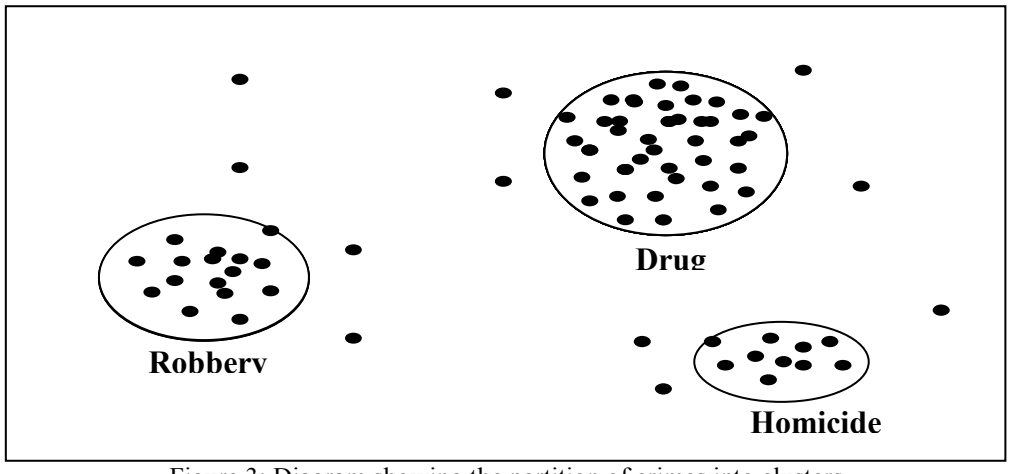

Figure 3: Diagram showing the partition of crimes into clusters

4.4 Outlier Detection

Outlier detection discovers data points that are significantly different from the rest of the data (Mansur et al, 2005). In crime data mining, outlier analysis can be used to detect outliers in crime data. Thus, outlier detection can be used to detect most threatening crime in an area to enable planning and correct allocation of security personnel and resources.

5. Proposed Data Warehousing Framework For Integrating Nigerian Prison Data

Data warehouse can span enterprise wide data needs or can be a collection of "conforming" data marts (Kimball and Ross, 2002). Data marts (subsets of data warehouses) are conformed by following a standard set of attribute declarations called a data warehouse bus. The data warehouse uses a metadata repository to integrate all of its components. The metadata stores definitions of the source data, data models for target databases, and transformation rules that convert source data into target data. The concepts of time variance and non volatility are essential for a data warehouse (Inmon, 2002). Inmon emphasized the importance of cross-functional slices of data drawn from multiple sources to support a diversity of needs (Inmon, 2002); the foundation of his subject-oriented design was an enterprise data model.

Kimball introduced the notion of dimensional modeling marts (Kimball and Ross, 2002), which addresses the gap between relational databases and multidimensional databases needed for OLAP tasks. These different definitions and concepts gave rise to an array of data warehousing methodologies and technologies. Data warehousing methodologies share a common set of tasks which includes business requirements analysis, data design, architecture design, implementation, and deployment which represents the top-down design approach of data warehousing (Kimbal et al, 1998). The top-down approach suits the vision of Bill Inmon, who considers that the data warehouse must respond to the requirements of all the users in the organization, and not of only a certain group (Inmon, 2002). Figure 4 below shows a flowchart for the proposed integration process of the Nigerian Prisons database.

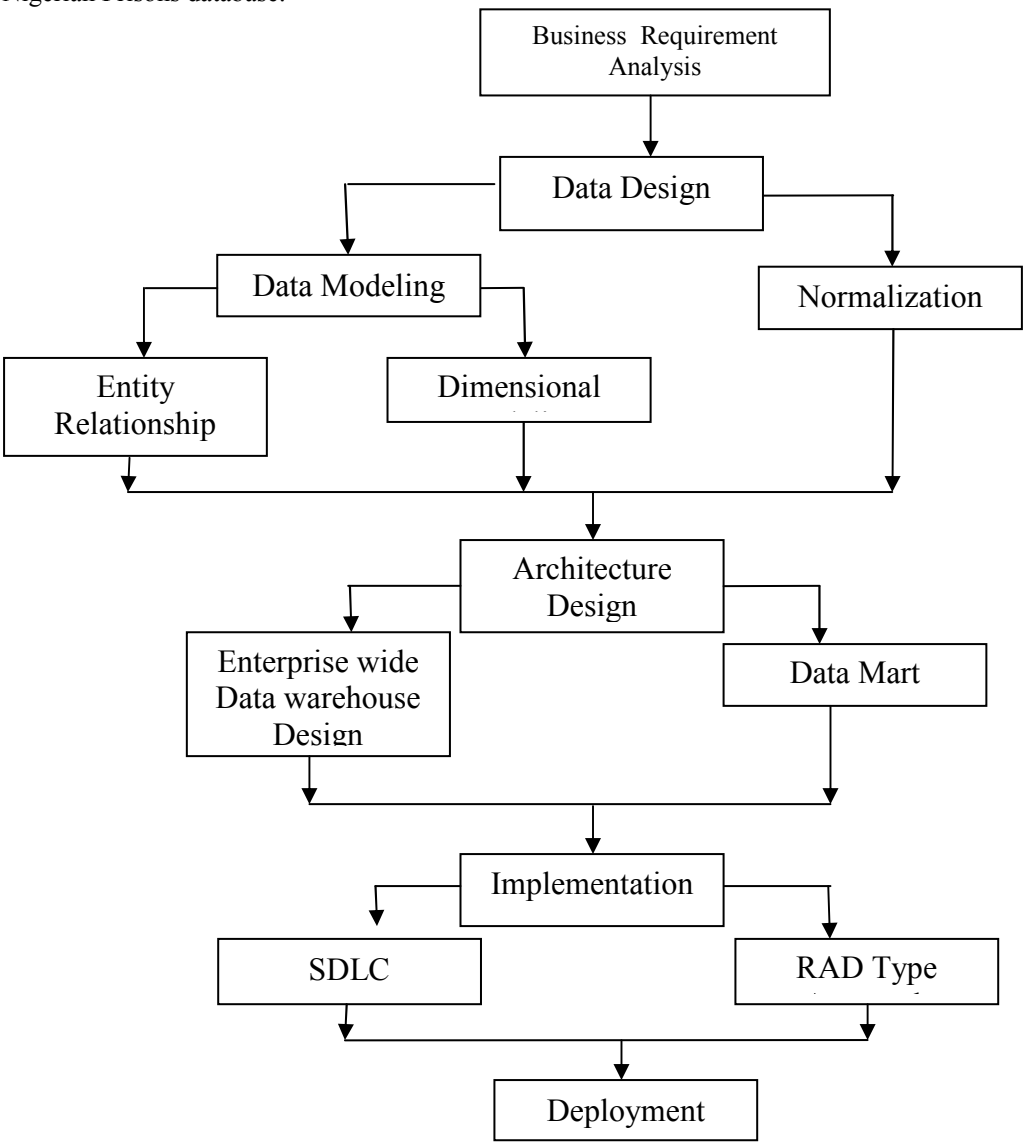

Figure 4: Flowchart for the proposed Top Down Design approach

5.1 Business Requirement

For business requirements analysis, techniques such as interviews, brainstorming, and JAD sessions are used to elicit requirements. Business requirements analysis is used to elicit the business questions from the intended users of the data warehouse. Business questions are decision support or analytic questions that managers

typically pose on project (Arun and Atish, 2005). After all the business questions are elicited, they are prioritized by asking the users to rate the questions, or by estimating the risk associated with the solutions needed for the questions. Next, a very high-level conceptual model (also known as the subject-area data model) of the solution for each of the business questions is created. The conceptual model serves as the blueprint for the data requirements of an organization.

5.2 Data Design

The data design task includes data modeling and normalization. The two most popular data modeling techniques for data warehousing are Entity-Relational and Dimensional modeling. The Entity-Relational modeling follows the standard OLTP database design process, starting with a conceptual entity-relationship (ER) design, translating the ER schema into a relational schema, and then normalizing the relational schema.

A dimensional model is composed of a fact table and several dimension tables (Kimbal and Ross, 2002). A fact table is a specialized relation with a multi-attribute key and contains attributes whose values are generally numeric and additive. A dimension table has a single attribute primary key (usually surrogate) that corresponds exactly to one of the attributes of the multi-attribute key of the fact table. The characteristic star-like structure of the physical representation of a dimensional model is called a star join schema, or simply a star schema. A dimensional model can be extended to a snowflake schema, by removing the low cardinality attributes in the dimensions and placing them in separate tables, which are linked back into the dimension table with artificial keys (Kimbal et al, 1998). Dimensional modeling is a data modeling techniques that have gained popularity and acceptance for data warehouse design (Arun and Atish, 2005).

Figure 4 presents a star schema for the proposed data warehouse design. It is made up of an inmate fact table and several dimension tables, the schema analyzes inmates by offenses, dates, place, time, age, courts, prison and skill aquision.

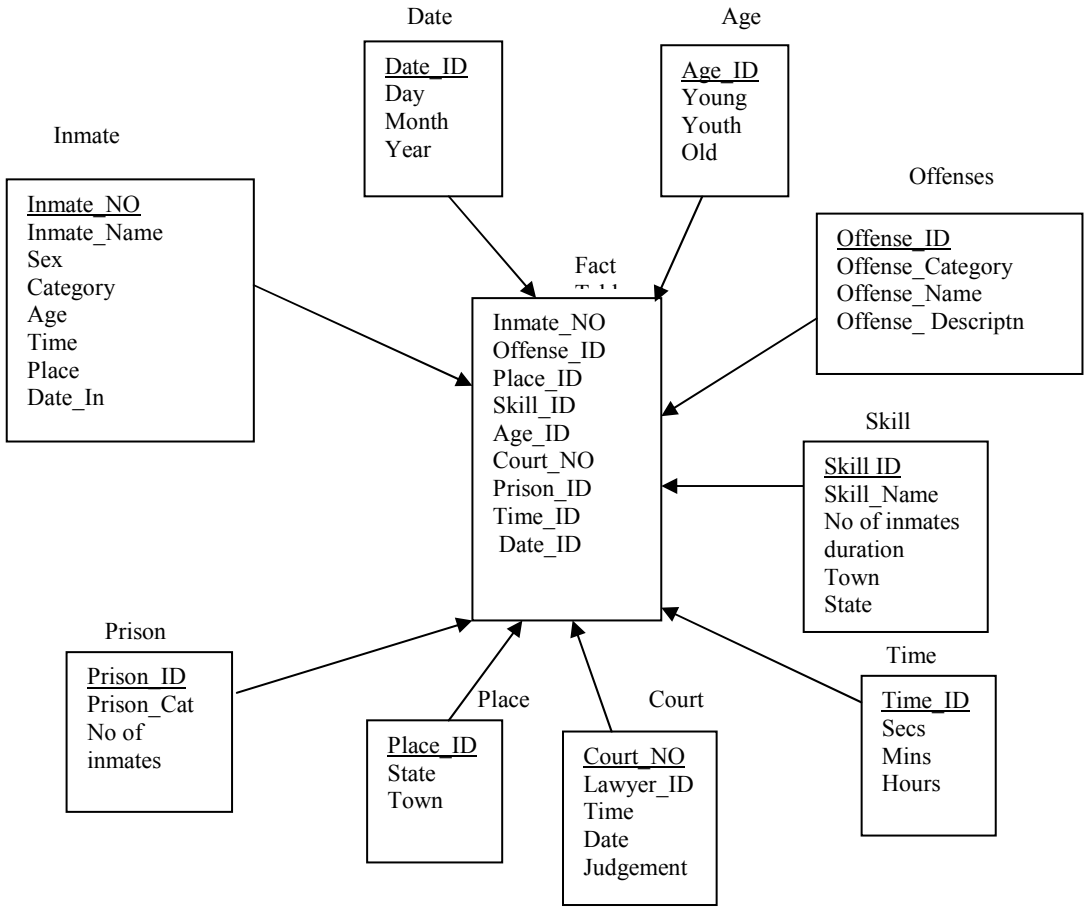

Figure 4: A Star Schema for the Proposed Data Warehouse Design for Nigerian Prison Services

5.3 Architecture

Architecture is a blueprint that allows communication, planning, maintenance, learning, and reuse. It includes different areas such as data design, technical design, hardware and software infrastructure design. The architecture design philosophy has its origins in the schema design strategy of OLTP databases. Several strategies for schema design exist, such as top-down, bottom-up, inside-out, and mixed (Batini et al, 1992). The

data warehouse architecture design philosophies can be broadly classified into enterprise wide data warehouse design and data mart design (Hackey, 1997). The data mart design, espoused by Kimball et al (1998) follows the mixed (top-down as well as bottom-up) strategy of data design. The goal is to create individual data marts in a bottom- up fashion, but in conformance with a skeleton schema known as the "data warehouse bus." The data warehouse for the entire organization is the union of those conformed data marts.

5.4 Implementation

Data warehouse implementation activities include data sourcing, data staging (ETL), and development of decision support oriented end-user applications. These activities depend on two things—data quality management and metadata management (Inmon,2000). As data is gathered from multiple, heterogeneous OLTP sources, data quality management is a very important issue. A data warehouse generates much more metadata than a traditional DBMS. Data warehouse metadata includes definitions of conformed dimensions and conformed facts, data cleansing specifications, DBMS load scripts, data transform runtime logs, and other types of metadata (Kimbal et al, 1998). Because of the size of metadata, every data warehouse should be equipped with some type of metadata management tool.

5.5 Deployment

The deployment task focuses on solution integration, data warehouse tuning, and data warehouse maintenance. Although solution integration and data warehouse tuning are essential, maintenance is cited as one of the leading causes of data warehouse failures. Warehouses fail because they do not meet the needs of the business, or are too difficult or expensive to change with the evolving needs of the business. Due to increased end-user enhancements, repeated schema changes, and other factors, a data warehouse usually goes through several versions.

6. Conclusion

It has been observed that in spite of the existence of databases with relatively large volume of data in Nigerian Prisons Service, the databases are scattered in various locations in the states. The commission has not yet integrated the databases in order to exploit the information embedded in the databases. The current system of the commission is still a manual system except the fact that the aforementioned databases are being used as repository. The crime prevention measures are being taken based on crime incidents although it would have been based on crime trends. This indicates that, currently, the crime prevention approach is based on the crime reports incoming to the commission, which is a reactive approach although it would have been proactive. The factors that may have contributed to this underutilization of the database systems of the commission may include: lack of knowledge on what could be done using these databases or deficiency of appropriate tools that could make use of these databases and the other problem may be due to lack of skilled manpower.

In areas where large volumes of data are found, like that of the Nigerian Prisons Services, information support technology tools have become mandatory to manage and process information for decision-making. Data mining is one such tool that has evolved to play a role as an instrument to discovery of patterns buried in large databases. Data mining is the "exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and relationships

REFERENCES

Arun S. and Atish P. S. (2005) "A Comparism of Data Warehousing Methodologies, ACM,Vol. 48, No. 3, pp. 79- 84.

Batini, C., Ceri, S., and Navathe, S.K. (1992) "Conceptual Database Design: An Enity- Relationship Approach". Benjamin/Cummings, Redwood City, CA

Brantingham, P. and Brantingham, P. (1984). *Patterns in Crime.* Macmillan Publishing Company, New York.

Berry, M. J. A and Linoff, G. S. (2004) " Data mining Techniques for Marketing and Sales and Customer Relationship Management" Wiley Publishing, Canada, pp. 36

Brown, D.E. (1998). The Regional Crime Analysis Program (ReCAP): A Framework for Mining Data to Catch Criminals. Proceedings of 1998 IEEE International Conference on Systems, Man, and Cybernetics, 3: 2848-2853.

Brown, D. (2003). The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals. http://vijis.sys.virginia.edu/publication/RECAP.pdf

Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., and Chau, M. (2004). Crime Data Mining: A General Framework and Some Examples. Computer, 37(4), 50-56.

Estivill-Castro, V. and Lee, I. (2001) " Data Mining Techniques for Autonomous Exploration of Large Volume of Geo-Referenced Crime Data", *Proceedings of the 6th International Conference on Geocomputation*

El-Halees, A. (2008) Mining Students Data to Analyze Learning Behavior: A Case Study, The 2008 international Arab Conference of Information Technology (ACIT2008) – Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18, 2008.

El-Halees, A. and Abu Tair, M. (2012) "Mining Educational Data to Improve Students Performance: A Case Study", International Journal of Information and Communication Technology Research, (2)2, 2012, pp. 140 – 146,.

Fayyad, U, Piatesky-Shapiro, G and Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. ftp://ftp.research.microsoft.com/pub/dtg/fayyad/kdd96/fayyad-intro.ps.

Fayyad, U.M. and Uthurusamy, R. (2002). Evolving Data Mining into Solutions for Insights. Communications of ACM, 45(8), 28-31.

Gartner Group (1995), Gartner Group Advanced Technologies and Applications Research Note http://www.gartner.com John N. (1986) " Megatrends", 6th ed., Warner Books, New York.

Julio, P. and Adem, K. (2009) Data Mining and Knowledge Discovery in Real Life Applications, ISBN 978- 3-902613-53-0, pp. 438, I-Tech, Vienna, Austria.

Hackney, D. (1997) "Understanding and Implementing Successful Data Marts", Addison- Wesley, Reading, MA.

Humphries, H. (1998) "Data Warehousing Architecture and Implementation", First Edition, Prentice Hall, USA.

Inmon, W.H. (2002) "Building the Data Warehouse", 3rd edition", Wiley, New York.

Inmon, W. (2000) "Metadata in the data warehouse", White Paper, www.inmoncif.com/library/whiteprs/earlywp/ttmeta.pdf.

Kimball, R., Reeves, L., Ross, M., and Thronthwaite, W. (1998) " Data Warehouse Lifecycle Toolkit", Wiley, New York.

Kimball, R. and Ross, M.(2002) "The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling", 2nd edition, Wiley, New York.

Lin, S. and Brown, D.E. (2006). An Outlier-based Data Association Method for Linking Criminal Incidents. *Decision Support Systems,* 41(3), 604-615.

Mansur , M. O. Sap, M. and Noor, M. Outlier Detection Technique in Data Mining: A Research Perspective, In Postgraduate Annual Research Seminar, 2005.

Megaputer Intelligence, Inc. (2002). Crime Pattern Analysis: Megaputer Case Study. http://www.elon.edu/facstaff/mconklin/cis230/cases/crime_pattern_case.pdf

Regional crime analysis program: Small department edition (ReCAP SDE) http://www.sys.virginia.edu/capstone/past/cap2000/11-01ReCAP.doc

Russom P. (2009) " Next Generation Data Warehouse Platforms", TDWI Report, www.tdwi.org

Surajit C. and Dayal U. (1997) " An Overview of Data Warehousing and OLAP Technologies" , ACM Sigmod

Witten, I. and Frank, E. (2000). Data mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann publishers

Wilson , O.W. (1963). Police Administration. USA, McGraw Hill Company.

ZhaoHui Tang and Jamie MacLennan, 2005 "Data Mining with SQL Server 2005",Wiley Publishing Inc, Indianapolis, Indiana, 2005.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Recent conferences: http://www.iiste.org/conference/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar