A project using APIs, ETL, Application Integration, Application Lifecycle Management, & Data Analysis

Award Data Project

AN AUTOMATED DATA SOLUTION

Andria Emmons

SCENARIO

One might want to know:

- How much money was awarded for all State of Texas agencies for 2020, 2021, and 2022
- How much of the awarded money went to Historically Underutilized Businesses (HUB)
- How much went to non-HUB.

Details about the HUB program can be found here:
https://comptroller.texas.gov/purchasing/vendor/hub/

PROBLEM

- At the Texas Open Data portal, data can be exported and visualized, but only one dataset at a time.
- Each year is a separate dataset.

SOLUTION

1. Extract the data using an API.
2. Load and Transform using Power BI
3. Create relationships to link the same vendor data together over multiple years.
4. Automate the process and data refreshes.

PROCESS

"The solution architecture and design process phase (see Figure 6.7) determines how to best design the solution, choosing the depth of the project, the schedule, hardware, and software to best satisfy both project and business requirements. This is also where the team determines how the integration (hardware and/or software) of any required physical security subsystem into a single interface can improve efficiencies and decisions and reduce cost and risk. An architectural team is assigned, with architects, subject matter experts, and engineers from various disciplines, hardware and software vendors, and the customer" (Caputo, 2014)

"As previously mentioned, there should be a lead architect—an individual who understands all the pieces of the puzzle to some capacity and who can listen to the wealth of information available from the team…" (Caputo, 2014)

Since most projects consist of more than just one team, understanding different processes is essential, when working in solo project or very small teams. The following Processes were used in this project.

SDLC – PROCESS

"The stages of the software development life cycle are which describe how to develop and maintain software. Each phase has different processes and activities." (Clarusway, 2021)

1. Planning & Analysis
2. Design
3. Development

4. Testing
5. Deployment
6. Maintenance

KANBAN PRINCIPLES

"*Kanban* is a Japanese term that means signboard or billboard. An industrial engineer named Taiichi Ohno developed Kanban at Toyota Motor Corporation to improve manufacturing efficiency." (Microsoft, 2022)

1. In Queue
2. In Progress
3. Completed

DATA ENGINEERING PROCESS

1. Collect the data
2. Cleanse the data
3. Transform the data
4. Process the data
5. Monitor data refreshes

WORK

EXTRACTING

Three data sets will be the focus of the current project.

Current Project

1. 2020 SRWDBTSSBF Post Award Summary
2. 2021 SRWDBTSSBF Post Award Summary
3. 2022 SRWDBTSSBF Post Award Summary

TESTING DATA VIA API

After extracting the API data, it was evaluated, at Postman. A collection folder was created, Award Summary. The data was added and evaluated to for accuracy. The test was successful.

LOADING

Loaded the data into Power BI to transform and build the reporting visualizations.

1. Choose "Get Data" and then" Blank query".
2. A pop-up window opens in the query editor.
3. Choose "Advanced Editor: and write the code to load the data.
4. Select "Anonymous" for the permissions.

TRANSFORMING DATA

To transform the data.

1. Under List Tools
2. Select convert the list "To Table."
3. When a pop-up appears, use the default settings.

4. Click the double arrows to the right of Column1 to expand the data
5. Deselect “Use original column name as prefix.”
6. Then select “OK.”
7. Rename the column headers to something more appropriate for business use.
8. Update the data types as needed.

TROUBLESHOOTING ERRORS

Updating the data types caused an error. I selected the previously applied step to identify the location of the error.

IDENTIFYING AND FIXING THE ERROR

After determining that not all the quantities are numerical values. I created two columns, Quantity1 and Check Quantity.

Criteria for Quantity1

- The first line, if null, stay null
- Line two, if there is a number in the quantity column, use the quantity column
- Line three addresses text and says if the quantity column contains text, then make it (““) blank
- The final line looks for a number, otherwise null.

Criteria for Check Quantity

- Duplicate the Quantity column.
- Split the column digit to non-digit.

EVALUATING NULLS AND IRREGULAR VALUES

Non-Digit Quantity Evaluation

Determine if non-digit values are ready for removal or if they need further transformations. Upon evaluation, it is determined that the validity of the data will not be affected upon removing the non-digit values.

Null Values Evaluation

The Solicitation number will be the Unique ID, so any null values in the Solicitation column are discarded. Before removing, I filtered and viewed them to make sure removing them does not skew the data.

Null Values in the Quantity Column

After reviewing, I have determined that removing the null values in the quantity column will not negatively affect the data. Deleted the columns Quantity and Check Quantity and renamed the Quantity1 column to Quantity.

DATA ANALSYSIS

There are a few ways to combine data through relationships and filtering based on joins, merging the queries, or appending queries. Analyzing the data before combining is essential to ensure that the integrity of the data stays intact. I analyzed the data using the random sample method. It is important when sampling data to get a large enough sample to be statistically sound. I use the 10% or 100 rows method, whichever is greater.

"The code above loads approximately 10% of the content of a CSV file, randomly selecting the rows to read. This method is good enough to preserve the statistical properties of the sample." (Rodríguez, 2022)

PROJECT CONSISTS OF

Project - APIs, ETL, Application Integration, Application Lifecycle Management, & Data Analysis

- 1 Scenario
- 1 SharePoint Communication site
- 1 Microsoft List used as a work in progress tracker.
- Power BI (PBI)
    - 1 PBI Service
    - 1 Power BI desktop application
        - 3 data sets
            - Merged into 1 Query
        - 4 Custom Transformed and Cleaned Queries
        - 2 Pages
- 1 Website

APPENDIX A – PROJECT IMAGES

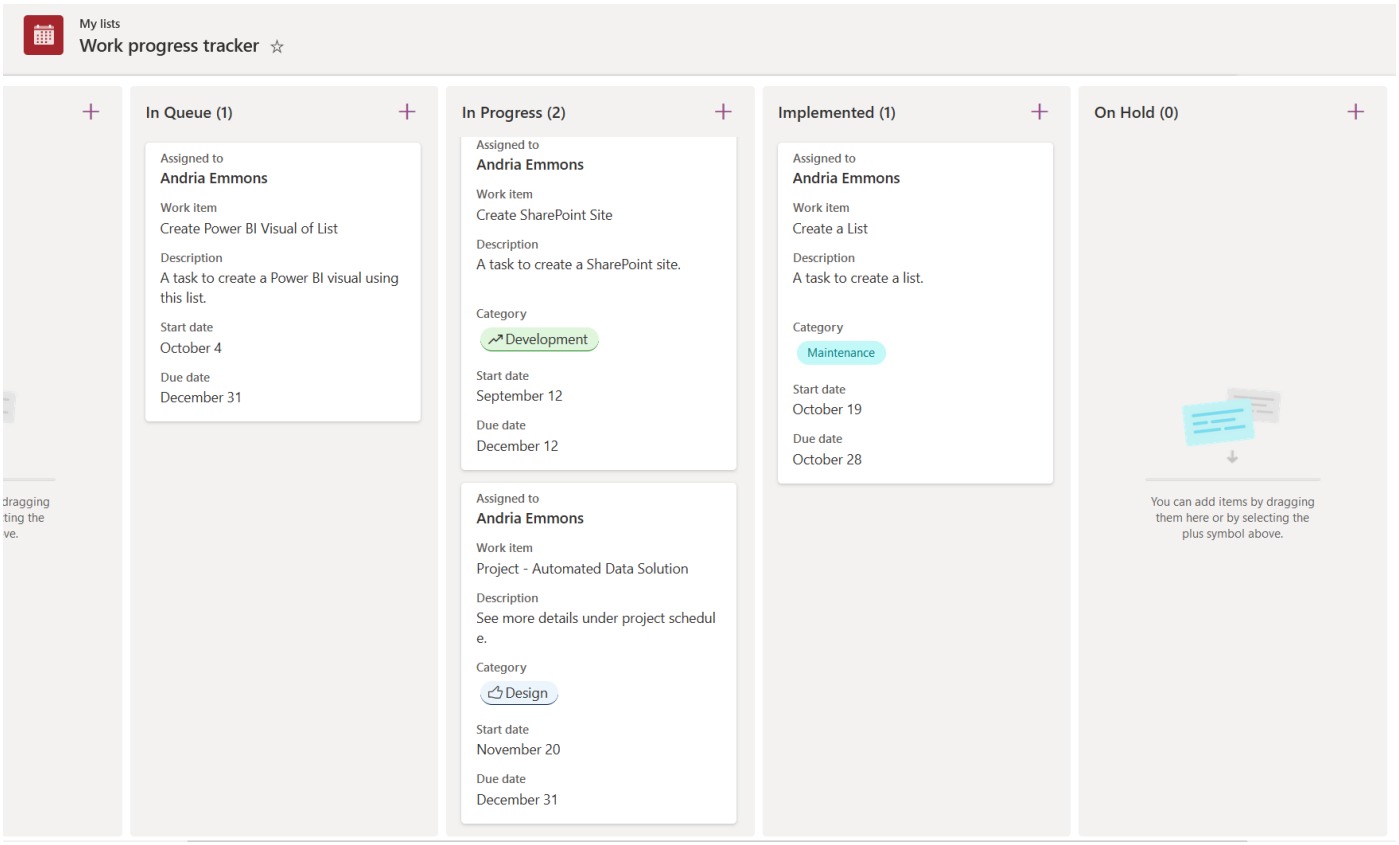


Figure 1 Work In Progress Tracker

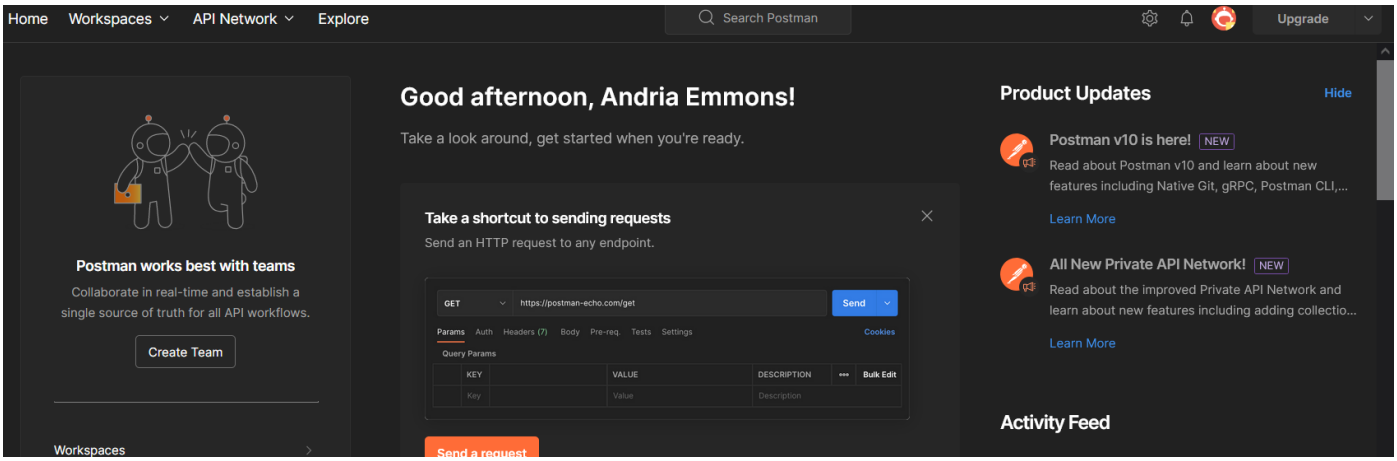


Figure 2 Postman Site

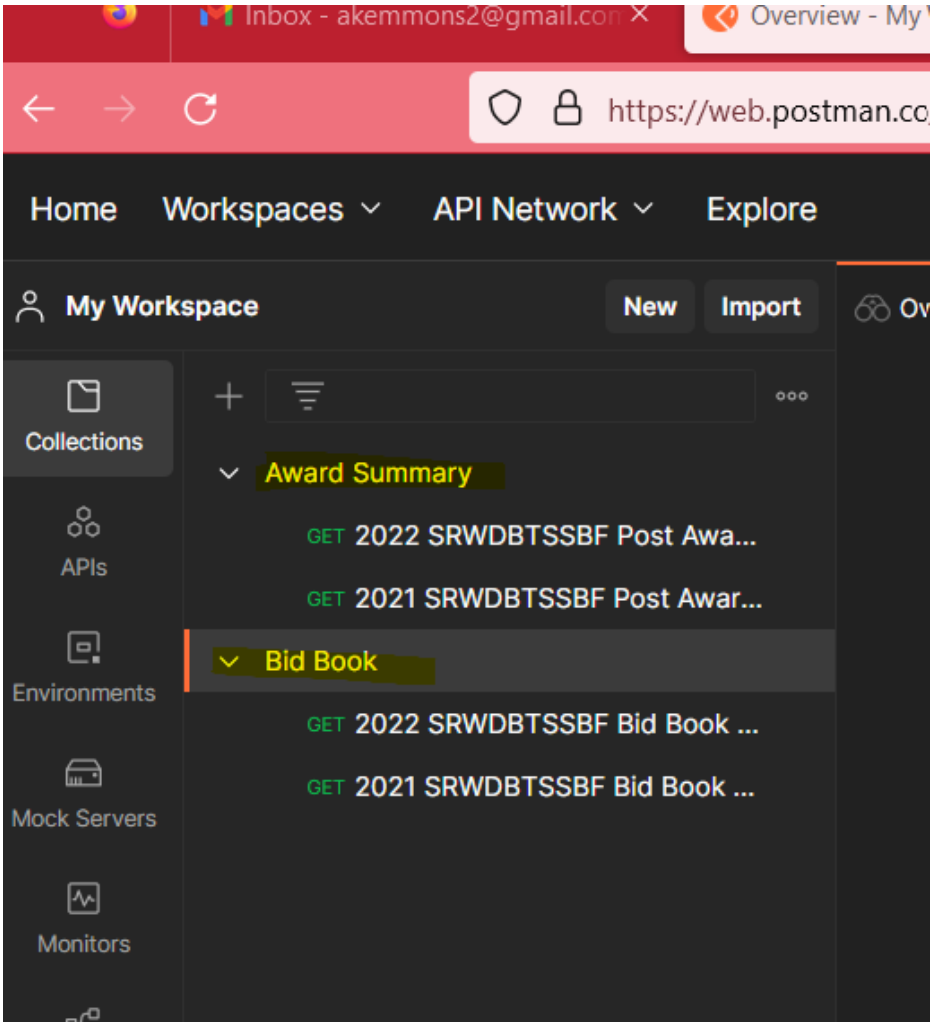


Figure 3 Loading APIs in Postman

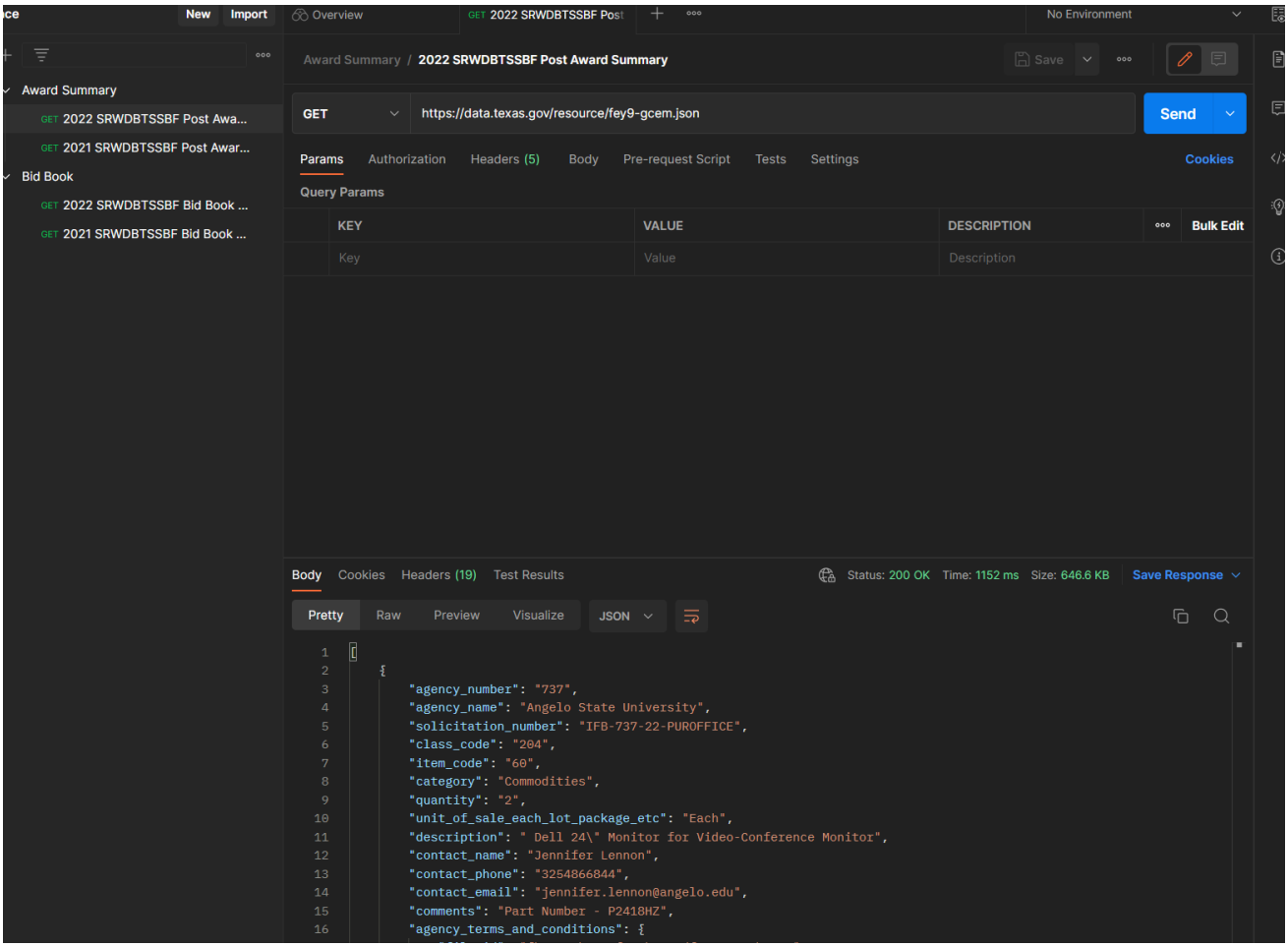


Figure 4 Test Successful

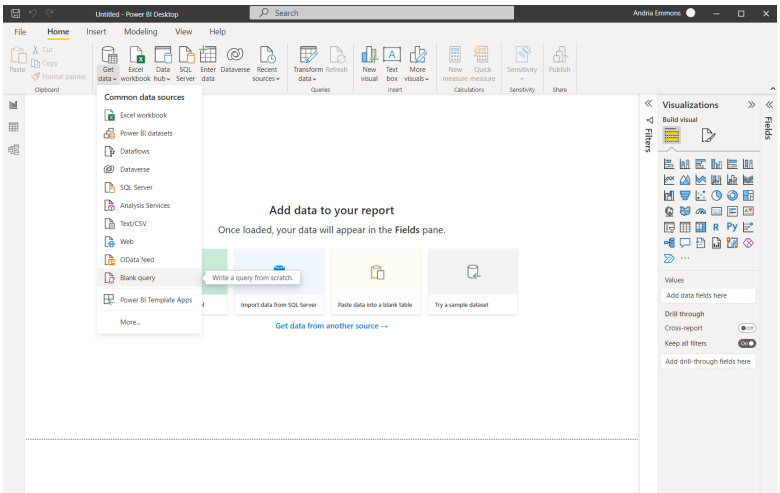


Figure 5 Getting Data

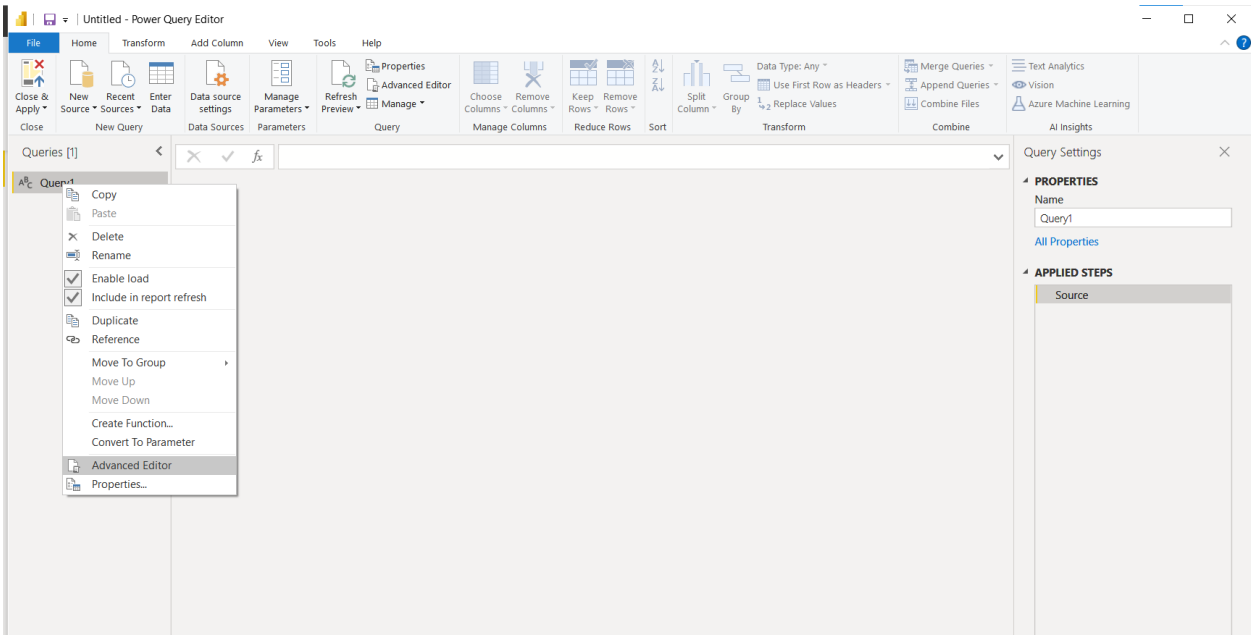


Figure 6 Navigating to Advanced Editor

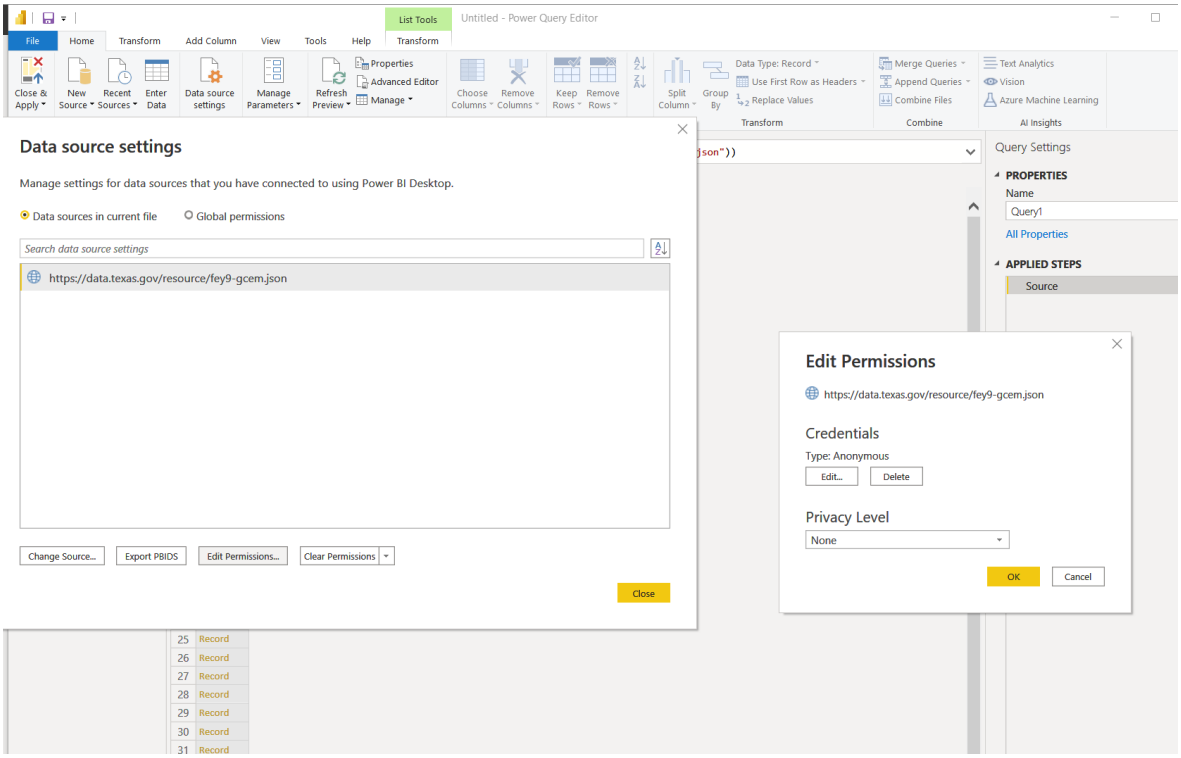


Figure 7 Setting Permissions

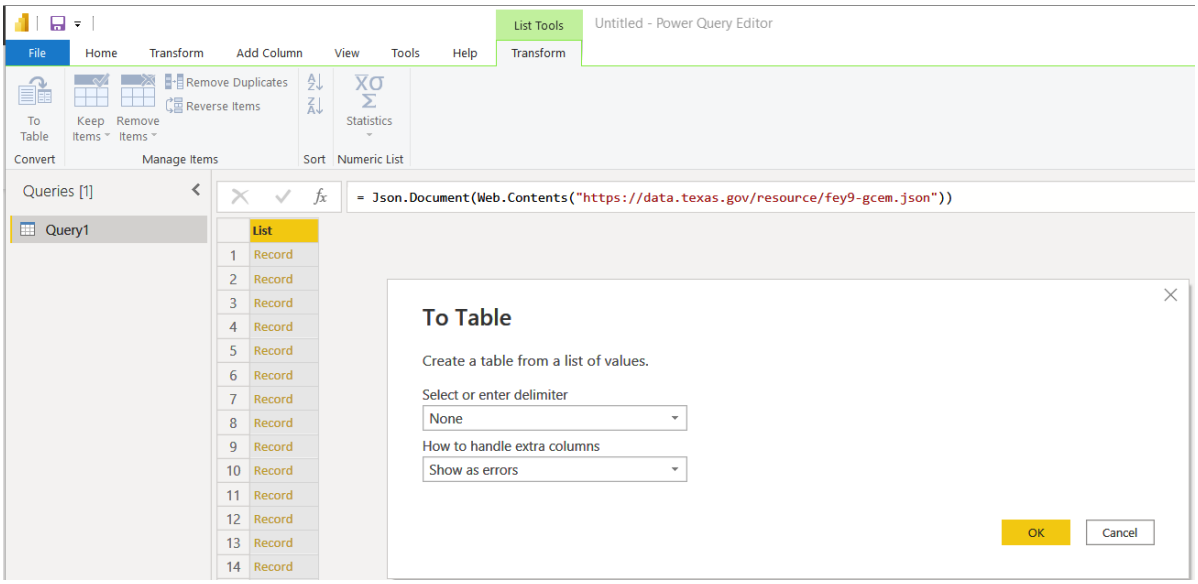


Figure 8 Changing to Table

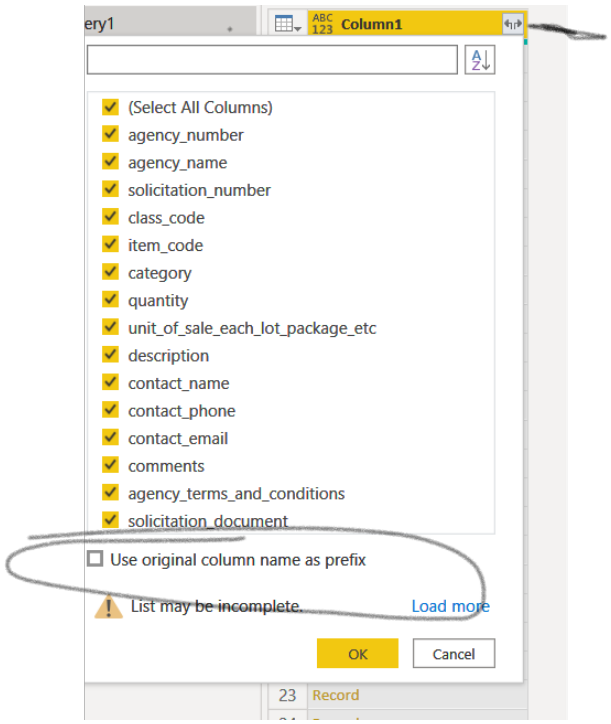
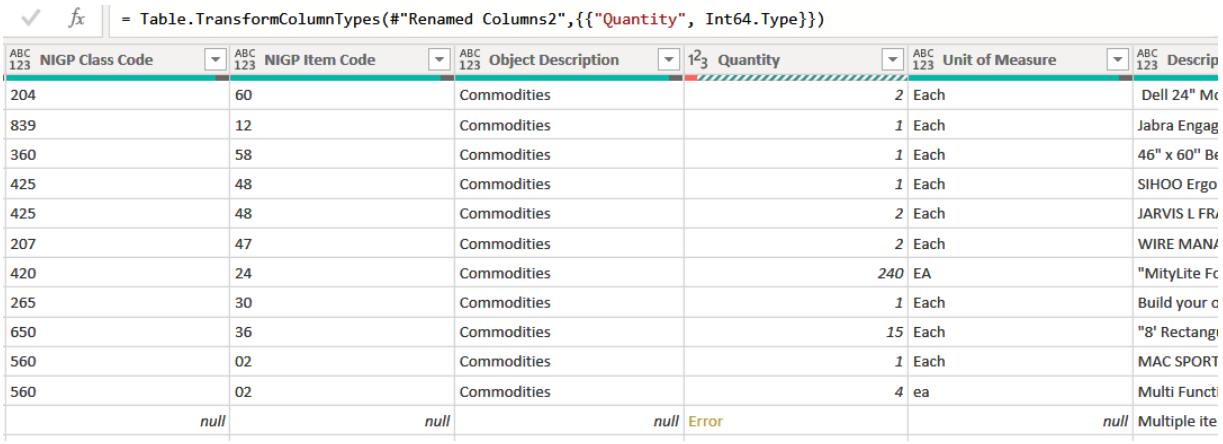


Figure 9 Extracting fields

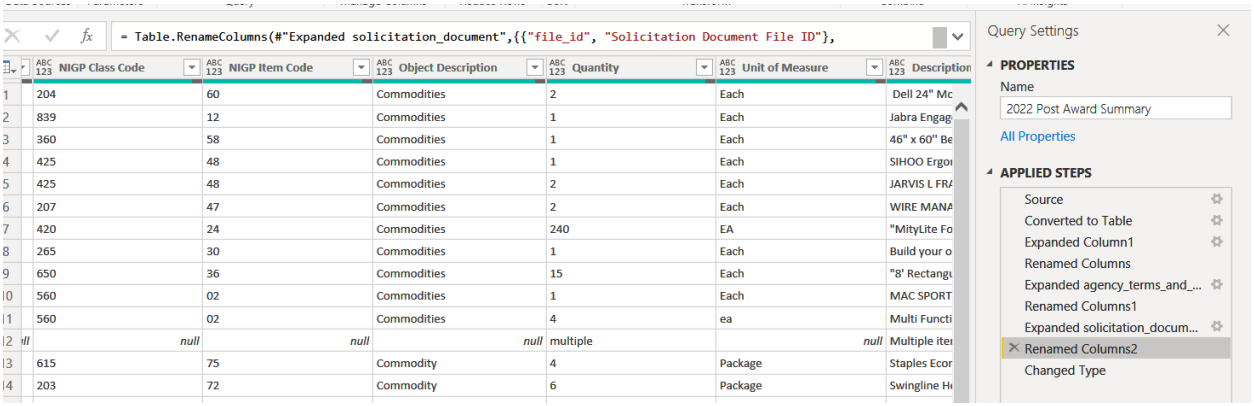
class_code	item_code	category	quantity	unit_of_sale_each_lot_package_etc	
204	60	Commodities	2	Each	De
839	12	Commodities	1	Each	Jab

Figure 10 Analyzing Data (1)



NIGP Class Code	NIGP Item Code	Object Description	Quantity	Unit of Measure	Descrip
204	60	Commodities	2	Each	Dell 24" Mc
839	12	Commodities	1	Each	Jabra Engag
360	58	Commodities	1	Each	46" x 60" Be
425	48	Commodities	1	Each	SIHOO Ergo
425	48	Commodities	2	Each	JARVIS L FR/
207	47	Commodities	2	Each	WIRE MANA
420	24	Commodities	240	EA	"MityLite Fc
265	30	Commodities	1	Each	Build your c
650	36	Commodities	15	Each	"8' Rectangu
560	02	Commodities	1	Each	MAC SPORT
560	02	Commodities	4	ea	Multi Functi
null	null	null	Error	null	Multiple ite

Figure 11 Analyzing Data (2)



	NIGP Class Code	NIGP Item Code	Object Description	Quantity	Unit of Measure	Description
1	204	60	Commodities	2	Each	Dell 24" Mc
2	839	12	Commodities	1	Each	Jabra Engag
3	360	58	Commodities	1	Each	46" x 60" Be
4	425	48	Commodities	1	Each	SIHOO Ergo
5	425	48	Commodities	2	Each	JARVIS L FR/
6	207	47	Commodities	2	Each	WIRE MANA
7	420	24	Commodities	240	EA	"MityLite Fo
8	265	30	Commodities	1	Each	Build your o
9	650	36	Commodities	15	Each	"8' Rectangu
10	560	02	Commodities	1	Each	MAC SPORT
11	560	02	Commodities	4	ea	Multi Functi
12	null	null	null	multiple	null	Multiple iter
13	615	75	Commodity	4	Package	Staples Ecor
14	203	72	Commodity	6	Package	Swingline H

Figure 12 Analyzing Data (3)

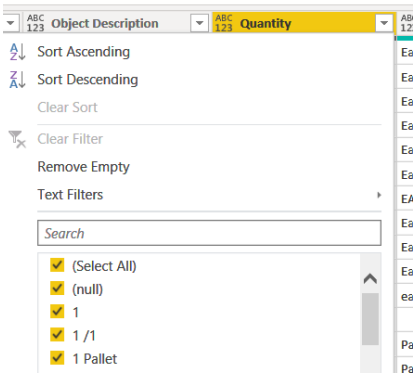


Figure 13 Filtering Quantity Field

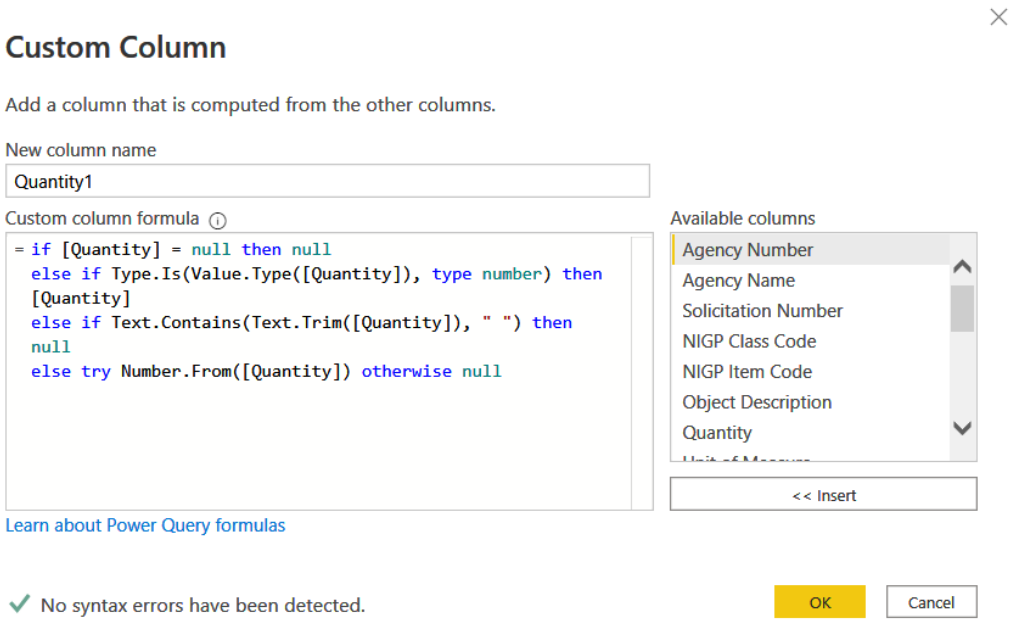


Figure 14 Building a Custom Column

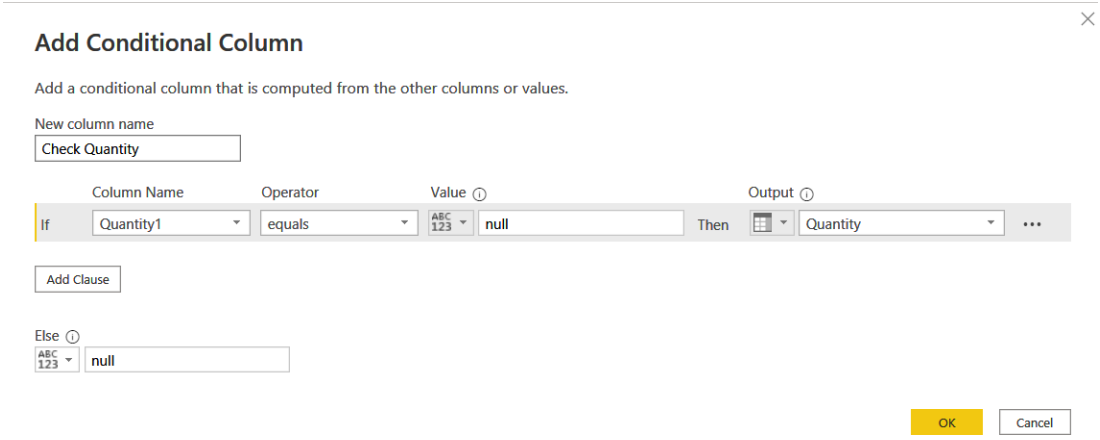


Figure 15 Building a Conditional Column

Quantity	Quantity1	Check Quantity
2	2	null
1	1	null
1	1	null
1	1	null
2	2	null
2	2	null
240	240	null
1	1	null
15	15	null
1	1	null
4	4	null
Error	null	multiple

Figure 16 Changing Data Types (1)

null	*null*	*null*	to
null	Multiple	Each	Ve
null	Multiple	Dz., Pkg., Ea., Cs.	Ot
null	*null*	Each	PE
null	*null*	*null*	SE
null	*null*	Each	Ol
null	various	various	Co
null	various	various	32
null	various	various	Ca
null	various	various	In
null	various	EA	Ur
null	various	EA	Cu
null	various*	various*	32
null	various*	various*	Co
null	various*	various*	Ca
null	various*	various*	In
null	various	EA	Gl
null	various	EA	Sc
null	various	EA	R1
null	various	EA	De
null	various	EA	20

Figure 17 Reviewing Data

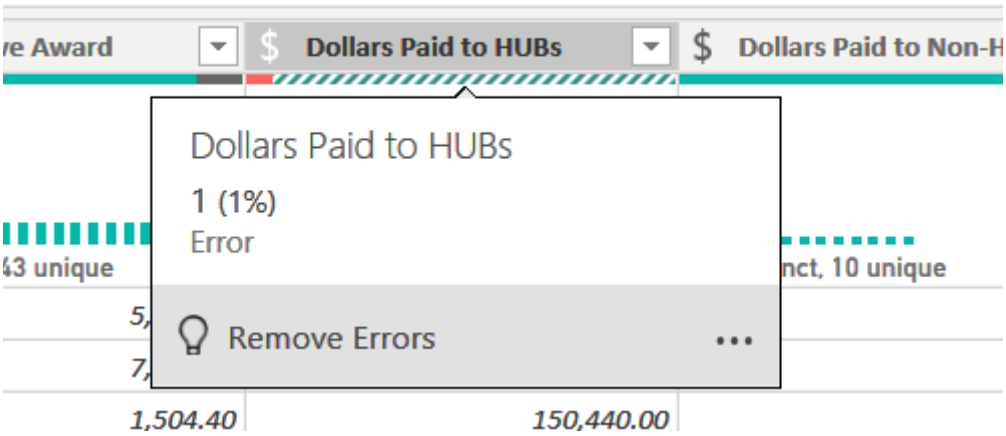


Figure 18 Identifying Errors

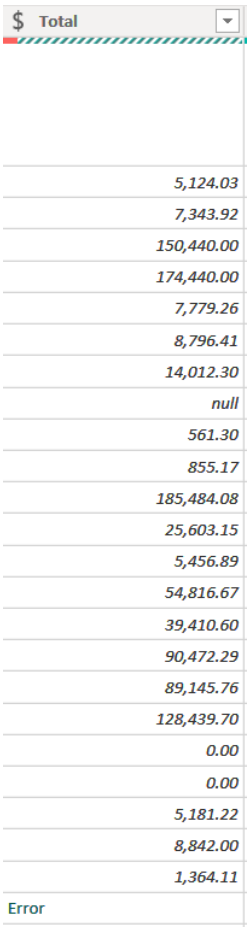


Figure 19 Finding Errors

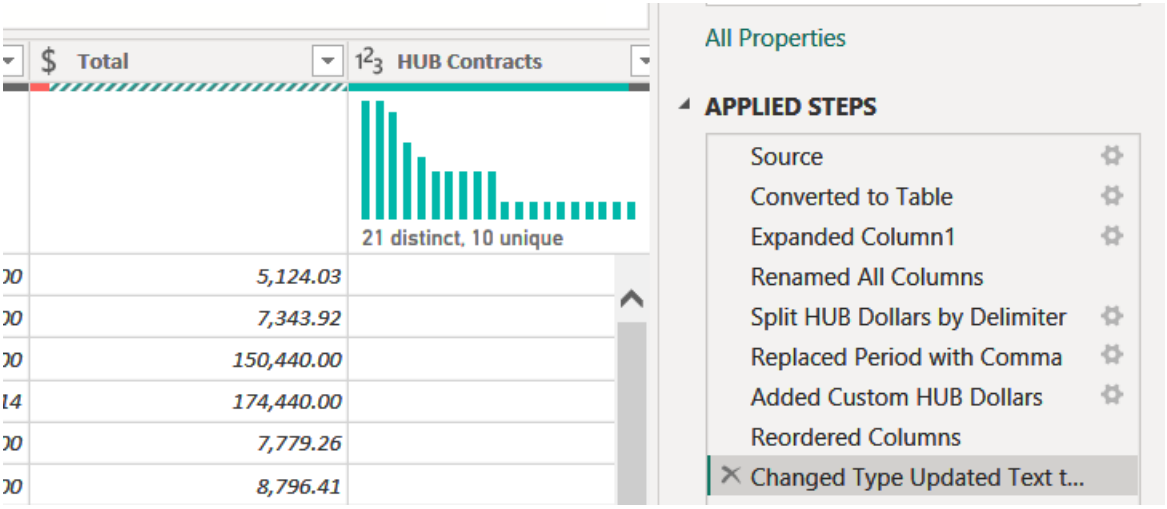


Figure 20 Changing Data Types (2)

	Agency Number	Agency Name	Tentative Award	Dollars Paid to HUB	Dollars Paid to Non-HUB	$
			1 distinct, 1 unique	1 distinct, 1 unique	1 distinct, 1 unique	1 di
1	458	Texas Alcoholic Beverage Commission	745.00	null	null	
2	458	Texas Alcoholic Beverage Commission	745.00	561.30	0.00	

Figure 21 Checking for Duplicates

Merged	Total
53 distinct, 49 unique	
5124.03	5124.03
7,343.92	7,343.92
$1,504,40.	$1,504,40
174,440.00	174,440.00
7779.26	7779.26
$8,796.41	$8,796.41
$14,012.30	$14,012.30
.	

Figure 22 Identifying Multiple Unique IDs

	Agency Number	Agency Name	Tentative Award	Dollars Paid to HUB	Dollars Paid to Non-HUB	$
			4 distinct, 3 unique	1 distinct, 0 unique	1 distinct, 0 unique	1
1	458	Texas Alcoholic Beverage Commission	745.00	null	null	
2	742	UNIVERSITY OF TEXAS PERMIAN BASIN	349,633.03	null	null	
3	802	null	null	null	null	
4	504	null	null	null	null	
5	713	Tarleton State University	16,800.00	null	null	
6	504	Texas State Board of Dental Examiners	null	null	null	
7	504	null	null	null	null	
8	320	null	null	null	null	

Figure 23 Identifying and Evaluating Nulls