

Responsável: Rodrigo Barbosa de Cerqueira
Silvânia Ferreira Conceição
Samuel Veríssimo Portela Mendes

Empresa: Superintendência de Estudos Econômicos e Sociais da Bahia

Workshop CEMEAI: *Desafio da Padronização e Classificação de Itens em Notas Fiscais Eletrônicas*

A SEI

A Superintendência de Estudos Econômicos e Sociais da Bahia – SEI tem como objetivo desenvolver estudos e projetos que subsidiem a formulação e a avaliação de políticas públicas, bem como de planos e programas de desenvolvimento do Estado. Com a missão de gerir o conhecimento sobre a realidade física, ambiental, econômica e social da Bahia, a SEI realiza pesquisas e análises, baseadas em dados, sobre o estado e seus municípios. Além disso, esta superintendência tem desenvolvido ações voltadas para o aperfeiçoamento da capacidade de análise de grandes bancos de dados. Atualmente, em articulação com a Secretaria da Fazenda do Estado da Bahia (SEFAZ-BA), a SEI tem conduzido estudos sobre a dinâmica econômica estadual, no âmbito das transações financeiras formais, por meio dos registros das notas fiscais eletrônicas.

Contexto

As notas fiscais eletrônicas ao consumidor final (NFC-e) constituem registros administrativos digitais que documentam operações comerciais formais de compra e venda, fornecendo informações detalhadas sobre produtos e serviços, como descrições, quantidades e valores. Para a SEI, esses dados são de grande relevância para analisar a movimentação econômica do estado e de seus municípios, identificar padrões de consumo, avaliar o impacto de eventos programados sobre o consumo local e subsidiar a formulação e a avaliação de políticas públicas. A abrangência e a granularidade das NFC-e

permitem a realização de análises robustas em grandes bases de dados, consolidando-as como uma fonte estratégica para o planejamento, monitoramento e a gestão pública baseada em evidências.

Problema

A realização de estudos a partir de notas fiscais eletrônicas depende da classificação e padronização adequadas dos itens comercializados. Entretanto, grande parte das informações sobre a descrição dos produtos é inserida diretamente pelos emissores das notas, gerando múltiplas variantes para um mesmo item em razão de nomenclaturas não padronizadas, variações regionais da linguagem, abreviações, descrições incompletas ou erros ortográficos e de digitação. Essa ausência de uniformidade compromete a classificação consistente dos itens e dificulta a realização de análises precisas sobre padrões de consumo, tornando imprescindível o desenvolvimento de procedimentos adicionais para padronizar e validar as descrições em grandes bases de dados. Embora cada produto possua um código de Nomenclatura Comum do Mercosul (NCM), esse código isoladamente é insuficiente, já que mercadorias com características distintas podem estar agrupadas no mesmo NCM, como é o caso do código “2801.30.00 - Flúor; bromo”.

Desafio

Diante das inconsistências e da falta de padronização nas descrições de produtos e serviços, o desafio consiste em desenvolver estratégias capazes de identificar e agrupar corretamente itens equivalentes presentes nos registros. Em particular, espera-se que sejam propostas metodologias para o tratamento e padronização das informações textuais, tornando os dados mais consistentes e comparáveis. As soluções devem incluir a identificação de padrões nas descrições, a correção de variações de grafia e abreviações, unidade de medida e a associação dos registros aos códigos válidos da Nomenclatura Comum do Mercosul (NCM). Dentro de um mesmo NCM, as descrições referentes ao mesmo produto devem ser unificadas e atribuídas ao mesmo grupo, garantindo uma representação padronizada para todos os itens equivalentes, incluindo aqueles com erros ortográficos, abreviações ou descrições não convencionais. Por exemplo, diferentes variações para representar farinha de mandioca de 1kg como “FARIN KICALDO BRANCA 1KG MAND”, “FARINHA DE MANDIOCA KICALDO 1KG”, “FARINHA BDE MANDIOCA KICALDO 1KG” ou “FARINHA BRANCA FINA SECA 1 KG KICALDO” — devem ser reunidas em uma única descrição padronizada.

O objetivo final é produzir uma base de dados mais

organizada e confiável, capaz de viabilizar análises sobre as vendas registradas nas NFC-e com maior precisão, consistência e comparabilidade.

Dados

As Notas Fiscais de Consumidor Eletrônicas emitidas na Bahia foram disponibilizadas pela Sefaz-BA por meio de um Acordo de Cooperação Técnica firmado com a SEI. Os dados estão armazenados em formato Parquet no Hadoop Distributed File System (HDFS), na nuvem privada da SEI, garantindo segurança, alta disponibilidade e capacidade de processamento distribuído.

Esses registros documentam transações diárias de vendas realizadas em estabelecimentos do comércio formal e estão estruturados em colunas que incluem:

- Data de emissão da nota: registro diário das transações (formato YYYY-MM-DD);
- Código IBGE do município do emitente: localidade onde a transação foi realizada;
- Descrição do produto ou serviço: nome do item comercializado, conforme declarado pelo estabelecimento;
- NCM (Nomenclatura Comum do Mercosul): classificação de mercadorias em operações de comércio exterior e tributação, com oito dígitos que detalham características do produto (ex.: chocolate = 1806.31.10);
- GTIN (Global Trade Item Number): código global de identificação do produto (ex: código de barras EAN/UPC);
- unidade de medida: tipo de unidade utilizado (ex: litro, quilograma, unidade);
- quantidade vendida: volume comercializado por transação;
- valor unitário vendido: preço individual do produto comercializado.

Por questões de confidencialidade e de proteção à identificação indireta dos contribuintes, algumas variáveis não estão disponíveis nas bases divulgadas, como as colunas correspondentes ao número de identificação mascarado do emitente e ao bairro do estabelecimento. Assim, as análises e os resultados apresentados consideram apenas as variáveis autorizadas, preservando o sigilo das informações fiscais.