**Data Storage Solutions for Data Analytics**

Design and Development of data storage solutions for analysis

Dublin University Data Warehouse

Aishwary Singh Gaur - 10523463

Akinkunmi Ademola Mustapha - 10574085

Gulnaz - 10522484

**Table of Contents**

[**1.0.** **Introduction** 3](#_Toc77632824)

[**1.1** **Reasons for Selecting the subject Area and Data** 3](#_Toc77632825)

[**1.2** **Vision and Goals** 5](#_Toc77632826)

[**1.3** **Key Stakeholders** 5](#_Toc77632827)

[**1.4.** **Insights** 6](#_Toc77632828)

[**2.0.** **Schema** 8](#_Toc77632829)

[**3.0.** **Transforming and Loading of Data in Data Warehouse** 10](#_Toc77632830)

[**4.0.** **SSRS Reports & Visualization:** 15](#_Toc77632831)

[**4.1. SSRS Reports:** 15](#_Toc77632832)

[**4.2. Visualization:** 20](#_Toc77632833)

[**5.0.** **Differences Of Storing, Retrieving Data In, From Relational and Graph Databases** 22](#_Toc77632834)

[**References:** 29](#_Toc77632835)

# **Introduction**

**Dublin University** continues to grow and build on its reputation as Ireland's leading University, with pre-professional programmes, the university is rooted in the liberal arts and sciences. The University is led by a mission and set of values that emphasise the importance of the student experience. We try to embody our beliefs in everything we do as a community that both honours our history and defines our future. With over 6,000 students and the potential to conduct over 30 approved courses, it is one of the largest universities in the country.

We want to compile data collected through these management concepts and technologies into reports, visualisations, and dashboards that can be modelled using analytical tools. The mocking data utilised by [Mockaroo](https://www.mockaroo.com/) for the implementation and dashboards was produced.

A typical university often has many subsystems which are essential to its internal processes and activities. For example, the student registration system, accounting system, administration system for courses, staff system, and many more will be part of these subsystems. The purpose of this study is to suggest a data warehouse design for a university information system that will aid and assist decision making. The proposed design transforms the existing operational databases into an information database or data warehouse by cleaning and scrubbing the existing operational data. All the data that has been gathered and cleansed is imported and indexed in the warehouse, making it suitable for data understanding operations for useful insights.

## **Reasons for Selecting the Subject Area and Data**

A successful university requires a balance of positive grades for all its students in all its departments. Grading systems are still debatable as to whether they push students to do better or do psychological harm to those who have not been at their best with the studies. The university gathers the data for all the students, studying in various department with multiple courses and all the grades have been collected for each of the modules. Using these parameters, we may identify what types of areas of improvement the university may require. Different departments of faculties might work with their strategies to improve their teaching after determining the grades.

We have a robust financial strategy through which, we have invested considerably in our estates and will continue to do so, delivering a thriving and financially viable university. In addition to continuing to manage our resources efficiently, we will better plan our capital expenditure to maximise strategic benefits.

There are 32 different courses provided by the university, each of the course averages close to €20000 for tuition fee. We do need to have the knowledge of the total revenue being generated from the student fees, which can be used to the betterment of the university. We will be using the gathered report for all the students studying in the university to find out how much amount in total was generated. The obtained revenue can be used for later research or advancement of the university as whole.

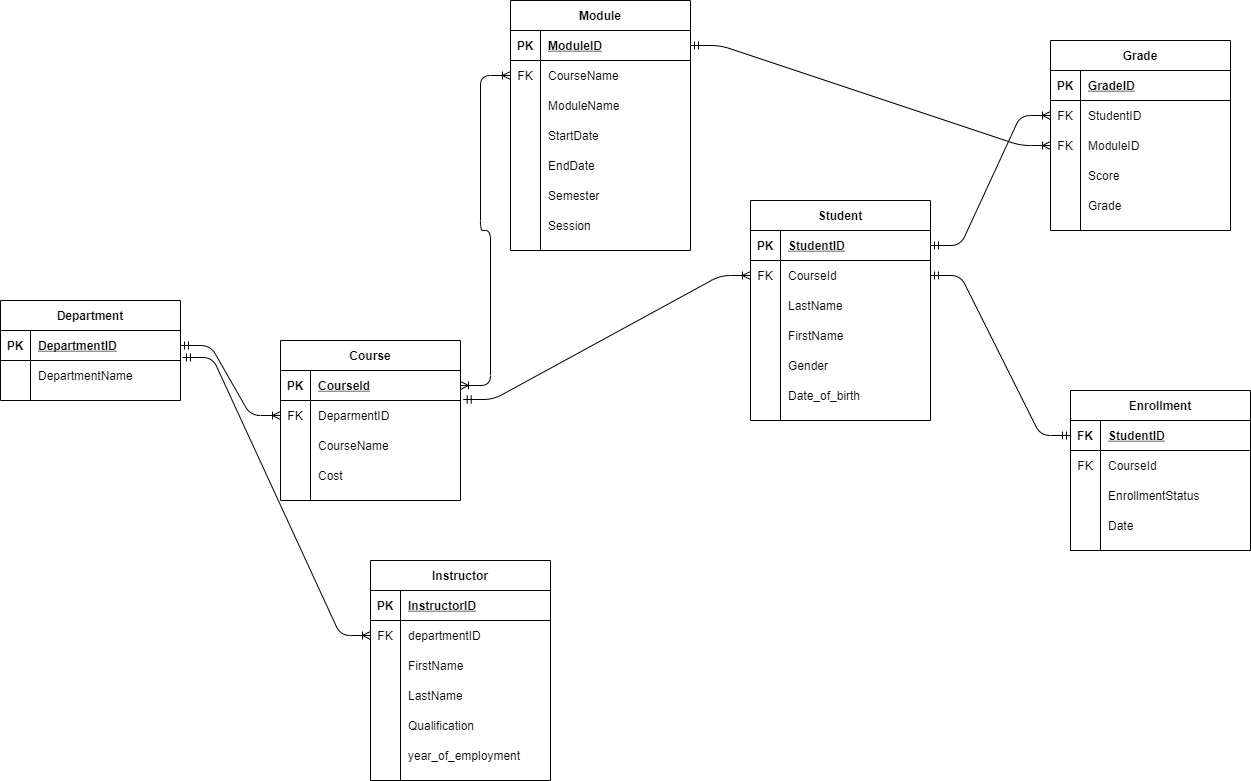


Figure 1: ER Diagram

The operational database that was utilised to create the data warehouse consists of seven different relations or tables that are linked together through relationships. It is a relational model database implemented on SQL Server**.** This database reflects the operations of a Dublin University, provides a front-end registration system for managing student enrolment processes, an accounting method for handling students’ fees payments, a course administration system for organising and assigning modules and instructors to courses, a grading system for students’ post examination records in the form of Grades and Score (GPA), Figure 1 depicts the conceptual ER Diagram of the operational database.

## **Vision and Goals**

Dublin University's mission is to give greater levels of intellectual and innovative talents in order to develop cutting-edge facilities for education, training, and research. The report is essential to assist the educational institute's decision makers in making an informed conclusion. Effective decisions are those that take an organisation closer to achieving a set of agreed-upon goals in a timely manner.

As part of such good decision-making, foundation and feedback information are required. As a result, we must disseminate business intelligence throughout the enterprise. Data warehouse is designed to answer questions being asked throughout the business on an everyday basis. These inquiries focus on the overall process rather than individual transactions and are mostly used to identify patterns and bottlenecks in the university. To address these concerns, the data warehouse should be designed in a way that reflects how management see their enterprise. That is, it should include the business's most important measurements as well as the criteria used to evaluate them.

The goal of an educational organisation is to become more student-centred, improve the campus atmosphere, increase research, and academic activities, among other things. Although, we will be collaborating with the faculty departments to help the students get their grades up and we will also work with the accounting department in order to help the institution with financial needs. Before providing the needs, we will have to understand what the university lacks in terms of its growth.

## **Key Stakeholders**

In education, the term stakeholder typically refers to anyone who is invested in the welfare and success of a school or college and its students. stakeholders have a “stake” in the school and its students, meaning that they have personal, professional, civic, or financial interest or concern.

The higher education literature is filled with lists of stakeholders, which do not really provide a framework to identify who the stakeholders are and how to recognize them.Persons, groups, public, private and government organizations, institutions, societies, and natural environment can be stakeholders of an organization (Stakeholder Definition, 2021). The following are the stakeholders for Dublin university:

**Government:** Government organizations have had a strong hold on university in the past and continue to have strong influences on schools today. The state government has always shown a helping hand to the university in all terms of research and granting them to perform them. Provisioning acreage for experiments and extensive study to the University is also one of the reasons why Government has been one of the major stakeholders for the University (THEORIES OF STAKEHOLDERS, 2021).

**Board of Trustees:** In the management of higher education institutions, governing boards are imperative. They oversee the appointment of university president, Director, as well as having the power to fire them. They also have influence over academia by being able to make faculty appointments. Governing boards are also in charge of an institution's financial health as well as difficulties with federal and non-federal regulators (THEORIES OF STAKEHOLDERS, 2021).

**Communities:** Universities have learnt that in order to achieve their aims and objectives, solid relationships and significant attention must be given to communities. Community support is essential for institutions to function well, and it can be gained through good deeds and effective public relations. No institution can function effectively while remaining disconnected from the life of the community in which it operates (THEORIES OF STAKEHOLDERS, 2021).

## **Insights**

There are a variety of insights that can be gained by working with acquired data, the most important area of interest will be the student, Instructor, Module for the courses under a department and keeping track of the university's finances, which are majorly generated from the students' education fees. although the most important points for us to gain the information for, is going to be student grades and finances of the university.

The university's main goal is to ensure that all students receive a better education in terms of the duration of time each course lasts and to evaluate them based on their grades. Working with grades will not only help with the improvement of student, but the department as whole, which eventually improve the students’ quality of education. Instructors also play an important role which can help the university, we have information about the total experience of all our instructors which helps the board with assigning the right educator to a course.

Financial stability is essential for any firm, an educational establishment will also require a structure for which the funds will be spent at the institution, one approach to achieve so is to concentrate on earned revenue. Using our dataset, we can assess how well a department is performing in terms of the revenue it generates, by determining how many enthusiast students are enrolling for the courses under a particular department, The funds raised will eventually be used to help establish research or experiments at the university. Keeping track of revenue over time utilising historical data might help the university board become more knowledgeable and help with the expanses.

# **Schema**

Dublin University is a large college that requires a database system to keep track of their operations. The University contains many Departments, Departments are described by a unique code (DepartmentID) and a name. Each Department has a certain number of Courses under its wings. Departments provide a variety of courses to fulfil the needs of the student community. The following data about courses are of interest: Course ID (CourseID), course name and course cost. Each Course consists of at least two modules but can be up to 10 modules. Modules are uniquely identified by a module id, and all have module titles, lists of duration is also given for which a module will last.

Students are uniquely identified by their ID (StudentID) with their first and last name, with generalizing of the gender they belong to. The received grades for all the students are available in our database system in the form of total score for a particular Module and the Grade (GPA).

In this Schema we have developed 2 fact tables for Grade and Revenue and there are 6-dimension tables which will be connected with the each of those fact tables to provide an insight. Student, Grade and Module dimensions relate to the Grade Fact Table to provide information on the grades, which can be used to gather information in terms of each attribute listed within them.

Cost being the most important attribute in the Course dimension will come in useful to Revenue Fact Table in order to provide information on the raised funds, the courses will be the part of the departments. The number of students assigned to several courses will help with total amount earned over the years. Date dimension will play a crucial part in segregating the revenue earned over the years.

Diagram

Description automatically generated

Figure 2: Star Schema

# **Extract, Transform and Loading of Data in Data Warehouse**

In order to load the data from the operational Data source to Data Warehouse, SQL Server Management Studio v18.9.1 and Visual Studio 2019 were used.

7 different tables were created to store the source data within them, as explained in the figure 1. Once the source data was collected, the below queries were used to create different dimension and fact tables.

CREATE DATABASE B9DA102\_UNIVERSITY\_DW

USE [B9DA102\_UNIVERSITY\_DW]

CREATE TABLE Student\_Dim (

student\_key INT NOT NULL IDENTITY,

student\_id INT,

first\_name VARCHAR(50),

last\_name VARCHAR(50),

gender VARCHAR(50),

PRIMARY KEY (student\_key),

);

GO

CREATE TABLE Course\_Dim (

course\_key INT NOT NULL IDENTITY,

course\_id VARCHAR (6),

course\_name VARCHAR(61),

cost DECIMAL(6,1),

PRIMARY KEY (course\_key),

);

GO

CREATE TABLE Department\_Dim (

department\_key INT NOT NULL IDENTITY,

department\_id INT,

department\_name VARCHAR(100),

PRIMARY KEY (department\_key),

);

GO

CREATE TABLE Module\_Dim (

module\_key INT NOT NULL IDENTITY,

module\_id VARCHAR (5),

module\_name VARCHAR(79),

PRIMARY KEY (module\_key),

);

GO

CREATE TABLE Enrollment\_Dim (

enrollment\_key INT NOT NULL IDENTITY,

enrollment\_id INT,

enrollment\_status VARCHAR(9),

PRIMARY KEY (enrollment\_key),

);

GO

CREATE TABLE Grade\_Dim (

grade\_key INT NOT NULL IDENTITY,

grade\_id INT,

score INT,

grade CHAR (1),

PRIMARY KEY (grade\_key),

);

GO

CREATE TABLE Date\_Dim (

date\_key INT NOT NULL IDENTITY,

full\_date DATE,

day\_of\_week CHAR (15),

day\_type CHAR (20),

day\_of\_month INT,

Month\_ CHAR (10),

Quarter\_ CHAR (2),

Year\_ INT,

PRIMARY KEY (date\_key),

);

GO

CREATE TABLE Grade\_Fact (

student\_key INT,

module\_key INT,

grade\_key INT,

grade INT,

PRIMARY KEY (grade\_key),

FOREIGN KEY (student\_key) REFERENCES Student\_Dim (student\_key),

FOREIGN KEY (module\_key) REFERENCES Module\_Dim (module\_key),

FOREIGN KEY (grade\_key) REFERENCES Grade\_Dim (grade\_key),

);

GO

CREATE TABLE Revenue\_Fact (

date\_key INT,

student\_key INT,

course\_key INT,

department\_key INT,

enrollment\_key INT,

amount DECIMAL (6, 1),

PRIMARY KEY (enrollment\_key),

FOREIGN KEY (date\_key) REFERENCES Date\_Dim (date\_key),

FOREIGN KEY (student\_key) REFERENCES Student\_Dim (student\_key),

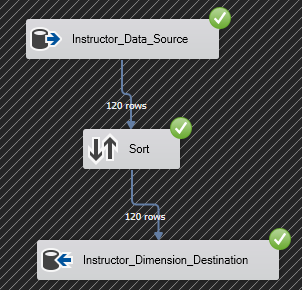
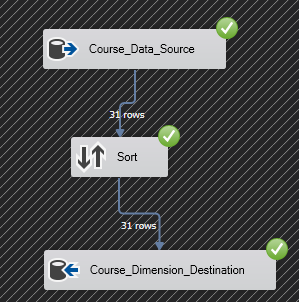
FOREIGN KEY (department\_key) REFERENCES Department\_Dim (department\_key),

FOREIGN KEY (enrollment\_key) REFERENCES Enrollment\_Dim (enrollment\_key),

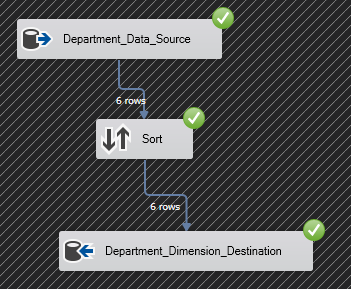
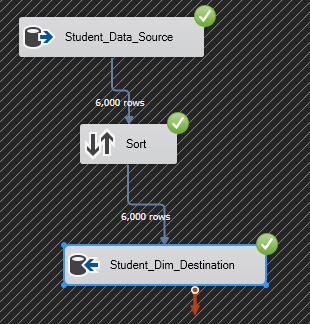
FOREIGN KEY (course\_key) REFERENCES Course\_Dim (course\_key),

);

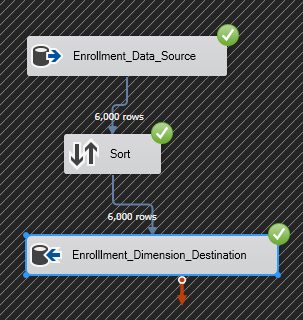
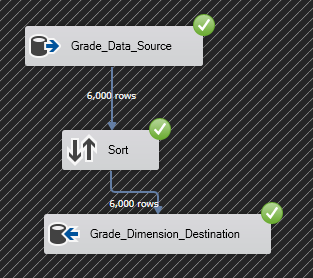
GO



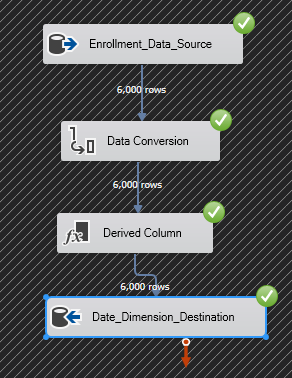
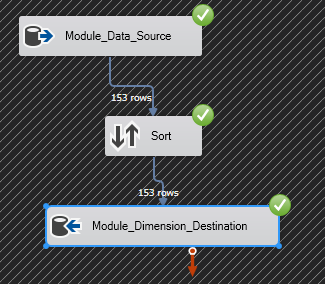
**Populated Course Dimension** **Populated Instructor Dimension**

****

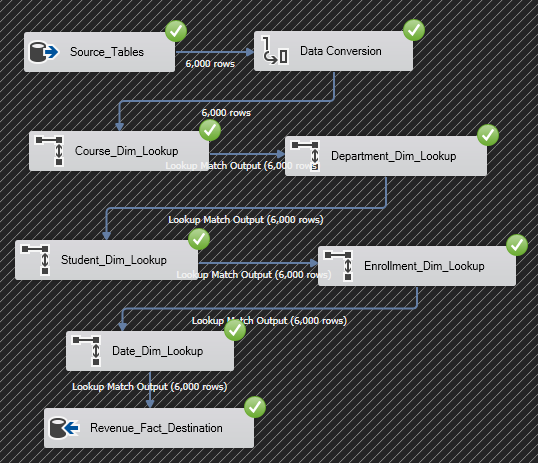
**Populated Student Dimension Populated Department Dimension**

****

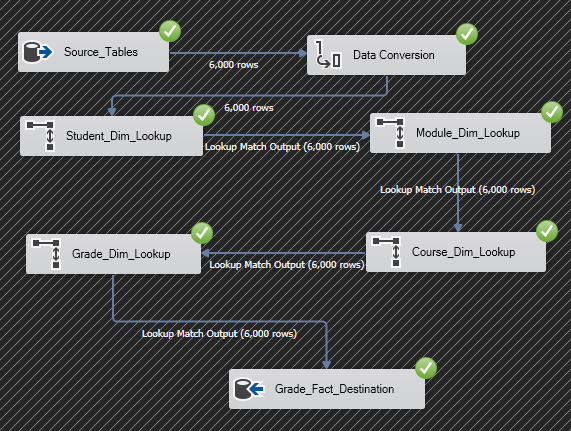
**Populated Enrolment Dimension Populated Grade Dimension**

** **

**Populated Date Dimension Populated Module Dimension**

****

**Revenue Fact table populated from multiple sources (Course, Student, Department & Enrolment)**

****

**Revenue Fact table populated from multiple sources (Student, Module, Grade & Course)**

# **SSRS Reports & Visualization:**

## **4.1. SSRS Reports:**

#### **Tabular Report:**

The report displays thorough information on 6000 students throughout a 41-year period from 1980 to 2021, including the linked departments and courses each student has taken, as well as their grade distribution.

Table

Description automatically generated

#### **Drill Up and Down Report**

This report depicts the annual summaries of enrolment and department revenue from 1980 through 2021, with 1995 being the year with the highest revenue for each department. The plus and minus signs are employed to disguise the data's complexity. To get more information, we can enlarge the + sign.

Table

Description automatically generated

#### **Parameterized Report**

The department parameter in this report displays the courses available under each department as well as the money generated by each course in a single year.

Table, treemap chart

Description automatically generated with medium confidence

##### **SQL Query for Parameterized Report**

USE [B9DA102\_UNIVERSITY\_DW]

GO

CREATE PROC QueryReport4

@id char(3)

AS

SELECT Department\_Dim.department\_id, Department\_Dim.department\_name, Course\_Dim.course\_name, Revenue\_Fact.amount, Date\_Dim.Year\_

FROM Course\_Dim

INNER JOIN Revenue\_Fact

ON Course\_Dim.course\_key = Revenue\_Fact.course\_key

INNER JOIN Department\_Dim

ON Revenue\_Fact.department\_key = Department\_Dim.department\_key

INNER JOIN Date\_Dim

ON Revenue\_Fact.date\_key = Date\_Dim.date\_key

WHERE department\_id=@id

GROUP BY Department\_Dim.department\_id, Department\_Dim.department\_name, Course\_Dim.course\_name, Revenue\_Fact.amount, Date\_Dim.Year\_

GO

EXEC QueryReport4 '100'

#### **Outer Report**

The average score, or grade, received by students throughout the course is depicted in the outer report.

Table

Description automatically generated

##### **SQL Query for Outer Report**

USE [B9DA102\_UNIVERSITY\_DW]

GO

CREATE PROC uspOuterQuery

AS

SELECT Student\_Dim.student\_id, AVG(Grade\_Fact.grade) AS [Average Grade]

FROM Grade\_Fact

INNER JOIN Student\_Dim

ON Grade\_Fact.student\_key = Student\_Dim.student\_key

GROUP BY Student\_Dim.student\_id

GO

EXEC uspOuterQuery

## **4.2. Visualization:**

**Insights for Grade:**

As shown in the Figure 3, we have visualized the data in 2 different types in the Tableau Dashboard, the first image is plotted as the Bar Chart, while the second image is a Pie Chart visual.

The first chart illustrates the average score received by students and the bars are dissected into two parts to generalize the gender, the bars are separated for each course student is enrolled for. The scores attained for courses, do not have much significant different between the genders, although we can also confirm Doctor of Philosophy in Electrical Engineering has the lowest scores received.

However, the pie chart shows the grade distribution between 6000 students, after making an understanding of the chart, we can tell, there is a significant number of students who received the Grade A, which does help the university set a standard.

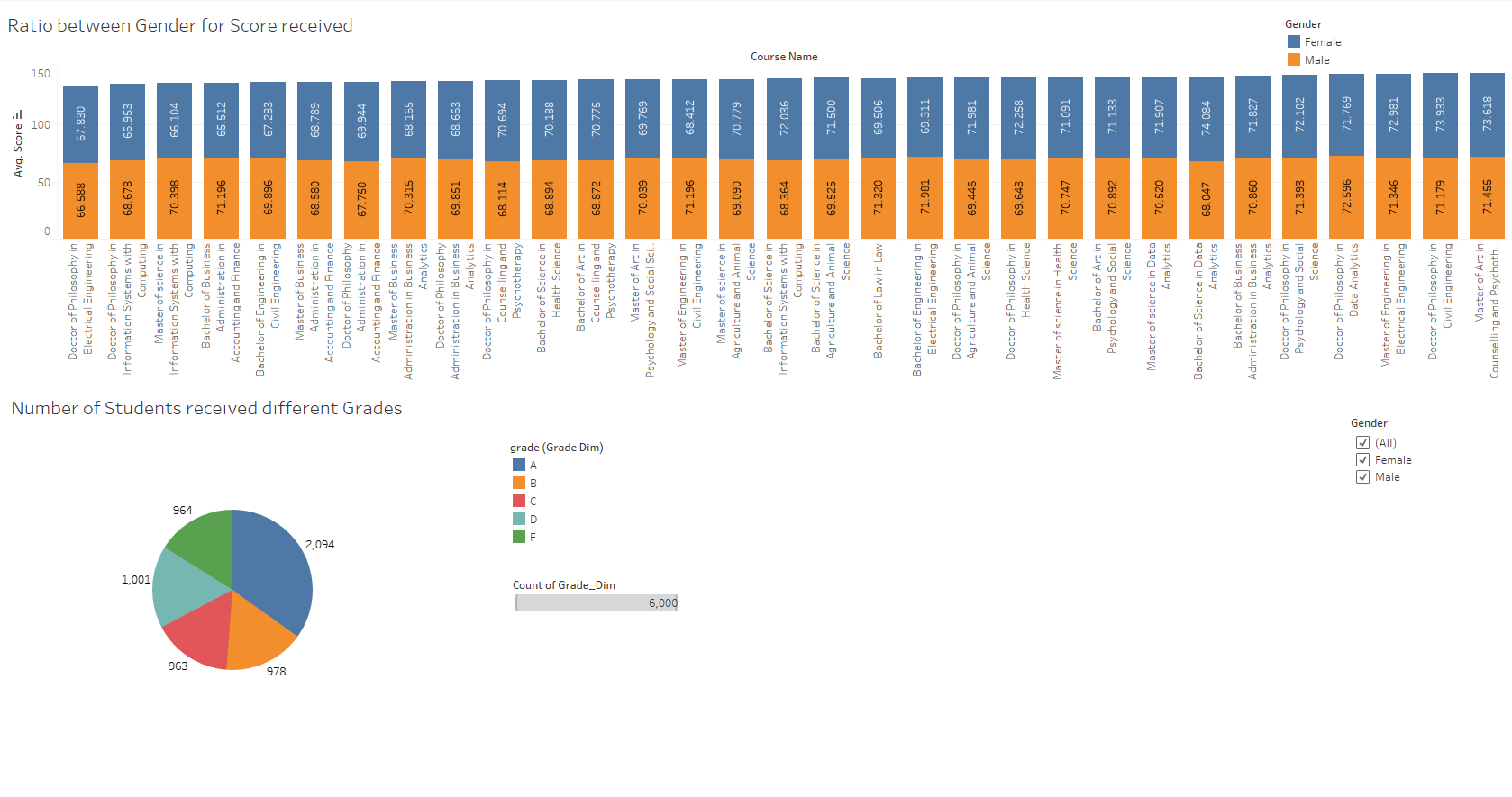


Figure 3: Dashboard of Bar Chart & Pie Chart

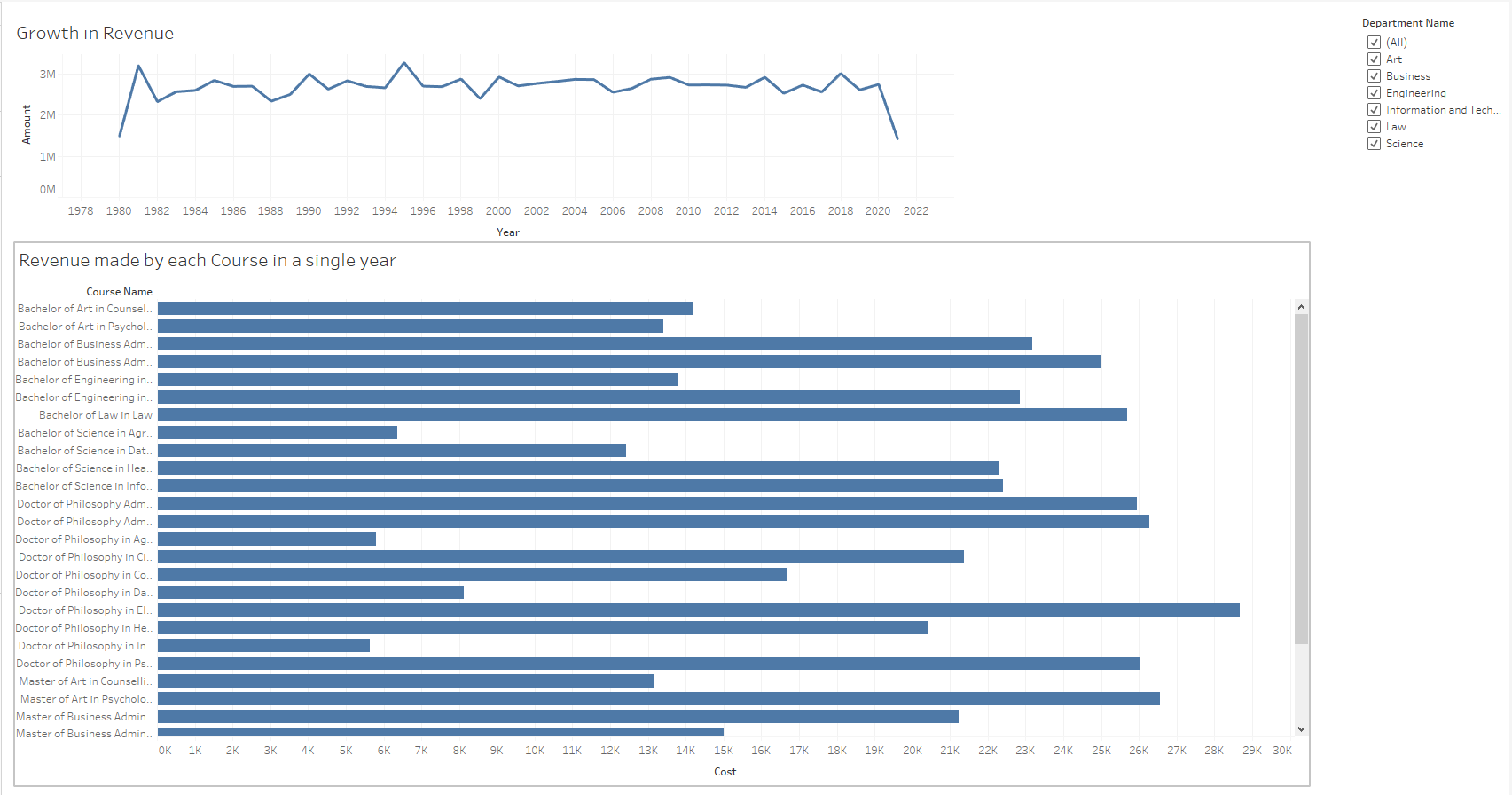
**Insights for Revenue:**

In the Tableau Dashboard Figure 4, we have presented the data in two different ways: the first image is a Line Chart, while the second image is a Bar Graph.

The line graph compares the revenue generated for the total number of years against the years, which are recorded for 41 years from 1980 to 2021. This illustrates, as time progresses from left to right, we can confirm the stability with income, except for the year 2021, there is a decline in revenue which must be monitored upon to improve for the years to come and help the board work with the expanses of university.

Although, the supplied bar graph compares total revenue of a single year against the different courses which come under 6 different departments.

Bachelor courses are by far the most generating for their respective departments in terms of revenue.

Figure 4: Dashboard of Line Chart & Bar Graph

# **Differences Of Storing, Retrieving Data In, From Relational and Graph Databases**

Structured query language (SQL) have gained widespread usage for querying relational databases since its inception, because it allows both for easy ad-hoc querying of a database as well as specifying use-case related queries, even other nonrelational databases make use of it under the hood, like object-relational mappers. However, SQL has some major performance challenges when navigating through connected data.

Furthermore, Cypher query language (CQL) is Neo4j’s graph query language which allows users to store and extract data from graph database. CQL borrows from SQL syntax, however its often considered to be more efficient than SQL. SQL queries are often lengthy, this increases the complexity of the code and the likelihood of human coding error, also the time it takes to run them.

In addition, graph database makes it easy to model data, unlike relational databases. The easy of model data with CQL and graph database has a lot of business benefits. In relational database, there is a major difference between analysis and implementation because the data modelling is abstracted from the actual day-to-day SQL queries that. While CQL and graph databases communicates how the data is related.

The following query screenshots illustrate the differences between SQL and CQL:

Graphical user interface, website

Description automatically generated

Figure 5.1: SQL query to insert a new department to the deparment table

Graphical user interface, text, application, email

Description automatically generated

Figure 5.2: CQL query to insert a new department to Department node

The two screenshot above illustrates the differences between the insertion queries of SQL and CQL.

Graphical user interface

Description automatically generated with low confidence

Figure 5.3: SQL query illustrating how to join tables

Table

Description automatically generated

Figure 5.4: CQL query illustrating how relationships can be used to associate different nodes

These two query in figure 5.3 and figure 5.4, helps illustrate how SQL queries can easily get complex when it is navigating through different data. The CQL query easily match up the nodes based on the exiting relationships. Thus, the likelihood of making mistakes when writing the SQL code is much higher compared to the CQL code because it’s lengthy.

Table

Description automatically generated with low confidence

Figure 5.5: SQL query to sort data

Text

Description automatically generated with medium confidence

Figure 5.6: CQL query to sort data

Graphical user interface, application

Description automatically generated

Figure 5.7: SQL query to perform aggregation in data.

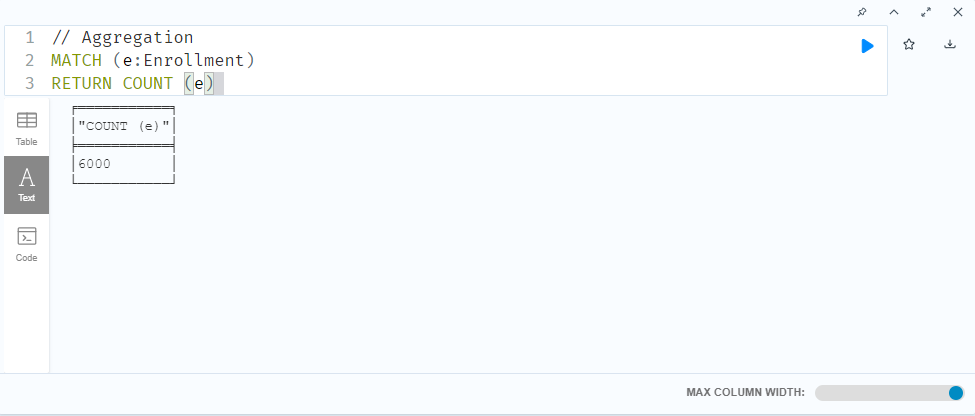


Figure 5.8: CQL query to perform aggregation.

Graphical user interface, application

Description automatically generated

Figure 5.9: SQL query to aggregate data and find unique values.



Figure 5.10: SQL query to aggregate data and find unique values.

A computer screen capture

Description automatically generated with medium confidence

Figure 5.11: SQL query for filtering data

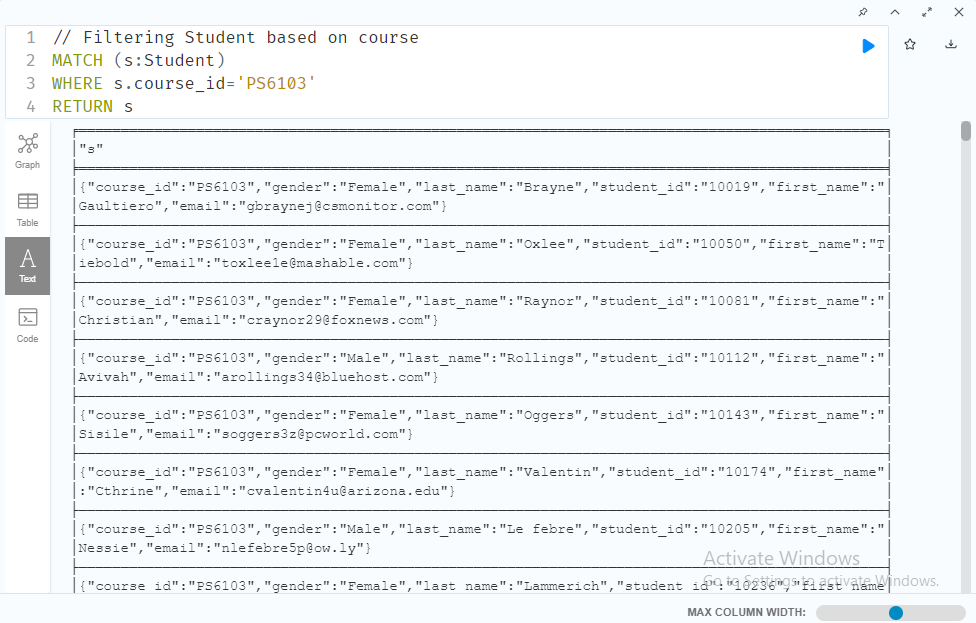


Figure 5.12: CQL query for filtering data

A computer screen capture

Description automatically generated with medium confidence

Figure 5.13: SQL query for selecting data.



Figure 5.14: CQL query for selecting nodes.

# **References:**

Dergipark.org.tr. (2021). *THEORIES OF STAKEHOLDERS*. [online]

Available at:

<https://dergipark.org.tr/tr/download/article-file/323147>

[Accessed 9 July 2021].

The Glossary of Education Reform. (2021). Stakeholder Definition. [online]

Available at: <https://www.edglossary.org/stakeholder/>

[Accessed 9 July 2021].