FACULTY OF ENGINEERING

SCHOOL OF COMPUTING

SEMESTER 2/20192020

SCSP3213: Business Intelligence

LECTURER: Dr. NOORFA HAZLINA MUSTAFA

SECTION: 02

TITLE: Project 1 ETL

GROUP MEMBERS:

1. AIMAN SHAQUAN BIN MOHD YUSOF (A18CS0021)
2. AMIR FIRDAUS BIN KHORUDIN (A18CS0027)
3. AHMAD SYAMIL AIMAN BIN MEZUKI (A18CS0019)
4. NUR ATIKAH BINTI SAWAL (A18CS0188)
5. MOHAMMAD AMIRUN HAZIQ BIN MOHAMMAD FADZLI (A18CS0109)

Abstract

To personalize the learner's experience, the adaptive e-learning system (AE-LS) research has long focused on the learner paradigm and learning experiences. There are several unanswered problems which make it difficult for trainee teachers to gain suitable knowledge about the behavior of the learner. The evolution of Learning Analytics (LA) creates new avenues for solving AE-LS problems. In this paper, we proposed a Business Intelligence system for AE-LS to more accurately track and handle the learner's performance. A data warehouse model that responds to these concerns is indicated by the proposed ALS architecture. This describes unique indicators and measurements that allow teachers and educational administrators to assess and interpret the behaviors of the learner. The adaptive e-learning analytical method (AE-LAS) has the ability to provide a predictive view of possible issues by evaluating these interactions. These forecasts are used to measure the adaptation of the presentation of the information and optimize the learning process efficiency.

Table of Content

Chapter 1: Introduction

- **Introduction**

Nowadays, live online learning stuff including webinars and virtual classrooms has been a new normal in our life. The distance between the lecturers and students was not a barrier anymore since online learning has started to become a norm and many online courses continue to grow. The Virtual Learning Environment (VLE) has given a lot of opportunities to students and lecturers to explore more libraries without going to a physical library. Elliot Maise said that we need to bring learning to people instead of people to learning. Hence, VLE is the perfect platform to make sure everyone is accessible to all the learning resources and expanding the knowledge virtually without meeting each other.

- **Background of problem**

It is challenging to advocate for and develop programs that respond to the needs of crisis-affected learners without quality data and evidence. Accurate, timely, reliable, and usable data is required not only to retort to crises but also to inform evidence-based planning for preparedness and prevention. Without quality data and evidence, national and native governments' ability to effectively harden and reply to crises is hindered, data and evidence are essential for planning processes the smallest amount bit levels of education systems. It is essential to monitor the student involvement and evaluate their progression over a period of time. Data analysis is very important for the number of active students at a university. A university produces graduates every year and has increased the number of students. That way in terms of interpreting important information, the amount of data stored would increase and affect. Until now, top management spent so much time acquiring structured, precise, and complete information to evaluate student data, particularly about total active students. This is because data retrieved from many different sources is generated by the data source. In support of the strategic decision-making process, active student data is required.

For this project, we have used an anonymized Open University Learning Analytics Dataset (OULAD) that contains the data about students, courses, and their interaction on the Virtual Learning Environment (VLE). This dataset will consist of thousands of student information including the courses that they took and also the data on the courses as well as the data student activities on the courses. We can see that the data are not in the highest quality because of the many noises in the data. Thus we need to use ETL architecture and implement it in Talend.

- **Objective**

The objective of this case study is mainly

I. To determine a suitable proposal for the case study
II. To propose a new Business Intelligence (BI) Architecture, Extract, Transform, and Load (ETL) architecture, as well as a Use Case diagram
III. To deepen the understanding of all related methodologies such as data gathering, database designing, and practicality of data warehouse
IV. To implement the architecture of data warehouse design by using snowflake or star scheme

- **Scope**

During revising the case study, an anonymized Open University Learning Analytics Dataset (OULAD) is given to us for the purpose of designing the main processes for the data warehouse and implementing our ETL knowledge around the data. An appropriate data warehouse schema is being designed based on our new proposed Use Case diagram, ETL architecture, BI Architecture. As a means, our data warehouse processes will use Talend as a medium to ensure the smoothness of the work.

Chapter 2: Literature Review

- **What is a data warehouse?**

Data warehouse is a storage area where all the data in an organization is kept in a single place. This involves information from multiple sources as well as current and historical data of an organization. Data warehouse is also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis. Most of the data that passes through will go through a process of data cleansing to ensure data quality. Data warehouse is also used to centralized historical data before going through objective reporting, decision making, and ad hoc searches. Characteristics of a data warehouse are subject-oriented, integrated, time-variant, and non-volatile.

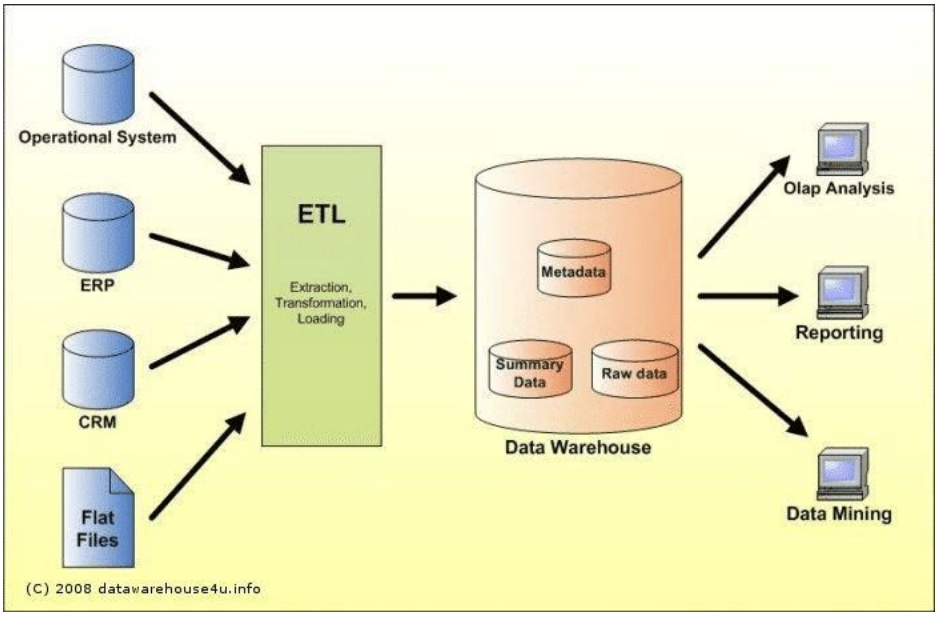


Figure 1.0: Data Warehouse Architecture

- **Previous study on Business Intelligence Architecture**

Business intelligence architecture is a framework to run business intelligence and analytics applications for an organization. Its main functions and processes are used to collect, integrate, store and analyze data. It is also used to present information on business operations and trends to the business users. The components of a business intelligence architecture are data integration and cleansing tools, including extract, transform and load (ETL), data virtualization, and data cleansing, analytics and storing data which includes Data Warehousing, data mart, and data lake, business Intelligence tools including queries, Online Analytical Process (OLAP) and data visualization and lastly information delivery which includes dashboards and reports.

Based on the previous study about Analytic Information Systems in the context of Higher Education is also interpreted in this project. [3] The same methods in the Business Intelligence Architecture are used in the BI system for the teachers which collect and analyze academic data of the Virtual Learning Environment (VLE). The business intelligence system for teachers is composed of ETL processes which consist of student information, vle interaction and course information, Data Warehousing which stores all the data needed for decision-making, and Data dashboard which provides insights to the data. This is also very related to our case study which is to provide more insights to the teachers based on the same area of data which is VLE interaction, academic data about students, and their interactions.

- **Previous study on Extract, Transform, Load ETL Architecture**

Extraction-transformation-loading (ETL) tools are pieces of software responsible for cleansing, customization, reformatting, integration, and insertion of data from various sources into a data warehouse. Building the ETL process is theoretically one of the main warehouse building tasks. It is complicated, time-consuming, and requires much of the efforts, expenses, and resources of developing the data warehouse project. The design of a data warehouse includes a tight emphasis on three main areas of understanding which are the source area, the destination area, and the mapping area (ETL processes). The source area has standard models such as the entity-relationship diagram, and there are standard models such as the star schema in the destination area, but so far the mapping area does not have a standard model.

Based on the previous study [3], the author highlighted that the data that is stored in the data warehouse is needed for the decision-making of the teachers. The data comes from different sources such as academic, virtual classroom, and learning resources, and also from different types (structured and unstructured). ETL processes are needed to extract, transform, and load, the academic data that is composed of student, teacher, and subject information. This process is the main processes in the data warehouse that extract, transform, and integrate data into the data warehouse. The large volume of data, extracted from the interactions of the students with the VLE when learning, can be used to provide such information to teachers.

Chapter 3: Methodology

- **Example SDLC**

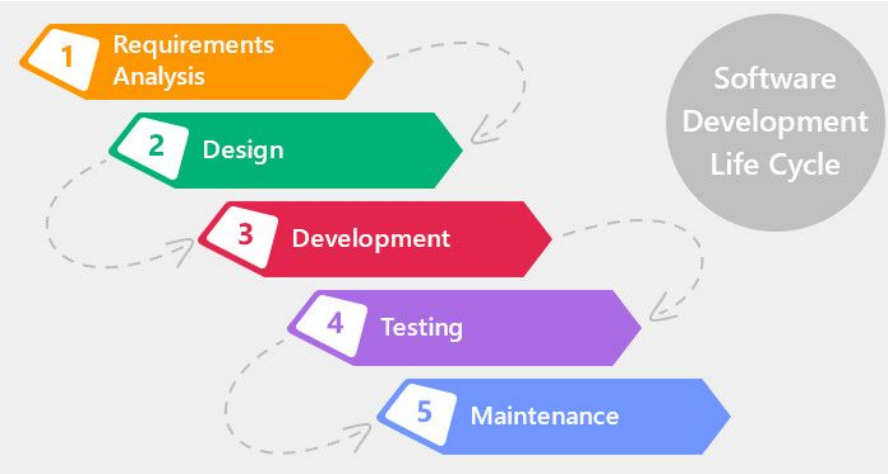


Figure 2.0: System Development Life Cycle

- Requirement gathering and analysis:
  Gather and figure out all requirements that are needed for this data warehouse project. Once all the requirements are gathered, we then can analyze the requirement for their validity in the project. This step is important as it will be the guidelines for the next step of the model.

- Design :
  In this phase, we need to define all the hardware and software needed for the project. This phase acts as a trail for us to ensure all our needs are fulfilled for the project.
  - Software:
    - Talend
    - Tableau
    - MS Excel (as a text editor)

  - Hardware:
    - (Minimum) Laptop 256 GB Storage Capacity 4 GB Ram

- Development:
  For this phase, we will use the ETL process to develop this project. We will be using Talend software to do this process.

- Testing:

After the development phase, testing is a must to ensure all the objectives needed are solved. There will be some testing implemented to ensure the work is ready for production.

- Maintenance:
  As for the last phase, the work or data warehouse will be monitored in case some problems are occurring. There also will be some precautions such as data backup, data validation, and data logs to ensure unnecessary problems are prevented.

- **Planning (include Gantt Chart and team working handle)**

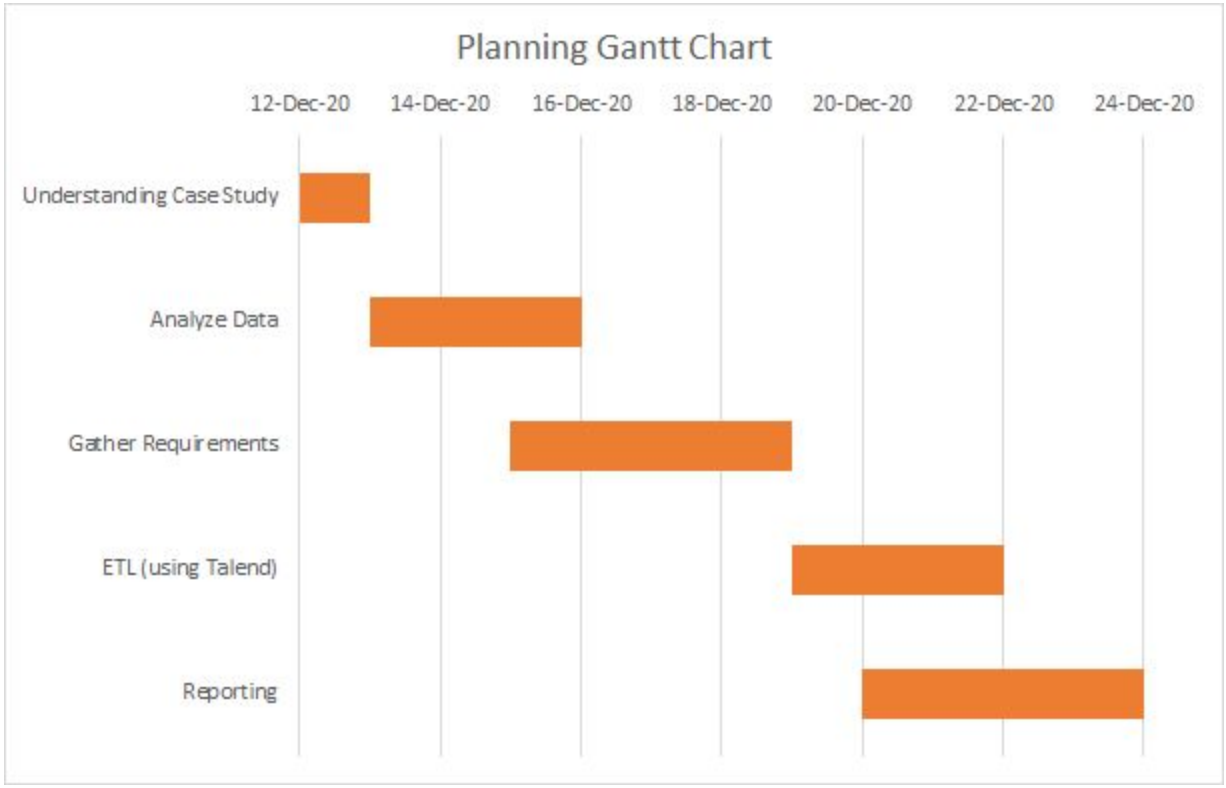


Figure 3.0 : Planning Gantt Chart

Team Working Handling

1. Understanding the case study and given data are done individually and then we discussed our understanding and added any information related.
2. Since everything is online right now, we were having our discussion using Discord as a platform for our meeting.
3. Requirements gathering has been done simultaneously with all group members.
4. Data pre-processing such as removing whitespaces, filtering unnecessary values, keep the integrity of data and data type standardized, and replacing the data with appropriate ones has been done by two members, while the other three, proceed with preparing the report, setting workflows for the group projects, analyzing cleaned data, preparing data visualization, and monitoring the project process.

Chapter 4: Design

- **Business Intelligence Architecture (case study)**

  Nowadays, we can see business intelligence is not just applied to the business industry, but also in academic universities and higher institutions management. This shows the range of business intelligence applications and the importance of applying them to different sectors. In Malaysia, we can see many Higher Educational Institutions(HEI) implementing business intelligence and information systems through their application of information technology systems. These are important steps to increase efficiency and optimizing performance (Owusu, Acheampong, et al., 2017). Many challenges will be faced such as handling the system competency, especially in this fast-growing era. To tackle all the problems that are appearing, adoption of business intelligence to make sure the system competency can handle dynamic data and scaling the size of the system accordingly. The emergence of the research framework Technology-Organization-Environment (TOE) framework and the Diffusion of Innovations (DOI) theory have increased the interest of all universities to indulge in business intelligence applications in their administration. One of the applications of a business warehouse is data warehouse building. The architecture of the system will keep the data handling systems and ensure the end-user to get a more clear view of the data for the analyzing process (Leo and Yulia, 2017).

- **ETL Architecture (case study)**

  Extract, Transform, Load or ETL processes are one of the major processes in business intelligence success. One of the subprocesses of ETL are data integration or DI. DI can be understood differently from a different perspective. Such conventional ways are data consolidation, data propagation, data virtualization, data disparate, and data warehousing which are slightly different in some aspects (J. Sreemathy, et al., 2020). For our proposal, we decided to use data warehousing as a medium to reach our goal. With data warehousing, a holistic process involving cleansing, standardizing, and keeping data in a proper format can be achieved. To have a better success rate in ETL processes, we need to make sure we have enough understanding about technical terms such as database, data warehouse, data integration, and data preprocessing. It is a must to keep the project on track. There are three basic extraction levels which are full extraction, half extraction (up to date alert), half extraction (no update alert). For our data warehouse, we use full extraction at a logical level (“Data Extraction Technique”, 2020). For a conclusion, data integration is an important step to ensure a correct business decision can be made from the BI process.

- **Use Case diagram**

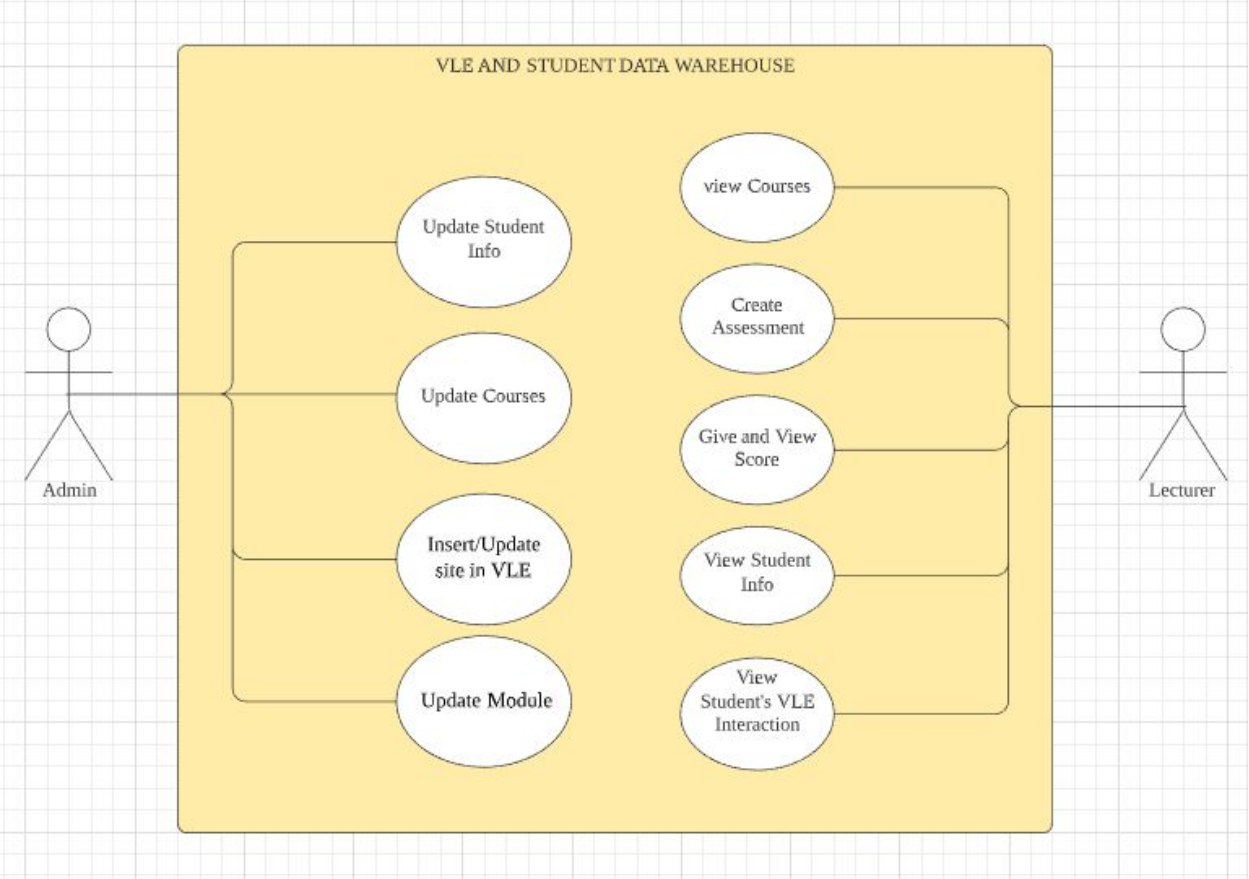


Figure 4.0: Use Case Diagram

- **Data Warehouse Design (Star or SnowFlake Schema)**

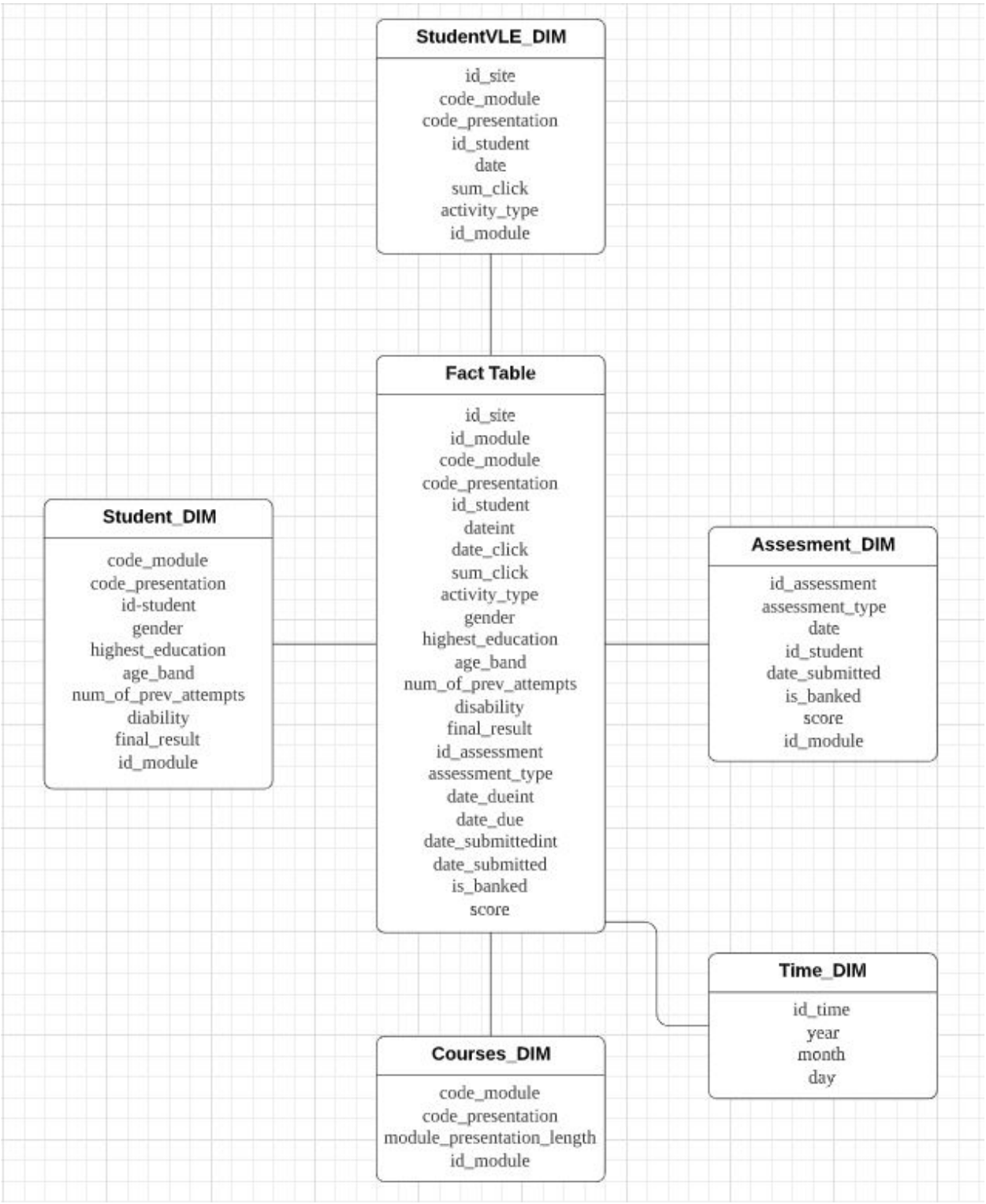


Figure 4.1 : Star Scheme

Chapter 5: Implementation

Transformation Table

File Sources	Changes	Implementation
Assessments.csv StudentAssessment.csv Courses.csv	Return module presentation length if due date for assessment is null	tmap: Relational.ISNULL(due_date)?module_presentation_length:due_date
Courses.csv	Remove whitespaces and add modules id	tmap: StringHandling.TRIM(code_module) code_module+code_presentation
StudentInfo.csv StudentRegistration.csv	Remove Students that has been unregistered	tmap: Relational.ISNULL(row2.date_unregistration)
StudentVLE.csv	Return 0 if sum click is null	tmap: Relational.ISNULL(row4.sum_click)?0:row4.sum_click
VLE.csv StudentAssessment.csv Assessment.csv StudentVLE	Change date value from int to actual date by adding to actual date since the start of the module	tmap: TalendDate.addDate(TalendDate.parseDate("yyyy-MM-dd",code_presentation),date,"dd")

The main function use in Talend

1. StudDIM

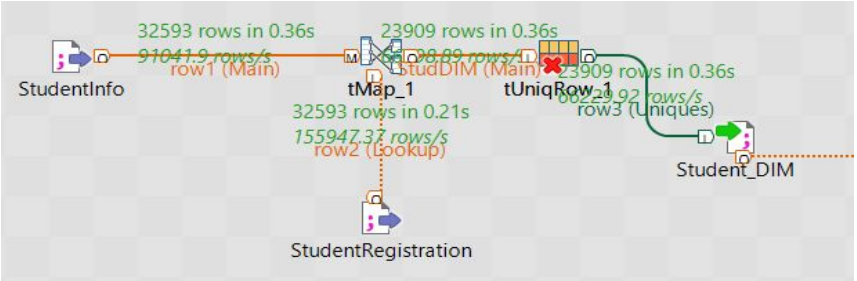


Figure 5.0: StudDIM implementation

a. A few tmap components have been implemented. The tMap_1 is used to join between two data which are StudentInfo and StudentRegistration. The output for registered students only is filtered using the below expression.

Figure 5.1: StudDIM Filter Expression

b. tUniqROw is used to filter duplicate data using the unique values which are code_module, code_presentation and id_student.
c. tFileOutputDelimited is used to write the result for the student dimension(StudDIM)

2. vle_DIM

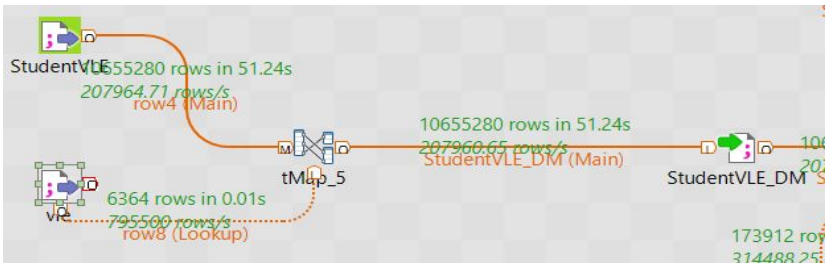


Figure 5.2: vle_DIM implementation

a. tMap is used to join between StudentVLE and vle data.Some data are extracted from both sources. The vle_DIM contains id_site,code_module,code_presentation,id_student,date,sum_click,activity _type and id_module.
b. Replacing null values of sum_click with 0

Figure 5.3: vle replacing the null value

c. tFileOutputDelimited is used to write the result for the vle dimension(vle_DIM)

3. Assessment_DIM

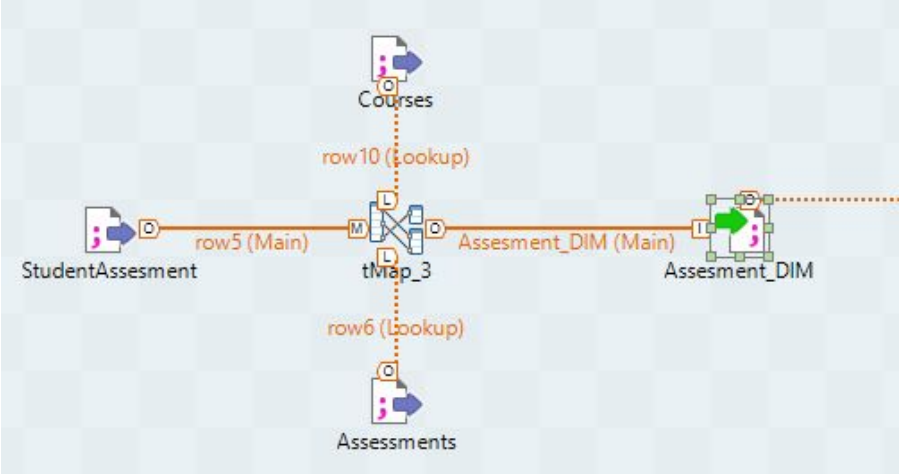


Figure 5.4: Assessment_DIM implementation

a. tMap was used to join together StudentAssessment, Assessment and courses data. The date for the assessment will enter the module presentation length based on the module if there is null. Expression below was used to fill the null value.

```
Relational.ISNULL(row6.date)?
row10.module_presentation_length:row6.date
```

Figure 5.5: Replacing null value of submission

b. Id_module was also generated by merging the module code and presentation code.

```
row6.code_module+row6.code_presentation
```

Figure 5.6: generating Id_module

c. tFileOutputDelimited is used to write the result for the assessment dimension(Assessment_DIM)

4. Courses_DIM

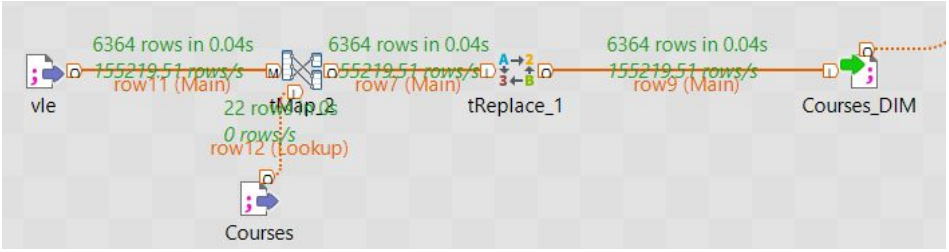


Figure 5.7: Courses_DIM implementation

a. tMap used to join between vle and courses
b. tReplace was used to find a specific code presentation and replace it with the first date of the semester.

InputColumn	Search	Replace with
code_presentation	"2013B"	"2013-02-01"
code_presentation	"2013J"	"2013-10-01"
code_presentation	"2014B"	"2014-02-01"
code_presentation	"2014J"	"2014-10-01"

Figure 5.8: Replacing new date for the reference in the fact table

c. tFileOutputDelimited is used to write the result for the courses dimension(Courses_DIM)

5. Fact table

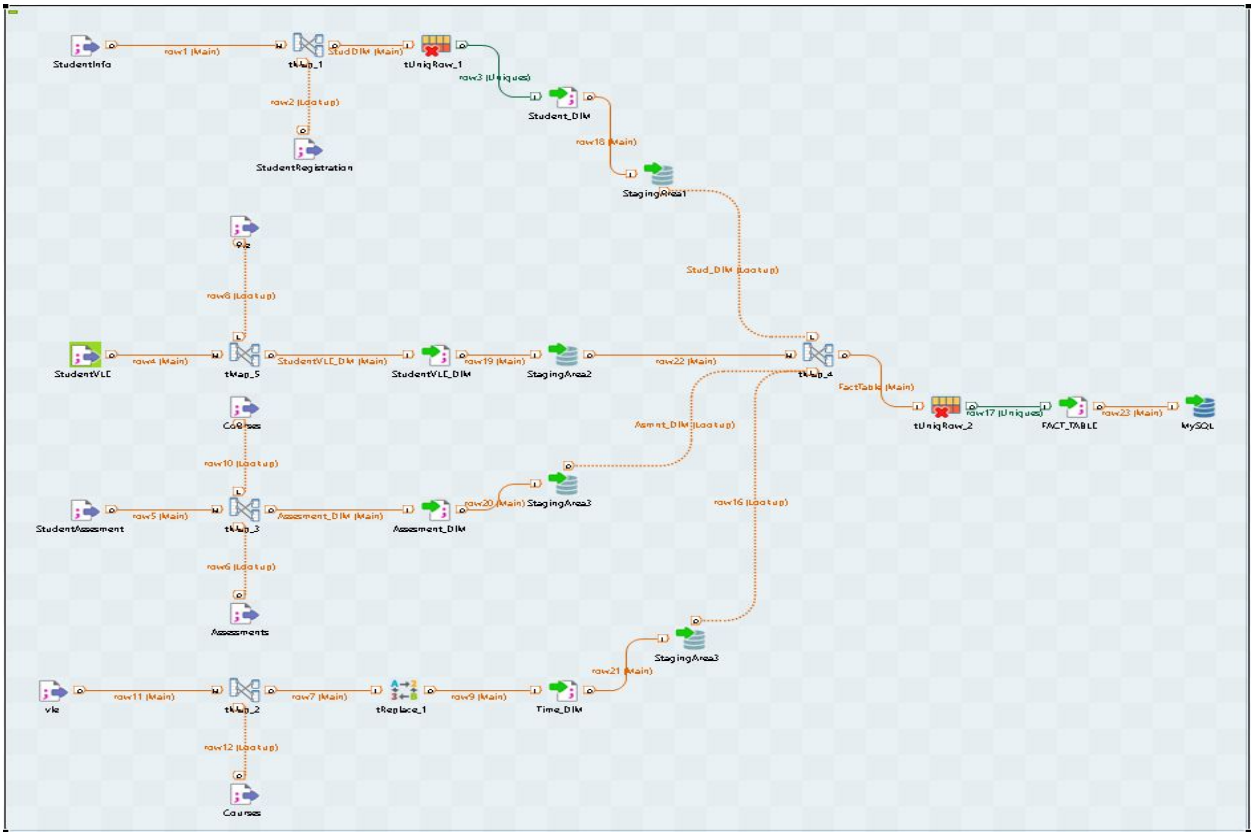


Figure 5.9: FactTable implementation

a. tMap and tUniqRow are used to join all the dimensions for the unique output with no duplicates after the process of extract and transform from the ETL architecture.

b. A formula used in the tMap output of the fact table, to translate integer data type date to actual date data type from every dimension that has

integer date values. Date also used the course module presentation as a reference for the year and the semester.

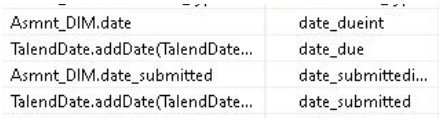
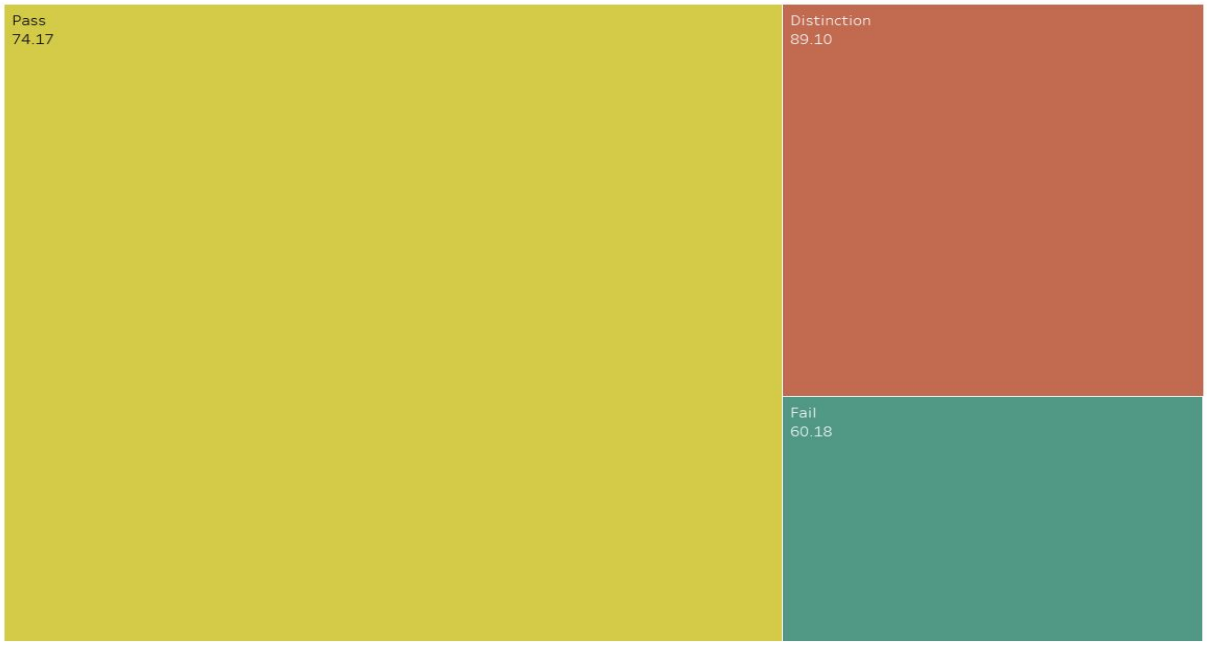


Figure 5.10: Replacing date integer with a real date

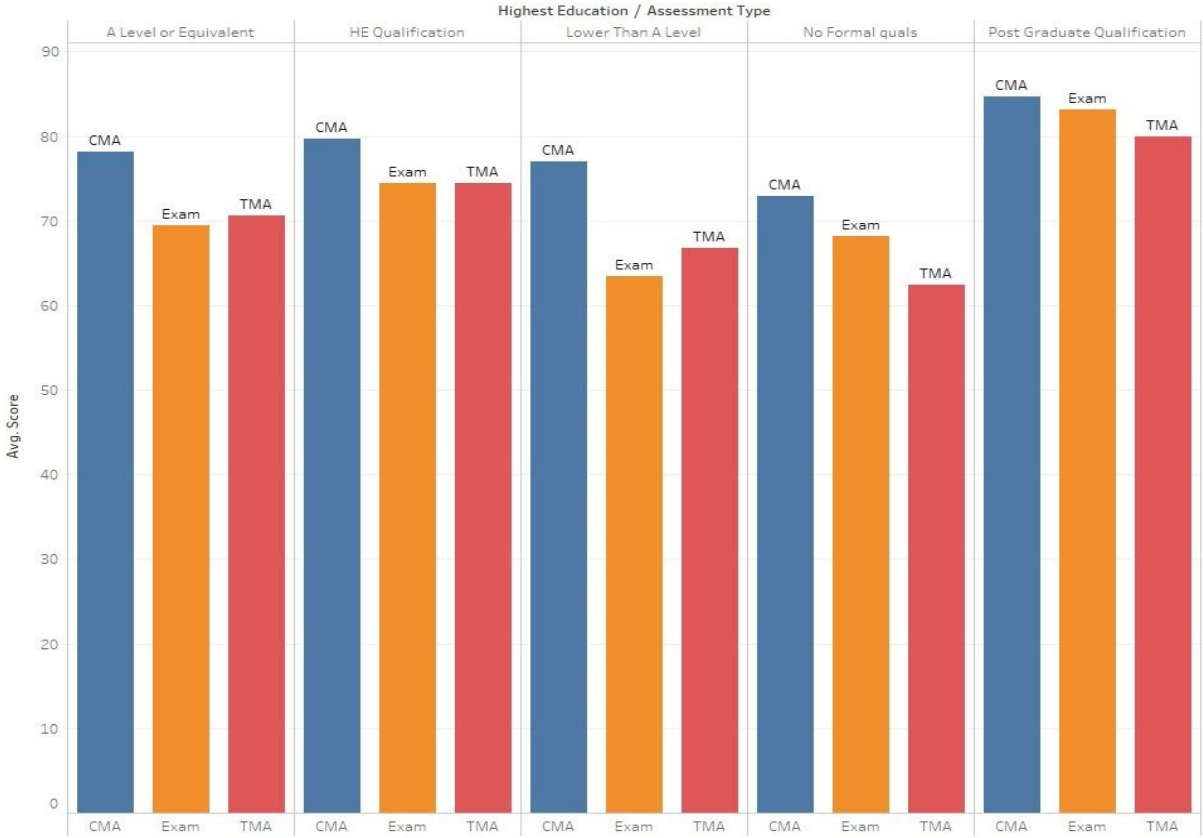
c. tFileOutputDelimited is used to write the result for the Fact Table (FactTable_DIM) for the further reporting in the Data Warehouse Architecture.

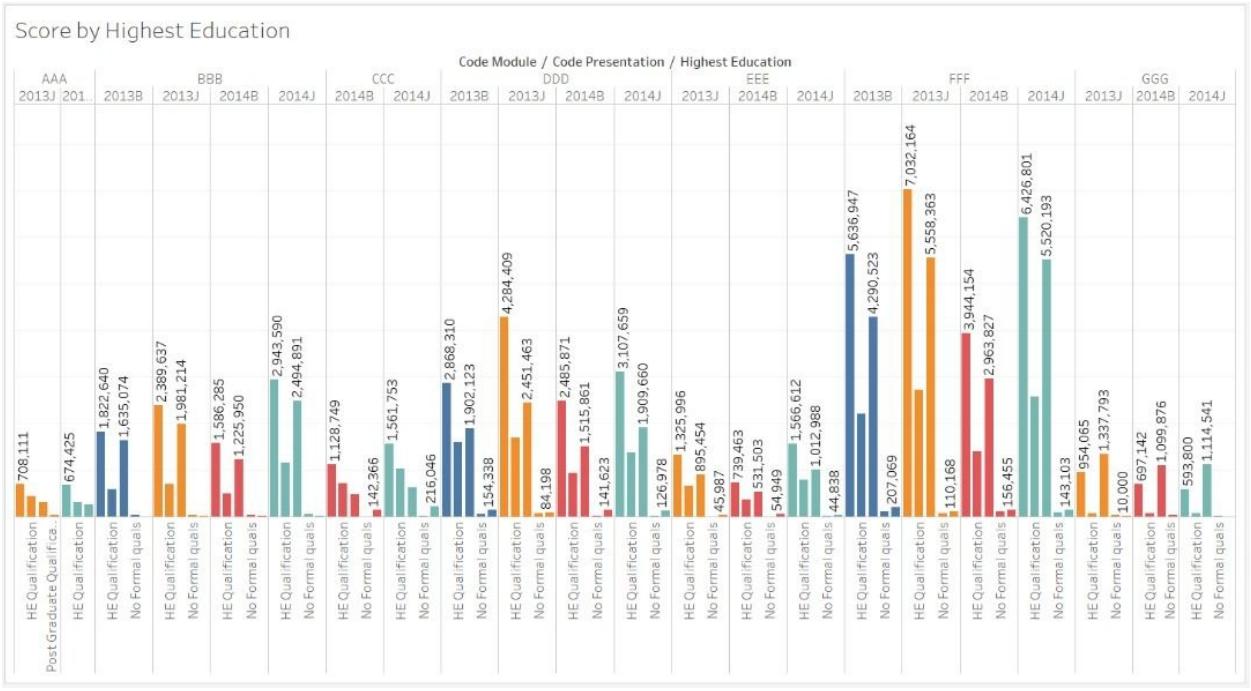
Final Result/Reporting

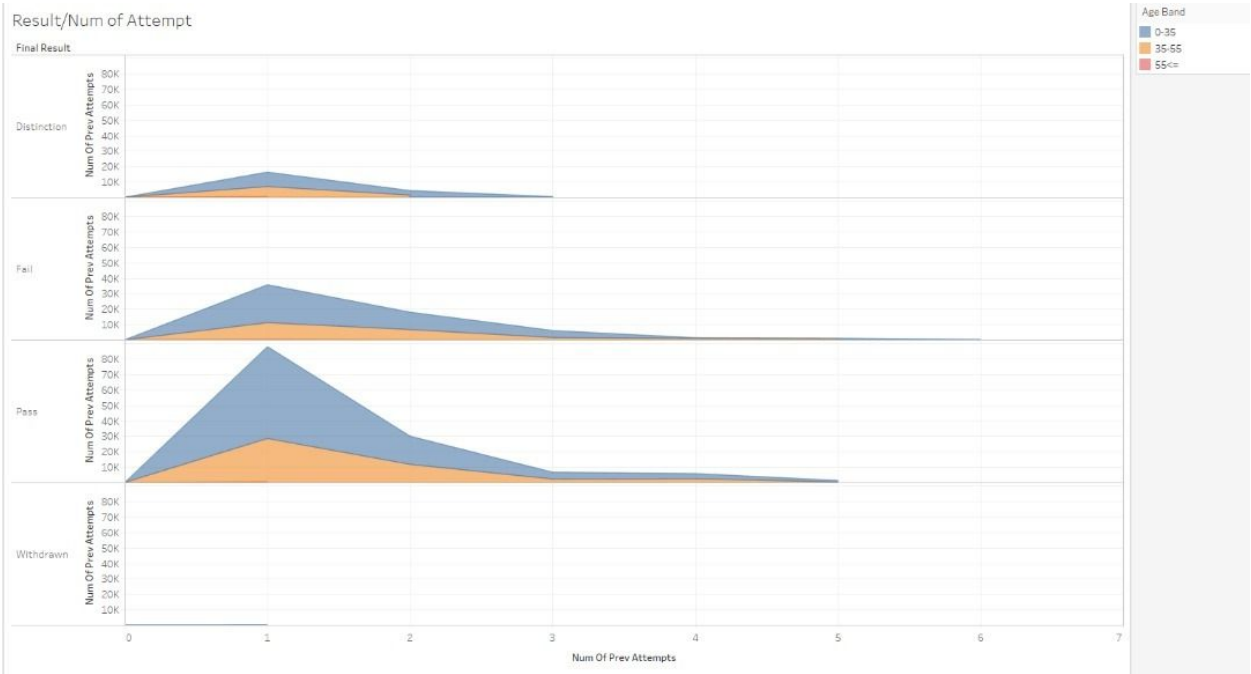
Final Result



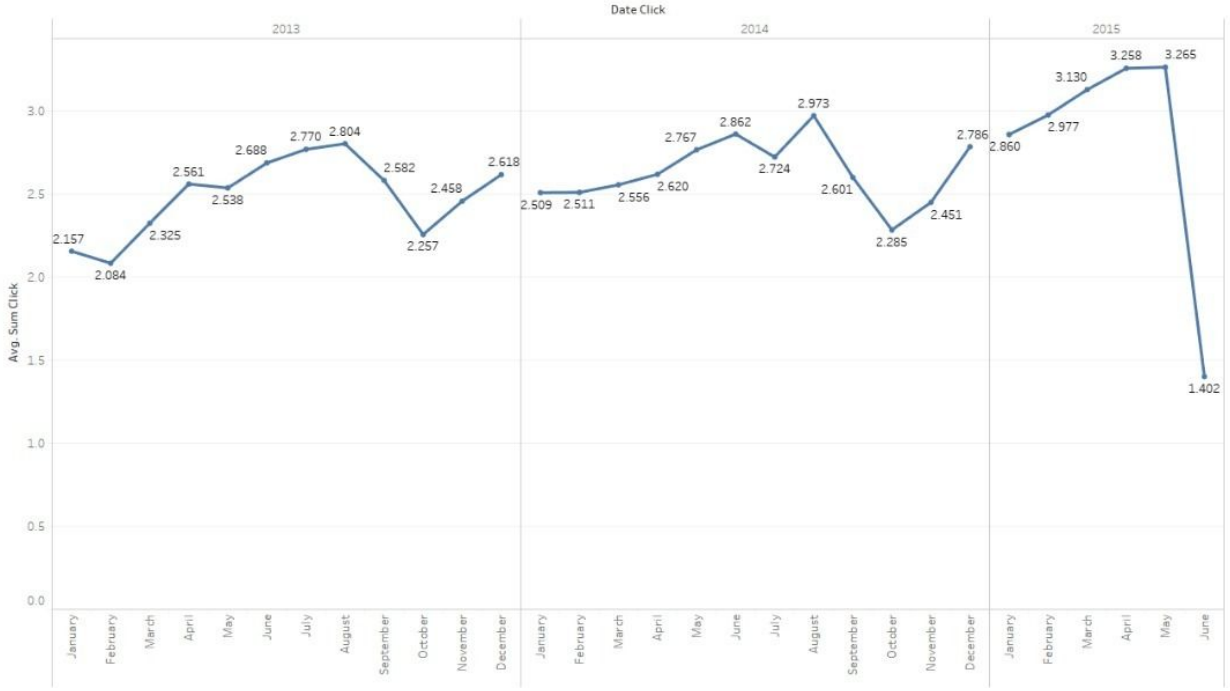
Average Score based on education



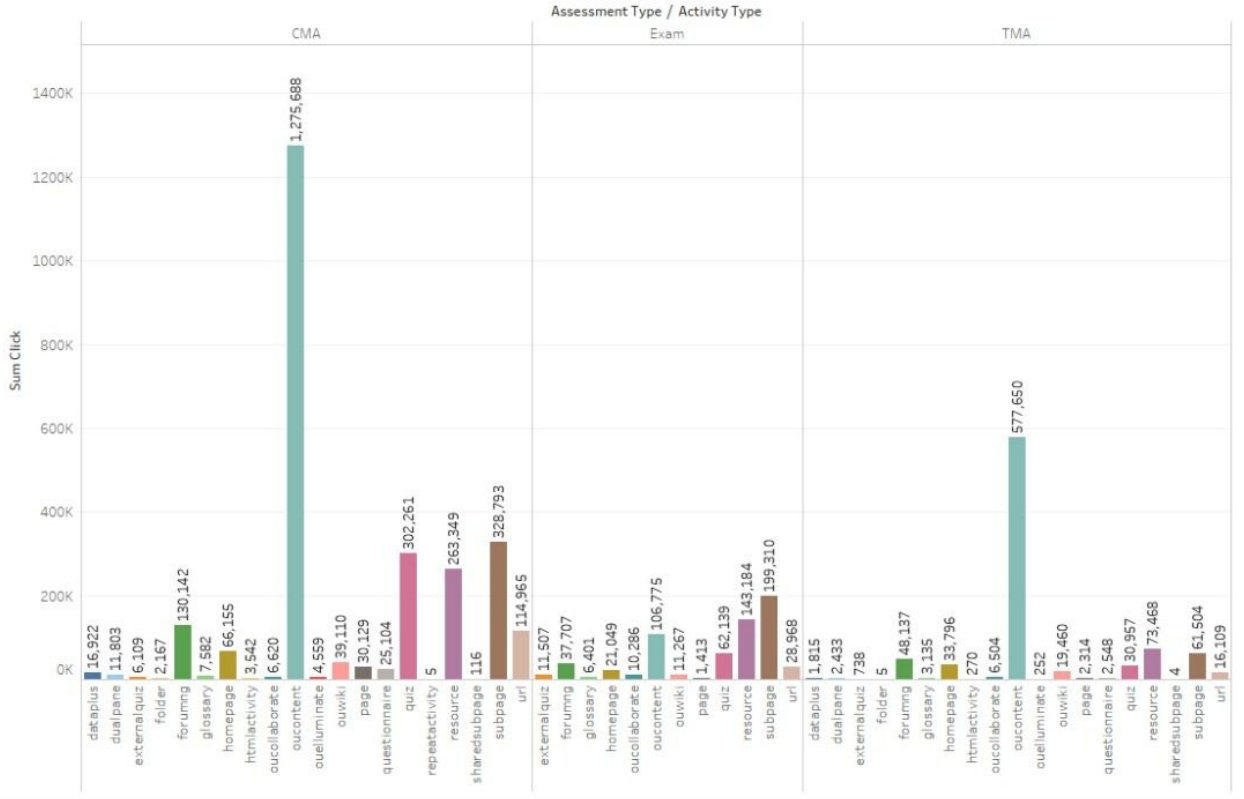

Score by Highest Education
Code Module / Code Presentation / Highest Education
AAA
BBB
CCC
DDD
EEE
FFF
GGG
2013J
2014B
2014J
HE Qualification
No Formal quals
Post Graduate Qualifica..
708,111
674,425
1,822,640
1,635,074
2,389,637
1,981,214
1,586,285
1,225,950
2,943,590
2,494,891
1,128,749
142,366
1,561,753
216,046
2,868,310
1,902,123
154,338
4,284,409
2,451,463
84,198
2,485,871
1,515,861
141,623
3,107,659
1,909,660
126,978
1,325,996
895,454
45,987
739,463
531,503
54,949
1,566,612
1,012,988
44,838
5,636,947
4,290,523
207,069
7,032,164
5,558,363
110,168
3,944,154
2,963,827
156,455
6,426,801
5,520,193
143,103
954,065
1,337,793
10,000
697,142
1,099,876
593,800
1,114,541

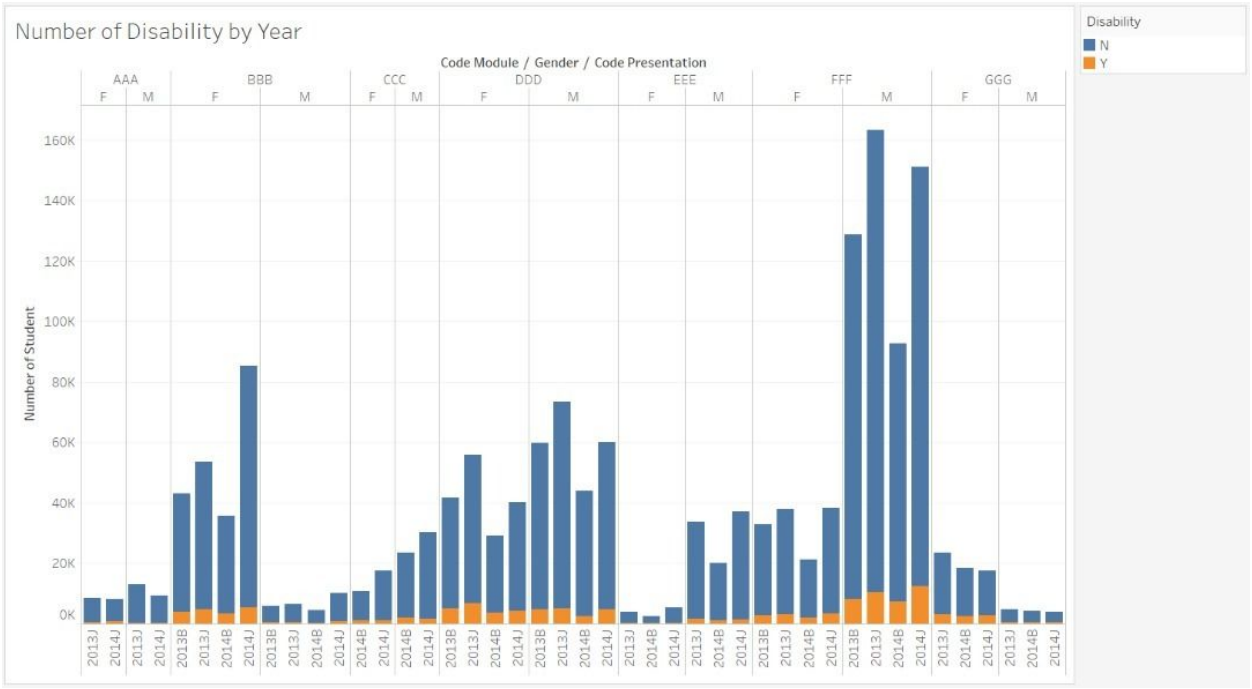

Result/Num of Attempt
Final Result
Distinction
Fail
Pass
Withdrawn
Num Of Prev Attempts
Num Of Prev Attempts
Age Band
0-35
35-55
55<=

Date vs SumClick



SumClick and Assessment/Activity Type




Number of Disability by Year
Code Module / Gender / Code Presentation
AAA
BBB
CCC
DDD
EEE
FFF
GGG
F
M
Number of Student
160K
140K
120K
100K
80K
60K
40K
20K
0K
2013J
2014J
2013B
2014B
Disability
N
Y

Bibliography

Abilitie Team. “3 Keys To Impactful Virtual Classroom Training.” abilitie, 11 March 2020, https://www.abilitie.com/3-keys-virtual-training/

Dedic, Nedim and Clare Stainer. “An Evaluation of the Challenges of Multilingualism in Data Warehouse Development”, STORE - Staffordshire Online Repository, 26 April 2016, An evaluation of the challenges of Multilingualism in Data Warehouse development - STORE - Staffordshire Online Repository (staffs.ac.uk)

El-Sappagh, Shaker H. Ali, et al. “A Proposed Model for Data Warehouse ETL Processes.” Journal of King Saud University - Computer and Information Sciences, Elsevier, 8 May 2011, A proposed model for data warehouse ETL processes - ScienceDirect .

Guitart, I., & Conesa, J. “Analytic Information Systems in the Context of Higher Education: Expectations, Reality and Trends”, IEEE Xplore, 2 November 2015, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7312087

J. Sreemathy, et al. “Data Integration in ETL Using TALEND”, IEEE Xplore, 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 7 March 2020, https://ieeexplore.ieee.org/abstract/document/9074186

Leo Willyanto Santoso, and Yulia. “Data Warehouse with Big Data Technology for Higher Education”, Procedia Computer Science, Elsevier B.V., 26 December 2017, http://www.sciencedirect.com/science/article/pii/S1877050917329022

Margaret Rouse. “What is Business Intelligence Architecture (BI Architecture)?”,SearchBusinessAnalytics, 13 July 2020, https://searchbusinessanalytics.techtarget.com/definition/business-intelligence-architecture

Mousa, Ayad Hameed, et al. “Virtual Data Mart for Measuring Organizational Achievement Using Data Virtualization Technique (KPIVDM)” Jurnal Teknologi, May 2014, https://www.sciencedirect.com/science/article/pii/S131915781100019X#:~:text=Research%20in%20the%20field%20of,modeling%20based%20on%20UML%20environment.

Owusu, Acheampong, et al. “Investigating the factors affecting business intelligence systems adoption: A case study of private universities in Malaysia.” *International Journal of Technology Diffusion (IJTD)*, June 2017, https://rtis2.ut.ac.ir/media/?activity_id=97754036&f_name=paper_upload&path=uploads/articles/papers/Investigating_the_Factors_Affecting_Business_Intelligence_Systems_Adoption_A_Case_Study_of_Private_Universities_in_Malaysia.pdf

“Data Extraction Techniques”, Rosoka, 26 May 2020, Data Extraction Techniques | Rosoka

References

[Article about Data Extraction Technique], (2020, May 26). *Data Extraction Techniques*. Rosoka. Retrieved December 25, 2020. Data Extraction Techniques | Rosoka

Abilitie Team(2020). 3 Keys To Impactful Virtual Classroom Training. Retrieved December 17, 2020, from https://www.abilitie.com/3-keys-virtual-training/

Dedic, Nedim and Clare Stainer(2016, April). An Evaluation of the Challenges of Multilingualism in Data Warehouse Development. Retrieved December 20, 2020 An evaluation of the challenges of Multilingualism in Data Warehouse development - STORE - Staffordshire Online Repository (staffs.ac.uk)

El-Sappagh, Shaker H. Ali., Hendawi, Abdeltawab M. Ahmed and El Bastawissy, Ali Hamed(2011, July). A proposed model for data warehouse ETL processes. Retrieved December 21, 2020 https://www.sciencedirect.com/science/article/pii/S131915781100019X#:~:text=Research%20in%20the%20field%20of,modeling%20based%20on%20UML%20environment.

Guitart, I., & Conesa, J. (2015, September). Analytic information systems in the context of higher education: Expectations, reality, and trends. In 2015 international conference on intelligent networking and collaborative systems (pp. 294-300). IEEE from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7312087

Hameed Mousa, Ayad., Shiratuddin, Norshuhada and Abu Bakar,Muhammad(2014, May). Virtual Data Mart for Measuring Organizational Achievement Using Data Virtualization Technique (KPIVDM). Retrieved December 20, 2020 https://www.researchgate.net/figure/Data-warehouse-architecture_fig1_275068752

J. Sreemathy, I. Joseph V., S. Nisha, C. Prabha I. and G. Priya R.M., "Data Integration in ETL Using TALEND," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 1444-1448, doi: 10.1109/ICACCS48705.2020.9074186. Retrieved December 22, 2020 from https://ieeexplore.ieee.org/abstract/document/9074186

Leo Willyanto Santoso, Yulia,Data Warehouse with Big Data Technology for Higher Education,Procedia Computer Science,Volume 124,2017,Pages 93-99,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2017.12.134. Retrieved December 22, 2020 from http://www.sciencedirect.com/science/article/pii/S1877050917329022

Owusu, A., Ghanbari-Baghestan, A., & Kalantari, A. (2017). Investigating the factors affecting business intelligence systems adoption: A case study of private universities in Malaysia. *International Journal of Technology Diffusion (IJTD)*, *8*(2), 1-25. Retrieved December 23, 2020 from https://rtis2.ut.ac.ir/media/?activity_id=97754036&f_name=paper_upload&path=uploads/articles

/papers/Investigating_the_Factors_Affecting_Business_Intelligence_Systems_Adoption_A_Case_Study_of_Private_Universities_in_Malaysia.pdf
TechTarget By Margaret Rouse. What is Business Intelligence Architecture. Retrieved December 24, 2020, from https://searchbusinessanalytics.techtarget.com/definition/business-intelligence-architecture