December 2021

Data Warehousing & Business Intelligence

CMS Medicare & Medicaid in United States of America

Professor: Dr. Nhan Le Thi

20C14001 – Tuan Anh Le Duong

20C14002 – Dat Vo Tien

ACKNOWLEDGEMENT

We would like to take this opportunity to acknowledge and express our sincere appreciation to the individuals who have helped us finish this project effectively.

Firstly, we would like to thank Professor Nhan Le Thi, Doctor of Scientific knowledge, University of Science, for guiding us through the entire process of report preparation by providing case study and valuable feedback at all stages of development and for teaching us the necessary tools required to efficiently tackle problems and overcome setbacks during the implementation of the project.

Secondly, we would like to thank our peers for providing a conducive atmosphere to understand, learn, contemplate and apply concepts and ideas to our individual projects.

Last but not least, we would like to thank Mr. Nguyen Thanh Toan and Mr. Nguyen Minh Truong, for all the sharing, tips and tricks, and constructive criticism provided through the course of the semester ensuring that we are proficient in the concepts and fundamentals rather than only the implementation.

TABLE OF CONTENTS

ABSTRACT

Analytics has become the technology driver of this decade. US Government and Companies such as IBM, Oracle, Microsoft, and others are creating new organizational units focused on analytics that help businesses become more effective and efficient in their operations. Decision makes are using more computerized tools to support their work. Even consumers are using analytics tools directly or indirectly to make decisions on routine activities such as shopping, healthcare, and entertainment. The field of Business intelligence (BI) is rapidly becoming more focused on innovative applications of data streams that were not even captured some time back, much less analyzed in any significant way. New applications turn up daily in healthcare, sports, entertainment, supply chain management, utilities, and virtually every industry imaginable.

A data warehouse is a system that retrieves and consolidates data periodically from the source systems into a dimensional or normalized data store. It usually keeps years of history and is queried for business intelligence or other ana- lytical activities. It is typically updated in batches, not every time a transaction happens in the source system. A data warehouse consists of many parts, such as the data model, physical databases, ETL, data quality, metadata, cube, application, and so on.

The purpose of this report is to describe the exploration of applying knowledge of building Data Warehouse and a Business Intelligence system to the Healthcare case study in the United States of America. This includes analyzing case study, design the Data Warehouse Architecture, describing the Cubes, and report visualization.

Keywords: *Data mining, data model, physical databases, ETL, data quality, metadata cube, business intelligence, application, DDS, NDS, cubes, dashboards, business intelligence, visualization data.*

1. Introduction

1.1. Introduction:

Healthcare in the United States is expensive! An accident causes not just physical, emotional and psychological damage, but economic too. Approximately 15% of the population is over 65 years of age. To aid people who are 65 and above, the government provides “Medicare” – A health program. Approximately 12% of the population falls below the poverty line. To aid people who fall in the poverty category, the government provides “Medicaid” – A health program with varying coverage based on the state.

Medicare and Medicaid are two separate, government-run programs. They are operated and funded by different parts of the government and primarily serve different groups.

- **Medicare** is a federal program that provides health coverage if you are 65+ or under 65 and have a disability, no matter your income.
- **Medicaid** is a state and federal program that provides health coverage if you have a very low income.
- If you are eligible for both Medicare and Medicaid (dually eligible), you can have both. The US Government will work together to provide you with health coverage and lower your costs.

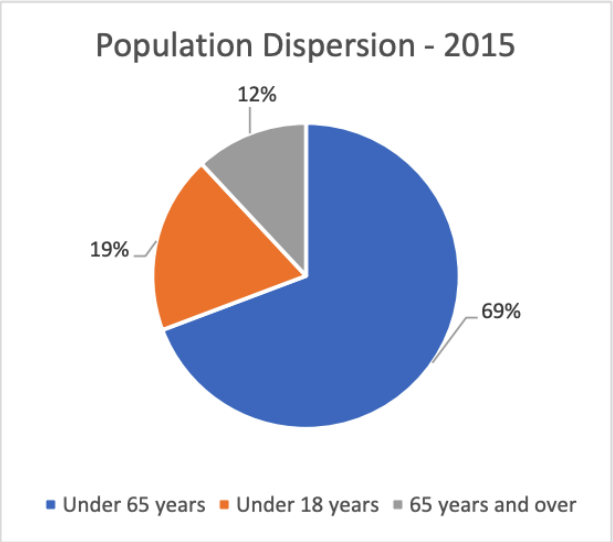


Figure 1: *Population dispersion in the year 2015 by age group*

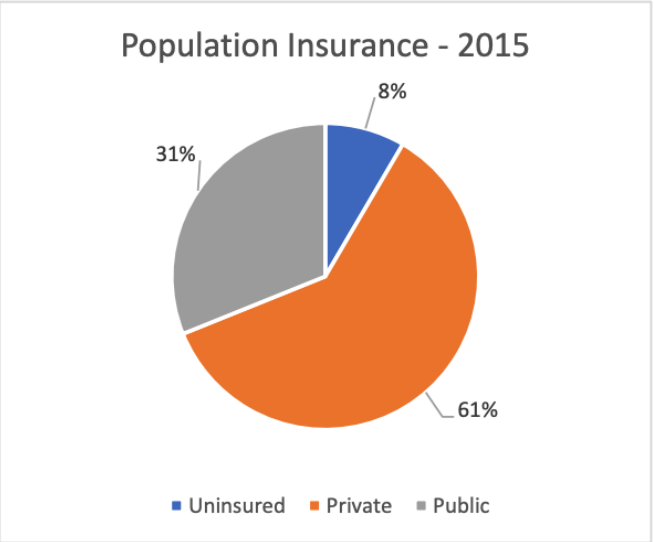


Figure 2: *Type of Population Insurance in the year 2015 by age group*

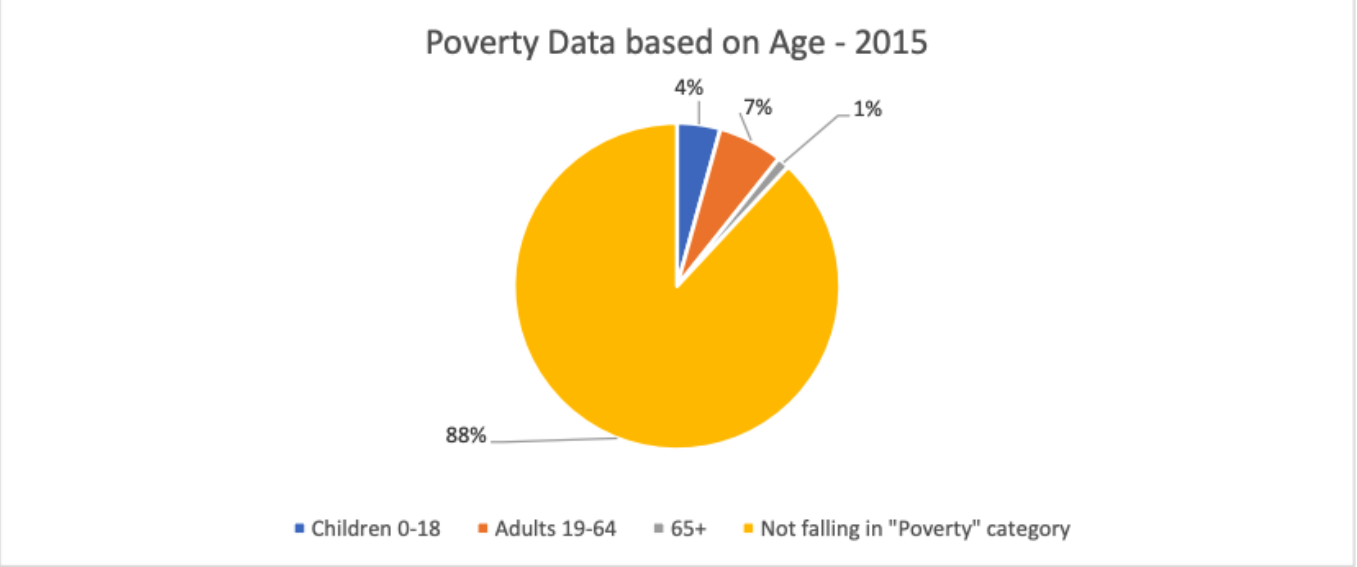


Figure 3: *Total Poverty in the year 2015 by age group*

1.2. The Problems:

The healthcare industry is seeing an increase in Data it has available and the expectation from customers and shareholders that this data be used to improve patient care, and return a higher value to shareholders.

The increase in data comes from several sources. In fact, it is estimated that nearly 80% of all data is unstructured, and medical records can be some of the hardest to work with. Due to its own set of shorthand and abbreviations, extra care must be taken in organizing this information.

Medicare Advantage plans with many participants across the United States, this is a health insurance plan that is offered by private providers to add benefits, or provide an advantage, over the standard Medicare plan. As we known that:

- Healthcare in the United States is expensive! An accident causes not just physical, emotional and psychological damage, but economic too.
- Nearly 80% of all data is unstructured, and medical records can be some of the hardest to work with.
- Approximately 15% of the population is over 65 years of age.
- Approximately 12% of the population falls below the poverty line.

We have made some hypotheses like below:

- States where the number of people whose age is >65 is high → Medicare spending is also high.
- States which have high levels of poverty → Medicaid spending is high.
- Government health spending is uniform between 2 genders Male and Female.
- States which have high income should have lower Medicaid spending → People are able to afford Private Insurance.

1.3. Our Goals:

Our goal is to test out the efficiency of dispersion of funds for these two government health programs and test hypotheses with visual indicators which throws light on areas of improvement. We want to visually understand the US Government spending on Medicare and Medicaid across the 50 States and provide recommendations for resource allocation.

What will we deliver?

- The report gives a high level view of Data, method of analysis and outcome.
- Visualization dashboard providing a holistic view of hypothesis testing in a visual manner.

2. Flow Chart

This section gives us a general flow of the entire project:

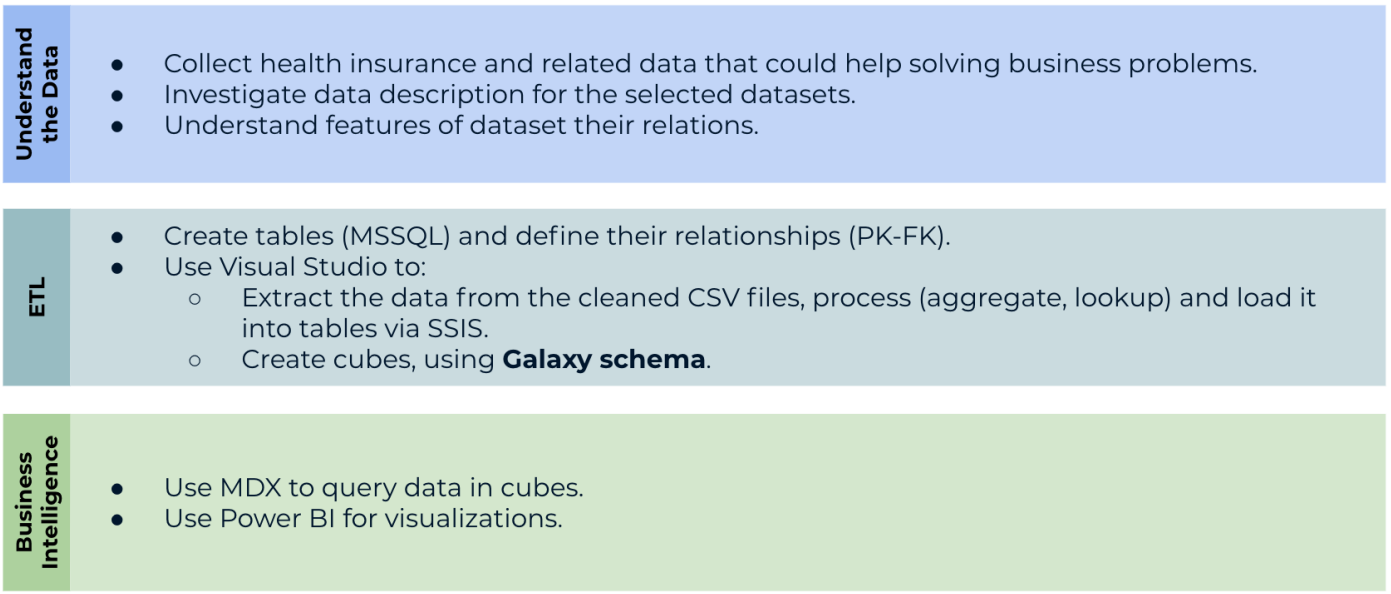


Figure 3: *The general flow chart of the Project*

3. Data Selection

3.1. Health Insurance Data

Figure 4: *Health Insurance Dispersion Data*

This entity contains data on the type of health care coverage that the people in the United States have categorized based on Age, State, Coverage and Year. It has information about the number of people who are Uninsured and Insured through various sources like Medicare, Medicaid, Public Insurance and Private Insurance. The data is identified by the User ID which has been encrypted.

3.2. Income Data

Figure 5: *Income Data*

This table describes the income for the year 2015 and 2016. It has the fields like Personal Income, Per Capita Income, Regional Price Parity, State and Year.

3.3. Poverty Data

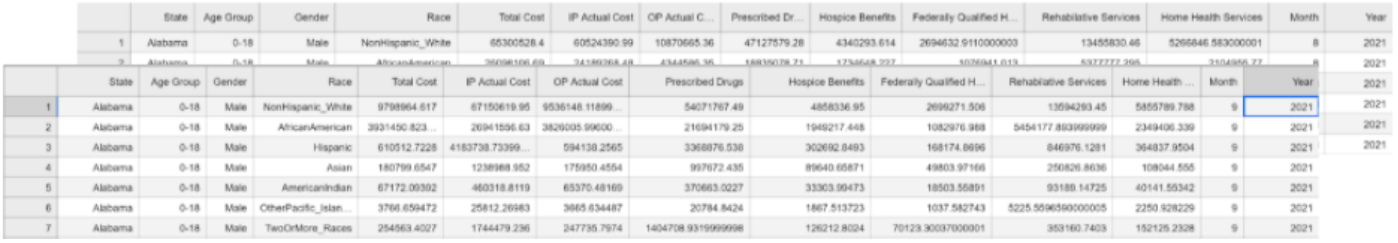
	State	Age Group	Gender	Race	Poverty Rate	Month	Year
1	Alabama	0-18	Male	NonHispanic_White	0.0018	9	2021
2	Alabama	0-18	Male	AfricanAmerican	0.00468	9	2021
3	Alabama	0-18	Male	Hispanic	0.00504	9	2021
4	Alabama	0-18	Male	Asian	0.0018	9	2021
5	Alabama	0-18	Male	AmericanIndian	0.00216	9	2021
6	Alabama	0-18	Male	OtherPacific_Islander	0.00396	9	2021
7	Alabama	0-18	Male	TwoOrMore_Races	0.0027	9	2021
8	Alabama	0-18	Female	NonHispanic_White	0.0024	9	2021

Figure 6: *Poverty Data*

This table describes the poverty levels in each state of the U.S for the year 2015 and 2016. It has the fields like State, Age Group, Race, Poverty Rate and Year.

Note: The poverty rate is the ratio of the number of people (in a given age group) whose income falls below the poverty line; taken as half the median household income of the total population. It is also available by broad age group: child poverty (0-17 years old), working-age poverty and elderly poverty (66 year-olds or more).

3.4. Medicare / Medicaid Data



	State	Age Group	Gender	Race	Total Cost	IP Actual Cost	OP Actual Cost	Prescribed Drugs	Hospice Benefits	Federally Qualified H...	Rehabilitative Services	Home Health ...	Month	Year
1	Alabama	0-18	Male	NonHispanic_White	9798964.617	67150619.95	9536148.11899...	54071767.49	4858336.95	2699271.506	13594293.45	5855789.788	9	2021
2	Alabama	0-18	Male	AfricanAmerican	3931450.823...	20941556.63	3826005.99600...	21694179.25	1949217.448	1082976.988	5454177.893999999	2349406.339	9	2021
3	Alabama	0-18	Male	Hispanic	610512.7228	4183738.73399...	594138.2565	3368876.538	302692.8493	168174.8696	846976.1281	364837.9504	9	2021
4	Alabama	0-18	Male	Asian	180799.6547	1238988.952	175950.4554	997672.435	89640.65871	49803.97166	250826.8636	108044.555	9	2021
5	Alabama	0-18	Male	AmericanIndian	67172.09302	460318.8119	65370.48169	370663.0227	33303.99473	18503.55891	93189.14725	40141.55342	9	2021
6	Alabama	0-18	Male	OtherPacific_Islan...	3766.659472	25812.26983	3665.634487	20784.8424	1867.513723	1037.582743	5225.559659000005	2250.928229	9	2021
7	Alabama	0-18	Male	TwoOrMore_Races	254563.4027	1744479.236	247735.7974	1404708.9319999998	126212.8024	70123.90037000001	353160.7403	152125.2328	9	2021

Figure 7: *Medicare / Medicaid Data*

The Medicare Data table contains the beneficiaries of the Medicare program. It has information on demographics, spending, and service utilization for Medicare beneficiaries in different parts of the country. It has some fields like Total cost, Prescribed Drugs, Hospice Benefits, Federally Qualified Health Center, Rehabilitative Services, Home Health Services.

The Medicaid Data table also has the fields just like Medicare but it describes the state-by-state total expenditures by program for the Medicaid Program, Medicaid Administration and CHIP programs.

3.5. External Data

	State	Abbreviation
1	Alabama	AL
2	Alaska	AK
3	Arizona	AZ
4	Arkansas	AR
5	California	CA
6	Colorado	CO
7	Connecticut	CT

ID	Race
1	NonHispanic_White
2	AfricanAmerican
3	Hispanic
4	Asian
5	AmericanIndian
6	OtherPacific_Islander
7	TwoOrMore_Races

ID	Coverage Type
1	Employment Based
2	Direct Purchase
3	Covered by TRICARE
4	Medicaid
5	Medicare
6	Uninsured

ID	Age Group
1	0-18
2	19-64
3	65+

ID	Gender
1	Male
2	Female

id	month	year
1	8	2021
2	9	2021

Figure 8: *External Data*

These external data tables are collected from multiple sources to support the main data more meaningfully.

4. Data Warehouse

4.1. Overview:

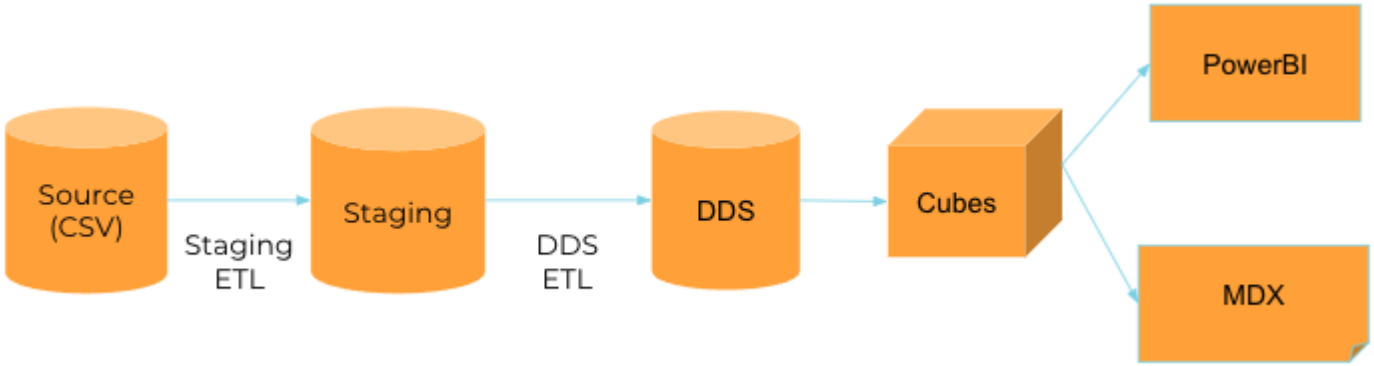


Figure 9: The overview of the data flow in Data Warehouse.

The process starts with the importation of Sources from multiple CSV files which are mentioned above. Then we deploy a copy of the source to Staging by following the ETL process and we are using DDS architecture. After the ETL process completes, the Cubes are generated, and it's ready for Visualization on PowerBI and ready for querying in MDX.

4.2. Warehouse Relationship:

4.2.1. Warehouse Relationship overview:

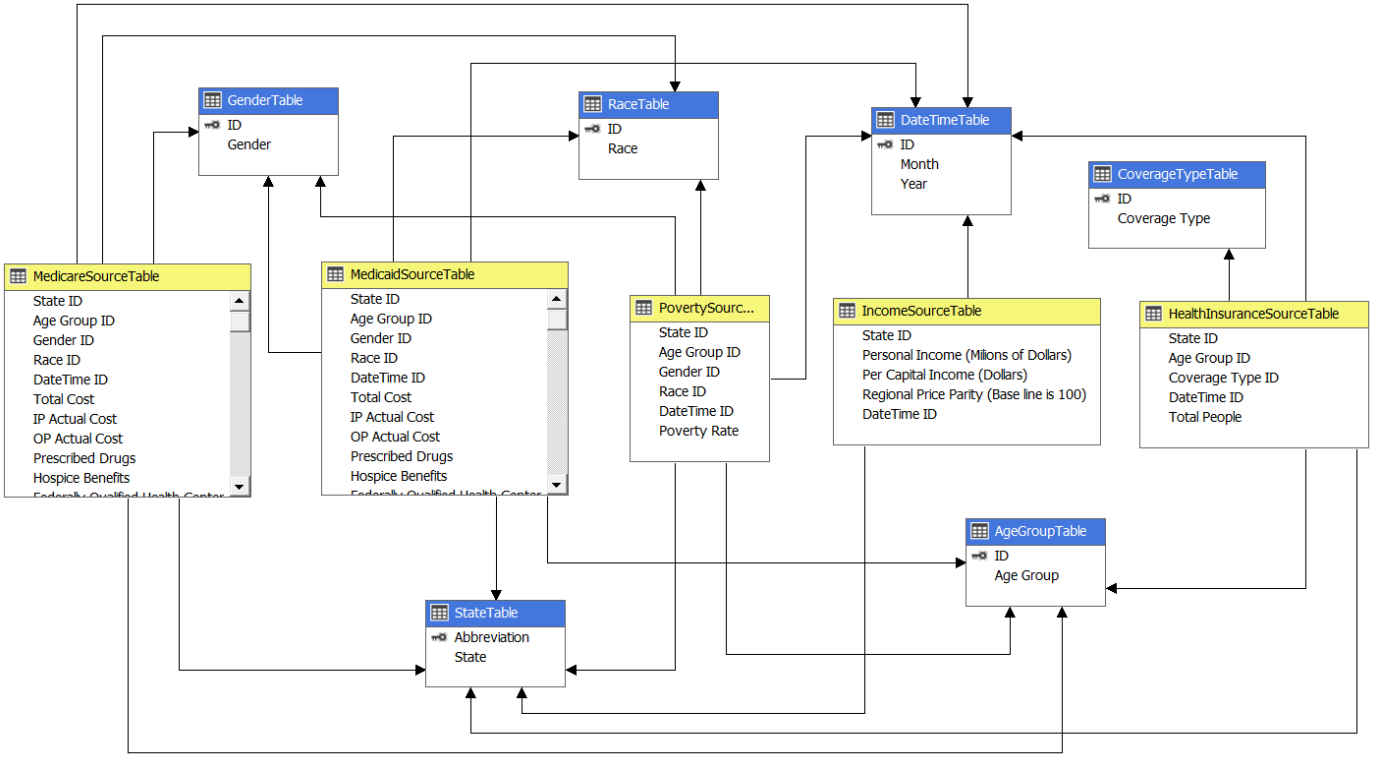


Figure 10: *The overview of the Data Warehouse relationship*

4.2.2. Star Schema:

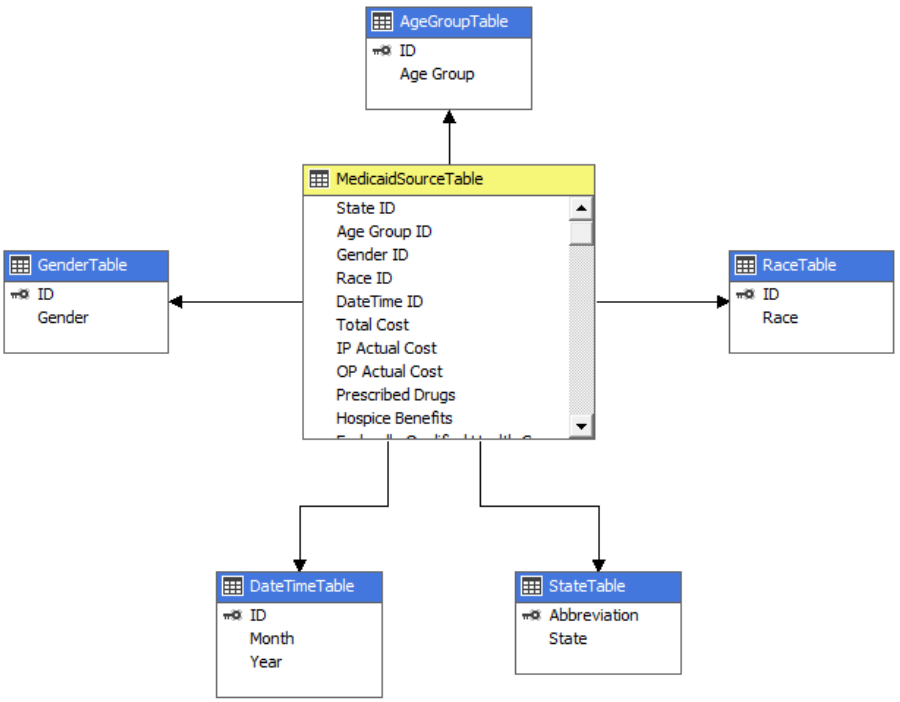


Figure 11: *Medicaid star schema*

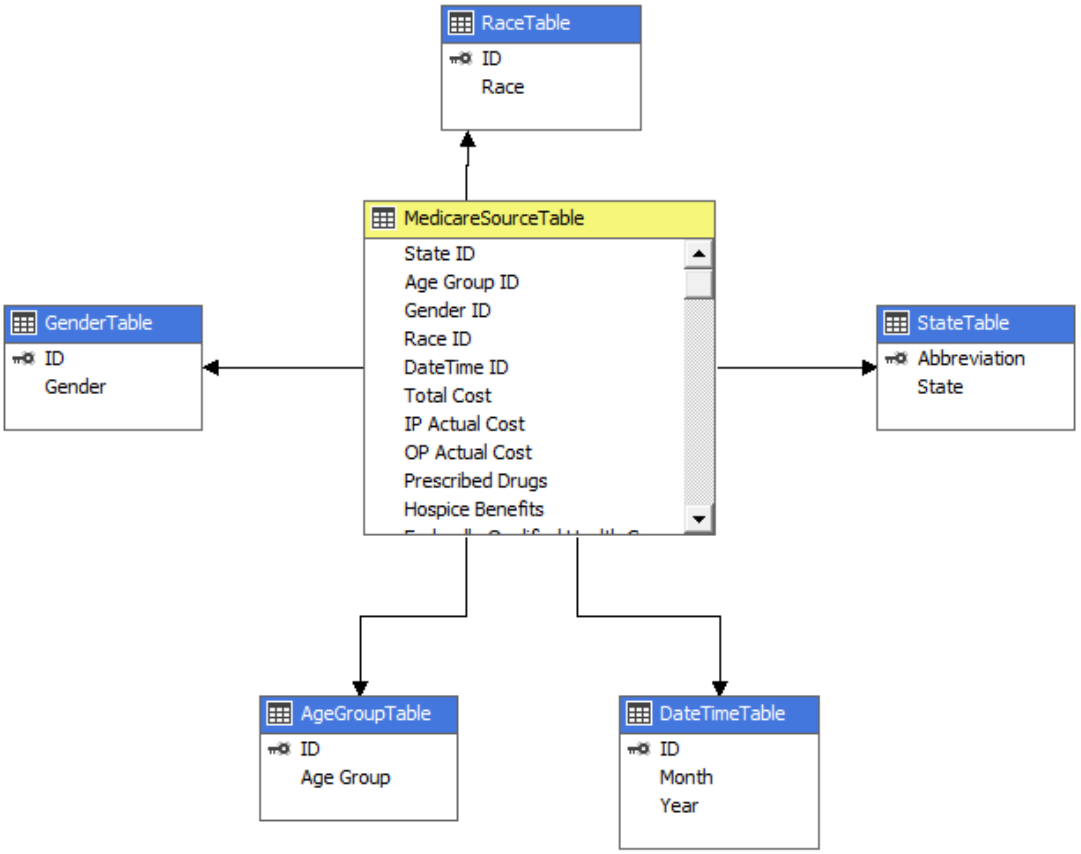


Figure 12: *Medicare Source star schema*

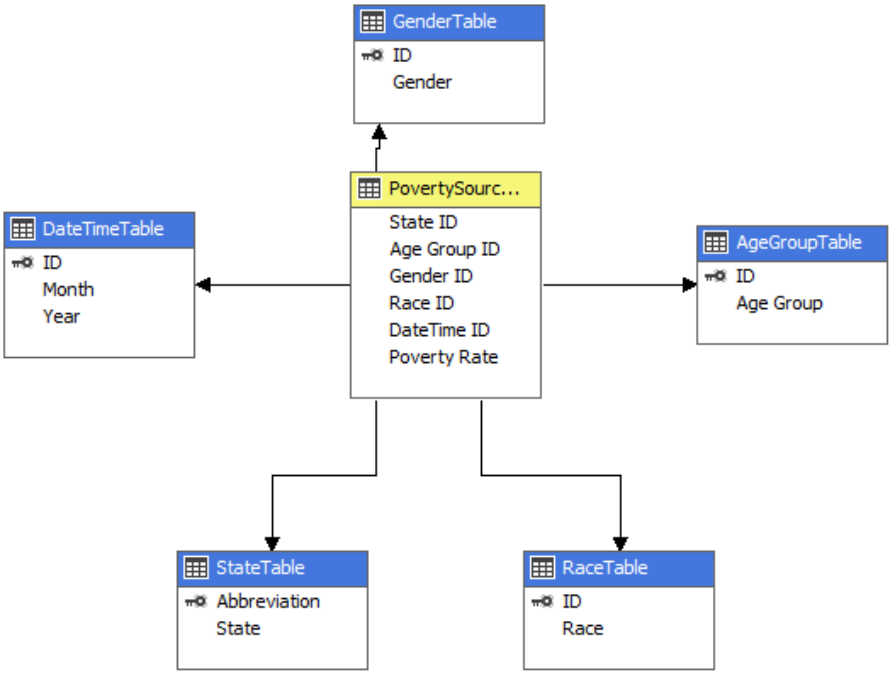


Figure 13: *Poverty source star schema*

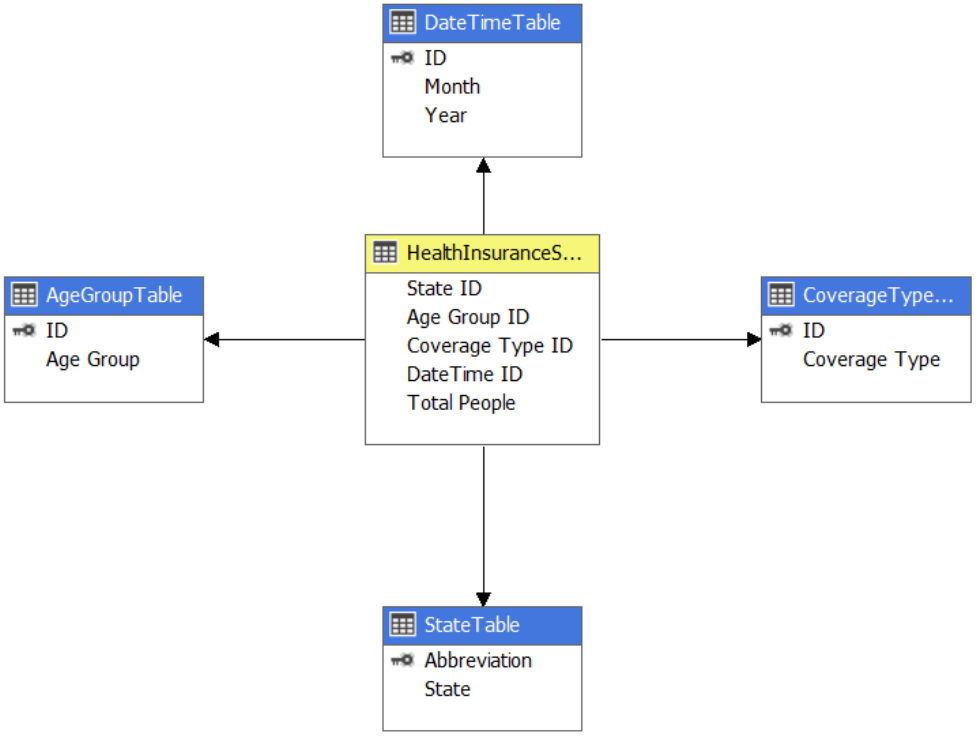


Figure 14: *Health Insurance star schema*

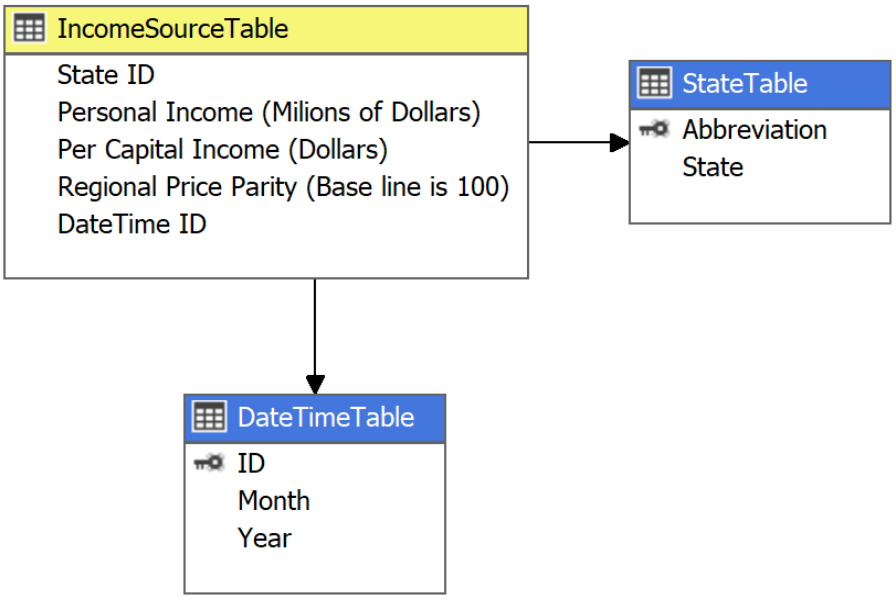


Figure 15: *Income Source table star schema*

5. ETL

We use the *Integration Service Project* (VS2019 Community Version). This project may be used for building high performance data integration and workflow solutions, including extraction, transformation, and loading (ETL) operations for data warehousing.

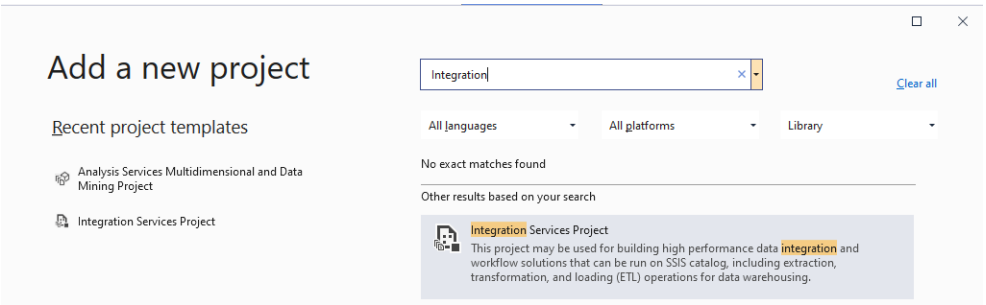


Figure 16: *Integration Service Project (VS2019 Community Version)*

5.1. Main Flow:

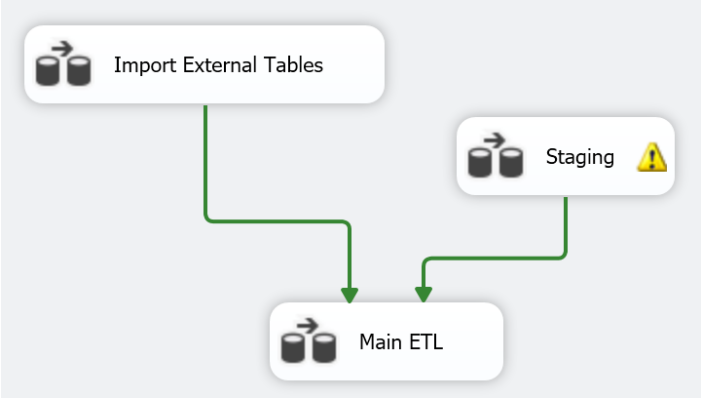


Figure 17: *The main flow of ETL process*

First of all, we loaded all data into staging and then to the destination on Main ETL. We inserted the information of the files which are being loaded in the respective tables.

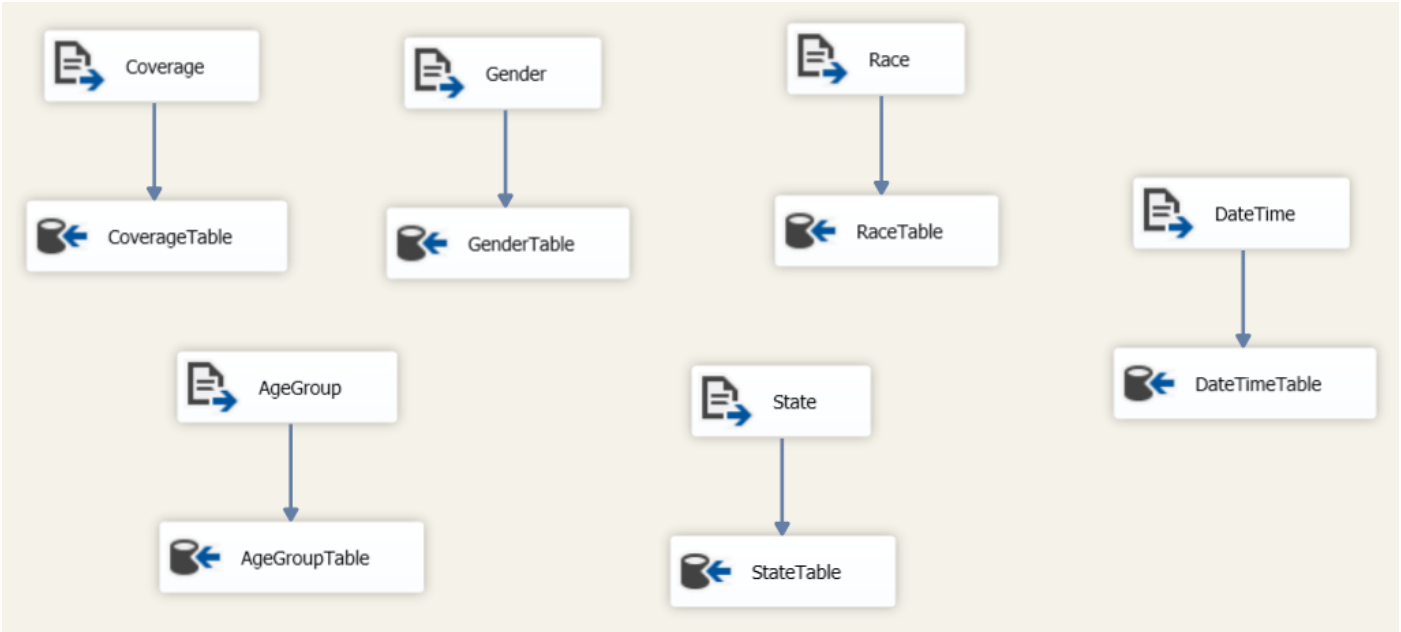


Figure 18: *Import external data flow*

5.2. Staging

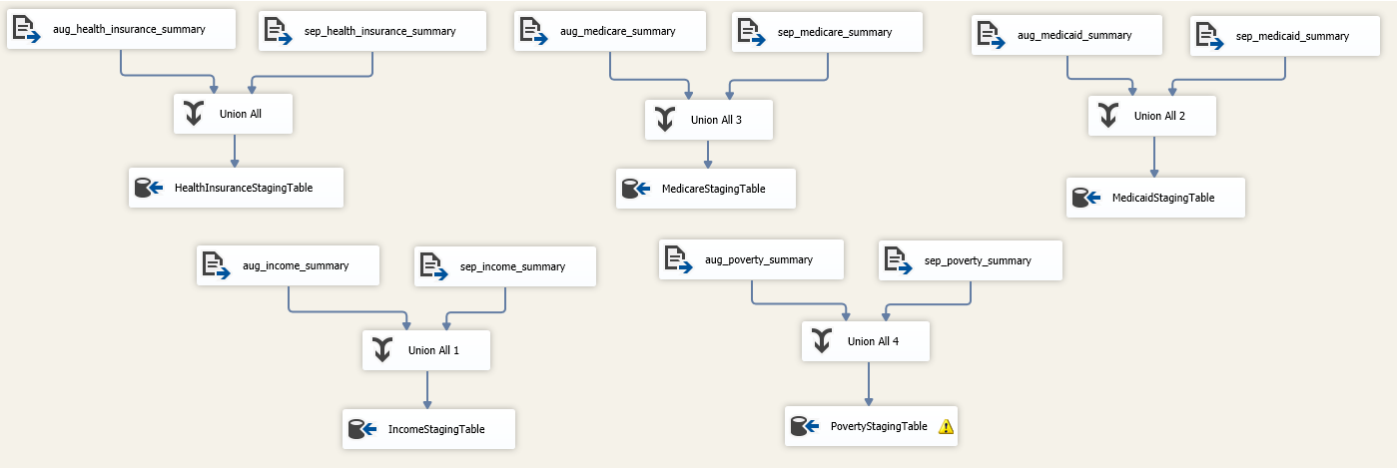


Figure 19: *Staging area*

A staging area is an intermediate storage area used for data processing during the extract, transform and load (ETL) process. The data staging area sits between the data source(s) and the data target(s), which are often data warehouses, data marts, or other data repositories.

5.3. DDS

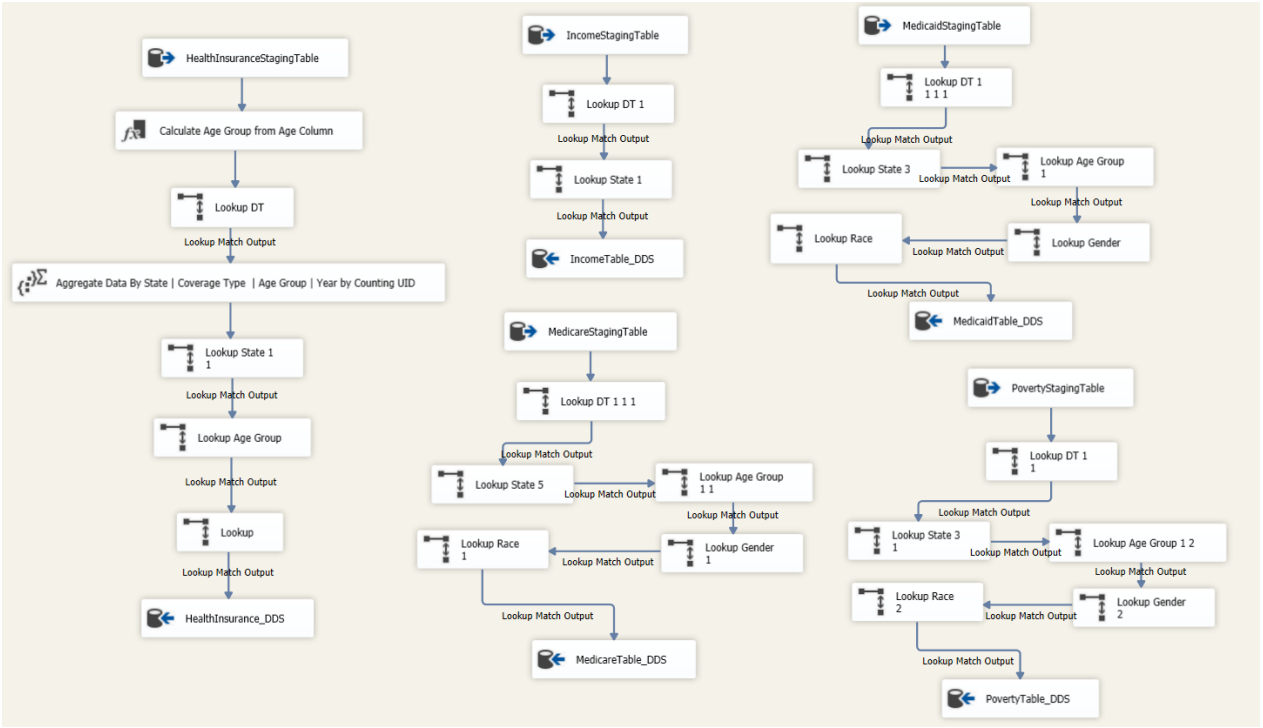


Figure 20: *Dimension Data Store*

A dimensional data store (DDS) is a user-facing data store, in the form of one or more relational databases, where the data is arranged in dimensional format for the purpose of supporting analytical queries.

6. SSAS

After the ETL, we moved on to SSAS where we have to generate the cubes. This is the flow that the cubes are generated:

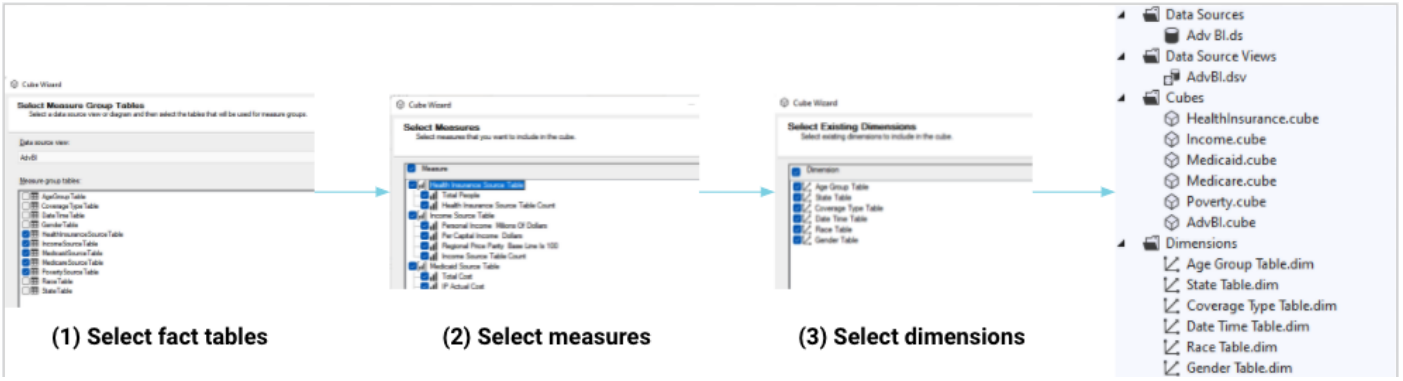


Figure 21: *The flow to generate the cubes*

Then to create a new Hierarchy on Dimension Structure, we need to add the Insight: DateTime → [Year –> Month] → Edit hierarchies

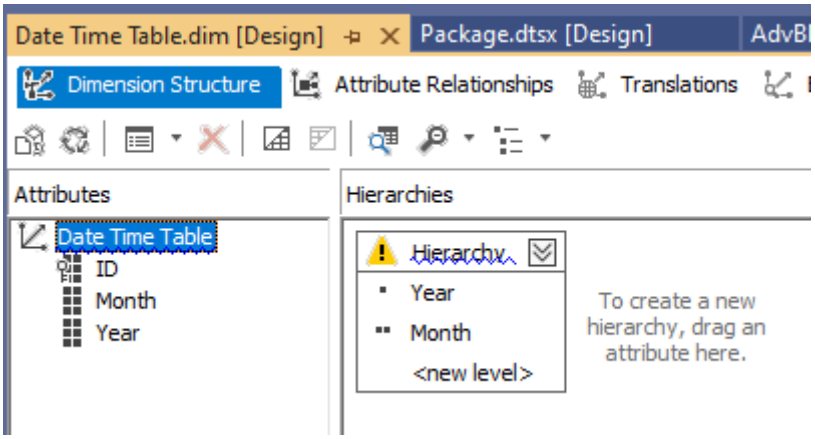


Figure 22: *Hierarchy on dimension structure*

Finally, we deploy Cubes to Microsoft Analysis Server:

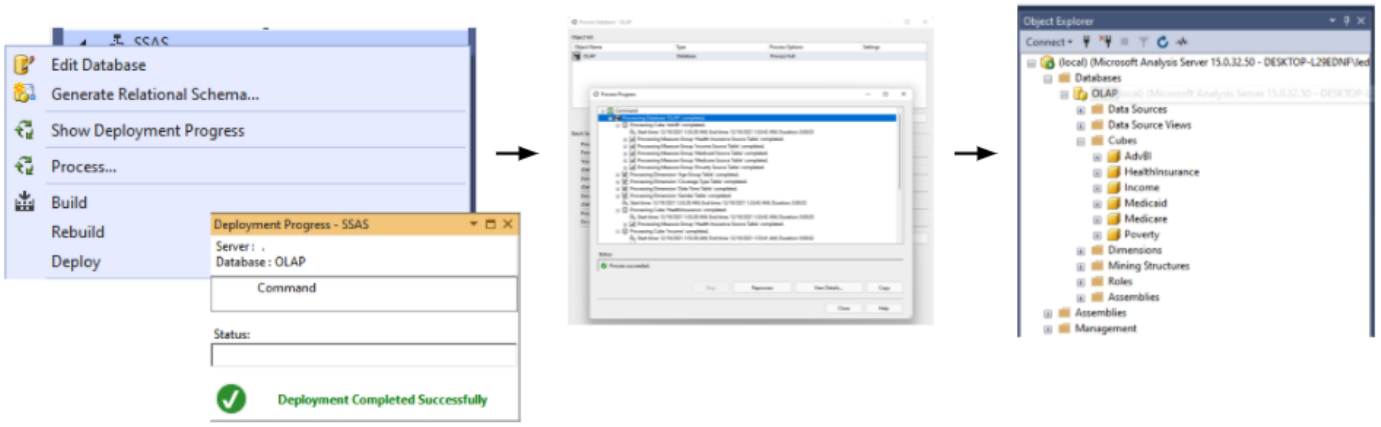


Figure 23: *Deploy the Cubes to Microsoft Analysis Server*

7. Visualization

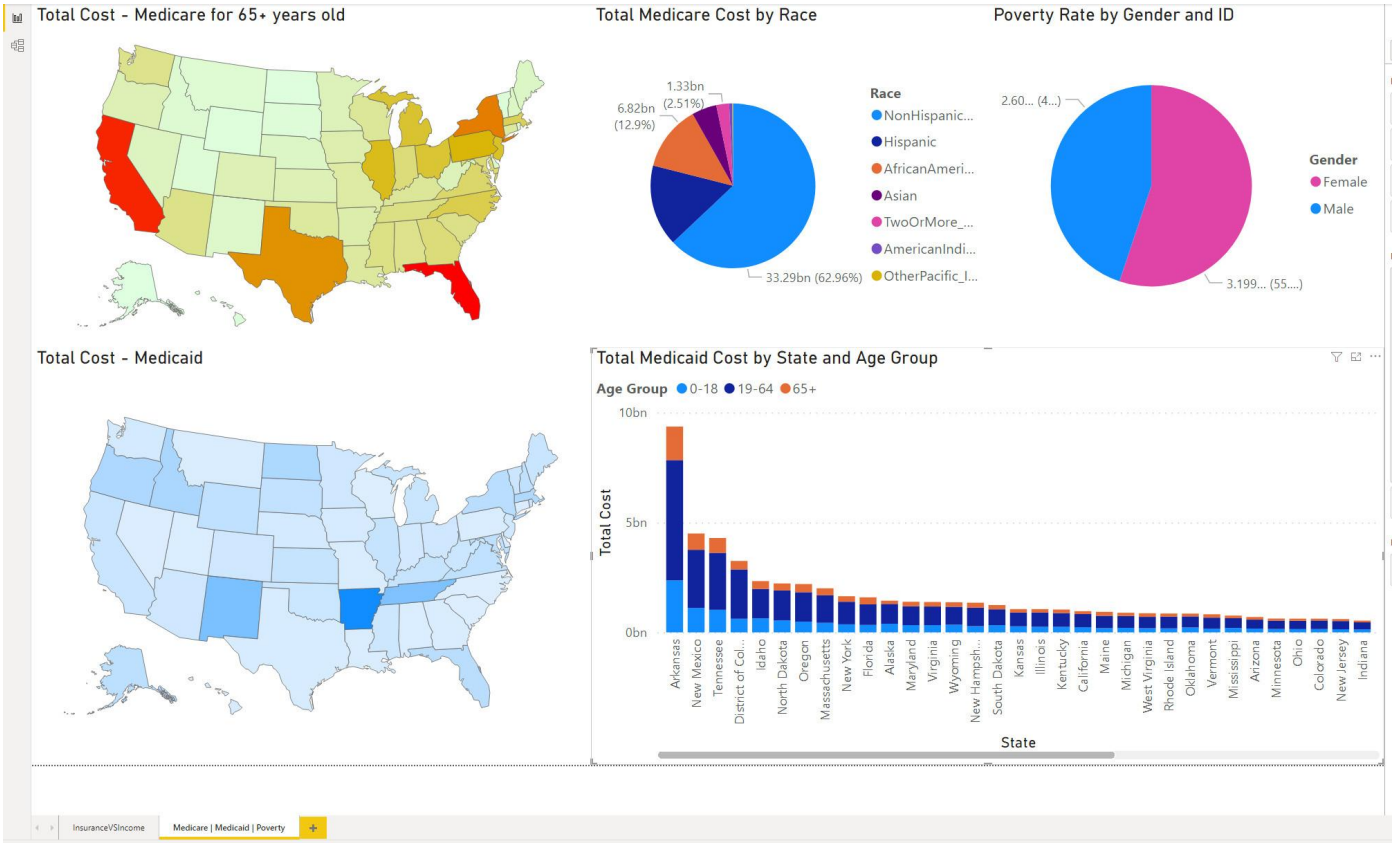


Figure 24: *Total cost spending for each state by Race and Gender Dashboard.*

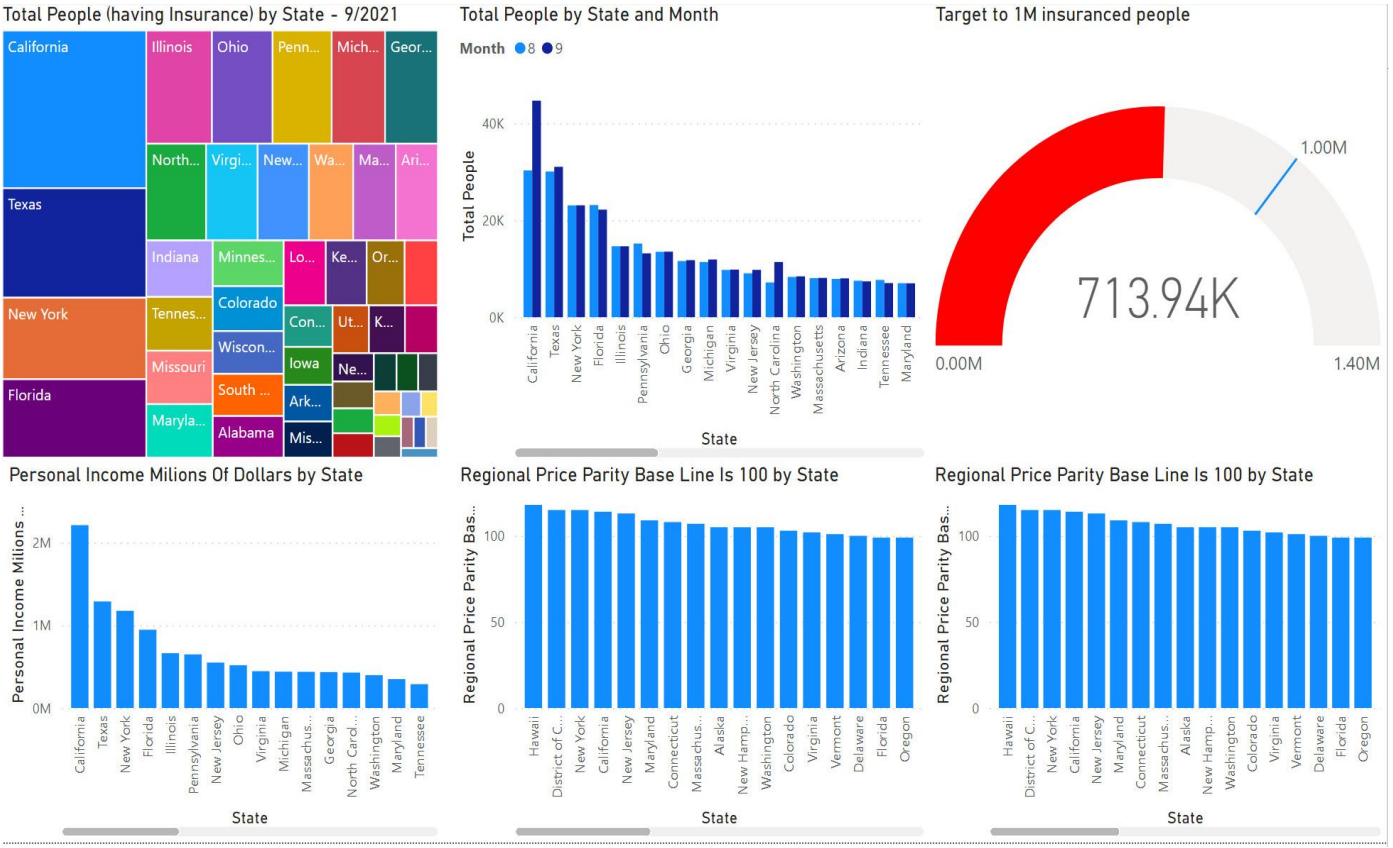


Figure 25: *The number of people having insurance by States and the reflection of income / regional price to the demand of having insurance.*

CONCLUSION

In conclusion, the Medicare and Medicaid spending is related to living standards and individual income across states.

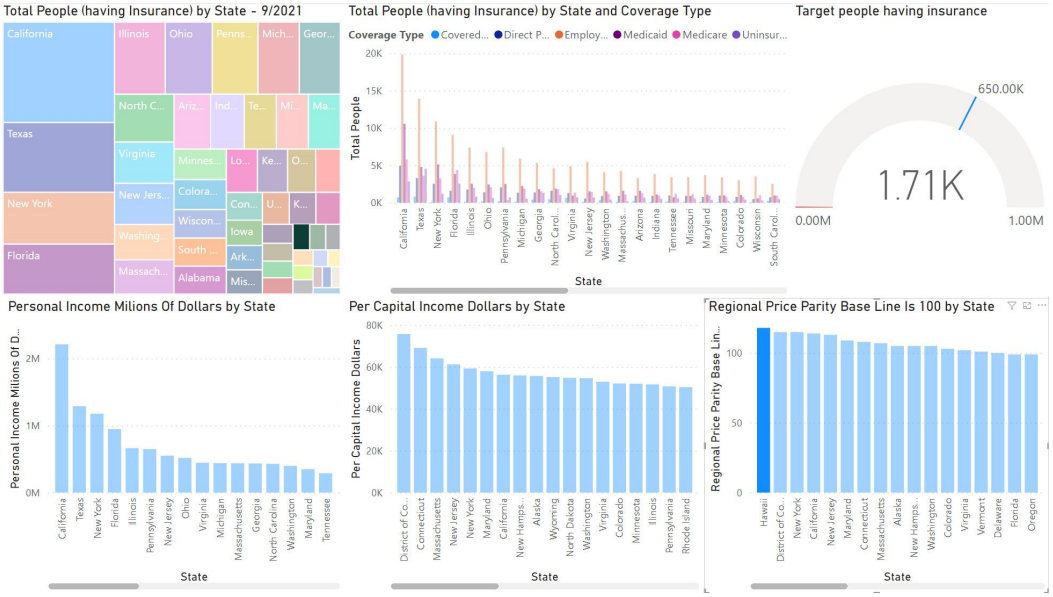


Figure 26: *Total people having insurance in Hawaii*

For example in Hawaii which has a high Regional Price Parity but the number of people having insurance is low (only 1.71K people having insurance). Otherwise, their people's incomes only come from Travel services.

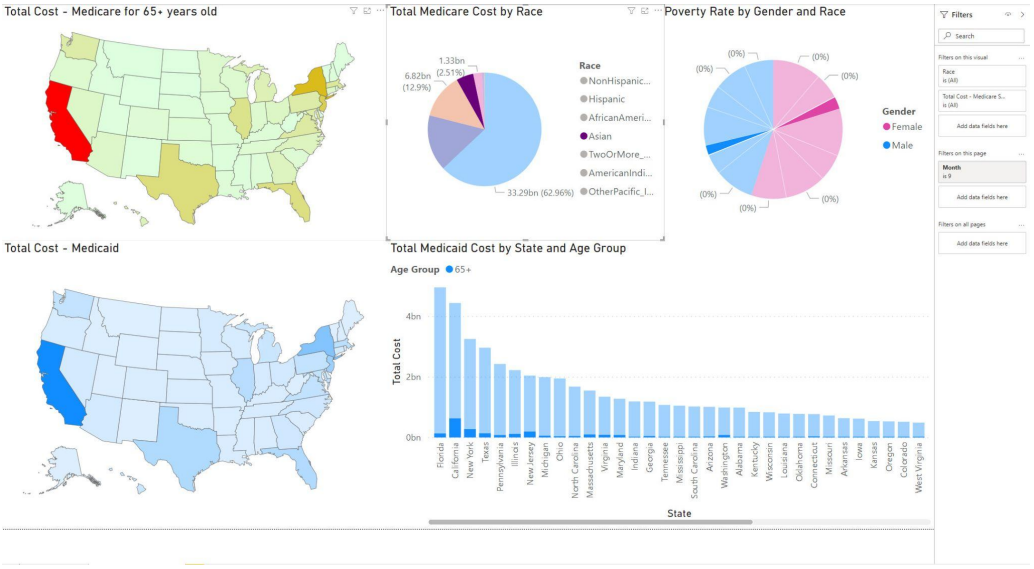


Figure 27: *Total health expenditure by Race*

Total health expenditure shows the rationality of distribution by race and state. Asia race is mainly concentrated in states such as California, Florida, Texas, ... So we can see, the total cost of health is reasonably displayed on the Dashboard.

The government should move the fund of spending in Medicaid to states with lower incomes and age groups 19-64 to improve the health of the main labor force.

REFERENCES

[1] Business Intelligence: A Managerial Perspective on Analytics, 3rd ed. Edition, Ramesh Sharda, Dursun Delen, Efraim Turban.
[2] Building a Data Warehouse: With Examples in SQL Server (Expert's Voice) 1st ed. Edition by Vincent Rainardi.
[3] Data Mining: Techniques, Applications and Issue - Rupali, Gaurav Gupta, International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) Volume 2, Issue 2, February 2013.
[4] CMS Research, Statistics, Data & Systems
[5] Poverty in the United States in 2015: In Brief
[6] ETL Process in Data Warehouse - GeeksforGeeks
[7] Modeling ETL Process for Data Warehouse: An Exploratory Study | IEEE Conference Publication | IEEE Xplore
[8] Microsoft SQL Server OLAP Solution
[9] The Data Warehouse Lab: A step-by-step guide using SSIS and SSAS 2017 | Semantic Scholar