Concepts about BI, data mining and datawarehouse.

Adriano Oliveira[1], Debora Alves[2] , Felipe Santana[1].
[1] *Escola Politécnica da Universidade de São Paulo.*
[2] *Instituto Federal de Educação Ciência e Tecnologia de São Paulo.*

Abstract:

Business Intelligence is a term that is used in many ways, that's why is really hard to define it. But many times it is used not properly for IT professionals, that for ignoring the right concepts around it, end up thinking that everything that has some reference to Business Intelligence is indeed Business Intelligence. This article seeks to clarify some of the most important concepts about Business intelligence, regarding some process and technologies involved in the implementation of such technique.

Introduction:

If there was any expression that could define what is Business intelligence and its characteristics, it would be simply "IT into business". That means that Business Intelligence is based on applying software technology to improve business performance.

With the current fast advance in technology fields and the changes going on the way of doing business; it is getting obvious the need to make decisions and evaluate performance of the companies in a more agile and efficient way. Nowadays, many companies have lots of different data source, storing important data, such as sales statistics and historical transaction registers that could be used to assess business conditions and financial performance. But this big amount of data is usually spread in many different data sources, like OLTP databases, excel sheets and other different data storage devices.

This heterogeneity of business data sources, most of times turns this big amount of data into meaningless stored objects, impossible to be used to evaluate company performance and business process status. That is when business intelligence technologies do their jobs; using computer based technologies to extract the data from all these different sources and process them, in order to provide meaningful statistical and historical data, that could be used by the company to analyze its current status and making the right decisions.

Business intelligence process and concepts:

Extract, transform and load:

Most of times the implementation of a Business intelligence solution requires the extraction of data from heterogeneous data sources, as said before, as well as the transformation of the data to fit the new data source requirements, and then loading the transformed data into a new data source. This process is called ETL(extract transform and load), and it is one of the first things to do when performing a BI solution, so all those different data could be put together into one unique data source.

For example, a company has a sheet in an Excel file that contains the sales amount corresponding to each employee, it also has a table in a data base that contains the region of the countries corresponding to each employee. In an ETL process, it is possible to retrieve the information from these data sources, join them together and then insert into a new data source, creating for example a table containing information about sales, employees and country regions.

Data analysis and data warehousing:

After perfoming the extracting, transforming and loading, the resulting data should be stored into an appropriate data storage system. A very important aspect of a BI solution is the design and implementation of the way the system data will be stored, in order to keep a structure that enables the user to retrieve meaningful and efficient business information; and improve the company capability of making decisions. That is when a Data warehouse is very useful.

A Data Warehouse is an OLAP (Online analytical process) data storage device that provides a flexible and analytical manner of storing data. In a data warehouse, data are stored in multidimensional structures, named cubes or datamarts; that contain a specific piece of data from the database. This cubes also contains dimensions that specify context for the data retrieved.

A cube is a so called multidimensional structure that contains information about a well defined piece of data from your data source. A cube is composed by a fact table and many different dimensions, organized in many different forms. The fact tables contains the main information of the piece of data you want to store; the dimensions specify a context for the main information to be shown.

For example, if the main information you want to store is the sales register of some company; so the fact table of the cube has to be the sales table. And imagine you want to organize these registers considering the products that were sold, the employees who sold it and the period of the year; so the dimensions of the cube would be the tables Product, employee and may be month/year. That is an example that well illustrates the scenario of building a cube for a data warehouse.

The process to get those data from the database and build the necessary cubes is called data analysis. There may be hundreds of cubes in a data warehouse, each of them are composed by a fact table and its dimensions. This structure of the data storage mechanism into a data warehouse enables the visualization of the data by so many means, as long as the cubes have their dimensions well projected.

Data mining:

Another process that is really important in business intelligence solutions is made after gathering all those different data in one data source, and it is called data mining; where the data is processed under statistical based algorithms, that could retrieve useful information from it.

Dealing with the significance of amounts of data, instead of single information, is a very familiar activity to every form of human knowledge. In that way, using statics as basis for data analysis becomes more and more relevant for generating profitable and helpful knowledge in business. Data mining is the faculty of applying rational analytical process - statical techniques - to amount of data in order to make patterns emerge. Data mining takes advantage of computational ability to dive deeper in meaning of patterns than pure statics. When allied with a data warehouse storage capability and the concepts of BI, data mining can give to a business or company, precision and security while having to make decisions.

Large amounts of data, like a census of a country or the profile of clients of a big company, would be almost impossible to extract reasonable information without the help of precise techniques and processes. Understanding the dimensions of problems like these, wise dealing with this data demands proper data mining processes.

In the actual business context a data mining process can be divided into four fundamental stages:

- Classification: Is the part which raw data must be separated into groups, in order to separate the natural patterns (the visible ones).
- Clustering: With the help of technology and specific organizational techniques, data is managed to compose smaller groups (less visible to human's eyes).
- Statistical fitting: In this stage, is searched for knowing mathematical patterns in data characteristics.
- Association: Useful information normally concur many variables, which bonds can be found in association processes.

Business intelligence technologies and tools:

There are many technologies and tools developed by different IT companies, used to implement Business Intelligence solutions. Companies like IBM, Microsoft, Oracle and SAP have put some efforts to provide useful tools for performing some of those BI processes described above.

Microsoft have developed tools for performing basically the first two process above, as well as developing reports based on the data stored; all of them are based on Visual Studio IDE. For ETL, Microsoft has the Integration Services framework that can extract data from Excel sheets, text files or many different databases (including DB2). For Data analysis, there is Analysis Services, that is capable of building the cubes, its dimensions and attributes using an OLTP data source; and storing these structures into an OLAP database (data warehouse). After all, there is Microsoft reporting Services, used to create reports based on the data in the OLAP or OLTP database.

There is also Cognos 8, one of the most important tools used for Business intelligence solutions implementations. Cognos 8 is a SOA based tool set, that provides a full range of capabilities; such as Analysis, reporting, dashboarding and scorecards.

These tool set is composed by many easy to use, web based interfaces that enables the creation of business reports, construction of datawarehouse solutions. As well as measuring performance to accomplish business objectives. In other words, IBM Cognos 8 provides functional capabilities necessary to build a whole Business Intelligence solution.

References:

[1] – **Curso TechNet**, Academia BI – Brasil.

[2] – **Analysis Services Overview** - Microsoft – December 2007

[3] – **IBM Cognos 8 Business Intelligence Analysis Data Sheet** - IBM – Canada – 2009

[4] - **IBM Cognos 8 Business Intelligence Reporting Data Sheet** - IBM – Canada – 2009

[5] – **Data mining: Concepts, Models, Methods and Algorithms** - MEHMED. K – John Wiley & Sons – 2003

[6] - **From Data Mining to Knowledge Discovery in Databases** – FAYYAD. U; PIATETSKY. G and SMYTH. P - 1996