Nobel Prize Data

Warehouse Project

Marie Hasegawa

mhasegawa7045@floridapoly.edu

Rosely Machado

rmachado9966@floridapoly.edu

Jasmine Brown

jbrown7763@floridapoly.edu

Professor Ray Ready

rready@floridapoly.edu

Data Warehouse CAP3774 - Spring 2022

1 Introduction

1.1 The Nobel Foundation Description

The Noble Foundation was founded by Alfred Nobel, who left a majority of his fortune to the institution, so Nobel Prizes are awarded to laureates " who, during the preceding year, shall have conferred the greatest benefit to humankind" (The Noble Foundation). The Nobel Prizes are yearly distributed to laureates across six categories: physics, chemistry, medicine, literature, peace, and economics (The Nobel Foundation). Each Nobel Prize comes with a USD 1 million cash reward shared amongst the laureates either being individuals or an organization involved with the research project(The Nobel Foundation).

1.2 The Nobel Prize Data Warehouse Description

This project will create a complex Nobel Prize Data Warehouse that consists of 8 dimensions and 3 facts. These dimensions and facts come from 8 CSV files (*Category, Country, Institution, Institution Type, Laureate, Motivation, Overall Motivation, and Prize*) that were manipulated and created based on the 3 original files (*Prize, Country, and Laureate*) from the Harvard Dataverse dataset, *Nobel Prize - Dataset with Information about Prizes, Laureates and Countries*, authored by Kuzmenko, Maryna.

The original dataset was made with the intention of being a "starting point of research in social sciences about quantitative and other characteristics of Nobel prize" (Kuzmenko, Maryna). The dataset was further expanded and manipulated by our team with the intent to create a more complex Nobel Prize Data Warehouse that will help further studies of Nobel Prizes for future researchers compared to previous datasets that had simpler solutions in making the Nobel Prize Data Warehouse. The data warehouse was designed through a BEAM template, Microsoft SQL

Server Management, Visual Studio, and Tableau, which will be further explained in the next section.

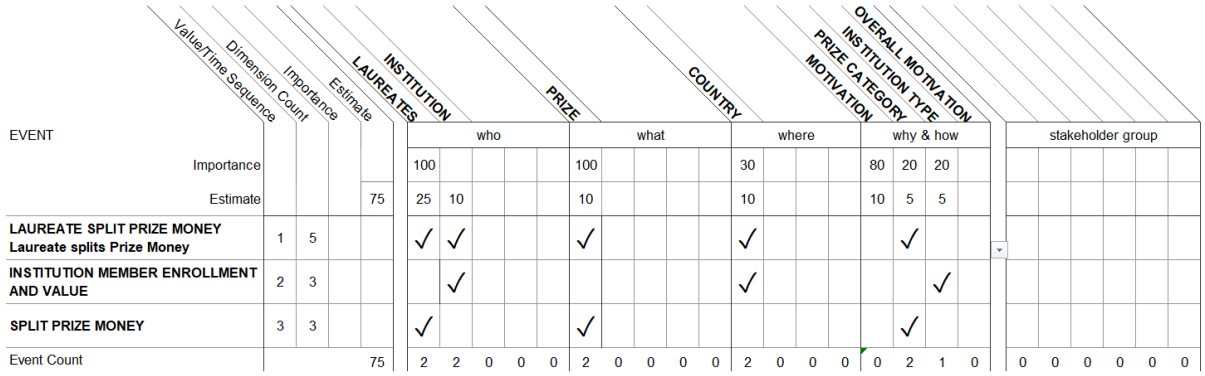
The data warehouse's 8 dimensions include: *Country* (where the specific Nobel Prize is won by the laureate's country for that year), *Nobel Prize* (the specific Nobel Prize won that year), *Nobel Prize Category* (the 6 categories of the Nobel Prizes), *Institution* (the institution where the laureates' research is supported and done at), *Institution Type* (includes the 6 types of institutions: Educational, Religious, Medicinal, Government, Economic, and Research), *Laureate* (the Nobel Prize winner), *Motivation* (the motivation of the research project), and *Overall Motivation* (the overall motivation of the research project).

The data warehouse's 3 facts include: *Distributed Prize Money*, *Institution Value Members*, and *Laureate Shared Prize Value*. The *Distributed Prize Money* Fact contains numerical and descriptive data on how much researchers share the Nobel Prize amongst the other researchers and how much is the prize money distributed per researcher. The *Institution Value Members* Fact provides numerical and descriptive data on the monetary value of the institution and the number of members that are present in the institution. The *Laureate Shared Prize Value* Fact contains numerical and descriptive data of the laureates' monetary gain from the Nobel Prize and how many people including themselves have to share the Nobel Prize for the same research.

As mentioned before, this project was implemented through Microsoft SQL Server Management Studio, Microsoft Visual Studio, and Tableau. The SQL Server Management Studio built the Nobel Prize Data Warehouse and the views, while Visual Studio was used to extract data from the 8 CSV files and create the ETL process. Afterward, the data visualizations based on the processed data are made through Tableau and the SQL queries are made in SQL Server Management Studio. This project will be further explained in the upcoming sections.

2 BEAM Data Warehouse Planning

Before anything got started, proper planning using the BEAM template had to get done. Various events were first identified like laureates receiving prize money, institution enrollment and value, and distributing prize money. From these events, various straightforward dimensions were created: laureate dimension, institution dimension, prize dimension, and country dimension. There arose a need to classify and understand the “why” behind some of these events so additional dimensions were created to satisfy this: Motivation, Overall Motivation, Prize Category, and Institution Type.



EVENT	Value/Time Sequence	Dimension Count	Importance	Estimate	LAUREATES	INSTITUTION	PRIZE	COUNTRY	MOTIVATION	PRIZE CATEGORY	INSTITUTION TYPE	OVERALL MOTIVATION	stakeholder group
					who			what		where		why & how	
Importance					100			100		30		80	20	20	
Estimate				75	25	10		10		10		10	5	5	
LAUREATE SPLIT PRIZE MONEY Laureate splits Prize Money	1	5			✓	✓		✓		✓			✓		
INSTITUTION MEMBER ENROLLMENT AND VALUE	2	3				✓				✓				✓	
SPLIT PRIZE MONEY	3	3			✓			✓					✓		
Event Count				75	2	2	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0

2.1 Nobel Prize BEAM Matrix

The Laureate dimension describes the laureates, their prize, and their affiliations with any countries and/or institutions. Since organizations could win Nobel Prizes, it was considered acceptable to allow NULL values for last name, birth date, death date, birth country, birth city, and gender. NULL values were also considered acceptable for institution name and institution country as some laureates didn’t belong to an institution. The country dimension provides a reference between country name and country abbreviation. Unlike the country name, country abbreviation allows for NULL values because some countries have dissolved. The prize dimension keeps individual records of the year and category for all Nobel Prizes ever won while

the institution dimension keeps individual non-duplicate records of all institutions. Now, the additional dimensions that were created to further organize the data warehouse include the prize category dimension and the institution type dimension. The prize category dimension contains non-duplicate values as it would just classify what type of category a prize belongs to. Since there are six types of Nobel Prizes, there are six categories: physics, chemistry, medicine, literature, peace, and economics. The institution type dimension also contains non-duplicate values as it categorizes the institution. It only has rows for educational, religious, medicinal, government, economic, and research institutions. The last two dimensions created were Motivation and Overall Motivation. These dimensions contain information and also categorize the type of motivation the laureate had, however, it was decided not to implement them into any fact tables as their information could be added to the laureate dimension instead. Additionally, all dimensions included a key and an ID column. The ID column was used to create the key column, which was then used to facilitate the surrogate key pipeline necessary to map the dimensions to the facts.

LAUREATES HV

Laureate ID	First Name	Last Name	Date of Birth	Date of Death	Birth Country	Birth City	Gender	Year Prize Received	Prize Category	Prize Amount	Motivation	Share	Institution Name	Institution Country
BK, NN	(LAUREATE.CSV)	(LAUREATE.CSV)	FV (LAUREATE.CSV)	(LAUREATE.CSV)	FV (LAUREATE.CSV)	FV (LAUREATE.CSV)	(LAUREATE.CSV)	FV, NN (LAUREATE.CSV)	FV, NN (LAUREATE.CSV)	NN (LAUREATE.CSV)	(LAUREATE.CSV)	NN (LAUREATE.CSV)	(LAUREATE.CSV)	(LAUREATE.CSV)
I	C100	C100	D	D	C100	C100	C10	D	C20	N	C300	I	C150	C100
1	Wilhelm Conrad	Rantgen	1845-03-27	2/10/1923	Prussia (now Germany)	Lennep (now Remscheid)	male	1901	physics	1000000	"in recognition"	1	Munich University	Germany
2	Hendrik Antoon	Lorentz	1853-07-18	2/4/1928	the Netherlands	Arnhem	male	1902	physics	500000	"for discovery"	2	Leiden University	the Netherlands
3	Pieter	Zeeman	1865-05-25	10/9/1943	the Netherlands	Zonnemaire	male	1902	physics	500000	"for recognition"	2	Amsterdam University	the Netherlands
4	Antoine Henri	Becquerel	1852-12-15	8/25/1908	France	Paris	male	1903	physics	500000	"in recognition"	2	Polytechnique	France
5	Pierre	Curie	1859-05-15	4/19/1906	France	Paris	male	1903	physics	250000	"for discovery"	4	(Municipal School of Industrial Physics and Chemistry)	France
6	Marie	Curie Sklodowska	1867-11-07	7/4/1934	Russian Empire (now Poland)	Warsaw	female	1903	physics	250000	"for recognition"	4	Frankfurt-on-the-Main University	Germany
6	Marie	Curie, Sklodowska	1867-11-07	7/4/1934	Russian Empire (now Poland)	Warsaw	female	1911	chemistry	1000000	"in recognition"	1	Sorbonne University	France
8	Lord Rayleigh	John William Strutt	1842-11-12	6/30/1919	United Kingdom	Langford Grove, Maldon, Essex	male	1904	physics	1000000	"for discovery"	1	Royal Institution of Great Britain	United Kingdom
9	United Nations	NULL	0000-00-00	0000-00-00	NULL	NULL	NULL	1905	peace	1000000	"for recognition"	1	NULL	NULL

2.2.1 Laureate Dimension

OVERALL MOTIVATION

Overall Motivation Key	Overall Motivation ID	Overall Motivation
SK, NN	BK, NN	NN
	{OverallMotivation.CSV}	{OverallMotivation.CSV}
I	C5	C100
1	OM001	for pioneering experimental contributions to lepton physics
2	OM002	for contributions to the developments of methods within DNA-based chemistry
3	OM003	for basic work on information and communication technology
4	OM004	for the development of methods for identification and structure analyses of biological macromolecules
5	OM005	for discoveries concerning channels in cell membranes
6	OM006	for pioneering contributions to the development of neutron scattering techniques for studies of condensed matter

2.2.2 Overall Motivation Dimension

MOTIVATION

Motivation Key	Motivation ID	Motivation
SK, NN	BK, NN	NN
	{Motivation.CSV}	{Motivation.CSV}
I	C5	C300
1	M001	in recognition of his work in thermochemistry
2	M002	for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material
3	M003	for their discoveries concerning the structural and functional organization of the cell
4	M004	for the discovery of the quantized Hall effect
5	M005	for their pathbreaking contribution to the theory of international trade and international capital movements
6	M006	for having created new poetic expressions within the great American song tradition
7	M007	for his discovery of the organizer effect in embryonic development
8	M008	for his development of nuclear magnetic resonance spectroscopy for determining the three-dimensional structure of biological macromolecules in solution
9	M009	for their discoveries relating to the hormones of the adrenal cortex, their structure and biological effects

2.2.3 Motivation Dimension

COUNTRY

Country Key	Country ID	Country Name	Country Abbreviation
SK, NN	BK, NN	NN	
	{Country.CSV}	{Country.CSV}	{Country.CSV}
I	C5	C100	C3
1	C100	Alsace, then Germany	DE
2	C101	Alsace	DE
3	C102	Germany	DE
4	C103	Argentina	AR
5	C104	Australia	AU
6	C105	Austria	AT
7	C106	Belgium	BE
8	C107	Burma	MM
9	C108	Old Republic	NULL

2.2.4 Country Dimension

PRIZE

Prize Key	Prize ID	Prize Year	Prize Category
SK, NN	BK, NN	NN	NN
	Prize.CSV}	{Prize.CSV}	{Prize.CSV}
I	C5	C4	C20
1	P001	2016	physics
2	P002	2016	physics
3	P003	2016	physics
4	P004	2016	chemistry
5	P005	2016	chemistry
6	P006	2016	chemistry
7	P007	2016	medicine
8	P008	2016	literature
9	P009	2016	peace

2.2.5 Prize Dimension

PRIZE CATEGORY

Category Key	Category ID	Category
SK, NN	BK, NN	ND
	{Category.CSV}	{Category.CSV}
I	C5	C20
1	NPC1	physics
2	NPC2	chemistry
3	NPC3	medicine
4	NPC4	peace
5	NPC5	literature
6	NPC6	economics

2.2.6 Prize Category Dimension

INSTITUTION

Institution Key	Institution ID	Institution Name	Institution City	Institution Country
SK, NN	BK, NN	NN, ND		
	{Institution.CSV}	{Institution.CSV}	{Institution.CSV}	{Institution.CSV}
I	C6	C100	C100	C100
1	IN001	Kiel University	Kiel	Germany
2	IN002	University of Cambridge	Cambridge	United Kingdom
3	IN003	University of Chicago	Chicago, IL	USA
4	IN004	Sorbonne University	Paris	France
5	IN005	Marconi Wireless Telegraph Co. Ltd.	London	United Kingdom
6	IN006	Strasbourg University	Strasbourg	Alsace (then Germany, now France)
7	IN007	Amsterdam University	Amsterdam	the Netherlands
8	IN008	Harvard	Cambridge, MA	United States
9	IN009	Swedish Gas-Accumulator Co.	Stockholm	Sweden

2.2.7 Institution Dimension

INSTITUTION TYPE

Institution Type Key	Institution Type ID	Institution Type	Institution Type Description
SK, NN	BK, NN		
	(InstitutionType.CSV)	(InstitutionType.CSV)	(InstitutionType.CSV)
I		C30	C200
1	INT001	Educational	involves schools and universities either being private or public
2	INT002	Religious	involves religious groups i.e. Christianity, Judaism, Islam, Hinduism, Buddhism, and Sikhism
3	INT003	Medicinal	involves hospitals and other health care institutions either being private or public
4	INT004	Government	involves governments that are either monarchy, oligarchy, dictatorship, and democracy
5	INT005	Economic	involves foundations focused on collecting economic data with the idea of providing a good or service that is important to the economy
6	INT006	Research	involves establishments that speciaize in advancing scholarly activity through research and experimental development

2.2.8 Institution Type Dimension

Identifying the events and the dimensions laid the foundation for creating the fact tables. With the first event, laureates receiving prize money, the laureate, country, institution, prize, and prize category dimensions were needed to create an effective picture of the following numerical data: share and prize amount in USD. A second fact table was created to fulfill the institution enrollment and value event. This fact table contains numeric information about the institution value in USD and information about the member count. To gain insight into this information, it was decided that the institution, institution type, and country dimensions would be linked through surrogate keys. Lastly, the distributed prize event led to the creation of a third fact table that had the prize dimension, the laureate dimension, and the prize category dimension linked. The numeric information in this fact table included the prize share and the prize distribution amount based on the winner in USD.

LAUREATE PRIZE

LAUREATE KEY	originates COUNTRY KEY	belongs INSTITUTION KEY	receives PRIZE KEY	in CATEGORY KEY	SHARE	earned PRIZE AWARD
[who]	[where]	[what]	[what]	[what]	[how]	[$]
53	87	88	124	2	1	1000000
88	81	86	11	3	2	500000
30	17	47	7	2	2	500000
5	70	89	13	1	2	500000
59	21	75	134	2	4	250000
26	47	14	37	1	4	250000
84	22	80	76	2	1	1000000
91	87	4	188	6	1	1000000
73	37	77	196	6	1	1000000

2.2.9 Laureate Prize Fact

INSTITUTION VALUE & MEMBERS

Institution Key	Institution Type Key	Country Key	has Institution Value in Millions	with Institution Member Count
[who/what]	[what]	[where]	[$]	
88	1	87	16.6	424
86	2	81	723.4	7744
47	3	17	40.3	399333
89	4	70	51.7	234
75	5	21	3.2	32
14	6	47	1.9	6533
80	1	22	500	66250
4	2	87	221.2	29309
77	3	37	3.3	437

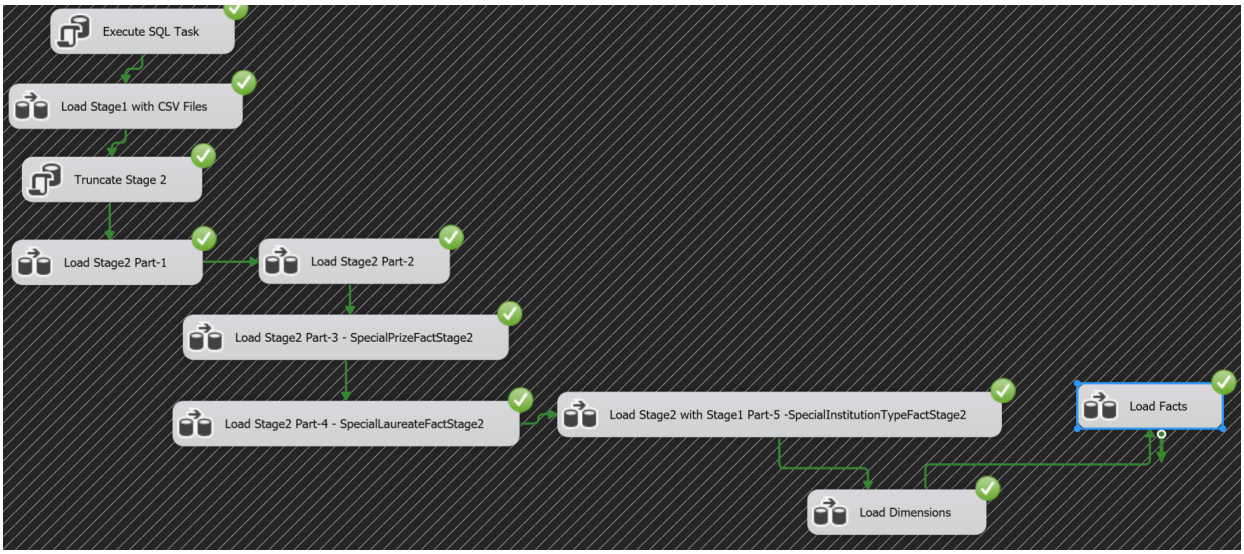
2.2.10 Institution Value and Members Fact

DISTRIBUTED MONEY

Laureate Key	recieves Prize Key	in Category Key	has to Share	results Prize Distribution per Winner
[who]	[what]		[how]	[$]
53	124	2	1	500000
88	11	3	2	250000
30	7	2	2	250000
5	13	1	2	333333.3333
59	134	2	4	333333.3333
26	37	1	4	333333.3333
84	76	2	1	1000000
91	188	6	1	1000000
73	196	6	1	1000000

2.2.11 Distributed Money Fact

3 ETL Process

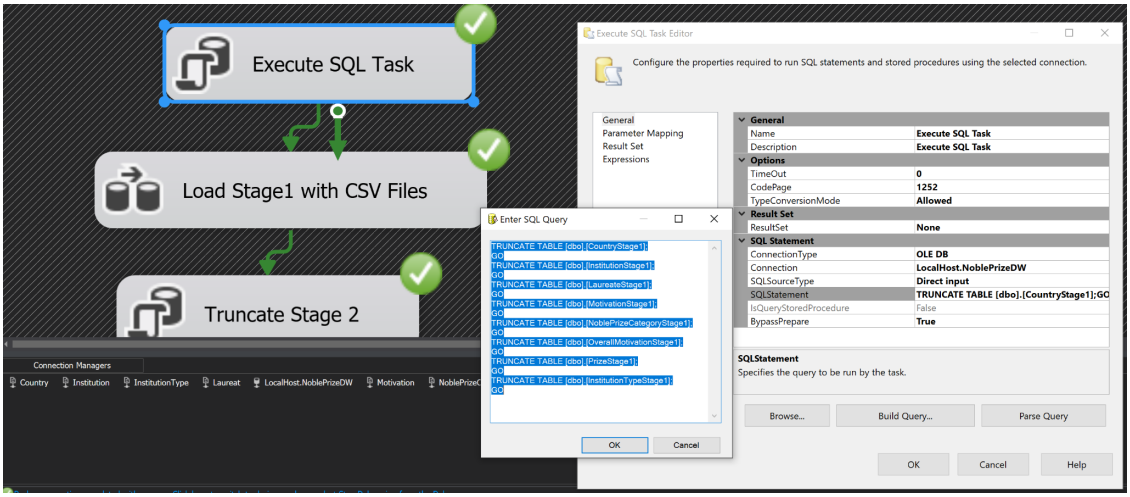


3.1 ETL Process: Whole Diagram

The diagram above shows the entire ETL Process's steps from executing the SQL task to loading the Facts. There are currently two *Execute SQL Task* steps, the *"Execute SQL Task"* and "*Truncate Stage 2"*. Their purpose is to create SQL statements that truncate the *Stage 1* SQL Server Destinations in the *"Load Stage1 with CSV files"* Data Flow Task and truncates the *Stage 2* SQL Server Destinations in the "*Load Stage2 Parts 1 to 5*" Data Flow Tasks.

The Data Flow Tasks, *"Load Stage1 with CSV Files"* and *"Load Stage2 - Parts 1 to 5"*, consist of transforming and loading the data from the *CSV file Extraction Stage* to *Stage 1* and *Stage 1* to *Stage 2*.

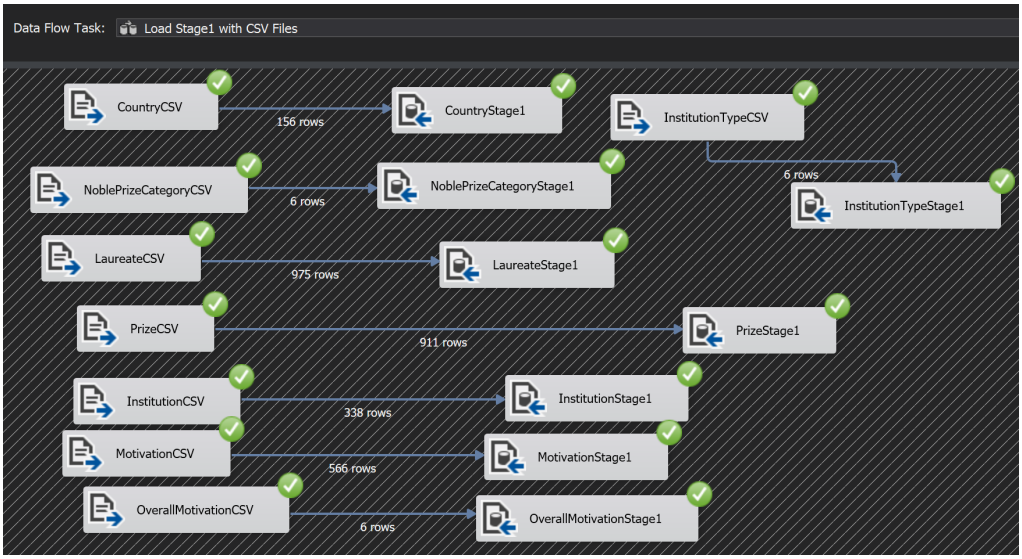
The *"Load Dimensions"* and *"Load Facts"* Data Flow Tasks load and transform the data from *Stage 2* to *Dimensions* and from *Stage 2* to *Facts*, where the *"Load Facts"* Data Flow Task has surrogate key pipelines to connect the facts with the dimensions through the surrogate keys that were made in the *"Load Dimensions"* Data Flow Task.

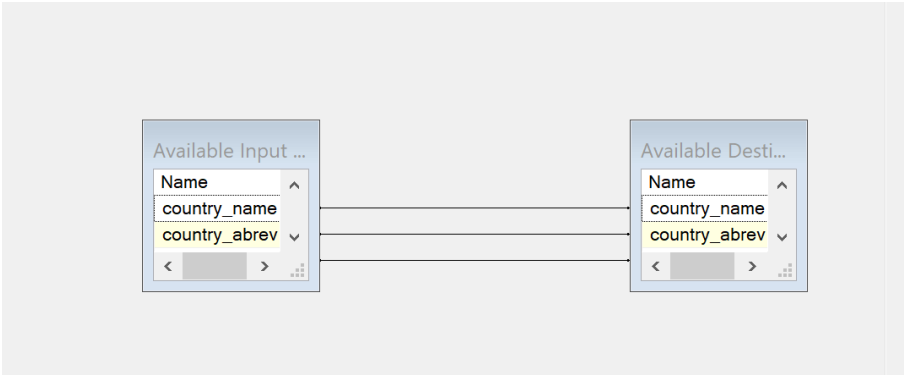


3.2 ETL Process: Execute SQL Task

The *"Execute SQL Task"* step contains a SQL statement in the connection, *LocalHost.NoblePrizeDW*, which truncates and executes the *Stage 1* SQL Server Destinations in

the *"Load Stage1 with CSV Files"* Data Flow Task. The SQL statement will help implement the *"Load Stage1 with CSV Files"* Data Flow Task to load the extracted 8 CSV files into the 8 stages: *CountryStage1*, *InstitutionStage1*, *LaureateStage1*, *MotivationStage1*, *NoblePrizeCategoryStage1*, *OverallMotivationStage1*, *PrizeStage1*, and the *InstitutionTypeStage1*. This step is necessary to help load and transform data from the CSV files to the Facts and Dimensions which will be the building blocks of making the marts in the Nobel Prize Data Warehouse in the later steps.

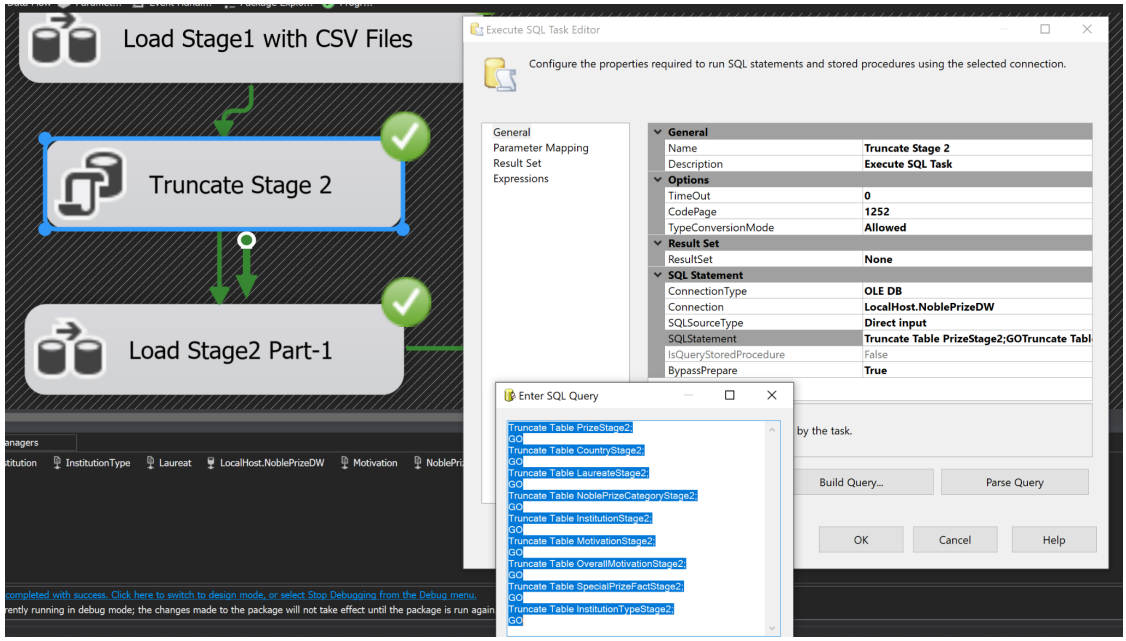




3.3.1 ETL Process: Load Stage1 with CSV Files of Country, Nobel Prize Categories, Laureates, Prizes, Institutions, Motivations, Institution Type, and Overall Motivations

3.3.2 Country Example of the CSV Available Inputs Variables being Mapped to the Stage 1 Available Destination Variables

In the *"Load Stage1 with CSV Files"* Data Flow Task extracts and loads the data from the 8 CSV files (*Country, Nobel Prize Categories, Laureates, Prizes, Institutions, Motivations, Institution Type, and Overall Motivations*) to the *Stage 1* SQL Server Destinations, which lets the *Stage 1* SQL Server Destinations to obtain the variables from the CSV files by mapping, which is further illustrated by the image above that shows an example of the *"CountryCSV"* available input variables being mapped to the available destination variables in *"CountryStage1"*. This step is necessary since this is obtaining data from the CSV files that will be further extracted by the *Stage 2* SQL Server Destinations in the *"Load Stage2 Parts 1 to 5"*, which will be further explained in the next sections.

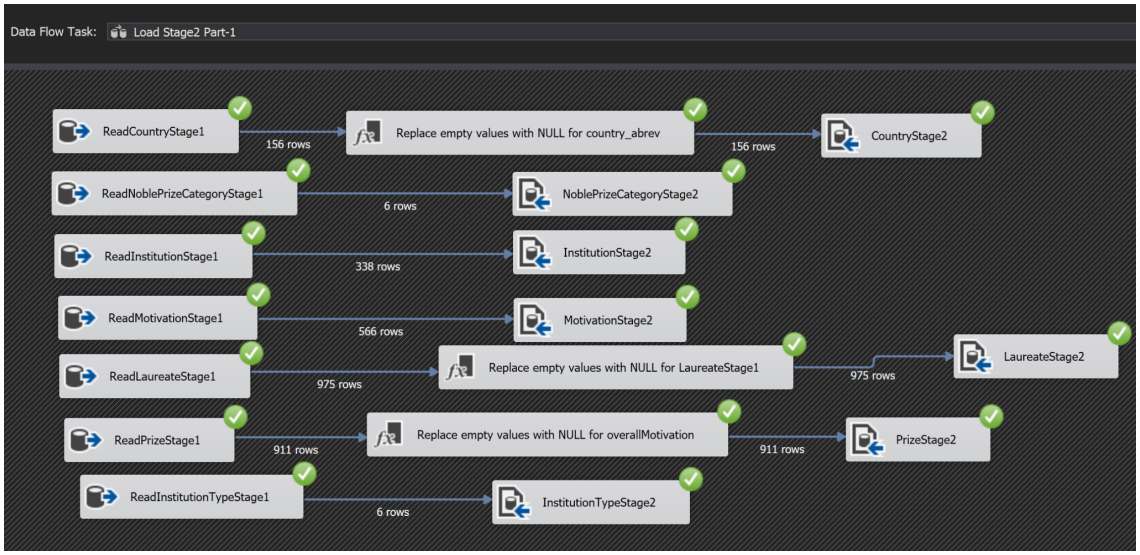


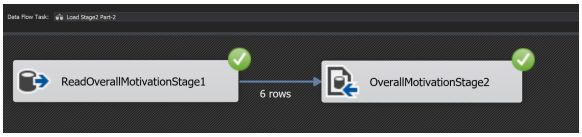
3.4 ETL Process: Truncate Stage 2

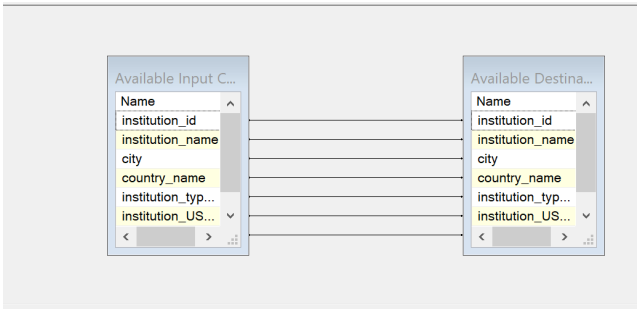
The *"Truncate Stage 2"* Execute SQL Task shown above contains a SQL statement in the connection, *LocalHost.NoblePrizeDW*, which truncates the *Stage 2* SQL Server Destinations in

the *"Load Stage2 Parts 1 to 5"* Data Flow Tasks. The SQL statement will help implement the *"Load Stage2 Parts 1 to 5"* Data Flow Tasks, which is divided into 5 parts, to load the transformed and extracted data from the *Stage 1* OLE DB Sources with the *Stage 2* OLE DB Sources: *"CountryStage2", "InstitutionStage2", "LaureateStage2"*, *"MotivationStage2"*, *"NoblePrizeCategoryStage2"*, *"OverallMotivationStage2", "PrizeStage2", and "InstitutionTypeStage2"*.

This step is necessary to help load and transform data from the *Stage 1* OLE DB Sources to *Stage 2* SQL Server Destinations that will make the Facts and Dimensions which will be used to make the marts needed in the Nobel Prize Data Warehouse.







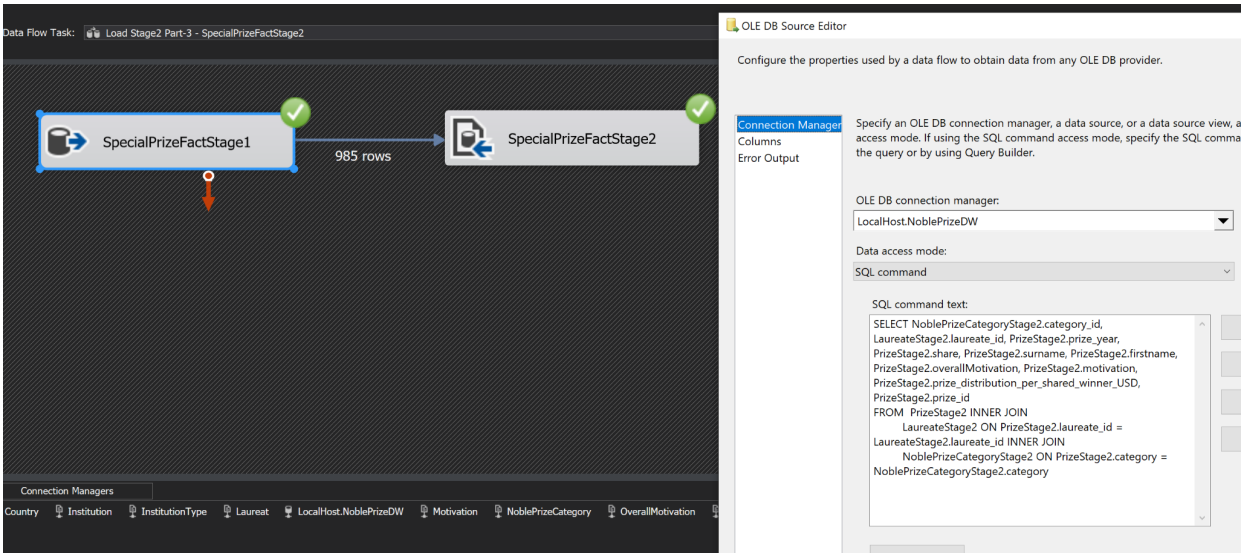
Input Column	Destination Column
institution_id	institution_id
institution_name	institution_name
city	city
country_name	country_name
institution_type_id	institution_type_id
institution_USD_value_in_millions	institution_USD_value_in_millions
institution_member_count	institution_member_count

3.5.1 ETL Process: Load Stage2 with Stage1 of Country, Nobel Prize Category, Institution, Motivation, Laureate, Prize, Institution Type, and Overall Motivation which will become Dimensions in the Next Steps- Part 1 and 2

3.5.2 Institution Example of the Stage1 Available Inputs Variables being Mapped to the Stage 2 Available Destination Variables

As you can see from the previous image, *3.1 ETL Process: Whole Diagram*, the *"Load Stage2"* Data Flow Tasks are divided into 5 parts. The reason for this strategy is due to the amount of excessive data overflow crashing in Visual Studio undergoes every time all the flows are implemented under one Data Flow Task. Dividing the Data Flow Task into 5 parts greatly reduced the runtime of the execution of the program's code and prevented my computer from being overwhelmed by the massive data being transferred. The *Country*, *Laureate*, *Prize*, and *Institution* flows have derived columns that replace the empty values with NULLs so that it does not affect the data analysis later on since empty values can greatly affect the performance of a data warehouse and its data analysis.

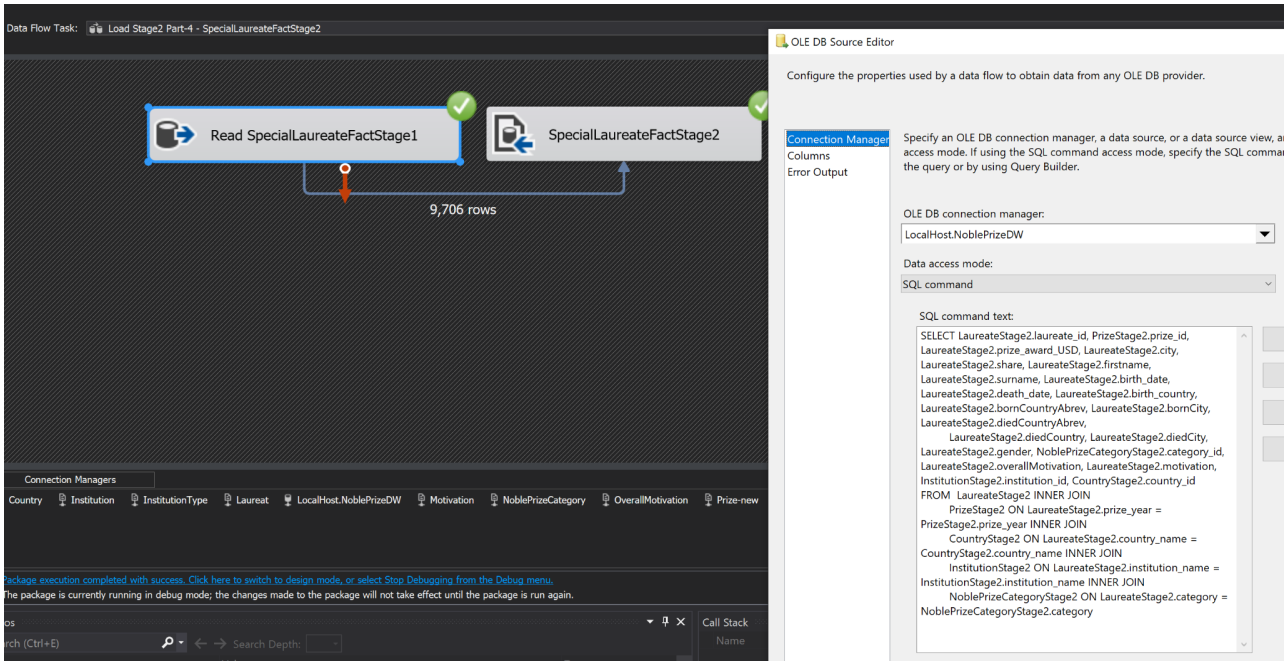
In the *"Load Stage2 Parts 1 to 2"* Data Flow Tasks extracts and loads the data from the *Stage 1* OLE DB Sources to the *Stage 2* SQL Server Destinations, which lets the *Stage 2* SQL Server Destinations obtain the variables from the *Stage 1* OLE DB Sources through mapping. This is further illustrated by the image above that shows an example of the *"ReadInstitutionStage1"* available input variables being mapped to the available destination variables in *"InstitutionStage2"*. The flows from the *"Load Stage 2 Parts 1 to 2"* Data Flow Tasks will be further manipulated to become dimensions in the *"Load Dimensions"* Data Flow Task, which will be further explained in the later sections. The next 3 sections will show the data flows of *"Load Stage 2 Part 3 to 5"* Data Flow Tasks, which will later become Facts in the *"Load Facts"* Data Flow Task.



3.6 ETL Process: Load Stage2 of Special Prize Fact with data from Stage 1 - Part 3

In the *"SpecialPrizeFactStage1"* OLE DB Source of the *"Load Stage2 Part3 - SpecialPrizeFactStage2"* Data Flow Task, a SQL command is used to create a table that obtains the variables, *category_id, laureate_id, prize_id, prize_year, share, surname, firstname, overallMotivation, motivation,* and *prize_distribution_per_shared_winner_USD*, by selecting

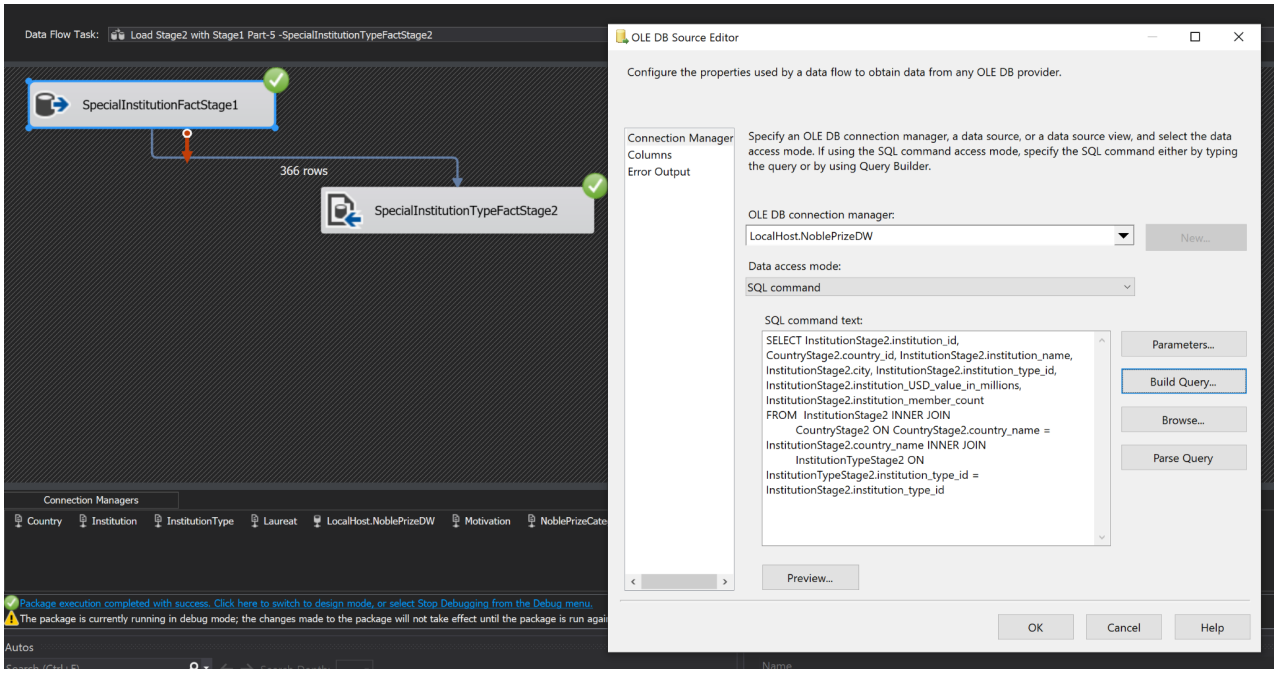
from *"PrizeStage2"* table and inner joining with the *"NobelPrizeCategoryStage2"* table and *"LaureateStage2"* table from the OLE DB connection, *"LocalHost.NoblePrizeDW"*. Typically, in order to inner join another table you have to obtain the shared keys between the two tables. However, the original dataset didn't have natural keys or IDs for most of the tables. The SQL statement was only able to be implemented by inner joining other variables like the *"NobelPrizeCategoryStage2"* and *"PrizeStage2"* tables' shared variable, *category*, and selecting the *category_id* that exists in the *"NobelPrizeCategoryStage2"* table, which the ID variables for each of the connection table is needed to allow surrogate key pipelines to work with creating Facts with the dimensions in the *"Load Facts"* Data Flow Task. Once *Stage 1* is completed, the data from the *Stage 1* OLE DB Source gets extracted by *Stage 2* through mapping.



3.7 ETL Process: Load Stage2 of Special Laureate Fact with Stage 1 - Part 4

In the *"SpecialLaureateFactStage1"* OLE DB Source of the *"Load Stage2 Part4 - SpecialLaureateFactStage2"* Data Flow Task, a SQL command is used to create a table that obtains the variables, *laureate_id, prize_id, prize_award_USD, city, share, firstname, surname,*

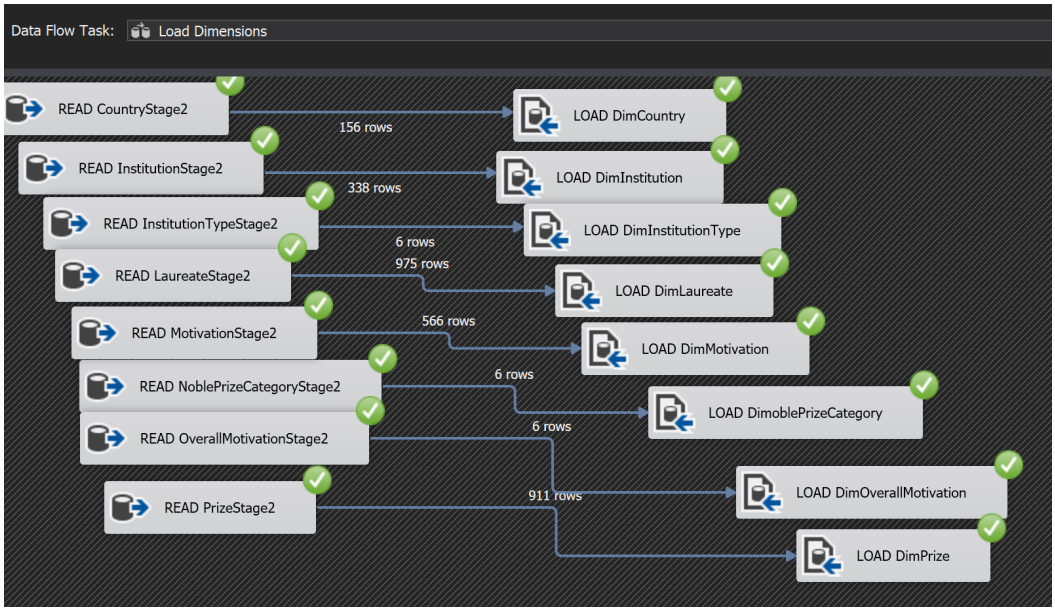
birth_date, death _date, birth_country, bornCountryAbrev, bornCity, diedCountryAbrev, diedCountry, diedCity, gender, category_id, overallMotivation, motivation, institution_id, and country_id, by selecting from the *"LaureateStage2"* table and inner joining with the tables, *"NobelPrizeCategoryStage2"*, *"InstitutionStage2"*, *"CountryStage2"*, and *"PrizeStage2"*, from the OLE DB connection, *"LocalHost.NoblePrizeDW"*. The original dataset didn't have natural keys or IDs for most of the tables, so the SQL statement was only able to be implemented by inner joining other variables like the *"NobelPrizeCategoryStage2"* and *"LaureateStage2"* tables' shared variable, *category*, and selecting the *category_id* that exist in the *"NobelPrizeCategoryStage2"* table, which the ID variables for each of the connection table is needed to allow surrogate key pipelines to work with creating Facts with the Dimensions in the *"Load Facts"* Data Flow Task. Once *Stage 1* is completed, the data from the *Stage 1* OLE DB Source gets extracted by *Stage 2* through mapping.

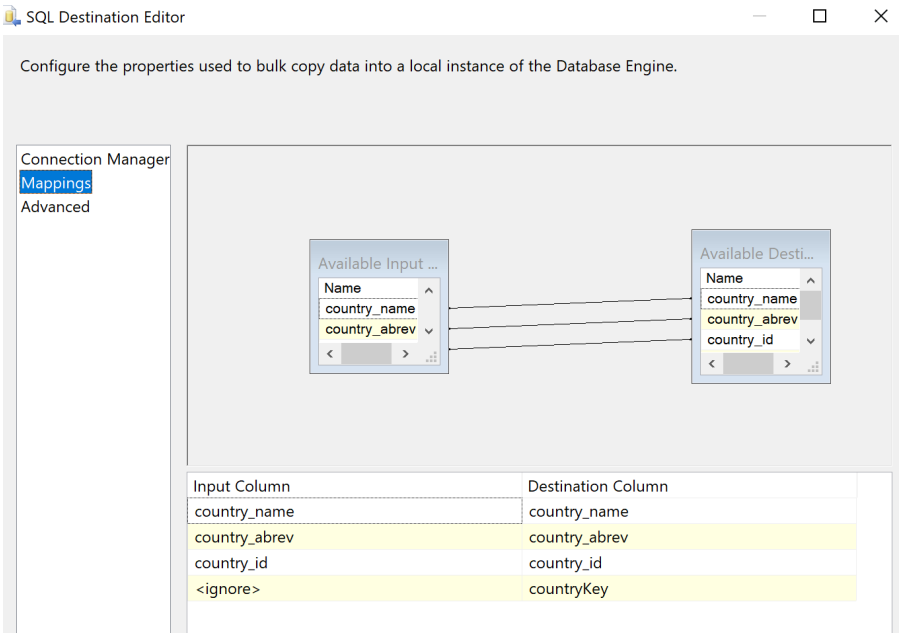


3.8 ETL Process: Load Stage2 of Special Institution Fact with Stage 1 - Part 5

Extracts the data from the Stage1 OLE DB Source to Stage 2 SQL Server Destination with Stage 1 made through a SQL Command

In the *"SpecialInstitutionFactStage1"* OLE DB Source of the *"Load Stage2 Part4 - SpecialInstitutionFactStage2"* Data Flow Task, a SQL command is used to create a table that obtains the variables, *institution_id, country_id, institution_name, city, institution_type_id, institution_USD_value_in_millions,* and *institution_member_count*, by selecting from *InstitutionStage1* table and inner joining with the tables, *"InstitutionTypeStage2"* and *"CountryStage2"*, from the OLE DB connection, *"LocalHost.NoblePrizeDW"*. The original dataset didn't have natural keys or IDs for most of the tables, so the SQL statement was only able to be implemented by inner joining other variables like the *"CountryStage2"* and *"InstitutionStage2"* tables' shared variable, *country*, and selecting the *country_id* that exist in the *"CountryStage2"* table. Once *Stage 1* is completed, the data from the *Stage 1* OLE DB Source gets extracted by *Stage 2* through mapping.



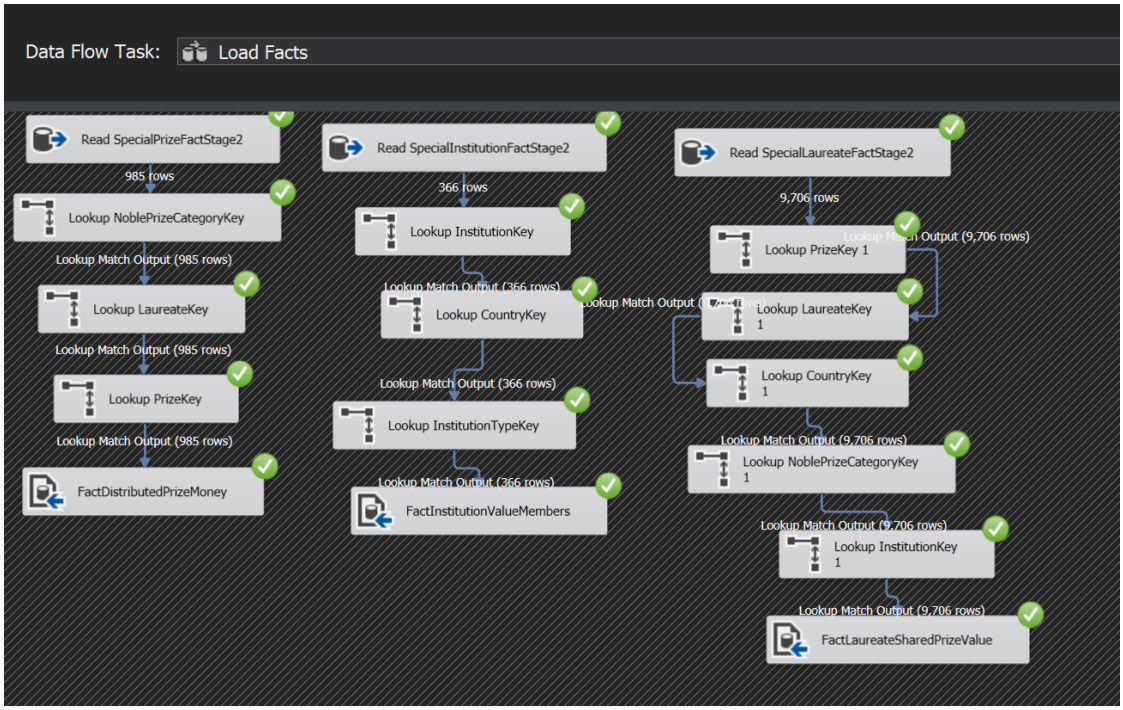


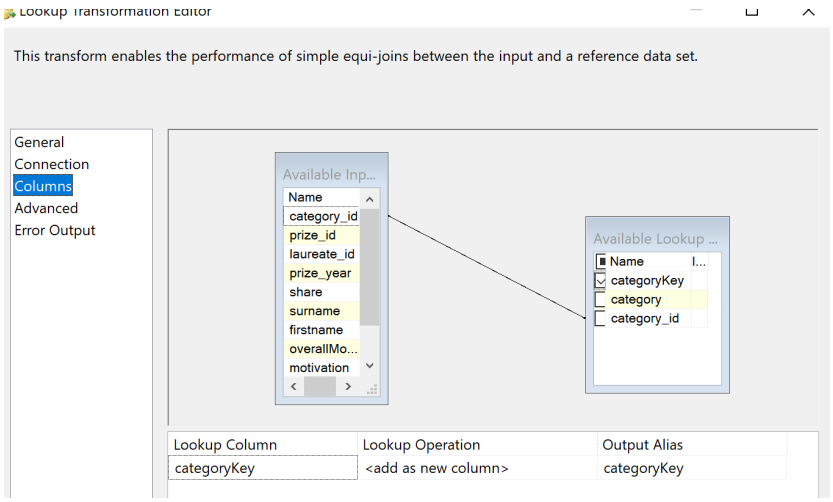
3.9.1 ETL Process: Load Dimensions of Country, Institution, Institution Type, Laureate, Motivation, Nobel Prize Category, Overall Motivation, and Prizes

3.9.2 Load Dimensions Mapping between the Available Input Variables and the Available Destination Variables for the DimCountry as an Example

The Data Flow Task, *"Load Dimensions"*, extracts and reads the data from the *Stage2* OLE DB Sources to the *Load Dimension* SQL Server Destinations which creates 8 Dimensions: *DimCountry, DimLaureate, DimInstitution, DimMotivation, DimOverallMotivation, DimPrize, DimNoblePrizeCategory,* and *DimInstitutionType*. Every flow's *"Load Dim"* SQL Server Destinations creates Surrogate keys that replace the Natural Keys for every Dimension.

As you can see above, image *3.9.2* shows the mappings between the *Country Stage 2* OLE DB Source's Available Input Variables to *DimDCountry* SQL Server Destination's Available Destination Variables. Also, the *countryKey* variable is currently not mapped to an input variable, because it was recently made in *"Load DimCountry"* SQL Server Destination. The mapping structure of the *Country* flow applies the same to the other flows. The Dimensions will be linked to 3 Facts, which will be further discussed in the later sections.



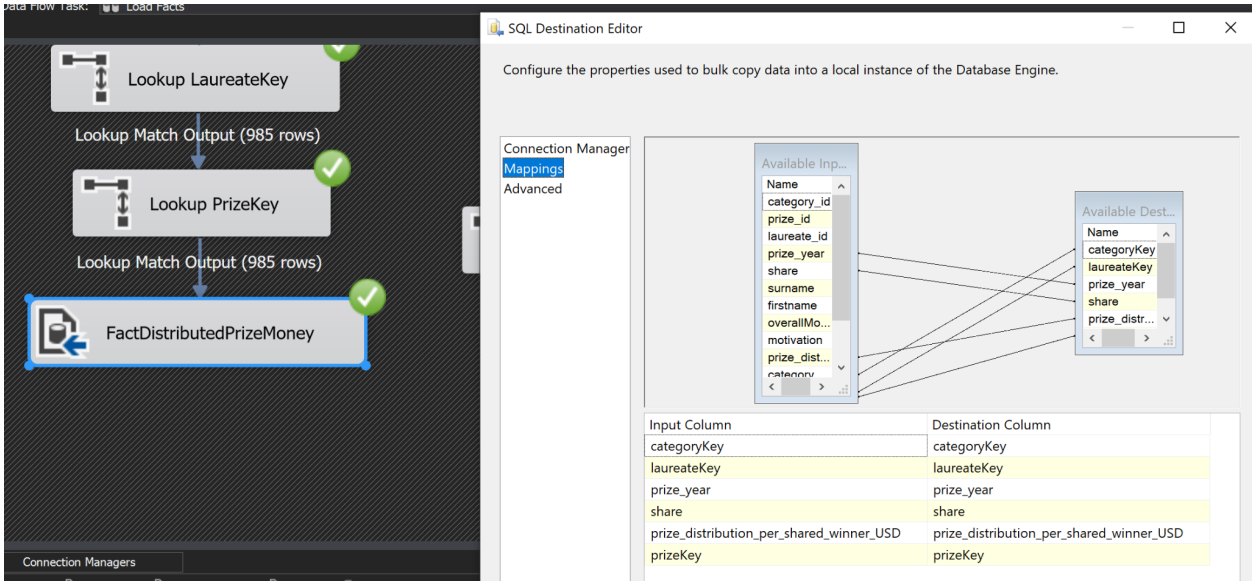


3.10.1 ETL Process: Load Facts of Distributed Prize Money, Institution Value Members, and Laureate Prize Value

3.10.2 Lookup Surrogate Key Pipeline of Dimension Category as Example

The *"Load Facts"* Data Flow Task has 3 Fact flows, *"FactDistributedPrizeMoney", "FactInstitutionMembers",* and *"FactSharedPrizeValue"*. Every Fact flow has a Surrogate Key Pipeline through Lookups that links the Facts with their Dimensions' Surrogate Keys instead of the original Natural Keys so that the Dimensions' Keys are represented in numerical values (i.e. Surrogate Keys) instead of random string values (i.e. Natural Keys).

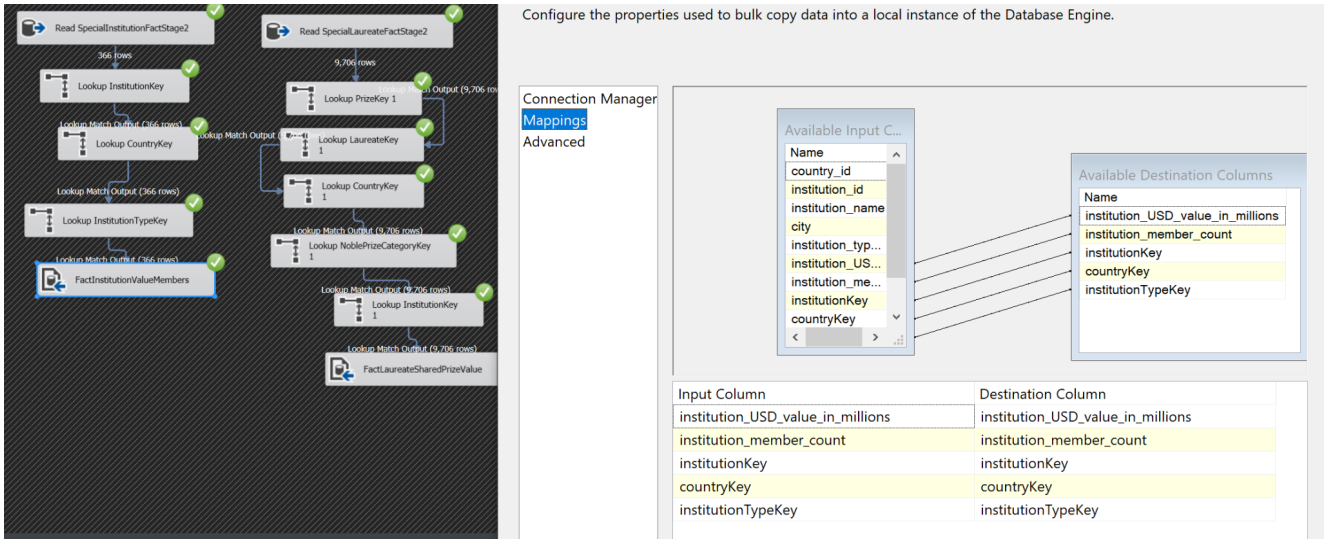
Treat image *3.10.2* as an example that represents all the lookups' structures. The Surrogate Key Pipeline's Dimension Category Lookup links the shared *category_id* variable between *"Read SpecialPrizeFactStage2"* OLE DB Source and the *"NoblePrizeCategoryKey"* Lookup. The point of linking the two tables together is to obtain the Lookup column, *categoryKey*. The Lookup mapping shown above helps the Fact table replace the Natural key with the Surrogate Key. In the next 3 sections, it will show how the numerical variables and the Dimension Surrogate Keys are mapped in each Fact table.



3.11 ETL Process: Load Facts - Distributed Prize Money Fact

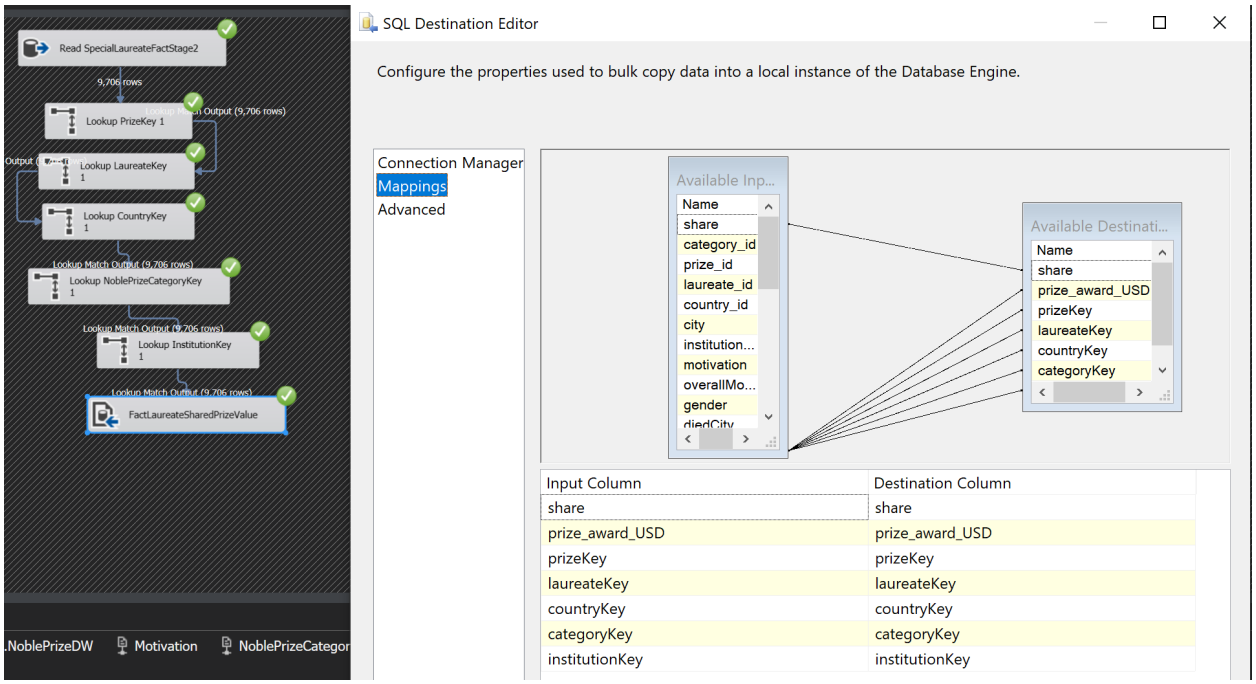
The *"FactDistributedPrizeMoney"* is a Fact table that measures the Noble Prize money distribution and shared credit between the co-laureates for a specific Nobel Prize won in a certain

year for their given Nobel Prize category. The *Fact Distributed Prize Money* SQL Server Destination extracts the data from the *Stage 2 OLE DB Source* with a Lookup Surrogate Key pipeline. As you can see, only the variables, *categoryKey, laureateKey, prize_year, share, prizeKey, and the prize_distribution_per_shared_winner_USD*, were used in this Fact table, because the Facts only need the numerical variables and the Surrogate Key variables in order to create the *Distributed Prize Money* Mart that allow data analysis.



3.12 ETL Process: Load Facts - Institution Value Members Fact

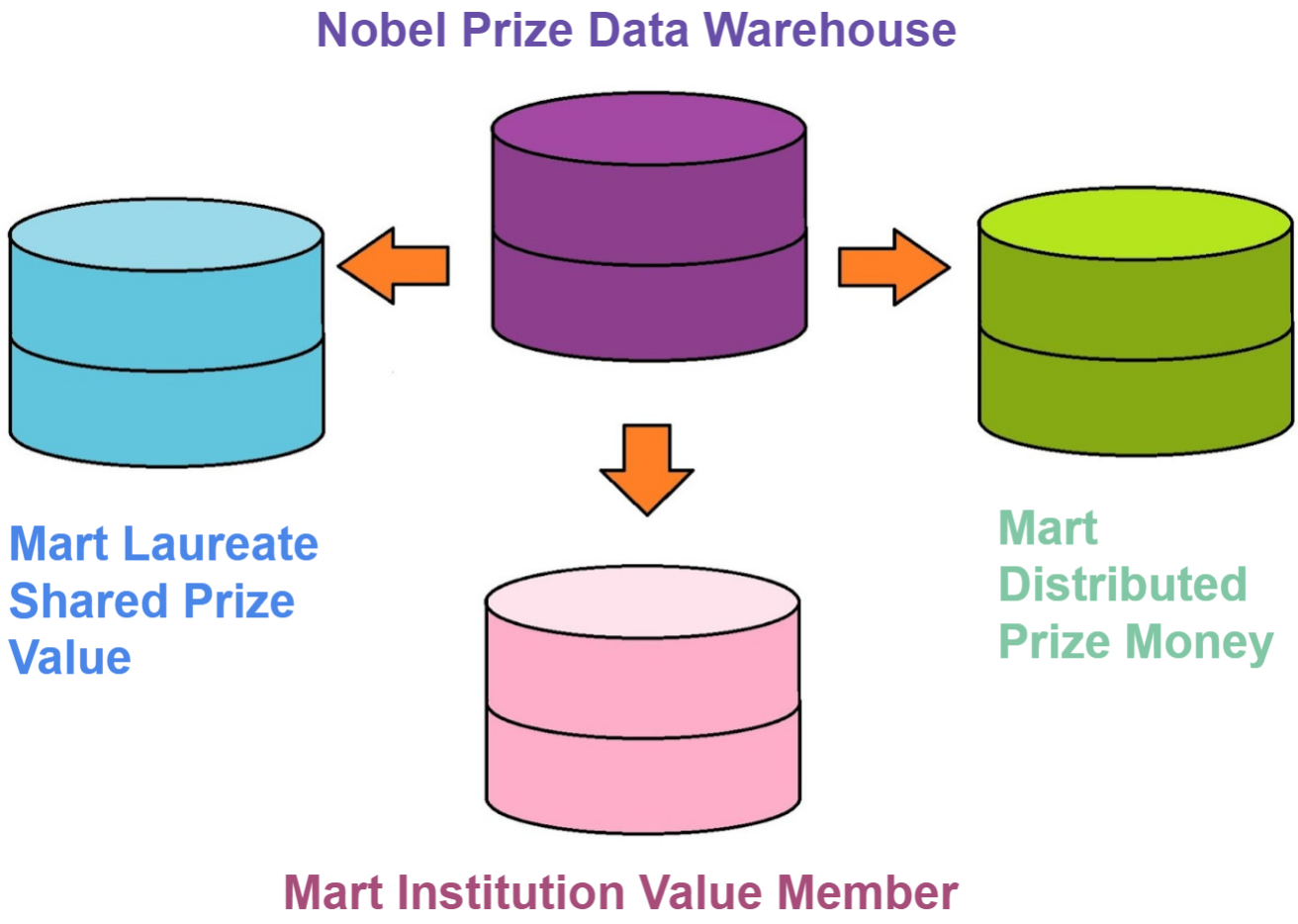
The *"FactInstitutionValueMembers"* is a Fact table that measures the Nobel Prize Laureates' Institutions' worth in millions of USD and the Institutions' number of registered members and identifies the Institutions' country and institution type. The *Fact institution Value Members* SQL Server Destination extracts the data from the *Stage 2 OLE DB Source* with a Lookup Surrogate Key pipeline. As you can see, only the variables, *institutionKey, countryKey, institutionTypeKey, institution_USD_value_in_millions,* and *institution_member_count* were used in this Fact table, because the Facts only need the numerical variables and the Surrogate Key variables in order to create the Institution Value Members Mart that allow data analysis.



3.13 ETL Process: Load Facts - Laureate Shared Prize Value Fact

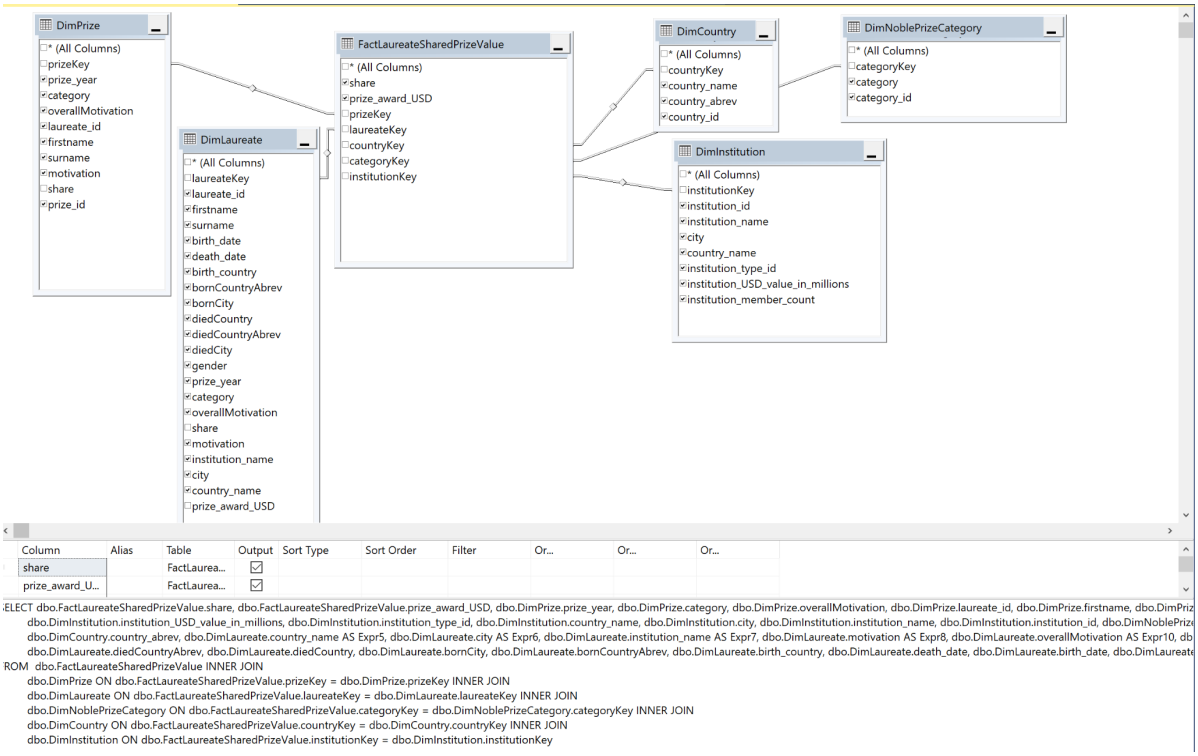
The *"FactLaureateSharedPrizeValue"* is a Fact table that measures the Laureate's Noble Prize money distribution in USD and shared credit between the other co-Laureates for a specific Nobel Prize won in a certain year for their given Nobel Prize category. The Fact table identifies the Laureate's represented institution and country. The *Fact Laureate Shared Prize Value* SQL Server Destination extracts the data from the *Stage 2 OLE DB Source* with a Lookup Surrogate Key pipeline. As you can see, only the variables, *prizeKey, laureateKey, countryKey, categoryKey, institutionKey, share,* and *prize_award_USD,* were used in this Fact table, because the Facts only need the numerical variables and the Surrogate Key variables in order to create the *Laureate Shared Prize Value* Mart that allow data analysis.

4 Data Warehouse Views



4.1 Diagram of the Nobel Prize Data Warehouse with the Laureate Shared Prize Value Mart, Institution Value Member Mart, and Distributed Prize Money Mart

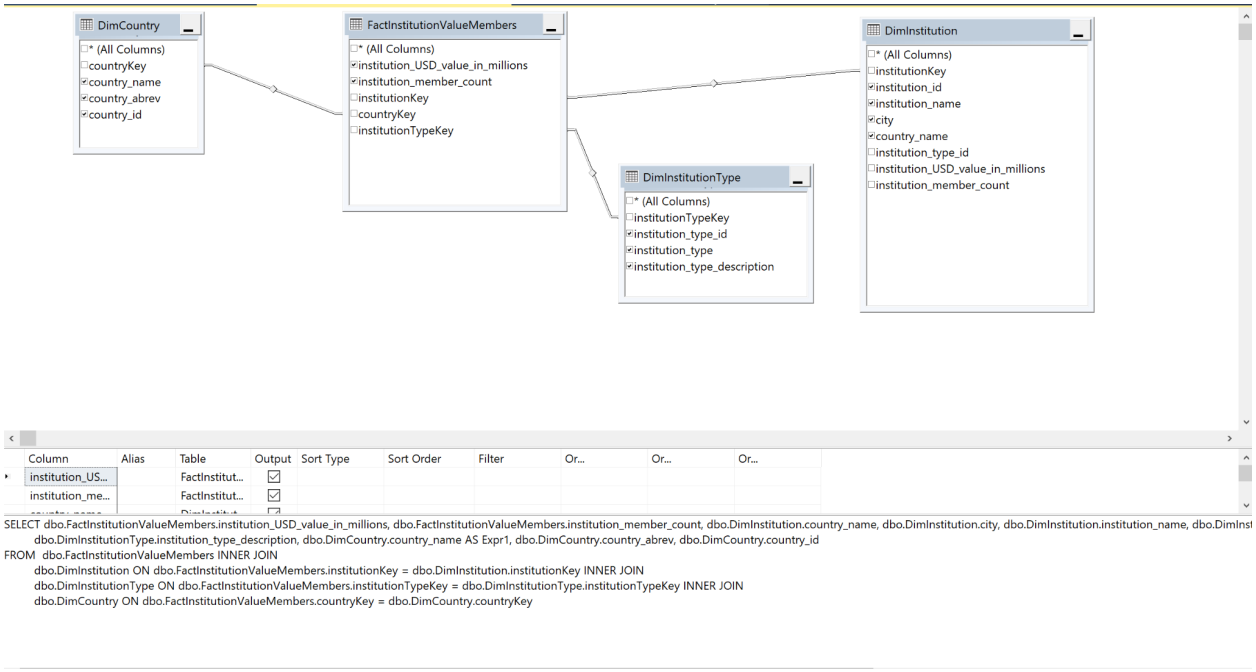
The point of this diagram shown above is to illustrate the *Nobel Prize Data Warehouse* being divided into 3 Marts: the *Laureate Shared Prize Value Mart*, the *Institution Value Member Mart*, and the *Distributed Prize Money Mart*. This is used as a reference for the members to use to understand the structure of the *Nobel Prize Data Warehouse* and how to build it in the Microsoft SQL Server Management.



4.2 Nobel Prize Data Warehouse View of the Laureate Shared Prize Value Mart with the Laureate Shared Prize Value Fact Table and the Prize, Laureate, Country, Institution, and Noble Prize Category Dimension Tables

The *Laureate Shared Prize Value Mart* has the Fact table, *FactLaureateSharedValue*, and 5 Dimensions: *DimCountry, DimLaureate, DimPrize*, *DimInstitution,* and *DimNoblePrizeCategory*. These Dimensions are obtained by the Surrogate Keys they share with the Fact table. The numerical variable, *share,* identifies how many Laureates the specific Laureate shares the specific Nobel Prize for that year. The numerical variable, *prize_award_USD,* identify the monetary value of the Nobel Prize Award the Laureate received. Those two numerical variables are checked because they are needed to make numerical data analysis. However the Surrogate Keys are unchecked because they only serve the purpose of linking the Fact table with the Dimension tables, so they are not needed in the end result of the

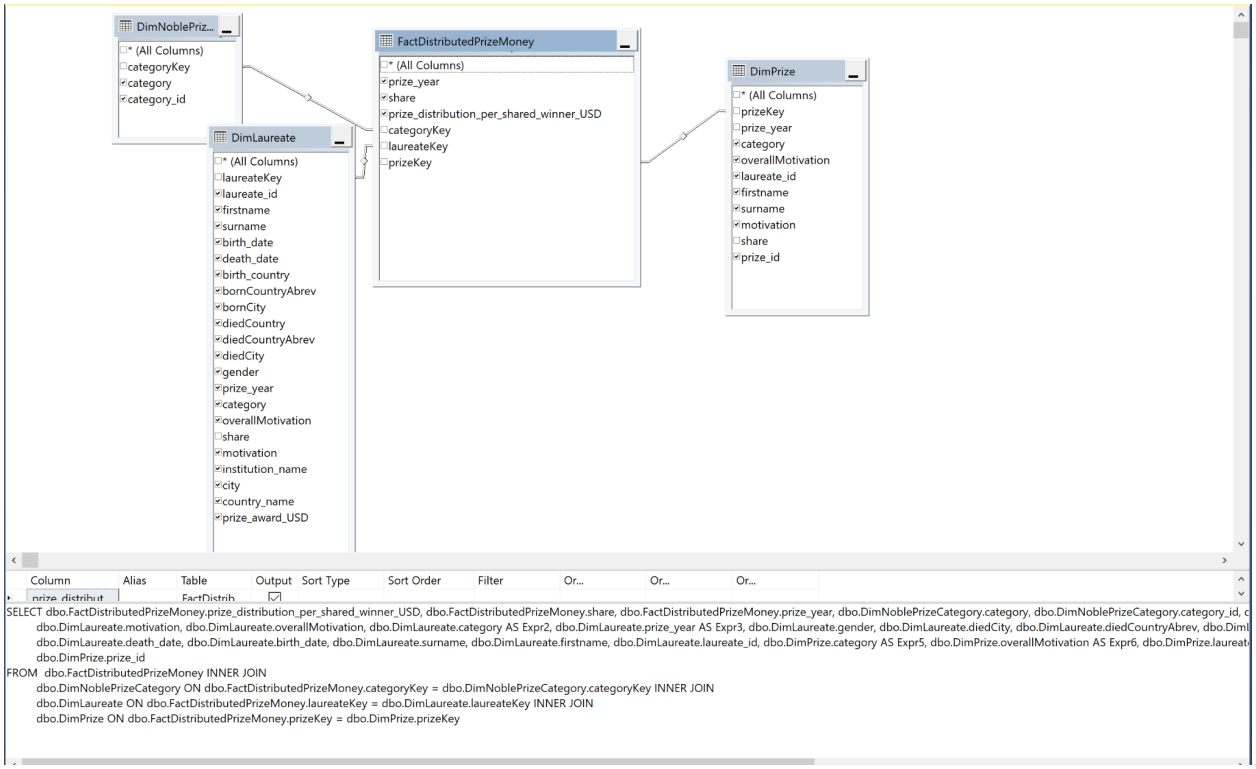
Mart afterward. The other variables in the Dimensions besides the Surrogate Keys are checked because they will be used to create more in-depth data analysis. The other variables in the Dimensions besides the Surrogate Keys are checked because they will be used to create deeper-level data analysis. Take note that some variables that are not Surrogate Keys in the Dimensions are unchecked because those variables either already exist in the Fact or other Dimensions.



4.3 Nobel Prize Data Warehouse View of the Institution Value Members Mart with the Institution Value Members Fact Table and the Country, Institution Type, and Institution Dimension Tables

The *Institution Value Members Mart* has the Fact table, *FactInstitutionValueMembers*, and 3 Dimensions: *DimCountry, DimInstitution,* and *DimInstitutionType*. These Dimensions are obtained by the Surrogate Keys they share with the Fact table. The numerical variable, *institution_USD_value_in_millions,* measures the monetary value of the institution, and the *institution_member_count*, identifies the number of registered members under that institution.

Those two numerical variables are checked because they are needed to make numerical data analysis. However the Surrogate Keys are unchecked because they only serve the purpose of linking the Fact table with the Dimension tables, so they are not needed in the end result of the Mart afterward. The other variables in the Dimensions besides the Surrogate Keys are checked because they will be used to create deeper-level data analysis. The other variables in the Dimensions besides the Surrogate Keys are checked because they will be used to create deeper-level data analysis. Take note that some variables that are not Surrogate Keys in the Dimensions are unchecked because those variables either already exist in the Fact or other Dimensions.

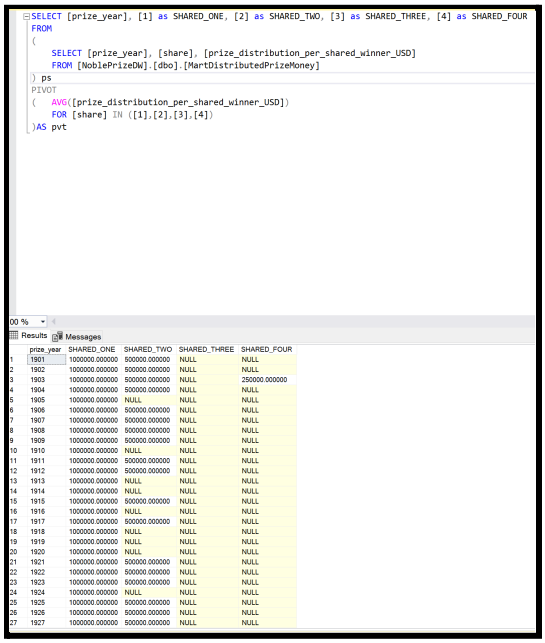


4.4 Nobel Prize Data Warehouse View of the Distribution Prize Money Mart with the Distributed Prize Money Fact Table and the Laureate, Nobel Prize Categories, and Prize Dimension Tables

The *Distributed Prize Money Mart* has the Fact table, *FactDistributedPrizeMoney*, and 3 Dimensions: *DimNoblePrizeCategory, DimLaureate,* and *DimPrize*. These Dimensions are obtained by the Surrogate Keys they share with the Fact table.

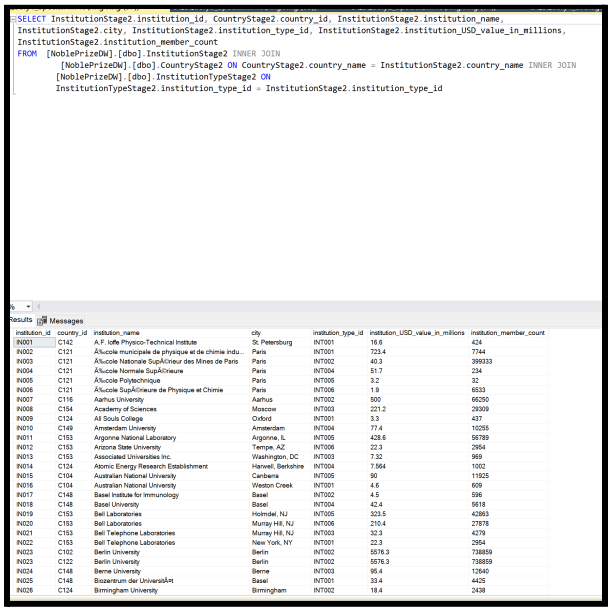
The numerical variable, *prize_distribution_per_shared_winner_USD,* measure the divided monetary gain of the Nobel Prize per co-Laureate. The *share* variable identifies how many co-Laureates are sharing the same Nobel prize for that specific year and Noble Prize category, and the *prize_year* identifies what year the Nobel prize was won at. Those three numerical variables are checked because they are needed to make numerical data analysis. However the Surrogate Keys are unchecked because they only serve the purpose of linking the Fact table with the Dimension tables, so they are not needed in the end result of the Mart afterward. The other variables in the Dimensions besides the Surrogate Keys are checked because they will be used to create deeper-level data analysis. Take note that some variables that are not Surrogate Keys in the Dimensions are unchecked because those variables either already exist in the Fact or other Dimensions.

5 The Basic SQL Queries and The Data Warehouse Tableau Visualizations



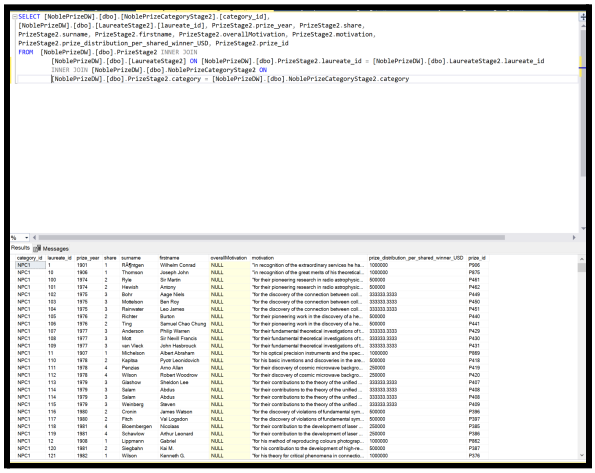
```
SELECT [prize_year], [1] as SHARED_ONE, [2] as SHARED_TWO, [3] as SHARED_THREE, [4] as SHARED_FOUR
 FROM
 (
     SELECT [prize_year], [share], [prize_distribution_per_shared_winner_USD]
     FROM [NoblePrizeDW].[dbo].[MartDistributedPrizeMoney]
 ) ps
 PIVOT
 (   AVG([prize_distribution_per_shared_winner_USD])
     FOR [share] IN ([1],[2],[3],[4])
 )AS pvt
```

prize_year	SHARED_ONE	SHARED_TWO	SHARED_THREE	SHARED_FOUR
1901	1000000.000000	500000.000000	NULL	NULL
1902	1000000.000000	500000.000000	NULL	NULL
1903	1000000.000000	500000.000000	NULL	250000.000000
1904	1000000.000000	500000.000000	NULL	NULL
1905	1000000.000000	NULL	NULL	NULL
1906	1000000.000000	500000.000000	NULL	NULL
1907	1000000.000000	500000.000000	NULL	NULL
1908	1000000.000000	500000.000000	NULL	NULL
1909	1000000.000000	500000.000000	NULL	NULL
1910	1000000.000000	NULL	NULL	NULL
1911	1000000.000000	500000.000000	NULL	NULL
1912	1000000.000000	500000.000000	NULL	NULL
1913	1000000.000000	NULL	NULL	NULL
1914	1000000.000000	NULL	NULL	NULL
1915	1000000.000000	500000.000000	NULL	NULL
1916	1000000.000000	NULL	NULL	NULL
1917	1000000.000000	500000.000000	NULL	NULL
1918	1000000.000000	NULL	NULL	NULL
1919	1000000.000000	NULL	NULL	NULL
1920	1000000.000000	NULL	NULL	NULL
1921	1000000.000000	500000.000000	NULL	NULL
1922	1000000.000000	500000.000000	NULL	NULL
1923	1000000.000000	500000.000000	NULL	NULL
1924	1000000.000000	NULL	NULL	NULL
1925	1000000.000000	500000.000000	NULL	NULL
1926	1000000.000000	500000.000000	NULL	NULL
1927	1000000.000000	500000.000000	NULL	NULL



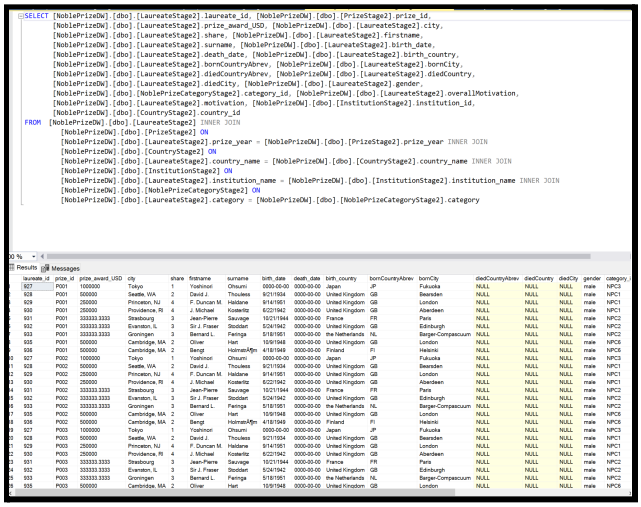
```
SELECT InstitutionStage2.institution_id, CountryStage2.country_id, InstitutionStage2.institution_name,
InstitutionStage2.city, InstitutionStage2.institution_type_id, InstitutionStage2.institution_USD_value_in_millions,
InstitutionStage2.institution_member_count
FROM  [NoblePrizeDW].[dbo].[InstitutionStage2 INNER JOIN
        [NoblePrizeDW].[dbo].[CountryStage2 ON CountryStage2.country_name = InstitutionStage2.country_name INNER JOIN
      [NoblePrizeDW].[dbo].[InstitutionTypeStage2 ON
      InstitutionTypeStage2.institution_type_id = InstitutionStage2.institution_type_id
```

institution_id	country_id	institution_name	city	institution_type_id	institution_USD_value_in_millions	institution_member_count
IN001	C142	A.F. Ioffe Physico-Technical Institute	St. Petersburg	INT001	16.6	424
IN002	C121	Ã‰cole municipale de physique et de chimie indu...	Paris	INT001	723.4	7744
IN003	C121	Ã‰cole Nationale SupÃ©rieur des Mines de Paris	Paris	INT002	40.3	399333
IN004	C121	Ã‰cole Normale SupÃ©rieure	Paris	INT004	51.7	234
IN005	C121	Ã‰cole Polytechnique	Paris	INT005	3.2	32
IN006	C121	Ã‰cole SupÃ©rieure de Physique et Chimie	Paris	INT006	1.9	6533
IN007	C116	Aarhus University	Aarhus	INT002	500	66250
IN008	C154	Academy of Sciences	Moscow	INT003	221.2	29309
IN009	C124	All Souls College	Oxford	INT001	3.3	437
IN010	C149	Amsterdam University	Amsterdam	INT004	77.4	10255
IN011	C153	Argonne National Laboratory	Argonne, IL	INT005	428.6	56789
IN012	C153	Arizona State University	Tempe, AZ	INT006	22.3	2954
IN013	C153	Associated Universities Inc.	Washington, DC	INT003	7.32	969
IN014	C124	Atomic Energy Research Establishment	Harwell, Berkshire	INT004	7.564	1002
IN015	C104	Australian National University	Canberra	INT005	90	11925
IN016	C104	Australian National University	Weston Creek	INT001	4.6	609
IN017	C148	Basel Institute for Immunology	Basel	INT002	4.5	596
IN018	C148	Basel University	Basel	INT004	42.4	5618
IN019	C153	Bell Laboratories	Holmdel, NJ	INT005	323.5	42863
IN020	C153	Bell Laboratories	Murray Hill, NJ	INT006	210.4	27878
IN021	C153	Bell Telephone Laboratories	Murray Hill, NJ	INT003	32.3	4279
IN022	C153	Bell Telephone Laboratories	New York, NY	INT001	22.3	2954
IN023	C102	Berlin University	Berlin	INT002	5576.3	738859
IN023	C122	Berlin University	Berlin	INT002	5576.3	738859
IN024	C148	Berne University	Berne	INT003	95.4	12640
IN025	C148	Biozentrum der UniversitÃ¤t	Basel	INT001	33.4	4425
IN026	C124	Birmingham University	Birmingham	INT002	18.4	2438



```
SELECT [NoblePrizeDW].[dbo].[NoblePrizeCategoryStage2].[category_id],
[NoblePrizeDW].[dbo].[LaureateStage2].[laureate_id], PrizeStage2.prize_year, PrizeStage2.share,
PrizeStage2.surname, PrizeStage2.firstname, PrizeStage2.overallMotivation, PrizeStage2.motivation,
PrizeStage2.prize_distribution_per_shared_winner_USD, PrizeStage2.prize_id
FROM  [NoblePrizeDW].[dbo].PrizeStage2 INNER JOIN
      [NoblePrizeDW].[dbo].[LaureateStage2] ON [NoblePrizeDW].[dbo].PrizeStage2.laureate_id = [NoblePrizeDW].[dbo].laureateStage2.laureate_id
      INNER JOIN [NoblePrizeDW].[dbo].NoblePrizeCategoryStage2 ON
      [NoblePrizeDW].[dbo].PrizeStage2.category = [NoblePrizeDW].[dbo].NoblePrizeCategoryStage2.category
```

category_id	laureate_id	prize_year	share	surname	firstname	overallMotivation	motivation	prize_distribution_per_shared_winner_USD	prize_id
NPC1	1	1901	1	RÃ¶ntgen	Wilhelm Conrad	NULL	"in recognition of the extraordinary services he ha...	1000000	P906
NPC1	15	1906	1	Thomson	Joseph John	NULL	"in recognition of the great merits of his theoretical...	1000000	P976
NPC1	100	1974	2	Ryle	Sir Martin	NULL	"for their pioneering research in radio astrophysics:...	500000	P481
NPC1	101	1974	2	Hewish	Antony	NULL	"for their pioneering research in radio astrophysics:...	500000	P482
NPC1	102	1975	3	Bohr	Aage Niels	NULL	"for the discovery of the connection between collect...	333333.3333	P449
NPC1	103	1975	3	Mottelson	Ben Roy	NULL	"for the discovery of the connection between collect...	333333.3333	P450
NPC1	104	1975	3	Rainwater	Leo James	NULL	"for the discovery of the connection between collect...	333333.3333	P451
NPC1	105	1976	2	Richter	Burton	NULL	"for their pioneering work in the discovery of a he...	500000	P440
NPC1	106	1976	2	Ting	Samuel Chao Chung	NULL	"for their pioneering work in the discovery of a he...	500000	P441
NPC1	107	1977	3	Anderson	Philip Warren	NULL	"for their fundamental theoretical investigations of t...	333333.3333	P429
NPC1	108	1977	3	Mott	Sir Nevill Francis	NULL	"for their fundamental theoretical investigations of t...	333333.3333	P430
NPC1	109	1977	3	van Vleck	John Hasbrouck	NULL	"for their fundamental theoretical investigations of t...	333333.3333	P431
NPC1	11	1907	1	Michelson	Albert Abraham	NULL	"for his optical precision instruments and the spec...	1000000	P889
NPC1	110	1978	2	Kapitsa	Pyotr Leonidovich	NULL	"for his basic inventions and discoveries in the are...	500000	P418
NPC1	111	1978	4	Penzias	Arno Allan	NULL	"for their discovery of cosmic microwave backgro...	250000	P419
NPC1	112	1978	4	Wilson	Robert Woodrow	NULL	"for their discovery of cosmic microwave backgro...	250000	P420
NPC1	113	1979	3	Glashow	Sheldon Lee	NULL	"for their contributions to the theory of the unified...	333333.3333	P407
NPC1	114	1979	3	Salam	Abdus	NULL	"for their contributions to the theory of the unified...	333333.3333	P408
NPC1	114	1979	3	Salam	Abdus	NULL	"for their contributions to the theory of the unified...	333333.3333	P408
NPC1	115	1979	3	Weinberg	Steven	NULL	"for their contributions to the theory of the unified...	333333.3333	P409
NPC1	116	1980	2	Cronin	James Watson	NULL	"for the discovery of violations of fundamental sym...	500000	P396
NPC1	117	1980	2	Fitch	Val Logsdon	NULL	"for the discovery of violations of fundamental sym...	500000	P397
NPC1	118	1981	4	Bloembergen	Nicolaas	NULL	"for their contribution to the development of laser...	250000	P385
NPC1	119	1981	4	Schawlow	Arthur Leonard	NULL	"for their contribution to the development of laser...	250000	P386
NPC1	12	1908	1	Lippmann	Gabriel	NULL	"for his method of reproducing colours photograp...	1000000	P882
NPC1	120	1981	2	Siegbahn	Kai M.	NULL	"for his contribution to the development of high-re...	500000	P347
NPC1	121	1982	1	Wilson	Kenneth G.	NULL	"for his theory for critical phenomena in connectio...	1000000	P376

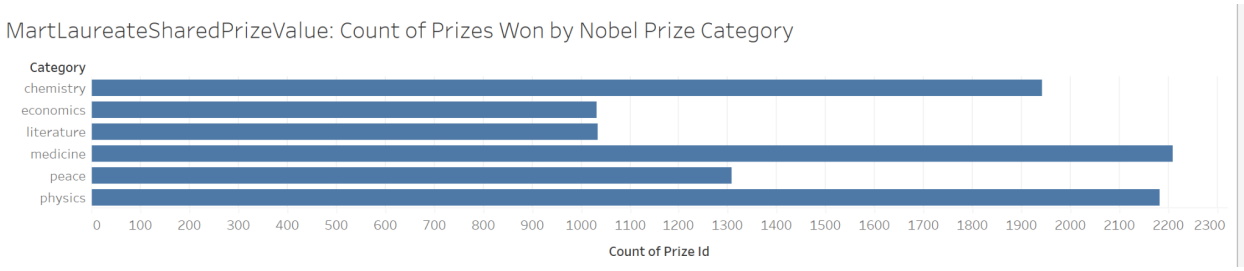


```
SELECT [NoblePrizeDW].[dbo].[LaureateStage2].laureate_id, [NoblePrizeDW].[dbo].[PrizeStage2].prize_id,
    [NoblePrizeDW].[dbo].[LaureateStage2].prize_award_USD, [NoblePrizeDW].[dbo].[LaureateStage2].city,
    [NoblePrizeDW].[dbo].[LaureateStage2].share, [NoblePrizeDW].[dbo].[LaureateStage2].firstname,
    [NoblePrizeDW].[dbo].[LaureateStage2].surname, [NoblePrizeDW].[dbo].[LaureateStage2].birth_date,
    [NoblePrizeDW].[dbo].[LaureateStage2].death_date, [NoblePrizeDW].[dbo].[LaureateStage2].birth_country,
    [NoblePrizeDW].[dbo].[LaureateStage2].bornCountryAbrev, [NoblePrizeDW].[dbo].[LaureateStage2].bornCity,
    [NoblePrizeDW].[dbo].[LaureateStage2].diedCountryAbrev, [NoblePrizeDW].[dbo].[LaureateStage2].diedCountry,
    [NoblePrizeDW].[dbo].[LaureateStage2].diedCity, [NoblePrizeDW].[dbo].[LaureateStage2].gender,
    [NoblePrizeDW].[dbo].[NoblePrizeCategoryStage2].category_id, [NoblePrizeDW].[dbo].[LaureateStage2].overallMotivation,
    [NoblePrizeDW].[dbo].[LaureateStage2].motivation, [NoblePrizeDW].[dbo].[InstitutionStage2].institution_id,
    [NoblePrizeDW].[dbo].[CountryStage2].country_id
FROM  [NoblePrizeDW].[dbo].[LaureateStage2] INNER JOIN
      [NoblePrizeDW].[dbo].[PrizeStage2] ON
      [NoblePrizeDW].[dbo].[LaureateStage2].prize_year = [NoblePrizeDW].[dbo].[PrizeStage2].prize_year INNER JOIN
      [NoblePrizeDW].[dbo].[CountryStage2] ON
      [NoblePrizeDW].[dbo].[LaureateStage2].country_name = [NoblePrizeDW].[dbo].[CountryStage2].country_name INNER JOIN
      [NoblePrizeDW].[dbo].[InstitutionStage2] ON
      [NoblePrizeDW].[dbo].[LaureateStage2].institution_name = [NoblePrizeDW].[dbo].[InstitutionStage2].institution_name INNER JOIN
      [NoblePrizeDW].[dbo].[NoblePrizeCategoryStage2] ON
      [NoblePrizeDW].[dbo].[LaureateStage2].category = [NoblePrizeDW].[dbo].[NoblePrizeCategoryStage2].category
```

laureate_id	prize_id	prize_award_USD	city	share	firstname	surname	birth_date	death_date	birth_country	bornCountryAbrev	bornCity	diedCountryAbrev	diedCountry	diedCity	gender	category_i
927	P001	1000000	Tokyo	1	Yoshinori	Ohsumi	0000-00-00	0000-00-00	Japan	JP	Fukuoka	NULL	NULL	NULL	male	NPC3
928	P001	500000	Seattle, WA	2	David J.	Thouless	9/21/1934	0000-00-00	United Kingdom	GB	Bearsden	NULL	NULL	NULL	male	NPC1
929	P001	250000	Princeton, NJ	4	F. Duncan M.	Haldane	9/14/1951	0000-00-00	United Kingdom	GB	London	NULL	NULL	NULL	male	NPC1
930	P001	250000	Providence, RI	4	J. Michael	Kosterlitz	6/22/1942	0000-00-00	United Kingdom	GB	Aberdeen	NULL	NULL	NULL	male	NPC1
931	P001	333333.3333	Strasbourg	3	Jean-Pierre	Sauvage	10/21/1944	0000-00-00	France	FR	Paris	NULL	NULL	NULL	male	NPC2
932	P001	333333.3333	Evanston, IL	3	Sir J. Fraser	Stoddart	5/24/1942	0000-00-00	United Kingdom	GB	Edinburgh	NULL	NULL	NULL	male	NPC2
933	P001	333333.3333	Groningen	3	Bernard L.	Feringa	5/18/1951	0000-00-00	the Netherlands	NL	Barger-Compascuum	NULL	NULL	NULL	male	NPC2
935	P001	500000	Cambridge, MA	2	Oliver	Hart	10/9/1948	0000-00-00	United Kingdom	GB	London	NULL	NULL	NULL	male	NPC6
936	P001	500000	Cambridge, MA	2	Bengt	HolmstÃ¶m	4/18/1948	0000-00-00	Finland	FI	Helsinki	NULL	NULL	NULL	male	NPC6
927	P002	1000000	Tokyo	1	Yoshinori	Ohsumi	0000-00-00	0000-00-00	Japan	JP	Fukuoka	NULL	NULL	NULL	male	NPC3
928	P002	500000	Seattle, WA	2	David J.	Thouless	9/21/1934	0000-00-00	United Kingdom	GB	Bearsden	NULL	NULL	NULL	male	NPC1
929	P002	250000	Princeton, NJ	4	F. Duncan M.	Haldane	9/14/1951	0000-00-00	United Kingdom	GB	London	NULL	NULL	NULL	male	NPC1
930	P002	250000	Providence, RI	4	J. Michael	Kosterlitz	6/22/1942	0000-00-00	United Kingdom	GB	Aberdeen	NULL	NULL	NULL	male	NPC1
931	P002	333333.3333	Strasbourg	3	Jean-Pierre	Sauvage	10/21/1944	0000-00-00	France	FR	Paris	NULL	NULL	NULL	male	NPC2
932	P002	333333.3333	Evanston, IL	3	Sir J. Fraser	Stoddart	5/24/1942	0000-00-00	United Kingdom	GB	Edinburgh	NULL	NULL	NULL	male	NPC2
933	P002	333333.3333	Groningen	3	Bernard L.	Feringa	5/18/1951	0000-00-00	the Netherlands	NL	Barger-Compascuum	NULL	NULL	NULL	male	NPC2
935	P002	500000	Cambridge, MA	2	Oliver	Hart	10/9/1948	0000-00-00	United Kingdom	GB	London	NULL	NULL	NULL	male	NPC6
936	P002	500000	Cambridge, MA	2	Bengt	HolmstÃ¶m	4/18/1948	0000-00-00	Finland	FI	Helsinki	NULL	NULL	NULL	male	NPC6
927	P003	1000000	Tokyo	1	Yoshinori	Ohsumi	0000-00-00	0000-00-00	Japan	JP	Fukuoka	NULL	NULL	NULL	male	NPC3
928	P003	500000	Seattle, WA	2	David J.	Thouless	9/21/1934	0000-00-00	United Kingdom	GB	Bearsden	NULL	NULL	NULL	male	NPC1
929	P003	250000	Princeton, NJ	4	F. Duncan M.	Haldane	9/14/1951	0000-00-00	United Kingdom	GB	London	NULL	NULL	NULL	male	NPC1
930	P003	250000	Providence, RI	4	J. Michael	Kosterlitz	6/22/1942	0000-00-00	United Kingdom	GB	Aberdeen	NULL	NULL	NULL	male	NPC1
931	P003	333333.3333	Strasbourg	3	Jean-Pierre	Sauvage	10/21/1944	0000-00-00	France	FR	Paris	NULL	NULL	NULL	male	NPC2
932	P003	333333.3333	Evanston, IL	3	Sir J. Fraser	Stoddart	5/24/1942	0000-00-00	United Kingdom	GB	Edinburgh	NULL	NULL	NULL	male	NPC2
933	P003	333333.3333	Groningen	3	Bernard L.	Feringa	5/18/1951	0000-00-00	the Netherlands	NL	Barger-Compascuum	NULL	NULL	NULL	male	NPC2
935	P003	500000	Cambridge, MA	2	Oliver	Hart	10/9/1948	0000-00-00	United Kingdom	GB	London	NULL	NULL	NULL	male	NPC6

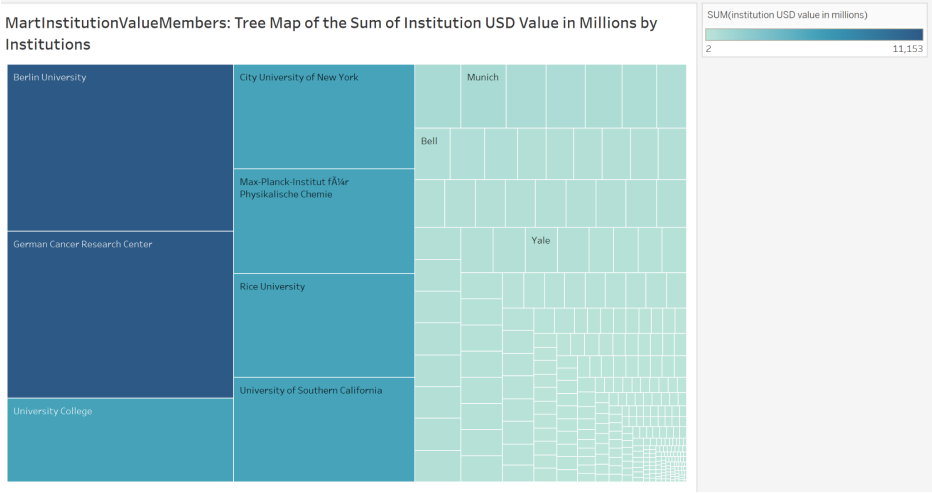
5.1.1 Basic SQL Queries - From Top Left to the Bottom Right: Average Prize Money Per Laureate Based On How Many Laureates They Share The Same Nobel Prize, SpecialInstitutionFactStage1, SpecialPrizeFactStage1, and SpecialLaureateFactStage1

The 1st query on the top left illustrates a cross-tab query that calculates the average prize monetary distribution per laureate based on how many laureates they co-authored the same Nobel Prize. The last three SQL queries were used in creating the *SpecialInstitutionFactStage1, SpecialPrizeFactStage1, and SpecialLaureateFactStage1,* which is shown in the previous sections. The basic SQL Queries show that the Data Warehouse creation was successful and functions well.



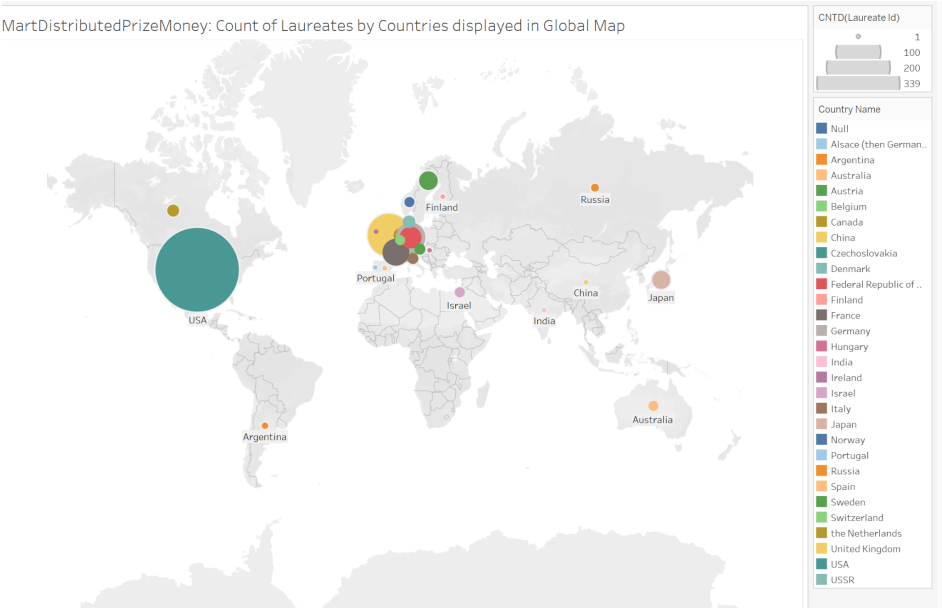
5.1.2 Tableau: MartLaureateSharedPrizeValue - Count of Nobel Prizes Won by Nobel Prize Category

The Horizontal Bar Chart shown above derives data from the *Laureate Shared prize Value Mart,* and it shows the number of Nobel Prizes by the 6 Nobel categories. The diagram shows that the *Medicine* and *Physics* Nobel Prize Categories have the highest Nobel Prize frequency of 2,209 and 2,183 compared to the other categories since the founding of the Noble Foundation. The *Economics* and the *Literature* Nobel Prize categories have the lowest Nobel Prize frequency of 1,031 and 1,034. This diagram implies that *Medicine* and *Physics* have more emphasis and importance to the Nobel Foundation compared to the other 4 Nobel Prize categories.



5.1.3 Tableau: MartInstitutionValueMembers - TreeMap of the Sum of Institution USD Value in Millions by Institutions

The TreeMap shown above displays the Institutions' monetary value in USD in Millions by established Institutions that were involved with Laureates who won Nobel Prizes, which the Institutions' USD Value in Millions is illustrated by the darkness of the parallelogram's color and size. As you can see, the Berlin University and German Cancer Research have the highest institution values of both 11,153 million USD. These two institutions are the richest institutions involved with Laureates who won the Nobel Prize compared to the other institutions which are represented by lightly colored, small parallelograms.



5.1.4 Tableau: MartDistributedPrizeMoney - Count of Laureates by Countries displayed in Global Map

The Map above shows the number of Laureates who won the Nobel Prize by Countries, which the size of the circles is represented by the count of laureates, and the color of the circles represents the countries involved with those laureates. As you can see, the USA has the highest Laureate frequency of 339 compared to any other country. The runner-up is the United Kingdom with only 88 laureates. This indicates that the USA has a higher success rate in winning the Nobel Prize compared to any other country.

6 Conclusion

This data warehouse was created to gain insight into the Noble Foundation and its laureates. Before anything was performed, the data warehouse had to be planned and data needed to be profiled. Documentation with the BEAM templates helped establish the events that were going on and the necessary fact and dimension tables. Further description and planning of the data

warehouse were accomplished by determining data types, creating column constraints, and mapping surrogate keys. The ETL implementation was achieved using SSIS and splitting the process into 5 sections: loading stage 1, loading stage 2, loading dimensions, and loading facts. Meanwhile, views and queries of the marts were created using SQL Server Management Studio. The fully functional data warehouse could finally be used for the insight it was intended for. Using Tableau, data visualizations were produced that allowed end-users to understand the data being presented. For this project three data visualizations were produced: Count of Prizes won by Nobel Prize Category, Value of the Institutions in Millions, and Count of Laureates by Country. Although these are the only visualizations provided, the Nobel Prize data warehouse is not limited to them and can provide insight for other possible questions.

The ETL Process implementation was the most difficult step throughout the project, since it required a lot of editing to fix the constant data overflow crashes that occured in the code. However, the easiest step would have to be the Tableau Data Visualizations since Tableau is a very user-friendly program compared to most data visualization programs. The project taught us that even with massive preparation with the BEAM Matrix and the BEAM modelstorming, there will be trials that our group was not prepared to handle like the excessive amount of data overflow crashes that occur in Visual Studio which cost so much time to fix. If we had the opportunity to do it again, we would have researched more on how to avoid crashes in our Visual Studio code, so we won't lose as much production as we had.

The new system our group created extends the original dataset from the Harvard Dataverse as a Nobel Prize Data Warehouse with 5 more CSV files made compared to the original dataset of 3 CSV files. This Data Warehouse provides a more in-depth analysis of the Nobel Prize data compared to the original research and can further provide better analysis of the Nobel Prize and

the Nobel Foundation as a whole. The Noble Data Warehouse can also be used by the Noble Foundation to organize their institution's big data.

7 References

Kuzmenko, Maryna. (2016). *Nobel Prize - Dataset with Information about Prizes, Laureates and Countries*. Harvard Dataverse, V1, UNF:6:McdDh+ldUTGgZDs5XVOQUA== [fileUNF]. https://doi.org/10.7910/DVN/AGAFAQ

The Nobel Foundation. (2022, April 13). *The Nobel Prize*. NobelPrize.Org. https://www.nobelprize.org/