Business Intelligence in the Electric Industry

Alfredo Ochoa[1], Diego Uribe[2]

[1] CENACE Área Norte
Departamento de Aplicaciones de la Laguna
Gómez Palacio, Dgo., C.P. 26100
alfredo.ochoa@cfe.gob.mx

[2] Instituto Tecnológico de la Laguna
División de Postgrado e Investigación
Blvd. Revolución y Cuauhtémoc s/n
Torreón, Coah., C.P. 27000
duribe@itlalaguna.edu.mx
Paper received on 11/08/08, accepted on 11/09/08

Abstract. In this study we show how the technology of Business Intelligence has been implemented at the area of Control Norte in order to support the multiple necessities of analysis which demand a diverse sort of users. In this specific case, we have designed and developed a Data Warehouse for the optimum analysis of the massive information which is produced every day. Once the repository was created, we applied Data Mining techniques to support the making decision processes; in particular, the forecast of the electric power consumption.

1 Introduction

From year 2001, the Comision Federal de Electricidad (CFE) was committed with the federal government to implement high quality standards in the institution and consequently initiated the definition and development of the Programa Institucional de la Calidad Total (PICT). On the basis of it, the Centro Nacional de Control de Energia (CENACE) participates in the implementation of this program since 2002 and continues with the contribution of advances. The different computer-based information systems which operate in the area of Control Norte from years 80s and 90s and that support the key functions for the decision making process on the management, supervision and control of the Sistema Electrico Nacional (SEN), have an important accumulation of valuable data on the history of operation of the national mains. These data that doubtlessly represent one of the strengths of the company, reside either in different online data bases available for their processing or in storage media with a wide diversity of facilities to recover them for their optimum analysis for the institution. Until now, however, the potential of accumulated data has not been exploited in order to obtain new knowledge applicable to the administration of the SEN. The aim of this article is to describe the design of an effective strategy of administration of the Information Technologies (IT) in the area of Control Norte of the CFE with the purpose of optimizing the decision making process. Said in a dif-

©M. G. Medina Barrera, J. F. Ramirez Cruz, J. H. Sossa Azuela. (Eds.).
Advances in Intelligent and Information Technologies.
Research in Computing Science 38, 2008, pp.213 -.223

ferent way, we intend to make a more refined use of the available information with the intention of discovering the hiding knowledge of the organization in order to become a highly competitive industry [1]. It is indeed here where the technology of Business Intelligence is of our interest. In fact, since the appropriate integration of this technology allows us to transform the operative data of the organization into resources of knowledge indispensable for the decision making process [2], the incorporation of this technology in the area of Control Norte has been an urgent measurement.

In this case study, the particular problematic to be solved is the inappropriate management of the huge amount of information provided by the multiple and diverse power plants of the company. For such effect, the present system named Hoja de March (HM), a typical OLTP system, is described as well as the systems which supply of information to HM commonly called in this context measurement systems. We describe in the section 3 the architecture in which the design and development of the data store (Data Warehouse) of this project is based on. More specifically, in this section we first describe the metadata and the various measurement systems (OLTP systems). We then describe in section 4 the processes of extraction, transformation and loading (ETL) developed to integrate the information to the data repository. Finally, in section 5 we show the evaluation results which exhibit the quality and the opportunity of the data that a measurement system contributes.

2 Present situation

The area of Control Norte supervises the states of Chihuahua, Durango and the Region Lagunera of Coahuila. Also, it coordinates five subareas of Control: Laguna, Durango, Camargo, Chihuahua and Cd. Juarez. Each one of these subareas controls different electrical substations and generating units, concentrating all the information related to energy in the HM system. To be more precise, the CENACE at this moment compiles hourly energy information from the generators and mains in order to determine the hourly demand by subareas and areas of control.

2.1 Operation processes

The processes of generation, transmission and distribution of energy are made in real time. For this reason it is necessary to know the behaviour of the different components which integrate the electrical power system. The HM system contains all the information produced by the three key processes of the company: generation, transmission and distribution of energy. There is a process named *balance of energy* whose purpose is to provide information about the origin of the generation of energy in our own plants (and the generated in external plants) as well as the consumption by area and subareas. All this information is hourly described so the data provided by this report is expressed in kWatt/hour terms. In this way, this process is relevant for the analysis of trends in the demand, consumption and generation of energy. Said in another way, in order to check the total consumption of energy corresponding to the area and subareas, the HM system reports the gross generation of the power sta-

tions and the foreign plants, the interchanges of energy between the area of Control Norte and the areas of Control Noreste and Noroeste, and the energy which is imported in hourly terms. The HM system is then crucial for the monitoring task of the administration: through it we are able to monitor the energy which is obtained, transmitted and finally sent to the client in hourly terms, that is, each hour.

2.2 The HM system

The HM system is the information system used to integrate, throughout the different geographic zones, the information supplied by the multiple sources of energy and the transmission mains. From the compiled information the system reports the crucial points for the company: the production and consumption of energy. The hierarchy of concepts which is operated by the HM system is illustrated in Figure 1. Figure 1 represents the relation between the different concepts derived from basic data such as the production of the power stations and the mains. The demand is the top level concept in the HM hierarchy and represents the relation between generation and the flow of energy between the mains. In the case of the area of Control Norte, the demand is determined at subarea level so that the sum of the different subareas (Laguna, Camargo, Chihuahua, Durango and JuÂ¡rez) establishes the total demand of the area. Also, the Figure 1 shows the different types of energy technologies in which the electric power production of the multiple plants is based on. On the other hand, the connections or mains represent the transmission lines that join two subareas or areas of control. In order to identify the flow of energy between two areas or subareas of control, a connection is made up of a line of input and another one of output. When in a particular point the production (generation) exceeds the consumption, there is a flow of energy towards the output line of that particular point and vice versa: there is a flow of energy in the input line when the consumption exceeds the production (increase of the demand). All this under the basic principle of conservation of energy: *the energy can only be changed from one form to another*. As we previously said, multiple data sources provide relevant information about the energy generation and the interchange of energy between the different subareas (connection lines). In this project, the multiple data sources are named *measurement systems* and all the relevant information supplied to the HM system is expressed in hourly terms. In this way, a measurement system contributes with either generation or connection data to the HM system. We describe in the next section each of these data sources.

2.3 Data sources for HM

The data sources which provide the information to the HM system are measurement systems supported by a particular database management system under the OLTP concept. These measurement systems represent the primary source of data for the HM system. It is important to point out that the data are validated by the control operator who is responsible to supervise the information during the day. The measurement systems that supply data to the HM system are shown in the Table 1.

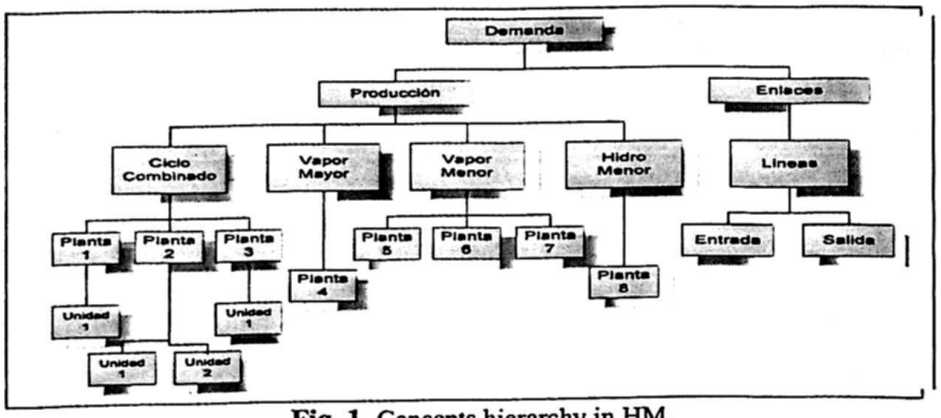

Fig. 1. Concepts hierarchy in HM.

Table 1. Multiple data sources: measurement systems.

System	Unity	DBMS	Server	OS
SIMER	Kilowatt/hr	Informix	Alpha Server	Unix
ESA	Megawatt/hr	Oracle	Intel	Windows
ION PEE	Kilowatt/hr	External	External	External
SIMO	Kilowatt/hr	Informix	Alpha Server	Unix
PI Osisoft	Megawatt/hr	Oracle	Intel	Windows

2.4 Feeding the HM system

There are two processes to integrate data in the HM system: automatic integration and manual integration. The automatic integration is a program that obtains magnitudes from the different measurement systems necessary to elaborate the report of the demand in hourly terms. This program is like a cron process which has been configured to automatically run for every 25 minutes. When by some external reason it is not possible to make use of automatic measurements, these must be estimated by the control operator or to look for an alternating source of data. The manual integration is carried out by the control operator who not only is responsible to estimate valid data but also to correct deviations observed in the incoming data. This process is executed by making use of an interface on the CFE intranet. The information of the HM system has been stored by long time in plain text files (code ASCII) which have been processed by making use of Perl programs. Figure 2 shows the components, and their relationships, which integrate the HM system. All these elements have had to be analyzed to design the data repository, that is, our Data Warehouse. As we said, the measurement systems provide the generation and transmission data which are to be gathered into a data warehouse in order to assist in the extraction, analysis, and reporting of information.

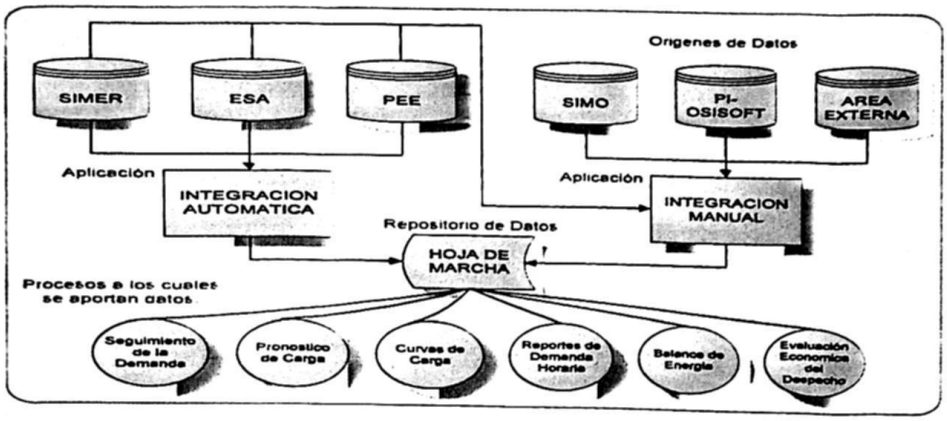

Fig. 2. The HM system.

3 Data Warehouse architecture

The description in the previous section of the present scheme of the HM system makes obvious the necessity to think about a technology that integrates different platforms and different data structures with the purpose of facilitating the data management. It is in this point where it becomes essential, and therefore strategic, the study of the proper architecture for the implementation of a data warehouse. For example, not only must we consider the dimension and fact tables which allow the analysis of the information as a cube of data, but we also look for more specific benefits as the hiding of details involved in the measurement systems in such a way that the user can focus his attention on his analysis task rather than on data sources issues. On the basis of the present scheme of the HM system, from the three more common architectures [3] we selected the architecture with an area of processing known as *staging area* (area in which the processes of extraction, transformation and loading taking place1), being the heterogeneity of the data sources the key factor for its selection. Said in another way, given that in our case of study the measurement systems exhibit a great heterogeneity, the architecture with staging area is the groundwork for the design of our data repository. Figure 3 displays this architecture.

3.1 Dimensional modeling

Due to its analytical orientation, the data warehouse technology entails a different thinking which relies on an intrinsic modeling of data bases known as *multidimensional modelling*. This perspective based on multiple dimensions intends to offer a high level perspective to the user about the company business [4]. The basic structure of a data warehouse which relies on multidimensional modeling is defined by two basic elements: fact and dimension tables [5].

The *fact* table is the central table in a dimensional design since it is in this table where the numerical measurements of the business, that in our particular case are the hourly measurements, are stored. On the other hand, the *dimensions* tables contain the detail of the values that are associated to the fact table so that the granularity of

the information of these tables plays a crucial role in the quality of the data repository. Figure 4 shows the fact and dimensions tables defined in our particular study that as can be seen represents the characteristic *starlike* structure. In fact, the Figure 4 shows the N_HECHO_HM table as the center of the data repository. It is the crucial table since it contains the attributes which describe the collected data for the comprehensible analysis and reporting of information provided by the HM system.

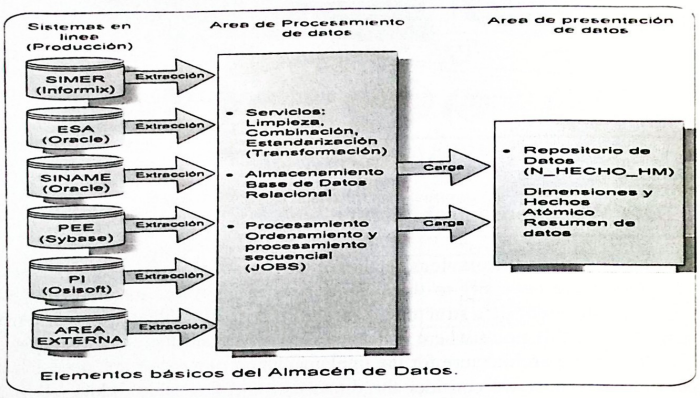

Fig. 3. Data repository and ETL processing used in the experiments.

4 Extraction-Transformation-Loading process

The use of tools for the extraction, transformation and loading of data are the starting point of any strategy launched by a company which intends to discover the hidden value of its information [6]. The access to a huge volume of information that it is produced in the area of Control Norte is not an easy task. As the organization grows, the company acquires different applications, operating systems, hardware platforms and data bases. All these resources are distributed through different departments, offices or processes. In this way, these areas of information are eventually more isolated and becoming more unusable by the rest of the company. The accessibility to the multiple and heterogeneous measurement services involved in this project has been carried out by making use of the Oracle facilities as we can gain access to different repositories from a unique platform, which also represents a valuable advantage since in this way we use a unique language to extract the data. SQL and PL-SQL of Oracle represent the base for the development and implementation of this stage of the project.

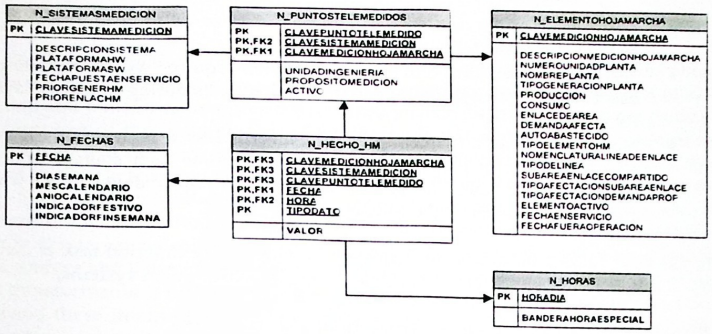

Fig. 4. Fact and dimensions tables in our dimensional modeling.

4.1 Extraction of data

Taking into account the diversity and complexity of the structures which constitute our measurement systems (our data sources), the implementation of the ETL process in this project considers two methods for the data extraction task that are described next:

1. Full extraction: the extraction of the whole information stored in the data source is carried out. That is, all the data set available in the data source is transferred thus it is not necessary to maintain a register of changes in the data source (later modifications do not exist).
2. Incremental extraction: only new or edited data are transferred. That is, data which has been modified or new records at the source will be extracted and transferred to the target repository.

We also reviewed other alternatives which Oracle offers for the extraction of data. For example, it is interesting the extraction based on the monitoring of changes known as Change Data Captures [7], which consists of an architecture which take care of the alteration of the data. Nevertheless, this type of extraction, which is ideal when the data sources are uniform, was discarded because of the diversity previously mentioned of our data sources. In short, and taking into account the diversity and complexity of our data sources, the following schemes of extraction were implemented:

- Initial: in which it is considered to extract historical information. This process will be unique for some measurement systems.
- Hourly: since the required information is hourly information, this scheme of extraction will be the most used in the implementation of the ETL process.

- Daily: some of the data sources add new data or edit existing data at the end of a day or the following day.
- Weekly: another scheme of extraction which requires weekly updating of data because some systems attend integration tasks during different days of the week.
- Monthly: it is not a frequent process. Nevertheless it is necessary to consider a process for monthly loading because some data sources undergo every day/every hour. This process is common in those data sources which are not administered by the company.

In this way, Table 2 illustrates the frequency in which the extraction task is carried out at the different data sources, that is, the different measurement systems.

Table 2. Schemes for the extraction task.

Data sources	Initial	Hourly	Daily	Weekly	Monthly
SIMER	X	X			
SIMO	X		X	X	X
ESA	X	X			
IONPEE	X	X	X		
PI OsiSoft		X			

4.2 Transformation of data

One of the challenges of any implementation of a data warehouse is the problem of transformation of the data. The transformation takes basically care of the activities of cleaning, combination and standardization of the data so that the inconsistencies which characterize to heterogeneous data sources may be integrated efficiently in a common repository. In our particular case, it was necessary to implement aggregate operations in some systems for which it was required to estimate hourly averages or conversions between different measurement units. Another important transformation is carried out when the records of a particular data table must be transposed to turn them to hourly information i.e. to produce 24 records from a record of 24 hours. Figure 5 illustrates the use of one of the transformation techniques that Oracle makes available: the *table functions*, which allow implementing aggregate operations on a set of records.

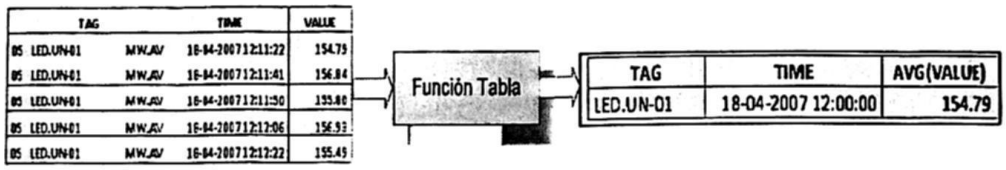

Fig. 5. Oracle's table functions.

4.3 Loading of data

The loading phase is the step in which the resulting data of the transformation phase are deposited on the target repository. Depending on the requirements of extraction and transformation, this process may include different tasks. In our particular case, one of the requirements of the project is to maintain historical records in such a way that the user can also obtain historical information. In this project, the use of an Oracle technique [3] for handling errors during the loading phase was crucial. Such technique allows the definition of the loading tasks by making use of powerful SQL instructions and procedures. When inadequate data for their integration to the target repository occur (it happens when the data come from different data sources and these inconsistencies are not detected by the processes of extraction and transformation), the demanding data loading operations are interrupted. In order to avoid these interruptions and to guarantee the load of trustworthy data onto the repository, we make use of the extended OracleÂ´s data manipulation language (DML) which allows defining a table for the management of detected errors for its later processing.

5 Experimental Evaluation

After some years of experience working with traditional systems in the company, it is possible to mention that the implementation of this project has made a major contribution to the handling of information. To be more precise, the foremost achievement of the project is the implementation of the ETL process which is the foundation for the creation of a data repository with reliable and opportune information, aspects that can be evaluated in different integrated OLAP systems.

For evaluation purpose, we focus our attention on the availability of the information. In other words, we want to determine whether the data which come from the measurement systems are available when they should be. Figure 6 shows whether the data are available to be deposited on the target repository. In this particular example, by making use of a period of 20 minutes as parameter, the information is considered opportune as long as the data are on time. With almost 86% of availability of the data in a margin of 20 minutes, and taking into account the heterogeneity of the data sources, we can argue the performance of the data repository is rather than acceptable.

6 Conclusion and directions for further research

The incorporation of Business Intelligence in the electrical industry has been a rewarding experience. The enormous volumes of information represent a great challenge due to the heterogeneity of our data sources: the electric power measurement systems. Working on compilation and cleaning tasks, as well as the transformation of the data, which come from the operational systems (no-structured information), into structured information, was fundamental for the development of the data repository for the analysis and support of the decision making process at the CENACE.

In fact, once the data warehouse was created, we proceeded with the integration of data mining techniques to identify useful patterns for the organization, or for the optimization of the processes related to prediction.

In our particular case, we have worked in the energy demand forecast with satisfactory results by making use of the Oracle Miner Data tool [8]. In short, an effort has been made to transcend the simple and traditional handling of data. According to the relations between the components of the informational chain [9], the data warehouse created has made possible the top-level analysis of the information as well as the elicitation of knowledge.

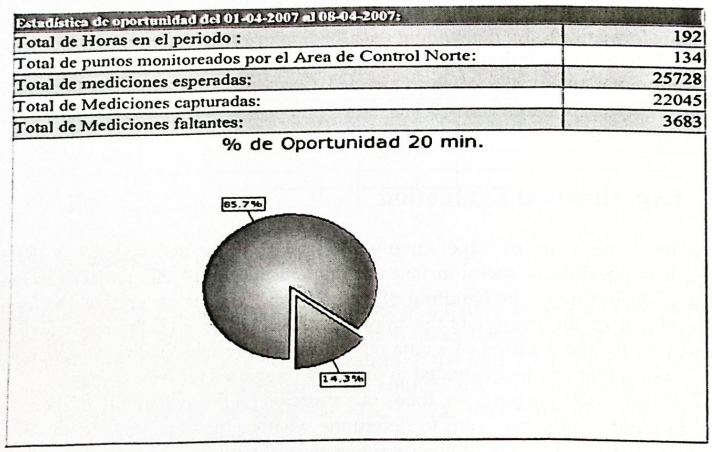

Fig. 6. Evaluation report.

References

1. Becerra Fernandez, I., Gonzalez, A. In: Knowledge Management. Pearson, Prentice Hall (2004).
2. Williams, S., Williams, N. In: The Pro¯t Impact of Business Intelligence. Morgan Kaufmann (2006).
3. Oracle: Oracle database data warehousing guide 10g release 2 (10.2), http://www.oracle.com/solutions/businessintelligence/dwhome.html (2005) Oracle Corporation.
4. Kimball, R.: The anti-architect. how not to design and roll out a data warehouse. In: Intelligent Enterprise. (2002).
5. Kimball, R. In: The Data WareHouse Toolkit. Wiley (2002) Second Edition.
6. Mimno, P.R.: How to select an extraction/transformation/loading (etl) tool. In: 101 Data Intelligence Solutions. (2005).

7. Nagarkar, N.: Change data capture implementation in oracle data warehouses. In: Database Journal. (2003).
8. Oracle: Oracle data mining guide 11g, http://www.oracle.com/solutions/businessintelligence/data-mining.html (2007) An Oracle White Paper.
9. Bertalan®y, L.v. In: Tendencias en la Teor¶a General de Sistemas. Morgan Kauf-mann (1990).