Enabling Economic Statistics Modernization

Jessica Wellwood, U.S. Census Bureau

October 24, 2023

2023 Research and Policy Conference

Federal Committee on Statistical Methodology

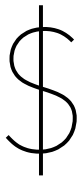
United States Census Bureau

Note: Any opinions and conclusions expressed herein are those of the author 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 (Disclosure Review Board (DRB) approval number: CBDRB-FY23-ESMD002-031).

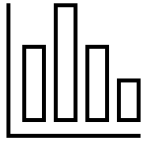
Transformation at the Census Bureau

PROBLEM

Declining Response Rates



Shrinking Budgets



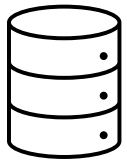
Increased demand for data

United States® Census Bureau

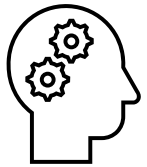
Transformation at the Census Bureau

SOLUTION

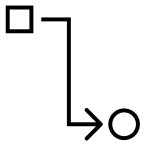
DATA CENTRIC



Leverage existing data



Promote Innovation



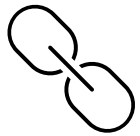
Simplify Processes

United States® Census Bureau

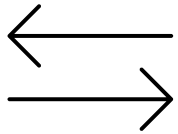
Transformation at the Census Bureau

VISION

To create Enterprise-wide frames that are…



Linkable in nature



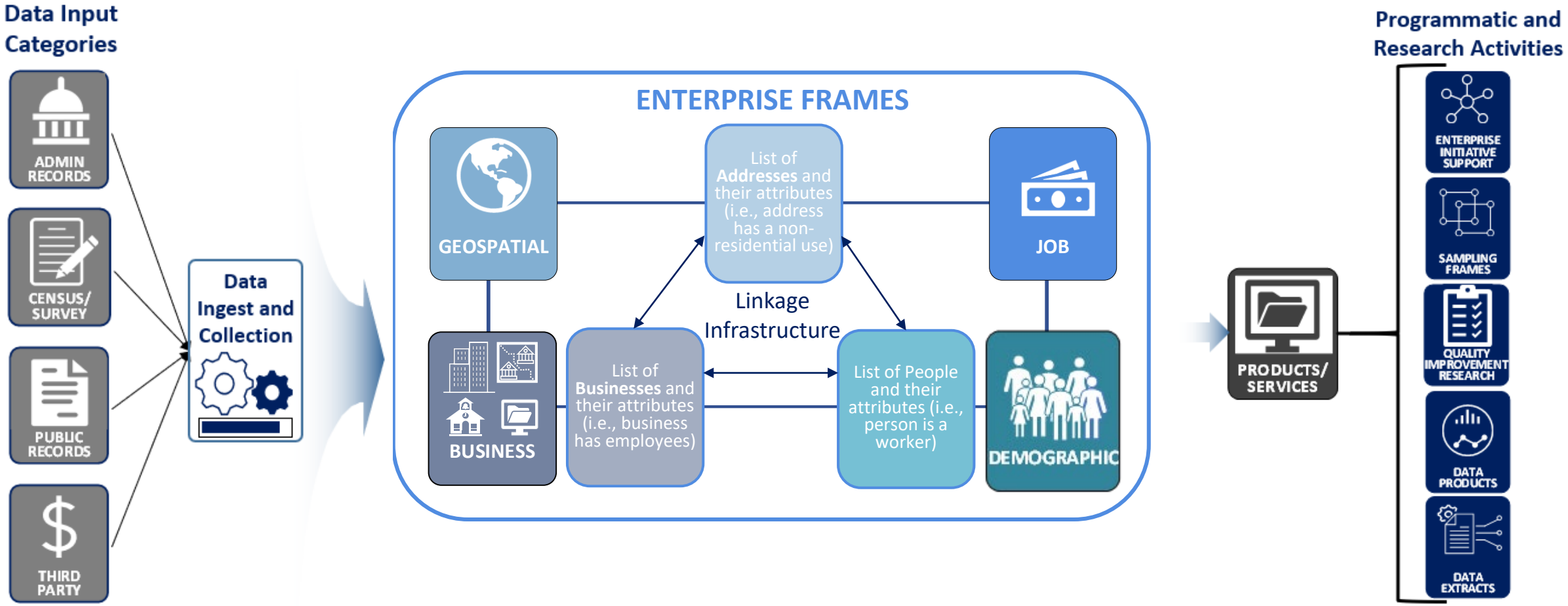
Agile in Structure

Accessible for research or production

United States® Census Bureau

Frames Program

Creating an Infrastructure to Modernize the Census Bureau's Statistical Foundation



United States® Census Bureau

The Business Frame is **NOT** the Business Register!

Business FRAME

Collection of auxiliary data, harmonized business data across multiple sources, linked to the BR, and stored in a central location

- In-depth data specific to a defined sub-population of Business Entities*
- Linkages across data sources and time (LBD)
- Data owned by designated data provider
- All data sources welcome!

Business REGISTER

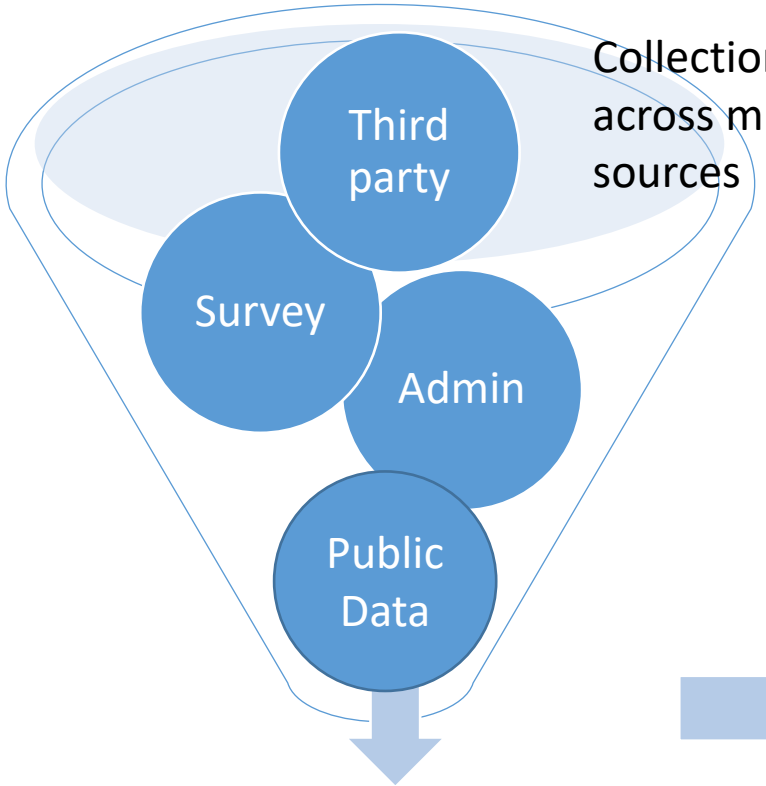
Master list of businesses with associated core attributes

- Full Coverage of business population
- Linkages between BR statistical units (e.g. Establishments, Enterprises)
- Updated from administrative data (IRS, SSA, BLS) and select census programs
- Data are owned by Business Register Staff

United States® Census Bureau

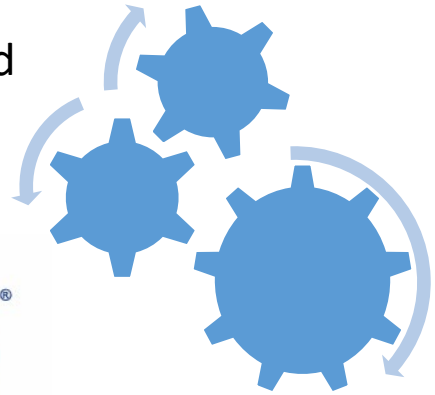
*Examples of Business Entities are governments, publicly traded companies, records in a specific NAICS

What is the Business Frame?



Collection of Data across multiple sources

Harmonized and linked to the BR



Business Register

- Survey Data
- Third Party Data
- Admin Data
- Public Data
- And More!

Stored in a central location

United States® Census Bureau

Business Frame

What?

Collection of rich, harmonized business data across multiple sources, **linked to the BR**, and stored in a central location

Why?

TO FACILITATE THE DEVELOPMENT OF NEW AND IMPROVED DATA PRODUCTS

How?

Create a relational database that links data together, utilizing probabilistic matching

United States® Census Bureau

Scope the Work – Phase 1: Prototype

Goal: Leverage existing data more effectively

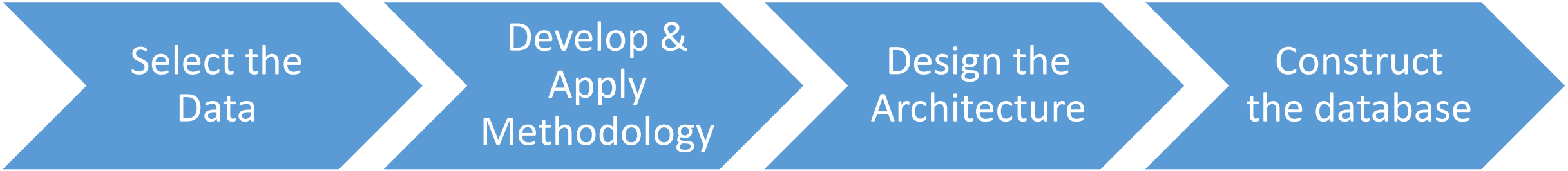
Objective: Demonstrate link-ability and utility

Can we link data to the Business Register in a reliable and useful way?

Acceptance Criteria*:* We have database, with all data loaded and connected in a meaningful way.

United States® Census Bureau

How are we going to do this?



Select the Data Sources

Point-of-sale data provided by Third Party*

What: Monthly credit card transactions aggregated to the product level for select retailers

Why: Timeliness of data (more current than BR), new level of granularity

Longitudinal Business Database (LBD)

What: Links BR establishments over time, calculates firm age and size

Why: Longitudinal dimension for use in sampling parameter, measuring business dynamism

Non-Employer Statistics by Demographics (NES-D)

What: Administrative data product assigning business level demographic characteristics to non-employer companies

Why: Demographic dimension to business data, Linking mechanism to Demographic Frame

Governments Master Address File (GMAF)

What: List of state & local governments, and core attributes

Why: Known overlap with BR, share/validate coverage, classification, financial data and contact information

United States® Census Bureau

*Data purchased from Circana, a private sector market research company

Develop Methodology

123=123
ABC=ABC
5XY=5XY

Common Identifier

- Match records based on a shared identifier in both data sets
- High Accuracy, Minimal Resources

Probabilistic Matching

- Utilizes machine learning to conduct pair-wise matching
- Varied Accuracy, High Learning Curve

Analyst Review

- Analyst matches records between data sources
- High Accuracy, Resource Intensive

United States® Census Bureau

Apply Methodology

Point-of-sale data provided by Third Party*

What: monthly credit card transactions aggregated to the product level for select retailers

Why: Timeliness of data (more current than BR), new level of granularity

Analyst Review

Longitudinal Business Database (LBD)

What: Links BR establishments over time, calculates firm age and size

Why: Longitudinal dimension for use in sampling parameter, measuring business dynamism

Common Identifier

Non-Employer Statistics by Demographics (NES-D)

What: Administrative data product assigning business level demographic characteristics to non-employer companies

Why: Demographic dimension to business data, Linking mechanism to Demographic Frame

Common Identifier

Governments Master Address File (GMAF)

What: List of state & local governments, and core attributes

Why: Known overlap with BR, share/validate coverage, classification, financial data and contact information

Probabilistic Matching

United States® Census Bureau

*Data purchased from Cirana, a private sector market research company

Design the Data Architecture

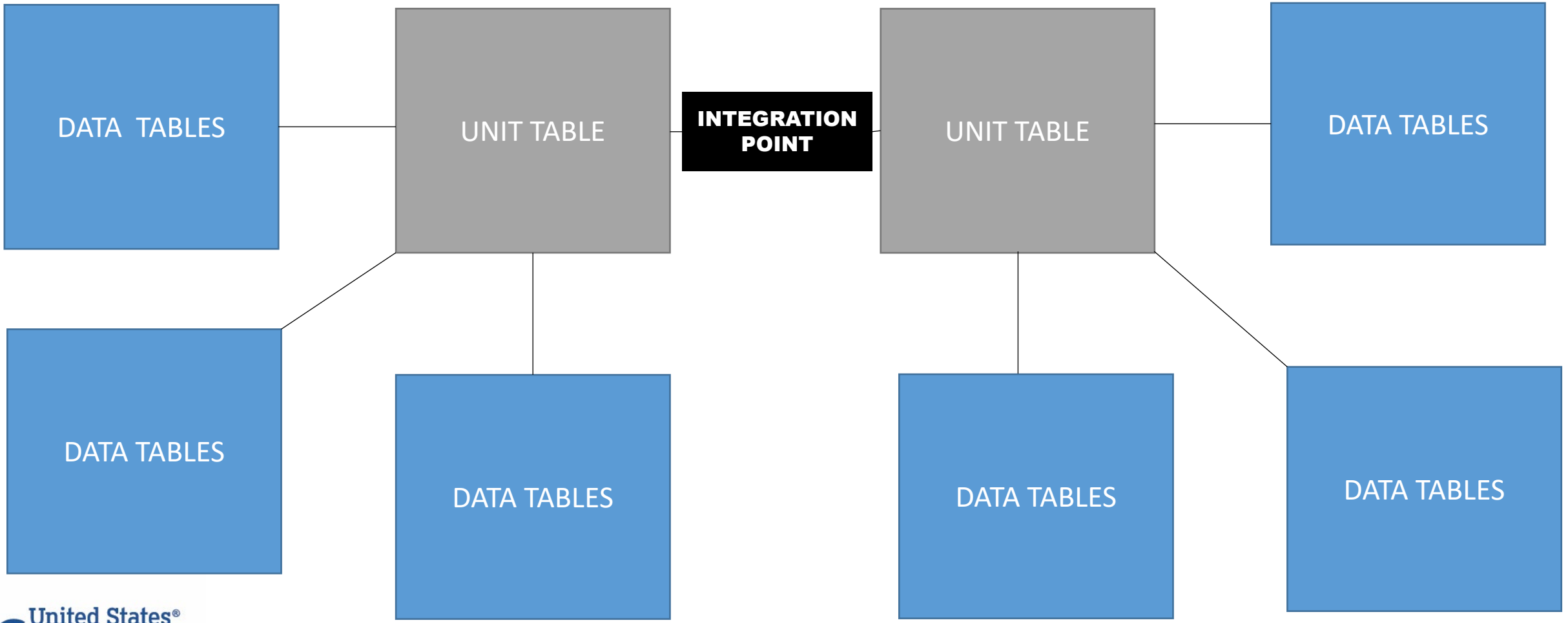
Expandable Adaptable Sustainable

What does every data source have in common?

UNITS & DATA

United States® Census Bureau

Design the Data Architecture



United States® Census Bureau

Design the Data Architecture

Point-of-sale data provided by Third Party*

Point of Sale Data

THIRD PARTY UNIT

Analyst Review

Longitudinal Business Database (LBD)

Common Identifier

LBD UNIT

LBD DATA

BR UNITS

Non-Employer Statistics by Demographics (NES-D)

Common Identifier

NESD DATA

NES-D UNIT

Governments Master Address File (GMAF)

Probabilistic Matching

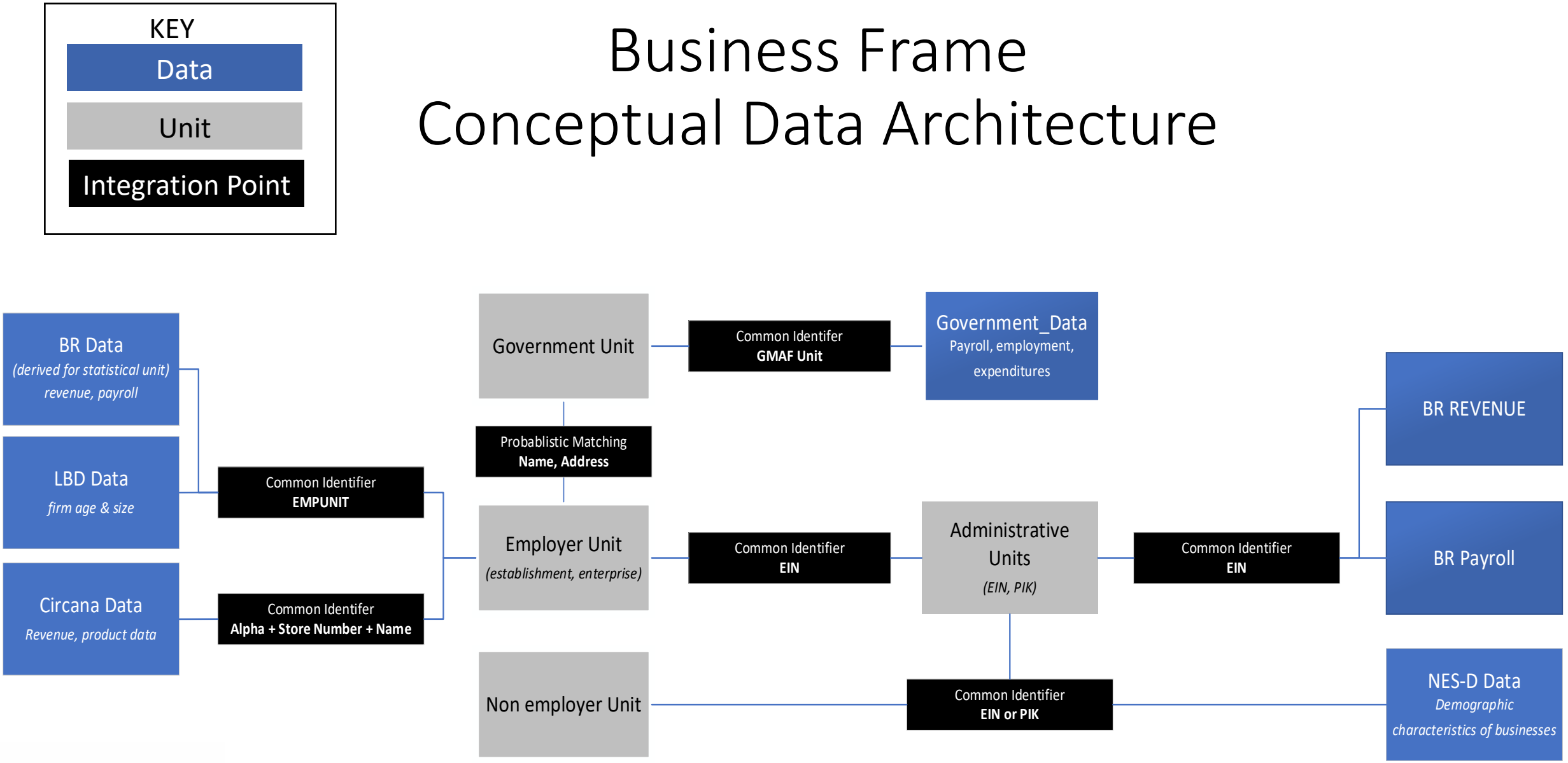
GMAF UNIT

GMAF DATA

United States® Census Bureau

*Data purchased from Cirana, a private sector market research company

Business Frame Conceptual Data Architecture



United States® Census Bureau

Data Architecture Design Decisions

Normalized Entities

- Linked records provided by source systems deconstructed
- Duplication reconciled and referential integrity built into model
- Optimal storage partitioning enabled for query performance

Extensible Data Model

- Foreign key links to other enterprise frames (Jobs, Demographic, Geospatial)
- Core characteristics assigned to UNIT entities to readily accommodate new data/measures
- Enterprise integration points established and agreed upon by all business data owners

Business Rules

- Cardinality* provided visually intuitive business rule representation for users
- Consistency of data relationships confirmed with source system data owners
- Source system relationships preserved for enterprise cohesiveness

Technology Agnostic

- ANSI SQL database constructs (tables, columns, data types, keys, constraints, indexes)
- Scripted, repeatable data population into prototype

United States® Census Bureau

* The numerical relationship between records across entities in a data model is called “cardinality”

Construct the Database

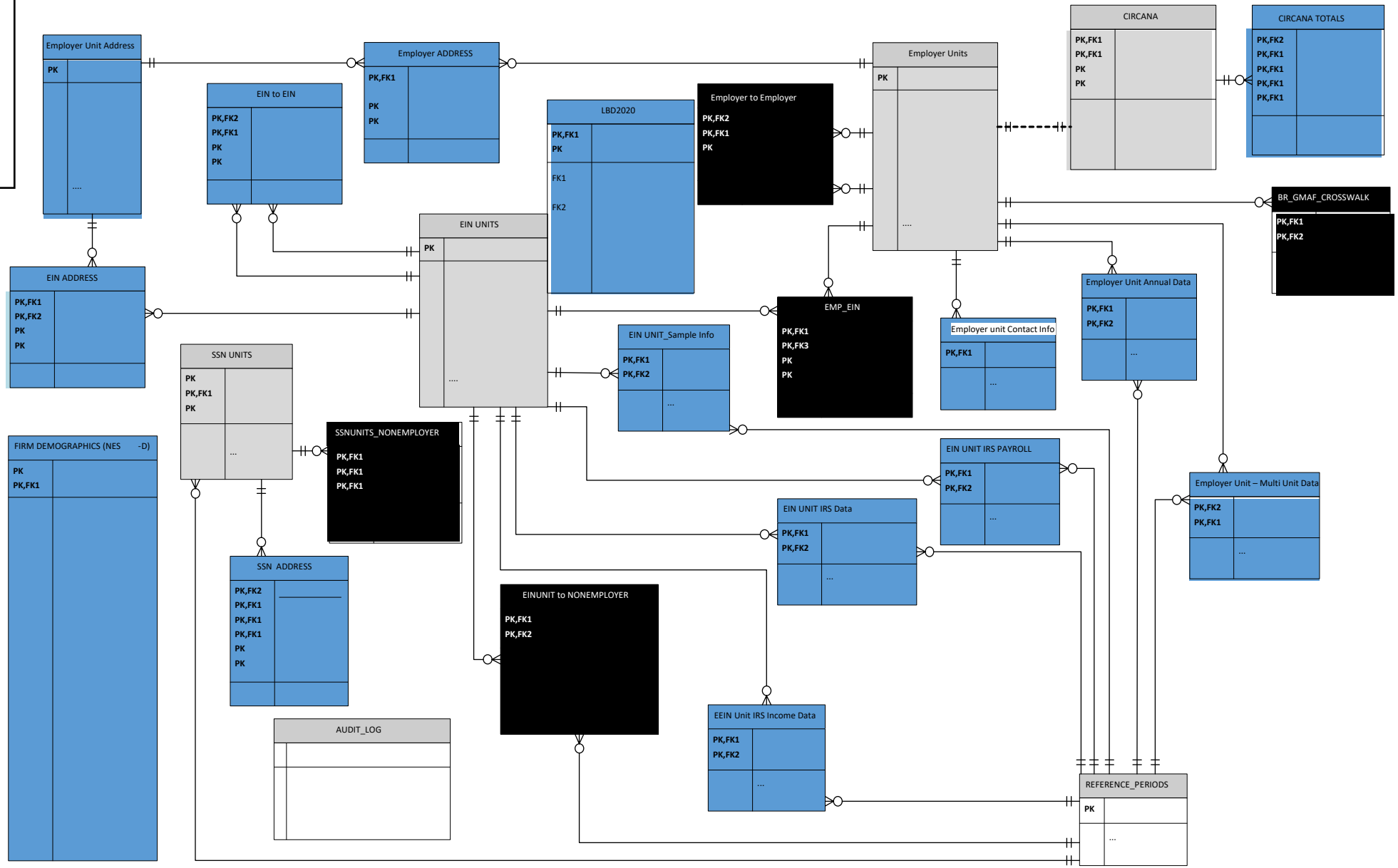
KEY

Data

Unit

Integration Point

- 5 Data Sources
- 43 Tables
- 1141 Columns



United States® Census Bureau

Assessing the Challenges

Tangible

- Differences between data sources
  - Physical differences between files
  - Conceptual differences in definitions
- Laws and regulations vary across data sources
- Managing data access

Non-Tangible

- Maintaining Data Integrity
- Ensuring data is used responsibility

United States® Census Bureau

Questions

?

Jessica.L.Wellwood@census.gov
301-763-7211

United States® Census Bureau

Federal Committee on Statistical Methodology

- 2023 FCSM Research & Policy Conference
- October 24th – 26th , 2023
- College Park Marriott Hotel & Conference Center, Hyattsville, MD

Enabling Economic Statistics Modernization

Jessica Wellwood, U.S. Census Bureau
Erica Marquette, U.S. Census Bureau
Ali Obaidi, MITRE
Adrienne Chen-Young, MITRE

The U.S. Census Bureau is prototyping a vision of an integrated infrastructure containing a comprehensive list of government and non-government businesses, with their core attributes, in a Business Frame that will enable linkages to respondent data, administrative data, and other enterprise demographic, geospatial, and job data. The Business Frame will assess challenges with merging records from multiple sources using identity matching algorithms to disambiguate data and capitalize on the similarities in data currently distributed across the enterprise, to create a robust centralized repository defining the complex population and relationships among businesses. Existing statistical programs can leverage this integrated business environment as they look to meet the demand for cross-domain data and increased granularity of attributes needed for economic data products. This presentation will describe the motivation behind the vision to strengthen research capabilities and promote innovation via the Business Frame. Further it will discuss the approach and findings in designing the data architecture for this modernized environment that will expand the capacity to answer critical questions about the nation's economy.

United States® Census Bureau