Updating Data Editing and Review for Large-scale Commodity Flow Survey Response Data

2023 Federal Committee on Statistical Methodology

October 24, 2023
College Park Marriott Hotel & Conference Center
3501 University Blvd E, Hyattsville, MD 20783

Gritiya Tanner

United States® Census Bureau

Disclaimers

Any conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau.

The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Project No. P-7504831, Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD002-002).

Background

Challenges:

- 16 times data increase
- Computational costs
- Storage requirements
- Data review

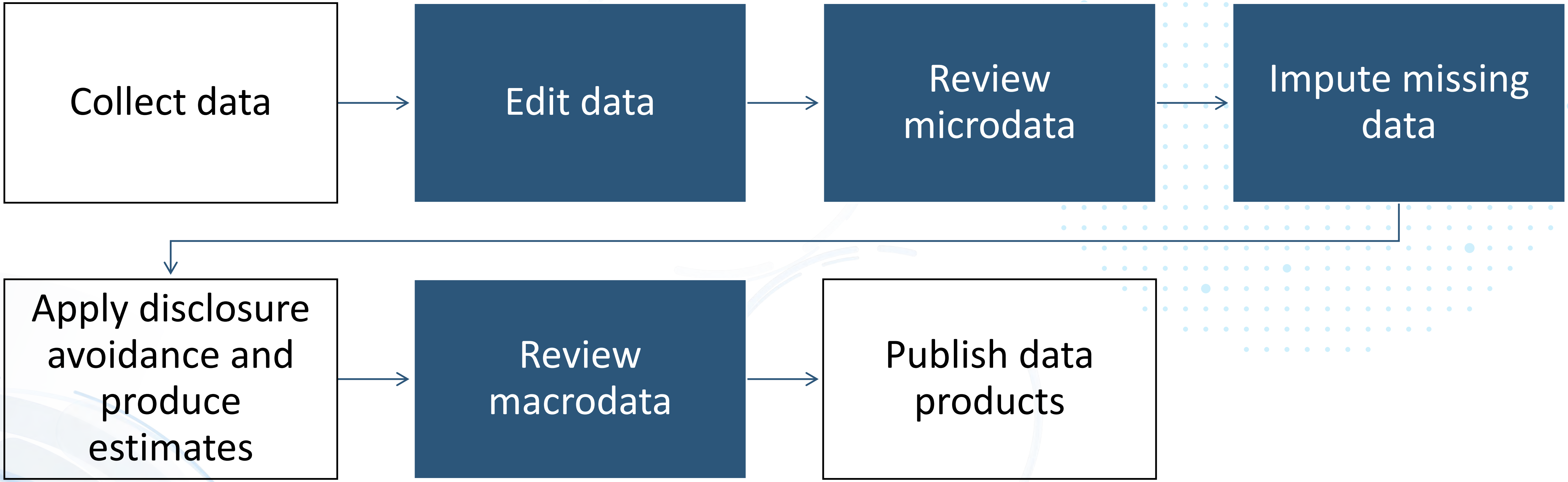
Solution:

- Redesign and automate the process as much as you can

Today:

- Highlight the implemented changes from 2017 CFS to 2022 CFS

2022 CFS Data processing pipeline



Collect data

Edit data

Review microdata

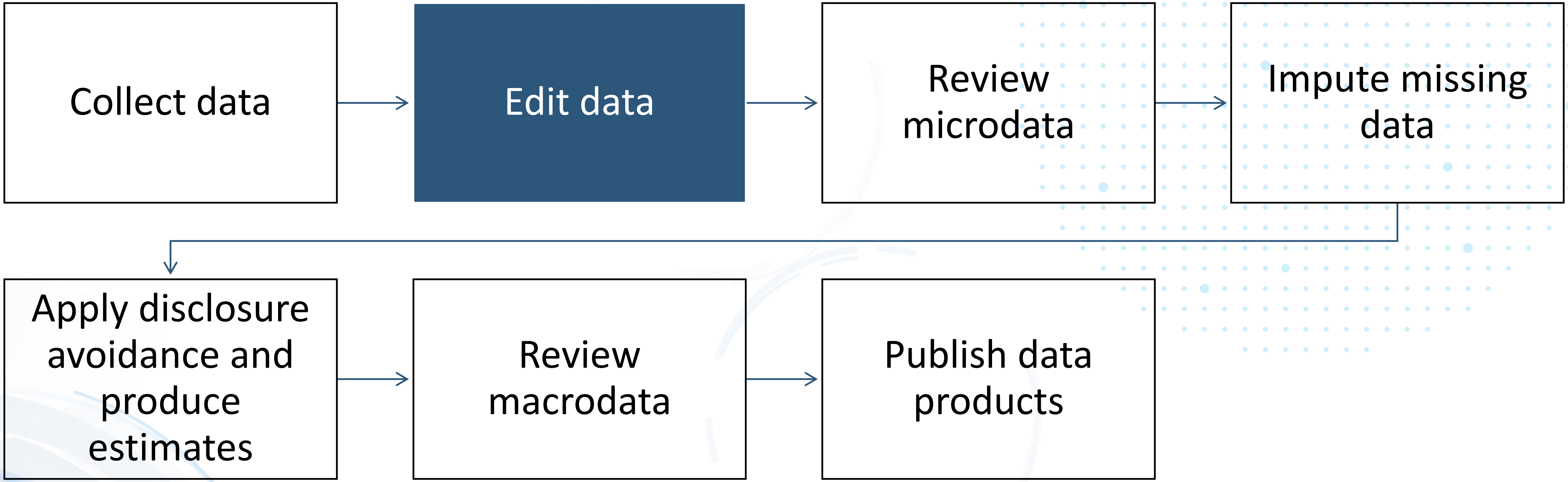
Impute missing data

Apply disclosure avoidance and produce estimates

Review macrodata

Publish data products

2022 CFS Data processing pipeline



Collect data

Edit data

Review microdata

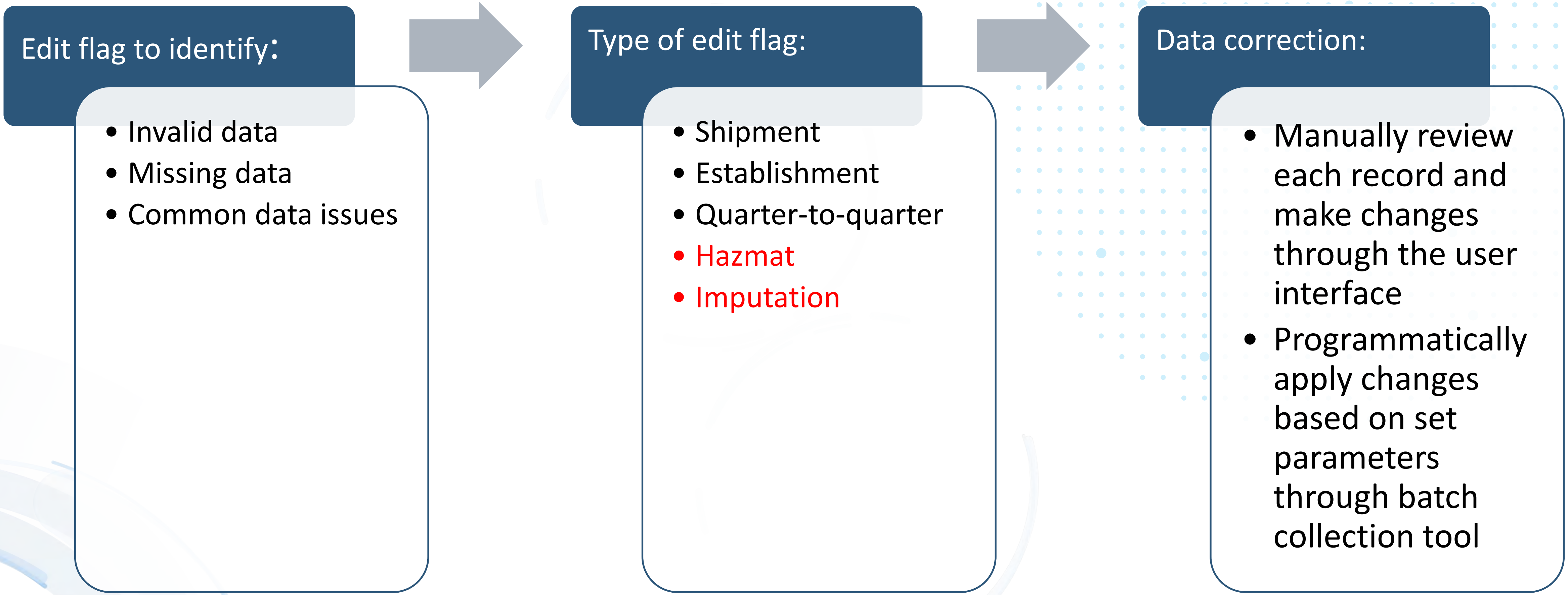
Impute missing data

Apply disclosure avoidance and produce estimates

Review macrodata

Publish data products

Data editing process



Edit flag to identify:

- Invalid data
- Missing data
- Common data issues

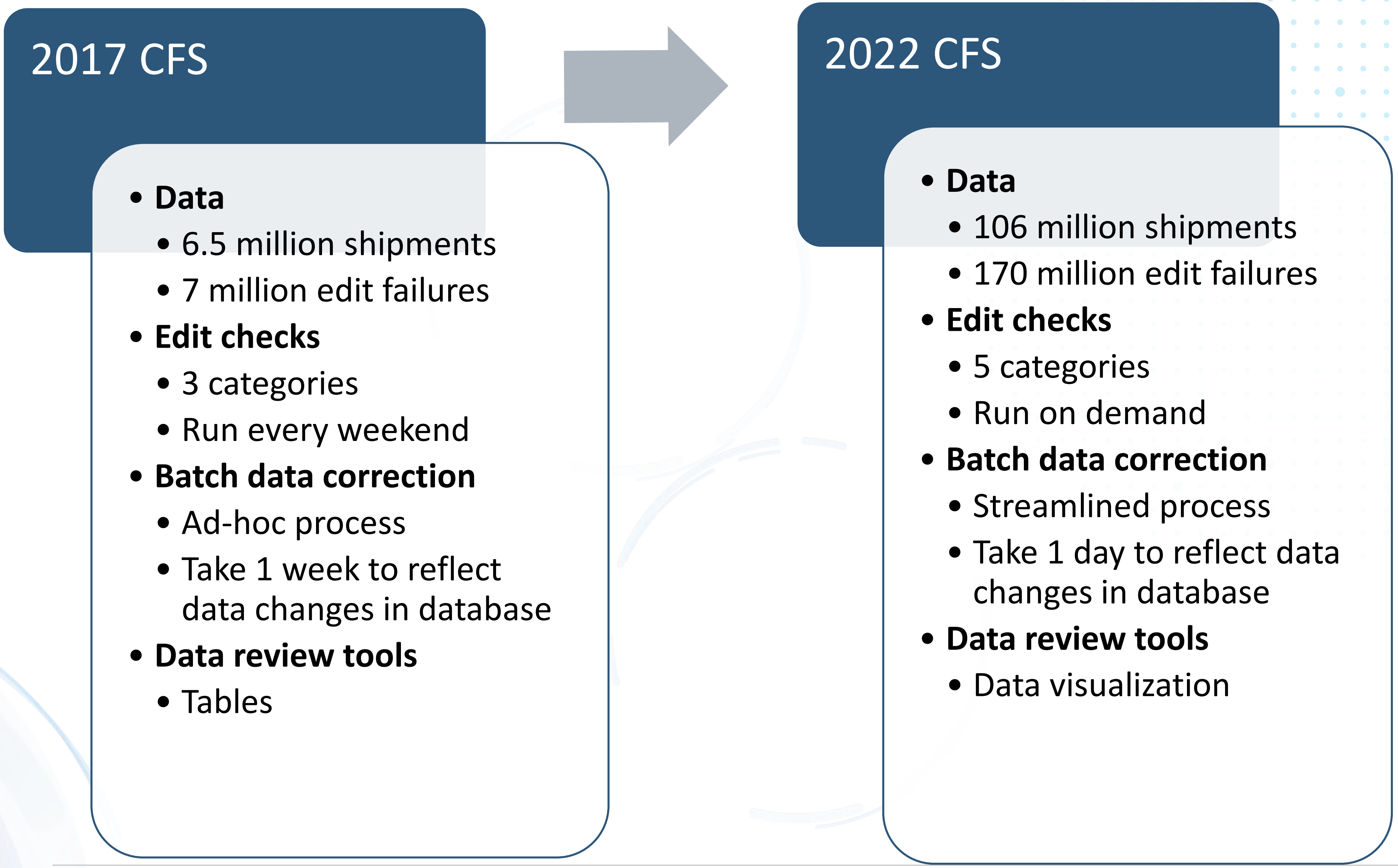
Type of edit flag:

- Shipment
- Establishment
- Quarter-to-quarter
- Hazmat
- Imputation

Data correction:

- Manually review each record and make changes through the user interface
- Programmatically apply changes based on set parameters through batch collection tool

Changes in data editing process



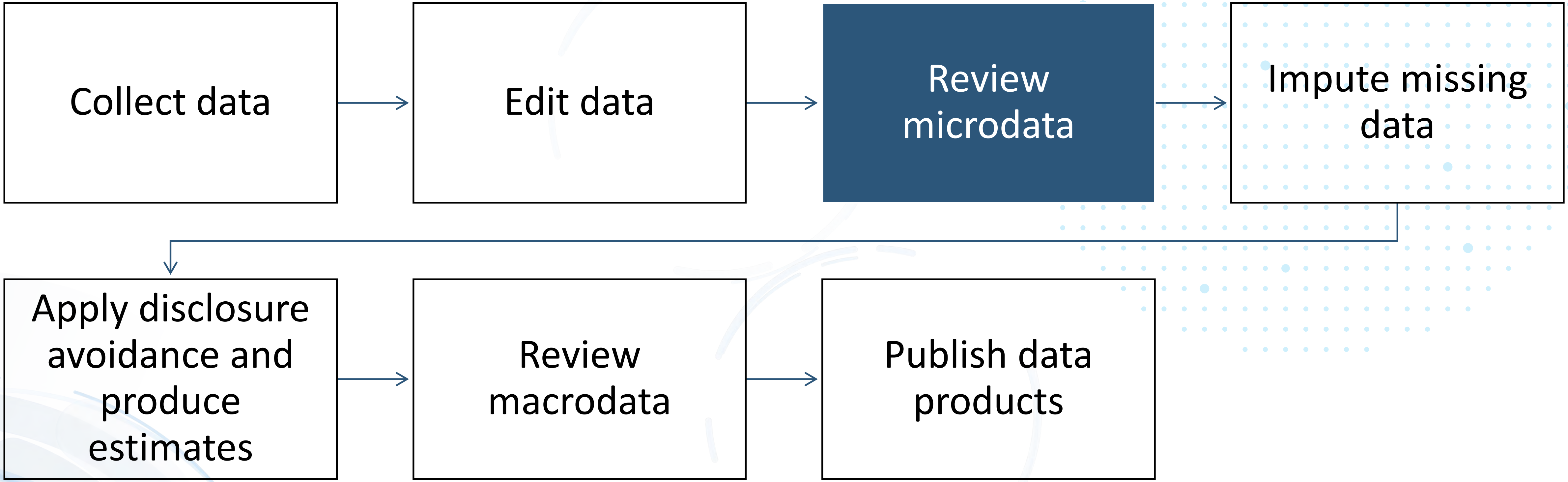
2017 CFS

- **Data**
  - 6.5 million shipments
  - 7 million edit failures
- **Edit checks**
  - 3 categories
  - Run every weekend
- **Batch data correction**
  - Ad-hoc process
  - Take 1 week to reflect data changes in database
- **Data review tools**
  - Tables

2022 CFS

- **Data**
  - 106 million shipments
  - 170 million edit failures
- **Edit checks**
  - 5 categories
  - Run on demand
- **Batch data correction**
  - Streamlined process
  - Take 1 day to reflect data changes in database
- **Data review tools**
  - Data visualization

2022 CFS Data processing pipeline



Collect data

Edit data

Review microdata

Impute missing data

Apply disclosure avoidance and produce estimates

Review macrodata

Publish data products

Microdata review process

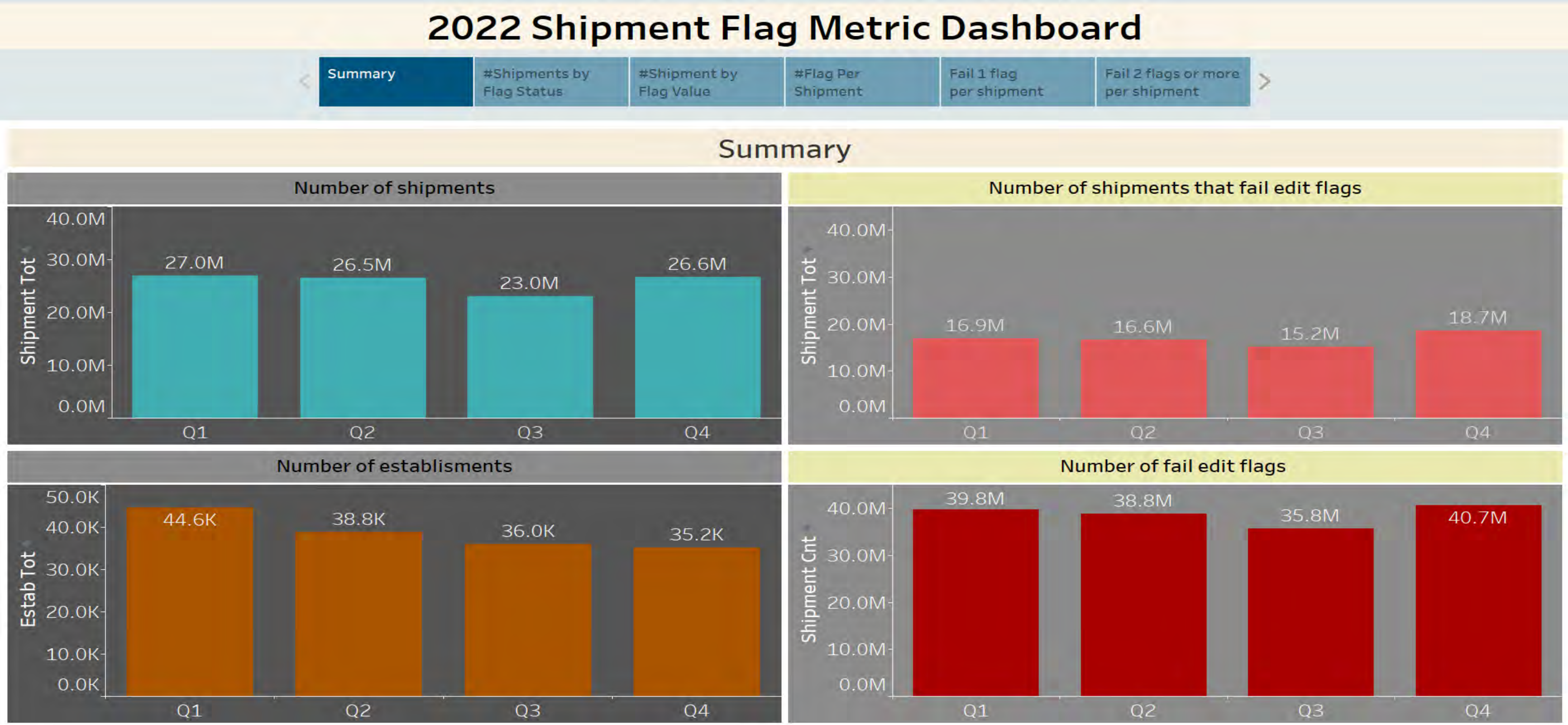
Review:

- Visualization tool
- Prioritize the review
  - Quarter
  - Make it to tabulation
- Figure out the approach to make a correction

Make any correction or changes:

- Manually change each record through the user interface
- Programmatically apply changes through batch correction

Microdata review tool – Visualization summary (New for 2022 CFS)



2022 Shipment Flag Metric Dashboard

Summary

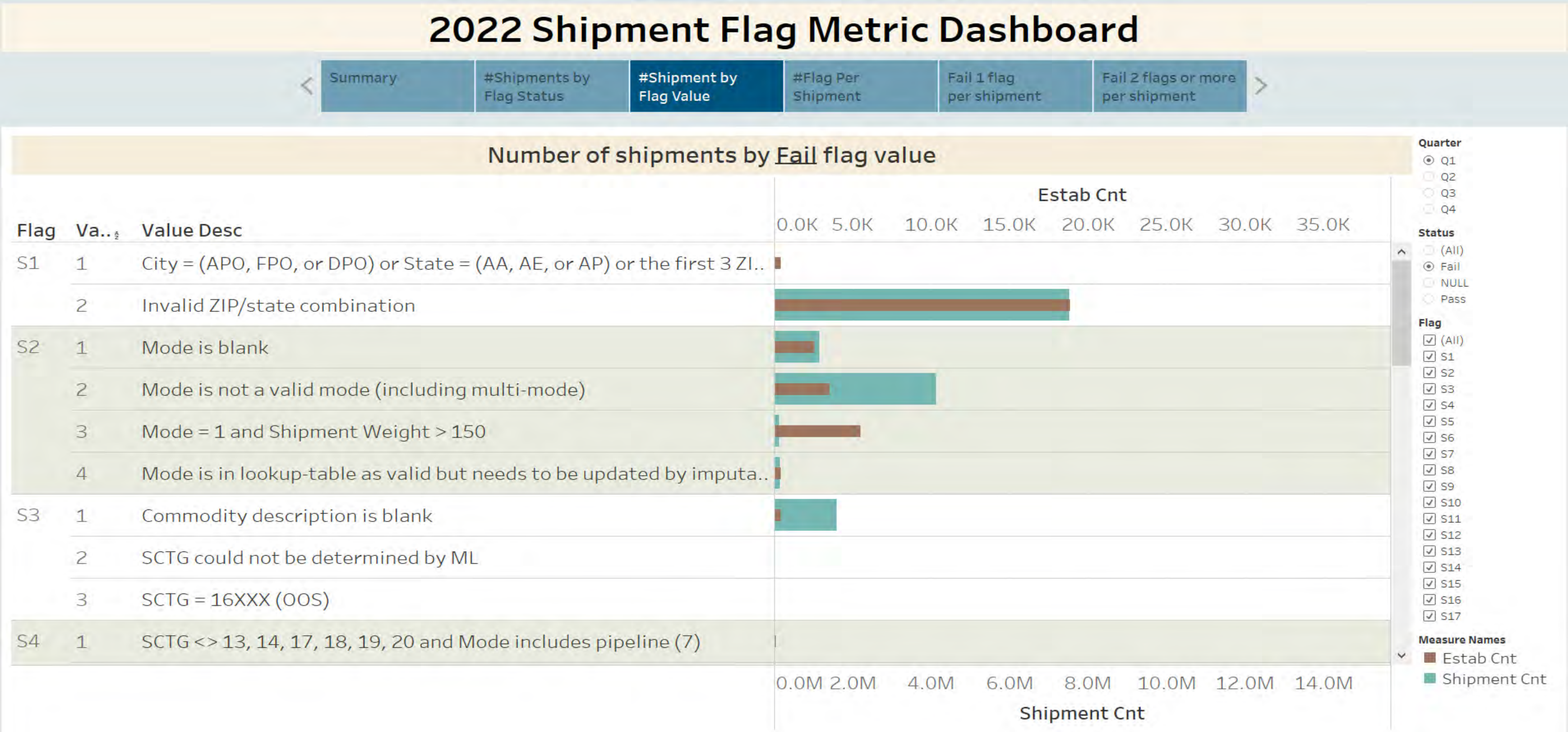
Number of shipments

Number of shipments that fail edit flags

Number of establisments

Number of fail edit flags

Microdata review tool – Visualization detail (New for 2022 CFS)



2022 Shipment Flag Metric Dashboard

Number of shipments by Fail flag value

Flag	Va..	Value Desc
S1	1	City = (APO, FPO, or DPO) or State = (AA, AE, or AP) or the first 3 ZI..
	2	Invalid ZIP/state combination
S2	1	Mode is blank
	2	Mode is not a valid mode (including multi-mode)
	3	Mode = 1 and Shipment Weight > 150
	4	Mode is in lookup-table as valid but needs to be updated by imputa..
S3	1	Commodity description is blank
	2	SCTG could not be determined by ML
	3	SCTG = 16XXX (OOS)
S4	1	SCTG <> 13, 14, 17, 18, 19, 20 and Mode includes pipeline (7)

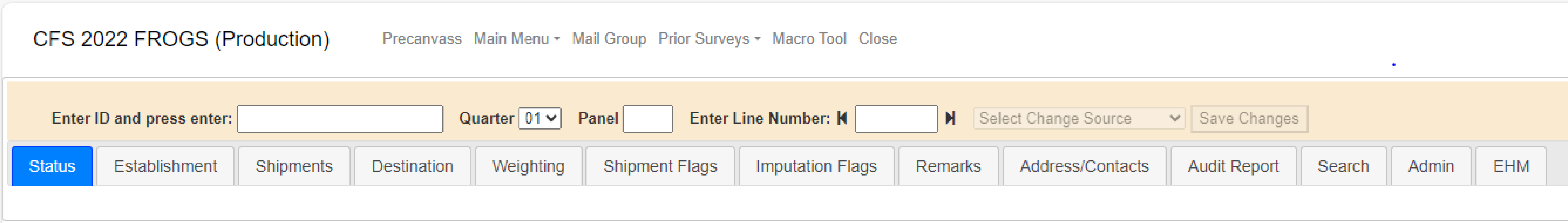
Quarter
● Q1
○ Q2
○ Q3
○ Q4

Status
○ (All)
● Fail
○ NULL
○ Pass

Flag
☑ (All)
☑ S1
☑ S2
☑ S3
☑ S4
☑ S5
☑ S6
☑ S7
☑ S8
☑ S9
☑ S10
☑ S11
☑ S12
☑ S13
☑ S14
☑ S15
☑ S16
☑ S17

Measure Names
🟫 Estab Cnt
🟩 Shipment Cnt

CFS Production user interface



CFS 2022 FROGS (Production)

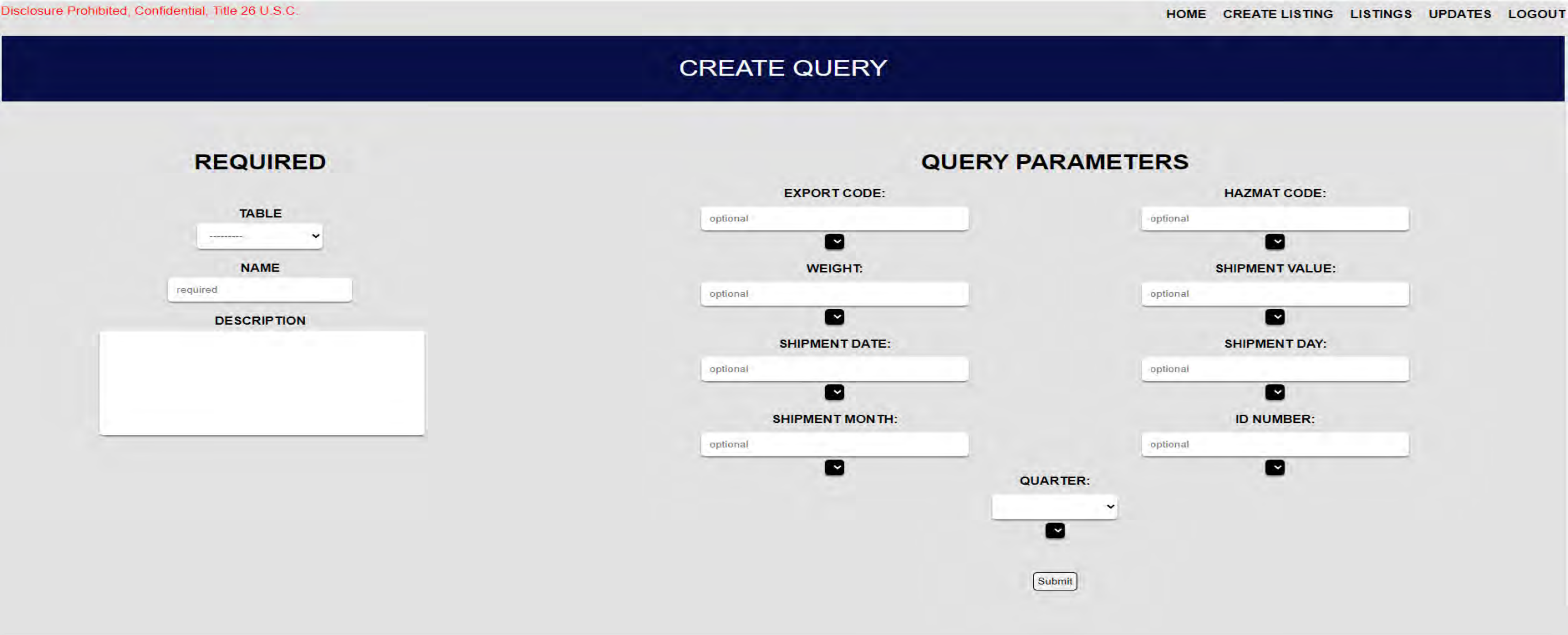
Enter ID and press enter:

Quarter 01

Panel

Enter Line Number:

Batch correction tool



Disclosure Prohibited, Confidential, Title 26 U.S.C.

CREATE QUERY

REQUIRED

TABLE: ---------

NAME: required

DESCRIPTION:

QUERY PARAMETERS

EXPORT CODE: optional

HAZMAT CODE: optional

WEIGHT: optional

SHIPMENT VALUE: optional

SHIPMENT DATE: optional

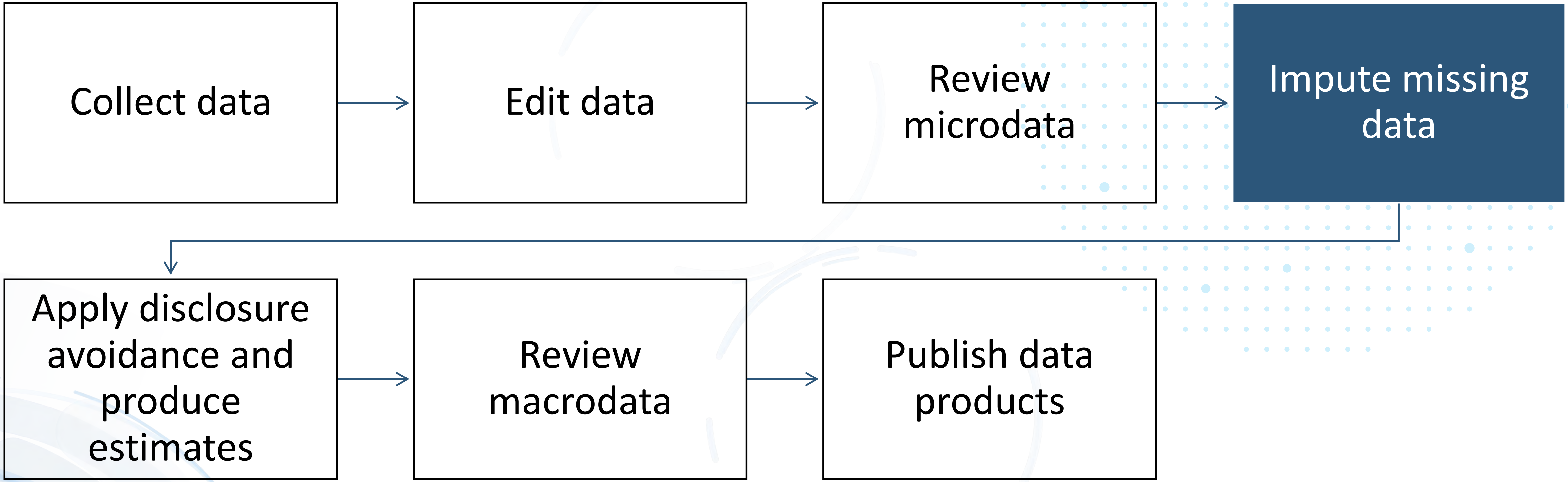
SHIPMENT DAY: optional

SHIPMENT MONTH: optional

ID NUMBER: optional

QUARTER:

2022 CFS Data processing pipeline



Collect data

Edit data

Review microdata

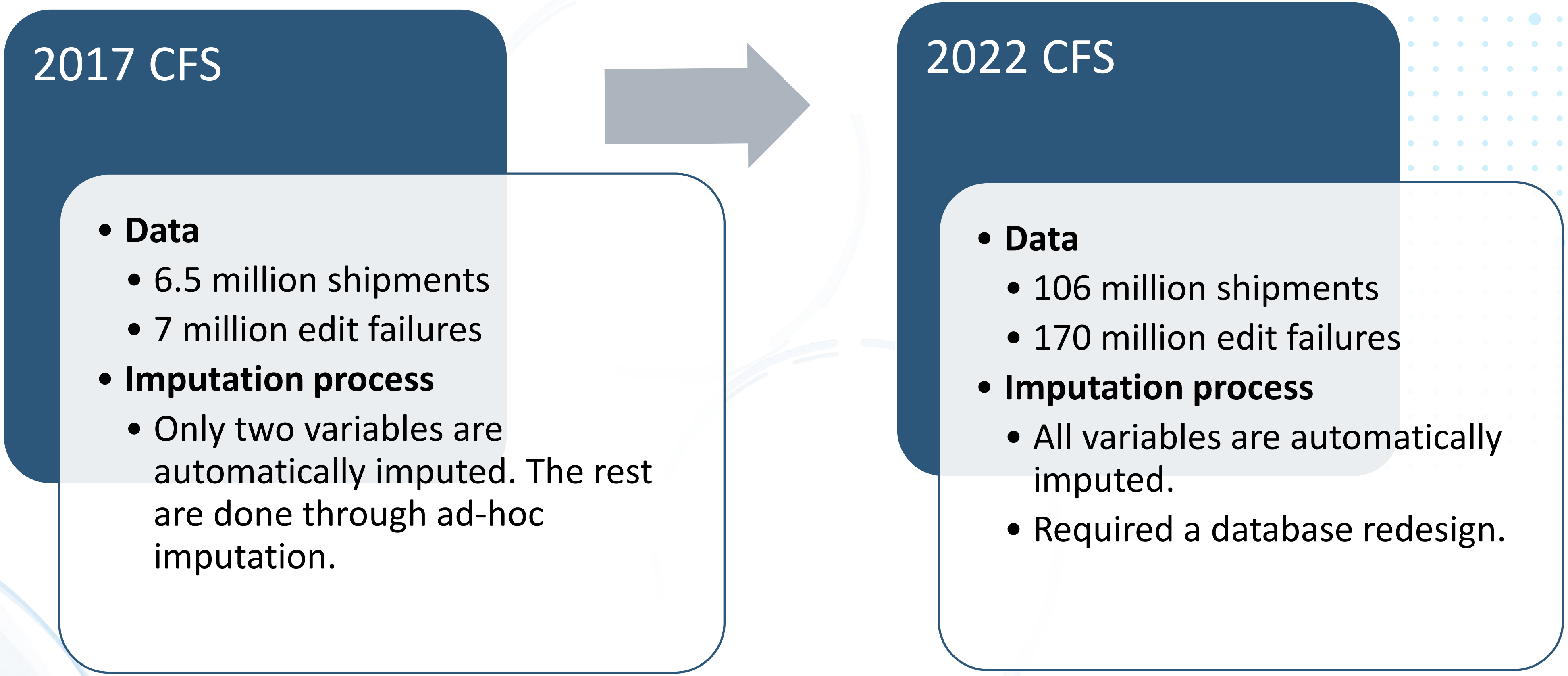
Impute missing data

Apply disclosure avoidance and produce estimates

Review macrodata

Publish data products

The changes on imputation process



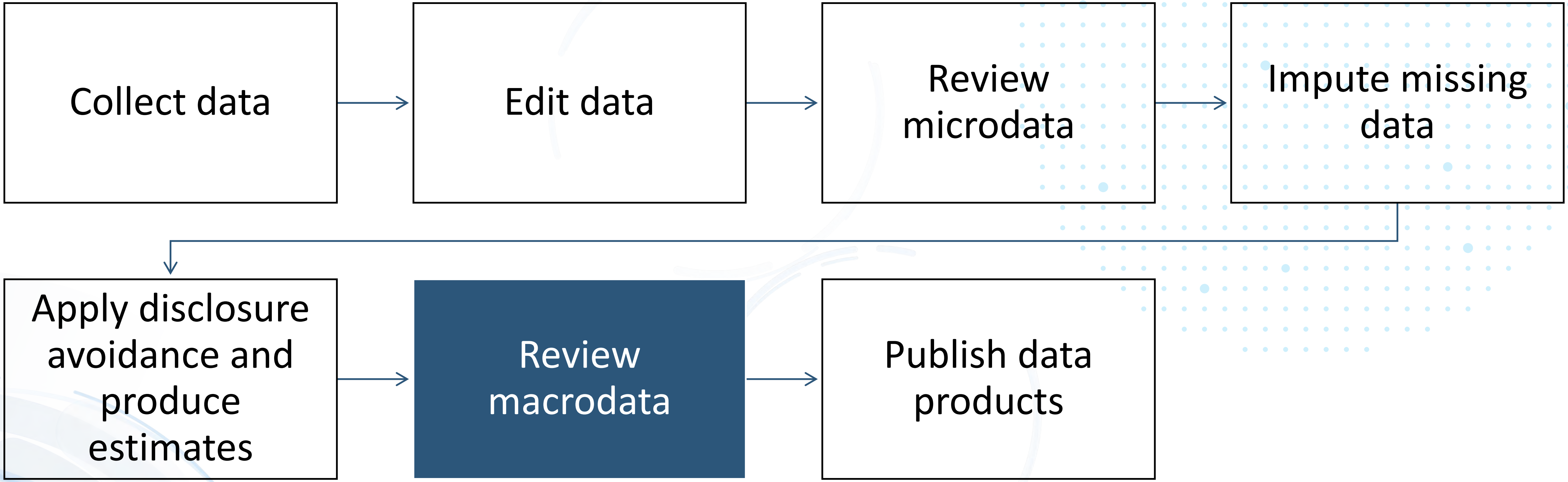
2017 CFS

- **Data**
  - 6.5 million shipments
  - 7 million edit failures
- **Imputation process**
  - Only two variables are automatically imputed. The rest are done through ad-hoc imputation.

2022 CFS

- **Data**
  - 106 million shipments
  - 170 million edit failures
- **Imputation process**
  - All variables are automatically imputed.
  - Required a database redesign.

2022 CFS Data processing pipeline



Collect data

Edit data

Review microdata

Impute missing data

Apply disclosure avoidance and produce estimates

Review macrodata

Publish data products

Macrodata review process

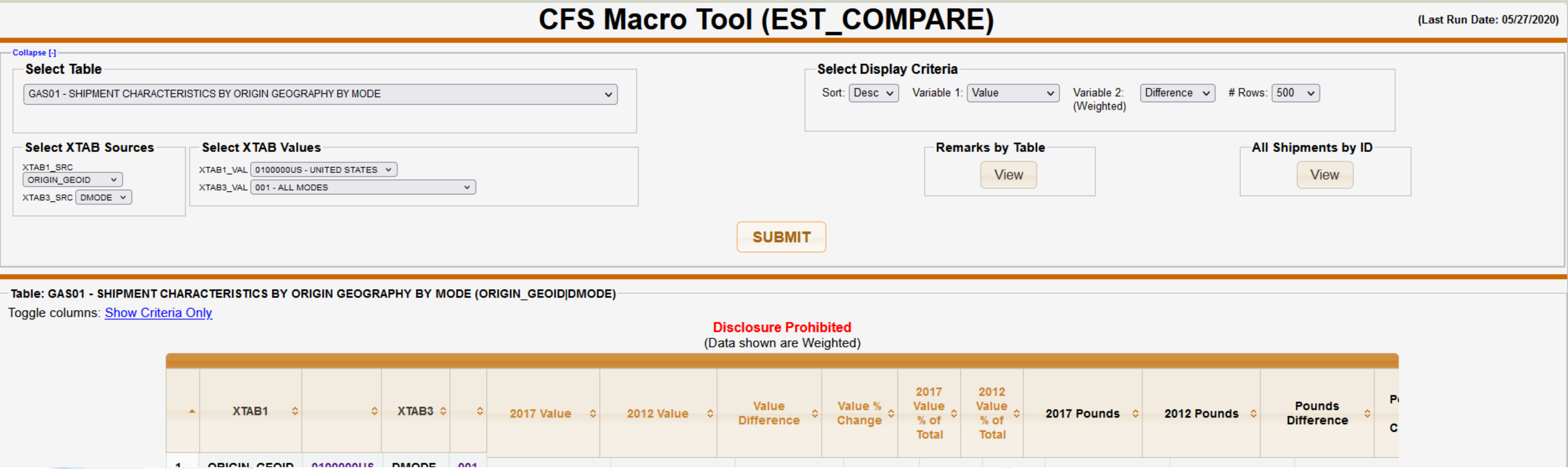
Review:

- Visualization tool
- Review and learn macrodata
- Figure out the approach to make a correction
- Perform macrodata review prior to micro data review when estimation available
- Estimation available 1.5 year sooner than last cycle

Make any correction or changes:

- Manually review record and make changes through the user interface
- Programmatically apply changes through batch correction

Macrodata review tool - 2017 CFS version



CFS Macro Tool (EST_COMPARE)

(Last Run Date: 05/27/2020)

Collapse [-]

Select Table

GAS01 - SHIPMENT CHARACTERISTICS BY ORIGIN GEOGRAPHY BY MODE

Select Display Criteria

Sort: Desc | Variable 1: Value | Variable 2: (Weighted) Difference | # Rows: 500

Select XTAB Sources

XTAB1_SRC ORIGIN_GEOID

XTAB3_SRC DMODE

Select XTAB Values

XTAB1_VAL 0100000US - UNITED STATES

XTAB3_VAL 001 - ALL MODES

Table: GAS01 - SHIPMENT CHARACTERISTICS BY ORIGIN GEOGRAPHY BY MODE (ORIGIN_GEOID|DMODE)

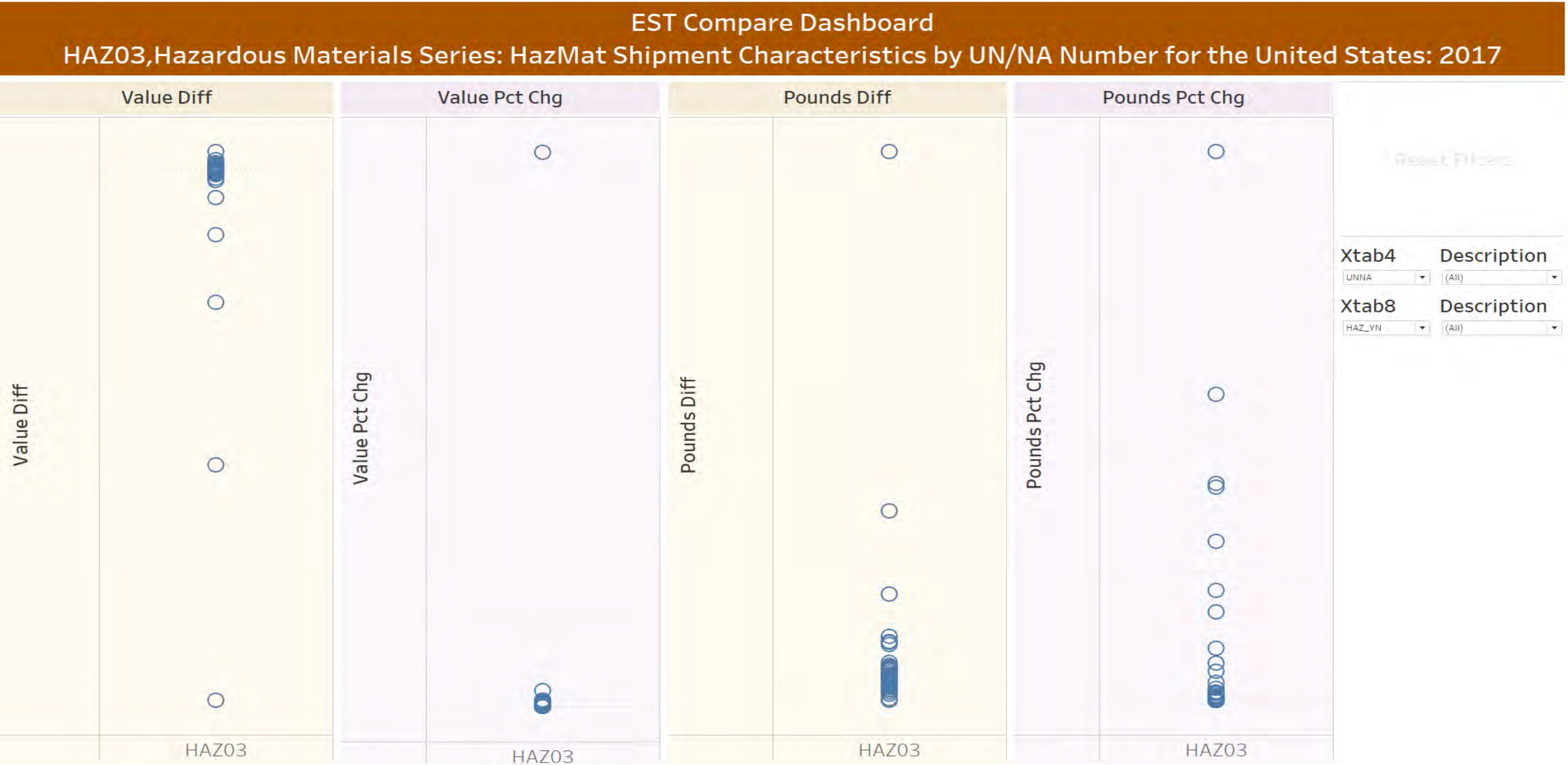
Toggle columns: Show Criteria Only

Disclosure Prohibited

(Data shown are Weighted)

	XTAB1		XTAB3		2017 Value	2012 Value	Value Difference	Value % Change	2017 Value % of Total	2012 Value % of Total	2017 Pounds	2012 Pounds	Pounds Difference	P C
1	ORIGIN_GEOID	0100000US	DMODE	001										

Macro data review tool – new in 2022 CFS



EST Compare Dashboard

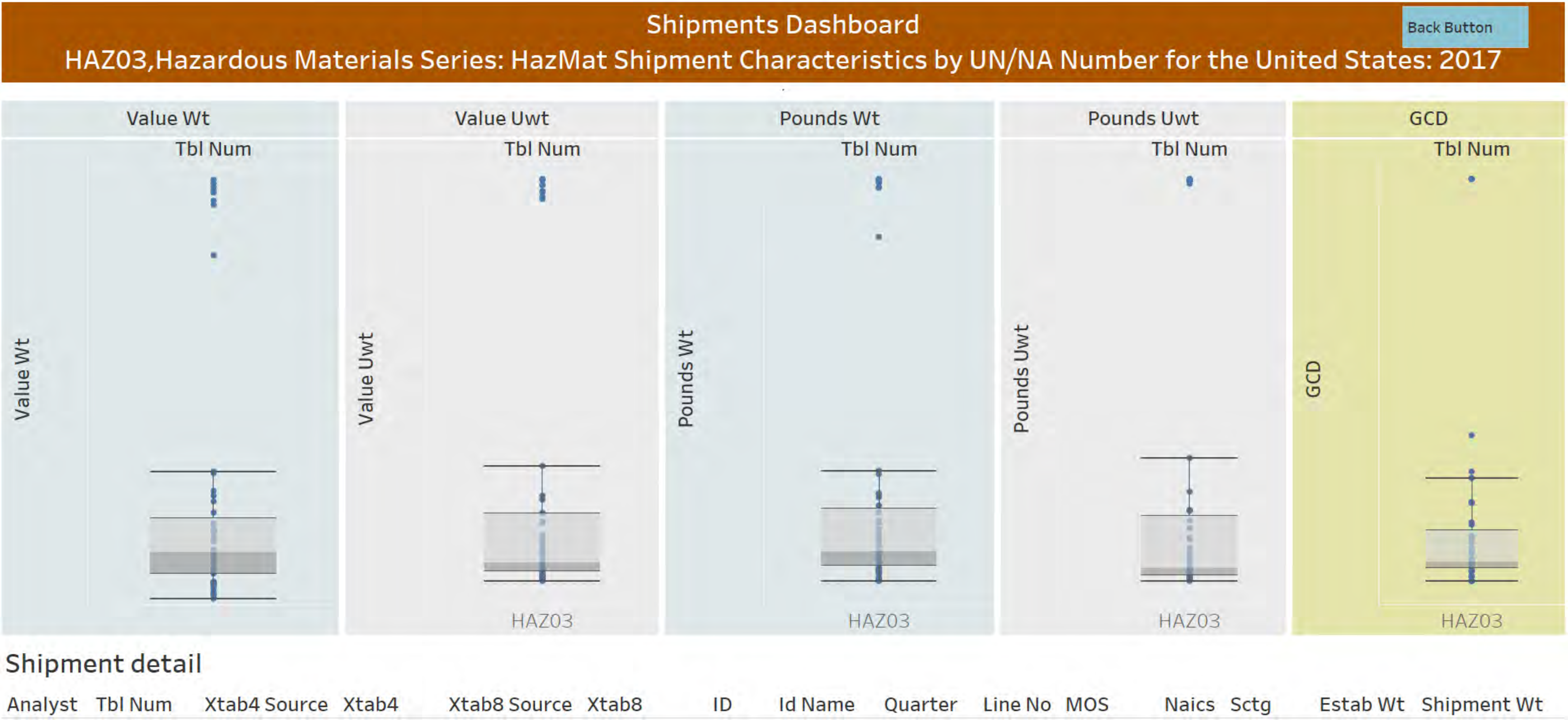
HAZ03,Hazardous Materials Series: HazMat Shipment Characteristics by UN/NA Number for the United States: 2017

Reset Filters

Xtab4: UNNA
Description: (All)

Xtab8: HAZ_YN
Description: (All)

Macrodata review tool - new in 2022 CFS



Shipments Dashboard

HAZ03,Hazardous Materials Series: HazMat Shipment Characteristics by UN/NA Number for the United States: 2017

Shipment detail

Analyst	Tbl Num	Xtab4 Source	Xtab4	Xtab8 Source	Xtab8	ID	Id Name	Quarter	Line No	MOS	Naics	Sctg	Estab Wt	Shipment Wt

Summary

To address the challenges associated with the large-scale data processing pipeline, including:

- Automated the process as much as we can
- Redesign the database to support the automated process effectively
- Using the data visualization tool to speed up the data review
- Review macrodata before microdata when possible

These improvements can be adopted by other surveys whether dealing with a large-scale data set or small sets of data.

Updating Data Editing and Review for Large-scale Commodity Flow Survey Response Data

QUESTIONS?

Contacts

Gritiya Tanner

Survey Statistician/Data scientist
Business Development Staff

Economic Reimbursable Surveys Division

U.S. Census Bureau

gritiya.tanner@census.gov