Brief Description

The following documents will reflect our data comparing the impacts of education to see if there is a drastic difference between the different levels of education earned and specifically the impacts it could have on poverty. We will look at poverty levels within the levels of education, those with less than a high school diploma, a high school diploma only, 1-3 years of after high school education which may include an Associate’s Degree, and those with an education level of 4 years or higher, including at least a Bachelor's Degree.

The Impacts of Education

Business Intelligence

Author:

Shane Varnum

# Data Mart Design Description:

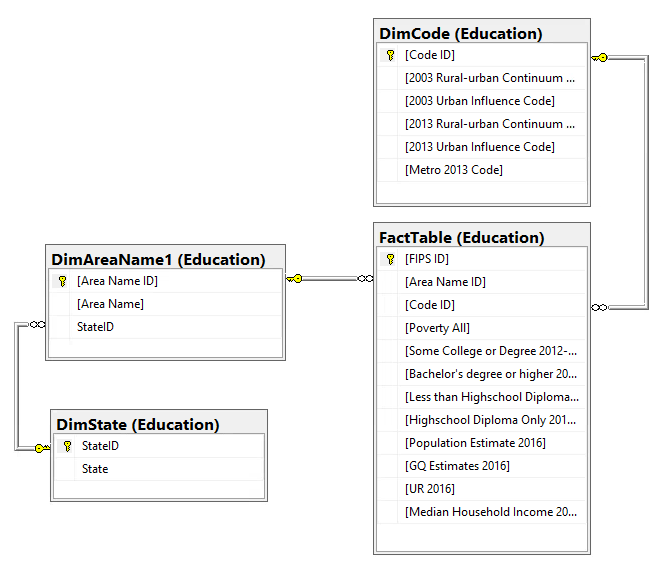
This document attempts to describe a proposed data mart design that can be used to analyze the Socioeconomic indicators: *Education, Population, Unemployment, and Poverty Estimates* data set provided by ers.usda.gov. The design will be implemented and replicated in Microsoft SQL Server Management Studio. SQL Server MS will be used to implement the analytical design in order to accommodate a variety of analytical technologies which could be used to perform the analysis. The analytical design is based off of the flat data structure included in the CSV and Excel files provided by Ers.Usda.gov and include county-level rural/urban codes and urban influence codes. They examine the Poverty, Population, Unemployment and Median Household Income, and Education estimates recorded by the United States Department of Agriculture, Economic Research Service and we will be using the data information dated between 2000 and 2017, specifically focusing on the years between 2012 and 2016, and compared by area/state.

This document will examine the proposed analytical tables and relationships based on structure seen in this partial sample. **NOTE: not all columns and rows are included in this sample**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIPS ID | Area Name ID | Code ID | Poverty All | Some College or degree 2012-2016 | Bachelor's degree or higher, 2012-2016 | Less than a high school diploma, 2012-2016 |
| 1001 | 100 | 100 | 7444 | 10451 | 8960 | 9811 |
| 1003 | 101 | 101 | 24005 | 44486 | 25876 | 11718 |
| 1005 | 102 | 102 | 6787 | 4475 | 3240 | 2078 |
| 1007 | 103 | 103 | 4099 | 4234 | 3475 | 2513 |
| 1009 | 104 | 104 | 8033 | 13436 | 10735 | 8777 |

In the above data sample taken from one of the source CSV files, the proposed analytical design will consider the measurable columns such as “Poverty All”, “Bachelor’s degree or higher, 2012-2016”, “Less than a high school diploma, 2012-2016”, as “Facts”. The “Area Name ID” and “Code ID”, columns will be used to derive “Dimensions”. These tables will be used together to form an analytical schema design. This document will illustrate the schema and describe its proposed design. The *Education, Population, Unemployment, and Poverty Estimates* data set will be extracted, transformed, and loaded (ETL) into the analytical schema design using a custom-built SQL Server Integration Services (SSIS) package developed with SQL Server Data Tools (SSDT) in Visual Studio. Additionally, analysis will be performed on the Education, Population, Unemployment, and Poverty Estimates data using the database labeled Socioeconomic Project.

# Data Mart Schema Diagram:



# Data Mart Meta Data:

|  |  |  |  |
| --- | --- | --- | --- |
| **FACT TABLE - Education.FactTable** | | | |
| Column Name | Data Type | Description | Key |
| FIPS ID | int | Primary Key | Primary |
| Area Name ID | int | Foreign Key | Foreign |
| Code ID | int | Foreign Key | Foreign |
| Poverty All | int | Estimate of people of all ages in poverty 2016 | None |
| Some College or Degree 2012-16 | int | People with Some College/Degree in the years 2012-2016 | None |
| Bachelor's Degree or < 2012-16 | int | People with Bachelor’s Degree or higher in the years 2012-2016 | None |
| < H.S. Diploma 2012-16 | int | People without a High School Diploma | None |
| H.S. Diploma Only 2012-16 | int | People with a High School Diploma but no Degree | None |
| Population Estimate 2016 | int | Total size or composition of populations for 2016 | None |
| GQ Estimates 2016 | int | Total Size of People living in group housing facilities in 2016 | None |
| UR 2016 | int | Unemployment Rate Percentage of people unemployed but actively looked for work in 2016 | None |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **DIMENSION TABLE Education.DimAreaName1** | | | |  | **DIMENSION TABLE Education.DimState** | | | | | |
| **Column Name** | **Data Type** | **Description** | **Key** |  | **Column Name** | **Data Type** | | **Description** | **Key** |
| Area Name ID | int | Primary Key | Primary |  | State ID | int | | Primary Key | Primary |
| Area Name | varchar(8000) | Name of Area | None |  | State | varchar(8000) | | Name of State | None |
| State ID | int | Foreign Key | Foreign |  |  |  |  | |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **DIMENSION TABLE Education.DimCode** | | | |
| **Column Name** | **Data Type** | **Description** | **Key** |
| Code ID | int | Primary Key | Primary |
| Rural Urban Continuum Code 2003 | int | 9-level classification of counties by metro-nonmetro status, location, and urban size 2003 | None |
| Urban Influence Code 2003 | int | 12-level classification of counties by metro-micro-nonmetro status, location, and size of largest place 2003 | None |
| Rural Urban Continuum Code 2013 | int | 9-level classification of counties by metro-nonmetro status, location, and urban size 2013 | None |
| Urban Influence Code2013 | int | 12-level classification of counties by metro-micro-nonmetro status, location, and size of largest place 2013 | None |
| Metro 2013 Code | int | Classification of counties by metro or non-metro; 1 = metro county and 0 = non-metro county | None |

# Data Mart ETL Description:

# 

This ETL process was built with SQL Server Data Tools, based on the SQL Server Integration Services template provided by Visual Studio. For our ETL process we took our cleaned Excel data and developed a staging table for our fact table data and dimensional table data. The staging table data was sourced from a flat file and the data types were transformed and values were then mapped to align with our database designs in SQL Server. From that point a simple combination of SSIS tasks were utilized to systematically load and alter the fact and dimension tables. These steps allow for dynamic changes to the data, table structure and simple and efficient loading and reloading of schemas. For basic loading of the dimension tables we used an OLE DB source and OLE DB destination data flow task, pulling data from our fact staging table. To map and assign the primary key values of the dimensions as surrogate keys, execute SQL tasks were utilized along with query building functions to update the fact staging table data to reflect the assigned surrogate keys. From that point the fact staging table served as a source for the final fact table, titled: Education.FactTable. This final load used the same basic loading process mentioned previously.

## SSIS Package Details:

|  |  |  |  |
| --- | --- | --- | --- |
| **CONTROL FLOW NAME** | **TYPE OF TASK** | **DATA FLOW NAME** | **DESCRIPTION** |
| **Get Socioeconomic data from CSV** | Data Flow Task | STEP1-Get source data from CSV | Connects to the .csv source data file and extracts the data |
| **Get Socioeconomic data from CSV** | Data Flow Task | STEP2-Transform staging data | Transforms columns to integers and decimal values |
| **Get Socioeconomic data from CSV** | Data Flow Task | STEP3-Load fact data into SQL Server | Moves data into final fact Staging Table. Loads data to Fact Staging Table |

|  |  |  |  |
| --- | --- | --- | --- |
| **CONTROL FLOW NAME** | **TYPE OF TASK** | **DATA FLOW NAME** | **DESCRIPTION** |
| **Load State Dimension** | Data Flow Task | STEP1-Grab States from Staging Table | Connects to SQL Server data regarding States |
| **Load State Dimension** | Data Flow Task | STEP2-Load States into State dimension | Transforms columns to integers and decimal values Loads data to State Dimension Table |

|  |  |  |  |
| --- | --- | --- | --- |
| **CONTROL FLOW NAME** | **TYPE OF TASK** | **DATA FLOW NAME** | **DESCRIPTION** |
| **Update State ID into Fact Staging Table** | Execute SQL Task | STEP1-Update State ID to change null values | Connects the database source and builds update query |
| **Update State ID into Fact Staging Table** | Execute SQL Task |  | Executes UPDATE query and loads data to Fact Staging Table and assigns surrogate key |

|  |  |  |  |
| --- | --- | --- | --- |
| **CONTROL FLOW NAME** | **TYPE OF TASK** | **DATA FLOW NAME** | **DESCRIPTION** |
| **Load Area Name Dimension** | Data Flow Task | STEP1-Grab Area Name and StateID from Staging Table | Pulls SQL server data for Area Name dimension |
| **Load Area Name Dimension** | Data Flow Task | STEP2-Load Area Name and StateID into DimAreaName1 | Moves source data into queue for SQL Server Loads data to Area Name Dimension(DimAreaName1) |

|  |  |  |  |
| --- | --- | --- | --- |
| **CONTROL FLOW NAME** | **TYPE OF TASK** | **DATA FLOW NAME** | **DESCRIPTION** |
| **Update Area Name ID into Fact Staging Table** | Execute SQL Task | STEP1-Update Area Name ID into Fact Staging Table | Connects to database and builds Update query |
| **Update Area Name ID into Fact Staging Table** | Execute SQL Task |  | Executes UPDATE query & loads data to Fact Staging Table & assigns surrogate key |

|  |  |  |  |
| --- | --- | --- | --- |
| **CONTROL FLOW NAME** | **TYPE OF TASK** | **DATA FLOW NAME** | **DESCRIPTION** |
| **Load Code Dimension** | Data Flow Task | STEP1-Get Code data | Connects to database source (Fact Staging Table) |
| **Load Code Dimension** | Data Flow Task | STEP2-Load Code Dimension | Maps and moves SQL server data to Code Dimension Loads data to DimCode dimension |

|  |  |  |  |
| --- | --- | --- | --- |
| **CONTROL FLOW NAME** | **TYPE OF TASK** | **DATA FLOW NAME** | **DESCRIPTION** |
| **Execute SQL Task** | Data Flow Task | STEP1-Update Code ID into Staging | Connects to database source and builds update query |
| **Execute SQL Task** | Execute SQL Task |  | Executes UPDATE query to load data and assign Surrogate Keys to Fact Staging Table |

|  |  |  |  |
| --- | --- | --- | --- |
| **CONTROL FLOW NAME** | **TYPE OF TASK** | **DATA FLOW NAME** | **DESCRIPTION** |
| **Load Fact Socioeconomic** | Data Flow Task | STEP1-Get Fact Staging Data | Connects to database source (Fact Staging Table) |
| **Load Fact Socioeconomic** | Data Flow Task | STEP2-Load Fact Table | Maps and moves SQL server data to Code Dimension and Loads data to Education.FactTable |