Applying Data Warehousing Technology to Support Planning and Control on Mass Transit Companies

João Mendes Moreira * Jorge Freire de Sousa *

* INEGI – Instituto de Engenharia Mecânica e Gestão Industrial
Rua Dr. Roberto Frias s/n; 4200-465 Porto – Portugal
FEUP – Faculdade de Engenharia da Universidade do Porto
Tel.:+351-225081639, Fax: +351-225081797
{jmoreira, jfsousa}@fe.up.pt

Abstract

This paper describes the methodology used in the construction of a Data Warehousing project to support planning and control on mass transit companies. The data sources used were the ones from the GIST98/EUROBUS system - a computer application for supporting operational planning in public transport companies. Firstly, the main Data Warehousing concepts and definitions are introduced. After the problem description and a brief analysis of different Data Warehousing methodologies, the analysis and development phases of the Data Warehousing project are presented. Finally, in the conclusions, the main advantages to implement such a system in the situation described are emphasized.

Keywords: Data Warehousing, OLAP, Mass Transit, Transportation Management

1 Introduction

When computation began to be used massively in the sixties/seventies, the data moved from paper files to device drivers but, in the main, the role that data had in the companies did not change so much. The manager of a company made use of the same information no matter if it was on paper or in a device driver. Along time, information and time won an increasing importance, and the pressure to get more and faster information also increased. In the nineties, Data Warehousing technology was an answer to that tendency that is still going on, particularly with the possibility

Apdio © 2006 *Associação Portuguesa de Investigação Operacional*

to use the data to extract hidden information. It is the era of the data mining and knowledge discovery.

In this paper the construction of a Planning Indicators Board (PIB) is described. The PIB is a top level oriented module of the GIST98/EUROBUS system, a Decision Support System (DSS) for mass transit companies. The proposal is to collect and to show the planning and control information to support the decision-making process using Data Warehousing technology.

Firstly, the main Data Warehousing concepts and definitions are introduced, followed by the description of the GIST98/EUROBUS system and the presentation of a Data Warehousing development methodology. The PIB's construction is described by using the structure of that methodology.

2 Data Warehousing concepts and definitions

Traditionally, the operational databases are:

- Data oriented: the database design depends on data itself and not on the way the data is red by end-users applications. Roughly we can say that the paradigm is: ”The less is the space needed to store the data, the better is the database design”.
- Not integrated: the data is spread across several databases not necessarily consistent.
- Volatile: the data can be changed.
- Time invariant: often, it is not possible to know the same information at different moments in time.

Databases with these characteristics are not adequate to support management's decisions because queries are time consuming, information is not integrated, and it is not guaranteed that the information is accurate and stored along time. The answer to these problems is the Data Warehouse: “a subject oriented, integrated, non-volatile and time variant collection of data in support of management's decisions” [6]. This concept became quite popular on Information Technology in the nineties, because, on one hand, the information won an increasing strategic importance in the companies and, on the other hand, hardware became cheaper and technologically much better.

The definition of Data Warehouse is data-oriented and does not include all the processes connected with the Data Warehouse technology. In order to get a process-oriented definition, the term Data Warehousing became more popular: “Warehousing refers to a set of processes or an architecture that merges related data from many operational systems to provide an integrated view of data that can span multiple business divisions” [16].

Connected with the Data Warehouse concept appears in 1993 the term OLAP (OnLine Analytical Processing) that is “a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user (. . .)” [8].

2.1 Data Warehousing reference architecture

As it can be seen in Figure 1, the Data Warehousing architecture can be divided, typically, in three big components:

- The sources: operational databases and/or data files.
- The Data Warehouse and all the software needed to load, administer and manage it: the Data Warehouse administration and management software is like a Data Base Management System but with more capacity to manage huge amount of data and to data reading operations. The data acquisition is the software responsible by the data migration from the sources to the Data Warehouse.
- The decision support tools: the frontend & report tools – they are easy to use tools that allow enduser to see and manipulate data in a useroriented way; the OLAP tools – they allow to organise and manipulate data in a multidimensional way; and vertical and Data Mining solutions – they are able to find, in a semi-automatic way, patterns, associations, changes and disturbances, using statistical methods and a big variety of algorithms.

Data Marts can be seen as small Data Warehouses. If the Data Marts load data directly from operational systems they are said to be independent. If they load data from the Data Warehouse, they are defined as dependent. The main difference between a Data Warehouse and an independent Data Mart is that the Data Mart is departmentally or functionally structured, while the Data Warehouse is organisationally structured.

The Data Warehouses and the Data Marts can be stored using the relational model (ROLAP), the multi-dimensional model (MOLAP) or a hybrid model that uses both relational and multi-dimensional models (HOLAP). The relational model [3] is the most popular of all database models. The multi-dimensional model [15] can be seen as a set of multi-dimensional arrays. For each array, the number of results (the NULL is a possible result) is the Cartesian product of the several possible arguments. A multi-dimensional array has the name of Multi-Dimensional Structure or Cube.

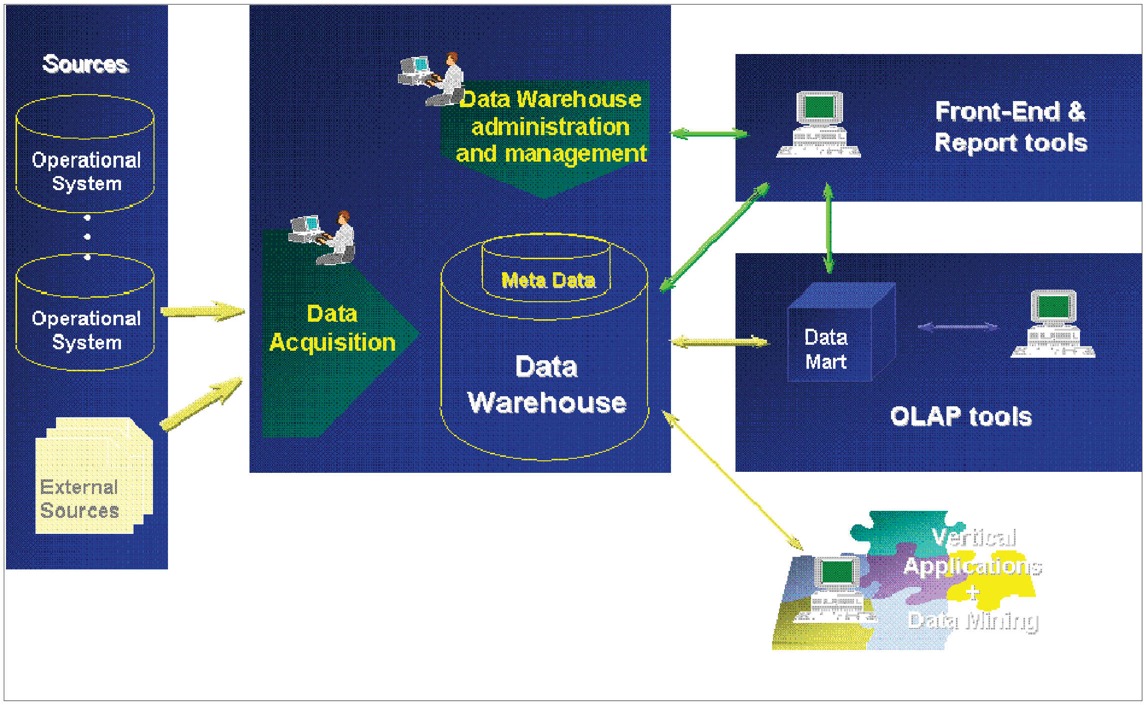

Figure 1: Data Warehousing reference architecture (translated from [12])

3 The GIST system

The GIST system [2], [13] is a computer application for supporting operational planning in public transport companies. It was developed as a decision support system that aims to help mass transit companies to improve the operation of critical resources, such as vehicles, drivers and planning staff. The system is also an important software tool to support tactical and strategic management studies regarding companies operations. A consortium of 5 leading Portuguese transport companies (CARRIS, STCP, Horários do Funchal, Empresa Barraqueiro and Vimeca) and 2 R&D institutes (INEGI-FEUP and ICAT-FCUL) is responsible for the GIST system.

The GIST system was successfully installed in those companies in 1996. In 2001, this application was developed under the name of GIST98/EUROBUS. This evolution represents improvements both to the GIST functions and extensions of its functionality. Nowadays, it is being used by eight Portuguese companies that operate, daily, about 2800 vehicles, corresponding broadly to 21% of the road public transport market in Portugal, including Madeira and Azores.

3.1 The GIST98/EUROBUS system

The GIST98/EUROBUS system (Figure 2) contains the following modules:

- Network Module, allowing the definition of the transportation network;
- Gist-Line Module, the route information module;
- Trip and Vehicle Scheduling Module, allowing the trip timetable definition and the vehicle scheduling information management and optimisation;
- Crew Scheduling Module, the crew scheduling information management and optimisation module;
- Crew Management and Rostering Module, which defines the daily tasks to every employee and where various optimisation algorithms are applied to the rostering rules.
- User Information Module, an user-oriented module producing information to the users of public transports;
- Performance Indicators Board Module, a top level oriented module providing planning and control indicators to support the decision making process.

The first four modules are upgrades of the GIST modules. The Crew Scheduling Module, the Crew Management and Rostering Module, the User Information Module and the Performance Indicators Board Module belong to the EUROBUS project that has been financially supported by a public institution named ‘Agência de Inovação’. All these modules form the GIST98/EUROBUS system.

3.2 The Performance Indicators Board Module (PIB)

The PIB Module filters the information derived from the other modules so that the toplevel managers can access, in an easy way, the relevant information for the decision making process. The main problem in structuring the PIB Module is to identify those indicators that are significant to top-level managers. All other information is included in the corresponding module. As the GIST98/EUROBUS is a decision support system to the operational planning, the information that it can provide it is essentially the planning and control information.

4 The methodology

There are several Data Warehousing methodologies [1], [16]. In the following, instead of describing them, it will be highlighted what is common to those methodologies (Figure 3), namely the Hadden-Kelly method [5], the methodology from NCR Corporation [4] and the methodology from DataWing Consulting Services, LLC [17]. In fact, it is possible to say that these methodologies have three main phases, even if in some of them, those phases are splitted in more than three. The phases are analysis, development and implementation. The analysis phase includes all the tasks related to the process definition, the requirement analysis and the software and hardware architecture design. The development phase corresponds to the database

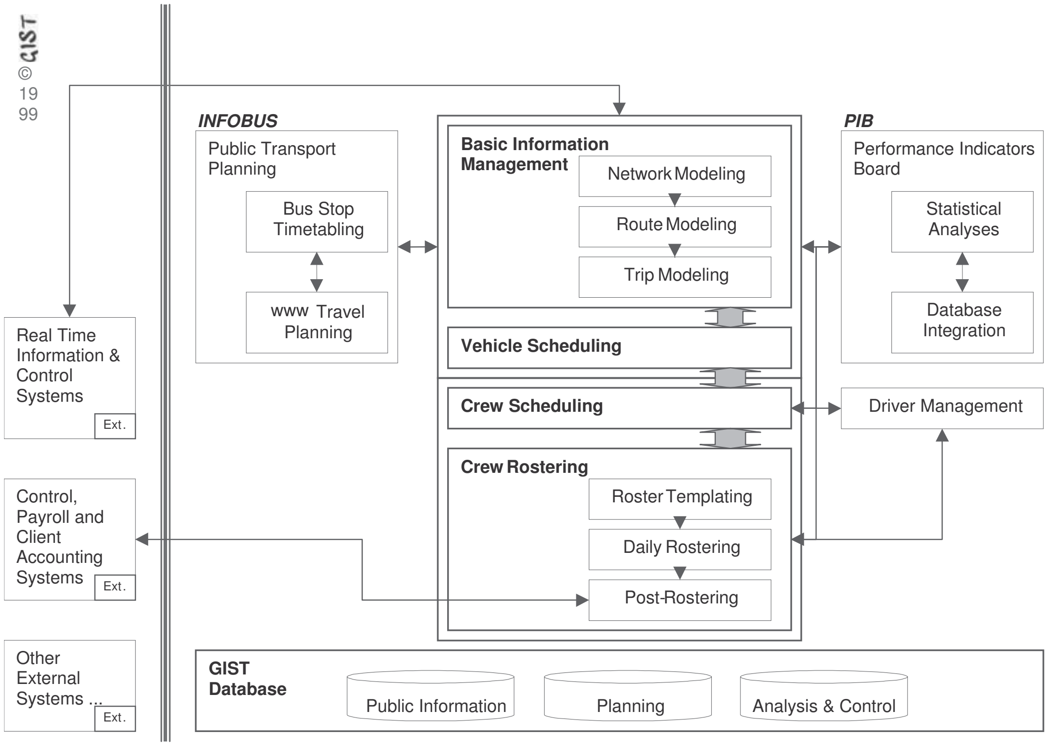

Figure 2: The GIST98/EUROBUS system architecture model

design and development, as well as the transformation and integration programs development and the Cube's definition. The last phase is the system implementation and configuration that is done in the company.

Another common aspect to the main Data Warehousing methodologies is their iterative nature. An iteration, that is every set of three phases, has between three to six months of duration, allowing a good expectation management. So, it can be a never stop process.

4.1 The methodology used to build the PIB module

In the methodology used to build the PIB module (Figure 4) it is clear the sequential structure of the iterations. The only exception is in the analysis phase. In fact, the study of the indicators that the company needs and the study of the data sources are done almost in parallel.

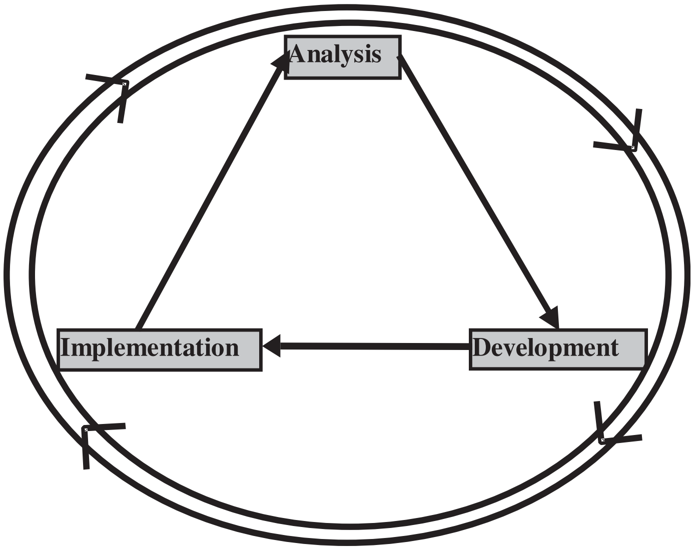

Figure 3: Data Warehousing reference methodology

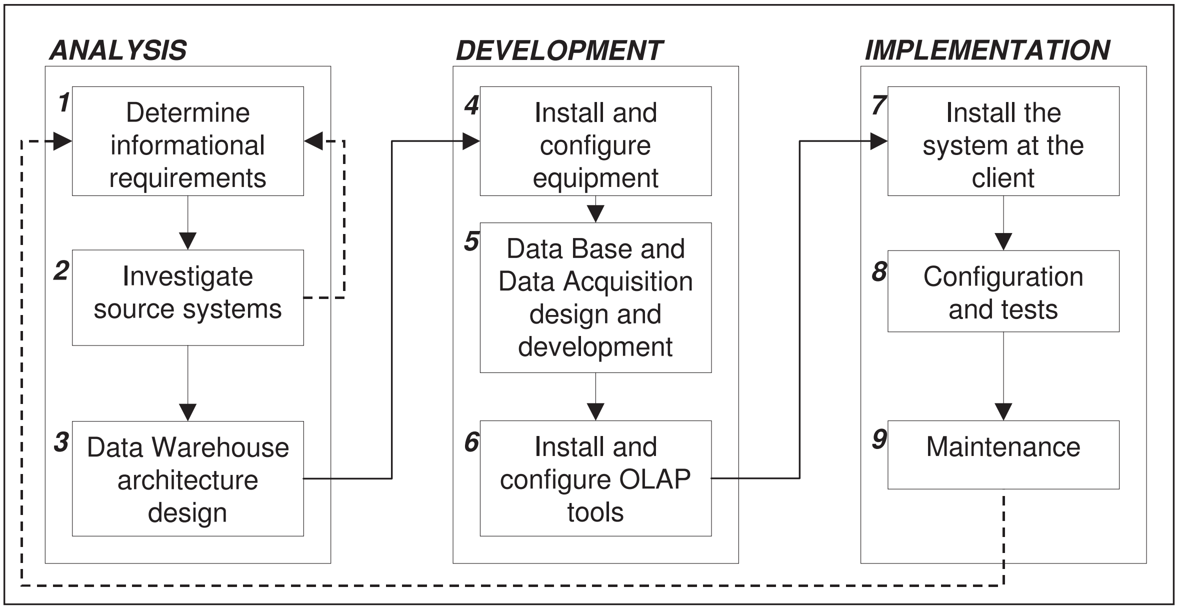
ANALYSIS
DEVELOPMENT
IMPLEMENTATION
1 Determine informational requirements
2 Investigate source systems
3 Data Warehouse architecture design
4 Install and configure equipment
5 Data Base and Data Acquisition design and development
6 Install and configure OLAP tools
7 Install the system at the client
8 Configuration and tests
9 Maintenance

Figure 4: The methodology used to build the PIB module

5 The PIB module: Analysis, Development and Implementation

By definition, a Data Warehouse integrates all the data of the company that is relevant to support management's decisions. Since the data used on PIB module is exclusively the one that comes from one or more GIST98/EUROBUS systems, we cannot say that the PIB's database is a Data Warehouse. In fact, it is more correct to define the PIB's database as an independent Data Mart. Independent because the data sources are operational systems (one or more GIST98/EUROBUS databases) and Data Mart because it is functionally structured (it uses the data relevant to the planning and control activities).

In this section it is described the analysis and development phases of the first iteration.

5.1 The analysis phase

The analysis phase has, typically, a duration of three months.

5.1.1 Informational requirements

The first questions we tried to answer were: "Which are the big groups of indicators that managers need?" and "Which indicators must be included in each group?"

The results presented in this text are derived from the study accomplished at the STCP Company (the Oporto Public Transport Authority).

"Which are the big groups of indicators that managers need?"

Taking attention to the information that GIST98 contains and to the big areas where top-level managers need planning and control information, we have defined two main groups of performance indicators: service indicators and crew indicators. The first group gives information about the level of service that the company is offering, which is especially important to the operations management department, while the crew indicators give information relevant to the human resource department.

"Which indicators must be included in each group?"

The first thing to do was to collect information from the companies. It was a long, interactive process between the analysts and the companies' staff. The service reports produced monthly by the companies were another major source of information.

Based on the information collected it was possible to notice that each potential indicator has no more than 4 parameters. This means that it is possible to identify a specific indicator by answering to 4 questions: "Which is the indicator denomination?", "Which entity does the indicator refers to?", "Which is the aggregation level used?" and "Which aggregation function does it represent (in other words, is it an average, a maximum or any other kind of function)?" As an example: if we want to know the average distance of trips per route, the indicator is the distance, the entity is the trip, the aggregation level is the route, and the aggregation function is the average. Using this methodology it was possible to identify the indicators group, the entities group and the aggregation function group. The aggregation level was more complex to define. In fact, using the example above, the trips can be aggregated by route, by line, by day type, by route and by day type at the same time, etc., i.e., they can be aggregated by one entity or by a combination of entities. The first thing we observed was that when an indicator can be aggregated by one entity it can also be aggregated by more generic entities. In the example, if routes can aggregate trips, lines can also aggregate trips (notice that a line is a set of routes). The step forward was to define the different dimensions to be used. Each dimension refers to a sequence of entities. The order by which they appear on table 1 is related with the degree of detail, i.e., the first entity of each dimension is the most generic one and the last one is the most specific. In other words, the entity referred at the first column of each dimension is a set of the entities referred at the next one, and so on.

In tables 1, 2 and 3 the results obtained by this methodological approach are presented (the aggregation functions were ignored just for the sake of simplicity).

Table 1: Dimensions

Dimensions	Levels			
Dates	Year	Quarter	Month	Day
Network Dates	Network Date			
Timetable Dates	Timetable Date			
Companies	Company			
Year Seasons	Year Season			
Depot	Depot			
Free Days Groups	Free Days Groups			
Periods of the day	Period of the day			
Network	Line	Route		
Day Types	Day Type			
Route Types	Route Type			
Network Types	Network Type			
Situation Types	Situation Type			
Trip Types	Trip Type			
Crew	Category	Crew		
Vehicle	Logic Vehicle	Stretch		

With these tables we can easily identify four groups of indicators: the route extension and the number of routes in the first group of indicators, the number of logic vehicles in the second group, all the others indicators from table 2, in the third group, and the indicators presented in table 3, in the fourth group. The identification of these groups of indicators derives from the indicator's entity and from the set of dimensions used. Each group will be implemented using the Cube concept. Once the cubes are defined tables 4 and 5 are used in substitution of tables 2 and 3 because they present the same information in a more concise and practical way.

Table 2: Service Indicators

	Dimensions												
Service Indicators	Companies	Network Types	Network	Network Dates	Timetable Dates	Year Seasons	Depots	Periods of the Day	Dates	Route Types	Day Types	Trip Types	Vehicle
Route Extension	x	x	x	x						x			
Number of Routes	x	x	x	x						x			
Number of Trips	x	x	x		x	x	x		x		x	x	x
Trip Distance	x	x	x		x	x	x		x		x	x	x
Trip Driving Time	x	x	x		x	x	x		x		x	x	x
Trip Driving Time with Support Time	x	x	x		x	x	x		x		x	x	x
Trip Commercial Speed	x	x	x		x	x	x		x		x	x	x
Trip Exploration Speed	x	x	x		x	x	x		x		x	x	x
Number of Logic Vehicles	x	x	x		x	x	x	x	x		x	x	x

Table 3: Crew Indicators

	Dimensions						
Crew Indicators	Companies	Depots	Situation Types	Free Days Groups	Network Types	Crew Crew	Dates Dates
Number of Drivers	x	x	x	x	x	x	x
Number of Planned Situations	x	x	x	x	x	x	x
Number of Real Situations	x	x	x	x	x	x	x
Driver Planned Actual Duration	x	x	x	x	x	x	x
Driver Planned Extra Duration	x	x	x	x	x	x	
Driver Planned Night Duration	x	x	x	x	x	x	x
Driver Real Actual Duration	x	x	x	x	x	x	x
Driver Real Extra Duration	x	x	x	x	x	x	x
Driver Real Night Duration	x	x	x	x	x	x	x

Table 4: Dimensions by Cube

Cubes Dimensions	Commercial Network	Vehicle's Services	Vehicle's Services along the day	Crew
Dates	x	x	x	
Network Dates	x			
Timetable Dates		x	x	
Companies	x	x	x	x
Year Seasons		x	x	
Depots		x	x	x
Free Days Groups				x
Periods of the Day			x	
Network	x	x	x	
Day Types		x	x	
Route Types	x			
Net Types	x	x	x	x
Situation Types				x
Trip Types		x	x	
Crew				x
Vehicle		x	x	

Table 5: Indicators by Cube

Commercial Network	Vehicle's Services	Vehicle's Services along the day	Crew
Route Extension	Number of Trips	Number of Logic Vehicles	Number of Drivers
Number of Routes	Trip Distance		Number of Planned Situations
	Trip Driving Time		Number of Real Situations
	Trip Driving Time with Support Time		Driver Planned Actual Duration
	Trip Commercial Speed		Driver Planned Extra Duration
	Trip Exploration Speed		Driver Planned Night Duration
			Driver Real Actual Duration
			Driver Real Extra Duration
			Driver Real Night Duration

5.1.2 Source systems

In order to know how to obtain the identified indicators from the available data, maps that relate the sources (modules) with the dimensions and indicators have been done (table 6).

The "New" column exists because there is some information that is created during the data acquisition process. An example is the company owner of each source. That information is implicit to each source but it must be explicit in the Data Warehouse.

In order to obtain that information with more detail, tables have been created that describe, for each level of every dimension and for each indicator of every Cube, which are the tables from the sources that are needed to obtain them. Table 7 shows this information for the different levels of each dimension.

5.1.3 Data Warehouse architecture design

To define the Data Warehouse architecture the first step is to preview the database dimension. Secondly, it is necessary to choose the software tools that will be evaluated. Once the evaluation is done, the architecture is established.

To preview the database dimension it was assumed the use of the star schema [11] since it is the most space consuming database design used for these proposals. Since the PIB module is part of the GIST98/EUROBUS system, the scalability problem is a minor one. The formula used to preview the database space was:

fmidx x *fmdiv* x *(ech* x *ndim* + *eind* x *nind)* x *nregmax* x *fmgra (1)*

Table 6: Indicators by Cube

From To	New	Network	Gist-Lines	Trip & Vehicle Scheduling	Crew Management & Rostering
Dates	x				x
Network Dates	x				
Timetable Dates				x	
Companies	x				
Year Seasons		x			
Depots					x
Free Days Groups					x
Periods of the Day			x		
Network			x		
Day Types	x				
Route Types		x			
Net Types	x				
Situation Types					x
Trip Types				x	
Crew					x
Vehicle				x	
Commercial Network Indicators		x	x		
Vehicle's Services Indicators		x	x	x	
Vehicle's Services along the day Indicators		x	x	x	
Crew Indicators					x

Table 7: Data sources for each level of each dimension

Dimension →Level	Data source
Dates	Data acquisition process and *Shifts_schedules*; *Daily_rosters*
Network Dates →Network Date	Data acquisition process
Timetable Dates →Timetable Date	Data acquisition process
Companies →Company	Direct introduction
Year Sessions →Year Session	*Year_seasons*
Depots →Depot	*Depots*
Free Days Groups →Free Days group	*Freeday_groups*
Periods of the Day →Period of the Day	*Periods*
Network →Line	*Lines*
Network →Route	*Paths* e *Lines_paths*
Day Types →Day Type	*Day_types*
Route Types →Route Type	Direct introduction
Net Types →Net Type	Direct introduction
Situation Types →Situation Type	*Situation_types*
Trip Types →Trip Type	Direct introduction
Crew →Category	*Categories*
Crew →Crew	*Crew_members*
Vehicle →Logic Vehicle	*Shifts_schedules*, *Shifts* e *Trips*
Vehicle →Stretch	*Shifts_schedules*, *Shifts*, *Trips*, *Trips_nodes*, *Glines_nodes_types* e *Nodes_types*

where:

fmidx – a factor to preview the space used by the indexes: 2.

fmdiv – a factor to preview the space used by metadata, or other no counted space: 1,5.

ech – size of each primary key field: 6 *Bytes*.

ndim – number of independent dimensions: marked with **X** on table 8.

eind – size of each indicator field: 4 *Bytes*.

nind – number of the Cube's indicator: see table 5.

nregmax – number of records (it was used the data from CARRIS - the biggest company): it is the Cartesian product of the number of members of the independent dimension's lowest levels (marked with **X** on table 8).

fmgra – is a granularity factor. It is the average of records' number needed for each value of *nregmax*. It is assumed that the data detail level is already defined. It is a critical decision once it is a compromise between the data space and the information detail level. For example, to the vehicle's services Cube, it is important to save the trips information because, even if indicators are not aggregated by trip, this information allows, in the future, the calculation of other indicators with a low cost. The average of trips by vehicle's stretches is around 15. This is the *fmgra* to the vehicle's services Cube. To the other Cubes the *fmgra* will be 1.

The values of *fmidx* and *fmdiv* will be studied using a more reliable sample.

Looking to table 8 it can be seen that around 80% of the database dimension is needed to obtain the indicator 'Number of logic vehicles'. It is possible to get this indicator by calculus once the beginning and the end of each stretch are stored. With this change, the database dimension for a year will be around 1,9 Giga Bytes. For three years – the time considered as necessary by the companies to keep the information – it will be around 5,7 Giga Bytes.

The software tools chosen to be evaluated were: *MicroStrategy 7* (a reference between the ones that use the ROLAP architecture); *Express* Server from *Oracle Corporation* (a reference between the ones that use the MOLAP architecture); i*TM1* from *Applix, Inc* (a very versatile tool); and *SQL Server7/OLAP Services* from *Oracle Corporation* (a tool with a very competitive ratio price/quality).

The weights used for each criteria were the ones that seem to be more adequate for this particular case (see table 9). The points were based in some tests and mainly in [10]. *MicroStrategy7* was classified with 8 in the 'Database management system robustness' despite it has not an own Database management system. That classification is the one attributed to *Oracle8 Enterprise Edition* once it is the Database management system used by all the other modules of the GIST98/EUROBUS system and it would be naturally the one to use with the *MicroStrategy 7* option.

Figure 5 presents the chosen architecture. An important feature of *SQL server 7' OLAP services* is the possibility to choose between ROLAP, MOLAP or HOLAP architecture. The *DTS – Data Transformation Services*, the *SQL Server 7*, and the

Table 8: Prevision study of the database dimension for a year

Dimensions	Number	Commercial Network	Vehicle's Services	Vehicle's Services along the day	Crew
Dates	365		**X**	**X**	**X**
Network Dates	10	**X**			
Timetable Dates	10		x	x	
Companies	1	**X**	**X**	**X**	**X**
Year Seasons	3		x	x	
Depots	6		x	x	x
Free Days Groups	10				x
Periods of the Day	96			**X**	
Network	350	**X**	x	x	
Day Types	3		x	x	
Route Types	2	x			
Net Types	1	**X**	**X**	**X**	**X**
Situation Types	100				x
Trip Types	2		x	x	
Crew	2500				**X**
Vehicle	2000		**X**	**X**	
	Number of records	3.500	10.950.000	70.080.000	912.500
	Space size (MB)	0,32	1.503,75	6.817,02	146,20

Table 9: OLAP tools evaluation

	Weight	MicroStrategy 7	Express	iTM1	SQL Server / OLAP Services
Capacity	10	10	8	6	9
Database structural simplicity	50	4	5	9	8
Database structural flexibility and calculation functionality	50	7	9	7	9,5
Data access functionality	70	7	10	7	9
Technical support and documentation	50	7	9	7	7
Database management system robustness	70	n.a. (8)	8	8	10
Decision support tools	70	6	8	6	9
Price	90	6	2	5	10
	4600	**3010**	**3230**	**3130**	**4175**
Classification (%)		**65,43%**	**70,22%**	**68,04%**	**90,76%**

n.a. - not appliable because it has not a own database management system

Ponts from 1 to 10

OLAP services are all part of the *SQL Server / OLAP Services* option. The *Knosys ProClarity 2.0* is an easy to use front-end tool as the quite known *Microsoft Excel 2000*.

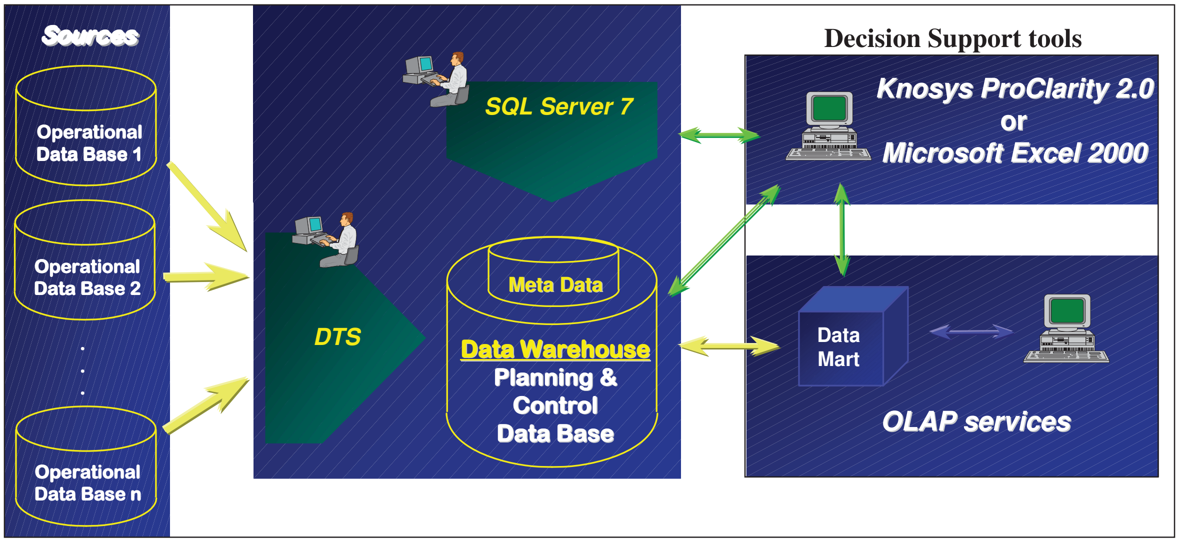

Figure 5: PIB architecture

5.2 The development phase

The development phase has, typically, a duration of three months.

5.2.1 Install and configure equipment

It is a technical work without any special difficulty.

5.2.2 Data Base and data acquisition design and development

In the development phase of a Data Warehousing system there are three main steps to consider [14]:

- The database design: As shown in table 3 and figure 6, the data is stored according to the indicators and dimensions required. The data is pre-processed before the user can access it; so, the concern with efficiency when designing the database has impact, especially on the processing time. It has no visible impact on the efficiency as seen by the end-user. The efficiency for the end-user is mainly determined by the architecture used. The ROLAP architecture is more scalable (manages more data) but is slower while MOLAP architecture is less scalable but faster. The PIB module uses the HOLAP architecture. Figure

6 shows the database design for the crew indicators. As it can be seen, it is directly obtained from tables 1 and 3. The database design was done according to the dimensional model [7].

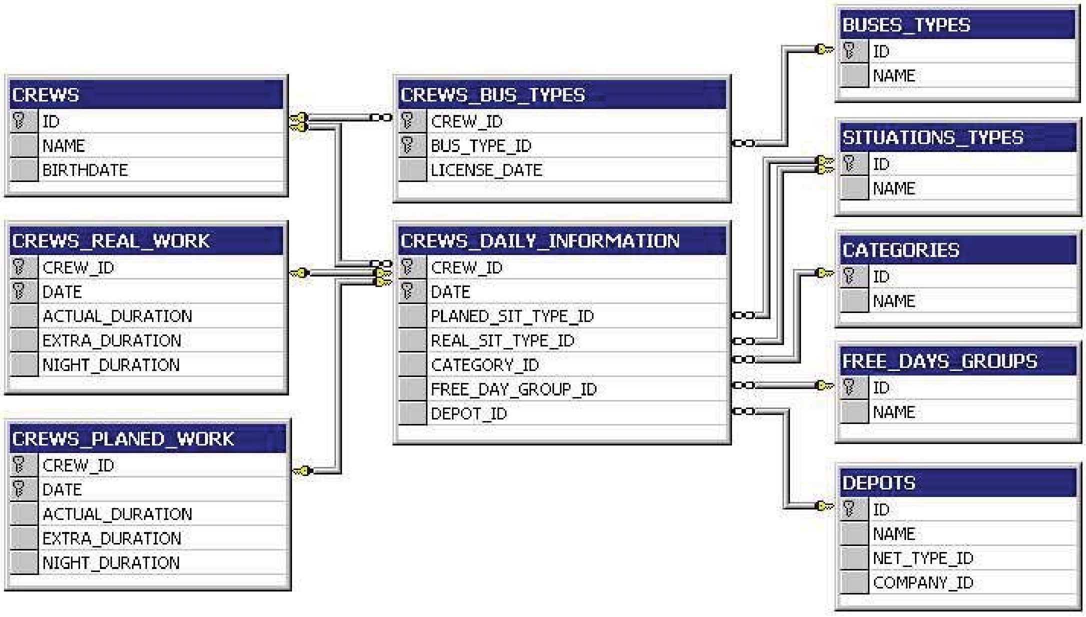

Figure 6: database design for the crew indicators

- The transformation and integration programs [16]: The development of all these programs takes typically around 80% of all the development effort. It includes the extraction, transformation, transport and load processes. Previously, it is important to analyse the source data and clean it. This is a very important issue because without reliable data it is not possible to guarantee the results. The source of information can be multiple: operational databases, other Data Warehouses/Marts or external data. In PIB's case, the source data can be one or more GIST98/EUROBUS databases. The execution of these programs is done every night assuring that the Data Warehouse is always up to date. This step was done using the *DTS – Data Transformation Services*.

- Cubes definition [15]: A cube is a structure that can be seen as a huge table (fact table) where the primary key is the set of all dimensions' identifiers and the attributes are the indicators. Taking attention to tables 1, 4 and 5, it is possible to define the cubes easily. The first thing to do is to define the dimensions as in table 1. The second step is to implement the cubes defining the indicators and selecting the related dimensions. This step was done using the *OLAP Services.*

5.2.3 Install and configure OLAP tools

Finally, how can the information be useful to the decision making process? There are several non-expensive analysis and reporting tools that allow the user to do all kind of analysis, just executing basic commands with the mouse. In figures 7 and 8 some examples using the Knosys – ProClarity 2.0 are presented. They show two specific types of analysis. Figure 7 shows the decomposition tree with the information about the distance planned to the day 28-05-1999 with different levels of detail: by net type; by logic vehicle; by stretch; by travel's type and by route. Figure 8 shows the relationship between the planned distance and the number of trips by logic vehicle for the day 28-05-1999.

Using this tool it is possible to do much more types of analysis such as, to compare indicators for the same month of consecutive years, or for consecutive months, to see line, column, pie, bar or other chart types, to create other indicators using formulae, etc. All this may be done using the aggregation level chosen by the user. To aggregate or detail data, it is enough to drill up or down.

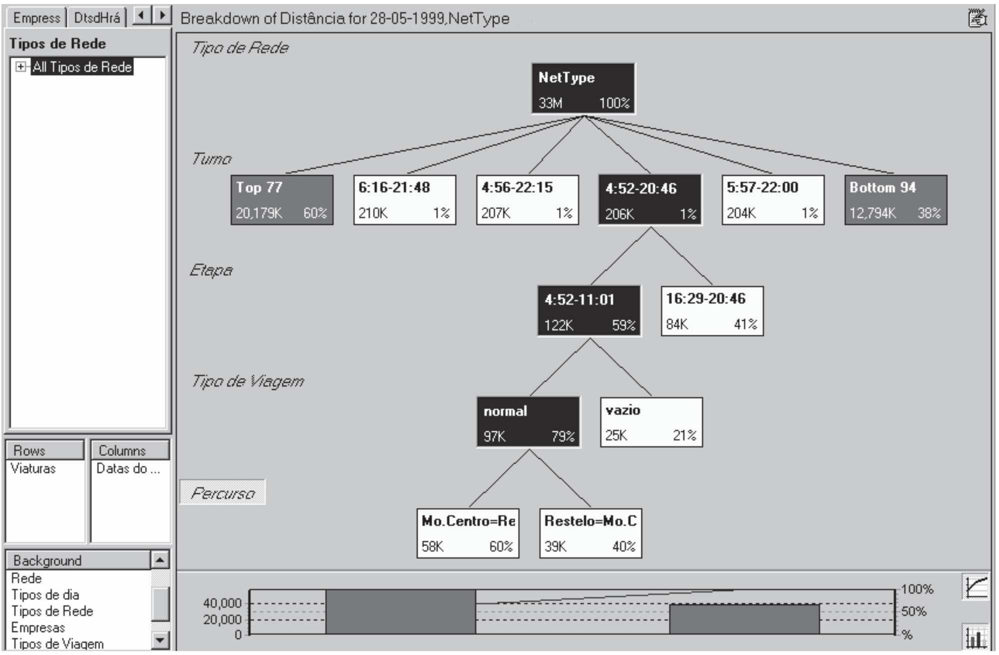

Figure 7: Decomposition Tree

5.3 Some considerations about the implementation phase

The implementation phase has a set of tasks that are done once and other tasks that must be done periodically. In the first group are:

- To install and configure hardware and software at the client: like the install and configure equipment task, it is a technical work without any special difficulty.

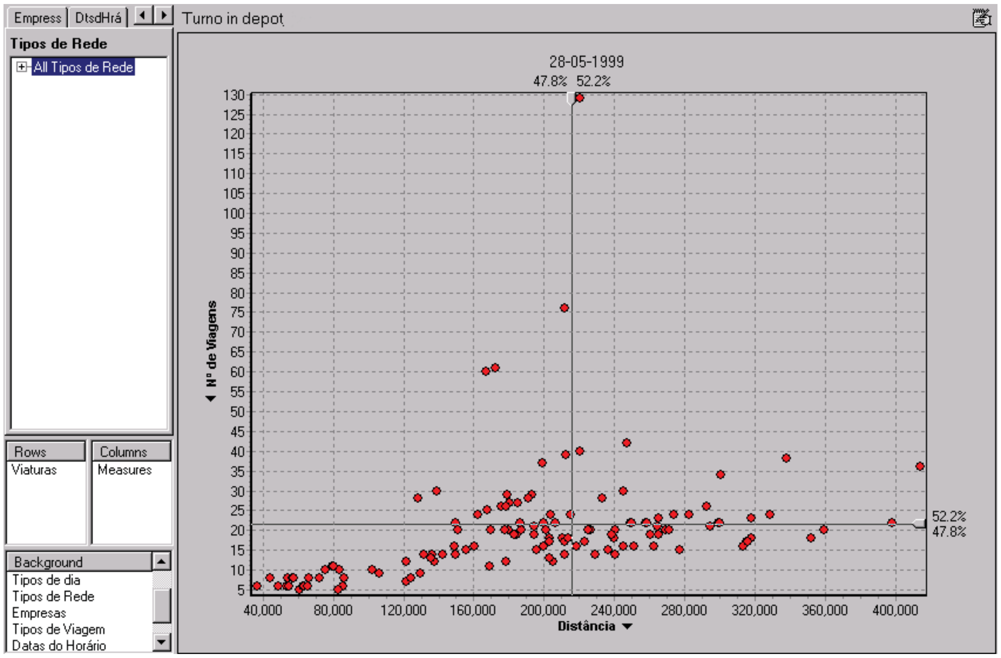

Figure 8: Perspective chart

- To create database, users and access policy: it is a typical task for a database manager.
- To verify and configure OLAP Services: when the OLAP Services will be tested with real data.
- To create a backup and data acquisition policy: it is a very important issue where the acquisition and backup periodicities are defined. The data acquisition policy depends on the aggregation level previously chosen. The backup policy is a basic task for every database system.

In the second group are:

- To verify log files: it is a regular task to do at least after every acquisition procedure since it is, usually, the main source of errors.
- To verify indicators evolution along time to detect eventual data errors: it is important to control the indicators evolution because there are several possible errors that are difficult to detect by software.

6 Conclusions

There are three main reasons to implement a Data Warehousing system [9]:

- There is the perception that the information exists but it is not so useful as it could be;
    - Each department has its own language and communication is difficult;
    - The reports production is expensive and inefficient.

To transform data in information is the main goal of a Data Warehousing system. Traditionally, when a manager wants information he asks it to the information department in order to get a report. If the manager wants additional information he needs to wait longer. A Data Warehousing system lets the manager free to pick up the information he wants, with the type of display he prefers, just using the mouse. This flexibility leads to a higher concentration on information instead on data, giving more time to analyse information. The report production is also reduced since time is not spent doing queries to the database. Another important feature is that all the information is centralised, with everybody using the same indicators with the same definitions, which makes communication easier. But the main advantage of the Data Warehousing systems is to turn data into a competitive advantage because it can be used to analyse relationships between variables, to analyse trends, in a word, to put the information at the service of the decision making process.

7 References

[1] Sid Adelman, Joe Oates, *Data Warehouse Project Management,* Data Management Review, May 1998

[2] João Falcão e Cunha, Jorge Pinho de Sousa, Teresa Galvão, José Luis Borges.*, A Decision Support System for the Operational Planning at Mass Transport Companies*, presented at the 6th International Workshop on Computer-Aided Scheduling of Public Transport, July 1993, Lisbon, Portugal.

[3] C. J. Date, *An Introduction to Database Systems (6th Edition),* Addison-Wesley, 1995

[4] Stephen R. Gardner, *Building the Data Warehouse*, Communications of the ACM, September 1998

[5] Hadden & Company, http://www.hadden-kelly.com, 1999

[6] Inmon W.H., *Building the Data Warehouse*, Wiley, 1996, pp.33

[7] Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite, *The Data Warehouse Lifecycle Toolkit*, Wiley, 1998, pp. 137-314

[8] OLAP Council, http://www.olapcouncil.org/research/glossaryly.htm, 1997

[9] John Onder, Todd Nash, *Building a business-driven Data Warehouse*, Data Management Review, October 1998

[10] Nigel Pendse, Richard Creeth, *The OLAP Report*, http://www.olapreport.com*, red in October 2000*

[11] Neil Raden, *Data Modeling – Star Schema 101*, http://www.archerdecision.com, 1995/96

[12] João Ranito, seminar presentation at Faculdade de Engenharia da Universidade do Porto, 1999.

[13] Jorge Pinho de Sousa, Jorge Freire de Sousa and Rui Campos Guimarães, *Un système informatique d'aide à la génération d'horaires de bus et de chauffeurs, in: Gestion de l'économie et de l'entreprise - l'approche quantitative*, Editions CORE, Série Balises, De Boeck, Brussels, 1988, pp. 477-492 (in French).

[14] The Data Warehousing Institute, *1998 Data Warehousing Buyer's Guide*, http://www.dwinstitute.com/buyersguide/, 1998

[15] Erik Thomsen, *OLAP Solutions*, Wiley, 1997, pp. 229-264

[16] Karen Watterson, *Attention, Data-Mart shoppers*, BYTE Magazine - July 1997

[17] J. D. Welch, http://www.datawing.com/BOOKS_F.HTM, 1998