EasyChair Preprint

№ 4729

Data Warehousing and Data Mining Architecture and Concepts

Stephen Paul Almarez and Janelli Mendez

December 9, 2020

Data Warehousing and Data Mining Architecture and Concepts

Stephen Paul Almarez
DIFZ
Dubai City, UAE
(971)56 682 4151
Stephen.almarez@live.com

Janelli Mendez
Lorma Colleges
Lingsat, San Fenando City ,
Philippines, La Union
(63)920 849 6918
Janelli.mendez@lorma.edu

ABSTRACT

In today's modern technology, the world generates billions and billions of data each day. We have been given an unlimited amount of data source to tap in. The aim of this research is to get existing and the best way to extract this data via data mining processes and at the same time look in to data warehousing models to save this data for future use and analysis.

Thus, the first part of this research shows the phase in designing and development of data warehouse, And for the later part of this research shows the data mining algorithms for the purpose of deducting rules, patterns and knowledge as a resource for future use.

Keywords

Data Warehouse, Data Mining Architecture, Data Mining Concepts, Data Mining, Data Mining Techniques, Data Mining Architecture, Data Mining Concepts

INTRODUCTION

I. INTRODUCTIONS

In today's industry data can be assumed as currency, for in it contains unlimited amount of information that when properly manage and analyze can give us foresight and directions. Every industry and or anyone can capitalize with rich and sanitized data, with this the challenge at task is how do we keep this vast amount of data being generated? And how are we going to leverage this data to derived usable quality data from it? The Ability to react quickly and efficiently to trends in any industry becomes overcrowded with complicated data and if they we're able to transform them into useful information, we will have the advantage of being competitive and informed.

II. What is Data warehousing and Data Mining?

Data warehousing - Data warehouse is a technique for collecting and managing data from varied sources to provide meaningful business insights. It is a blend of technologies and components which allows the strategic use of data.

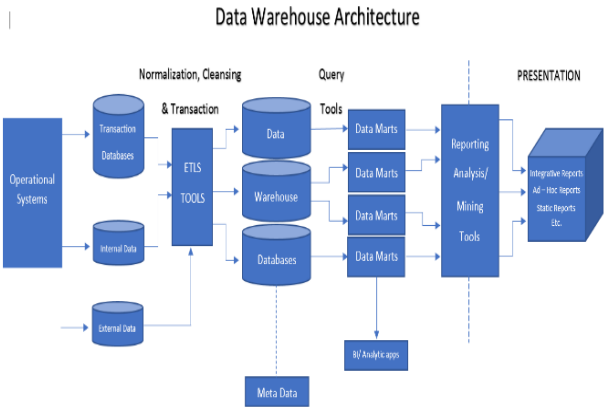
Data Warehouse is an electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users for analysis. Data warehousing also includes the process of aggregating data from various database sources into one place for efficient access and analysis. Another aspect of data warehousing is the architecture of the data—that is, how it's structured so that it can be joined, even if the sources have different fields and schema. As a record of a company's past operational and transactional information, especially given the proliferation of data sources, volume, and density, data warehousing is a key strategic differentiator in today's global marketplace.

Data mining - Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.

It is a multi-disciplinary skill that uses machine learning, statistics, A.I. database technology and Big Data.

It is the process of extracting value from the data stored in the data warehouse. Data mining includes the process of transforming raw data sources into a consistent schema to facilitate analysis; identifying patterns in a given dataset, and creating visualizations that communicate the most critical insights.

III. DATA WAREHOUSE ARCHITECTURE AND COMPONENTS



A. THREE TYPES OF DATAWAREHOUSE ARCHITECTURE

1. Single-tier architecture

- Single-tier architecture is mostly used for basic and or minimize data storage. It's main goal or functions is to remove data redundancy. This architecture is not used frequently in practice.

2. Two-tier architecture

- Two-tier architecture separates are designed to separate available sources and data warehouse. This concept is not expandable and doesn't support large numbers of end-users. Connectivity issue is also a rampant problem due to network limitations.

3. Three-tier architecture

- A widely used architecture that combines;

a. **Bottom Tier** – This is where the data warehouse servers reside. Usually this are relational database system and it is where data is cleaned, transformed, and loaded thru the front end.

b. **Middle Tier** – This can be an OLAP server that is either ROLAP or MOLAP implemented model. This represent the abstract view of the database, it is also where mediation between end user and the database occurs.

c. **Top-Tier** – This is where data entry and extractions happens, it's the front end where user can connect and interact with the database warehouse for tools and or API used for query, reports, analysis and data mining.

B. FIVE DATAWAREHOUSE COMPONENTS

1. DATA WAREHOUSE DATABASE

- The central database that builds the foundation of the data warehousing environment. Most likely this is implemented on a relational database management system (RDBMS) technology. Implementation of this kind of data warehousing is constrained by the fact that traditional RDBMS is optimized for transactional database processing against data warehousing. Samples instances can be ad-hoc query, multiple-table joins, and aggregates are resource hog and slows down performance.

2. SOURCING, ACQUISITION, CLEAN-UP and TRANSFORMATION TOOLS (ETL).

- This tools are used to perform conversions, summarizations, and all changes needed to transform the data into a unified format for the data warehouse. They are also knows as Extract, Transform and Load (ETL) Tools.

Common functionality

- Anonymize data as per regulatory stipulations.
- Eliminating unwanted data in operational databases from loading into the warehouse.
- Search and replace common attributes and definitions for data arriving from difference sources.
- Calculating summaries and derived data
- Default population for cases of missing data.
- Data de-duplication for repeated data arriving from multiple sources.

ETL Tools can generate cron jobs, background jobs, Cobol programs, shell scripts etc. that regularly update the warehouse. This tools are also helpful in maintaining the Meta data.

This tools have to deal with challenges of the Database and Data heterogeneity.

3. META DATA

- Can be summed as data about data or simply data characteristics that defines it and the data warehouse. It plays an important role in a data warehouse as it specifies the source, usage, values and features of the data warehouse. It also defines how it can be change and processed.

Example: Database barcode 123 K789 300

The data is meaning less until we consult the Meta data that tells it is.

Model: 123

Agent ID: K789

Unit Price: $300

Therefore we can assume that Meta data is essential in transforming data into knowledge. It also helps to answer the following questions.

- What tables, attributes and keys does the data warehouse contain?
- Where did the data come from?
- How many times does it get reloaded?
- What cleansing transformation is applied?

3.1 METADATA CLASSIFICATION

1. Technical Meta Data

- This data contains information about the warehouse that is used by the Data designers and administrators.

2. Business Meta Data

- This contains details that gives end-user to easily understand the stored information in the data warehouse.

4. QUERY/ ACCESS TOOLS

- Providing information is a primary object of data warehousing to business and industries alike to make a strategic decision or outcome. Query tools allows this to interact with the data warehouse system.

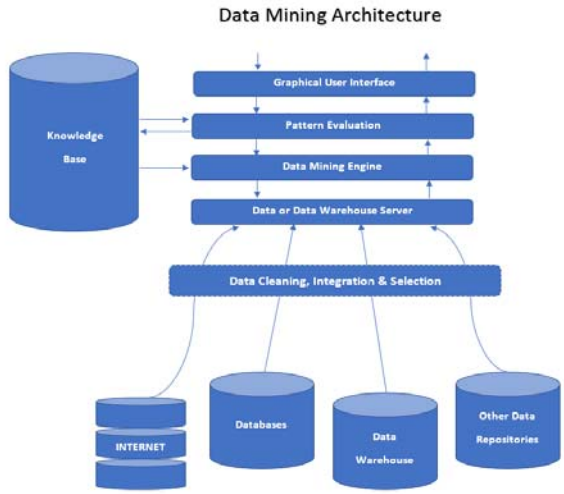
- Tool categories
  - Query and reporting tools
  - Application Development tools
  - Data mining tools
  - OLAP tools

5. DATA MARTS

- It is an access layer that is used to get data out to the users. It is presented as an option for larger size data warehouses as it takes less time and money to build. It can be said that it is a subsidiary

of a data warehouse, it can be used for data partition created for a specific group of users. It can be created within the same database as the data warehouse or on a physically separate database.

IV. DATA MINING ARCHITECTURE AND COMPONENTS



1. **DATA SOURCES**

- This are data's or historical data coming from different sources like the internet, a database, data warehouse, big data, text file and or documents. A large volume of data are more suitable for data mining to be successful.

This data must be first cleaned and normalize before integration or passing it to the database or data warehouse.

2. DATA STORAGE

The data storage can either be a Database or Data Warehouse Server and or both. This is the actual area where the data is stored and save for retrieving or processing during data mining by a user. The data stored should already be cleaned and integrated.

3. DATA MINING ENGINES

- This are the core components of any data mining systems. It's made up of different number of modules or apps to process and extract the quantifiable and qualitative data. Task such as data association, classification, characterization, clustering, prediction, reports, analysis, unification etc. are done here.

4. PATTERN EVALUATION MODULES

- Pattern evaluation is looking into regular series of data algorithm or it can also be said as a repetitive occurrence of data to form a measurable or common threshold.

5. GRAPHICAL USER INTERFACE

- This is where the user interacts with the data mining systems, it is where task and queries are entered to form a readable result from the stored data.

6. KNOWLEDGE BASE

- A knowledge base is a helpful guidance to look into for searches, previous data and or pattern where the resulting processes had been done before at which can help the data mining process to get a more accurate and reliable outcome without re-inputting the same pattern or code again to create the same output.

V. DATA MINING TECHNIQUES

1. DECISION TREES

- A common technique used for data mining because of its simple structure. The root of decision tree acts a condition which leads to specific data that helps us to arrive to a final decision.

2. SEQUENTIAL PATTERN

- As the technique implies it follows or try to discover a regular series of events in the historical data or on the data to help identify a common theme.

3. CLUSTERING

- It is used to define classes of data that has the same characteristic to form a clustered objects to form a common blocks of data.

4. PREDICTION

- It is a method to which we try to define the relationship between different instances whether common or uncommon to evaluate or arrive to an answer.

5. ASSOCIATION

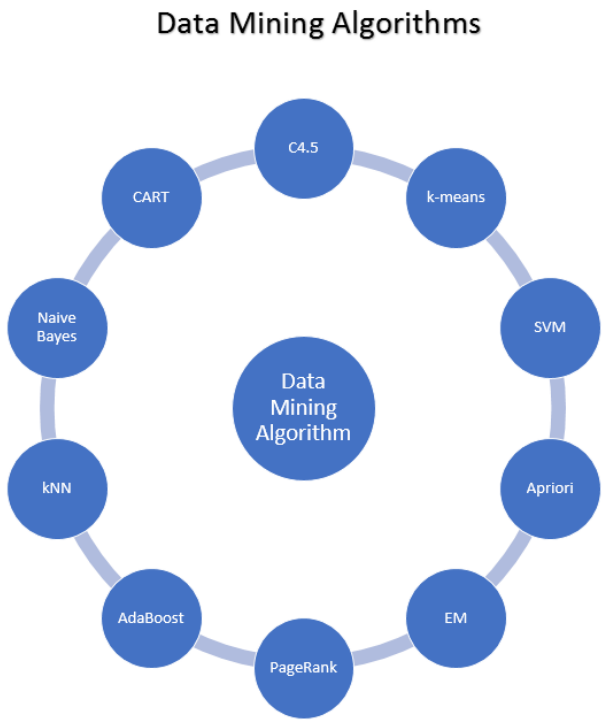
- Association can be termed as relation technique thru this we try to recognize a pattern based on a single transaction or data series.

6. CLASSIFICATION

- Based on Machine Learning, it's used to classify a particular set of each item into a particular group. This method adopts a mathematical techniques such as neural networks, linear programming and decision trees and so on.

IV. COMMON DATA MINING ALGORITHMS

An algorithm is a set of heuristics and calculations that creates model from a data or set of information. It is way to analyze specific types of patterns or trends, it can also be a mathematical model that forecast or predicts an outcome thru analysis. Below are 10 of the most commonly known Data Mining algorithm that we can apply or used in data mining in no particular order.



1. C4.5 DATA MINING ALGORITHM

- This algorithm constructs a classifier in the form of a decision tree. To do this a given a set of data representing things that are already classified is inputted to the C4.5 algorithm

2. k-means DATA MINING ALGORITHM

- The k-means algorithm creates k groups from a set of objects so that the members of a group are more similar. It's a popular cluster analysis technique for exploring a dataset.

3. SVM DATA MINING ALGORITHM

- Support vector machine (SVM) are algorithm that learns a hyperplane to classify data into 2 classes. At a high-level, SVM performs a similar task like C4.5 except SVM doesn't use decision trees at all.

4. APRIORI DATA MINING ALGORITHM

- This algorithm learns association rules and is applied to a database containing a large number of transactions.

5. EM DATA MINING ALGORITHM

- The expectation-maximization (EM) algorithm is generally used as a clustering algorithm (like k-means) for knowledge discovery.

6. **PAGERANK DATA MINING ALGORITHM**

- This algorithm uses a link analysis algorithm designed to determine the relative importance of some object linked within a network of objects.

7. ADABOOST DATA MINING ALGORITHM

- It is a boosting algorithm which constructs a classifier. As you probably remember, a classifier takes a bunch of data and attempts to predict or classify which class a new data element belongs to.

8. kNN DATA MINING ALGORITHM

- k-Nearest Neighbors, is a classification algorithm. A lazy type learning or instance-based learning. It is considered as a non-parametric method that is used for classification and regression.

9. NAIVE BAYES DATA MINING ALGORITHM

- It is a family of classification algorithms that share one common assumption: Every feature of the data being classified is independent of all other features given the class. They are considered to be highly scalable in machine learning and are known to be a family of simple probabilistic classifiers that are based on the application of Bayes' theorem with the help of strong independent assumptions between the features.

10. CART DATA MINING ALGORITHM

- This algorith stands for classification and regression trees. It is a decision tree learning technique that outputs either classification or regression trees. Like C4.5, CART is a classifier and both of them are decision tree learning techniques and features like ease of interpretation and explanation are also applied to CART as well.

IIV. LITERARY REVIEWS

Data warehouse as a modern technological concept, actually has the role to incorporate related data from vital functions of companies in the form that is appropriate for implementation of various analyses. In order to make data warehouse more useful it is necessary to choose adequate data mining algorithms. Those algorithms are described further in the paper for the purpose of describing the procedure of transforming the data into business information i.e. into discovered patterns that improve decision making process [1]. Data mining is a set of methods for data analysis, created with the aim to find out specific dependence, relations and rules related to data and making them out in the new, higher-level quality information [2]. As distinguished from the data warehouse, which has unique data approach, Data mining gives results that show relations and interdependence of data. Mentioned dependences are mostly based on various mathematical and statistic relations [3].

A Data warehouse is a repository of integrated information available for querying and analysis [4]. It is an in-advance approach to the integration of data from multiple, possibly very large, distributed, heterogeneous databases and other information sources [5]. It can be seen as a set of materialized views defined over the sources. When a query is posed, it is evaluated locally, using the materialized views, without accessing the original information sources. The information stored at the Data warehouse can be used by organizations for decision support. [6].Even though there has been a lot of work on various aspects of materialized views with respect to Data warehouses, there is little or no theoretical work at all on providing a method for configuring a Data warehouse. As a consequence the design of a

Data warehouse is haphazard and the quality of data is often dubious [7].

Data Mining and Knowledge Discovery in Databases (KDD) are rapidly evolving areas of research that are at the

Intersection of several disciplines, including statistics, databases, pattern recognition/AI, optimization, visualization, and high-performance and parallel computing [8]. It is concerned with the secondary analysis of large databases in order to find previously unsuspected relationships which are of interest or value to the database owners [9]. Data Mining is mainly concerned with methodologies for extracting data patterns from large data repositories. The extracted patterns are evaluated based on some interestingness measures that identify patterns representing knowledge. [10].

IIIV. CONCLUSION

This paper tackles the foundation on which Data Warehousing and Data Mining are intertwined and used to bring out ways to store and analyze large volume of data in order to bring out qualitative and quantitative usable information to benefit any industry or individual to help them come out with the best model of implementing them and further enhances their decision making tools or approach to a certain degree or a problem.

IX. REFERENCES

[1] Milija SUKNOVIĆ, Milutin ČUPIĆ, Milan MARTIĆ & Darko KRULJ...Faculty of Organizational Sciences & Trizon Group, Belgrade, Serbia and Montenegro, February 2005

[2] Berry, M.J.A., and Linoff, G., "Mastering data mining", The Art and Science of Customer Relationship Management, 1999.

[3] Bhavani, T., Data Mining: Technologies, Techniques, Tools and Trends, 1999.

[4] J.Widom, editor. Data Engineering, Special Issue on Materialized Views and Data Warehousing, volume 18(2). IEEE, 1995.

[5] J.Widom. Research problems in data warehousing.In Proc. CIKM, pages 25-30, Nov. 1995.

[6] Theodoratos Dimitri, Sellis Timos. Data Warehouse Configuration. Proceedings of the 23rd VLDB Conference Athens, Greece, 1997,pp 126

[7] Theodoratos Dimitri and Sellis Timos , Proceedings of the 23rd VLDB Conference Athens, Greece, 1997,pp 126

[8]Bradley P. S., Fayyad Usama M., Mangasarian O. L., Mathematical Programming for Data Mining: Formulations and Challenges. Journal on Computing 11, 1999, 217-238

[9] HAND David J., Data Mining: Statistics and More?. The American Statistician, May 1998 Vol. 52, No. 2

[10] M.Vazirgiannis, M.Halkidi, D.A. Keim , I.Ntoutsi, A. Pilrakis, S. Theodoridis, Y. Theodoridis , G. Tsatsarois, E. Vrachnos . “Recent Advances on Pattern Representation and Management” PANDA Technical Report Series , 2003 , pp. 4– 5, http://dke.cti.gr/panda/

X. ATTRIBUTIONS

DATA WAREHOUSE & DATA MINING CONCEPTS

https://www.zentut.com/data-mining/data-mining-processes/

https://addepto.com/implement-data-warehouse-business-intelligence/

https://www.guru99.com/data-warehouse-architecture.html#7

https://techburst.io/data-warehouse-architecture-an-overview-2b89287b6071

https://data-flair.training/blogs/data-mining-architecture/

https://hackerbits.com/data/top-10-data-mining-algorithms-in-plain-english/

https://www.techleer.com/articles/438-a-list-of-top-data-mining-algorithms/