BrightQuery®
The Factual AI Company™

AI Readiness Case Study: Bureau of Economic Analysis

AI-Ready Data Products to Facilitate Discovery and Use

AI-RD–24
Jan 3, 2025

Report Authors
Jose M. Plehn-Dujowich, Ph.D., CEO and Founder, (310) 448-3157, jplehn@brightquery.com

Chris Shaffer, VP of Technology
chris@brightquery.com

Peyton Marsh, VP of Communications
peyton@brightquery.com

Table of Contents

America's DataHub Consortium (ADC), a public-private partnership, implements research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). These results document research funded through ADC and are being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed in this report do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. NCSES has reviewed this product for unauthorized disclosure of confidential information and approved its release (NCSES-DRN26-055).

Glossary

- **AI (Artificial Intelligence):** A field of computer science that enables machines to perform tasks that typically require human intelligence, such as understanding natural language, recognizing patterns, or making decisions.
- **AI Readiness:** The process of preparing datasets to be machine-readable and machine-understandable, enriched with metadata, and organized in a standardized format to optimize their usability by AI technologies.
- **API (Application Programming Interface):** A set of rules and tools that allows different software applications to communicate and share data. The BEA API enables users to retrieve BEA datasets programmatically.
- **Croissant Tagging:** An advanced metadata framework designed for AI and machine learning applications. It helps datasets become more discoverable and usable by enabling context-aware data interpretation.
- **Data Dictionary:** A centralized document or resource containing definitions of data items, descriptions of parameters, and metadata to help users understand the structure and meaning of a dataset.
- **Digitization:** The process of converting information from physical or analog formats into digital formats that can be accessed and analyzed using computers.
- **Granularity:** The level of detail in a dataset. High granularity means data is broken down into finer details, such as by geography (e.g., county level) or time (e.g., monthly).
- **Interactive Data Tools:** Web-based tools that allow users to explore datasets dynamically by applying filters or generating visualizations, such as tables, charts, or graphs.
- **Interoperability:** The ability of different systems, datasets, or tools to work together seamlessly, often by adhering to standardized formats or metadata structures.
- **Metadata:** Data that provides information about other data, such as the title, description, units, and date of publication. Metadata helps users and AI systems understand the content and context of datasets.
- **NAICS (North American Industry Classification System):** A standard used by federal agencies to classify businesses and industries. It organizes sectors hierarchically for consistent reporting and analysis.
- **PII (Personally Identifiable Information):** Information that can identify an individual, such as a name, address, or Social Security number. BEA does not publish PII, ensuring privacy.
- **Schema.org Tagging:** A semantic vocabulary that provides metadata for web content. It enhances search engines' ability to understand the context of datasets and improve their discoverability.
- **Semantic Interoperability:** The ability of different datasets to be linked and understood in terms of their meaning and context, often facilitated by metadata standards like Schema.org.
- **Structured Data:** Data organized into a defined format, such as rows and columns in a table or fields in a database, making it easily machine-readable.
- **Unstructured Data:** Data that does not have a predefined structure, such as text documents or reports, requiring additional processing to extract information.
- **Version Control:** The process of tracking and managing changes to datasets over time, including initial releases, revisions, and final versions.
- **Visualization:** The graphical representation of data, such as charts, graphs, or maps, to make patterns and trends easier to interpret.
- **Web Scraping:** The automated process of extracting data from websites. It can be challenging for tools to process data not explicitly structured or tagged.
- **XML (Extensible Markup Language):** A markup language that defines rules for encoding data in a structured format, often used for sharing information between systems.

Executive Summary

The Department of Commerce describes machine-understandable, "AI-ready" data as "data that [are] enriched with contextual metadata and organized in interpretable standard formats." By making its data AI ready, "AI models can then better interpret data, link them to similar data, and return accurate results from authoritative sources. This report provides an in-depth landscape analysis of the Bureau of Economic Analysis (BEA) functional content—APIs, Reports & Unstructured Data, Tables, Interactive Charts, and Website & Metadata, evaluating their readiness for integration into Artificial Intelligence (AI) and Machine Learning (ML) frameworks. As the role of AI continues to expand in transforming data into actionable insights, ensuring BEA content meets the stringent requirements of AI systems is critical to unlocking their full potential.

The objective of this evaluation is twofold: first, to analyze the current state of BEA's information published in terms of structure, metadata, interoperability, and other critical AI readiness criteria; and second, to identify actionable steps that can enhance their utility in AI applications. By doing so, the report will demonstrate how we intend to bridge the gap between traditional data practices and the advanced capabilities of AI technologies, **enabling search engines such as Google, and Large Language Models like ChatGPT to cite BEA's statistical products more accurately.** Title II of the Evidence Act (or the "OPEN Government Data Act") "requires public government data assets to be published as machine-readable data and also machine understandable. In order to accomplish this, BrightQuery will help the BEA place statistical data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost.

BrightQuery's efforts are supported by a team of esteemed consultants and strategic partnerships with leaders in AI and data technologies. BrightQuery also partners with Google Data Commons to enhance data linkage capabilities. Through the AI Alliance, BrightQuery works with IBM and Meta as part of the Open Trusted Data Initiative to establish standards for secure and authoritative AI-ready data. These collaborations underscore BrightQuery's ability to deliver innovative and robust solutions for advancing BEA's AI readiness.

BrightQuery will incorporate BEA's high-priority datasets into its case study to showcase actionable solutions for AI readiness. These datasets and their related tools and reports, including Corporate Profits, GDP at various geographic levels, Personal Income and Outlays, and U.S. International Trade and Transactions, will serve as the foundation for demonstrating best practices in metadata enrichment, semantic tagging, and interoperability improvements.

Throughout this assessment, we outline the strengths of BEA published information while identifying opportunities for improvement to align them with the standards necessary for advanced AI tools. This work also introduces key strategies and methodologies that can aid BEA in transitioning its published content toward a more AI-ready framework, ensuring seamless integration and optimal performance in AI-driven environments.

The Importance of AI Readiness. Advancements in artificial intelligence (AI) are transforming how data is accessed and utilized. For the Bureau of Economic Analysis (BEA), ensuring its datasets are AI-ready enhances their usability, accessibility, and relevance in today's data-driven world.
AI readiness involves two key elements:

- **Machine-Readability:** Structuring data in formats like APIs and tables that can be processed automatically.
- **Machine-Understandability:** Enriching data with metadata, such as definitions and units, to enable accurate interpretation by AI systems.

BEA's datasets underpin critical economic analyses, from GDP to trade statistics. By improving metadata and adopting frameworks like Schema.org and Croissant tagging, BEA can make its data more discoverable and interoperable with other systems, increasing its value for policymakers, researchers, and the public.

This report outlines a roadmap for BEA to enhance AI readiness by addressing gaps in metadata, structure, and accessibility. By adopting these changes, BEA can set a standard for transparent, AI-optimized federal data systems that empower economic decision-making.

Sources and Types of Information Analyzed. This report evaluates a wide range of datasets and methodologies related to AI readiness, including BEA's existing datasets, metadata standards, and data structures. Key information sources include time-series datasets, relational databases, and semi-structured and unstructured data formats. In addition, the report considers methodologies like triplification for semantic data organizationand tie-series modeling strategies for temporal data alignment. External reference materials and industry best practices were also analyzed to provide comparative insights, including data interoperability frameworks, metadata schema guidelines, and AI/ML feature engineering techniques.

The analysis leverages insights from non-governmental datasets and private-sector practices, such as financial data integration platforms, knowledge graph implementations, and AI-driven data processing pipelines. These provide a benchmark for understanding the current landscape and setting actionable objectives for BEA's data readiness.

Restatement of Objectives. The primary objective of this project is to explore how a future National Secure Data Service (NSDS) could provide shared information and tools for making statistical data products more readily ingestible by AI technologies. Generative AI (GenAI) is transforming industries and influencing the daily decisions of Americans. Typically, GenAI models are trained on a limited set of data, often excluding the most recent information. To make GenAI truly resourceful with the latest data, it is necessary to enhance these models with additional datasets using technologies such as Retrieval-Augmented Generation (RAG), Text2SQL, and model fine-tuning. These technologies significantly expand the knowledge base of GenAI models. However, these processes impose stringent requirements on the datasets used for augmentation, including data format, schema, context, provenance, and metadata. The same challenges apply to statistical datasets produced by government agencies, including BEA. The BrightQuery (BQ) team proposes developing an open-source platform called **Government Data Agent (GDA)** with two components: **GDA Evaluator** to assess the AI readiness of a dataset; and **GDA Transformer** to perform the necessary transformations recommended by the evaluator. This multi-agent system powered by a Large Language Model (LLM) will intelligently assess if a dataset is AI-ready and then transform problematic datasets to meet AI standards. GDA will primarily serve federal agencies by evaluating, transforming, and upgrading their datasets to meet the demands of the rapidly growing GenAI applications. This ensures that the information retrieved by GenAI applications is accurate, trustworthy, and traceable to its original source.

Background on AI Readiness. To be AI-ready, datasets and tools must meet several critical criteria that ensure compatibility with advanced AI and ML frameworks. AI readiness encompasses the following key attributes:

- **Structured Data Formats**: Data should be structured in formats that enable seamless integration with AI systems, such as relational databases, JSON-LD, or RDF formats for knowledge graphs.
- **Comprehensive Metadata**: Metadata provides essential context, including provenance, schema details, and update logs, ensuring datasets can be accurately interpreted and processed by AI models.
- **Interoperability**: Datasets must integrate easily with other systems or datasets, often achieved through standardization and adherence to widely accepted data exchange formats.
- **Granularity and Scalability**: AI systems require data at granular levels to support diverse use cases, along with scalability to accommodate increasing data volume and complexity.
- **Temporal and Contextual Consistency**: For time-series data, such as economic indicators, datasets must represent both the time of data capture and the time to which the data applies. This ensures accuracy in temporal analyses.
- **Semantic Enrichment**: Techniques like triplification enable datasets to include explicit relationships between entities, enhancing machine understanding.

Examples of Successful Implementations. Outside of government applications, several organizations have successfully implemented AI-ready datasets and tools, including:

- **Google Knowledge Graph**: Utilizes structured triples to connect and enhance search results, demonstrating the power of semantic data organization.
- **FactSet**: A leading provider of extensive financial data and analytics, **FactSet relies on BrightQuery's structured and real-time data feeds** to power its platforms. By integrating BrightQuery's comprehensive datasets, FactSet enhances its ability to deliver actionable insights and enable decision-making in capital markets..
- **LinkedIn Economic Graph**: Maps global workforce dynamics through structured and semi-structured datasets, supporting AI-driven recommendations and insights.

For further insights, recommended resources include:

"The Role of Knowledge Graphs in AI" (https://www.croissant.ai/knowledge-graphs)
"Entity-Attribute-Value Model for Semantic AI" (https://en.wikipedia.org/wiki/Entity%E2%80%93attribute%E2%80%93value_model)
Google's Data Commons Framework (https://datacommons.org/).

About This Report.. This report covers a human-prepared evaluation of the readiness of BEA data sets for ingestion by AI and LLM tools. In addition to offering this assessment and providing a starting point for estimation of the effort required to transform and prepare BEA data sets, this will serve to inform the design of the automated tool that we'll be developing in Q2 of 2025.

Request for Feedback. We are actively seeking validation of the approach we've taken by hand before we automate it and roll that methodology out across additional agencies. We expect the scoring methodology described below to change slightly over the next 3 months. We hope, *primarily*, to achieve completeness – covering all aspects that are necessary or useful for machine ingestion and model training – and, *secondarily*, to rank and score those aspects in order of importance.

BEA High-Priority Datasets and Related Tools

The following high-priority datasets identified by the BEA of national and regional economic analysis, provide critical insights into the economic health of the nation, its states, regions, and territories. These datasets, along with the associated reports, tools, and methodologies, are essential for understanding and addressing the complexities of economic performance, policy impacts, and resource allocation.

Key Datasets and Their Significance

1. **Corporate Profits:** Offers detailed insights into business earnings, helping analysts evaluate industry performance and economic resilience. Related tools include historical trend visualizations and profit distribution models.
2. **GDP (Gross Domestic Product):** The cornerstone of economic measurement, GDP quantifies the overall economic activity of the nation. Supporting reports include GDP revisions and projections, accompanied by tools for time-series analysis.
3. **GDP by County and Metro:** Provides a granular view of economic activity at the local level, vital for regional planning and investment decisions. Interactive maps and regional economic modeling tools are integral to leveraging this dataset.
4. **GDP by State:** Highlights state-level economic performance, enabling comparisons and policy analysis. Complementary resources include state economic reports and benchmarking tools.
5. **Personal Consumption Expenditures by State:** Tracks spending trends across states, offering critical data to understand consumer behavior. Tools like expenditure comparison dashboards enhance usability.

6. **Personal Income and Outlays:** Captures household earnings and expenditures, critical for evaluating the standard of living and economic stability. Accompanying reports include income trends and policy impact assessments.
7. **Personal Income by County, Metro, and Other Areas:** Provides detailed income data for smaller geographic units, essential for local policy development. Related resources include small-area income estimators and visualization tools.
8. **Personal Income by State:** Assists in understanding state-level income distributions and trends, informing state budget planning and economic initiatives. Income growth analytics tools often accompany this dataset.
9. Puerto Rico GDP: Critical for assessing the economic status of U.S. territories, providing insights into growth, challenges, and opportunities. Supporting tools include cross-territory economic comparison platforms.
10. **Regional Price Parities:** Enables cost-of-living comparisons across regions, aiding policymakers and businesses in decision-making. Tools for inflation adjustment and purchasing power parity analysis are key additions.
11. **U.S. International Trade in Goods and Services:** Tracks trade flows and balances, providing insights into the nation's global economic engagement. Valuable here are trade flow dashboards & trend analysis tools.
12. **U.S. International Transactions:** Highlights balance-of-payments data, critical for understanding financial interactions with global markets. Complementary tools include transaction mapping and international finance models.

The BQ AI Readiness Scorecard

The BQ AI Readiness Score Table below provides a comprehensive evaluation of BEA's key functional content areas—APIs, Reports & Unstructured Data, Tables, Interactive Charts, and Website & Metadata. Each area is assessed against critical dimensions, including Scope, Data Structure & Format, Metadata & Context, Interoperability & Granularity, Technical Implementation, AI/ML-Specific Features, and Governance & Compliance.
These functional areas form BEA's data ecosystem, supporting a wide range of users, from policymakers and analysts to developers and AI systems. Evaluating their readiness for AI ensures that BEA's datasets and tools can meet the growing demands of data-driven innovation. By addressing gaps and leveraging opportunities for improvement, BEA can enhance the usability, accessibility, and impact of its data assets.
Here we delve into each functional area, offering insights into their current capabilities, challenges, and potential for transformation. This structured analysis provides actionable recommendations to help search engines such as Google, and Large Language Models like ChatGPT cite BEA's data more accurately.

Functional Content Areas

1. APIs

APIs are the backbone of modern data interoperability, offering programmatic access to datasets for seamless integration into external applications. They enable users to query, retrieve, and manipulate data dynamically, making them indispensable for building scalable, AI-powered systems. For BEA, APIs serve as a critical gateway for external developers and analysts for faster integration into third-party platforms for real-time economic forecasting, allowing them to interact with structured economic data. Their effectiveness lies in their ability to support complex queries while maintaining data integrity and performance.

- **Scope**: APIs provide programmatic access to BEA datasets, enabling integration into external applications and platforms.
- **Data Structure & Format**: The data accessed via APIs is primarily structured, using standardized formats such as JSON and XML for seamless machine consumption.
- **Metadata & Context**: Metadata is included but requires enrichment, such as detailed field descriptions and update frequencies, to enhance usability.
- **Interoperability & Granularity**: APIs are interoperable with various systems but could benefit from finer granularity to allow more specific queries (e.g., down to individual metrics).

- **Technical Implementation**: APIs demonstrate robust functionality with endpoint reliability but require more advanced features such as batch data retrieval and real-time updates.
- **AI/ML-Specific Features**: Current APIs lack advanced AI features such as semantic search or predictive query suggestions, limiting their utility in AI/ML applications.
- **Governance & Compliance**: APIs meet security standards but would benefit from improved compliance tracking for changes in data schemas or updates.

Learn more: Understanding RESTful APIs for foundational knowledge. Google Data Commons API for an example of effective public APIs.

2. Reports & Unstructured Data

Reports, such as narrative documents and policy analyses, capture valuable qualitative insights which complement quantitative data. These reports often include contextual explanations, detailed interpretations, and domain-specific knowledge critical for a comprehensive understanding of economic trends. However, their unstructured nature poses challenges for AI systems, which thrive on structured data. Transforming these reports into AI-ready assets, through methods like triplification and semantic tagging, can unlock their potential for machine-driven insights and enhanced decision-making.

- **Scope**: Reports & unstructured data include economic analyses, policy briefs, and other narrative documents relevant to BEA's datasets.
- **Data Structure & Format**: These reports are primarily text-based, with minimal structure, often lacking standardized formats for machine readability.
- **Metadata & Context**: Metadata for unstructured reports is minimal, with limited details on authorship, timestamps, or topic tagging.
- **Interoperability & Granularity**: Poor interoperability due to a lack of semantic annotations or structured formats, making it difficult to extract granular insights from these reports.
- **Technical Implementation**: Reports lack integration into structured repositories or platforms that facilitate indexing and searching for AI applications.
- **AI/ML-Specific Features**: Minimal AI readiness; triplification or knowledge graph enrichment would significantly enhance their utility for machine learning systems.
- **Governance & Compliance**: Requires improvement in documenting provenance and ensuring compliance with accessibility and security standards.

Learn more: Introduction to Semantic Annotation for enriching text documents. The Role of Knowledge Graphs in AI.

3. Tables

Tables represent one of the most fundamental and widely used data formats in BEA's ecosystem. They organize quantitative and categorical data into structured formats, making them easily accessible for traditional analyses and machine learning models. Tables provide a foundation for AI applications, as their structure inherently supports efficient data processing. Enhancing their metadata and granularity can elevate their utility, allowing for richer, more nuanced insights to be derived from AI systems.

- **Scope**: Tables serve as structured datasets containing quantitative and categorical data, fundamental for analysis and modeling.
- **Data Structure & Format**: Highly structured, using formats like CSV or relational databases, but may lack consistency across tables.
- **Metadata & Context**: Metadata provides basic schema descriptions but needs expansion to include data lineage, assumptions, and update cadences.
- **Interoperability & Granularity**: Tables support interoperability but require granular data elements (e.g., cell-level metadata) to enhance AI compatibility.
- **Technical Implementation**: Tables are technically sound but could benefit from advanced querying capabilities, such as relational links to other datasets.
- **AI/ML-Specific Features**: High potential for AI applications if enriched with semantic tagging or relationships to related datasets.

- **Governance & Compliance**: Well-governed but requires stronger adherence to open data standards for broader reuse and compliance with interoperability mandates.

Learn more: Data Cleaning Techniques for improving structured data. Schema.org for metadata standards.

4. Interactive Charts

Interactive charts offer dynamic visualization tools which bring BEA's datasets to life. By enabling users to explore trends, patterns, and relationships in real time, these charts transform raw data into an engaging, user-friendly experience. For AI applications, however, these visualizations often lack the structure and accessibility required for direct integration. Improving the underlying data connections and embedding AI-ready features into these charts can enhance their role as a bridge between data exploration and actionable insights and improve public engagement through AI-driven predictive visualizations.

- **Scope**: Interactive charts provide dynamic visual representations of BEA data for exploration and analysis.
- **Data Structure & Format**: Underlying data is structured, but visualization outputs are often proprietary or embedded in non-machine-readable formats.
- **Metadata & Context**: Metadata is limited, with little information on how charts are generated or linked to underlying datasets.
- **Interoperability & Granularity**: Charts lack interoperability as they are often static or embedded, restricting granularity and machine-driven manipulation.
- **Technical Implementation**: Current implementations focus on user interactivity but lack APIs or export functions for integration into AI workflows.
- **AI/ML-Specific Features**: Limited AI readiness; enabling automated insights or predictive features based on chart data could improve utility.
- **Governance & Compliance**: Requires enhancements to ensure charts adhere to accessibility guidelines and maintain traceability to data sources.

Learn more: D3.js Documentation for charting interactivity. Charting and AI Integration for bridging visualization and machine learning.

5. Website & Metadata

The BEA website serves as the central hub for disseminating data to the public and stakeholders. It hosts datasets, tools, and metadata essential for both human users and AI systems. High-quality metadata is the linchpin of discoverability and usability, providing the context necessary for effective data integration and application. By enhancing the website's metadata and structural elements, BEA can ensure its digital assets are both accessible and actionable, driving innovation and adoption across AI-powered systems.

- **Scope**: The BEA website acts as the central hub for dataset distribution, public interaction, and metadata publication.
- **Data Structure & Format**: The website uses mixed formats, ranging from structured (HTML metadata tags) to unstructured content.
- **Metadata & Context**: Metadata quality varies; while some areas are well-documented, others lack sufficient descriptions or semantic tags for machine consumption.
- **Interoperability & Granularity**: The website supports basic interoperability, but limited granularity hampers precise data discovery and integration into external systems.
- **Technical Implementation**: A technically robust platform but would benefit from semantic enhancements (e.g., schema.org tags) and machine-readable sitemap extensions.
- **AI/ML-Specific Features**: Lacks advanced metadata or semantic features to support AI-based indexing or reasoning.
- **Governance & Compliance**: Generally compliant with accessibility standards but requires better governance for metadata updates and schema changes.

Learn more: Schema.org Markup Guide for semantic web best practices. Open Government Data Principles for aligning with accessibility standards.

Table 1. BQ AI Readiness Scorecard

✓ Complete / Good

✗ Missing / Not usable

◔ ◑ ◕ Partial (slightly, somewhat, mostly)

NA Not applicable

? Unable to determine

	Functional Area				
	APIs	**Reports & Unstructured Data**	**Tables**	**Interactive Charts**	**Website & Metadata**
Scope					
Breadth	◑	✓	✓	◑	◔
History	◑	◔	◑	◑	◔
Data Structure & Format					
Digitization	✓	✓	✓	✓	◕
Standardized Schema	✓	NA	◔	NA	◕
Standardized Layout	NA	◕	NA	✓	◕
Tabular Layout	NA	✓	✓	✓	✗
Field/Column Labels	◑	✓	✓	✓	✗
Metadata & Context					
Units Discernable	✓	✓	✓	✓	NA
Data Dictionary	✓	◕	◑	◑	NA
Methodology Described	◑	✓	◑	◑	NA
Release Date/Time	◔	✓	◔	◔	◔
Interoperability & Granularity					
Geographical	◑	◕	◑	◑	NA
Industry/Sector	◑	◕	◑	◑	NA
Reporting Period	◔	◔	◔	◔	NA
Methodology	NA / ◑	NA / ◕	NA / ◑	NA / ◕	NA / ◕
Technical Implementation					
Direct Access Links	✓	✓	✓	◑	◕
Documentation	✓	NA	NA	✓	NA
Rate Limiting	◑	?	?	?	?
Authentication Methods	◑	✓	✓	✓	✓
AI / ML Specific Features					
Schema.org Tagging	✗	✗	✗	✗	✗
Croissant Tagging	✗	✗	✗	✗	✗
Training Data Labels	?	?	?	?	?
Bias Documentation	◔	◔	◔	◔	◔
Governance & Compliance					

Version Control	✗	✓ ≥ 1996	✓ ≥ 2002	✗	✗/✓
Usage Rights	✓	✓	✓	✓	✓
Privacy Annotations	NA	NA	NA	NA	NA
Individual Linkages Possible	✗	✗/?	✗/?	✗	NA
PII Action Required	NA	NA	NA	NA	NA

Table 2. BEA's AI Readiness - Summary KPIs

By Functional Area		By Category	
Overall	66%	**Overall**	68.14%
APIs	67%	**Scope**	69%
Reports & Unstructured Data	Recent data: **77%** prior to 1996: **72%**	**Data Structure & Format**	83%
Tables	Recent data: **72%** prior to 2002: **67%**	**Metadata & Context**	76%
Interactive Charts	69%	**Interoperability & Granularity**	70%
Website & Metadata	42-49%	**Technical Implementation**	94%
		AI / ML Specific Features	7%
		Governance & Compliance	Recent data: **78%** prior to 1996: **56%**

By Objective	
Overall	67.75%
Search Engines *Website & Metadata, Reports & Unstructured Data* *Scope (All), Digitization, Field/Column Labels, Data Dictionary, Direct Access Links, Rate Limiting, Authentication Methods, Schema.org Tagging, Privacy, PII*	62%
LLMs *Website & Metadata, Reports & Unstructured Data* *Scope (All), Data Structure & Format (All), Metadata & Context (All), Direct Access Links, Rate Limiting, Authentication Methods, AI/ML Specific Features (All), Usage Rights, Privacy, PII*	62%
BrightQuery Chatbot Project (DAA-24) *All functional Areas* *Scope (All), Data Structure & Format (All), Metadata & Context (All), Interoperability & Granularity (All), Technical Implementation (All), Training Data Labels, Bias Documentation, Governance & Compliance (All)*	Recent data: **75%** prior to 1996: **72%**

By Objective

ML & Predictive Analytics *APIs, Tables* *Scope (All), Data Structure & Format (All), Metadata & Context (All), Interoperability & Granularity (All), Bias Documentation, Governance & Compliance (All)*	Recent data: **72%** prior to 2002: **69%**

Description and Importance of Criteria

Scope

Breadth. Being able to gather all necessary data from one functional area makes the ingestion process easier; having to collate data from multiple increases effort required. In addition, having all of the same data specified in multiple formats increases the ability of machine-learning systems to check their own work.Of course, if some data isn't available online at all, it can't be ingested.
Does the functional area cover all data sets made available? This evaluates whether all datasets in one format are also available in others. For example, the BEA's Corporate Profits dataset is available via API and as downloadable tables, ensuring uniform access.
Learn more: BEA API Access.

History. Determines if datasets include their complete historical range. For instance, BEA's GDP data includes figures dating back to 1929 in some formats, but original reports may only be available from 1996.
Learn more: BEA GDP Data Archive.

Data Structure & Format

Digitization. Data not digitized must be processed using OCR (optical character recognition) software. While this technology is relatively mature, it does require additional effort and processing power, while introducing a small chance of transcription errors. Verifies if data is alphanumeric and not an image. For example, historical BEA GDP reports from before 1996 often require OCR for digitization.
Learn more: Digitized BEA Reports

Standardized Schema. Checks if data adheres to a consistent schema. A consistent data schema which rarely or never varies allows for processing code to utilize a single, simpler, algorithm. This greatly increases testability, accuracy, and reliability. BEA's API employs a uniform endpoint structure (e.g., GetData), making it easy to process data programmatically.
Learn more: BEA API User Guide.

Standardized Layout. Evaluates whether datasets are consistently organized.A consistent layout and document/HTML structure which rarely or never varies allows for processing code to utilize a single, simpler, algorithm. While not quite as useful as a structured schema, this still greatly increases testability, accuracy, and reliability compared with no standardization at all. For instance, BEA's Personal Income tables typically use horizontal headers for years and vertical labels for regions, ensuring layout uniformity.
Learn more: BEA Personal Income Data.

Tabular Layout. Assesses whether data is presented in table format. Tabular layouts are simpler to process and verify, particularly when the layout is not otherwise standardized. BEA's international trade data uses clear tabular layouts for easy machine processing.
Learn more: BEA Trade Data Tables.

Field/Column Labels. Ensures clear naming conventions for columns. Clear labels can be read unambiguously by a machine, as opposed to labeling schemes requiring context and cross-referencing. For example, the "GeoFips" field in BEA tables corresponds to geographical codes but requires documentation for clarity.
Learn more: BEA API Field Descriptions.

Metadata & Context

Units Discernable. Verifies if units (e.g., billions of dollars, percentages) are clearly specified. This is particularly important when making comparisons or calculations across data sets, or when a natural language question requires conversion. For instance, BEA's Real GDP by State tables label units consistently, aiding accurate interpretation.
Learn more: BEA Real GDP Metadata.

Data Dictionary. Determines if comprehensive descriptions of terms are available. A data dictionary will greatly improve the performance and accuracy of lexical searches when a user doesn't know the exact terminology, or would like a list of choices, by providing descriptions (which can be further enriched with synonyms). BEA provides parameter lists and descriptions through its API, which can be compiled into a data dictionary.
Learn more: BEA API Parameters.

Methodology Described. Checks for detailed explanations of data collection methods. This is useful for simple presentation and explanation to users; it can also be used by more advanced LLM-based systems to speculate as to possible biases and the relevance of certain comparisons between data sets or across time. BEA's methodology papers outline survey questions and sampling for its datasets, like GDP.
Learn more: BEA Methodology Resources.

Release Date/Time. Assesses if the publication date and revision history are clear. This is crucial to any predictive or decision-making process. A reactive policy proposal ("do X when Y") can be simulated, and thus validated, based on past situations in which Y happened (What actions would we have taken had this policy existed historically? Are they the actions we'd want to take in similar circumstances in the future?). However, only by accurately modeling *when* Y was known can you be certain when (or if) your policy proposal would have resulted in X actually happening. For instance, BEA's news release calendar specifies dates for updates and revisions.
Learn more: BEA Release Schedule.

Interoperability & Granularity

Interoperability allows for making comparisons and drawing conclusions across multiple data sets. All other things considered, an interoperable data set along more dimensions at a higher granularity will be more useful in conjunction with others. It's unlikely there will be much to be done about non-interoperable data in the near-term. While a lack of interoperability doesn't affect the ability to ingest and analyze data in isolation, it limits the ability to ask questions like "how does [national trend] affect the economy in [city]" and get meaningful and precise answers.

Geographical. Is the geographic breakdown compatible across all data points? Between this data set and those from other agencies? For example, county-level data can be aligned with state-level data, but metropolitan statistical area -level data cannot (as MSAs may include parts of multiple states).

Industry/Sector. Is the industry/sector breakdown compatible across all data points? Between this data set and those from other agencies? For example, NAICS is compatible with SIC, but not with GICS. A free-text industry description would not be compatible with either (unless it corresponds directly to codes specified in a data dictionary).

Reporting Period. Is the reporting period breakdown compatible across all data points? Between this data set and those from other agencies? For example, monthly or quarterly data can be aligned with the Federal Government's fiscal year; annual data aligned to the calendar year (covering Jan-Dec) cannot.

Methodology. Is the methodology compatible across all data points? Between this data set and those from other agencies? Across time? For example, the Census has changed or added race/ethnicity choices on multiple occasions; this creates fissures across which data cannot be compared on a like-for-like basis.

Technical Implementation

Direct Access Links. Is it possible to link directly to a data item or table, or is source information buried in a page with numerous other tables and/or tests? What about a single number? Direct links will be necessary for a system to cite its sources. It also greatly helps the QA process.

Documentation. Specifically, technical documentation. Is it possible to use an API effectively without technical support from agency personnel, extensive guessing, or trial and error? BEA's API includes a detailed user guide explaining how to retrieve datasets, use endpoints, and handle parameters. For instance, the guide outlines the process for querying GDP data without requiring external support or trial and error.
Learn more: BEA API User Guide.

Rate Limiting. Are rate limits clearly described and high enough to allow full ingestion of the data set? The BEA API specifies rate limits, such as a maximum number of requests per time period, to ensure fair use while allowing adequate data ingestion for most users. For example, rate limits enable researchers to download comprehensive datasets without overloading the system.
Learn more: BEA API Rate Limiting Information.

Authentication. Are there access controls? If so, is the sign-up free and quick? Is there a waiting period? These all relate to the ease of data ingestion and processing, and the amount of engineering time and effort required.

AI / ML Specific Features

Schema.org Tagging. Schema.org is a widely adopted semantic vocabulary used to add structured metadata to web content. By embedding Schema.org tags into BEA's datasets and web pages, this metadata enhances search engines' understanding of the data's context, structure, and relevance. Is Schema.org metadata present? Is it complete? This is a standard set of metadata which various commercial web crawlers look for and use to understand documents. While it's not sufficient for our purposes, it will mean more relevant search results for people coming from the outside internet.
Learn more: Schema.org Docs

Croissant Tagging. Croissant is an advanced metadata framework tailored to AI and machine learning applications. Unlike Schema.org, which focuses on search engine optimization, Croissant tagging is designed to make datasets more discoverable and usable by AI systems. Is Croissant metadata present? Is it complete? This is a standard set of metadata which various commercial web crawlers look for and use to understand tabular data. While it's not sufficient for our purposes, it may improve the performance of commercial LLMs over the long term.
Learn more: MLCommons Croissant

Training Data Labels. These are question/answer pairs designed to aid in developing and validating lexical search results and natural-language interfaces, for example, "how is the restaurant industry doing in my area?" These question-and-answer pairs are crucial for translation of complex economic data into meaningful answers, either to retrieve the proper data set in a lexical search, or carry out a conversation in a chat-style interface.

Bias Documentation. Is an explanation of likely statistical bias provided? For example, a land-line phone survey in the early 2000s would have under-sampled young adults. This will be useful for predictive modeling or for more advanced use cases, such as those in which the system draws conclusions. More prosaically, it should be presented to the humans who will make those determinations, in order to raise awareness of potential blind spots.

Governance & Compliance

Version Control. Are updates and revisions to published data and their contents tagged and auditable? *Not typically relevant for data sets (e.g., economic indicators) in which the published date is a more fundamental property of the data.* This will be important for testing and ensuring the accuracy of our data retrieval, or simply keeping our system up-to-date. Records of when what changed, how, why, and by whom can also build public trust.

Usage Rights. Are there restrictions on use or publication of the data? *Not typically relevant for government data sets; this is more commonly applied to data compiled by for-profit companies.*

Privacy Annotations. Is private or personal data clearly marked? It's crucial to avoid publishing or using (even for training) private data inappropriately. More prosaically, data without open usage rights can't be published in an openly accessible tool.

Individual Linkages Possible. Is non-anonymized information present at the individual level, in such a way as to enable individual linkages across data sets? When possible, it opens entirely new categories of question for study or analysis which previously would have required dedicated one-off research efforts. How does an *individual's* response to a consumer sentiment survey relate to their spending patterns? Their educational outcomes? Their health? What effect does losing a driver's license have on an individual? Does this differ across geographies? Age groups?

PII Action Required. Must private or personal data be redacted before publication? May it be shared publicly or between agencies? May data be shared if it undergoes an anonymization process? An aggregation process? *Action will almost always be required if PII is present; exceptions would be made for data which is already public (for example, FEC donations).* This greatly increases the amount of effort needed for security, testing, and responsible publishing of data. Data with PII (which isn't otherwise public information) can't be fed into machine-learning systems at all, in many cases.

Analysis Details

Access to BEA Data

- The BEA API is available at BEA API Signup.
- Reports, unstructured data, and tables can be accessed at BEA Data.
- Interactive data tables are available at BEA Interactive Tables.

Scope

Breadth.

The breadth of BEA's offerings is substantial, with most datasets, such as GDP and Personal Income, accessible via multiple channels, including the API, interactive tools, and structured downloads. However, certain datasets, like those under "Special Topics," are not consistently available across all platforms. Despite this, the majority of data appears to be presented in both structured table formats and as reports or unstructured data. Notably, some content remains exclusively on the website, a typical arrangement for data-intensive systems but one that can pose challenges for commercial web scraping tools and AI/ML applications.

History.

BEA datasets generally include extensive historical data, particularly in structured tables, the API, and the interactive data application. For example, GDP figures are available back to 1929 through these platforms. However, original source reports are only available in the "Archive" section, dating back to 1996. Furthermore, preliminary releases (prior to revisions) exist as structured downloads from 2002 onwards and in the archive for the 1996-2002 period. For users seeking older data, the limitations are more pronounced. A consumer examining a GDP figure from 2001 in the API, interactive tables, or Excel downloads would only access the final, revised value. To retrieve the preliminary

version, they would need to consult the archive. For figures from 1987, neither the preliminary value nor the original report is accessible, reflecting a gap in historical reporting that could affect longitudinal analysis or back-testing. These insights highlight the strengths and limitations of BEA's data ecosystem. While the API and other platforms provide substantial coverage and accessibility for modern datasets, gaps in the availability of historical reports and preliminary data present challenges for certain analytical use cases. Addressing these gaps would enhance BEA's ability to support both traditional data consumers and AI-driven applications.

Data Structure & Format

Digitization.

All data available on BEA's website is fully digitized, making it accessible for modern applications. However, certain historical resources, such as reports prior to 1996, are not available on the BEA's website and are often only accessible as scanned paper documents from external sources. Some older methodology papers, like "Economic Theory and BEA's Alternative Quantity and Price Indexes," are also not fully digitized, creating gaps in the accessibility of foundational documentation.

Standardized Schema.

The BEA API operates through a single endpoint, GetData, with parameters and datasets managed via additional endpoints (GetDatasetList, GetParameterList, and GetParameterValues). This consistent schema supports seamless programmatic access. Downloadable tables, while generally consistent within datasets, show variations across datasets. For example, tables like "U.S. International Trade by Selected Countries and Areas" present data with years listed vertically and regions horizontally, differing from the more common layout of years horizontally and categories vertically. Minor inconsistencies, such as variations in headers and footnotes, are also present. On the website, the use of a content management system (CMS) ensures a largely uniform structure.

Standardized Layout.

Reports and unstructured data maintain a high degree of consistency over time, though occasional breaks occur. For instance, some reports transition from plain text files to PDFs with enhanced formatting, including images and logos. These shifts are managed effectively in interactive tables, which compensate for such inconsistencies. In the API, this concern is irrelevant, as all data is already structured rather than presented as documents or images. The website itself benefits from a CMS, maintaining a largely consistent layout across pages.

Tabular Layout.

Across all formats, including unstructured data, key information is consistently presented in a tabular format rather than buried within prose. Even the oldest plain-text documents display data with a clear tabular layout, making it easy to interpret. The API also adheres to a structured format, with no data presented as images or unstructured text. On the website, tabular data is typically embedded within downloads and tools rather than being present directly in the HTML.

Field/Column Labels.

Field and column labels are present across all functional areas. However, some labels in the API, such as "GeoFips" and "CL_UNIT," require external documentation to interpret correctly, posing a barrier to seamless usability. On the website, tabular data is generally confined to downloadable resources and interactive tools rather than being displayed directly.

Metadata & Context

The metadata and contextual information provided across BEA's functional areas play a critical role in making its datasets interpretable and actionable. While units and definitions are generally clear and accessible, challenges remain in standardizing methodologies and release date metadata.

Units Discernable.
Units are consistently marked across all BEA platforms, ensuring that users can easily interpret whether data is presented in thousands, millions, billions, dollars, or ratios. For users of the API, unit conversions can be requested directly, further enhancing accessibility and reducing manual efforts.

Data Dictionary.
In the API, a comprehensive data dictionary can be constructed using the endpoints GetDatasetList, GetParameterList, and GetParameterValues, which provide definitions for all data items. However, for other functional areas, such as downloadable tables and reports, the data dictionary and methodologies are combined into unstructured documents. This intertwining introduces challenges for users trying to parse specific definitions, especially in the absence of standardized formatting.

Methodology Described.
BEA provides detailed methodology descriptions at BEA Methodologies. These descriptions cover the processes behind data compilation and interpretation but are not directly integrated into the API or interactive tools. Moreover, the formats of these descriptions vary, lacking a standard structure that could simplify navigation and usability for analysts and developers.

Release Date/Time.
The handling of release date and time metadata across BEA's platforms is inconsistent. Timestamps on downloadable data tables can refer to the date of download, the date the file was generated, or the publication date, making it difficult to discern the true timeline of the data. While a schedule is available for cross-referencing, it requires users to step outside the API, tools, or downloads for confirmation. For example, the file "Table 1.2A. Real Gross Domestic Product, Chained (1937) Dollars" covering annual data from 1929 to 1947 is timestamped as created in 2002 and published in 2000, yet it represents restated figures originally compiled decades earlier.

News releases and reports available on the BEA website generally embed a clear release date in the text. However, this information is not consistently presented as metadata across all pages, limiting its utility for automated processing or AI applications. A schedule is provided here.

Interoperability & Granularity
Geographical.
Many BEA datasets are available at multiple geographic levels, including state, county, and metropolitan area. For example, datasets such as Personal Income frequently include all three levels, while others, like "Real Personal Income for States and Metropolitan Areas," are limited to state-level data. International datasets, while comprehensive, often aggregate smaller regions into broad categories like "all other countries," reducing their granularity for global analyses.

Industry/Sector.
SIC and NAICS codes are widely employed across BEA datasets, offering standardized industry classification. However, certain datasets lack detailed industry breakdowns. For instance, international goods trade data is categorized into broad groups, each encompassing multiple NAICS codes. This aggregation can obscure more granular trends within specific industries, potentially limiting the precision of sectoral analyses.

Reporting Period.
BEA datasets are reported at varying frequencies. While some datasets provide monthly breakdowns, many are only available on a quarterly or annual basis. This variability impacts their applicability for high-frequency analysis or time-sensitive decision-making, where more granular data might be required.

Methodology.
For the most central BEA datasets, such as GDP, methodological interoperability is not a concern, as BEA itself provides the authoritative definitions. However, discrepancies arise in datasets like trade data, where differences between BEA and USITC methodologies may result in irreconcilable variations. While international GDP comparisons or alternative U.S. GDP definitions fall outside the scope of this analysis, such differences are worth noting for broader interoperability considerations.

Technical Implementation
Direct Access Links.
BEA datasets are accessible through structured and unstructured downloads, each identifiable by a unique URL. Interactive data tables support hyperlinked navigation, with URLs dynamically updating to reflect applied filters or changes in data views. However, this functionality does not extend to the charts embedded in the same interface; while the link can direct users to the correct data table, it does not retain applied chart views or filters.

The API allows users to retrieve entire datasets through parameterized queries, with individual figures identifiable using static line numbers or series codes. While the CMS supports links to specific pages, it does not provide element names or IDs for more granular hash linking, limiting its usability for highly targeted queries.

Documentation.
Comprehensive API documentation is provided in the BEA API User Guide, which offers detailed guidance on endpoints, parameters, and retrieval methods. This ensures users can navigate the API effectively without extensive trial and error.

Rate Limiting.
The API has clear rate limits that may constrain the speed at which large-scale ingestion tasks can be completed, though these limits are well documented. For other functional areas, rate limits are undocumented, making it difficult to evaluate their impact until ingestion testing is performed. Testing limits through responsible interaction is critical, as aggressive testing methods, like simulating a DDoS attack, are neither feasible nor appropriate.

Authentication.
Authentication is required for the API but involves a straightforward process of email verification and CAPTCHA completion, ensuring accessibility without significant barriers. No authentication is required for other functional areas, enabling easier public access.

Schema.org Tagging.
Schema.org is a widely adopted semantic vocabulary used to add structured metadata to web content. By embedding Schema.org tags into BEA's datasets and web pages, this metadata enhances search engines' understanding of the data's context, structure, and relevance. Adding Schema.org metadata would not only improve the prominence and accuracy of Google search results referencing BEA data but also direct more traffic to the BEA website. This increased visibility could help members of the public access authoritative data, bolstering education and awareness of economic trends and insights.

BrightQuery is collaborating with the creator of Schema.org. This partnership allows BEA to benefit from cutting-edge expertise in metadata enrichment, enabling its datasets to achieve maximum visibility and utility across both search engines and AI applications.
Learn more: Schema.org.

Croissant Tagging.
Croissant is an advanced metadata framework tailored to AI and machine learning applications. Unlike Schema.org, which focuses on search engine optimization, Croissant tagging is designed to make datasets more discoverable and usable by AI systems. By adding Croissant metadata to downloadable tables, BEA can ensure seamless integration with tools and models developed by organizations such as ML Commons, Microsoft, Google, Meta, and OpenML. This tagging standard enhances AI systems' ability to accurately interpret, relate, and utilize the data, paving the way for more sophisticated applications and analyses in areas such as predictive modeling, data fusion, and semantic search.

BrightQuery's partnership with MLCommons, brings unparalleled expertise in implementing Croissant tagging to optimize BEA datasets for AI and machine learning applications. This collaboration also leverages MLCommons' network, including Microsoft, Google, Meta, and OpenML, to ensure the tagging framework is robust, interoperable, and future-proof, empowering BEA to deliver cutting-edge data solutions across a wide array of analytical domains.
Learn more: MLCommons Croissant

Training Data Labels.
Pre-built training data labels are not available for BEA datasets, which is consistent with the broader context of government data. BrightQuery is working with BEA personnel to develop these labels, which are essential for training AI systems effectively.

Bias Documentation.
While methodology papers and descriptions provide insights into potential biases, they are not explicitly documented. Making biases explicit would support fair and transparent use of BEA data, particularly in AI and machine learning applications.

Governance & Compliance
Version Control.
Access to previous data releases varies by platform. Earlier versions of data, prior to the final or latest release for a reporting period, are not accessible via the API or interactive data tools. However, structured downloads provide access to historical data releases from 2002 onward, while reports and unstructured data in the archive extend this access back to 1996. Although the BEA's website does not display an audit history, its Drupal content management system (CMS) likely maintains an internal audit log, even if this is not accessible to external users. For more details, refer to the discussion on release date and time inconsistencies in metadata.

Usage Rights.
BEA data is entirely unrestricted in terms of usage rights, ensuring open access for researchers, policymakers, businesses, and the public. This open-use policy supports a wide range of applications, from academic studies to AI-driven analyses, without the need for licensing or additional permissions.

Privacy Annotation.
Privacy concerns are minimal, as the BEA does not publish data at the individual level. The absence of personally identifiable information (PII) ensures that datasets can be used freely without the need for anonymization or additional safeguards.

Individual Linkages Possible.
Although survey respondents provide company names to the BEA, individual-level responses are not published and are protected by law. This safeguards confidentiality while allowing for potential internal linkages. For instance, BEA could integrate external datasets internally, perform linkages, and release aggregated results without ever sharing sensitive data. This approach mirrors models used by certain state agencies that are legally restricted from sharing raw data, even with other government entities.

PII Action Required.
No actions related to PII are necessary, as BEA does not publish or release individual-level data. This policy minimizes risks and eliminates the need for compliance measures specific to individual privacy.

AI Readiness Action Items

The **AI Readiness Action Items Table** outlines tailored solutions to address the gaps identified in the BQ AI Readiness Scorecard, focusing on enhancing BEA's data ecosystem to meet the demands of AI-driven insights. This table serves as a roadmap, detailing actionable steps designed to improve each functional area's performance across key dimensions: *Scope, Data Structure & Format, Metadata & Context, Interoperability & Granularity, Technical Implementation, AI/ML-Specific Features, and Governance & Compliance.* Our solutions prioritize BEA's high-value datasets, which are critical to national economic analysis and public policy.

Framework for the Solutions

The solutions in the **AI Readiness Action Items Table** are designed to:

- **Close Gaps**: Address specific shortcomings identified in the AI Readiness Headline Results.
- **Enhance Usability**: Improve accessibility, interoperability, and contextual clarity for end-users and AI systems.
- **Future-Proof Datasets**: Equip datasets with AI-specific features and compliance frameworks to meet evolving demands.
- **Maximize Impact**: Prioritize efforts on high-value datasets to demonstrate measurable improvements in BEA's AI readiness.

Each functional area within the **AI Readiness Action Items Table** is mapped to these datasets, ensuring the proposed solutions are both actionable and strategically aligned with BEA's goals.

Guiding Principles

The solutions proposed adhere to the following principles:

- **Scalability**: Solutions must scale across datasets and functional areas to maximize efficiency and reduce redundancy.
- **Interoperability**: Enhancements should align with federal and international data standards, such as schema.org and W3C.
- **AI Optimization**: Proposed actions emphasize features such as semantic tagging, metadata enrichment, and advanced querying capabilities.
- **Compliance and Governance**: Solutions address data security, bias mitigation, and adherence to privacy standards, ensuring ethical AI use.

Expected Outcomes

Implementing these action items will result in:

- Enhanced AI readiness for BEA's datasets, enabling more accurate and timely insights.
- Improved discoverability and usability for a broader audience, including policymakers, researchers, and AI systems.
- Demonstrated leadership in adopting AI-ready standards for federal economic datasets.

Table 3. BQ AI Readiness Action Items

Low Effort, High Reward		High Effort, High Reward	
Low Effort, Low Reward		High Effort, Low Reward	
Moderate Effort, Worthwhile Reward		NA	No Action Recommended

	Functional Area				
	APIs	**Reports & Unstructured Data**	**Tables**	**Interactive Charts**	**Website & Metadata**
Scope					
Breadth		NA	NA		/
History					
Data Structure & Format					
Digitization	NA	NA	NA	NA	
Standardized Schema	NA	NA		NA	NA
Standardized Layout	NA		NA	NA	NA
Tabular Layout	NA	NA	NA	NA	NA
Field/Column Labels		NA	NA	NA	NA
Metadata & Context					
Units Discernable	NA	NA	NA	NA	NA
Data Dictionary	NA				NA
Methodology Described		NA			NA
Release Date/Time		NA			
Interoperability & Granularity					
Geographical					NA
Industry/Sector					NA
Reporting Period					NA
Methodology	NA	NA	NA	NA	NA
Technical Implementation					
Direct Access Links	NA	NA	NA		
Documentation	NA	NA	NA	NA	NA
Rate Limiting	NA				
Authentication Methods	NA	NA	NA	NA	NA
AI / ML Specific Features					
Schema.org Tagging					
Croissant Tagging					
Training Data Labels					
Bias Documentation					
Governance & Compliance					
Version Control					
Usage Rights	NA	NA	NA	NA	NA
Privacy Annotations	NA	NA	NA	NA	NA
Individual Linkages Possible					
PII Action Required	NA	NA	NA	NA	NA

Action Items Details

Scope

Breadth.

Expanding the breadth of data availability is a long-term goal but should not come at the expense of historical data access. Prioritizing the availability of older datasets and reports will provide greater value for longitudinal studies and historical analyses. Efforts should focus on ensuring comprehensive access to past data before addressing the full breadth of current and future releases.

History.

To improve historical data accessibility, two key recommendations stand out:

- Archive Expansion: All reports should be included in the BEA archive, even if provided as scanned PDFs with minimal metadata, such as title, product, reporting period, advance/preliminary/final designation, and publication date.
- Structured Historical Data: At least one structured data platform (e.g., the API or downloadable tables) should support fetching advance, preliminary, and final figures for every reporting period, along with their respective publication dates.

Data Structure & Format

Digitization.

For older research documents, preprocessing with optical character recognition (OCR) can enhance accessibility if native digital versions are unavailable. However, this is considered a lower priority compared to other action items.

Standardized Schema.

Creating "pivoted" versions of downloadable tables, where years are listed horizontally and headers and footers are standardized across files, could improve usability. While helpful, this enhancement is also ranked as a low priority.

Standardized Layout.

Standardizing the layout of reports and unstructured data is not recommended, as retaining their original formatting preserves the integrity of the original publications. Any inconsistencies are already mitigated by other functional areas, such as interactive tools and APIs.

Tabular Layout.

No action is recommended for improving tabular layout, as all key data is already presented in a structured format across functional areas.

Field/Column Labels.

For the API, introducing user-friendly “display names” for fields that currently require external documentation (e.g., “GeoFips” and “CL_UNIT”) is highly recommended. This improvement would make the API more accessible for non-technical users. No further action is necessary for other areas.

Metadata & Context

Units Discernable.

No further action is recommended for units, as they are consistently clear and well-documented across all functional areas.

Data Dictionary.
Providing a consolidated, human-readable data dictionary is recommended as a minor but impactful enhancement. This could be implemented as a single document or web page that consolidates definitions for all data items, making them more accessible. The API already supports this functionality, so extending it to other areas would be a straightforward process.

Methodology Described.
To make detailed methodology documents easier to use, they should be organized for seamless cross-referencing, particularly from the data dictionary. The optimal implementation would involve hyperlinking the data dictionary and methodology documents, creating an interconnected system of definitions and detailed explanations.
Simpler implementations could include:

- Interactive Data Charts: Each line item could link directly to the data dictionary for a definition, with onward links to policy papers detailing the calculation methodologies.
- API Data Items: Each unique key or code could be associated with a unique URL pointing to the corresponding policy paper.

A more complex challenge arises when methodologies change over time. Links would need to be tied not only to the data item but also to the relevant year, ensuring historical accuracy and preventing ambiguity.

Release Date/Time.
To improve metadata on release dates, we recommend embedding structured HTML metadata, such as:
<meta property="article:published_time" content="YYYY-MM-DDTHH:MM:SSZ">

This would supplement the existing human-readable labels and enable automated tools to identify release dates more efficiently. This enhancement complements the recommendations under **Scope > History**, ensuring comprehensive tracking of data releases across platforms.

Interoperability & Granularity
Recommendations here are likely obvious but unrealistic. We can reasonably assume any lack of interoperability or granularity is the result of a considered decision which evaluated the tradeoffs.

Most of what is described in this section should be thought of as “things to be aware of when building any sort of model, training AI/ML, or running analysis” rather than “things to be addressed”. The BEA’s statisticians are, of course, already well-versed in these topics!

Geographical.
BEA datasets offer significant geographical detail, with many datasets already broken down to the county or metro area level. Expanding this granularity further—such as to individual countries in international data or below the county level for U.S. data—could enhance the precision of certain analyses but is likely constrained by data availability and collection costs.

Industry/Sector.
While BEA data uses standardized SIC and NAICS codes, the current granularity often stops short of the leaf level in the NAICS hierarchy. Providing data at this level could allow for more detailed sectoral analyses, though such granularity would require significant additional resources and may not be practical for all datasets.

Reporting Period.
Although some datasets are reported monthly, many are available only on a quarterly or annual basis. Monthly breakdowns across more datasets could support higher-frequency analysis but are limited by the nature of data collection processes and reporting schedules.

Methodology.
Achieving perfect methodological comparability across historical datasets is inherently unrealistic. For example, updating early 20th-century figures to conform to modern methodologies would require retroactive data collection and assumptions that are neither feasible nor historically accurate. Analysts and AI/ML systems must account for these unavoidable discrepancies when working with long-term datasets.

Technical Implementation

Direct Access Links.
The interactive data tool already updates its URL dynamically to reflect changes made in the table view. Extending this functionality to the graph view would enhance usability by preserving filters and configurations applied to visualized data. On the website, adding HTML names or IDs to headers and figures could enable hash linking, improving navigation and direct access to specific elements.

Documentation.
No further action is recommended for documentation, as the API user guide and other resources are sufficient for current use cases.

Rate Limiting.
While the API's rate limits are well-documented, there is uncertainty about rate limits for other functional areas. If such limits exist, a "Crawl-delay" directive should be added to the robots.txt file at https://apps.bea.gov/robots.txt to guide automated access appropriately and prevent server overload.

Authentication.
No changes are recommended for authentication, as the current system balances accessibility and security effectively.

Schema.org Tagging.
Currently, Schema.org tagging is not implemented within BEA's data infrastructure. The addition of this metadata framework would significantly enhance discoverability and semantic integration, particularly for search engines and AI applications.

Croissant Tagging.
Croissant tagging is also absent from BEA datasets. Implementing this advanced metadata framework would improve usability for machine learning systems by enabling richer, context-aware data interpretation.

Training Data Labels.
We're asking every agency we work with to provide a list of question/answer pairs for our natural language search tool and chatbot. This could also be made publicly available for other teams, including commercial AI chat tools.

Bias Documentation.
An explicit discussion of algorithmic bias could be incorporated into methodology papers in the future. This will inform AI researchers and developers working with those data sets. Even rudimentary discussions may help, here – while it may be second nature to a statistician to consider who is being left out of a phone or mail survey, computer scientists don't always have the training (and thus often neglect to build systems with that in mind).

Governance & Compliance

Version Control.
Recommendations regarding version control align with those detailed under **Scope > History and Metadata & Context > Release Date/Time**. Enhancing access to historical versions and embedding clear release dates in metadata would improve data traceability and reliability for analytical and AI-driven applications.

Usage Rights.
BEA data is unrestricted, with no limitations on its use. This open-access policy ensures maximum utility for researchers, policymakers, businesses, and the public, fostering innovation and evidence-based decision-making.

Privacy Annotation.
Privacy concerns are minimal, as BEA does not publish individual-level data. No further actions are required in this area.

Individual Linkages Possible.
As a long-term objective, BEA could explore opportunities for administrative or legislative action enabling the linkage of individual records using personally identifiable information (PII). Such an initiative, managed by BEA or a politically neutral agency, could facilitate the creation of aggregated statistics on topics that cannot be addressed with already aggregated datasets. This approach mirrors practices used by some state agencies, where aggregated insights are generated without exposing individual-level data.

PII Action Required.
No actions are necessary regarding PII, as BEA's current practices ensure that no individual-level data is published or shared externally.

Conclusion

This report serves as a step forward in the Bureau of Economic Analysis's (BEA) journey toward AI readiness, reflecting BrightQuery's commitment to enabling seamless integration of BEA's high-priority datasets into generative AI frameworks. By transforming the BEAs statistical products into machine-readable and machine-understandable formats, BEA can unlock new opportunities for enhancing data discovery, usability, and trustworthiness in AI-driven applications. This transformation will equip generative AI technologies to return accurate, authoritative insights, benefiting policymakers, researchers, and the public alike.

Key Findings

Our assessment identified critical gaps and opportunities across functional areas such as APIs, Reports & Unstructured Data, Tables, Interactive Charts, and Website & Metadata. These areas form the foundation of BEA's data ecosystem but require enhancements in metadata quality, interoperability, and technical implementation to fully support AI applications.

Using BrightQuery's AI Readiness Scorecard, we evaluated each functional area across several dimensions, assigning an overall readiness score of **67.3%**. The individual scores below provide a clearer picture of where improvements are most needed:

- **APIs**: **67%** readiness, showcasing strong functionality but requiring enhancements in metadata richness and advanced AI features.
- **Reports & Unstructured Data**: Recent data scored **77%**, while historical data scored **72%**, reflecting challenges in structure and accessibility.
- **Tables**: Scored **72%** for recent data and **67%** for historical data, underscoring the need for consistent metadata and interoperability improvements.
- **Interactive Charts**: **69%** readiness, highlighting the potential for improving AI compatibility and data granularity.
- **Website & Metadata**: Scored between **42%** and **49%**, emphasizing the critical need for schema.org tagging and metadata enrichment.

BrightQuery's focused analysis of BEA's high-value datasets—including Corporate Profits, GDP at various levels, Personal Income metrics, and U.S. International Trade—has highlighted their potential for transformation into AI-ready assets, enabling deeper insights and broader accessibility. This detailed scoring approach not only identifies areas for enhancement but also provides a roadmap for aligning BEA's data ecosystem with advanced AI and ML frameworks.

Recommendations and Action Items

Key recommendations focus on enriching metadata with **Schema.org** and **Croissant tagging** to improve **searchability** and **machine understanding**, while aligning datasets with industry standards for **interoperability** and **accessibility**. BrightQuery's **Government Data Agent (GDA) Evaluator** and **Transformer** will automate these enhancements, reducing manual efforts and enabling the BEA to efficiently meet the needs of generative AI systems. Providing a consolidated, human-readable **data dictionary** is also recommended as a minor but impactful enhancement. Priorities include expanding historical data availability by incorporating scanned archival reports with essential metadata and enabling structured access to preliminary, advance, and final data releases. Integrating **user-friendly display names** for API fields **and linking data items** to detailed methodology documents will **streamline usability**, while **embedding standardized metadata**, such as release dates, and **enabling dynamic URLs** for interactive charts will **improve automation and navigability**. These targeted actions will position BEA's data as a benchmark for AI-ready, interoperable government datasets.

Broader Impacts

The BEA's commitment to AI readiness can serve as a model for other federal agencies, demonstrating the feasibility and value of transitioning datasets into machine-understandable formats. This work aligns with the OPEN Government Data Act and the Department of Commerce's AI and Open Government Data Assets Working Group, showcasing BEA as a leader in leveraging federal data to drive innovation and evidence-based decision-making.

The Federal statistical system comprises over 125 agencies or units that engage in statistical activities. A substantial portion of official statistics are produced by the 13 agencies that have statistical work as their principal mission. BrightQuery is building a comprehensive framework to support these agencies in transitioning their datasets to AI-ready formats, ensuring that statistical products are enriched with metadata, interoperable, and optimized for machine learning applications. This initiative will enhance the accessibility, accuracy, and utility of federal statistics for policymakers, researchers, and the public.

—Bureau of Economic Analysis (Department of Commerce);
—Bureau of Justice Statistics (Department of Justice);
—Bureau of Labor Statistics (Department of Labor);
—Bureau of Transportation Statistics (Department of Transportation);
—Census Bureau (Department of Commerce);
—Economic Research Service (Department of Agriculture);
—Energy Information Administration (Department of Energy);
—National Agricultural Statistics Service (Department of Agriculture);
—National Center for Education Statistics (Department of Education);
—National Center for Health Statistics (Department of Health and Human Services);
—National Center for Science and Engineering Statistics (National Science Foundation);
—Office of Research, Evaluation, and Statistics (Social Security Administration);
—Statistics of Income Division (Department of the Treasury);
—Microeconomic Surveys Unit, (Board of Directors of the Federal Reserve System);
—Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration (Department of Health and Human Services); and
—National Animal Health Monitoring System, Animal and Plant Health Inspection Service (Department of Agriculture).

Appeal

To realize this vision, BEA should implement the outlined recommendations, pilot the GDA tools, and engage with BrightQuery in continued collaboration. Expanding the scope of this transformation to include replicability testing with other federal agencies will further validate and refine the methodologies and tools developed in this project. Establishing a roadmap for long-term AI readiness will ensure sustained progress and position BEA at the forefront of AI-enhanced data dissemination..

Looking Ahead

BrightQuery is ready to support the BEA to enhance high-priority datasets and build scalable solutions to ensure data integrity, accessibility, and utility. Together, we can transform BEA's data ecosystem into a beacon of innovation and accessibility, setting a new standard for public data in the age of artificial intelligence.

Appendix A: Triplification

Triplification refers to the process of converting structured or semi-structured data into a format suitable for the **Semantic Web** or other AI-driven systems work with **knowledge graphs**. This typically involves organizing data into **triples**, which are the foundational units of a **Resource Description Framework (RDF)**. A triple consists of three components and are structured as **"subject-predicate-object"** statements:

1. **Subject**: The entity being described (e.g., ""John hasAge 30"").
2. **Predicate**: The relationship or property of the subject (e.g., "hasAge").
3. **Object**: The value or another entity related to the subject (e.g., "30").

Learn More: EAV (entity-attribute-value) data models

Why Triplification for AI?

AI systems, especially those leveraging **natural language understanding (NLU)** or **knowledge graphs**, benefit from structured, interconnected data because it enables:

- **Semantic understanding**: Triplification enriches data with explicit meaning by defining entities and their relationships.
- **Interoperability**: Standardized triples can integrate seamlessly with other datasets and systems.
- **Machine reasoning**: AI can infer new facts or relationships using rules and ontologies applied to triples.

Steps in Triplification

1. **Data Extraction**: Identify structured, semi-structured, or unstructured data sources (e.g., relational databases, spreadsheets, XML, JSON, etc.).
2. **Entity Mapping**: Map data fields to entities, predicates, and values based on an ontology or schema.
3. **Triple Creation**: Transform data into the "subject-predicate-object" format.
4. **Serialization**:Store triples in RDF or compatible formats (e.g., Turtle, JSON-LD, N-Triples).
5. **Integration**: Merge triples into a knowledge graph or database for AI consumption.

Example 1: Triplifying a Spreadsheet

Name	Age	City
John	30	New York
Alice	25	London

Triplification Output:

- **John hasAge 30**
- **John livesIn NewYork**
- **NewYork isPartOf USA**
- **Alice hasAge 25**
- **Alice livesIn London**
- **London isPartOf UK**

This representation enables AI to understand relationships and infer new knowledge, such as "John and Alice are both humans" or "John's city is in the USA."

Benefits of Triplification for AI

- Facilitates **context-aware AI systems**.
- Supports advanced **querying and reasoning**.
- Makes data reusable across domains and applications.
- Enhances AI's ability to generalize knowledge and make decisions.

By adopting triplification, you prepare data in a way that aligns with how AI systems understand and process information, creating a foundation for smarter, more connected applications.

Example 2: Triplification for a classic person-company-employment data model.

1. Update the Ontology

Introduce an **Employment** entity as a first-class object:

- **Person**: Represents the individual.
- **Company**: Represents the employer.
- **Employment**: Represents the relationship between a person and a company, including the role, start date, end date, and any other attributes specific to the job.

Example Ontology Properties:

- **Person**:
  - :hasName, :hasAge
- **Company**:
  - :hasName, :locatedIn
- **Employment**:
  - :involvesPerson (connects to the person)
  - :involvesCompany (connects to the company)
  - :hasRole, :startDate, :endDate

2. Triplify the Data

Example Relational Data:

Person			Company		
ID	**Name**	**Age**	**ID**	**Name**	**Location**
P001	John	30	C001	Acme Corp	New York
P002	Alice	25	C002	Beta Inc	London

Employment					
EmploymentID	**PersonID**	**CompanyID**	**Role**	**StartDate**	**EndDate**

E001	P001	C001	Engineer	2020-01-01	2023-01-01
E002	P001	C002	Manager	2023-03-01	NULL
E003	P002	C002	Designer	2021-06-01	NULL

Resulting RDF Triples:

People	Companies	Employments
:P001 :hasName "John" ; :hasAge "30"^^xsd:integer	:C001 :hasName "Acme Corp" ; :locatedIn "New York" .	:E001 :involvesPerson :P001 ; :involvesCompany :C001 ; :hasRole "Engineer" ; :startDate "2020-01-01"^^xsd:date ; :endDate "2023-01-01"^^xsd:date .
:P002 :hasName "Alice" ; :hasAge "25"^^xsd:integer	:C002 :hasName "Beta Inc" ; :locatedIn "London" .	:E002 :involvesPerson :P001 ; :involvesCompany :C002 ; :hasRole "Manager" ; :startDate "2023-03-01"^^xsd:date .
		:E003 :involvesPerson :P002 ; :involvesCompany :C002 ; :hasRole "Designer" ; :startDate "2021-06-01"^^xsd:date .

3. Explanation of Structure

- Each **Employment** instance (e.g., :E001) represents a unique job and acts as a bridge between the **Person** and the **Company**.
- Properties like :hasRole, :startDate, and :endDate are attached to the **Employment** entity, not the **Person** or **Company** directly.
- This allows a single person (e.g., :P001) to have multiple employment relationships, each with distinct attributes.

4. Benefits

- **Scalability**: The model can accommodate multiple roles for a person at different companies.
- **Flexibility**: Additional attributes like salary, job description, or performance ratings can be added to the **Employment** entity without affecting the **Person** or **Company** structures.
- **Queryability**: AI or graph queries can focus on:
  - All roles held by a person.
  - Employees of a company during a specific time range.
  - Overlapping employment or career progression of individuals.

Some Tools for Triplification

Triplifying unstructured text data involves extracting meaningful entities, relationships, and facts from the text and transforming them into subject-predicate-object triples. This process requires tools and techniques from natural language processing (NLP), knowledge graph construction, and semantic web technologies. Here are some commonly used tools and approaches for this purpose:

1. Pre-trained NLP Frameworks

These frameworks extract entities and relationships from unstructured text and can be configured to output triples.

a. spaCy with extensions

- Description: A powerful Python NLP library for tokenization, named entity recognition (NER), part-of-speech tagging, and dependency parsing.
- How to use: Combine spaCy's NER and dependency parsing with custom rules or extensions (like spaCy-Transformers) to identify subject-predicate-object triples.
- Output: Use the parsed relationships to generate RDF triples.

b. Stanford CoreNLP

- Description: Java-based NLP toolkit offering NER, sentiment analysis, and relation extraction.
- How to use: Apply its OpenIE module to extract facts and relations, which can be directly converted to triples.
- Output Example: From "John works at Google," it might extract (John, worksAt, Google).

2. Open Information Extraction (OpenIE) Tools

These tools are specifically designed to extract structured relationships (facts) from unstructured text.

a. OpenIE

- Description: A Stanford NLP tool for extracting subject-predicate-object triples directly from sentences.
- Features:
  - Handles complex sentences.
  - Outputs triples in text or RDF format.
- Example: From "Microsoft was founded by Bill Gates in 1975," OpenIE extracts:
  (Microsoft, was founded by, Bill Gates)
  (Microsoft, was founded in, 1975)

b. OLLIE (Open Language Learning for Information Extraction)

- Description: Focuses on extracting triples from sentences, including ones with implicit relationships.

3. Semantic Annotation and RDF Conversion Tools

These tools annotate text with semantic tags and convert them into RDF triples.

a. Apache Jena

- Description: A framework for building and querying RDF data.
- How to use: Pair it with NLP tools to store extracted triples in a semantic graph.

b. DBpedia Spotlight

- Description: An open-source tool for annotating and linking entities in text to DBpedia.
- How to use:
  - Input unstructured text.
  - Spotlight links entities in the text to concepts in DBpedia.
  - Output RDF triples that describe the linked entities and their relationships.

c. FRED (Semantic Graph Extraction)

- Description: A tool that converts text into RDF/OWL graphs.

- Features:
  - Combines semantic role labeling, NER, and ontology linking.
  - Outputs triples in multiple RDF formats.
- Example: From "Barack Obama was born in Hawaii," FRED generates:(Barack_Obama, bornIn, Hawaii)

4. Machine Learning-based Relation Extraction Tools

These use pre-trained or custom models to extract relationships for triplification.

a. Hugging Face Transformers

- Description: Provides pre-trained language models like BERT, RoBERTa, and T5 for relation extraction and entity recognition.
- How to use:
  - Fine-tune a transformer for relation extraction tasks.
  - Pair it with a triple generator script to convert detected relationships into triples.

b. DeepDive

- Description: A framework for statistical relational learning which extracts structured information from text.
- Use Case: Ideal for large-scale information extraction pipelines.

5. Knowledge Graph Construction Platforms

These platforms integrate NLP pipelines and RDF triple generation.

a. Neo4j with NLP Plugins

- Description: A graph database with NLP integration for knowledge graph creation.
- How to use:
  - Extract entities and relationships with NLP libraries.
  - Store the results as triples in Neo4j's graph format.

b. Amazon Comprehend

- Description: AWS NLP service for entity recognition and relationship extraction.
- Output: Results can be transformed into triples for storage in RDF-compatible formats.

c. IBM Watson Knowledge Studio

- Description: A cloud-based tool for training custom models to extract entities and relationships.
- How to use: Train the system to detect domain-specific entities and relationships, then export as triples.

6. Ontology-driven Tools

a. Protégé with Plugins

- Description: Ontology editing and management tool which can annotate and structure text-based data.
- How to use:
  - Define an ontology.
  - Annotate text with the ontology terms to extract triples.

b. AlchemyAPI (discontinued but alternatives exist)

- Description: Previously provided entity and relation extraction services. Alternatives like Google Cloud NLP now offer similar capabilities.

7. Custom Pipelines

For domain-specific triplification, you might need to build a custom pipeline combining:

- Pre-processing: Tokenization, entity recognition (e.g., spaCy, NLTK).
- Relation extraction: Using tools like Stanford OpenIE or custom models.
- Triple generation: Scripts to transform extracted relationships into RDF format.

Example Pipeline Using OpenIE and Jena

Input Text:	"Alice joined Acme Corp in 2021 as a designer."
OpenIE Output:	(Alice, joined, Acme Corp) (Alice, in, 2021) (Alice, as, designer)
RDF Triple Generation:	:Alice :joined :AcmeCorp . :Alice :joinedIn "2021"^^xsd:date . :Alice :hasRole "designer" .

Appendix B: Modeling Time Series data with Revisions

Appendix B explores the challenges and opportunities in triplifying time-series data, focusing on the unique complexities of incorporating temporal dimensions into AI-ready datasets. While there are numerous tools available for approximating triplification from unstructured text, integrating the time component presents distinct difficulties that require innovative approaches.

One promising strategy involves treating the **reporting time period** as a first-order entity in the data model. Time in this context is inherently two-dimensional: it encompasses the time at which the data was produced and the time to which the data refers. Furthermore, time-series datasets often include multiple data points—such as estimates, revisions, and finalized values—that pertain to the same reference period. These intricacies make reporting periods a unique challenge, often more complex than a simple date stamp.

In the financial data domain, time-series modeling incorporates these nuances to account for revisions and overlapping periods effectively. Given the limitations of current language models (LLMs) in handling time-series data, testing this approach within an AI framework may yield insights and improvements, especially as LLMs currently struggle to manage these complexities. This methodology aligns with strategies used in financial analysis and could serve as a benchmark for improving AI systems' interaction with temporal data.

What do we aim to solve?

A relational database offers robust capabilities for managing and querying time-series data, enabling precise insights and historical context. Below are some of the key use cases that such a database could support:

1. **Identify Data Within a Specific Time Range**
   Easily retrieve data points that fall within a defined time range, facilitating focused analyses such as quarterly or yearly trends.
2. **Generate Line Graphs Across Time**
   Use structured time-series data to create line graphs that illustrate changes over time, providing clear visualizations of trends and patterns.
3. **Identify the Latest Revision for Each Time Period**
   Quickly pinpoint the most up-to-date data for any given time period (e.g., determining the final GDP figure for Q2 2020 after all revisions).
4. **Roll Back Charts to a Historical Data View**
   Recreate what the dataset looked like at a specific historical point in time (e.g., showing the GDP figure reported in July 2020 for Q2 2020 before subsequent revisions).
5. **Align Data Across Multiple Time Series**
   Compare and analyze related data points from different time series that overlap in time (e.g., correlating an unemployment rate of 4.2% with a GDP value of $29 trillion for the same period).
6. **Approximate Like-for-Like Comparisons Across Asynchronous Data**
   Adjust for misaligned reporting schedules to generate accurate comparisons (e.g., aligning federal government fiscal year-end debt with calendar year-end GDP, accounting for the three-month offset).

By structuring time-series data in a relational database, these use cases can be addressed with precision, providing analysts and decision-makers with accurate and actionable insights. This approach also facilitates consistent handling of revisions and temporal relationships, ensuring a comprehensive understanding of historical and current trends.

Example 3: Time Series Modeling

In a relational database, we'd make each of these its own table, of course, but it's a similar concept.

STAT1: hasName: "Unemployment Rate"
STAT2: hasName: "GDP"
STAT3: hasName: "Unemployment Rate, Men"
STAT3: hasParent: STAT1
STAT4: hasName: "Unemployment Rate, Women"
STAT4: hasParent: STAT1

LOC1: hasName: United States
LOC2: hasName: Florida
LOC2: isIn: LOC1

PERIOD1: hasYear: 2024
PERIOD1: hasMonth: November
PERIOD1: hasType: Calendar Month
PERIOD1: precedes: PERIOD2
PERIOD2: hasYear: 2024
PERIOD2: hasMonth: December
PERIOD2: hasType: Calendar Month
PERIOD2: follows: PERIOD1

STATINSTANCE1: isForPeriod: PERIOD1
STATINSTANCE1: isStat: STAT1
STATINSTANCE1: isForLocation: LOC1
STATINSTANCE1: hasValue: 4.1%

STATINSTANCE2: isForPeriod: PERIOD2
STATINSTANCE2: isStat: STAT1
STATINSTANCE2: isForLocation: LOC1
STATINSTANCE2: hasValue: 4.2%
(each period + stat + location has a value)

We could capture estimates/actuals/revisions with:

STATINSTANCE1: isForPeriod: PERIOD1
STATINSTANCE1: isStat: STAT1
STATINSTANCE1: isForLocation: LOC1
STATINSTANCE1: hasValue: 4.1%
STATINSTANCE1: releaseDate: 2024-Dec-06
STATINSTANCE1: revises: None
STATINSTANCE1: revisedBy: STATINSTANCE1a

STATINSTANCE1a: isForPeriod: PERIOD1
STATINSTANCE1a: isStat: STAT1
STATINSTANCE1a: isForLocation: LOC1
STATINSTANCE1a: hasValue: 4.0%
STATINSTANCE1a: releaseDate: 2024-Jan-06
STATINSTANCE1a: revises: STATINSTANCE1
STATINSTANCE1: revisedBy: None

More on Period Types

Here is a (mostly complete) list of period types:

- Calendar Year [periodic]
- Calendar Quarter [periodic]
- Calendar Month [periodic]
- Fiscal Year [periodic]
- Fiscal Quarter [periodic]
- LTM (last twelve months), also referred to as TTM (trailing twelve months) [periodic]
- NTM (next twelve months) [periodic]
- Point in time [*not* periodic]

Why?

Understanding and accurately modeling time concepts is essential for interpreting data effectively. Various temporal dimensions, such as calendar periods, fiscal periods, and point-in-time versus periodic metrics, influence how data is structured, reported, and analyzed. Below are key distinctions and their implications:

1. Calendar Periods

These are the most intuitive timeframes, corresponding to standard calendar dates like "December 2024" or "the year 2000." They are what most people think of when referencing time and are universally understood. Calendar periods serve as the baseline for many datasets, providing a common framework for comparing data.

2. Fiscal Periods

Fiscal periods are tailored to organizational needs, diverging from calendar years to suit financial reporting purposes.

- **Corporate Examples**: Retailers often end their fiscal year on January 31 of the following year to account for holiday sales and returns. For instance, a retailer avoiding a December 31 year-end mitigates arbitrary financial fluctuations caused by factors like snowstorms affecting returns in the Northeast.
- **Government Examples**: The U.S. Federal Government begins its fiscal year in October, allowing time for Congress and the President to finalize the budget. This timing ensures that reporting aligns with the fiscal decision-making process rather than the calendar year.

3. LTM (Last Twelve Months) and NTM (Next Twelve Months)

LTM and NTM metrics offer a dynamic approach to annualized reporting, smoothing seasonal variations by providing rolling or forward-looking annual estimates.

- **Example**: If you want to understand “how many jobs the U.S. is creating annually” in July, relying solely on data from January-December of the previous year could misrepresent trends. LTM captures July-to-June data for a more timely and representative analysis.

4. Point-in-Time Metrics

These metrics capture discrete values at specific moments, akin to balance sheet data in accounting.

- **Examples**:
  - "What is the total household debt in 2024?" depends on whether you’re asking about January 1 or December 31, as the answers will differ.

- Stock prices, which fluctuate intra-second, exemplify the granularity possible with point-in-time data, though most government datasets are unlikely to report with such frequency.

5. Periodic Metrics

Periodic data aggregates values over a span of time, such as annual GDP, which represents the sum of goods and services produced during a year.

- **Examples**:
  - "What is U.S. GDP in 2024?" includes every day of the year, from January 1 to December 31, summed up.
  - Some questions don't semantically align with periodic data, such as "How many new jobs were created at 2pm on Friday?" These metrics inherently represent a broader timeframe, like monthly or quarterly data, even if interpreted as momentary.

Why These Distinctions Matter

Accurately distinguishing between these temporal dimensions helps avoid misinterpretations and ensures consistency in analysis. For example:

- A point-in-time metric like "temperature at 2pm" may refer to the exact value at 2:00, while a periodic metric like "rainfall at 2pm" implicitly spans a range (e.g., 1:45 to 2:00).
- Reporting GDP as an annual figure makes semantic sense, but attributing it to a single moment within the year, such as 2pm on December 31, does not.

By modeling time appropriately, analysts can ensure that the data reflects its true context and meaning, enabling more accurate and insightful interpretations across various use cases.

Estimates and Revisions

Estimates and revisions share a common framework: both are modeled with a **reference period** (the time the data pertains to) and a **reporting timestamp** (when the estimate or revision is published). This dual-timestamp system captures the temporal relationship between when the data is analyzed or revised and the period it describes.

Example Timeline: US GDP for 2024

1. **Initial Estimates**:
   - *January 7, 2019*: An economist predicts, "US GDP will be $a by 2024."
   - *July 14, 2020*: A revised estimate states, "US GDP will be $b by 2024."
   - *December 20, 2023*: Another revision predicts, "US GDP will be $c by 2024."
2. **Near-Term Estimates**:
   - *January 7, 2024*: An economist estimates, "US GDP was $d in 2024," anticipating BEA's official compilation.
3. **Initial BEA Report**:
   - *January 25, 2025*: BEA releases its first official figure: "US GDP was $e in 2024."
4. **Subsequent Revisions**:
   - *February 28, 2025*: BEA revises its figure to: "US GDP was $f in 2024."
   - *March 28, 2025*: BEA further refines its estimate: "US GDP was $g in 2024."

The Difference Between Estimates and Revisions

- **Estimates**: Typically published **before** the publisher releases an official value. These are often created by third parties, such as economists or financial analysts, and may use incomplete or alternative data sources to project outcomes.

- **Revisions**: Published **after** an initial official value is released, often by the publisher itself (e.g., BEA). Revisions refine the original data based on new information or methodologies.

The Complexity of Definitions

Even this distinction can blur. For example, the BEA refers to its March 2025 revision of 2024 GDP as an "estimate" because, in their words, "the real value of GDP is unknowable for certain." This highlights the inherent uncertainty in economic data, where even finalized values are approximations informed by the best available evidence and assumptions.

By understanding this nuanced relationship between estimates and revisions, analysts can better interpret how data evolves over time and account for its inherent uncertainties in decision-making.

Why do we even care about that second time component?

For most natural language queries, the second time component—the reporting timestamp—doesn't matter. If someone asks, "What was the US GDP in 1980?" they generally expect the latest revision unless they specifically request the value as it was reported at a particular moment in the past.

However, this time component becomes critically important in scenarios like **back-testing strategies**, where the accuracy of historical data as it was originally reported is essential.

Example: Back-Testing a Gold Investment Strategy

Imagine a strategy that dictates, "Buy gold if inflation exceeds **2.5%.**"

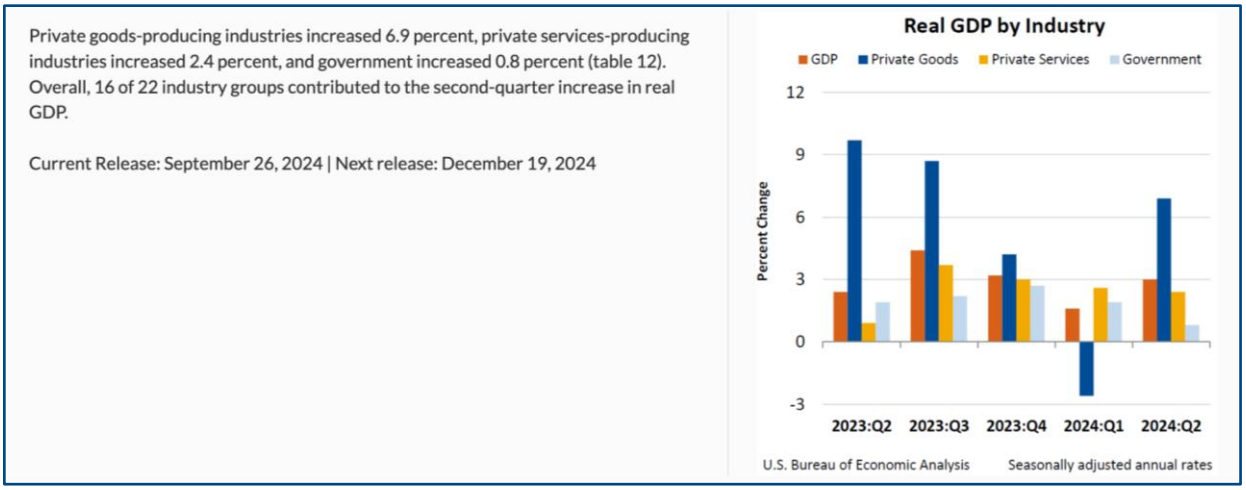
1. **Hypothetical Scenario**:
    - March: Inflation is reported as **2.6%** in the revised data. Following the strategy, you would buy gold in April.
    - April: Gold prices rise in May. The strategy appears successful.
2. **What Actually Happened**:
    - In April, the **original report** stated inflation was only **2.4%,** not **2.6%**. Based on this initial report, you wouldn't have bought gold in April.
    - The inflation figure wasn't revised to **2.6%** until September, months after you would have acted on the original report.
3. **Outcome**:
    - The back-test falsely assumes you would have acted on the revised inflation data **(2.6%)** rather than the initially reported figure **(2.4%)**.
    - As a result, you might conclude that your strategy was successful when, in reality, you would have made the wrong decision based on the information available at the time.

The Importance of Historical Context

This example illustrates why it's crucial to differentiate between **original reports** and **subsequent revisions**. Using only the latest data for back-testing can create a false sense of confidence in a strategy that wouldn't have worked in real-time. For scenarios requiring historical accuracy, such as financial modeling or policy evaluation, the second time component ensures decisions are evaluated against the data as it was known at the time, preserving fidelity to real-world conditions.

How?

The good news is that, at least for a human, making this distinction is pretty easy because it's one that the statisticians all tend to care to make. https://www.bea.gov/data/gdp/gdp-industry



It's very clear

- **releaseDate:** Sept 26, 2024. You can even get a timestamp if you click into the documents. 8:30am EDT
- Data in the chart
  - **isForPeriod:** 2023 Q2
  - **isForPeriod:** 2023 Q3
  - **isForPeriod:** 2023 Q4
  - **isForPeriod:** 2024 Q1
  - **isForPeriod:** 2024 Q2

END OF REPORT

Learn more

BrightQuery.com

BrightQuery.ai