Index

1. Introduction

The US Adult Income data gives us the information about the income of people depending on various factors such as education, occupation, age, location, race etc. The dataset has the data of 32561 people. This consists the anonymous information of people without their names. The dataset has people from different countries however most of the people has their native as United States.

Tasks to do

Major task:

Create a Data Warehouse for this project

1. Business questions that the data warehouse could help answer.
2. Draw Starnet with the aim to identify the dimensions and concept hierarchies for each dimension. This should be based on the lowest level information you have access to.
3. Use the Starnet footprints to illustrate how the business queries can be answered with the design.
4. Implement a star or snowflake schema using SQL Server Management Studio (SSMS).
5. Load the data from the csv files to populate the tables.
6. Use SQL Server Data Tools to build a multi-dimensional analysis service solution, with a cube designed to answer your business queries. Make sure the concept hierarchies match your Starnet design. Paste the cube diagram to your Power BI Dashboard.
7. Use Power BI to visualize the data returned from your business queries.

2. Business Queries

1. Finding number of people having income maximum 50000 USD per annum also the number of people who have income more than 50000 USD a year from US, India, China, Germany and Mexico of all age group.
2. Finding the number of young (18-30) people having income more than 50000 USD a year from United States.
3. Finding total young (age 18-30) private employees having earnings more than USD 50000 a year from Mexico and Canada.
4. Finding the number of people having income more than 50000 USD per year and who works 10-30 hours a week, from United States.
5. Finding the preferred education of people who are earning more than 50000 USD a year and along with that people who are getting maximum 50000 USD a year working for Federal-govt from United-States.
6. Getting all the males and females working as a sales representative, Exce-managerial and a Machine-op-inspector having maximum income USD 50000 a year from US, Canada, Cuba having race white.

3. Tables

a. Fact Table :

i. **Fact _user_info**: This table here in this project is our fact table.

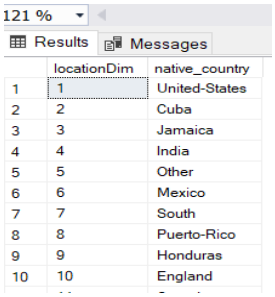
This table consists measures such as user_count, age, working hour, final weight. This table will have the foreign keys of the dimension table in order make the relationship with the dimension tables for this project .

Results Messages

	user_id_fact	income	age	finalWeight	totalworkHrs	location_dim_id	race_dim_id	occupation_dim_id	sector_dim_id	gender_dim_id	relation_dim_id	education_dim_id
1	1	<=50K	39	77516	40	1	3	11	4	1	1	1
2	2	<=50K	50	83311	13	1	3	2	5	1	2	1
3	3	<=50K	38	215646	40	1	3	12	9	1	3	2
4	4	<=50K	53	234721	40	1	5	12	9	1	2	3
5	5	<=50K	28	338409	40	2	5	13	9	2	2	1
6	6	<=50K	37	284582	40	1	3	2	9	2	2	4
7	7	<=50K	49	160187	16	3	5	7	9	2	4	5
8	8	>50K	52	209642	45	1	3	2	5	1	2	2

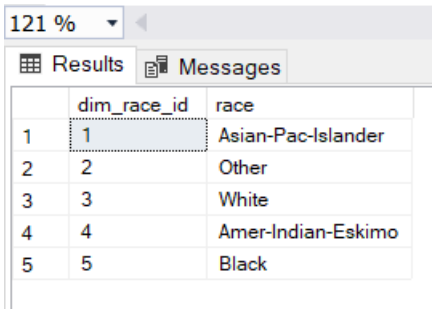
b. Dimension Tables:

i. **LocationDim**: This tables will consist the countries of the users having unique ID for each country.



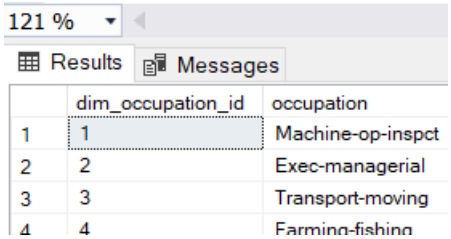
	locationDim	native_country
1	1	United-States
2	2	Cuba
3	3	Jamaica
4	4	India
5	5	Other
6	6	Mexico
7	7	South
8	8	Puerto-Rico
9	9	Honduras
10	10	England

ii. **RaceDim:** This dimension table will have the race such as Black, White ,Asia-Pacific etc.



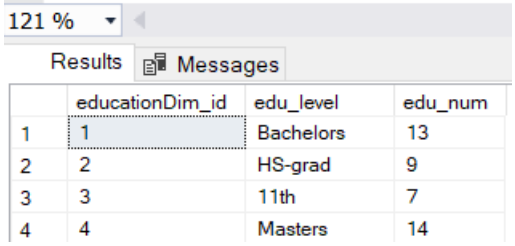
	dim_race_id	race
1	1	Asian-Pac-Islander
2	2	Other
3	3	White
4	4	Amer-Indian-Eskimo
5	5	Black

iii. **OccupationDim:** This table will have the occupation for all the user with a unique primary key. Example : Machine-op-incpct



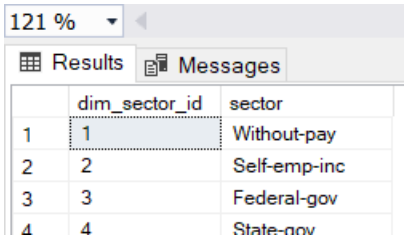
	dim_occupation_id	occupation
1	1	Machine-op-inspct
2	2	Exec-managerial
3	3	Transport-moving
4	4	Farming-fishing

iv. **EducationDim:** This dimension table education will have the education qualification along with the education number of a particular user.



	educationDim_id	edu_level	edu_num
1	1	Bachelors	13
2	2	HS-grad	9
3	3	11th	7
4	4	Masters	14

v. **SectorDim:** This tables holds the information of different sector where people work or have a business of their own.



121 %

Results | Messages

	dim_sector_id	sector
1	1	Without-pay
2	2	Self-emp-inc
3	3	Federal-gov
4	4	State-gov

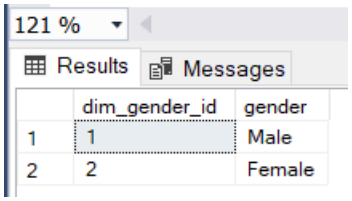
vi. **RelationDim:** Relation dimensional tables gives us the data of marital status of people. Example : Married, Never-married, Divorced.

121 %

Results | Messages

	relationDim_id	marital_status
1	1	Never-married
2	2	Married-civ-spouse
3	3	Divorced

vii. **GenderDim:** This dimension table holds the gender of people. Example : Male, female.

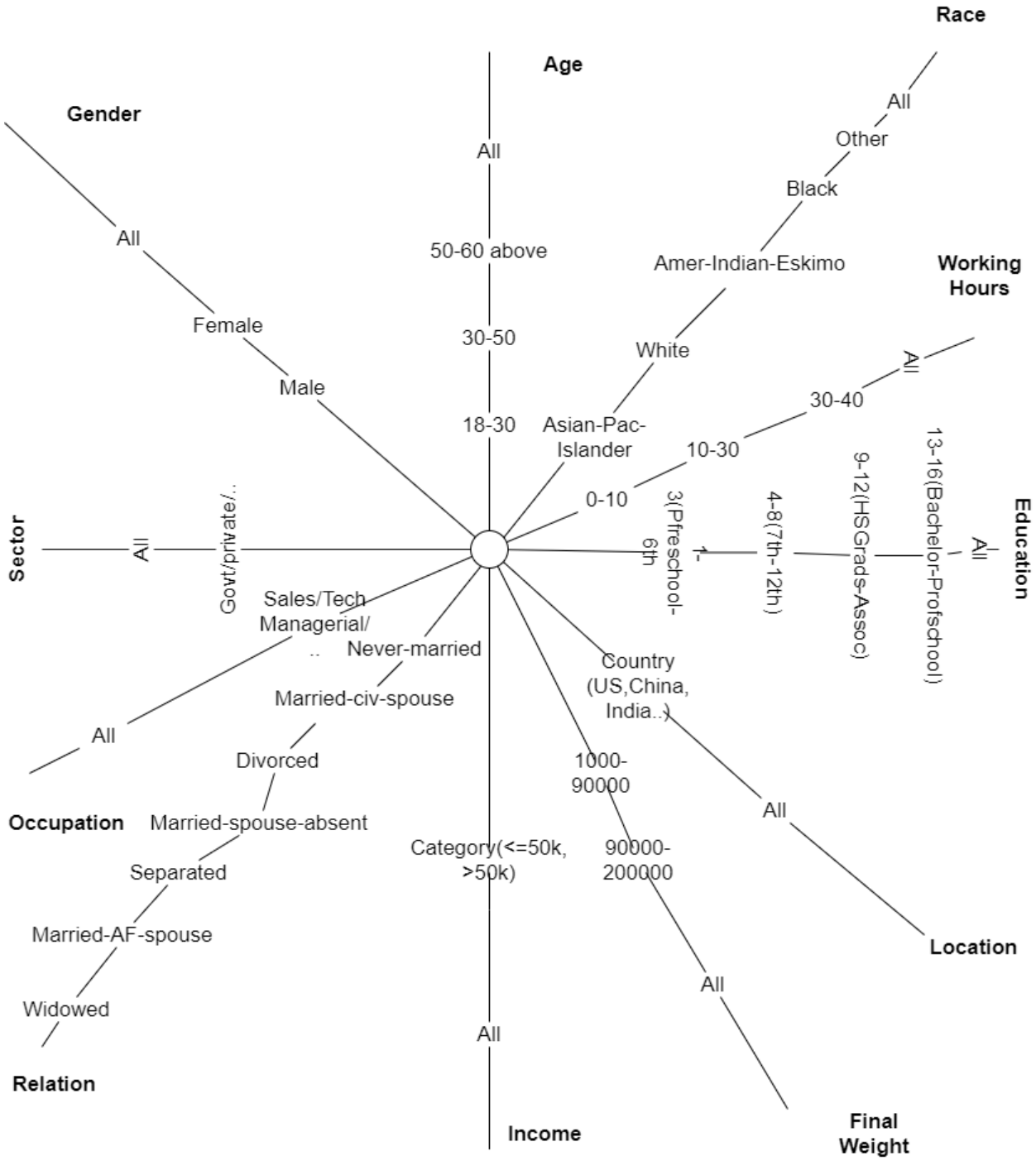


121 %

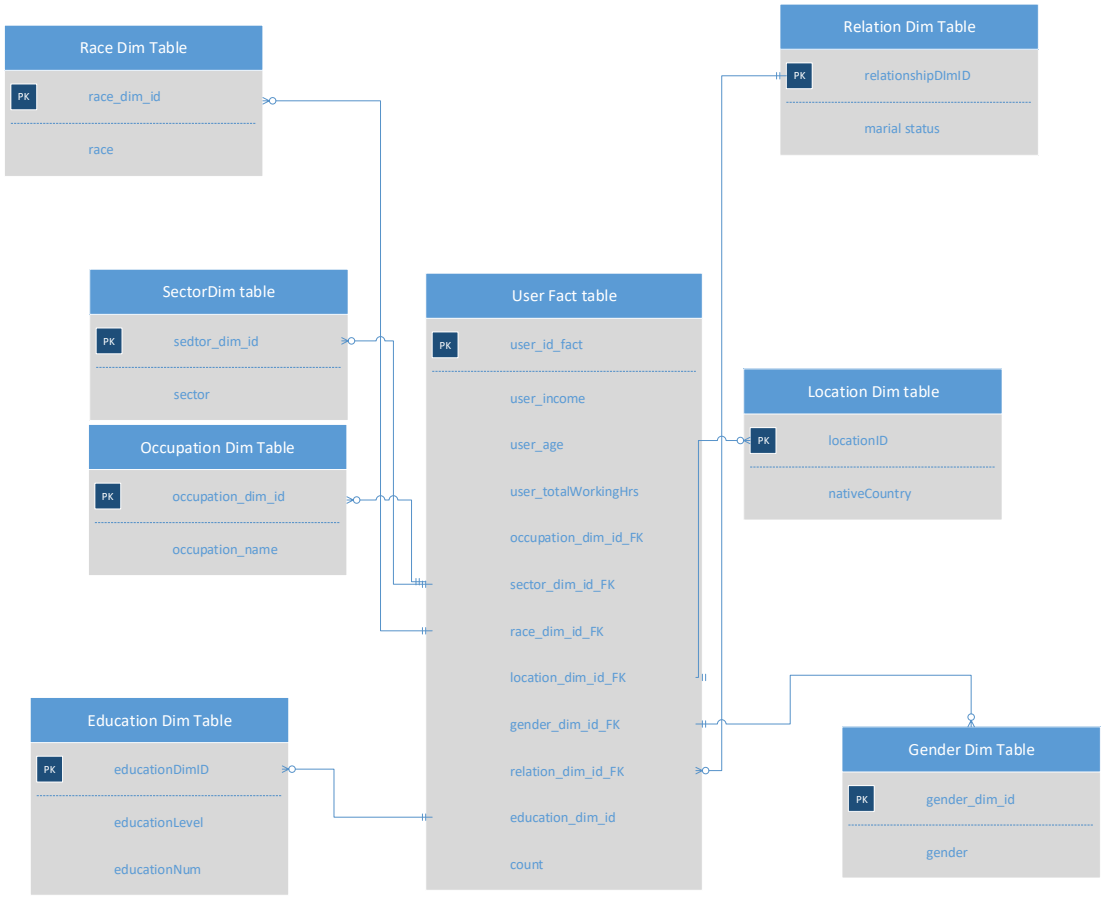
Results | Messages

	dim_gender_id	gender
1	1	Male
2	2	Female

4. Starnet schema and the diagram



Startnet Diagram

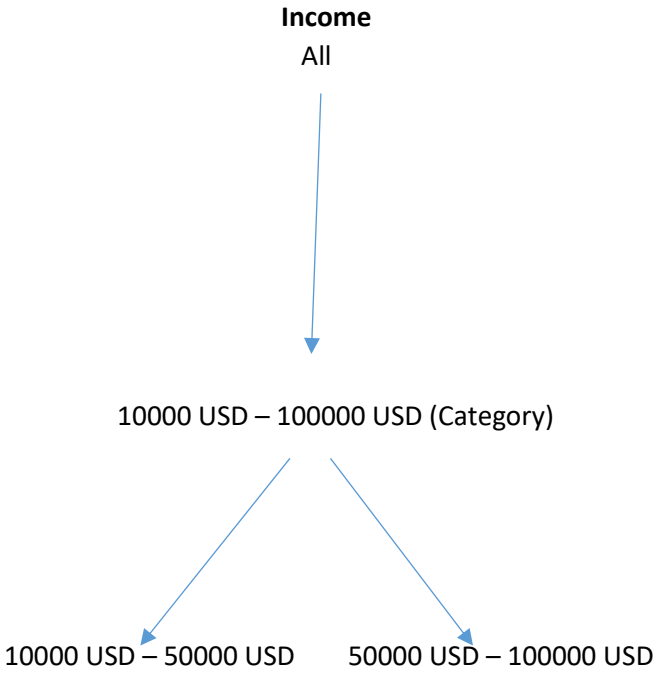


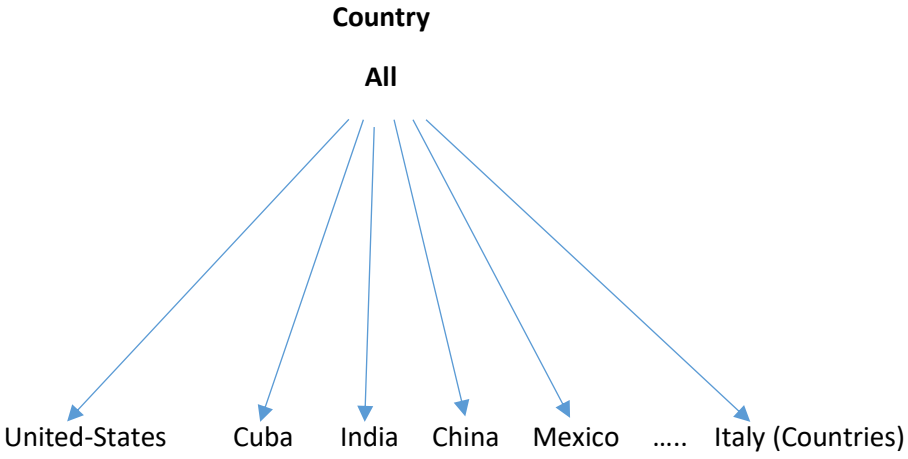
Numeric Measures of fact tables:

2. Age
3. Final Weight
4. Total working hours
5. User fact ID count

5. Concept hierarchy

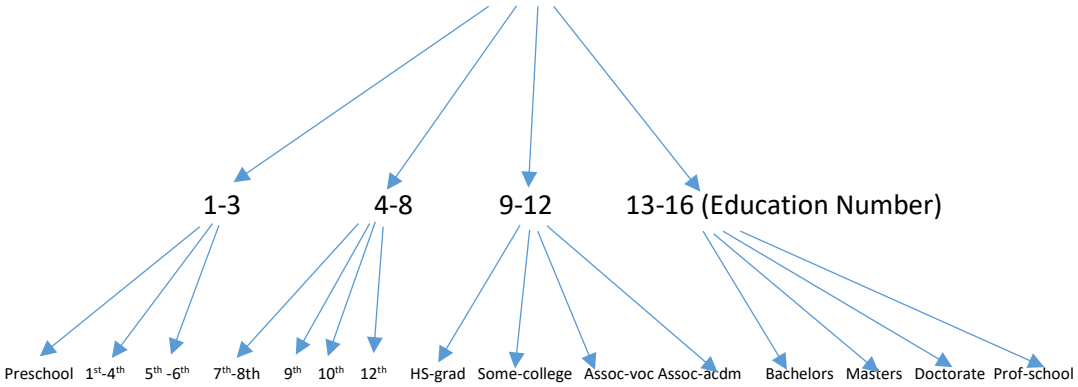
Here, the Drill-down process has been used to find the concept hierarchy for the US-Adult-Income. We will get information by using the lowest level hierarchy.





Education

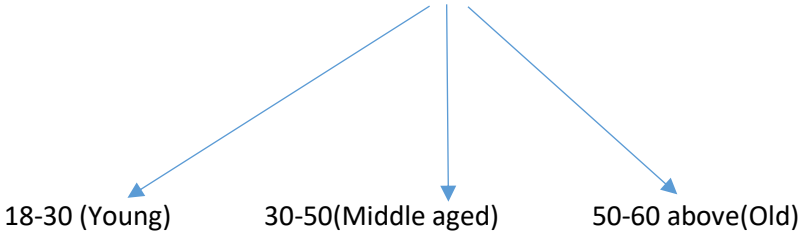
1-16 (All)



(Education Level)

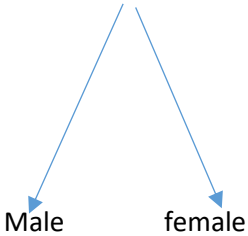
Age

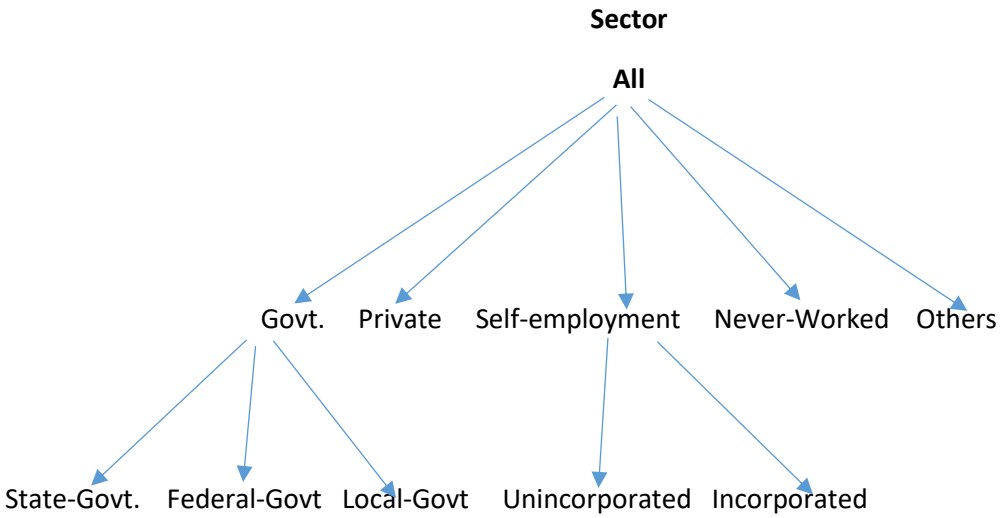
18-80

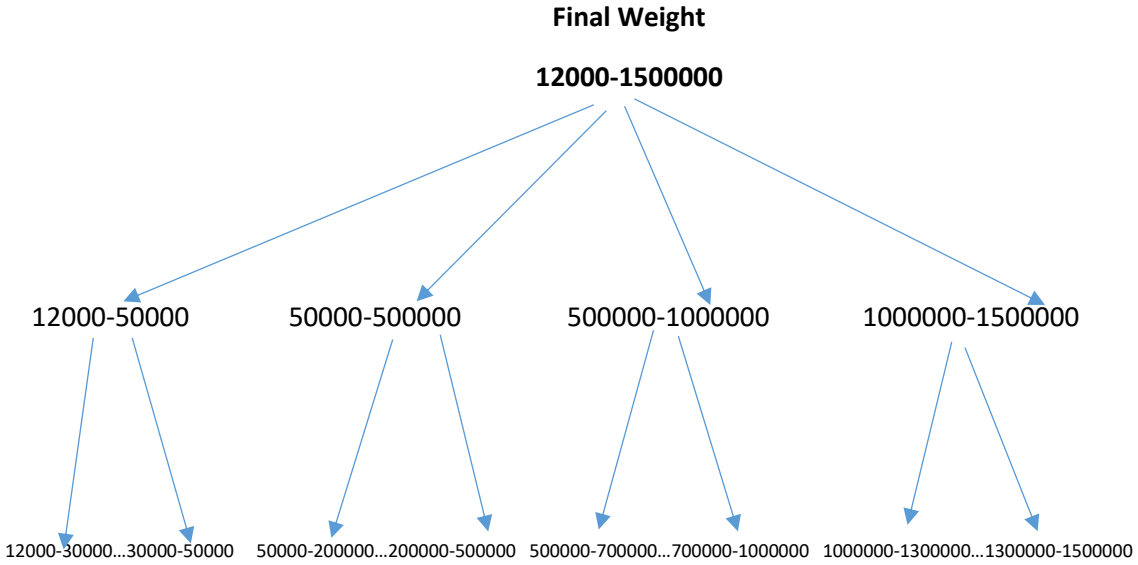


Gender

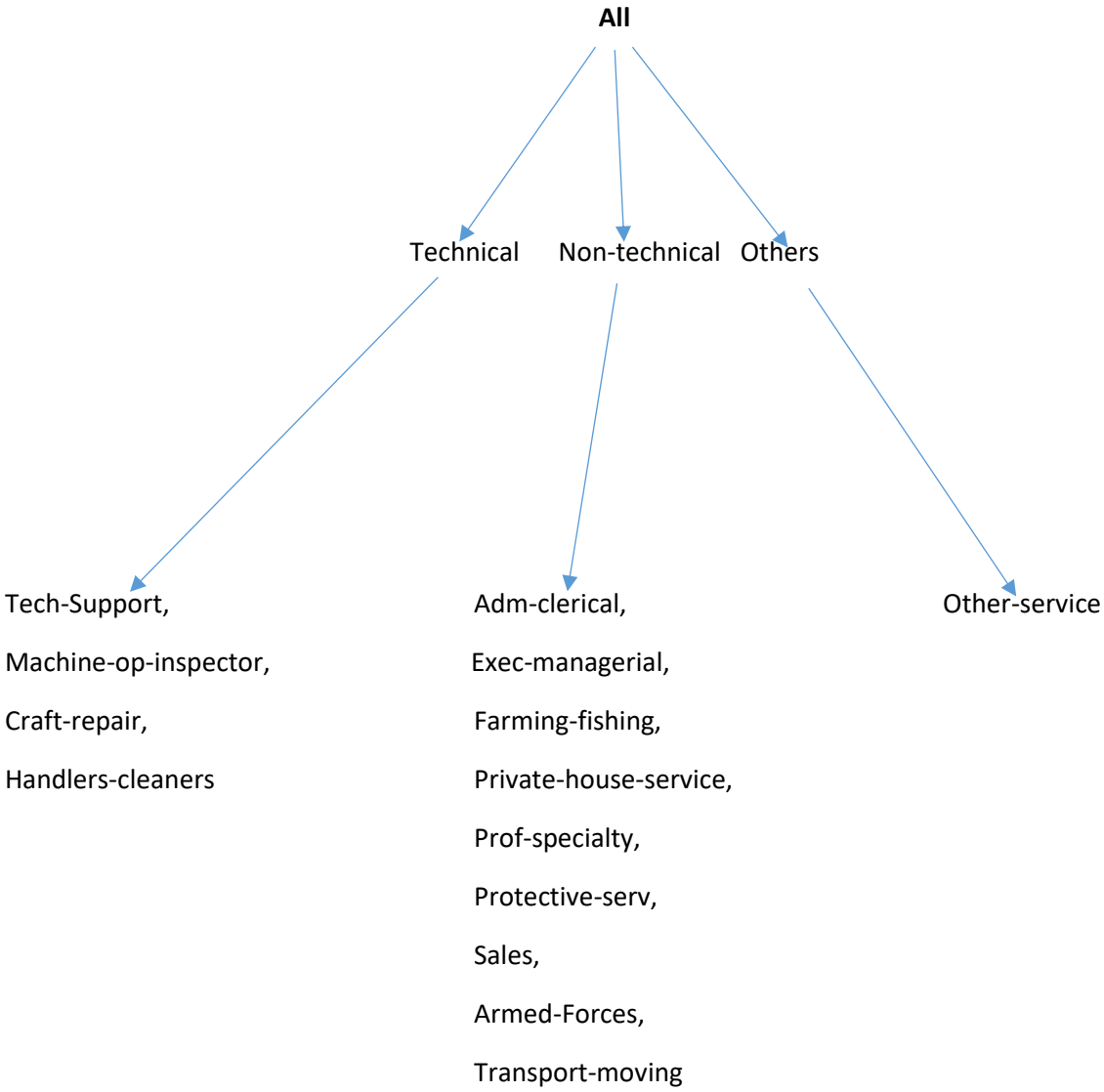
All



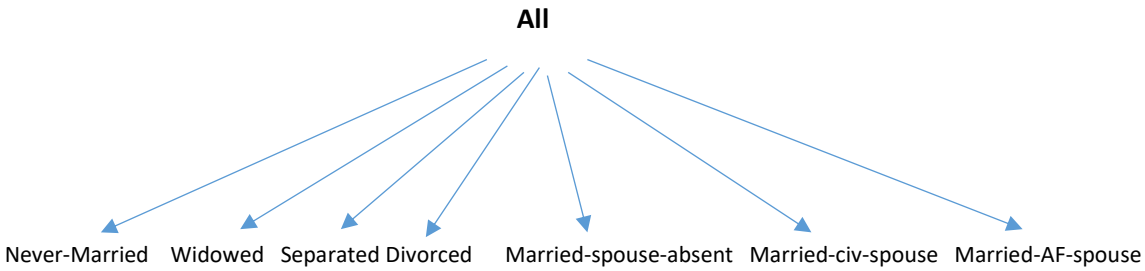

Sector
All
Govt.
Private
Self-employment
Never-Worked
Others
State-Govt.
Federal-Govt
Local-Govt
Unincorporated
Incorporated


Final Weight
12000-1500000
12000-50000
50000-500000
500000-1000000
1000000-1500000
12000-30000...30000-50000
50000-200000...200000-500000
500000-700000...700000-1000000
1000000-1300000...1300000-1500000

Occupation

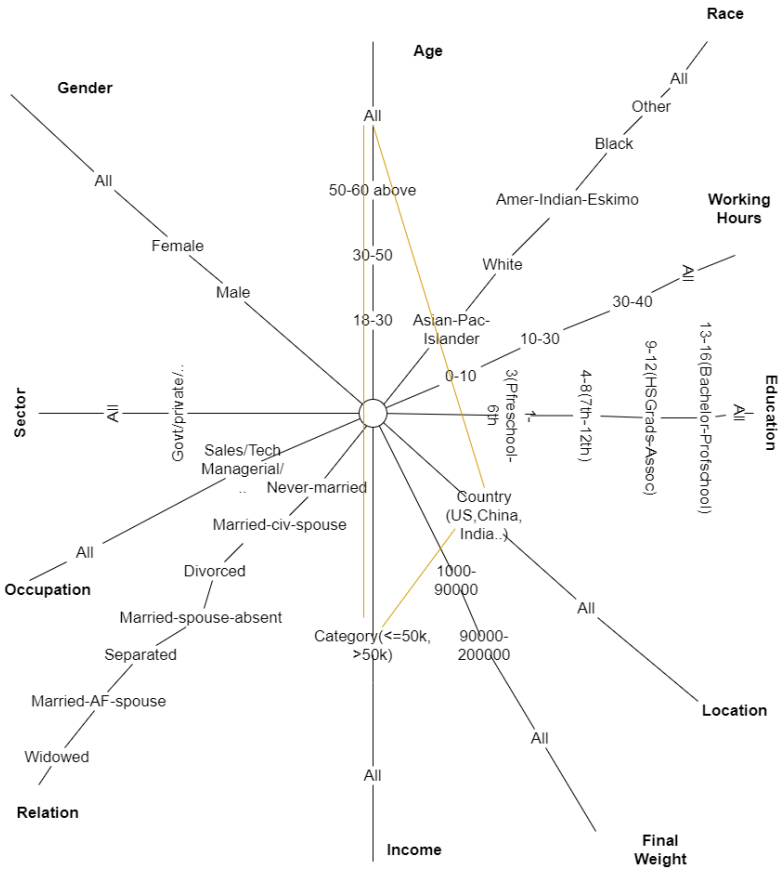


Marital status

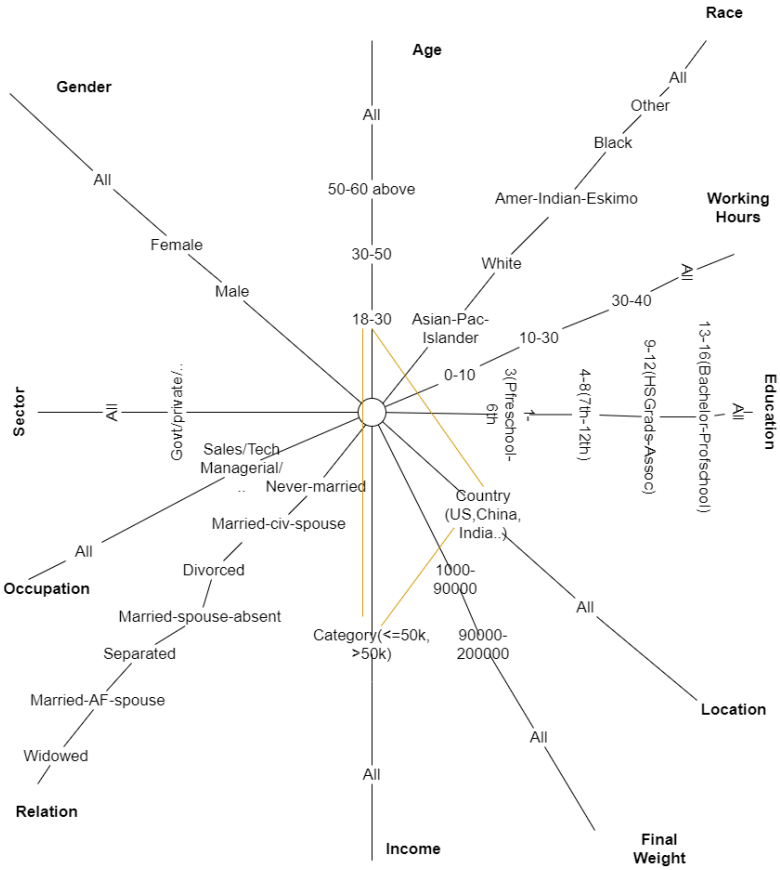


6. Starnet footprints for business queries

1. We will go with the mentioned footprint in order to solve the first business query. This footprint helps in finding the number of people having income maximum 50000 USD per annum also we are finding the number of people who have income more than 50000 USD a year in the same business query from US, China, India, Germany and Mexico of all age group.



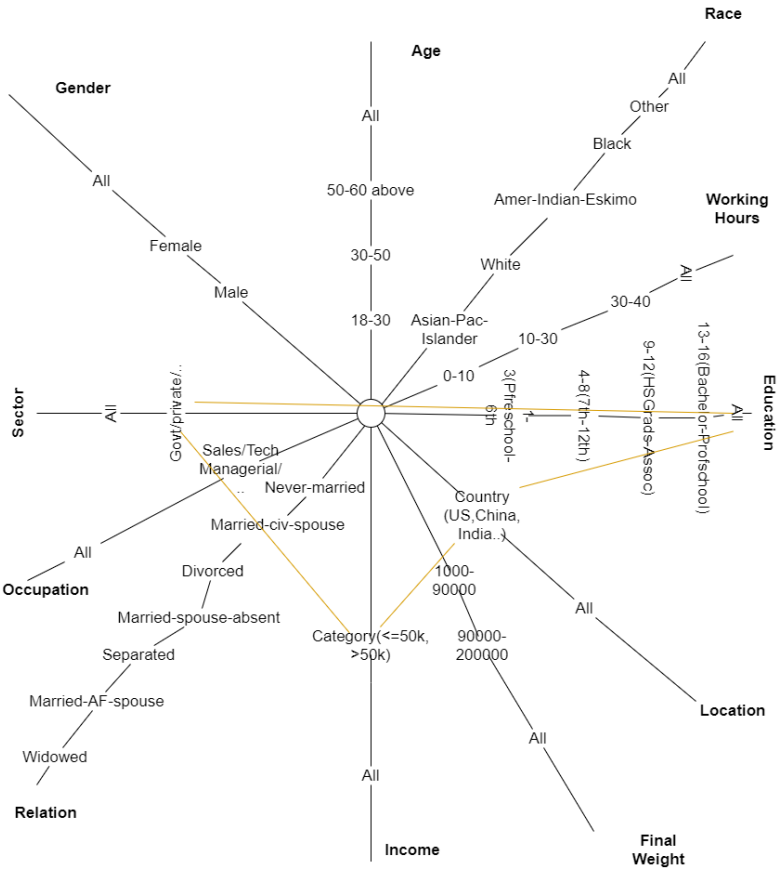
2. This foot print focuses on second business query which finds the number of young people having income more than 50000 USD in a year in United States.



3. Finding the number of young employees of a private sector companies in Mexico and Canada who is having income over 50000 USD per annum.

4. 4th business query also helps in find the people having highest wages rate per week because in this query the working hour is less and income is high. So, we are finding the people having income 50000 USD per annum having shifts from 10- 30 hours a week for people from US.

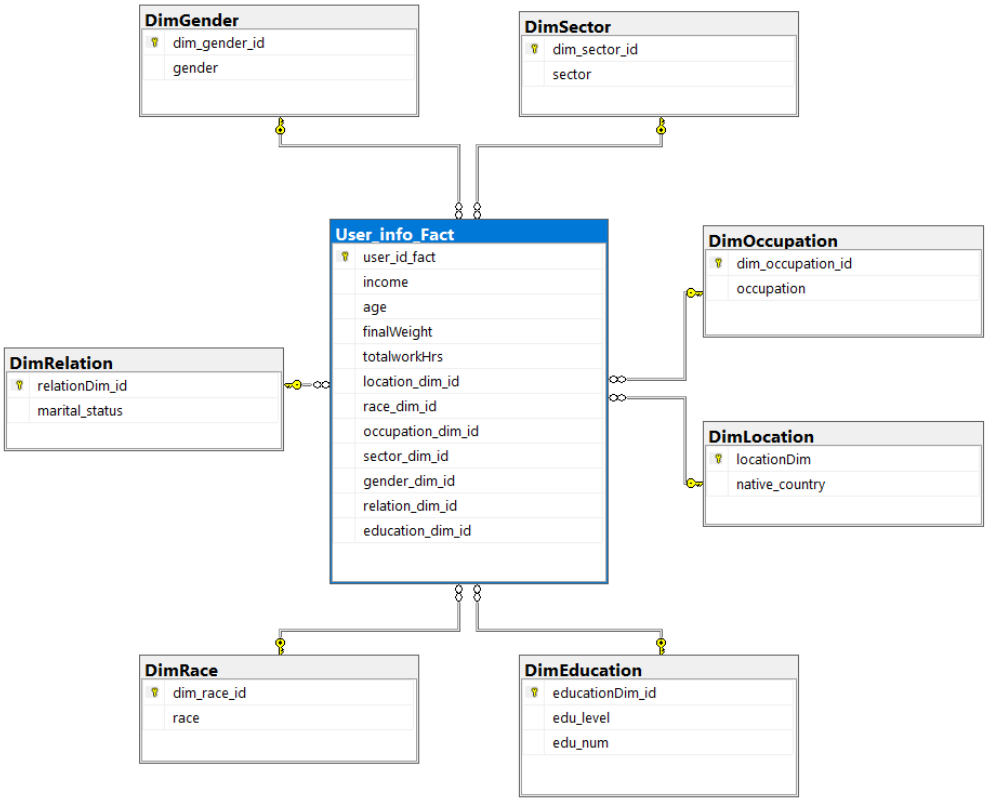
5. Finding the preferred education people who is earning more than 50000 USD a year and maximum 50000 USD a year from United-Stated.



7. This starnet foot print helps in getting all the males and females working as a sales representative, executive manager and a machine-op-inspector(Occupation: sales,exce-managerial, machine-op-inspector) having maximum income USD 50000 in a year from US, Canada and Cuba having race white.

7. Starnet database diagram (ER diagram) :

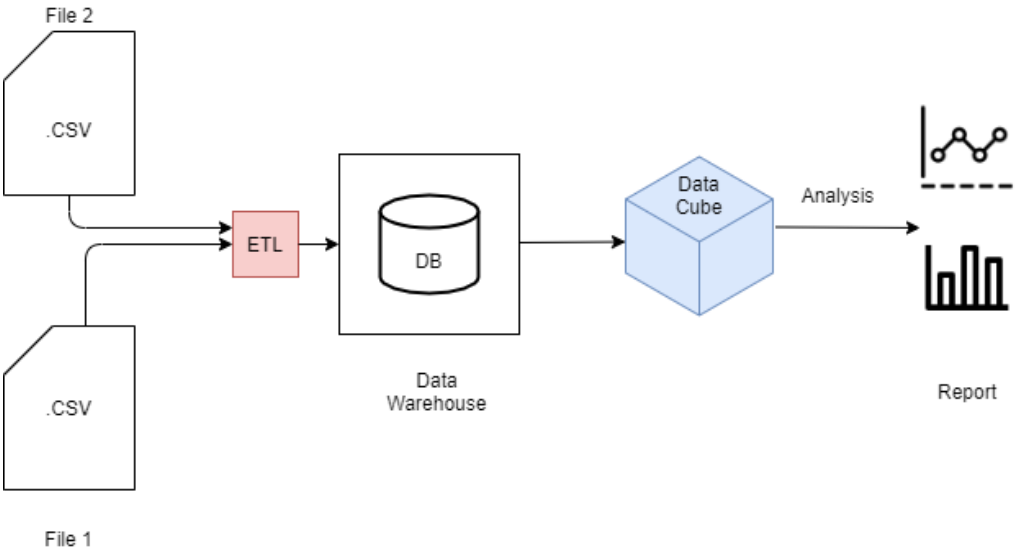
Fact tables	Dimension tables
User	Gender
	Location
	Education
	Occupation
	Relation
	Race
	Sector



8.ETL process

Extract Transform Load gets the data from different sources to a destination. Here, in this case our source is the .csv files. The .csv files will contain the User_info_fact.csv as fact table the rest of the file will be the dimension tables, dimensional tables are DimSector.csv, DimRelation.csv, DimRace.csv, DimSector.csv, DimEducation.csv, DimGender.csv and the DimLocation.csv.

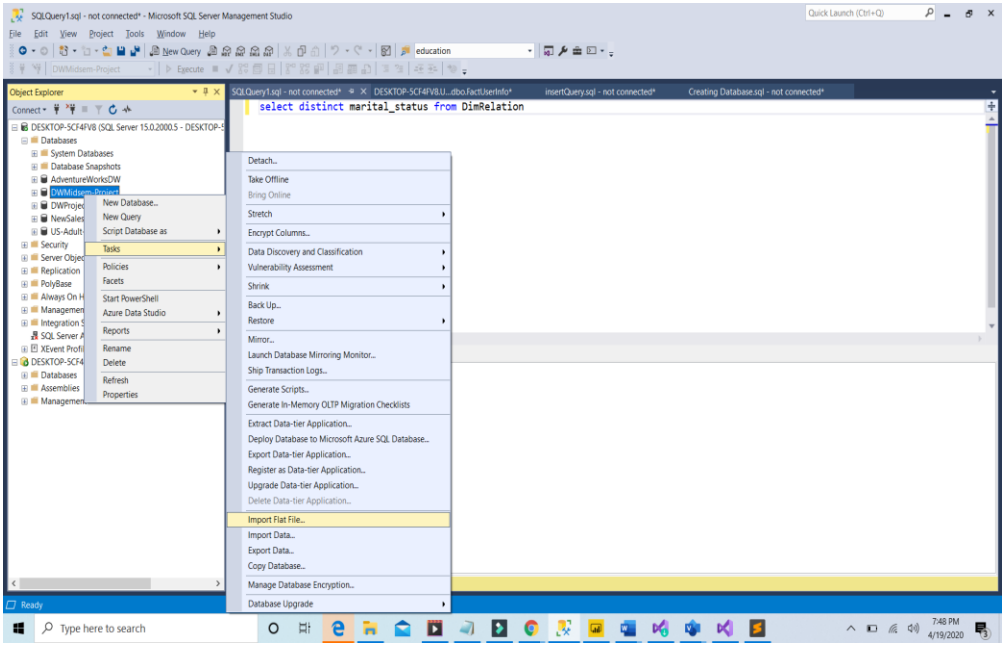
A diagram to show the ETL process for this project



How to import the .csv files and populate the database:

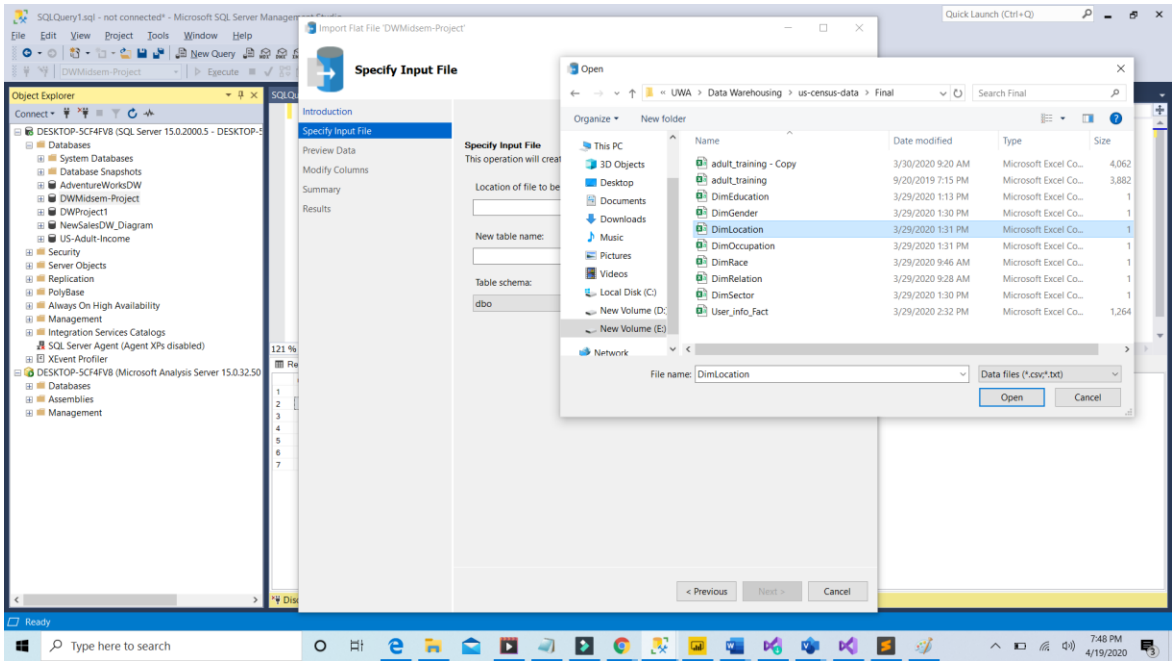
Step 1 :

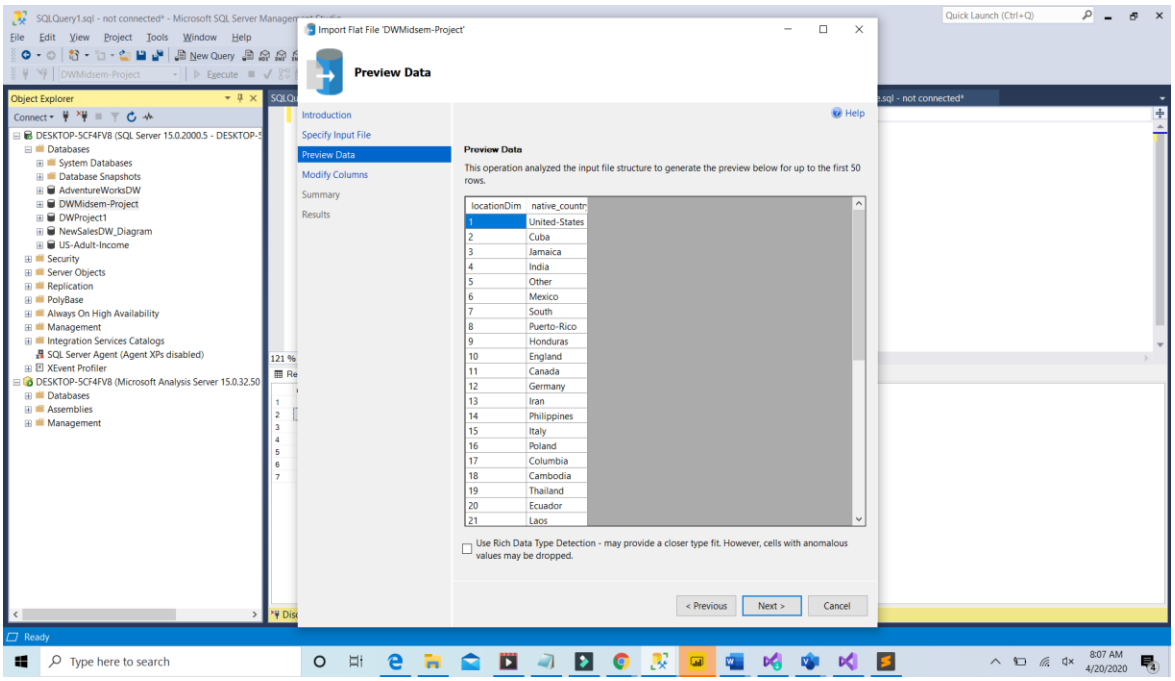
1. Go to object explorer select the name of the database where will populate the files.
2. Right click on to the Database go to Tasks.
3. Then go to import flat file at the bottom.



Step 2:

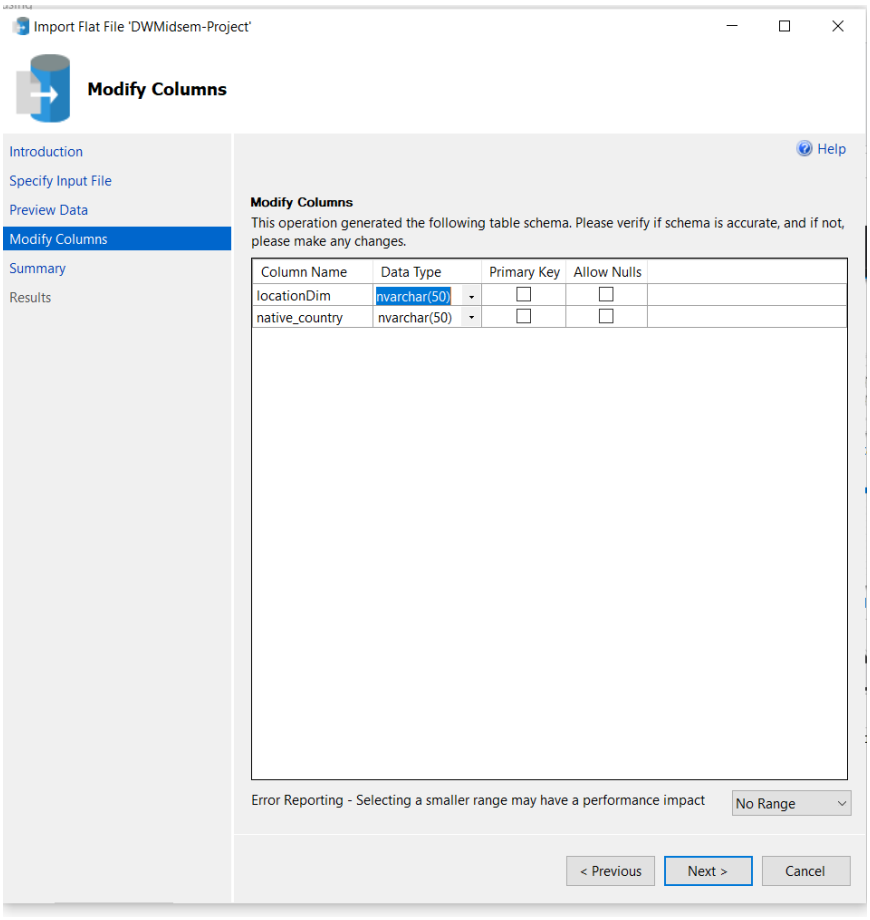
1. Then select the location where the csv file located. In my case the location is "E:\UWA\Data Warehousing\us-census-data\Final"
2. Load the .CSV

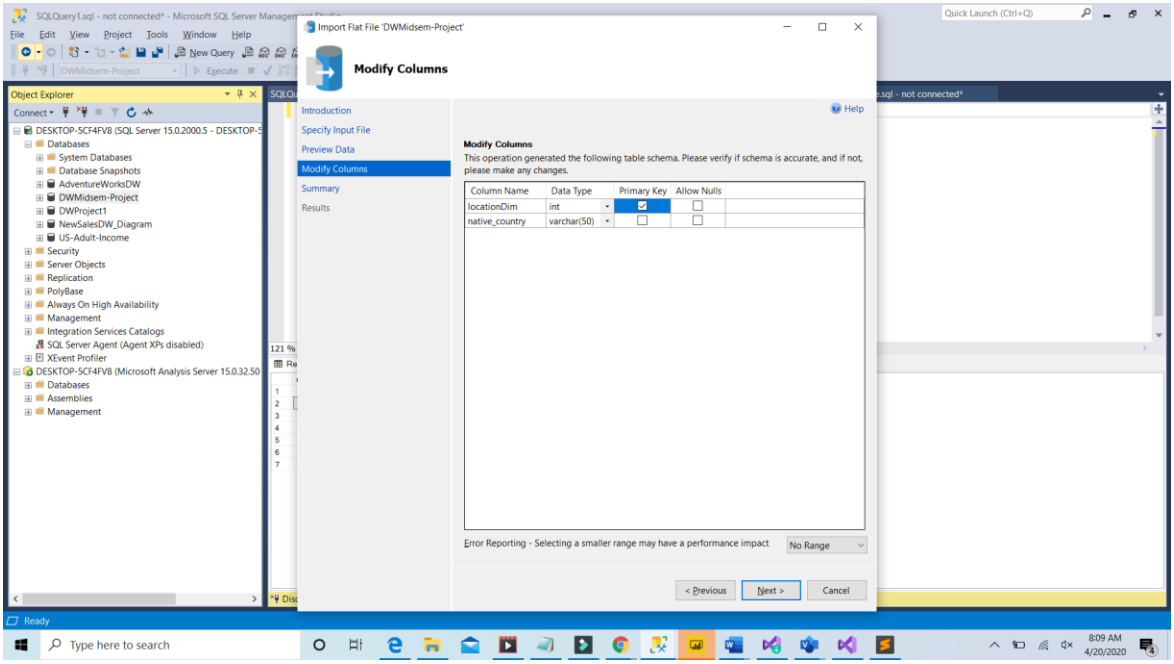




Step 3:

1. Change the datatypes. The IDs should be integer and the name column will be varchar because all the IDs of dimension table will work as a foreign key in our fact table and the data type of surrogate keys in the fact table is integer so both the datatype should match in order to create the relationship between two tables.


Import Flat File 'DWMidsem-Project'
Modify Columns
Introduction
Specify Input File
Preview Data
Modify Columns
Summary
Results
Help
Modify Columns
This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.
Column Name
Data Type
Primary Key
Allow Nulls
locationDim
nvarchar(50)
native_country
nvarchar(50)
Error Reporting - Selecting a smaller range may have a performance impact
No Range
< Previous
Next >
Cancel


Import Flat File 'DWMidsem-Project'
Modify Columns
Introduction
Specify Input File
Preview Data
Modify Columns
Summary
Results
Help
Modify Columns
This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.
Column Name
Data Type
Primary Key
Allow Nulls
locationDim
int
native_country
varchar(50)
Error Reporting - Selecting a smaller range may have a performance impact
No Range
< Previous
Next >
Cancel
Object Explorer
Ready
Type here to search

Step 4:

1. Now, click next and finished. The .csv fill will now populate the database.
2. Do the same process form all the csv file, including all the dimension and fact tables.

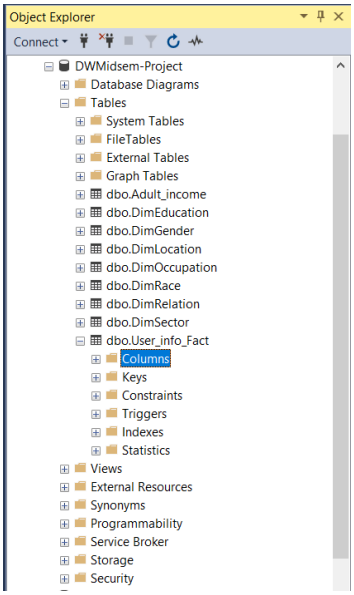
Step 5:

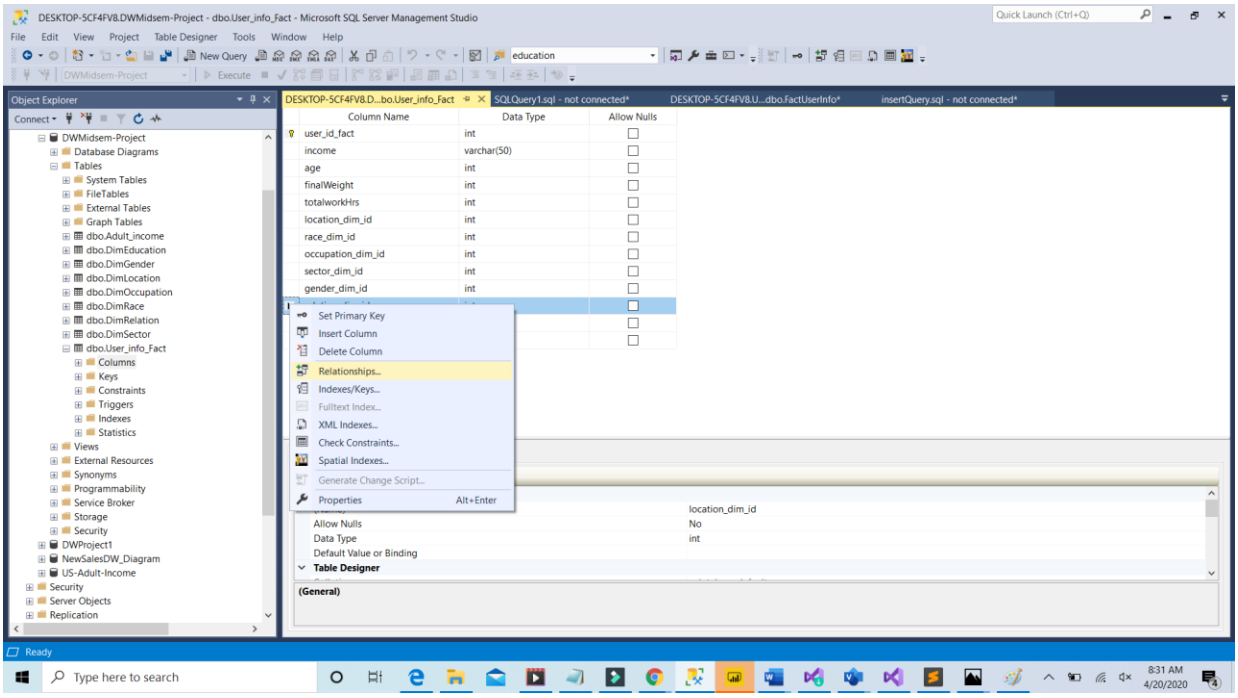
1. Now creating the relationship between the fact and the dimension tables.

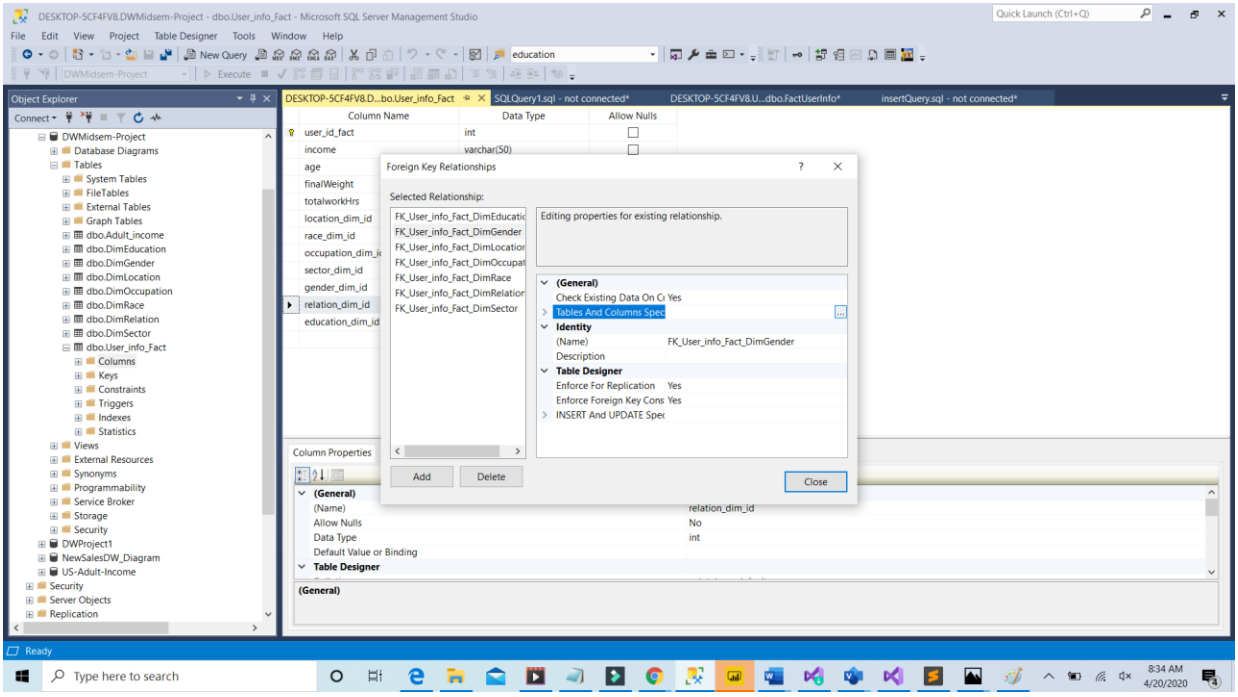
Step 6:

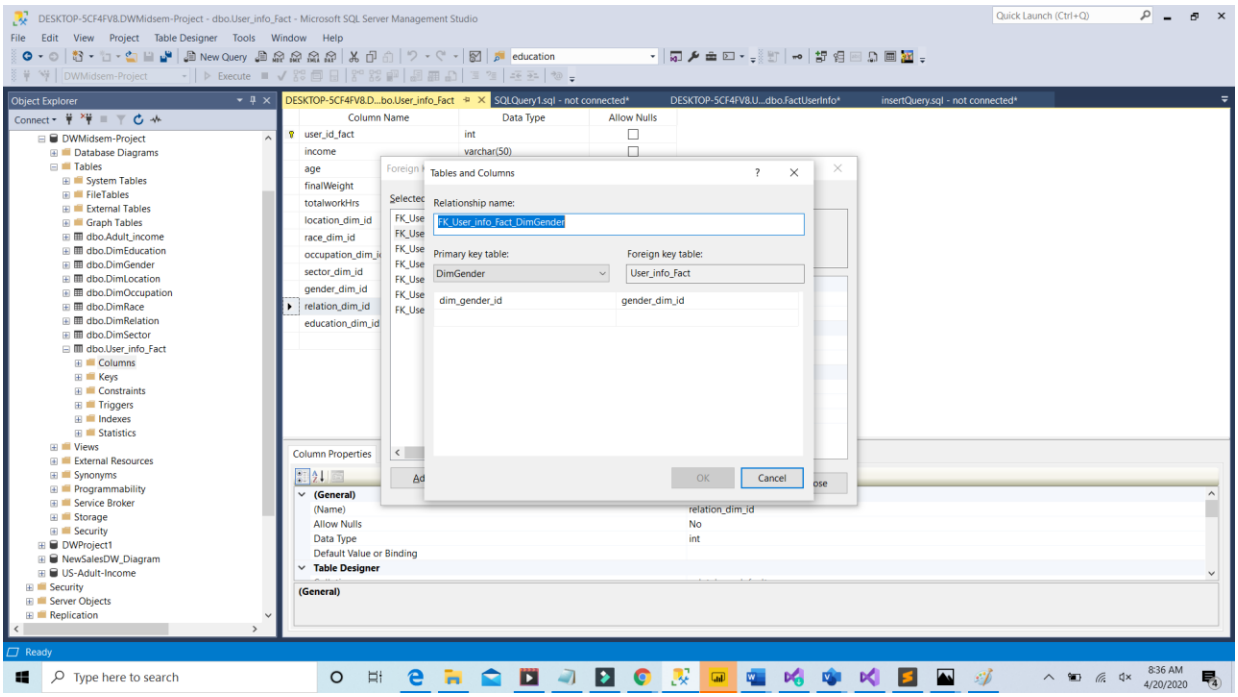
1. Select the fact table Go to columns, right click and select the new column.
2. Once the new window opens for the columns of the fact table which is User_fact_info in this case. Then select the columns of the dimension table Right Click and choose relationships. (Do the same process for all the dimension table one by one).
3. Now, the new window will open then click "Add". This will add a relation to the particular column.
4. Now, open Tables and column Space as shown in the screenshot.
5. Now, in the Primary key table section choose your dimension table in the dropdown list and choose the column which is the **primary key** of that particular table.
6. Go to the Foreign key table section and select the column which makes the foreign key in the fact table of that particular dimension table.

The whole process to generate a foreign key




DESKTOP-5CF4FV8.DWMidsem-Project - dbo.User_info_Fact - Microsoft SQL Server Management Studio
Object Explorer
DESKTOP-5CF4FV8.D...bo.User_info_Fact
SQLQuery1.sql - not connected*
DESKTOP-5CF4FV8.U...dbo.FactUserInfo*
insertQuery.sql - not connected*
Column Name
Data Type
Allow Nulls
user_id_fact
income
age
finalWeight
totalworkHrs
location_dim_id
race_dim_id
occupation_dim_id
sector_dim_id
gender_dim_id
Set Primary Key
Insert Column
Delete Column
Relationships...
Indexes/Keys...
Fulltext Index...
XML Indexes...
Check Constraints...
Spatial Indexes...
Generate Change Script...
Properties
Alt+Enter
Allow Nulls
No
Data Type
int
Default Value or Binding
Table Designer
(General)
Ready
8:31 AM
4/20/2020


DESKTOP-5CF4FV8.DWMidsem-Project - dbo.User_info_Fact - Microsoft SQL Server Management Studio
Foreign Key Relationships
Selected Relationship:
FK_User_info_Fact_DimEducatic
FK_User_info_Fact_DimGender
FK_User_info_Fact_DimLocatior
FK_User_info_Fact_DimOccupat
FK_User_info_Fact_DimRace
FK_User_info_Fact_DimRelatior
FK_User_info_Fact_DimSector
Editing properties for existing relationship.
(General)
Check Existing Data On Cr Yes
Tables And Columns Spec
Identity
(Name) FK_User_info_Fact_DimGender
Description
Table Designer
Enforce For Replication Yes
Enforce Foreign Key Cons Yes
INSERT And UPDATE Spec
Add
Delete
Close
Column Properties
relation_dim_id
education_dim_id
Ready
8:34 AM
4/20/2020



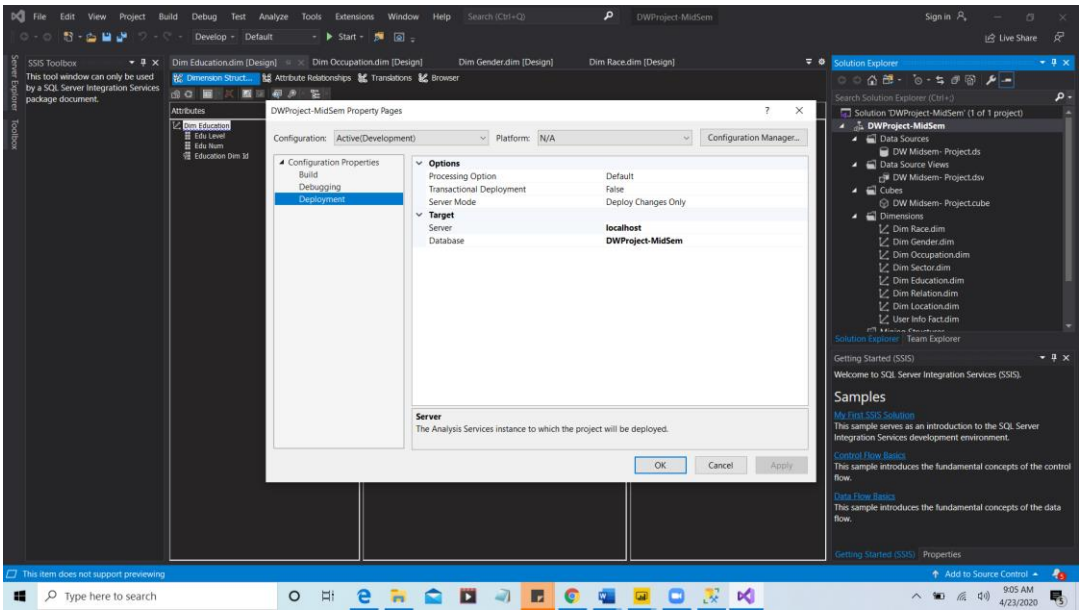
Do the same for all the dimension tables.

9. Process of creating Data Cube:

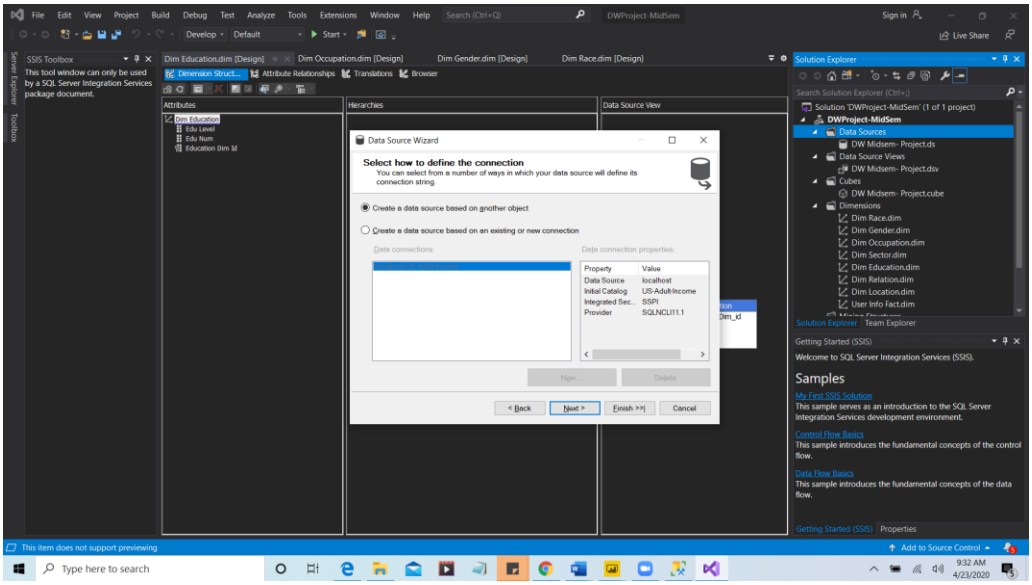
In order to design a data cube we use visual studio 2019 Enterprise. Visual should have the extension for

Analysis Service Multidimensional and Data Mining project (ASMD). Then we create a project in ASMD in visual studio.

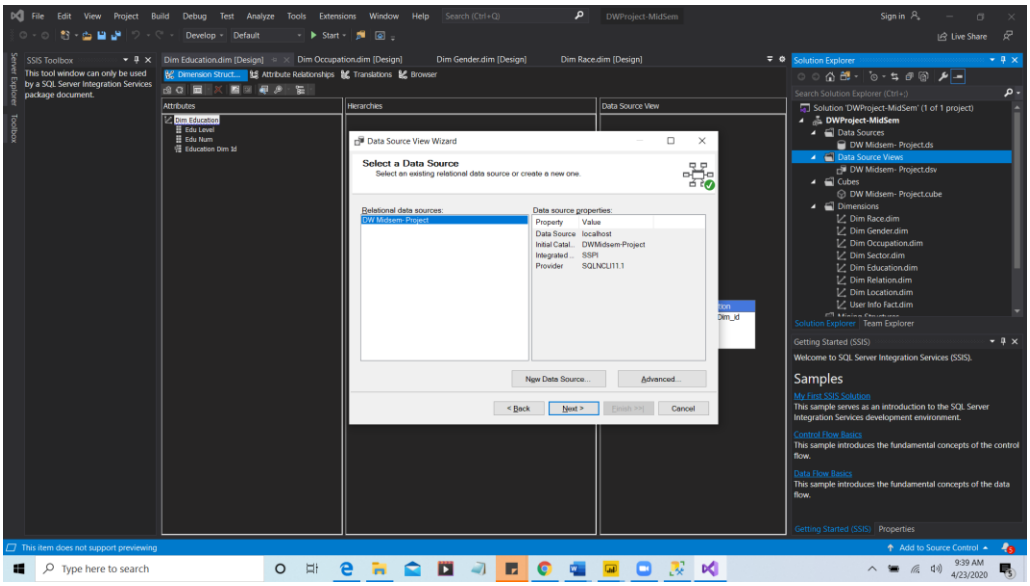
Step 1. Make sure in the properties of the project “DW Midsem-Project” the sever should be connected with the “localhost”.



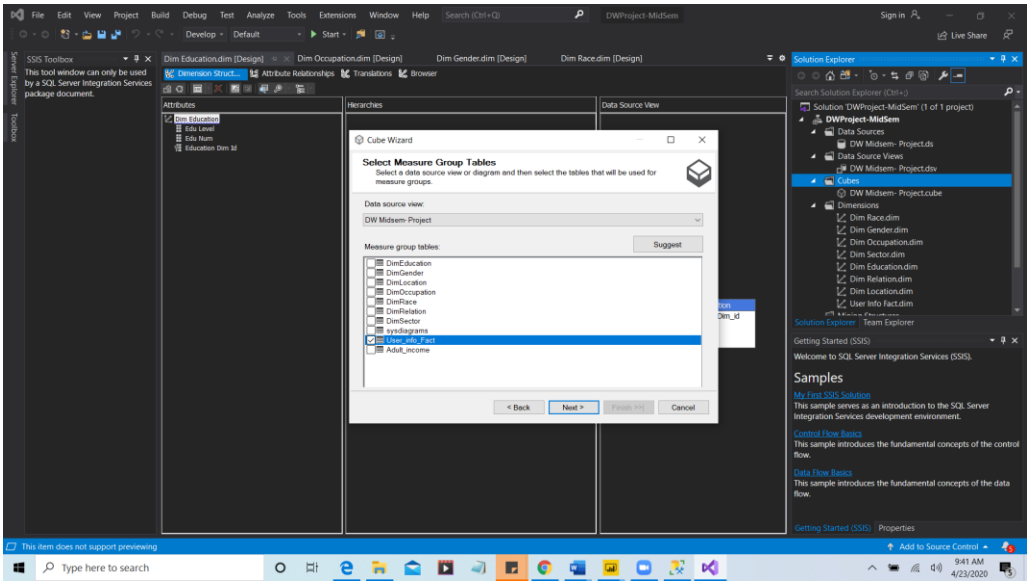
Step 2: Right click on the project solution (.ds file) and connect the Datasource.

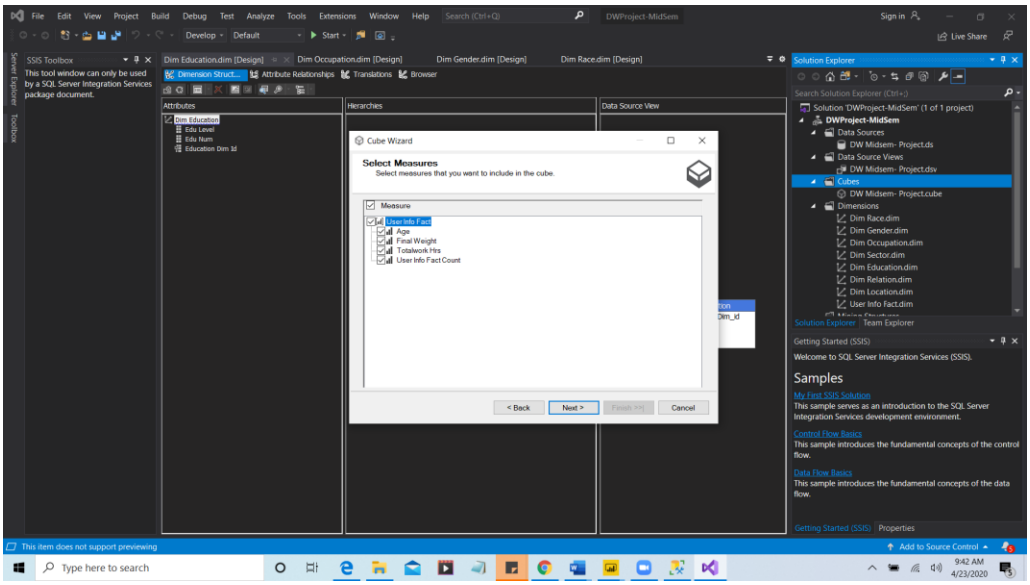


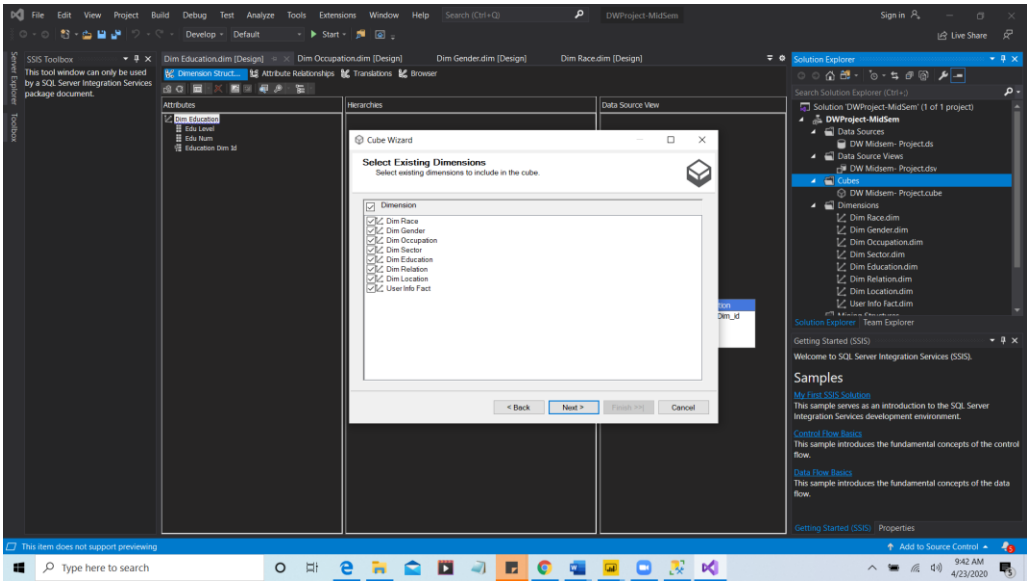
Step 3: Then go to the Data Source View folder and create a data source view.



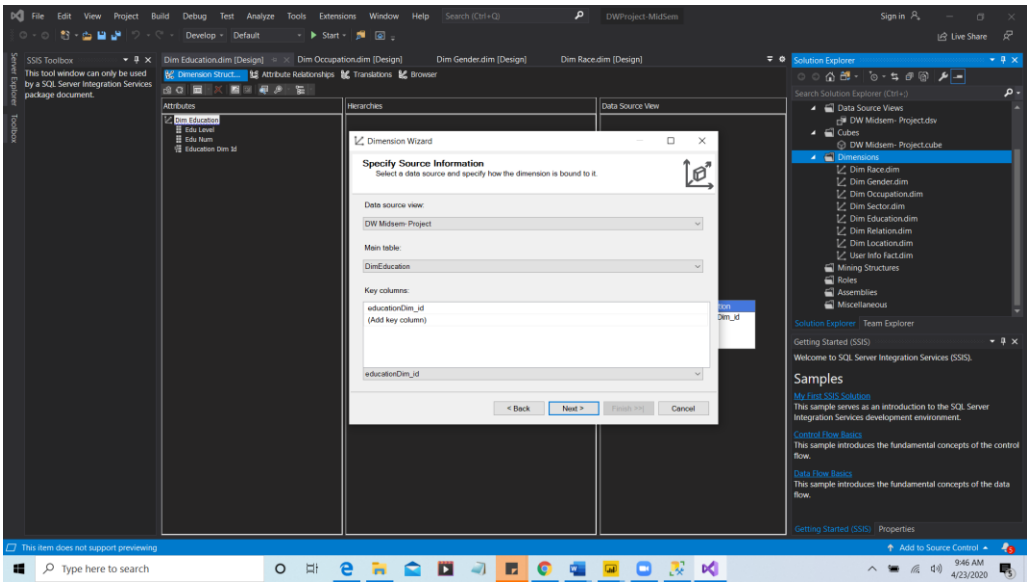
Step 4: Now, create the cube by right clicking on “Cubes”. Select only the fact table which is User_fact_info, once the fact table is selected the measures and the dimension tables linked with the fact table will automatically be selected because creates a relationship using surrogate keys. Now, load the cube.

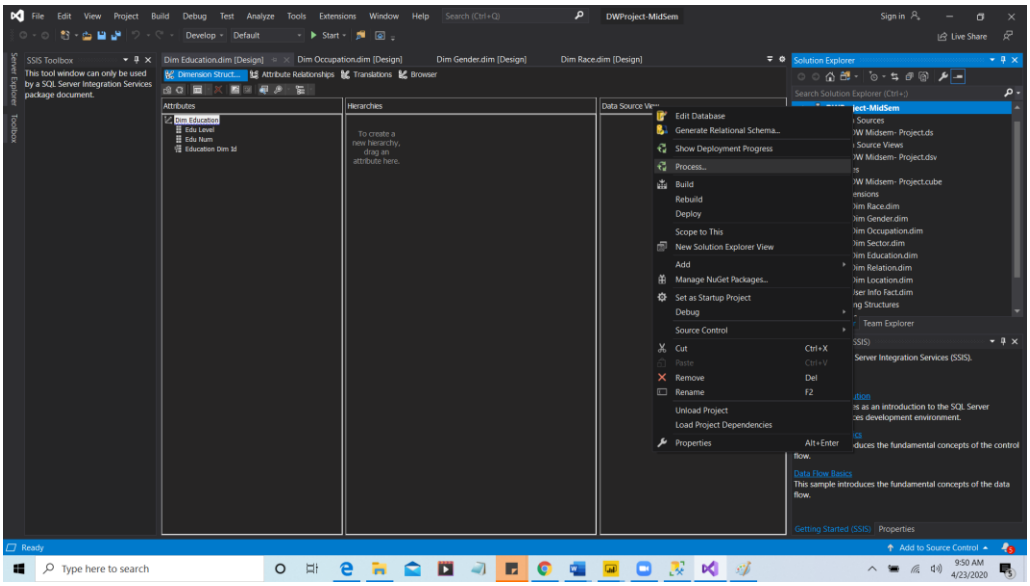



Cube Wizard
Select Measures
Select measures that you want to include in the cube.
Measure
User Info Fact
Age
Final Weight
Totalwork Hrs
User Info Fact Count
< Back
Next >
Finish >>|
Cancel
Samples
My First SSIS Solution
This sample serves as an introduction to the SQL Server Integration Services development environment.
Control Flow Basics
This sample introduces the fundamental concepts of the control flow.
Data Flow Basics
This sample introduces the fundamental concepts of the data flow.


Cube Wizard
Select Existing Dimensions
Select existing dimensions to include in the cube.
Dimension
Dim Race
Dim Gender
Dim Occupation
Dim Sector
Dim Education
Dim Relation
Dim Location
User Info Fact
< Back
Next >
Finish >>|
Cancel
Samples
My First SSIS Solution
This sample serves as an introduction to the SQL Server Integration Services development environment.
Control Flow Basics
This sample introduces the fundamental concepts of the control flow.
Data Flow Basics
This sample introduces the fundamental concepts of the data flow.

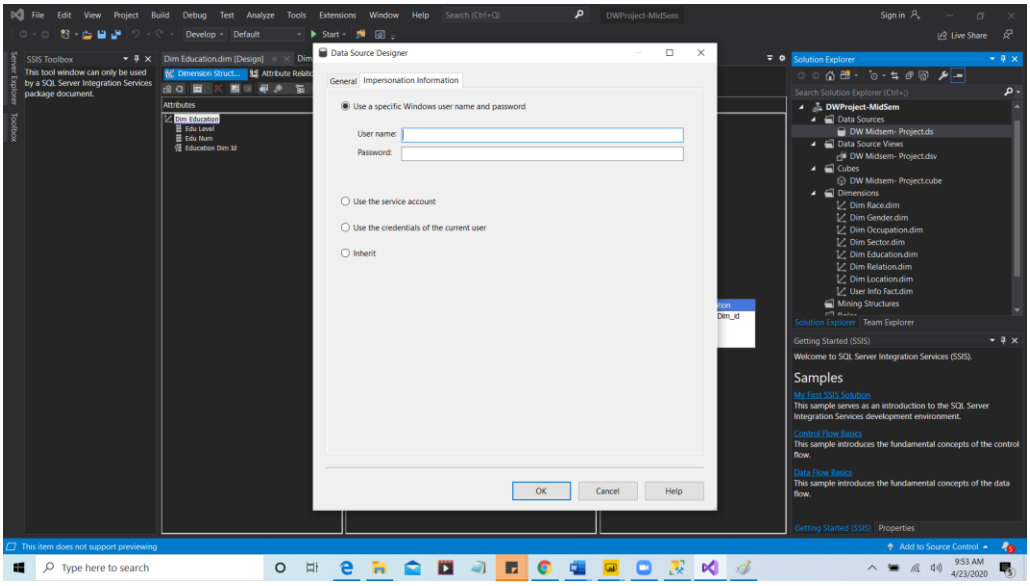
Step 5: Now, create the dimension by right clicking on “Dimensions” and then process it.



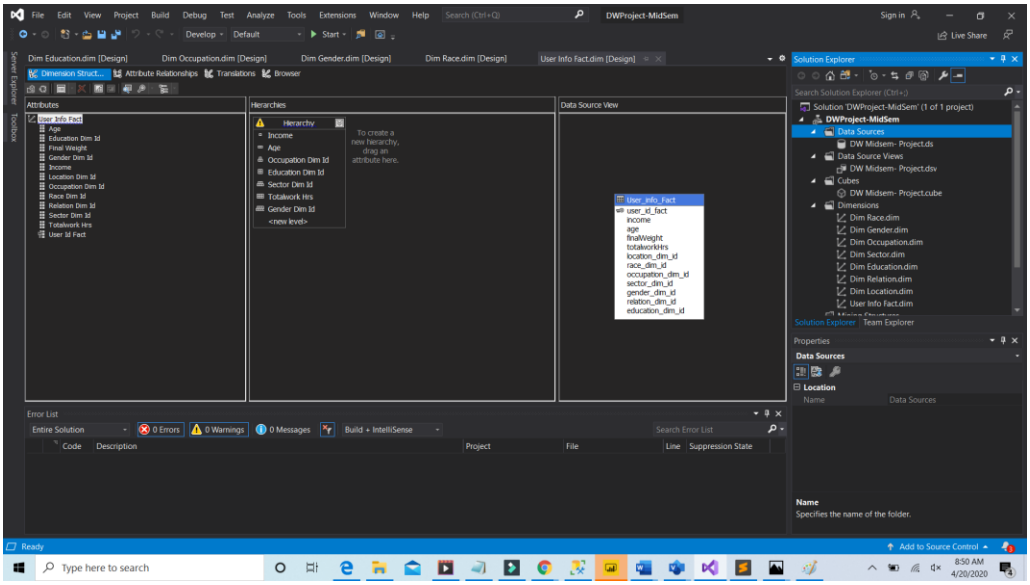


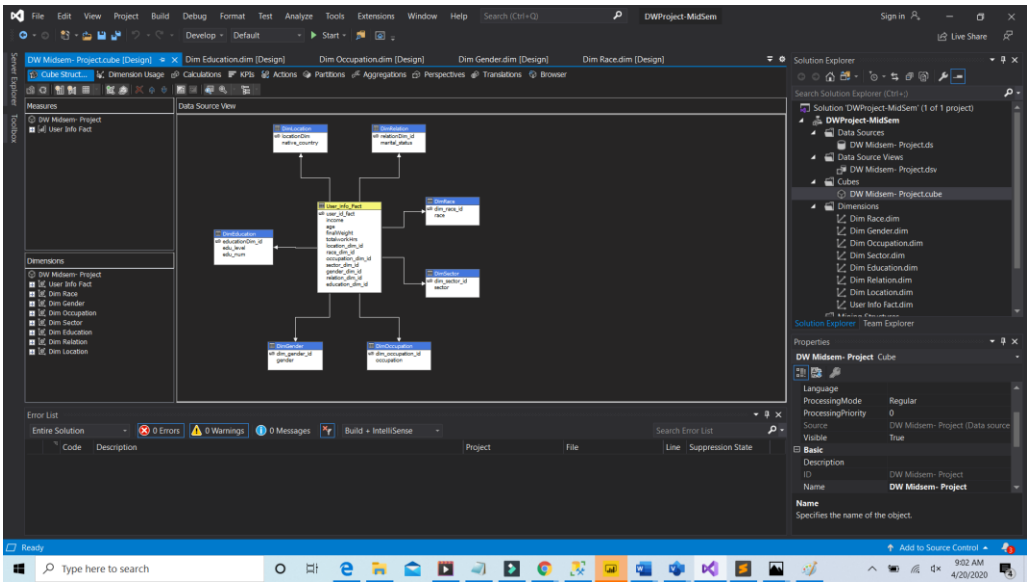
If any error is occurred while processing the data then we have to follow the 6th step.

Step 6: Make sure you have authentication of your system that means before processing the data double click on the .ds file in Data Sources then go to **Impersonation information** enter the username of the system and the password then click OK and then process again.



10. Data cube hierarchy in Visual Studio

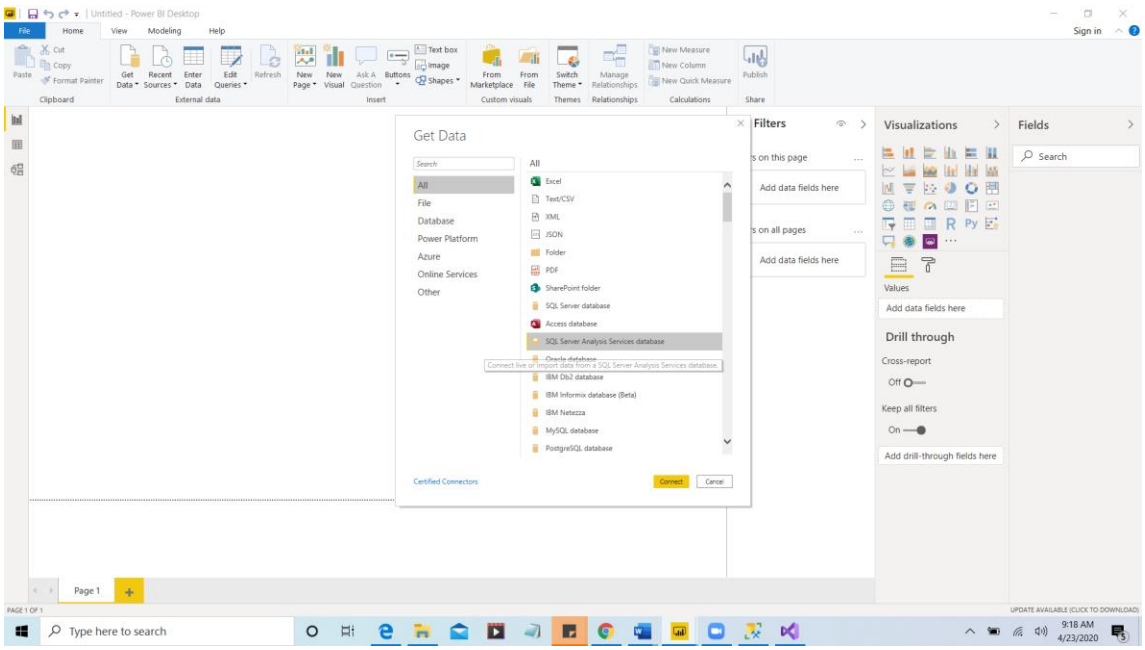




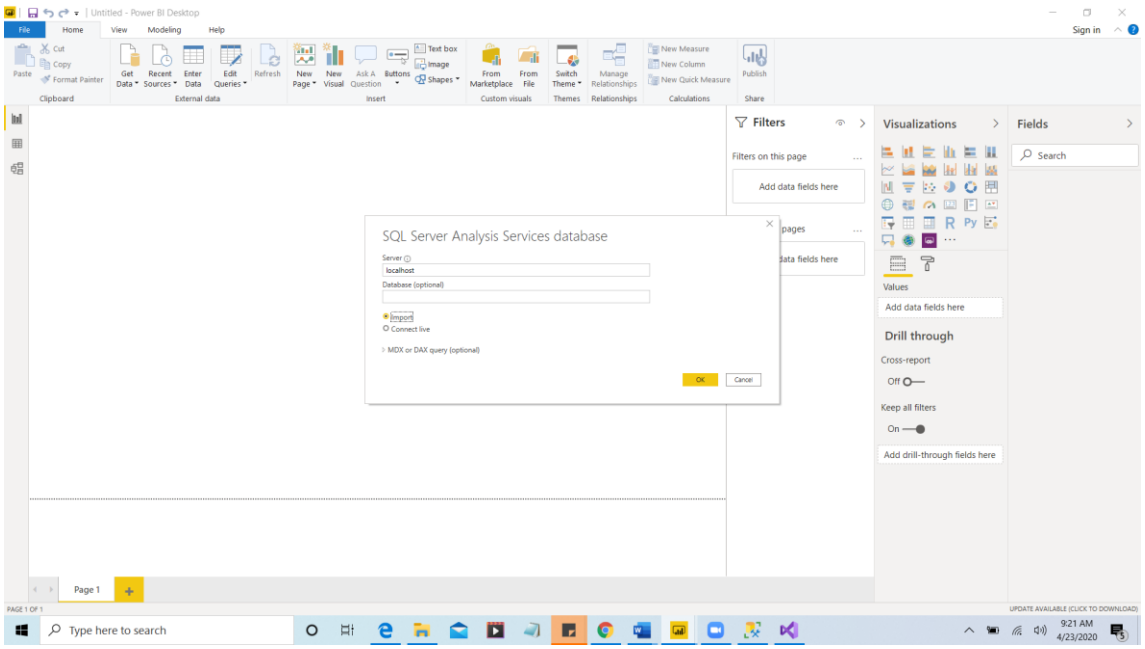
11. SQL Analysis service in Power BI

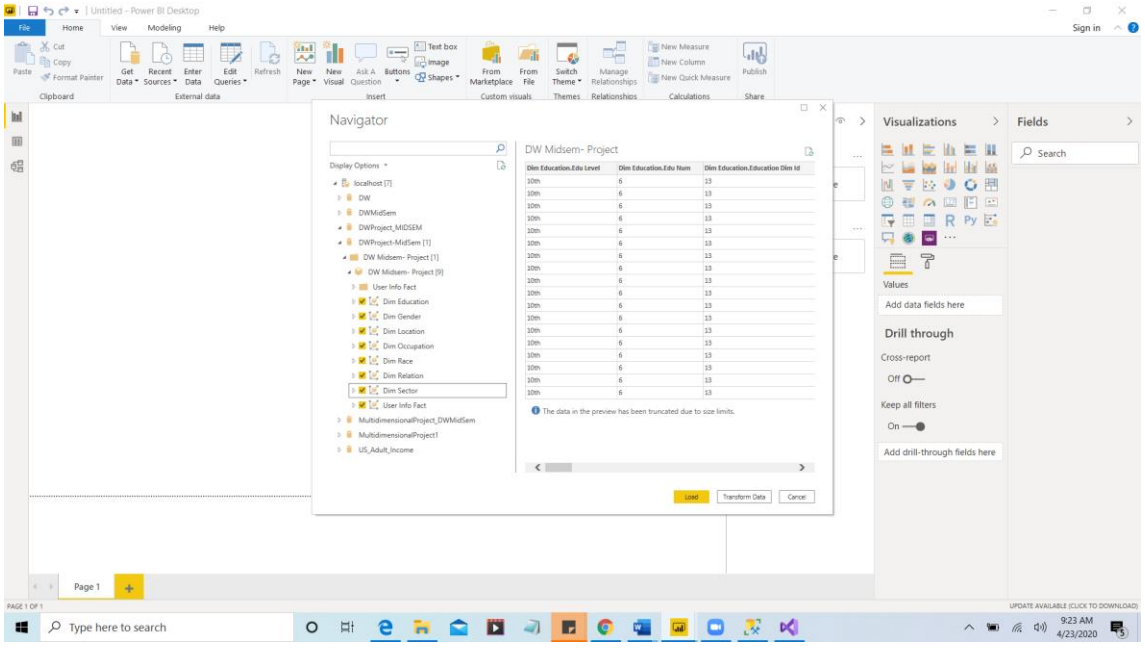
Process of importing the data cube in the power BI is shown below.

1. Open Microsoft PowerBI and click on **Get Data** select **Server Analysis Services database** and then **connect.**



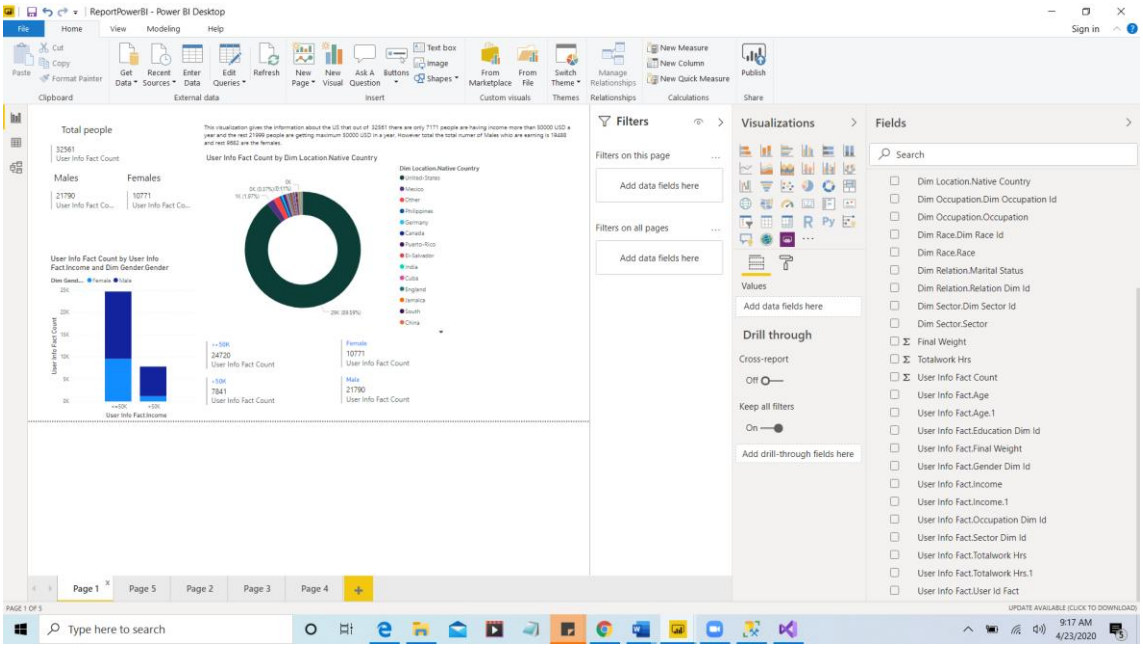
2. Enter your server name “localhost“ and choose import then OK.
3. Select the relevant data cube which is created by the visual studio from the server and select the tables (Fact and the dimension) and then load.


Untitled - Power BI Desktop
File
Home
View
Modeling
Help
Sign in
Paste
Cut
Copy
Format Painter
Clipboard
Get Data
Recent Sources
Enter Data
Edit Queries
Refresh
External data
New Page
New Visual
Ask A Question
Buttons
Text box
Image
Shapes
Insert
From Marketplace
From File
Custom visuals
Switch Theme
Themes
Manage Relationships
Relationships
New Measure
New Column
New Quick Measure
Calculations
Publish
Share
SQL Server Analysis Services database
Server
localhost
Database (optional)
Import
Connect live
MDX or DAX query (optional)
OK
Cancel
Filters
Filters on this page
Add data fields here
Visualizations
Values
Add data fields here
Drill through
Cross-report
Off
Keep all filters
On
Add drill-through fields here
Fields
Search
Page 1
PAGE 1 OF 1
UPDATE AVAILABLE (CLICK TO DOWNLOAD)
Type here to search
9:21 AM
4/23/2020


Untitled - Power BI Desktop
File
Home
View
Modeling
Help
Sign in
Navigator
Display Options
localhost [7]
DW
DWMidSem
DWProject_MIDSEM
DWProject-MidSem [1]
DW Midsem- Project [1]
DW Midsem- Project [9]
User Info Fact
Dim Education
Dim Gender
Dim Location
Dim Occupation
Dim Race
Dim Relation
Dim Sector
User Info Fact
MultidimensionalProject_DWMidSem
MultidimensionalProject1
US_Adult_Income
DW Midsem- Project
Dim Education.Edu Level
Dim Education.Edu Num
Dim Education.Education Dim Id
10th
6
13
The data in the preview has been truncated due to size limits.
Load
Transform Data
Cancel
Visualizations
Values
Add data fields here
Drill through
Cross-report
Off
Keep all filters
On
Add drill-through fields here
Fields
Search
Page 1
PAGE 1 OF 1
UPDATE AVAILABLE (CLICK TO DOWNLOAD)
Type here to search
9:23 AM
4/23/2020

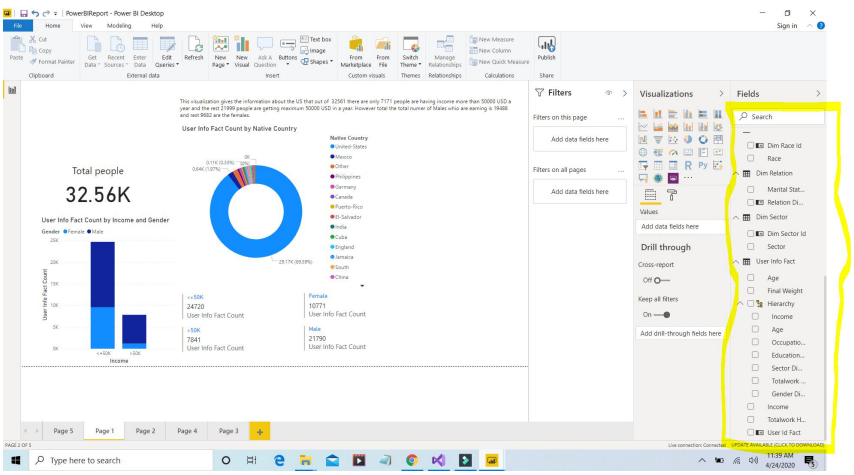
4. After the loading the data cube in Microsoft PowerBI it looks something like this. All the tables (Fact and dimension tables) and measures are situated at the top right side and we can do the analysis and visualization using the functions dragging from the visualization section.

Using Import Connection

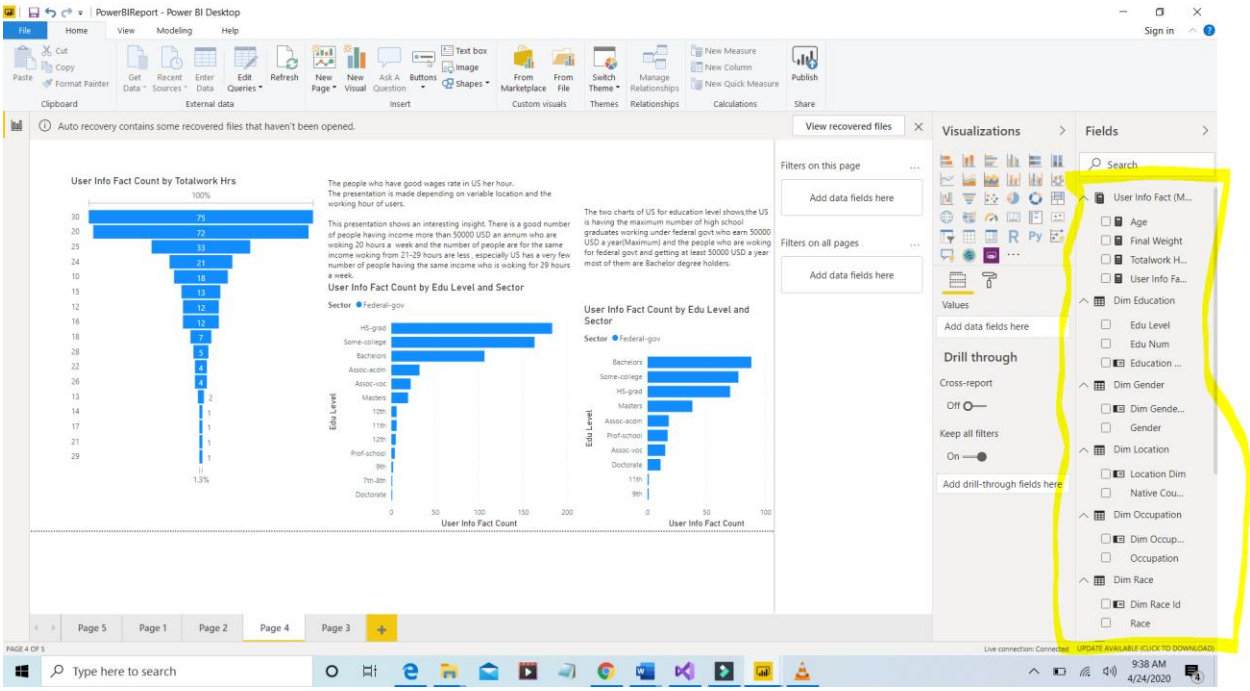


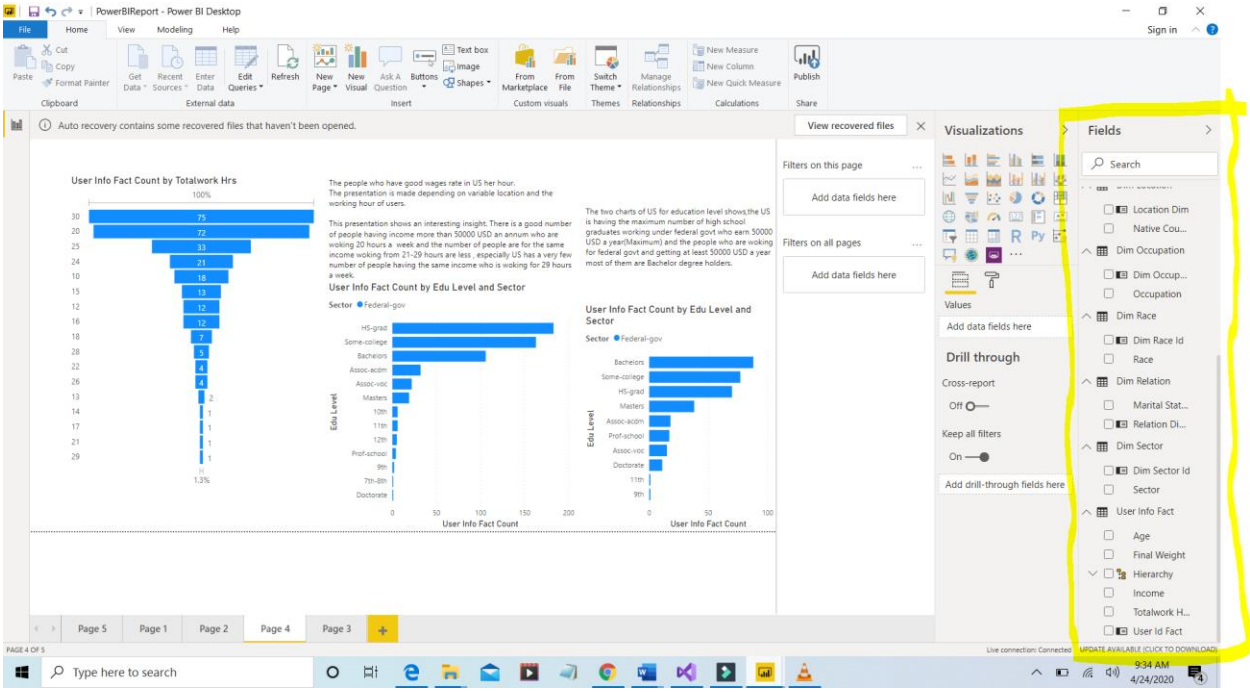
12.Analysis Report screenshot using live connection :

Here, the screenshots are added to show the hierarchy of the cube using live connection with the database. But the final report has been made using import method in PowerBI. The purpose this is to show the hierarchy of the data cube as is not shown in PowerBI using the live connection. The proper analysis report has been added below using import method.



The hierarchy using live connection





13. Final analysis report :

US-Adult-Income Dataset
Over all description

Total people

32561
User Info Fact Count

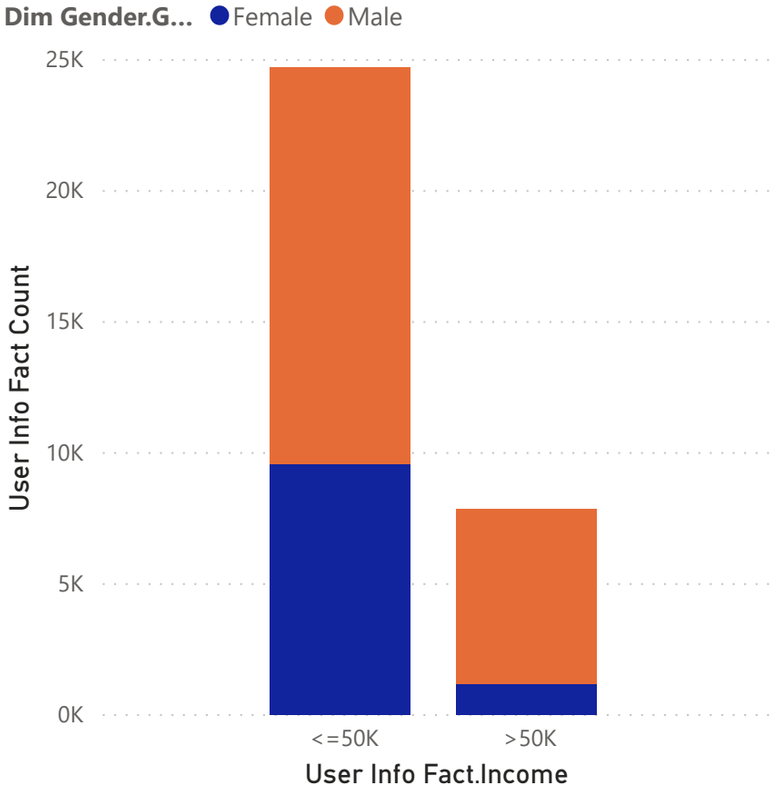
Males

21790
User Info Fact Co...

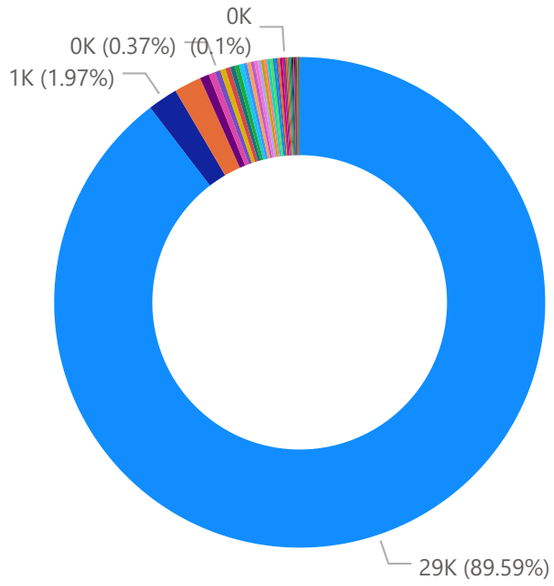
Females

10771
User Info Fact Co...

User Info Fact Count by User Info Fact.Income and Dim Gender.Gender



User Info Fact Count by Dim Location.Native Country



<=50K
24720
User Info Fact Count

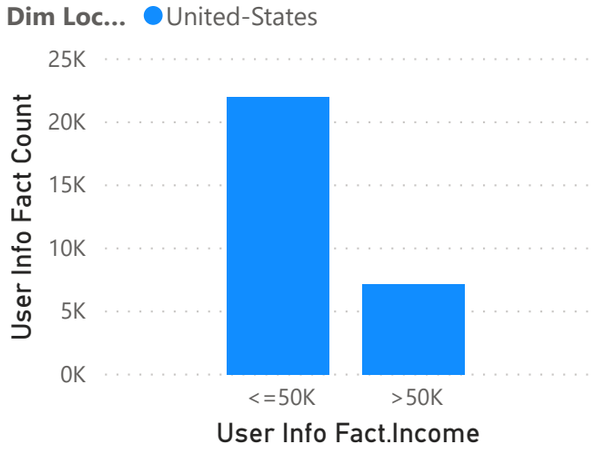
>50K
7841
User Info Fact Count

Female
10771
User Info Fact Count

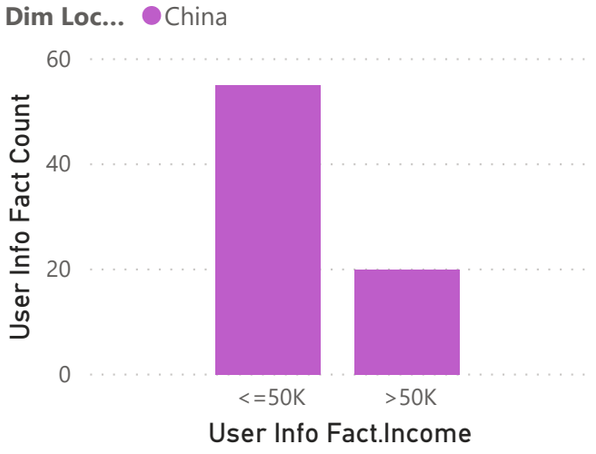
Male
21790
User Info Fact Count

Income of pople who earns more than USD 50000 also maximum USD 50000 a year from US, China, Germany, India and Mexico. (Solution to business Query 1)

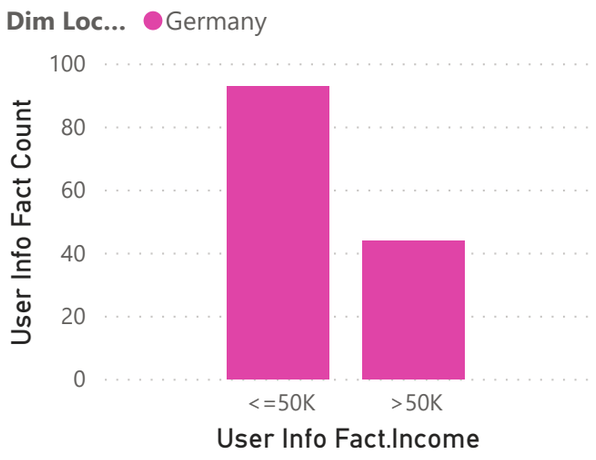
User Info Fact Count by User Info Fact.Income and Dim Location.Native Country



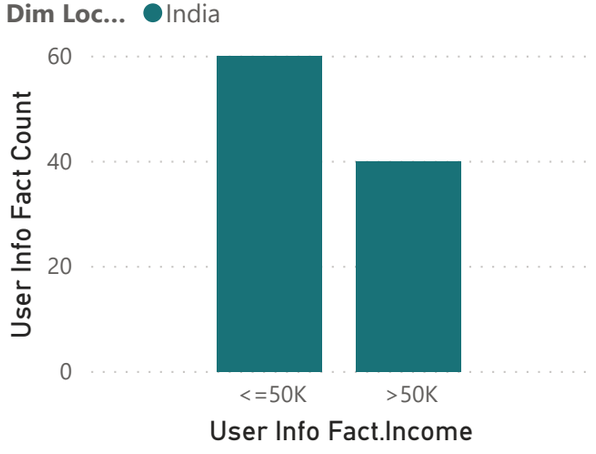
User Info Fact Count by User Info Fact.Income and Dim Location.Native Country



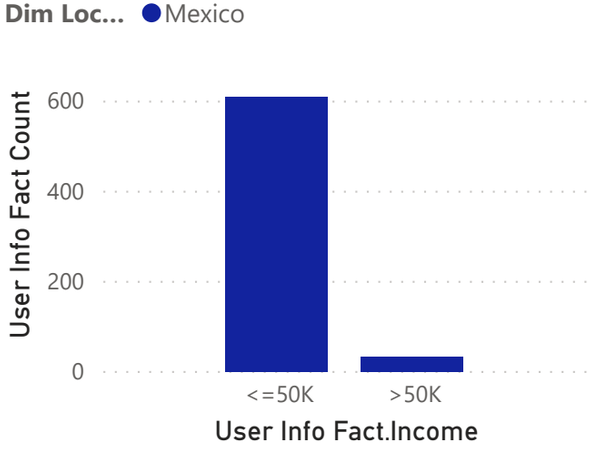
User Info Fact Count by User Info Fact.Income and Dim Location.Native Country



User Info Fact Count by User Info Fact.Income and Dim Location.Native Country

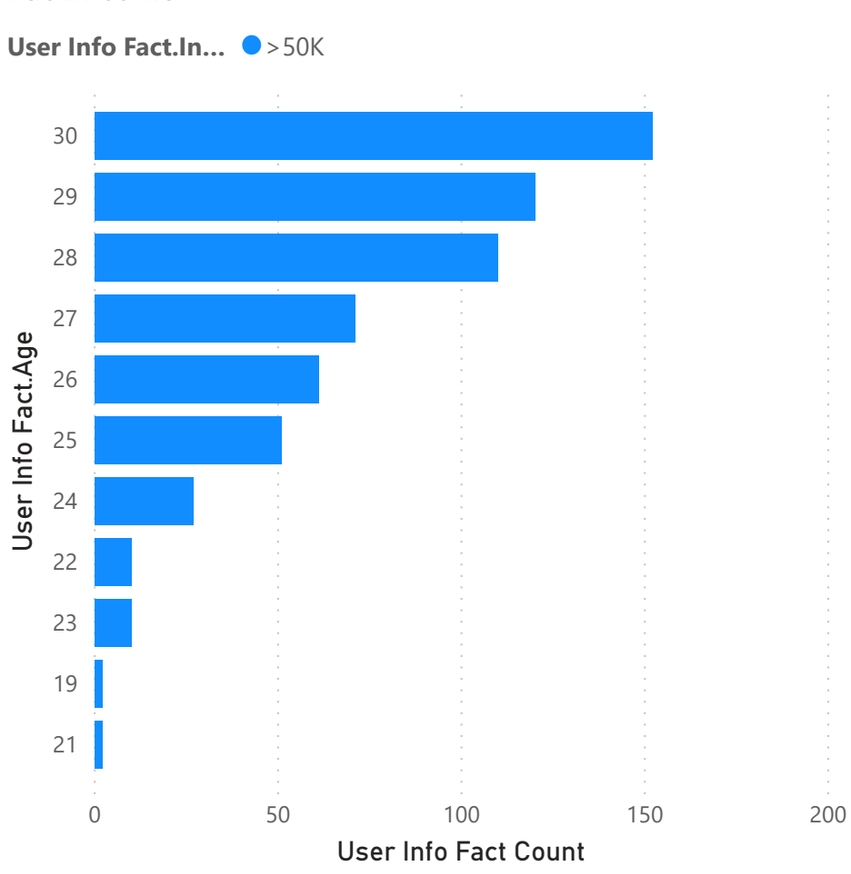


User Info Fact Count by User Info Fact.Income and Dim Location.Native Country

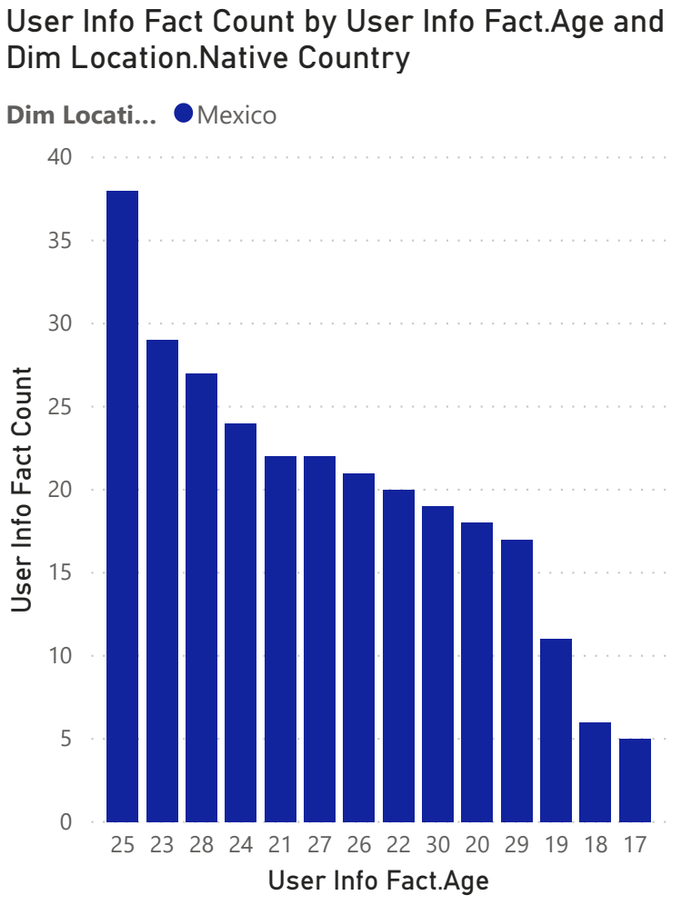


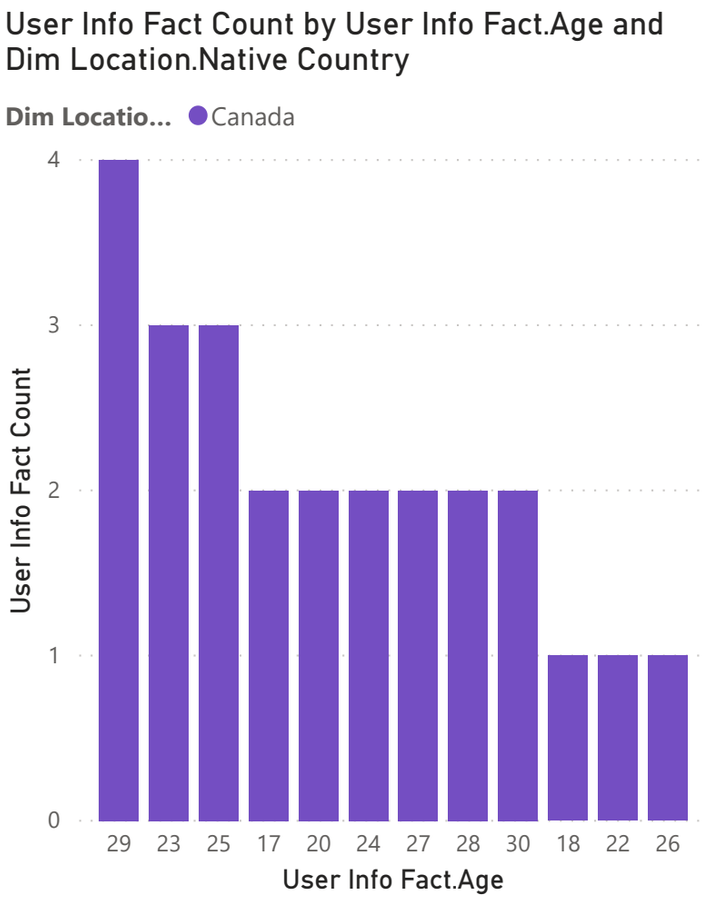
A very few people have got income more than USD 50000 a year in Mexico

The number of young people of age group 18-30 having income
at least 50000 USD a year in the US (Solution to business Query 2)

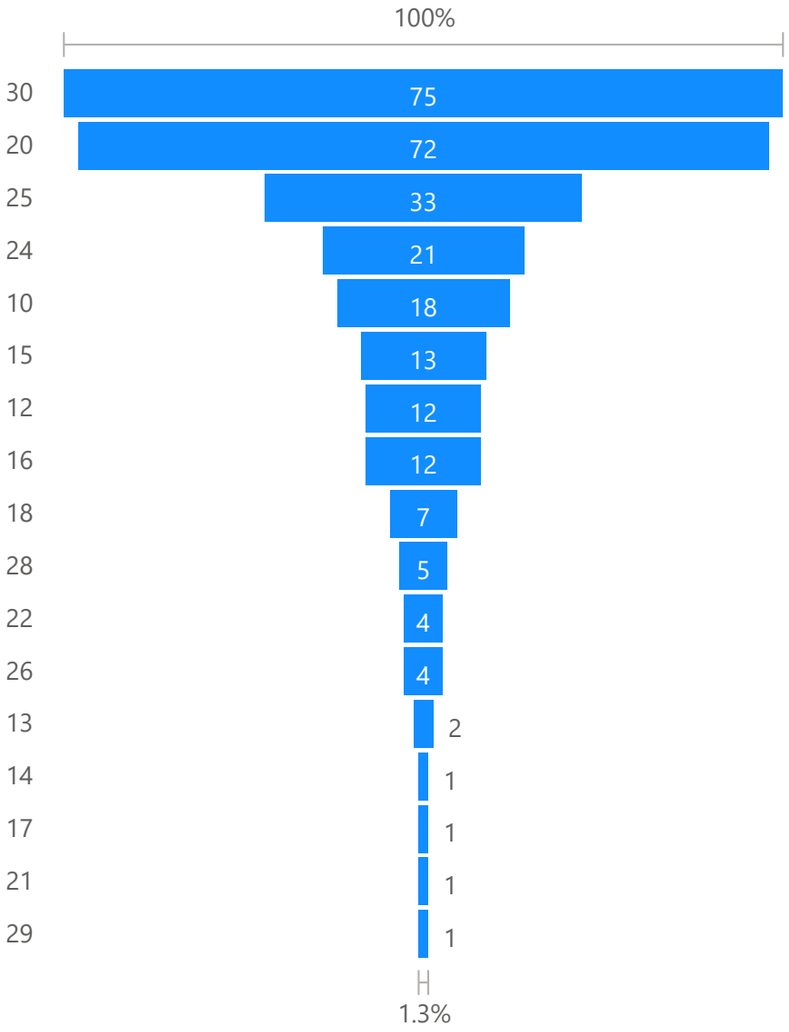


Young people working for private companies in Mexico and Canada having income more than USD 50000 per annum. (Solution to business Query 3)



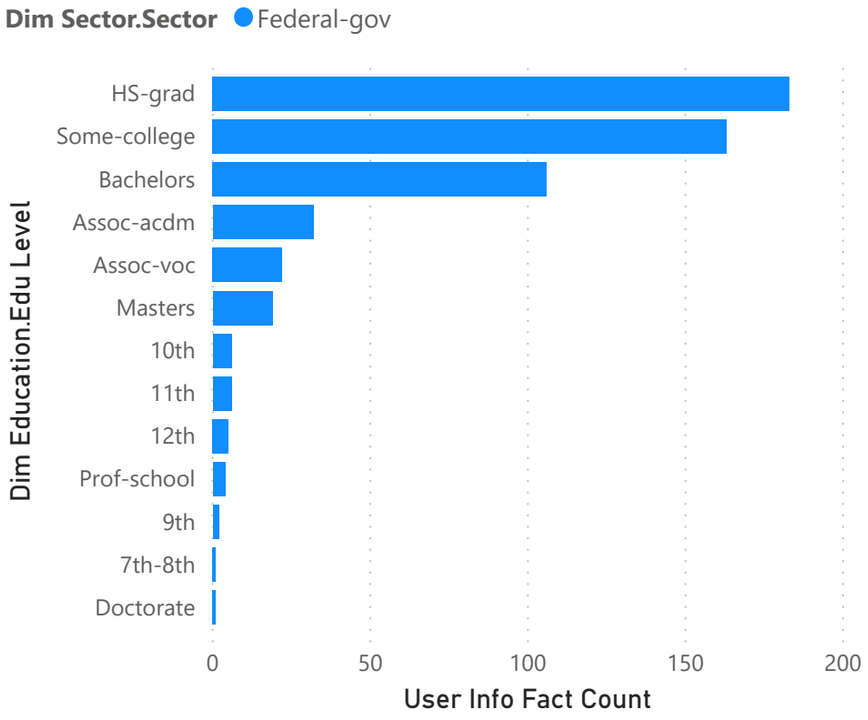


User Info Fact Count, First Dim Location.Native Country and First User Info Fact.Income by User Info Fact.Totalwork Hrs



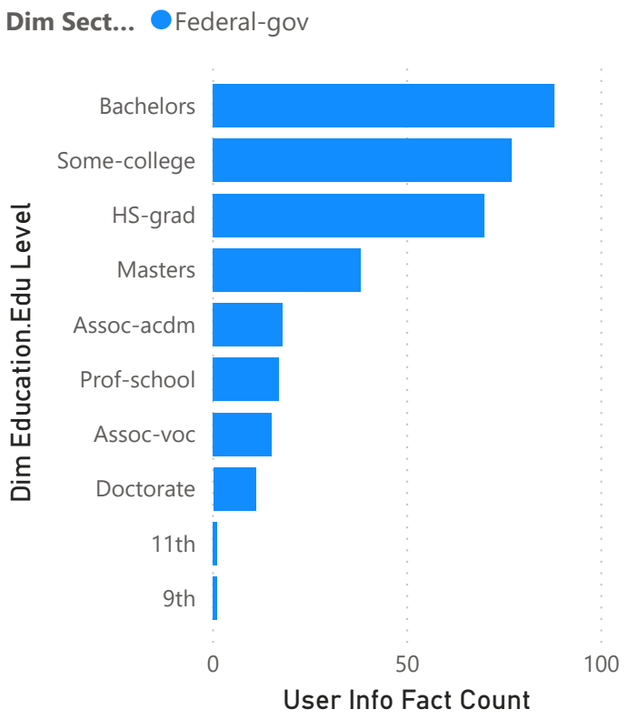
The people who have good wages rate in US her hour.
The presentation is made depending on variables such as location and the working hour of users.
This presentation shows an interesting insight. There is a good number of people having income more than 50000 USD an annum who are woking 20 hours a week and the number of people are for the same income woking from 21-29 hours are less , especially US has a very few number of people having the same income who is woking for 29 hours a week. (Solution to business Query 4)

User Info Fact Count by Dim Education.Edu Level and Dim Sector.Sector



The two charts of US for education level shows,the US is having the maximum number of high school graduates working under federal govt who earn 50000 USD a year(Maximum) and the people who are woking for federal govt and getting at least 50000 USD a year most of them are Bachelor degree holders. (Solution to business query 5)

User Info Fact Count by Dim Education.Edu Level and Dim Sector.Sector



(Solution to business Query 6)