

Retail and Wholesale Trade: Practical Experience in the Use of Corporate Data

Michelle Chapman and Katie Frawley

U.S. Bureau of Labor Statistics

40th Meeting of the Voorburg Group on Services
Statistics

The Hague, Netherlands

September 2026

Introduction: Why Corporate Data Matters

The Bureau of Labor Statistics (BLS) depends on survey data collection to publish official economic measures of the U.S. economy. While traditional collection methods continue to provide high-quality statistical outputs, there are limitations. Efforts to mitigate respondent burden while maximizing data collection are constantly evaluated. The successful collection of responses to voluntary surveys faces many challenges. The BLS has actively worked to expand alternative methods for collecting data in the face of these challenges.

For establishment surveys that provide prices for calculating the Producer Price Index (PPI), manual reporting activities can impose a substantial burden on participating firms. Additionally, periodic instead of continuous sampling limits representativeness of product changes over time. Maintaining representative samples within dynamic industries requires significant resources and effort. Industries characterized by large transaction volumes, broad product inventories, and rapidly changing inventories are especially prone to these limitations.

At the same time, advances in cloud-based enterprise resource planning systems, real-time inventory management platforms, and transaction processing programs have resulted in many firms maintaining increasingly detailed operational and transactional records. The growing availability of these platforms has created new opportunities for statistical agencies to evaluate alternative collection approaches to support official statistics.

For the BLS, corporate data refers to a large set of transaction-level data provided by a sampled firm. The purpose of collecting corporate data for the Producer Price Index (PPI) is to reduce respondent burden by eliminating manual reporting while increasing the amount of data available for analysis and price index calculation. Typically, corporate data is requested in the form of a report that the firm already maintains within its existing operational system. This data should contain information related to sales prices, costs, quantities, product identifiers, and other transaction-level details. Access to large transactional datasets allows economists to develop more representative transaction groupings, improve insight into price movements, and better evaluate the drivers of index changes.

The successful use of corporate data in the official statistics depends on more than just access. Corporate data is often developed for operational and accounting purposes rather than statistical production, differing across firms in structure, granularity, and reporting practices. Statistical agencies need clearly defined operational procedures before including corporate data into the calculation of official statistics.

This paper discusses the collection and evaluation of corporate data for use in calculating PPI measures. It examines integration of large transactional datasets into production-oriented workflows that meet established statistical quality standards. In addition, the paper explores the operational procedures, respondent engagement strategies, and ongoing maintenance necessary to transform transactional records provided by firms into statistical inputs capable of supporting PPI production.

Corporate Data within the PPI Retail and Wholesale Trade Sector

Producer Price Indexes (PPIs) measure changes in the selling prices received by domestic producers for their output. Calculation of PPIs requires data to be ingested in complex customized software systems to support monthly price index calculation and analysis.

PPI data collection requirements differ across sectors depending on the economic activity being measured. Manufacturing industries primarily rely on transaction prices associated with the sale of goods produced by the firm. In contrast, retail and wholesale trade industries measure gross margins, representing the difference between the selling price received by the firm and the cost associated with acquiring the product for resale.

Because PPIs require continuity across reporting periods, data must have stable product definitions, recurring observations, and sufficient detail to evaluate price or margin movement over time. These requirements are particularly important in the dynamic retail and wholesale trade industries which are driven by large transaction volumes as well as broad and constantly changing product inventories.

Retail and wholesale trade industries represent strong candidates for corporate data collection. Most firms today maintain highly granular records of individual transactions that include dates, quantities, prices, costs, discounts, promotions, customer segments, and product descriptions. These data elements are often available in near real-time because of automated cloud-based systems.

Within retail and wholesale trade industries, the BLS generally prefers the use of comparable product lines rather than individual products when measuring margins for PPIs. Comparable product lines are groups of similar products sold under similar market conditions and provide a representative measure of the trade service performed by the firm.

For example, a grocery wholesaler may sell hundreds of brands and package sizes of bottled water. Rather than measuring margins for a single product, BLS concepts for

calculating PPIs allow for grouping similar bottled water products into a comparable product line. Corporate datasets can provide access to transaction-level pricing information across a substantially larger share of the firm's activity than is typically available through traditional collection. Access to these records allows economists to evaluate product groupings, develop these comparable product lines, and calculate average margins using a broader set of observations. The result is a more representative average margin measurement that provides greater insight into margin movements for the entirety of the firm.

By requesting corporate data, BLS can also reduce respondent burden. Firms that provide corporate data leverage information already available within wholesalers' and retailers' operational systems, rather than working on customized BLS requests that require firms to identify and calculate representative margins for comparable product lines manually.

These factors made retail and wholesale trade industries strong candidates for evaluating the use of corporate data to provide robust prices for PPI calculation.

Pilot Implementation: Grocery and Related Products Merchant Wholesalers

Selection of the Pilot Industry

The Grocery and Related Products Merchant Wholesalers industry (NAICS 424400) was selected for the corporate data pilot because it represented a significant component of wholesale trade activity and was expected to generate transactional datasets suitable for evaluating corporate data collection. Firms within this industry typically maintain extensive inventories that are managed and monitored by systems capable of producing detailed electronic records.

Firms in this industry often manage large inventories spanning thousands of products and deal with frequent product turnover and substitution. These characteristics created an opportunity to evaluate whether corporate data could be used in the PPI production process and to assess whether large transactional datasets could be transformed consistently into inputs for official price index production. These challenges were likely to be encountered in future corporate data collection efforts in all other retail and wholesale trade industries.

The pilot's objective was to assess both the feasibility of acquiring large transaction-level datasets from respondents and the ability of economists to process, evaluate, and use that data.

Respondent Engagement and Data Acquisition

Firms targeted by the pilot were offered flexibility in how they reported price information. They could continue providing average margins on a set number of sampled transactions or choose to submit transaction-level corporate datasets derived from existing business systems. Offering multiple reporting options allowed the pilot to evaluate both respondent willingness to provide corporate data and the practical challenges associated with acquiring large operational datasets.

Successful implementation of the pilot required engagement with personnel with direct knowledge of firm pricing, inventory, accounting, or information systems. In many cases the most effective contacts were individuals responsible for managing sales records, financial systems, or enterprise data platforms. Early discussions between economists and respondents focused on understanding how data was maintained within the firm, identifying available information, and determining whether existing records could support statistical requirements. Other topics included understanding how products were organized by the respondent and the level of granularity.

For example, economists may have asked whether sales and cost information are stored within the same system, how products are organized within inventory categories, whether historical transactional records can be extracted, and which personnel are responsible for generating reports.

Respondents who agreed to provide corporate datasets were asked to provide records covering the major product categories associated with their wholesale operations. The economists provided a list of major product categories and the respondents self-assigned products to this list. The categories reflected economic activity within the industry and served as the foundation for grouping and margin analysis.

These interactions demonstrated that data availability often varied across firms operating within the same industry. Differences in inventory systems, accounting, product categories, and reporting structures influenced both the content and the format of data. As a result, respondent engagement became a critical component of evaluating corporate data feasibility.

Data Structure and Processing

Corporate datasets submitted during the pilot generally contained product identifiers, descriptions, quantities sold, sales prices, and acquisition costs information. These elements provided the necessary data to calculate gross margins and evaluate product - level activity across large portions of a firm's operations.

Example:

Product	Cost	Sale Price	Margin
Product A	\$60	\$100	\$40
Product B	\$30	\$80	\$50

While traditional sampling and data collection might obtain only the single average margin for a product category, corporate data sets provide access to the underlying transactions used to calculate those margins.

Whenever possible, BLS economists provided mockups to respondents of acceptable formats for corporate data file submission. Respondents were encouraged to provide as much detailed transaction information as possible. Access to underlying sales and cost data allowed economists to apply consistent calculation procedures, evaluate data quality, and develop representative product groupings using a broader set of observations than is typically available through traditional data collection.

The use of transaction-level records provided greater detail and understanding of product composition within product categories. Rather than relying on a limited number of selected products, economists could examine a substantially larger share of the firm's activity when evaluating representative product lines and calculating average margins. This expanded view of firm operations represents one of the primary advantages of using corporate data for PPI measures.

Operational Challenges

While the pilot demonstrated the potential value of corporate data, it also highlighted several operational challenges associated with integrating business records into statistical production.

One challenge involved the dynamic nature of product inventories within Grocery and Related Products Merchant Wholesalers industry. Product introductions, discontinuations, substitutions, and changing sales patterns can affect the composition of product groupings over time and require constant review of the microdata to ensure a level of uniformity. Also, during the pilot, respondents were often asked to categorize products into specific reporting categories. Despite their extensive knowledge of their own inventory, product assignments did not always line up with PPI classifications of the same products. These differences in classification required additional review and follow-up with respondents to ensure proper calculations.

For example, a respondent may classify egg products within a meat category while PPI classification procedures require assignment in a dairy category. These differences required economist review and follow-up discussions before data could be incorporated into production.

Additional challenges resulted from differences in firms' data structures and file formats. File formats, including product categories, inventory systems, and accounting methods, differed substantially across respondents. While data was readily available in one firm's systems it might be organized differently or unavailable in another. These differences increased the level of review and processing required before incorporating data into production.

The pilot also highlighted an unexpected limitation in existing BLS data collection systems. While most respondents provide data through the established BLS web portal in the form of a spreadsheet or .csv file, some firms preferred to transmit data through secure file platforms already used within their business operations. These systems were not supported by existing BLS collection paths, meaning establishment of technological infrastructure became a challenge to efficiently collect data from respondents.

Pilot Findings

The pilot demonstrated that firms could provide transaction-level data from existing business systems, and these datasets supported the calculation of margin prices. However, the pilot also revealed that obtaining data was only one component of successful implementation.

Participating firms provided data in a variety of formats and structures, requiring considerable effort to review, standardize, and process the information before use because existing BLS production systems were not designed to incorporate large transactional datasets for price index calculation and analysis.

In many ways, the pilot shifted the primary challenge from data acquisition to development of operational processes capable of transforming large transactional datasets into reliable statistical inputs.

Operationalizing Corporate Data for Production

From Pilot to Production

The pilot confirmed that corporate datasets could support PPI data collection within trade and other industries, providing access to more detailed information than traditional collection methods. However, the pilot also emphasized that successful integration required more than obtaining access to datasets. Differences in respondent systems, file structures, product categories, and reporting practices created challenges that should not be addressed through ad hoc processing alone.

As a result, the focus shifted from evaluating the feasibility of corporate data collection to developing repeatable production procedures for supporting respondent initiation, data acquisition, data preparation, quality review, and ongoing monthly updates. The objective was to establish procedures that balanced statistical quality, respondent burden, and operational needs while allowing corporate data to be integrated into existing PPI production processes.

These procedures were designed to move corporate data beyond a pilot environment into routine production. Rather than treating each corporate dataset as a unique collection effort, standardized procedures created consistency in data collection procedures, while allowing flexibility to accommodate differences in business operation data systems. The resulting procedures provided a scalable approach for evaluating corporate data opportunities and supporting future expansion into additional industries.

Balancing Standardization and Flexibility

One primary lesson learned during the pilot was that respondents maintain operational data in a wide variety of systems and formats. Firms differed substantially in how they organized products, managed inventory, and stored transaction records. Requiring a single reporting structure would have limited participation, while accepting unrestricted submissions would have increased processing complexity.

To address this, the procedures focus on defining minimum data requirements rather than prescribing a single reporting format. Required information generally includes product identifiers and descriptions, quantities, sales prices, and acquisition costs. Respondents can submit data in commonly used business formats, including Microsoft Excel (xlsx) and comma separated value (CSV) files, provided the required information is available.

The procedures also establish preferred reporting periods. Month- to-date transactional data is preferred because it allows for calculating current month price indexes. When this is not operationally feasible, shorter reporting periods within the pricing month are

accepted, followed by complete transaction files from the previous month. This approach balances production needs with respondent capabilities to maximize participation.

By defining minimum data requirements rather than complete file structure, economists collecting data can accommodate a broader range of respondent systems while maintaining the ability to collect the necessary data for production. The balance between standardization and flexibility is a foundational principle of the production procedures.

Integrating Corporate Data into Respondent Initiation

The transition to collecting corporate data requires modifications to traditional initiation procedures. In addition to identifying products and services represented by a firm, economists are required to evaluate whether respondents maintain datasets capable of supporting monthly reporting.

For firms identified as potential corporate data candidates, traditional collection methods and corporate data methods are presented on equal footing. Respondents are provided information about both reporting approaches and allowed to select the option that best aligns with their operational capabilities. This approach helps to continually assess respondents' willingness to participate in corporate data collection while maintaining established collection alternatives for a successful outcome.

A critical early step in initiation involves confirming that the respondent's revenue-generating activities align with the industry categories being measured. Economists work with respondents to identify and verify major product categories associated with the survey unit, ensuring that those categories reflect the economic activity represented by the industry classification. This process establishes the foundation for organizing and evaluating subsequent data.

Initiation discussions frequently involve personnel responsible for pricing, accounting, inventory management, information technology, and enterprise data systems. These personnel are not always traditional firm contacts. Additional discussions, however, are key to incorporating corporate data evaluation into the initiation process, identifying potential reporting challenges early, and reviewing reporting formats to make sure they align with respondent capabilities and contain minimum data requirements.

Transforming Operational Data into Statistical Inputs

Once the economist determines the validity and usability of the corporate data, rather than developing new procedures, implementation has focused on adapting existing BLS processes to operate efficiently with that data.

Once corporate data is received, BLS economists must transform the dataset into statistical inputs suitable for PPI production and publication. This process begins with validating that required fields such as product identifiers, quantities, sales values, and acquisition costs are present and consistently reported. Products are then assigned to PPI reporting categories, transaction-level margins are calculated, and review procedures are applied to identify unusual observations or significant changes in product composition. Finally, transactions are aggregated into comparable product lines that support monthly price index calculation for publication.

The use of the transaction-level datasets requires the capability to process large volumes of data while maintaining data quality. For example, in trade industries, transaction-level review occurs before data is aggregated into comparable product lines. This approach allows economists to identify transactions with missing costs, zero margins, negative margins or other unusual characteristics before calculating average margins. Performing review at the transaction level improves consistency and helps prevent atypical observations from influencing aggregate prices.

Review procedures are in place to assess data comprehensiveness, identify unusual observations, evaluate changes in product composition, and monitor continuity over time. These review procedures help ensure that submitted datasets remain representative of firms' activity. It is particularly important to maintain stable product groupings, evaluate the impact of changing product mixes, and identify observations that warrant additional review or follow-up with respondents.

Implementation has also demonstrated the continued importance of analytical and automation tools. The size of transactional datasets makes manual review and processing increasingly resource intensive. To mitigate this additional work, economists can use tools such as R and Python to support data validation, transformation, aggregation, quality assurance, and analytical review.

Despite these automations, economist review remains a critical component of the production process. Economists continue to play an essential role in evaluating product classifications, understanding respondent reporting practices and assessing changes in product composition that may affect price index quality.

The need for scalable review procedures and supporting analytical tools is critical to transforming business records into reliable statistical outputs. The implementation experience has demonstrated that successful integration of corporate data depends not only on data availability, but processes and infrastructure needed to transform transactional records into inputs for official statistics.

Challenges, Lesson Learned, and Ongoing Considerations

While significant progress has been made in developing procedures and operational frameworks for the use of corporate data for PPI measures, many challenges continue to evolve as new respondents, industries, and data structures are incorporated.

Experience to date suggests that scaling corporate data collection will require continued attention to workloads and resource allocation. While procedures have improved consistency and efficiency in collecting and processing large datasets, the work remains resource intensive. For example, some participating firms maintain inventories consisting of thousands of products and generate files containing hundreds of thousands of records per reporting period. While these datasets provided substantially greater detail than traditional collection methods, they also increased the complexity of ongoing data validation, review, and processing activities.

As implementation continues, additional experiences will help refine approaches for maintaining representative measurement and leveraging the increased level of detail available within corporate datasets. More data does not automatically produce better statistics. Statistical quality continues to depend on the ability to transform detailed business records into stable and representative measures of economic activity.

Implementation has also reinforced the importance of continued investment in technical training, documentation, analytical capabilities, and automation. As the size and complexity of corporate datasets increase, the ability to efficiently review, process, and analyze large volumes of information becomes increasingly important. Tools such as R and Python have already played an important role in supporting these activities.

Corporate data implementation requires the development of operational capabilities that support sustainable production procedures. Corporate datasets can enhance statistical products and reduce respondent burden. However, achieving these benefits requires more than just access to the data. Statistical agencies must develop operational frameworks capable of supporting respondent engagement, data transformation, quality review, statistical processing, and ongoing maintenance. The long-term value of corporate data depends not only on data availability, but also on the ability to build and sustain the operational capabilities needed to use the data effectively.

Conclusion

The growing availability of corporate datasets presents significant opportunities for statistical agencies seeking to improve the detail, timeliness, and representativeness of official statistics. However, as demonstrated by the BLS experience with PPIs, the successful use of corporate data requires substantially more than access to large volumes of data. Corporate data must be evaluated against established methodological requirements, integrated into existing production, and supported by operational procedures capable of sustaining recurring collection and processing.

This paper illustrates the evolution from a pilot effort focused on testing the feasibility of corporate data collection to the development and implementation of operational procedures supporting ongoing production. Through respondent engagement, data review, collection procedures, quality controls, and data processing, BLS has worked to incorporate corporate data into official PPI measures. Additionally, the implementation efforts demonstrated that traditional collection and corporate data collection can co-exist within the same operational process, allowing respondents to provide data through the method most appropriate to their business operations.

As additional firms and industries are incorporated, new challenges continue to emerge related to respondents, data structures, product changes, automation, and long-term maintenance. Each implementation provides additional insight into what types of data can support production requirements and how collection, review, and processing procedures can be further refined. Continued investment in analytical tools, automation, data management capabilities, and staff expertise will remain important as corporate data collection expands to support data collection needs to continue publication of a quality PPI measure.