A logo with a red and black design

Description automatically generated

**Data Mining and Warehousing**

*Module Leader: Assoc.Prof. Marwa Salah*

*Ta:**Omnia Elsodany*

**Phase 1**

|  |  |
| --- | --- |
| Student Name | ID |
| Abdelrahman Nader | 227110 |
| Mohamed Sameh | 228727 |
| Abdelrahman Youssef | 224774 |
| Ruba Mobasher | 222335 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **ETL** | **Data conversion** | **Extracting data** | **Star schema** | **Union all editor** |
| **Abdelrahman 227110** | **AVG Score**  **Age Category** | **Data type conversion 1** |  | **Star schema 2** | **Union all transformation** |
| **Mohamed 228727** | **Split Student name**  **Student Year** |  | **Flat file source** | **Star schema 1** |  |
| **Abdelrahman 224774** | **Predicted\_graduation\_year**  **Parent\_Education\_background** | **Data type conversion 2** |  | **Star schema 2** |  |
| **Ruba 222335** | **Round up subjects**  **Education Type** |  | **Excel file source** | **Star schema 1** |  |

**Description of Data sets and Columns**

This dataset provides detailed information on students Education in Egypt, including their parents’ education levels, types of educational system, and academic performance for 10 subjects. The dataset was created for exploring what can influence student performance based on multiple factors.

Column details:

1. **Student Demographics**:

* **Student Name**: student identifier.
* **Student Age**: Provides an understanding of age variation through students in different academic years, will be useful in knowing their performance relative to age.
* **Student Year**: indicating the current academic year for each student.

1. **Parental Education**:

* **Father Degree** and **Mother Degree**: These columns indicate the highest degree achieved (High School, Bachelor’s, and PhD) by each parent and if they do not have any degree, it will be NONE -NONE here indicates to an info.

it will help indicating the relation between parental education level and students’ performance.

1. **Educational System**:

* **Education Type**: This column specifies the educational system the student is enrolled in, such as Thanweya (an Egyptian secondary education program), IB (International Baccalaureate), or IGCSE (International General Certificate of Secondary Education). Different systems may highlight different skills or subjects, impacting overall academic performance.

1. **Academic Performance**:

* **Subject\_1**: **Subject\_10**: These columns contain student scores for each of the 10 subjects. This will allow identifying student strengths and weaknesses in many academic areas.

**ETLs description**

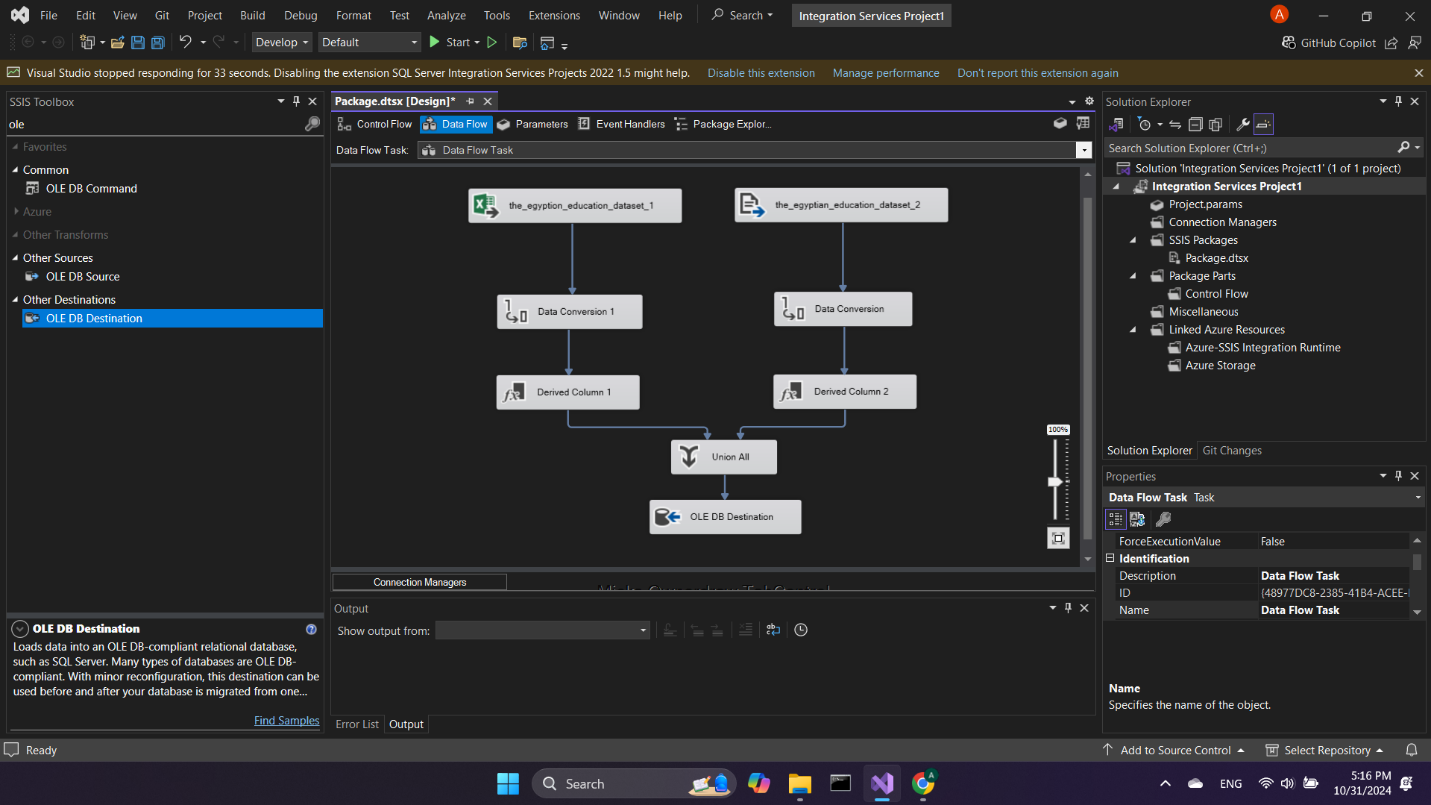
|  |  |
| --- | --- |
| AVG Score | This transformation is to sum the average score for each student by getting the sum of the subjects and dividing it by the 10 (number of subjects) and then the ROUND function added. |
| Age category | This transformation is to categorize the student ages to +16 and 16-. Knowing that info may help in the data mining phase. |
| Split Student name | This transformation is to split the student’s name to first name and second name via the space condition between both as before the space is the first name and after the space is the second name. |
| Student year | This transformation is to add underscore sign instead of the space between the year and number. (Year 8) 🡪 (Year\_8) |
| Predicted graduation year | This transformation is to predict the graduation year of each student considered the 12 years of education. |
| Parent education background | This transformation is to determine if there at least a mother or a father has any education background. If both have, then available. If one has, then available. If no one has, then unavailable |
| Round up subjects | This transformation is to round up the subjects’ marks to be easily read. It also will ease any calculations. |
| Education Type | This transformation is to add “Egyptian” beside “Thanaweya” to be “Thanaweya (Egyptian)”. It’s for anyone who doesn’t know the Thanaweya system. |

**Star schema description**

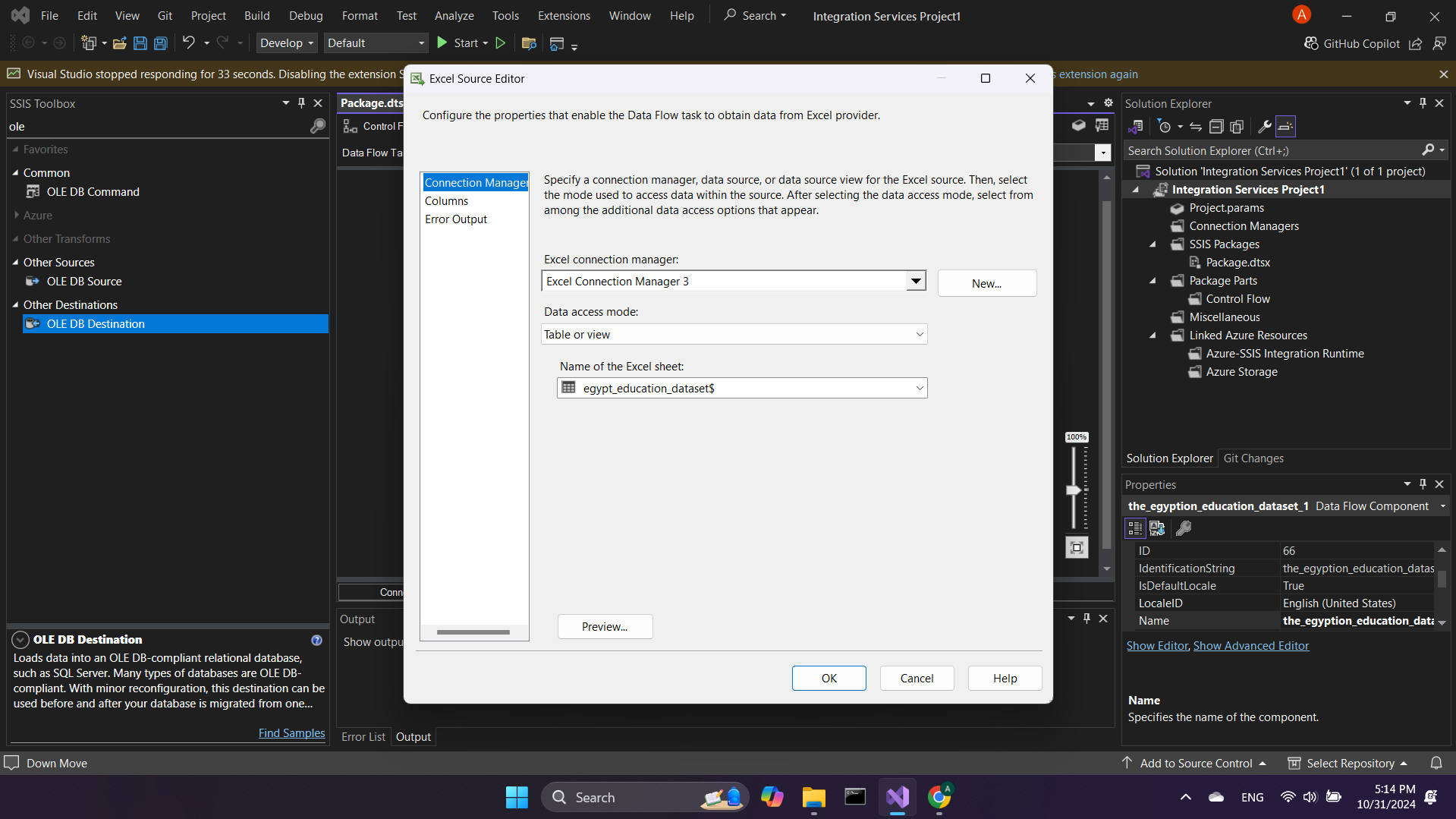
* **Star schema 1(FactStudentPerformance**) This schema enables educational institutions to analyze:
  + **Overall Student Performance and Demographics**: Evaluate student performance by comparing average scores and grades.
  + **Performance Trends by Subject**: Identify performance trends and strengths in specific subjects.
  + **Education Type Analysis**: Analyze and compare student performance across different educational programs to know their impact on student results.
* **Star schema 2(FactStudentPerformanceByParentalEducation)** This schema enables educational institutions to analyze:
  + **Impact of Parental Education on Performance**: Compare student scores and evaluate the relation between the parents’ education level with the student performance.
  + **Performance Trends by Subject and Parental Background**: See if parental education influences performance in specific subjects.
  + **Education Type Analysis**: Understand if education types have different performance patterns based on parental background.

Screen shots:

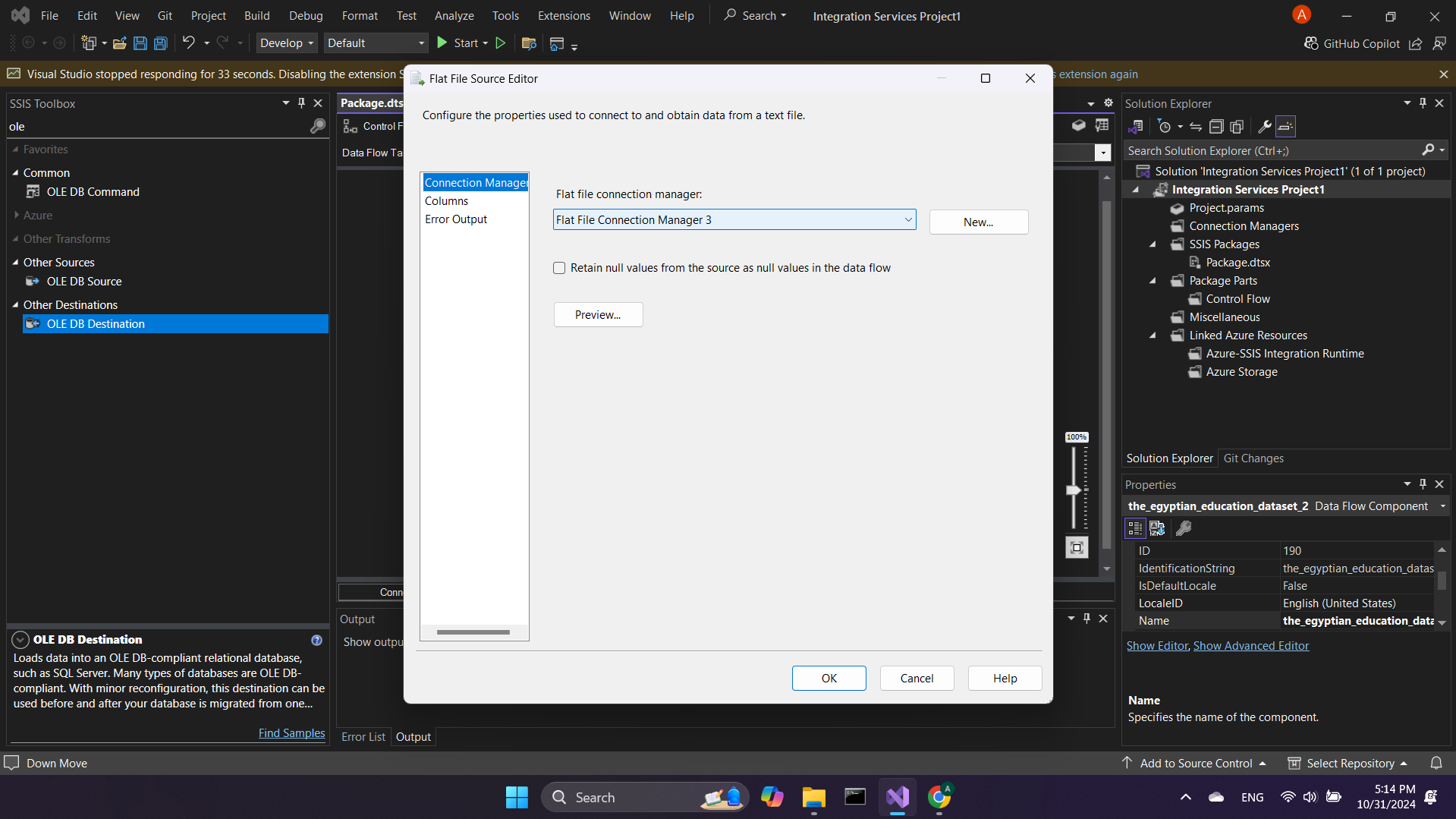
Overview of the component's connections



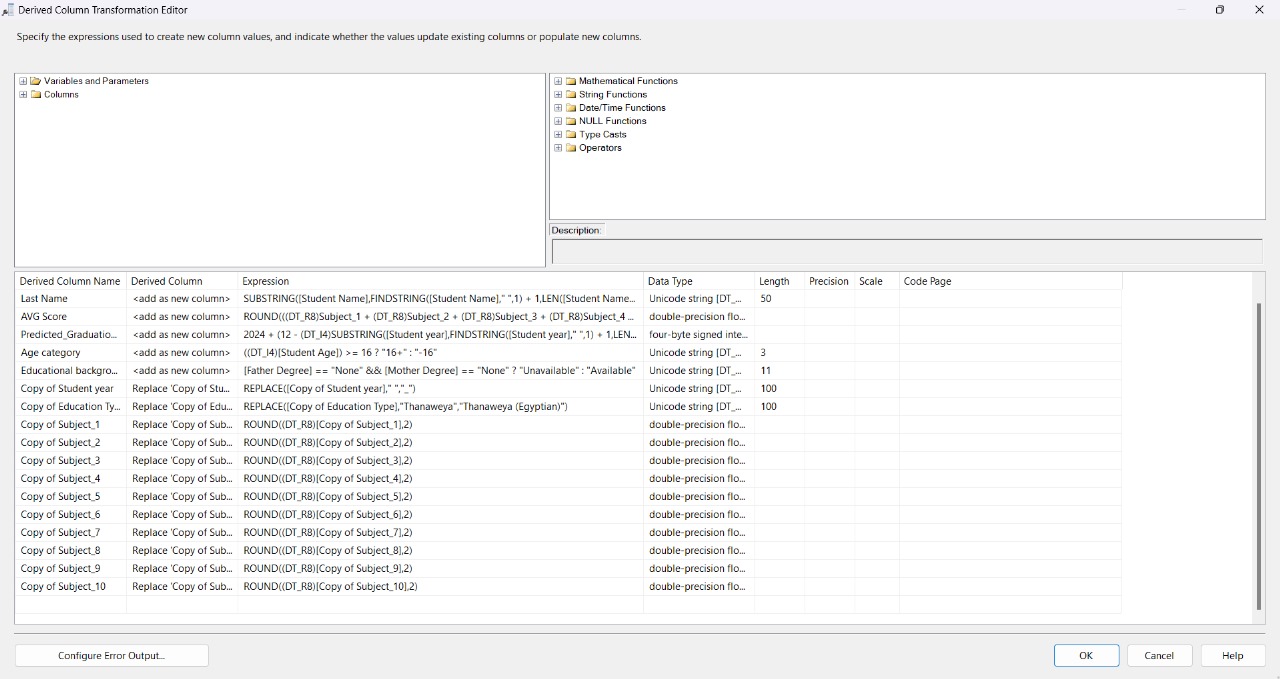
File 1 connection: excel file source connecting for importing data

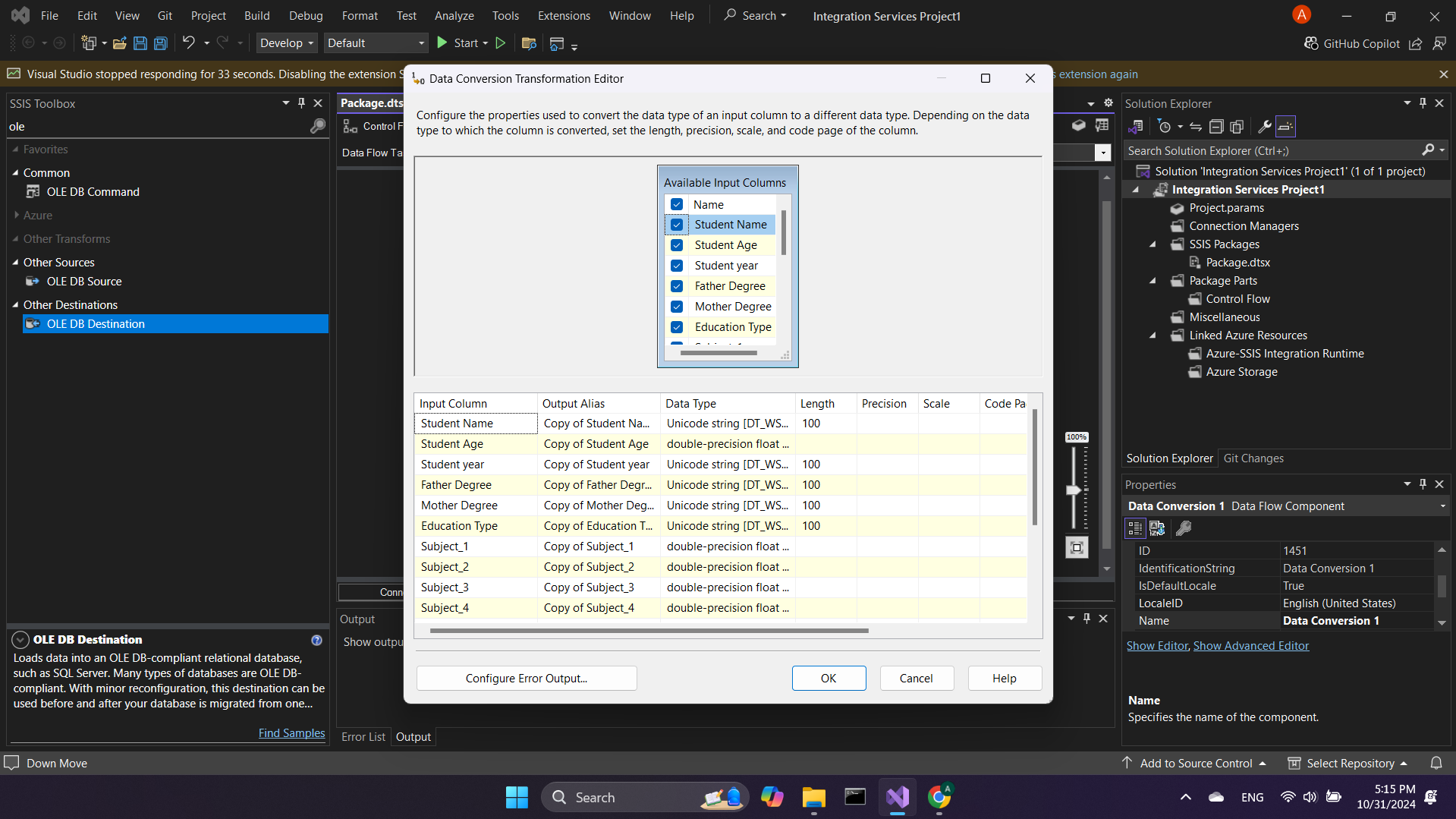


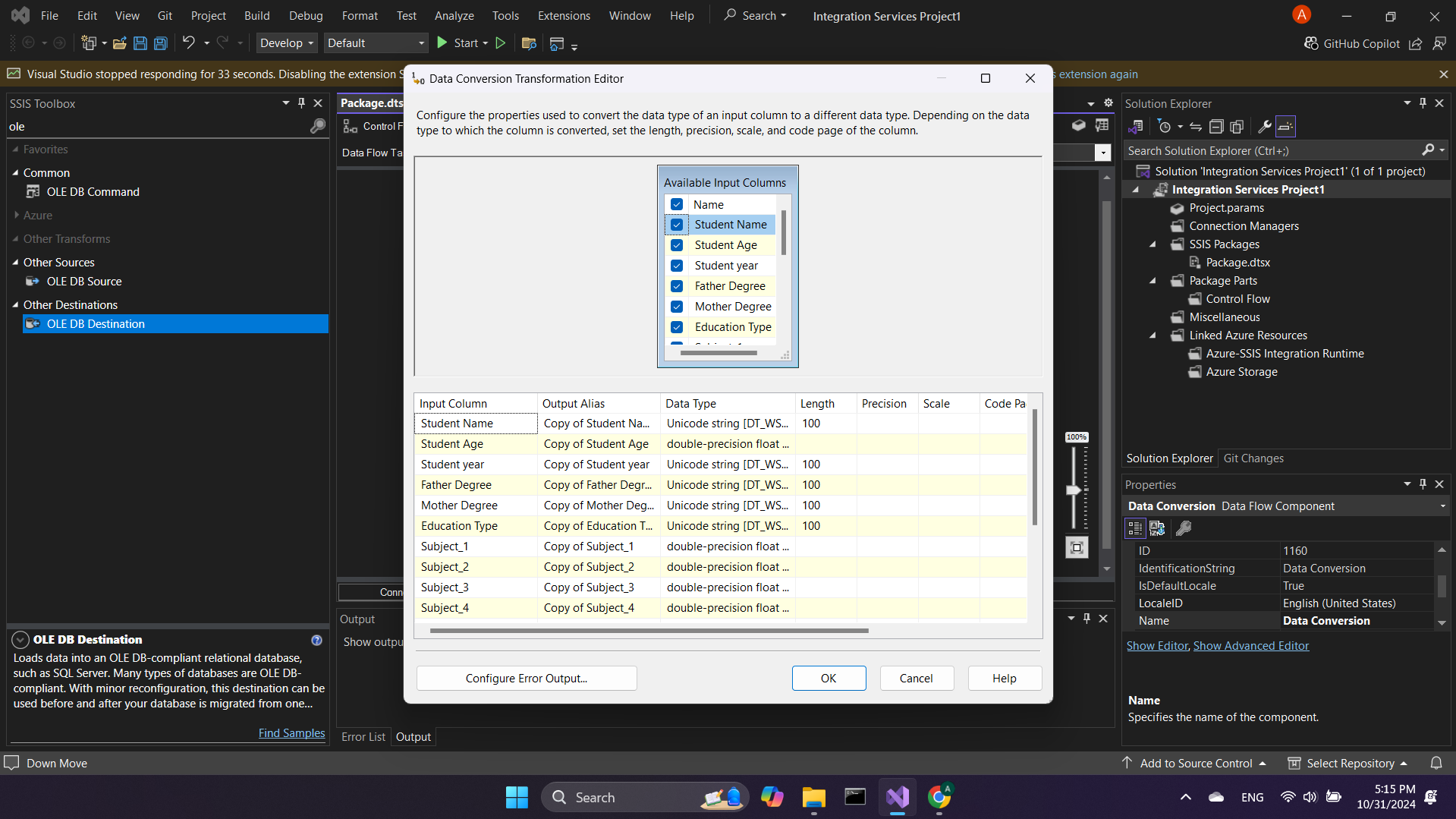
File 2 connection: Flat file source (csv file) connecting for importing data

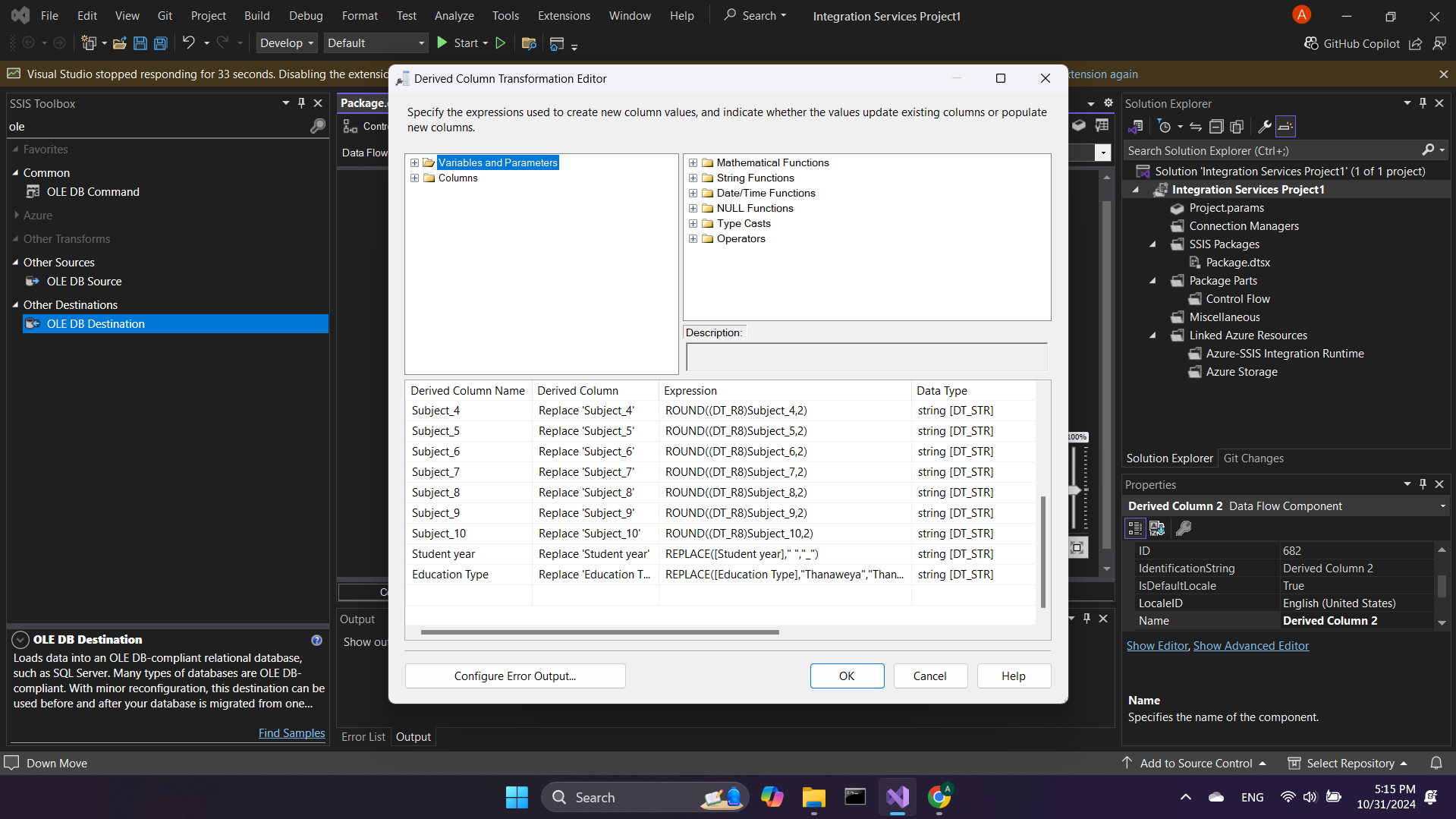
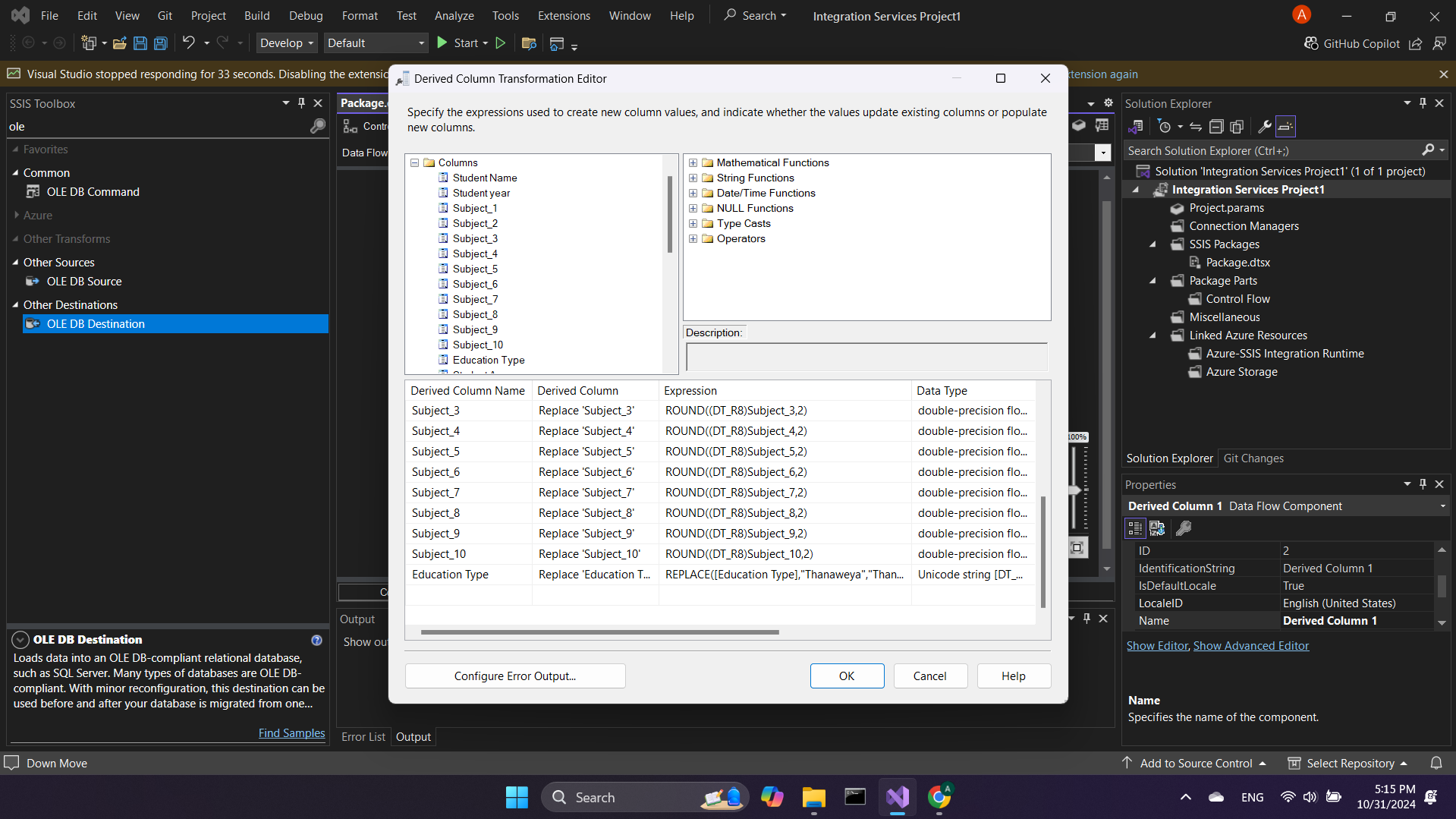


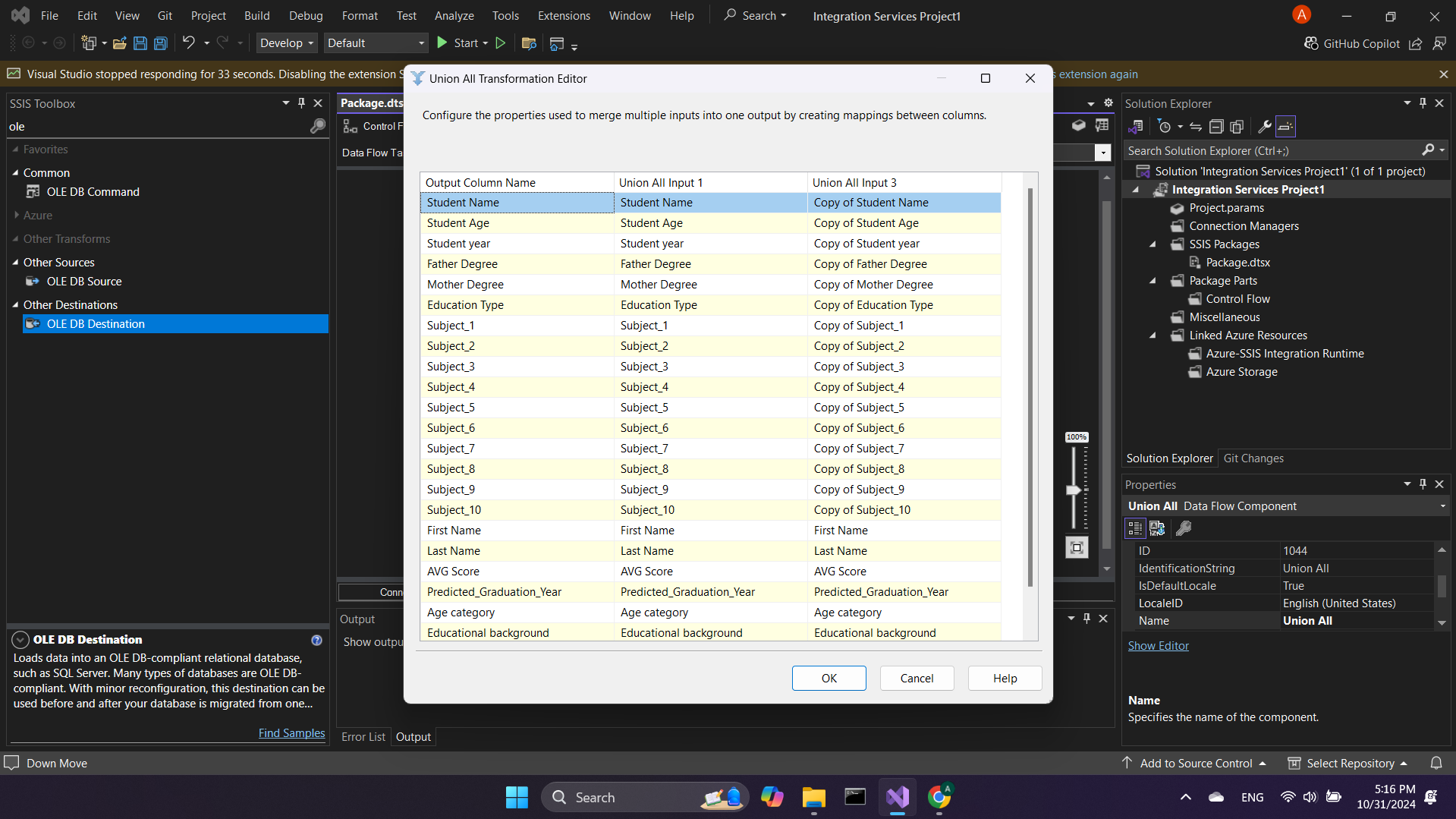
The ETLs expressions/Logic

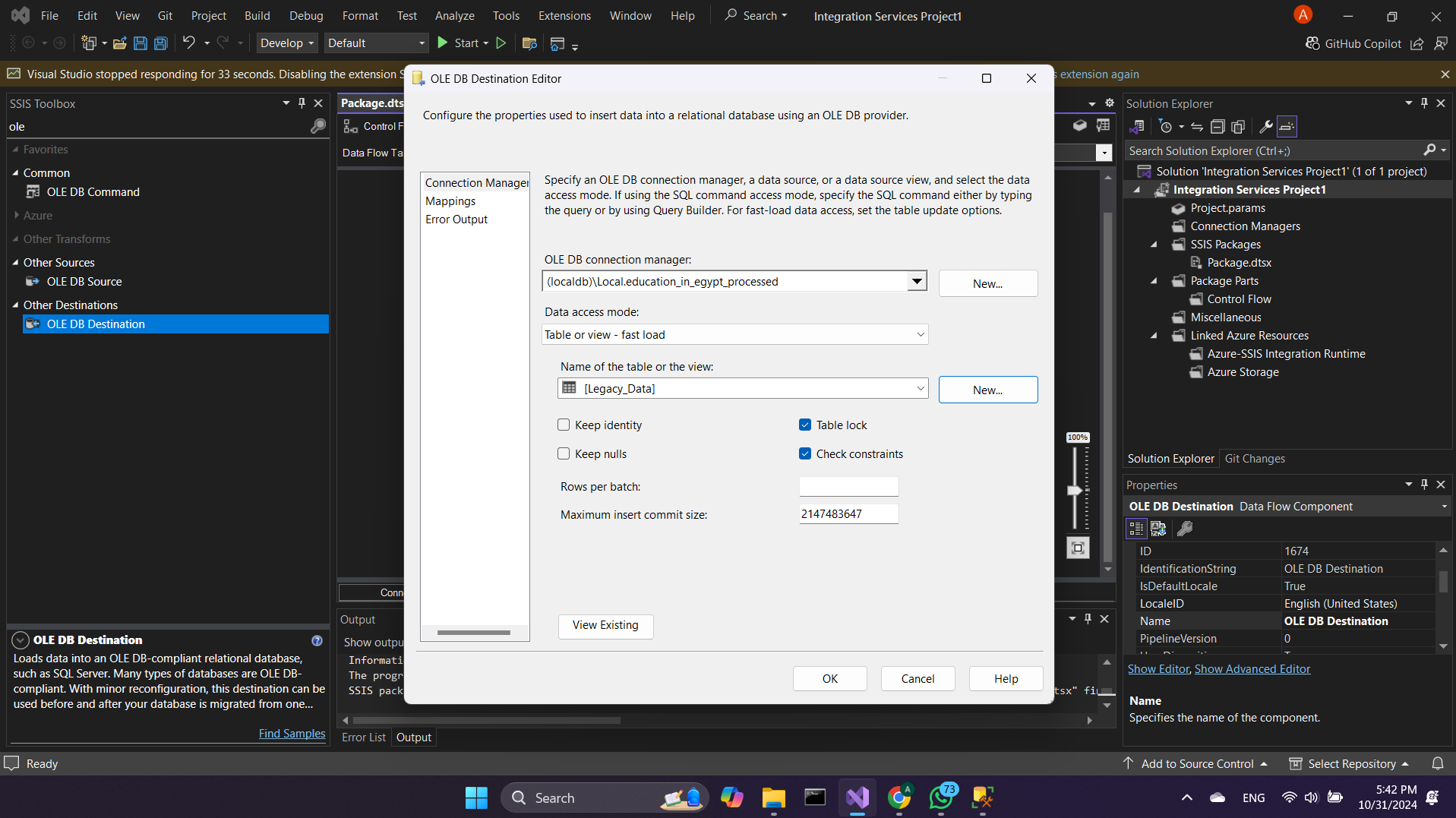


Data types transformation for the first file source (excel file) so both files' data types will be exactly same so we can union them into the same destination

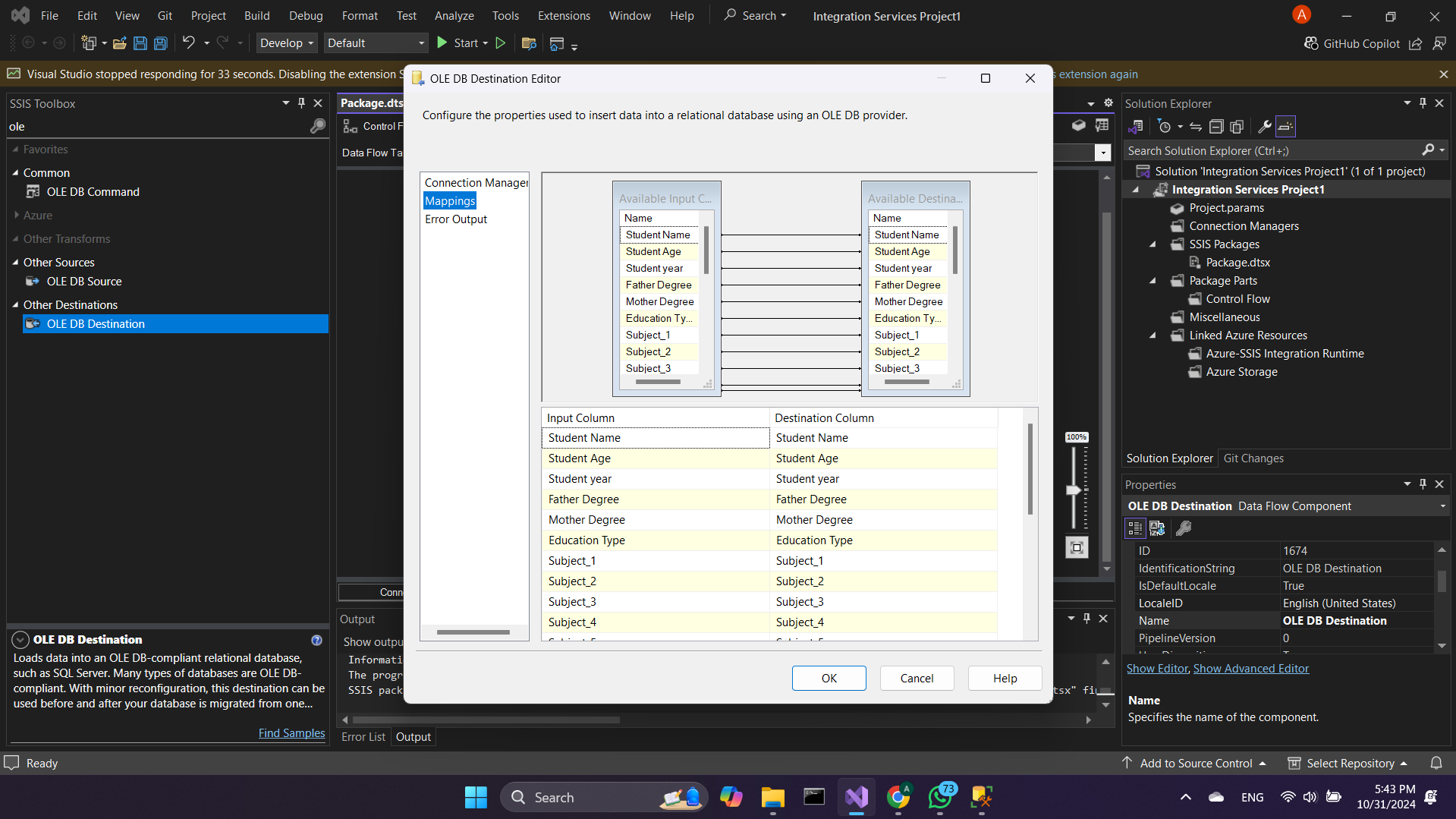
Data types transformation for the first file source (excel file) so both files' data types will be exactly same so we can union them into the same destination

The same derived column transformation edits applied for both data sources 

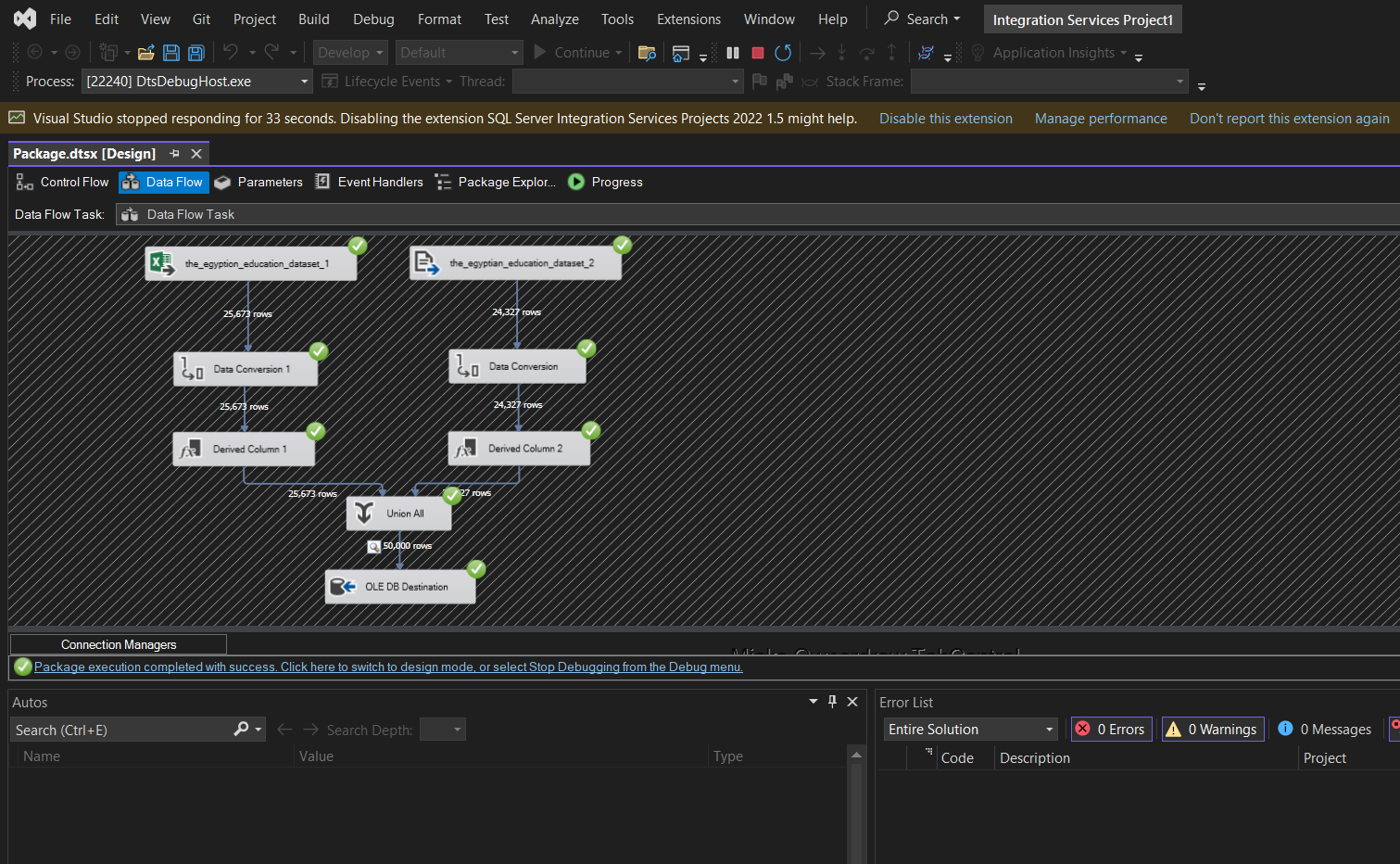
Union all tool for collecting the derived data from both sources and arrange them to enter the destination database 

OLE DB destination connection: connecting to the database destination 

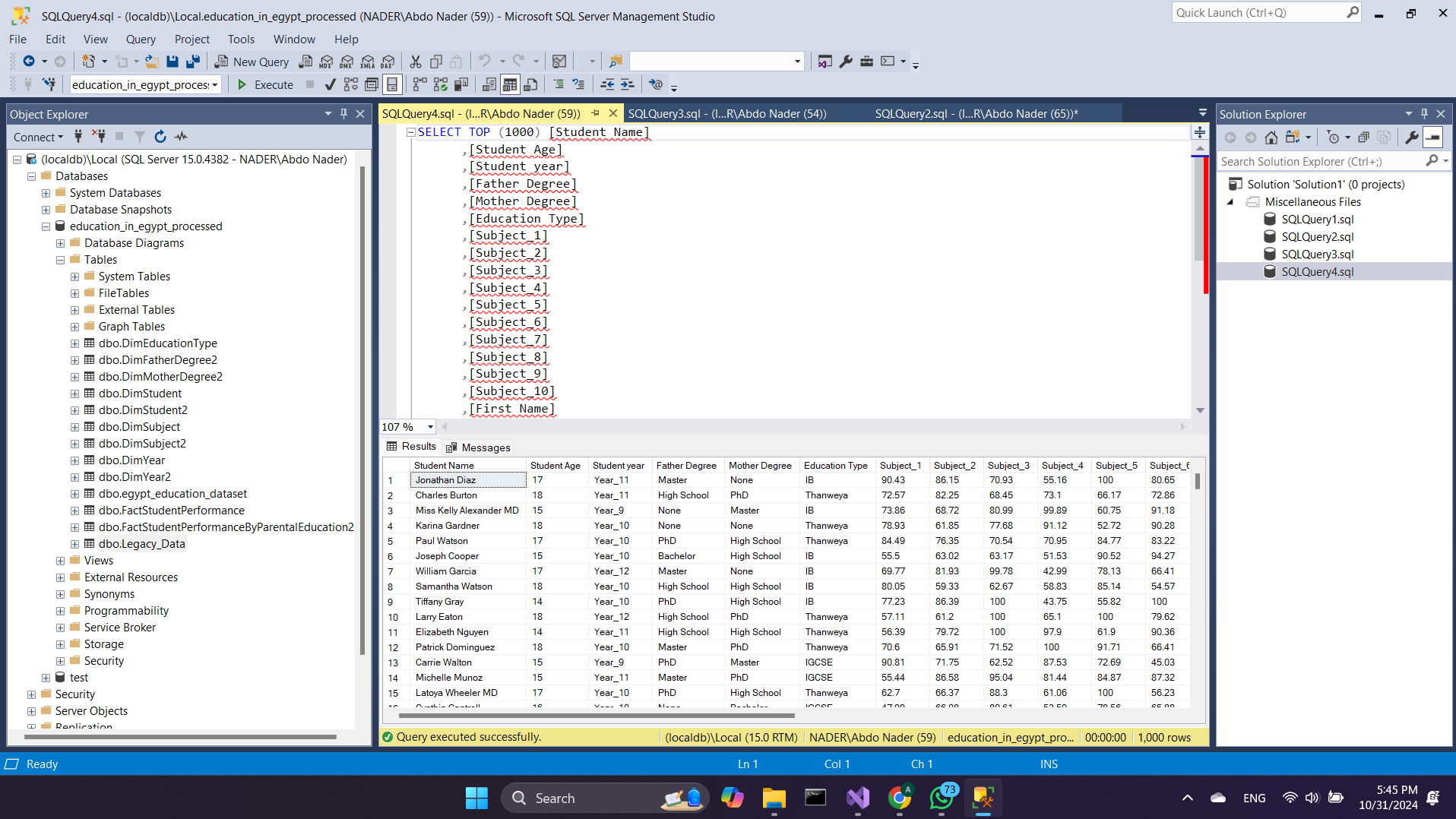
Checking if the data is mapped correctly for the destination database

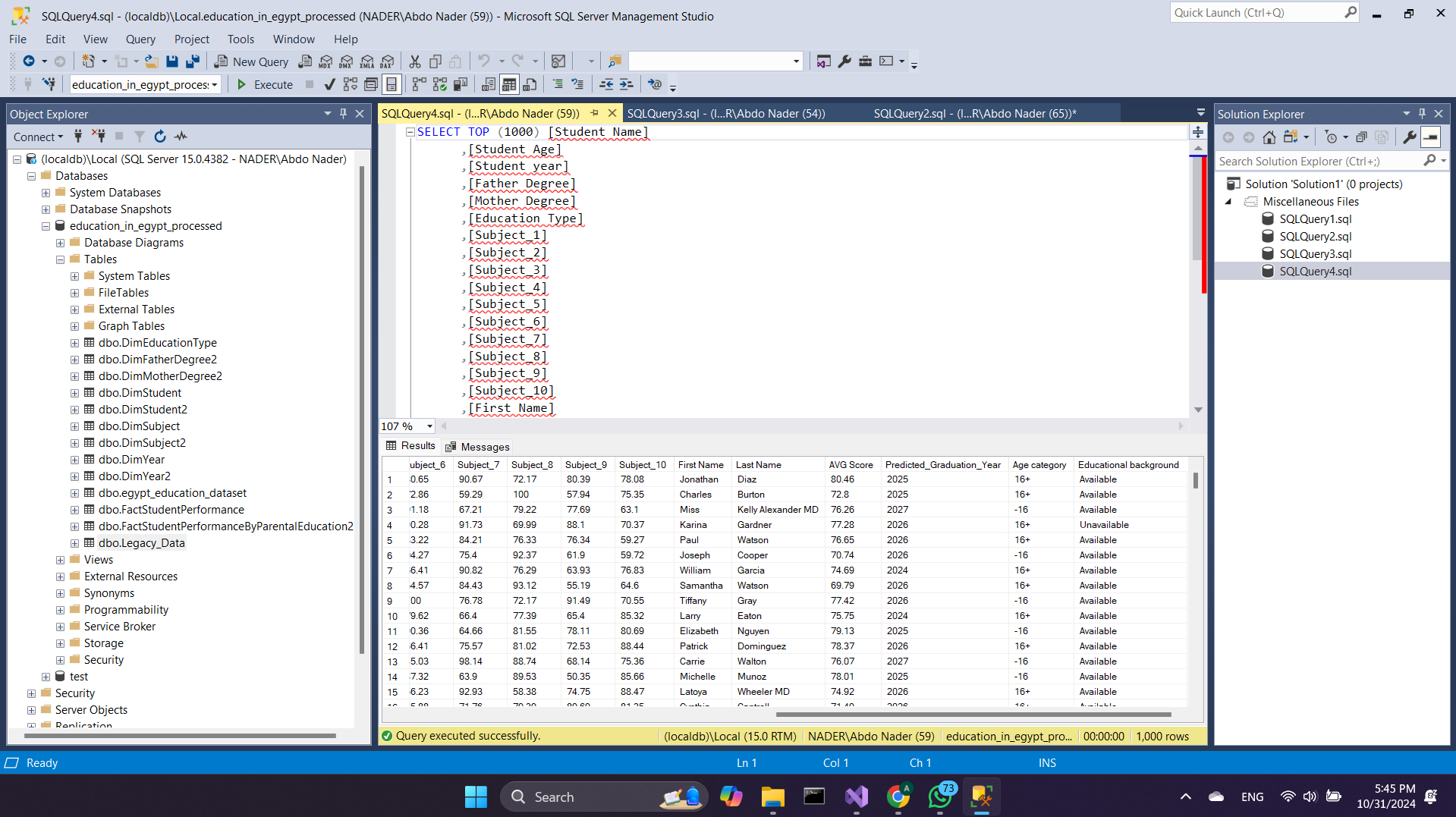


Running perfectly without any errors

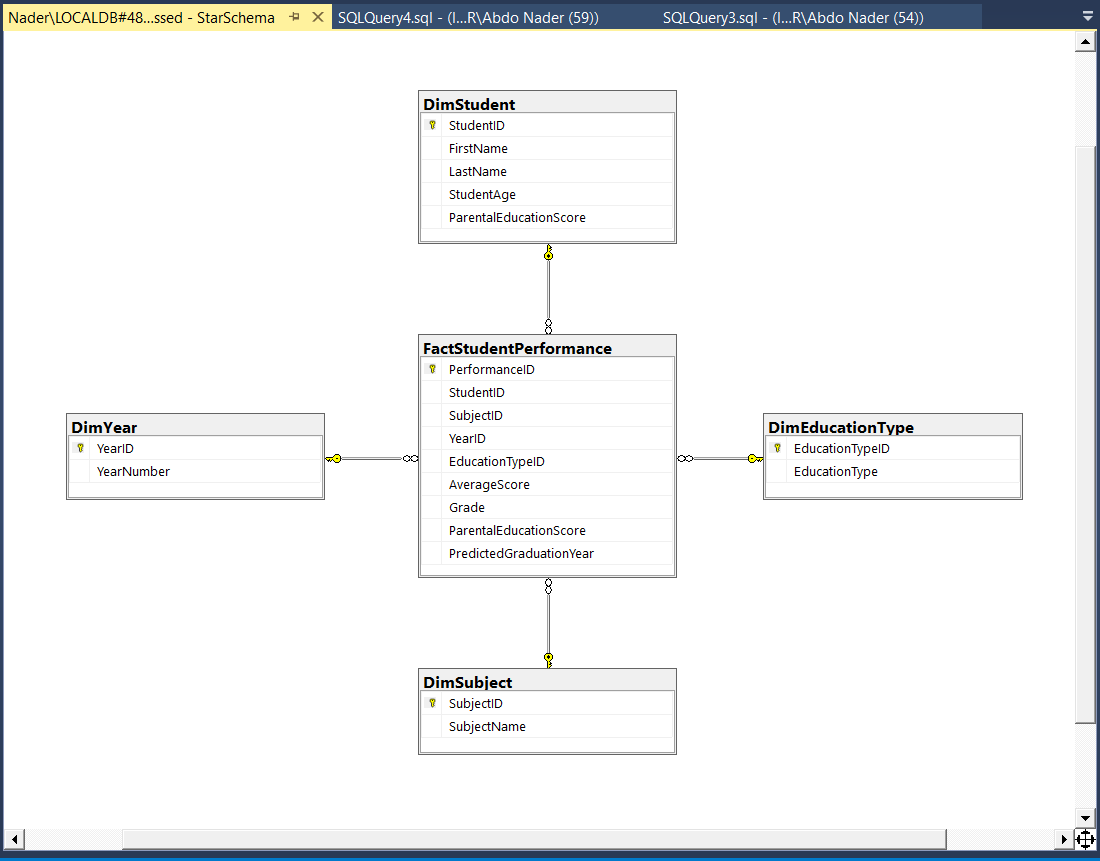


Checking the data in the databse

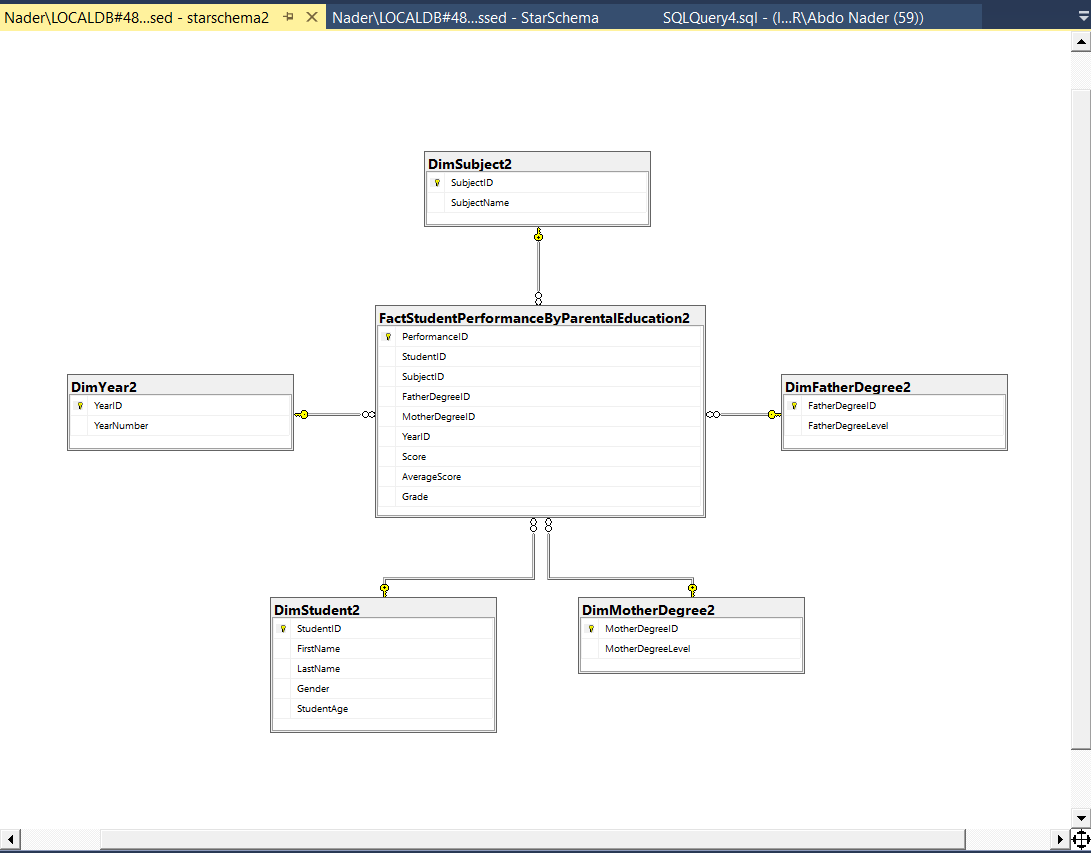




Star schema 1:



Star schema 2:



**Preprocessing:**

* **Checking nulls:**

There aren’t any nulls in the dataset.

There are ~10K cells at father degree and ~10K cells at mother degree that have **‘None’** but it indicates for an info which is that father or mother doesn’t have any educational degree. So, we will not remove them

**A screenshot of a computer

Description automatically generated**

* **Checking duplicates:**

**There are not any duplicates rows in the dataset. As it’s impossible to find two students who have the same marks in the 10 subjects.**

**The last line in the terminal**

**A screenshot of a computer program

Description automatically generated**

* **Removing columns:**

**After determining the targeted feature that we will make in the mining, We found that ‘First name’ and ‘Last name’ columns not important to be there and it’s not affected on the mining phase. We have ‘full name’ column which is enough. Removing useless data is a good thing in terms of the model performance and prediction.**

**Note that we made good progress in the warehousing phase, we edited and added some columns supposed to be in the preprocessing. I think the dataset now is ready for applying the algorithms**