Major Project Documentation

(Anamitra Maji(MT19112), Pragya Dara(MT19126), Swati Verma(MT19073))

1. **Statement of the decision problem**

   Which educational institute will you select to do your B.Tech ?

2. **Representation as DIEM schema**

   All the screenshots of DIEM schema are present in the minor project report.

I. Building DIEM Schema

a. **Decision:** Which institute to choose for B.Tech?

b. **Uncertainties:**

- Fees
- Extracurricular Activities
- Quality of Education
- Reputation of the college
- Placements

c. **Action:**

- Select Institute

d. **Objectives:**

- Minimize Expenditure
- Maximize Education Satisfaction
- Maximize Placement Satisfaction
- Maximize Career Progression

e. **Object:**

- **Action Object:**

  Select_Institute(College, Total fees, Placements, Academics, Extracurricular activities, Reputation**)**

- **Uncertainty Object:**

1. **Fees(** College**,** , Year, Hostel Fees, Mess Fees, Tuition Fees, Security Fees, others**)**
2. **Extracurricular_Activities (**College, year, Number of Clubs, Number of Committees, Fests, Sports Complex, Auditorium**)**
3. **Academics (**College, Year, Branch, Number of Seats, Cutoff, Faculty **)**
4. **Placements (**College, Year, Student, Placement**)**

- **Objective Object:**
    1. **ENDS Objective:** Career Progression of the student.
    2. **MEANS Objective:** Minimize Expenditure, Maximize Education Satisfaction, Maximize Placement Satisfaction.
    3. **Critical Success Factor:** Good Placements
- **Environmental Factors**:

  Reputation (College,Year, Rating**)**

II. The BI elicitation process

1. **Select Institute (**College, Fees, Placements, Academics, Extracurricular Activities, Reputation**)**

   There is only one action **Select Institute** in our project and this would be affected by changes in Fees, Placements, Academics and Extracurricular activities of the college.
2. **Fees (**College**,** Year, Total Fees, Hostel Fees, Mess Fees, Tuition Fees, Security Fees, others**)**

   Total fees is a derived attribute and depends on Hostel Fees, Mess Fees, Tuition Fees, Security Fees, others
3. **Placements (** College, Year, Student, Placement**)**
4. **Placement Stats** (College, Year, Average Package, Highest Package)

   Average Package and Highest Package would be derived from Placement data.
5. **Academics (**College, Year, Branch, Number of Seats, Cutoff, Faculty **)**
6. **Extracurricular Activities (**College, year, Number of Clubs, Number of Committees, Fests, Sports Complex, Auditorium**)**

7. **Reputation (**College,Year, Rating**)**

III. The choice elicitation process

The attributes for computing the various Objective objects are as follows:

1. **Career Progression (**Student, Year, College, Overall Satisfaction**)**
2. **Placement (**Student, Year, College, Package, Company, Profile**)**
3. **Quality Education (**Student, Year, College, Satisfaction **)**
4. **Expenditure (**College, Year, Total fees)

3. Conversion to GOM4DW schema

I. Functional dependency:

Defining functional dependencies to arrive at data objects and category objects for the following DIEM Objects

1. **Select_Institute (**College, Fees, Placements, Academics, Extracurricular Activities, Reputation**)**

   College ➜ Fees, Placements, Academics, Extracurricular Activities, Reputation
2. **Fees (**College**,** Year, Hostel Fees, Mess Fees, Tuition Fees, Security Fees, Other**)**

   College, Year ➜ Hostel Fees, Mess Fees, Tuition Fees, Security Fees
3. **Extracurricular_Activities (**College, year, Number of Clubs, Number of Societies, Fests, Sports Complex, Auditorium**)**

   College, year ➜ Number of Clubs, Number of Societies, Fests, Sports Complex, Auditorium
4. **Academics (**College, Year, Branch, Number of Seats, Cutoff, Faculty **)**

   College, year, Branch ➜ Number of Seats, Cutoff, Faculty
5. **Reputation (**College,Year, Rating**)**

   College, Year ➜ Rating
6. **Placements (**College, Year, Student, Company, Profile, Package**)**

   College, Year, Student ➜ Company, Profile, Package

7. **Placement Stats** (College, Year, Average Package, Highest Package)

   College, Year ➜Average Package, Highest Package

8. **Career Progression(**Student, Year, College, Overall Satisfaction**)**

   Student, Year, College ➜ Overall Satisfaction

9. **Quality Education(**Student, Year, College, Satisfaction **)**

   Student, Year, College ➜ Satisfaction

10. **Minimize Expenditure(** College, Year, Total Fees**)**

    College, Year➜ Total Fees

II. Objects, attribute, contain, history

S. No.	Category Objects	Data Objects	Aggregate Objects	Categories over which aggregated	History
1.	**College(** college_id,college_name, college_address,college_state)	**Select_Institute(**Total Fees, Placements, Academics, Reputation, Extracurricular Activities)			
2.	**College(** college_id,college_name, college_address,college_state) **Year**	**Fees**(Hostel Fees, Mess Fees, Tuition Fees, Security Fees, others)			Yearly, for 5 years
3.	**College(** college_id,college_name, college_address,college_state)	**Extracurricula_Activities (** Number of Clubs, Number of Societies, Fests, Sports Complex , Auditorium**)**			
4.	**College(** college_id , college_name, college_address, college_state) **Branch(**branch_id**,**branch_name**)**	**Academics (** Number of Seats, Cutoff, Faculty**)**			

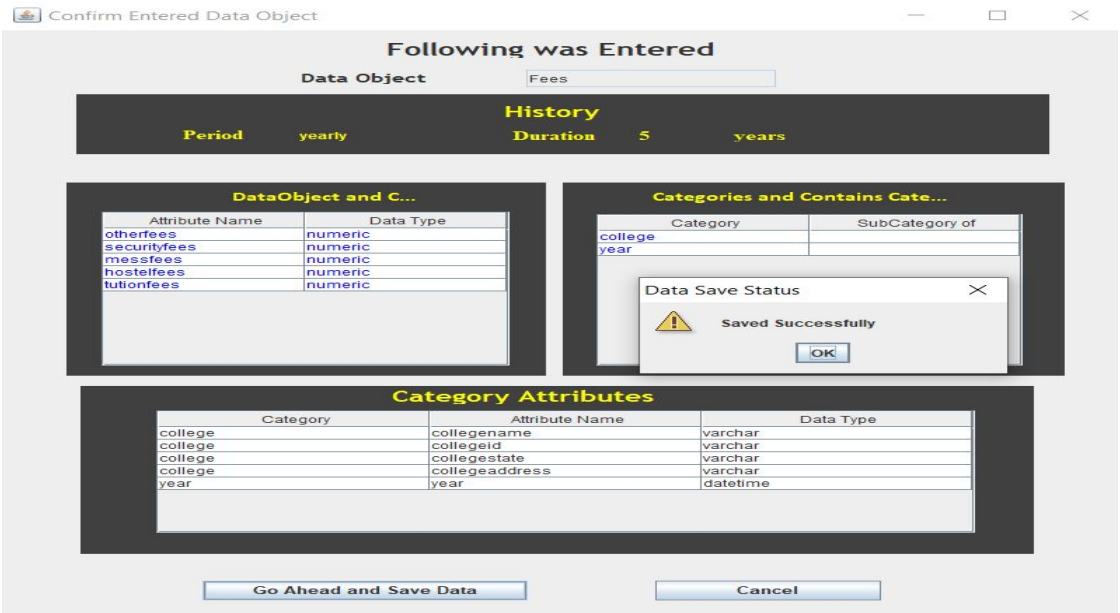
5.	**College(** college_id,college_name, college_address,college_state) **Year**	**Reputation (**Rating**)**			
6.	**College(** college_id,college_name, college_address,college_state) **Year** **Student(** Roll number, student_name, degree, passing_year**)**	**Placement** (Company, Package, Profile)	**Placement Stats**(Average Package, Highest Package)	**College, Year**	Yearly, for 5 years
7.	**Student(** Roll number, student_name, degree, passing_year**)** **College(** college_id,college_name, college_address,college_state) **Year**	**Career Progression(** Overall Satisfaction**)**			
8.	**Student(** Roll number, student_name, degree, passing_year**)** **College(** college_id,college_name, college_address,college_state) **Year**	**Quality of Education(**Satisfaction **)**			
9.	**College(** college_id,college_name, college_address,college_state) **Year**	**Expenditure (**Total fees**)**			

***** The student category contains college and year subcategory

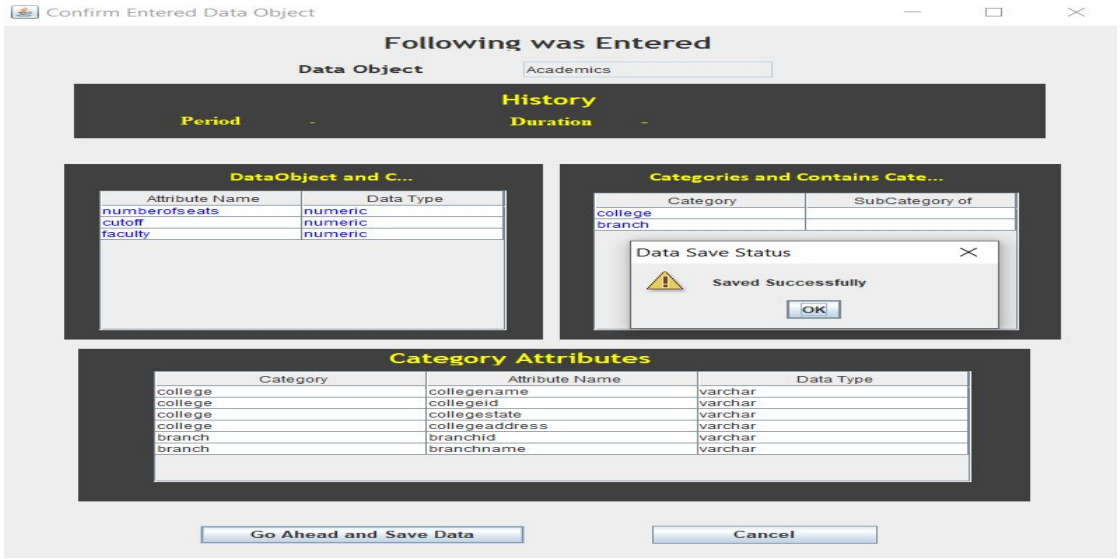
4. **Conversion to Star Schema using Conversion Algorithm**

Some screenshots of the GUI tool where data objects, category objects and their are entered is shown below:

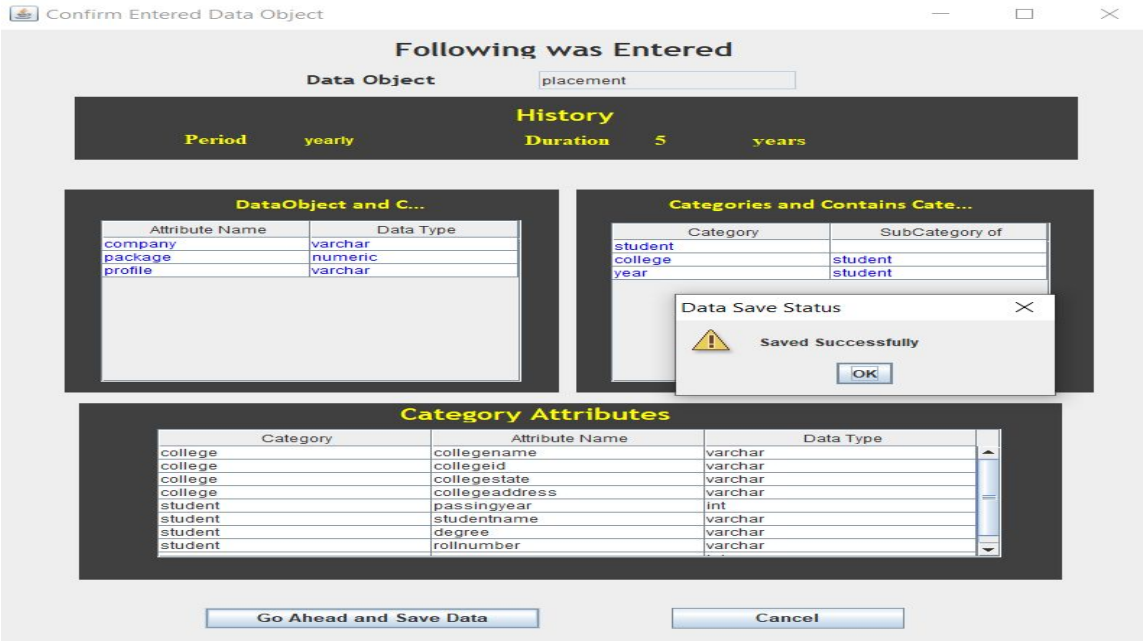
Fees:



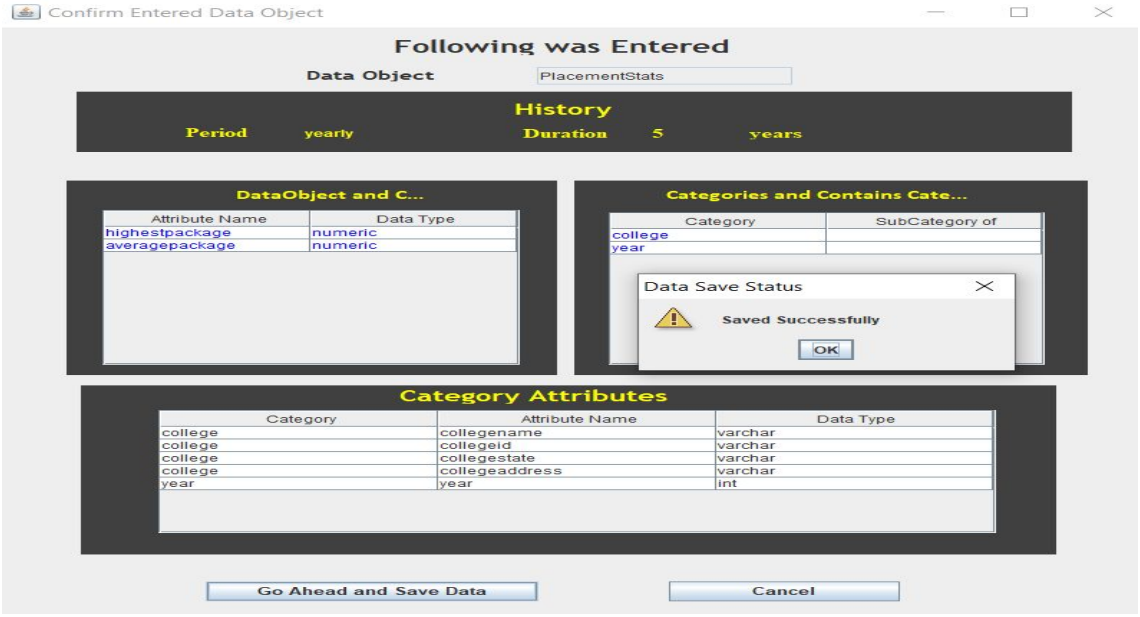
Academics:



Categories under category is shown for **Student dimension** for **Placement fact:**



Placement Stats:



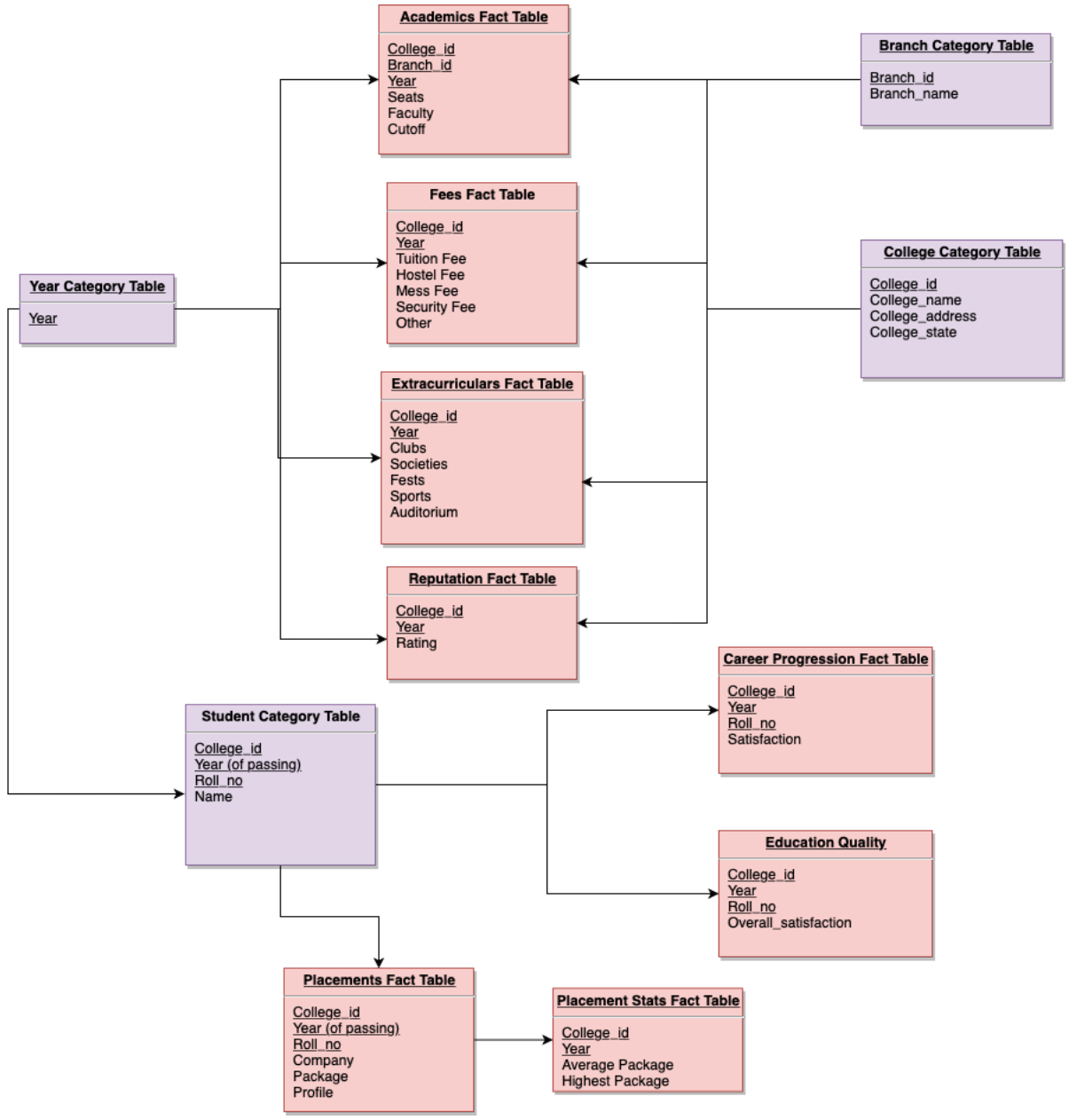
Following output is generated on converting DIEM model to GOM4DW:

```
Fact:
academics (cutoff,faculty,seats)
Dimensions:
branch (branch_id,branch_name)
Dimensions:
college (college_address,college_id,college_name,college_state)
Dimensions:
year (year)
Fact:
careerProgression (satisfaction)
Dimensions:
student (name,roll_no)
Subdimensions:
college (college_address,college_id,college_name,college_state)
Subdimensions:
year (year)
Fact:
educationQuality (overallsatisfaction)
Dimensions:
student (name,roll_no)
Subdimensions:
college (college_address,college_id,college_name,college_state)
Subdimensions:
year (year)
Fact:
extracurricular (auditoriums,clubs,fests,societies,sports)
Dimensions:
college (college_address,college_id,college_name,college_state)
Dimensions:
year (year)
Fact:
Fees (hostelfee,messfee,other,securityfee,tuitionfee)
Dimensions:
college (college_address,college_id,college_name,college_state)
Dimensions:
year (year)
Fact:
placement (company,package,profile)
Dimensions:
student (name,roll_no)
Subdimensions:
college (college_address,college_id,college_name,college_state)
Subdimensions:
year (year)
Fact:
reputation (rating)
Dimensions:
college (college_address,college_id,college_name,college_state)
Dimensions:
year (year)
```

5. Designing Star Schema:

I. **Facts:** Academics , Fees, Extracurricular Activities, Reputation, Career Progression, Placement, Placement Stats, Education Quality.

II. **Dimensions:** College, Branch, Student, Year.



6. The SQL tables Schema for the Star:

The following SQL tables were created using the GOM4DW to Star Schema Algorithm

100% 13:2

Result Grid Filter Rows:

Tables_in_dw2
academics
branch
carrer_progression
college
extracurricular
fees
placement_stats
placements
quality_education
reputation
student
year

a. Fees

Select * from fees;

100% 15:3

Result Grid Filter Rows: Search Edit: Expor

id	college_id	year	tuition_fee	hostel_fee	mess_fee	security_fee	other
1	c_0	2017	85000	30000	20000	10000	4000
5	c_2	2017	45000	30000	12000	8000	6000
6	c_3	2017	25	45000	25000	10000	8000
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

b. Extracurricular

Select * from extracurricular;

100% 24:3

Result Grid Filter Rows: Search Edit: Expor

id	college_id	year	clubs	societis	fests	sports	audi
1	c_3	2017	15	26	3	1	1
2	c_1	2018	34	23	12	5	2
3	c_1	2019	34	23	7	6	3
4	c_2	2017	-1	26	-1	1	1

c. Academics

Select * from academics;

100% 19:3

Result Grid Filter Rows: Search Edit: Expor

id	college_id	year	branch_id	seats	faculty	cutoff
1	c_0	2017	b_1	650	23	12000
2	c_0	2018	b_1	650	20	12000
3	c_1	2017	b_1	800	46	9000
4	c_1	2018	b_1	850	54	0
6	c_1	2019	b_1	850	53	9500
NULL	NULL	NULL	NULL	NULL	NULL	NULL

d. Reputation

Select * from reputation;

100% 19:4

Result Grid Filter Rows: Search

id	college_id	year	rating
1	c_1	2019	4
2	c_0	2019	2
3	c_1	2017	4

e. Placement

Select * from placement;

100% 26:10

Result Grid Filter Rows: Search Edit: Expor

id	roll_number	college_id	passing_year	company_name	profile	package
1	MT19112	c_3	2019	Amazon	Data Scientist	1800000
2	MT19073	c_1	2017	Adobe	SDE	2700000

f. Placement Stats

Select * from placement_stats;

00% 13:18

Result Grid Filter Rows: Search Edit: Expor

d	college_id	year	avg_package	highest_package
1	c_3	2019	1800000	1800000
2	c_1	2017	2700000	2700000

g. Career Progression

Select * from career_progression;

00% 25:17

Result Grid Filter Rows: Search Edit: Expor

id	roll_number	college_id	passing_year	overall_satisfacti...
1	MT19073	c_1	2017	4
2	101411028	c_3	2018	5

h. Quality of Education

Select * from quality_education;

Result Grid Filter Rows: Search Edit: Expor

id	roll_number	college_id	passing_year	satisfaction
1	MT19073	c_1	2017	4
2	101411028	c_3	2018	5

i. College

Select * from college;

Result Grid Filter Rows: Search Edit: Expo

id	college_id	college_name	college_address	college_state	time_stamp
1	c_0	IIITD	Harkesh Nagar Okhla	Delhi	1590403601
3	c_1	Thapar Institute	Patiala	Punjab	1590403766
4	c_2	NIT Raipur	Raipur	Chattisgarh	1590405389
5	c_3	NIT Kurukshetra	Kurukshetra	Haryana	1590405389

j. Branch

Select * from branch;

id	branch_id	branch_name	time_stamp
2	b_3	ElectronicsEngineering	1590406528
3	b_1	Computer Engineering	1590453250

k. Student

Select * from student;

Result Grid Filter Rows: Search Edit: Export,

id	roll_number	college_id	passing_year	student_name	degree	time_stamp
5	MT19073	c_1	2017	Swati	Mtech	1590451895
6	MT190112	c_3	2019	Anamitra	Mtech	1590451911
7	101411028	c_3	2018	Pragya	Btech	1590451928

7. The ETL process including for Type I and II changes

ETL stands for Extraction, Transformation and Load.

Extraction: This step includes extracting all the data which would be inserted in our DW. The data would have different sources like cloud storage, local etc. and may be present in different formats like XML, JSON, CSV, text files. Etc. The goal of this step is to extract all the data from various sources.

Transformation: This step includes preparing the data so that it can be stored. Data from different sources would be of different types. To ensure consistency of DW transformation is important.

The transformations done by us are:

- Data type conversion.
- Handling null values by replacing with default values.
- Removing duplicates.
- Validating the key data for facts.

Loading: This phase involves loading the data into the DW. Since we implemented a ROLAP schema, we used MySQL to implement the DW. To load data into the DW we fired various insert queries. To maintain consistency of keys and type 1, type 2 changes we used various other DML queries also.

Categories have type 1 and type 2 changes. **Type 1** changes indicate change in values which may occur because of typing errors. When a type 1 change occurs, the value for all occurrences of that attribute is updated. In our schema we have the following type 1 change categories:

college - college_address,college_state

branch- branch_name

student - student_name, degree

```
18 •  select * from branch;
19
```

00% 15:18

Result Grid Filter Rows: Search Edit: Export

id	branch_id	branch_name	time_stamp
1	b_1	Computer Science Engineering	1590405944
2	b_3	ElectronicsEngineering	1590406528

b_1 branch_id has the value Computer Science Engineering. When we try to insert a record with b_1 branch_id but with an updated value of branch_name(here Computer Engineering) , the existing value of branch_name for b_1 is updated. Timestamp also changes.

Result Grid Filter Rows: Search

id	branch_id	branch_name	time_stamp
2	b_3	ElectronicsEngineering	1590406528
3	b_1	Computer Engineering	1590453250

Type 2 changes are those attributes where the change in value is because of some update and both the old value and new values are important. In our schema we have the following type 2 change categories:

College - college_name

Result Grid Filter Rows: Search Edit: Expo

id	college_id	college_name	college_address	college_state	time_stamp
1	c_0	IIITD	Harkesh Nagar Okhla	Delhi	1590403601
3	c_1	Thapar Institute	Patiala	Punjab	1590403766
4	c_2	NIT Raipur	Raipur	Chattisgarh	1590405389
5	c_3	NIT Kurukshetra	Kurukshetra	Haryana	1590405389

For college_id c_1 the college_name is Thapar Institute. When we try to insert a record with college_id c_1 and some changed value of college_name(here Thapar University) a new record would be inserted. Timestamp helps in differentiating the old value and new value.

id	college_id	college_name	college_address	college_state	time_stamp
1	c_0	IIITD	Harkesh Nagar Okhla	Delhi	1590403601
3	c_1	Thapar Institute	Patiala	Punjab	1590403766
4	c_2	NIT Raipur	Raipur	Chattisgarh	1590405389
5	c_3	NIT Kurukshetra	Kurukshetra	Haryana	1590405389
6	c_1	Thapar Univeristy	Patiala	Punjab	1590452766