Data Mining Techniques For the Integrated Postsecondary Data System

Wilfredo Jiménez Vargas
Jeffrey Duffany, Ph.D.
Electrical & Computer Engineering/Computer Science Department

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 department's needs.
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 that often reveal patterns that shape the future organizational choices. Some of the different techniques and concepts used will be discussed in further detail on this paper.

Introduction

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 of all institutions that take part in any federal assistance program authorized the title IV of the Higher Education Act of 1965. Usually, the data requested for this kind of report involves a population 'Cohort.' This population will be used to verify data; one example is the average amount of grants or the average amount of time a student finished their degree in the first 150 percent of the time.

Background

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; 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.

Problem

This kind of data collection fragmentation creates problems when the data is loaded to the system, and some of it needed to be explained. One central data collection for the distinct institutional resources is required to preserve the integrity and the traceability of the information.

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 being of high quality, being a target to prior processing and transformation. The very nature of this process poses a disadvantage, giving space to OLAP, which stands for (Online Analytical Processing) this, is a method by which multidimensional analysis works.
The data must be sourced from the different institutional departments, this implies that a process for data requirements gathering that must include various meetings with the department directors or delegates where arranged in order to understand the relationship between the different sources of information that will be used in order to create an accurate schema design. The relationship that the information have 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.

What is MOLAP

Multidimensional OLAP (MOLAP) is a 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 this the version that is used for this project is the MOLAP because it has all the possible combinations of data already stored in a multidimensional array.

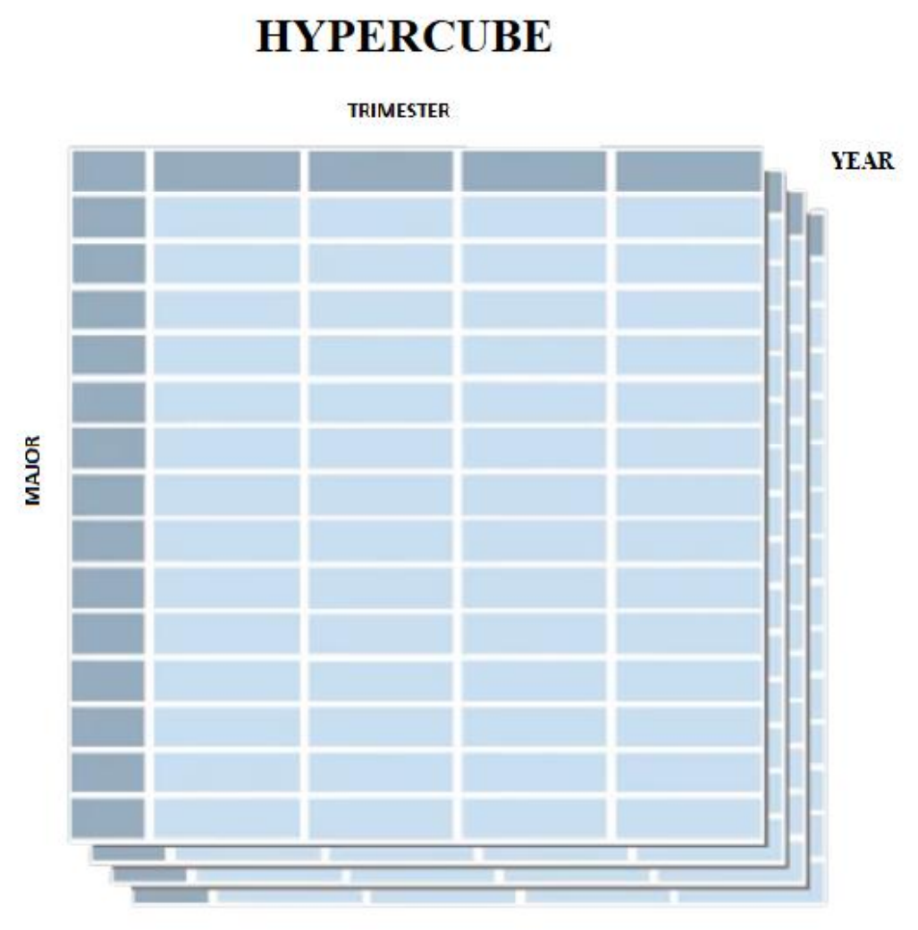


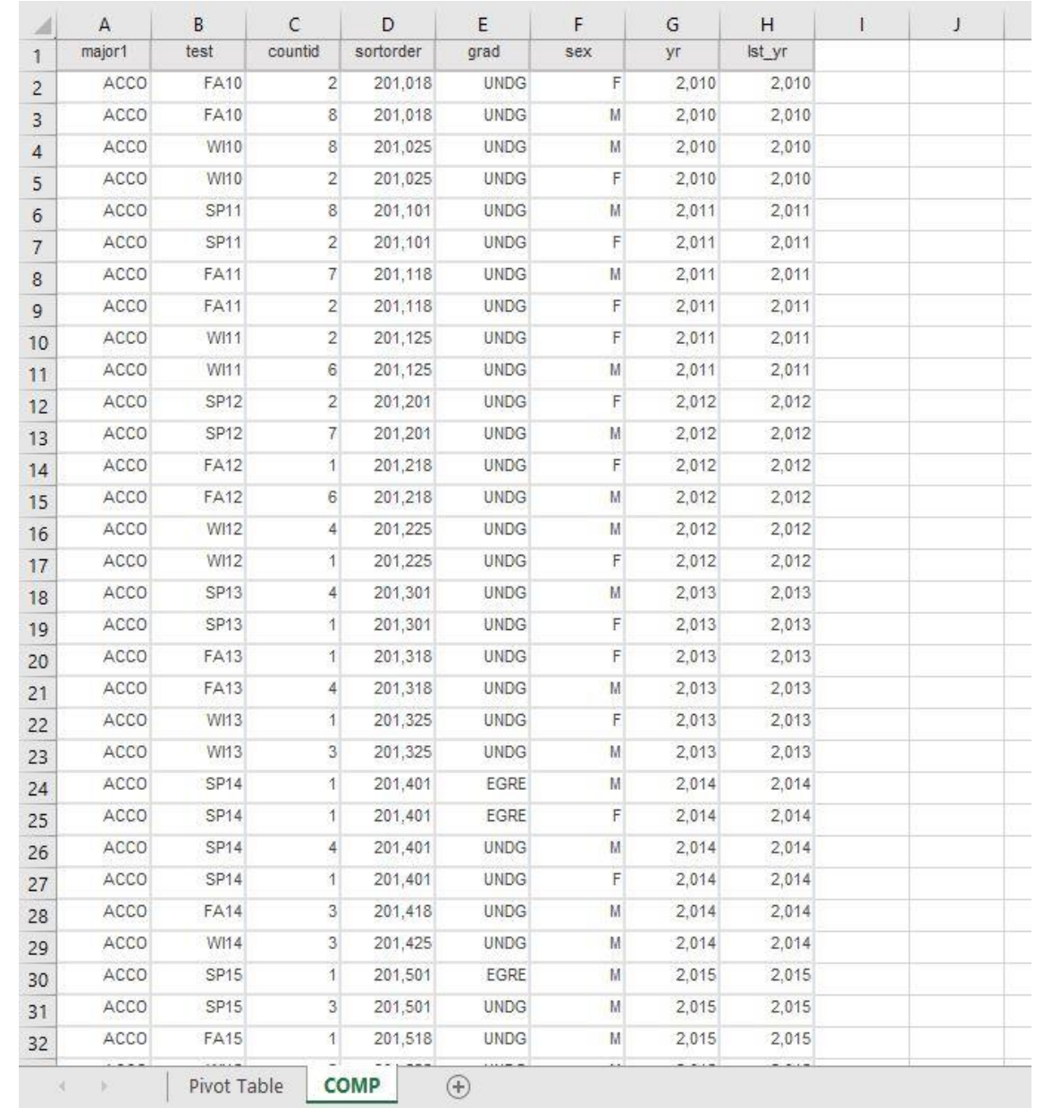
Figure 1
IPEDS Hypercube

Key Points

- Using MOLAP to process information with the same time irrespective of the level of summarizing.
- It also remove the need for design a relational database to store data for analysis.
- Facts are stored in multidimensional arrays and dimensions used to query them.

Results and Discussion

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.



	A	B	C	D	E	F	G	H	I	J
1	major1	test	countid	sortorder	grad	sex	yr	lst_yr		
2	ACCO	FA10	2	201,018	UNDG	F	2,010	2,010		
3	ACCO	FA10	8	201,018	UNDG	M	2,010	2,010		
4	ACCO	WI10	8	201,025	UNDG	M	2,010	2,010		
5	ACCO	WI10	2	201,025	UNDG	F	2,010	2,010		
6	ACCO	SP11	8	201,101	UNDG	M	2,011	2,011		
7	ACCO	SP11	2	201,101	UNDG	F	2,011	2,011		
8	ACCO	FA11	7	201,118	UNDG	M	2,011	2,011		
9	ACCO	FA11	2	201,118	UNDG	F	2,011	2,011		
10	ACCO	WI11	2	201,125	UNDG	F	2,011	2,011		
11	ACCO	WI11	6	201,125	UNDG	M	2,011	2,011		
12	ACCO	SP12	2	201,201	UNDG	F	2,012	2,012		
13	ACCO	SP12	7	201,201	UNDG	M	2,012	2,012		
14	ACCO	FA12	1	201,218	UNDG	F	2,012	2,012		
15	ACCO	FA12	6	201,218	UNDG	M	2,012	2,012		
16	ACCO	WI12	4	201,225	UNDG	M	2,012	2,012		
17	ACCO	WI12	1	201,225	UNDG	F	2,012	2,012		
18	ACCO	SP13	4	201,301	UNDG	M	2,013	2,013		
19	ACCO	SP13	1	201,301	UNDG	F	2,013	2,013		
20	ACCO	FA13	1	201,318	UNDG	F	2,013	2,013		
21	ACCO	FA13	4	201,318	UNDG	M	2,013	2,013		
22	ACCO	WI13	1	201,325	UNDG	F	2,013	2,013		
23	ACCO	WI13	3	201,325	UNDG	M	2,013	2,013		
24	ACCO	SP14	1	201,401	EGRE	M	2,014	2,014		
25	ACCO	SP14	1	201,401	EGRE	F	2,014	2,014		
26	ACCO	SP14	4	201,401	UNDG	M	2,014	2,014		
27	ACCO	SP14	1	201,401	UNDG	F	2,014	2,014		
28	ACCO	FA14	3	201,418	UNDG	M	2,014	2,014		
29	ACCO	WI14	3	201,425	UNDG	M	2,014	2,014		
30	ACCO	SP15	1	201,501	EGRE	M	2,015	2,015		
31	ACCO	SP15	3	201,501	UNDG	M	2,015	2,015		
32	ACCO	FA15	1	201,518	UNDG	M	2,015	2,015		

Figure 2
Data for Multidimensional Analysis

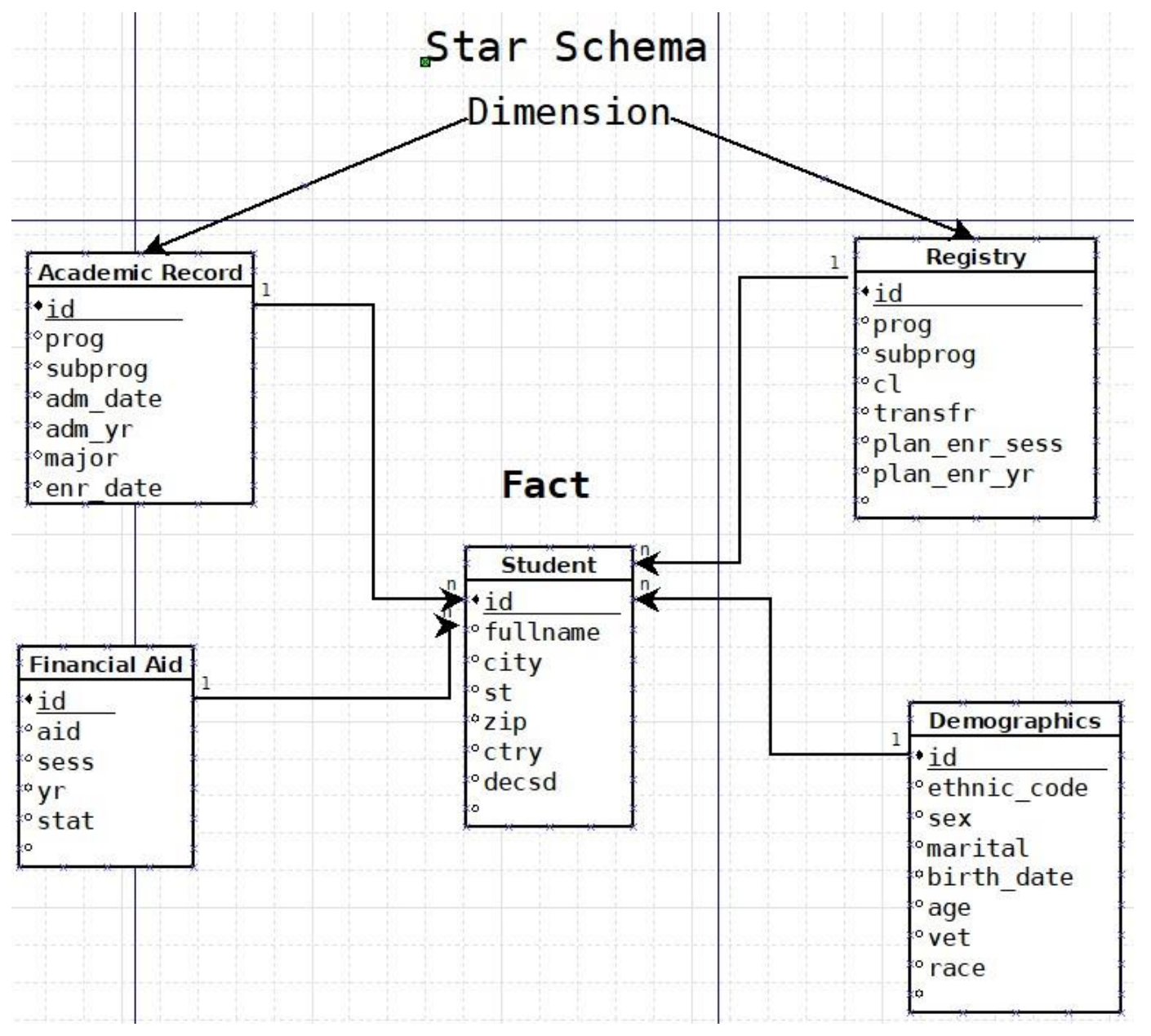
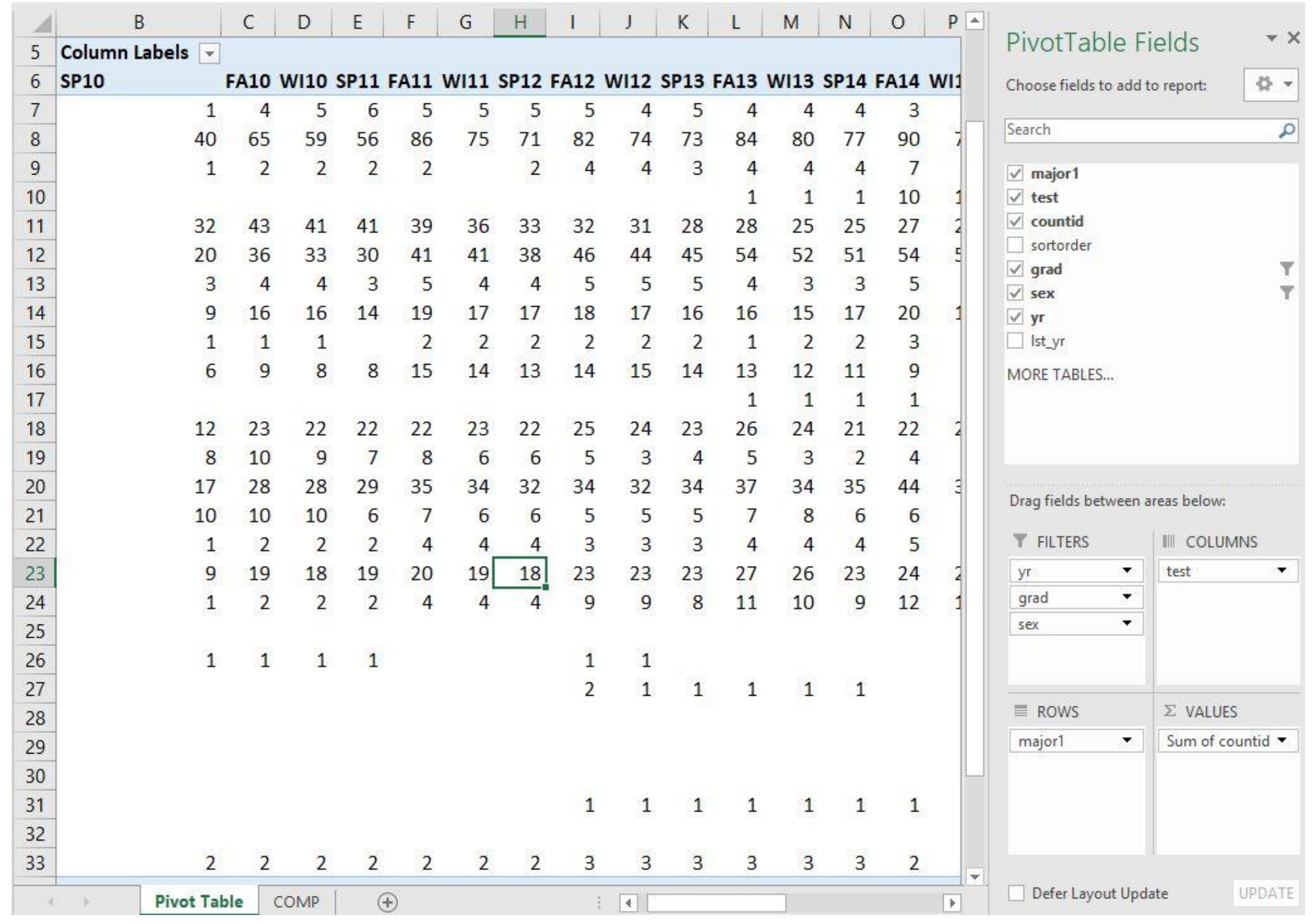


Figure 3
Star Schema

However, OLAP tools that exist in the decision-making area 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.



	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	Column Labels													
6	SP10	FA10	WI10	SP11	FA11	WI11	SP12	FA12	WI12	SP13	FA13	WI13	SP14	FA14
7	1	4	5	6	5	5	5	5	4	5	4	4	4	3
8	40	65	59	56	86	75	71	82	74	73	84	80	77	90
9	1	2	2	2	2		2	4	4	3	4	4	4	7
10											1	1	1	10
11	32	43	41	41	39	36	33	32	31	28	28	25	25	27
12	20	36	33	30	41	41	38	46	44	45	54	52	51	54
13	3	4	4	3	5	4	4	5	5	5	4	3	3	5
14	9	16	16	14	19	17	17	18	17	16	16	15	17	20
15	1	1	1		2	2	2	2	2	2	1	2	2	3
16	6	9	8	8	15	14	13	14	15	14	13	12	11	9
17											1	1	1	1
18	12	23	22	22	22	23	22	25	24	23	26	24	21	22
19	8	10	9	7	8	6	6	5	3	4	5	3	2	4
20	17	28	28	29	35	34	32	34	32	34	37	34	35	44
21	10	10	10	6	7	6	6	5	5	5	7	8	6	6
22	1	2	2	2	4	4	4	3	3	3	4	4	4	5
23	9	19	18	19	20	19	18	23	23	23	27	26	23	24
24	1	2	2	2	4	4	4	9	9	8	11	10	9	12
25														
26	1	1	1	1				1	1					
27								2	1	1	1	1	1	
28														
29														
30														
31								1	1	1	1	1	1	1
32														
33	2	2	2	2	2	2	2	3	3	3	3	3	3	2

PivotTable Fields
Choose fields to add to report:
Search
☑ major1
☑ test
☑ countid
☐ sortorder
☑ grad
☑ sex
☑ yr
☐ lst_yr
MORE TABLES...
Drag fields between areas below:
FILTERS: yr, grad, sex
COLUMNS: test
ROWS: major1
VALUES: Sum of countid
☐ Defer Layout Update
UPDATE

Figure 4
Modeled And Filtered Data

Conclusions

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, the Microsoft Excel was selected to display the information, because the familiarity and wide spread use prevented that end user 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.

Future Work

At this point is uncertain what will be the changes for the next year reporting period. In addition, that 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

Acknowledgements

I would like to thank Professor Dr. Jeffrey Duffany for his expert advice and encouragement through this difficult project, as well to Professor Nelliud D. Torres for the lessons learned in the course Data Mining & Data Warehousing. One last acknowledgement to my colleagues for their wonderful collaboration. Your supported me greatly and were always willing to help me.

References

[1]Codd, Edgar F., et al., "Beyond decision support," in Computerworld. Oxford University Press, [online document], 1993. 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]Ralph 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]BM 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 ," Jan, 2017. [Online].Available:https://nces.ed.gov/collegenavigator/?q=politecnica&s=all&id=243577. [Accessed Jan. 22, 2019]