v. 3, n. 40, 2023

Journal of
**Engineering
Research**

BUILDING A DATA WEBHOUSE TO SUPPORT CLICKSTREAM DATA ANALYSIS

Wendel Góes Pedrozo
Universidade Tecnológica Federal do Paraná – Coordenação de Engenharia de Computação
Apucarana, PR, Brasil
Pontifícia Universidade Católica do Paraná - Programa de Pós-Graduação em Informática
Curitiba, PR, Brasil
https://orcid.org/0000-0002-5343-3198

Leonardo Henrique Pereira
Pontifícia Universidade Católica do Paraná - Programa de Pós-Graduação em Informática
https://orcid.org/0000-0002-7786-1816
Curitiba, PR, Brasil

DOI 10.22533/at.ed.3173402323116

Abstract: This work intends to show the main points for building a Data Webhouse – a tool to support clickstream analysis. Currently, people consume more and more Web products and services. Information technology now has the mission of making products and services available through a browser-type interface. In this market where the unit of value is the relationship with the customer, marketing focused only on products no longer meets the needs of the current economy. Therefore, it is necessary to increasingly intensify customer relationship management activities on the Web. This work shows a case study that consists of outlining the steps for building a Data Webhouse in an e-commerce environment, to assist in data analysis clickstream (sequence of clicks). Finally, it is showing the first observations about the implantation of a Data Webhouse at the clickstream's analysis and its importance at the process of decision taking.

Keywords: Data Warehouse, Data Webhouse, Data Mining, Clickstream.

INTRODUCTION

The impact of the Web in recent years has definitely changed the mission of information technology (IT), which now aims to provide content, information and transaction processing using a browser-type interface.

Observing the history, it appears that some revolutions related to communication occurred, where the cost of this communication was reduced. As an example, we can mention: the newspaper, the telephone, the radio and the TV. In all these examples, when the cost reduction was significant, the volume of communication grew quickly and people became trained in using the new medium.

Starting in the 90s, the Web revolution, which is a descendant of past revolutions, has as its main difference the speed at which it is spreading. In the space of a few years, a significant fraction of the world has changed the way they communicate, the way they conduct business and the way they use information. The WEB has become much more than a technology for connecting distributed processing devices, it has become an infrastructure for providing products and services to anyone, anywhere on the planet.

Receiving more than 100 million hits per day, the most popular websites have an excellent opportunity to collect valuable data about their customers. User interaction with websites through clicks provides an immense source of behavioral data about them. Although this data, called clickstream - click sequence, in many cases is in a raw state and does not have an adequate appearance, it provides details never imagined about each gesture made by each person using the WEB. The click sequence is a chronological series of microscopic actions that can be grouped into sessions, representing the trajectory of actions that led to a purchase or other behavior in which we are interested and that can be analyzed and understood (Kimball, 2000).

It is observed that unfortunately many organizations are unable to take full advantage of this enormous amount of information, simply because they do not have the necessary resources to use it effectively. The evolution of IT has made it possible to generate resources that meet the needs of these organizations, that is, to understand the sequence of clicks, capture it and take it to a database environment to be analyzed effectively

Before the WEB revolution, IT companies learned how to publish company assets in the form of data for internal analysts and management. This publication is the central task of the Data Warehouse.

From another perspective, the Data Warehouse, in its mission to store vital information for the decision process, has become useful for storing and monitoring customer interaction with the company's website. In order to be paired with this much greater responsibility, the Data Warehouse must be adjusted. Its nature needs to be somewhat different from what it has been in the past. This rebirth of the Data Warehouse is called Data Webhouse.

CLICKSTREAM AS A DATA SOURCE

When customers enter a store in a shopping mall, they are immediately approached by a salesperson. During the entire time they remain there, they are being closely monitored, so they can be helped promptly at any time.

On the WEB, the websites initially face a barrier in communicating with their customers, because when entering the website, the customer is alone, interacting with a computer. The initiative to search for and identify the products/services of interest is up to him. To break down this barrier, organizations are investing in ways to provoke interaction between the website and the customer, whether through opening communication channels, such as e-mails and chats, or through customizing the website in response to past interactions, as a way of showing that the organization is recognizing its customer (Kimball, 2000).

By analyzing the Web scenario, it is clear that it represents a source of data relating to the behavior of individuals when interacting with websites through browsers. This data is literally a record of all gestures and interactions made by any visitor to a website.

APPLYING CLICKSTREAM DATA

If the organization has information about each click of its customers, that is, the path they follow within its website, it will probably be able to answer these questions (Kimball, 2000):

- What is the most visited location on the website?
- Which location has the highest number of sales on the website?
- Which page is seen as the “Final Session” where users typically leave the site?
- Where does the new user click on their first visits according to their profile?
- What is the browsing profile of an existing customer? Is the customer profitable?
- How to induce the customer to register on the website, so they can get to know you better?
- Before registering, do customers purchase any products or services?

If the company is unable to measure the degree of satisfaction of its customers and/or the relevance of the information available on the website, it will lack the support to serve its customers well. By using clickstream data, the organization will be able to better respond to its customers’ needs based on knowledge of their behavior.

DECISION MAKING PROCESS

Data by itself does not represent anything, it is necessary to work with this data to obtain meaningful information. This can be done through a coherent process of organization and identification of patterns, from there it is possible to make correlations, comparisons, cause and effect analyzes based on this information. There are different types of decisions that are facilitated considering the clickstream, as follows:

IDENTIFICATION AND COMMUNICATION WITH CUSTOMERS

The types of decisions possible when identifying the customer are:

- Customization of marketing activities:

it is possible according to the customer profile, considering, for example, frequency, time between accesses, tastes and income, to carry out personalized marketing campaigns for each customer or groups of customers;

- Choose, through the desired group of customers, which Websites the company should continue paying for for links to your Website (banners, for example).

The types of possible decisions in relation to communication with customers are:

- What type of browsers and operating system customers use, that is, what compatibility the Website must have;
- Detection of profitable or unprofitable promotions, not only due to the increase in sales, but due to the general impact on the business;
- Customization according to changes in clients' lives, for example, marriage, children, moving to another city and a marked change in health status;
- Website efficiency, such as adequacy of the interface, arrangement of information, choice of subjects, number of cascading screens for transactions and content.

DECISION SUPPORT SYSTEMS

Data processing and storage capacity has shown significant growth in recent years. As a result of this evolution, it appears that scientific or commercial applications have generated gigabytes and even terabytes of data in a few hours, and this volume far exceeds the capacity of researchers and market analysts to analyze it. To overcome these difficulties, techniques and tools are adopted that enable analyzes to be carried out on large amounts of data, supporting the decision-making process (Campos, 2002).

DATA WAREHOUSE

The implementation of a Data Warehouse is one of the first steps towards enabling analysis of large amounts of data in decision support systems.

The Data Warehouse environment consists of a Database, created to structurally store data from different sources in one or more repositories. This data can be from local or non-local systems, spreadsheets, text files, etc.

The main characteristics of a Data Warehouse are well defined and the most renowned authors in the area agree on them. According to Inmon (1997), the Data Warehouse is characterized by being subject-based, integrated, non-volatile and variable in relation to time. More information can be found in Inmon (1997).

DATA MART

In a Data Warehouse, the physical area can be organized not only into a single Data Warehouse, but into several, called Data Marts, logical subsets of the Data Warehouse, generally treated as a sectoral Data Warehouse (Kimball, 1996). Data Marts are often seen as an alternative to using a Data Warehouse, as they take less time to develop and implement.

DATA WEBHOUSE

Kimball (2000) shows two proposals for building the Data Webhouse. In the first of them, the objective is to bring the WEB into the Data Warehouse through the study of user behavior on the WEB. In the second proposal, the objective is to bring the Data Warehouse to the WEB, through the availability of Data Warehouse data on the WEB. The focus of this work is on the first proposal, which is to bring the WEB to the Data Warehouse.

The Data Webhouse has a central and crucial role in the operations of a WEB-enabled business, and to fulfill this potential, the Data Webhouse (Kimball, 2000):

- Stores and publishes clickstream data and other WEB behavioral data, which guides an understanding of customer behavior;

- It is an adaptable and flexible source of information. As new business questions arise and new data sources become available, Data Webhouse elegantly responds to them;

- It is extendable to new WEB media, including graphic images, audio and video;

- It is a secure medium that publishes data to customers, business partners and employees appropriately, but at the same time protects the company's data assets against unintended use;

- It is the basis for decisions regarding conversions to the WEB. The Data Webhouse must allow users to make decisions about the WEB while using the WEB.

THE WEB IN THE DATA WAREHOUSE

For the WEB to be brought to the Data Warehouse, clickstream data is used to explore WEB access information. For the operation of this technique, data is obtained through logs of all accesses made and maintained by the Web Server as shown in figure 1. It is also obtained through click sequences from referral partners (referring) or ISPs (Information Service Providers), or through WEB statistics services, which are used to place control over WEB pages that alert when a user accesses the page.

This collected data is important to provide rich information about access dynamics, helping to improve the quality of interactions with users, which can lead to greater loyalty and, consequently, increased revenue (Campos, 2002).

However, this data cannot be used indiscriminately, firstly because it does not contain enough information, and secondly because a raw click sequence is not a useful description of behavior and can lead to hasty conclusions.

Cleaning and transforming this data are fundamental actions and require knowledge of the structure of the website and the application. In fact, at this point, the construction of the Data Webhouse is very similar to the construction of the Data Warehouse, since from the moment the data was obtained, it will be loaded into the Data Webhouse only after its transformation, first going through the Extraction steps, Data Transformation and Loading (ETL).

According to Kimball (2000), the ETC steps in the Data Webhouse, which is called "click sequence post-processor", has the function of extracting information, as well as being responsible for the following tasks:

- Filtering of unnecessary records: where associated data are merged and records that will not be passed to the Data Webhouse are excluded, reducing the volume of transactions present as much as possible, without compromising the integrity and completeness of the data necessary to support the granularity of the Data Webhouse project;

- Session identification: where associated clickstream records are marked with a unique session identifier. It is also checked whether event times are logically consistent with each other and between the records that describe the session;

- User identification: where the user is matched with an existing user identifier, if possible. Otherwise, a unique anonymous user identifier is assigned if the identity is unknown;

- Host identification: where the IP

(Internet Protocol) addresses of clients and connection sources are converted (to the desired granularity);

- Consolidation of data into a uniform format: where clickstream data is placed in a defined format acceptable to the Data Webhouse loading software.

As soon as the data is uploaded to the Data Webhouse, it must be analyzed. To this end, we have OLAP (On-Line Analytic Processing) Tools, as they provide views of the data from different perspectives and different conceptual levels, answering questions such as: Which components or services are the most and least used? What is the distribution of network traffic over time? What are the differences in access between users from different geographic regions? etc. (Campos, 2002).

When it is necessary to discover trends in the database and relationships between objects, such as customers and products, the Data Mining tool is often used. According to TAN (2009), Data Mining is the discovery of interesting knowledge, but hidden in large databases. Data Mining uses statistical and machine learning techniques to build models capable of predicting customer behavior. Nowadays, technology can automate the mining process, integrate it with Data Warehouses and present them in a relevant way to Business Intelligence users. Data Mining seeks to discover patterns and relationships in a database, so that the organization can have a better basis for making decisions (Reis, 2017). It is used to answer questions such as: Under what circumstances are components or services used? What are the typical sequences of events? Are there patterns of behavior among all users? Does user behavior change over time and how? (Campos, 2002).

PROPOSED ARCHITECTURE

To help understand Data Webhouse, a case study was carried out and will be discussed below. It is worth noting that this case study was implemented following the steps described in this chapter, and the results will be commented on in the session entitled conclusion.

DIMENSIONAL MODELING

A standard model for a Data Warehouse or Data Webhouse project is dimensional. The dimensional model is the preferred format for presenting data in the Data Warehouse. Dimensional modeling is an alternative to traditional entity-relationship (E/R) modeling.

All dimensional models are built around the concept of measured facts. A fact could be the sale of a product, the price of a stock at a particular point in time, or the change in your salary as a result of your promotion.

We collect measured facts from our computer systems and place them in tables called fact tables. For example, if the measured event is the sale of a product, then the following context for this measurement may be known:

- The date of this sales transaction;
- The Customer;
- The employee;
- The product sold;
- The price;
- The payment method (cash, installments, or a special type of financing.

Each of the items in the context list are called dimensions. Six dimensions are identified for this example. When you think about these dimensions, it becomes clear that they are not independent numerical facts, like the price of the product. Rather, such dimensions are rich textual descriptions of something that exists at the time the fact table record is defined. Even the date is a rich textual description. It's a day of the week, the name of a month, a holiday (yes/no), a season, for example.

Given the richness and openness of dimensional descriptions it is clear that one does not want to place all dimension descriptions in each record that represents a measured fact. Then, you place all textual descriptions in separate dimension tables and a foreign key in the actual record of the fact. Each foreign key connects to its corresponding primary key in a dimension table.

DEFINING THE ORIGIN OF THE FACTS TO BE MEASURED

The data to be stored while browsing the Website comes from two sources. The first source refers to the clickstream data contained in the WEB protocols and also stored in the WEB Server logs. The second source refers to additional information about the user's activities, from the moment he enters and starts a session on the Website, and which is captured by applications such as order entry, search on the Website, order viewing, etc. others.

The operations that must be recognized are:

- Entry points;
- Usage time per page;
- Query expressions;
- Navigation within the Website;
- Exit point.

DEFINING THE DIMENSIONS

An important activity to model the clickstream Data Mart is to think and identify possible dimensions that are relevant and at an appropriate granularity.

TIME DIMENSION

Like practically every Data Warehouse, the clickstream Data Mart has a time dimension. The time dimension, in its greatest granularity, normally has a record for each day of the calendar, that is, it is the date dimension. If it is necessary to store the time of the event, there is also the time dimension, that is, time is divided into two dimensions.

CUSTOMER DIMENSION

In a Data Webhouse, the data to make up the customer dimension is not easy to obtain, due to issues of privacy and anonymity when using the WEB. For this reason, the customer dimension can actually be the user, visitor or machine dimension, depending on the level of detail of the data that will be loaded. The first group are fields that are always known through navigation, they are:

- Customer key: substitute key;
- Customer type: describes the known degree of identification, for example, unknown, regular, not applicable, fixed IP, variable IP, fixed cookie, variable cookie, identified customer;
- Access provider address: multivalued, client connects from home, work, etc.;
- Cookie Identification: identification generated for the customer, the same for home, work, etc.

The second group of fields assumes that some name and location information is known and customers can be identified when browsing, that is, a customer key can be generated and is not simply cookie identification. Some fields are: full or partial name, nationality, gender, city, state and country.

The third group is the highest level of detail, which includes: type of customer (residential or commercial), occupation, company, department, function, telephone, fax, e-mail, age, income, marital status, interests, etc.

PAGE DIMENSION

The page dimension represents the context of WEB pages. The fields normally existing in this dimension are the following:

- Page key: substitute key;
- Page origin: static, dynamic;
- Function of the page: homepage, search, product description, etc.;
- Page model: sparse, dense, etc;
- Type of item: product code, book ISBN, etc.;
- Type of graphic, animation or sound: GIF, JPG, etc.;
- Page file: HTML file name, CGI details, ASP, etc.

EVENT DIMENSION

The event dimension describes events, that is, particular events on certain pages at certain moments in time. Some interesting events are opening, link selection and data entry.

The fields in this dimension are:

- Event key: substitute key;
- Type of event: opening, recharge;
- Event content: form data, etc.

SESSION DIMENSION

The session dimension indicates one or more diagnostic levels of the user session as a whole. A local context might be selecting product, while a general context might be purchasing. This state indicates the progression of the activity the client is doing.

Furthermore, you can momentarily characterize the client by session levels, for example, new client, filling in identification data, registered client, trusted client, client thinking about giving up, standard client, etc.

This dimension can be very useful for carrying out certain types of analysis, for example, number of views of product details before purchase, number of purchase cancellations after viewing payment conditions, number of unfinished orders and the stopping point .

Typically this dimension has the following fields:

- Session key: substitute key;
- Session type: classified, unclassified, corrupted, etc;
- Local context: context relative to the current page;
- General context: general context relating to the transaction that was made;
- Sequence of actions: summary description of the operations carried out in the session;
- Success status: success or failure of the session;
- Client state: default, trusted, etc.

REFERENCE DIMENSION

The reference dimension describes how the customer arrived at the current page. This information comes from the Web Server log: the URL of the previous page and some possible additional information. The fields in this dimension are as follows:

- Reference key: substitute key;
- Type of reference: Intranet, Internet, search engine, corrupt or not applicable;
- URL: URL that references the current page;
- Site: site que referencia a página atual; Website: website that references the current page;
- Domain: domain that references the current page;

- Type of search: simple or advanced;
- Specification: expression, useful if searching for simple text;
- Target: where the search found the expression, whether meta-tag, header or title, for example.

PRODUCT (OR SERVICE) DIMENSION

The product dimension describes the product or service that is the subject of the page or target of the event. Typically, this is a dimension that contains a large set of descriptor attributes. The fields in this dimension are as follows:

- Product key: replacement key;
- Type of product;
- Manufacturer;
- Brand;
- Category;
- Price.

BUILDING THE FACT TABLES

After studying the relevant dimensions, fact tables must be constructed. An initial decision is regarding the fact to be measured and its granularity. In the case of clickstream analysis, there are two obvious granularities, the first represented by the Fact Table for analyzing complete sessions, and the second represented by the Fact Table for analyzing page usage.

FACT TABLES FOR FULL SESSION ANALYSIS

In the complete session fact table, as shown in figure 2, each complete session will have a record in this table. Information about the session is known through the Web Server log and the facts to be measured will be: session time in seconds, number of pages visited, number of orders placed, number of units ordered and the order amount in the desired currency

This fact table is appropriate for some types of analysis. Some examples are:

- Directing marketing activities according to the classification and grouping of customer profiles;
- Banner return evaluation;
- Why does the customer use our Website?
- How did the customer reach him?
- How much time do customers spend visiting our Website?
- How many pages do they view?
- What is our sales volume via the WEB?

FACT TABLE FOR PAGE USAGE ANALYSIS

The fact table for page usage analysis, as shown in figure 3, has the following dimensions: date, time, customer, page, event, session, product, reference and we add a degenerate dimension, session identification. In addition to the analyzes already carried out previously, other questions can be answered: how to identify the customer's possible intention to withdraw; how to identify whether advertising on the WEB is working; how to identify whether a thank you message or offer is working; how to identify whether promotions are profitable; how to increase the efficiency of the Website; decide which services and products to offer via the WEB.

CLICKSTREAM DATA MART

With the Fact tables described previously and the dimension tables, the clickstream Data Mart is built. This data model looks a bit like a star because the fact table is in the center and all the dimension tables are organized around it. Database engineers call it a star join. Star

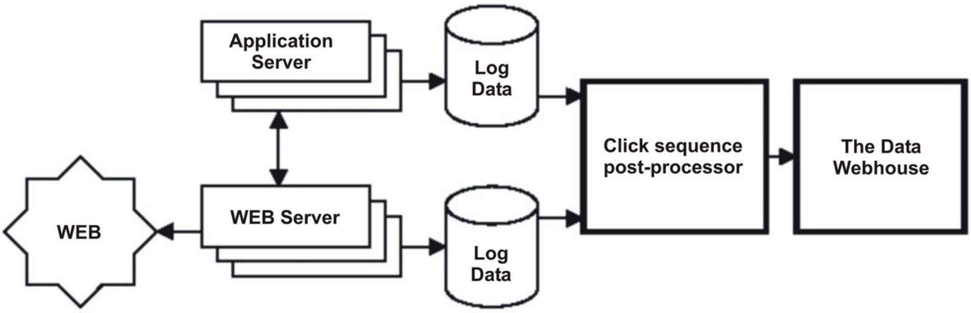

Figure 1 - Data Webhouse creation mechanism via the WEB

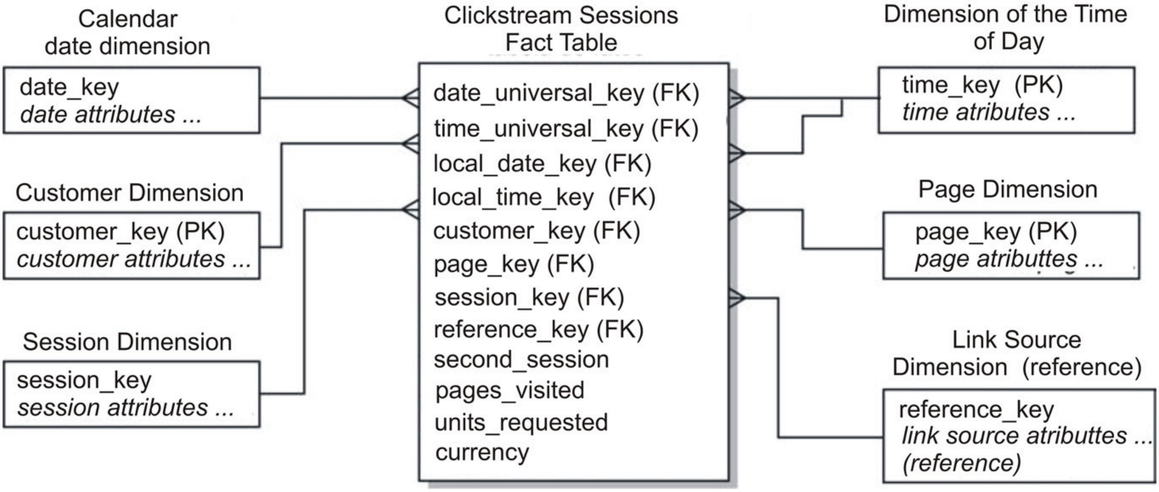

Figure 2 – Full Sessions Fact Table

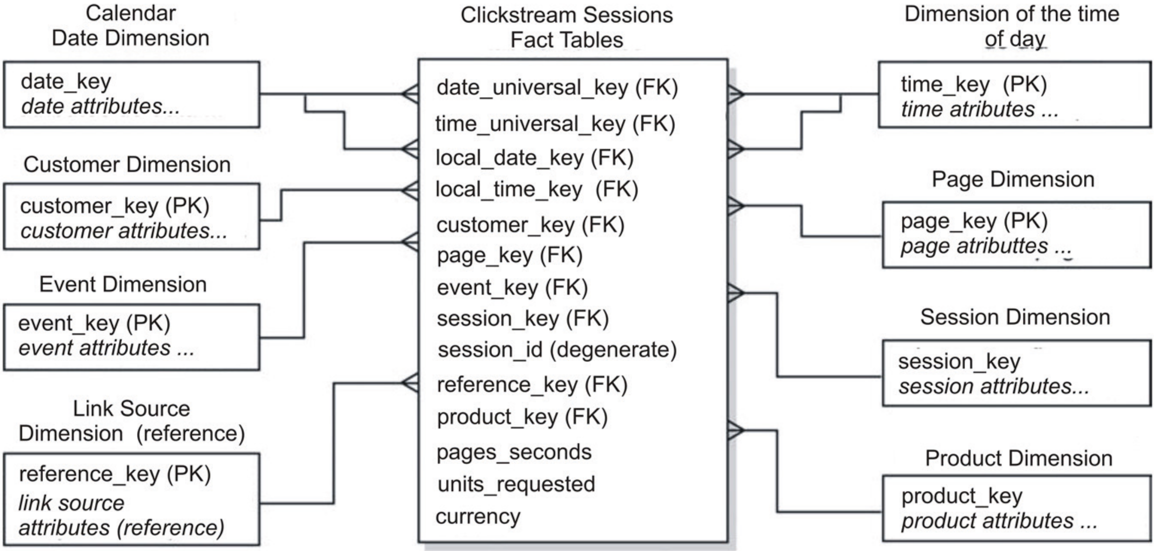

Figure 3 – Fact Table for Page Usage Analysis

schemas are linked through the common use of dimensions.

The most important fact table in this Data Mart is the full sessions fact table. This is due to the fact that it already helps with various analyzes without compromising performance and without needing a lot of storage space.

CONCLUSIONS

The Data Webhouse developed in this work was provisionally implemented in a given business. In a first data collection, it was noticed that the Data Webhouse to support clickstream data analysis brought some advantages to marketing professionals through access to information about customer behavior in e-commerce, enabling quick and targeted action to individualized customer profiles. It is expected that with the more effective use of Data Webhouse, other analyzes can be carried out, such as: predicting the increase or reduction in sales in the WEB environment of a given product due to the variation in its prices, determining the optimal price in order to maximize the virtual company's profit, determine the influence of a marketing activity on sales, know the influence of competitors' price variations on its sales and know the influence of its price variations on competitors' sales. It will also be possible to obtain knowledge of competitors' marketing activities and their influence on sales, segment the market depending on the behavior of its customers, in order to enable specific marketing campaigns for each segment and extract knowledge of competitors' strategy.

Considering that the Data Webhouse is modeled and developed appropriately, it can become the central and cohesive element of the modern, customer-focused company, providing essential information to managers and those responsible for strategic decisions.

REFERENCES

CAMPOS, M.L. Data Warehouse. In: Simpósio Brasileiro de Banco de Dados, 2002, Gramado. **Minicurso de Data Warehouse**. Porto Alegre: Instituto de Informática da UFRGS, 2002.

INMON, W.H. **Como Construir o Data Warehouse**. Rio de Janeiro: Editora Campus, 1997. KIMBALL, R; The Data Warehouse Toolkit. John Wiley & Sons, 1996.

KHALAF, D, K.; HAMAD, M. M. **Data Webhouse for Monitoring the Use of Enterprise Information System,** *2019 12th International Conference on Developments in eSystems Engineering (DeSE)*, Kazan, Russia, 2019, pp. 278-283, doi: 10.1109/DeSE.2019.00059.

KIMBALL, R.; MERZ, R. **Data Webhouse: construindo o Data Warehouse para a WEB**. Rio de Janeiro: Editora Campus, 2000.

REIS, W. A. D. **Um Método de Identificação de Emoções Baseado na Mineração de Padrões Sequenciais**. 2017. 109 p. Dissertação de Mestrado - Pontifícia Universidade Católica do Paraná, Curitiba, 2017.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining**. Rio de janeiro: Editora Ciência Moderna Ltda., 2009.