**ISYS 637 - Data Warehouse - Group Project Report**

****

**SECTION 603 - GROUP 7 – Brazos University**

**SAIRAM JANAKIRAMAN (924001504)**

**BHARADWAJ VENKATESHWARAN (823009748)**

**RATNADEEP SIMHADRI (424000610)**

**DINESH KAKARLA (723008494)**

**COMPLETED ON 4th May 2015**

Contents

[Introduction: 3](#_Toc418478153)

[Need for data Warehouse: 3](#_Toc418478154)

[Business Requirements: 4](#_Toc418478155)

[Business Dimensions: 4](#_Toc418478156)

[Key Business Metrics: 6](#_Toc418478157)

[Architectural Components of the DW: 7](#_Toc418478158)

[Source Data: 7](#_Toc418478159)

[Data Staging: 7](#_Toc418478160)

[Data Storage: 11](#_Toc418478161)

[Information Delivery: 14](#_Toc418478162)

[Dimensional Model for Brazos University Data Warehouse 14](#_Toc418478163)

[OLAP Cubes and Reports: 16](#_Toc418478164)

[Defining Data Source and Data Source View: 16](#_Toc418478165)

[Create a new Project: 17](#_Toc418478166)

[Add New Data Source: 17](#_Toc418478167)

[Create Data Source View: 18](#_Toc418478168)

[Creating Cubes and Reports: 18](#_Toc418478169)

[Facilities: 18](#_Toc418478170)

[Admissions and Financial aid: 22](#_Toc418478171)

[Course Registration: 25](#_Toc418478172)

[Placements: 28](#_Toc418478173)

[Appendix 32](#_Toc418478174)

[A. Infrastructure: 32](#_Toc418478175)

[B. Metadata: 33](#_Toc418478176)

[C. Estimate of Size of DW: 35](#_Toc418478177)

[D. Peer Evaluation Forms: 36](#_Toc418478178)

[References: 37](#_Toc418478179)

# Introduction:

Founded in 1945, Brazos University (BU) is a well-regarded private university. It has an enrollment of roughly 7,000 students, with current areas of study that include liberal arts, music, design and social sciences, as well as certificate programs. Traditionally, there has been less focus on revenue and profit in this arena, but with the ever-escalating costs and competition associated with higher education, universities and colleges are very interested in attracting and retaining high-quality students. Brazos University is no exception to this. They are interested in getting the best of the students into their undergraduate, graduate and doctoral programs.

In order to understand their inflow of revenue, facility usage, alumni connections, finance status and applicants, BU is looking for a team to design and implement a data warehouse that would provide them strategic information about all their important processes.

## Need for data Warehouse:

The concept of data warehousing is not hard to understand. The notion is to create a permanent storage space for the data needed to support reporting, analysis, and other BI functions. On the surface, it may seem wasteful to store data in more than one place.

Brazos University is looking to attain the highest quality of students into their programs for which they require strategic information to make critical decisions. The historic data is available but not being used in the right manner. By using a data warehouse, BU would be able to make on the fly decisions regarding admissions, course capacity increase and new facility buildings which would give them competitive edge over other competing universities.

The best data warehouses do some predigesting of the raw data in anticipation of the types of reports and inquiries that will be requested. This is done by developing and storing metadata (i.e., new fields such as averages, summaries, and deviations that are derived from the source data). There is some art involved in knowing what kinds of metadata will be useful in support of reporting and analysis. The best data warehouses include a rich variety of useful metadata fields. It is important for BU to create the requirements of their data warehouse and the types of reports they require which would help the design team in incorporating accurate data that can be analyzed.

Once a data warehouse is made operational, it is important that the data model remain stable. If it does not, then reports created from that data will need to be changed whenever the data model changes. New data fields and metadata need to be added over time in a way that does not require reports to be rewritten. Therefore, it is highly essential for the implementation team to make note of measures which would help BU create precise reports.

There is a need to integrate many different sources of data in near real-time. This will allow for better business decisions because BU users will have access to more data. Plus this will save users lots of time because they won’t waste precious time retrieving data from multiple sources. Tons of historical data that is needed to gather would be easily accessible and will have common formats, common keys, common data model, and common access methods. Data can be restructured, tables and fields renames so that it makes more sense to users. In traditional database systems, users are running reports directly against operational systems, causing performance problems. Instead, a data warehouse would help users run reports off of that. Data warehouse is optimized for read access, resulting in faster report generation.

There is a risk that BI users might misuse or corrupt the transaction data. Having an easy to use data warehouse allows users to create their own reports without having to get IT involved. Leading to “Self Service BI”. A data warehouse is a convenient place to create and store metadata. Improve data quality by cleaning up data as it is imported into the data warehouse (providing more accurate data) as well as providing consistent codes and descriptions. Having a data warehouse makes it easy to create business intelligence solutions on top of it, such as SSAS cubes.

# Business Requirements:

BU has provided a list of requirements for the data warehouse in relation to the type of data, measures and reports to be obtained. The requirements are from various domains.

The first step is to define the various business dimensions that need to be tracked along with their relevant hierarchies. Business dimensions form the underlying basis of the new methodology for requirements definition. The business dimensions and their hierarchical levels form the basis for all further phases.

## Business Dimensions:

BU has identified important domains of the university that need data to be tracked on a daily basis. These domains require dimensions to be defined in order to track measures. The dimensions are listed below along with their usage.

Applicant: The core information of BU is about its applicants, who apply from various countries and during different times in the year. The important attributes of the applicant that need to be tracked are:

* Applicant Name – First name and last name of the applicant are required as per federal regulations. Applicants can fill out applications through online web portal and at the university. The first and last name of the applicant can be retrieved from their application form.
* Applicant Address – All details related to the physical and contact location of the applicant are required. Details like country, state, city, street address and zip code need to be entered into the DW.
* Applicant Phone – The contact number of the applicant at which he/she can be contacted
* Applicant Birthday – The age can of the applicant can be helpful for planning coursework and more importantly understanding the age group of applicant pool.
* Marital Status and Gender – Demographic information is key to the university DW and this information can be retrieved from the applicant form.
* Examination Scores – Entrance exam scores of the applicant which would be used to make the final decision of admission is important to determine the quality of applicants.

Course: BU is very interested in understanding patterns related to course registration and the kind of students that register for a particular course. Reports related to tuition generated by a course and which professor takes that course are important for planning course enrollments. The important attributes of the course that need to be tracked are:

* Course Size: The maximum capacity/registrations that a course can have to be tracked as room requirements, professor availability and exam schedule all depend on this attribute.
* Course Fee: The cost of attending this course to be an important dimension.

Degree and Department: Information regarding degrees offered and the relevant departments offering them are to be tracked and measured. Department Head are officials in charge and one department can offer multiple degrees. The important attributes of the course that need to be tracked are:

* Degree Duration: The minimum and maximum duration of the degree will provide information regarding the active years of education of students.

Placement Domain: BU requires details in relation to student placement, graduate placement rate, base salary etc… The information would help BU keep in touch with alumni and ensure a strong alumni presence. From a marketing point of view, the salary statistics could influence future students to join BU. The important attributes of the course that need to be tracked are:

* Domain Category: The domain of the company where the students get a job offer
* Role: For which position/role the student has taken the job.

Facility: Since BU is looking to expand its student base, it also needs to ensure that its facilities are not overcrowded. Another aspect is maximum utilization of existing facilities and allocation of resources to building new required facilities. BU has requested for reports related to facility usage in relation with amount of revenue and the number of students attending/using the facility. The important attributes of the facility that need to be tracked are:

* Facility Capacity: Maximum number of people that can be accommodated in a facility.
* Facility Type: Each facility at BU is categorized into a number of facility types.

Faculty: BU has been known for its reputed faculty and are looking to track measures using their faculty and the courses they teach. The important attributes of the faculty that need to be tracked are:

* Position: The designation of the faculty is an important attribute that determines the courses they take and also the salaries of the faculty.
* Experience: The number of years of experience will help BU determine the growth and reputation of each faculty and the number of students that want to take up research or another opportunities under the faculty.

Program: Each department in BU offers a variety of programs. The programs are offered across different degrees.

Date: An important dimension to be tracked is date and time. Different metrics need to be measured at a daily, weekly, semester and annual level. The date dimension would be the foundation dimension along which all metrics would be measured.

## Key Business Metrics:

The approach for designing the data warehouse is such that each domain of the university would be modeled into a data mart with connecting dimensions, and an overall centralized schema would be designed by merging all data marts.

Admissions and Finance Metrics: As BU is looking for high quality students they require reports and analysis on the basis of applicants admission scores, acceptance rate based on scholarship offering, amount of federal loan dispersed etc… By understanding their applicant pool better, BU would be able to provide influence applicant decision by providing choice of semester enrollment, instate waiver for international and out-of-state applicants, and support of federal loan process. Important metrics to be tracked:

* Amount of Scholarship: This amount given to each applicant, tracked at a daily level and across programs would provide strategic information about the quality of the student and interest in different programs.
* Amount of Federal Loan: This is the amount borrowed by the student and can be compared against the amount of scholarship received per semester.

Facility Usage: Each facility needs to be tracked on a daily basis for the amount of revenue generated and the number of attendees. This would provide information regarding the utility of a particular facility and can give critical information to BU about which facility is overburdened and which is underutilized. Important metrics to be tracked:

* Number of Students: The total number of students (aggregated) that attend a particular event/stay at a facility.
* Revenue: Total amount of money received from ticket sales or rent of a particular facility.

Placement Statistics: Information regarding the companies that students get hired and their relevant roles would be an important measure from a marketing and administration perspective. Important metrics to be tracked:

* Total Salary: The amount of salary (aggregated) at a program level on a daily basis.
* NoofPlacements: The count of the number of student placements at a daily level across programs.

Course Registration: Since BU is intending to track course enrollment and the amount of tuition per course, measuring tuition received by a course taught by a particular faculty is a key measure.

* Total Tuition: The total amount of tuition received by a particular course.
* NumberOfStudents: The total number of students that have enrolled for a particular course.

Information Gathering Methodologies: Data would need to be collected through various sources like online web portals, existing OLTP databases, and conducting interviews with IT staff.

# Architectural Components of the DW:

## Source Data:

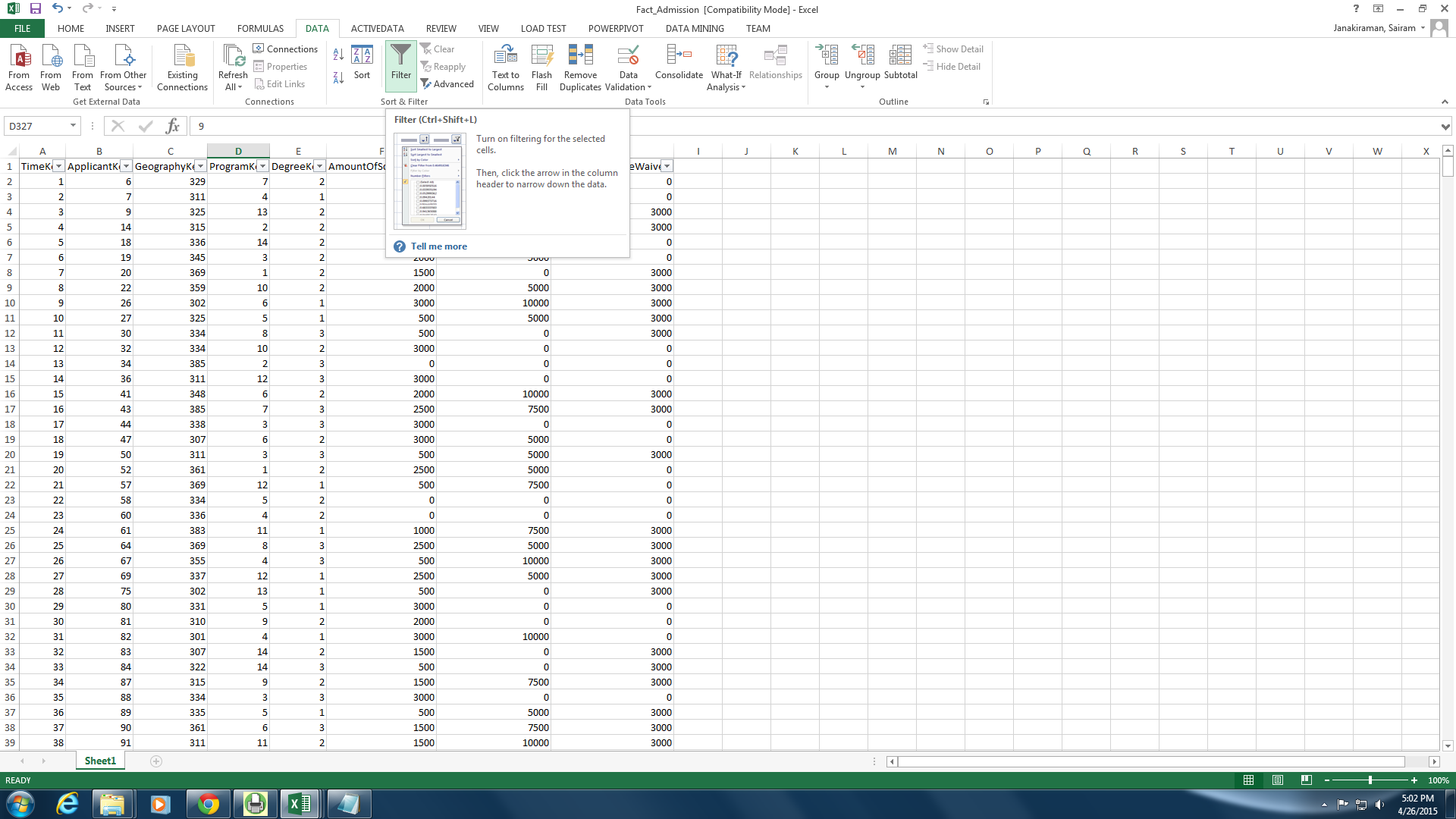
In general, different data sources feed data into a university data warehouse. Applicant data could be from operational data sources while faculty data could be from internal data sources such as files, spreadsheets or documents. While implementing the prototype for Brazos University Data Warehouse, data sources were not readily available. As a result, we had to create our own data sources which could then be used to feed the data for BU data warehouse.

There are many data generators that are available on the internet. We used www.mockaroo.com data generator to generate data for admissions, faculty, applicant, facility, registration tables etc. The data for the courses and department tables were extracted from the Texas A&M University website since “mockaroo” was unable to generate the related data.

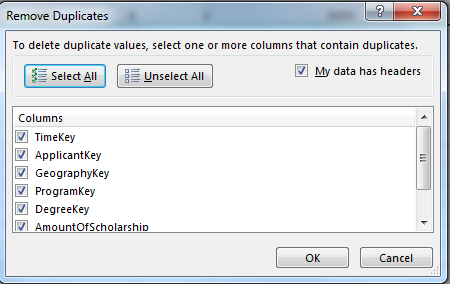
## Data Staging:

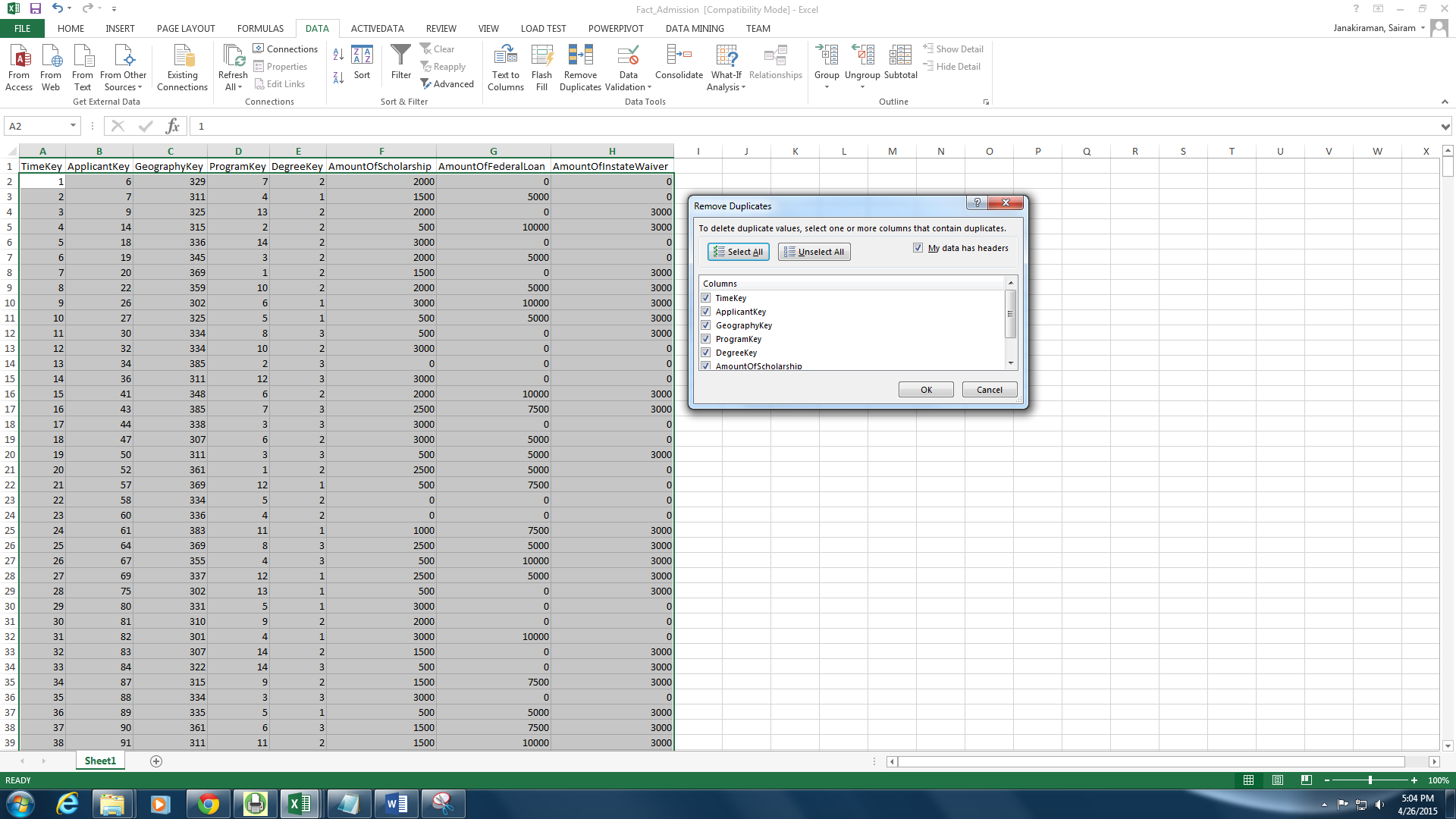
After the data sources were identified, a set of ETL processes were performed to enable the data to be loaded into the data warehouse. Data was extracted from the source system and stored in excel spreadsheets. While following the extraction process, business rules were followed to make sure that the structure of the data follows data warehouse schema. For example, each dimension table was assigned a surrogate key that would be the unique identifier for dimension table.

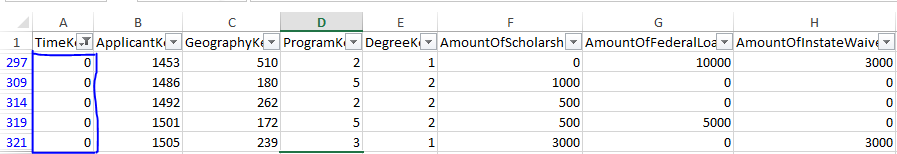
After the extraction process, each table was checked for data and referential integrity in case of a foreign key dependency. As part of the transformation process, aggregate data for the fact tables were added. Since the data was created by the team using external data generators, aggregate data created did not follow business rules. To mitigate this issue, formulas were created in excel, which would follow the business rule parameters.

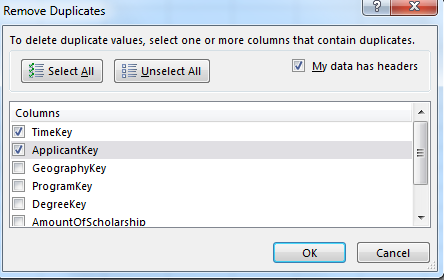


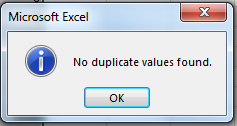
Once all the data were collected, the data was checked for duplicates and missing value using Excel’s remove duplicate functionality. The sequence of diagrams below illustrate the data staging process followed for Brazos University Data Warehouse





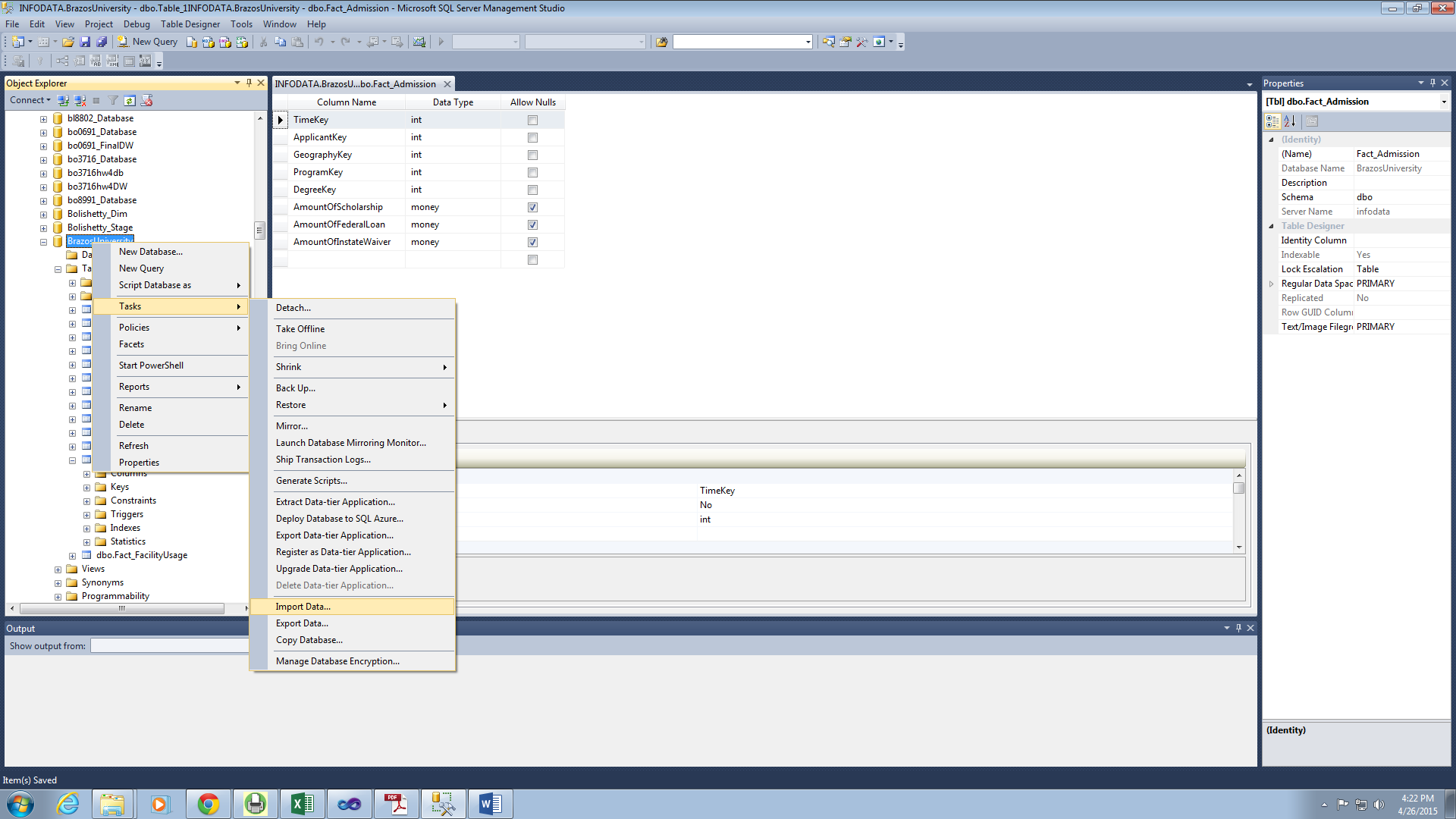




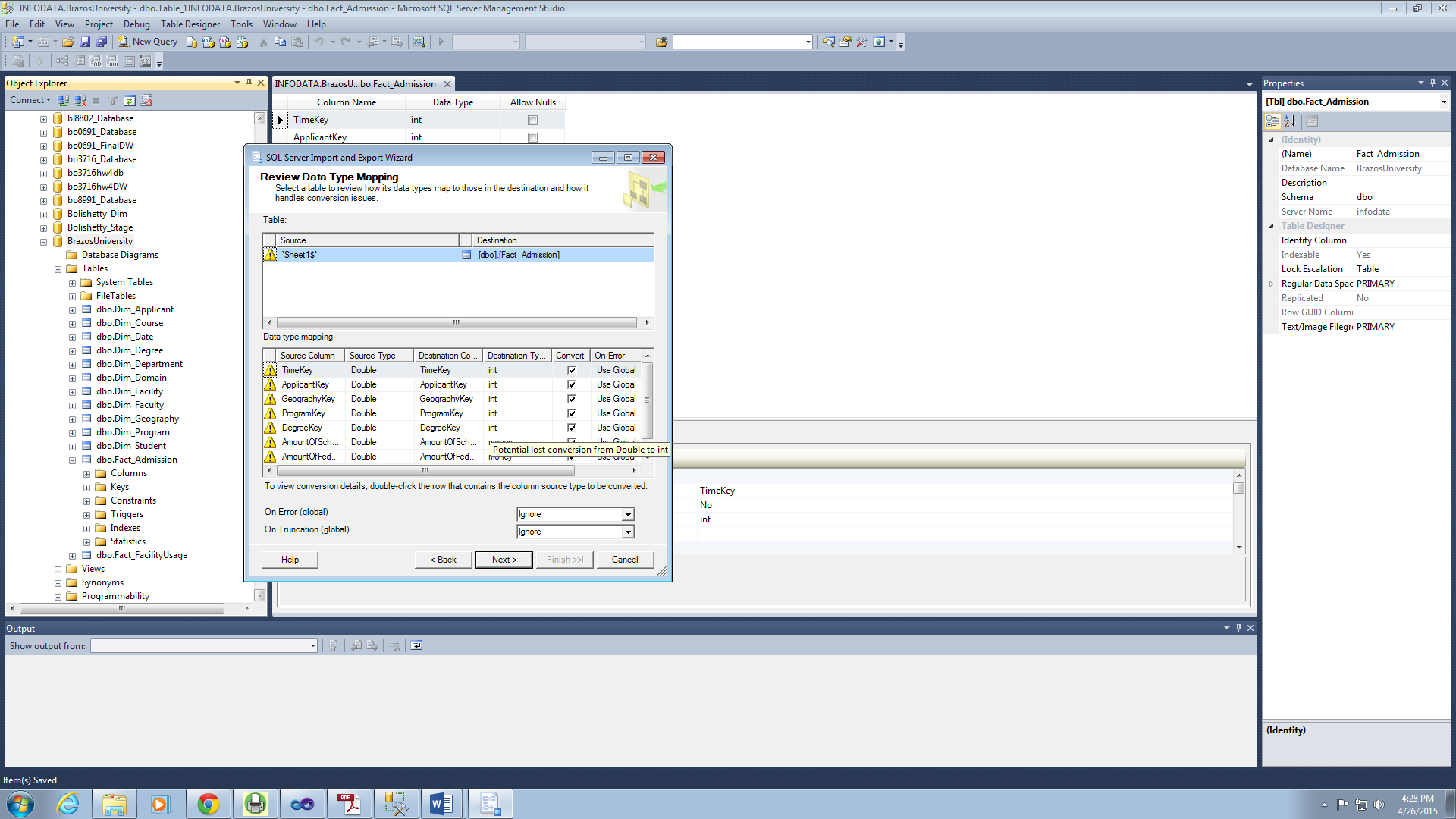


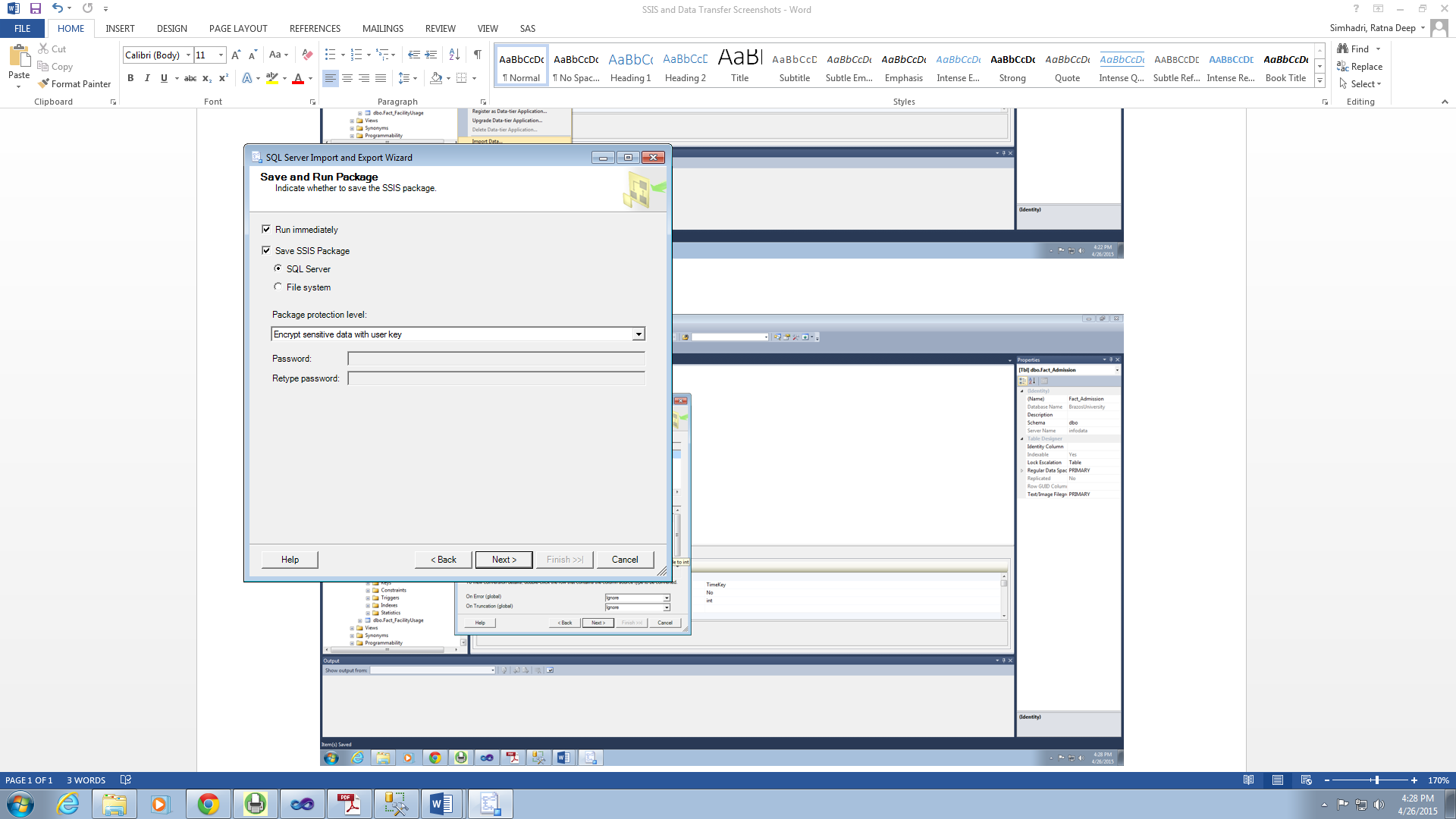
## Data Storage:

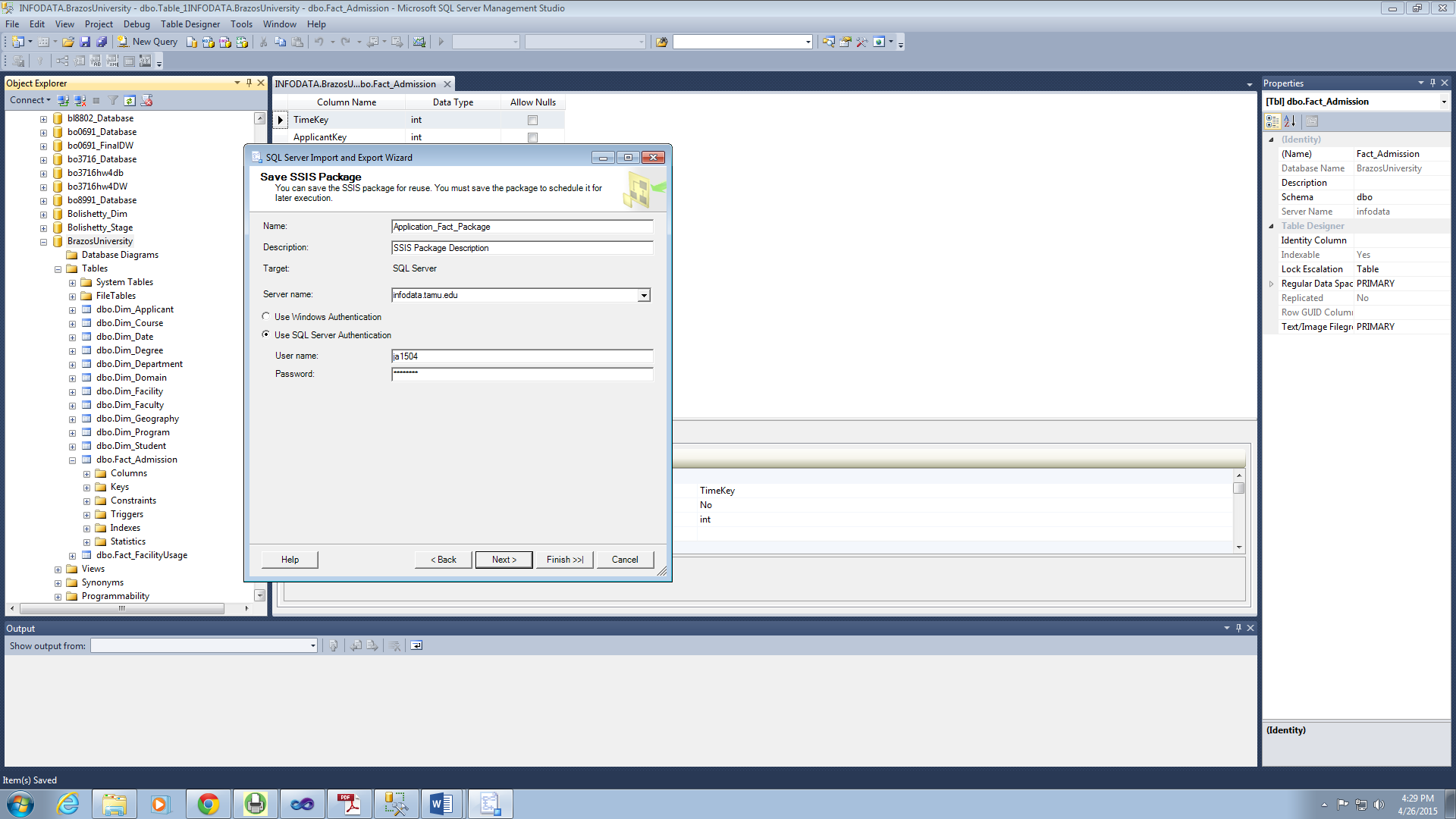
Once the transformation process is complete, duplicates of each data file is created to provide backup and recovery options if needed. Before loading the data into the data warehouse any remaining primary and foreign key referential are identified and resolved. The data sets are created and consolidated as excel sheets. Now these excel files are imported through the import option in the SQL Server Management Studio. Below are the screenshots of the steps that were followed to load the data into the data warehouse.



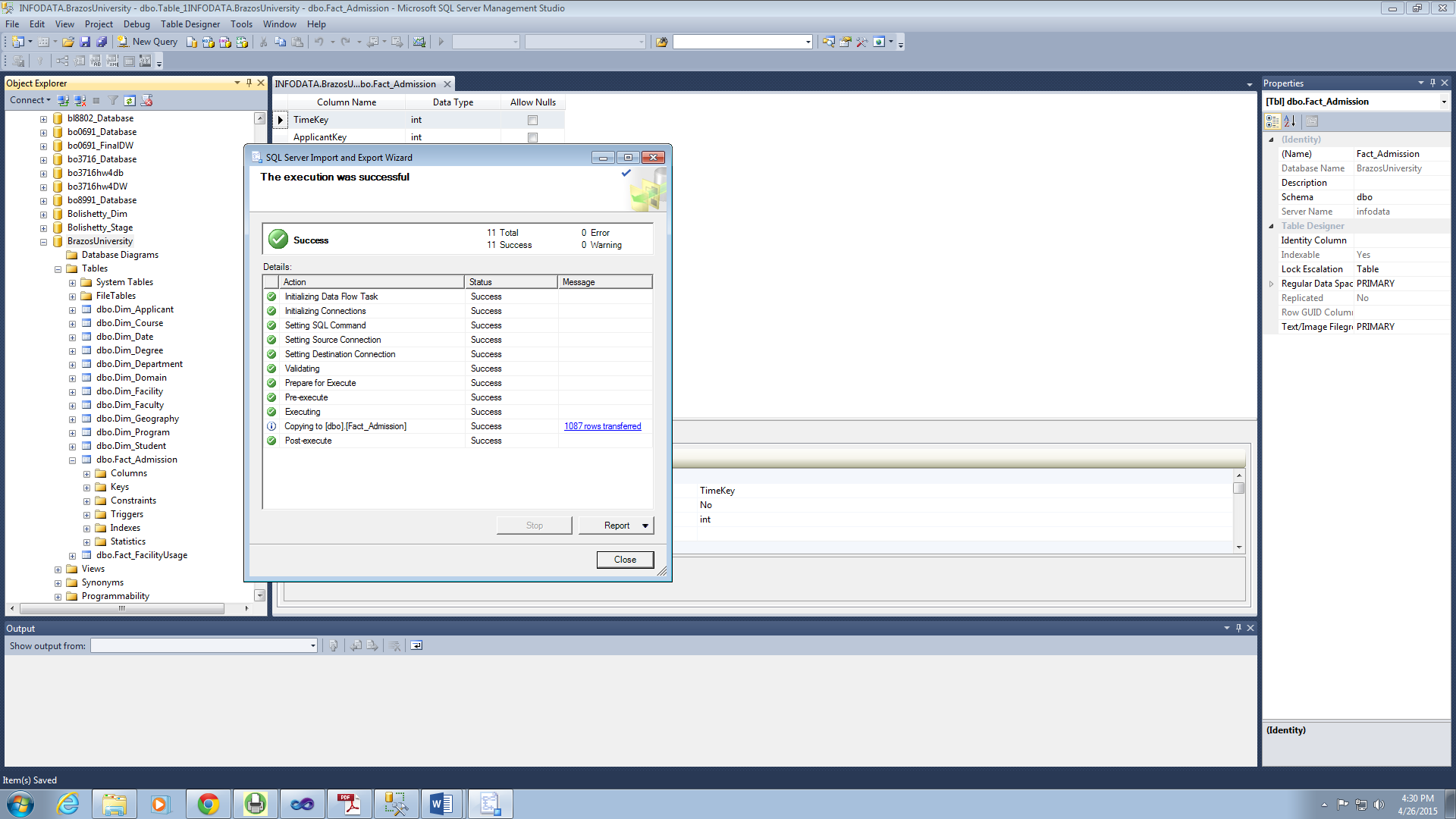
After importing the files, for each table a data type mapping is performed. This helps in defining the data types for each attribute of a table as well as assign foreign key referential for a particular attribute. The tasks option under the Data Warehouse is used to import data into the selected table of the DW. The data must be formatted prior to loading.







Choose the correct SSIS package for loading the data. SQL authentication is used for security purposes.

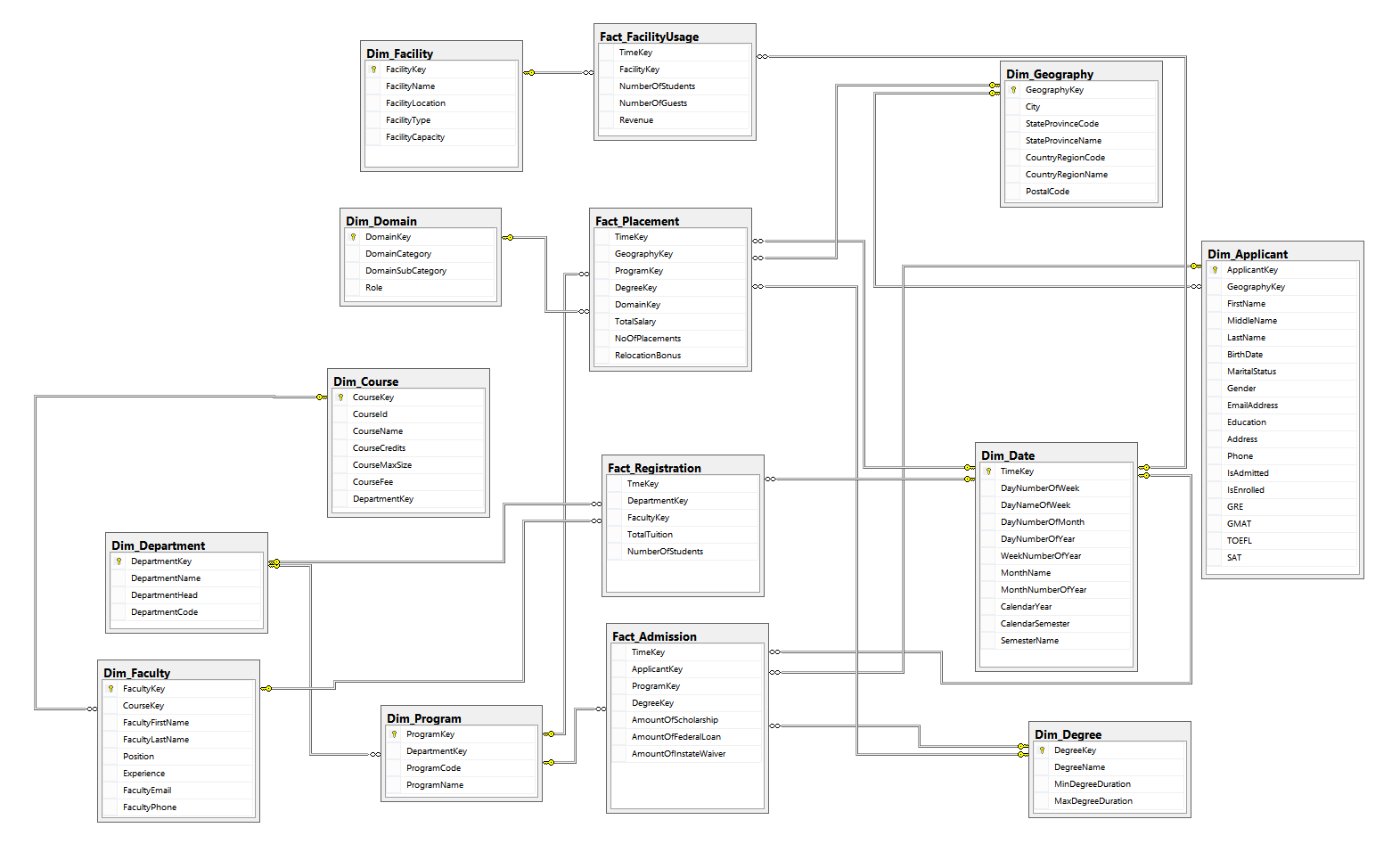


## Information Delivery:

Once the data is loaded into the data warehouse, there has to be a process which allows the users to access the data. The strength of any data warehouse architecture is manifested through the robustness and flexibility of the information delivery component. To achieve this, we have provided online analytical processing (OLAP) to help the users to generate reports that would be useful in making business decisions for Brazos University.

# Dimensional Model for Brazos University Data Warehouse

Below is the dimensional model for Brazos University. There are 4 fact tables and 10 dimensions.



The four main business processes involved in the Brazos University data warehouse are Admissions, Registration, Placements and Facility Usage. The schema above depicts the fact constellation for the data warehouse and as shown in the figure above, the corresponding facts for these business processes are *Fact\_Admissions, Fact\_Registration, Fact\_Placements* and *Fact\_FacilityUsage* respectively. The star schemas for these business process are integrated along the conformed dimensions viz. *Dim\_Date, Dim\_Degree and Dim\_Program*. These three dimensions will be shared by more than one data marts and they should be synchronized between the data marts. The other dimensions involved in the schema are *Dim\_Facility, Dim\_Geography, Dim\_Domain, Dim\_Applicant, Dim\_Course, Dim\_Department* and *Dim\_Faculty*.

|  |  |  |
| --- | --- | --- |
| **Business Process** | **Fact Table** | **Dimensions** |
| Admissions | Fact\_Admissions | Dim\_Program, Dim\_Degree, Dim\_Date, Dim\_Applicant |
| Registration | Fact\_Registration | Dim\_Department, Dim\_Faculty, Dim\_Date |
| Placements | Fact\_Placements | Dim\_Domain, Dim\_Program, Dim\_Geography, Dim\_Date, Dim\_Degree |
| Facility Usage | Fact\_FacilityUsage | Dim\_Facility, Dim\_Date |

The table shows the fact tables of each business process and the corresponding dimensions associated with each fact. Since our business questions do not require the grain to be at its lowest level, i.e. for each student, the measures in each fact table are generally aggregated values such as number of students, Total Tuition, Total Salary etc.

# OLAP Cubes and Reports:

For Brazos University, the business questions under consideration can be categorized into various functional areas as follows. This section illustrates in detail each of the business question and the associated OLAP reports from SSAS.

• Facilities

• Admissions and Finance

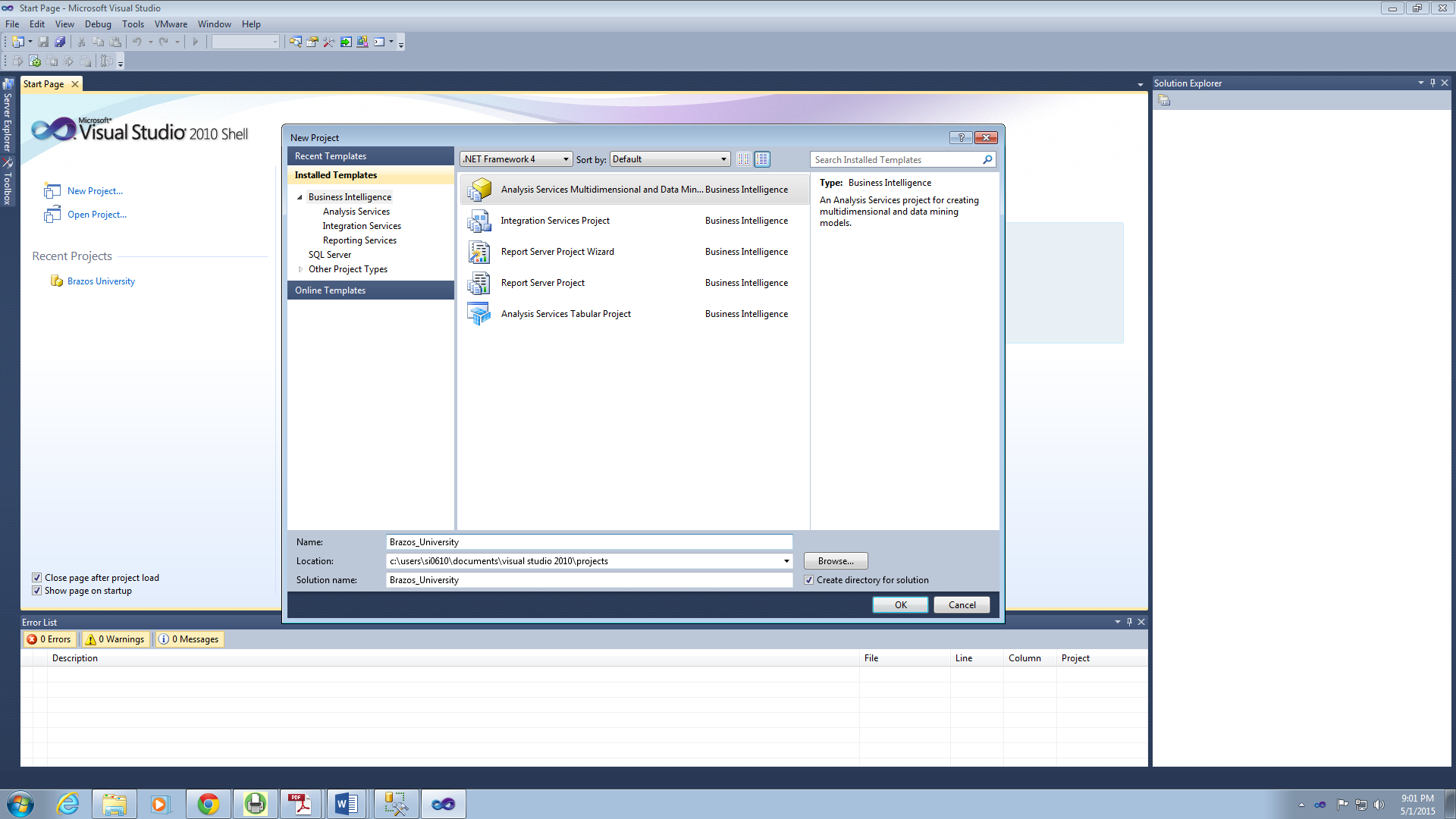
• Course Registration and Revenue

• Placements

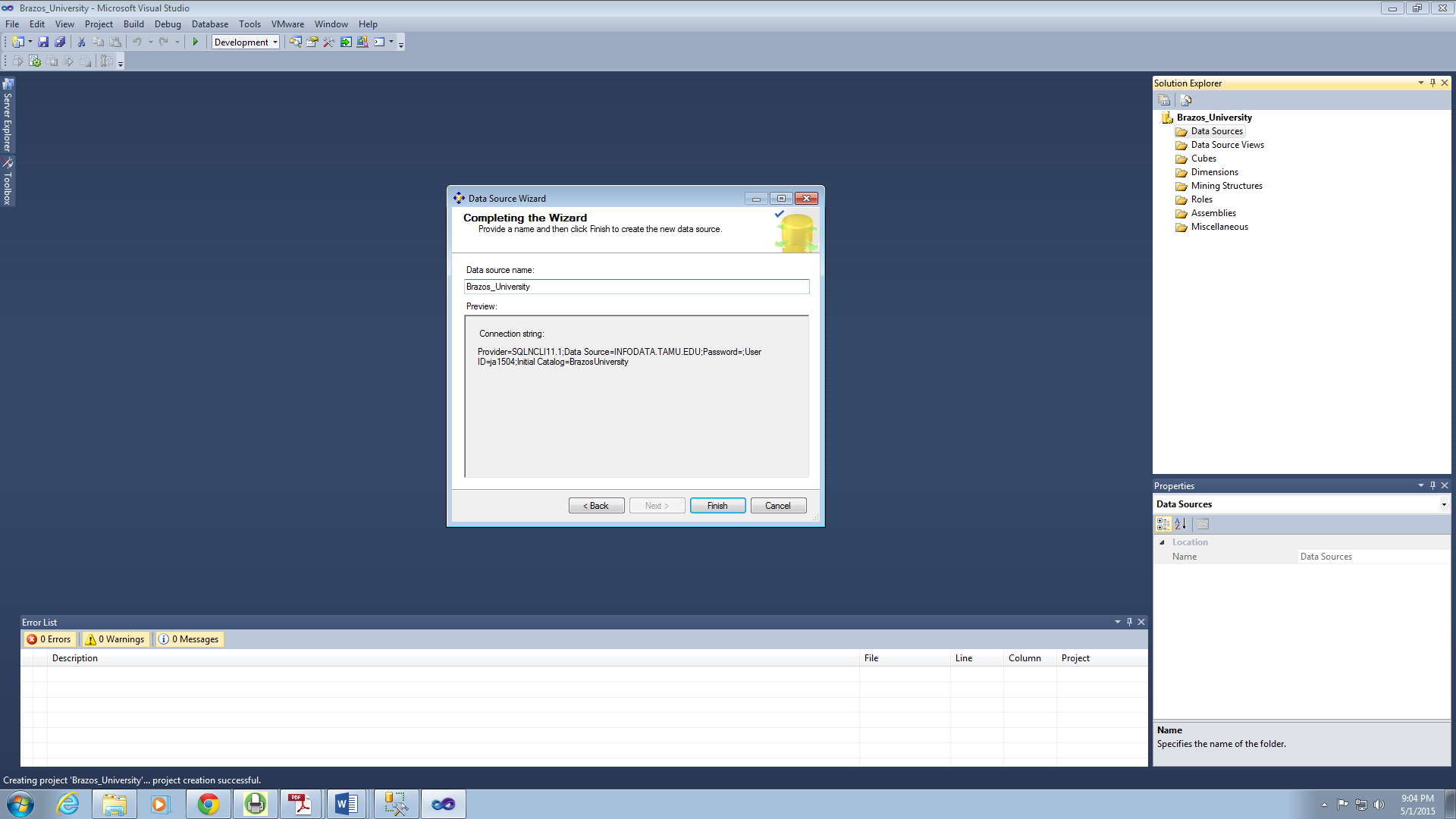
Before going into the details of individual cube and report development. The following section details the creation of Data Source and Data Source View.

## Defining Data Source and Data Source View:

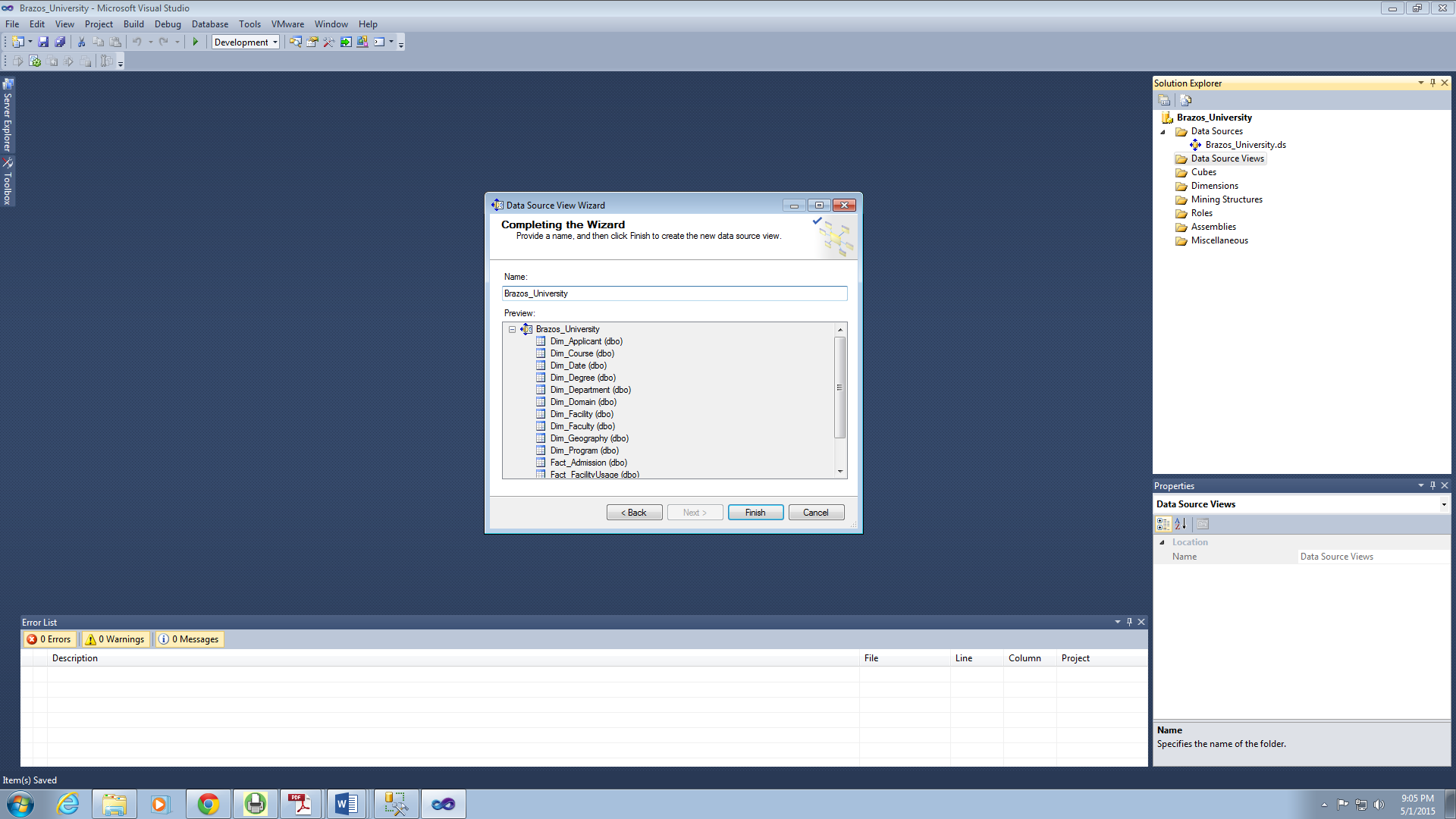
### Create a new Project:



### Add New Data Source:



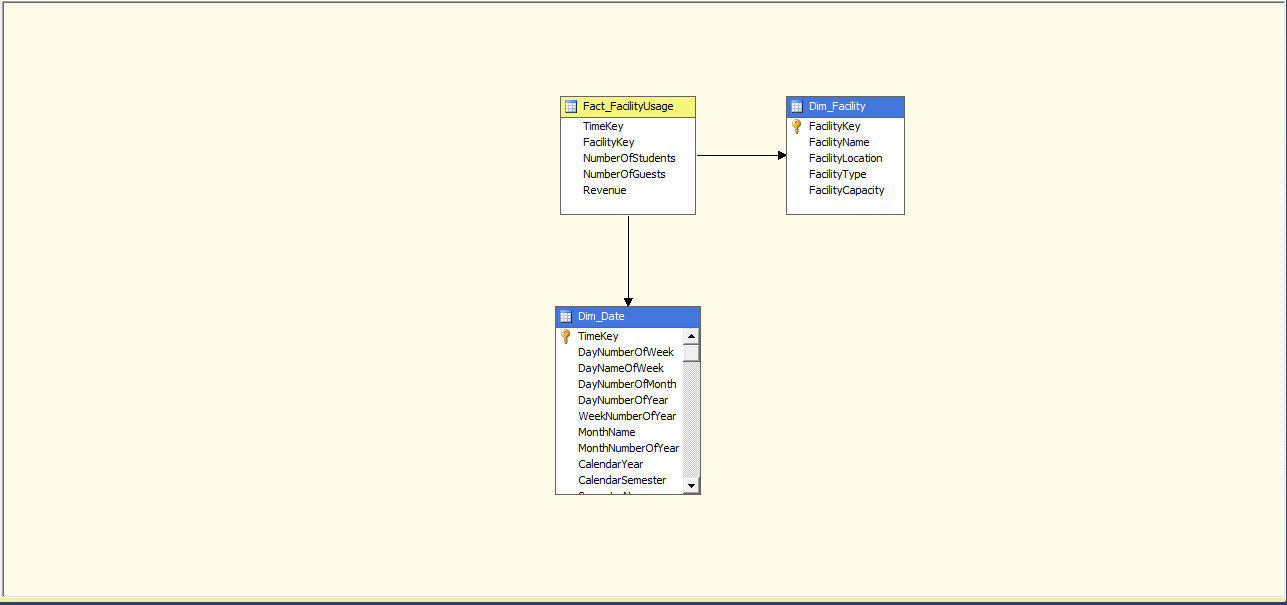
### Create Data Source View:



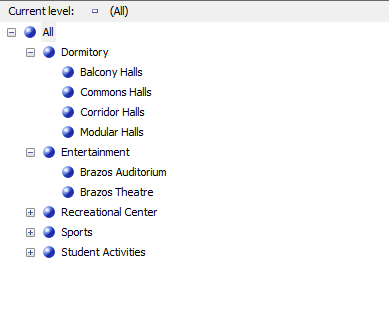
## Creating Cubes and Reports:

### Facilities:

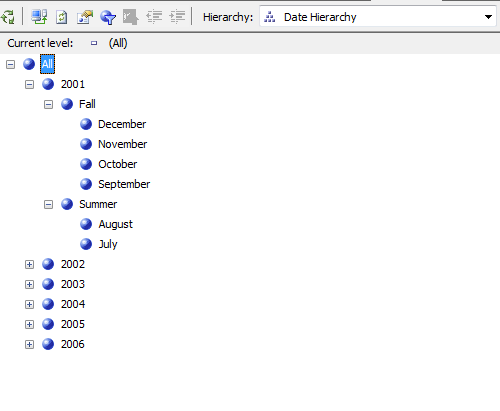
#### Cube Structure:



#### Hierarchies:



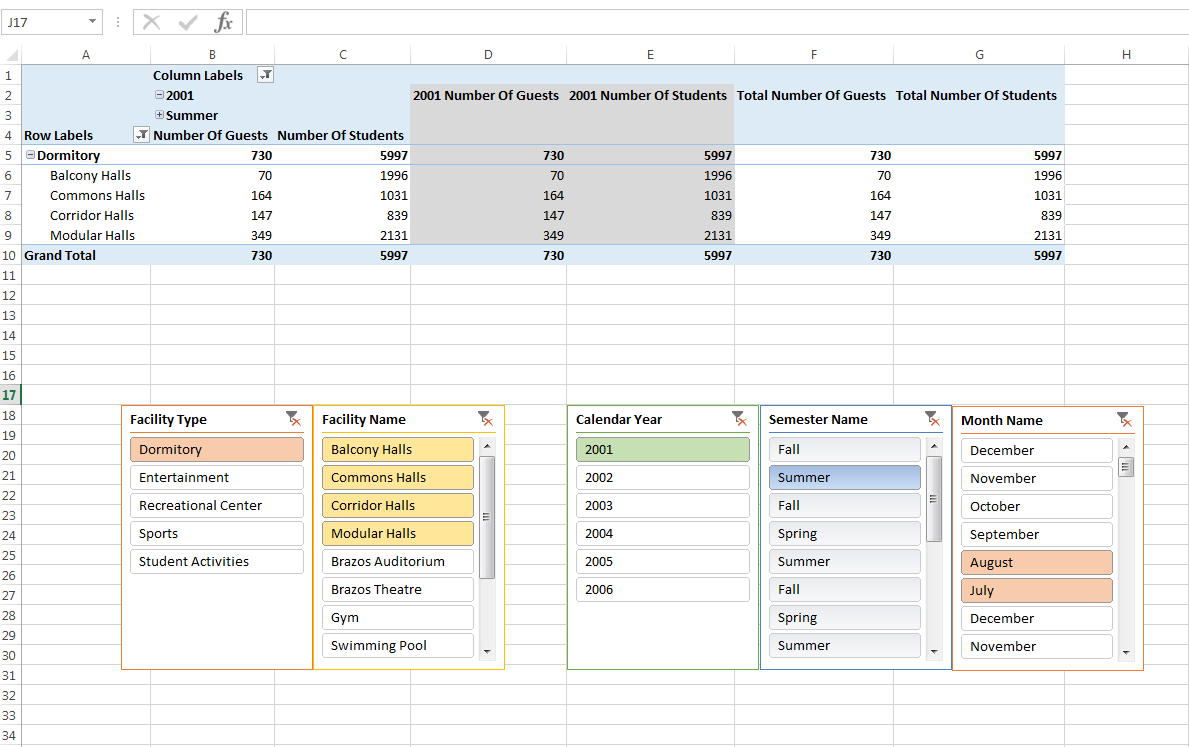
Facility Hierarchy

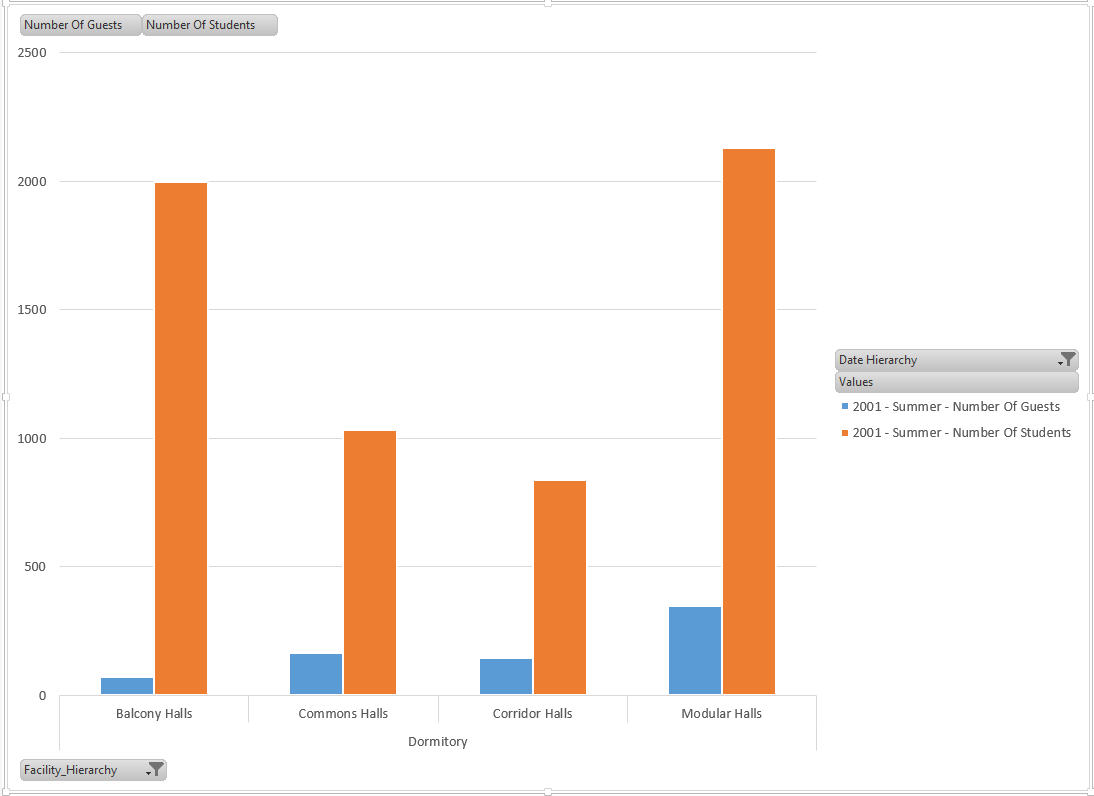


Date Hierarchy

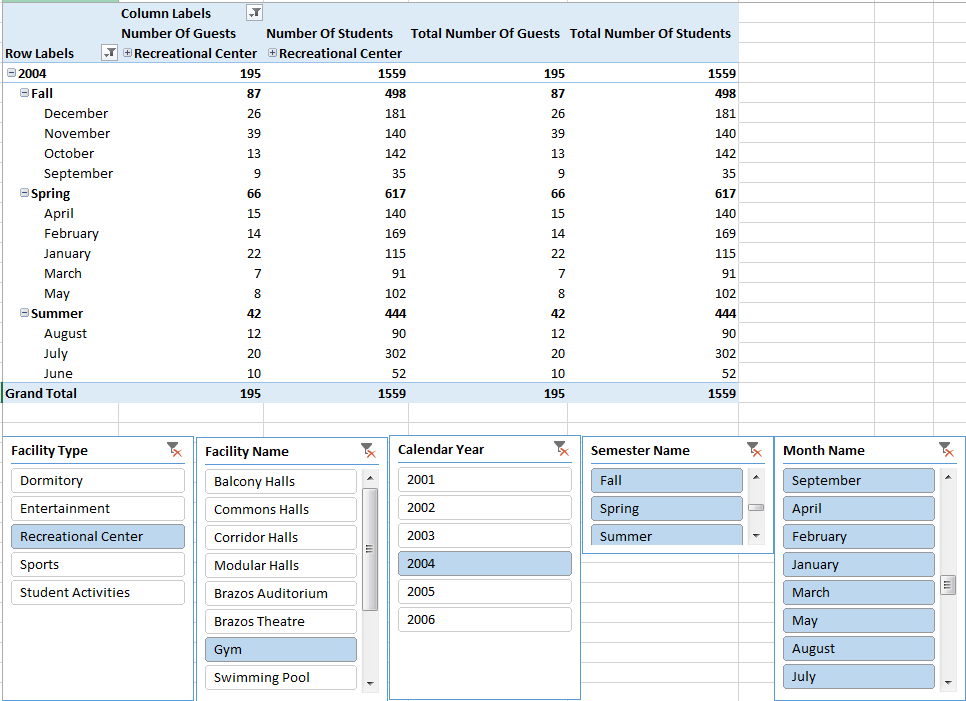
#### Reports:

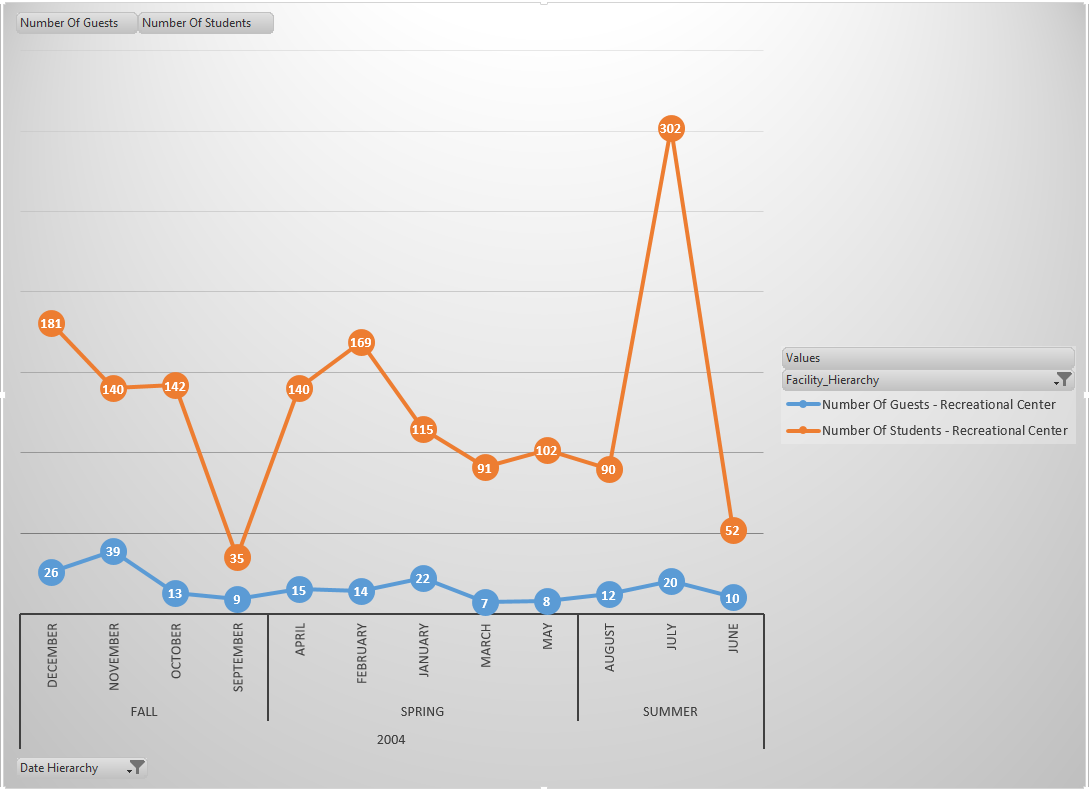
**Business Question:** Number of Students and Guests who have used Dormitory facilities during summer of 2001 and the comparison between each of the dormitory halls





**Business Question:** Identify thetrends in Gym usage for the year 2004

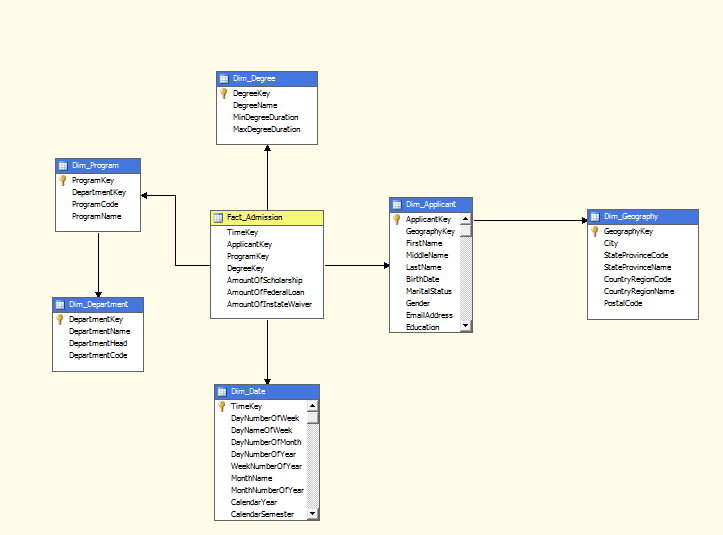




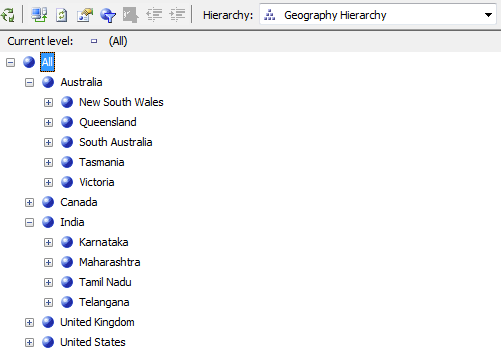
The above reports on facility usage are helpful for the management in understanding the trends in facility usage. Based on this, the university management will be able to make strategic decision on how when and where to improve the infrastructure of the facilities. For example from the first graph it can be observed that though there are greater number of students who prefer to stay in Balcony halls, relatively only a very few number of guests choose to stay at these halls. This insight will help into analyzing the underlying problems, for example facilities, associated with these dormitory halls.

### Admissions and Financial aid:

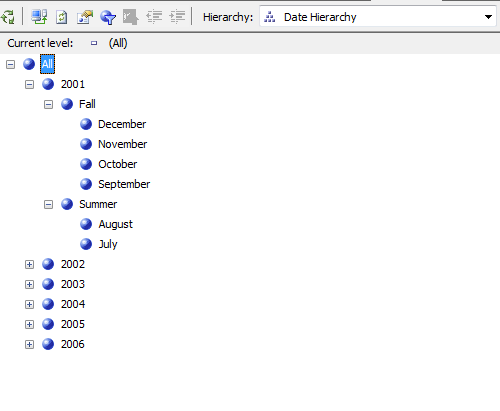
#### Cube Structure:



#### Hierarchies:



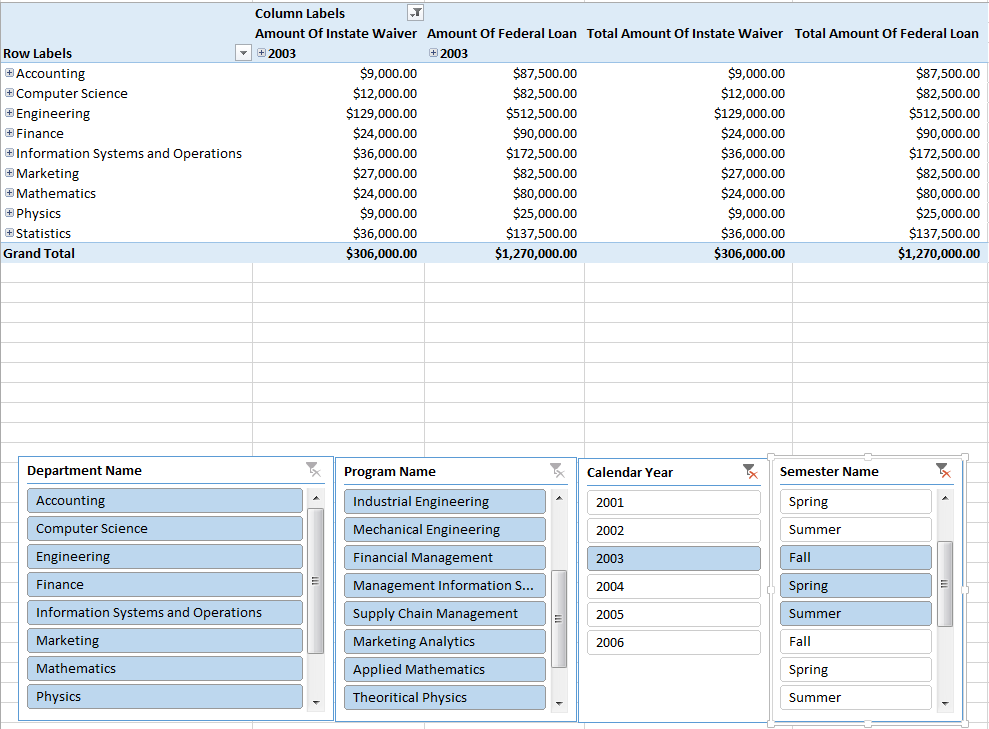
Geography Hierarchy

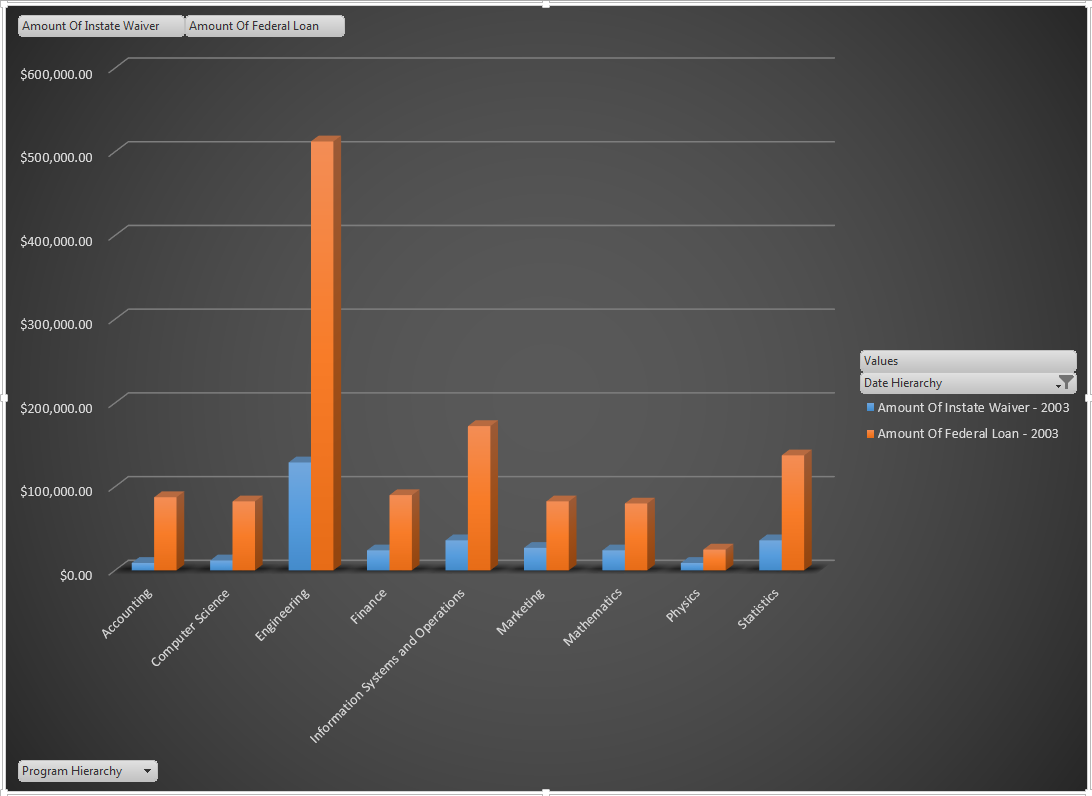


Date Hierarchy

#### Reports:

**Business Question:** Compare the amount of Instate Wavier provided and amount of Federal Loan granted among various departments during the year 2003

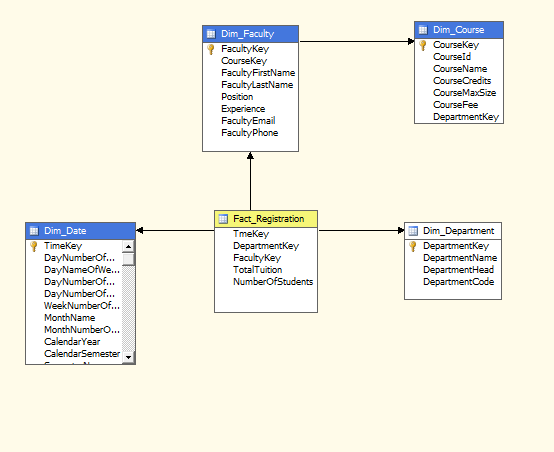




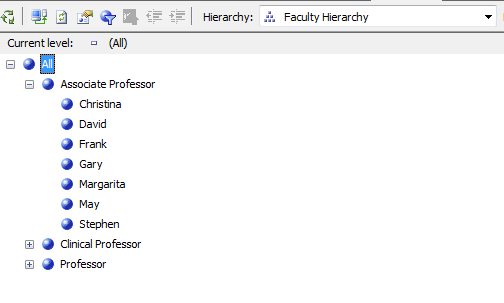
The above report helps in identifying the departments that sanction the highest amount of instate wavier. Also the management will be able to compare the amount of federal loan sanctioned to students of each department. These statistics can also be useful for strategic decision making on scholarship and funding allocation to departments.

### Course Registration:

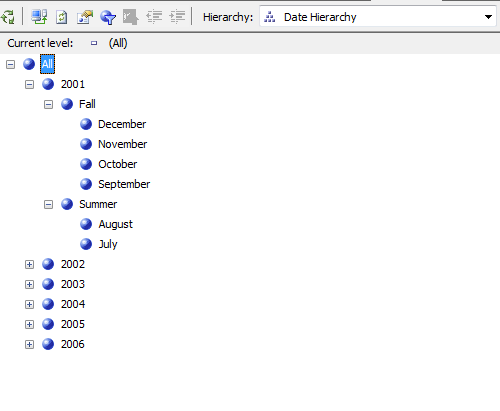
#### Cube Structure:



#### Hierarchies:



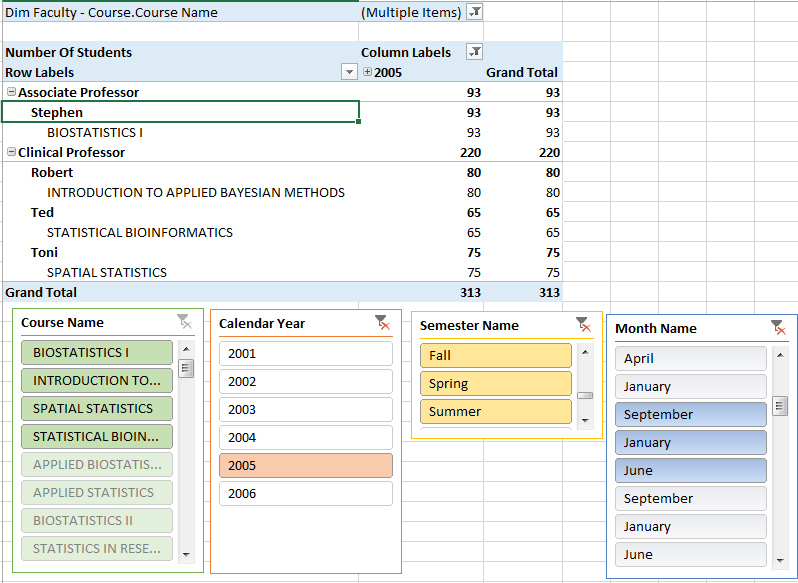
Faculty Hierarchy

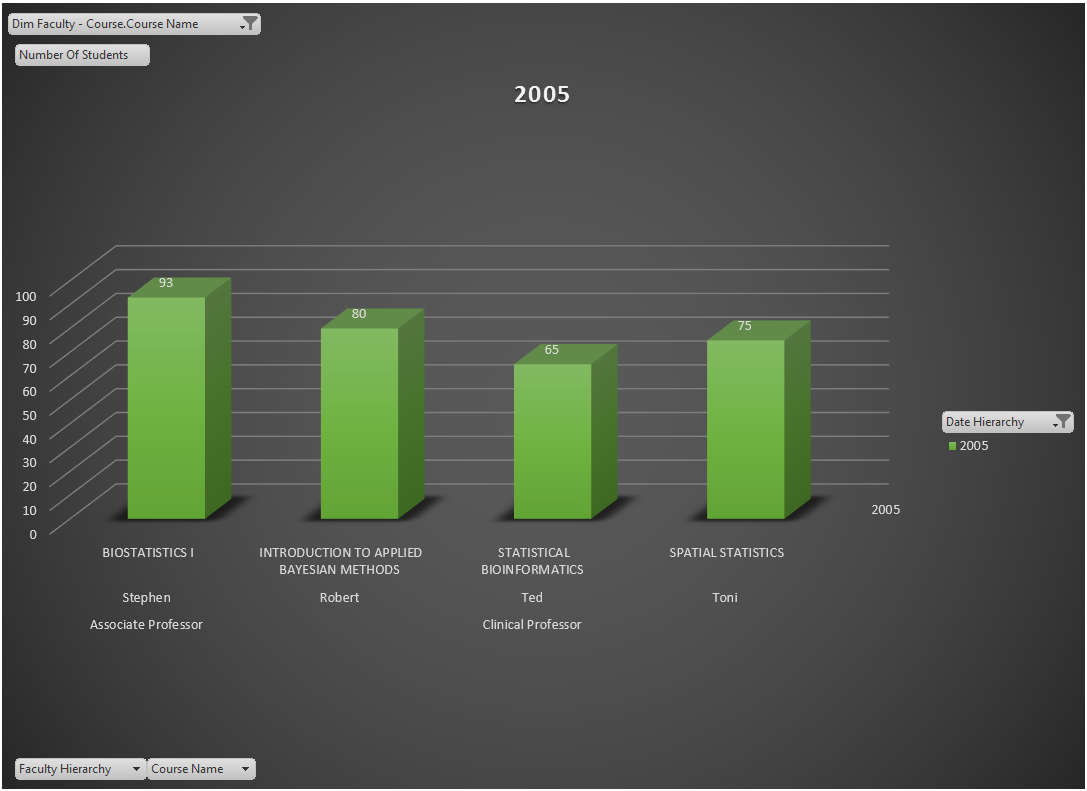


Date Hierarchy

#### Reports:

**Business Question:** Find the number of students who have enrolled for Biostatistics-I, Introduction to Applied Bayesian Methods, Statistical Bioinformatics and Spatial Statistics under each professor taking the course during the year 2005.

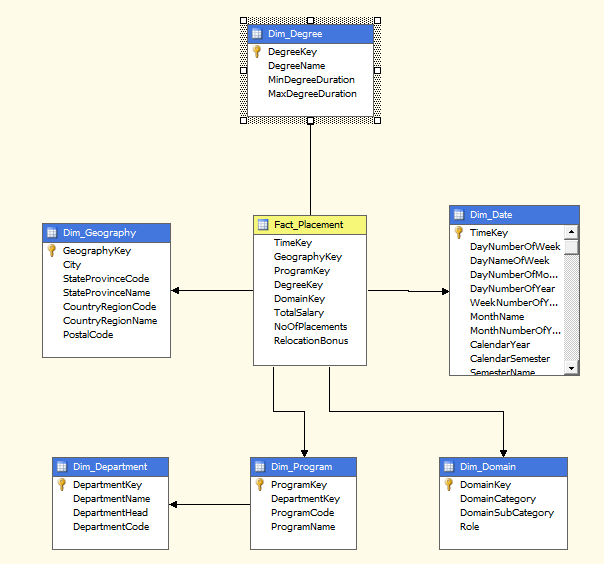




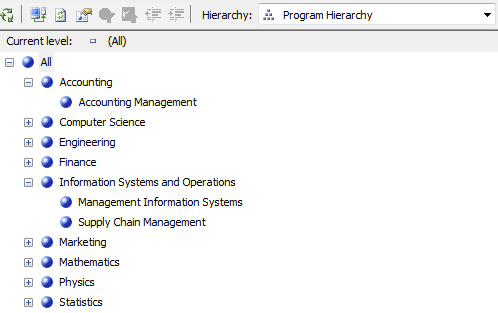
The above report gives a holistic view of the number of students who are registering for a particular course during a particular year. This piece of information can be really useful in understanding the preferences of students when it comes to course selection. The university will be able to allocate greater resources and personnel to courses that are more in demand. Besides, based on the preferences of students, the University can achieve a better academic planning.

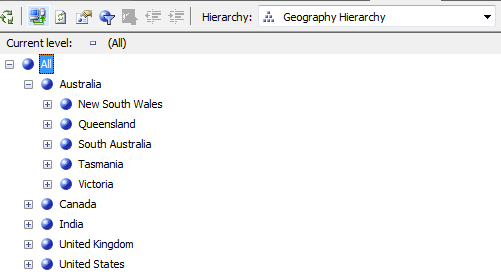
### Placements:

#### Cube Structure:



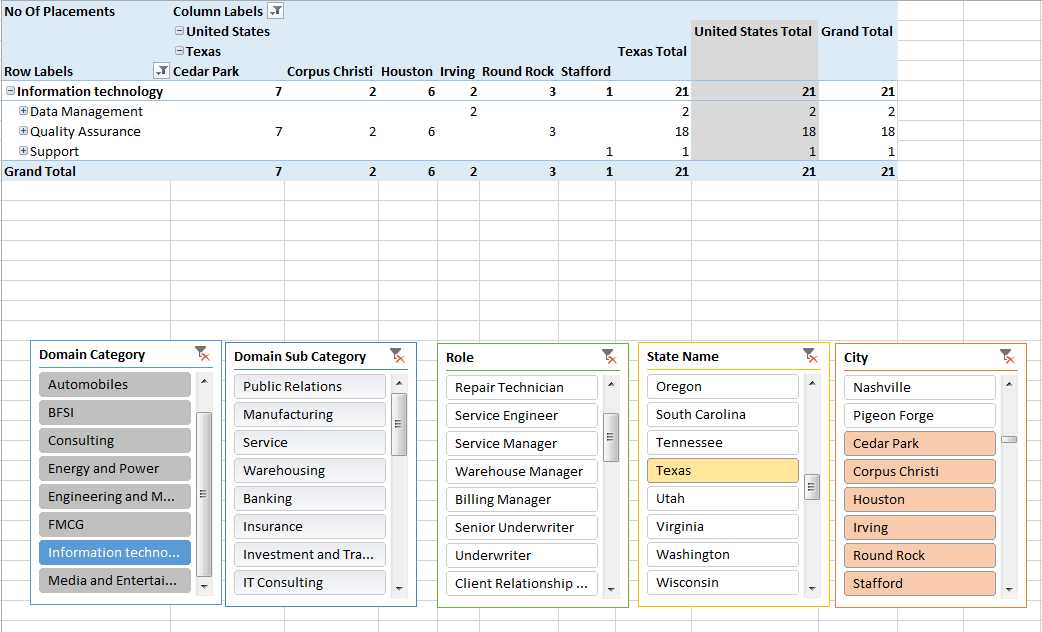
#### Hierarchies:

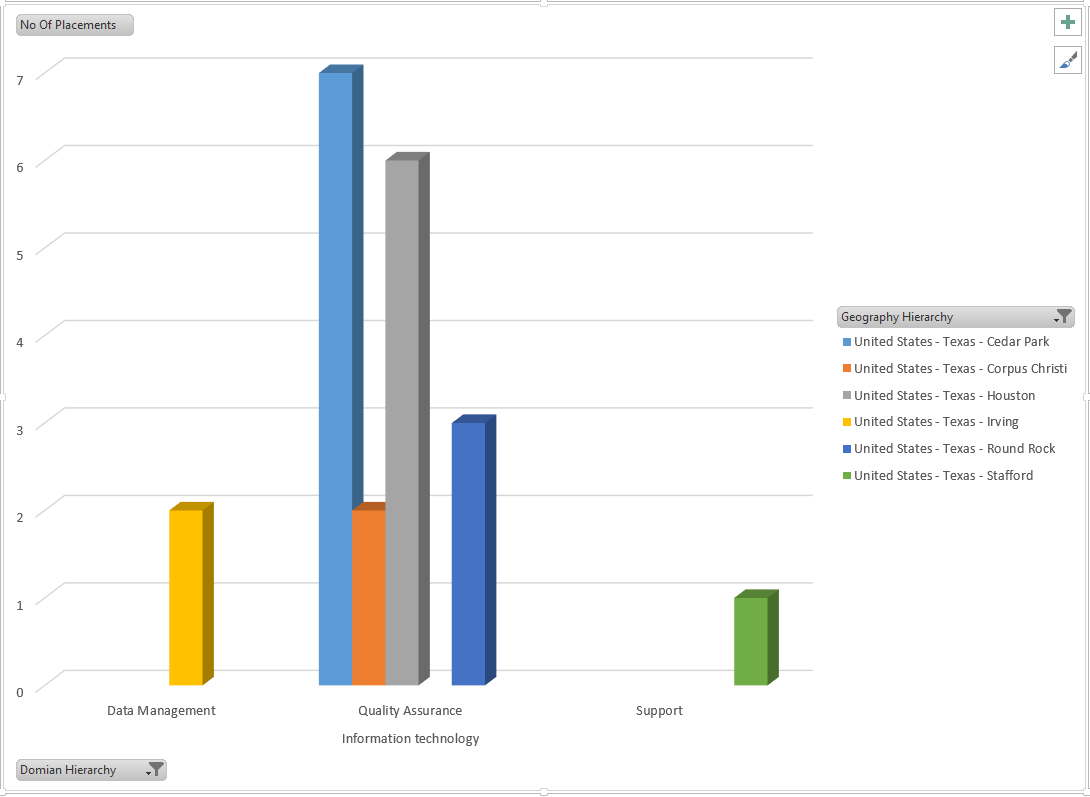
Program Hierarchy

Geography Hierarchy

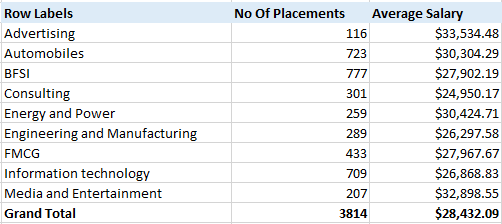
#### Reports:

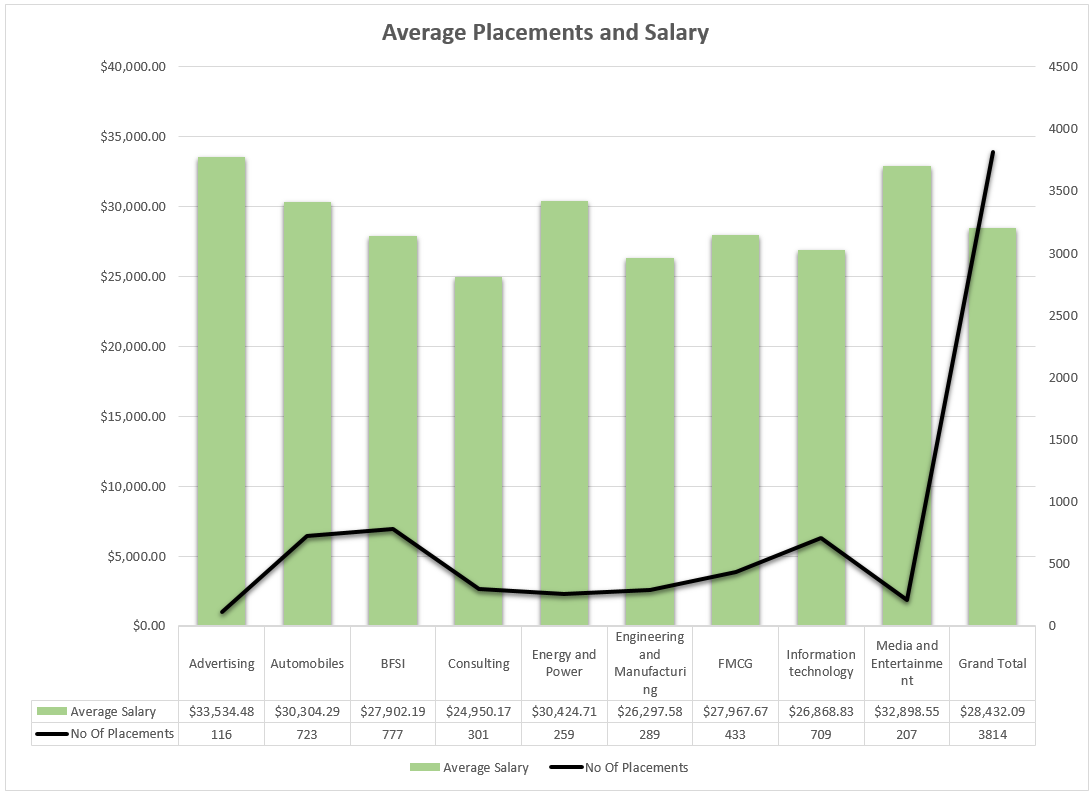
**Business Question:** Find the numberplacements in the field of **Information Technology** in the state of **Texas**





**Business Question:** Find and compare the average salary based on placements in each industry domain





The above reports give an idea on the trends of placements. The university can also compare the placements and employment opportunities in various domain areas and accordingly improve their career guidance cell. Also these reports are also helpful for the analysis at each department level, for example, the number of placements and the average salary in the field of Information Technology is illustrated in the first graph. This graph provides insights into geographical locations and domain etc. This information can turn out to be a very valuable piece of information for prospective students as well.

# Appendix

## Infrastructure:

The fundamental purpose of a data warehouse infrastructure is to support the overlying architectural components. In other words, the infrastructure includes all the foundational elements that enable the data warehouse architecture to be implemented. For example, elements such as server hardware, operating system, network software, LAN and WAN, people, procedures and training are all part of infrastructure.

In a broad classification the elements of a data warehouse infrastructure can be divided into two major categories: operational infrastructure and physical infrastructure.

Operational infrastructure consists of people, procedures, training and management software

Physical infrastructure has components such as Hardware, Operating Systems, and DBMS etc.

The following section will list in detail the infrastructure used for each of architectural components in our project.

**Operational Infrastructure:**

**People:**

Project Team and Internal Information Technology team of Brazos University

**Management Software:**

Microsoft Project to schedule, monitor and administer data transformation tasks.

**Training:**

Scheduled trainings between Project Team, Brazos University IT Team and its Executive Management

**Physical Infrastructure**

**For Data Acquisition and Staging Tools:**

* **Hardware:** Brazos University Operational Databases
* **Software:** Microsoft Excel, Microsoft Access

**For Data Storage:**

* **Server:** SQL Server 2012 Enterprise Edition
* **Location (Network)**: (INFODATA.TAMU.EDU) **Server type:** Database Engine

**For Information Delivery through OLAP:**

* + - **Hardware**: SQL Server 2012 Analytical Services Server at **INFODATA.TAMU.EDU**
    - **Software:**
      * **SQL Server Data Tools -** Microsoft Visual Studio 2010 Shell (Business Intelligence).
      * Microsoft Excel and PowerPivot

## Metadata:

|  |  |  |
| --- | --- | --- |
| Metadata | | |
| **Attribute** | **DataType** | **Description** |
| TimeKey | int | Surrogate key for time dimension |
| DayNumberOfWeek | int | Denotes the day number for a week |
| DayNameOfWeek | varchar(50) | Denotes the name of the day |
| DayNumberOfMonth | int | Denotes the day number for a month |
| DayNumberOfYear | int | Denotes the day name of the month |
| WeekNumberOfYear | int | Denotes the week number in a year |
| MonthName | varchar(50) | Denotes the name of the month |
| MonthNumberOfYear | int | Denotes the month number for a year |
| CalendarYear | int | Denotes the year |
| CalendarSemester | int | Denotes the number of the semester (1= Fall, 2=Spring, 3=Summer) |
| SemesterName | varchar(50) | Denotes the name of the semester |
| DegreeKey | int | Surrogate key for degree dimension |
| DegreeName | varchar(50) | Name of the degree |
| MinDegreeDuration | int | Minimum duration of the degree |
| MaxDegreeDuration | int | Maximum duration of the degree |
| DepartmentKey | int | Surrogate key for department dimension |
| DepartmentName | varchar(50) | Name of the department |
| DepartmentHead | varchar(50) | Name of the head of department |
| DepartmentCode | varchar(50) | Department code for a department |
| DomainKey | int | Surrogate key for domain dimension |
| DomainCategory | varchar(50) | Industry category of domain (jobs) |
| DomainSubCategory | varchar(50) | Industry sub-category of domain (jobs) |
| Role | varchar(50) | Role of the job attained by student |
| FacilityKey | int | Surrogate key for facility dimension |
| FacilityName | varchar(50) | Name of the facility |
| FacilityLocation | varchar(50) | Name of the place where facility is located |
| FacilityType | varchar(50) | Type of facility |
| FacilityCapacity | int | Maximum allowed people in a facility |
| FacultyKey | int | Surrogate key for Faculty dimension |
| CourseKey | int | Surrogate key for Course dimension |
| FacultyFirstName | varchar(50) | First Name of the faculty |
| FacultyLastName | varchar(50) | Last Name of the faculty |
| Position | varchar(50) | Designation of the faculty |
| Experience | int | No of years of experience of faculty |
| FacultyEmail | varchar(50) | Email Address of the faculty |
| FacultyPhone | varchar(50) | Phone number of the faculty |
| GeographyKey | int | Surrogate key for geography dimension |
| City | varchar(50) | Name of the city |
| StateProvinceCode | varchar(50) | State code of the state |
| StateProvinceName | varchar(50) | Name of the state |
| CountryRegionCode | varchar(50) | Country code |
| CountryRegionName | varchar(50) | Name of the country |
| PostalCode | varchar(50) | Zip code of the location |
| AmountOfScholarship | money | Amount of money given as scholarship |
| AmountOfFederalLoan | money | Amount of money given as federal loan |
| AmountOfInstateWaiver | money | Amount of money given as instatewaiver |
| NumberOfStudents | int | Number of students attending an event |
| NumberOfGuests | int | Number of guests attending an event |
| Revenue | money | Total amount of money earned |
| TotalSalary | money | Total amount of salary earned |
| NoOfPlacements | int | Total numbe of placements |
| RelocationBonus | money | Amount of relocation bonus given |
| TotalTuition | money | Total amount of tuition received |
| CourseId | varchar(50) | Course ID of a course |
| CourseName | varchar(50) | Name of the course |
| CourseCredits | int | Number of credits of a course |
| CourseMaxSize | int | Maximum size of the course |
| CourseFee | money | Fees of a course |
| ApplicantKey | int | Surrogate key of applicant |
| FirstName | varchar(50) | First Name of the applicant |
| MiddleName | varchar(50) | Middle Name of the applicant |
| LastName | varchar(50) | Last Name of the applicant |
| BirthDate | date | Date of birth of applicant |
| MaritalStatus | nchar(1) | Marital status of applicant |
| Gender | varchar(50) | Gender of applicant |
| EmailAddress | varchar(50) | Email address of applicant |
| Education | varchar(50) | Education qualification of applicant |
| Address | varchar(50) | Street Address of applicant |
| Phone | varchar(50) | Phone number of the applicant |
| IsAdmitted | nchar(1) | 1(Admitted) or 0 (Not Admitted) - Indicates if applicant admitted or not |
| IsEnrolled | nchar(1) | 1(Enrolled) or 0 (Not Enrolled) - Indicates if applicant enrolled or not |
| GRE | int | GRE score of applicant |
| GMAT | int | GMAT score of applicant |
| TOEFL | int | TOEFL score of applicant |
| SAT | int | SAT score of applicant |

## Estimate of Size of DW:

The data warehouse dimensions have been classified into 3 types:

* **Small**: This kind of dimensional table usually has around 10000 historical records. This dimension is usually a slowly changing dimension with type 1 changes, which corresponds to errors. As a result there is not much change in the dimension table row
* **Intermediate**: This kind of dimensional table usually has around 100000 historical records. This dimension is usually a slowly changing dimension with type 2 changes, which corresponds to preservation of history. As a result there are few changes in the dimension table rows.
* **Large**: This kind of dimensional table usually has around 1000000 historical records. This dimension is usually a slowly changing dimension with type 2 changes, which corresponds to frequent revisions. As a result there are frequent changes in the dimension table rows as new records are added every time there is a revision change.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table Name** | **Type** | **Record Size** | **No of records** | **Total Table Size (In GB)** |
| Dim\_Applicant | Large | 438 | 10000000 | 4.38 |
| Dim\_Course | Medium | 124 | 100000 | 0.0124 |
| Dim\_Date | Large | 182 | 10000000 | 1.82 |
| Dim\_Degree | Small | 62 | 10000 | 0.00062 |
| Dim\_Department | Small | 154 | 10000 | 0.00154 |
| Dim\_Domain | Small | 154 | 10000 | 0.00154 |
| Dim\_Facility | Small | 158 | 10000 | 0.00158 |
| Dim\_Faculty | Medium | 262 | 100000 | 0.0262 |
| Dim\_Geography | Large | 304 | 10000000 | 3.04 |
| Dim\_Program | Small | 108 | 10000 | 0.00108 |
|  |  |  | Total Size | 9.28496 |

The source data of the data warehouse is present in several different formats and need to be transformed and loaded onto these tables.

#### Fact Table Size Estimation:

There are 4 fact tables that have measures and they contribute towards the size of the DW.

The total size of all dimension tables: 9.28 GB

Assume that 5-8% (lower level) of this data would be present in fact tables after the required conditions and rules have been met.

5% of 9.28GB = 0.464GB

#### Total Size:

0.464GB is the storage required for 1 day. Since records would be added daily, the size for 365 days is 0.464\*365 = 169.36GB

For a 10 year period of this project, the estimated storage would be 169.36\*10 = 1693.6 GB

## Peer Evaluation Forms:

The peer evaluation forms of all team members have been attached at the end of this document.

# References:

1) Data Warehouse Fundamentals by Paulraj Ponniah

2) Estimate Size of DW - http://www.exactsoftware.com/docs/DocView.aspx?DocumentID=%7B3d5ea053-4aae-4bb8-80fc-d79ceb05a98d%7D

3) Microsoft SSAS Tutorials - <https://msdn.microsoft.com/en-us/library/hh231701.aspx>

4) Why you need a Data Warehouse? - <http://datalyticstechnologies.com/wp-content/uploads/2014/01/2013-03-Why-You-Need-a-Data-Warehouse.pdf>

5) Excel filtering and removing duplicates - https://support.office.com/en-au/article/Filter-for-unique-values-or-remove-duplicate-values-ccf664b0-81d6-449b-bbe1-8daaec1e83c2