IJMRT: Volume (5), Issue 5, 2023
International Journal of Multidisciplinary Research Transactions
(*A Peer Reviewed Journal*)
www.ijmrt.in

ISSN (Print):2663-2381
ISSN(Online):2663-4007

The Overview of Data Warehouse and Data Mining in an Organization

Pooja Kumbhar[1], Sayali Sarode[2], Prashant Tandale[3*], Dr. S D Bhoite[4]

[1,2] MCA Students, Bharati Vidyapeeth Deemed University Institute of Management, Kolhapur, India.
[3]Assistant Professor, Bharati Vidyapeeth Deemed University Institute of Management, Kolhapur, India.
[4]Chhatrapati Shahu Institute of Business Education & Research, Kolhapur, India.

[*]Corresponding author

DoI: https://doi.org/10.5281/zenodo.7784945

Abstract

Every organization has data. That data is collected from day to day transactions, inputs and outputs. Organizational data is stored by using different database techniques like Database management system, relational database management system, object oriented database management system or by using unstructured format. That organizational data is converted in summarized form by using different algorithms and rules depending upon the requirement . That summarized data is stored in data warehouse. Data mining is technique that uses different algorithms to convert data into useful information in the form of reports. These reports are basically designed on specific patterns with knowledge which are beneficial for decision making for organization.

Keywords: Data, Purification, Data Warehouse, Data Mining'.

1. Introduction

A data in the organization is collected from day to day transactions, inputs and outputs. Organizational data is stored by using different database techniques like Database management system, relational database management system, object oriented database management system or by using unstructured format. The organizational data is converted in summarized form by using different algorithms and rules depending upon period. The summarized data is stored in

data warehouse. Data mining is technique that uses different algorithms to convert data in the warehouse into meaningfull reports. These reports are basically designed on specific patterns with knowledge which are beneficial for decision making for organization.

2. Journey of Information of Organization from Data to Knowledge

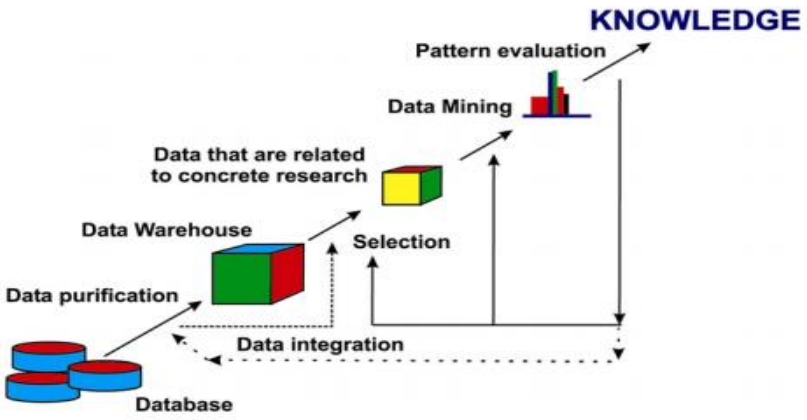

Figure.1. Journey of Information of Organization from Data to Knowledge

Database A database is an organized collection of data.[3] Databases are divided into following subtypes databases.

End user: These databases are shared by users and contain information meant for use by the end-users like managers at different levels.

Operational: These databases store data relating to the operations of the enterprise. Generally, such databases are database of day to day transactions of organization.

Centralized: These databases store the entire information and application programs at a central computing facility. The users at different locations access the central database to make processing.

Distributed: These databases have contributions from the common databases as well as the data captured from the local operations. The data remains distributed at various sites in the organization.

Personal: The personal databases are maintained, generally, on Personal computers. They contain information that is meant for use only among a limited number of users, generally working in the same department.

Commercial database: The database to which access is provided to users as a commercial venture is called a commercial database. These databases contain information that external users would require but by themselves would not be able to afford maintaining such huge databases. These databases are subjected specific and access to these databases is sold as a paid service to its user.[4]

3. Data Purification

Data purification is also called as Data scrubbing, or data cleansing. It is the process of revising or removing data in a database that is incorrect, incomplete, improperly formatted, or duplicated. Different data scrubbing tool to systematically examine data for flaws by using rules, algorithms, and look-up tables. Using a data scrubbing tool saves a significant time of a database administrator and can be less costly than fixing errors manually. Generally data purification is also used to convert transactional database into summarized form. [5]

1. **Data warehouse :** Data Warehouse is different that operational environments. It contains integrated data. Data integration is done on transactional data by using different algorithms and rules. It contains historical data over a long period of time but in the form of summery. Data is a snapshot data captured at a given point in time. Data is subject-oriented task specific. “A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. Subject-Oriented means a data warehouse can be used to analyze a particular subject area. Integrated means a data warehouse integrates data from multiple data sources.

Time-Variant means historical data is kept in a data warehouse. Non-volatile means once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered- Bill Inmon says , [1]

2. **Data Mining :**Data Mining is the extraction or "Mining" of knowledge from a large amount of data or data warehouse. For the extraction, data mining combines artificial intelligence, statistical analysis and database management systems to attempt to pull knowledge form stored data. Data mining is the process of applying intelligent methods to extract data patterns. This is done using the front-end tools. [2] There are number of data mining software available. Some of those are as follows:

I)Weka Data Mining : Weka is collections of machine learning algorithms for data mining tasks.T he algorithms can applied directly on dataset or called from java code. Weka features includes machine learning, data mining, preprocessing, classification, regression, clustering, association rule, attribute selection etc.

II) Apache Mahout : The Apache Mahout™ project's goal is to build an environment for quickly creating scalable preformat machine learning applications. Apache Mahout introduces a new math environment call Samsara, for its theme of universal renewal. It reflects a fundamental rethinking of how scalable machine learning algorithms are built and customized. Mahout-Samsara is used to help people to create their own math while providing some off-the-shelf algorithm implementations. At its core are general linear algebra and statistical operations along with the data structures to support them. You can use is as a library or customize it in Scala with Mahout-specific extensions that look something like R. Mahout-Samsara comes with an interactive shell that runs distributed operations on a Spark cluster. This makes prototyping or task submission much easier and

allows users to customize algorithms with a whole new degree of freedom. Mahout Algorithms include many new implementations built for speed on Mahout-Samsara. They run on Spark 1.3+ and some on H2O, which means as much as a 10x speed increase. You'll find robust matrix decomposition algorithms as well as a Naive Bayes classifier and collaborative filtering. The new spark-item similarity enables the next generation of Co-occurrence recommenders that can use entire user click streams and context in making recommendations.[7]

III) RapidMiner : RapidMiner is one of the most widely used analytics platforms in the world, with over 250,000 users. Organizations of all sizes use RapidMiner, and its range of application is very broad. The fact that many predictive models can be built without resorting to program code is one reason for its popularity, the other being very reasonable pricing. This is a sophisticated offering with over 1500 drag-and-drop operators. Novice users can quickly get up to speed with RapidMiner's 'Wisdom of Crowds' online repository of best practices, and this is quite unique among analytics platforms. Big data is well accommodated through its Radoop platform – insulating users from the complexities and volatility of big data technologies. The importance of this facility cannot be over-emphasized, as organizations struggle to keep up with rapid developments in big data technologies. Organizations of all sizes looking for a cost effective, powerful analytics platform, will find that RapidMiner is a speedy, scalable environment in which to develop and deploy predictive models.[8]

IV) Rattle : Rattle the R Analytical Tool to Learn Easily - "Rattle is a tab-oriented user interface that is similar to Microsoft Office's ribbon interface. It is a popular GUI for data mining using R. It presents statistical and visual summaries of data, transforms data that can

be readily modelled, builds both unsupervised and supervised models from the data, presents the performance of models graphically and scores new datasets. [9]

V) GNU Octave : GNU Octave is a high-level interpreted language, primarily intended for numerical computations. It provides capabilities for the numerical solution of linear and nonlinear problems and for performing other numerical experiments. It also provides extensive graphics capabilities for data visualization and manipulation. Octave is normally used through its interactive command line interface . It can also be used to write non-interactive programs. The Octave language is quite similar to Matlab, so that most programs are easily portable.

4. Conclusion

In modern era, organizational data is converted into knowledge by using different steps. First collected organizational data from day to day activates and functions are purified by using different techniques. These purified data is stored in data warehouse. Depending upon concentered research data from data warehouse is used for mining. Mining creates patterns and reports which are used for decision making process of organization.

REFERENCES

[1]. http://www.1keydata.com/datawarehousing/data-warehouse-definition.html
[2]. Data Warehousing, Data Mining, OLAP and OLTP Technologies Are Essential Elements to Support Decision-Making Process in Industries S.Saagari, P.Devi Anusha, Ch.Lakshmi Priyanka, V.S.S.N.Sailaja, International Journal of Innovative Technology and Exploring Engineering (IJITEE).
[3]. "Database - Definition of database by Merriam-Webster". merriam-webster.com.
[4]. Database: Six Important Types of Databases | Business Management, by Saritha Pujari.
[5]. http://searchdatamanagement.techtarget.com/definition/data-scrubbing.
[6]. Data Mining: A prediction Technique for the workers in the PR Department of Orissa (Block and Panchayat) Neelamadhab Padhy, and Rasmita Panigrahi, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.5, October 2012.
[7]. http://mahout.apache.org/

[8]. RapidMiner Review, post on site http://www.butleranalytics.com.
[9]. Rattle: A Graphical User Interface for Data Mining using R, posted on http://rattle.togaware.com/

