THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

A Road Map for Detecting Financial Frauds in Enterprises through Data Mining

Adekunle Joshua Akinjobi
Lecturer / Head of Department, Department of ICT/ Computer Science,
Crawford University, Igbesa, Ogun State, Nigeria

Abstract:
Enterprises face the problems of combating financial frauds in their various locations. Although, most of these enterprises engage full time internal auditors, the manual process carried out particularly to detect frauds, have not been completely assured. This is as a result of the large volume of financial transactions existing such that only random sampling of the audit is presently done.
A financial fraud detection is efficiently conducted if the process successfully passes through journal entries as the primary source of financial statements.
This road map having reviewed the previous attempts of researches such as Benford's law, Neural networks and Self organizing maps to detect financial frauds as modest contributions, describes a data mining process using decision tree algorithm that builds pattern models. An SQL Server Integration Services (SSIS) are utilized to affect the decision tree data mining to reduce anticipated frauds and enables a complete auditing. This resulted in providing complete, accurate, confident and reliable audit results to the enterprise managers.
The Journal Entries data mining revealed patterns of attempted frauds in enterprises particularly patterns gleaming unapproved entries, entries of amounts above approval limits, entries made during holidays and weekends.

Keywords: *auditing, data mining, decision tree, frauds, journal entries, SQL Server*

1. Introduction

It is noteworthy that enterprises face the problems of combating financial frauds in their various locations. Although, most of these enterprises engage full time internal auditors, the manual process carried out particularly to detect fraud, have not been completely assured. This is as a result of the large volume of financial transactions existing such that only random sampling of the audit is done. There has not been a guarantee that the remaining volumes not audited (those outside the sampled volumes) do not contain frauds.
Financial frauds are professionally detected through auditing by auditors of an enterprise. Such auditors are expected to be familiar with all accounts segments originating from journal entries through the general ledger up to the balance sheet.
The center for Audit Quality document in its Practice Aid for Testing Journal Entries section, recommends that "It is important in testing journal entries and other adjustments, that the auditor considers the entire population of the journal entries".
To effectively detect financial frauds, consideration has to be given to the primary source of financial statements. This primary source is the journal entries from where general ledger, trial balances and balance sheet are prepared. See figure 1
In situations where internal control is lacking most especially where management override is prevalent, fraudulent manipulation of accounts is carried out through journal entries.

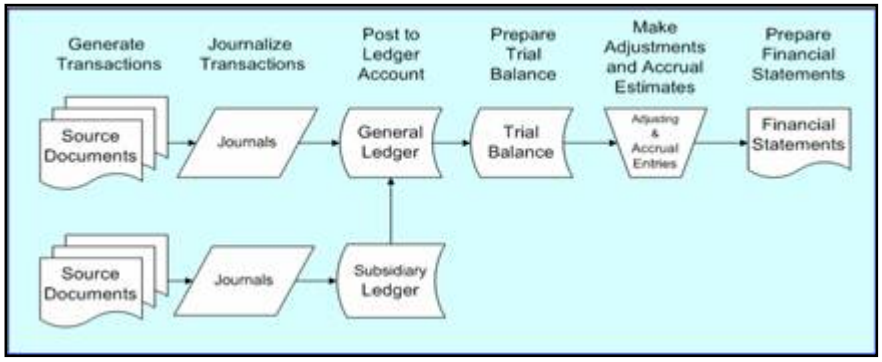

Figure 1: Overview of the accounting process
Source: (Bay, Kumaraswamy, Anderle, Kumar & Steier, 2012)

The Center for Audit Quality recognises this in its assertion that 'evidence has shown that fraudulent financial reporting frequently involves the recording of fraudulent journal entries. As a result, auditors are to presume that the risk of management override of controls is always present and to test journal entries for indications of possible material misstatement due to fraud'.

Data mining is the answer to complete auditing of large volumes of financial transactions through which fraud patterns are revealed

Data mining Journal entries for the purpose of fraud detection is the process of gaining insights and identifying fraudulent patterns from the journal entries data stored in databases. Data mining is also an analytical tool that can assist the auditors in analyzing fraudulent patterns for further investigation and management decision making.

Data mining falls into categories of supervised learning and unsupervised learning. A supervised learning focuses on a target variable with known values and classes as input about which prediction will be made. Unsupervised learning such as clustering, are employed on data without a target variable and classes but with known values as input variables.

The Journal Entries supervised learning is classed into 'Fraud' and 'No Fraud' which makes it easy to use a decision tree algorithm. The algorithm builds model(s) or rules that predict frauds in a journal entries data set containing several journal entries as input. Once the model accuracy is successfully established, it is applied to new journal entries that contain both fraudulent and non-fraudulent data. The model discovers patterns of frauds and no frauds that are stored in the journal entries database for subsequent decision making.

Once detection of fraudulent pattern is established through the data mining algorithm of decision tree carried out in this work, a model evolves that help in a related fraud prediction.

2. Literature Review

2.1. Data Extraction, Transformation and Loading (ETL)

A review of a typical ETL process otherwise called Data Warehouse (DW) Business Intelligence (BI) tools consists of the Pentaho Community Business Intelligence (BI) used to construct the DW based on the Hefesto methodology. The Pentaho BI is an integrated platform that includes ETL (Extraction, Transformation and Load), data integration capabilities, data mining, reporting, OLAP (On-Line Analytical Processing) services.

The Hefesto methodology approach allows tackling the design of the DW from different detailed levels, and reducing risks of failure and dissatisfaction by involving end-users early in the design process through four steps:

i. Requirement analysis,
ii. OLTP analysis,
iii. Building the Logical Model (this represents the structure of the DW, defining the type of implementation schema with the dimension and fact tables) and
iv. Data Integration using cleansing techniques, data quality control, and ETL processes) (Bernabeu, 2010).

2.2. Data Warehouse (DW) Construction

i. The Inmon methodology, or top-down approach transfers the information from various Online Transactions Processing (OLTP) systems to a centralized DW, given that the DW is subject-oriented, integrated, time-variant and nonvolatile (Inmon, 2005).
ii. The Kimball methodology, or bottom-up approach.

This is the union of smaller data marts, where every data mart represents a business process or dimensional mode (Kimball, 2006). A data mart is a subset of the DW based on the same principles, but with a more limited scope.

iii. The Hefesto methodology

The Hefesto methodology is a hybrid approach that integrates the Inmon and Kimball methodologies.
(Bernabeu, 2010).

iv. The Server Integration services

SQL Server 2005 introduced SQL Server Integration Services for performing Extract, Transform, and Load (ETL) operations enabling the merging and consolidation of data from heterogeneous sources.

The SQL Server 2008 implementation of Integration Services builds upon the strengths of the previous releases by enabling data integration of files stored in different formats in multiple geographical locations.
(Ellis,2008).

2.3. Data Mining Audit of Journal Entries(JE)

Argyrou (2013) carried out Journal Entries audit using Extreme Value Theory and Bayesian analysis of Poisson. He assessed the veracity of a bipartite model that contained the JE based on extreme value theory and distributions via a series of experiments on a dataset. It can detect journal entries with a low probability of occurring and a monetary amount large enough to cause fraud and assist auditors to form opinions about JE.

Debreceny & Gray (2010) carried out a pilot study data mining Journal Entries for fraud detection. The paper explored emerging research issues related to the application of statistical data mining technology to fraud detection in Journal Entries. It set out the underlying issues that will guide effective and efficient data mining of Journal Entries and reviewed the standards from auditing regulators and guidance from the professional audit community. It also explored the potential for statistical data mining of large sets of Journal Entriesb by testing the statistical properties of Journal Entries.

There is a clear and pressing need for research on a variety of interrelated areas in data mining Journal Entries.
Based on the successful applications of data mining to other domains, it would appear that data mining holds the potential to improve both the effectiveness and efficiency of the auditors in their analysis of Journal Entries and fraud detection (Debreceny & Gray, 2010).

3. Method

Journal Entries Data mining tasks are comprised of the following Extraction, Transformation and Loading (ETL) processes:

i. Identifying and selecting the type of Journal Entries to data mine by gaining access to the enterprise accounts database. Journal entries exist both in manual and electronic form and some are integrated within the company's accounting information system.
ii. Extracting the Journal Entries data from source.
iii. Cleaning the journal entries data.
iv. Transforming the Journal Entries data.
v. Creating tables and attributes for the Journal Entries data warehouse.
vi. Loading the Journal Entries data into a data warehouse.
vii. Partitioning the journal entries data as training data for the model building, testing data and validating data, so as to ensure that the model is accurate.
viii. Carrying out Model exploration (visualization) through graphs and charts.
ix. Implementing results obtained from the data mining as aids to management decision making.

Items vii and viii above are outside the scope of this paper.

3.1. Types of Journal Entries to Data Mine.

Although all types of entries that record information into the general ledger and, in turn, the financial statements are qualified for auditing, the following are potentials for fraud.

i. Summarized journal entries (JE) by General ledger (GL) account to identify repetitive, unique account sequence and top occurring amount.
ii. Journal entries posted outside the working hours on weekends and holidays, summarized by day, month, year and time.
iii. Journal entries made immediately following a fiscal year to the prior year (Post closing entry).
iv. Journal entries posted to seldom used and/or unusual accounts.
v. Journal entries made to adjust, reclassify, and reverse suspense accounts or reserves allowance and expenses.
vi. Journal entries that equate to round multiples of N10,000, N100,000 and N1,000,000.
vii. Journal entries made below set approval limits.
viii. Journal entries whose debits less credits do not net to zero (i.e. debits to equal credits)
ix. Journal entries posting by unauthorized or casual staff.
x. Journal entries with gaps in number sequence and without description.
xi. Journal entries posting made to unreconciled accounts.

3.2. Building the Journal Entries Data Warehouse

The journal entries fraud detection system through a decision tree algorithm utilizes the SQL analysis services facility of SQL server 2008 to implement the data mining tasks as follows:
An architecture that consists of two subsystems, the database interface, and the Fraud Detection Engines.

3.2.1. A Database Interface Subsystems

This is the entry point through which the transactions are read into the system.
The journal entry database having successfully passed through the ETL processes described previously, is extracted to an SQL server (the host server) to obtain the SQL server database using SQL server integration services (SSIS). The SSIS is used to define the schema for the journal entries server database otherwise called the Analysis services database.
This analysis database contains the journal entries data to be mined in addition to its structures, the ultimate mining models, data sources and the data source views.

3.2.1.1. Data Source

A database source consists of a connection string plus information on how to connect the server by indicating the server name; the journal entries database location and connection process.

3.2.1.2. The Data Source View

A database source view (DSV) serves as the client interface where the selection, organization, exploration and manipulation of the journal entries data in the source are carried out. It enables Analysis services to view the data source. In addition, it enables the modification of journal entries data structure and selection of tables relevant to the fraud detection task. It can be used to find relationships among the tables, to add columns, to create calculated columns without modifying the original journal entries data source. It is also used to create the mining structures with the chosen mining model of decision tree algorithm.

3.2.2. The Fraud Detection Engines

The engines comprise of a Business Intelligent Development Studio, editors and an SQL management studio. The BI intelligent studio and editors are used to create a mining structure, create test view and examine models, using the custom viewers and accuracy charts. The SQL Management Studio is used to manage the mining models. provide the tools for security, process back up and restoring databases with other management functions. The viewers, accuracy charts and prediction builder are utilized in the SQL Management Studio.

4. Finding and Discussion

With the journal entries data properly stored in the database, a connection between the Data mining tool and the database is established before the data mining commences. The data mining engine uses the server and database information provided.

4.1. Data Cleaning, Transformation and Loading

4.1.1. Data Cleaning

Data cleaning and transformation are the most resource consuming steps in a data mining project, Data cleaning removes errors and irrelevant information from the journal entries data set.

The cleaning process is also used to correct inconsistent values; to identify and remove outliers that may affect the modeling results, affect the classification and prediction precision of the models.

4.1.2. Transformation

Data transformation modifies the source data to make it useful for mining.

The journal entries data is put into a format that the decision tree data mining algorithm will accept having completed the cleaning process that also eliminate unwanted columns and rows.

Journal entries transaction input consists of the following data attributes. This is used to construct the Journal entry tables for the data mining task.

Input Variables	Data Type	Content Type
Journal Entry Number	Numeric	Discrete
Key	Alphanumeric	Discrete
Posting date	Alphanumeric	Discrete
General Ledger Acct Type	Alphabetic	Discrete
(e.g., suspense, receivable, revenue, expense, etc.)		
Entry date/time	Alphanumeric	Discrete
Posting Period	Alphanumeric	Discrete
General Ledger Acct code	Alphanumeric	Discrete
Journal Entry Serial No	Numeric	Discrete
Journal Entry Prepared by	Alphabetic	Discrete
Journal Entry Adjusted by	Alphabetic	Discrete
Journal Entry Approved by	Alphabetic	Discrete
Approval Limit	Currency	Continuous
Journal Entry Acct Descr	Alphabetic	Discrete
Journal Entry Amount	Currency	Continuous
Journal Entry transact desc	Alphabetic	Discrete
Fin Statement Mapping	Binary	Discrete
Debit presence indicator	Binary	Discrete
Credit presence indicator	Binary	Discrete
Journal Entry Line number	Numeric	Discrete
Currency Sign	Text	Discrete

Table 1: Journal entries data attributes transformation/encoding

The Data Mining Engine would detect whether a column is discrete (categorical) or continuous by sampling and analyzing the source data and choosing an appropriate content type. If a continuous type is determined and a selected algorithm does not support continuous columns, the content type will be specified as DISCRETIZED (the continuous values broken into discrete ranges). It will be verified that the content types of the journal entry table (see table 1) were assigned correctly, and any that is not will be modified.

4.1.3. Typical Journal data Entries key Variables (underlined) used for the Construction of the Decision Trees are as Listed below:

i. Journal entries summarized by general ledger account for repetitive and unique account sequences
ii. Summarized general ledger activity on the amount field

iii. Journal entries posted on weekends and holidays
iv. Journal entries of prior year posted after fiscal year-end .
v. Journal entries summarized by day, month and year.
vi. Journal entries to suspense accounts
vii. Journal entries errors corrected
viii. Revenue debits summarized into general ledger account.
ix. Exceeded general ledger average transaction amounts by a specified percentage.
x. Journal entries equating to round multiples of 10,000, 100,000 and 1,000,000.
xi. Journal entries made below accounting approval limits
xii. Journal entries that don't net to zero (debits less credits).

The selected journal entries' data attributes and variables consist of both discrete and continuous attribute types. In the typical journal data entries key variables reproduced below, the attributes underlined in iii,iv,v,above respectively, are Journal entries posted on weekends and holidays, of prior year posted after fiscal year-end, summarized by day, month and year are discrete, while the attributes i,ii,vi,vii,viii,ix,x,xi and xii: Journal entries summarized by general ledger account for repetitive and unique account sequences, Summarized general ledger activity on the amount field,. posted to suspense accounts, errors corrected, Revenue debits summarized into general ledger account, Exceeded general ledger average transaction amounts by a specified percentage, equating to round multiples of 10,000, 100,000 and 1,000,000., made below accounting approval limits, and that don't net to zero (debits less credits). are continuous type.

The continuous type attributes are converted to discrete values by applying the Supervised Discretized method module available in SQL Server Integration Services (SSIS) software. This method also helps to remove the outliers not completely removed during the data cleaning phase.

4.1.4. Loading the Journal entries into the Data warehouse

Microsoft SQL Server Management Studio is the platform on which the database(s) would be created. It is a relational database management system developed by Microsoft.

The Journal Entry data is extracted to an SQL Server database (the host server and the corporate data warehouse) using SQL Server Integration Services (SSIS). The SQL Server Analysis Services is thereafter used to define the schema for the journal entries objects in a single Analysis Services database created from the SQL Server data base. An Analysis Services journal entries database would contain the mining journal entries, its structures, the ultimate mining models, and the supporting objects such as the data sources and the data source views.

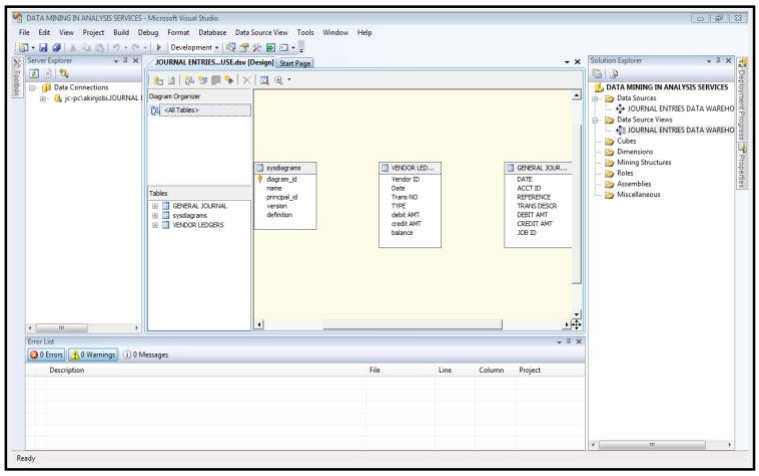

Figure 2: The Journal Entry Data Source View Created With Related Schema

4.1.5. Connecting to the SQL Server Database

With the data properly stored in the data base through the processes earlier listed, a connection is set up between the Data Mining Tool and the SQL Server database before beginning to define, adding columns to and creating the model. This requires working with the SQL Server holding and organizing the raw data, and the Analysis Server where the models are built and administered. The Data Mining Tool uses the server and database information provided to create a connection to SQL Server and Analysis Services database.

Figure 3: Connecting to a Data base server.

4.2. The Data Mining

4.2.1. Creating and editing the Journal Entry Fraud Mining Model and Structure

The first step in creating a Journal Entry Fraud Mining Model is using the Data Mining engine of the SQL Server Data Tools (SSDT) to create a new JE mining structure for models based on Decision Trees. The mining structure describes the columns and training data that will be used for mining, and optionally a mining model, which takes those columns, applies an algorithm, and defines the usage of each column for that algorithm.

4.2.2. Data Mining Algorithm for fraud of non-approved and non-authorized Adjusted Journal Entries

If JE was not approved but contains an adjusted entry then label as Fraud. If JE was not approved but without an adjusted entry written off to bad debts then label as fraud

If JE was not approved, does not contain an adjusted entry, not written off to bad debts but reversed for a particular customer then label as fraud

If JE was not approved, does not contain adjusted entry, not written off to bad debts and not reversed for a particular customer then label as fraud

If JE was approved by a non-authorized accounting officer then label as fraud

If JE was approved by an authorized accounting officer then label as non-fraud

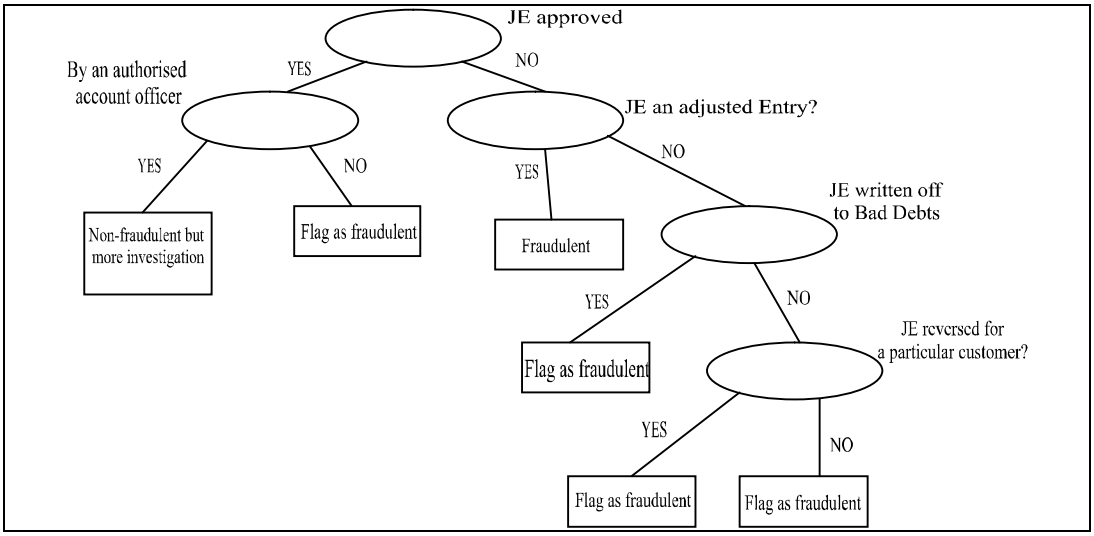

Figure 4: Decision Tree Model: Adjusted Journal Entries

This algorithm used for the data mining, identifies JE amounts adjusted, reversed and written off without official approval

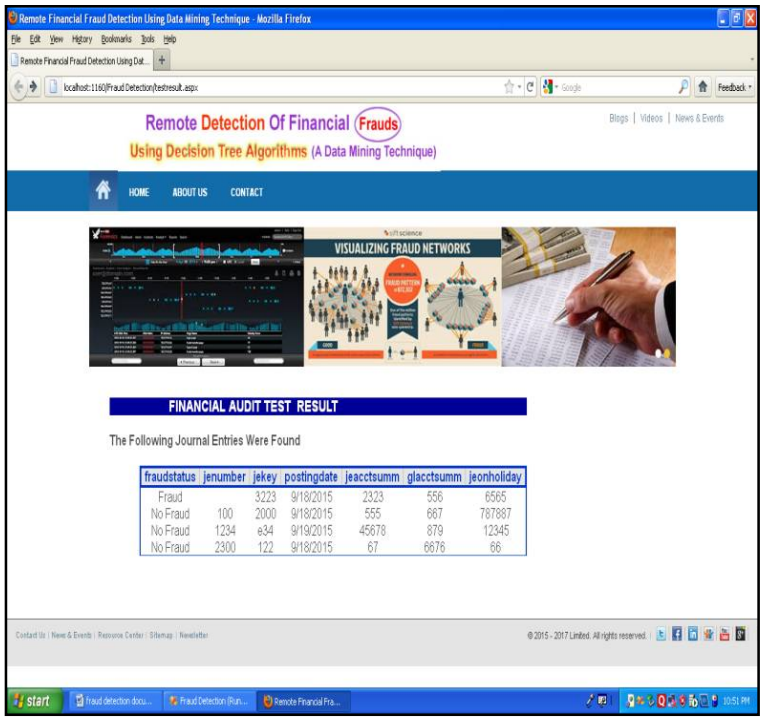

Figure 5: Results of the Decision Tree Analysis

5. Conclusion and Suggestion

According to the study carried out by Debreceny & Gray (2010), "There was no literature that models the statistical properties of populations of journal entries. Nor is there a literature that takes exemplar databases of journal entries and tests the statistical properties of those databases". The study further said that "it is questionable that direct auditor assessment of small samples of journal entries will effectively and efficiently detect likely patterns of fraudulent activity. Although there are large bodies of literature regarding data mining in other domains, a broad search of audit literature did not locate any research literature on the data mining of journal entries. Yet, auditing standards require that auditors consider fraud in their financial audits and those standards specifically require that auditors examine journal entries".

There was therefore a clear and pressing need for research on a variety of interrelated areas in data mining journal entries.

The research conducted in this framework filled those gaps identified by Debreceny & Gray, (2010) and others, to serve as contribution to the literature to enhance the body of knowledge in the area of detecting journal entries fraud in multidivisional enterprises.

In Nigeria, many frauds are never discovered while those discovered are concealed by concerned enterprises management in order not to damage their reputation and avoid costs associated with investigation.

The journal entries fraud detection system through a decision tree algorithm utilizes the SQL analysis services facility of SQL server 2008 to implement the data mining tasks through data cleaning, transformation and loading into a data warehouse.

6. References

i. Adejuwon A. J (2011). The Relevance of International Financial Reporting Standards to the Nigerian Economy. Lagos, J. A. Adejuwon & Co.**D**

ii. Aral,K.D., Güvenir, H.A., Sabuncuog,T., & Akar, A.R.(2011). A prescription fraud detection model.An Elsvier journal publication D

iii. Argyrou, A.,(2013). Auditing Journal Entries Using Extreme Value Theory. Proceedings of the 21st European Conference on Information Systems. 22, 00101 Helsinki, Finland,

iv. Argyrou, A.,(2013). Developing Quantitative Models for Auditing Journal Entries. A PhD Theisis, Hanken School of Economics.

v. Bay, S., Kumaraswamy,K., Anderle, M.G., Kumar, R., & Steier, D.M(2012). Large Scale Detection of Irregularities in Accounting Data. Center for Advanced Research, PricewaterhouseCoopers LLP

vi. Bernabeu, R. D. (2010). Data Warehousing: Research and Concept Systematization - HEFESTO: Methodology for the Construction of a Data Warehouse. Cordova, Argentina.

vii. Center for Audit Quality.(2008). Practice Aid for Testing Journal Entries and Other Adjustments Pursuant to AU Section 316. A publication of the Center for Audit Quality.

viii. Chaudhuri, S., & Dayal, U. An Overview of DataWarehousing and OLAP Technology.Microsoft Research, Redmond Hewlett-Packard Labs, PaloAlto.D

ix. Chintalapati S., Jyotsna G. (2013). Application of Data Mining Techniques for Financial Accounting Fraud Detection Scheme, International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 11. D

x. Chorba, R. W., &Bommer, M. R. W. (1983).Methodology for the Construction of a Data Warehouse. Cordova, Argentina. D

xi. Codd, E.F., Codd, S.B., & Salley, C.T.(2003). Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate", E.F. Codd and Associates, 1993 (sponsored by Arbor Software Corporation). D

xii. Debreceny, R.S.,& Gray, G.L.(2010). Data Mining Journal Entries f or Fraud Detection:A Pilot Study Source:

xiii. Ellis, M.(2008). Connectivity Options for Microsoft SQL Server 2008 Integration Services. An SQL Server Technical Article.

xiv. ETL Tools.com.(2014).Data Extraction in the ETL. Retrieved from http://www.etltools.org/ extraction.html

xv. Ghanbari, M.K.,& Einakian, M.(2014). Using "Data Mining" to Detect Frauds of Internal Audits. Proceedings of 9th International Business and Social Science Research Conference , Dubai, UAE, ISBN: 978-1-922069-41-2 D

xvi. Gray,G.L.,& Debreceny,R.S.(2014). A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits. International Journal of Accounting Information Systems
D

xvii. Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Indianapolis, IN: Wiley Publishing, Inc.

xviii. Kimball, R. (2006). The data warehouse toolkit. John Wiley & Sons.

xix. Kirkos, E., Spathis, C., &Manolopoulos,Y. (2007). Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, 32, 995-1003.