# Data Warehousing

EDUCATION SECTOR

Team- Ammen Sanjay Siingh, Latefa Mahmoud, Vaibhav Nigam
SYRACUSE UNIVERSITY | GROUP 6

Company: Education
Name: Syracuse University - School of information Studies
Location: 343 Hinds Hall, Syracuse, NY 13210

## Background and Introduction

In simple terms, a data warehouse (DW) is a pool of data produced to support decision making; it is also a repository of current and historical data of potential interest to managers throughout the organization. Data are usually structured to be available in a form ready for analytical processing activities (e.g. online analytical processing [OLAP], data mining, querying, reporting and other decision supporting applications). A data warehouse is a subject-oriented, integrated, time -variant, non-volatile collection of data in support of management's decision-making process. The day-to-day operations of an organization are done by using the OLTP system. The lack of historical data in OLTP makes it unsuitable to provide a comprehensive information about the operations of the business. DW on the other hand, provides a central repository of historical data which provides an integrated platform for historical analysis of data. With a data warehouse and Online Analytical Processing (OLAP), users can perform better data analysis.

In today's ambitious environment, educational institutions have been privatized and are competitive throughout all top universities. Currently, one of the top schools, the School of Information Studies at Syracuse University has an unorganized data system that is difficult to manage and maintain properly. In order for the institution to maintain their reputation and rankings the institution needs to be more organized and needs to take better decisions while handling heavy data and reports. By creating a data warehouse educational institutions could overcome their problems. Not only will it provide a centralized and organized database but also offer good quality management solutions. It is important to have top management in the educational institutes because it often need timely analysis reports of students, faculty, course records, and so forth. In addition, they need to provide timely analysis reports to assist in making long-term decisions. It has been shown that most of the reporting and analysis, time was spent on collecting data from the various systems before the analysis can be made. Therefore, the purpose of creating the data warehouse is to provide good quality, convenient and accessible data in one centralized system to the educational institution providing top level management.

## Proposed Solution / Justification

The case study used as a model for the project is a student course management system in Syracuse University. The school registers students every semester and students take courses and exams. These are being managed by an online transaction processing (OLTP) system. A simplified representation of the logical design of the OLTP system is shown in Figure
After analyzing the current situation of the educational institution our team proposed to create a Data Warehouse to help the institution with making their database more organized, easily retrieving different types of data, user friendly interface, and providing reports efficiently and effectively. The solution we hope to implement is a centralized Data warehousing technology to help collect large amount of data from many kinds of databases and unify them under a star schema in order to be used by Online Analytical Processing (OLAP) to help users utilize the

system. We would be retrieving raw data from flat files (txt or csv), SQL tables, and excel files and cleaning them before loading them into our star schema. We could then use the star schema to answer different questions that the institutions may have.

## Scope

To create a data warehouse for the educational system. It will provide a functioning data warehouse that contains all the entries for the students, courses and the employees. The features of a data warehouse are to easily access student as it will be a large dataset and improving the performance and provide reports.

The scope of our project includes extracting, transforming, and loading all the OLTP data into our data warehouse from different sources like flat files and other SQL databases. We will use SSIS ETL tool for this purpose and we will also provide visualizations and a plan to obtain intelligence from the data warehouse via a BI plan.

In Scope

- Create a functional star schema data warehouse that contains students, faculty, and course records.
- We will do data profiling and use ETL tools to load the data into our data warehouse
- We will also give out visualizations and a plan for obtaining intelligence from our system.

Out of Scope

- Developing a system interface.

Sponsor/Champion

- Inhouse
  - Dean
  - Head of departments
  - Faculty/staff
  - Board members

## Business Reasons

| Direct benefits: | Indirect benefits: |
|---|---|
| • Simplification of data access<br>• Enhanced system performance | • Enhance business knowledge |

| | |
|---|---|
| • Allows end users to perform extensive analysis | • Present competitive advantage<br>• Facilitate decision making |

## Interview and answers

1. **What is the business objective of the project? What are they trying to achieve?**

The purpose of this project is to warehouse the data of the school's course management system so that the school can use the data to obtain some intelligence from it at a later stage.

2. **What are the different sources of data for the warehouse?**

There are numerous sources of the data like csv files and also a sql database table. This is further explained in this document.

3. **What is our data loading frequency (daily/weekly/monthly)?**

The data would be loaded into the data warehouse on a monthly basis as there is no greater need for a higher frequency since the data would not change much during the course of the semester.

4. **Who will be the users of the system?**

The system would be used by the school administration for long term data analysis for decision making purposes.

**Some other questions that we thought were pertinent and were asked during the interviews that were conducted:**

1. What are input sources? Where are they going to get there data from?
2. What is the technology and version of input sources?
3. What is our data loading frequency (daily/weekly/monthly)?
4. How will the application be rolled out?
5. Who will support the application?
6. What all documents are required by Support team?
7. What is the distribution mechanism of re ports
8. What are the security requirements?

## Management Team

| Roles: | Responsibilities: | Skills: |
|---|---|---|
| Project manager | Tracks the progress and acts as a mediator among the team members | • Communication<br>• Leadership<br>• Team management<br>• Risk management<br>• Negotiation |
| DBA | Creates the logical design of the database | • Database Design<br>• Metadata management<br>• ERP and business knowledge<br>• Backup and recovery<br>• Performance management and tuning |
| Technical Architect | Relays the technical architecture of the company to the DBA | • data Modeling<br>• Understanding Framework<br>• Design tools<br>• Knowledge of UML<br>• Analytic problem solving |
| ETL Developer | Researches about the different sources of data for the target database | • SQL<br>• Debugging<br>• ETL tools/software<br>• Scripting Language |
| Front End Developer | Uses .Net or Access to connect with SQL server and have a GUI for users to query the database | • Programming<br>• SQL<br>• Command Line<br>• GUI |
| OLAP Developer | Analyzes the data to see what insights can be gained from it via reports | • Oracle<br>• Siebel Analytics<br>• SSRS<br>• Report Builder<br>• Crystal Reports |
| Trainer | Providing resources and training sessions for users to make them comfortable with the new system | • Communication<br>• SQL data<br>• Reporting<br>• Analytics |

| | | |
|---|---|---|
| Data Modeler | Work with ETL developer | • Expertise in data modeling principles/methods including conceptual, logical & physical Data Models |
| QA Group | Maintenance of quality throughout the development lifecycle | • Problem Solving skills<br>• Identify Areas of Improvement<br>• Ability to Code Automated Tests |

## Issues List

| Issue No. | Description | Priority | Reported By | Assigned To | Status |
|---|---|---|---|---|---|
| 1 | Cost of construction of a database. | High | Ammen | Ammen | Close |
| 2 | Data Quality-Unstructured, vague, and undefined source data | Medium | Latefa | Latefa | Open |
| 3 | Change Management-Training the concerned people | Medium | Vaibhav | Vaibhav | Close |
| 4 | Lack of technical skills | High | Ammen | Ammen | Open |
| 5 | Tme constraints | Medium | Latefa | Latefa | Open |
| 6 | Fuzzy testing scenario | Medium | Vaibhav | Vaibhav | Close |
| 7 | Ensuring acceptable performance | High | Ammen | Ammen | Close |

## Data Dictionary

| **Students** | **Description** | **Data Type** |
|---|---|---|
| Student ID | The ID of the student | INT |
| Last Name | Last Name of the student | VARCHAR |
| First Name | First name of the student | VARCHAR |
| Student Permanent Address | Permanent Address of the student | VARCHAR |
| Student City | City of the student | VARCHAR |
| Student Date of birth | Date of birth of student | DATE |
| Student Gender | Gender of Student | CHAR |
| Student Address | Address of the student | VARCHAR |
| Student Email | Emails of the student | VARCHAR |

| **Staff/Faculty** | **Description** | **Data Type** |
|---|---|---|
| First Name | First Name of the staff/fac | VARCHAR |
| Last Name | Last Name of the staff/fac | VARCHAR |
| Address | Address of the staff/fac | VARCHAR |
| Gender | Gender of the staff/fac | |
| Email | Email of the staff/fac | VARCHAR |
| Address | Address of the staff/fac | VARCHAR |
| ID | ID of the staff/fac | INT |
| Titles | titles of the staff/fac | VARCHAR |
| **Courses** | **Description** | **Data Type** |
| Course ID | Course ID | INT |
| Credits | Number of Credits | INT |
| Staff ID | Staff number ID | INT |
| School | The Name of the School | VARCHAR |
| Course Name | The Name of the course | VARCHAR |
| Total students | Total number of students | INT |
| Prerequisite | Perquisite course | INT |

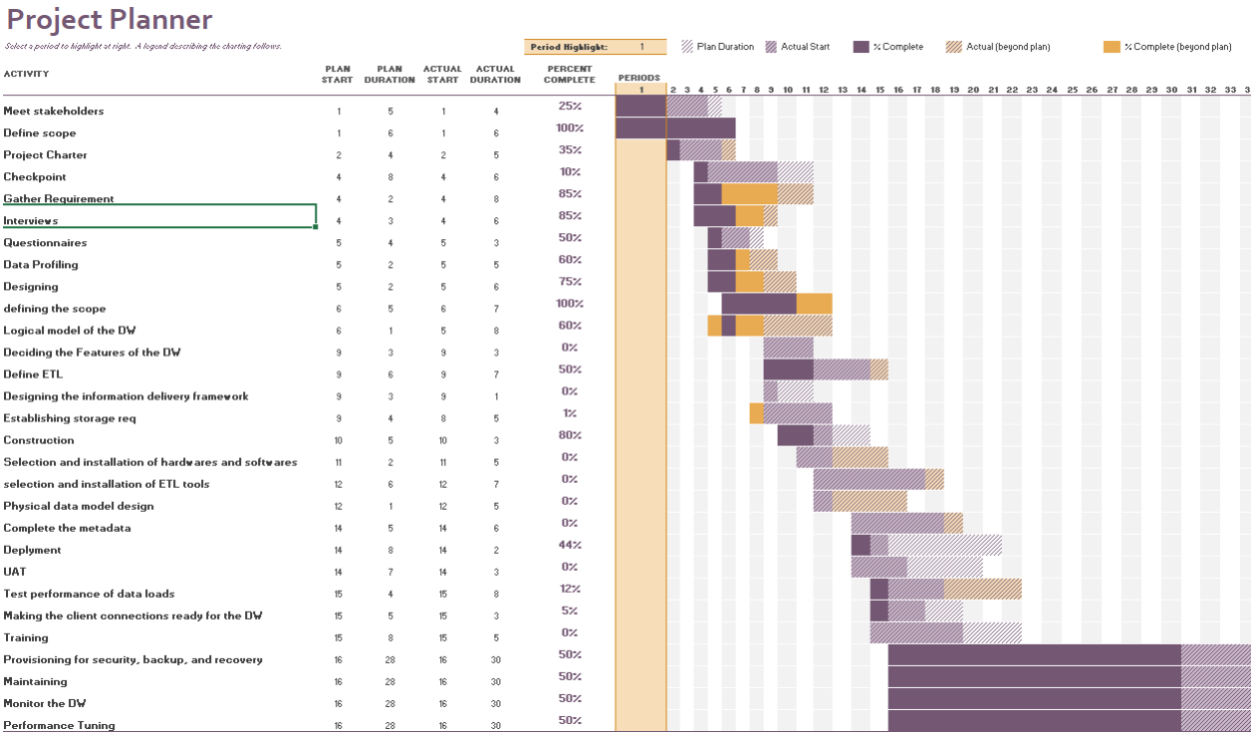
## Data Marts

We will have 3 data marts:

1. Students
2. Faculty
3. Courses

Each data mart would include all of the attributes of their concerned tables and the user would be able to choose from any of the attributes to find the rest of them using a key.

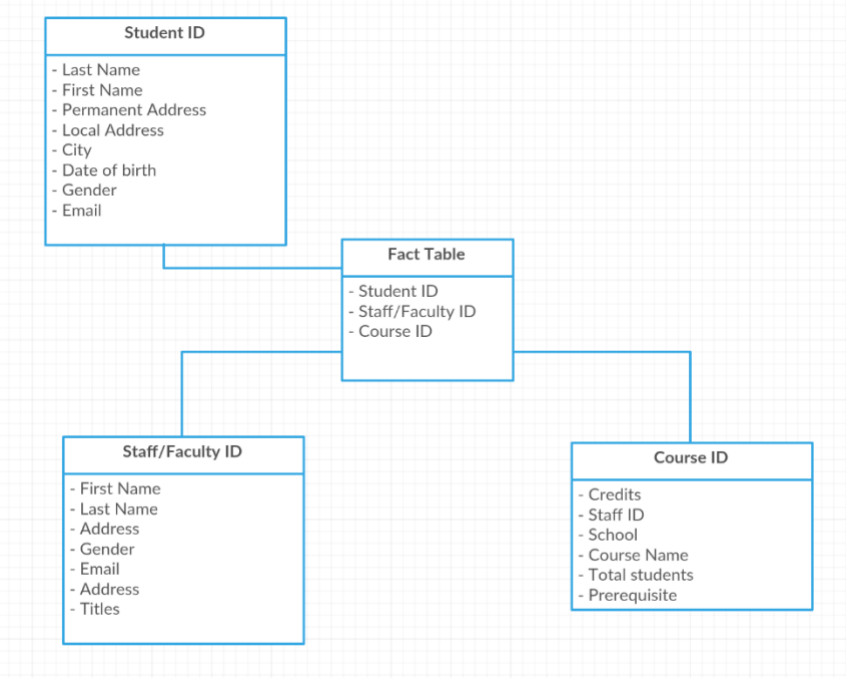
## Work Breakdown Structure



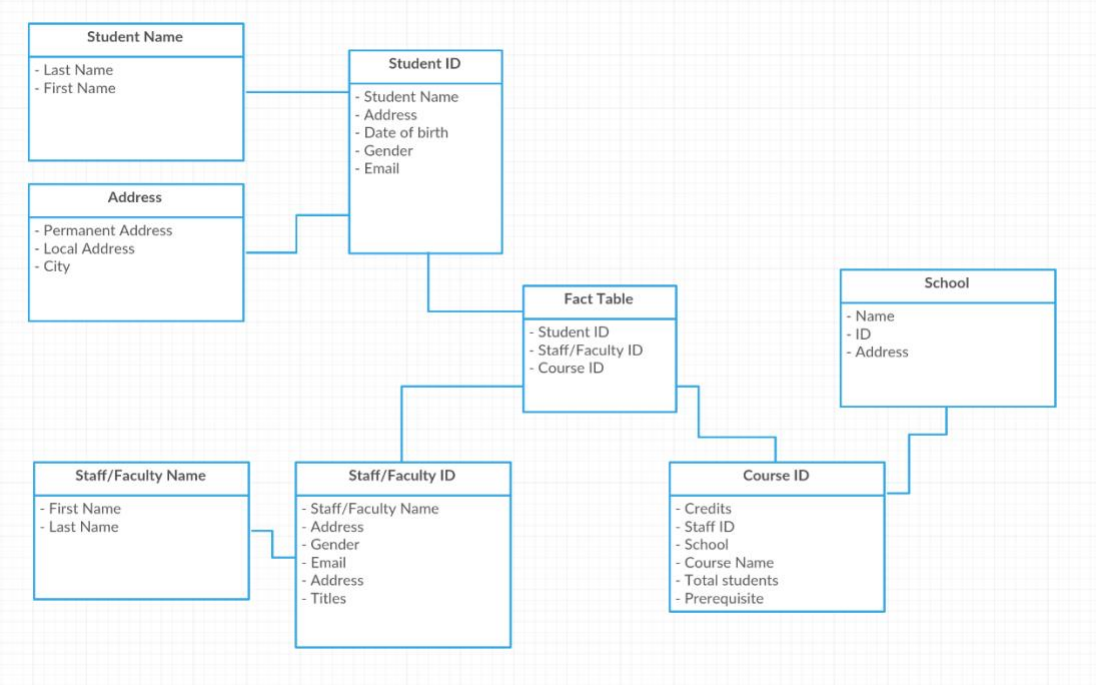

## Estimated Cost and ROI

| Roles | Number | Amount | Duration | Cost |
|---|---|---|---|---|
| Data Modeler | 1 | $ 90,000.00 | 0.1 | $ 9,000.00 |
| DBA | 1 | $ 80,000.00 | 0.5 | $ 40,000.00 |
| ETL Developer | 1 | $ 90,000.00 | 0.9 | $ 81,000.00 |
| Front End Developer | 1 | $ 80,000.00 | 0.5 | $ 40,000.00 |
| OLAP Developer | 1 | $ 70,000.00 | 0.5 | $ 35,000.00 |
| Project manager | 1 | $ 100,000.00 | 1 | $ 100,000.00 |
| QA Group | 1 | $ 90,000.00 | 0.5 | $ 45,000.00 |
| Technical Architect | 1 | $ 110,000.00 | 0.8 | $ 88,000.00 |
| Trainer | 1 | $ 80,000.00 | 0.1 | $ 8,000.00 |

## Star Schema




## Snowflake Schema



## SQL for Create

SQLQuery4.sql - ist...w (AD\asiingh (59)) SQLQuery3.sql - ist...w (AD\asiingh (58))

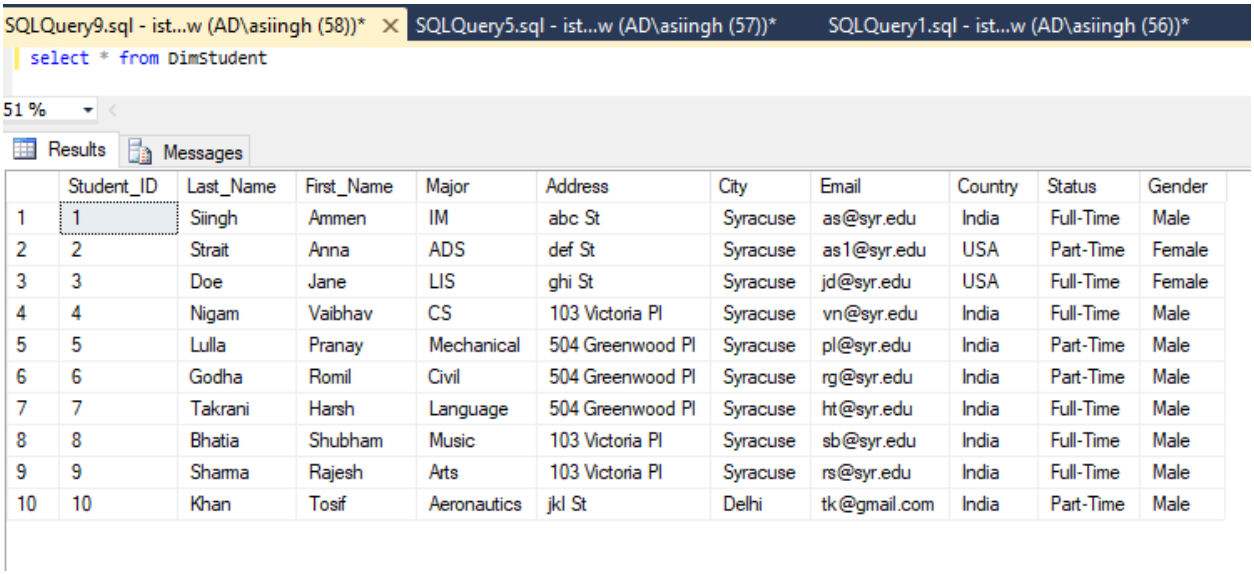
```
USE [ist722_asiingh_dw]


CREATE TABLE [dbo].[DimCourse](
    [Course ID] [int] NULL,
    [Course Name] [varchar](500) NULL,
    [Credit] [int] NULL,
    [Semester] [varchar](50) NULL,
    [DayofWeek] [varchar](100) NULL,
    [Time] [varchar](200) NULL,
    [Classroom No ] [varchar](50) NULL,
    [Factulty_ID] [int] NULL
) ON [PRIMARY]

CREATE TABLE [dbo].[DimFaculty](
    [Faculty_ID] [int] NULL,
    [First Name] [varchar](100) NULL,
    [Last Name] [varchar](100) NULL,
    [Address] [varchar](500) NULL,
    [Gender] [varchar](50) NULL,
    [Email] [varchar](500) NULL,
    [Citizenship] [varchar](200) NULL,
    [Experience Years] [int] NULL
) ON [PRIMARY]

CREATE TABLE [dbo].[DimStudent](
    [Student_ID] [int] NULL,
    [Last_Name] [varchar](100) NULL,
    [First_Name] [varchar](100) NULL,
    [Major] [varchar](500) NULL,
    [Address] [varchar](500) NULL,
    [City] [varchar](100) NULL,
    [Email] [varchar](100) NULL,
    [Country] [varchar](200) NULL,
    [Status] [nvarchar](9) NULL,
    [Gender] [nvarchar](6) NULL
) ON [PRIMARY]

CREATE TABLE [dbo].[Fact](
    [FactID] [int] NULL,
    [StudentID] [int] NULL,
    [CourseID] [int] NULL,
    [FacultyID] [int] NULL,
    [Marks] [int] NULL,
    [Grade] [varchar](10) NULL
) ON [PRIMARY]

GO
```

## Example SQL for Inserting

SQLQuery5.sql - ist...w (AD\asiingh (57))* SQLQuery1.sql - ist...w (AD\asiingh (56))*

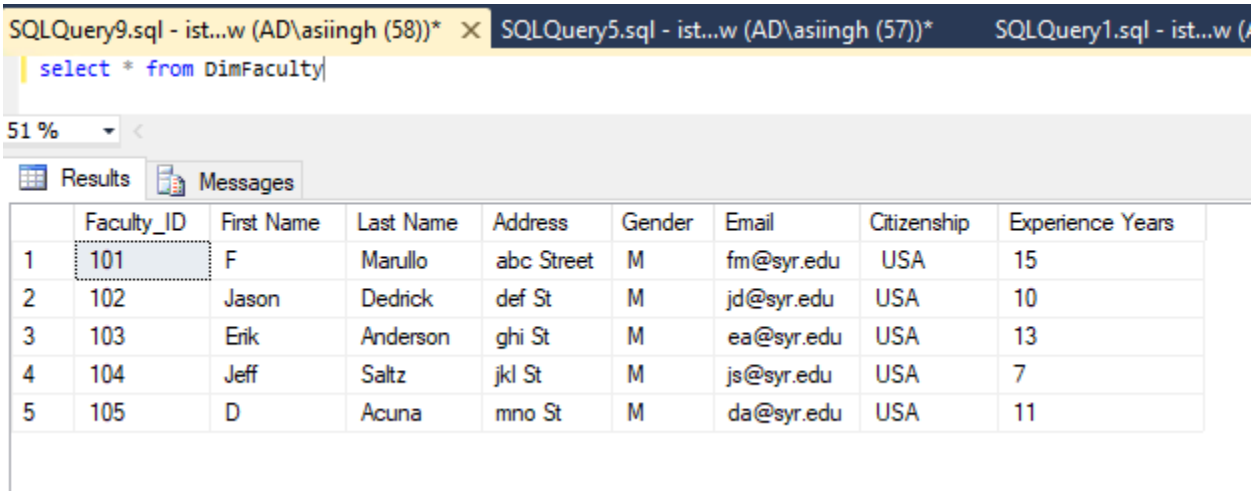
```
USE [ist722_asiingh_dw]


INSERT INTO [dbo].[DimCourse]
           ([Course ID]
           ,[Course Name]
           ,[Credit]
           ,[Semester]
           ,[DayofWeek]
           ,[Time]
           ,[Classroom No ]
           ,[Factulty_ID])
     VALUES
           (1,'InfoSec',3,'Fall','Monday','9AM-3PM',101,1)

INSERT INTO [dbo].[DimFaculty]
           ([Faculty_ID]
           ,[First Name]
           ,[Last Name]
           ,[Address]
           ,[Gender]
           ,[Email]
           ,[Citizenship]
           ,[Experience Years])
     VALUES
           (1, 'Frank','Marullo', 'abc Street',1,'fm@syr.edu''USA',5)


INSERT INTO [dbo].[DimStudent]
           ([Student_ID]
           ,[Last_Name]
           ,[First_Name]
           ,[Major]
           ,[Address]
           ,[City]
           ,[Email]
           ,[Country]
           ,[Status]
           ,[Gender])
     VALUES
           (1,'Siingh','Ammen','Information Mgmt','1048 Lancaster Ave','Syracuse','asiingh@syr.edu','India',1,1)
```

## SQL Dimension Tables- DimStudent



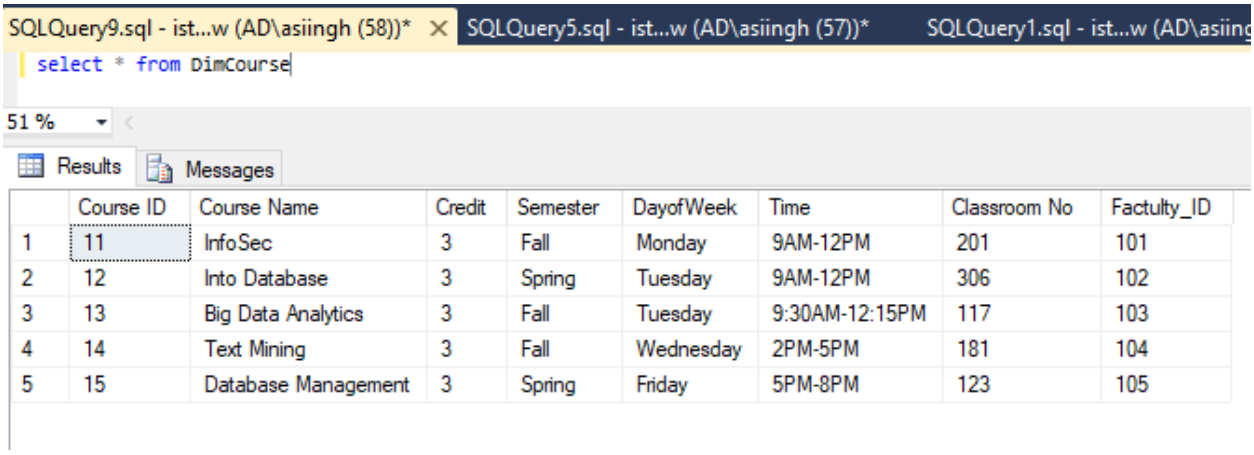

| | Student_ID | Last_Name | First_Name | Major | Address | City | Email | Country | Status | Gender |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Siingh | Ammen | IM | abc St | Syracuse | as@syr.edu | India | Full-Time | Male |
| 2 | 2 | Strait | Anna | ADS | def St | Syracuse | as1@syr.edu | USA | Part-Time | Female |
| 3 | 3 | Doe | Jane | LIS | ghi St | Syracuse | jd@syr.edu | USA | Full-Time | Female |
| 4 | 4 | Nigam | Vaibhav | CS | 103 Victoria Pl | Syracuse | vn@syr.edu | India | Full-Time | Male |
| 5 | 5 | Lulla | Pranay | Mechanical | 504 Greenwood Pl | Syracuse | pl@syr.edu | India | Part-Time | Male |
| 6 | 6 | Godha | Romil | Civil | 504 Greenwood Pl | Syracuse | rg@syr.edu | India | Part-Time | Male |
| 7 | 7 | Takrani | Harsh | Language | 504 Greenwood Pl | Syracuse | ht@syr.edu | India | Full-Time | Male |
| 8 | 8 | Bhatia | Shubham | Music | 103 Victoria Pl | Syracuse | sb@syr.edu | India | Full-Time | Male |
| 9 | 9 | Sharma | Rajesh | Arts | 103 Victoria Pl | Syracuse | rs@syr.edu | India | Full-Time | Male |
| 10 | 10 | Khan | Tosif | Aeronautics | jkl St | Delhi | tk@gmail.com | India | Part-Time | Male |

## SQL Dimension Tables- DimFaculty




| | Faculty_ID | First Name | Last Name | Address | Gender | Email | Citizenship | Experience Years |
|---|---|---|---|---|---|---|---|---|
| 1 | 101 | F | Marullo | abc Street | M | fm@syr.edu | USA | 15 |
| 2 | 102 | Jason | Dedrick | def St | M | jd@syr.edu | USA | 10 |
| 3 | 103 | Erik | Anderson | ghi St | M | ea@syr.edu | USA | 13 |
| 4 | 104 | Jeff | Saltz | jkl St | M | js@syr.edu | USA | 7 |
| 5 | 105 | D | Acuna | mno St | M | da@syr.edu | USA | 11 |

## SQL Dimension Tables- DimCourse




| | Course ID | Course Name | Credit | Semester | DayofWeek | Time | Classroom No | Factulty_ID |
|---|---|---|---|---|---|---|---|---|
| 1 | 11 | InfoSec | 3 | Fall | Monday | 9AM-12PM | 201 | 101 |
| 2 | 12 | Into Database | 3 | Spring | Tuesday | 9AM-12PM | 306 | 102 |
| 3 | 13 | Big Data Analytics | 3 | Fall | Tuesday | 9:30AM-12:15PM | 117 | 103 |
| 4 | 14 | Text Mining | 3 | Fall | Wednesday | 2PM-5PM | 181 | 104 |
| 5 | 15 | Database Management | 3 | Spring | Friday | 5PM-8PM | 123 | 105 |

## SQL Fact Table

SQLQuery9.sql - ist...w (AD\asiingh (58))* | SQLQuery5.sql - ist...w (AD\as

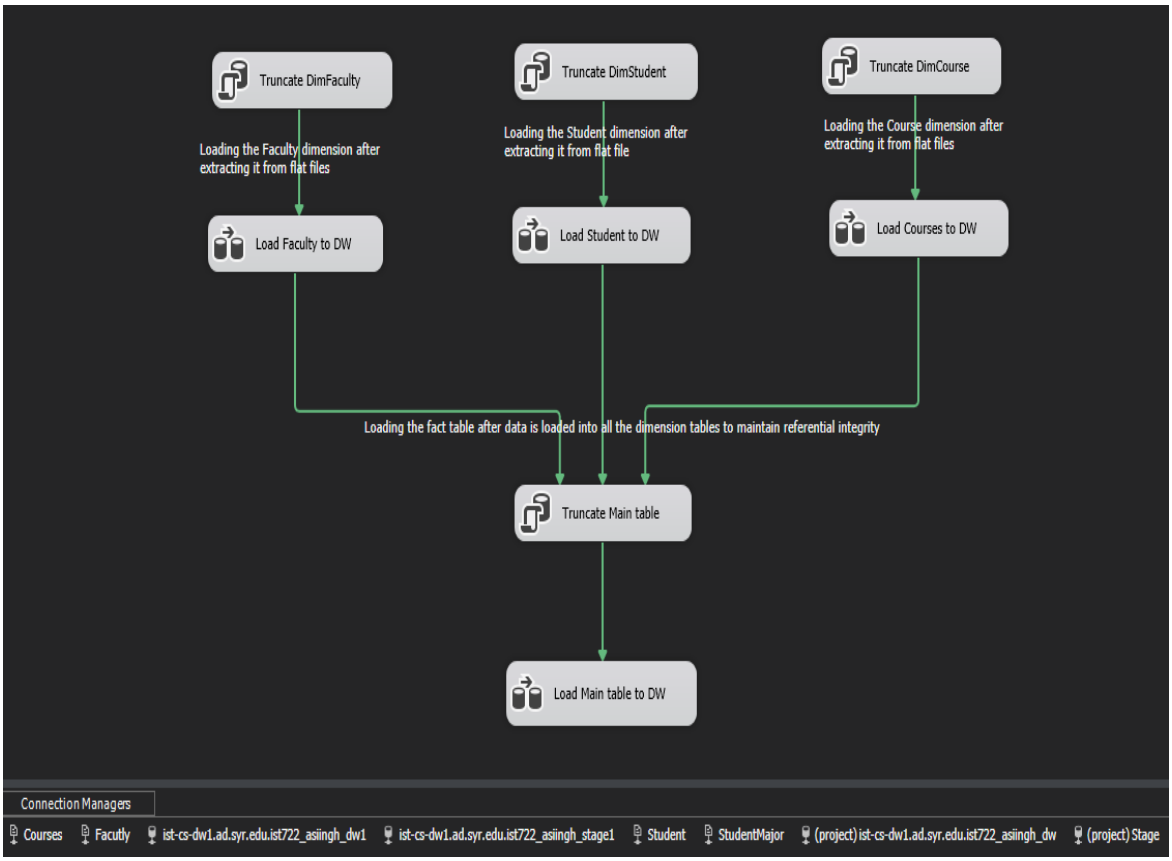
```
select * from Fact
```

51 %

Results | Messages

| | FactID | StudentID | CourseID | FacultyID | Marks | Grade |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 11 | 101 | 99 | A |
| 2 | 2 | 1 | 12 | 102 | 99 | A |
| 3 | 3 | 2 | 11 | 103 | 89 | B |
| 4 | 4 | 3 | 13 | 104 | 69 | C |
| 5 | 5 | 4 | 14 | 103 | 90 | A- |

## Implementation plans (SSIS)

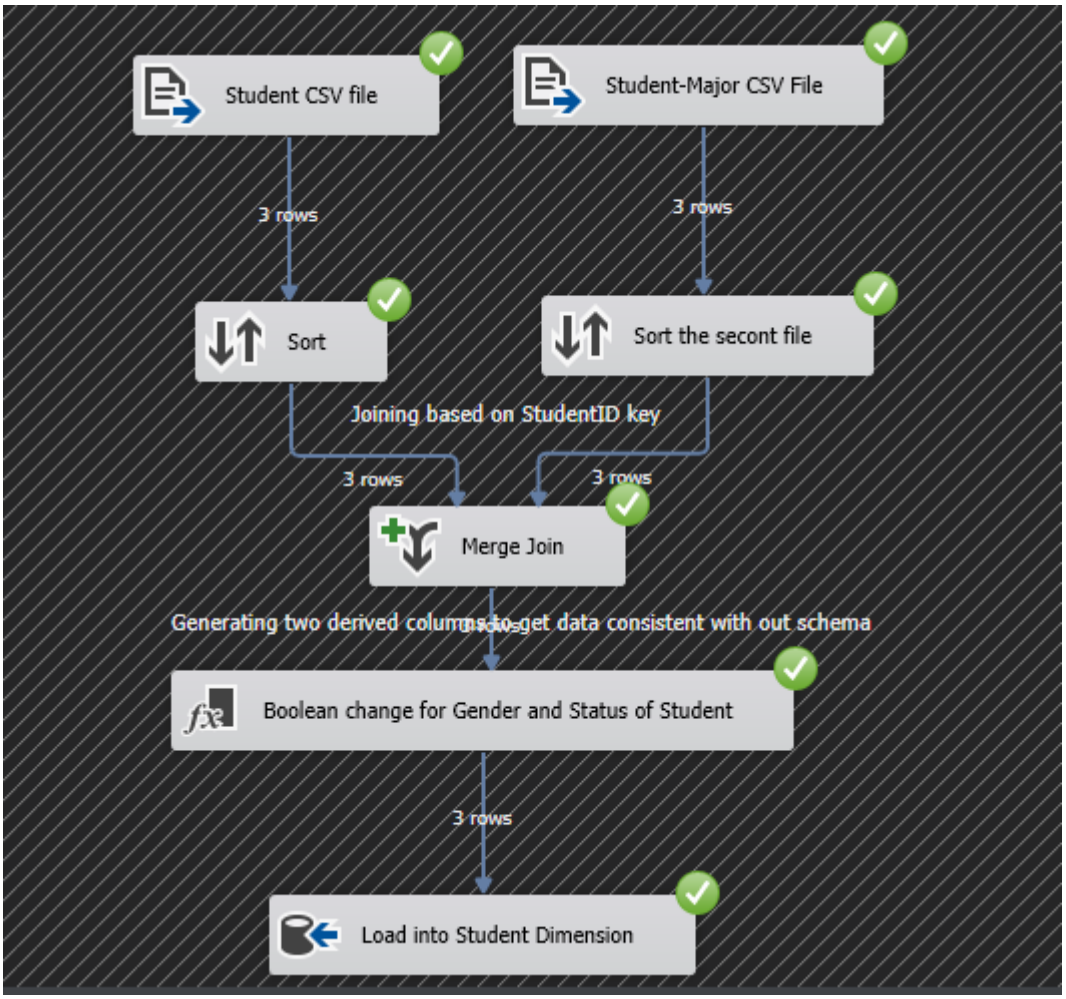


## Source Target Analysis and ETL

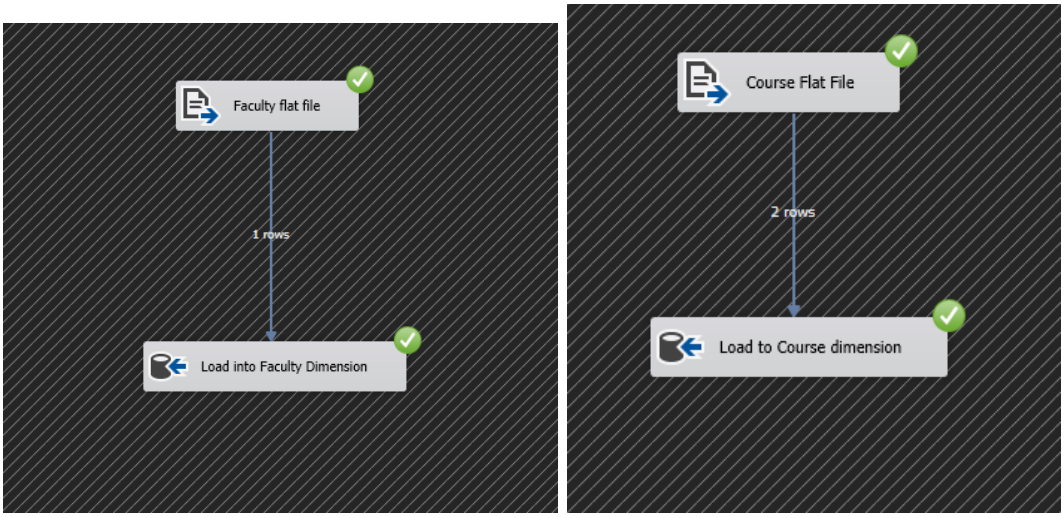
We have created three CSV files namely “Student”, “Faculty” and “Course” which we processed using SSIS tool to get the main table. Then we loaded the truncated dimensional files which we extracted from the respective flat files. Then we integrated all the fact tables which had all the data loaded into all the dimensional tables to create a truncated main table and to maintain referential integrity, after which we loaded the main table to the data warehouse.

The SSIS images for the three files are shown below in the three images. It shows the way we downloaded the different files onto the data warehouse from different sources like csv files and sql database.

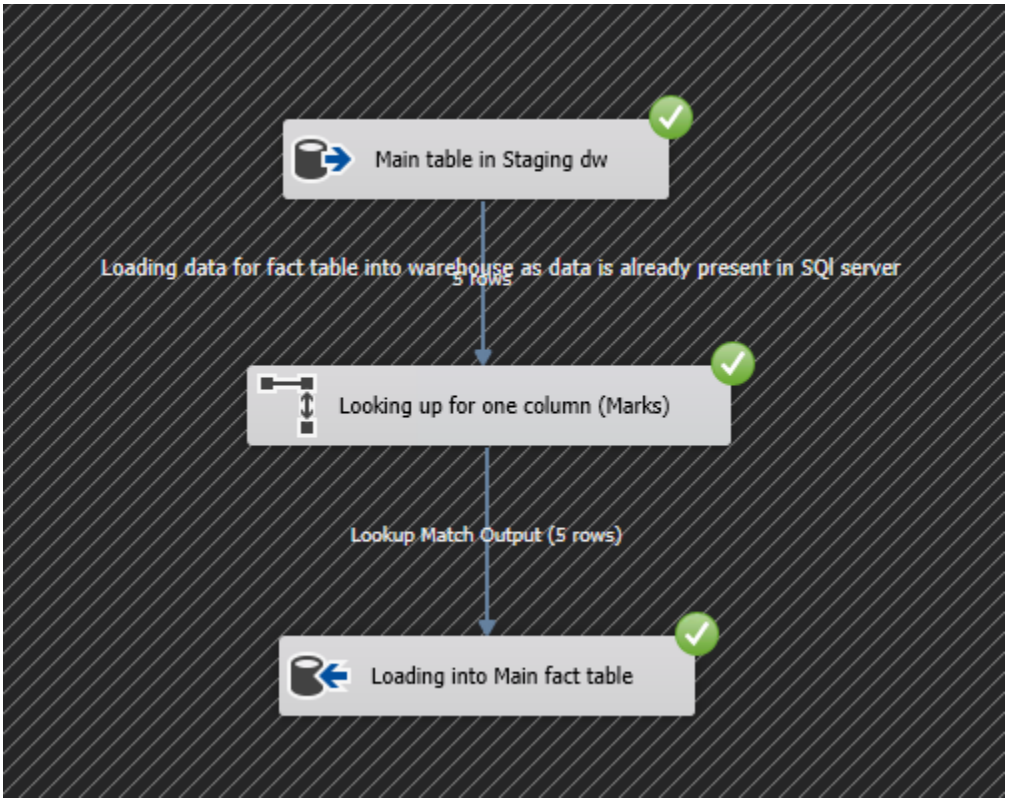
### ETL for DimStudent dimension



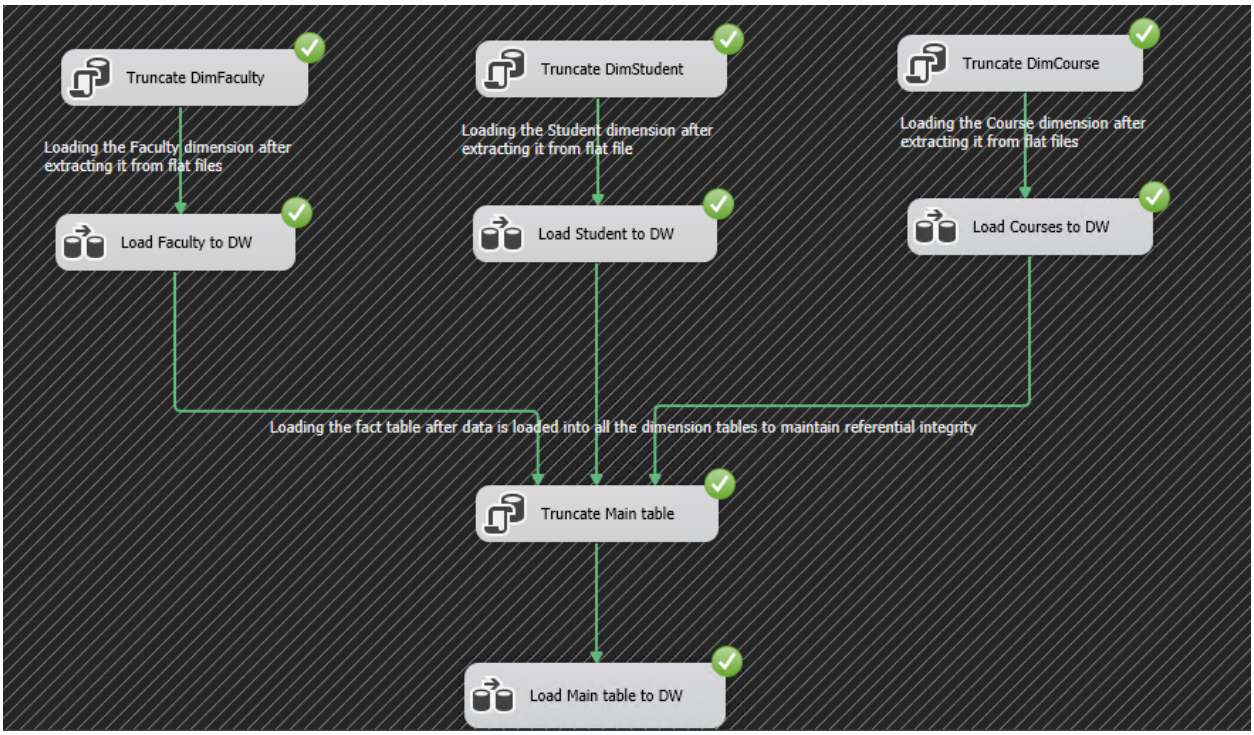
## ETL for DimFaculty and DimCourse




## ETL for Fact Table



### Whole ETL System execution




## Work Distribution in the team

We were a team of three members and we divided the work accordingly. Latefa was responsible for combining the report and bringing it all together. Ammen was responsible for implementation of the ETL tool and Vaibhav was responsible for data profiling and other documents like the issues log and Work breakdown structure. This distribution of work ensured that we were always abreast with what was needed to be done and therefore managed to finish the project in time.

## Maintenance and support of the Data Warehouse

- Monitoring the realization of expected benefits
- Providing ongoing support to users (see deployment)
- Training new staff
- Assisting with the identification and cleansing of dirty data
- Maintaining both feeds & meta-data as source systems change over time;
- Tuning the warehouse for maximum performance (this includes managing indexes and aggregates according to actual usage)
- Purging dormant data
- Recording successes and using these to continuously market the warehouse