Securing Detected Hidden Associations from Big Data using Encryption
A case Study on: Breast Cancer Patients’ Data*

Heba Ayeldeen¹, Kareem Kamal A.Ghany², and Olfat Shaker³
1 Cairo University, Faculty of Computers and Information, Cairo, Egypt
2 Faculty of Computers and Information, Beni-Suef University, Egypt
3 Department of Medical Biochemistry and Molecular Biology, Faculty of Medicine, Cairo, Egypt

www.isirlab.net

Abstract— Big Data can solve business and world problems. For most organizations and complex systems big data is the reality of doing business. Of the major crucial questions concerning companies and organizations are what if big data is difficult to work? What if handling big and huge amount of data is hard to do? Taking the right decision at the right time is considered the main and critical issue most of the organizations are facing nowadays. Not just taking the right decision at the right time and any time but also the data handling and consistency of data. Understanding the data, transforming, and shaping the data is a key strategy for any organization to achieve competitive advantage. To support operations, management and decision making, organizations are in need of information systems and strategic tools. Business Intelligence as a strategic tool aids in problem solving on the right time and every time. In the way of complex data analyzing and patterns identifying it is very important to store, manage and share large amounts of complex data in a securely way. So we used an encryption technique to secure the patients’ data because of its sensitivity.

Keywords—Soft Business Intelligence, Big Data, Associations, Encryption.

*All the breast cancer patients involved in the study were diagnosed at department of Biochemistry and Molecular Biology of Kasr Alainy Hospital of Cairo University. For all participants in this study, written informed consent was obtained as delineated by the protocol which was approved by the Ethical Committee of Cairo University.
I. INTRODUCTION

By time the volume of data increases, handling data and data consistency are the main success of any organization [9]. Back in 1960, the data was simple and can be easily handled in the form of master files. While in 1965, the complexity of data grew which in return made it a bit difficult to handle. The synchronization of data and hardware used to store the data increased eventually. By 1970, the concept of having a single repository to store data evolved. The databases management systems were used as single source of data for all processing.

Organizations started to grow which in return the volume of data grew eventually. In 1975, organizations started to look for online and high performance processing. In the 1980, the 4GL technology started to take place and huge data evolved. At the end handling this huge amount of data was the main critical concern for all organizations and complex systems.

Decision makers and Executive managers were seeking for a single database serving all purposes paradigm [7], [9]–[11].

The concept of data warehouse to be a single repository that can serve all kind of data popped up. The data warehouse solved the problem of data handing, data integration, data consistency, extraction and data mining; and even online analytical processing of any kind of huge data [2], [12].

Different systems, tools and techniques are used to deal with the big data. Depending upon the different levels within the organization as well as the business needs of the systems, Information systems and decision support systems differ [10], [13].

Several industry specific areas are in need of system analysis strategic tool as well as data mining and symmetry tool to cope with the environmental and technological changes [14].
Business intelligence tools emerged. The number of existing Business Intelligence structures today and the functionalities they offer might be considered as an evolution. However, the concept of Business Intelligence (BI) has existed since the 1970’s more popularly known as Decision Support Systems (DSS) back then [15].

Big Data is the term used to describe huge volumes of unstructured and structured data that are so large and it is very difficult to process this data using traditional databases and software technologies.

Since the data is present in the machines in a cluster, a hacker can steal all the critical information. Therefore, all the data stored should be encrypted. Different encryption keys should be used on different machines and the key information should be stored centrally behind strong firewalls. This way, even if a hacker is able to get the data, he cannot extract meaningful information from it and misuse it.

The rest of the paper is organized as follows: section II discusses the evolution of Business intelligence, section III shows the architecture and main components of BI. While section IV discusses the importance of data mining showing the main data mining methods used to determine significant patterns and hidden associations from big data. In section V we present our proposed model, and in section VI the results of implementing SAP BI on the selected data set are shown. The last section is the conclusion.

II. EVOLUTION OF BUSINESS INTELLIGENCE

The history of information systems coincides with the history of computer science that began long before the modern discipline of computer science emerged in the twentieth century. An Information System is any combination of information technology and people’s activities using
that technology to collect, process, create, and distribute data to support operations, management and decision-making [8].

Handling large amounts of information, performing complex calculations, and controlling business processes are of the main important task any organization needs. An information system consists of four main components that are people, equipment, procedures and data. In other words, information systems focus on the use of equipment such as networks of hardware and software in order to help people and organizations collect, filter, process, create, and handle data and take the right decision at the right time [10].

Depending upon the different levels within the organization as well as the business needs of the systems, Information systems differs. In the 1980s organizations started to categorize the information systems based on the different levels of hierarchy in an organization. As shown in figure 1, the Transaction Processing System (TPS) is at the bottom of the pyramid, followed by management information systems (MIS), decision support systems (DSS), and ending with executive information systems at the top (EIS) [10], [16].

Fig. 1. Architecture pyramid of Information system [16].
A transaction processing system is an information system that supports the operational level and transaction processing of the business within an organization. TPS collects, stores, and organizes the day-to-day activities for supplying data for higher-level management decision. The transactions differ from organization to organization. A transaction can be any activity of the organization. TPS provides high speed and accurate processing of record keeping of basic operational processes [17].

Previous to 1965, it was very expensive to build large-scale information systems. At about this time, the concepts of Management Information System was developed in many companies that made it more practical and effective when using the data. A MIS usually takes large quantity of data from the transaction processing systems and stored in databases. Data is then summarizes it into a series of management reports for middle managers to aid top managers take the right decision [6], [10]. In the late 1960s, decision support systems evolved through time and became a practical model-oriented after the information system. DSS is another type of information system that are flexible, adaptable to the changes within the organization as well as the environment changes [7], [18], [19]. Decision taken by the decision makers and top managers should be decisive and taken on the right time [20].

The Decision support systems have various subcategories, of which is executive information system and support system. Executive information system (EIS), also known as an Executive support system (ESS) which is used by top executive/strategic managers. ESS designed to provide the senior management level information in a readily accessible and interactive format to be used in strategic decisions. In contrast to EIS, ESS provides additional application such as analysis support, communication, office automation and intelligence support. It involves lots of data analysis and modeling tools [4], [5].
The term business Intelligence is sometimes used interchangeably with briefing books and executive information systems. The concept of business intelligence has made significant advances over the last decades, since the emergence of decision support systems had taken shape with other versions of analytical software packages [21].

By the early 1980s, the capabilities of DSS evolved to include dynamic, multi-dimensional reporting systems with tools that could be used for forecasting, analyzing trends, drilling down, and others [10].

Fig. 2. Evolution of Business Intelligence

As the volume of data gets much bigger organizations are in need of a system or a tool that can easily handles this vast amount of data. Business intelligence (BI) systems deal with the vast amounts of data. Business Intelligent systems extract useful information from this data in the form of patterns, trends and present this in an understandable way to decision makers [22].
III. BI ARCHITECTURE AND COMPONENTS

This section explains briefly the main architecture and components of business intelligent systems. As in figure 3 the Business Intelligence process is clearly illustrated starting from getting the data into the data warehouse till the end of the process with a well-organized report [23], [29], [33], [34].

Figure 3 overviews many of the major BI processes. Looking at the stages:

- In stage 1: data is extracted from different data sources through the Extraction Transformation and Loading Process (ETL). ETL tools are responsible for identifying and extracting relevant data from operational systems, customizing and integrating it into a common format for updating the data warehouse area. Types of data sources vary. Data sources can be operational databases, historical data, and external data for the data sources can be also relational databases or any other data structure that supports the line of business applications. They also can reside on many different platforms and can contain structured information, such as tables or spreadsheets, or unstructured information, such as plain text files or pictures and other multimedia information.

- Then comes the role of the data warehouse (DW) in the second stage where data are then loaded into the cubes/multi-dimensional sources.

- In stage 3: within the BI server queries are performed as well as columns and rows are built for analysis and reporting.

- At last comes the user interface section where reports are built by which all business modules are supported with it.
BI workloads have very different characteristics than the traditional transaction processing workloads. These major differences include that the detailed level of granularity of data is more summarized and consolidated, the database size eventually is huge compared to the operational database and complex queries can be generated easily (35). Thus, BI frameworks are accordingly evolved to support such needs this involves broadening organizations horizons so that making decisions are faster, more informed and more accurate [35]. Components of BI can be summarized into Data sources, data warehouse and data marts, the OLAP processor, advanced analysis for query generation. Data sources which are temporarily the tables
the fields are stored in, are considered to be the first layer that data is store in the BI system. Data sources can be a feature of many different platforms and can contain structured information, such as tables or spreadsheets, or unstructured information, such as plaintext files or pictures and other multimedia information [35].

As the business success factor for any enterprise is finding ways to bring vast amount of data that is flowing within and across the business processes together and making sense out of the data and here comes the importance of the data warehouse which is a significant component of business intelligence [36]. To support multidimensional analysis OLAP technology is used. With OLAP technology, business users can slice and dice their way through data using sophisticated tools that allow for the navigation of dimensions such as time or hierarchies [37].

OLAP provides user with summarized views of business data and provides quick answers to ad-hoc queries against the data warehouse. When talking about the multidimensional analysis comes the role of InfoCubes in which data are stores in cubes for further analysis [1], [37].

![Multidimensional InfoCube](image)

Fig. 4. Multidimensional InfoCube
InfoCubes can be either by using the star schema or the snowflake schema. For example, assume that Water Quality Index data has been integrated into a Cube. One dimension would be the “where” that is the point where the measurements were made (Key Figures). The time dimension would be the “when”, that is the date of measurement (Time Dimension). In the same set we would like to have the “version” dimension in which we could compare the actual value against the desired or allowed value and a variance in percentage (Dimension). With the above mentioned dimensions we can get the what-if analysis as well as compare between the actual data and the planned data.

Of the main component of BI is the analysis part where advanced analysis is referred to as data mining, forecasting or predictive analytics. Data mining is a set of advanced decision support systems that aids in statistical analysis and data prediction. Data mining algorithms find patterns in the data and report models back to the user [38–40]. BI tools are widely used for reporting, dashboard and analysis of data [25].

IV. The Proposed Model

In our proposed model we made a combination between the knowledge management system as well as the software engineering system. This model consists of four main phases as shown in fig. 5 and described as follows:

Phase 1: Data processing

The data is being preprocessed where the reduction of attributes is taken place like the factors/parameters that do not directly affect in the results are being removed. For instance the
historical data of the patients and the age factor. Data is then indexed to ease the process of representation and retrieval.

**Phase 2: Data storage**

The data is saved in the business warehouse for further processing actions. Each case study we have is well stored in the warehouse with ID for the patient (healthy cases or carcinoma patient) with the relative data corresponding to the case.

**Phase 3: Analysis**

Several analysis steps can take place in the analysis phase where correlations and associations between different cases take place. The KPIs (Key performance Indicators) are well defined according to each case and according to the business need. In this step decision makers (physicians) are being able to take the right decision on time and all the times.

**Phase 4: Data visualization**

Nowadays, one of the main problems concerning many decision makers is knowledge representation. Data visualization and knowledge representation tools aid executive and strategic managers easily predict and take the right decision. SAP Business Object tool (Xcelcius 2008) is used in the model to visualize the data, showing the correlation between proteins (LAPTM-4B and OPG).

Finally, we should indicate that we applied an encryption algorithm in the security layer after the preprocessing phase to encrypt the patient’s data to protect his/her privacy during other phases.
Patient’s data will be stored securely in an encrypted manner. An MD5 [44] hash is created by taking a string of any length and encoding it into a 128-bit fingerprint. Encoding the same string using the MD5 algorithm will always result in the same 128-bit hash output.

Fig. 5. The proposed model

V. CASE STUDY: BREAST CANCER IN EGYPTIAN PATIENTS

In Egypt, breast cancer is the most frequent cancer disease spread (28% of all cancers) and the most fatal form of malignancy among women, accounting for 15% of cancer deaths. Breast cancer is when cancer develops from breast tissue where a change in breast shape, the skin changes, fluid coming from the nipple, or a red scaly patch of skin. When the disease spread more, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin [41].
A. Dataset

The data set used within this study is collected and diagnosed at the Department of Biochemistry and Molecular Biology of Kasr Alainy Hospital of Cairo University. To detect the relationship between LAPTM-4B polymorphism and breast cancer vulnerability, one hundred three breast cancer patients and eighty cancer-free healthy controls who were recruited from patients undergoing annual physical examination at Kasr Alainy Hospital of Cairo University were investigated.

The studied subjects were divided into three groups as follows: Group I: (n=80) healthy females as a control group. Group II: (n=40) patients with fibroadenoma. Group III: (n=88) patients with breast carcinoma, they were classified according to TNM grading system into 11 cases in stage II, 57 cases in stage III and 20 cases in stage IV. This group included 68 non metastatic breast cancer patients and 20 metastatic subjects. We mainly used two of the groups (Control and Breast carcinoma patients).

The protein was extracted from whole blood serum of both patients and control group with QIAamp DNA mini kit (Qiagen, Hilden, Germany), following the manufacturer’s instructions. LAPTM-4B abbreviated for Lysosomal-associated transmembrane protein 4B is a protein that in humans is encoded by the LAPTM4B gene. Clinical significance results in Increased expression of LAPTM4B has been found in breast, liver, lung, ovarian, uterine as well as gastric cancers.

Elevated LAPTM4B level contributes to chemotherapy resistance in breast cancer [42].

Osteoprotegerin (OPG) is a protein that in humans is encoded by the TNFRSF11B gene. Not just in breast cancer cases, elevated OPG levels has been reported in heart diseases as well [43].
B. Mathematical and Statistical Analysis

All statistical analysis in our study were carried out with SAP BusinessObjects Dashboards and WEKA 3.7.9 (WEKA, The University of Waikato).

C. Implementation steps

Most likely to use SAP BusinessObjects Dashboards, it is recommended to build the data model of the dataset by SAP Business Warehouse. SAP BusinessObjects Dashboards tool gets the data from different DataSource including SAP source system, Databases, Mobile Applications, File systems or even XML files.

Below are the steps used within the study:

1) Data staging
2) Dashboard connectivity
3) Dashboard creating
4) What-if Analysis

Data is extracted using CSV file added into SAP BusinessObjects Dashboards tool. Data is then cleansed to avoid any special characters. Connection between the CSV file with the tool to start building the required dashboards. The main aim of the dashboards is to aid decision makers find the associations and correlation between the data to make the right decision on time.

VI. INTERPRETING RESULTS
The main aim of the study is to find the correlation between proteins to aid decision maker early detect breast cancer disease. As shown in the table within the dashboard shown in figures 6 to 11, family historical data as well as diabetes and hyper tension data was collected.

Fig. 6. Master table analysis

The data can be used to detect whether family history is considered as a factor in breast cancer detection. Other data was collected like details about menstrual history, tumor size and type whether it is inductive duct i, ii, or iii.

Based on the business needs and the capabilities of the users using the dashboard, many analysis can be easily done with the functionalities of the tool.
Fig. 7. Drop down menu properties

Fig. 8. Push button analysis properties
There is a significance correlation between LAPTM-4B protein and OPG protein, with a cutoff value 710 and 231.3 respectively. This can conclude that these proteins highly aid in the early detection of breast cancer.

Fig. 9. Statistical graph analysis for LAPTM4B and OPG proteins

Fig. 10. Correlation between LAPTM4B and OPG proteins
For the patients with result “Cancer” we have an encryption key associated with their data to avoid any privacy attack, the following is some examples of the MD5 generated keys:

For patient no. CC1 we have \( 8a5a70b866e5cdf3d281b75c173ecdd8 \)

For patient no. CC2 we have \( 7720d299d306a5c813c6bcfa330ac26a \)

For patient no. CC3 we have \( 91cba18826f46b110fa2aa018cc4a322 \)

VI. CONCLUSION

Finally it can be concluded that the evolution of information systems directly effects decisions by identifying what should be done and ensure that the chosen criterion is relevant which affects the whole success of any organization. The usage of data mining and visualization tools highly affect
the performance of organization and decision makers. We succeeded in applying the SAP BusinessObjects to automatically detect the hidden associations from the big data (Breast Cancer Patients’ data), with respect with the security issues by protecting the patient’s privacy using the encryption technique.

REFERENCES

[27] Gangadharan and Swamy S.N. "Business intelligence systems: design and implementation strategies” Information Technology Interfaces, 26th International Conference, vol. 1, 2004, pp. 139 - 144.