PROJECT REPORT

On

**STUDENT REPORT**

Submitted For the Partial Fulfillment Of

The Data Integration Full Stack Developer Intern

At Cognizant

**By**

Molla Shammir Baig 2158546

Manish Thapa 2179805

**From the Cohort**

INTCDB22DW158

**Under the guidance of**

Mrs.Bera Ghosh Munmun

Mentor

Ms. Nandini Keerthi

Coach



**ACKNOWLEDGEMENT**

We must not forget to acknowledge everyone who has provided constant support to us during our internship. First and foremost, we would like to express sincere gratitude to our supervisor Mrs. Bera Gosh Munmun, and Our Coach Ms. Nandini Keerthi for their continuous support and motivation in fulfilling the pursuance of carrying out this project endeavor. Without their guidance and persistent encouragement, this project work would not have been possible. she has been a tremendous mentor for us throughout this Internship journey. Most of her advice about our career growth has been priceless.

A special thank is due to the entire cognizant family. Words cannot express how grateful we are to all our session teachers and batchmates for all the support that they have given throughout the internship.

Finally, we would like to thank everyone who has provided assistance, in whatever little form, towards the successful realization of this project but with an apology that we could not mention everybody’s name individually. We are very grateful to be associated with such a great company and to got to learn under the guidance of such amazing mentors.

# Introduction

The Student Report is a system that takes care of the complete internal details of a student which includes personal details, academic performance and many more. It allows curate the students list according to their performance. Also departments can bhi distinguished according to the different criteria, which may be student performance and so on.

## About this document

## Purpose & Scope of the document

The purpose of the software requirements document is to systematically capture requirements for the project and the system “Student report” to be developed. Functional requirements of this system are captured in this document. It also serves as the input for the project scoping.

The scope of this document is limited to addressing the requirements from a user, quality, and non-functional perspective. It is recommended that design aspects are not added in this document

## About the Software System

The software system is used to load a sample Data Warehouse for an Educational institution. Data from the source system are provided in a flat file extract and to be loaded to Data Warehouse in a relational Database. The educational institution will be using the data loaded in to Data Warehouse for their decision making.

**1.2.1 Scope of the system**

The scope of the system is explained through its modules as follows

St

* **Student Details**—This contains the details of all the students enrolled which consists of Student ID, Student Name, Student Address, Pincode, Student Phone number, Department Id.
* **Student Marks –** This contains the marks of the students in five subjects and student Id.
* **Department Details –** This contains the information of the student with the deparment they are studying.

### Exclusions

### 1. The system will operate only on the modules discussed above and will not include any additional functionality.

### 2. Disaster recovery and failure recovery are not part of the scope

### 3. Any future changes to the source and target structures are not in scope

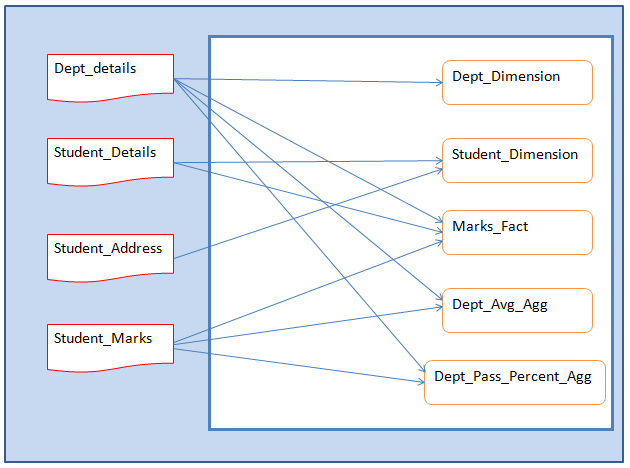
### System Perspective

The Student Report System is an independent software system developed to manage the activities like Registering New Student, curating the student details, etc. using the architecture.

### Architecture diagram

#### *Physical Architecture:*

A physical architecture is an arrangement of physical elements, (system elements and physical interfaces) that provides the designed solution for a Student and Department. It is intended to satisfy logical architecture elements and system requirements. Auto Identification Process follows a three layered architecture namely presentation layer, business logic layer and data access layer.



***Logical Architecture:***

The Logical Architecture defines the Processes (the activities and functions) that are required to provide the required student details. Many different [Processes](http://www.iteris.com/itsarch/html/pspec/pspecs.htm) must work together and share information to provide a student details. The Processes can be implemented via software, hardware, or firmware. Logical Architecture is independent of technologies and implementations.

### Impact of the System

This makes the operatibility and the accessiblity of the student and department details more easier, faster and secured.

### Assumptions, Risks / Constraints

**Assumptions:**

### All source files will be available in specified format

### Target database will be available to create Data Warehouse tables

### Informatica server will be able to connect to Data Warehouse database

### The system will not be available if there are any outages to Data Warehouse Database

### Design Constraints – from the template

* All source files should be ‘tab’ delimited files
* All target should be Data Warehouse Tables
* Every Dimension, Fact and Aggregate tables should be loaded using individual non-reusable sessions
* All session loading Data Warehouse tables should be present in one single workflow
* All sessions should be connected in a sequential manner in workflow.
* If there are any failure to one of the session, then following sessions should not run
* Workflow should succeed only if all sessions have completed successfully

**2 Underlying Concept**

**2.1 Informatica**

* Informatica is introduced as a software development company in the market. It provides a complete data integration solution and data management system. It launched multiple products that mainly focused on data integration.
* Informatica is used to extract required data from the operation of all systems and transforms the same data on its server and load it to the data warehouse.
* Informatica is also introduced as a data integration tool. This tool is based on the ETL architecture. It provides data integration software and services for different industries, businesses, and government organizations, as well as telecommunication, healthcare, insurance and financial services.
* It has a unique property to connect, process, and fetch the data form a different type of mixed sources.
* For example, we can connect with more than one server database.

**2.2 Informatica constraints:**

* **Data Extraction:** The process of reading and extracting the data from multiple source systems into the Informatica server is called data extraction. Informatica data can extract or read different methods such as SQL Server, Oracle, and many more.
* **Data Transformation:** Data transformation is a process of converting the data into the required format. Data transformation supports the following activities ,such as:

1. **Data Merging:** It integrates the data from multiple sources.
2. **Data Cleansing:** It cleans the data from unwanted or unnecessary information.
3. **Data Aggregation:** It aggregates the data using the aggregate function such as Sum(), min(), max(), count(), etc.
4. **Data Scrubbing:** It is used to derive the new data.

* **Data Loading:** Data loading is used to insert the data into a target system. There are two types of data loading, such as:

1. **Initial load or Full load**: It is the first step; it adds or inserts the data into an empty target table.

2. **Delta load or Incremental load or Daily load**: This step takes place after the initial load, and it is only used to load new records or changed the old records.

Informatica is an easy-to-use ETL tool, and it has a simple visual primary interface. You drag and drop the different objects and design process flow for data extraction transformation and load.

These designed process flow diagrams are called mappings. Informatica has the property to communicate with all major data sources such as RDBMS, mainframe, Flat files, XML, SAP, VSM, and many more.

**2.3 Need for Informatica**

* To perform some operations on the data at the backend in a data system, then we need Informatica.
* To modify, and clean up the data based on some set of rules, we need Informatica.
* By using Informatica, it is accessible to the loading of bulk data from one system to another.
* It provides a broad set of features such as integration of the data from multiple unstructured, semi-structured, or structured systems, operations at row level on data, and scheduling operation of the data operation.
* It also supports the features of metadata, so it keeps preserved the information of the process and data operations.

**3. Implementation**

## 3.1 Module Information

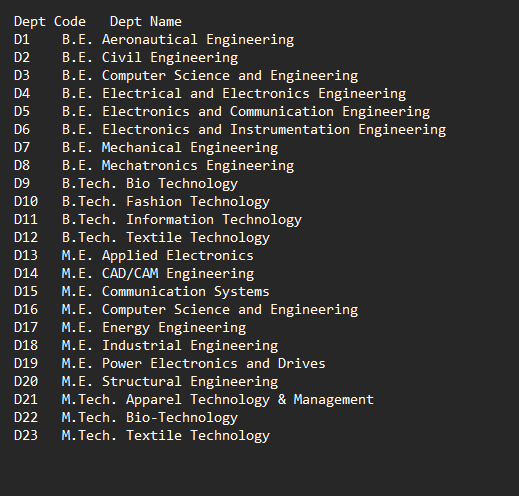
|  |  |
| --- | --- |
| **Data Warehouse Table load Logic** |  |
| **Department Dimension Table** | Department Dimension table should be loaded from Department Details source file as a one to one load |
| **Student Dimension Table** | Student Dimension table should be loaded from Student Details and Student Address source files  Postal code column should be cleansed to remove any extra space and special characters  Phone Number column should be validated for valid 10 digit phone number, if there are any invalid number found then phone number has to be loaded as NULL  All characters in address column has to be converted to upper case. |
| **Marks Fact Table** | Marks Fact table should be loaded from Student Details, Student Address and Student marks source files  Total has to be calculated as sum of all subject marks  Average column has to be calculated as average of all marks  Result has to be calculated as Pass or Fail, Students scoring 50 and above marks in all subjects are termed as Pass |
| **Subject Average Aggregate Table** | Subject Average Aggregate Table should be loaded from Department Details and Student Marks table  Table should contain average marks scored by students in all subjects by department |
| **Subject Pass Percentage Aggregate Table** | Subject Pass Percentage Aggregate Table should be loaded from Department Details and Student Marks table  Table should contain pass percentage for each subject by department |

**3.2 Creation of Database**

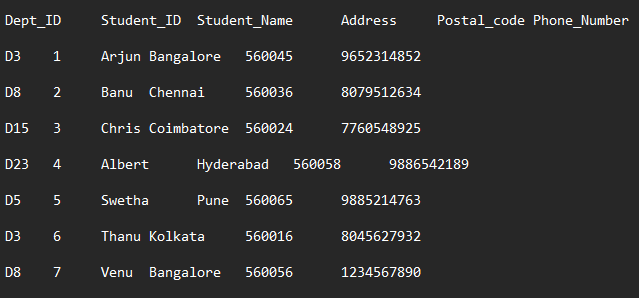
Create database orcl1;

**3.3 Creation of Source**

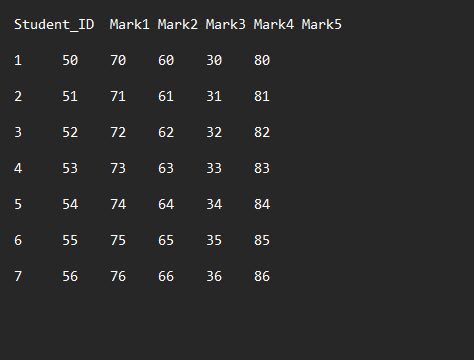
* **Department Details Strucuture**



* **Student Details Structure**



* **Student Marks Structure**



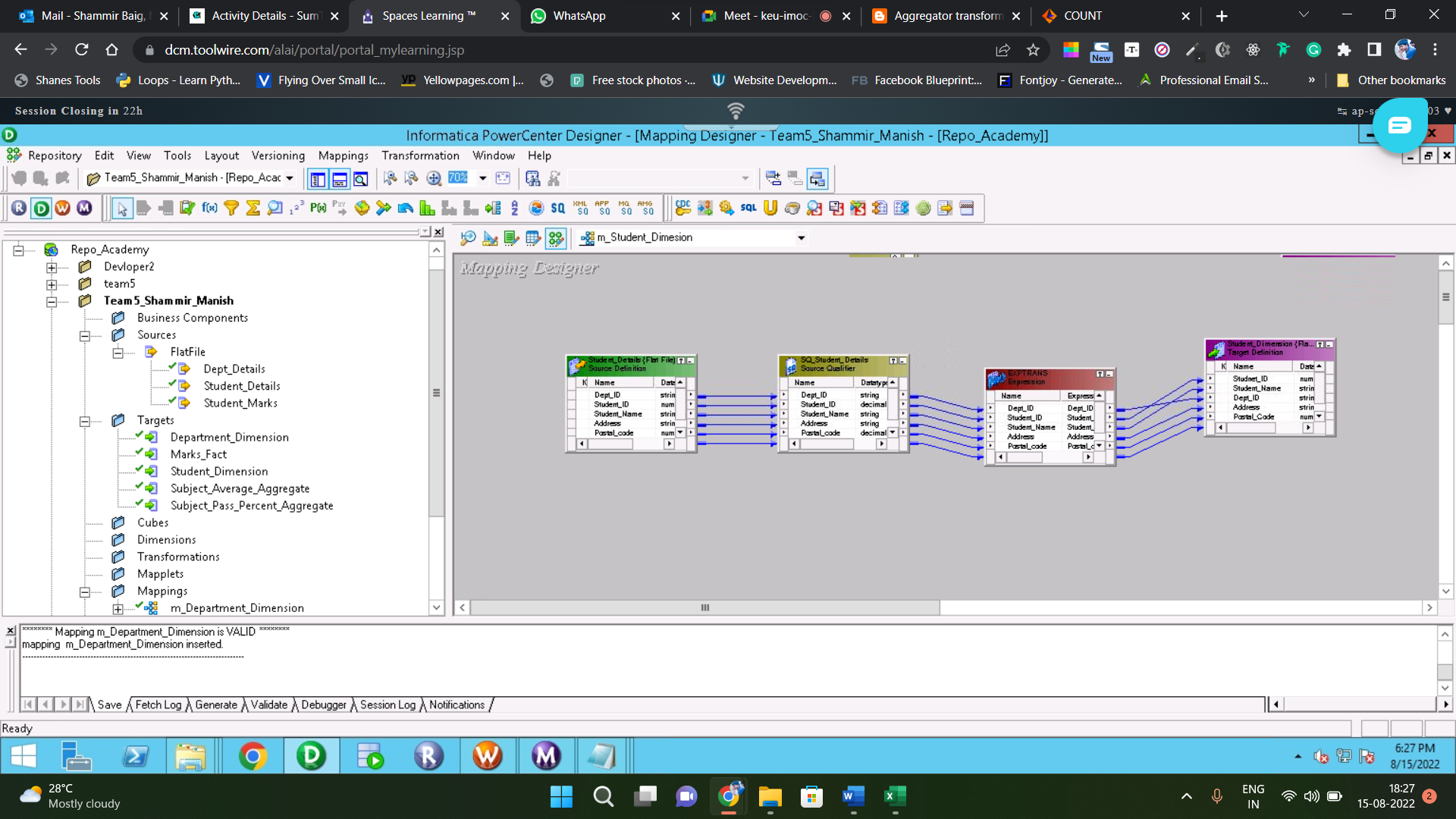
**3.5 Mappings:**

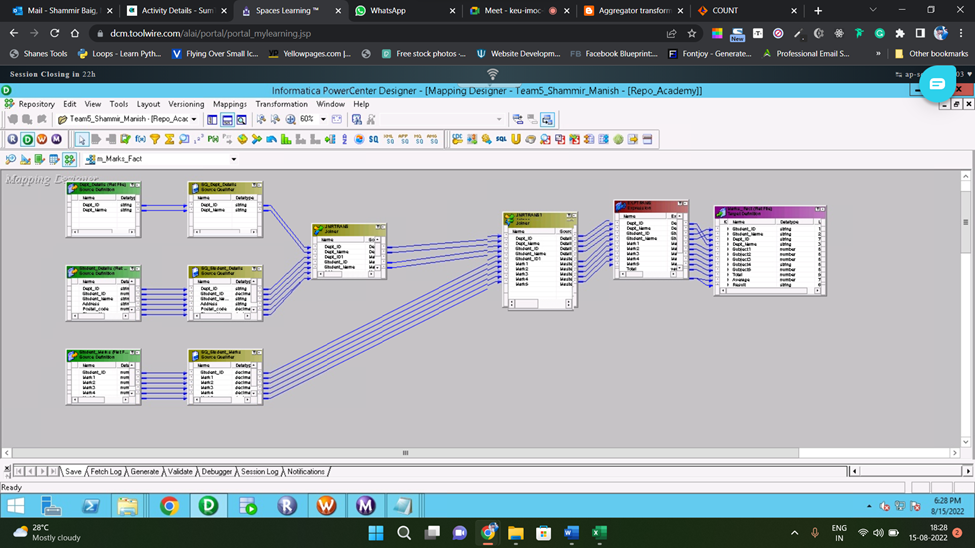
Mapping is a collection of source and target objects which is tied up together through a set of transformations. These transformations are formed with a set of rules that define how the data is loaded into the targets and flow of the data.

Mapping in Informatica includes the following set of objects, such as:

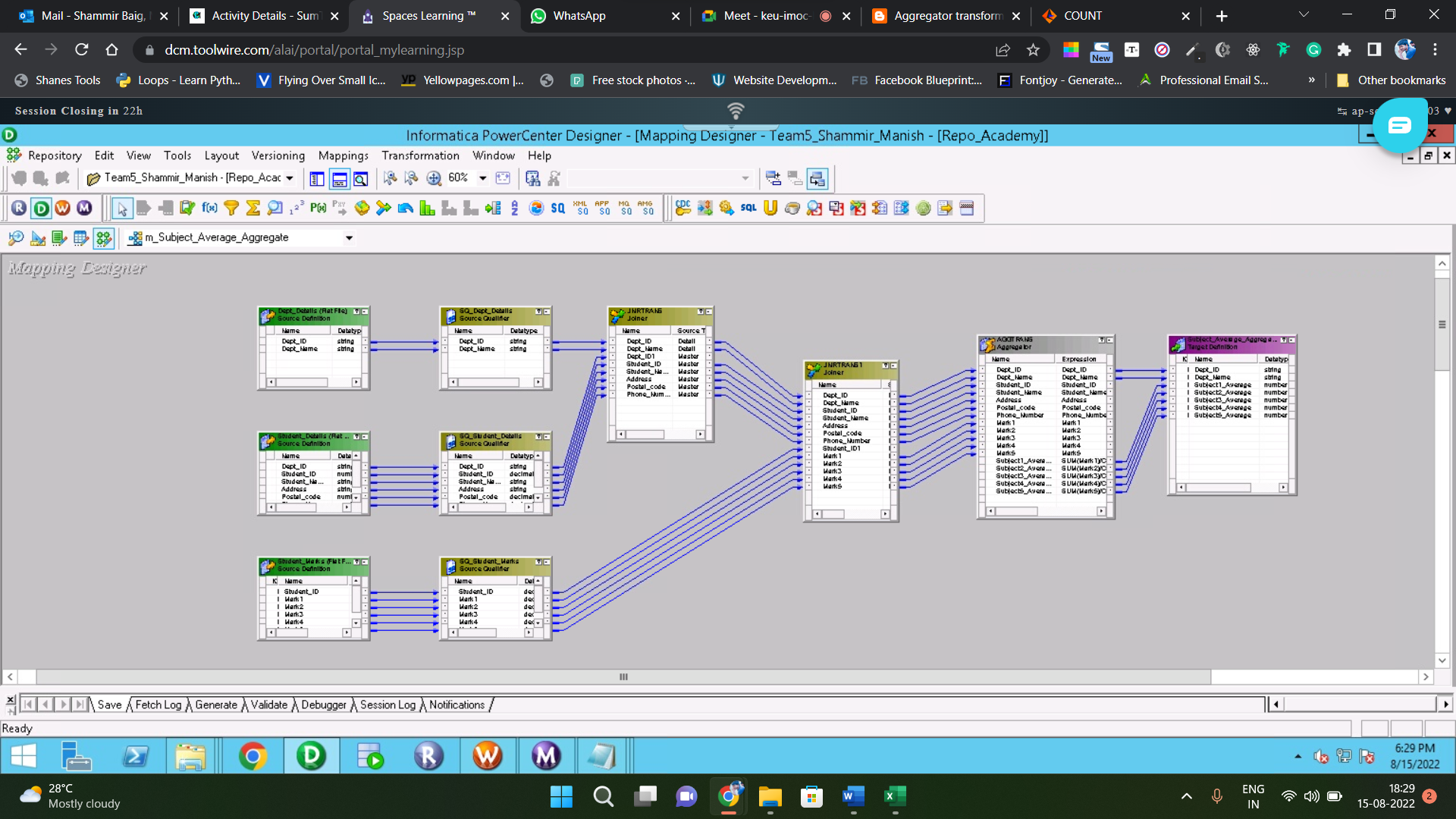
1. **Source definition:** The source definition defines the structure and characteristics of the source, such as basic data types, type of the data source, and more.
2. **Transformation:** It defines how the source data is changed, and various functions can be applied during this process.
3. **Target Definition**: The target definition defines where the data will be loaded finally.
4. **Links**: Link is used to connect the source definition with target tables and different transformations. And it shows the flow of data between the source and target.

To execute the mappings designed in PowerCenter, workflow is used and to create workflow Informatica workflow manager is used.

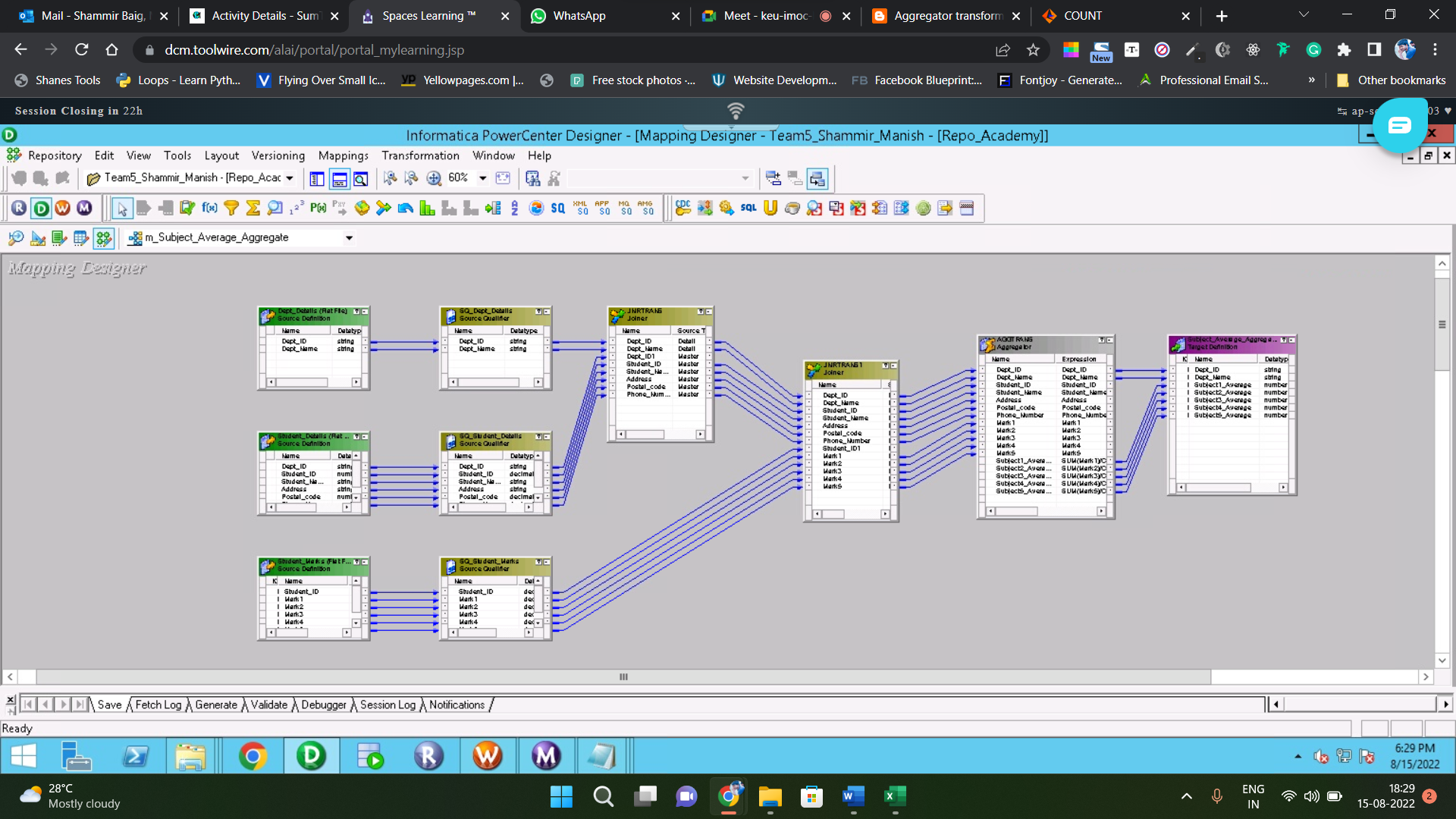
* **MAPPING OF STUDENT DIMENSION** 
* In PowerCenter designer ,We have imported a flat file as a source in the source analyzer and imported a target table from the database in the target analyzer and then created mappings according to the requirements.
* For the mapping of the Branch table, we have used transformations like expression, filter, look up, and have applied functions in order to meet the requirements.
* Once we finish creating the Mapping, we have to create the workflow and session for it in Workflow manager. Save and execute the workflow. After execution selected columns will be loaded to the target.
* **MAPPING OF MARKS FACT**

****

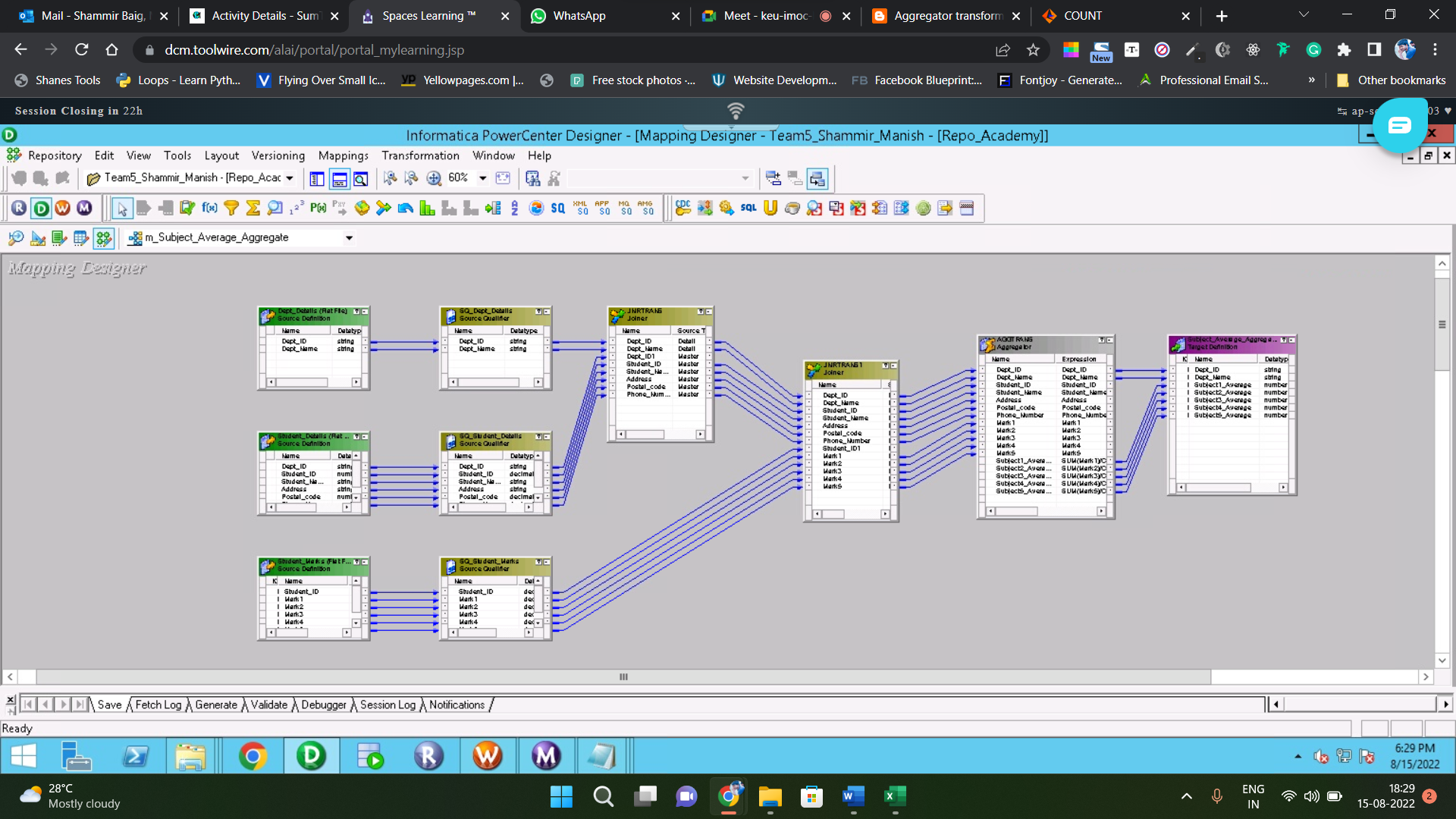
* In PowerCenter designer ,We have imported a flat file as a source in the source analyzer and imported a target table from the database in the target analyzer and then created mappings according to the requirements.
* For the mapping of the Student Marks table, we have used expression transformation and have applied functions in order to meet the requirements.
* Once we finish creating the Mapping, we have to create the workflow and session for it in Workflow manager. Save and execute the workflow. After execution selected columns will be loaded to the target.
* **Create mapping for Student Avg Aggregate**



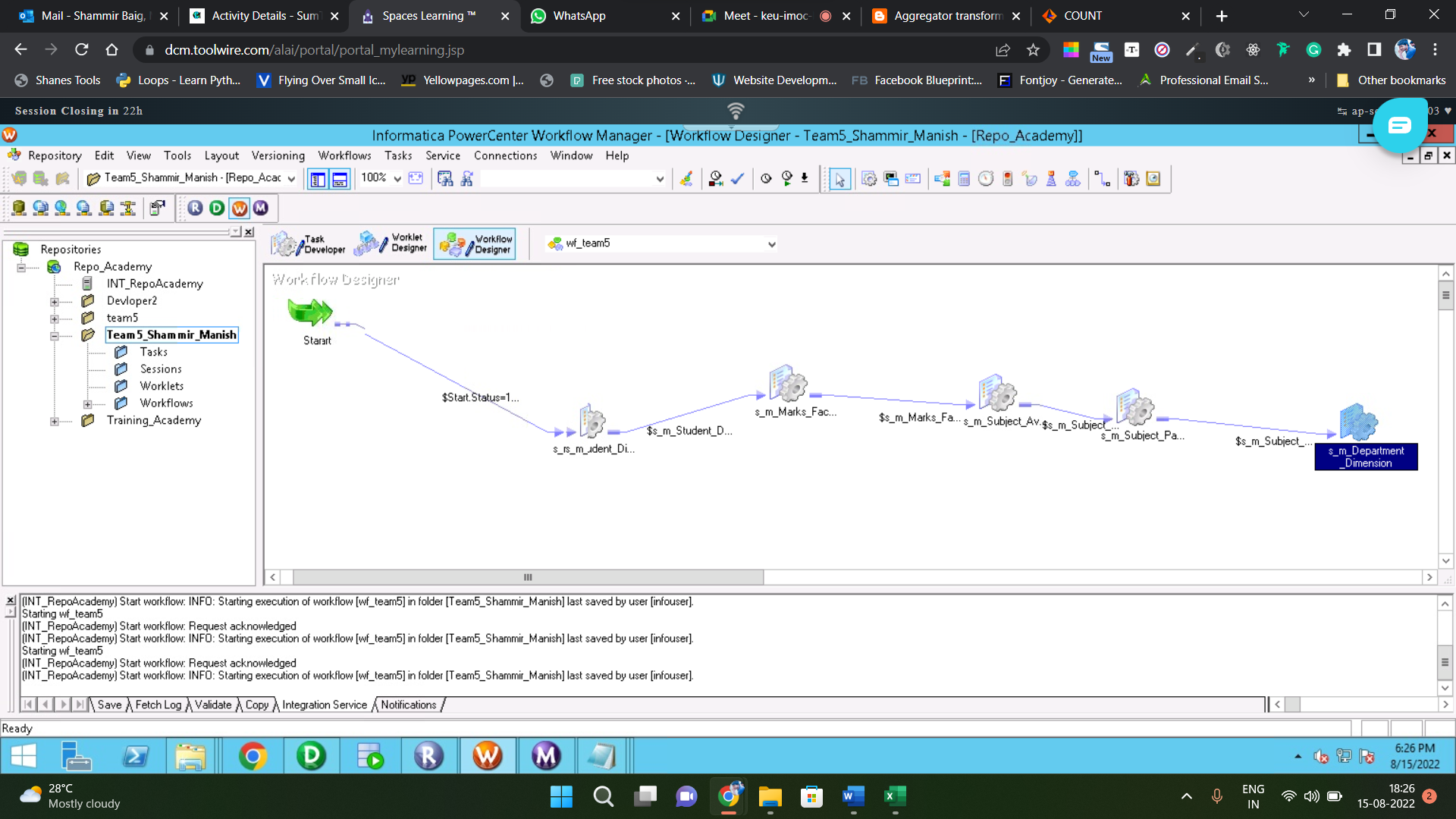
* In PowerCenter designer ,We have imported a flat file as a source in the source analyzer and imported a target table from the database in the target analyzer and then created mappings according to the requirements.
* For the mapping of the Student Average Aggregate, we have used expression transformation and have applied functions in order to meet the requirements.Once we finish creating the Mapping, we have to create the workflow and session for it in Workflow manager.
* **Create mapping for Subject Pass Percentage**



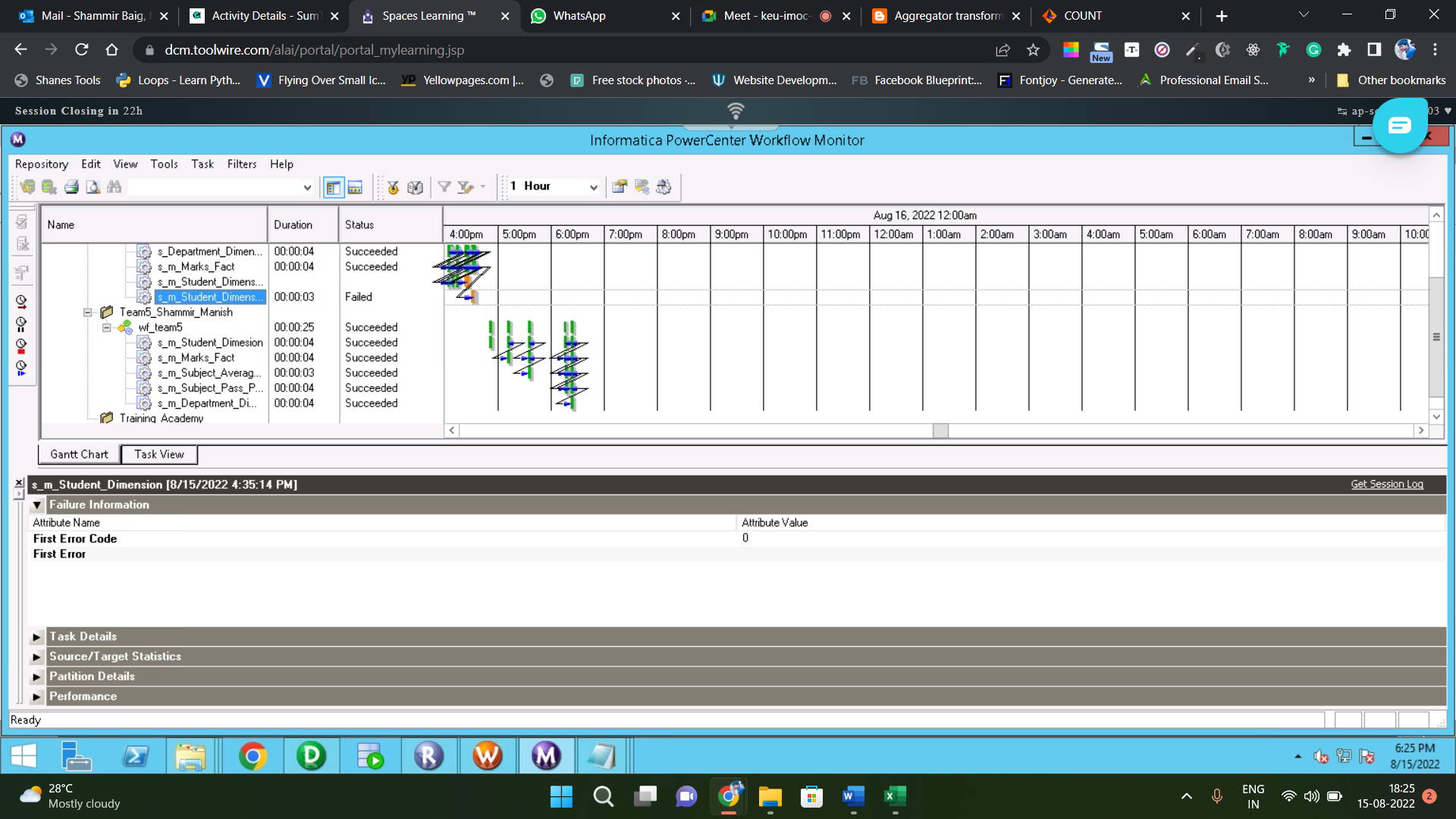
* In PowerCenter designer ,We have imported a flat file as a source in the source analyzer and imported a target table from the database in the target analyzer and then created mappings according to the requirements.
* For the mapping of the Subject Pass Percentage Aggregate, we have used expression transformation and have applied functions in order to meet the requirements.Once we finish creating the Mapping, we have to create the workflow and session for it in Workflow manager.
* **Create mapping for Department Dimension**



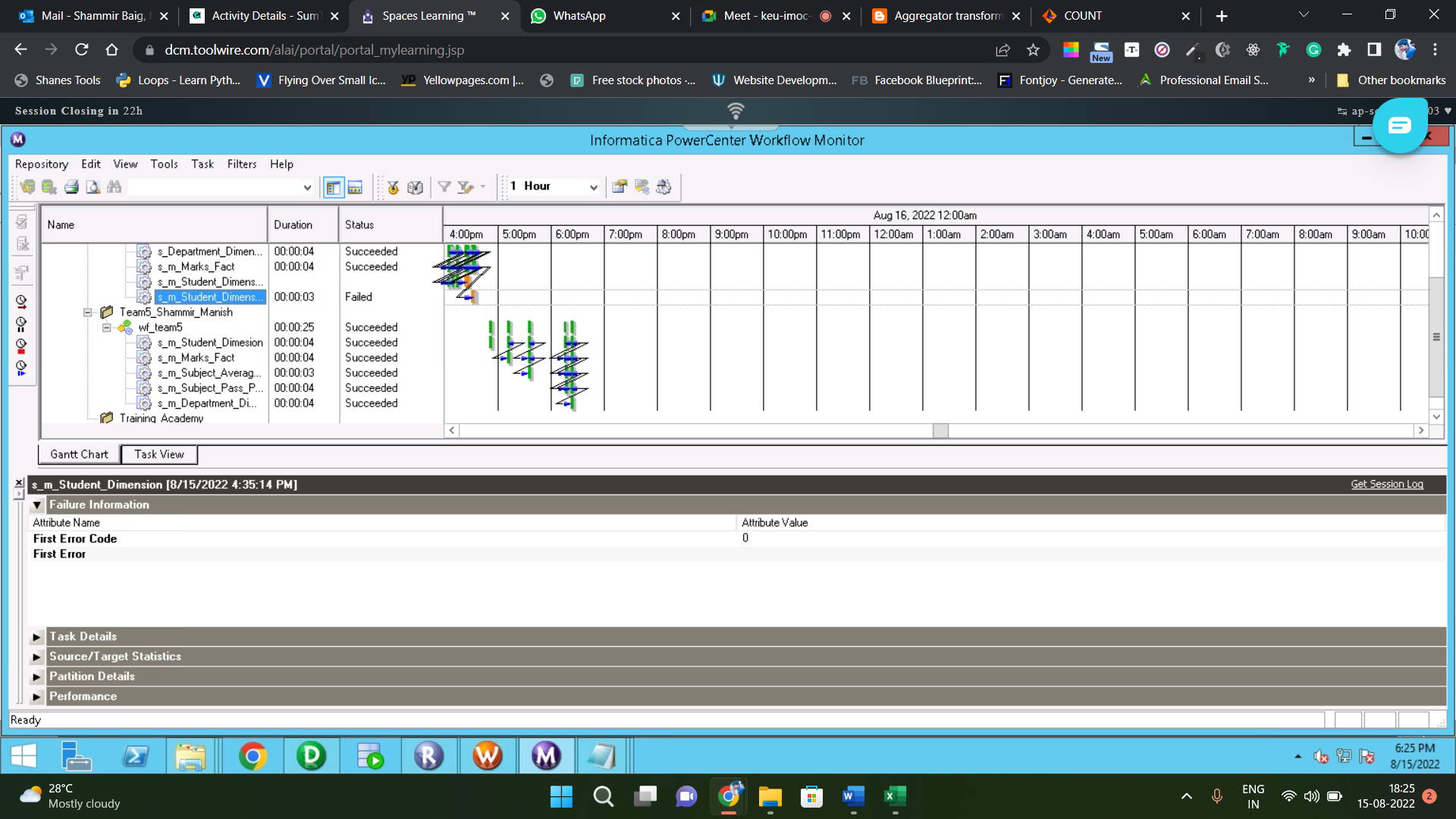
* In PowerCenter designer ,We have imported a flat file as a source in the source analyzer and imported a target table from the database in the target analyzer and then created mappings according to the requirements.
* For the mapping of the Subject Pass Percentage Aggregate, we have used expression transformation and have applied functions in order to meet the requirements.Once we finish creating the Mapping, we have to create the workflow and session for it in Workflow manager.
* **WorkFlow Manager**



* Save and execute the workflow. After execution selected columns will be loaded to the target.
* **WorkFlow Monitor**

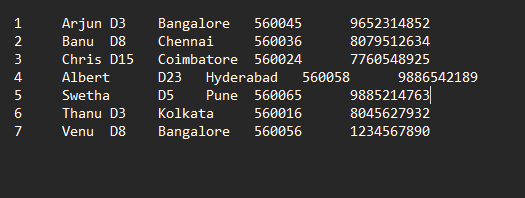


* **Repository Manager**

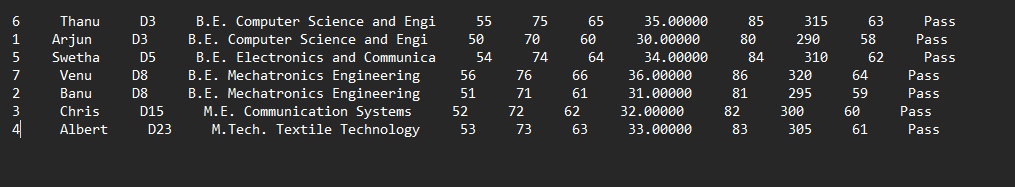


**4.Target Outputs:**

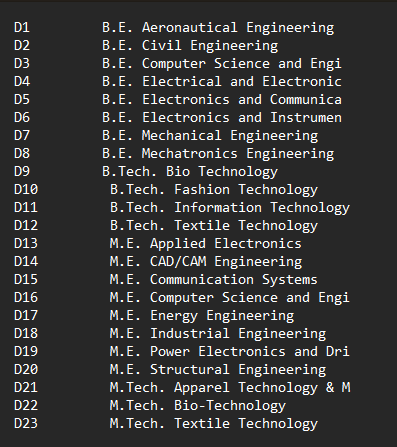
* **Output of Student Dimension:**

****

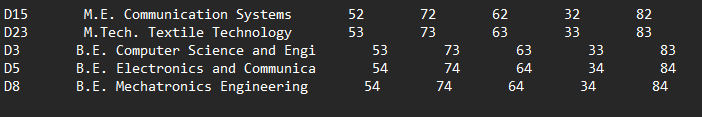
* **Output of Marks Fact:**

****

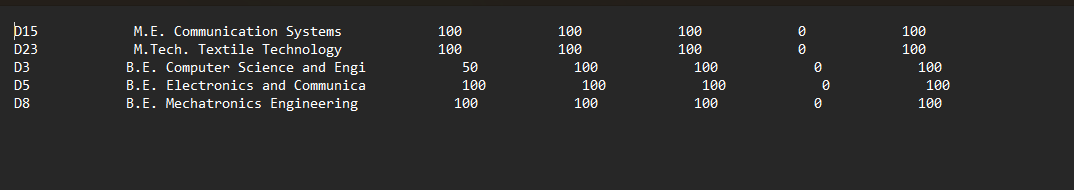
* **Output of Department Dimension:**

****

* **Output of Student Avg Aggregate:**

****

* **Output of Subject Pass Percent Aggregate:**

****

**Conclusion:**

Informatica PowerCenter is a robust solution for data integration activities and the creation of data warehouses for industries. The technology performs data transformation operations quickly and delivers data to industries so that a data warehouse may be built and utilized in projects.