VIVEKANANDA INSTITUTE OF PROFESSIONAL STUDIES

Master Computer Applications (MCA)

BATCH: 2017 – 2020

Data Warehousing & Data Mining

Practical Lab

SUBMITTED TO:

Dr. Deepali Kamthania

PROFESSOR

VIPS

SUBMITTED BY:

Akshit Tripathy (01017704417)

Ayush Chauhan (01817704417)

MCA 4-A

STUDENT PERFORMACE ANALYSIS SYSTEM

We consider a real time application, a warehouse that functions within a time frame that the users senses as immediate or current for the top management to analysis the student academic performance in their institutions.

A large number of academic institutions work on operational database for their day to day update. This system fully satisfies the complex quality requests of OLTP system, but it also shows significant OLAP failures. Data are not adequately prepared for complex report forming. The system uses operational database that can't provide broad range of possibilities for creating complex reports. Operational database does not have special tools for creating queries that are defined by users.

The significant benefit from this solution of information and knowledge retrieval in databases is that the user does not need to possess knowledge concerning the relational model and the complex query languages. This approach in data analysis becomes more and more popular because it enables OLTP systems to get optimized for their purpose and to transfer data analysis to OLAP systems.

The information from the system of academics institutions can be rapidly assessed to find the performance of students in that institution.

The data & information gained from the system can be use as a substantial indicator for monitoring of the potential failure & improvements. Furthermore,

- Alerts can be sent to the parent & academic staff to intimate them about the performance of the student.
- Counselling can be given to students who struggle with their performances before they lose their grounds.
- Insights can be sense in future, so students benefited themselves to avail the advantages in placements.

- Tools :

  - SSDT : SQL Server Data Tools (SSDT) is a toolset that allows professional database and application developers to carry out all their database design work for SQL Server and SQL Azure within Visual Studio.

  - SSAS : SQL Server Analysis Services (SSAS) is the technology from the Microsoft Business Intelligence stack, to develop Online Analytical Processing (OLAP) solutions. In simple terms, you can use SSAS to create cubes using data from data marts / data warehouse for deeper and faster data analysis.

- Selecting fact table, dimensional tables and appropriate schemas :

The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for on-line transaction processing. A data warehouse, however, requires a concise, subject-oriented schema that facilitates online data analysis.

We choose 3 facts for analysis of student performance:

- SCORE
- PLACEMENT
- ATTENDENCE

Let's take the 'PLACEMENT' fact for elaboration of its respective dimensions, hierarchies/ categories & measure facts/ metrics.

1. FACT : SCORE

- Dimensions:

a) Student :
   1) Stu_id
   2) Stu_name
   3) Stu_dob
   4) Stu_gpi
   5) Stu_qualifiactions

b) Institute :
   1) Ins_id
   2) Ins_name
   3) Ins_loc

c) Company :

1)Comp_loc
2) Comp_name
3)Comp_mv
4)Comp_position
5) Comp_id

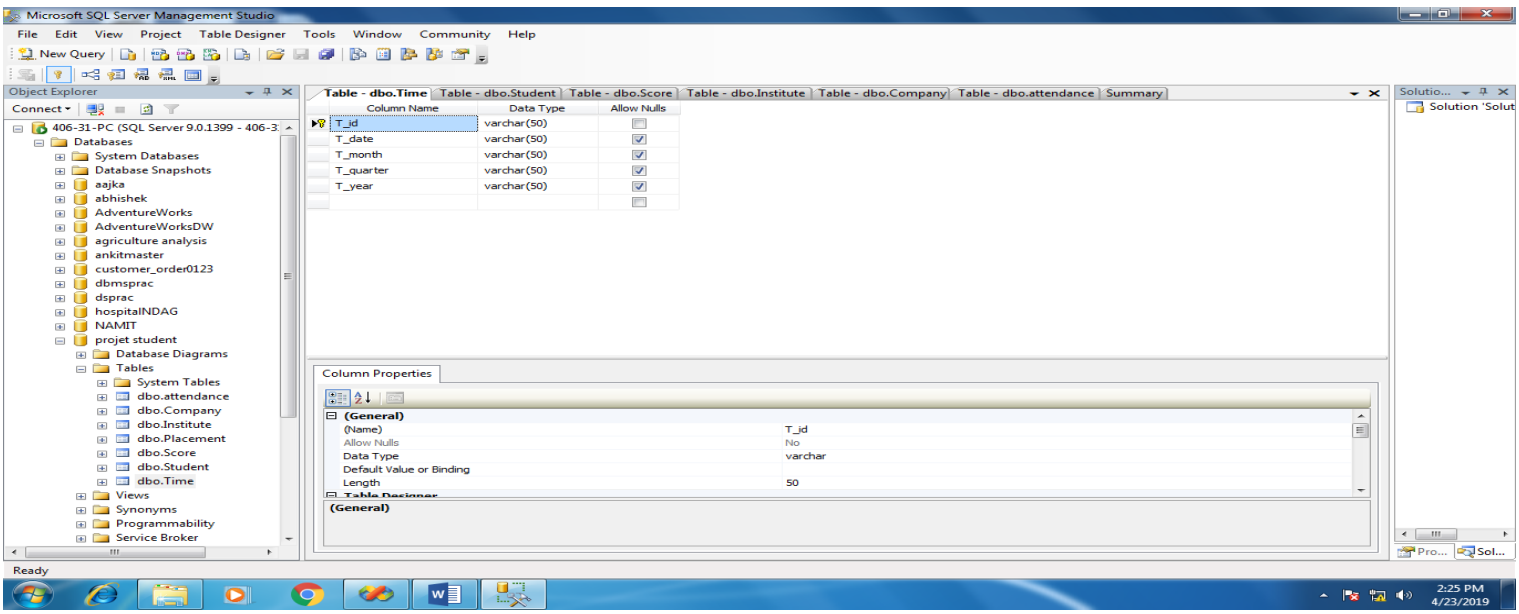
d) Time :

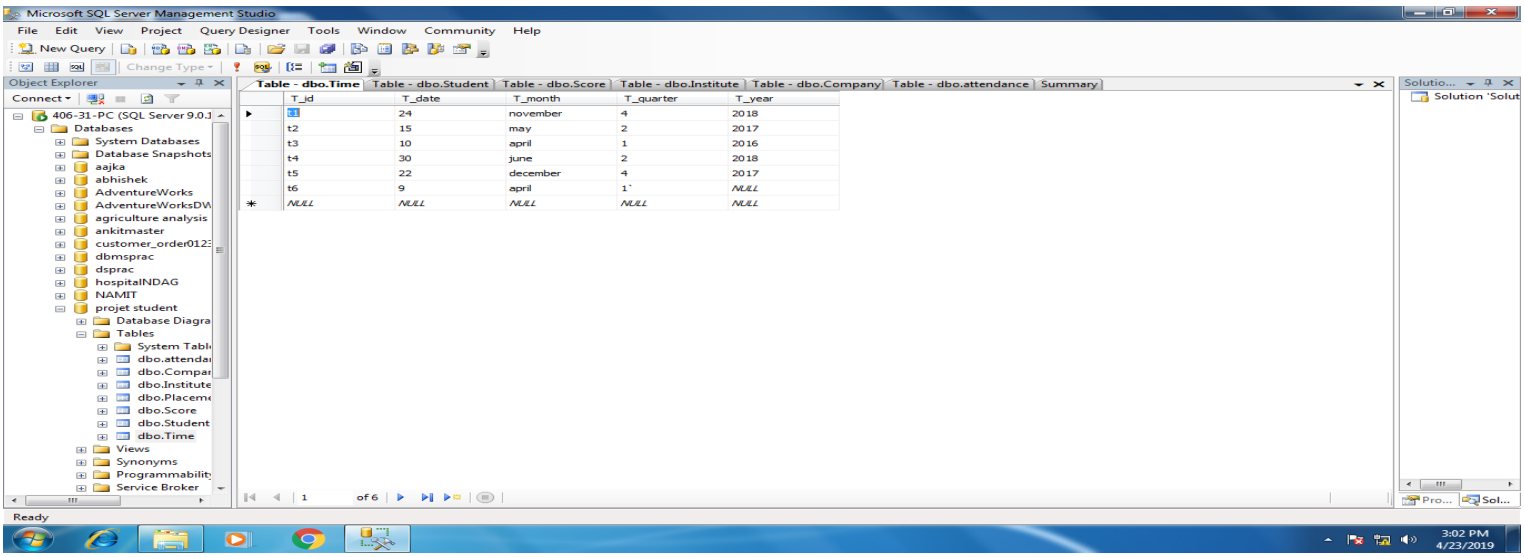
1) t_id
2) t_date
3) t_month
4) t_quarter
5) t_year

- Time is always a must dimension for every fact.

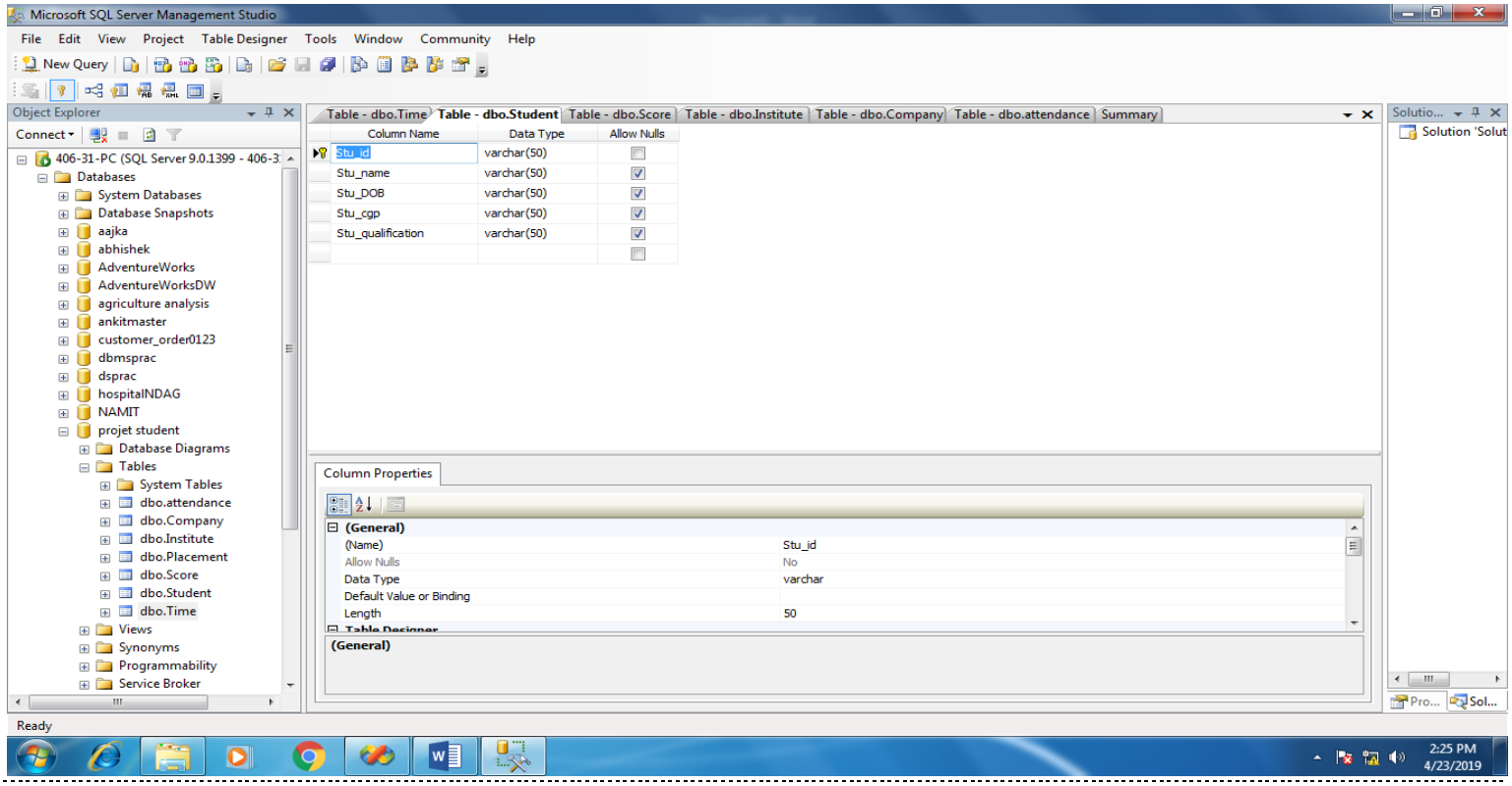
Schema Designing –

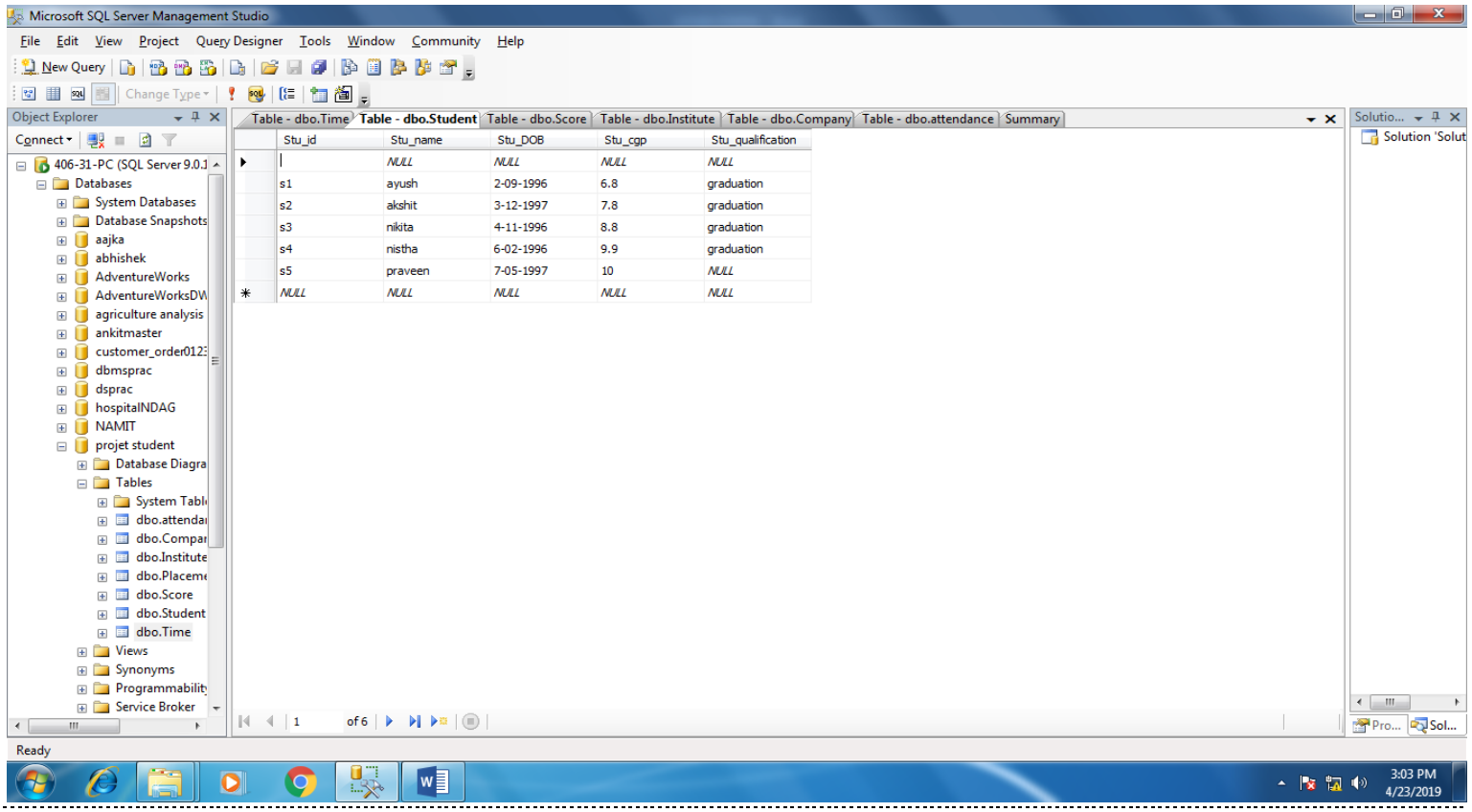
- Time



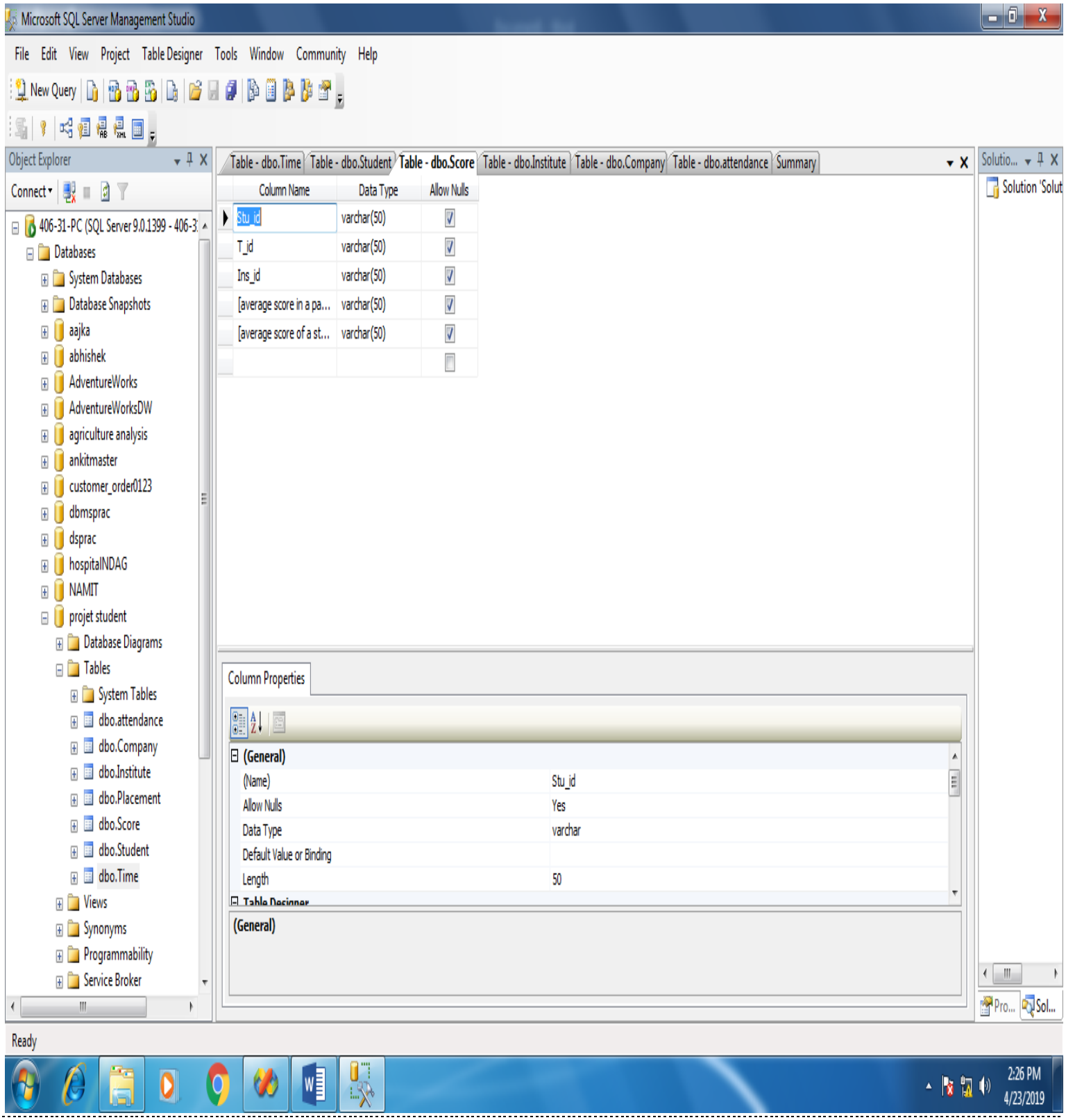


- Student

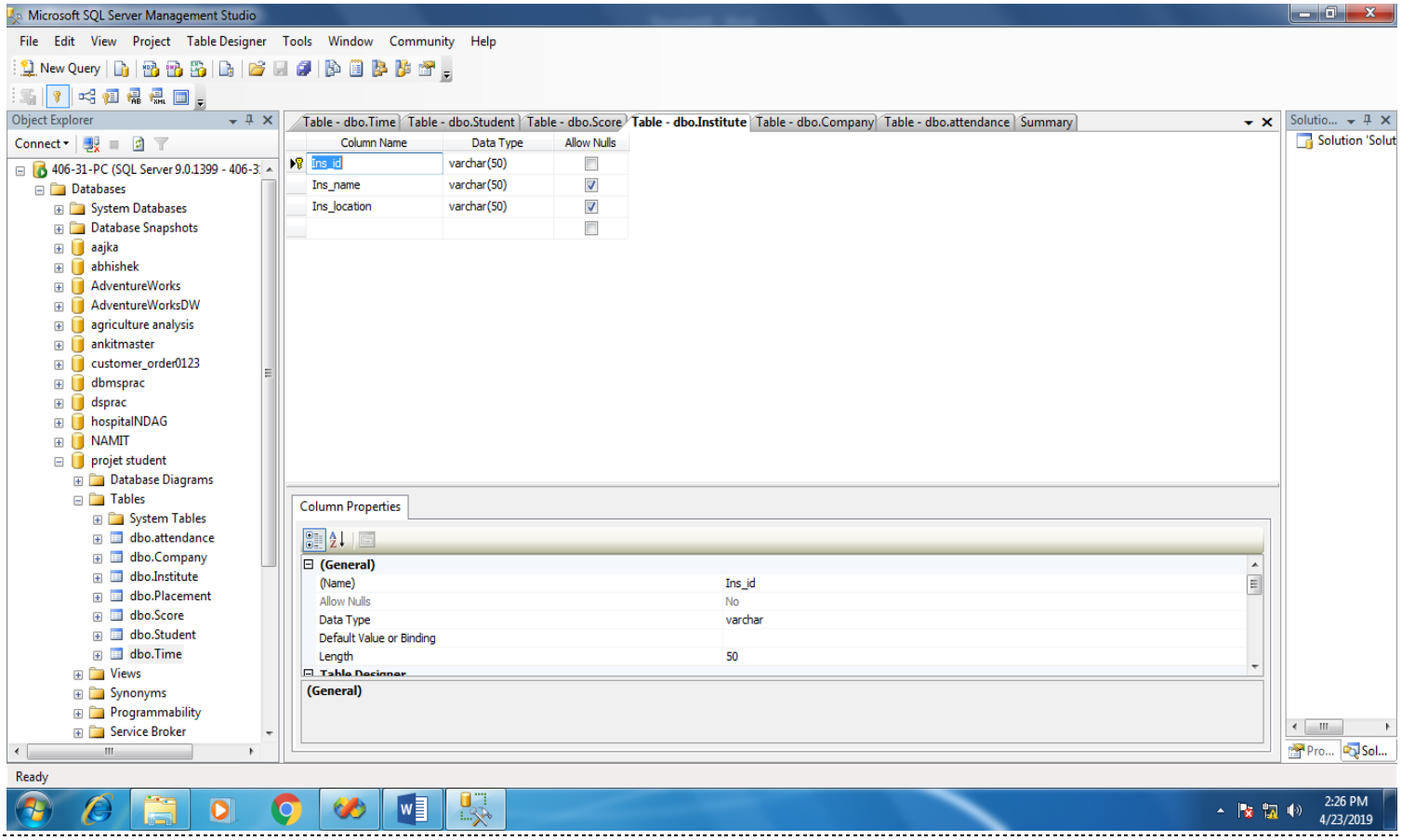

Microsoft SQL Server Management Studio
File Edit View Project Table Designer Tools Window Community Help
New Query
Object Explorer
Connect
406-31-PC (SQL Server 9.0.1399 - 406-3
Databases
System Databases
Database Snapshots
aajka
abhishek
AdventureWorks
AdventureWorksDW
agriculture analysis
ankitmaster
customer_order0123
dbmsprac
dsprac
hospitalNDAG
NAMIT
projet student
Database Diagrams
Tables
System Tables
dbo.attendance
dbo.Company
dbo.Institute
dbo.Placement
dbo.Score
dbo.Student
dbo.Time
Views
Synonyms
Programmability
Service Broker
Table - dbo.Time
Table - dbo.Student
Table - dbo.Score
Table - dbo.Institute
Table - dbo.Company
Table - dbo.attendance
Summary
Column Name
Data Type
Allow Nulls
Stu_id
varchar(50)
Stu_name
varchar(50)
Stu_DOB
varchar(50)
Stu_cgp
varchar(50)
Stu_qualification
varchar(50)
Column Properties
(General)
(Name)
Stu_id
Allow Nulls
No
Data Type
varchar
Default Value or Binding
Length
50
Table Designer
(General)
Solution 'Solut
Ready
2:25 PM
4/23/2019

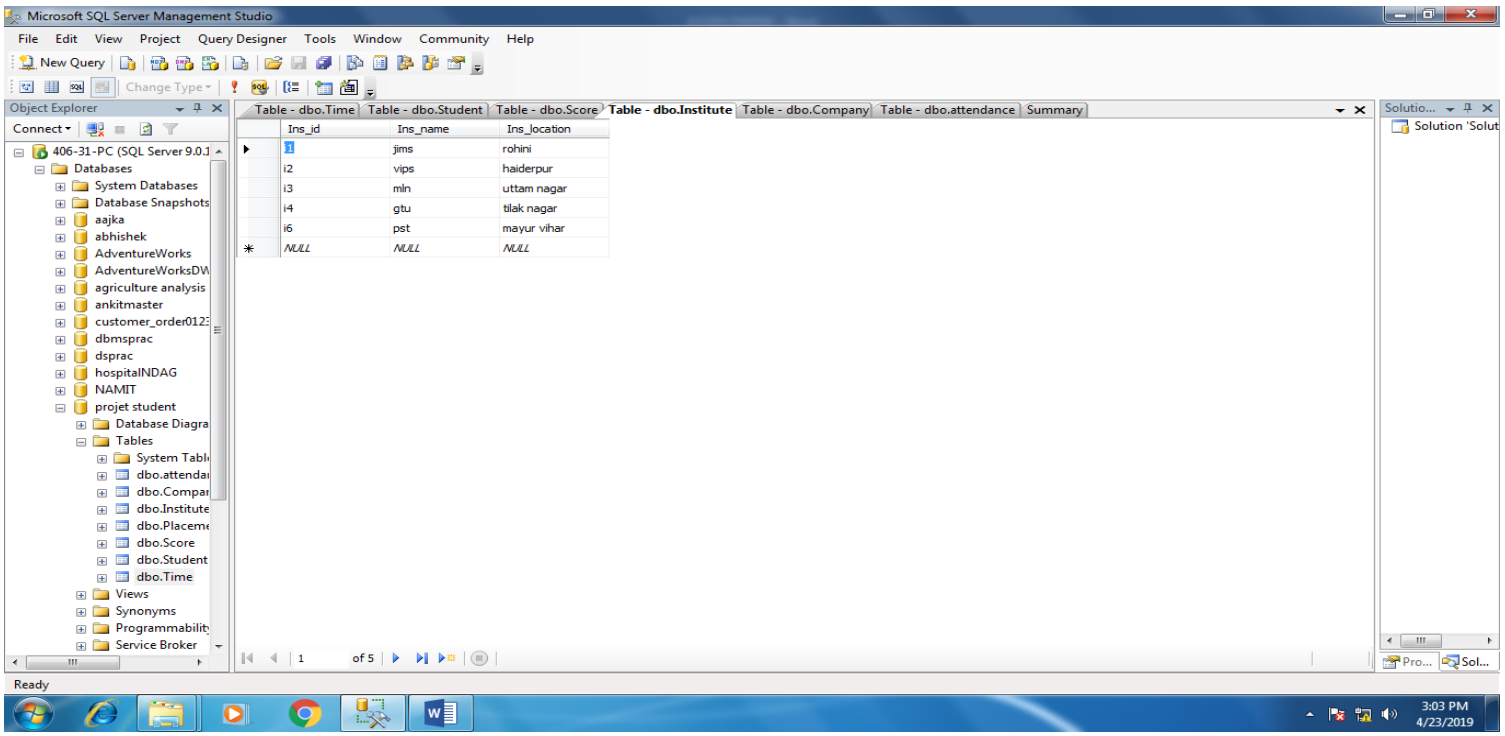

Microsoft SQL Server Management Studio
File Edit View Project Query Designer Tools Window Community Help
New Query
Change Type
Object Explorer
Connect
406-31-PC (SQL Server 9.0.1
Databases
System Databases
Database Snapshots
aajka
abhishek
AdventureWorks
agriculture analysis
ankitmaster
dbmsprac
dsprac
hospitalNDAG
NAMIT
projet student
Tables
Views
Synonyms
Service Broker
Table - dbo.Time
Table - dbo.Student
Table - dbo.Score
Table - dbo.Institute
Table - dbo.Company
Table - dbo.attendance
Summary
Stu_id
Stu_name
Stu_DOB
Stu_cgp
Stu_qualification
s1 ayush 2-09-1996 6.8 graduation
s2 akshit 3-12-1997 7.8 graduation
s3 nikita 4-11-1996 8.8 graduation
s4 nistha 6-02-1996 9.9 graduation
s5 praveen 7-05-1997 10
1 of 6
Solution 'Solut
Ready
3:03 PM
4/23/2019

- Score

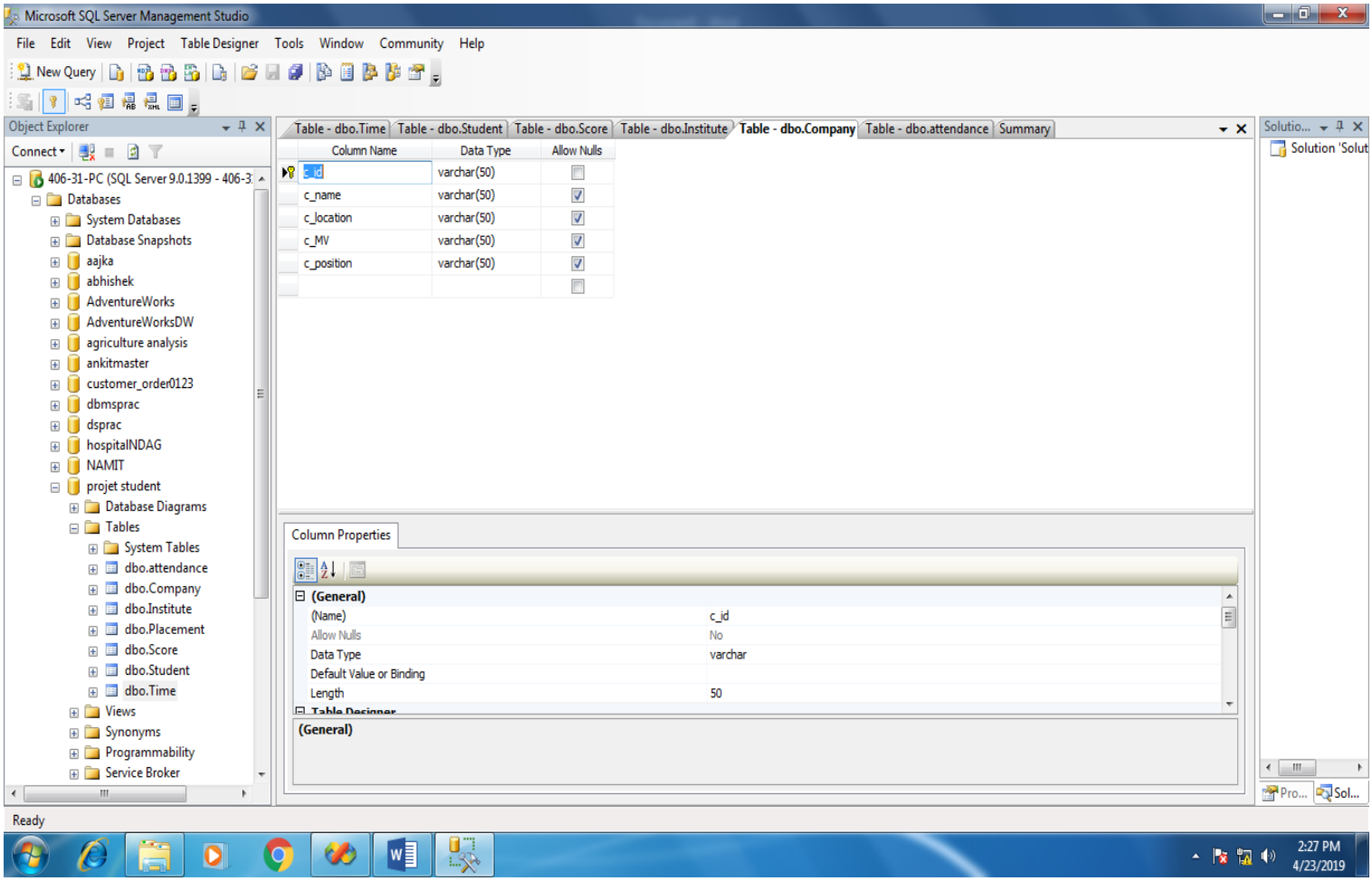

Microsoft SQL Server Management Studio
File Edit View Project Table Designer Tools Window Community Help
New Query
Object Explorer
Connect
406-31-PC (SQL Server 9.0.1399 - 406-3
Databases
System Databases
Database Snapshots
aajka
abhishek
AdventureWorks
AdventureWorksDW
agriculture analysis
ankitmaster
customer_order0123
dbmsprac
dsprac
hospitalNDAG
NAMIT
projet student
Database Diagrams
Tables
System Tables
dbo.attendance
dbo.Company
dbo.Institute
dbo.Placement
dbo.Score
dbo.Student
dbo.Time
Views
Synonyms
Programmability
Service Broker
Table - dbo.Time
Table - dbo.Student
Table - dbo.Score
Table - dbo.Institute
Table - dbo.Company
Table - dbo.attendance
Summary
Column Name
Data Type
Allow Nulls
Stu_id
varchar(50)
T_id
varchar(50)
Ins_id
varchar(50)
[average score in a pa...
varchar(50)
[average score of a st...
varchar(50)
Column Properties
(General)
(Name)
Stu_id
Allow Nulls
Yes
Data Type
varchar
Default Value or Binding
Length
50
Table Designer
(General)
Solutio...
Solution 'Solut
Ready
2:26 PM
4/23/2019

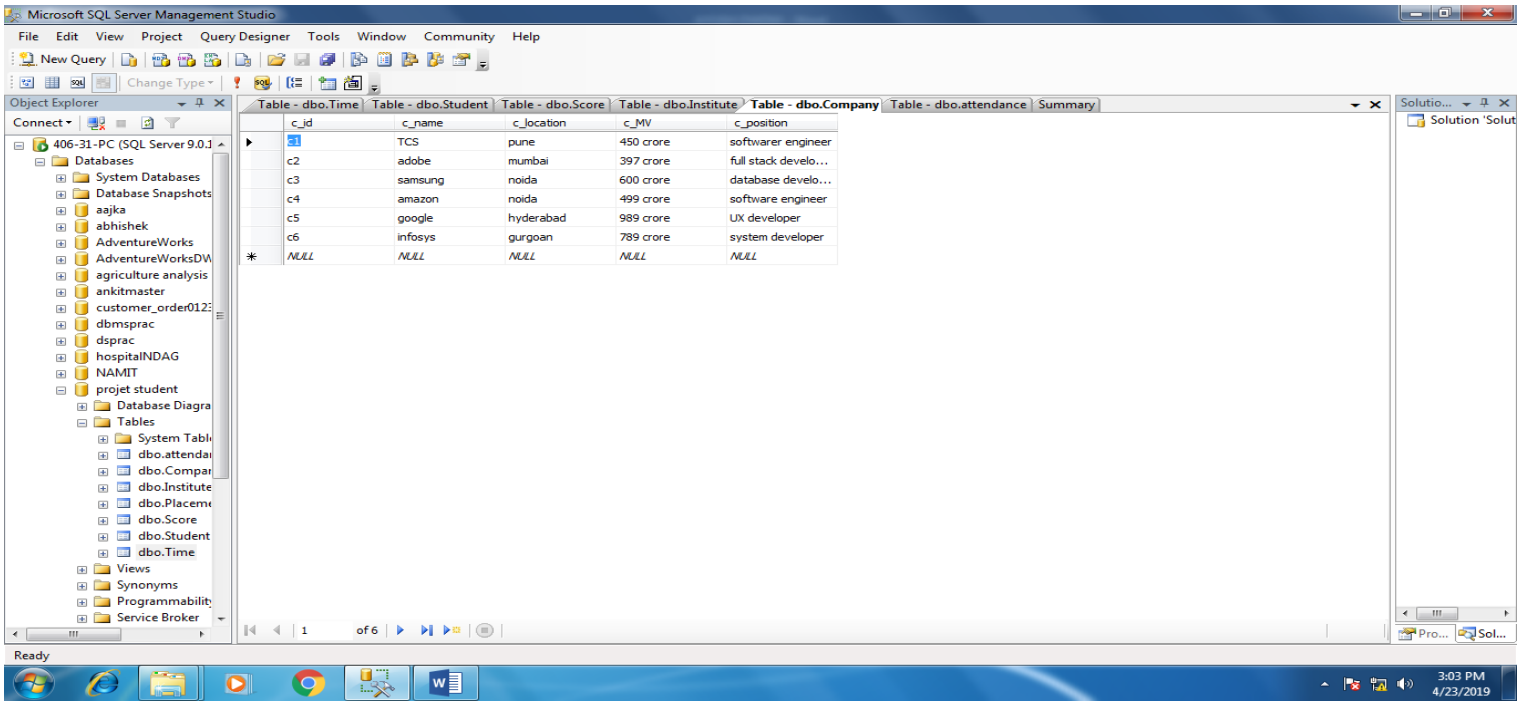
- Institute


Microsoft SQL Server Management Studio
File Edit View Project Table Designer Tools Window Community Help
New Query
Object Explorer
Connect
406-31-PC (SQL Server 9.0.1399 - 406-3
Databases
System Databases
Database Snapshots
aajka
abhishek
AdventureWorks
AdventureWorksDW
agriculture analysis
ankitmaster
customer_order0123
dbmsprac
dsprac
hospitalNDAG
NAMIT
projet student
Database Diagrams
Tables
System Tables
dbo.attendance
dbo.Company
dbo.Institute
dbo.Placement
dbo.Score
dbo.Student
dbo.Time
Views
Synonyms
Programmability
Service Broker
Table - dbo.Time
Table - dbo.Student
Table - dbo.Score
Table - dbo.Institute
Table - dbo.Company
Table - dbo.attendance
Summary
Column Name
Data Type
Allow Nulls
Ins_id
varchar(50)
Ins_name
varchar(50)
Ins_location
varchar(50)
Column Properties
(General)
(Name)
Ins_id
Allow Nulls
No
Data Type
varchar
Default Value or Binding
Length
50
Table Designer
(General)
Solution 'Solut
Ready
2:26 PM
4/23/2019

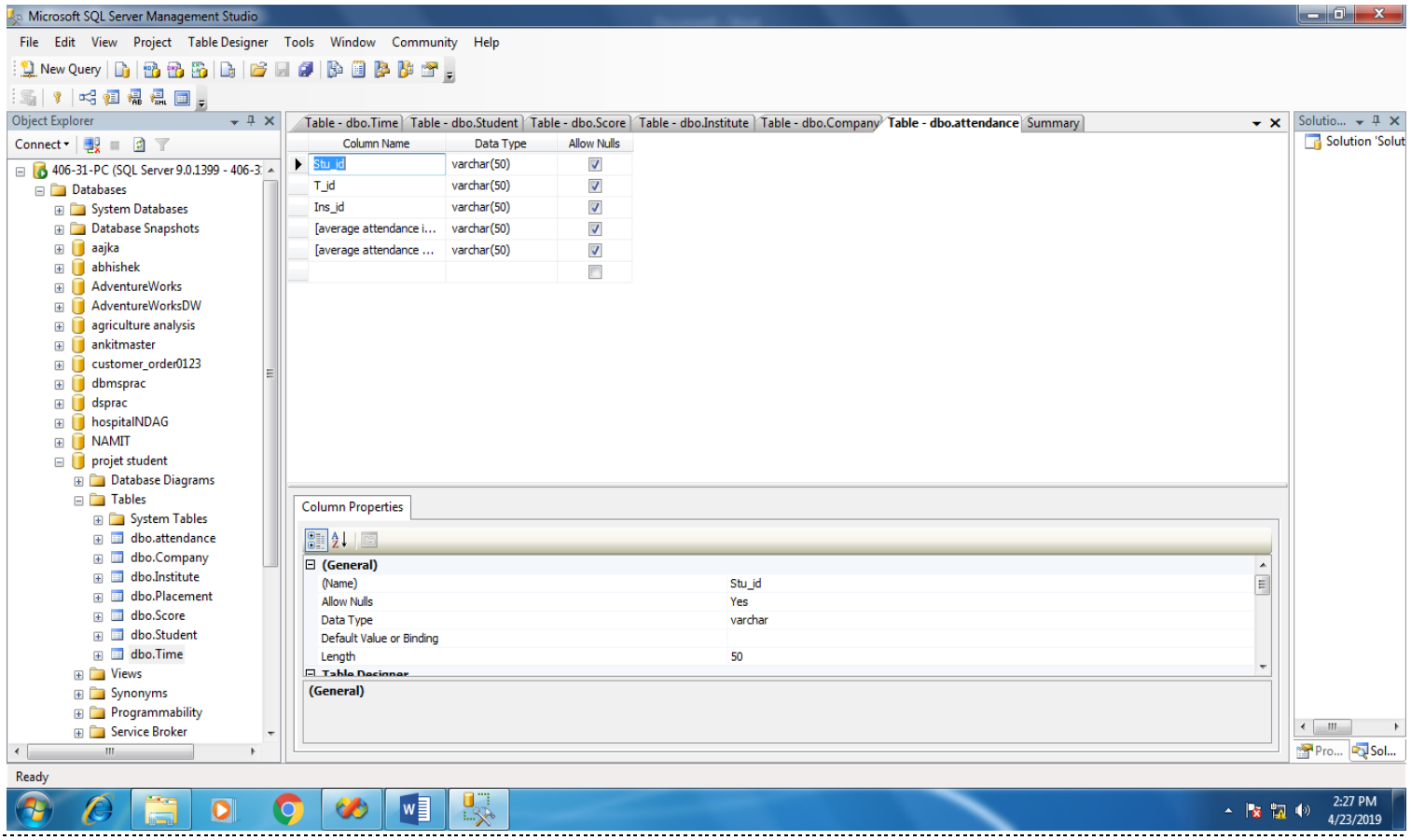

Microsoft SQL Server Management Studio
File Edit View Project Query Designer Tools Window Community Help
New Query
Change Type
Object Explorer
Connect
Table - dbo.Time
Table - dbo.Student
Table - dbo.Score
Table - dbo.Institute
Table - dbo.Company
Table - dbo.attendance
Summary
Ins_id
Ins_name
Ins_location
1
jims
rohini
i2
vips
haiderpur
i3
mln
uttam nagar
i4
gtu
tilak nagar
i6
pst
mayur vihar
NULL
NULL
NULL
1
of 5
Solution 'Solut
Ready
3:03 PM
4/23/2019

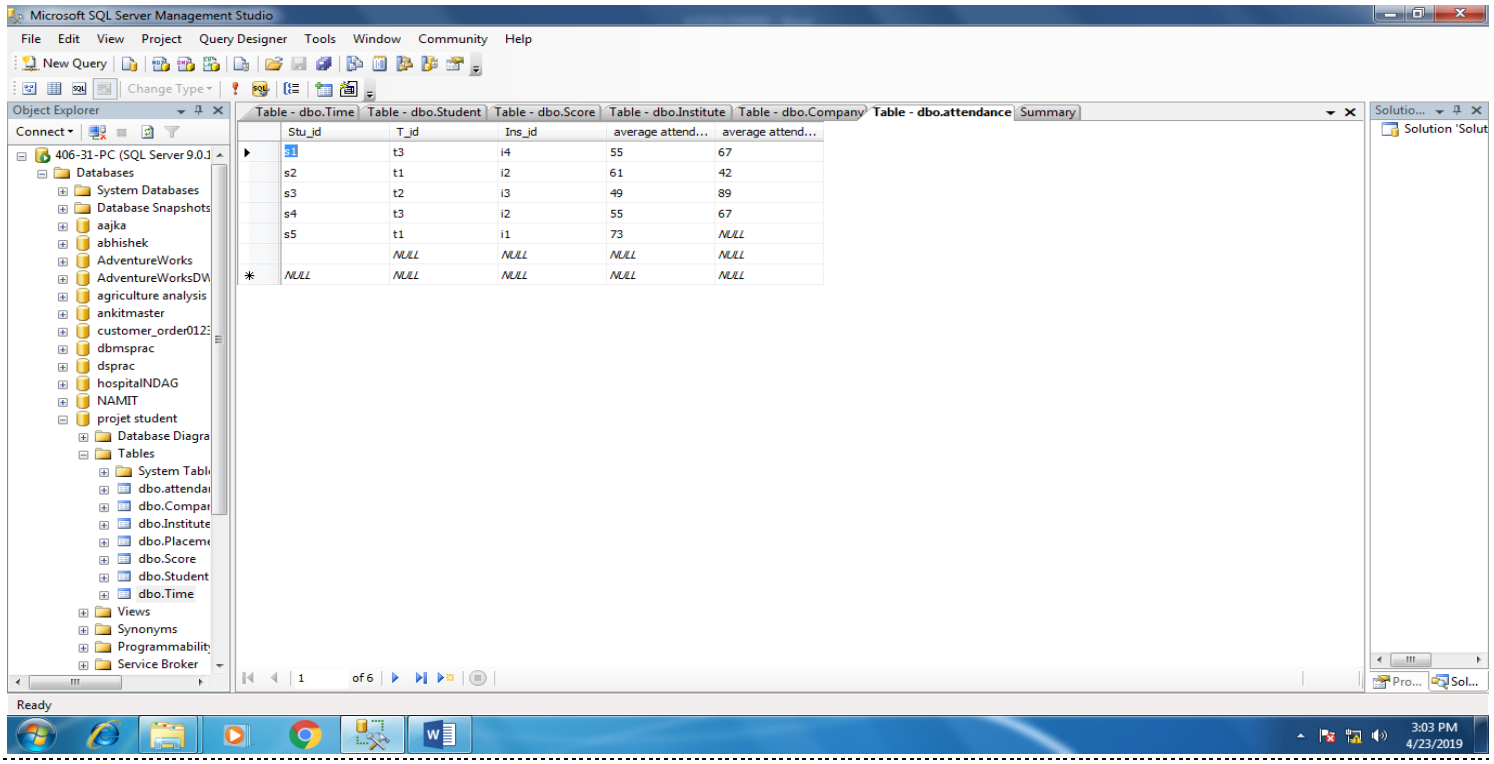
- Company


Microsoft SQL Server Management Studio
File Edit View Project Table Designer Tools Window Community Help
New Query
Object Explorer
Connect
406-31-PC (SQL Server 9.0.1399 - 406-3
Databases
System Databases
Database Snapshots
aajka
abhishek
AdventureWorks
AdventureWorksDW
agriculture analysis
ankitmaster
customer_order0123
dbmsprac
dsprac
hospitalNDAG
NAMIT
projet student
Database Diagrams
Tables
System Tables
dbo.attendance
dbo.Company
dbo.Institute
dbo.Placement
dbo.Score
dbo.Student
dbo.Time
Views
Synonyms
Programmability
Service Broker
Table - dbo.Time
Table - dbo.Student
Table - dbo.Score
Table - dbo.Institute
Table - dbo.Company
Table - dbo.attendance
Summary
Column Name
Data Type
Allow Nulls
c_id
varchar(50)
c_name
varchar(50)
c_location
varchar(50)
c_MV
varchar(50)
c_position
varchar(50)
Column Properties
(General)
(Name)
c_id
Allow Nulls
No
Data Type
varchar
Default Value or Binding
Length
50
(General)
Solution 'Solut
Ready
2:27 PM
4/23/2019

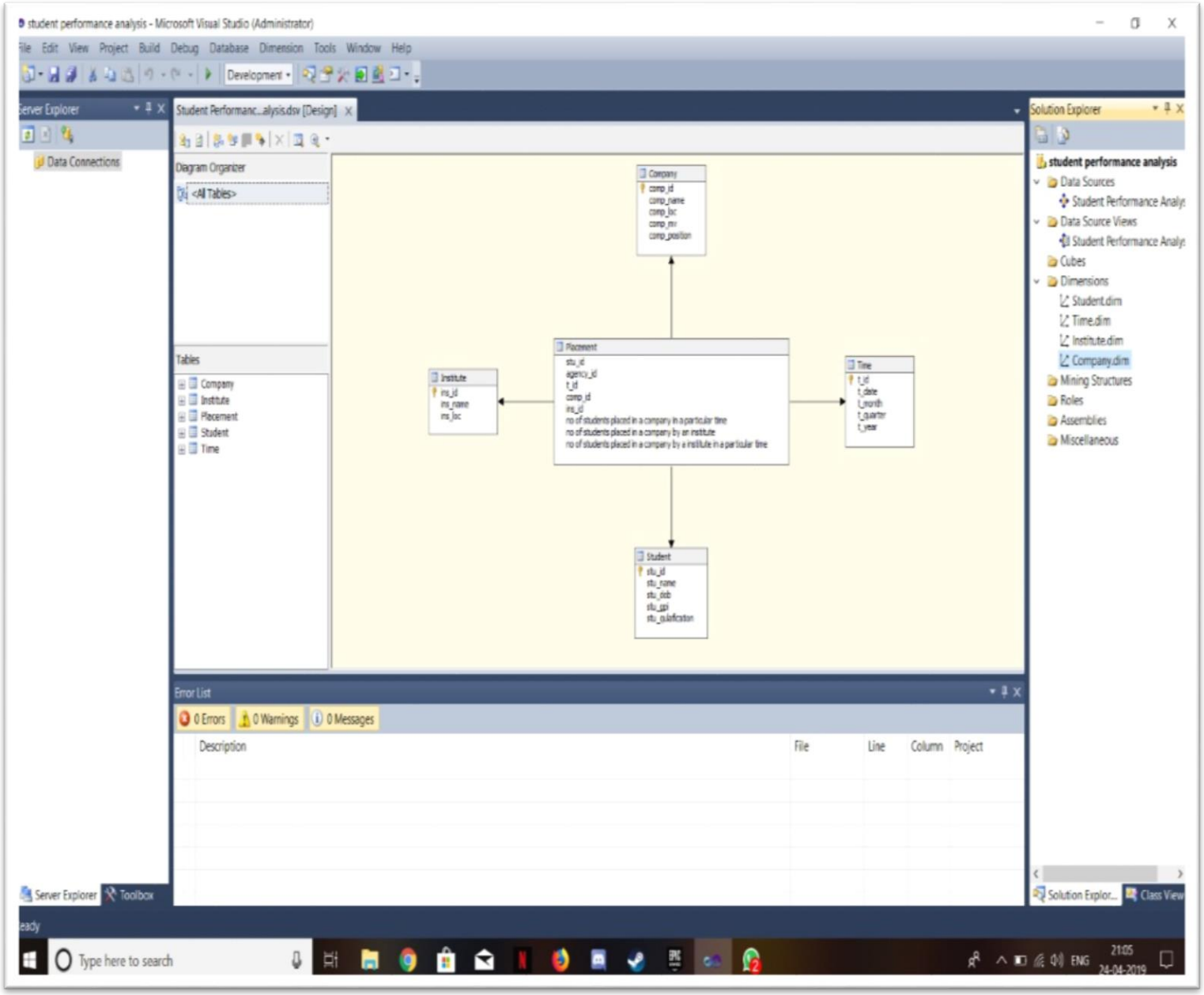

Microsoft SQL Server Management Studio
File Edit View Project Query Designer Tools Window Community Help
New Query
Change Type
Object Explorer
Connect
Table - dbo.Time
Table - dbo.Student
Table - dbo.Score
Table - dbo.Institute
Table - dbo.Company
Table - dbo.attendance
Summary
c_id
c_name
c_location
c_MV
c_position
c1 TCS pune 450 crore softwarer engineer
c2 adobe mumbai 397 crore full stack develo...
c3 samsung noida 600 crore database develo...
c4 amazon noida 499 crore software engineer
c5 google hyderabad 989 crore UX developer
c6 infosys gurgoan 789 crore system developer
NULL
of 6
Solution 'Solut
Ready
3:03 PM
4/23/2019

- Attendance


Microsoft SQL Server Management Studio
File Edit View Project Table Designer Tools Window Community Help
New Query
Object Explorer
Connect
406-31-PC (SQL Server 9.0.1399 - 406-3
Databases
System Databases
Database Snapshots
aajka
abhishek
AdventureWorks
AdventureWorksDW
agriculture analysis
ankitmaster
customer_order0123
dbmsprac
dsprac
hospitalNDAG
NAMIT
projet student
Database Diagrams
Tables
System Tables
dbo.attendance
dbo.Company
dbo.Institute
dbo.Placement
dbo.Score
dbo.Student
dbo.Time
Views
Synonyms
Programmability
Service Broker
Table - dbo.Time
Table - dbo.Student
Table - dbo.Score
Table - dbo.Institute
Table - dbo.Company
Table - dbo.attendance
Summary
Column Name
Data Type
Allow Nulls
Stu_id
varchar(50)
T_id
varchar(50)
Ins_id
varchar(50)
[average attendance i...
[average attendance ...
Column Properties
(General)
(Name)
Stu_id
Allow Nulls
Yes
Data Type
varchar
Default Value or Binding
Length
50
Table Designer
(General)
Solution 'Solut
Ready
2:27 PM
4/23/2019


Microsoft SQL Server Management Studio
File Edit View Project Query Designer Tools Window Community Help
New Query
Change Type
Object Explorer
Connect
Table - dbo.Time
Table - dbo.Student
Table - dbo.Score
Table - dbo.Institute
Table - dbo.Company
Table - dbo.attendance
Summary
Stu_id
T_id
Ins_id
average attend...
average attend...
s1 t3 i4 55 67
s2 t1 i2 61 42
s3 t2 i3 49 89
s4 t3 i2 55 67
s5 t1 i1 73 NULL
NULL NULL NULL NULL NULL
1 of 6
Solution 'Solut
Ready
3:03 PM
4/23/2019

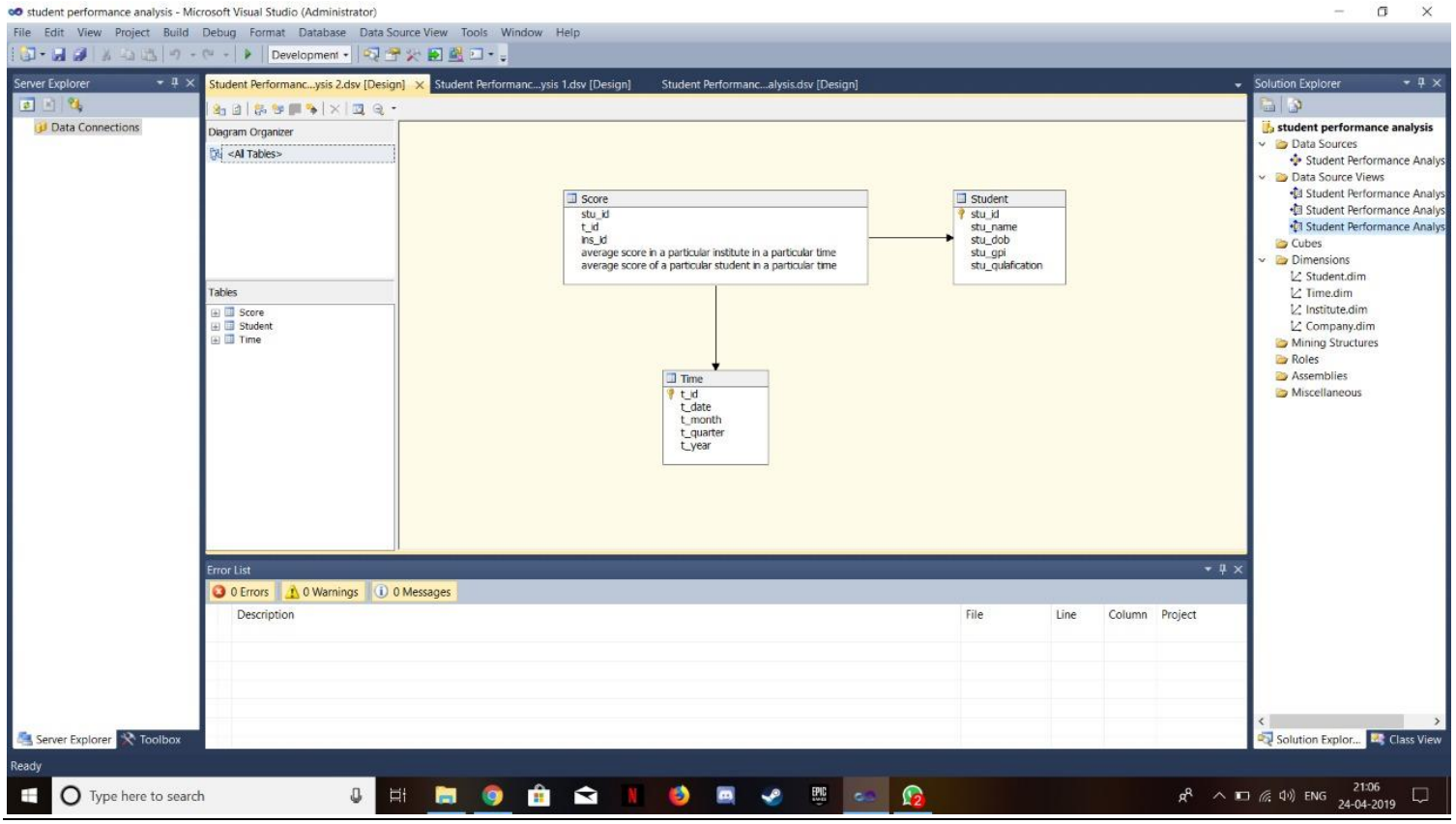
- The simplest scheme is a single table scheme, which consists of redundant fact table. The most common modelling paradigm is star schema, in which the data warehouse contains a large central fact table containing the bulk of data, with no redundancy, and a set of smaller attendant tables (dimension tables), one for each dimension.



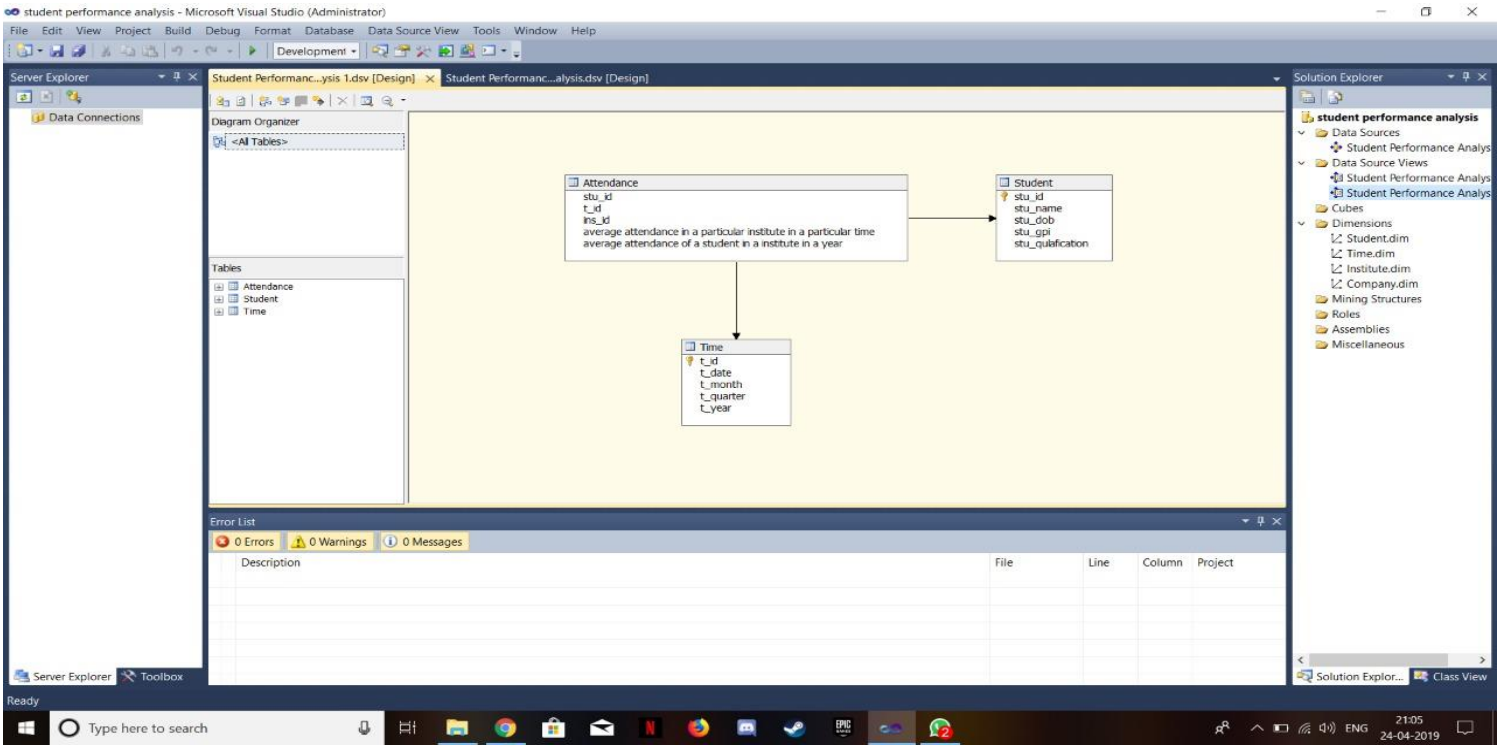
(Placement star schema)

- Measure facts: Average placements for a particular quarter, Average placements from a particular gpi range, etc.

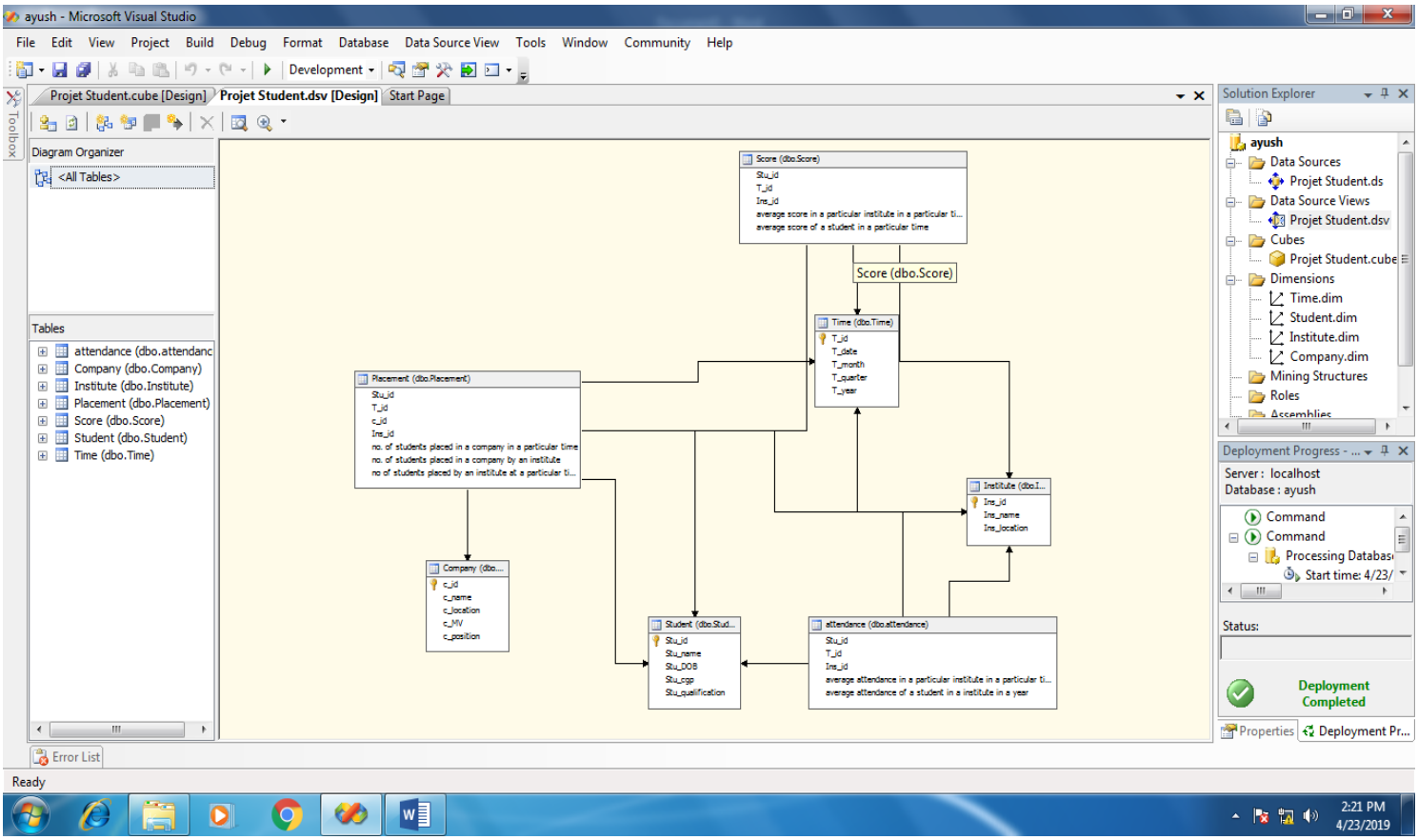
2. FACT : SCORE



3. FACT : ATTENDENCE

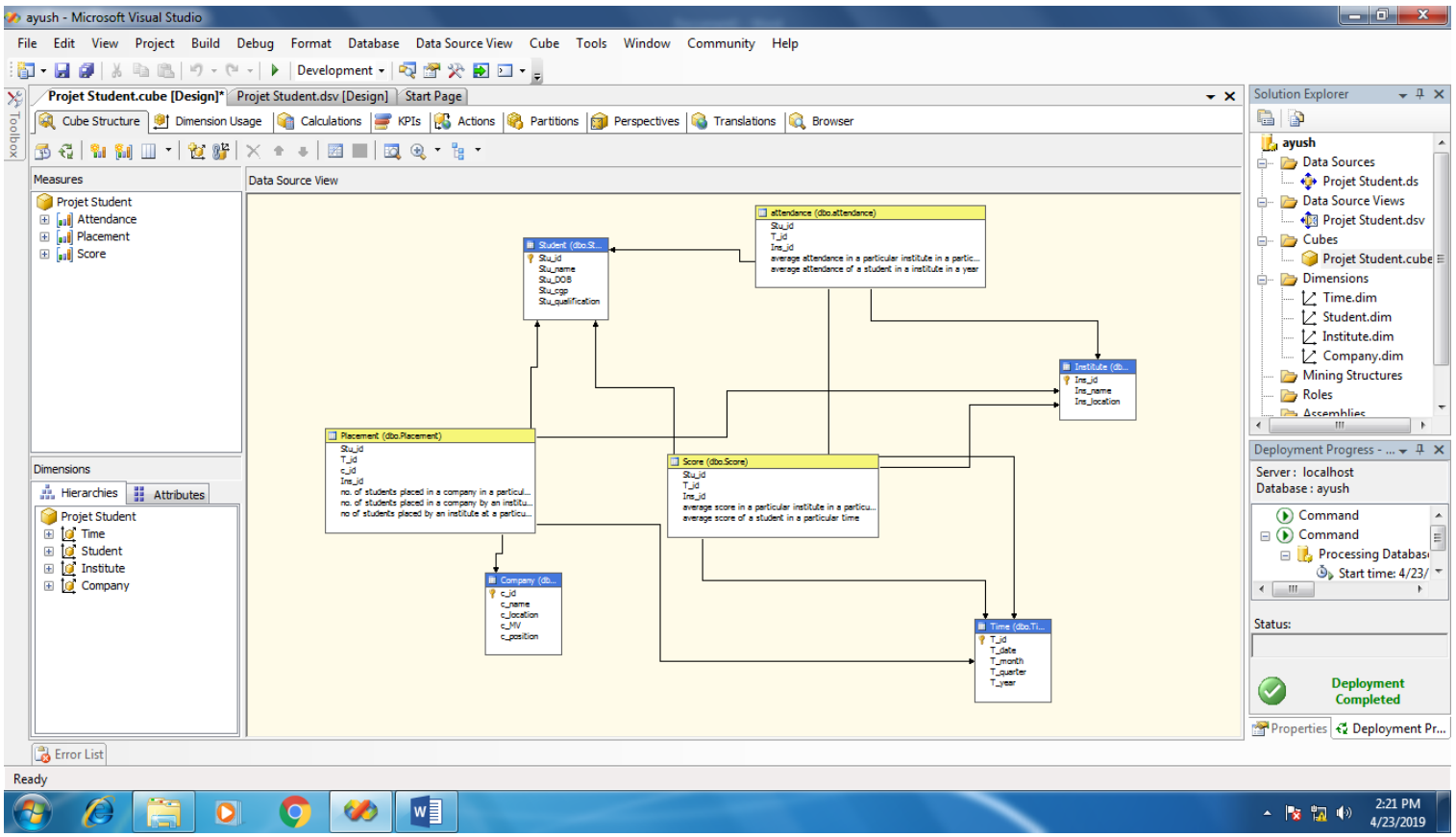


- FACT CONSTELLATION : Multiple fact tables share dimension tables. This schema is viewed as collection of stars hence called galaxy schema or fact constellation.

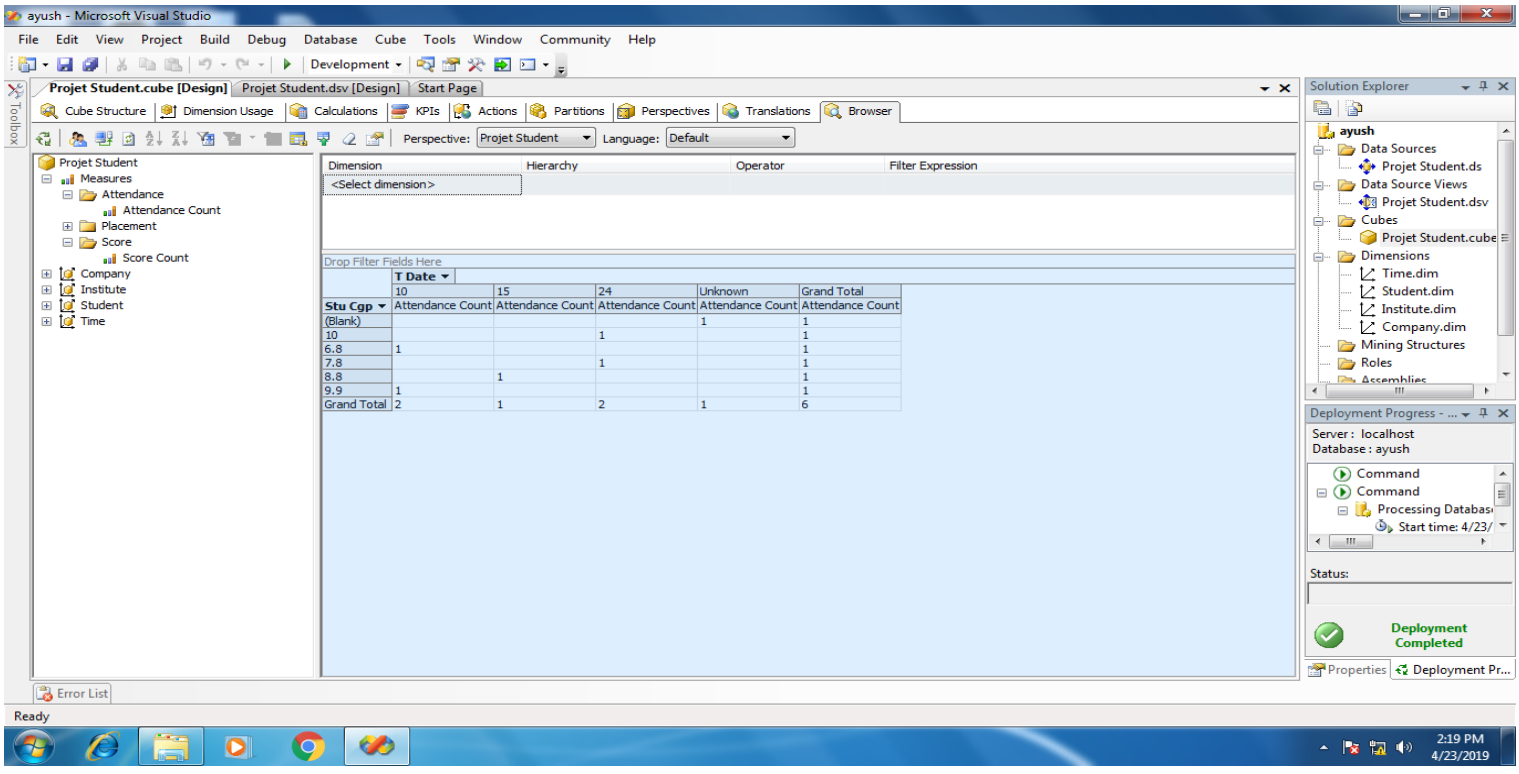


- OLAP CUBE :

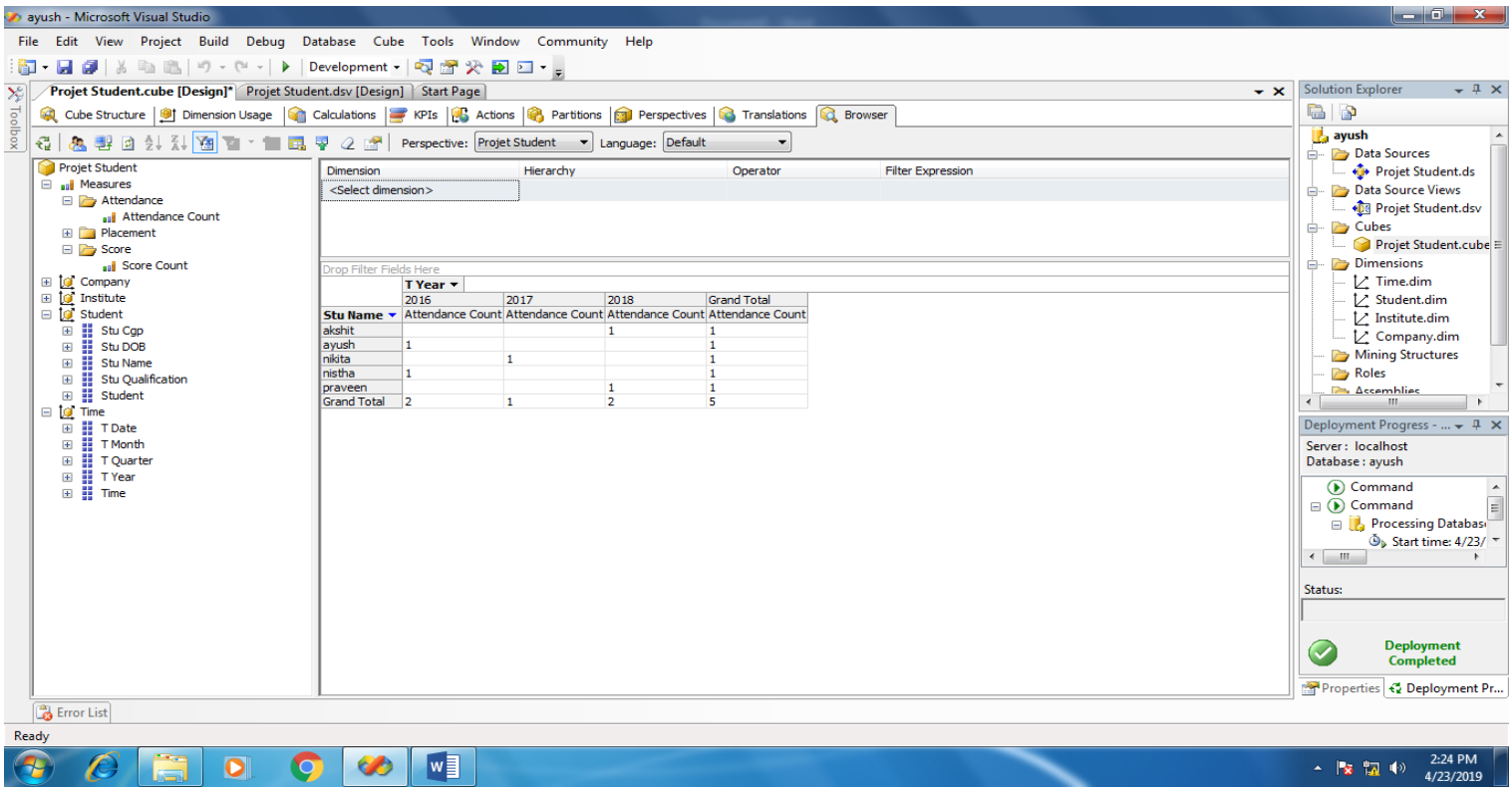
An OLAP cube is a multidimensional database that is optimized for data warehouse and online analytical processing (OLAP) applications. An OLAP cube is a method of storing data in a multidimensional form, generally for reporting purposes. In OLAP cubes, data (measures) are categorized by dimensions. OLAP cubes are often pre-summarized across dimensions to drastically improve query time over relational databases. The query language used to interact and perform tasks with OLAP cubes is multidimensional expressions (MDX). The MDX language was originally developed by Microsoft in the late 1990s, and has been adopted by many other vendors of multidimensional databases. Although it stores data like a traditional database does, an OLAP cube is structured very differently. Databases, historically, are designed according to the requirements of the IT systems that use them. OLAP cubes, however, are used by business users for advanced analytics. Thus, OLAP cubes are designed using business logic and understanding. They are optimized for analytical purposes, so that they can report on millions of records at a time. Business users can query OLAP cubes using plain English.



(An OLAP Cube)



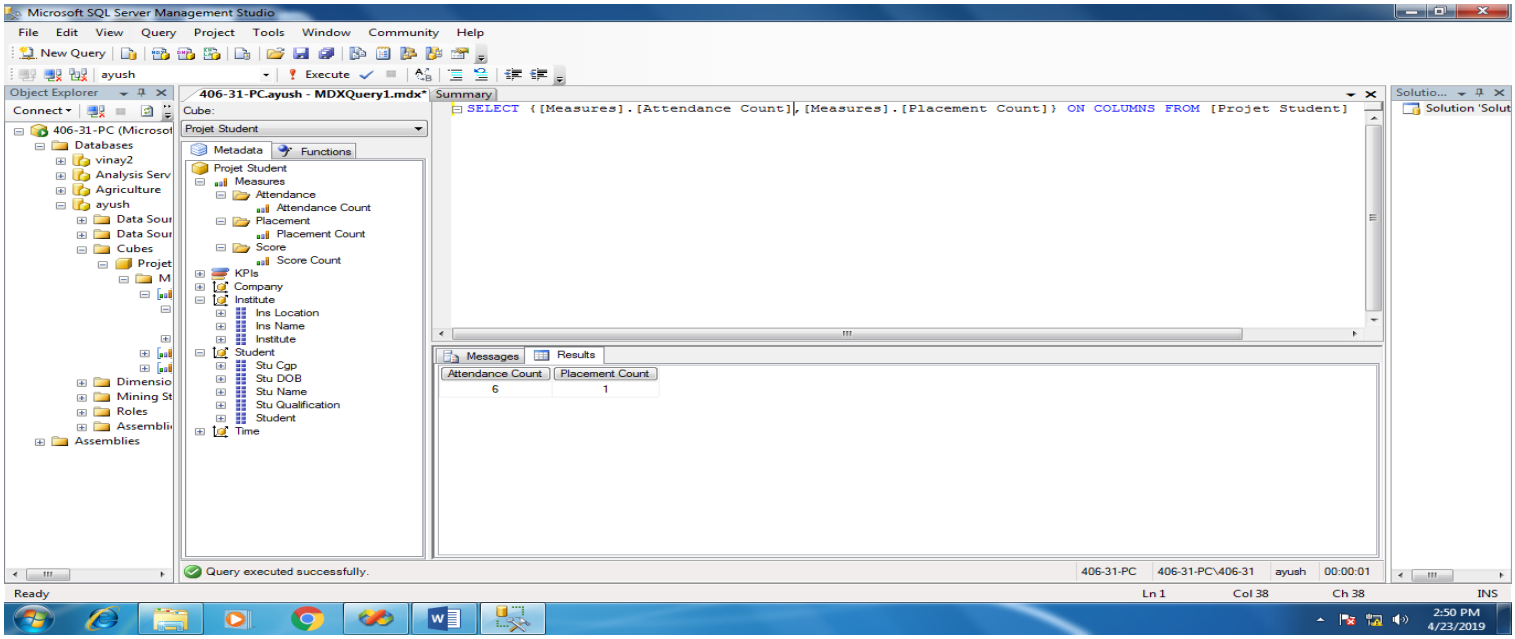
(Cgp - Attendance Dimensions)



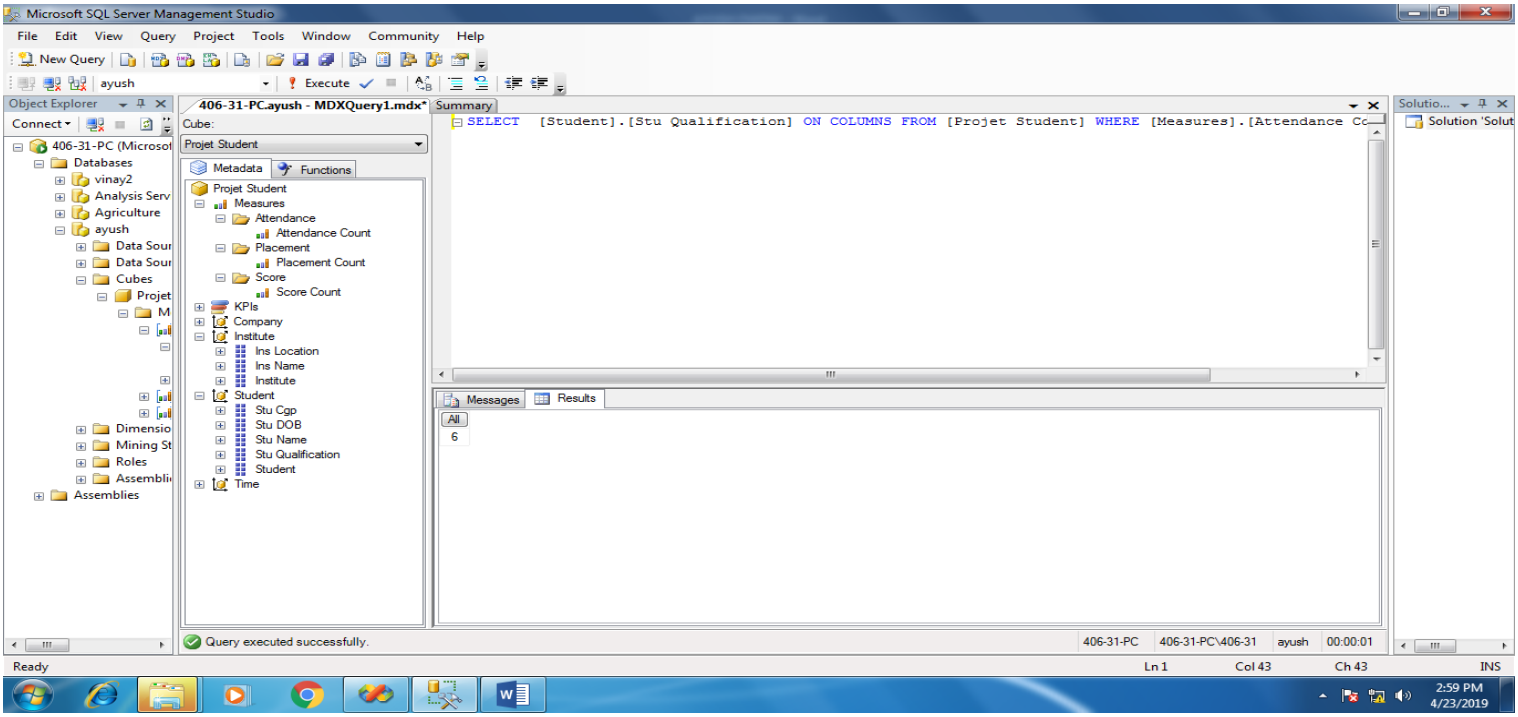
(Student –Attendance Dimension)

- Multidimensional Expression

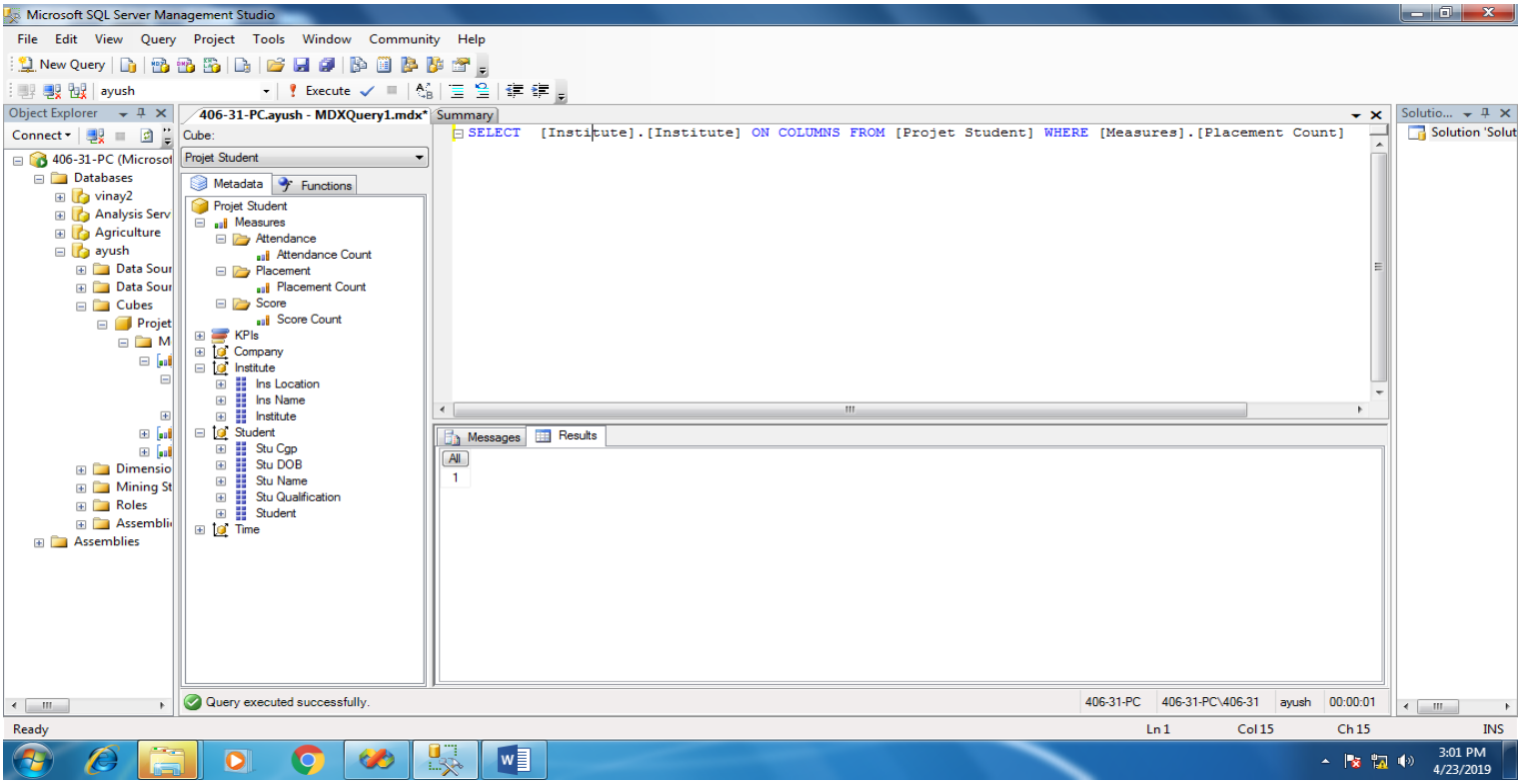
MDX is a query language for online analytical processing (OLAP) using a database management system. Much like SQL, it is a query language for OLAP cubes. It is also a calculation language, with syntax similar to spreadsheet formulas. The MultiDimensional eXpressions (MDX) language provides a specialized syntax for querying and manipulating the multidimensional data stored in OLAP cubes. While it is possible to translate some of these into traditional SQL, it would frequently require the synthesis of clumsy SQL expressions even for very simple MDX expressions. MDX has been embraced by a wide majority of OLAP vendors and has become the standard for OLAP systems.



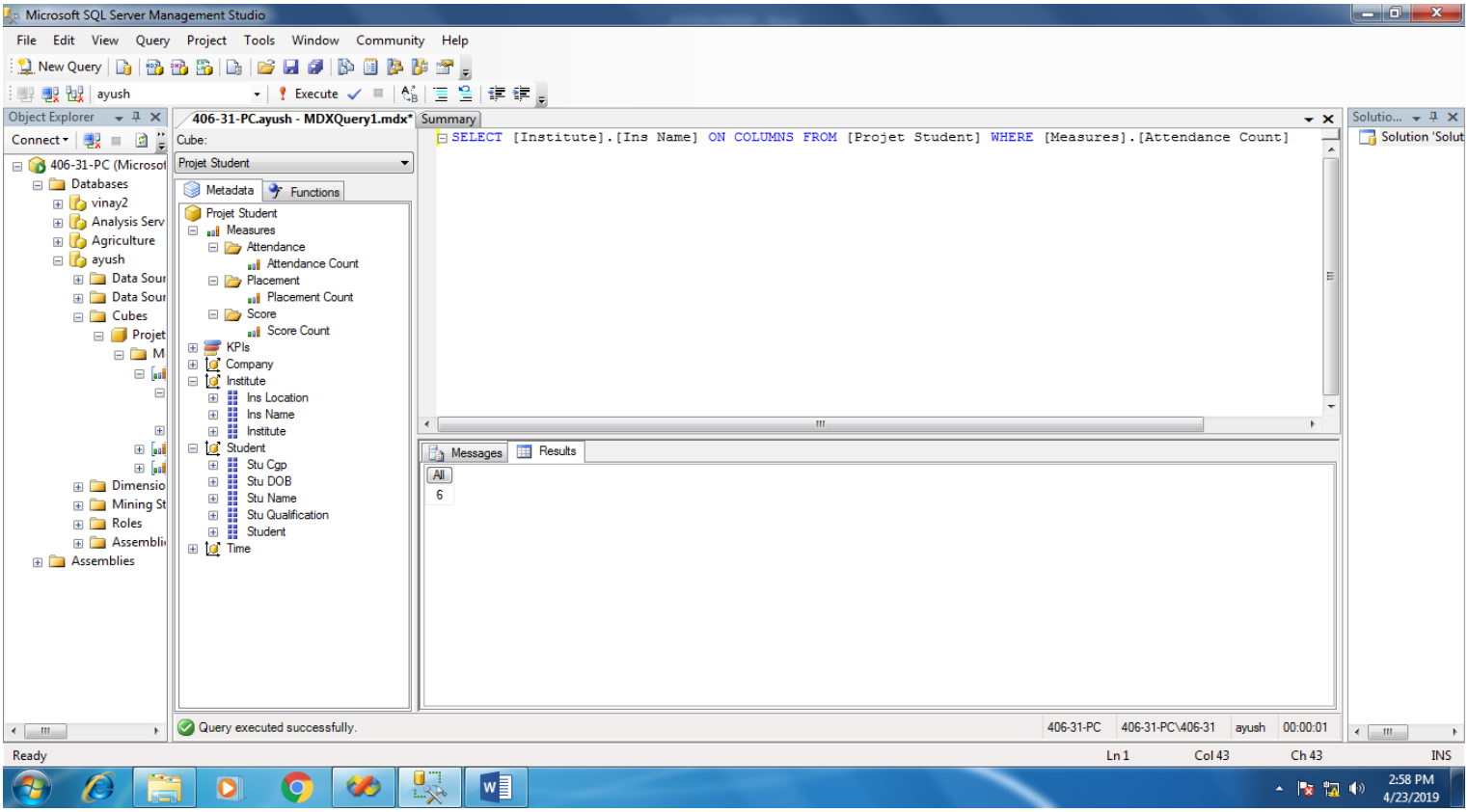
- 1) Select [Student].[Stu.Qualification]
  2) ON COLUMNS FROM [Project Student]
  3) WHERE [Measures]. [Attendance Count]



- 1) Select[Institute].[Institute] ON COLUMNS
  2) FROM [Project Student]
  3) WHERE [Measures].[Placements Count]



- 1) SELECT [Institute].[Ins Name]
  2) ON COLUMNS FROM [Project Student]
  3) WHERE [Measures].[Attendance Count]



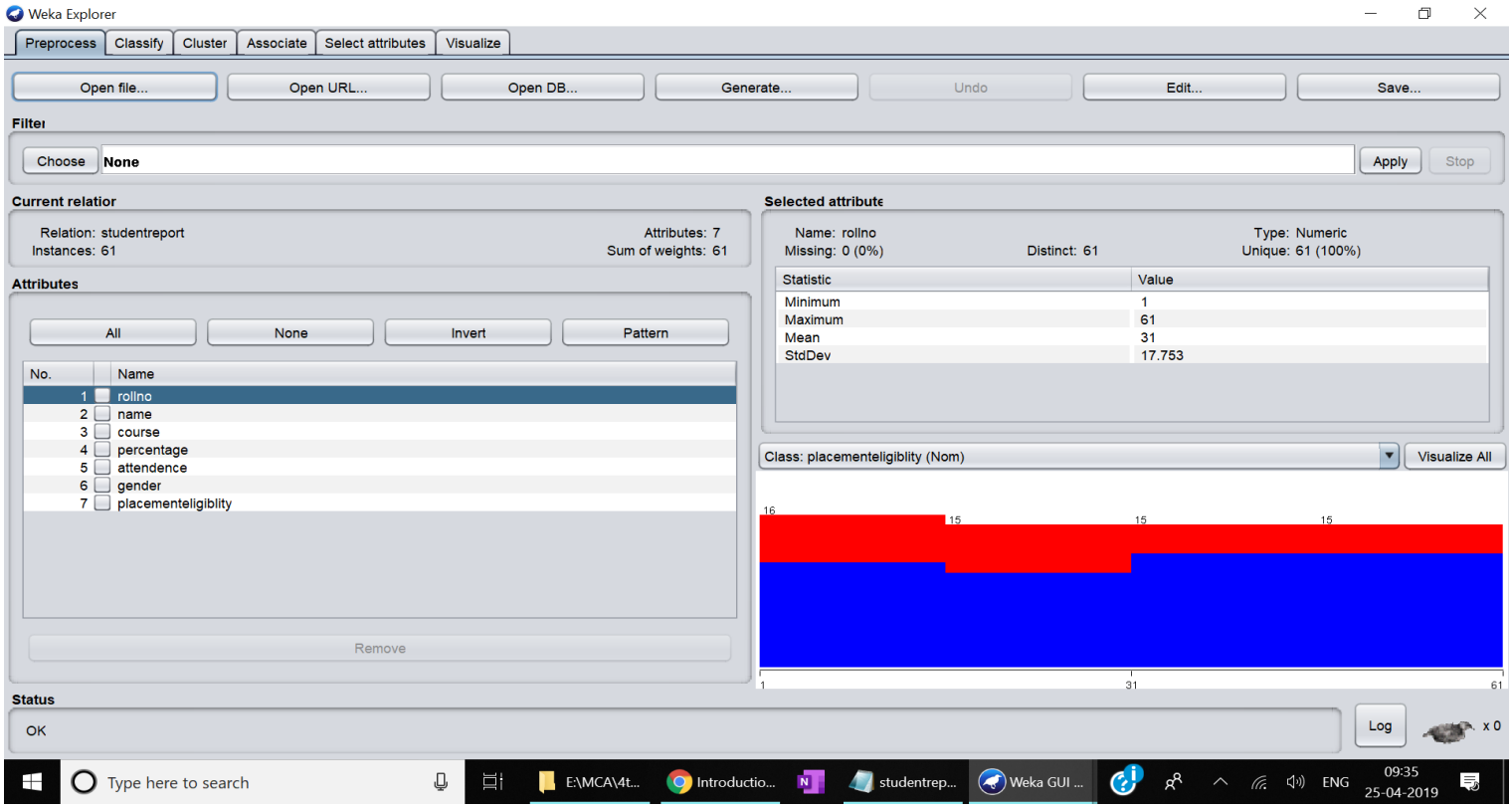
WEKA

We have used a data mining software named as WEKA for this project. For the purposes of this study, we select WEKA (Waikato Environment for Knowledge Analysis) software that was developed at the University of Waikato in New Zealand. WEKA tool supports to a wider range of algorithms & very large data sets. The WEKA (pronounced Waykuh) workbench contains a collection of visualization tools & algorithms. WEKA is open source software issued under the GNU General Public License. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. The original non-java version of WEKA was a TCL/TK, but the recent java based version is WEKA 3(1997), is now used in many different application areas, in particular for education & research. WEKA's main user interface is Explorer. The Experimenter is also there by which we can compare WEKA's machine learning algorithms' performance. The Explorer interface has many panels by which we can access to main components of workbench. The Visualization tab allows visualizing a 2-D plot of the current working relation, it is very useful. In this study WEKA toolkit 3.8.1 is used for generating the association rules and prediction of result. WEKA supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of WEKA's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). WEKA provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using WEKA.

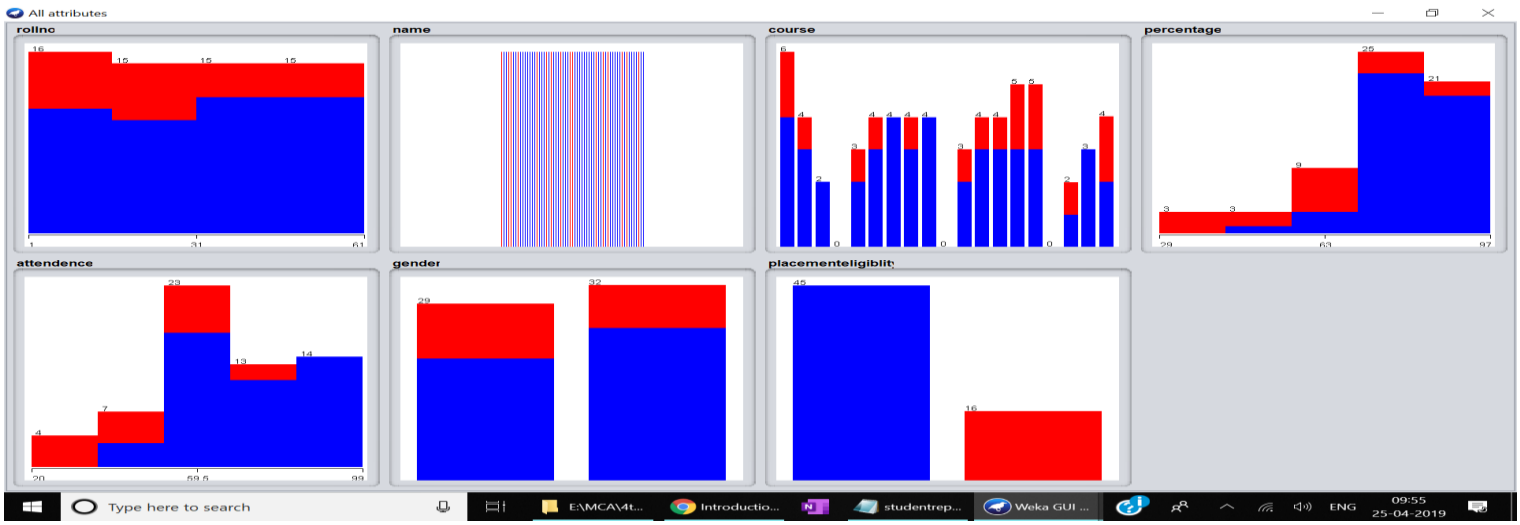
- ## Design and Implementation

The followings are the step by step process of our project evaluation. Dataset and attribute selection- We have collected a dummy dataset contains the result of students of last semester. The dataset contains 61 instances and 7 attributes. It has some missing values also. The data file has to be in either in 'CSV' format or 'ARFF' format. Here is the sample of our dataset which is in 'ARFF' format.

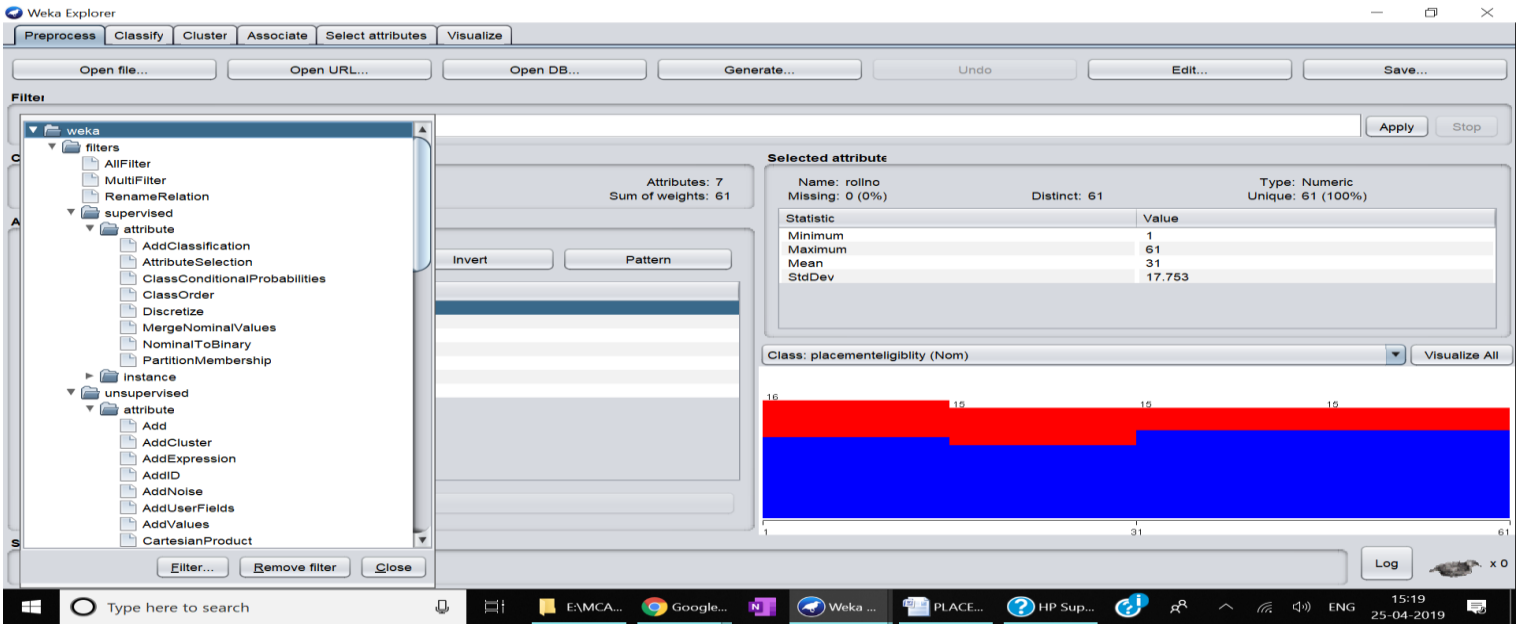
- **Preprocessing Data** : Preprocessing is the first step of evaluation of this project. For our project we will choose WEKA Explorer interface. Here the source data file is selected from local machine. After loading the data in Explorer, we can refine the data by selecting different options which is known as 'Data Cleaning' and can also select or remove attributes as per our need. The following is the preprocessed of our dataset. Left hand side of the above screen shows detail of relation name, number of attributes and number of records. Right hand side gives details of attribute values, type, and number of distinct values. Specification of every attribute is displayed in the right bottom of the screen.



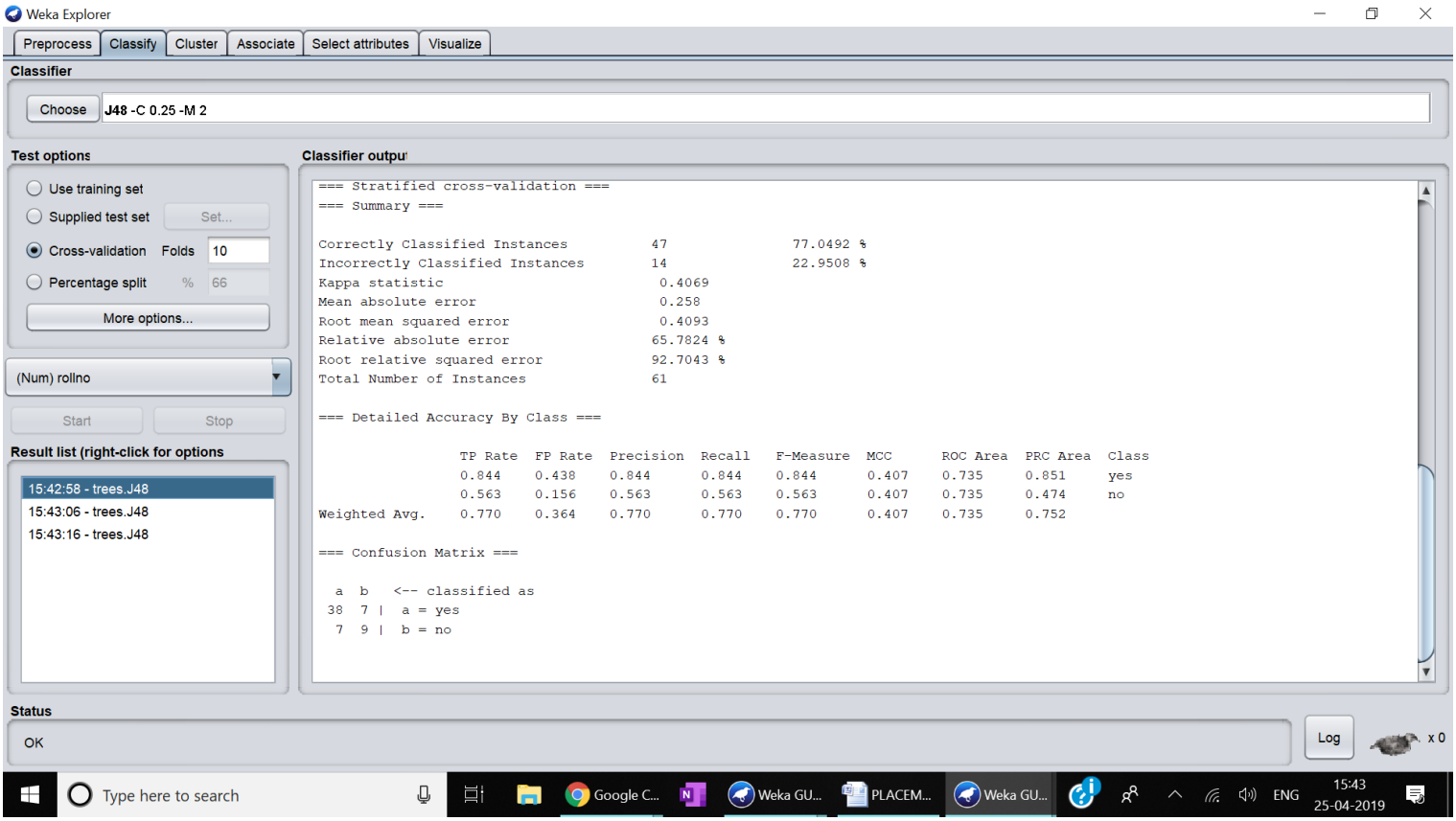
Here, we can select different attributes among 7 and visualize them through histogram on right panel. We can also see the distinct & unique attributes in our data set, as roll no of every student is different that's why here its show 61 distinct & unique attributes. We can also visualize them all at once :



- **Filters -**The preprocess section allows filters to be defined that transform the data in various ways. The Filter box is used to set up the filters that are required. There are mainly two categories of filters-Supervised and Unsupervised. In unsupervised category filters the dataset is contained with any numeric values we have to covert it nominal values(as Association in WEKA can only support nominal values) by using 'Numeric To Nominal' filter under attribute section of Unsupervised filters. Another one filter we will apply named as 'Replace Missing Values' which will replace all missing values of our dataset and will make the dataset able to perform 'Approximate Association Rule Generation'.



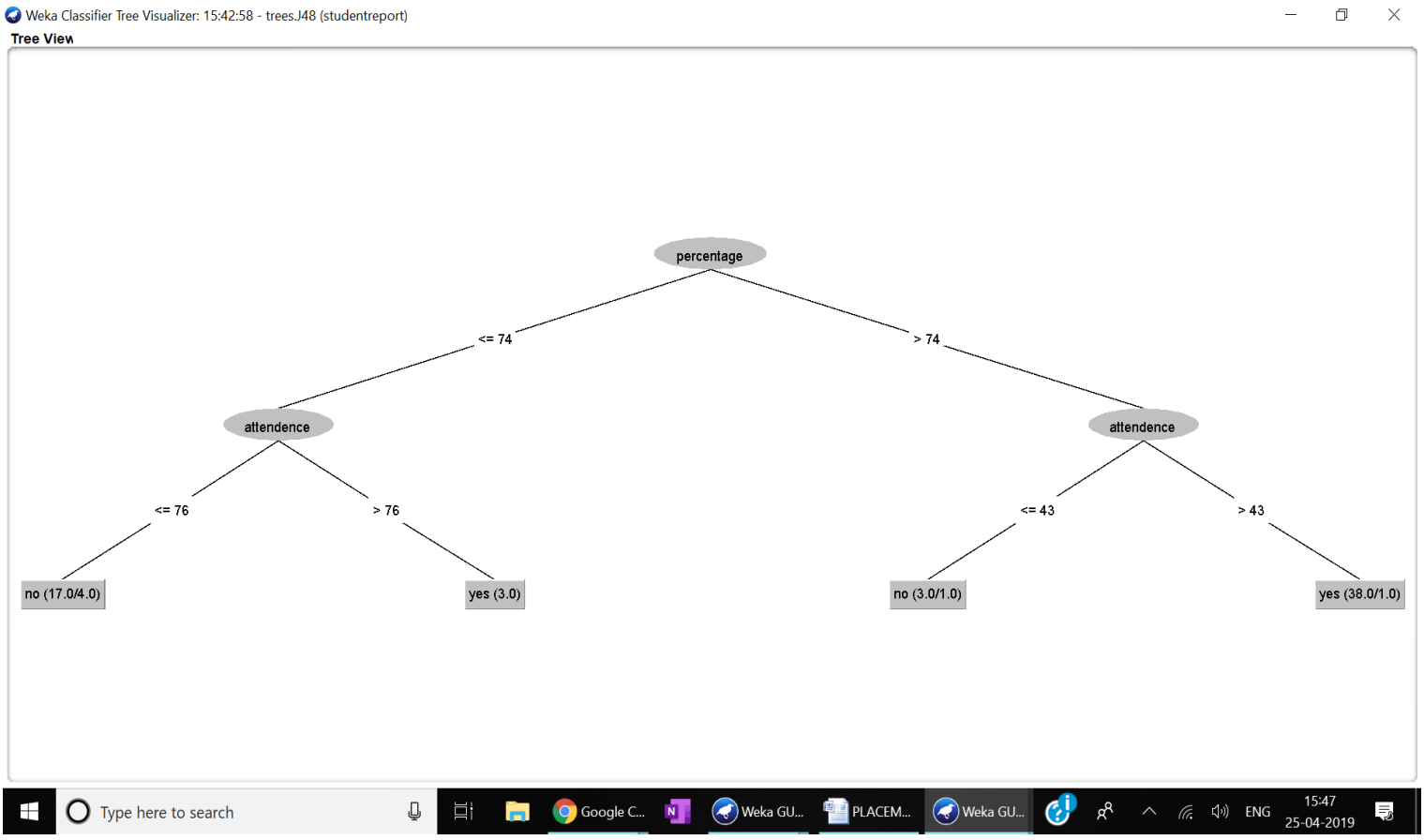
- **Classification -** To predict nominal or numeric quantities we have classifiers in WEKA. For our prediction purpose we have to choose a classifier. We will select a standard classifier named as J48 for classification.



From the above example we can say J48 is a good classifier as it gives an accuracy of 77.04% because the percentage of correctly classified instances is often called accuracy or sample accuracy. The correctly and incorrectly classified instances show the percentage of test instances that were correctly and incorrectly classified. The raw numbers are shown in the confusion matrix, with a , b, c and d representing the class labels. Here are some others factor in classifier output-

- TP Rate : rate of true positives (instances correctly classified as a given class).
- FP Rate : rate of false positives (instances falsely classified as a given class).
- Precision : proportion of instances that are truly of a class divided by the total instances classified as that class .
- Recall : proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate).
- F-Measure : A combined measure for precision and recall calculated as 2 * Precision * Recall / (Precision + Recall)

Tree Visualizes

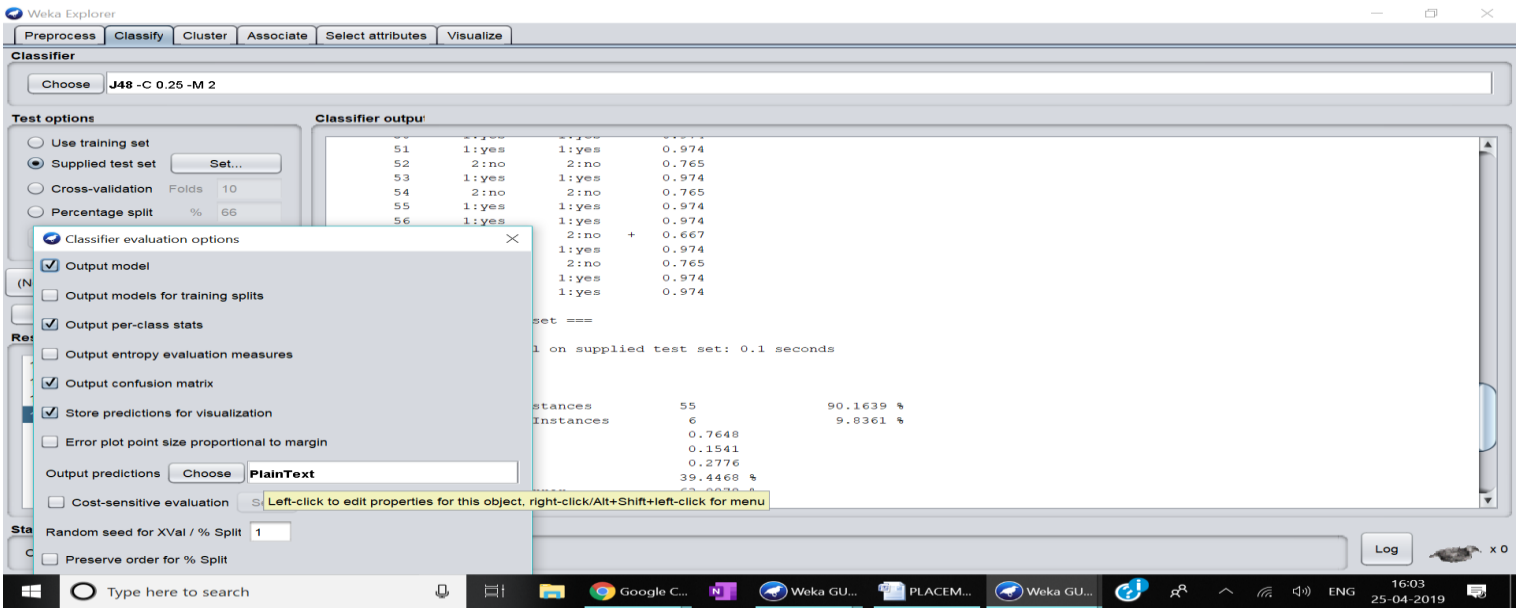


Inbuilt WEKA Prediction :

i. At first we have to load our dataset into WEKA Explorer.
ii. After loading our dataset go to classify tab and start classification by J48 classifier. In classify tab Test options can be 'Cross Validation'.
iii. Then change the test option into 'Supplied Test set and load the same dataset as test file.

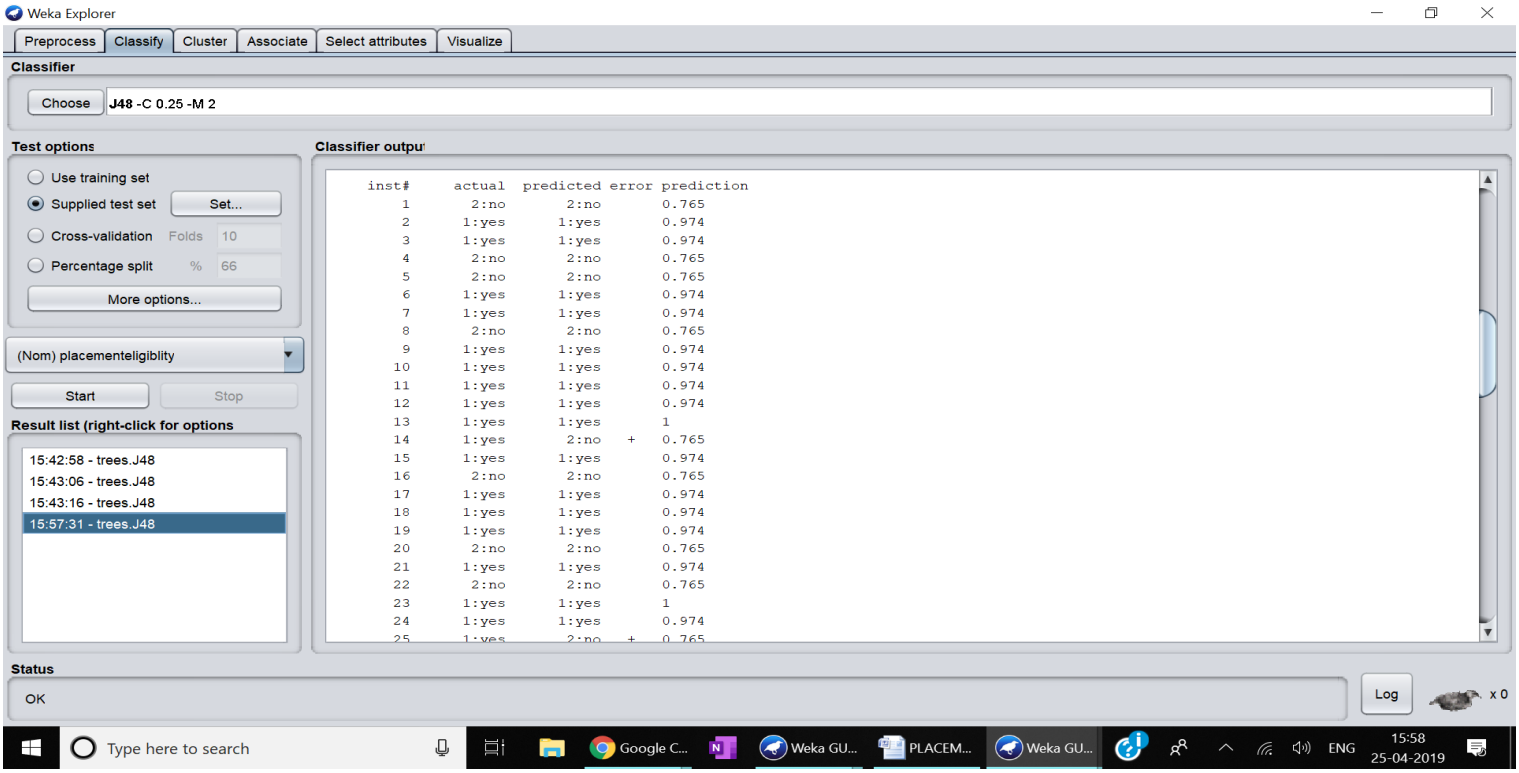
iv. After loading the test file in classify tab under test options select more options to go to the 'classifier evaluation options'.

v. Now in classifier evaluation options select the output predictions and choose 'Plaintext' as prediction.



vi. Now perform the classification of test set by J48 Classifier. Now in classifier output it will be seen that WEKA performs predictions on test set. In the result the 'Predicted error' column contains predicted value of Result attribute which is the predicted result of original result of train data set. Thus WEKA performed prediction.

Here one more column is generated named as 'Prediction' which has some certain values for all instances. The 'Prediction is defined as the difference between the probability predicted for the actual result and the highest probability predicted for the other results. One hypothesis as to the good performance of boosting algorithms is that they increase the margins on the training data and this gives better performance on test data. In the following picture for some instances the '+' sign signifies that WEKA prediction fails to match the actual result.



- References/Bibliography

1. Samrat Singh, Dr. Vikesh Kumar , "Performance Analysis of Engineering Students for Recruitment Using Classification Data Mining Techniques ",IJCSET February 2013.
2. [2] M. Goyal and R. Vohra, “Applications of Data Mining in Higher Education”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue2, No 1, March 2012.
3. [3] Jason Brownlee ,"How to Save Your Machine Learning Model and Make Predictions in Weka", August 3, 2016.
4. [4] Neelam Naik & Seema Purohit, “Prediction of Final Result and Placement of Students using Classification Algorithm”International Journal of Computer Applications (0975 – 8887) Volume 56– No.12, October 2012
5. [5] Alaa M.El-Halees,Mohammed M. Abu Tair, “Mining Educational Data to Improve Students ‘Performance: A Case Study”, International Journal of Information and Communication Technology Research, 2012.
6. [6] B.K. Bharadwaj and S. Pal,“Data Mining: A prediction for performance improvement using classification”, International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.
7. [7] Suchita Borkar, K. Rajeswari, "Predicting Students Academic Performance Using Education Data Mining ", IJCSMC,Vol. 2, Issue. 7, July 2013, pg.273– 279. [8] Randhir Singh, M.Tiwari, Neeraj Vimal,"An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education", 2013.
8. https://docs.microsoft.com/en-us/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-2017
9. https://docs.microsoft.com/en-us/sql/ssdt/previous-releases-of-sql-server-data-tools-ssdt-and-ssdt-bi?view=sql-server-2017
10. https://www.mssqltips.com/sqlservertutorial/2000/sql-server-analysis-services-ssas-tutorial/
11. https://www.cs.waikato.ac.nz/ml/weka/documentation.html
12. https://pdfs.semanticscholar.org/2456/a979fbe8eea47b90d625c1a064162be5382e.pdf