The Development of Academic Data warehouse as a Basis For Decision Making

A Case Study at XYZ University

Paulina H.Prima Rosa, Ridowati Gunawan, Sri H. Wijono
Informatics Engineering Department
Sanata Dharma University
Yogyakarta, Indonesia
rosa@usd.ac.id, rido@usd.ac.id, tatik@usd.ac.id

***Abstract*— In this paper, the authors report the development of academic data warehouse which consist of students' performance in high school, their university entrance test scores, and their grade point average (GPA) in university. The data warehouse was developed based on Student Admission Information System (SAIS) and Academic Information System (AIS) of XYZ university. To validate the usefullness of the data warehouse as a basis for decision making, the warehouse has been analyzed using OLAP technique and tested to several users. In addition, the data resulted from the warehouse has been mined using Weka data mining tools to test whether the data can be utilized as a basis for decision making.**

***Keywords*— academic data warehouse; decision making**

I. Introduction

The result of National Examination (NE) for Indonesian high school students in year 2010 statistically decreased. There were 267 high schools in Indonesia whose students 100% failed [1]. These failures were responded variously by students. Among of them showed destructive responds either to themself or their environment [1]. It can be easily understood that in the context of some students, fail in NE is similar to fail in life. Therefore, they perceive that committed to suicides is a solution.

On the other side, several state and private universities declared that students who passed entrance tests in those universities but failed in NE are eligible to follow the lecturer after submitting credentials in a certain period [2]. The fact implicitly showed that these universities believe that students who were failed in NE potentially capable to continue their studies in higher educations.

The previous mentioned facts rise a question about the relation between high school students' performance and their academic performance in higher educations. Driana [3] proposed a study of policy in NE through several mechanisms, for example by organizing researchers from universities to study the implication of NE towards the readiness of students in pursuing higher educations.

This idea can be initiated by studying the relationship between high school students performance and their academic performance in higher educations by using data warehousing and data mining techniques. Previous studies in data warehousing and data mining showed that data warehousing and data mining techniques have been used to understand students and their academic environments better. Baranovic et.al. [4] reported an implementation of data warehouse for the Higher Education Information System in Croatia. Ramaswami and Bhaskaran [5] studied the usefullness of data mining to predict result of elementary and high school students in India using CHAID Prediction Model. Bravo and Ortigosa [6] elaborated their research on student activities in e-leaning media to detect symptoms of low performance using production rules. Vialardi et al. [7] presented a recommendation in higher education using data mining techniques. Azimah and Sucahyo [8] have developed a data warehouse and performed data mining of academic data in “Nasional” University. Ernawati [9] performed data mining to find quantitative association rules of students' academic performance and their genders as well as their entrance tests in higher education. Haryanto and Rosa [10] performed a prediction towards admitted students who are not enrolled. Wirati and Rosa [11] presented the classification of students' performance based on their university entrance test scores.

Eventhough several researches in students' performance have been carried out, none of it discusses about the relationship between high school students' performance and their success rate in higher educations due to the facts that the data are not available electronically.

XYZ University is a private University in Indonesia that has developed its own information system since 1996. Among of its systems are Student Admission Information System (SAIS) and Academic Information System (AIS). SAIS was developed to handle student admission from application process up to the process of deciding which applicants will be admitted. On the other hand, AIS was developed to handle academic process after students have been officially registered. Although the two systems were operated in the platform of university's local area network, they are not integrated. Due to this fact, authors' preliminary survey in XYZ University showed that Head of Study Programs had difficulties in finding comprehensive view of students' performance in high school and their academic records during study in higher educations, which is needed in making decisions related to students. For example, during admission process, Head of Study Program might need information about university students' success rate based on high school study area of students from a particular high school.

Therefore, this paper elaborates the development of data warehouse to handle high volume of data that were extracted from the Student Admission Information System and the Academic Information System of XYZ University. The result of the study is an integrated data warehouse which is usefull for users and ready to be utilized as a basis for decision making.

II. SUBJECT AND DATA

A. *Subject*

The subjects used in this research are the students batch 2007 and 2008 of four study programs in XYZ University who are registered in the second semester of 2010/2011. The students admitted through two types of admission namely: (1) outstanding student admission, and (2) regular admission. The outstanding student admission will select eligible students by using students' grades in high school, while regular admission will select students based on entrance test scores.

B. *Data*

The data were extracted from two sources of XYZ University: (1) the Student Admission Information System (SAIS) and (2) the Academic Information System (AIS). All data are gathered from the system in the form of database format. The two systems were not integrated as a single system. It causes difficulties for novice users to access comprehensive knowledge from the two systems. Table 1 describes the data used in this research.

Table 1. Data Source

Data Name	Period	Source
The entire database table of Student Admission System from all admission types, which contains all fields	Student Admission in 2007-2008	SAIS
Complete data of all active students in 4 Study Programs, which contains all fields	Active student in second semester 2009/2010	AIS
Course taken by all students	Course taken by all students from academic year 2007/2008 up to 2009/2010	SAIS
Data dictionary containing the meaning of all data in SAIS and AIS	Lifetime of SAIS and AIS.	SAIS and AIS

III. DATA WAREHOUSE DEVELOPMENT

The data warehouse design was performed based on nine-step methodology as proposed by Kimball [11]. Following are the step of the data warehouse development:

1. Choosing the process
   Data warehouse to be constructed is intended to conduct an analysis of students' performance. The subject of data warehouse is student. From the whole system of XYZ University, two processes are selected to develop the warehouse, namely student admission process and student assessment process.
2. Choosing the grain
   The selected grains in this case are: (1) student admission report which consists of students' performance in high school and university entrance test score report; (2) result of study in university which consits of students' Grade Point Average (GPA) for each semester.
3. Identifying and conforming the dimensions
   Following are possible dimensions that are chosen: gender, admission type, study program, high school name, high school district, and students' study area in high school.
4. Choosing the facts
   Following are the facts that are incorporated in the warehouse: (1) identity of admitted students, (2) students' score in high school, (3) students' entrance test score, (4) study program, (5) subjects taken by students, (6) students' grade for each subject.
5. Storing pre-calculations in the fact table
   In this case, a pre-calculation to find GPA of each student for each semester were required. The GPA will then be stored in the fact table together with other facts elaborated in step 4.
6. Rounding out the dimension tables
   The information in dimension table were completed if needed.
7. Choosing the duration of the database
   Data to be loaded and analyzed are students of academic year 2007/2008 up to 2009/2010.
8. Tracking slowly changing dimensions
   There is no changing dimensions in this case.
9. Deciding the query priorities and the query modes.
   After the fact table was defined, the following OLAP queries were designed to present data from the data warehouse:
   a. Examining the average of students' high school grades and GPA based on the study program and school dimensions for all students from outstanding track.
   b. Examining the average of entrance test score and GPA based on the study program and school dimensions for all students from regular admission track.
   c. Examining the average of students' high school grades and GPA based on the school dimension for all students from outstanding track.

d. Examining the average of entrance test score and GPA based on school dimensions for all students from regular admission track.
e. Examining the average of students' high school grades and GPA based on the study area dimension for all students from outstanding track.
f. Examining the a average of entrance test score and GPA based on study area dimensions for all students from regular admission track.

IV. Results and Analysis

A. *Online Analytical Processing (OLAP)*

Once the data warehouse is constructed, the author carried out an analysis using OLAP techniques. The atributes used as the analysis is the students' high school grades, NE grades, university entrance test score, GPA1, GPA2, GPA3, and GPA4, which will be viewed from the following dimensions: high school, admission type, study program, gender, study area, and high school district. Fig. 1 shows the star schema which is designed for OLAP.

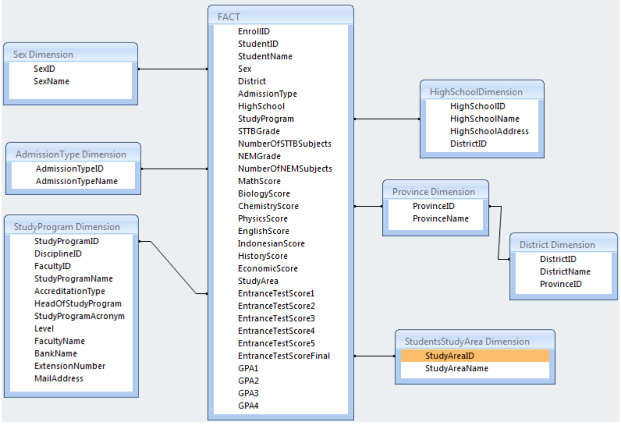


Fig 1. Star Schema Design

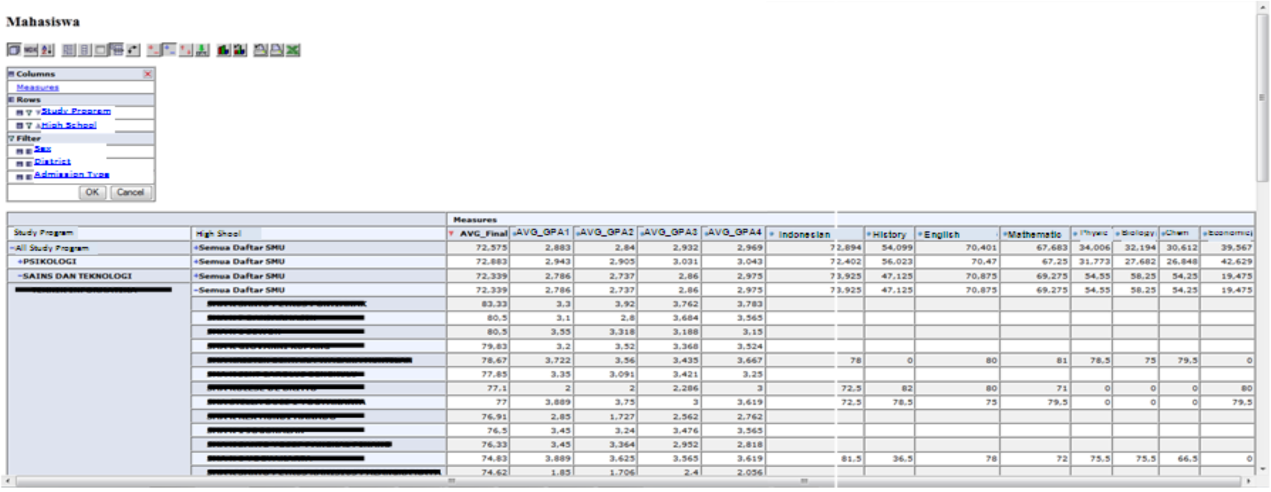
Analytical results obtained by using OLAP techniques which generate XML documents . Users could select the focus of their attention. The possible measures to be analyzed are the average GPA, the average university entrance test scores, the average students' high school grades on several subjects (Indonesian Language, History, English, Mathematics, Physics, Chemistry, Biology, and Economy). The analysis might based on high school, admission type, study program, gender, study area, and high school district. The following are examples of the interesting possibilities of multidimensional analysis:

a. Examining the average of students' high school grades and GPA based on the study program and school dimensions for all students from outstanding track.
b. Examining the average of entrance test score and GPA based on the study program and school dimensions for all students from regular admission track.
c. Examining the average of students' high school grades and GPA based on the school dimension for all students from outstanding track.
d. Examining the average of entrance test score and GPA based on school dimensions for all students from regular admission track.
e. Examining the average of students' high school grades and GPA based on the study area dimension for all students from outstanding track.
f. Examining the average of entrance test score and GPA based on study area dimensions for all students from regular admission track.

Fig. 2 shows a sample report of multidimensional analysis example point a aboved. Several parts of the report are blocked due to the confidentiality of the information.

To generate the report in Fig. 2, average of students' high school grades were calculated from the fields of fact table by using average (AVG) aggregate function. For example, the average of MathScore field in Fig. 1 was stored as Mathematic field in Fig. 2. While the average of each GPA was calculated by using average (AVG) aggregate function to be stored as the following fields: AVG_GPA1, AVG_GPA2, AVG_GPA3, and AVG_GPA4. Data aggregate were calculated based on six dimensions namely high school, admission type, study program, gender, study area, and high school district.

.



		Measures												
Study Program	High Shool	AVG_Final	AVG_GPA1	AVG_GPA2	AVG_GPA3	AVG_GPA4	Indonesian	History	English	Mathematic	Physic	Biology	Chem	Economic
-All Study Program	+Semua Daftar SMU	72.575	2.883	2.84	2.932	2.969	72.894	54.099	70.401	67.683	34.006	32.194	30.612	39.567
+PSIKOLOGI	+Semua Daftar SMU	72.883	2.943	2.905	3.031	3.043	72.402	56.023	70.47	67.25	31.773	27.682	26.848	42.629
-SAINS DAN TEKNOLOGI	+Semua Daftar SMU	72.339	2.786	2.737	2.86	2.975	73.925	47.125	70.875	69.275	54.55	58.25	54.25	19.475
[redacted]	-Semua Daftar SMU	72.339	2.786	2.737	2.86	2.975	73.925	47.125	70.875	69.275	54.55	58.25	54.25	19.475
	[redacted]	83.33	3.3	3.92	3.762	3.783								
	[redacted]	80.5	3.1	2.8	3.684	3.565								
	[redacted]	80.5	3.55	3.318	3.188	3.15								
	[redacted]	79.83	3.2	3.52	3.368	3.524								
	[redacted]	78.67	3.722	3.56	3.435	3.667	78	0	80	81	78.5	75	79.5	0
	[redacted]	77.85	3.35	3.091	3.421	3.25								
	[redacted]	77.1	2	2	2.286	3	72.5	82	80	71	0	0	0	80
	[redacted]	77	3.889	3.75	3	3.619	72.5	78.5	75	79.5	0	0	0	79.5
	[redacted]	76.91	2.85	1.727	2.562	2.762								
	[redacted]	76.5	3.45	3.24	3.476	3.565								
	[redacted]	76.33	3.45	3.364	2.952	2.818								
	[redacted]	74.83	3.889	3.625	3.565	3.619	81.5	36.5	78	72	75.5	75.5	66.5	0
	[redacted]	74.62	1.85	1.706	2.4	2.056								

Fig 2. Sample Report from OLAP

OLAP multidimensional analysis resulted from the data warehouse has been demonstrated to four Head of Study Programs. A questionnaire to evaluate the result were also distributed to them. The results of the questionnaire are presented on table 2.

Table 2. Users' Evaluation

Statement	Strongly Agree	Agree
The information generated from the data warehouse facilitates Head of Study Programs to thoroguhy view high school grades and GPA of the students from outstanding track, based on the school dimensions	80%	20%
The information generated from the data warehouse facilitates Head of Study Programs to thoroughy view university entrance test score and GPA of the students from regular admisssion track, based on the school dimensions	100%	
The information generated from the data warehouse facilitates Head of Study Program to get student profiles and to make decisions related to the student	80%	20%
The user interface of the data warehouse is easy to use	100%	

Thus in general it can be concluded that the data warehouse has been able to integrate the student academic record in high school, the university entrance test scores, and student academic performance in university.

B. *Data Mining*

Data extracted from the datawarehouse that has been saved in a spreadsheet format (. csv) were tested further using Weka data mining tools to evaluate whether it could be utilized as a basis for generating knowledge potentially usefull for decision making. Table 3 in the following are the examples of possible data mining towards the data.

Table 3.Result of Data Mining

Goal	Algo-rithm	Rules	Accuracy
Classify GPA based on NE scores	Id3	If 6.73≤ NE score < 7.63 then GPA < 2.5 If NE score ≥ 8.54 then GPA ≥ 3.0 If 7.63 NE score <8.54 then GPA< 2.5 If 5.82≤ NE score < 6.73 then GPA< 2.5 If 4.92 ≤ NE score then GPA< 2.5	53. 27%
Classify GPA based on universi-ty entrance test score	Id3	If entr_tes < 23 then GPA < 2.5 If entr_tes > 71.8 then GPA < 2.5 If 59.6≤entr_tes< 71.8 then GPA>3.0 If 47.4≤entr_tes< 59.6 then GPA<2.5 If 35.2≤entr_tes< 23.0 then GPA<2.5	55.14%
Classify universi-ty entrance test score based on NE scores	Id3	If NE score < 4.92 then 35.2≤ entr_tes < 23.0 If NE score ≥ 8.54 then 47.4≤ entr_tes < 59.6 If 7.63≤NE score <8.54 then 59.6≤ entr_tes < 71.8 If 75.82≤NE score <7.63then entr_tes < 23.0 If NE score <4.92 then entr_tes < 23.0	32.71 %

Eventhough the accuracy is not high enough, we could find rules relating NE scores, university entrance test scores, and GPA. To improve the accuracy, further refinement of the parameters need to be investigated. However, it has been proofed that the datawarehouse is ready to be mined. In addition to the above examples of data mining, other data mining techniques such as clustering and asssociation might be applied to the data extracted from the warehouse.

V. CONCLUSION

Academic data warehouse of XYZ University has been successfully designed and implemented using the star schema. Analysis of academic performance can be seen with various measure of the average GPA, average university entrance test score, average NE score, and the average students' high school grades in several subjects. Analysis can be viewed from various dimensions namely high school, admission type, study program, gender, study area, or high school district, based on the interest of the user. Users' evaluation showed that the information generated from the data warehouse facilitates Head of Study Program to get student profiles and to make decisions related to the student. Analysis results in the form of spreadsheet format can also be stored separately and subjected to data mining.

Further research to be done is performing more comprehensive data mining towards data extracted from the data warehouse to get valuable knowledge from it. In addition, the presentation of multi-dimensional analysis for novice users might be improved so that it will be more user friendly.

ACKNOWLEDGMENT

This work has been funded by Institution of Research and Community Services, XYZ University.

REFERENCES

[1] _______, "Hasil UN Mengejutkan, Sejumlah Sekolah 100 Persen Siswanya Tidak Lulus". Kompas Newspaper. 27 April 2010.

[2] _______, "Siswa Tak Lulus UN Perguruan Tinggi Berikan Kesempatan". Kompas Newspaper. 29 April 2010.

[3] Elin Driana, "Tengoklah Ruang-ruang Kelas Anak-anak Kita". Kompas Newspaper. 4 Mei 2010.

[4] M. Baranovic, M. Madunic, and Igor Mekterovic,"Data Warehouse as Part of The Higher Education Information System in Croatia", Proceedings of the 25th International Conference on, ….

[5] M. Ramaswami & R. Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining". IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010

[6] J. Bravo & Alvaro Ortigosa, "Detecting Symptoms of Low Performance Using Production Rules, Proceeding of The 2nd International Conference on Data Mining, Cordoba, Spain, 2009.

[7] C. Vialardi, Javier Bravo, Leila Shafti, Alvaro Ortigosa, "Recommendation in Higher Education Using Data Mining Techniques", Proceeding of The 2nd International Conference on Data Mining, Cordoba, Spain, 2009

[8] A. Azimah & Yudho Giri Sucahyo, ""Penggunaan Data warehouse dan Data Mining untuk Data Akademik: Sebuah Studi Kasus pada Universitas Nasional", Jurnal Sistem Informasi MTI UI Vol. 3 – No. 2 – Oktober 2007, Jakarta, Indonesia.

[9] Ernawati, 2007, "Penggalian Kaidah Asosiasi Kuantitatif Prestasi Akademik Mahasiswa dengan Jenis Kelamin dan NilaiTest Masuk Mahasiswa". Jurnal Teknologi Industri Vol. XI No.1 Januari 2007.

[10] L. Haryanto & P.H. Prima Rosa, "Prediksi Calon Mahasiswa yang Tidak Mendaftar Ulang dengan Metode Pohon Keputusan", Prosiding Digital Information and System Conference, Universitas Kristen Maranatha, Bandung, Indonesia, 2009.

[11] N.M.P. Wirati & P.H. Prima Rosa, "Klasifikasi Latar Belakang Mahasiswa Berdasarkan Prestasi Akademiknya dengan Metode Pohon Keputusan", Prosiding Konferensi Nasional Sistem Informasi 2008, Universitas Sanata Dharma Yogyakarta

[12] Thomas M. Connoly & Carolyn E. Beg., Database Systems A Practical Approach to Design, Implementation and Management, 5th edition, Scotland: Addison Wesley, 2005.