A Data Warehouse Development Model based on Mining, Decision and Operational Process Requirements

1. Introduction

In order to support the decision processes of an organization the main component is the Data Warehouse (DW). Inmon, W.H. (1996) defines a DW as a subject oriented, integrated, non volatile and time variant collection of data in support of management's decision. Kimball, R. (1998) defines a DW as the queryable source of data in the enterprise. Therefore, both authors define the DW as a data warehouse of all organizations whose primary purpose is to support the organization decision processes.

Associated to the DW concept there is the Data Mart (DM) concept. From a logical point of view, Kimball defines a DM as a subset of the DW oriented to the needs and requirements of an area, section or department of an organization. Then, it is logically viable to consider that a DW is made up of the union of all its DMs. On the other hand, Inmon, who focused on a physical perspective, considers the DMs components emanating from the DW. That is, the DW serves the DMs as a feeder system.

Besides these really different points of view as regards the DW-DMs relationship, the main inconvenience is that the development of DW-DMs is a very expensive project that takes a long time (Asen and Jacob, 1998), and it has to overcome resistance in the organization (Watson and Holey, 1998). For this reason, organizations have followed different strategies in the development of DW-DMs. In connection to this, Firestone, J. (1997) defines three informal patterns or models for the development of DW-DMs: the top down model, the bottom up model and the parallel development model. We will briefly describe each of them in the following section.

The three development models essentially take the same starting point: the fact that data are in the organization, in legacy systems and operational data bases, and they pose the project for the development of DW-DMs starting from these data. This generates a classic stage of every DW-DMs development project that Kimbal, R. (1998) calls Data Staging area, which involves the data extraction, translation, filtering and integration process. This stage is very expensive, requires a long time and great effort, and constitutes an important area to be considered from the academic point of view (Widom, J., 1995).

Another problem that these development models have in common is that many the data required for decision processes are not in the operational data bases and they are not supplied by the legacy systems.

To sum up, in these development models we can point out two significant problems: collection of scattered data in the organization and the lack of data.

In this work, we present a non-conventional DW-DMs development model. We have approached this development model with the aim of solving the two aforementioned problems posed by the conventional DW-DMs development strategies. In the first section, we will describe the main informal models conventionally used in the development of DW-DMs and then we will describe our DW development model. In the second section, we will describe the use of this model to design a DW within the framework of the SPU project, (1998) for developing a SSD for two academic institutions.

2. The data warehouse-data marts conventional development models

In the introduction, we have mentioned the Firestone (1997) proposal that states that if we analyse the different DW-DMs development projects carried out by organizations, we can define at least three informal development models, each being based on a substantially different development strategy. These are: the top down model, the bottom up model and the parallel development model. Following, we will briefly describe each of them.

The top down model: In this model the DW is developed from the source systems (legacy systems, operational data bases, etc.) though an appropriate integration process that executes the following operations: extracts, cleans, transforms, combines, removes duplication and stores data at the DW. In this way, the DW integrates in an appropriate format all sources of an organization's data that can be used to support the organization's decision processes. After the DW is implemented, the DMs are developed. That is, the DMs are derived from the DW. We graphically represent this model in Figure 1(a).

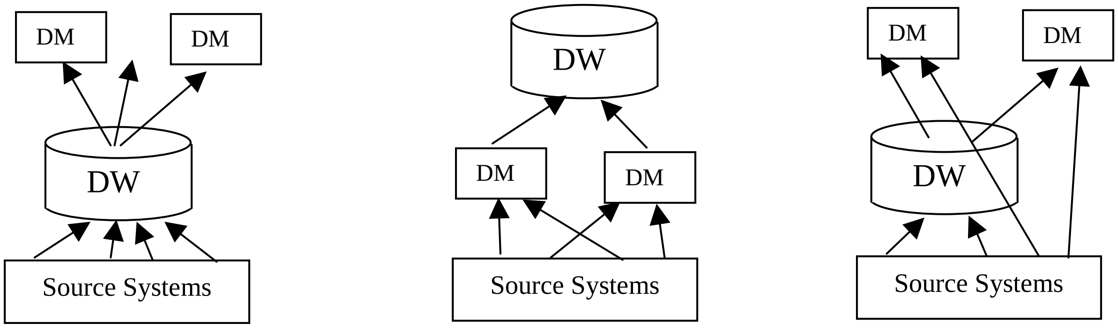

Figure 1: *(a) The top down development model, (b) The bottom up development model, (c) The parallel development model*

The bottom up model: In this model, the DMs are independently designed and implemented from the source systems. This policy results in the development of a set of DMs with non-conformed data structures, containing redundancy of data and possibly important information gaps from an organization's point of view. In other words, we get through a set of disintegrated data sources to another set of data sources that remain disintegrated from the global organization point of view. This creates the need of developing a data warehouse in common, which brings about the need to develop a DW from the DMs. The need for an appropriate integration process remains, since the DMs can keep so many incompatibilities among them just as the ones among the source systems. It is clear that in this development model we cannot consider the DW as the joint of the DMs, nor can we consider the DMs as subsets of the DW. Obviously, this development pattern can lead to catastrophic results. This model is represented in Figure 1(b).

The parallel development model: The model proposes to define a data model for the DW, which will be used as a guide for the development of DMs, which initially operate independently, as long as the DW, whose development is carried out in parallel, is concluded. The main advantage of this model is the fact that it solves the posed problem top down model, since it allows the satisfaction of the short-term decision support needs and it also allows the capitalization of experiences for the DW development. This is graphically shown in Figure 1(c).

A common characteristic of these DW-DMs development models is that they basically assume that data are in some place of the organization (in legacy systems, operational data base, etc) and they bring up the DW-DMs development project starting from this assumption. Conceptually, we could say that the paradigm of these models is to integrate existing data. As a consequence of this paradigm, a classic stage appears that is very important in every DW-DMs development project, which tackle the process of data extraction, translation, filtering and integration. It is an expensive stage that demands a long time and great effort.

Another direct consequence of this paradigm is that having all the data of the organization correctly integrated, does not mean that all the data required by the decision processes are in the DW-DMs. This is generally so because they are not in the operational data bases, they are not supplied by the legacy systems and their capture was not foreseen. Thus, if the DW development project is limited to this paradigm, we will have a data warehouse from which it may be not possible to generate all the information that the decision-maker requires. This lack could be more noticeable when trying to carry out data mining processes.

At the time of deciding DW-DMs development strategy for both academic institutions, within the framework of the SPU project, we have specially emphasized the consequences of support these development paradigm. Therefore, with a future view, we can count on the political will of the board of directors at these academic institutions to support a development model that differs from the previously described models since its aim is to solve these problems.

Conceptually, we have committed ourselves to changing the previous paradigm by a new DW-DMs development paradigm, which proposes to integrate the required data in the organization. The problem of this paradigm lies on the fact that it is necessary to determine, at the design time, the data required today and data that will be required in the future. Specially taking into account the current dynamics of the organizations decision processes.

Taking this paradigm as a goal we have proposed a DW-DMs development model in the framework of the SPU project, which we call global development model and that will be described in the following paragraph.

3. The global development model

The model is based on designing DW-DMs together with new operational systems to replace the current ones. This strategy involves two main stages: Data Structure Design and Implementation.
Figure 2 graphically shows the stage of the global data structure design of the operational systems and DW-DMs.

At this stage, the model brings about a flow of requirements resulting from three substantially different activities of the organization: the operational processes, the decision processes and the planned mining activity. From this flow of requirements we define the data structure design of the operational systems and the DW-DMs.

The operational requirements are defined from the “business” processes of the organization by means of functional and non-functional requirement analysis.

The information requirements to give support to the decision-makers of the different sectors and levels of the organization are defined by means of the analysis of current and future decision scenarios.

Finally, a third level of requirements is approached in our model, whose aim is to determine a set of possible patterns of behaviours that would be interesting to be analysed for the purpose of inferring possible data requirements by this activity.

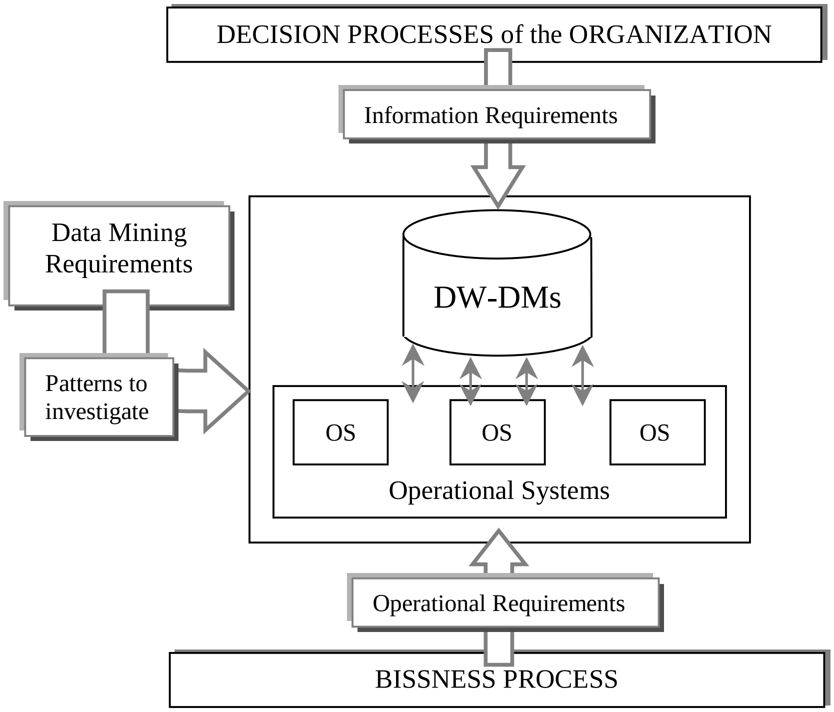

Figure 2: *The global development model*

Once the data structure is designed, the stage that follows is the implementation. For this, we could use an implementation strategy similar to the one proposed by the top down model or to the one proposed by the parallel development model, which were previously described.

With this development model, we have significantly simplified tasks and therefore, the respective process algorithms to integrate the data from the operative data base into the DW-DMs is also simpler. Moreover, it is normally expected that the decision-makers of the organization, as well as who carry out mining activities, can satisfy better their needs for information.

Finally, it is normally expected that the data architecture designed for both operational systems and DW-DMs can make it easier not only the maintenance process but also the evolution process that naturally demands the current dynamics of the organizations.

4. Application of the global development model

UTN-Concepción del Uruguay and *UTN-Santa Fe* are academic institutional branches of the National Technological University (UTN), whose goal is human resource education at bachelor and graduate level, appropriately suited to the environment requirements. These institutions have initiated a continuous improvement process with the purpose of improving both their efficiency and efficacy. This has resulted in the need of relying on an appropriate support to generate complex analytical information. For example, search for behavior patterns, future scenario analysis and multidimensional analysis of key indexes to evaluate the evolution of the improvement process. For this reason, the academic houses have initiated a contiguous project (SPU Project, 1998) to develop a Decision Support System.

In this section we will describe some details for applying the global development model to the DW-DMs design project to support the management of Academic Institutions within the framework of the SPU project.

The project tasks have been grouped into two main sets: Analysis and design of six new operational systems based on modern information technology to replace the current ones. Six work teams have been in charge of these tasks, each team being integrated by both system analysts and engineers. The second task group of the project includes the analysis and design of a DSS to support the management of the academic institutions. A team of researchers of GIDSATD (group of DSS research and development - UTN) has been in charge of these tasks.

The information requirements of decision processes were essentially determined from three main sources:

- Definition of current and future decision scenarios through group and individual work meetings with staff from every level and hierarchy at the academic institution.
- More than sixty indexes from institutional and academic performance evaluation, classified in seven main areas: academic; research and development; relationship with the surrounding environment (in the social, cultural and economic-productive aspects); equipment (library, laboratories, etc.); students welfare (sports, fellowships, medical assistance, etc.); financial administration; infrastructure. These indexes were defined in terms of the experience gained along ten years implementing the continuous improvement process within the framework of the quality program that the academic institution UTN-FRCU is carrying out.
- A set of medium and short term forecasting models for the planning activity of the academic units. These models define causal relationships among variables.

Data mining requirements: they were "inferred" from specifications made by different people at the institution, who stated their interest in searching certain behaviour patterns. For example: the student academic evolution vs. The student origin characteristics (economical, social, cultural, geographical ones, high school, etc.) vs. extra activities (sports, research, work practice, etc). Students that are more liable to quit studying. Factors that improve the student academic performance.

The operational requirements: six operational systems have been defined to support the activities involved in all operational processes of the academic houses.

Each module was assigned to a work team. The requirements have been defined by means of functional and non-functional analyses of the module.

4.1. DW-DMs data model

As regards the operational systems data model, we can point out the following outstanding characteristics:

- It contains all the data required to satisfy the needs of all the operative activities of the organization.
- It includes all the defined data to meet the information requirements of the decision processes, not defined by the operational requirements.

- In order to make the mining process easier, a field is defined for each elemental data. As an example we can mention the guidelines recommended by Kimbal, (1998) for structuring a client's information.

For the DW-DMs data structure design, we have used a dimensional modelling strategy, defining a star-like data structure with explicit attribute relationship hierarchies.

The primitive data (not aggregated) of the DW-DMs are organized in a finite set *FT* of fact tables and a finite set *DT* of dimension tables. The primitive data structure is used as a basis to define a set of aggregation hierarchies.

For each fact table $F \in FT$ associated to a finite set of dimension tables $D=\{d_j\} \subset DF$ we define a finite set of aggregation hierarchies $H=\{h_i\}$.

The aggregation hierarchy $h_i \in H$ represents a data aggregation from the fact table *F* in respect of a subset of dimension tables $D(h_i) \subset D$ and has a finite set of aggregation levels $L(h_i)=\{l_{i,k}\}$ so that, the following properties are verified:

- h_i and $h_j \in H$ $\forall i \neq j$ are two different hierarchies.
- each $l_{i,k} \in L(h_i)$ has all some dimensions as $h_{i.}$
- each $l_{i,k} \in L(h_i)$ satisfies a growth criteria by aggregating one or more dimensions $d_i \in D(h_i)$ in respect of the preceding aggregation level $l_{i,k-1} \in L(h_i)$, and maintaining all the other dimensions with the same aggregation level.
- each $l_{i,k} \in L(h_i)$ has a new fact table $F(l_{i,k})$ generated from the fact table $F(l_{i,k-1})$ following the aggregations of $l_{i,k}$. Then, $F(l_{i,k})$ has associated a new set of dimension tables $D(l_{i,k})$ containing the dimension tables generated from the aggregations and shares with the preceding level $l_{i,k-1}$ the other dimension tables that have not be aggregated.
- Each hierarchy $h_i \in H$ defines its first aggregation level $l_{i,1}$ from the primitive fact table *F*, then $F(l_{i,1})$ is generated from *F* following the aggregation of $l_{i,1}$.

The kind of multidimensional dimension that we want to observe defines the *hi* aggregation dimensions. Let us consider the following example:

Consider the fact table, shown in Figure 3, that defines the student academic yield in terms of the dimension tables: time, student, academic_position, course, high_school, student_household and student_sport, fellowship. We can define for example the aggregation hierarchy:

hi = aggregation of the student academic yield based on course, academic_position and high_school dimension tables, with the following aggregation levels:

- $l_{i,1}$: (course, academic_position, high_school) = (area, academic_level, maen_mark_mathematics).
- $l_{i,2}$: (course, academic_position, high_school)= (area, academic_name, maen_mark_mathematics).

Then, in this generic way we have represented the data model that we have used to design our DW-DMs.

5. Conclusions

Using the global development model presented in this work, we have designed a DW-DMs together with a new set of operational systems to replace the current ones based on requirements defined by operational processes, decision processes and planned data mining activities.

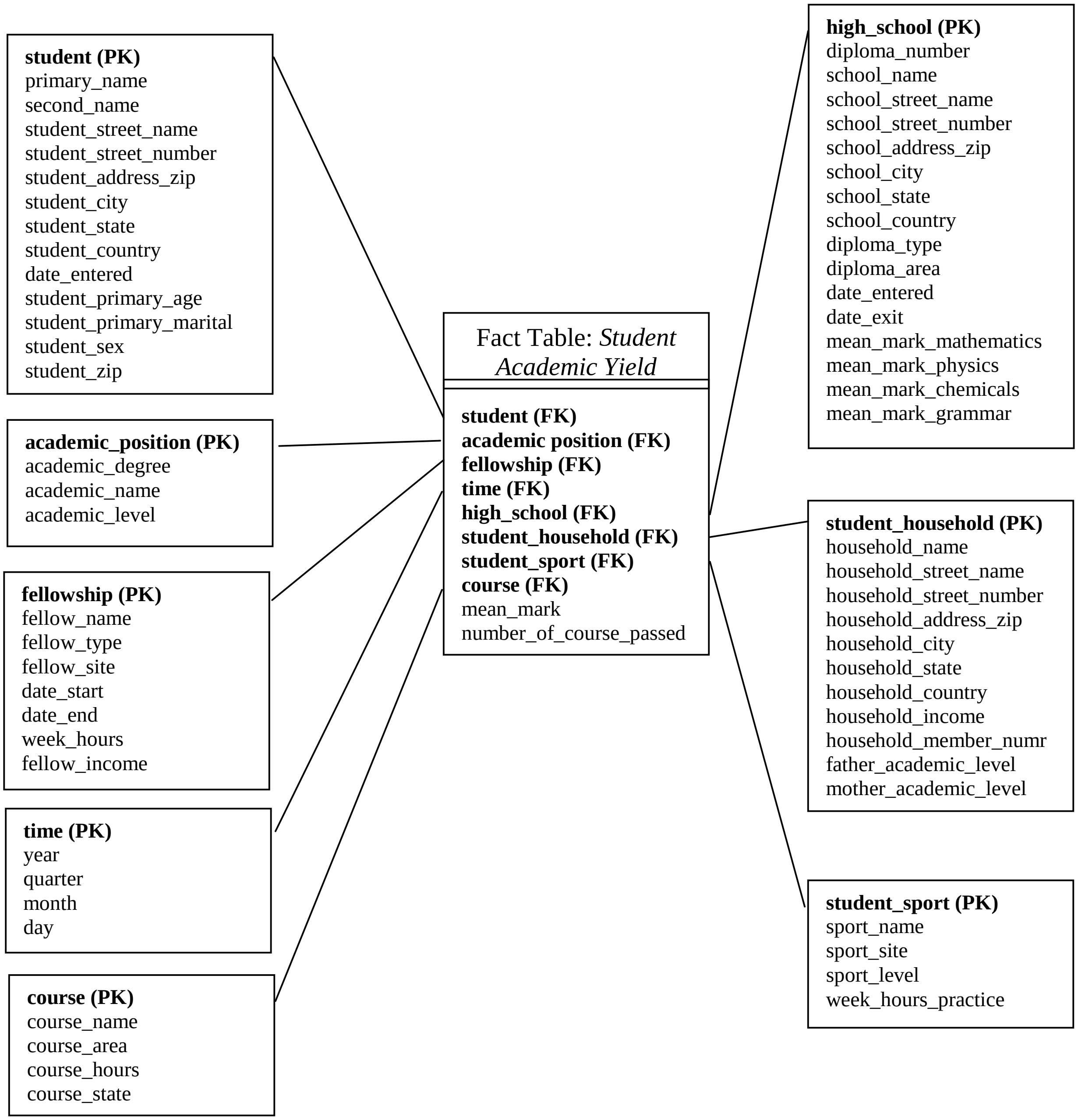

Figure 3: *Fact Table: student academic yield*

As the main advantages of this strategy, we can highlight the following: Firstly, with this development model, we have significantly simplified the tasks and therefore, the respective process of extraction, translation, filtering and integration of data from the operative data base into the DW-DMs. Secondly, it is normally expected that the decision-makers of the organization, as well as who carry out mining activities, can satisfy better their needs for information. Third, it is normally expected that the data architecture designed for both operational systems and DW-DMs can make it easier not only the maintenance process but also the evolution process that naturally demands the current dynamics of the organizations.

A drawback to consider is that this model implies a titanic task of analysis and design. A numerous work team is required to execute this task. In the SPU project, a work team of twenty-five professionals has worked for a year in the academic institution UTN-FRSF and a similar work team has simultaneously worked in the academic institution UNT-FRCU. Furthermore, a solid political decision is required from Institutions directors. And all people that integrate the institution should have a special predisposition towards supporting the work team.

Another factor to be considered is the need for replacing all current operational systems of the organization.

Finally, we must say that after a year of work we are currently working into the data design. The implementation will be carried out during the year 2000.

As regards to the software technology, we are using Informix and the hardware technology of Sun.

References

Berry, M and G.Linoff, *Data Mining Techniques*, Wiley, 1997.

Firestone, J.M, *Data Warehouse and Data Marts*: A Dynamic View, White paper #3,1997, http://www.softwarejobs.com/firestone.html.

Hammer,J., H.Garcia-Molina, J.Widom, W.Labio, Y.Zhuge, *The Stanfor Data Warehousing Project*, IEEE Data Engineering Bulleting, 18(2) pp 41-48, 1995.

Inmon, W.H., *Building the Data Warehouse*, Wiley, 1996.

Kimball, R., *The data Warehouse Lifecycle Toolkit*, Wiley, 1998.

Labio,W.J., Y.Zhuge, J.Wiener, H.Gupta, H.Garcia-Molina, J.Widom, *The WHIPS Prototype for Data Warehouse Creation and Maintenance*, Proceeding of ACM Workshop on Materilized Views: Techniques and Applications, pp 26-33, Montreal, Canada, 1996.

Project FOMEC-SPU#10011, *Development and Implementation of a DSS for Academic Institution UTN-FRSF and UTN-CdelU*. The University Politic Secretary -The National Education Ministery, 1998-2000.

Sen, A. and V. Jacob, *Industrial - Strength Data Warehousing*, Communications of the ACM, Vol. 41, 9, pp 29-31, Set 1998.

Watson, H, and B. Haley, *Managerial Considerations*, Communications of the ACM, Vol. 41, 9, pp 32-37, Set 1998.

Widom, J, *Research Problems in Data Warehousing,* Proceeding of 4th International CIKM, pp. 25-30, Baltimor, Maryland, Dec 1995.