Community Health Status Indicator Information Management Solution

Team Number: 12

Team Members:

- Rahil Kapadia (001300490)
- Kunj Mithapara (001400004)
- Shrishti Khanna (001315098)
- Abhilash Hemaraj(001305283)

Course: INFO 7290

Course Name: Data Warehousing and Business Intelligence

Guided By: Prof. Vincent Lattuada

Objectives:

- To find multiple sources of data to form a substantial SOR required for warehousing.
- To provide a brief walk through of building a data warehouse that will house Health-centric demographic data granularized at the 'county' level.
- To implement advanced Data- warehousing techniques while answering questions with a Business perspective.
- To provide an overview of the Data Flow Process.
- Serve as Version - log for future references

Version History:

Date	Submission	Description
11/05/2020	Project Proposal-1	First Draft
11/15/2020	Project Proposal-2	Modified the First draft with the changes instructed by professor
12/04/2020	Project Design Doc	Added some more design related details like data flow, fields, olap cubes, SSIS snapshots, Visualizations
12/16/2020	Final Draft	Updated dimensional model, and listed additional features

Table of Contents:

1.Description of the data to be stored and moved into the analytics system

1.1.Data Source Description

Community Health Status Indicators (CHSI) to combat obesity, heart disease, and cancer are major components of the Community Health Data Initiative.

1. **Leading Cause of Death:** This dataset contains details leading cause of deaths among various American counties, it gives detailed information regarding the age group and cause of death.
https://data.world/health/chsi-to-combat-obesity/workspace/file?filename=chsi_dataset%2FLEADINGCAUSESOFDEATH.csv

2. **Preventive Services Used across the counties:** This data source gives us the idea about various preventive measures in terms of vaccine taken across the counties. It includes data of vaccines like Hepatitis, Measles, Syphilis, Flu Vaccine. It also gives a bifurcation among the age groups and vaccines taken by them. This data can be joined with the previous set to analyse preventive services vs deaths.

https://data.world/health/chsi-to-combat-obesity/workspace/file?filename=chsi_dataset%2FPREVENTIVESERVICESUSE.csv

3. **Risk Factors and Access to Care:** This data source adds a new aspect to the analysis plan. It consists of data regarding the risk factor among counties like resident with high blood pressure, diabetes, obesity. The data source also contains data regarding the accessibility to care like elderly Medicare, insurance, disability care, dentists, etc.

https://data.world/health/chsi-to-combat-obesity/workspace/file?filename=chsi_dataset%2FRISKFACTORSANDACCESSTOCARE.csv

4. **Demographics:** This data source is our base source as it contains information regarding the population density of each county. This dataset also contains percentiles in respective to peer counties.
https://data.world/health/chsi-to-combat-obesity/workspace/file?filename=chsi_dataset%2FDEMOGRAPHICS.csv

1.2. Joining Data Sources

The datasets Leading Cause of Death and Preventive Services Used across the counties,Risk Factors and Access to Care, can be joined by a composite primary key formed by the State_FIPS_Code and County_FIPS_Code. We are forming a composite primary key here as

the State_FIPS_Code and County_FIPS_Code contains repeated values but if used together they form a unique key for each record.

1.3.Data Transformations

Our dataset contains following indicators:

	data_value	description
1	-9999	Indicate N.A. value from the source data for the Unemployed column on the VUNERABLEPOPSANDEN
2	-2222 or -2222.2 or -2	nda, no data available, see Data Notes document for details
3	-1111.1 or -1111 or -1	nrf, no report, see Data Notes document for details
4	1	Represent 'No' in the indicator columns
5	2	Represent 'Yes' in the indicator columns
6	3	Represent 'Favorable to peers' in the indicator columns
7	4	Represent 'Unfavorable to peers' in the indicator columns
8	5	Represent ''Favorable to peers and favorable the U.S. Rate' in the indicator columns
9	6	Represent 'Favorable to peers and unfavorable the U.S. Rate' in the indicator columns
10	7	Represent 'Unfavorable to peers and favorable the U.S. Rate' in the indicator columns
11	8	Represent 'Unfavorable to peers and unfavorable the U.S. Rate' in the indicator columns
12	-9998.9	Indicate no objective for the Healthy People 2010 Target data

Fig : 1- Data transformations

The dataset contains numeric value in place of indicators like “Yes” or “No” for example 1 represents “No” 2 represents “Yes”. We plan to transform those numeric values to “Yes”, “No”, “Favorable”, “Unfavorable” so we can use them as categories during analysis.

The dataset has been built keeping in mind the various missing values encountered in the real world. Some of these values look like extreme integers such as ‘-9999’, ‘-1111’, etc. For simplicity we will transform these values into ‘NULL’ values.

As shown above we need to assign a consistent value to these indicators because if they are not transformed they may produce erroneous results when aggregated leading to a false analysis.

1.4 Dataset Dimensions: These datasets have following dimensions.

- Age
- Race
- State and County
- Disease
- Cases

- Testing
- Risk factors

1.5 Fields and DataTypes

Dataset: Leading Causes of Death
Corresponding Staging Table: Leading_Causes_Staging
Corresponding Dimension Table: Dim.Leading_Causes

IN DATASET	FOR STAGING	DATATYPE	SIGNIFICANCE
STATE_FIPS_CODE	STATE_FIPS_CODE	INT	Two-digit state identifier
COUNTY_FIPS_CODE	COUNTY_FIPS_CODE	INT	Three-digit county identifier
CHSI_STATE_NAME	STATE_NAME	VARCHAR	Name of State
CHSI_COUNTY_NAME	COUNTY_NAME	VARCHAR	Name of county
CHSI_STATE_ABBR	STATE_ABBR	CHAR(2)	Two-character abbreviation for state name
B_WH_CANCER	CANCER_WH_A1	INT	County data, ages 1-14, cancer, White
B-BL-CANCER	CANCER_BL_A1	INT	County data, ages 1-14, cancer, Black
B_Ot_Cancer	CANCER_OT_A1	INT	County data, ages 1-14, cancer, Other

C_Wh_Cancer	CANCER_WH_A2	INT	County data, ages 15-24, cancer, White
C_BL_CANCER	CANCER_BL_A2	INT	County data, ages 15-24, cancer, Black
C-OT_CANCER	CANCER_OT_A2	INT	County data, ages 15-24, cancer, Other
D-WH-CANCER	CANCER_WH_A3	INT	County data, ages 25-44, cancer, White
D-BL-CANCER	CANCER_BL_A3	INT	County data, ages 25-44, cancer, Black
D_ot-CANCER	CANCER_OT_A3	INT	County data, ages 25-44, cancer, other
E-WH-CANCER	CANCER_WH_A4	INT	County data, ages 45-64, cancer, white
E_BL_CANCER	CANCER_BL_A4	INT	County data, ages 45-64, cancer, black
E_Ot_Cancer	CANCER_OT_A4	INT	County data, ages 45-64, cancer, other
F_WH_CANCER	CANCER_WH_A5	INT	County data, ages 65+, cancer, white
F_BL_CANCER	CANCER_BL_A5	INT	County data, ages 65+, cancer, black

F_Ot_Cancer	CANCER_OT_A5	INT	County data, ages 65+, cancer, white
D_Wh_HeartDis	HD_WH_A3	INT	County data, ages 25-44, heart disease, White
D_Bl_HeartDis	HD_BL_A3	INT	County data, ages 25-44, heart disease, Black
D_Ot_HeartDis	HD_OT_A3	INT	County data, ages 25-44, heart disease, Other
E_Wh_HeartDis	HD_WH_A4	INT	County data, ages 45-64, heart disease, White
E_BL_HeartDis	HD_BL_A4	INT	County data, ages 45-64, heart disease, Black
E_Ot_HeartDis	HD_OT_A4	INT	County data, ages 45-64, heart disease, other
F_Wh_HeartDis	HD_WH_A5	INT	County data, ages 65+, heart disease, White
F_BL_HeartDis	HD_BL_A5	INT	County data, ages 65+, heart disease, Black
F_OT_HeartDis	HD_OT_A5	INT	County data, ages 65+, heart disease, Other

D_Wh_HIV	HIV_WH_A3	INT	County data, ages 25-44, hiv/aids, White
D_BL_HIV	HIV_BL_A3	INT	County data, ages 25-44, hiv/aids, Black
D_Ot_HIV	HIV_OT_A3	INT	County data, ages 25-44, hiv/aids, Other
C_Wh_Suicide	SUI_WH_A2	INT	County data, ages 15-24, suicide, White
C_Bl_Suicide	SUI_BL_A2	INT	County data, ages 15-24, suicide, Black
C_Ot_Suicide	SUI_OT_A2	INT	County data, ages 15-24, suicide, Other
D_Wh_Suicide	SUI_WH_A3	INT	County data, ages 25-44, suicide, White
D_Bl_Suicide	SUI_BL_A3	INT	County data, ages 25-44, suicide, Black
D_Ot_Suicide	SUI_OT_A3	INT	County data, ages 25-44, suicide, Other
A_Wh_BirthDef	BD_WH_A	INT	County data, under age 1, birth defects, White

A_Bl_BirthDef	BD_BL_A	INT	County data, under age 1, birth defects, Black
A_Ot_BirthDef	BD_OT_A	INT	County data, under age 1, birth defects, Other

Dataset: Risk Factors And Access To Care
Corresponding Staging Table: stg_Riskfactors_Access_Care
Corresponding Dimension Table: Dim.RiskFactors

IN DATASET	FOR STAGING	**DATATYPE**	SIGNIFICANCE
STATE_FIPS_CODE	STATE_FIPS_CODE	INT	Two-digit state identifier
COUNTY_FIPS_CODE	COUNTY_FIPS_CODE	INT	Three-digit county identifier
NO_EXERCISE	NO_EXERCISE	INT	Deaths due to No Exercise
OBESITY	OBESITY	VARCHAR	Deaths due to Obesity
HIGH_BLOOD_PRESSURE	HIGH_BLOOD_PRESSURE	INT	Deaths due to High Blood Pressure
SMOKER	SMOKER	INT	Deaths due to smoking

DIABETES	DIABETES	INT	Deaths due to diabetes
UNINSURED	UNINSURED	INT	Deaths of the Uninsured
ELDERLY_MEDICARE	ELDERLY_MEDICARE	INT	Deaths of the Elderly who had Medicare
DISABLED_MEDICARE	DISABLED_MEDICARE	INT	Deaths of the Disabled who had Medicare

Dataset: Demographics
Corresponding Staging Table: Stg_Demographics
Corresponding Dimension Table: Dim.Demographics

IN DATASET	FOR STAGING	**DATATYPE**	SIGNIFICANCE
STATE_FIPS_CODE	STATE_FIPS_CODE	INT	Two-digit state identifier
COUNTY_FIPS_CODE	COUNTY_FIPS_CODE	INT	Three-digit county identifier
CHSI_STATE_NAME	STATE_NAME	VARCHAR	Name of State
CHSI_COUNTY_NAME	COUNTY_NAME	VARCHAR	Name of county
CHSI_STATE_ABBR	STATE_ABBR	CHAR(2)	Two-character abbreviation for state name
Number_Counties	n_counties	INT	Number of Peer Counties

Population_Size	population	INT	County data, population size
Population_Density	population_density	INT	'County data, population density (people per square mile)
Poverty	poverty	DECIMAL	County data, individuals living below poverty level
Age_19_Under	under_19	INT	County data, population under age 19
Age_19_64	between_19_64	DECIMAL	County data, population age 19-64
Age_65_84	between_65_84	DECIMAL	County data, population age 65-84
Age_85_and_Over	above_85	DECIMAL	County data, population age 85+
White	white	DECIMAL	County data, White
Black	black	DECIMAL	County data, Black
Native_American	native	DECIMAL	County data, American Indian
Asian	asian	DECIMAL	County data, Asian/Pacific Islander
Hispanic	hispanic	DECIMAL	County data, Hispanic origin

Dataset: Preventive Services Use
Corresponding Staging Table: Preventive_Services_Use
Corresponding Dimension Table: Dim_Preventive_Services

IN DATASET	FOR STAGING	**DATATYPE**	SIGNIFICANCE
STATE_FIPS_CODE	STATE_FIPS_CODE	INT	Two-digit state identifier
COUNTY_FIPS_CODE	COUNTY_FIPS_CODE	INT	Three-digit county identifier
CHSI_STATE_NAME	STATE_NAME	VARCHAR	Name of State
CHSI_COUNTY_NAME	County_Name	VARCHAR	Name of county
CHSI_STATE_ABBR	STATE_ABBR	CHAR(2)	Two-character abbreviation for state name
FlueB_Rpt	InfluenzaeB_reported_cases	INT	County data, Haemophilus Influenzae B reported cases
FlueB_Exp	InfluenzaeB_Expected_cases	INT	County data, Haemophilus Influenzae B expected cases
HepA_Rpt	HepatitisA_reported_Cases	INT	County data, Hepatitis A reported cases

HepA_Exp	HepatitisA_Expected_Cases	INT	County data, Hepatitis A expected cases
HepB_Rpt	HepatitisB_reported_Cases	INT	County data, Hepatitis B reported cases
HepB_Exp	HepatitisB_Expected_Cases	INT	County data, Hepatitis B expected cases
Meas_Rpt	Measles_reported_Cases	INT	County data, Measles reported cases
Meas_Exp	Measles_Expected_Cases	INT	County data, Measles expected cases
Pert_Rpt	Pertussis_reported_cases	INT	County data, Pertussis reported cases
Pert_Exp	Pertussis_Expected_cases	INT	County data, Pertussis expected cases
CRS_Rpt	Congenital_Rubella_Syndrome reported cases	INT	County data, Congenital Rubella Syndrome reported cases
CRS_Exp	Congenital_Rubella_Syndrome_expected_cases	INT	County data, Congenital Rubella Syndrome expected cases
Syphilis_Rpt	Syphilis_reported_cases	INT	County data, Syphilis reported cases

Syphilis_Exp	Syphilis_Expected_cases	INT	County data, Syphilis expected cases
ID_Time_Span	Time_period	INT	Time period of reported data for infectious disease cases in the preventive services use
Pap_Smear	pap_smears_tests_count	INT	County data, pap smears, age 18+
Mammogram	mammography_tests_count	INT	County data, mammography , age 50+
Proctoscopy	Sigmoidoscopy_tests_count	INT	County data, sigmoidoscopy , age 50+
Pneumo_Vax	pneumonia_vaccine_count	INT	County data, pneumonia vaccine, age 65+
Flu_Vac	flu_vaccine_count	INT	County data, flu vaccine, age 65+

2.1.Dataflow Process: A high level overview:

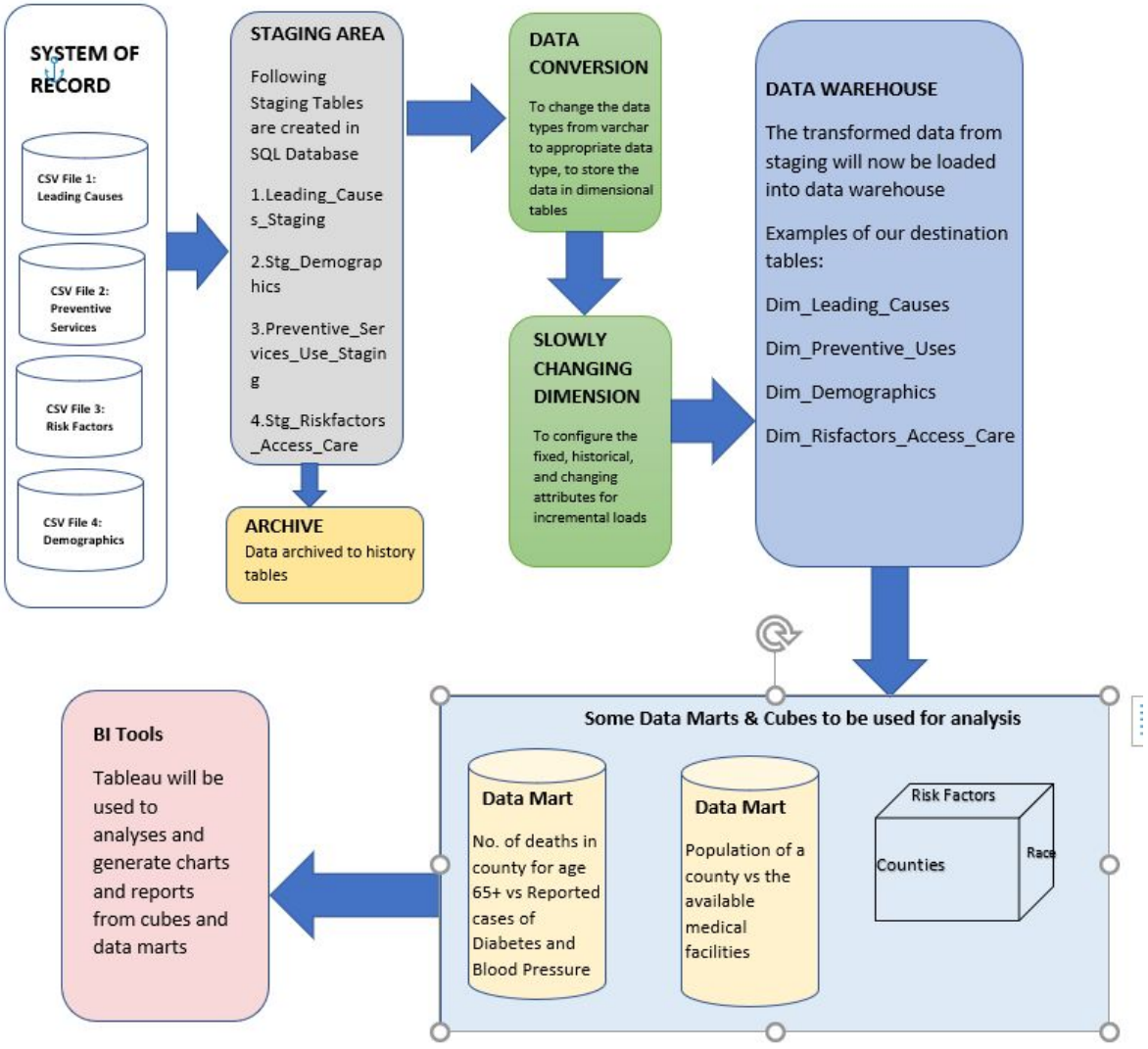


Fig : 2- Data Flow overview

2.2 Dimensional Model

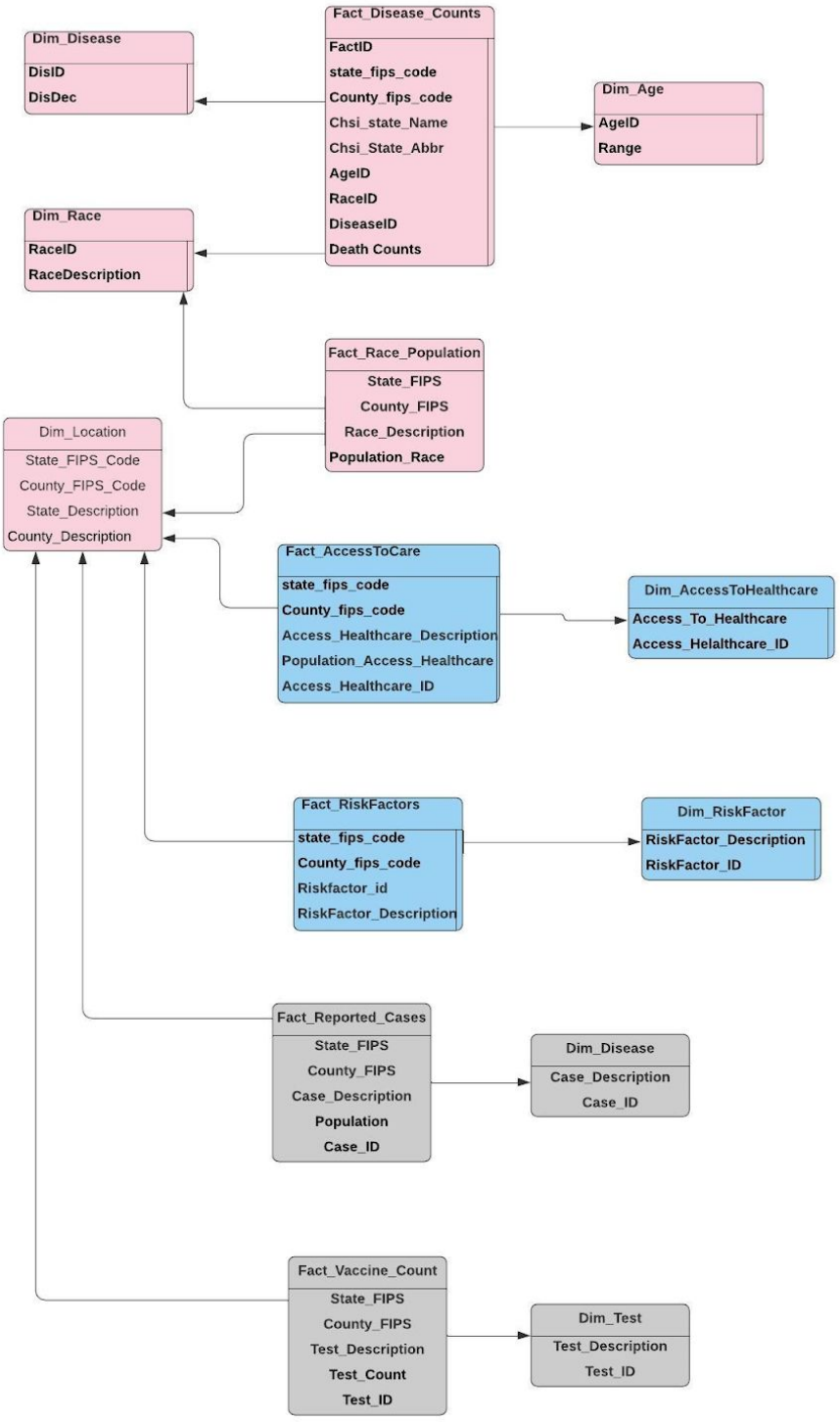


Fig : 3- Dimensional Model

2.3 Incremental loads (slowly changing dimensions)

For the first time loading the dataset can be loaded through SSIS to a staging area and then to the data warehouse. In this process we configure our SSIS package by mapping the input source fields to the destination fields. The initial SSIS package will flow as following:

Fig : 4- Data Flow Simplified

Monthly Loads: Considering the nature of our data sources the data we receive every month should update the existing data or add new records based on the changing field. We have divided all our dimension attributes in three categories namely : fixed, changing and historical. The value of fixed attributes cannot be manipulated in any case eg: County Code, State Code. The changing attributes can be overwritten and we can optionally maintain their prior values. The historical attributes if changed need to be added as an entirely new record, and flagging the old records as expired.

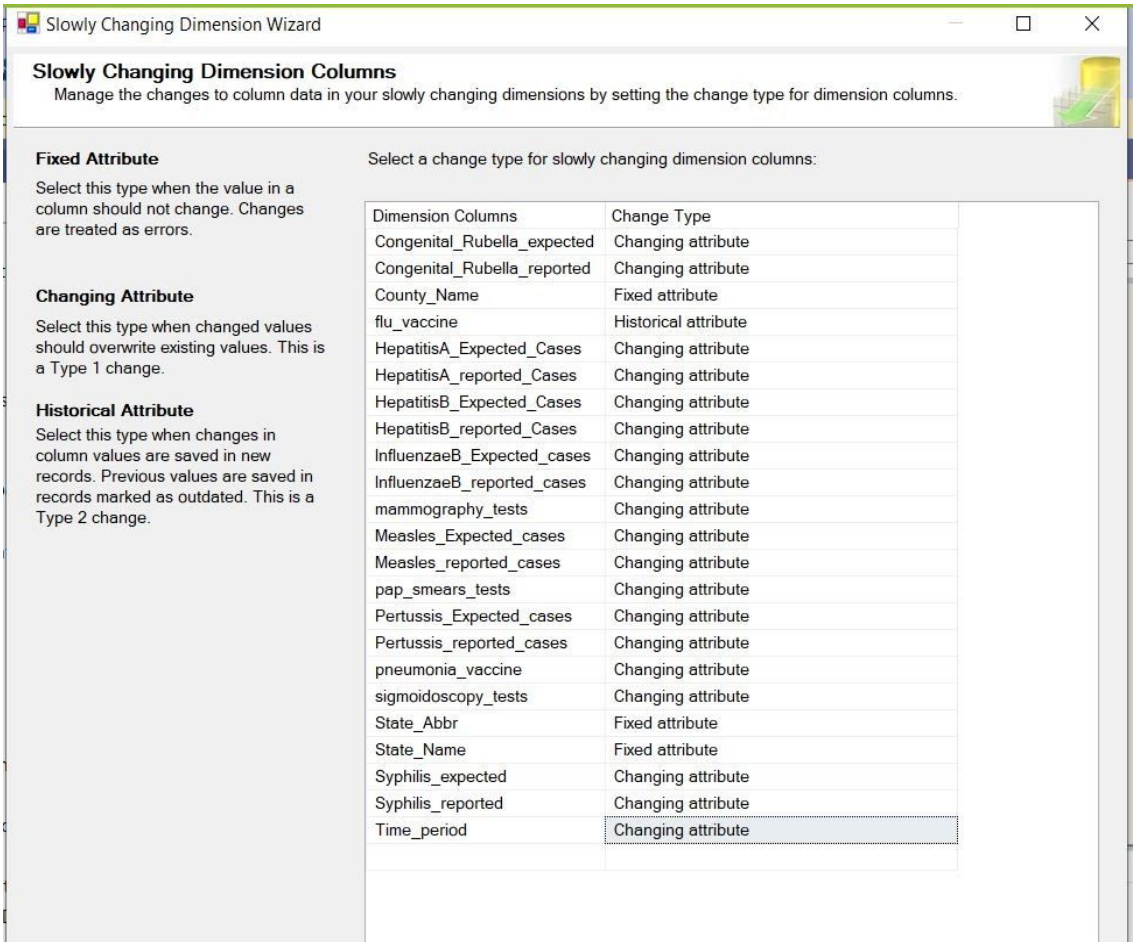


Fig 5: Slowly Changing Dimension Object from SSIS

2.4 Lookup Tables

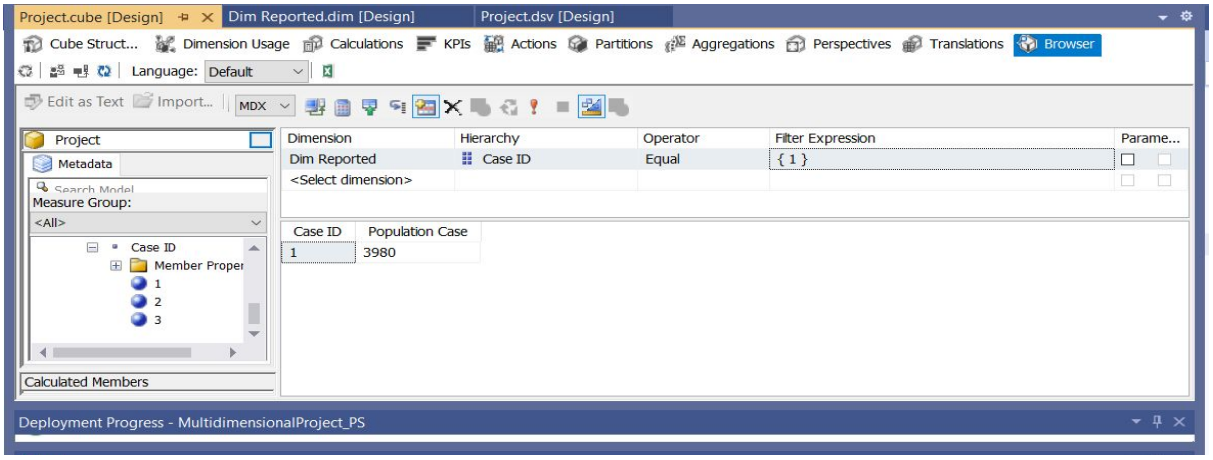
- Lookup tables are used as reference points for decisions regarding correct or incorrect data.
- Lookup tables can be used to check if the data in categorical fields match the listed categories.
- Lookup table for state abbreviation holds the abbreviation for 50 US states, and if data has an abbreviation that does not match then that data is sent to the error table.
- Lookup table for the race field holds values like white, black, hispanic, etc and if there is a race value which does not match it is directed to the error table.

2.5 OLAP Cubes

Keeping reporting and analysis in mind the data can be franchised into different data marts. This is done for snappy reporting and ease of access to relevant data. For e.g. the team that looks after risk factor analysis does not require data related to preventive measures.

A data mart will be created where it will aggregate the data of all the counties of an individual state so we can perform state wise analysis related to the leading cause of death. A sample use case - deaths caused by cancer in NewYork compared to deaths caused by Cancer in Florida.

The other data mart will consist of data comparing the deaths and population data at a county level granularity. This can be done by pulling relevant fields and aggregating to find effective death rate, poverty level, population density, etc. This provides the BI team to look at the overall status of the country and pin-point as to which areas require assistance in terms of medicare and other resources.



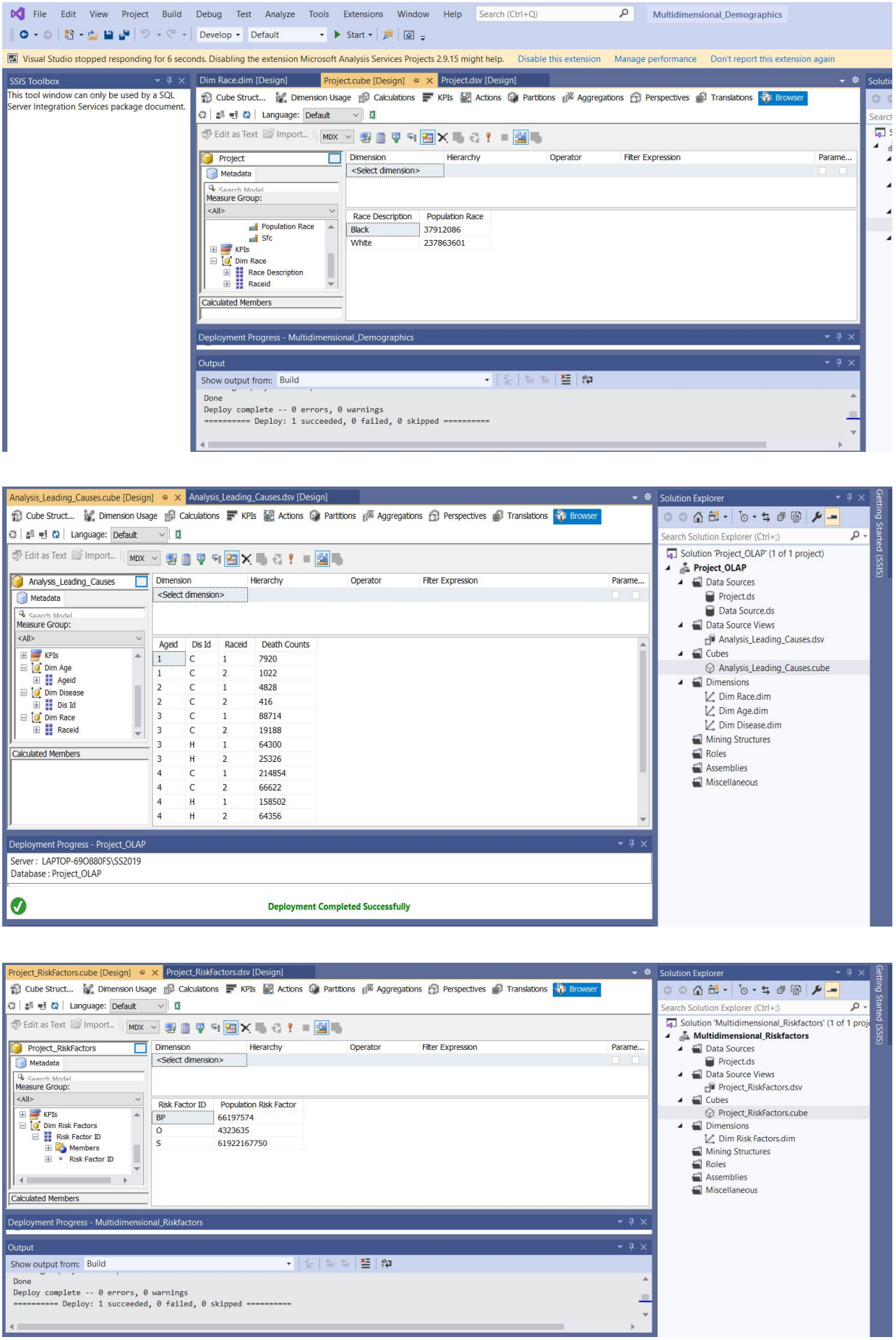
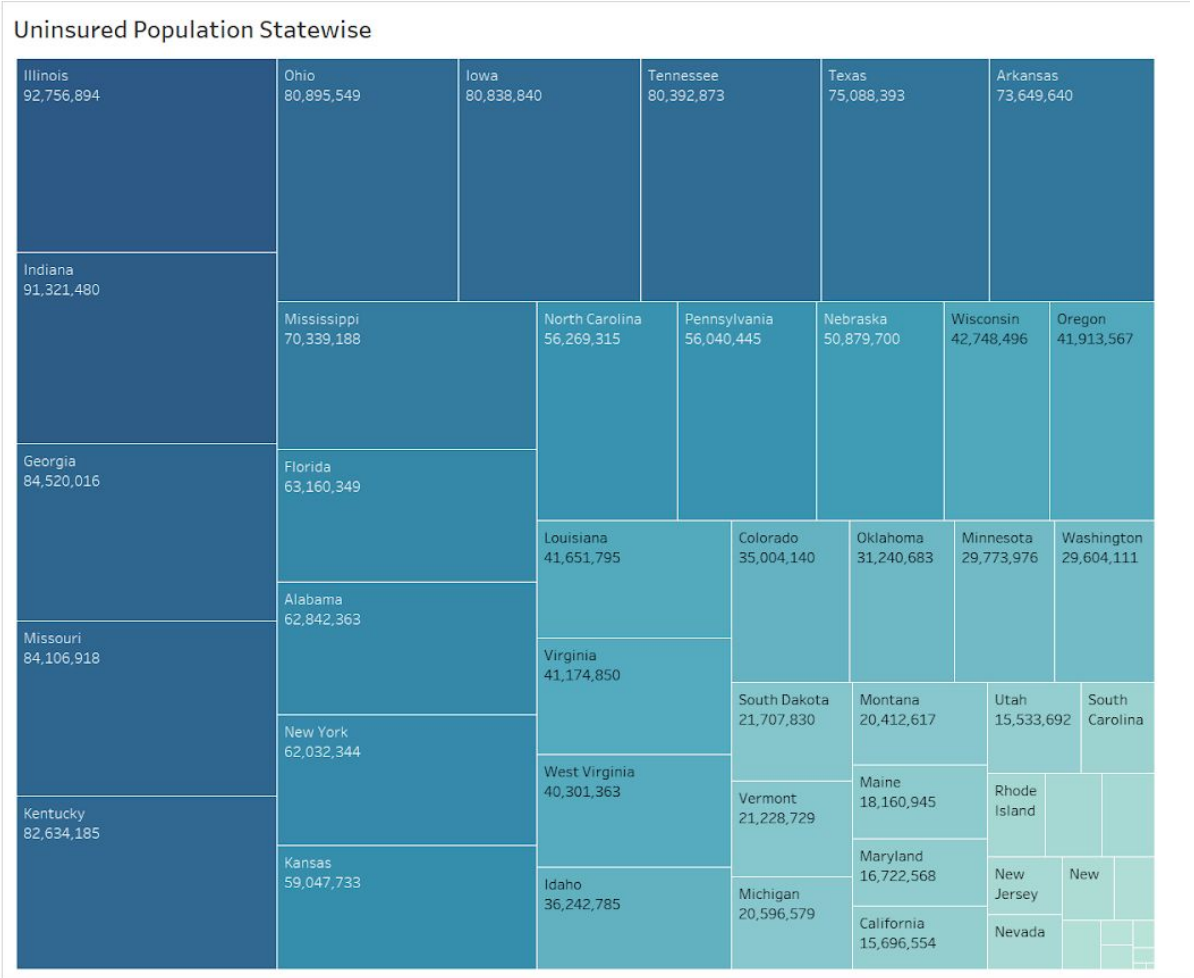


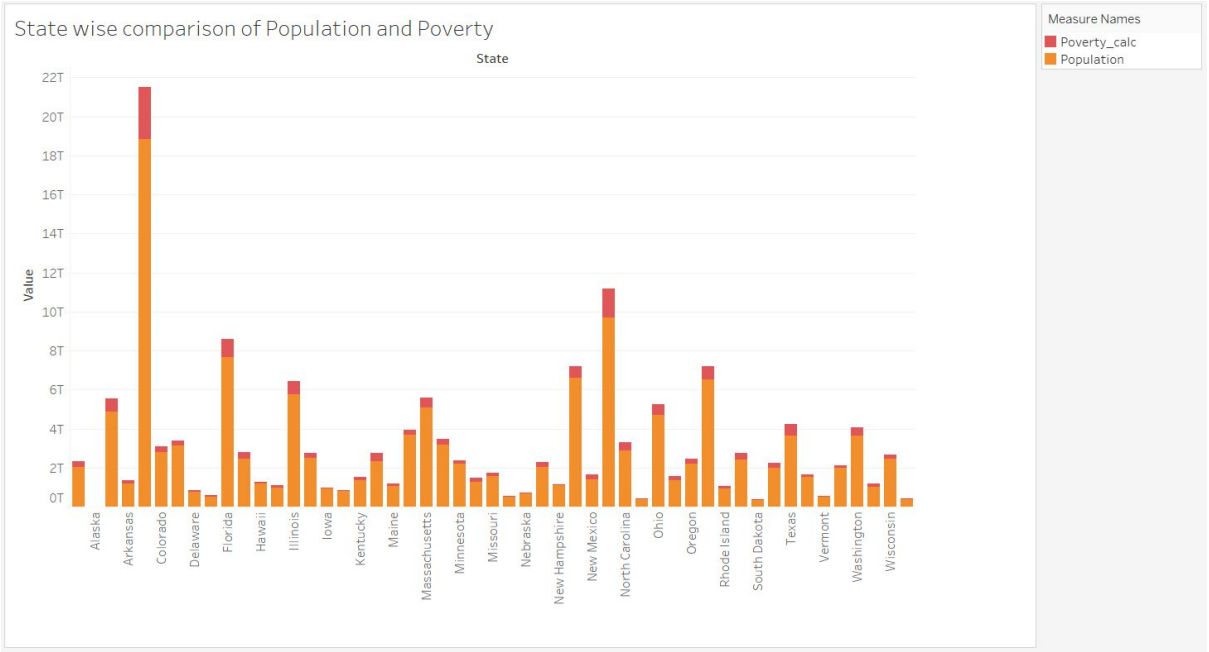
Fig 6: OLAP CUBES:

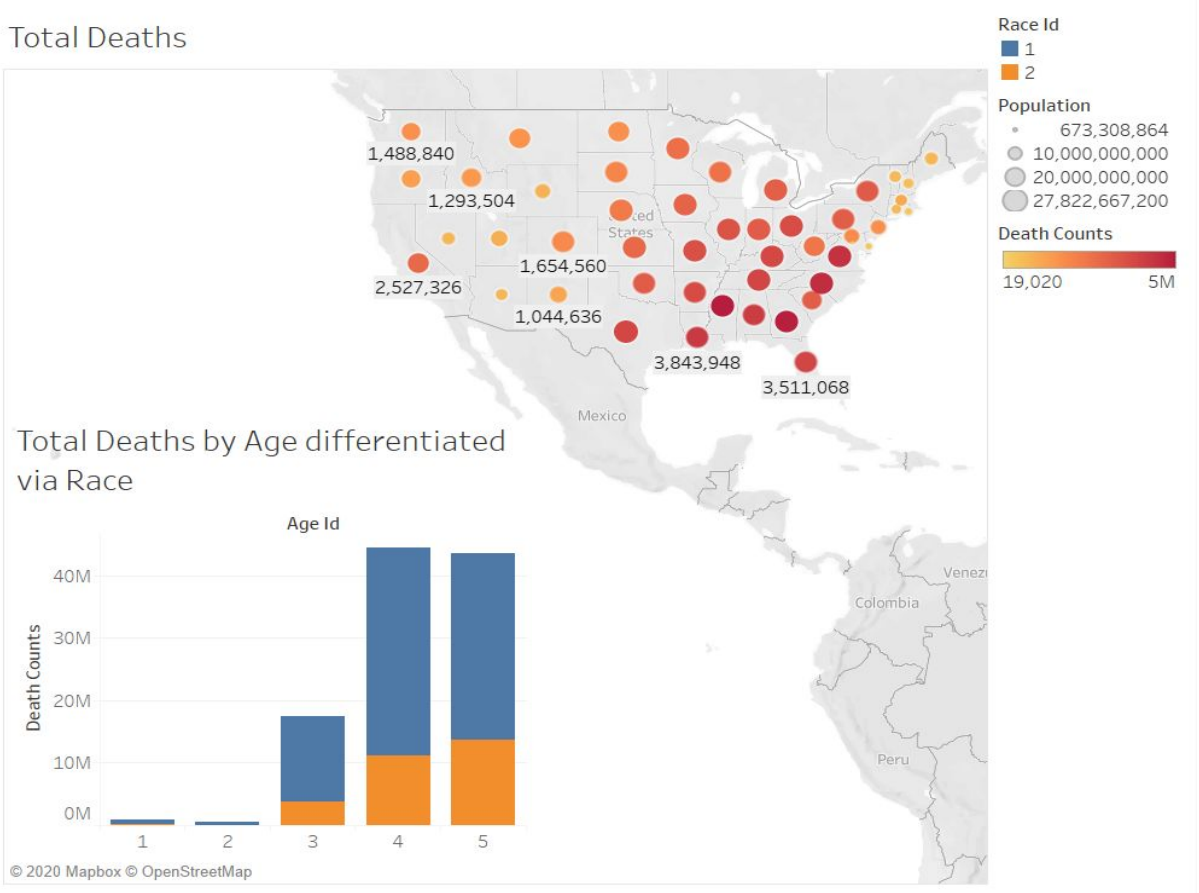
3.1Reporting and Visualization Tools

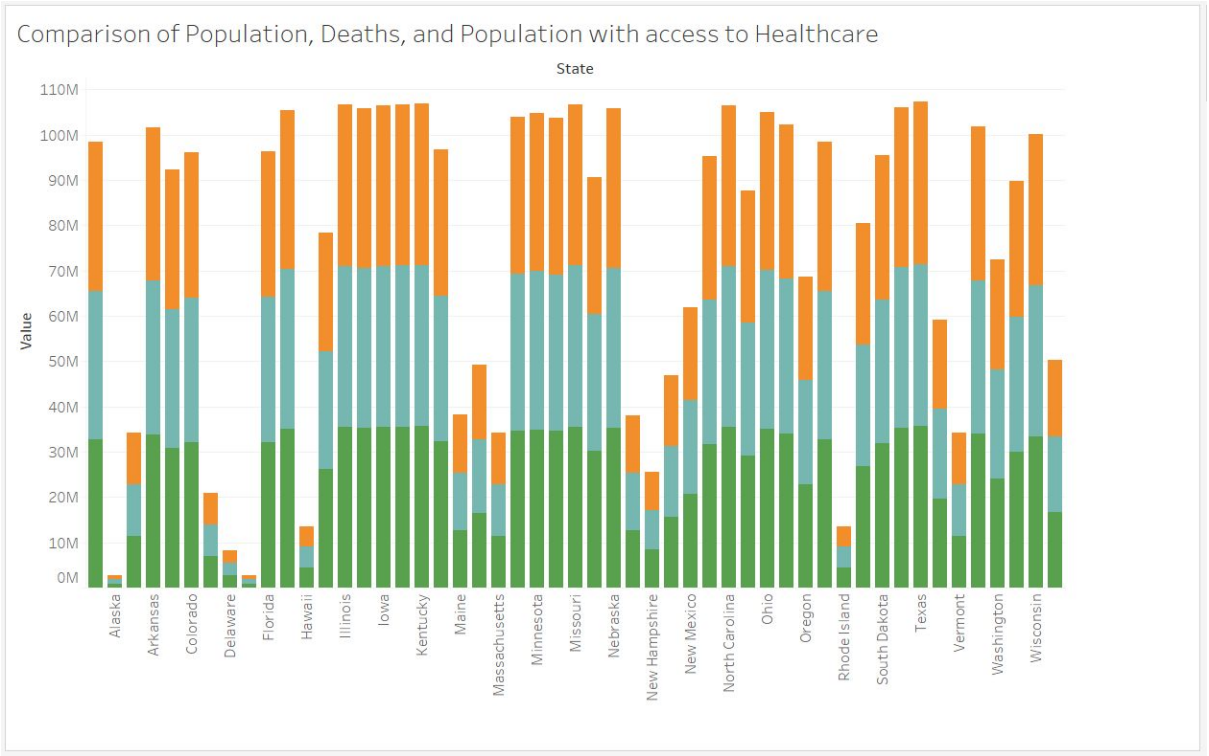
Tableau is an interactive data visualization tool. It's an application that is used for translating raw data into valuable insights. A few advantages of using tableau over using any other visualization tool are :

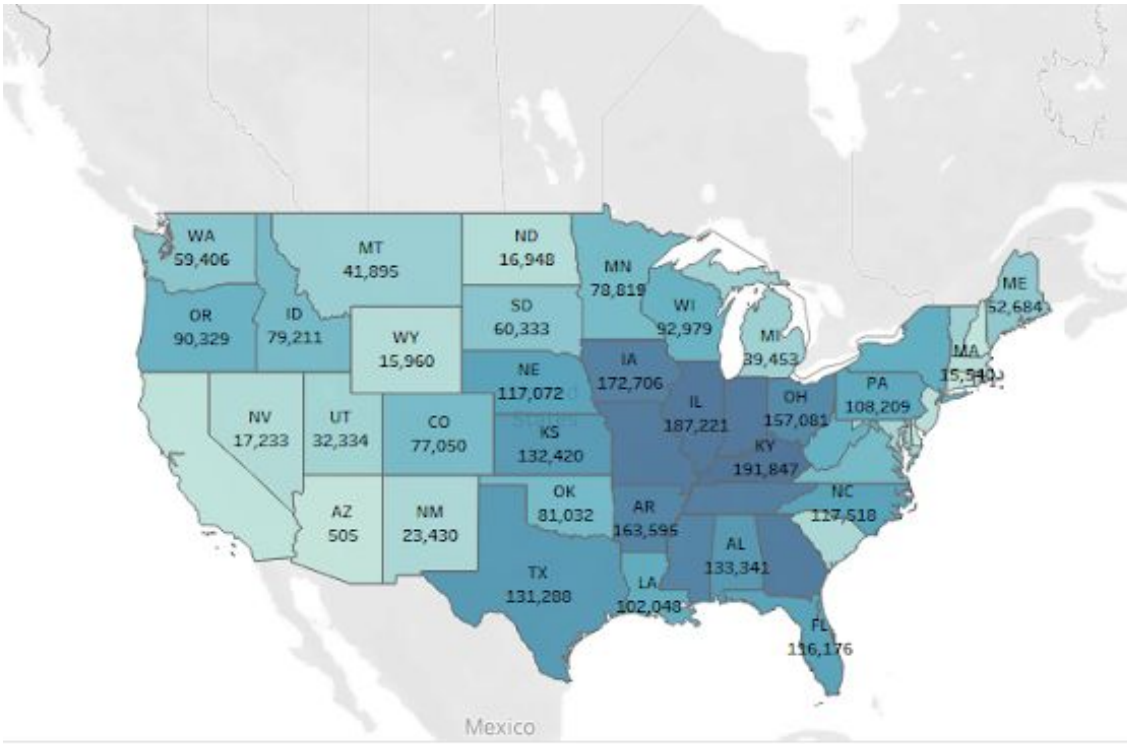
- Remarkable Visualization Capabilities
- Multiple Data Source Connections
- Ease of Use
- High Performance
- Mobile Friendliness




State wise comparison of Population and Poverty
State
Value
22T
20T
18T
16T
14T
12T
10T
8T
6T
4T
2T
0T
Alaska
Arkansas
Colorado
Delaware
Florida
Hawaii
Illinois
Iowa
Kentucky
Maine
Massachusetts
Minnesota
Missouri
Nebraska
New Hampshire
New Mexico
North Carolina
Ohio
Oregon
Rhode Island
South Dakota
Texas
Vermont
Washington
Wisconsin
Measure Names
Poverty_calc
Population


Total Deaths
1,488,840
1,293,504
1,654,560
2,527,326
1,044,636
3,843,948
3,511,068
Mexico
Colombia
Peru
Race Id
1
2
Population
673,308,864
10,000,000,000
20,000,000,000
27,822,667,200
Death Counts
19,020
5M
Total Deaths by Age differentiated via Race
Age Id
Death Counts
40M
30M
20M
10M
0M
1
2
3
4
5
© 2020 Mapbox © OpenStreetMap


Comparison of Population, Deaths, and Population with access to Healthcare
State
Value
110M
100M
90M
80M
70M
60M
50M
40M
30M
20M
10M
0M
Alaska
Arkansas
Colorado
Delaware
Florida
Hawaii
Illinois
Iowa
Kentucky
Maine
Massachusetts
Minnesota
Missouri
Nebraska
New Hampshire
New Mexico
North Carolina
Ohio
Oregon
Rhode Island
South Dakota
Texas
Vermont
Washington
Wisconsin


WA
59,406
MT
41,895
ND
16,948
MN
78,819
OR
90,329
ID
79,211
WY
15,960
SD
60,333
WI
92,979
MI
89,453
ME
52,684
NE
117,072
IA
172,706
MA
PA
108,209
NV
17,233
UT
32,334
CO
77,050
KS
132,420
IL
187,221
OH
157,081
KY
191,847
AZ
505
NM
23,430
OK
81,032
AR
163,595
NC
117,518
AL
133,341
TX
131,288
LA
102,048
FL
116,176
Mexico

4.1 Interesting facts

This dataset provides key health indicators for local communities and encourages dialogue about actions that can be taken to improve community health (e.g., obesity, heart disease, cancer). The CHSI report contains over 200 measures for each of the 3,141 United States counties. The dataset covers factors like obesity, tobacco use, diet, physical activity, alcohol and drug use, sexual behavior and others substantially contribute to these deaths.

The dataset consists of detailed information regarding patients with diabetes, blood pressure, no.of smokers in each county, accessibility to medicaid centers, etc. These parameters form a platform for detailed and meaningful analysis.

The dataset contains figures regarding populations size, minimum and maximum population density , poverty values. This measure helps us to analyse the relation between the deaths, available services and other demographics.

We believed this dataset being detailed and covering multiple aspects helps us effectively analyse the actions to improve community health.

Questions our analysis can address:

The dataset revolves around the health demographic at the county level of the United States. There are a number of insights that can be derived from the sources listed.

1. What diseases ail the age groups provided in the dataset?
2. At what amount are preventive measures being taken towards the various health risks the country is facing?
3. Geo-spatial distribution of various risk factors and access to care to quickly realize which areas are lagging behind.
4. The Error table could also tell us about data consistency coming from different counties. Suggesting which areas need to be more careful with data generation/collection.

Appendix

Screenshots from SSIS Implementation

Fig: Data from flat file(Leading_Causes) to staging table

	STATE_FIPS_CODE	COUNTY_FIPS_CODE	STATE_NAME	COUNTY_NAME	STATE_ABBR	CANCER_WH_A1	CANCER_BL_A1	CANCER_OT_A1
10	5	119	Arkansas	Pulaski	AR	17	0	0
11	6	1	California	Alameda	CA	14	13	17
12	6	13	California	Contra Costa	CA	19	17	0
13	6	19	California	Fresno	CA	15	0	0
14	6	29	California	Kern	CA	13	0	0
15	6	37	California	Los Angeles	CA	20	10	23
16	6	47	California	Merced	CA	17	0	0
17	6	53	California	Monterey	CA	19	0	0
18	6	59	California	Orange	CA	23	0	21
19	6	61	California	Placer	CA	30	0	0
20	6	65	California	Riverside	CA	17	0	0
21	6	67	California	Sacramento	CA	13	0	13
22	6	71	California	San Bernardino	CA	15	0	0
23	6	73	California	San Diego	CA	27	0	22
24	6	77	California	San Joaquin	CA	15	0	0
25	6	79	California	San Luis Obispo	CA	30	0	0

Fig: Data in staging table(SQL DB)

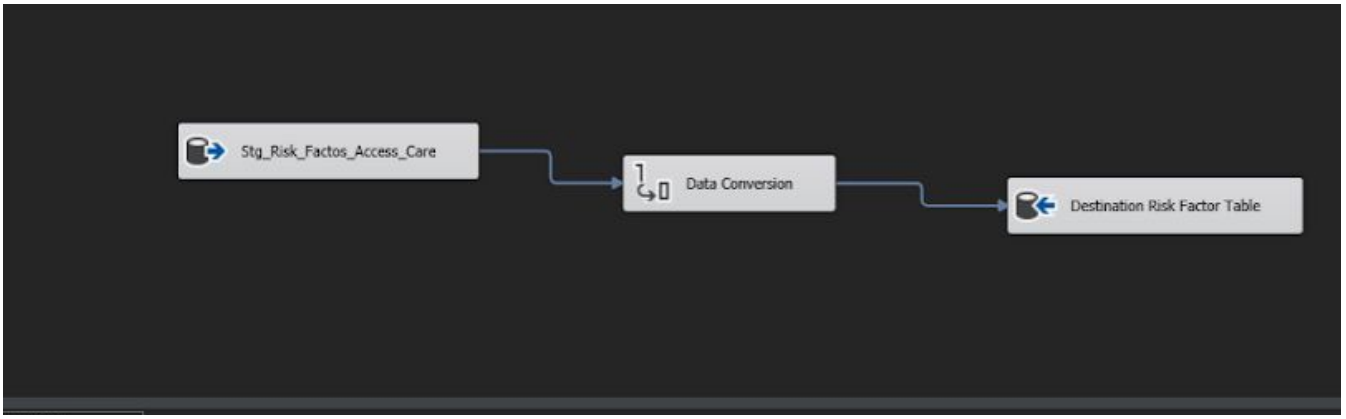


Fig: staging → Data Conversion → Destination

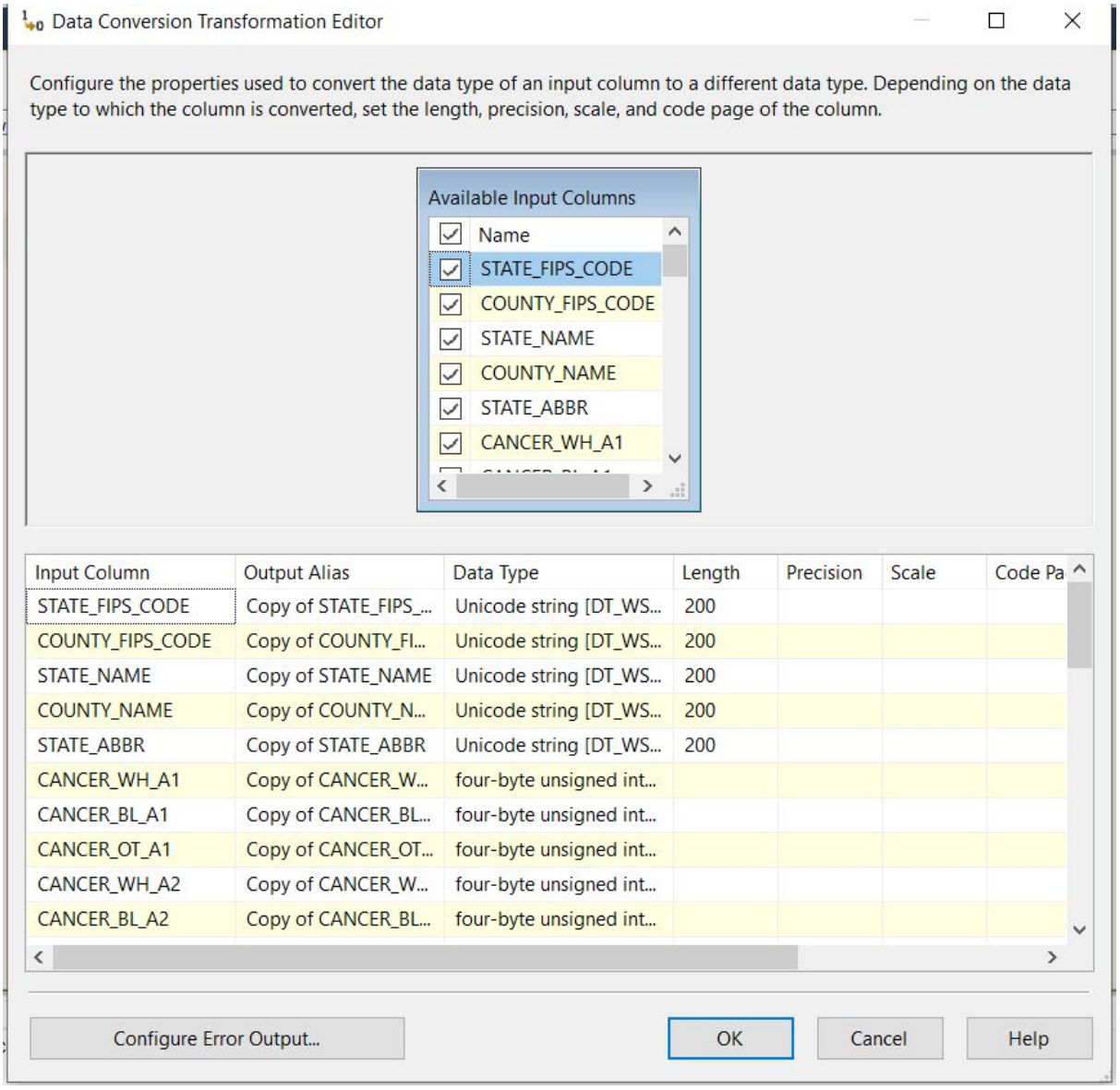


Fig: Data Conversion Transformation Editor

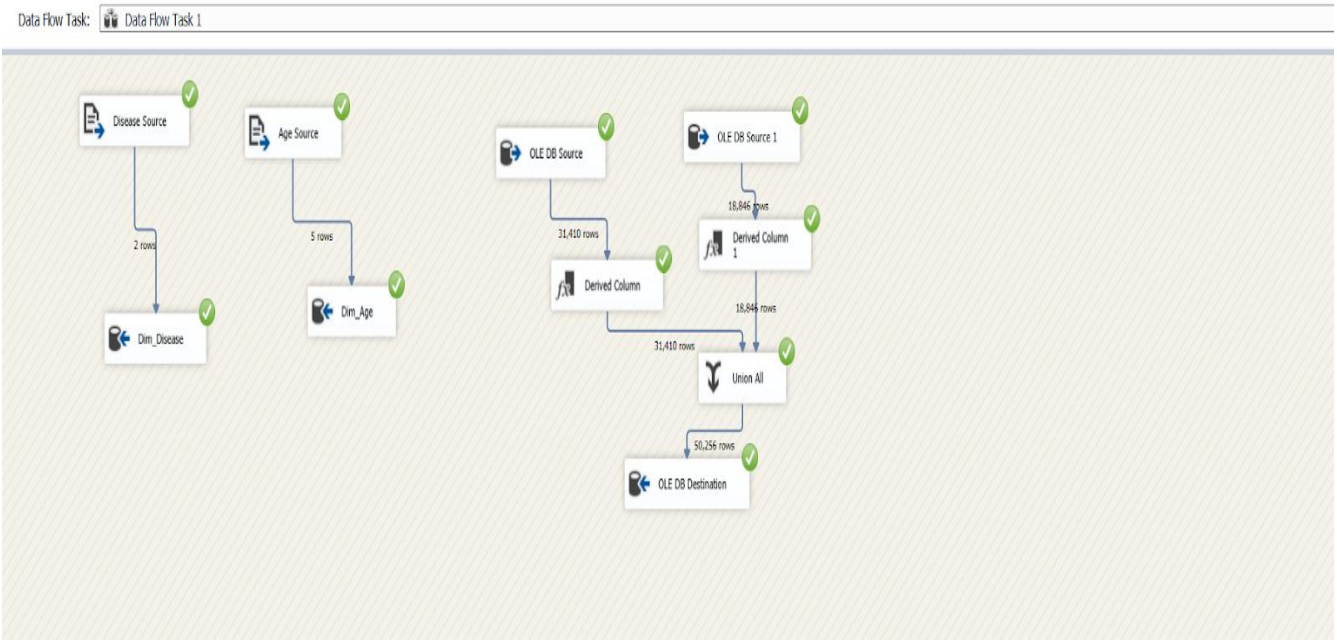


Fig: Dimension and Fact tables load - Leading causes of deaths

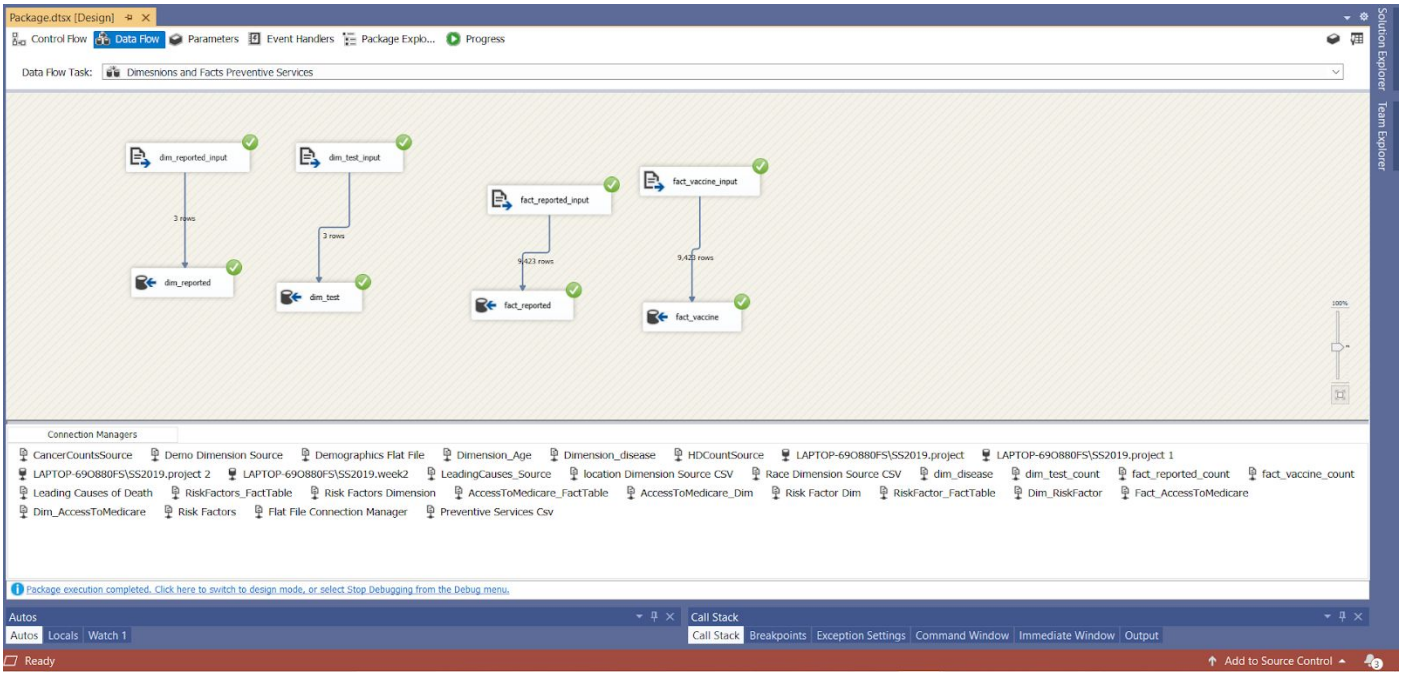


Fig: Fact and Dimension tables load for Preventive Measures.

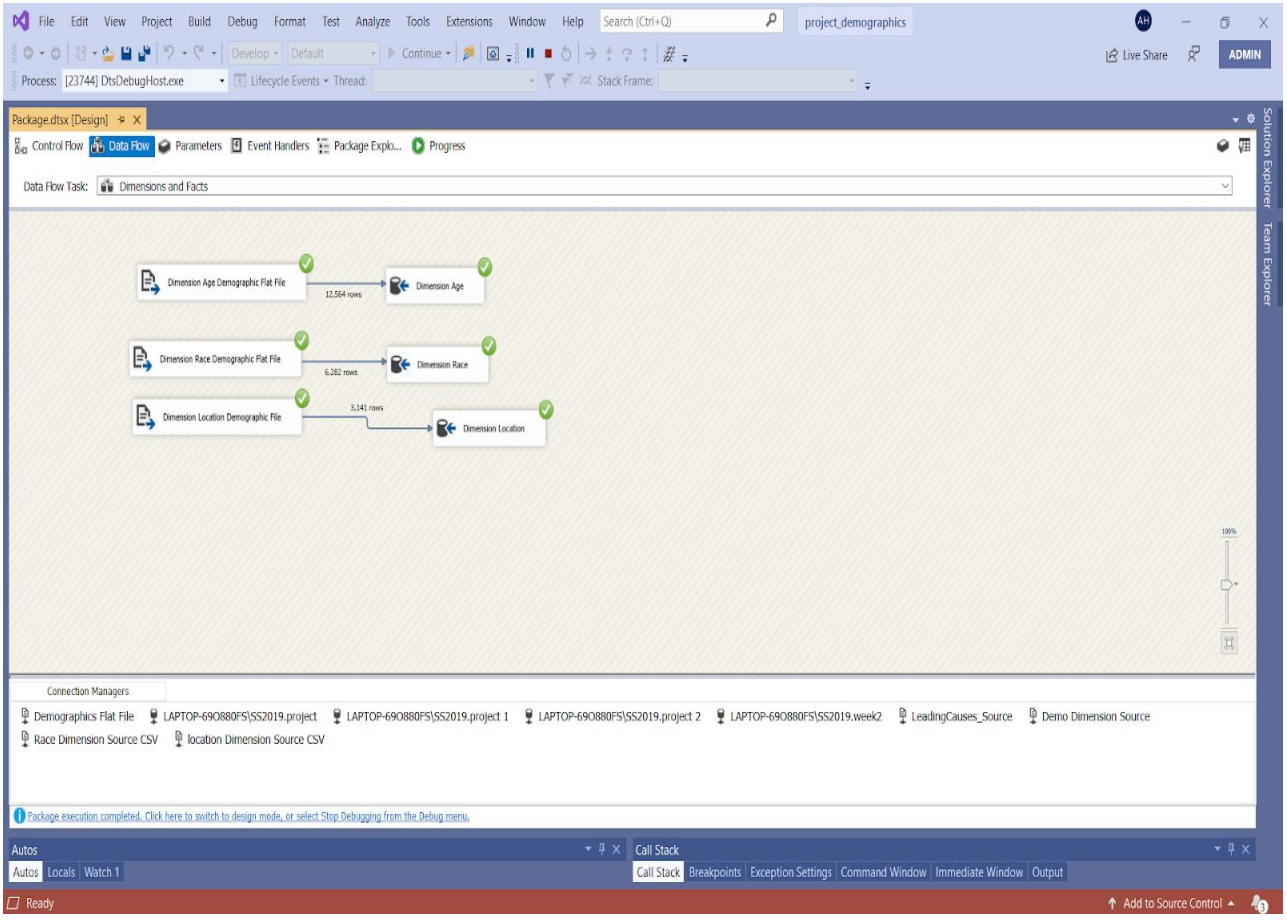


Fig: Fact and Dimension load- Demographics

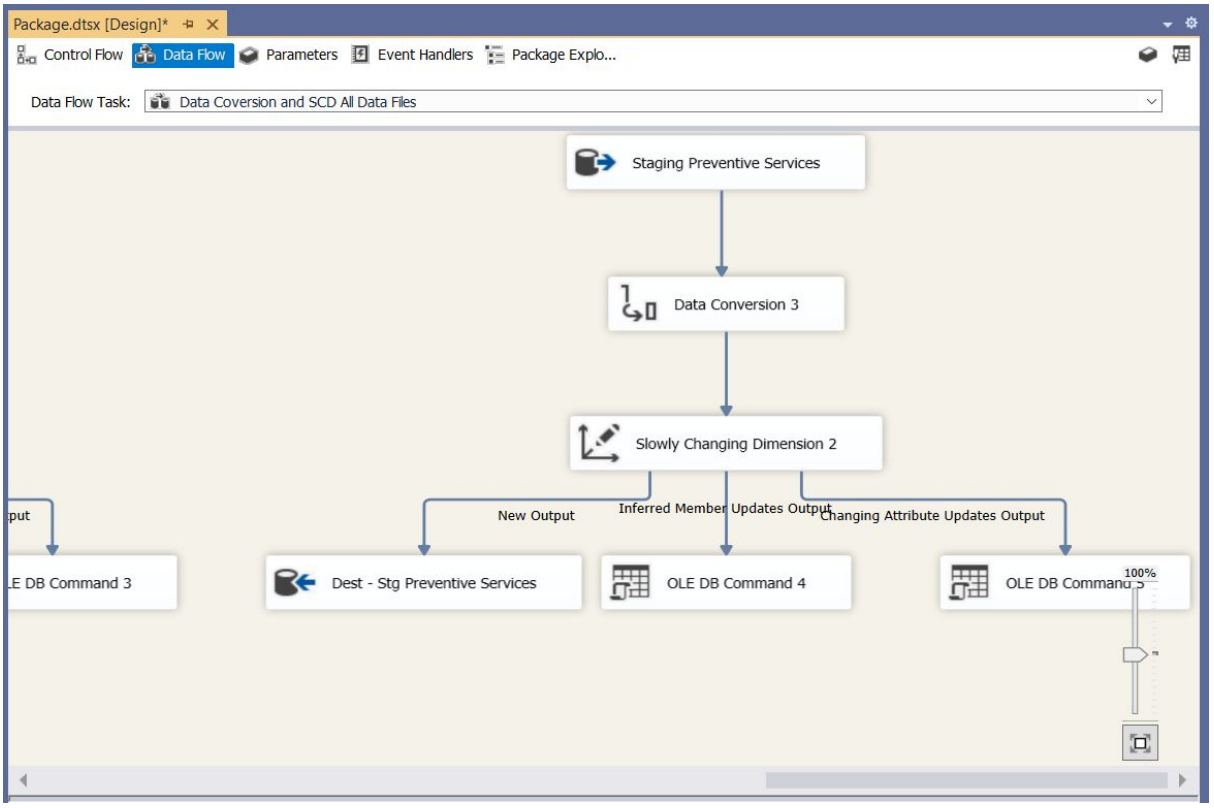


Fig: Slowly changing dimensions implemented for Preventive Services.

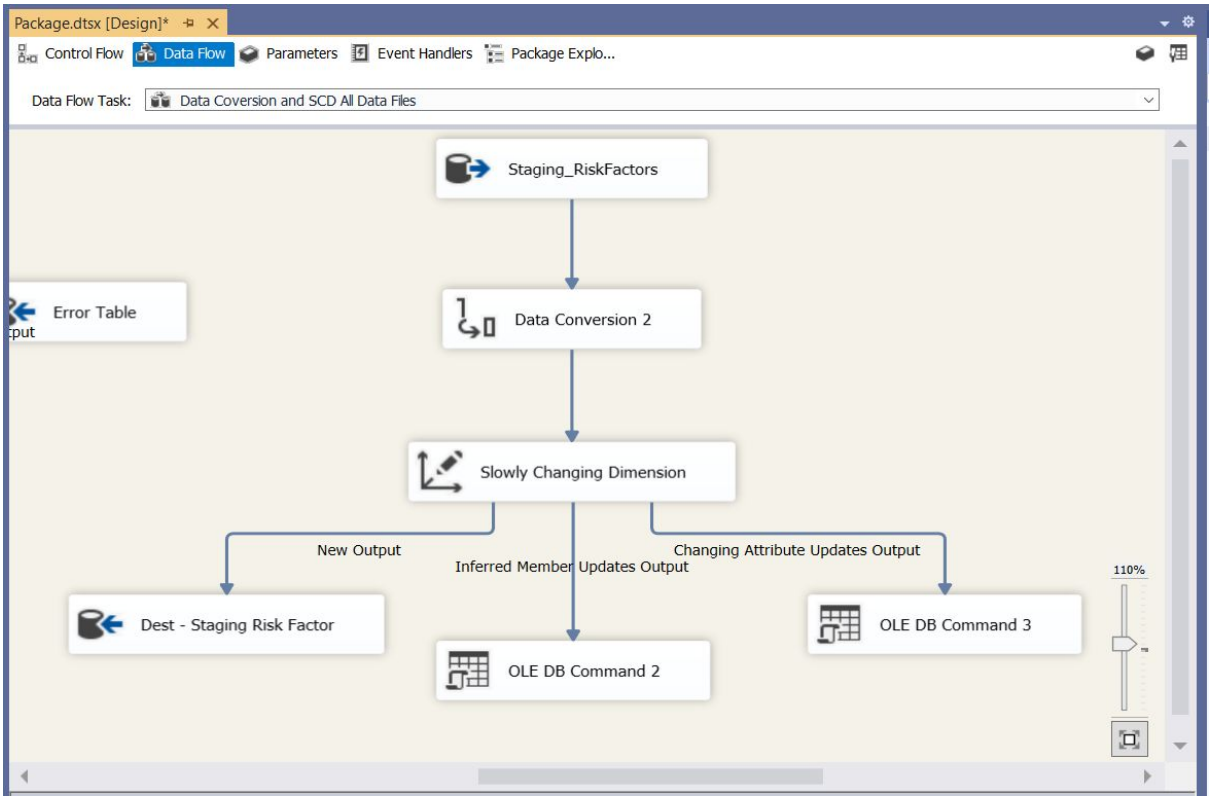


Fig: Slowly changing dimensions implemented for Risk Factors.

Professor's Comments:

Sr No.	Date	Proposal	Comments
1	11/11/2020	Project Proposal 1	Looks good Don't understand the transformations section
2	27/11/2020	Project Proposal 2	I wanted my comments in the end of the document as an appendix I don't expect to see my prompts in your document. You should use your own titles for the sections etc. Typically the cube will be built of the mart Race / Age are better as dimensions What is field / type
3	12/06/2020	Project Design Document	I still see my prompts as your titles You list dimension but sum of them will combine into a single dimensions age for example will not be 3 dimensions State code is a char(2) not a varchar In your final design it

			would be best to have an indicator of race and the counts associated with them as opposed to individual fields with race as part of the name… Data model is not consistent with regards to case and naming I don't think it is fair to say ssis lacks that feature Validation isn't meant to be a manual process I'm not sure of your intentions based on your document You visualizations don't really tell me what is happening you need to make sure you have proper labels etc. I expect more visualizations