Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

National Secure Data Service

Artificial Intelligence for Enhancing Data Quality, Standardization, and Integration

Data Transformation Toolkit

April 30, 2026

U.S. Department of Transportation
Office of the Secretary of Transportation

Bureau of Transportation Statistics

Project Overview

- **Project objective**: to develop tools and services, including the use of AI, to assist in the processing, formatting, standardization, and integration of data, as well as generating metadata, to support statistical agencies' production of high-quality statistics and data products. This project was in support of the National Secure Data Service (NSDS) Demonstration project.

- **Technical lead**: Bureau of Transportation Statistics

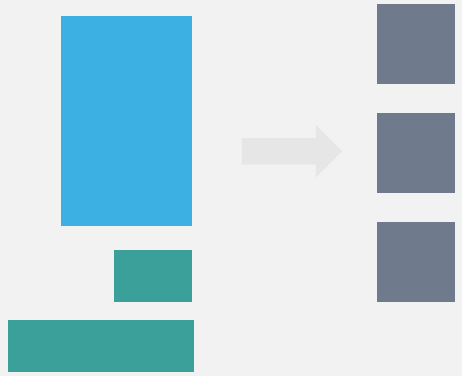
- **Vendor**: NORC / Subcontractor: Wolfram Research(University of Illinois Discovery Partners Institute

- **Period of Performance**: September 13, 2024, to April 1, 2026

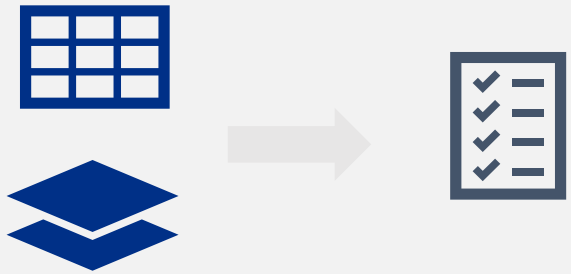
- **Main product/service**: Data Transformation Toolkit

Motivating Use Cases

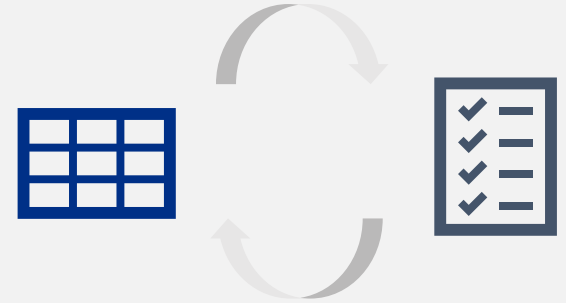
- *Case 1*: a statistician or analyst would like to use data that is in a difficult format (e.g., pdf file, non-tabular format)



- *Case 2*: a statistician or analyst would like to augment survey data with location-based data



- *Case 3*: a statistician or analyst has hundreds of data variables and would like to search for missing values, outliers, inconsistencies, etc.



Data Transformation Toolkit

- Five tools were developed in response to a needs assessment summarized in the framework plan report: https://www.americasdatahub.org/award-ai-dqsi-24/
  - Free Text Encoder (AI/LLM based)
  - Tabular Data Extractor (AI/LLM based)
  - Missing Data Assistant
  - Metadata Extractor (AI/LLM based)
  - Data Harmonizer (AI/LLM based)

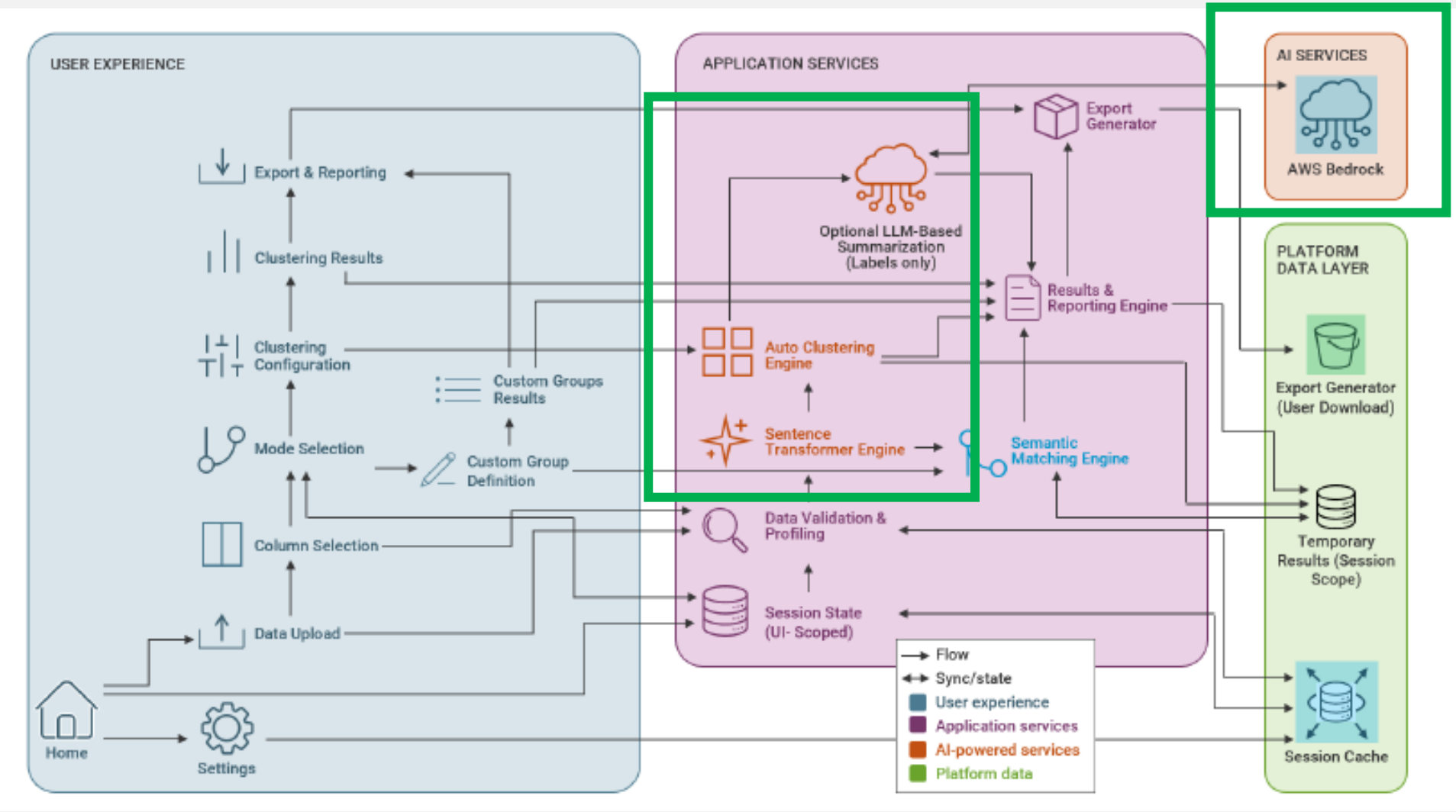
Free Text Encoder

Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

National Secure Data Service

- ***Problem***: difficulty of analyzing free-text fields in survey and administrative data at scale.
- Manual coding of these responses is time-consuming, inconsistent, and often infeasible for large datasets.
- ***Solution***: This tool uses AI to enable users to cluster or categorize free-text fields using either unsupervised clustering algorithms or user-defined custom groups.

Free Text Encoder



Bureau of Transportation Statistics

Tabular Data Extractor

Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

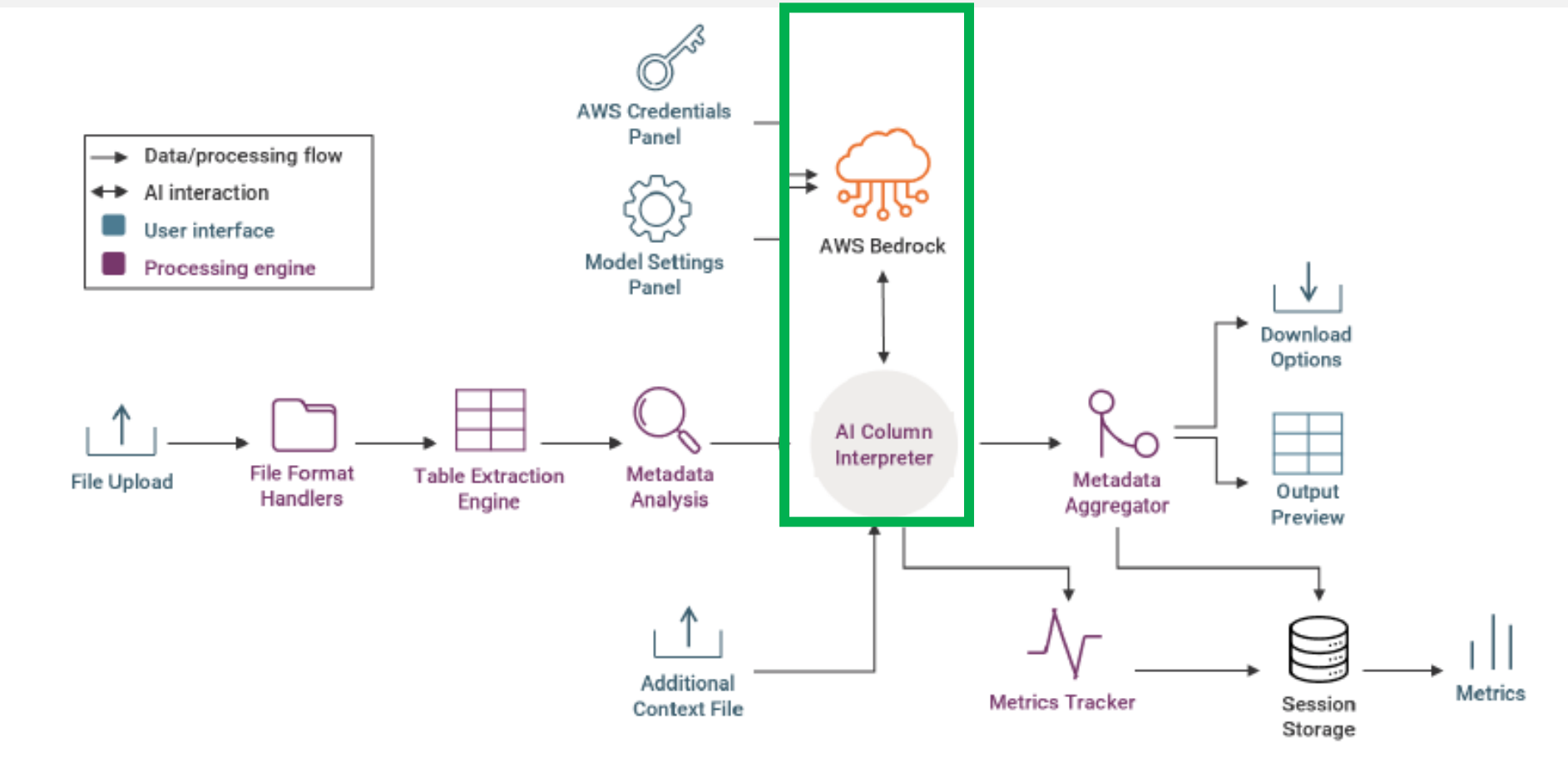
National Secure Data Service

- ***Problem***: the need to extract structured information from semi-structured or unstructured documents, such as regulatory filings and public reports.
- Federal statistical agencies often rely on manual review to identify relevant facts, figures, and metadata from these documents – a process that is labor-intensive, inconsistent, and prone to error.
- ***Solution***: This tool leverages AI to enable users to extract structured answers and tabular data from semi-structured documents (e.g., PDFs, TXT files).

Tabular Data Extractor

Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

National Secure Data Service



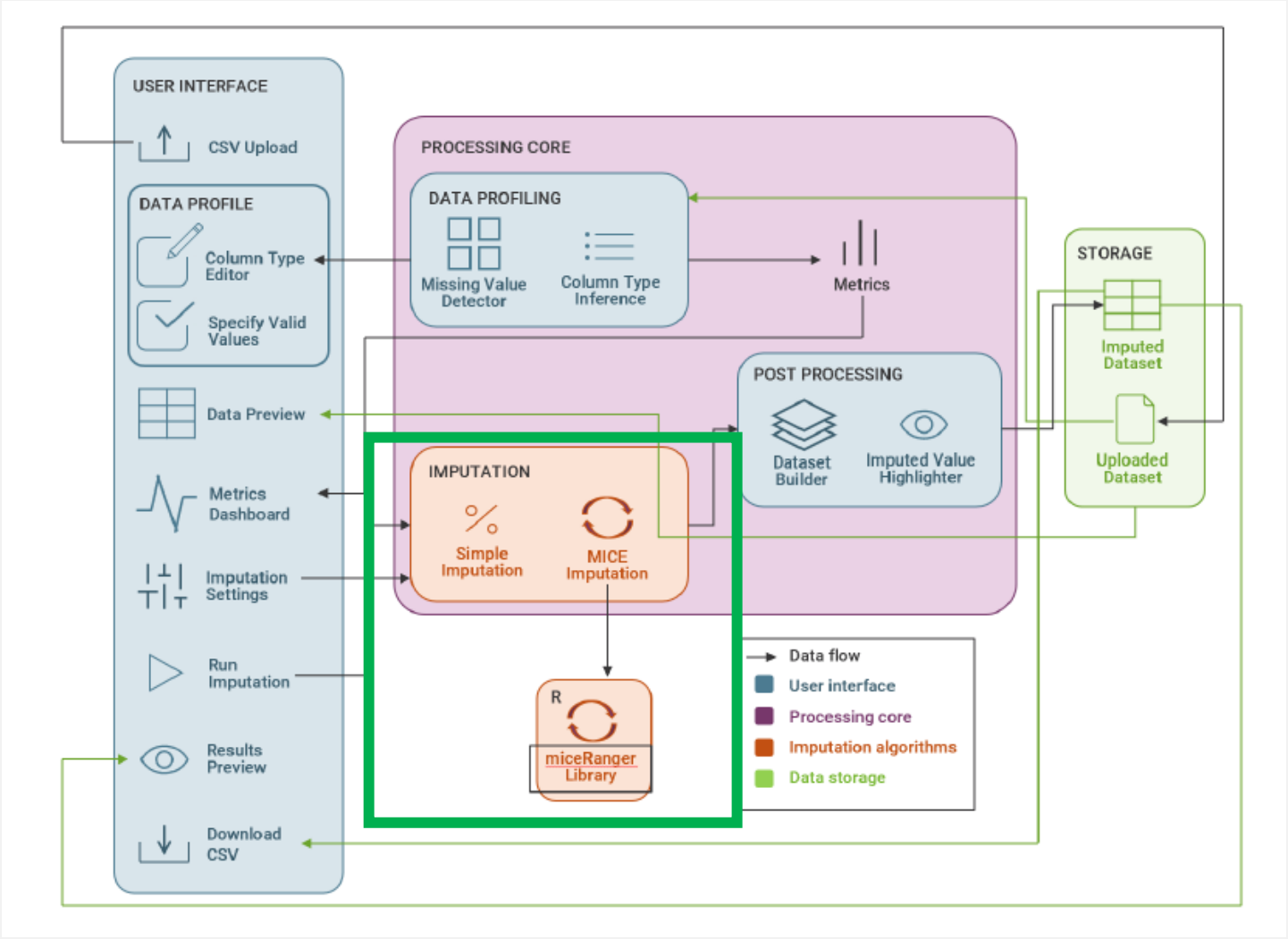
Missing Data Assistant

Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

National Secure Data Service

- ***Problem:*** the need for scalable, transparent, and accurate methods to handle missing data in administrative and survey datasets.
- Missing data is a near-universal issue in federal statistical work, and improper handling can introduce bias, reduce statistical power, and compromise the validity of analyses.
- ***Solution***: This tool uses a browser-based interface to offer more sophisticated imputation methods, while providing users with control over strategy selection and diagnostics to assess imputation quality.

Missing Data Assistant



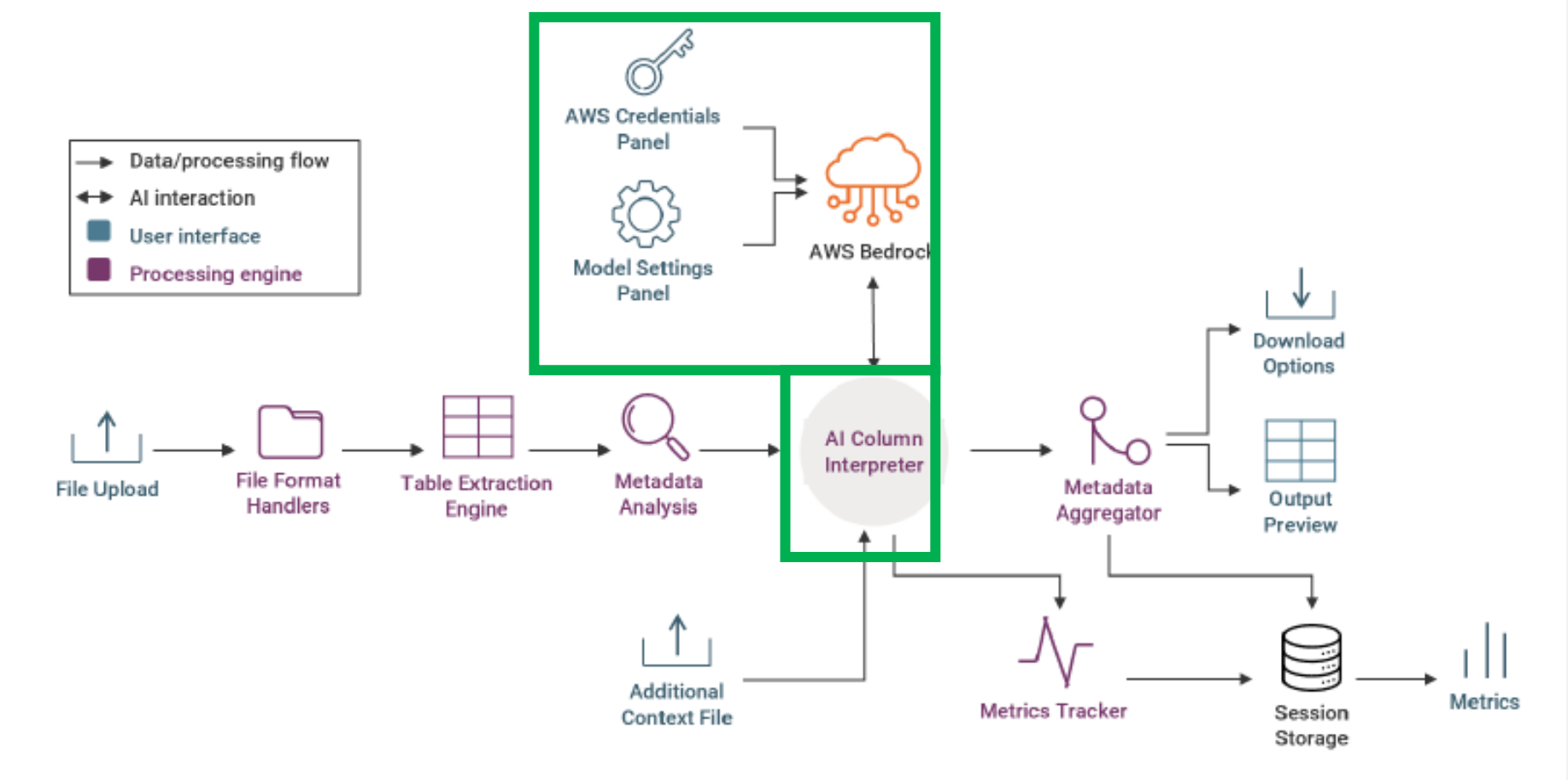
Metadata Extractor

Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

National Secure Data Service

- ***Problem***: lack of standardized, complete, and machine-readable metadata for tabular datasets, particularly from non-traditional sources.
- Interviews with federal statistical agencies revealed that extracting metadata is time-consuming, error-prone, and inconsistent.
- ***Solution***: This tool leverages AI to extract metadata from documents containing tabular data including Excel, CSV, and PDF files.

Metadata Extractor



Data Harmonizer

Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

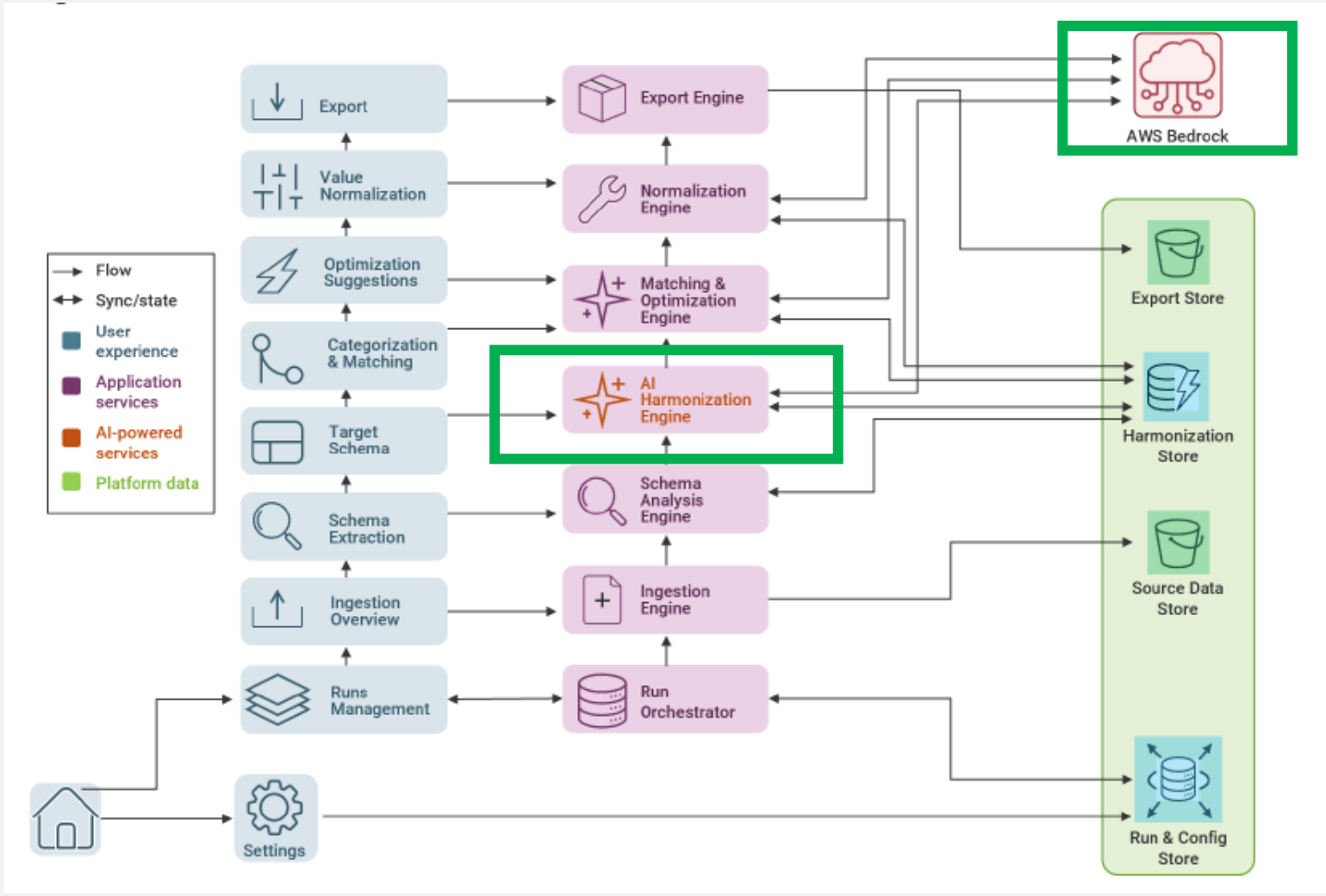
National Secure Data Service

- ***Problem***: often difficult to align and standardize tabular data from multiple sources that represent similar concepts but differ in structure, naming conventions, or formatting.
- Federal statistical agencies often receive administrative data from states, localities, or other partners in inconsistent formats.
- ***Solution***: This tool uses AI to enable users to align and harmonize multiple tabular datasets by extracting schemas, defining or generating a target schema, grouping similar fields, and matching values across files.

Data Harmonizer

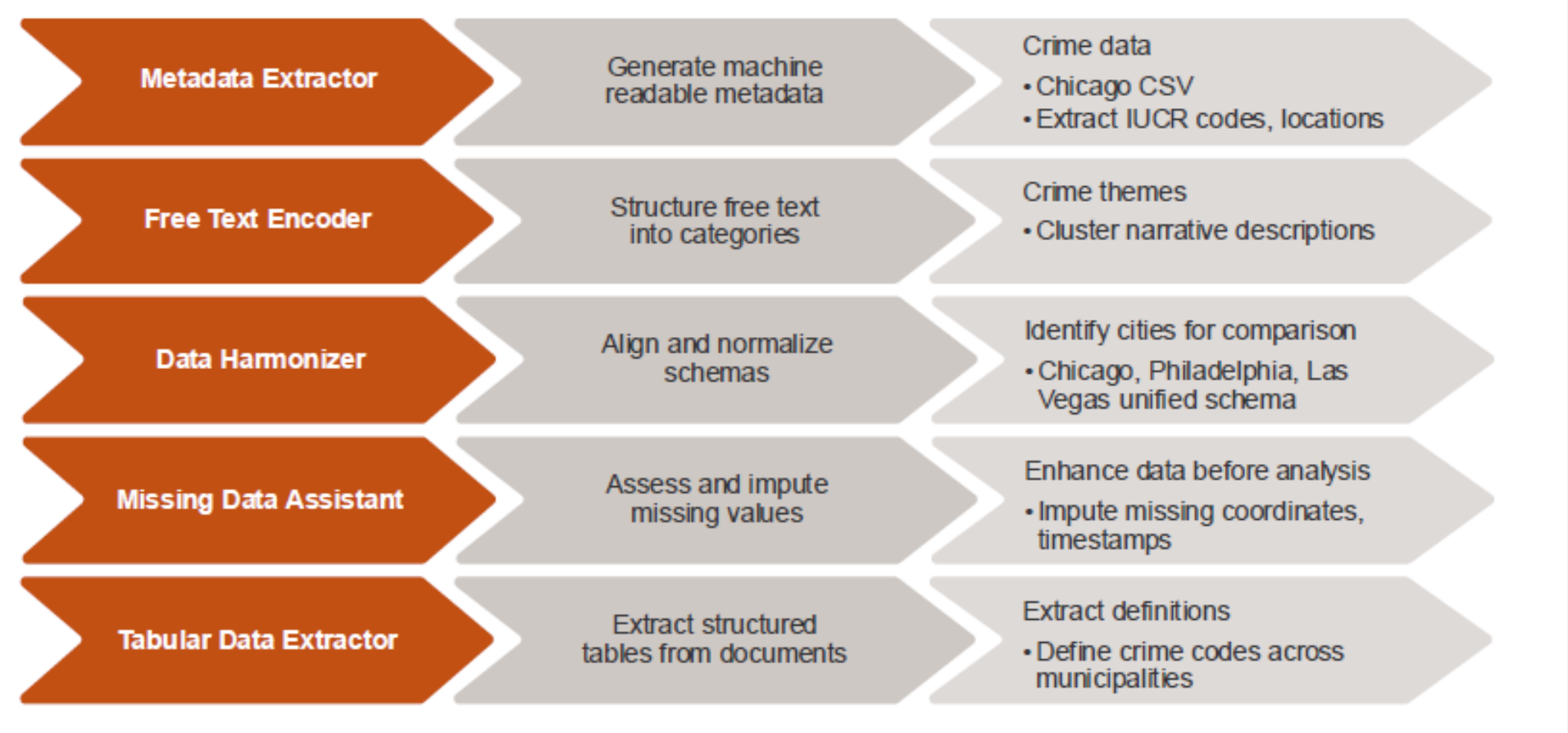
Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

National Secure Data Service



Bureau of Transportation Statistics

Fitting Together



Lessons Learned

Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

National Secure Data Service

- AI is most effective when applied upstream
- Metadata quality is foundational
- Human-in-the-loop, transparent design
- Modular, task-specific AI tools outperform general-purpose

Thank You

- Carlos Restrepo, PhD- Heather Madray, PhD- and Chris Mayfield (NCSES)

- Ramond Robinson, Hyun Kim, and Novin Ghaffari, PhD(BTS)

- Emily Wiegand, Mehmet Celepkolu, Zachary Seeskin, Lilian Huang, Sara Lafia, and Beth Fisher (NORC)

- Karl Isensee (Wolfram) and Anuj Tiwari (University of Illinois Discovery Partners Institute).

This project was supported by ADC on behalf of the National Center for Science and Engineering Statistics in the U.S. National Science Foundation.