Data Mining Techniques for the Integrated Postsecondary Data System

Wilfredo Jiménez Vargas
Master in Computer Science
Jeffrey Duffany, Ph.D.
Electrical & Computer Engineering and Computer Science Department
Polytechnic University of Puerto Rico

Abstract *— The Process of generating the necessary information for the Integrated Postsecondary Data System can be tedious; a standard way to collect the information for the different departments is needed. To process it, we use data mining techniques to obtain consistent and replicable results that can be consumed by the end user for multiple criteria analysis. This often reveal patterns that shape the future organizational choices. Some data mining techniques and concepts used will be discussed in further detail on this paper.*

Key Terms *— Data Mining, Database, Hypercube, Star Schema.*

INTRODUCTION TO IPEDS

The Integrated Postsecondary Education System is a collection of data from higher education institutions that are accredited by the United States Department of Education. The completion of this report is mandatory for all institutions that take part in any federal assistance program authorized by title IV of the Higher Education Act of 1965. The data requested for this kind of report involves a population cohort which will be used to verify data. Some examples are: the average amount of grants, the average amount of time for a student to finished their degree in the first 150 percent of the time, gender, the enrollment tracking of the selected cohort and the grade conferred.

Usually, the steps needed to collect and analyze the data required for this kind of report is a challenging and intricate process. This information is verified with previous year submissions and cross-checked with the current year data.

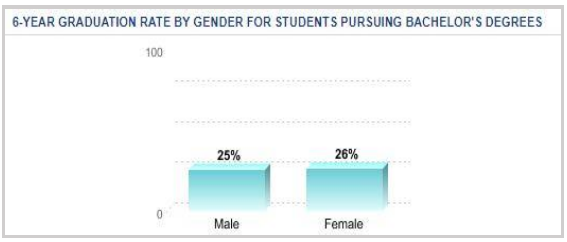

Figure 1
150% Degree Completion

Another challenge is that the majority of this information is collected manually or requested to the Information System office in multiple reports generated by different actors. This kind of data collection fragmentation creates problems when the data is loaded to the system and some of it needs to be explained. One central data collection for the distinct institutional resources is required to preserve the integrity and the traceability of the information.

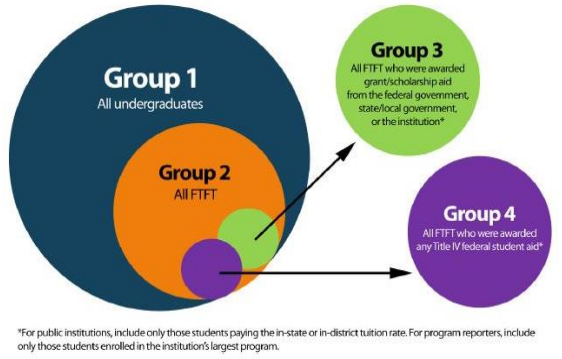

Figure 2
IPEDS Venn Diagram

METHODOLOGY

In terms of data processing the traditional statistical models have restrictions. This kind of processes require assumptions, sufficient familiarity of probabilities and distributions, however, the information must have the prerequisite of having

high quality, being a target to prior processing and transformation. The data must be sourced from different institutional departments. To be able to create an accurate schema design it is implied that a process for data requirements gathering (that must include various meetings with the department directors or delegates) was arranged in order to understand the relationship between the different sources of information that will be used. The relationship that the information has with different departments must be fully analyzed in order to design robust schemas and produce the necessary reports, this must be done in order to certify the information is accurate.

The very nature of this process poses a disadvantage, giving space to OLAP (Online Analytical Processing) a method by which multidimensional analysis works. It is a process that consists of collecting knowledge from databases to generate information that is not known beforehand. Multidimensional analysis is the manipulation of information using a variety of relevant categories or "dimensions" to facilitate analysis encompassing the understanding of underlying data.

BASIC OPERATIONS OF OLAP

Slice

For the slice procedure a dimension is selected, a new sub-cube is created. Example: Divide the Hypercube with 3 dimensions (Trimester, Quarter, Campus), and use only the first two dimensions to facilitate analysis.

Dice

The dice operation is similar to slice. The difference in this operation is that two or more dimensions are selected. Example: Divide the Hypercube with 3 dimensions (Trimester, Major, Campus), and use only the latter two dimensions, for a more in detail analysis.

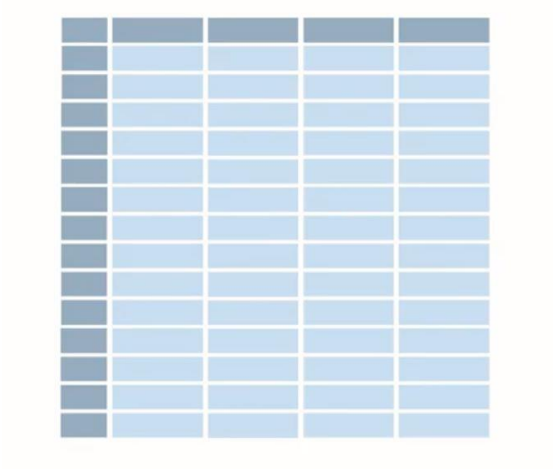

Figure 3
Dice Example

Pivot

For the pivot operation, the data axes are rotated to provide an alternative presentation for the data.

Roll-Up

The roll-up is also recognized as "consolidation" or "aggregation". The roll-up operation can be completed in 2 ways.

- Decreasing Dimensions
- Ascending up concept hierarchy, this a system of grouping objects based on their order level.

Example: Civil and Electrical Engineering can be merged into a single classification called engineering.

Drill-Down

In drill-down, data is fragmented into smaller segments. It is the contrary of the rollup process. This can be accomplished via:

- Moving down the concept hierarchy
- Increasing a dimension.

Example: An academic year can be drilled down into trimesters corresponding registered student will be presented in a more detailed manner.

WHAT IS MOLAP

Multidimensional OLAP (MOLAP) is an implementation of OLAP that facilitates data analysis by using a multidimensional data cube. Data is precomputed, pre-summarized, and stored. Using this method a stakeholder can use the multidimensional view to observe the data with different facets.

The Multidimensional data analysis can also be implemented in a relational database by querying multiple tables. However, the version that is used for this project is the MOLAP because it has all the possible combinationsof data already stored in a multidimensional array.

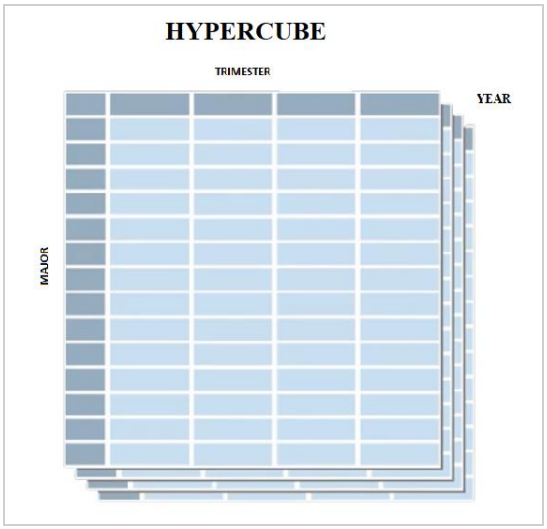

Figure 4
IPEDS Hypercube

KEY POINTS

Using MOLAP to process information with the same time, irrespective of the level of summarizing.

- MOLAP remove the need of designing a relational database to store data for analysis.
- Facts are stored in multidimensional arrays and dimensions used to query them.

The following architecture includes the following components:

- Database Server
- MOLAP server
- Front-end tool.

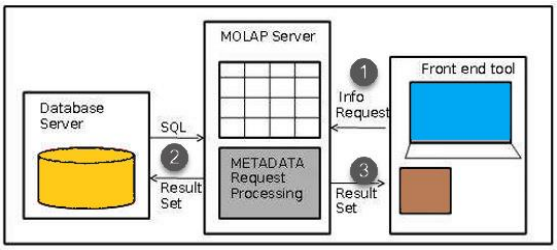

Figure 5
MOLAP Architecture

Above Given MOLAP Architecture

- The User Request Reports through the interface
- The application logic layer of the MDDB retrieve the stored from the Database
- The application logic layer forwards the result to the client user.

This architecture mainly reads the precompiled data. However, this architecture has limited capabilities to create aggregations dynamically or to calculate results that have not pre-calculated and stored.

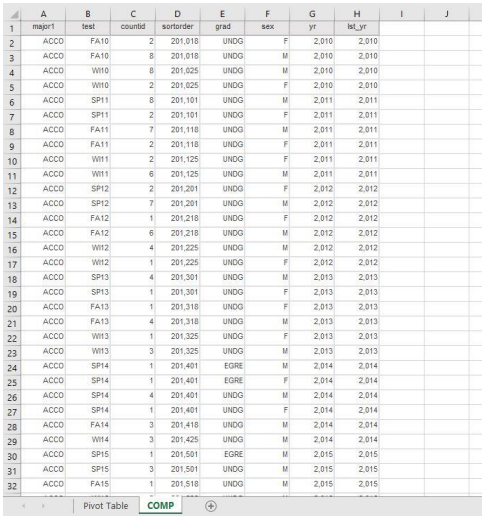

Figure 6
HyperCube Raw Data

Advantages of Using OLAP

- Information and calculations are consistent in a Hypercube, and this is a crucial benefit.
- Quickly create and analyze “What if” scenarios
- Easily search the OLAP database for broad or specific terms

- OLAP provides the building block for business modeling tools, data mining tools, and performance reporting tools.
- Allows the users to do slice and dice cube data all by various dimensions, measures, and filters.

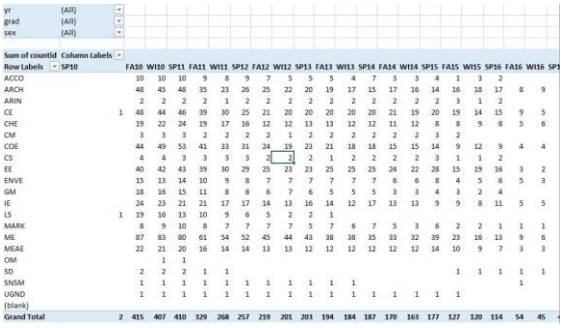

Figure 7
Single Dimension in Excel

Disadvantages of OLAP

OLAP have the requirement for organizing data into a star or snowflake schema. These schemas are complex to implement and administer.

- You cannot have a large number of dimensions in a single Hypercube.
- Transactional data cannot be accessed with the OLAP system.
- Any modification in a Hypercube needs a full update of the cube, this a complex and time-consuming process.

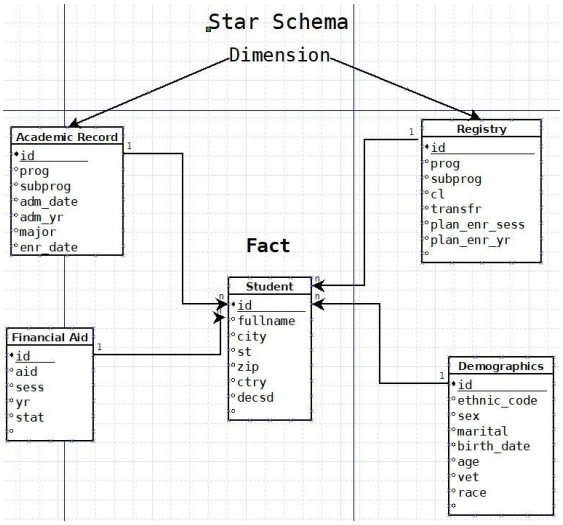

Figure 8
Single Star Schema for IPEDS

ORGANIZATIONAL IMPACT

In a decision-making context, the analysis is an intellectual process that allows generating knowledge from hypotheses and data [1]. This can be presented by explaining a phenomenon and proposing recommendations for the stakeholder decision making.

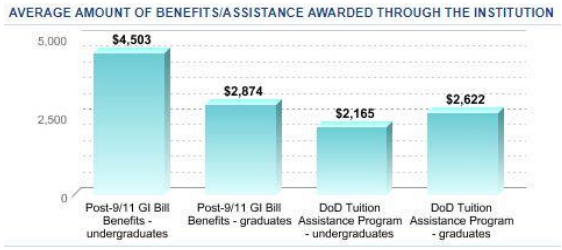

Figure 9
Results for Context Decision Analysis

The OLAP system as technology appears among the most suitable multidimensional analysis tools that are usually used by decision makers and analyst who need to transform data into usable information, which will enable the management to have a clear vision of their activities all the time. Moreover, the economic and business intelligence tools, especially OLAP systems, are considered among the best technologies, most eminent and versatile in the environment of decision support systems. In fact, OLAP systems are at the heart of many economic analysis applications and appear as complete systems providing useful and necessary services for efficient, rational, and analytic processing of data. The functionalities of these systems, based on a multidimensional database approach [2], are characterized by the ability to support an efficient and flexible exploration of multidimensional cubes.

Several studies have been conducted around the topic of OLAP technology reflecting its degree of relevance and effectiveness to be implemented in multiple business intelligence areas. In fact, in a decision-making context, the OLAP system is a known well-mastered technology when it comes to simple data, which explains its ability to be easily integrated with other environments.

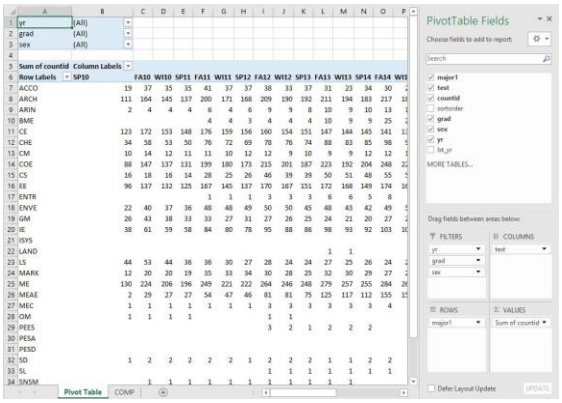

Figure 10
Multiple Criteria Data Analysis

The complex consequences of certain decision-making situations necessitate taking into account the multicriteria and conflicting aspects of data, as well as the consideration of several types of data (quantitative and qualitative) to represent all the necessary information for making decisions adequately.

OLAP tools that exist in the decision-making area however, still suffer from limitations related to the lack of technical means to consider the multiple criteria and imprecise nature of decision data in the analysis process. In fact, OLAP systems are the cornerstone of many analysis applications and present as a complete system providing useful and necessary services for efficient, rational, and analytic processing of data.

TOOLS & ENVIRONMENT

The IBM Informix Dynamic Server RDBMS serves as the intermediary between the user and the database. The database structure itself is contained as a collection of files, which can be accessed using it. It translates all application request and renders in the complex operations required to fulfill those request, in this specific project the database has over 130 tables with pertinent data, which will be processed.

It is imperative to know the version of SQL that this version of Informix is using for purposes of code maintenance and as a sound developing practice ensuring that the code us reusable for a reasonable amount of time. In this case, it is SQL-ISO-9075-1.

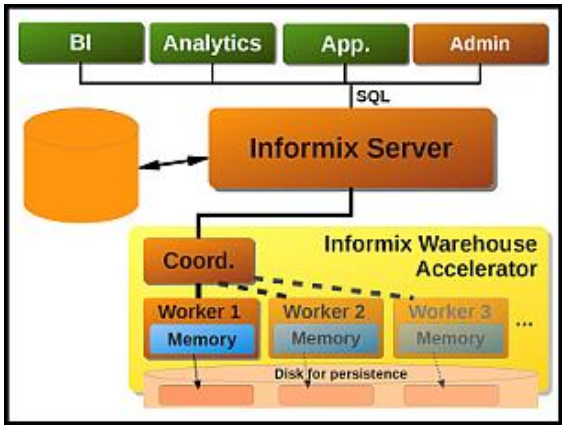

Figure 11
IBM Informix Typical Implementation [3]

The SUSE Enterprise Server 11 is a Linux based operating system developed by SUSE. It is designed for running in servers and significant versions are usually released every four years. This version was released precisely on March 24, 2009 and included the Linux kernel 2.6.27, and features an Oracle Cluster File System Release 2 and OpenAIS cluster communication protocol for server and storage clustering.

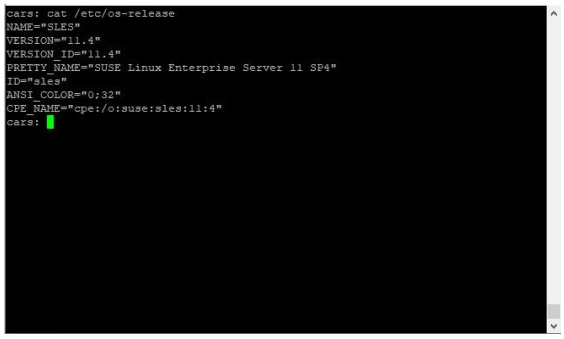

Figure 12
SUSE Enterprise on Putty Client

CONCLUSION

In overall, the success of this kind of data collection techniques include exhaustive communication with the stakeholder, and the data presentation must be in a manner that the end user can relate. As a result, Microsoft Excel was selected to display the information because the

familiarity and widespread use, prevented that end users felt alienated.

It is important to have in mind the necessities that the user might have because they are the subject manner expert in their field and there is a need that the data is understandable and useful by them in order to be verified and certified by the departments.

RESULTS

The IPEDS schemas generated from the collective cooperation from different institutional sources where the cornerstone for timely completion of the information requested. Still, the data generated can be used for other reports because the IPEDS demand a broad range of information that is applicable to other reporting needs. The design of this of data structure has future changes on its design by implementing a highly maintainable, documented and robust architecture.

FUTURE WORK

At this point it is uncertain what will be the data requirement changes for next year reporting period. In addition, sometimes the data source might change because of a modification on the system requirements. If this is the case the schemas must be updated and documented accordingly preventing future problems.

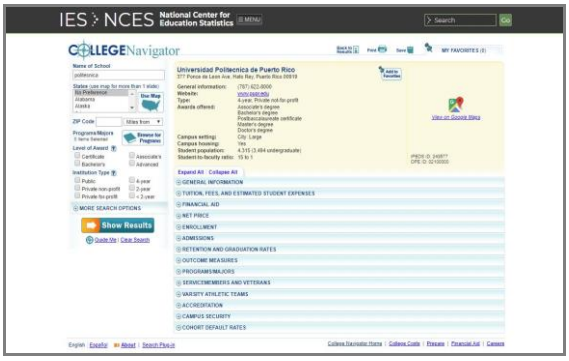

Figure 13
Result as Part of the NCES [4]

REFERENCES

[1] E. F. Codd, et al. "Beyond decision support," in *Computerworld*. Oxford University Press, 1993. [online document]. Available: Ebsco Host Online, http://ezproxy.pupr.edu:2055/login.aspx?direct=true&db=bth&AN=9311083206&site=ehost-live [Accessed: Jan 24, 2019].

[2] R. Kimball, *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*, 3rd ed. Reading, MA: Addison Wesley, 2013. [E-book] Available: Kindle e-book.

[3] IBM "Community," Dic, 2015. [Online]. Available: https://www.ibm.com/developerworks/community/blogs/2fa81a5c-cb30-4873-b775-1370151e3614?lang=en.htm. [Accessed Jan. 20, 2019].

[4] National Center for Education Statistics. "Universidad Politécnica de Puerto Rico College Navigator" Jan, 2017. [Online].Available:https://nces.ed.gov/collegenavigator/?q=politecnica&s=all&id=243577. [Accessed Jan. 22, 2019].