**Final Project**

**Analysis and Design Document**

**Course: ITMD – 526 Data Warehousing**

**Project Developers**

1. **Cristina Quiroga Valcarcel**
2. **Joris Ekpangbo**
3. **Manikandan Ganesh**
4. **Saahil Sofat**
5. **Shushmitha Anthay Suthakakaran**

**Objective:**

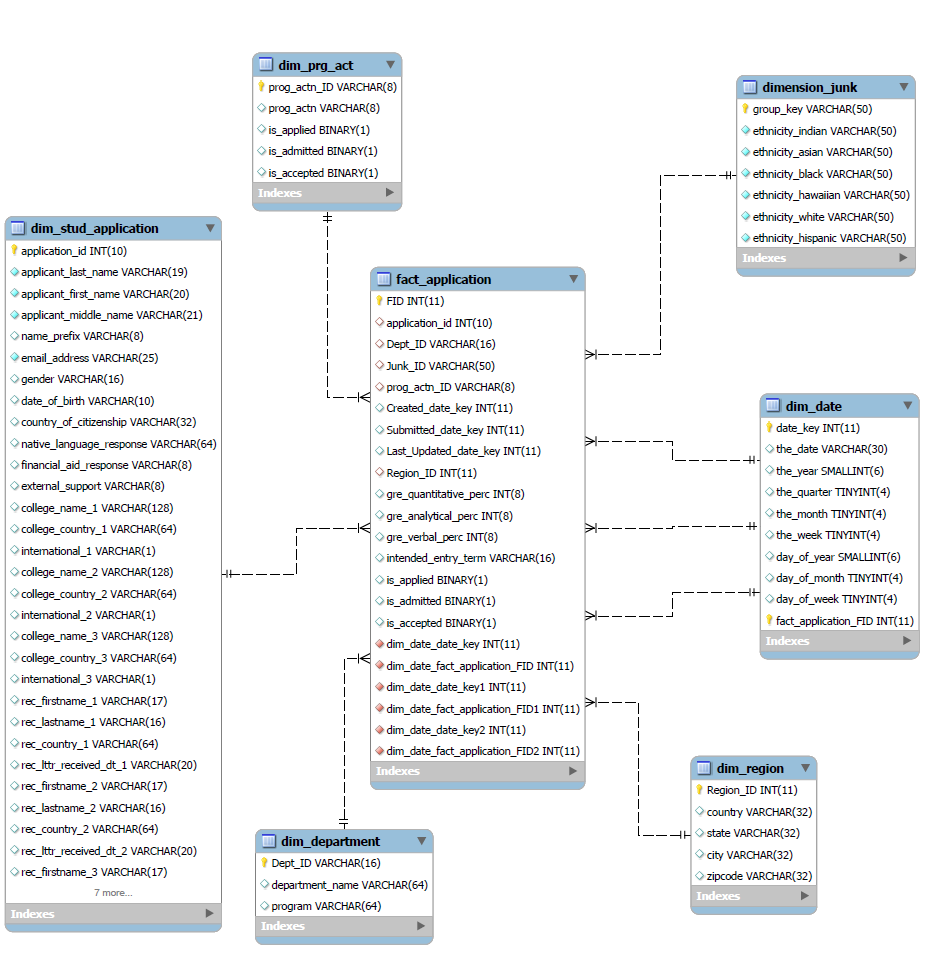
To Analyze and create a design document for the student application info source file.

This document consists of ERD, Tables and DDL scripts.

**Analysis:** We required 6 Dimension tables.

* + Create three views from Dim\_date for Created date view, last updated date view, submitted date view as Role Playing Dimensions.
  + Dim\_junk table is used as a junk dimension with ETHNICITY values.
  + Bridge table can be used (However we will not implement in our design)
* One Fact table.
* 5 reports (Tableau).

**ERD:**



**List of Dimension and Fact:**

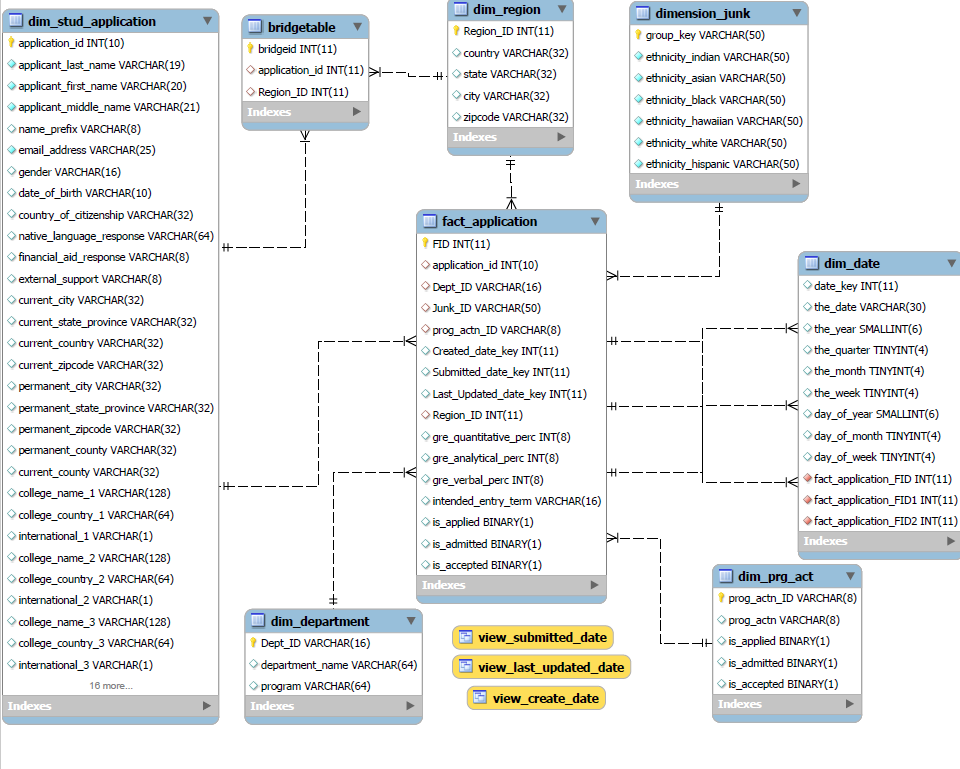
## Dimensions:

|  |  |  |
| --- | --- | --- |
| S.NO | Static Dimension | Insert-Update Dimension |
| 1. | Dim\_Junk | Dim\_Stud\_Application |
| 2. | Dim\_Prg\_Act | Dim\_Region |
| 3. |  | Dim\_Department |
| 4. |  | Dim\_Date |

**Fact\_Table:** Fact\_Application

**Bridge Table (Not part of implementation):**

We are using the bridge table concept to join the dim\_stud\_application table and dim\_region table.



**Design:**

**Dimension Attributes**

|  |
| --- |
| Dim\_Stud\_Application |
|  |
| `application\_id` VARCHAR(10) PRIMARY KEY, |
| `applicant\_last\_name` VARCHAR(19) NOT NULL DEFAULT '', |
| `applicant\_first\_name` VARCHAR(20) NOT NULL DEFAULT '', |
| `applicant\_middle\_name` VARCHAR(21) NOT NULL DEFAULT '', |
| `name\_prefix` VARCHAR(8) CHARACTER SET utf8 DEFAULT NULL, |
| `email\_address` VARCHAR(25) NOT NULL DEFAULT '', |
| `gender` VARCHAR(16) CHARACTER SET utf8 DEFAULT NULL, |
| `date\_of\_birth` VARCHAR(10) DEFAULT NULL, |
| `country\_of\_citizenship` VARCHAR(32) CHARACTER SET utf8 DEFAULT NULL, |
| `native\_language\_response` VARCHAR(64) CHARACTER SET utf8 DEFAULT NULL, |
| `financial\_aid\_response` VARCHAR(8) CHARACTER SET utf8 DEFAULT NULL, |
| `external\_support` VARCHAR(8) CHARACTER SET utf8 DEFAULT NULL, |
| `college\_name\_1` VARCHAR(128) CHARACTER SET utf8 DEFAULT NULL, |
| `college\_country\_1` VARCHAR(64) CHARACTER SET utf8 DEFAULT NULL, |
| `international\_1` VARCHAR(1) DEFAULT NULL, |
| `college\_name\_2` VARCHAR(128) CHARACTER SET utf8 DEFAULT NULL, |
| `college\_country\_2` VARCHAR(64) CHARACTER SET utf8 DEFAULT NULL, |
| `international\_2` VARCHAR(1) DEFAULT NULL, |
| `college\_name\_3` VARCHAR(128) CHARACTER SET utf8 DEFAULT NULL, |
| `college\_country\_3` VARCHAR(64) CHARACTER SET utf8 DEFAULT NULL, |
| `international\_3` VARCHAR(1) DEFAULT NULL, |
| `rec\_firstname\_1` VARCHAR(17) DEFAULT NULL, |
| `rec\_lastname\_1` VARCHAR(16) DEFAULT NULL, |
| `rec\_country\_1` VARCHAR(64) CHARACTER SET utf8 DEFAULT NULL, |
| `rec\_lttr\_received\_dt\_1` VARCHAR(20) CHARACTER SET utf8 DEFAULT NULL, |
| `rec\_firstname\_2` VARCHAR(17) DEFAULT NULL, |
| `rec\_lastname\_2` VARCHAR(16) DEFAULT NULL, |
| `rec\_country\_2` VARCHAR(64) CHARACTER SET utf8 DEFAULT NULL, |
| `rec\_lttr\_received\_dt\_2` VARCHAR(20) CHARACTER SET utf8 DEFAULT NULL, |
| `rec\_firstname\_3` VARCHAR(17) DEFAULT NULL, |
| `rec\_lastname\_3` VARCHAR(16) DEFAULT NULL, |
| `rec\_country\_3` VARCHAR(64) CHARACTER SET utf8 DEFAULT NULL, |
| `rec\_lttr\_received\_dt\_3` VARCHAR(20) CHARACTER SET utf8 DEFAULT NULL, |
| `gre\_quantitative\_perc` INT(8) DEFAULT NULL, |
| `gre\_analytical\_perc` INT(8) DEFAULT NULL, |
| `gre\_verbal\_perc` INT(8) DEFAULT NULL, |
| `intended\_entry\_term` VARCHAR(16) CHARACTER SET utf8 DEFAULT NULL |
| Dim\_Region |
|  |
| Region\_ID VARCHAR(32) CHARACTER SET utf8 PRIMARY KEY, |
| `country` VARCHAR(32) CHARACTER SET utf8 DEFAULT NULL, |
| `state` VARCHAR(32) CHARACTER SET utf8 DEFAULT NULL, |
| `city` VARCHAR(32) CHARACTER SET utf8 DEFAULT NULL, |
| `zipcode` VARCHAR(32) CHARACTER SET utf8 DEFAULT NULL |
|  |
| Dim\_ Junk |
|  |
| Junk\_Id VARCHAR(32) CHARACTER SET utf8 PRIMARY KEY,  Group\_key varchar(50), |
| `ethnicity\_indian` VARCHAR(32) CHARACTER SET utf8 DEFAULT NULL, |
| `ethnicity\_asian` VARCHAR(32) CHARACTER SET utf8 DEFAULT NULL, |
| `ethnicity\_black` VARCHAR(32) CHARACTER SET utf8 DEFAULT NULL, |
| `ethnicity\_hawaiian` VARCHAR(64) CHARACTER SET utf8 DEFAULT NULL, |
| `ethnicity\_white` VARCHAR(32) CHARACTER SET utf8 DEFAULT NULL, |
| `ethnicity\_hispanic` VARCHAR(8) CHARACTER SET utf8 DEFAULT NULL, |
|  |
|  |
| Dim\_department |
|  |
| Dept\_ID VARCHAR(16) CHARACTER SET utf8 PRIMARY KEY, |
| `department\_name` VARCHAR(64) CHARACTER SET utf8 DEFAULT NULL, |
| `program` VARCHAR(64) CHARACTER SET utf8 DEFAULT NULL |
|  |
| Dim\_Prg\_Act |
|  |
| `prog\_actn\_ID` VARCHAR(8) CHARACTER SET utf8 PRIMARY KEY, |
| `prog\_actn` VARCHAR(8) CHARACTER SET utf8 DEFAULT NULL, |
| is\_applied BINARY , |
| is\_admitted BINARY, |
| is\_accepted BINARY |

|  |
| --- |
| Dim\_dim\_date |
| `date\_key` int(11) DEFAULT NULL, |
| `the\_date` date DEFAULT NULL, |
| `the\_year` smallint(6) DEFAULT NULL, |
| `the\_quarter` tinyint(4) DEFAULT NULL, |
| `the\_month` tinyint(4) DEFAULT NULL, |
| `the\_week` tinyint(4) DEFAULT NULL, |
| `day\_of\_year` smallint(6) DEFAULT NULL, |
| `day\_of\_month` tinyint(4) DEFAULT NULL, |
| `day\_of\_week` tinyint(4) DEFAULT NULL |

Note: Views for the date will be populated using transformation.

**Fact Table Attributes:**

|  |
| --- |
| Fact\_admission |
|  |
| FID INT AUTO\_INCREMENT, |
| `Junk\_ID` VARCHAR(8) , |
| `Created\_date\_key` INT (11), |
| `Submitted\_date\_key` INT (11), |
| `Last\_Updated\_date\_key` INT (11), |
| `Region\_ID` INT , |
| `gre\_quantitative\_perc` INT(8) DEFAULT NULL, |
| `gre\_analytical\_perc` INT(8) DEFAULT NULL, |
| `gre\_verbal\_perc` INT(8) DEFAULT NULL, |
| `intended\_entry\_term` VARCHAR(16) DEFAULT NULL, |
| `is\_applied` BINARY(1) DEFAULT NULL, |
| `is\_admitted` BINARY(1) DEFAULT NULL,  `is\_accepted` BINARY(1) DEFAULT NULL,  **ETL Design Optimization:**  Initial Design for populating fact table is as follows. Below is what we did.   1. Selected the required columns from the source file. 2. Compared the current selected values with every dimension to gather the respective keys. 3. After gathering all this information, selected only the keys to be populated into the fact table. |

Post designing this, it was realized that this is not a very improved model for optimization. On realizing the same, have modified and used the current transformation which has helped to optimize the query time from 40sec to a mere 10 sec.

**Assumption:**

We are not considering few columns for our analysis and design process.

**Documents:**

DDL Scripts

****

Note – At times we have to run the table creation statements individually, as running them all at once gives an error. On running them individually it works perfectly.

S2T

