**Data Warehousing and Data mining**

**Charles Garalis**

**INFO 4307**

**Dr. Merlo**

**03/11/2018**

**Abstract**

A data warehouse stores historical data about an organization in an infinite number of databases. This is done through multiple different types of architectures which are also used in the analytics process. Data warehousing is used for query and analysis to create business intelligence, in order to drive a company’s business goals. Data mining is the act of using different methods to extract information (patterns) from the historical data in a data warehouse. There are serious issues that still exist today with data warehousing, most of which have to do with security such as theft and unauthorized access. These security issues are met with encryptions of historical data along with several sept processes in order to identify different forms of security threats.

**Introduction**

Historical data all being used for analytics purposes in order to drive an organization towards its goals. This is essentially what data warehousing is. If used correctly a data warehouse can create an extensive amount of business intelligence, in turn driving a great profit for the organization that it is a part of. Data mining is used extensively in knowledge creation and knowledge organization. Knowledge creation can be and often is the heart of an organizations competitive advantage. The practice, interaction and learning you see from knowledge creation inside of a company all stems from data mining. This is because of the massive amount of intelligence extracted from the historical data found in a data warehouse through data mining. Between data warehousing and data mining, the tasks to keep it running smoothly become very costly and can consume a massive amount of time. However, because of reasons previously explained, it is well worth the effort.

In this paper we will discuss the different forms a data warehouse can take, for example the different architectures (Basic, Basic with staging area, Basic with staging area and data marts) that can completely refine the type of data you are analyzing. We will also discuss the relationship between a data warehouse and data mining, as well as some of the issues that exist with both data warehousing and data mining. Furthermore, we will discuss the different types of methods that exist with data mining, what they are used for and how they are different, along with an example of each.

**Data warehousing and data mining**

Data warehousing and data mining go hand in hand. One cannot survive without the other. A data warehouse is a huge compilation of a nearly infinite amount of data bases, full of historical data. All of this historical data is used for query and analysis in order to create business intelligence for organizations to profit from. The definition of a data warehouse is a database designed to enable business intelligence activities, existing to help optimize their organizations performance. Data warehouses are subject oriented, integrated, non-volatile and time variant sources of analytical information. All of this information is consolidated and stored as historical data.

The aspects of data warehousing are vast. Most of them are applications that manage the process of gathering data. These types of processes that gather data are things such as; Extraction, transportation, statistical analysis, reporting and data mining capabilities. All of these processes work in transforming data into business intelligence that can be used to drive an organization towards its goals.

Architectures are also an important piece in the building process of a data warehouse. There are three common architectures that most organizations use regularly, these architectures consist of; A Basic architecture, an architecture with a staging area and lastly an architecture with a staging area and data marts. A basic architecture has end users that directly access data derived from several source systems through the data warehouse. An architecture with a staging area is a little more complicated and has different stops along the way. Your operational data must be cleaned and processed through an extra staging area before being stored in a data warehouse for analysis. This staging area simplifies data cleansing and consolidation for operational data coming from multiple sources. The last common architecture to talk about is the architecture with both a staging area and a data mart. Now that we already know what a staging area is, time to explain a data mart. A data mart is most simply defined as systems designed for a particular line of business, such as, sales, purchasing or inventory. In this architecture it classifies different types of raw data that an analyst can then use to make more specific predictions about specific lines of business rather than the overall historical data present.

Now that a little of what a data warehouse is has been explained, we can get into the tools of a data warehouse. One of the most important tools that we will talk about is data mining. Data mining is a very simple concept that can be used in a various amount of ways. To define data mining, it would be the practice of examining large databases in order to generate new information. More specifically it is discovering patterns in those large data sets and those patterns are what generates new information. The data that is extracted through the data mining process is like a summary that can be used to obtain accurate predictions in knowledge discovery. However, data mining is its own step and not to be confused with the knowledge discovery process. Data mining is used before a decision support system comes into play.

There are several different methods of pattern detection that exist within data mining. Each one of these methods has a different way to detect patterns and a different meaning for its use. The first is Anomaly detection, in this method they use statistics to determine if something is notably different from the pattern that exists. For example, the IRS could use a model layout of common tax returns and use this method to identify the outliers that differ.

The second method is Association learning, this type of pattern detection is used in finding relations between types of data. For example, people that buy a cell phone tend to buy a cell phone case along with it. This method it targeted at detecting this kind of relational data. This method is also most used but advertising.

The third method to explain is Cluster detection. This type of pattern recognition, identifies distinct clusters of sub-categories within a set of data. Machine learning algorithms detect all of these different subgroups within a data set that differ from each other. This happens in the same say that a fisher man will purchase differently than that of a designer. Or how a farmer with purchase differently than a librarian. This method picks up on these different sub-groups of data and differentiates them.

The fourth method is classification. This method only works if an existing data set is already known. If this is the case, then this data mining method can be used to classify new cases into these pre-determined categories. A good example of this method is a spam filter in an e-mail box. Filters have been enabled to notice differences in pre-determined word usage which allows them the determine between real messages and spam from a machine.

The fifth and final method to explain is called regression. This last type of pattern recognition uses historical data in order to predict future events. For example, amazon may use your past purchases from their website in order to predict what you might use their website for in the future. This encouraged behavior towards an organization and drives revenue.

All of these different methods show data mining in a very intricate and positive light. However, there are still flaws and inconsistencies in data mining that need to be dealt with. These flaws with data mining in turn greatly effect data warehousing, creating a domino effect. Some of the issues faced in data mining are things such as poor representations in data sampling. This can be poor quality data or inadequate data size, both of these can lead to a poor analysis of historical data, which can in turn cause an organization to suffer. Another problem with data mining is the time it takes to extract the data needed and the find patterns in the data. Data mining needs to be a very efficient process. If the efficiency and scalability of data mining algorithms to effectively extract the information from huge amounts of data in databases is not timely, it can cost the organization a lot of money.

Luckily in most cases, solutions have been brought up in data warehousing to combat these problems. For example, an operational data store or ODS has been created in data warehousing to deal with the latter of the two problems mentioned earlier. What an operational data store does is works as a subject oriented system that is optimized for looking up one or two data sets at a time for decision making. This creates a much more cost-efficient process for any organization, making the data warehouse as a whole run much smoother.

Other than problems with the analysis of data, there are serious issues with security in data warehouses. The information and historical data that a data warehouse holds are vital to an organization. If fallen into the wrong hands it can be catastrophic for any company. However, there are actions that organizations take to prevent this. Organizations will encrypt their data allowing only people with the right code access. Organizations will often pair something like this with a several step processes in order to identify and neutralize new forms of security threats. They use this process in order to keep their security from becoming static, because the world of theft doesn’t stand still either. There is even a type of security model called a metadata security model. This security model is unique in its way it uses the metadata that describes the warehouses contents. When a user goes to access data there is a security management layer that checks if access is allowed. This is done by verifying the corresponding access authorizations by analyzing security metadata. This method was made to avoid fundamental, cost intensive adaptations.

**Conclusion**

Data warehousing has become and always will be the centerpiece to data analytics in an organization. Without a data warehouse there would be no organization in information. There would be no able way to use data mining because there would be no historical data to detect patterns from.

After having discussed the different types of architectures that exist with a data warehouse and the different methods that exist with data mining, we can plainly see that there are many different ways to process information. Anything from an algorithm-based machine learning to classify different subgroups of data to a simple spam filter on an e-mail account. All of these types of technologies use data mining every day. However, as I talked about in this paper not quite every aspect of data mining and data warehousing is good. There are still major issues that need to be addressed in security, cost efficiency and the time it takes to find specific patterns in the data mining process.

In conclusion, data warehousing is the backbone of any major organization. Its architectures allow for the historical data to be refined and classified into different sub groups to find clearer patterns and data mining methods allow for organizations to be able to find different types of patterns from those very sub groups that the data warehouse creates. In essence one cannot survive without the other.

**References**

IBM Informix, Overview of data warehousing IBM knowledge center (2011) Pg.1; Retrievedfrom:https://www.ibm.com/support/knowledgecenter/SSGU8G\_11.50.0/com.ibm.whse.doc/ids\_ddi\_344.htm

Introduction to data warehousing concepts Data warehousing and business intelligence (2001) Pg. 6-10; 1(12.1) Retrieved from: https://docs.oracle.com/database/121/DWHSG/concept.htm#DWHSG8075

Sweety Patel, what is Data Warehouse International journal of advanced innovations, thoughts and ideas Pg.1 1(1) ISSN: 2277-1891

Inmon, W.H. "The data warehouse and data mining." Communications of the ACM, Nov. 1996, p. 49+. Academic OneFile, Accessed 8 Mar. 2018.

Daniel Harris Data warehouse problems: do you even need one? software advice (2006) Pg.1; Retrieved from: https://www.softwareadvice.com/resources/data-warehouse-problems/