ISYS 637 - Data Warehouse - Group Project Report

SECTION 603 - GROUP 7 – Brazos University

SAIRAM JANAKIRAMAN (924001504)

BHARADWAJ VENKATESHWARAN (823009748)

RATNADEEP SIMHADRI (424000610)

DINESH KAKARLA MATHAM (723008494)

COMPLETED ON 4th May 2015

Contents

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:

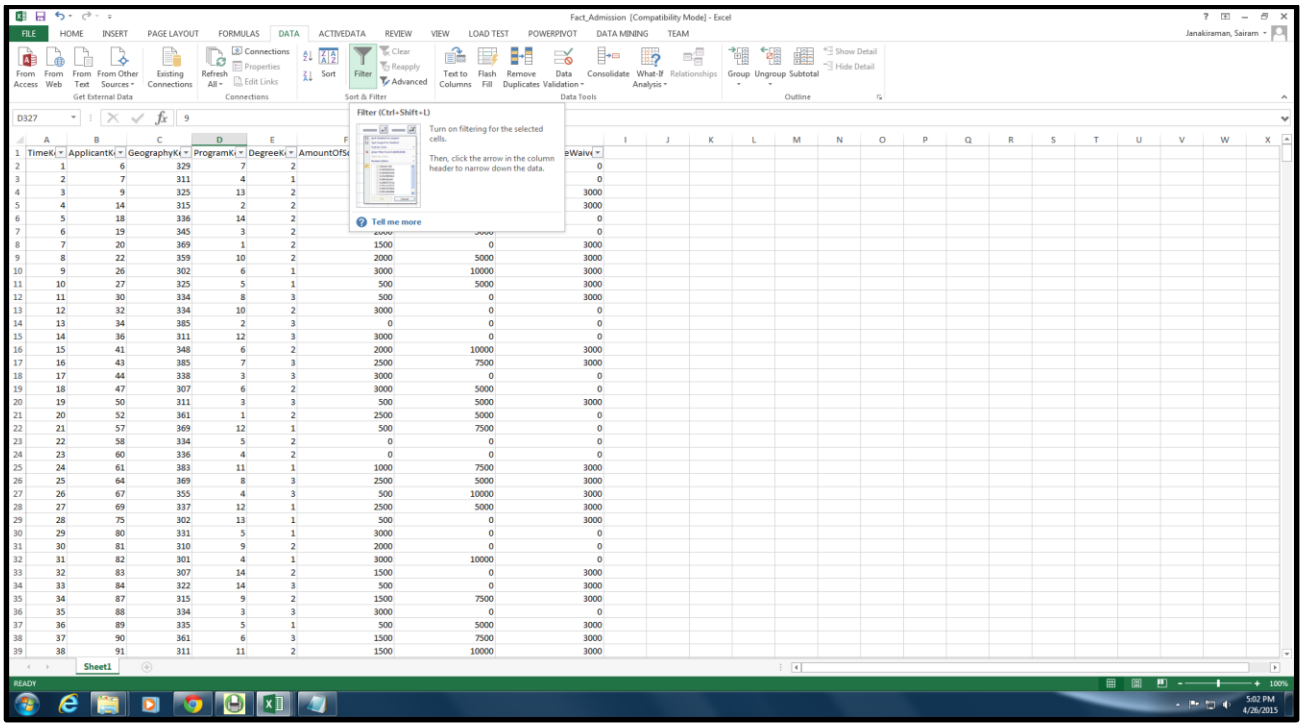
In general, different data sources feed data into a university data warehouse. Applicant data could be from operational data sources while faculty data could be from internal data sources such as files, spreadsheets or documents. While implementing the prototype for Brazos University Data Warehouse, 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.

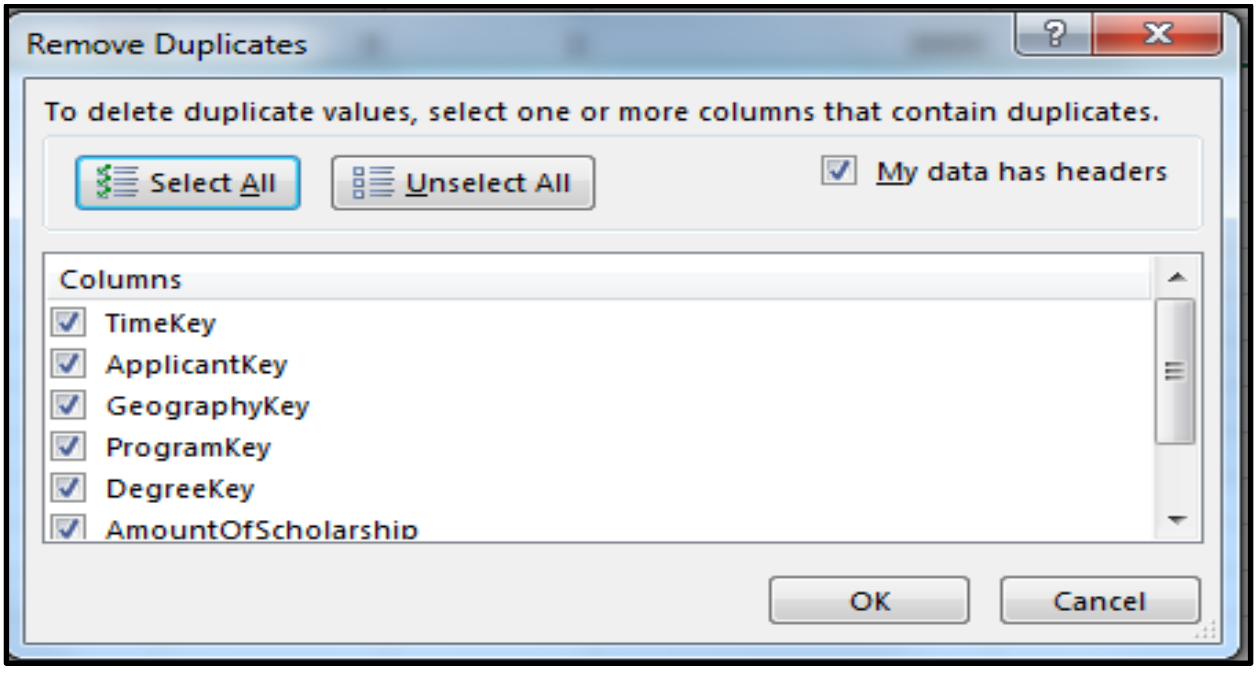
Data Staging:

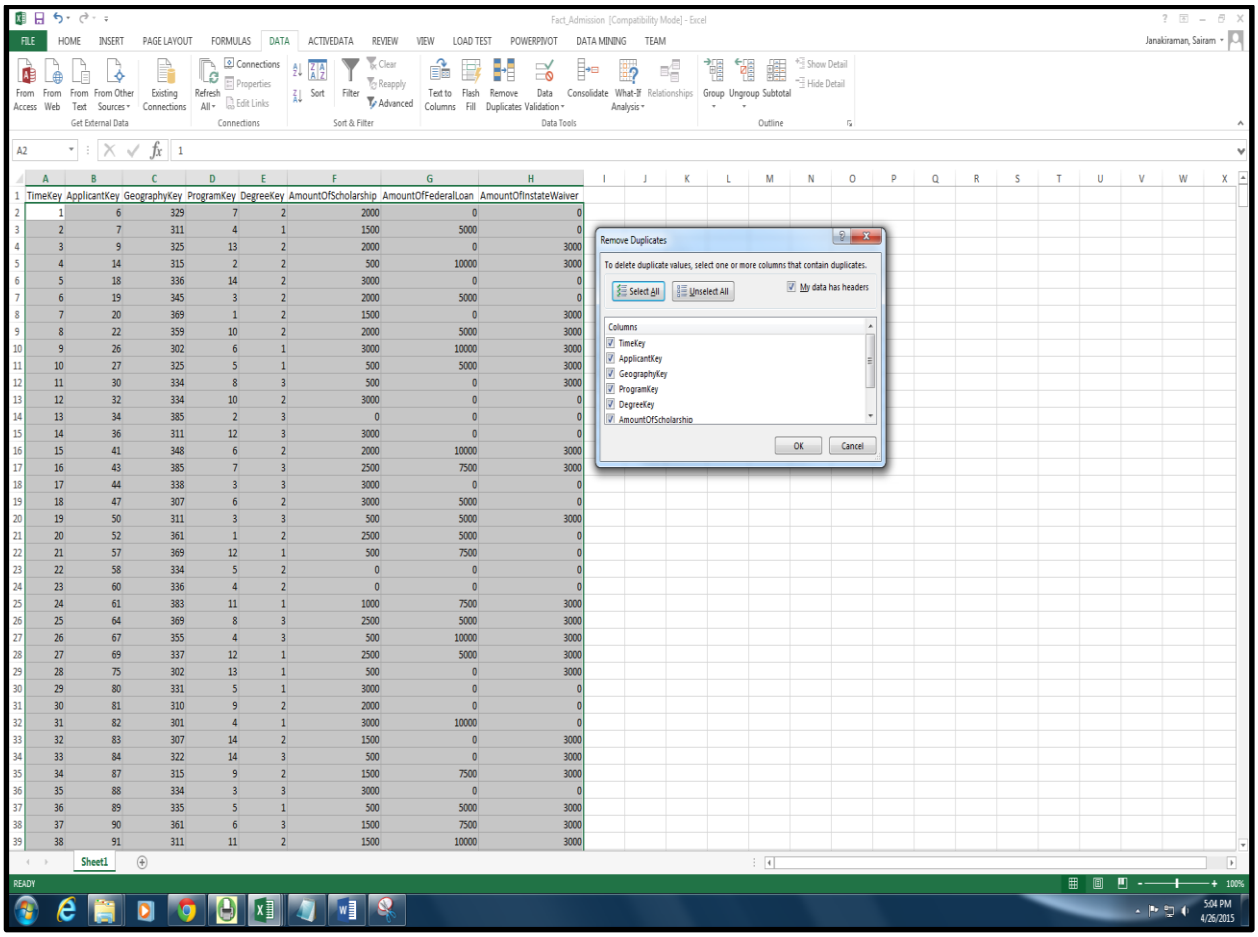
After the data sources were identified, a set of ETL processes were performed to enable the data to be loaded into the data warehouse. Data was extracted from the source system and stored in excel spreadsheets. While following the extraction process, business rules were followed to make sure that the 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.

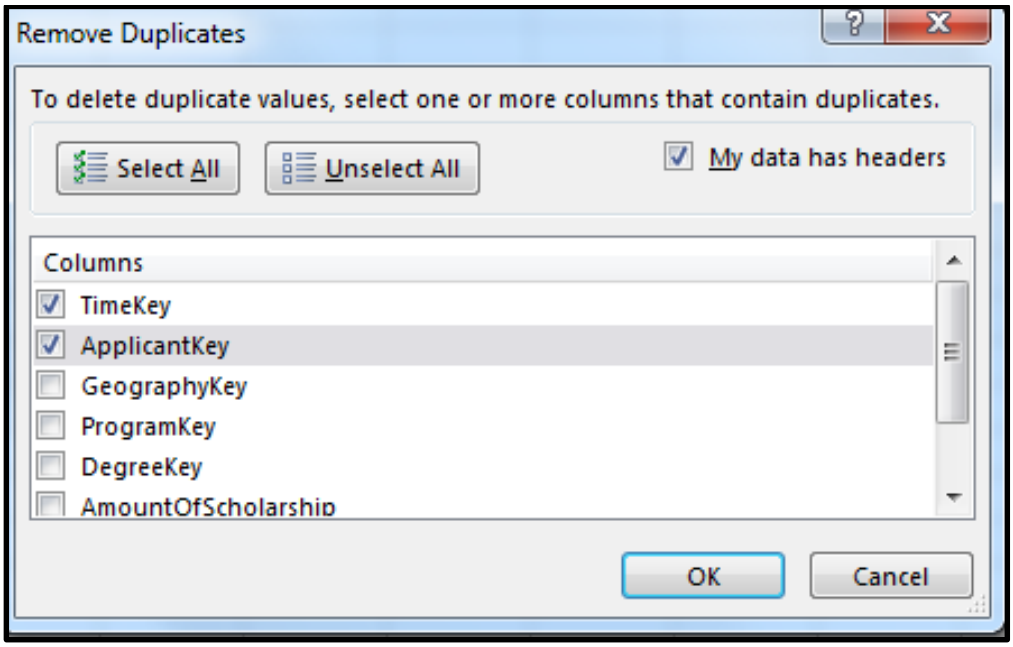


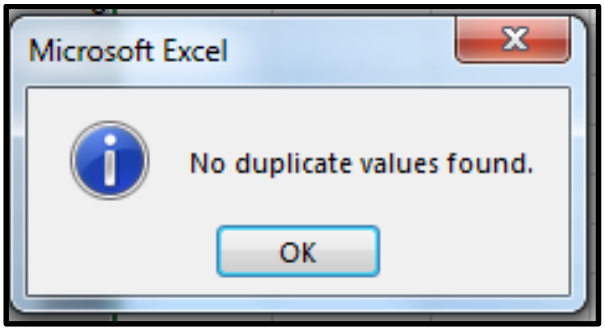
Once all the data were collected, the data was checked for duplicates and missing value using Excel's remove duplicate functionality. The sequence of diagrams below illustrate the data staging process followed for Brazos University Data Warehouse





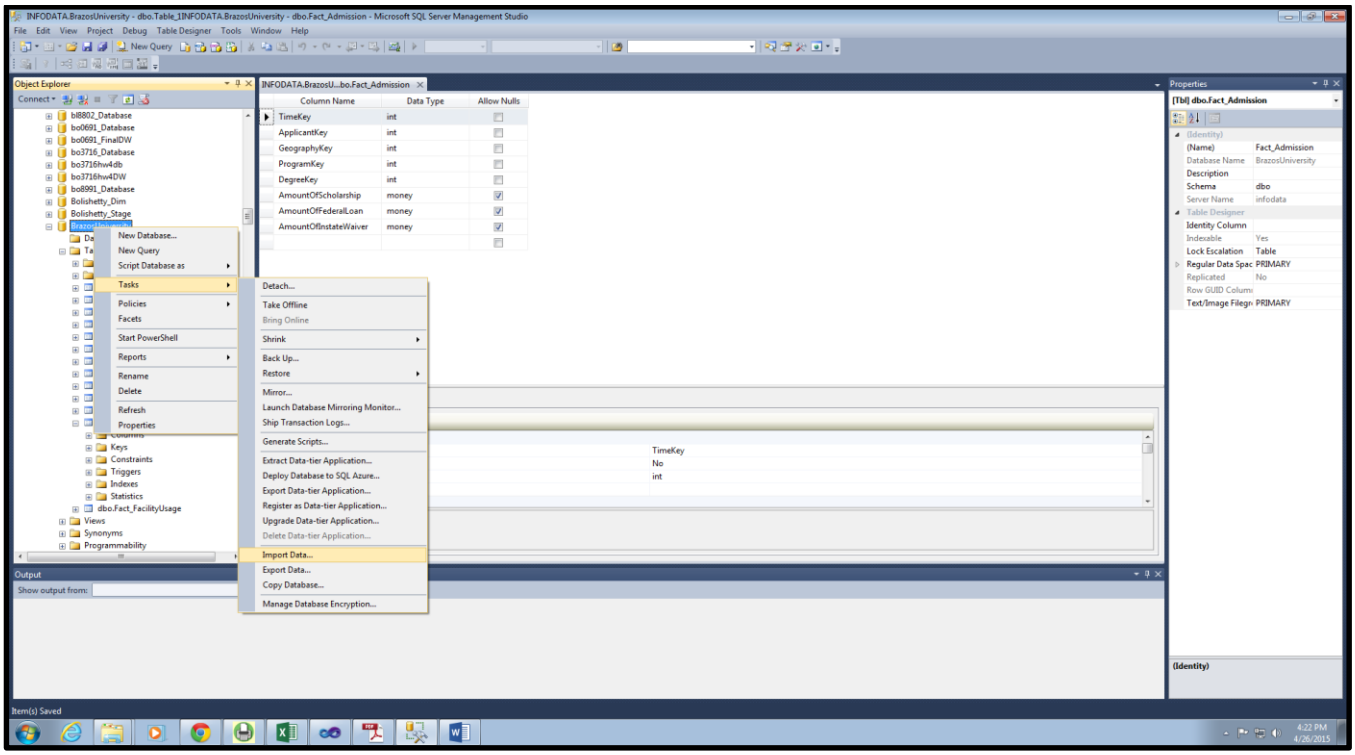
	A	B	C	D	E	F	G	H
1	TimeKey	ApplicantKey	GeographyKey	ProgramKey	DegreeKey	AmountOfScholarship	AmountOfFederalLoan	AmountOfInstateWaiver
297	0	1453	510	2	1	0	10000	3000
309	0	1486	180	5	2	1000	0	0
314	0	1492	262	2	2	500	0	0
319	0	1501	172	5	2	500	5000	0
321	0	1505	239	3	1	3000	0	3000



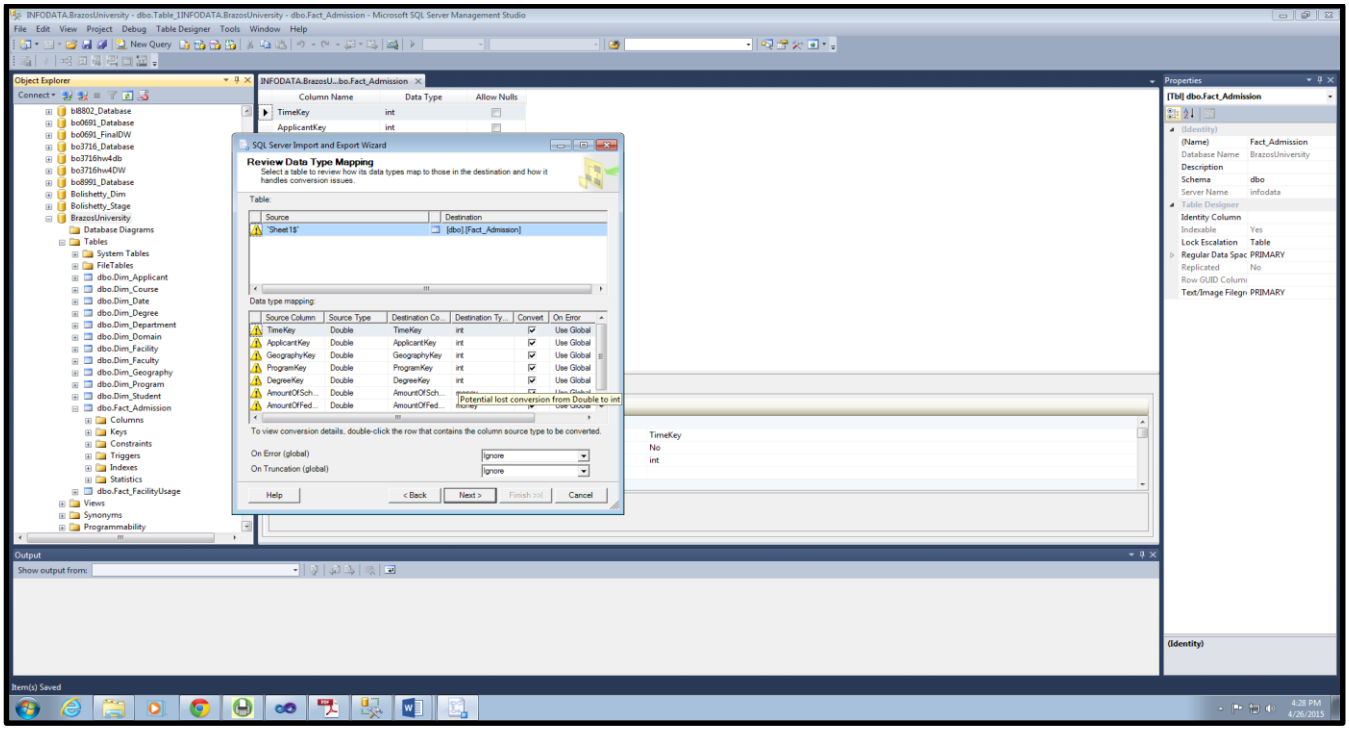


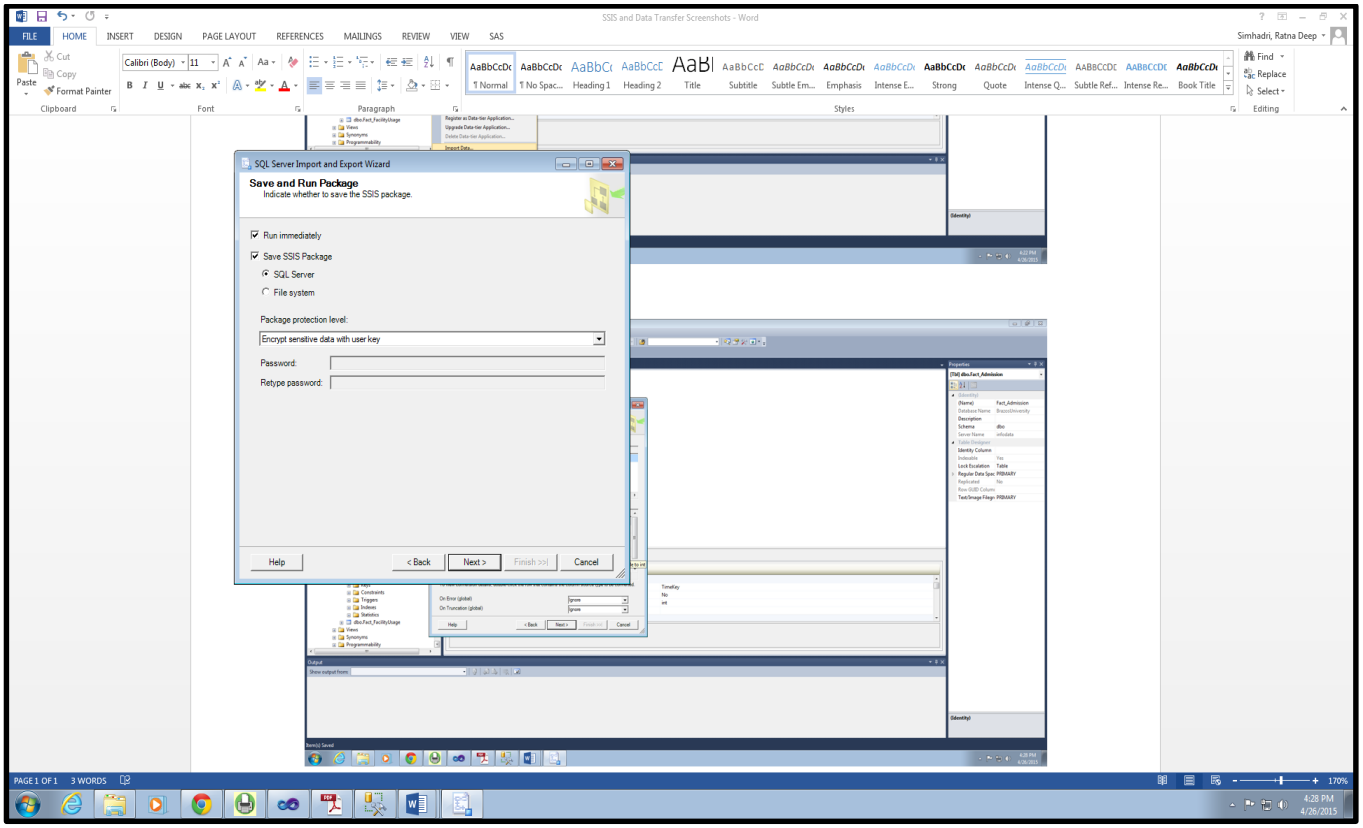
Data Storage:

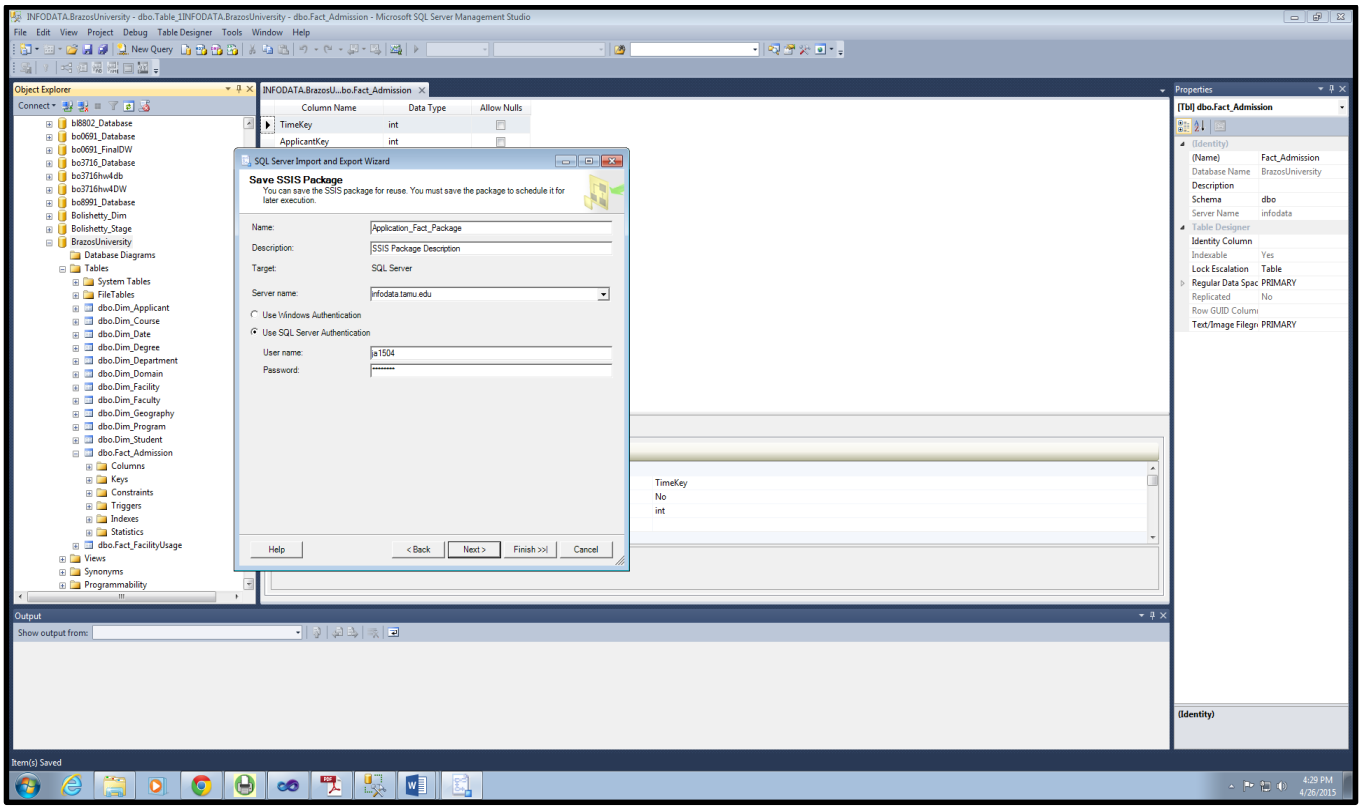
Once the transformation process is complete, duplicates of each data file is created to provide backup and recovery options if needed. Before loading the data into the data warehouse any remaining primary and foreign key referential are identified and resolved. The data sets are created and consolidated 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.



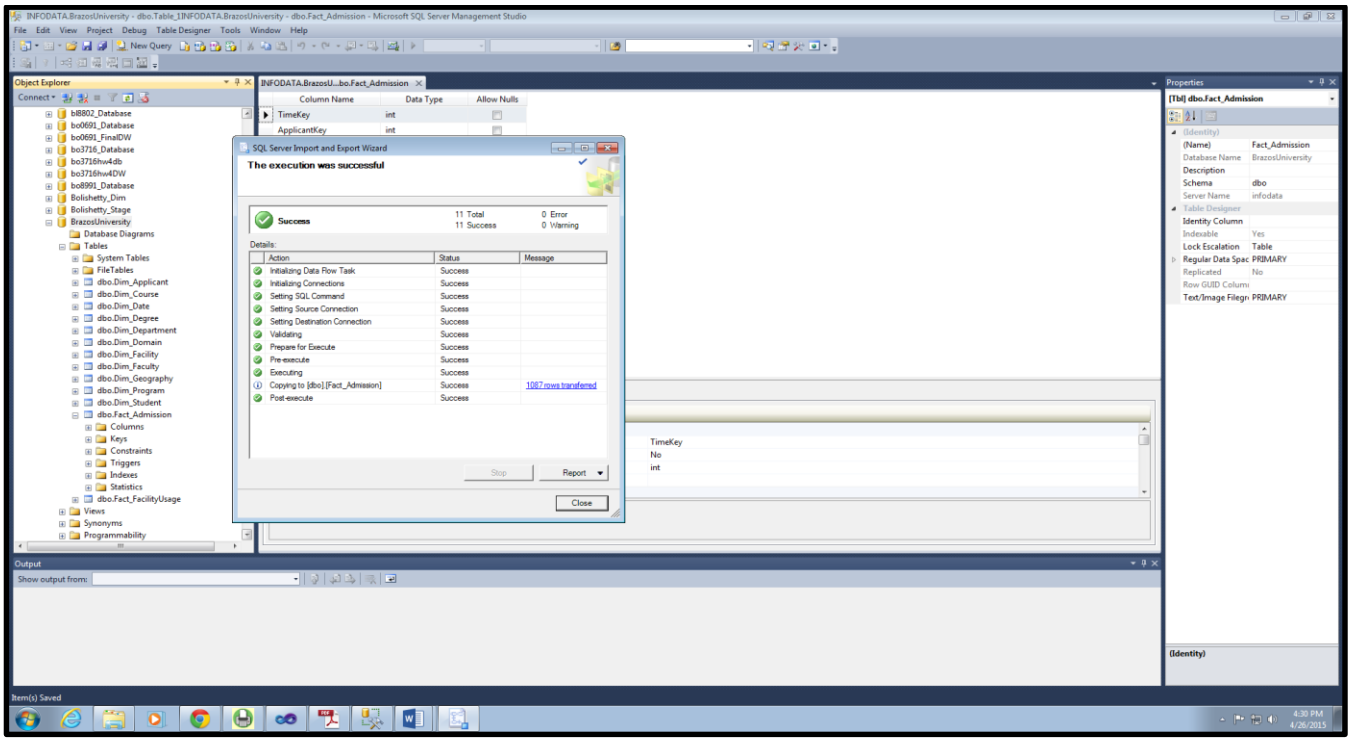
After importing the files, for each table a data type mapping is performed. This helps in defining the data types for each attribute of a table as well as assign foreign key referential for a particular attribute. The tasks option under the Data Warehouse is used to import data into the selected table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




SSIS and Data Transfer Screenshots - Word
SQL Server Import and Export Wizard
Save and Run Package
Indicate whether to save the SSIS package.
Run immediately
Save SSIS Package
SQL Server
File system
Package protection level:
Encrypt sensitive data with user key
Password:
Retype password:
Help
< Back
Next >
Finish >>|
Cancel


INFODATA.BrazosUniversity - dbo.Table_1INFODATA.BrazosUniversity - dbo.Fact_Admission - Microsoft SQL Server Management Studio
SQL Server Import and Export Wizard
Save SSIS Package
You can save the SSIS package for reuse. You must save the package to schedule it for later execution.
Name:
Application_Fact_Package
Description:
SSIS Package Description
Target:
SQL Server
Server name:
infodata.tamu.edu
Use Windows Authentication
Use SQL Server Authentication
User name:
ja1504
Password:
Help
< Back
Next >
Finish >>|
Cancel

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

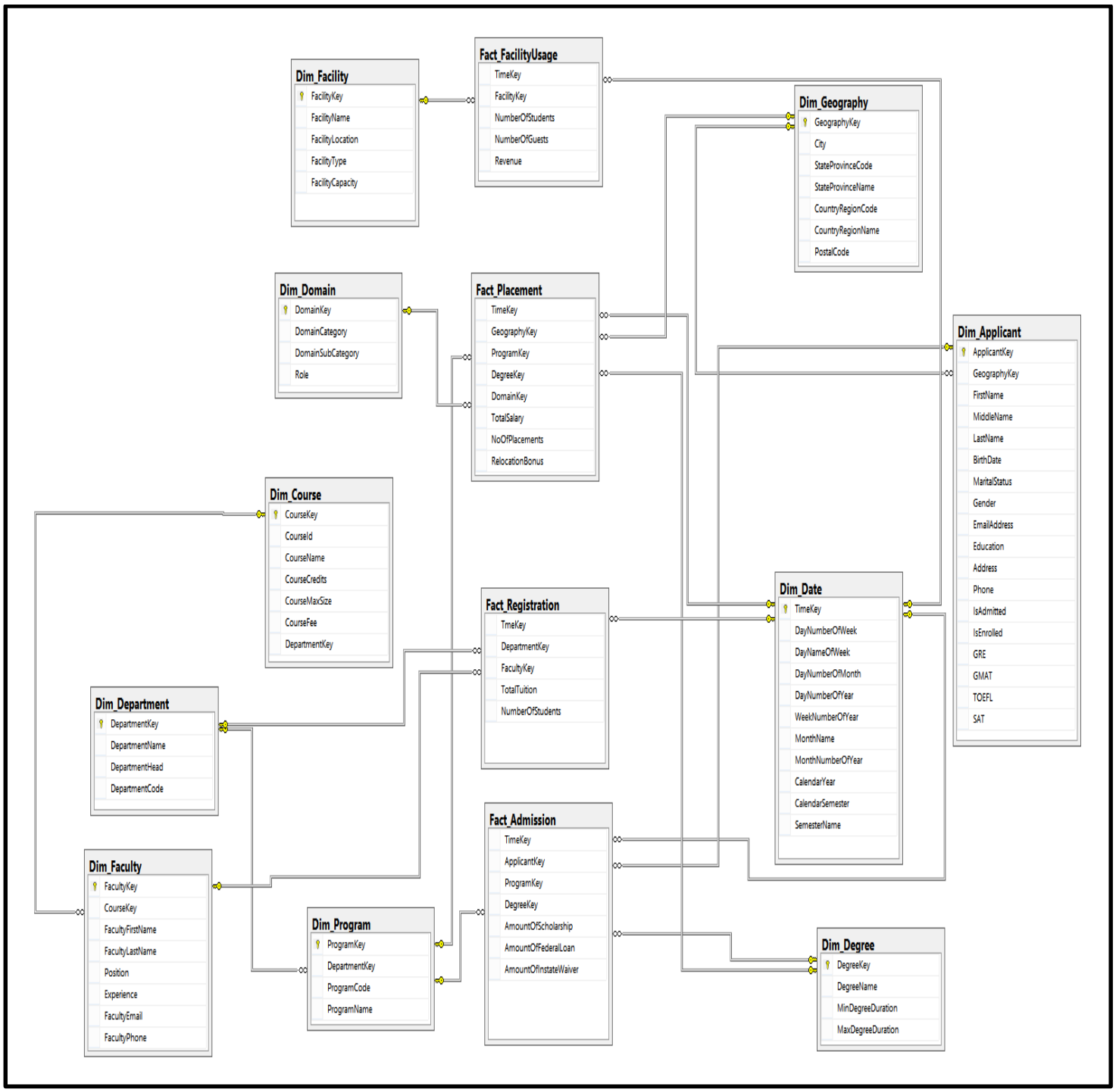


Information Delivery:

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:

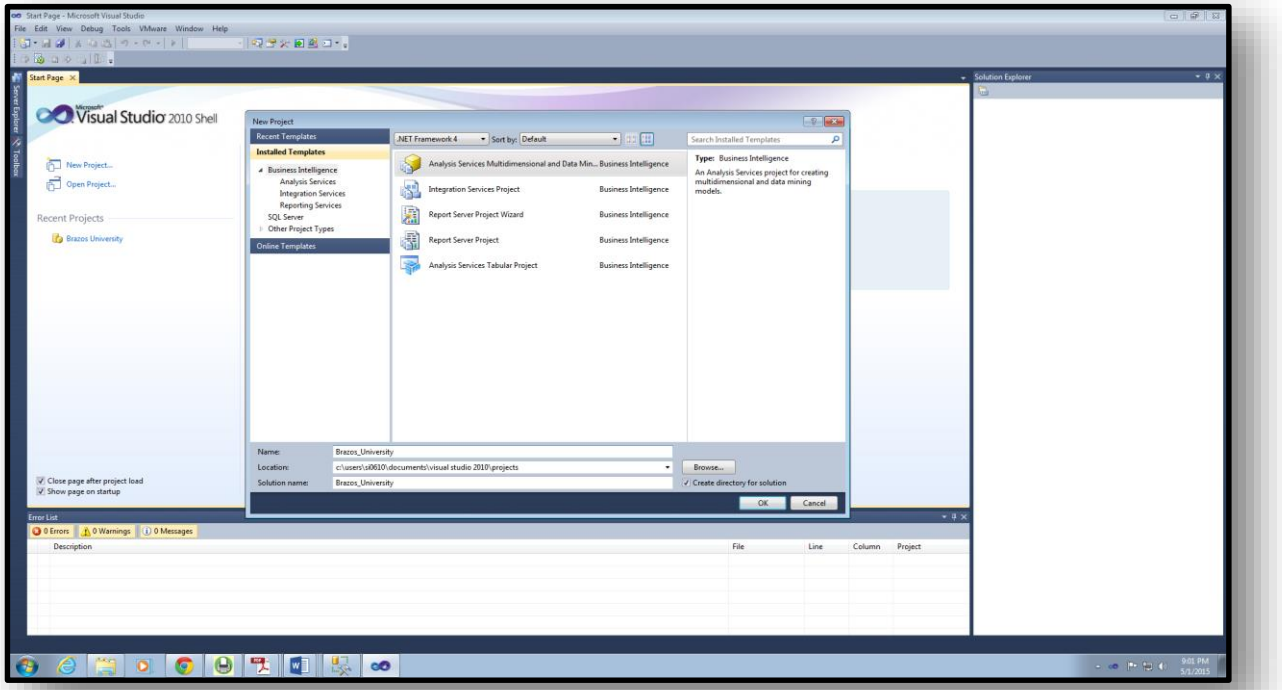
For Brazos University, the business questions under consideration can be categorized into various functional areas as follows. This section illustrates in detail each of the business question and the associated OLAP reports from SSAS.

- Facilities
- Admissions and Finance
- Course Registration and Revenue
- Placements

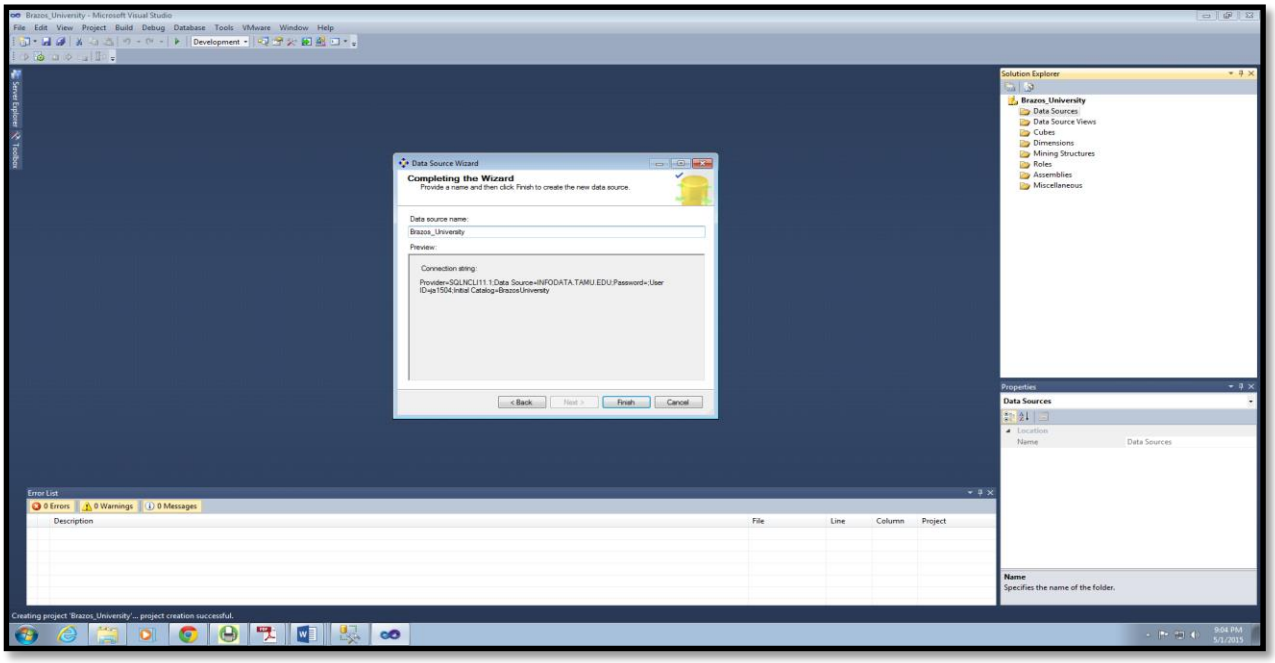
Before going into the details of individual cube and report development. The following section details the creation of Data Source and Data Source View.

Defining Data Source and Data Source View:

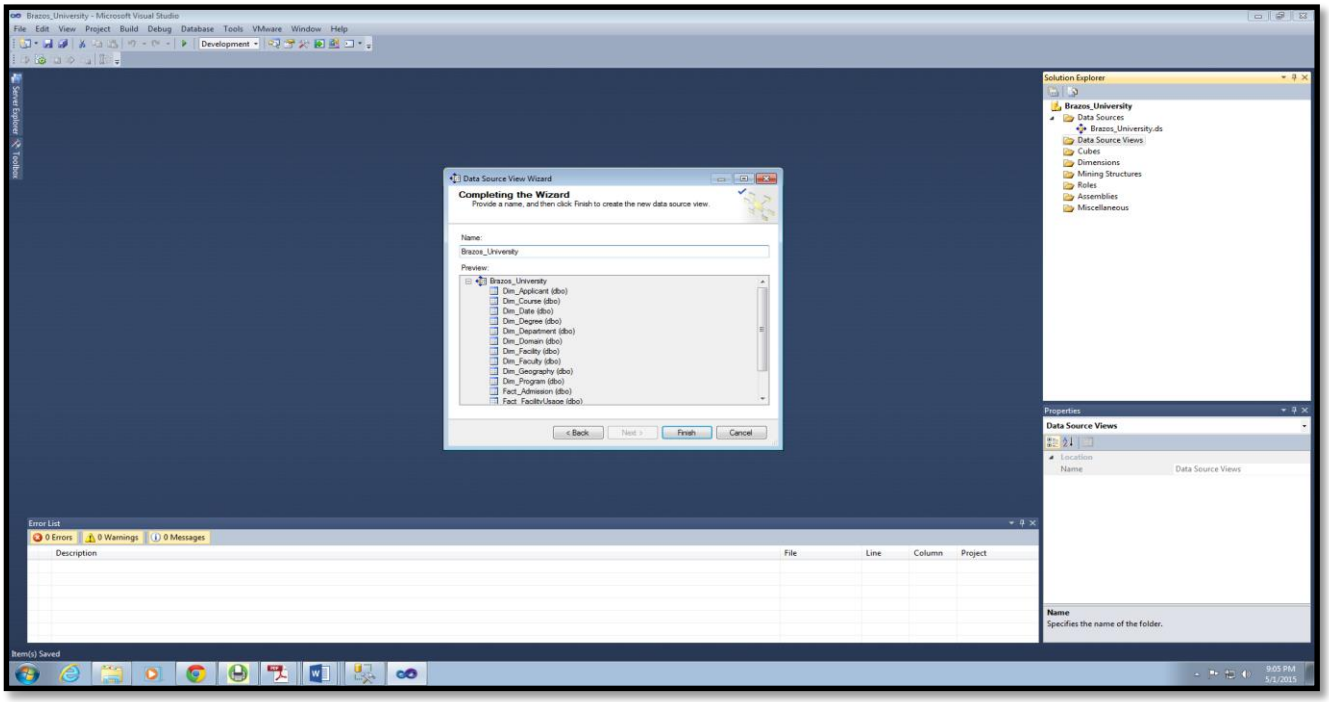
Create a new Project:



Add New Data Source:



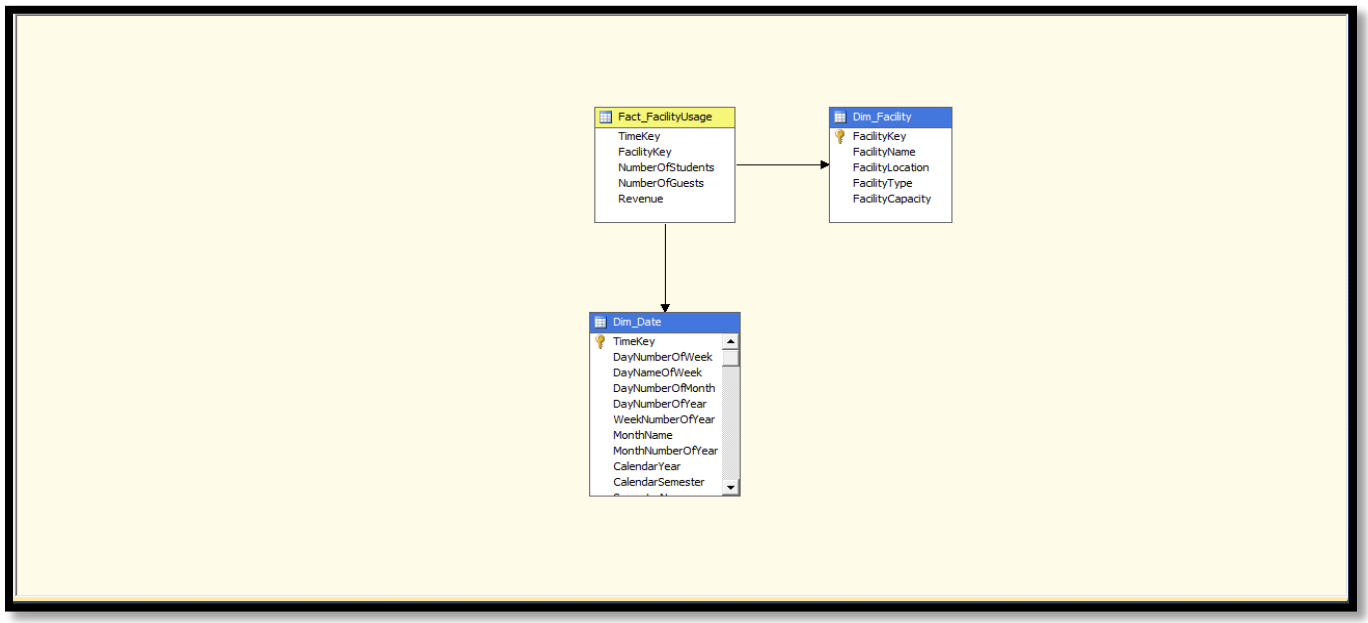
Create Data Source View:



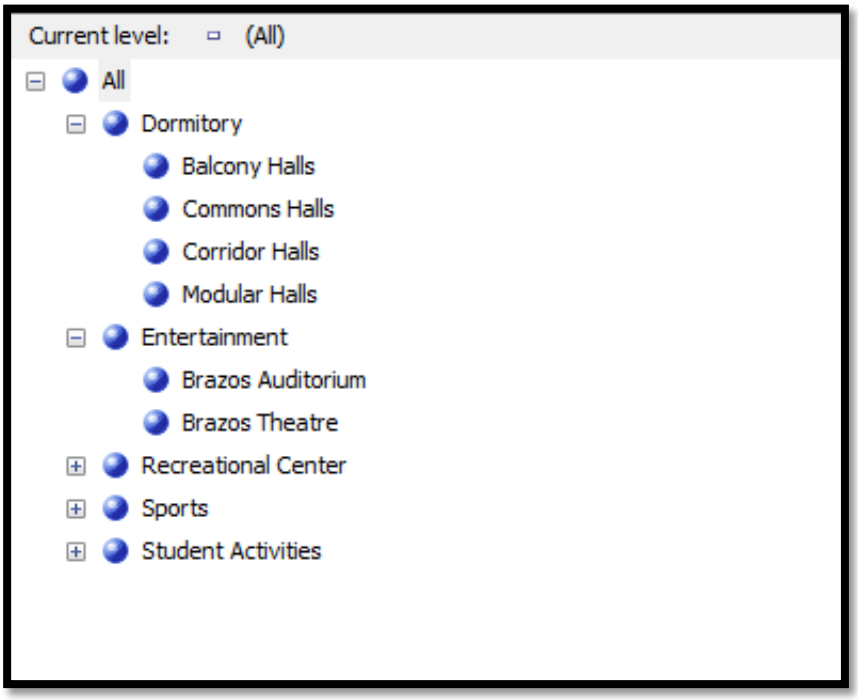
Creating Cubes and Reports:

Facilities:

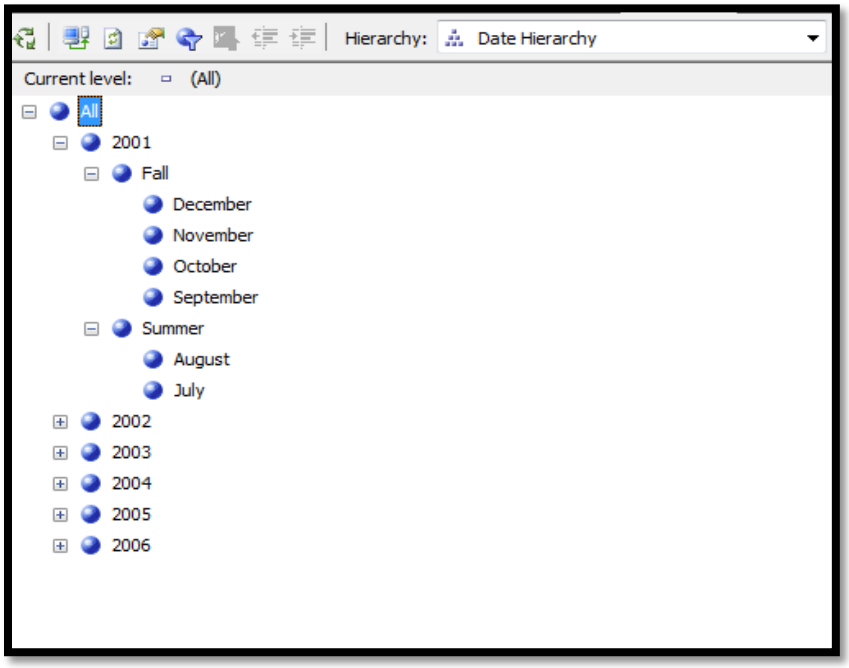
Cube Structure:



Hierarchies:



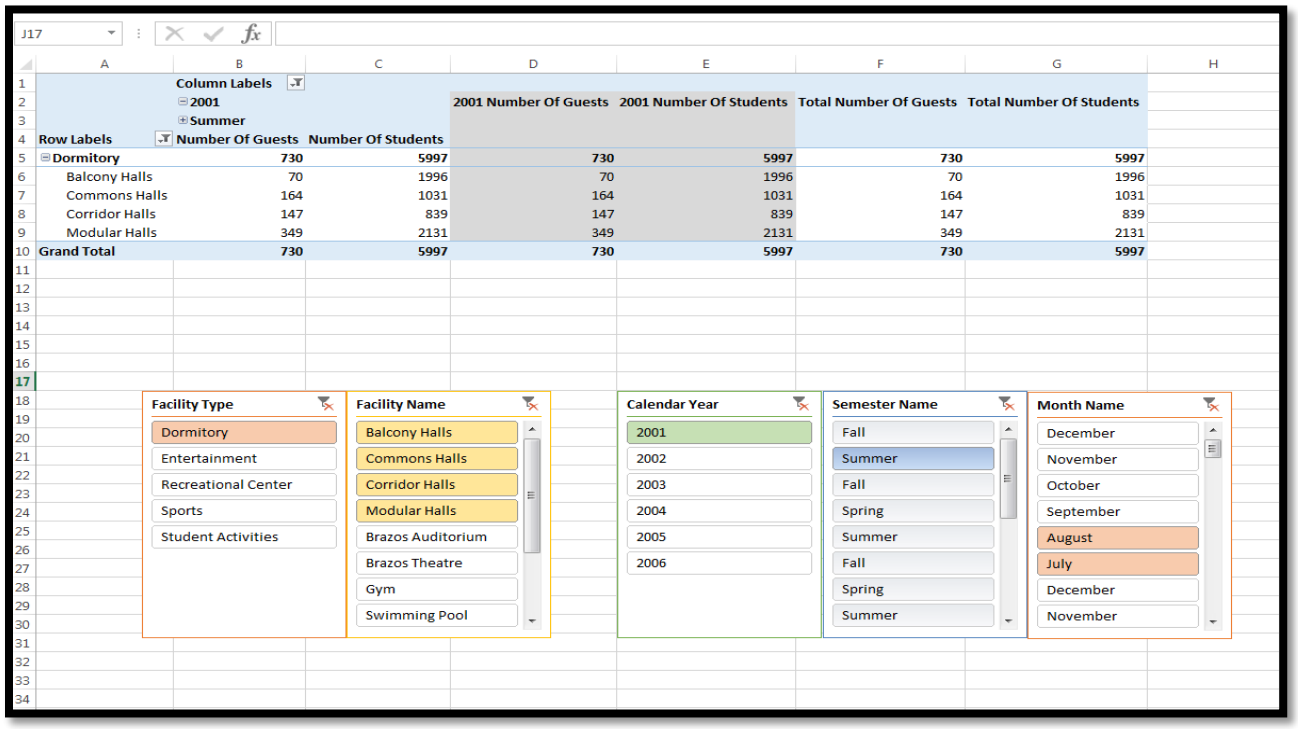
Facility Hierarchy



Date Hierarchy

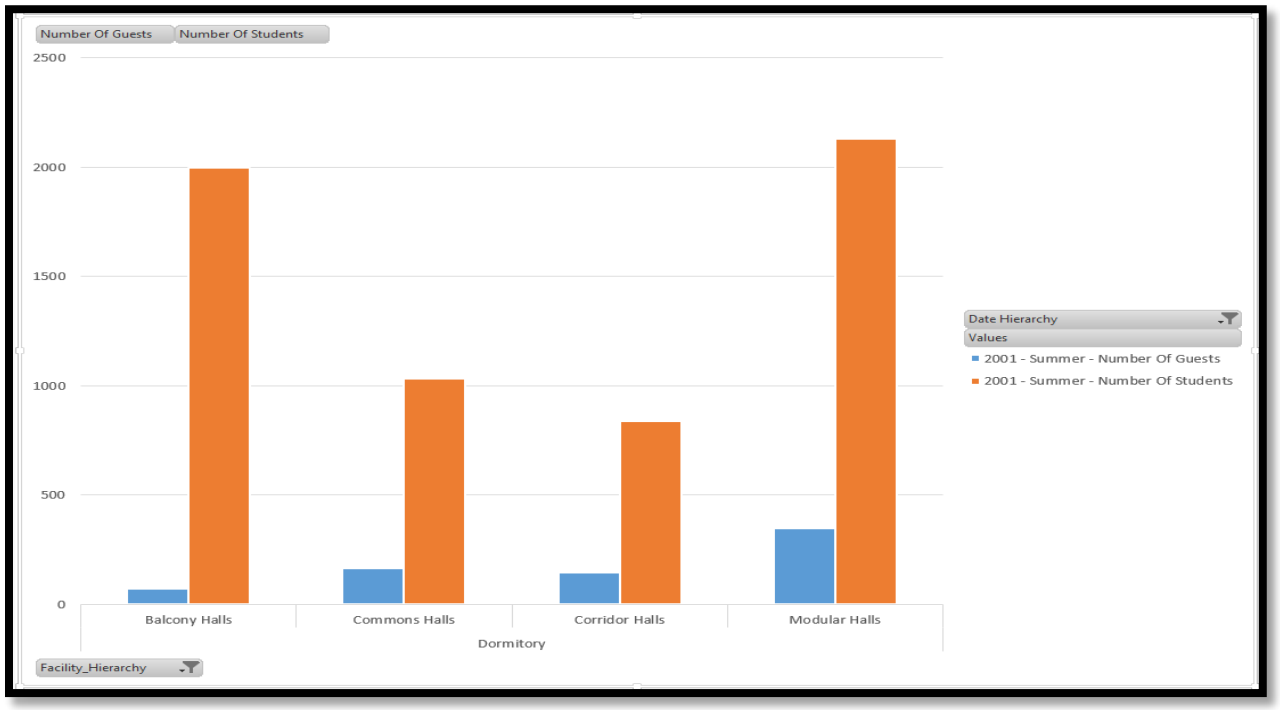
Reports:

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



Row Labels	Number Of Guests	Number Of Students	2001 Number Of Guests	2001 Number Of Students	Total Number Of Guests	Total Number Of Students
Column Labels						
2001						
Summer						
Dormitory	730	5997	730	5997	730	5997
Balcony Halls	70	1996	70	1996	70	1996
Commons Halls	164	1031	164	1031	164	1031
Corridor Halls	147	839	147	839	147	839
Modular Halls	349	2131	349	2131	349	2131
Grand Total	730	5997	730	5997	730	5997

Facility Type	Facility Name	Calendar Year	Semester Name	Month Name
Dormitory	Balcony Halls	2001	Fall	December
Entertainment	Commons Halls	2002	Summer	November
Recreational Center	Corridor Halls	2003	Fall	October
Sports	Modular Halls	2004	Spring	September
Student Activities	Brazos Auditorium	2005	Summer	August
	Brazos Theatre	2006	Fall	July
	Gym		Spring	December
	Swimming Pool		Summer	November



Business Question: Identify the trends in Gym usage for the year 2004

Row Labels	Number Of Guests - Recreational Center	Number Of Students - Recreational Center	Total Number Of Guests	Total Number Of Students
2004	195	1559	195	1559
Fall	87	498	87	498
December	26	181	26	181
November	39	140	39	140
October	13	142	13	142
September	9	35	9	35
Spring	66	617	66	617
April	15	140	15	140
February	14	169	14	169
January	22	115	22	115
March	7	91	7	91
May	8	102	8	102
Summer	42	444	42	444
August	12	90	12	90
July	20	302	20	302
June	10	52	10	52
Grand Total	195	1559	195	1559

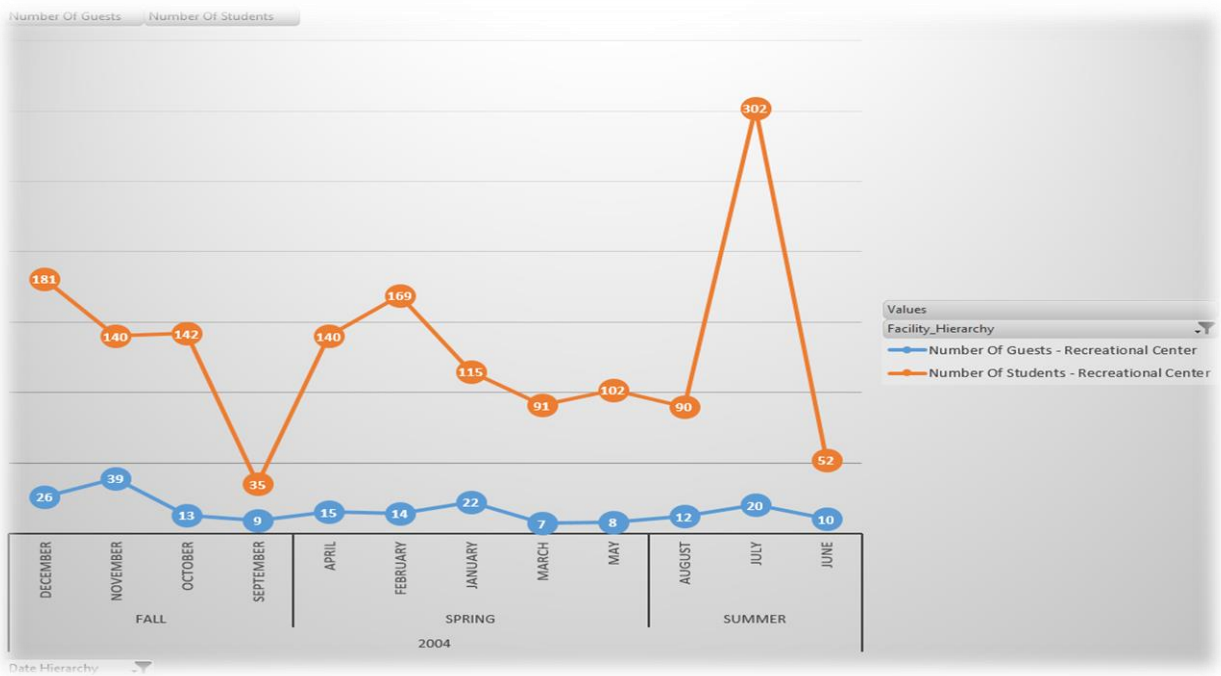
Facility Type: Dormitory, Entertainment, Recreational Center, Sports, Student Activities

Facility Name: Balcony Halls, Commons Halls, Corridor Halls, Modular Halls, Brazos Auditorium, Brazos Theatre, Gym, Swimming Pool

Calendar Year: 2001, 2002, 2003, 2004, 2005, 2006

Semester Name: Fall, Spring, Summer

Month Name: September, April, February, January, March, May, August, July

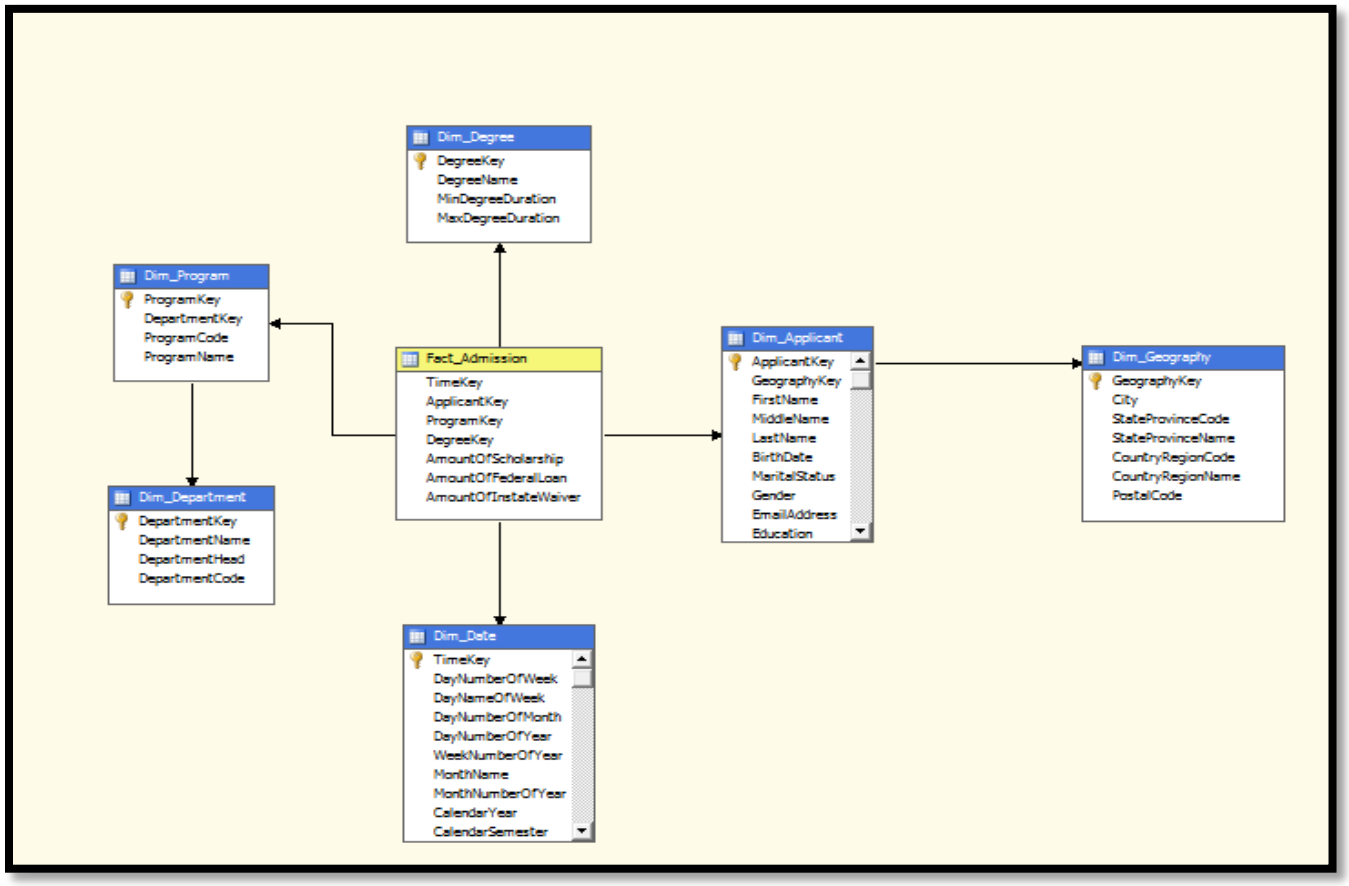


The above reports on facility usage are helpful for the management in understanding the trends in facility usage. Based on this, the university management will be able to make strategic decision on

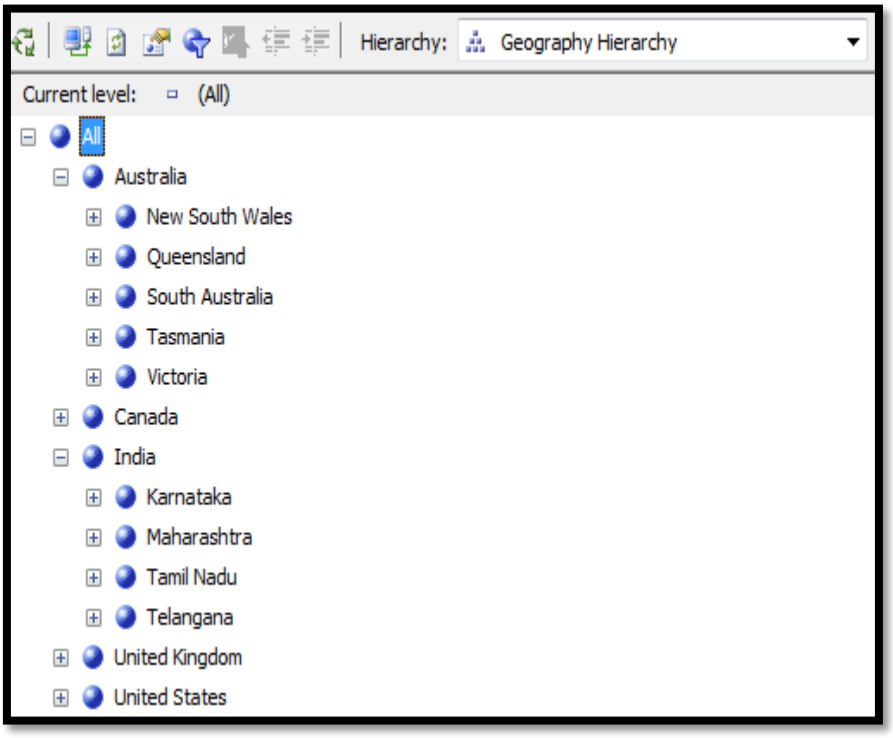
how when and where to improve the infrastructure of the facilities. For example from the first graph it can be observed that though there are greater number of students who prefer to stay in Balcony halls, relatively only a very few number of guests choose to stay at these halls. This insight will help into analyzing the underlying problems, for example facilities, associated with these dormitory halls.

Admissions and Financial aid:

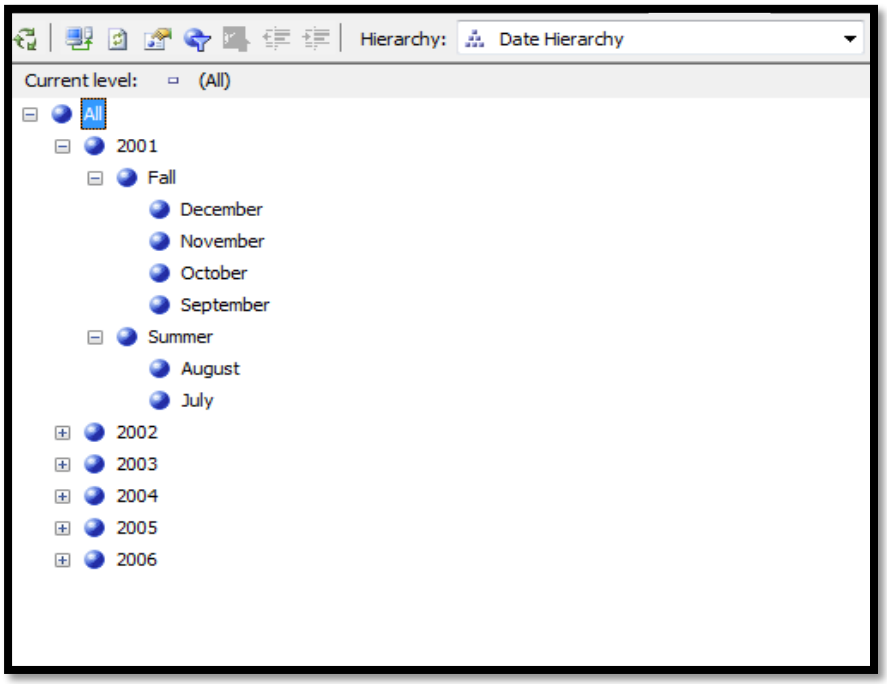
Cube Structure:



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

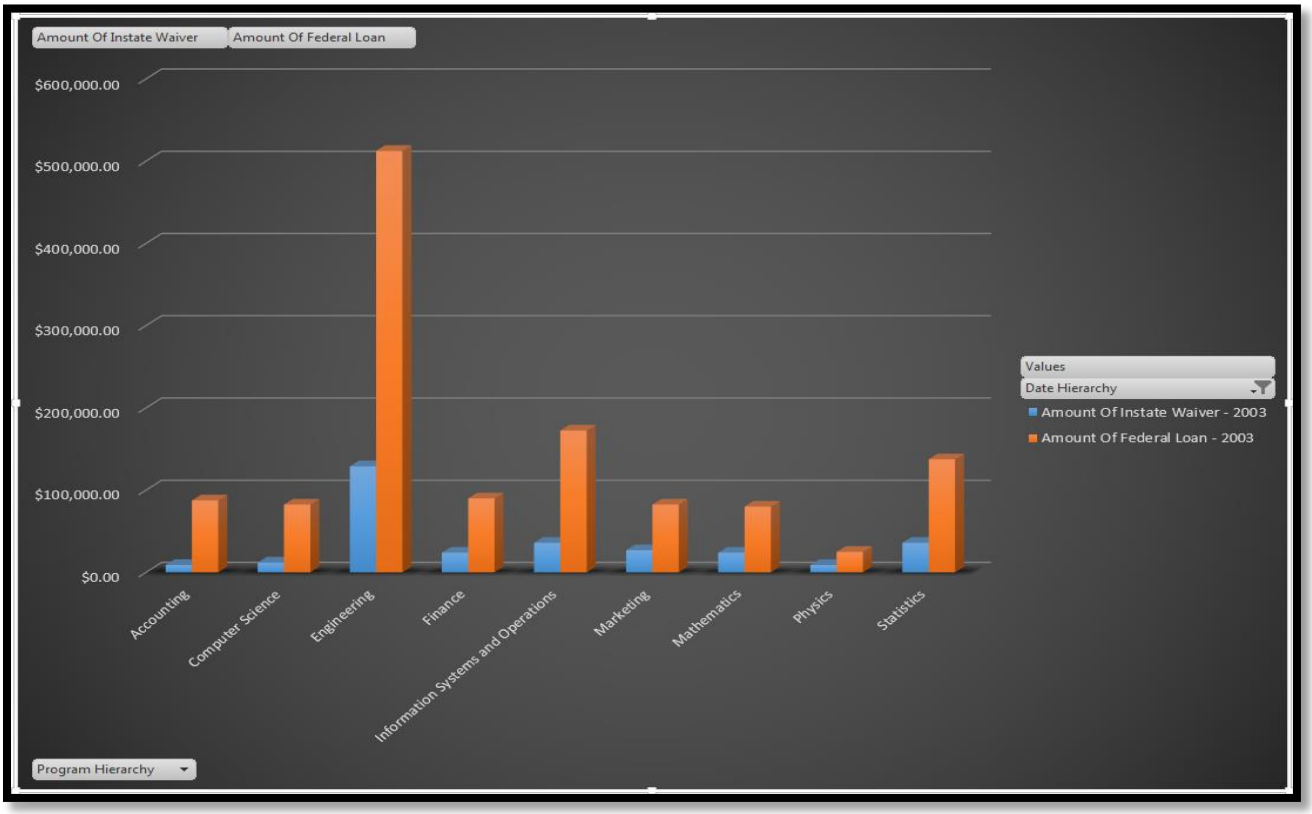
Row Labels	Amount Of Instate Waiver 2003	Amount Of Federal Loan 2003	Total Amount Of Instate Waiver	Total Amount Of Federal Loan
Accounting	$9,000.00	$87,500.00	$9,000.00	$87,500.00
Computer Science	$12,000.00	$82,500.00	$12,000.00	$82,500.00
Engineering	$129,000.00	$512,500.00	$129,000.00	$512,500.00
Finance	$24,000.00	$90,000.00	$24,000.00	$90,000.00
Information Systems and Operations	$36,000.00	$172,500.00	$36,000.00	$172,500.00
Marketing	$27,000.00	$82,500.00	$27,000.00	$82,500.00
Mathematics	$24,000.00	$80,000.00	$24,000.00	$80,000.00
Physics	$9,000.00	$25,000.00	$9,000.00	$25,000.00
Statistics	$36,000.00	$137,500.00	$36,000.00	$137,500.00
Grand Total	**$306,000.00**	**$1,270,000.00**	**$306,000.00**	**$1,270,000.00**

Department Name: Accounting, Computer Science, Engineering, Finance, Information Systems and Operations, Marketing, Mathematics, Physics

Program Name: Industrial Engineering, Mechanical Engineering, Financial Management, Management Information S..., Supply Chain Management, Marketing Analytics, Applied Mathematics, Theoritical Physics

Calendar Year: 2001, 2002, 2003, 2004, 2005, 2006

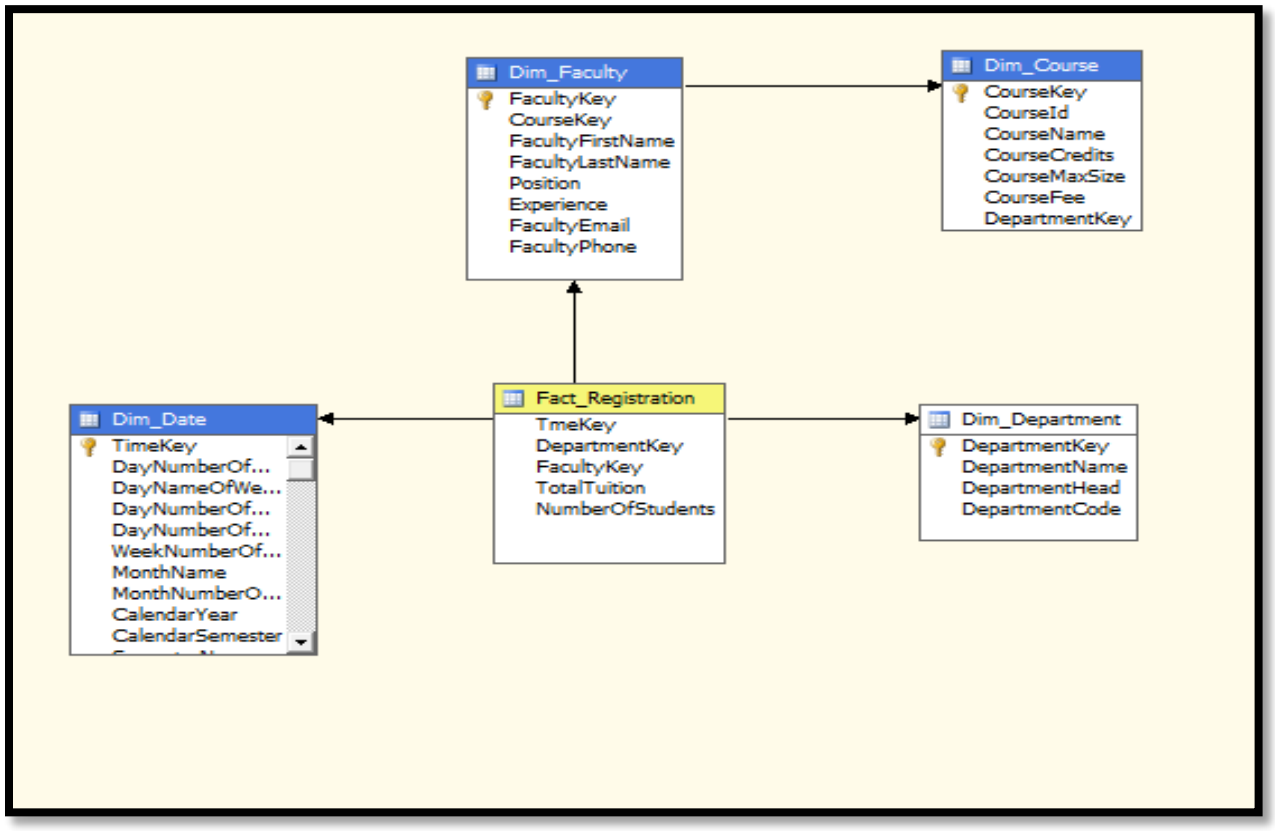
Semester Name: Spring, Summer, Fall, Spring, Summer, Fall, Spring, Summer



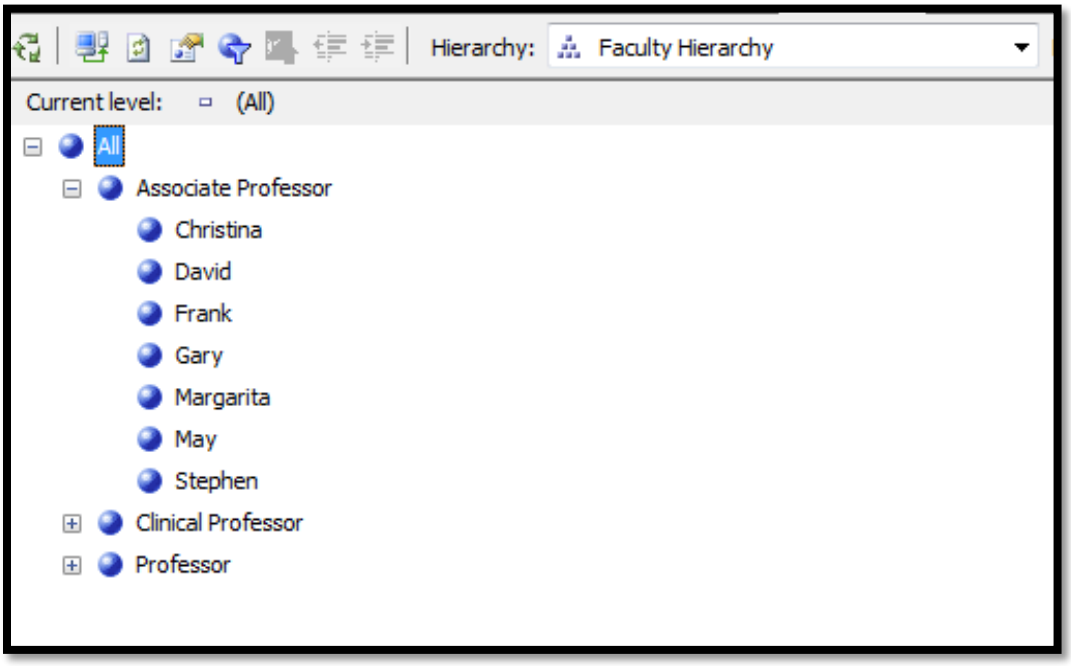
The above report helps in identifying the departments that sanction the highest amount of instate wavier. Also the management will be able to compare the amount of federal loan sanctioned to students of each department. These statistics can also be useful for strategic decision making on scholarship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

Course Registration:

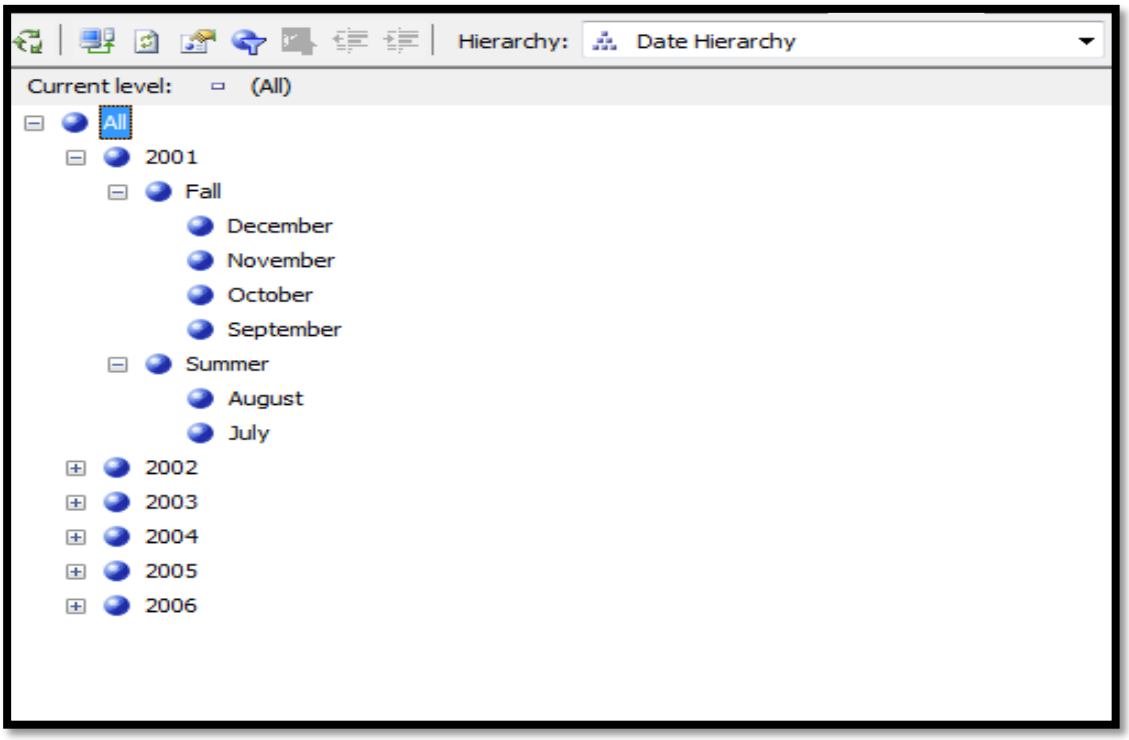
Cube Structure:



Hierarchies:



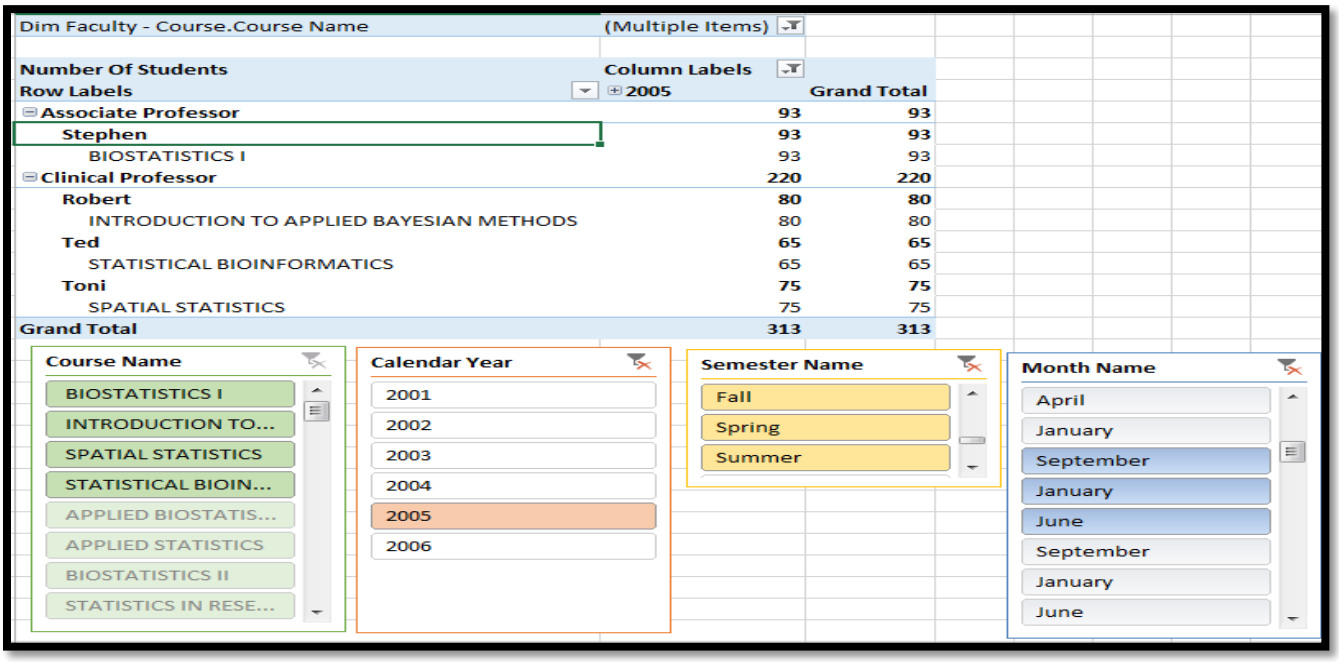
Faculty Hierarchy



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.



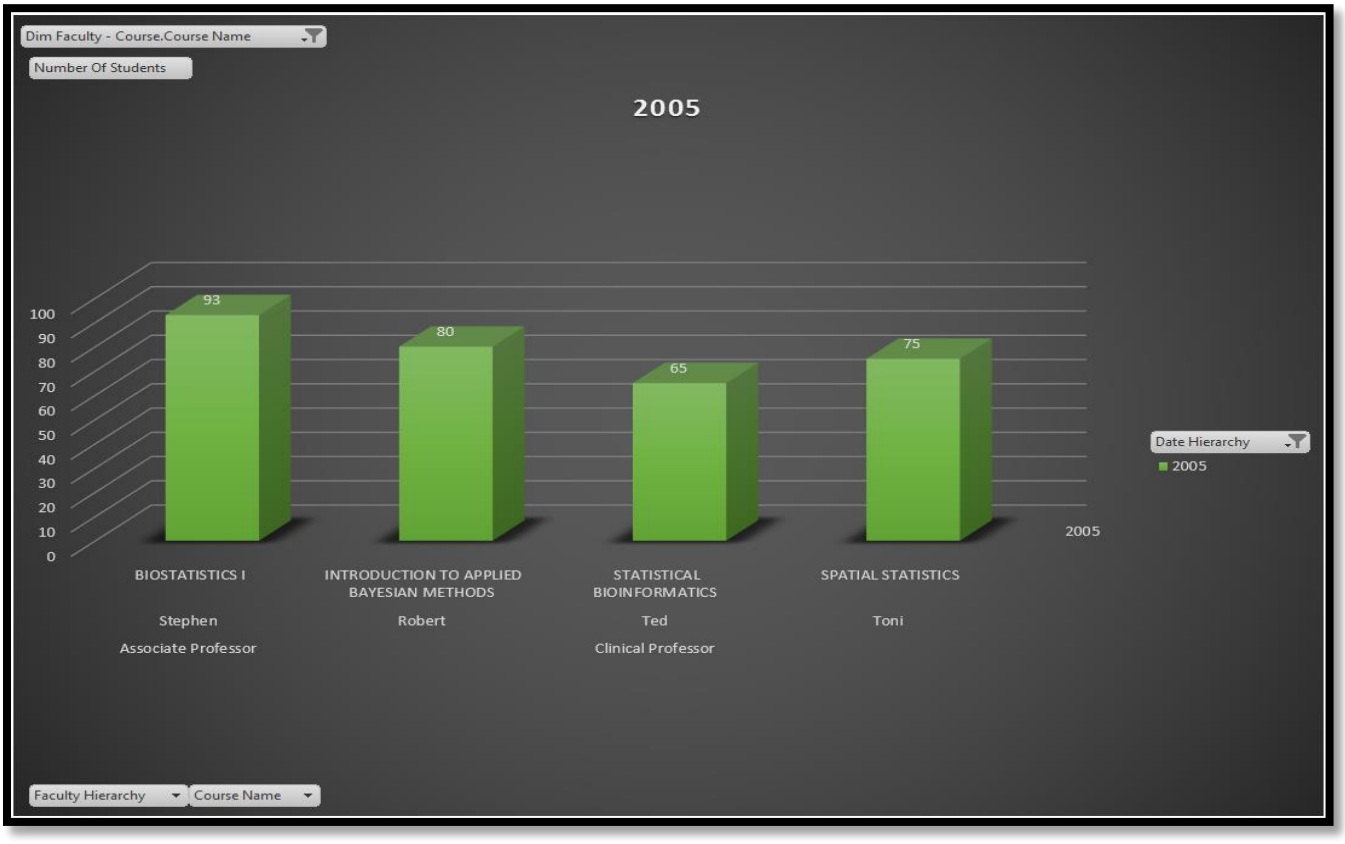
Dim Faculty - Course.Course Name	(Multiple Items)	
Number Of Students	**Column Labels**	
Row Labels	**2005**	**Grand Total**
Associate Professor	**93**	**93**
Stephen	**93**	**93**
BIOSTATISTICS I	93	93
Clinical Professor	**220**	**220**
Robert	**80**	**80**
INTRODUCTION TO APPLIED BAYESIAN METHODS	80	80
Ted	**65**	**65**
STATISTICAL BIOINFORMATICS	65	65
Toni	**75**	**75**
SPATIAL STATISTICS	75	75
Grand Total	**313**	**313**

Course Name: BIOSTATISTICS I; INTRODUCTION TO...; SPATIAL STATISTICS; STATISTICAL BIOIN...; APPLIED BIOSTATIS...; APPLIED STATISTICS; BIOSTATISTICS II; STATISTICS IN RESE...

Calendar Year: 2001; 2002; 2003; 2004; 2005; 2006

Semester Name: Fall; Spring; Summer

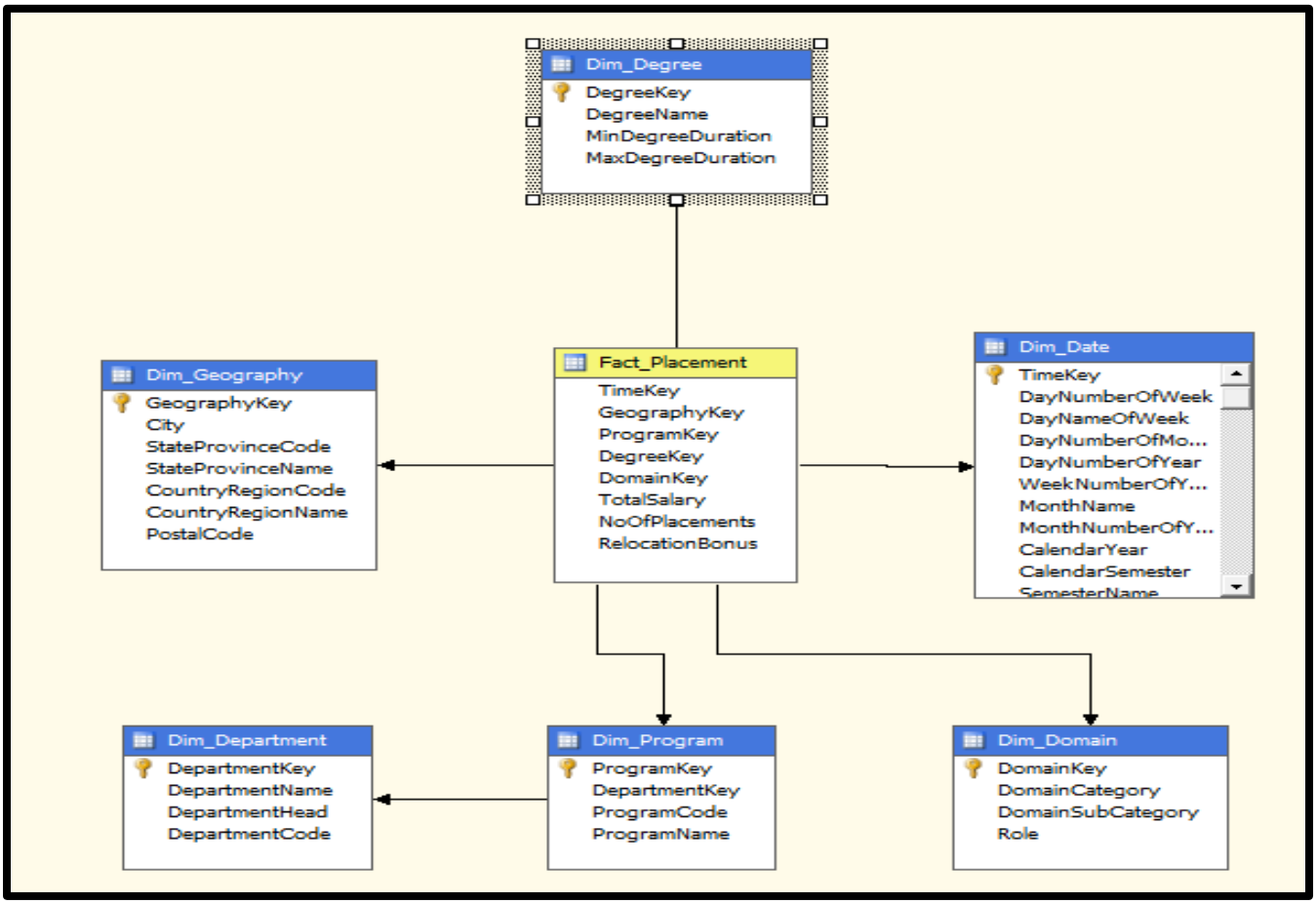
Month Name: April; January; September; January; June; September; January; June



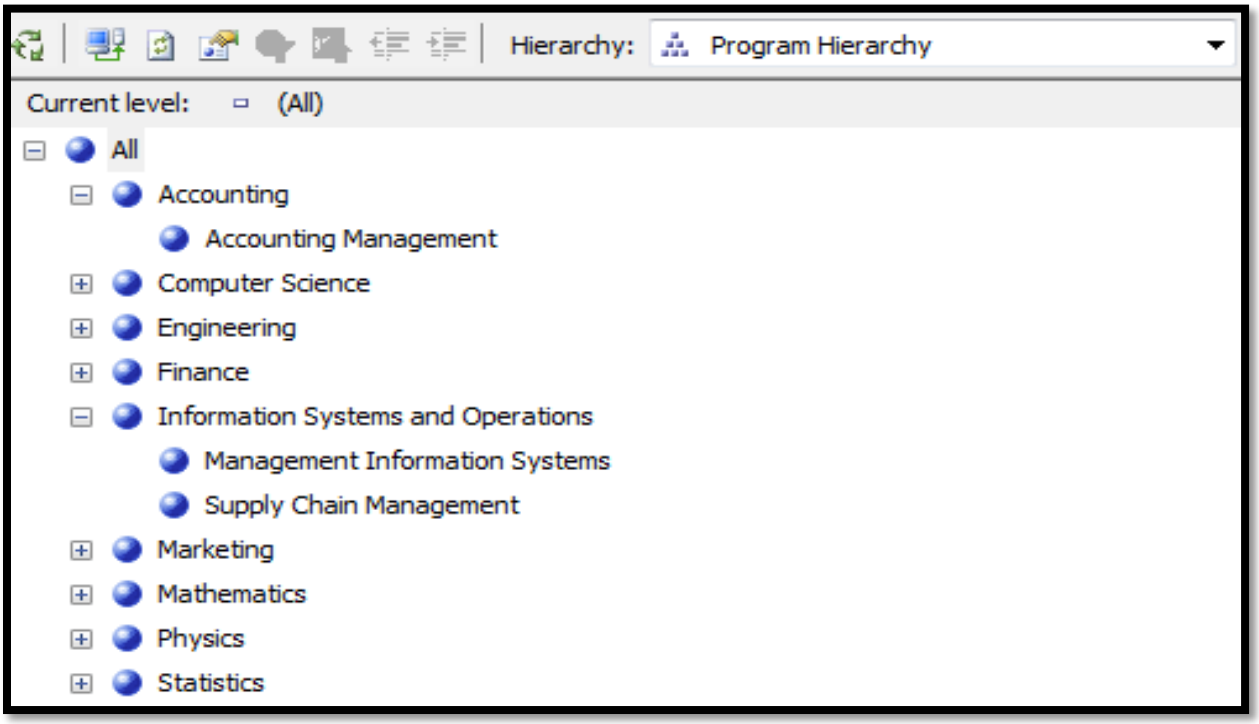
The above report gives a holistic view of the number of students who are registering for a particular course during a particular year. This piece of information can be really useful in understanding the preferences of students when it comes to course selection. The university will be able to allocate 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:

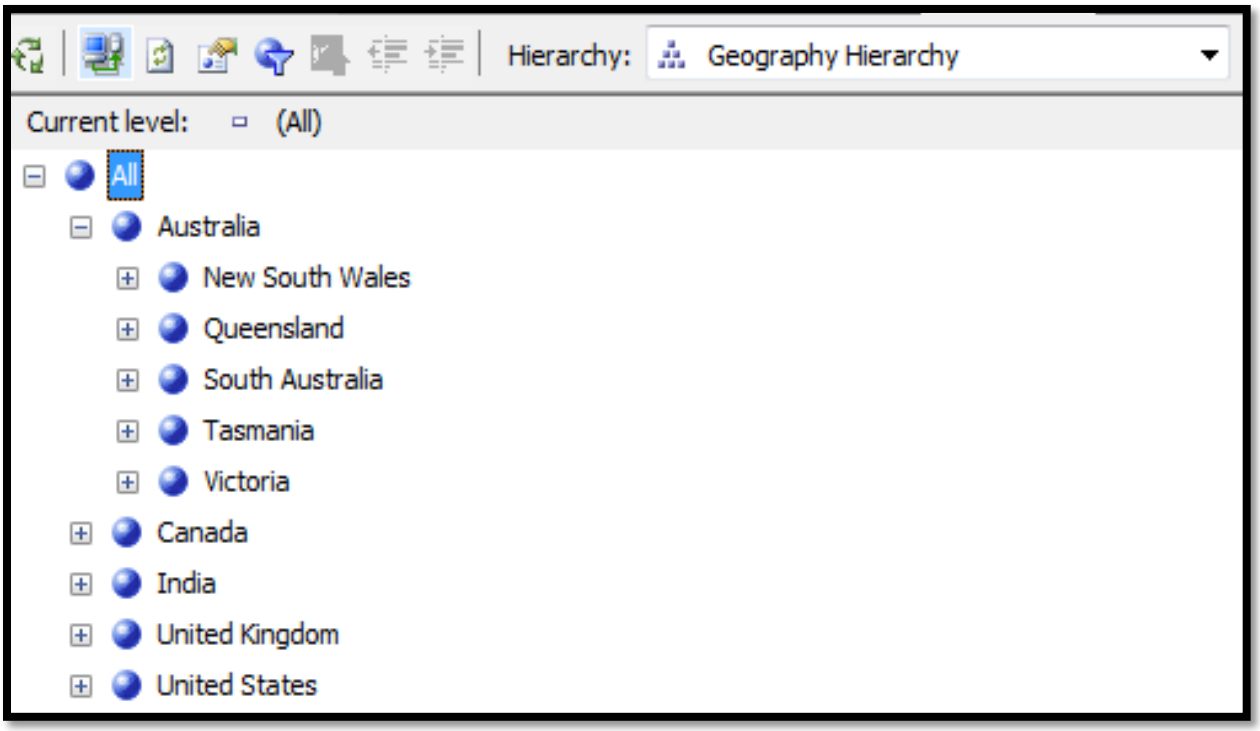
Cube Structure:



Hierarchies:



Program Hierarchy

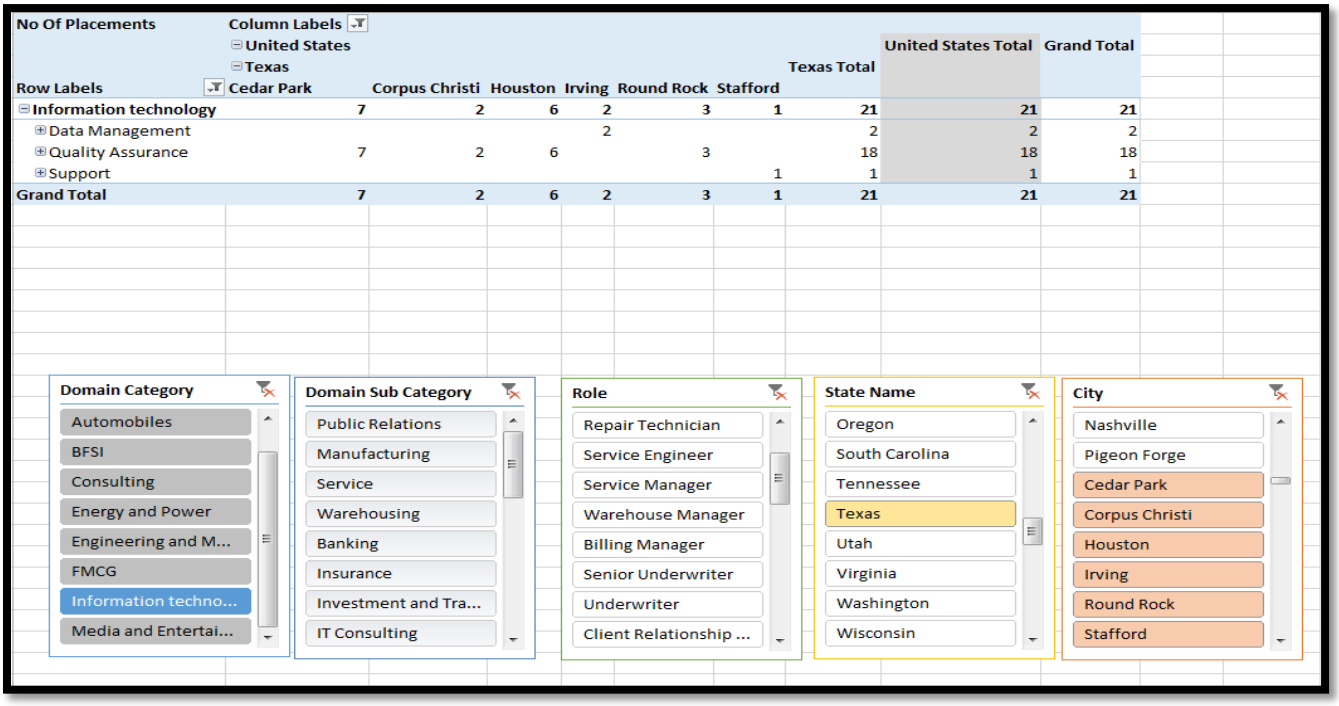


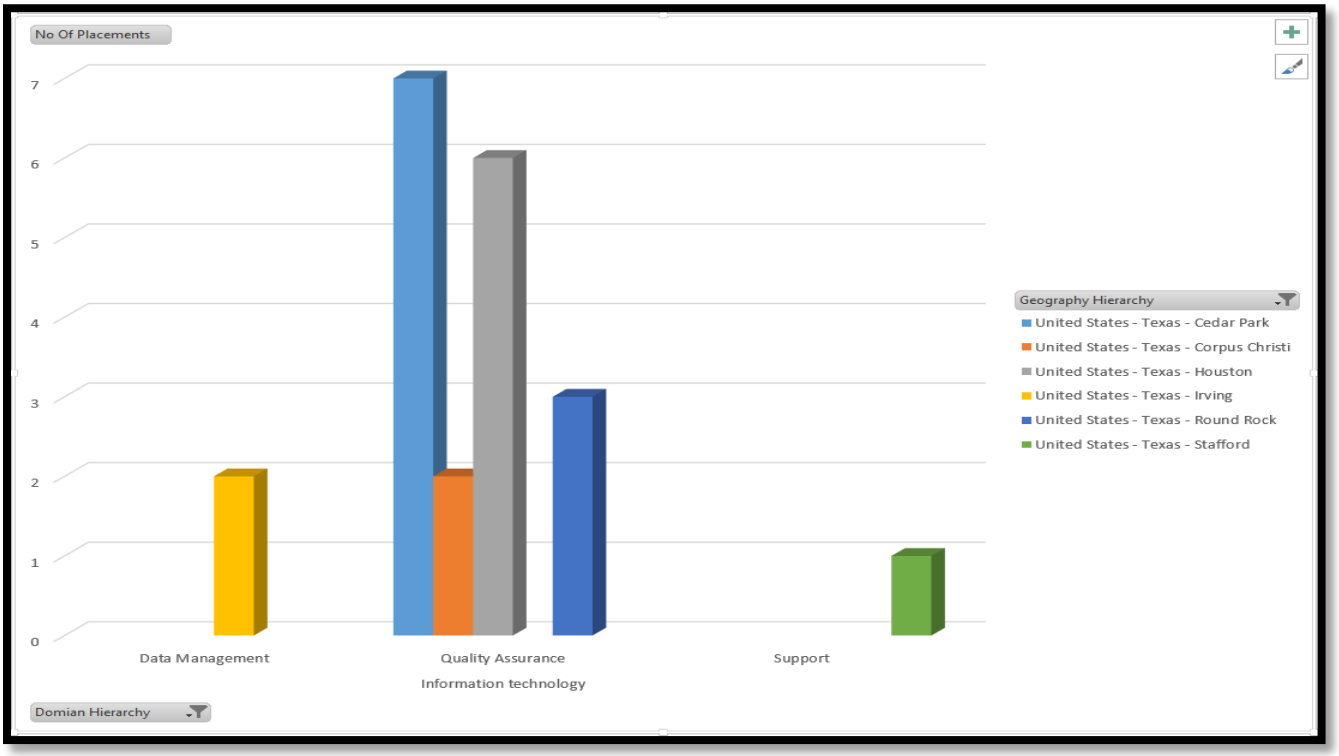
Geography Hierarchy

Reports:

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

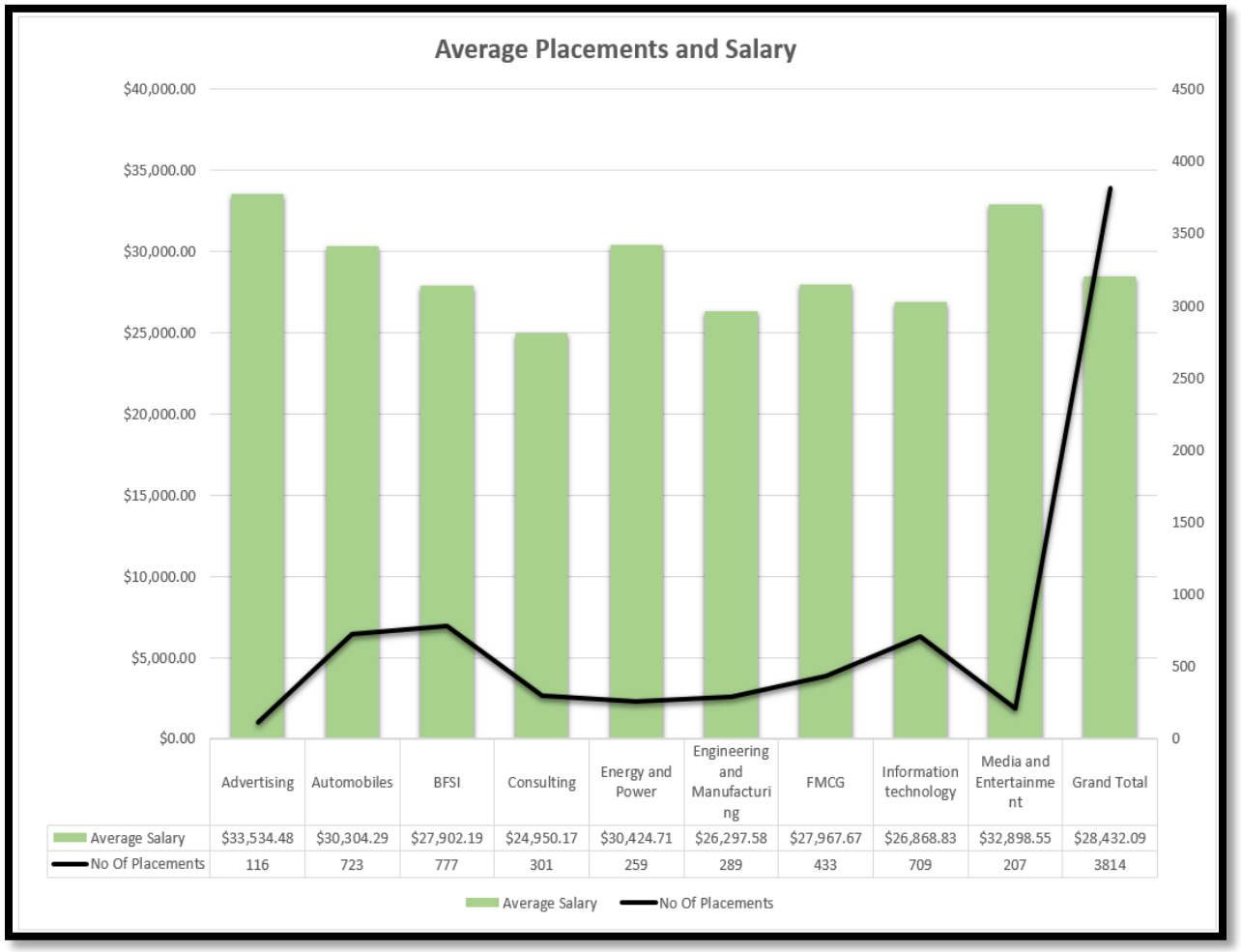
No Of Placements	Column Labels								
	United States							United States Total	Grand Total
	Texas						Texas Total		
Row Labels	Cedar Park	Corpus Christi	Houston	Irving	Round Rock	Stafford			
Information technology	7	2	6	2	3	1	21	21	21
Data Management				2			2	2	2
Quality Assurance	7	2	6		3		18	18	18
Support						1	1	1	1
Grand Total	7	2	6	2	3	1	21	21	21





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

Row Labels	No Of Placements	Average Salary
Advertising	116	$33,534.48
Automobiles	723	$30,304.29
BFSI	777	$27,902.19
Consulting	301	$24,950.17
Energy and Power	259	$30,424.71
Engineering and Manufacturing	289	$26,297.58
FMCG	433	$27,967.67
Information technology	709	$26,868.83
Media and Entertainment	207	$32,898.55
Grand Total	**3814**	**$28,432.09**



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

A. 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

B. 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

C. 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

D. 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