Procesamiento Analítico con Minería de Datos

Analytical Processing with Data Mining

Angelino Feliciano Morales
Universidad Autónoma de Guerrero, México
afmorales@uagro.mx

René Edmundo Cuevas Valencia
Universidad Autónoma de Guerrero, México
reneecuevas@uagro.mx

José Mario Martínez Castro
Instituto Tecnológico de Chilpancingo, México
jmariomtz@yahoo.com

Resumen

Este trabajo describe la utilidad e importancia de la herramienta OLAP en Business Intelligence con el fin de recomendarla a los administradores de empresas para su toma de decisiones. La tecnología OLAP permite el rápido acceso a datos mediante data warehouse, agilizando el analisis de la información. Los cubos proveen de un rápido mecanismo de búsqueda de datos y de un tiempo de respuesta uniforme, independientemente de la cantidad de datos o de la complejidad del procedimiento de búsqueda. Tomando en cuenta su funcionamiento y estructura, el sistema OLAP se clasifica en tres categorías: ROLAP, MOLAP y HOLAP. Actualmente el sistema OLAP que más se utiliza es el denominado ROLAP.

Palabras clave: OLAP, Multidimensional, Cubo OLAP.

Abstract

This work describes the utility and importance of the OLAP Business Intelligence tool in order to recommend it to the administrators of companies for their decision making. OLAP technology enables faster access to data using a data warehouse, speeding up the information analysis. Cubes provide a fast mechanism of search data and a uniform response time, regardless of the amount of data or the complexity of the search procedure. Taking into account its operation and structure, the OLAP system is classified into three categories: ROLAP, MOLAP and HOLAP. Currently the OLAP system most widely used is the so-called ROLAP.

Key words: OLAP, Multidimensional, Cube OLAP.

Fecha recepción: Agosto 2015 **Fecha aceptación:** Diciembre 2015

Introduction

This document highlights the importance of using this tool in databases to facilitate the consultation of large amounts of data, especially in the business world.

The achievement of competitiveness in the production is a primary task in Business Intelligence. For this it is essential to develop enterprise system in an innovative mentality; with a science and innovation strategy you can achieve an adequate level of technology management and maintaining the standards of quality and efficiency.

In the field of solutions, OLAP applications are one of the tools most commonly used by companies, since they have been created according to multidimensional databases that allow you to process large volumes of information in well-defined fields, and with immediate access to the data for your future reference.

OLAP tools provide companies a reliable system for processing data which then they will be used to carry out analyses and reports that allow to improve the productive operations, make smart decisions and optimize the competitiveness. Also, give support to the technologies of Data Warehouse. In general, these OLAP systems must:

- Support complex analysis requirements.
- Analyze data from different perspectives.
- Support complex analysis.

The main characteristic of OLAP tools is that they are environments designed for the execution of multidimensional analysis of corporate data from any supported user. They also provide possibilities of navigation, selecting information, allowing the analysis of segmented data allowing to reduce the set of data that have been reported. This type of selection is reflected in the visualization of multidimensional structure, through some selection fields that allow you to choose the level of aggregation (hierarchy) dimension, or the choice of a fact in particular.

The information is managed and processed in large blocks organization, such as the geographical structure or the academic, called dimensions. These dimensions of business are structured at the same time in different levels of detail.

Currently its application has been extended to all business areas and other types of organizations that analyze massive volumes of data —including SMEs, academia, Government and other public and private institutions—, requiring ever more than one dynamic, powerful analysis and online to make appropriate decisions, passing on the demand for this type of software.

Database

A database is a collection of data organized and structured according to a certain model of information reflecting not only the data itself, but also the relationships between them. A database is designed with a specific purpose and is organized with a coherent logic. The data may be shared by several users and applications, however, they must retain their integrity and

security outside the interactions of both. The definition and description of data must be unique to minimize redundancy and maximize independence in their use.

In a database, entities and attributes of the real world become records and fields. These entities can be both material objects such as books or pictures, but also people and even abstract concepts and ideas. Entities have attributes and maintain relationships between them. The databases can be classified according to the characteristics (Figure 1).

Figure 1. Classification Database

A database gives users access to information, you can view, enter or update, in accordance with the input rights have been granted.

A local database can be used by a single user on a computer or distribute the information to remote computers and access it through a network.

The main advantage of using databases is that multiple users can access them simultaneously.

Data management requires a process to become useful information; some of the tools you need this type of procedure presented below (Figure 2).

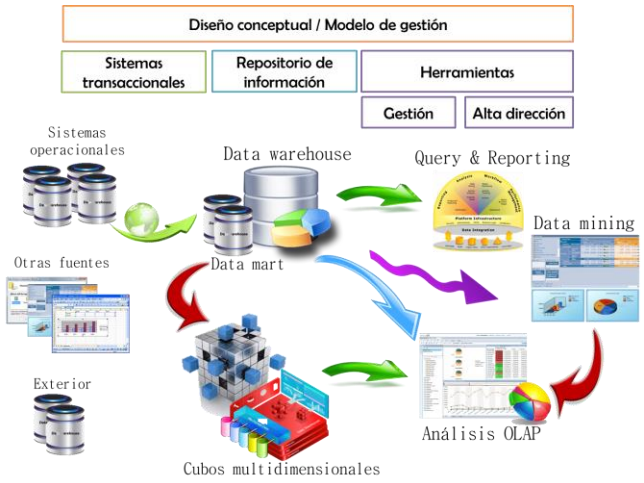

Figure 2. Process data

Data Warehouse

It is a combination of concepts and technologies to meet the requirements of an organization or company, in terms of improving management efficiency and ease of access.

It is a relational database designed for query and analysis rather than transaction processing. It usually contains historical data derived from a transaction, but it can include data from other sources. Separate analysis workload of transactions and enables an organization to consolidate data from several sources.

Data warehouses and OLTP systems (On-Line Transactional Processing) have very different needs. The following are some examples of the differences between typical data warehouses and OLTP systems:

- Workload

  Beforehand, the user can not know the workload of the data warehouse, which must be optimized to promote good performance of variety of possible query operations.

OLTP systems support predefined operations. Its applications can be tuned or designed specifically to support these operations.

• Modifying Data

A data warehouse is updated on a regular basis by the ETL process (run by nightly or weekly) using modification techniques bulk information. End users of a data warehouse are not updated directly from the warehouse.

In OLTP systems, end users routinely issue instructions for modifying individual information to the database. The OLTP database is always up to date and reflects the current state of each business transaction.

• Outline design

Data warehouses often use schemes that are not standardized or partially standardized (as a star schema) to optimize query performance.

OLTP systems often use fully normalized schemas to optimize update / insert / delete performance, and to ensure data consistency.

• Typical operations

A typical data storage query scans thousands or millions of rows. For example, "locate the total sales for all customers last month."

A typical OLTP operation accesses only a handful of records. For example, "retrieve the current order for this customer."

• Historical data

Data warehouses usually store many months or years of information. This is to support historical analysis.

OLTP systems usually store data from a few weeks or months. The OLTP system stores only historical data needed to successfully meet the requirements of the current transaction.

Besides a relational database, a data warehouse includes a solution extraction, transport, transformation and loading (ETL), an online analytical processing (OLAP) engine, analysis tools, clients and other applications that manage the process information collection and delivery to business users (figure 3).

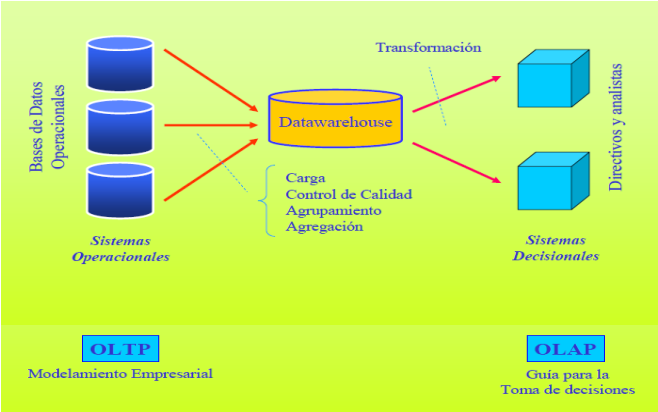

Figure 3. Procedure DATA WAREHOUSE

Datamart

It is a departmental database, specialized in storing information of a specific business. It is characterized by having an optimum data structure for analyzing information from various perspectives that affect the processes of the department. A Datamart can be fed data from a data warehouse or integrating itself a compendium of different sources of information.

OLAP datamart

They are based on popular OLAP cubes, which are built according to the requirements of each area or department, dimensions and the necessary indicators for each relational cube. The creation mode, operation and maintenance of OLAP cubes is heterogeneous, depending on the final tool used.

Datamart OLTP

The OLTP databases can be based on a simple extract data warehouse, however, it is common improvements in performance (and filtered aggregates are usually the most common operations) taking advantage of the particular characteristics of each area of the company. The common structures are the report tables, which are to be reduced fact-tables (which add appropriate dimensions), and materialized views, which are built with the same structure as the above, even with the aim of exploiting rewriting queries (although it is only possible in some advanced DBMS such as Oracle).

The Datamart endowed with these optimal structures analysis have the following advantages:

- Low volume data
- Faster consultation
- SQL queries and / or simple MDX
- Direct Validation of information

Process ETL

ETL processes are a part of data integration, but it is an important element whose full function the result of all the development of the cohesion of applications and systems.

The word corresponds to ETL stands for:

- Extraer: extract.
- Transformar: transform.
- Cargar: load.

With this, one can say that any ETL process consists precisely of these three phases: extraction, transformation and loading. Then defined what each one of them.

Phase extraction

To correctly perform the extraction process, follow these steps:

- Extract data from the source systems.
- Analyze the data extracted getting a checkup.
- Perform this check to verify that the extracted data meets the pattern or structure expected. If not, the data should be rejected.
- Convert the data into a format ready to begin the transformation process.

In addition, one of the most important to take into account during the extraction process would be preventions always require this task cause minimal impact on the source system. This requirement is based on practice, because if the data to be extracted are many the source system may slow or even collapse, causing could not be used again normally for daily use.

Phase transformation

The phase transformation of an ETL process applies a set of business rules or functions on the extracted data to convert data to be loaded. These guidelines can be declarative, based on exceptions or restrictions, but to enhance their effectiveness and pragmatism must ensure that they are:

- Declarative
- Independent
- Claras
- Intelligibles
- With a useful purpose for business

Charging process

At this stage, the data from the previous phase (phase transformation) are loaded into the target system. Depending on the requirements of the organization, this process can encompass a wide variety of different actions.

There are two basic ways to develop the charging process:

- Simple Accumulation: This way to load data is to perform a summary of all transactions within the selected period and carry the result as a single transaction to the data warehouse, storing a calculated value that typically consist of a sum or an average the magnitude considered. It is the simplest and most common way to perform the charging process.
- Rolling: this process would be recommended in cases where it seeks to maintain various levels of granularity. For this summary at various levels information is stored, corresponding to different groupings of the unit time or different hierarchical levels in one or more of the dimensions of the stored magnitude (eg total daily, weekly total, monthly totals, etc.).

Whatever the way to develop this process, keep in mind that this phase interacts directly with the target database, and so to do this all the restrictions that are defined in this shall apply. If they are well defined, the quality of the data in the ETL process is ensured (Figure 4).

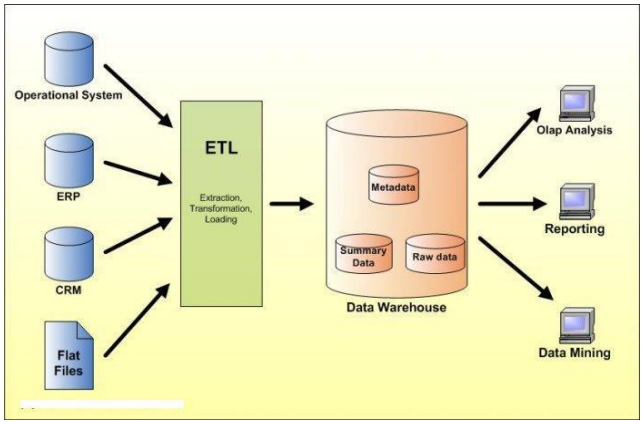

Figura 4. Proceso ETL

¿What is OLAP?

OLAP is an acronym for online analytical processing. It is a solution used in the field of business intelligence (or Business Intelligence) which aims to speed up the query of large amounts of data. It uses multidimensional structures (or OLAP cubes) that contains summary of large databases information. It is used in business reporting for sales, marketing, management reporting, data mining and similar areas

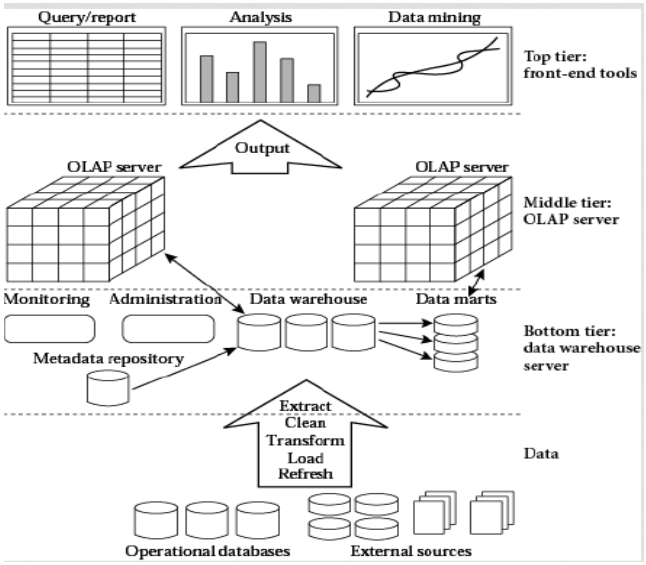

Figure 5. Structure of OLAP

The tool uses multidimensional data structures or OLAP cubes, which are multidimensional databases in which the physical storage of the information is performed on a multidimensional vector. OLAP cubes can be considered as an extension of the 2-dimensional spreadsheet, these contain summary of large databases or transactional systems information.

Classification and comparison of OLAP systems

According to the operation and structure, OLAP systems have been classified into different categories, such as ROLAP, MOLAP, HOLAP. The OLAP system most used today is called ROLAP.

Here are the different OLAP systems are described.

ROLAP

It means Relational Online Analytical Processing, OLAP is a tool built on a relational database. This system is important the fact table, where the history of the relevant information to the company that needs to be studied is stored.

In the OLAP industry, the ROLAP system is known for being able to scale large volumes of information, but its performance when running queries is unstable compared with another method of OLAP, MOLAP industry. However, after having conducted a study came to the following conclusion: that a company that uses ROLAP has lower performance than those using the MOLAP system (Figure 6).

MOLAP

Means Multidimensional online analytical processing, the system stores the data in a multidimensional array requires storage and processing and accumulation of information are contained in the OLAP cube.

HOLAP

Online analytical processing means Hybrid is a combination of ROLAP and MOLAP ordering systems allowing a portion of the data in a MOLAP while the rest makes it as a ROLAP (Figure 6).

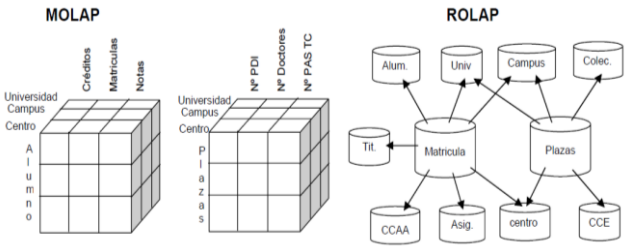

Figure 6. Vistas Molap y Rolap

Comparison of ROLAP and MOLAP systems

When comparing the two architectures can be made the following observations:

- The delegates ROLAP negotiation between response time and batch system design process, while MOLAP generally requires that their databases are pre compiled to achieve an acceptable query performance by increasing the batch requirements.
- Systems with high volatility of the data (those who change the rules of aggregation and consolidation) require an architecture that can perform this ad-hoc consolidation. ROLAP well this dynamic systems support consolidation, while MOLAP are oriented towards batch consolidations.
- The ROLAP support OLAP analysis against large volumes of elementary data, while MOLAP behave reasonably controlled volumes of data. The volume of information which are works which are implemented in the analysis cube, while the ROLAP environments at any time is available the entire database.
- The environment MOLAP stores information format structures with multidimensional arrays, while ROLAP managed using metadata information that map database schemas, Star, multidimensional views.

Just as organizations use a variety of tools for everyday OLAP may require different types of resources depending on the level or area of analysis. Environments planning, financial analysis and resource location may require MDDB environments. Meanwhile, the sales analysis environments or marketing campaigns that require data with millions of constant change, both product and customer or attributes, require ROLAP environments.

- ROLAP

  Data is entered directly from the data warehouse or other relational data source and are not stored separately.

- MOLAP

  The information is pre calculated and then is stored in multidimensional data cubes.

- HOLAP

  Maintains larger volumes of information in relational database and a separate MOLAP aggregations.

OLAP and OLTP

OLAP

These are applications that are responsible for analyzing business data to generate tactical and strategic information that serves as support for decision making. While OLTP transactions use relational databases or other files, OLAP achieves its maximum efficiency and flexibility operating on multidimensional databases.

Here are the salient features of these applications are set.

- Structure data transparently to the user.
- Only perform consultation work on the operational information generated by the OLTP systems.
- Consultations on large volumes of data predictable.
- Historic information.
- Batch update mode.
- Increased data resignation to facilitate the generation of queries and get good response times.
- Work with condensed summaries of thousands of records in a single response.

OLTP (On-Line Transactional Processing)

OLTP systems are oriented database transaction processing. A transaction generates an atomic process, and may involve operations insertion, modification and deletion of data. The transaction process is typical of operational databases.

- Access to data is optimized for frequent reading and writing tasks (for example, the enormous number of transactions that have to bear the BD banks or hypermarkets daily).

- The data are structured according to the application level (as program management, ERP or CRM implemented, departmental information system, etc.).
- The data formats are not necessarily uniform in the different departments (common the lack of compatibility and the existence of islands of data).
- The historical data is usually limited to current or recent data.

These are applications that define the normal behavior of an operational management environment and run the daily operations. The most common features of these transactions are:

- High / low / modifications
- Quick, terse and predictable Consultations
- Low volume of information and disjointed
- Fast transactions
- High level of concurrency
- Update mode on-line
- Low data redundancy

Examples of such applications are:

- Purchases
- Sales
- Inventory
- Wages

OLTP It has also been used to refer to the transformation in which the system responds immediately to user requests. A bank teller is an example of an application processing of commercial transactions.

The OLTP technology is used in many applications such as electronic banking, order processing, e-commerce, supermarkets or industry.

Concepts of multidimensional data (OLAP)

For its operation, OLAP applications use a type of database that has the peculiarity of being multidimensional, commonly called OLAP cube.

Basically, the OLAP cube, whose name comes from its multidimensional feature is a database that has several dimensions (Figure 7).

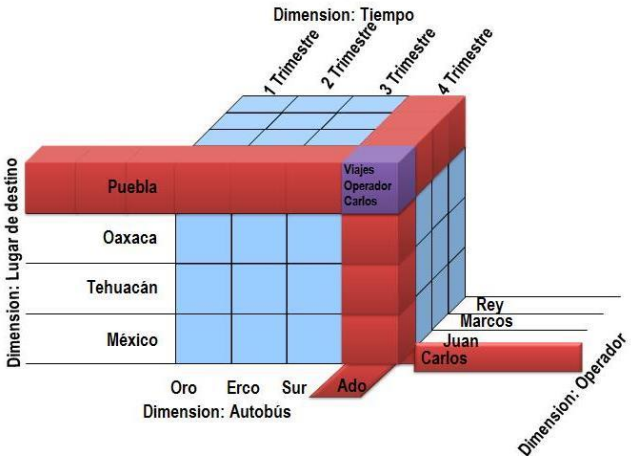

Figure 7. Cube OLAP

In the multidimensional analysis, data dimensions are represented by destination, bus, operator and time. In general, the dimensions are listed in hierarchies, for example, city, state, region, country and continent. Time is also a standard dimension with their own hierarchies such as day, week, month, quarter and year.

To have a simpler function OLAP cubes in a multidimensional database idea should be noted that each of the cube dimensions or scales basically corresponds to a data hierarchy.

other dimensions of the cube can be used to gather information about geographical locations, product classification categories, expenses incurred by the company, and others.

This confluence of information allows you to perform a complete analysis of various situations, to get the right solutions to business problems.

By incorporating these vectors or cubes, they have expanded the possibilities of relational databases, enabling processing of large volumes of information, otherwise it would be impossible to carry out such activity. Each of the dimensions that owns the database incorporates a particular field for a specific type, which can then be compared with the information contained in the other dimensions, to enable further evaluation and reporting of activities relevant to a company.

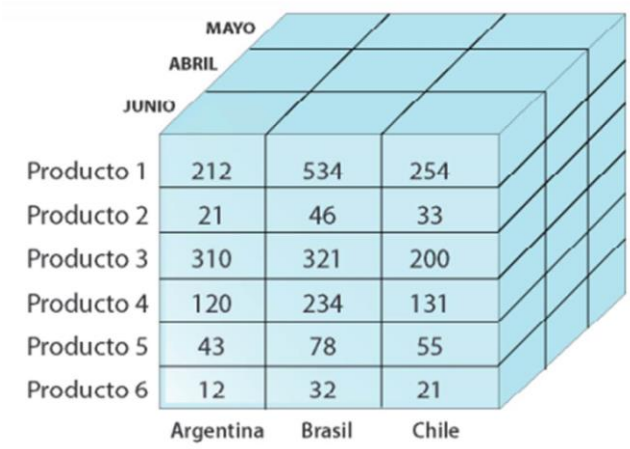

Figure 8. multidimensional structure

SMBD with OLAP

OLAP applications are one of the pillars of any solution Business Intelligence because that provides decision makers access to information sumarizada by convenient navigation methods that allow them to analyze and keep conversation flowing with data organization, optimal response times.

The OLAP database servers using multidimensional data structures to store information and the relationships between them. The best way to visualize the multidimensional structures is cubed. Each side of a cube is one dimension. The hub can be expanded to include another dimension, for example, product family. The hub also supports matrix arithmetic.

The OLAP technology supports common analytical operations, for example, consolidation, deepening and navigation.

The consolidation involves adding data, for example, simple or complex expressions totalizations involving interrelated data.

Deepening is the inverse operation involves the consolidation and display detailed information contained in the consolidated data.

Navigation refers to the ability to examine data from different viewpoints. Navigation often usually performed along a time axis, in order to analyze trends and patterns locate.

An example of how the OLAP system works, as to queries (Figure 9) is shown.

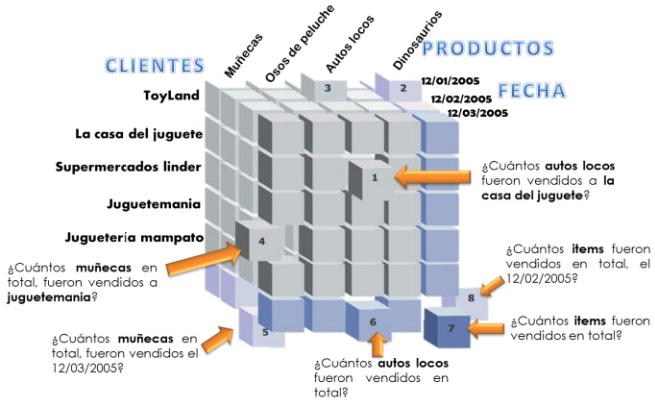

Figure 9. Example of an analysis with a bucket OLAP

Pentaho

Pentaho is a Business Intelligence tool, with the incorporation of the main tools of the Open Source market. It is currently the most comprehensive and widespread. It has a large development community that continually improves and extensions in the Platform.

It consists of an environment to which you can access via web safely. Each user displays all elements enabled profile, which includes reporting, OLAP analysis, and dashboards with indicators and tables. The user can create new views and new analytical reports and save them for later consult and export them to Excel, PDF or perform printing on paper.

It is a composite of different programs that meet the requirements of BI platform. It offers solutions for the management and analysis of information, including multidimensional OLAP analysis, reporting, data mining and creating dashboards for the user.

The platform has been developed under the Java programming language and has an environment based on the same tool implementation, Pentaho is making a flexible covering a high-end solution business needs. With Pentaho you can have a clearer view of the data (Figure 10).

Figure 10. OLAP with Pentaho

CONCLUSIÓN

Having done research regarding online analytical processing and the tool used to solve in the field of business intelligence, the following is concluded.

- Implement properly an OLAP system provides benefits of increased productivity for end users in your organization
- The controlled and timely access to strategic information enables a process in decision-making more effectively.
- Higher revenues and potential benefits are obtained by allowing the organization to respond quickly to market demands.
- Using OLAP rapid response in the consultation and better understood by the user view for decision making is guaranteed.
- Can exploit the advantages of multidimensional reports as OLAP cubes generated by the engine.
- PivotTables provide great impetus to the platform, providing users with the necessary for creating custom tools cubes.

Bibliography

Cervantes M., Ballesteros B.B., y Hernández F. (2012). "Programas para elevar la competitividad de las empresas en México" en Observatorio de la Economía Latinoamericana. Sitio web: http://www.eumed.net/cursecon/ecolat/mx/2012/vlc.html

Gonzalo H- Morales, Álvaro (2000). "Migrator OLAP", en Facultad de Ingeniería, Universidad de la República Oriental del Uruguay. Sitio web: http://www.fing.edu.uy/~ruggia/T5s/Migrador-OLAP.pdf

Hernández, E. N., y Mosquera, C. I. I. S. (2012). "Pentaho: software líder de Inteligencia de Negocio de código abierto", en Revista Telem@tica. Sitio web: http://revistatelematica.cujae.edu.cu/index.php/tele/article/view/44

Ibarra M, A. (2006). "Procesamiento Analítico en Línea (OLAP)" en Universidad Nacional del Noreste. Sitio web: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/OLAPMonog.pdf

Informática Hoy (2007-2012). "Cubo OLAP; una base de datos multidimensional". Sitio web: http://www.informatica-hoy.com.ar/telefonos-celulares/Cubo-OLAP-una-base-de-datos-multidimensional.php

Núñez F. D. J. (2011). "Sistemas ROLAP y MOLAP". Sitio web: http://repository.uaeh.edu.mx/bitstream/handle/123456789/14495

Pentaho (2013). "Pentaho BI - Software Libre para la Toma de Decisiones". Sitio web: http://gravitar.biz/pentaho/

Romucho P. (2010). "Bases de datos y sistemas de información", en Monografias. Sitio web: http://www.monografias.com/trabajos77/base-datos-sistema-informacion/base-datos-sistema-informacion.shtml

Sinnexus (2012). "Datamart", en Sinnexus. Sitio web: http://www.sinnexus.com/business_intelligence/datamart.aspx

Tamayo, M., y Moreno, F. J. (2006). "Comparing the MOLAP the ROLAP storage models", en Ingeniería e Investigación. Sitio web: http://www.scielo.org.co/scielo.php?pid=S0120-56092006000300016&script=sci_arttext&tlng=pt

Urquizo P. (2011). "¿Qué es OLAP?", de Business Intelligence. Sitio web: http://www.businessintelligence.info/definiciones/que-es-olap.html

Wikipedia (2013). "OLTP", Enciclopedia libre. Sitio web: http://es.wikipedia.org/w/index.php?title=OLTP&oldid=70248939.