ETL Preprocessing with Multiple Data Sources for Academic Data Analysis

Gant Gaw Wutt Mhon
Faculty of Information Science
University of Computer Studies,Yangon
Yangon, Myanmar
gantgawwuttmhon@ucsy.edu.mm

Nang Saing Moon Kham
Faculty of Information Science
University of Computer Studies,Yangon
Yangon, Myanmar
moonkham@ucsy.edu.mm

Abstract

More and more, the needs of academic data analysis are requiring in educational settings for the purpose of improving student learning and institutional effectiveness. In the education world, testing and data are driving decisions for what knowledge and skills students should be learning and how students' learning relates to their learning outcomes. On that occasion, a better option for building a machine learning model is to get effective data preprocessing concepts. For these reasons, this paper describes the very first step of the main research work which considers the correlations between students' academic performance, behavior and personality traits to reveal the presence of an intriguing way. Intuitively, this paper proposes the uses of Extraction-Transformation-Loading (ETL) in the preprocessing stage to collect and analyze of students' data from multiple data sources. In this system, data is collected from multiple data sources based on the structures which are used as a testbed. Students' demographic data and assessment results from Student Information System (SIS), logs of their interaction with Moodle are used for data collection. Then aggregating with Web logs also captures student behavior that is represented by daily summaries of student clicks based on courses and by their actions.

Keywords: *machine learning, data preprocessing, ETL, multiple data sources, demographic data, assessment results, Moodle logs, Web logs*

I. INTRODUCTION

Many new pathways and insights for institutional effectiveness and the learning sciences are opening up due to the growth of information and educational technologies like Learning Management Systems (LMS), SIS. These technologies captured as 'digital breadcrumbs' from sources such as personality profiles, learning outcomes and behaviors; hence various analytics systems are emerged for improvement in institutional decision making, advancements in learning outcomes for at-risk students, significant evolutions in pedagogy more accurately and easily. Over time, their digital records may be augmented with other information, including financial and awards, involvement on campus, disciplinary and criminal reports, and personal health information. This increasing amount of admitted student data available on various data sources, the new technologies for linking data across datasets, and the increasing challenges need to integrate structured and unstructured data are all driving new aspects. Then, data collection and preprocessing are the most important and essential stages to acquire the fine and final data from multiple data sources that can be lead to correct and suitable for further data mining tasks.

By seeing current student data situation, data preprocessing becomes more crucial part and plays as a key concept of a system .There are a number of data preprocessing techniques rely on requirements and features of system's data model. Since last decade, ETL process became fruitful to flow preprocessing step smoothly. After data preprocessing, consistent data will be pulled from multiple sources and loaded to data warehouse. After retrieving some meaningful and knowledge information, the system needed to do for understanding and identifying the correlation rules between student assessment, behavior and personality traits with machine learning algorithm. There are many aspects to achieve in academic system. But the main idea of the proposed research work is that clustered analysis result are firstly explored based on students' academic performance and behavior of each student. After that, with their personality results from online survey test are merged to conduct the correlated rules between them.

The introduction of today's educational technologies situation and advantages, ETL preprocessing with multiple sources are presented in this section. The remainder of this paper has been arranged in different sections. Section 2 describes the related researches in this area. Then section 3 implements terminology of ETL. Implementation and

Experiments will be presented in section 4. Conclusion is made in the final section.

II. Related Work

In recent past, most of the related works are implementation of predicting academic achievement based on personality model and then focusing on students' previous marks, other historical data and students' behavior analysis by using data mining techniques approaches. Due to the development of educational technologies, there are many statistics and evolution settings to provide better academic environment by tracking, aggregating and analyzing student profiles along with all of the digitalized data of students.

Some of the research works with ETL processing in higher education are reviewed in this section. There are various implements in the processing of data in ETL process to suit with the requirements of the system. To qualify for work in academic data analysis model with current data situation, building preprocessing step is one of the challenges and important parts of a model. In [1], the authors evaluated student behavior clustering method based on campus big data with density based clustering method which is parallelized on the Spark platform and applied to subdivide student behavior into different group using historical data are as source in digital campus shared database. This proposed approach is to study for universities to know students well and manage them reasonably and improved algorithm is also effective. And then, the authors proposed ETL as data acquisition and preprocessing by integrating multi-source data before loading to the target system. Because this has been the way to process large volumes of data that it can scale cost effectively.

In [2], the authors explored the framework of the modern data warehouse with big data technology to support decision making process in academic information system. To reduce difficulties associated with traditional data warehouse, they designed a decision support system for big data by involving Hadoop technology. They also discussed ETL architecture based on the characteristics of the traditional data warehouse technology which cannot handle unstructured data and modern data warehouse. In [3], the authors proposed a framework for development of flexible educational data mining application to facilitate self-discovery of rules and trends from educators with little technical skills to various user types. This framework is also demonstrated with utilizing tools which are used for data mining analysis within a LMS. The authors used ETL stage as preprocessing for extracting the information from the LMS or e-learning system to do analysis much easier.

In [4], the authors implemented the case study in business intelligence framework by using data integration and ETL and then described its significance for better higher education management. They detected the functions of data integration and ETL tools and described how the correlated to each other to develop business intelligence. Because of data integration is one of the important components and ETL is the common steps to integrate data and transform it to targeted system from different academic data sources for higher education. They demonstrated on the Graduate Studies Management System (GSMS) database of University Technologies of Malaysia to store basic information of student. To ensure the data is reliable and could make decision accurately with ETL process, this database must be integrated with various sources which have different platform and format.

III. ETL Terminology

There are solutions coming up for better data integration and data warehousing because of data management has been evolving rapidly. ETL is a popular architectural pattern and used for data process with necessary integration from heterogeneous and distributed data sources with different format. ETL procedures is needed to dedicate based on the design and implementation of the system because designed ETL process are expensive to maintain, alter, and upgrade, so it is crucial to make the right choices in terms the best innovation will certainly be used for developing and preserving the ETL procedures.

A. Data Warehouse and ETL

Data Warehouses are used in higher education for decision making to take an action and to predict risk and opportunities of students. It is a process of collecting and managing data from varied sources to achieve reliable, accurate and meaningful information from various data sources. There are three steps to follow before storing data in a data warehouse, which is called ETL. This three-step process takes place when information from one system needs to be moved into a data warehouse environment. Data extraction involves extracting data from homogeneous or heterogeneous sources; data transformation processes data by data cleansing and transforming them into a proper storage format/structure for the purposes of querying and analysis; finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, data lake or a data warehouse.

B. ETL with Big Data

The need for ETL has increased considerably due to modern data analytics operations often have to process with rise in data volumes as quickly as possible. Therefore traditional ETL approach has to develop to function of more constrained data storage and data availability because it can slow down the process significantly for systems optimization. Apache Hadoop provides a cost-effective and massively scalable platform for ingesting big data and preparing it for analysis [5]. Using Hadoop to offload the traditional ETL processes can reduce time to analysis by hours or even days.

Moreover, ETL on platform like Hadoop and Spark give ETL a new look because it changes the cost structure around harnessing big data and save time too. [6] From their study, they recognized each ETL process instance handles a partition of data source in parallel way to improve further ETL performance facing the big data by using parallel/distributed ETL approach (Big-ETL). The researchers developed ETL functionalities which can be run easily on a cluster of computers with Map Reduce (MR) paradigm. Apparently, by using this approach on very large data integration in a data warehouse to qualify in good performance of ETL process significantly.

IV. IMPLEMENTATION AND EXPERIMENT

In today's education world, for better or worse, testing and data are driving decision more and more as universities seek to be evidence-driven by using all of student related data. To fulfill their needs and know their weakness in time for their educational life, there are various new trends and aspects in need to collect, analyze, interpret, store, track, aggregate the increasing amount of admitted student data available from various data sources.

Apparently, there need to explore accurate and effective way of data collection and preprocessing for linking data across datasets from multiple data sources. Accurately, the proposed system collects student related data from three different data sources by using ETL as preprocessing stage for the collection analysis of students' data. Demographic data and assessment results are collected from SIS. It is very time saving and effective way to observe all students' related data and many suggestions and ideas can get from these data which are used to investigate the academic growth. Then logs of their interaction with Moodle are also collected which is one of the best data sources to know about students' behaviors based on their interested course and relevant teachers. To know their academic behavior more accurately, aggregating with Web logs which are represented by daily summaries of student clicks based on courses and by their actions. In above Fig 1, overview of proposed system for academic data analysis with ETL preprocessing is described in details. For research environment, University of Computer Studies Yangon (UCSY) is used for students' datasets as sample.

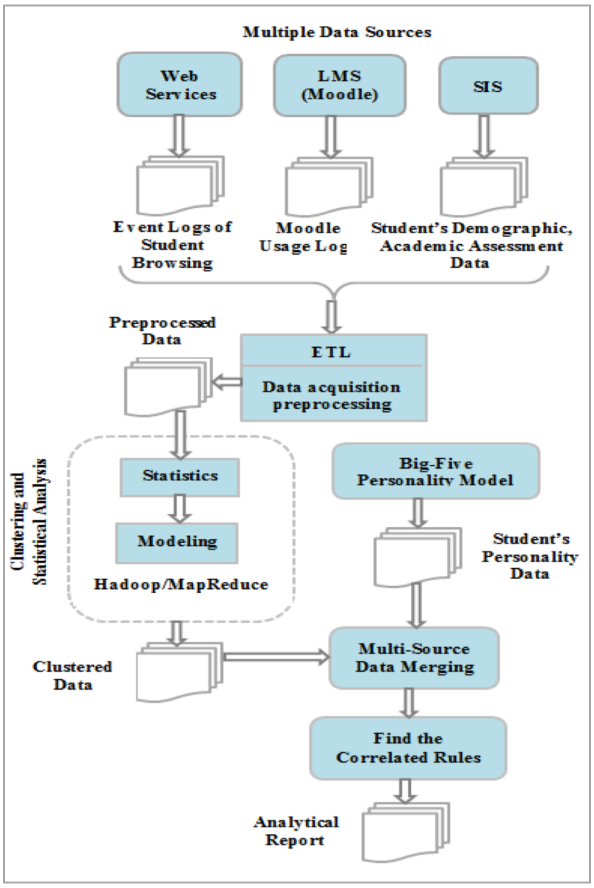
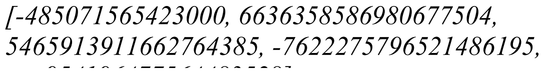

Multiple Data Sources
Web Services
LMS (Moodle)
SIS
Event Logs of Student Browsing
Moodle Usage Log
Student's Demographic, Academic Assessment Data
Preprocessed Data
ETL
Data acquisition preprocessing
Clustering and Statistical Analysis
Statistics
Modeling
Hadoop/MapReduce
Big-Five Personality Model
Student's Personality Data
Clustered Data
Multi-Source Data Merging
Find the Correlated Rules
Analytical Report

Figure 1. Proprosed System of Academic Data Analysis with ETL Preprocessing

A. Data Collection and Implementation of ETL

As in fact of the data processing of the system, each student represents the portion of data which are correlated and these are from multiple data sources of the university. Therefore, building ETL preprocessing based on the proposed system's used data scale detect a formal representation model for capturing the ETL process that map the incoming data from different data sources to be in a suitable format for loading to the target system. In this ETL process, the system prepares the data model for building datasets of machine learning algorithm. Therefore, all of the columns are not processed in transaction data.

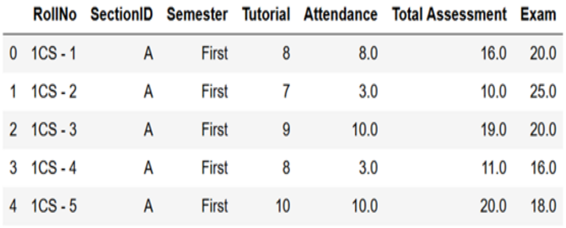
Although many other features are existed, the system defines some of attributes which are more dominant for data processing of the system especially helps to know an individual in academic life. In the extract phase of ETL process, the system will store the transaction data in staging. Based on the demographic data as a sample, only *EnrollmentNo, RollNo, ParentalEducation and MathScore* will be loaded into the warehouse model among the attributes such as *RollNo, Name, Email, SectionID, EnrollmentNo, NRC, Address, FatherName, ParentalEducation and MathScore*. *EnrollmentNo* is taken as unique key for the students and *RollNo* will be taken to link with other tables. Thereafter, anonymization is necessary for *EnrollmentNo* due to student information privacy. According to the ethical and privacy requirements, this system will also add annonimization for the data privacy. Experiment of ETL process with sample datasets is conducted on Python. After applying hash function to *EnrollmentNo* of some student, the forms of anonymized data are emerged as in below Fig 2:



[-485071565423000, 6636358586980677504,
5465913911662764385, -7622275796521486195,

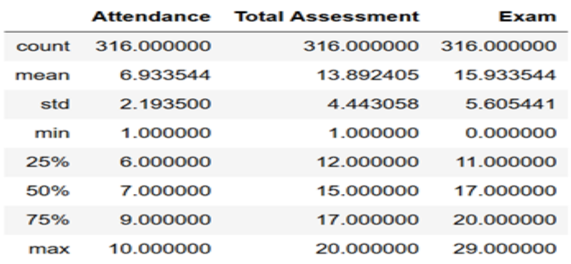
Figure 2. Example of Anonymized Data with Hash Method

The sample assessment datasets are received by course code. These datasets are extracted by each course which is described in Fig 3. From this data, attendance and other assessment results of each student are needed to compute in sum up and score by *RollNo.* Then these computed scores are to be taken into database which is also described in Fig 4.



	RollNo	SectionID	Semester	Tutorial	Attendance	Total Assessment	Exam
0	1CS - 1	A	First	8	8.0	16.0	20.0
1	1CS - 2	A	First	7	3.0	10.0	25.0
2	1CS - 3	A	First	9	10.0	19.0	20.0
3	1CS - 4	A	First	8	3.0	11.0	16.0
4	1CS - 5	A	First	10	10.0	20.0	18.0

Figure 3. Example of Students' Assessment Data by Course



	Attendance	Total Assessment	Exam
count	316.000000	316.000000	316.000000
mean	6.933544	13.892405	15.933544
std	2.193500	4.443058	5.605441
min	1.000000	1.000000	0.000000
25%	6.000000	12.000000	11.000000
50%	7.000000	15.000000	17.000000
75%	9.000000	17.000000	20.000000
max	10.000000	20.000000	29.000000

Figure 4. Computing score of a Student by Course

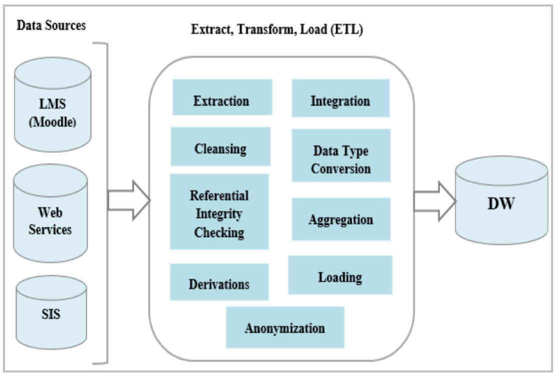


Figure 5. Steps within ETL Process

In transformation phase is cleaninging and confirmation steps of ETL. Its purpose is to gain accurate data which are correct, complete, unambiguous and consistent. In loading phase, warehouse model of this system is not like traditional warehouse and its purpose is to build the dataset for model building in analytical process. After that, the aggregation model (Data Warehouse) is explored but we need to change data model to suit for Machine Learning Algorithm. In Fig 5, ETL processing steps of the system is also described.

B. Density-Based k-means Clustering Method

After the system proposed ETL as data acquisition and preprocessing by integrating multi-source data, the system evaluates the cluster analysis results as a performance of students in different groups from students data along with their academic digitized record by using density-based k-means clustering method as improved k-means clustering algorithm on the Hadoop platform and applied to subdivide student academic performance and behavior into different group.

Then, the system extracts the valuable attributes that reflect in evaluation of cluster analysis results. With traditional k-means clustering and improved k-means clustering algorithm, the system analyses the statistical results and compare the accuracy between them. To provide more accurate implementation method, k-means algorithm based on density partitioning is needed. In fact, the system construct initial clustering center set based on density not k value. Thereafter improved k-means clustering algorithm is developed by applying the selected initial clustering center. The method first determines a similarity measure method suitable for student data to normalize data and then builds initial clustering center set based on the density.

The density of the data object X_i in the sample set $\boldsymbol{S=\{X_1, X_2, \dots, X_n \mid X_i \in R^t\}}$ is defined as the number of samples within the ***Eps*** neighborhood of $\boldsymbol{X_i}$. The density reflects the intensity of the sample points in the neighborhood. The density threshold ***minPts*** is the specified division of the core and the isolated point of the density range, which can be artificially set. The density parameter $\boldsymbol{N_{Eps}(X_i)}$ of the sample is calculated as in 1:

$$N_{Eps}(X_i)=\{X_j \in S|0 \leq D(X_i, X_j) \leq Eps, j=1,2,\dots,n\} \quad (1)$$

Where $D(X_i,X_j)$ represents the distance between two samples in ***S***. Actually, k-means doesn't allow development of an optimal set of clusters and so for effective results, the need to decide on the clusters before. Then, the statistic of traditional k-means clustering is difficult to predict k-value, different initial partitions can result in different final clusters.

C. The Big-Five Personality Test

Finally, the students' personality results and the clustered results are merged with their student identification number and then analyze with Apriori association rule algorithm, the most classical and important algorithm for mining frequent item sets, to conduct the correlated rules between them which are firmly and consistently associated. In this system, 50-questions (items) inventory is used to measures as individual on the Big-Five personality dimension that is recreated from the Big-Five Inventory (BFI). To achieve students' personality results, they need to take online personality survey test from Moodle.

The development of questionnaires measuring the Big-Five personality traits is common in psychology research for different reasons. There are countless personality tests designed with many formats. But the Big-Five personality model is one of the popular models and it also called as the Five-Factor Model (FFM). The personality survey test is designed to assess the Big-Five factors of personality: *Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N), and Openness to Experience (O)* [7,8]. In the system, the items on the BFI scale were also scored with a *Likert-scale* (5-point) from "strongly disagree" to "strongly agree" and each factor represents ten questions respectively. Some example phases are:

- *I am talkative. (E)*
- *I feel little concern for others. (A)*
- *I am always prepared. (C)*
- *I get stressed out easily. (N)*
- *I am an inventive. (O)*

Each personality trait is associated with a set of statements and the score of a trait is calculated for each user on the associated questions. After data collected through the questionnaires, Statistical Package for the Social Sciences (SPSS) is used for statistical analysis. Firstly, the system need to test the reliability and validity of the questionnaire in data collection by reason of the results of research quality. In addition, reliability estimates conduct the amount of measurement error and validity determines the questionnaire compiled it valid or not in a test. In this paper, 30-participants are only participated as sample to test the reliability and validity of system's personality survey test questionnaires.

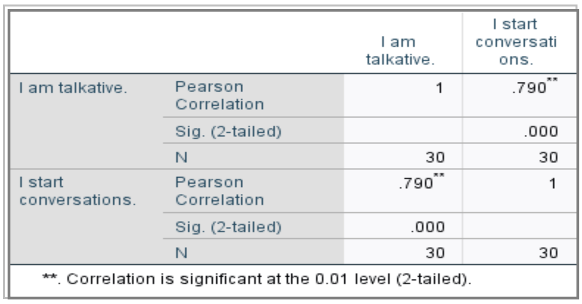
Cronbach's alpha is the most widely used objective measure of reliability or internal consistency and it is a simple way to measure whether or not a score is reliable. But in some questions often contain some items with negative sense which are need to be reversed score before run Cronbach's alpha in SPSS. Reverse scoring means that the numerical scoring scale runs in the opposite direction and so the system need to reverse-score all negatively-keyed items. For interpreting alpha for *Likert-scale* question: 0.70 and above is good, 0.80 and above is better, and 0.90 and above is best [9].

In Table 1, the reliability of the questionnaire (30-participants to complete the N= 50-item) in data collection are tested with compare result based on reverse-score are described. After reverse-score the negative worded questions, the scale had an internal consistency is α=.76. However, a high coefficient alpha does not always mean a high degree of internal consistency and it is also affected by the length of the test.

Table 1. Comparison of Reliability Statistics

	Cronbach's Alpha	**Cronbach's Alpha Based on Standardized Items**	**N of Items**
Without Reverse -Score	.588	.490	50
With Reverse-Score	.762	.756	50

As pointed out earlier, test the validity of the questionnaire is conducted using Pearson Product Moment Correlation (PPMC) using SPSS. It is the test statistics to measures the statistical relationship, or association, between the continuous variables and based on the method of covariance. This validity test is done by correlating each item questionnaire scores with the totally score. The sign of the correlation coefficient r indicates the direction of the relationship, while the magnitude of the correlation indicates the strength of the relationship in the range -1 to 1 [10]. In Fig 6, the validity of the questionnaire (N=30-participants with sample questions of Extraversion) in data collection are tested.



		I am talkative.	I start conversations.
I am talkative.	Pearson Correlation	1	.790**
	Sig. (2-tailed)		.000
	N	30	30
I start conversations.	Pearson Correlation	.790**	1
	Sig. (2-tailed)	.000	
	N	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 6. Pearson's Correlation among the Items

In fact of above figure, Pearson's correlation of two questions is (r=0.790) which indicate that the strength of association between the variables is very high. Then Sig(2-tailed) value is 0.000<0.05. This means there is a statistically significant correlations between two variables. After the testing of reliability and validity, with the factor analysis on scale items, each student has five traits which are scored on a continuum from high to low. According to calculated result, score description of each trait is as an example in extraversion, high scores tend to be very social while low scores prefer to work on their projects alone.

Finally, the main idea of the proposed approach is explored by merging based on students' identification number to find the correlations between the students' personality results and the clustered results. Then to conduct the correlated rules between them which are firmly and consistently associated by analyzing with Apriori association rule algorithm. Therefore, from the final resulted correlated rules, students' academic performance and behaviors are correlated or not on their personality results.

V. CONCLUSION

Higher education is working in a more and more complex and there need to explore accurate and effective way of data collection and preprocessing for linking data across datasets from multiple data sources. This paper described on the ETL preprocessing step for the collection and analysis of student academic digital record from multiple data sources to ensure the data is reliable and could contribute to decision making for work in academic data analysis model. Then, to qualify for the results of research quality in data collection, a part of measuring the personality test questionnaire are also described in this paper.

REFERENCES

[1] Ding, Dong; Li, Junhuai; Wang, Huaijun; Liang, Zhu; "Student Behavior Clustering Method Based on Campus Big Data"; IEEE, 2017.

[2] Santoso, Leo Willyanto; "Data warehouse with big data technology for higher education"; Procedia Computer Science; Elsevier, 124, 93-99, 2017.

[3] DeFreitas, Kyle; Bernard, Margaret; "A framework for flexible educational data mining"; Proceedings of the International Conference on Data Mining (DMIN); 2014.

[4] Rodzi, Nur Alia Hamizah Mohamad; Othman, Mohd Shahizan; Yusuf, Lizawati Mi; "Significance of data integration and ETL in business intelligence framework for higher education"; 2015 International Conference on Science in Information Technology (ICSITech), IEEE, 181-186, 2015.

[5] Big Data Analytics; White Paper - "Extract, Transform, and Load Big Data with Apache Hadoop"; 2013.

[6] Bala, Mahfoud; Boussaid, Omar; Alimazighi, Zaia; "Big-ETL: extracting-transforming-loading approach for Big Data"; Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), 2015.

[7] O'Connor, Melissa C, Paunonen, Sampo V,"Big Five personality predictors of post-secondary academic performance", Elsevier, 2007.

[8] Gosling, Samuel D; Rentfrow, Peter J; Swann Jr, William B; "A very brief measure of the Big-Five personality domains"; Elsevier, 2003.

[9] Tavakol, Mohsen; Dennick, Reg;" Making sense of Cronbach's alpha"; International journal of medical education, IJME, 2011

[10] Fosse, Thomas Hol; Buch, Robert; Säfvenbom, Reidar; Martinussen, Monica; "The impact of personality and self-efficacy on academic and military performance: The mediating role of self-efficacy"; Journal of Military Studies, 2015.