FERRAMENTA DE ETL - SELEÇÃO EM BIG DATA PARA CONDENSAÇÃO DE DADOS DE PRODUTOS E TRIBUTOS[1]

Paulo R. M. de Oliveira[2]

Resumo: Implantação de sistema e manutenção dos dados de produtos e tributos tem um alto custo para as empresas de varejo. Nos lançamentos manuais das informações básicas para o seu funcionamento são comuns dados incompletos ou imprecisos. Este trabalho propõe a seleção em grande número de cadastros de mercadorias para a geração de uma base de dados de supermercado completa para automatização de cadastros e qualificação das informações através da criação de uma ferramenta desktop de processos de ETL.

Abstract: System implementation and maintenance of product and tax data has a high cost for retail companies. In manual entries of basic information for its operation, incomplete or inaccurate data is common. This work proposes the selection of a large number of merchandise registers for the generation of a complete supermarket database for automation of registers and qualification of information through the creation of a desktop tool for ETL processes.

Palavras-Chave: DBA, SQL, Big Data, ETL, SGDB, ERP, Data Warehouse, Supermercado, Varejo, Produtos, Tributos, GTIN, NCM.

1 INTRODUÇÃO

A rotina diária de manutenção de base de dados costuma ser desgastante para as empresas. Em supermercados é visível tais dificuldades principalmente em cadastro de produto, devido à grande quantidade de itens e alta complexidade de seus respectivos tributos. Neste cenário, falhas humanas ou no ERP são recorrentes.

O Brasil é um dos países com a maior complexidade fiscal do mundo, sendo que mudanças podem ocorrer a qualquer momento (TAXWEB, 2019). De acordo com Pietro (2017):

> O Brasil é constituído por 27 estados, cada qual com sua legislação tributária específica. O cenário piora quando seus 5.564 municípios também possuem tributos próprios. A cereja do bolo são as frequentes mudanças nas leis criadas pelos governos Federal, Estadual e Municipal. Só de ICMS são publicadas, por ano, mais de 5 mil normas. A complexidade do caso só aumenta quando não há um banco de dados exclusivo para lidar com a publicação simultânea dessas leis.

É de extrema importância manter os registros precisos para o momento de uma transação de entrada ou saída de mercadoria. Assim declarando ao Fisco os valores corretos evitando prejuízos. Segundo Taxweb (2019), com a complexidade de nosso sistema tributário

erros em notas fiscais são comuns. Por isso atualmente muitas empresas preferem terceirizar o cadastramento e atualização de seus dados com ferramentas de validação tributária. Para Junqueira (2016):

> O único caminho é manter o seu cadastro de produtos atualizado. Acompanhar as novidades e mudanças, não acreditar que um cadastro correto irá se manter se não houver acompanhamento. Sabemos que é um trabalho exaustivo, cuja execução deve ser avaliada pela empresa. Qual é o melhor caminho: qualificar os funcionários da empresa e realizar internamente ou ter ajuda de empresas especializadas? Temos constatado que o custo de empresas compensa, em muito, quando comparados à necessidade de manter equipe de colaboradores constantemente "antenadas" nas constantes alterações legais.

Com a complexidade de nosso sistema tributário, como sistema ERP de empresas varejistas poderiam manter tantos dados fidedignos? A solução proposta neste trabalho é realizar constantes validações nos cadastros através de comparações à uma base de dados exemplar. Tal base pode ser periodicamente atualizada através da ferramenta OpenETL que faz a leitura em diversas fontes de dados passando principalmente, dentre várias transformações, por um algoritmo de condensação *most frequent*.

2 CONTEXTUALIZAÇÃO

2.1 BIG DATA

Big Data é um termo bastante usado atualmente, refere-se a análise de grandes volumes de dados transformando-os em informações úteis. Segundo um estudo da Gartner, o aumento das iniciativas de Big Data irá demandar contratação de milhares de profissionais capacitados e com habilidades na área (MACHADO, 2014, p.13).

As tecnologias passam por diversas transformações, mas grande quantidade de dados coletados precisam ser persistidos. Muito são os motivos pelo qual informações necessitam passar por migrações ou reorganizações de forma automatizada, seja por empresas trocando de sistema de gestão ou SGDB, atualizações de versão dos sistemas, "desnormalizações" de modelagens relacionais para soluções de BI e tomada de decisão, mudança de arquitetura cliente-servidor para multicamadas com dados em nuvem, entre outros. Para tais tarefas existem diversas soluções adotadas por DBA (administradores de bancos de dados) como programas especializados em extração de base de dados e geração de scripts SQL que serão mencionados posteriormente. Em 2001, Özsu e Valduriez descreveram:

> Sistemas de bancos de dados nos levaram de um paradigma de processamento de dados no qual cada aplicativo definia e mantinha seus próprios dados até um paradigma em que os dados são definidos e administrado de forma centralizada.

Segundo Inmoh (1997), há pelo menos duas razões para o uso de programa de extração:

a) Porque é possível retirar dados do caminho do processamento online de alta performance com um programa de extração, não havendo conflito em termos de performance quanto os dados precisam ser analisados coletivamente;
b) Quando os dados são retirados do domínio do processamento de transações online com um programa de extração, ocorre na mudança no controle dos dados. O usuário final acaba "tendo a posse" dos dados quando ele assume o controle sobre os mesmos.

2.2 ETL

Este projeto faz uso de conceitos difundidos no estudo de SAD (Sistemas de Apoio a Decisão) como Business Intelligence, Data Warehouse e principalmente o de ETL (alguns autores mais antigos denominavam de Camada de Integração e de Transformação - CIT).

Considera-se processo ETL, do inglês *Extract Transform Load*, um conjunto de processos para trazer dados de sistemas para uma base de dados, providos não só de sistemas, mas também de websites, bases de e-mails e de redes sociais, arquivos de texto dos mais variados contextos e bases de dados pessoais (TANAKA, 2015). Visa trabalhar com toda a parte de extração de dados origens de diversos sistemas, transformação desses dados conforme regras de negócios e por fim o carregamento dos dados geralmente para uma estrutura de Data Warehouse.

2.2.1 Data Warehouse e Data Marts

Um Data Warehouse é um conjunto de dados baseado em assunto, integrado, não-volátil, e variável em relação ao tempo, de apoio às decisões gerenciais (INMOH, 1997, p.33). Os projetos de Data Warehouse consolidam dados de diferentes fontes (NOVAIS, 2012).

Inmoh (1997) cita que o DW é o alicerce do processamento dos Sistemas de Apoio a Decisão. Ainda Segundo Inmoh (1997), em virtude de haver uma fonte única de dados integrados no Data Warehouse, e uma vez que os dados apresentam condições de acesso, a

tarefa do analista de SAD no ambiente de Data Warehouse é incomensuravelmente mais fácil do que no ambiente clássico.

Os Data Marts são estruturas moldados granulares encontrados no Data Warehouse corporativo. Os Data Marts pertencem aos departamentos específicos dentro de uma empresa, geralmente finanças, contabilidade, vendas ou marketing, e são moldados pelos requerimentos dos departamentos. Consequentemente o design de cada Data Mart é único (INMOH; TERDEMAN; IMHOFF, 2001).

A fase de ETL é considerada uma das mais críticas na geração de um DW, porém, nada impede que suas ferramentas sejam aplicadas para outros fins como dar suporte a banco de dados transacionais OLTP (Online Transaction Processing). Os principais objetivos do processo ETL são:

a) Remover erros e corrigir dados faltantes;

b) Assegurar a qualidade dos dados;

c) Capturar o fluxo de dados transacionais;

d) Ajustar dados de múltiplas origens e usá-los juntos;

e) Fornecer estruturas de dados para serem utilizadas por ferramentas pelos analistas responsáveis pelo desenvolvimento;

f) Fornecer dados em formato físico para serem usados por ferramentas dos usuários finais (DA COSTA, 2009).

> Um processo ETL pode ser feito de forma manual ou utilizando ferramentas próprias para tal. O processo feito por ferramentas possuem mais vantagens em relação ao processo manual: desenvolvimento mais rápido e simples, não é necessário ter profundos conhecimentos em programação, geram automaticamente os metadados, apresentam estágios nativos de conexão a banco de dados, entre outras. As ferramentas têm um nível de aprendizado acentuado, mas mesmo sendo uma ferramenta de difícil utilização, que exige investimentos em pessoal, são compensados com o desempenho e flexibilidade da mesma (DA COSTA, 2009).

2.2.2 Processo de ETL

2.2.2.1 Fase de Extração

A extração dos dados dos sistemas de origem (também chamados *Data Sources*) é o primeiro passo do processo ETL. Cada *Data Source* pode estar em um banco de dados diferente, de plataformas distintas, ou até planilhas, arquivos texto, etc.

O grau de dificuldade depende diretamente de como será o cenário a enfrentar nos sistemas de origem, que podem estar armazenados em esquemas comuns (homogêneos) ou em estruturas diferentes (heterogêneas); em um banco de dados comum ou com os dados espalhados por diferentes bancos (NETO, 2012).

Em estruturas heterogêneas a seleção de dados do ambiente operacional pode ser muito complexa pois, em geral, é necessário selecionar vários campos de um sistema transacional para compor um único campo no DW ou vice-versa (VARGAS, 2009).

> Podem existir várias fontes de dados diferentes para compor uma informação, que pode ser oriunda de uma planilha Excel, por exemplo, enquanto uma outra informação que serve para compor um mesmo fato vem de um arquivo texto. Quando há vários arquivos de entrada, a escolha das chaves deve ser feita antes que os arquivos sejam intercalados. Isso significa que, se diferentes estruturas de chaves são usadas nos diferentes arquivos de entrada, então se deve optar por apenas uma dessas estruturas. Os arquivos devem ser gerados obedecendo a mesma ordem das colunas estipuladas no ambiente de DW (VARGAS, 2009).

As rotinas de extração devem ser capazes de isolar somente os dados que foram inseridos e atualizados desde a última extração, sendo este processo conhecido como *Refresh*. A melhor política de *Refresh* deve ser avaliada pelo administrador do DW, que deve levar em conta características como as necessidades dos usuários finais, tráfego na rede e períodos de menor sobrecarga (DA COSTA, 2009).

Não existe um ponto determinado onde vão ocorrer as transformações, elas podem ocorrer durante a extração, na movimentação dos dados na staging area, ou até na carga dos dados no DW mesmo que raros, mas sempre com o objetivo de transformar os dados em informação (NETO, 2012). Existem diversas técnicas de Extração de Informação que podem ser enquadradas dentro da área de recuperação de informação, já que podem ser vistas como técnicas especiais de indexação. Entre as técnicas que podem ser aplicadas, podemos citar:

a) Sumarização, que identifica as palavras e frases mais importantes de um documento ou conjunto de documentos e gera um resumo ou sumário;
b) Clustering, é um método de descoberta de conhecimento utilizado para identificar correlacionamentos e associações entre objetos, permitindo a identificação de classes;
c) Classificação, identifica a que classe ou categoria a que determinado documento pertence, utilizando como base o seu conteúdo (ZORZO, 2009).

Em projetos de ETL é necessário escolher qual a área para estagiar os dados. A decisão depende do ambiente e dos requisitos do negócio. Se for feito em memória ganha-se

desempenho pois leva-se os dados da origem ao alvo final o mais rápido possível. Mas se for feito em disco terá as seguintes vantagens:

a) Recuperabilidade: é uma boa prática armazenar os dados assim que são colhidos da fonte e após cada uma das transformações mais importantes. Dessa forma, não é necessário entrar no sistema-fonte novamente caso ocorra algum erro. Isso é especialmente importante com dados na web, que podem não estar disponíveis novamente;
b) Backup: usualmente, os DWs contém volumes massivos de dados, o que impossibilita seu backup completo. Se os arquivos de carga forem guardados, o DW pode ser recuperado caso haja algum problema;
c) Auditoria: com os dados salvos em áreas de estagiamento fica muito mais fácil fazer uma auditoria do processo de ETL apenas comparando-se os dados de entrada com os dados de saída, levando-se em consideração as modificações no código ETL (ZORZO, 2009).

2.2.2.2 Fase de Transformação

Limpeza, Ajustes e Consolidação (ou também chamada transformação): É nesta etapa que realizamos os devidos ajustes, podendo assim melhorar a qualidade dos dados e consolidar dados de duas ou mais fontes (NOVAIS, 2012).

É o trabalho que concentra o maior esforço de análise, consequentemente gasta-se maior tempo, embora o grau de dificuldade dependa diretamente de fatores que são determinados pela qualidade dos dados de origem, das regras de negócios a serem aplicadas e da disponibilidade ou não de documentação atualizada das bases de dados transacionais (NETO, 2012).

> Na etapa de pré-processamento/limpeza é feito o tratamento de ausências de dados, eliminação de dados incompletos, repetição de registros, problemas de tipagem, tratamento de dados inconsistentes. Os tratamentos destes relevantes problemas na qualidade de dados podem ocorrer devido a omissões na entrada de dados, problemas na conversão entre bases de dados, falhas em mecanismos de leitura, entre outros. O tratamento consiste em, tipicamente, substituir os valores ausentes, incompletos ou inconsistentes por valores default, substituí-los por um valor médio ou simplesmente excluí-los (DA COSTA, 2009).

Como exemplos de limpeza de dados tem-se: a correção de erros de digitação, a descoberta de violações de integridade, a substituição de caracteres desconhecidos, a padronização de abreviações (DA COSTA, 2009).

Transformação é a etapa onde os dados sofrem uma determinada transformação, para que os dados fiquem em um formato padrão. Como exemplo de transformação tem-se a conversão de valores quantitativos em valores categóricos, ou seja, cada valor equivale a uma faixa. Idade entre 0 e 18 equivale a Faixa 1; idade entre 19 e 21 equivale a Faixa 2 e assim por diante (DA COSTA, 2009).

O passo seguinte é colocar os dados em uma forma homogênea, aplicando uma metodologia de comparação de representações, que inclui os critérios a serem utilizados na identificação de semelhanças e conflitos de modelagem. Conflitos de modelagem podem ser divididos em dois: semânticos e estruturais. Os conflitos semânticos são todos aqueles que envolvem o nome ou palavra associada às estruturas de modelagem, como por exemplo, mesmo nome para diferentes entidades ou diferentes nomes para a mesma entidade. Já os conflitos estruturais englobam os conflitos relativos às estruturas de modelagem escolhidas, tanto no nível de estrutura propriamente dito como no nível de domínios. Os principais tipos de conflito estruturais são os conflitos de domínio de atributo que se caracterizam pelo uso de diferentes tipos de dados para os mesmos campos. Conflitos típicos de domínio de atributo encontrados na prática são:

a) Diferenças de unidades: quando as unidades utilizadas diferem, embora forneçam a mesma informação. Exemplo: distância em metros ou quilômetros.
b) Diferenças de precisão: quando a precisão escolhida varia de um ambiente para outro. Exemplo: quando o custo do produto é armazenado com duas posições ou com seis posições decimais.
c) Diferenças em códigos ou expressões: quando o código utilizado difere um do outro. Exemplo: sexo representado por M ou F e por 1 ou 2.
d) Diferenças de granularidade: quando os critérios associados a uma informação, embora utilizando uma mesma unidade, são distintos. Exemplo: quando horas trabalhadas correspondem às horas trabalhadas na semana ou às horas trabalhadas no mês.
e) Diferenças de abstração: quando a forma de estruturar uma mesma informação segue critérios diferentes. Exemplo: endereço armazenado em um atributo único ou subdividido em rua e complemento) (DA COSTA, 2009).

Outros tipos de modificações nos dados podem ser desejadas, tais como:

a) Realizar somatório por um determinado campo;
b) Selecionar o primeiro ou o último registro de uma determinada coluna;
c) Selecionar o valor máximo ou mínimo de uma coluna;
d) Fazer a média dos registros de uma coluna;
e) Ordenar os dados por uma coluna ou mais (DA COSTA, 2009).

2.2.2.3 Fase de Carregamento

Entrega ou Carga dos dados: Consiste em fisicamente estruturar e carregar os dados para dentro da camada de apresentação seguindo o modelo dimensional (NOVAIS, 2012). Esta etapa possui uma complexidade alta e alguns fatores devem ser levados em conta:

a) Integridade dos dados: ao realizar a carga é necessário checar os campos que são chaves estrangeiras com suas respectivas tabelas para certificar que seus dados estão de acordo com a tabela primária.
b) Carga incremental ou a carga por cima dos dados: a carga incremental, normalmente, é feita em tabelas fatos, e a carga por cima dos dados é feita em tabelas dimensão onde o analista terá que excluir os dados existentes e inserir todos os dados novamente. Mas em alguns casos, poderá acontecer que as tabelas de dimensão terão que manter o histórico, então, o mesmo deverá ser mantido. A decisão para o tipo de carga a ser feita deve ser planejada com cuidado, pois a carga pode levar um tempo elevado.
c) Os arquivos a serem gerados devem obedecer a mesma ordem das colunas que foram estipuladas para o ambiente DW.
d) Criação de rotinas: apesar de ter ferramentas especializadas nesta etapa, muitas vezes é necessário criar rotinas de carga para atender determinadas situações que poderão ocorrer (DA COSTA, 2009).

Alguns Data Warehouses podem substituir as informações existentes semanalmente, com dados cumulativos e atualizados, ao passo que outro DW (ou até mesmo outras partes do mesmo DW) podem adicionar dados a cada hora. A latência e o alcance de reposição ou acréscimo constituem opções de projeto estratégicas que dependem do tempo disponível e das necessidades de negócios (NOVAIS, 2012). Processos ETL podem operar em dois modos básicos:

a) Modo On-line: o tempo a partir do momento da execução da transação no sistema legado é medido em pequenas unidades de tempo, como milissegundos, até que a transação seja refletida no ambiente do DW;
b) Modo Batch (processamento em lote): as atualizações nos ambientes legados são armazenadas em memória ou marcadas como modificadas, então, em algum momento conveniente (talvez durante a noite), o lote das transações legadas é rodado no processo de ETL, isso quer dizer que, desde o tempo de a transação ser processada no ambiente legado até o DW conhecê-la, 24 horas ou mais podem ter transcorrido (ZORZO, 2009).

2.2.3 Normalização vs. Desnormalização

A técnica de Normalização é um processo matemático formal, que tem seus fundamentos na teoria dos conjuntos (MACHADO; ABREU, 2011, p. 177), onde cada tabela representa um pequeno aspecto único do negócio, e cada tabela tem relação chave/chave estrangeira com outras tabelas (INMOH; TERDEMAN; IMHOFF, 2001, p. 41). No momento da criação do modelo dimensional surgem dúvidas de qual nível de normalização (1FN, 2FN, 3FN, FNBC, 4FN, 5FN) deve-se aplicar. Depende qual finalidade do modelo, no caso de ambiente operacional/legado a normalização é uma boa prática pois combate as chamadas "anomalias":

a) Anomalia de repetição (redundâncias);
b) Anomalia de atualização (inconsistências);
c) Anomalia de inserção (inconsistências);
d) Anomalia de eliminação (inconsistências) (ÖZSU; VALDURIEZ, 2001).

Mas a normalização, ou o Modelo *Snowflake*, não é adequado em Data Warehouses, pois as consultas acabam ficando complexas tanto no código quanto no processamento. Por isso surge a necessidade de migrar os dados para modelos desnormalizados (Modelo Estrela ou *Star Schema*), ideal para o uso nos sistemas OLAP (Online Analytical Processing).

2.3 INFORMAÇÕES GERENCIAIS E TRIBUTÁRIAS

Neste tópico estão sendo destacados os principais campos contidos no cadastro de um produto embalados, além da descrição, classe, valor de venda, entre outros.

2.3.1 Códigos Gerenciais

2.3.1.1 GTIN

Sigla de “Global Trade Item Number” é um identificador para itens comerciais desenvolvido e controlado pela GS1, antiga EAN/UCC. É ele que aparece abaixo dos códigos de barras, amplamente utilizados no varejo físico para identificação de produtos. Anteriormente chamados de códigos EAN, são atribuídos para qualquer item (produto ou serviço) que pode ser precificado, pedido ou faturado em qualquer ponto da cadeia de suprimentos. O GTIN é utilizado para recuperar informação pré-definida e abrange desde as matérias-primas até produtos acabados (GS1 BRASIL, 2020). Possui quatro padrões:

- GTIN-12 (UPC-A): número de 12 dígitos usado principalmente na América do Norte;
- GTIN-8 (EAN / UCC-8): número de 8 dígitos usado predominantemente fora da América do Norte;
- GTIN-13 (EAN / UCC-13): número de 13 dígitos usado predominantemente fora da América do Norte;
- GTIN-14 (EAN / UCC-14 ou ITF-14): número de 14 dígitos usado para identificar itens comerciais em vários níveis de embalagem (GTIN INFO, 2020).

A GS1 possui um Cadastro Nacional de Produtos (CNP) gerando para as marcas ou empresas um número exclusivo para cada produto, por isso torna-se ideal e foi usado como *primary-key* na entidade de produtos neste projeto.

O CNP é integrado com o Cadastro Centralizado de GTIN (CCG) que é o banco de dados da SEFAZ que contém um conjunto de informações sobre os produtos (GS1 BRASIL, 2020).

2.3.1.2 NCM/SH

A Nomenclatura Comum do Mercosul (NCM) é uma nomenclatura regional para categorização de mercadorias adotada pelo Brasil, Argentina, Paraguai e Uruguai desde 1995. Nos seis primeiros dígitos, toma por base o SH (Sistema Harmonizado de Designação e de Codificação de Mercadorias) mantido pela Organização Mundial das Alfândegas (OMA), que

foi criado para melhorar e facilitar o comércio internacional e seu controle estatístico. E seus dois últimos dígitos são definidos pelo Mercosul (DINOM, 2019).

Contém pouco mais de 10.000 códigos que seguem uma ordenação bem elaborada numa estrutura hierárquica em árvore:

a) 6 Regras Gerais para Interpretação do Sistema Harmonizado e 2 Regras Gerais Complementares;

b) Notas de Seção, de Capítulo, de Subposição e Complementares;

c) Lista ordenada de códigos em níveis de posição (4 dígitos), subposição (5 e 6 dígitos), item (7 dígitos) e subitem (8 dígitos), distribuídos em 21 Seções e 96 Capítulos (DINOM, 2019).

A NCM é fundamental para determinar as alíquotas de impostos envolvidos nas operações de comércio exterior e de diversos tributos internos nas operações com mercadorias, entre outras utilizações. Além disso, a NCM é base para o estabelecimento de direitos de defesa comercial, sendo também utilizada no âmbito do ICMS, na valoração aduaneira, em dados estatísticos de importação e exportação, na identificação de mercadorias para efeitos de regimes aduaneiros especiais, de tratamentos administrativos, de licença de importação, etc (DINOM, 2019). Então, posso relacionar todos os impostos na tabela de NCM? Quase, segundo Junqueira (2017), a análise pelo EAN é mais precisa, uma vez que produtos de mesmo NCM podem ter variações relevantes na tributação.

2.3.1.3 CEST

> A sigla CEST refere-se ao Código Especificador da Substituição Tributária. Este código estabelece uma identificação padrão para todas as mercadorias e bens sujeitos à Substituição Tributária. Este código deverá ser informado em cada produto, e deverá aparecer em todos os documentos fiscais emitidos com Substituição Tributária. O objetivo da utilização do código CEST é estabelecer uma padronização e identificação das mercadorias sujeitas ao regime de Substituição Tributária e de antecipação do recolhimento do ICMS (SIGE CLOUD, 2020).

O CEST é composto por sete dígitos numerais, onde:

a) 1º e 2º dígitos: representam o segmento da mercadoria;

b) 3º, 4º e 5º dígitos: correspondem ao item de um segmento de mercadoria;

c) 6º e 7º dígitos: relacionam-se à especificação do item (SIGE CLOUD, 2020).

2.3.1.4 CST

Código de Situação Tributária é um número inteiro positivo, índice de tabela, que indica que tipo de tributação se aplica a um produto ou serviço. Cada imposto: PIS, COFINS, IPI e ICMS, possui sua tabela de CST ou CSOSN (Código de Situação da Operação do Simples Nacional) no caso de ICMS de empresas optantes pelo Simples Nacional.

2.3.1.5 CFOP

O CFOP é uma sequência numérica de 4 dígitos que identifica a natureza de circulação de produtos e a prestação de serviços em todo o Brasil, e até mesmo no exterior. Em resumo, o CFOP indica se há ou não recolhimento de impostos sobre produtos transportados e como isso deve ocorrer (NASCIMENTO, 2020).

2.3.2 Impostos Federais

2.3.2.1 PIS

O Programa de Integração Social foi instituído em 1970 com o objetivo de pagar o seguro-desemprego e abonos salariais aos trabalhadores, além de benefícios a servidores públicos. A alíquota desse tributo pode variar de acordo com o regime tributário utilizado pela empresa, sendo que os optantes pelo Lucro Presumido, onde não há descontos de créditos, devem arcar com 0,65% sobre o faturamento. Já os optantes pelo Lucro Real, que têm direito a deduções da quantia a se pagar por meio de créditos, devem arcar com alíquotas de 1,65% (TAXWEB, 2019).

2.3.2.2 COFINS

Contribuição para o Financiamento da Seguridade Social: seu objetivo é custear o financiamento da seguridade social em todo o território nacional. Esse tributo incide sobre a receita bruta das empresas que prestam serviços e também pode ter uma alíquota variável de acordo com o regime tributário escolhido pela empresa. Para optante do Lucro Presumido, a

taxa é fixa em 3%, já no Lucro Real, o valor sobe para 7,6% sobre o total da nota (TAXWEB, 2019).

2.3.2.3 IPI

O Imposto sobre Produtos Industrializados, instituído pela União, incide sobre uma categoria específica de bens, neste caso, produtos industrializados nacionais e estrangeiros. Esse Imposto é obrigação tributária principal devida pelas indústrias e estabelecimentos equiparados (CAMARGO, 2017).

Portal Tributário (2020) explica que o campo de incidência do imposto abrange todos os produtos com alíquota, ainda que zero, relacionados na Tabela de incidência do Imposto sobre Produtos Industrializados (TIPI), observadas as disposições contidas nas respectivas notas complementares, excluídos aqueles a que corresponde a notação "NT" (não-tributado). A TIPI tem por base a NCM acrescida do número “EX” (Exceção) e das alíquotas do IPI.

2.3.2.4 IRPJ

O Imposto de Renda de Pessoa Jurídica incide sobre todas as organizações que mantém um CNPJ e sobre pessoas físicas equiparadas. Ele é calculado diretamente sobre a base de lucro obtida, sendo necessário verificar o regime tributário escolhido, Lucro Real ou Lucro Presumido (TAXWEB, 2019).

2.3.3 Impostos Estaduais

2.3.3.1 ICMS

O Imposto sobre Operações de Circulação de Mercadorias e Prestação de Serviços de Transporte Interestadual, Intermunicipal e de Comunicação tem seu recolhimento direto em nota fiscal. Como esse imposto é aplicável em vários casos, existem diversas regras acerca de seu cálculo e também variadas alíquotas de acordo com o serviço prestado. Além das observações acerca da legislação em cada um dos estados da federação nos quais o serviço será prestado (TAXWEB, 2019).

3 TRABALHOS CORRELATOS

3.1 O PROCESSO DE ETL NA CONSTRUÇÃO DE CONHECIMENTO EM UMA APLICAÇÃO DE UMA EMPRESA SEGURADORA

A monografia desenvolvida na Universidade Federal de Juiz de Fora no Curso de Bacharelado em Ciência da Computação realizou estratégias para auxiliar no processo de tomadas de decisão. Dentre essas estratégias, destacam-se o processo de KDD (Knowledge Discovery in Databases) e de DW (Data Warehouse). Para que o desenvolvimento do processo de KDD e do DW seja feito com sucesso é preciso realizar um tratamento nos dados das bases utilizadas. Este tratamento é conhecido como ETL (Extraction, Transformation and Load) e consiste em extrair os dados dos bancos de dados, realizar um processo de limpeza e transformação e, então, realizar a carga. Este é o foco principal deste trabalho, que além de apresentar os princiais conceitos de DW e KDD é feito um processo prático de ETL utilizando uma ferramenta própria para isto (DA COSTA, 2009, p. 6).

3.2 ETL 2.0: UMA PROPOSTA DE EXTENSÃO AO PROCESSO DE EXTRAÇÃO, TRANSFORMAÇÃO E CARGA VOLTADA À INTEGRAÇÃO DE DADOS ESTRUTURADOS E NÃO ESTRUTURADOS

Trabalho de Conclusão de Curso da Universidade Federal de Santa Catarina no Curso de Sistemas de Informação propõe uma extensão ao processo de ETL tradicional, integrando as duas visões (estruturada e não estruturada) em um modelo genérico para apoiar a tomada de decisão. Para atingir esses objetivos, foram desenvolvidas extensões a uma ferramenta de ETL de código aberto. Posteriormente, a ferramenta foi utilizada sobre dados reais para suportar o processo de ETL, produzindo como resultado um DW. Analisando-se os dados inseridos no DW a partir do processo completo, foi possível encontrar informações de correlação entre entidades pertencentes a ambas as classificações de dados. A principal contribuição do trabalho reside na extensão ao processo de ETL tradicional e na proposição, ainda que inicial, de um modelo de DW genérico para análise de relações entre entidades que promovem suporte a diversos cenários de tomada de decisão (ZORZO, 2009, p. 3).

3.3 CONSTRUÇÃO DE DATA WAREHOUSE PARA PEQUENAS E MÉDIAS EMPRESAS USANDO SOFTWARE LIVRE

O Trabalho de Conclusão de Curso da Universidade do Planalto Catarinense no Curso de Bacharelado de Sistemas de Informação busca demonstrar a viabilidade de desenvolvimento de um DW a partir da ferramenta Pentaho como pacote de software livre e demonstrar o uso do mesmo através de um estudo de caso. Contribuindo para que as empresas adotem uma postura de trabalho mais voltada à gestão da informação e à criação de estratégias competitivas (VARGAS, 2008, p. 7).

3.4 APLICATIVOS SEMELHANTES

Tabela 1 – Ferramentas ETL e Pacotes Business Intelligence pagos

Empresa	**Ferramentas**
IBM	InfoSphere Information Server (components incluídos: InfoSphere DataStage, InfoSphere QualityStage, InfoSphere Change Data Capture, InfoSphere Federation Server, InfoSphere Foundation Tools), InfoSphere Data Event Publisher, InfoSphere Replication Server WebSphere DataStage
Microsoft	SQL Server Integration Services, BizTalk Server
ORACLE	Data Integrator, Data Service Integrator, Warehouse Builder, GoldenGate

Fonte: MEDEIROS (2015).

Tabela 2 – Ferramentas ETL e Pacotes Business Intelligence Open Source

Empresa	**Ferramentas**
KETTLE	Pentaho Comunity Data Integration Pentaho de Business Inteligence
CloverDX	CloverETL
Talend	Talend

Fonte: MEDEIROS (2015).

4 MATERIAIS E MÉTODOS

4.1 FONTE DE DADOS

Os registros foram coletados exemplares de duzentos e trinta e nove bancos de dados MySQL homogêneos, ou seja, de mesma estrutura. A soma da quantidade de GTINs encontrado em todos os bancos de dados totalizou mais de setecentos e noventa e seis mil registros. E a quantidade de NCMs autênticos somados em todos os bancos de dados totalizou mais de duzentos e vinte e nove mil registros.

4.2 FERRAMENTAS

4.2.1 FireDAC

FireDAC é uma biblioteca para Delphi e C ++ Builder de conexão a bancos de dados corporativos. Com sua poderosa arquitetura universal, o FireDAC permite acesso direto nativo de alta velocidade ao InterBase/Firebird, SQLite, MySQL/MariaDB, Microsoft SQL Server, Oracle Database, PostgreSQL, IBM DB2 Server, Sybase SQL Anywhere, Advantage DB, Microsoft Access, Informix, Teradata (ODBC), DataSnap e mais, incluindo o NoSQL Driver para MongoDB (EMBARCADERO, 2020).

O FireDAC é uma camada de acesso poderosa, mas fácil de usar, que suporta, abstrai e simplifica o acesso a dados, fornecendo todos os recursos necessários para criar aplicativos de alta carga do mundo real. O FireDAC fornece uma API comum para acessar diferentes back-ends do banco de dados, sem abrir mão do acesso a recursos específicos específicos do banco de dados e sem comprometer o desempenho. Pode ser usado nos aplicativos Android, iOS, Windows e Mac OS X desenvolvendo para PCs, tablets e smartphones (EMBARCADERO, 2020).

O FireDAC fornece uma variedade de recursos que ajudam a abstrair as diferenças entre os sistemas de banco de dados, facilitando a gravação de código que não precisa se preocupar com diferentes dialetos do DBMS ou outras diferenças sutis (EMBARCADERO, 2020). As classes TFDQuery, TFDMemTable, TFDStoredProc e TFDTable são descendentes de TDataSet com funções para trabalhar com conjuntos de dados em memória como: ordenação, formatação, filtragem, agregados, “Local SQL” e muito mais.

4.2.2 OpenETL

O OpenETL é uma ferramenta desktop que foi desenvolvida para solução deste trabalho. Possui as três etapas de um ETL em um sistema inteiramente parametrizável em cada uma delas para adequação com diversas situações. O usuário poderá criar projetos de ETL salvando-os em arquivo no formato de projeto (.etl) para que os processos sejam executados periodicamente ou para que os parâmetros possam sofrer eventuais alterações.

É totalmente dedicado ao processo de transformação de dados, em relação a maioria dos softwares existentes no mercado atualmente que contém muitas outras funcionalidades de BI. Isso permite que o produto final seja prático para solução de diversas tarefas como:

a) Automatizar os cadastramentos;
b) Aumentar a qualidade das informações;
c) Alimentar Data Warehouse ou Data Marts;
d) Preparar dados para algoritmos de Data Mining;
e) Executar transferências ou replicações entre bancos de dados distribuídos;
f) Realizar backup de dados.

A montagem do projeto ETL se faz com interação gráfica através de *drag-and-drop*, arrastando componentes desejados da paleta de componentes para o contêiner (tela de fundo) e criando ligações entre eles. Os componentes estão divididos em três grupos respectivos às fases do ETL: Extract, Transform e Load.

4.2.2.1 Componente Query (Extract)

É configurado uma lista de conexões a banco de dados e um script SQL para seleção dos dados. Como saída gera um result-set podendo ser apresentado em uma grade ao clicar em Preview. Na instrução Select é possível inserir na cláusula From uma expressão regular (Regex) para acrescentar automaticamente cláusulas Union ou Union All para cada banco de dados “Schema” encontado.

4.2.2.2 Componente Files (Extract)

É configurado uma pasta e um nome de arquivo aceitando caracteres coringa “*” para selecionar uma lista de arquivos. No caso de encontrar arquivos nos formatos JSON, XML ou CSV uni todos gerando um único result-set podendo ser apresentado em uma grade ao clicar em Preview.

4.2.2.3 Componente Filter (Transform)

Eliminar elementos indesejáveis de um result-set que não atendem as condições configuradas.

4.2.2.4 Componente Conversion (Transform)

Faz diversas tradução em registros nas colunas configuradas. Por exemplo, se o sistema de origem armazena 1 para sexo masculino e 2 para feminino, e deseja-se padronizar M para masculino e F para feminino (MACHADO, 2006).

4.2.2.5 Componente Derivation (Transform)

Insere coluna de cálculo derivada de uma expressão matemática entre as colunas. Fórmulas para produzir dados virtuais com um novo valor calculado (montante_vendas = qtde * preço_unitário, por exemplo) (MACHADO, 2006).

4.2.2.6 Componente Join (Transform)

Este componente faz a junção de dois result-sets. Como sita Machado (2006), relaciona campos provenientes de diversas fontes.

4.2.2.7 Componente Condensation (Transform)

Condensação de Dados para Machado (2006) é resumo de várias linhas, com somatório, média, entre outras funções. É nesse ponto que entra a solução central deste projeto, pois o OpenETL possui, além das funções de agregação (Group By) mais comuns (Sum, Average, Max, Min, Count), o algoritmo Most Frequent (GEEKS FOR GEEKS, 2020) que seleciona os valores mais encontrados. Para exemplificar veja a Tabela 2 abaixo:

Tabela 2 – Exemplo de informações antes de sofrer condensação

Banco de Dados	**GTIN**	**NCM**
Empresa1	7895800305065	**21069050**
Empresa2	7895800305065	17049020
Empresa3	7895800305065	Null
Empresa4	7895800305065	Null
Empresa5	7895800305065	**21069050**
Empresa6	7895800305065	Null

Na tela de edição da condensação ao marcar a coluna GTIN como chave GroupBy e a colunas NCM como MostFrequent o resultado é a Tabela 3, pois para o GTIN igual a "7895800305065" a NCM mais popular é "21069050". Note também que os campos nulos são despresados.

Tabela 3 – Resultado após Tabela 2 sofrer condensação por MostFrequent

GTIN	**NCM**
7895800305065	**21069050**

4.2.2.8 Componente Script (Load)

A partir de um ou mais result-set ligados a ele, gera um script SQL de atualização de registros.

4.2.2.9 Componente Execute (Load)

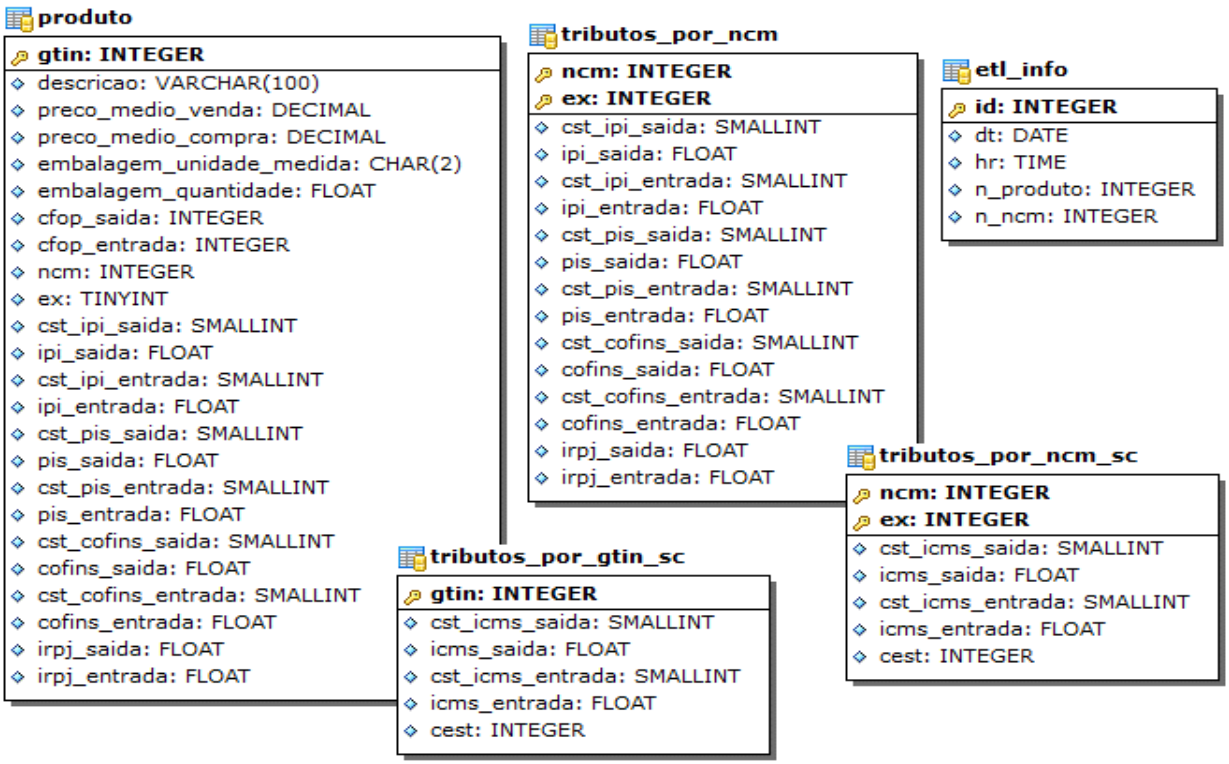
Executa os componentes Script ligado a ele. Quando ligado à um componente Files pode executar uma lista de arquivos no formato SQL.

5 METODOLOGIA

5.1 ESTRUTURA DO DATA WAREHOUSE

A função do Data Warehouse deste projeto não é movimentações financeiras ou de estoque e por isso tem apenas uma tabela Fato chamada “etl_info” que guarda as informações dos processos ETL já realizados. Tem quatro tabelas Dimensão, a “produto” contém além das informações básicas do produto os impostos federais. Foi separada da tabela “tributos_por_gtin_sc” numa relação um-para-um (GTIN são as chaves) porque a mesma possui informações referentes aos impostos estaduais do produto e sabe-se que para cada estado do Brasil podem haver diferenças, principalmente no ICMS ou em transações interestaduais. Ou seja, cada Estado cria-se uma tabela com mesma estrutura. Como demonstrado na Figura 1, neste trabalho será focado apenas no estado de Santa Catarina.

Figura 1 – Tabelas do Data Warehouse

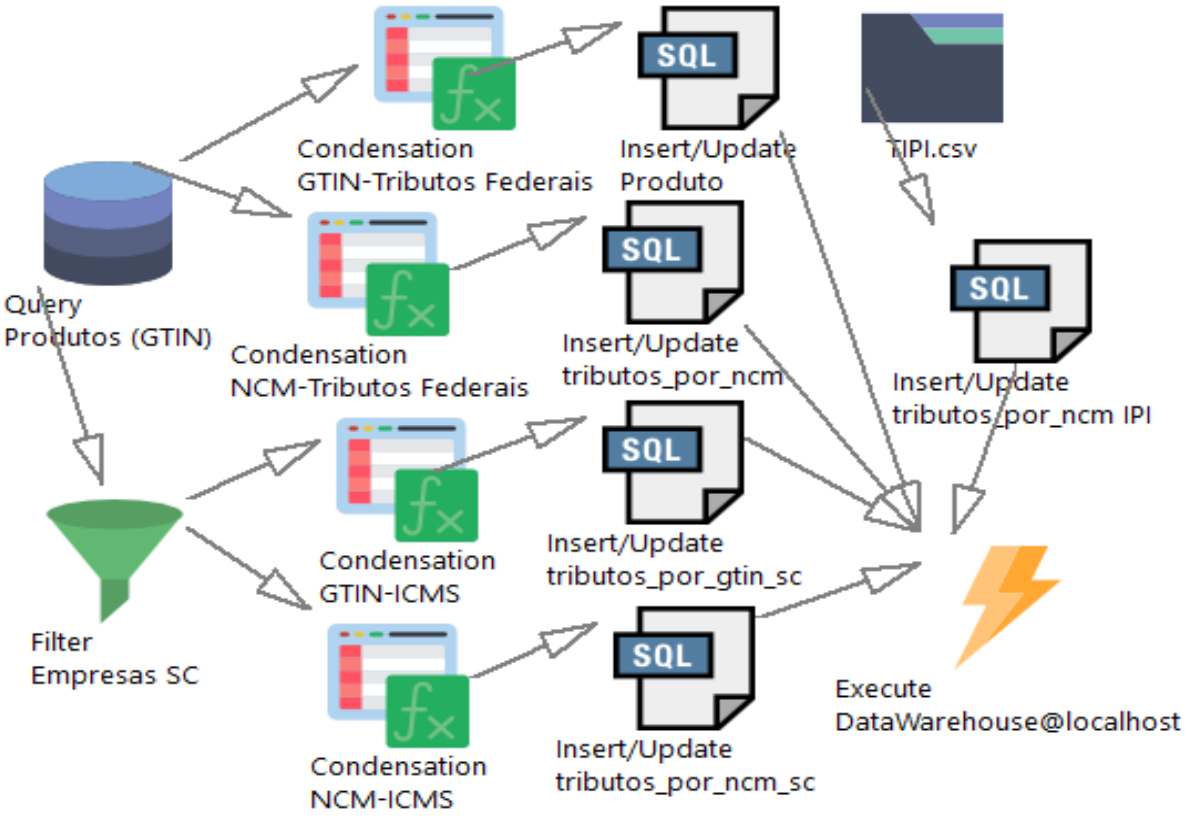


As duas últimas tabelas Dimensão chamam-se “tributos_por_ncm” e “tributos_por_ncm_sc”. Pode-se pensar como extensões da TIPI, contém as mesmas chaves primárias (NCM e EX) mas com a inclusão dos demais impostos. Poderão servir como solução paliativa um pouco menos precisa, se não existir o GTIN referente no Data Warehouse, para consultar os tributos a partir de um código NCM.

5.2 PROJETO ETL

A Figura 2 demonstra os componentes usados no projeto de ETL e o fluxo dos dados para solução deste trabalho e serão explicados com mais detalhes a seguir.

Figura 2 – Tela principal do OpenETL com os componentes do projeto



5.2.1 Extração dos Dados

Para ganho de performance nas consultas SQL na extração dos dados foi realizado inicialmente uma centralização de todos os bancos dados para a mesma máquina, possibilitando uma conexão *localhost*.

Primeiramente foi inserido ao projeto um componente Query. Fazendo uso do operador SQL “Union All” carrega-se os dados de produtos de todos os bancos de dados, formando um grande volume de dados no mesmo nível em uma única tabela bidimensional.

5.2.2 Transformação dos Dados

Para obter os impostos estaduais é necessário selecionar do Query apenas os dados das empresas catarinenses, para isso liga-se ao Query um componente Filter.

Foram vinculados dois componentes Condensation ao Query e dois ao Filter, para gerar condensação por GTIN e condensação por NCM. Na configuração dos componentes Condensation seleciona-se todas as colunas desejadas, que não são chave de agrupamento, como MostFrequent, a não ser as colunas referentes a preços de venda e de custo foi optado por selecionar como Average para gerar como resultado suas médias podendo servir como referência em cadastros de preço de venda.

5.2.3 Carregamento dos Dados

O carregamento dos dados se faz a partir arquivos de script SQL. Para a geração dos arquivos interliga-se a cada um componente Condensation um componente ScriptSQL.

Por último é possível fazer uma atualização dos dados de IPI utilizando a versão mais atual da Tabela de Incidência do IPI fornecido pela Receita Federal periodicamente através de arquivos no formato CSV. Para isso extrai-se o arquivo usando o componente Files ligado ao componente ScriptSQL para o carregamento.

6 RESULTADOS E DISCUSSÕES

No final da execução dos processos, o projeto ETL resultou dados mais concretos em aproximadamente cento e vinte mil registros de diferentes códigos de barras, com mais cinquenta por cento de preenchimento dos tributos e demais campos. São resultados satisfatórios que ainda podem ser incrementados posteriormente com uma fonte maior de dados e ajustes.

6.1 PONTOS FORTES

O objetivo deste trabalho foi atingido. A partir de centenas de bancos de dados incompletos gerou-se um Data Warehouse confiável para consumo no auxílio a cadastramento ou validação de dados de produtos e tributos.

A nova ferramenta de ETL é flexível e intuitiva na migração de dados. Mostrou-se de grande proveito para usuários administradores de bancos de dados.

6.2 PONTOS FRACOS

É necessário um computador com grande espaço de memória RAM para carregar e trabalhar com grande volume de dados. Além disso, um projeto de ETL complexo gera um alto custo computacional na execução dos processos, podendo levar horas para finalizar todas as operações.

7 CONCLUSÃO

Após um estudo mais aprofundado das normas referentes aos impostos sobre mercadorias, notou-se o quão complexo e instável são as leis vigentes no nosso país. Por isso foi modelado uma estrutura de banco de dados pronta para consultas performáticas, assim como nos processos de ETL para atender as constantes atualizações.

É fundamental, para a saúde financeira das empresas de varejo, automatizações em um sistema ERP para o rigoroso cadastro de produtos e tributos, consequentemente a eficácia nas declarações de impostos, poupando esforços de funcionários, contadores ou terceiros. Neste trabalho, através da criação de uma ferramenta ETL e o uso dos conceitos de BI, foi possível aproveitar o compartilhamento de um grande número de bases de dados transacionais para geração de informações consistentes, atendendo as empresas envolvidas.

Após várias utilizações notou-se como ferramentas de ETL podem ser útil no dia a dia para as mais diversas tarefas referente as persistências e alterações nas bases dados. Fica como ideia de implementações futuras ao OpenETL recursos na criação de API RESTFul para o consumo do Data Warehouse ou a adição de um terceiro componente de extração que faça autenticações e leitura de APIs genéricas nas nuvens.

REFERÊNCIAS

BARBIERI, Carlos. **BI – Business Intelligence**: MODELAGEM & TECNOLOGIA. RJ: Rio de Janeiro, 2001, p. 74-75.

BRACKETT, Michael H. **The Data HarehouseChallenge**: Taming Data Chaos. Wiley Computer Publishing, 1996.

CAMARGO, Renata Freitas. **Saiba tudo sobre o IPI**: Imposto sobre Produtos Industrializados. Treasy. 18 abr. 2017. Disponível em: <https://www.treasy.com.br/blog/ipi-imposto-sobre-produtos-industrializados/>. Acesso em: 12 mai. 2020.

DATE, C. J. **Introdução a Sistemas de Bancos de Dados**. 8ª Ed., Rio de Janeiro: Campus, 2004.

DA COSTA, Marcelo A. S. **O processo de ETL na construção de conhecimento em uma aplicação de uma empresa seguradora**. Juiz de Fora, MG: dez. 2009. Disponível em: <http://monografias.nrc.ice.ufjf.br/tcc-web/downloadPdf?id=76>. Acesso em: 25 set. 2018.

DINOM, COSIT. **NCM**: Saiba Quais São. Receita Federal do Brasil. 22 nov. 2019. Disponível em: <http://receita.economia.gov.br/orientacao/aduaneira/classificacao-fiscal-de-mercadorias/ncm/>. Acesso em: 15 mai. 2020.

EMBARCADERO. **FireDAC**: Multi-Device Data Access Library. 2020. Disponível em: <https://www.embarcadero.com/br/products/rad-studio/firedac>. Acesso em: 20 mai. 2020.

GEEKS FOR GEEKS. **Most frequent element in an array**. Disponível em: <https://www.geeksforgeeks.org/frequent-element-array/>. Acesso em: 01 mai. 2020.

GITHUB. Open ETL. Disponível em: <https://github.com/paulo-oliv/open-etl.git/>. Acesso em: 14 julho. 2020.

GS1 BRASIL. **O que é o GTIN?** Disponível em: <https://www.taxweb.com.br/impostos-incidentes-sob-nfs/>. Acesso em: 12 mai. 2020.

GS1 BRASIL. **CNP:** Cadastro Nacional de Produtos. Disponível em: <https://www.gs1br.org/servicos-e-solucoes/Paginas/CNP---Cadastro-Nacional-de-Produtos.aspx/>. Acesso em: 12 mai. 2020.

GTIN INFO. **GTIN Definition**: Information**.** Disponível em: <https://www.gtin.info/>. Acesso em: 12 mai. 2020.

INMON, W. H. TERDEMAN, R. H. IMHOFF, Claudia. **Data Warehousing**: Como transformar informações em oportunidades de negócio. São Paulo: Berkeley, 2001.

INMOH, W. H. **Como Construir o Data Warehouse**. Tradução da segunda edição. Rio de Janeiro: Ed. Campus, 1997, p. 269-275.

JUNQUEIRA, Gabriel. **A importância da classificação tributária no Cadastro de Produtos**. Info Varejo. 07 nov. 2016. Disponível em:

<https://confeb.liveuniversity.com/2017/02/21/como-manter-se-atualizado-sobre-mudancas-nas-leis-que-regem-seu-negocio/>. Acesso em: 30 abr. 2020.

JUNQUEIRA, Gabriel. **Validação tributária dos produtos no supermercado**: O que é e quais as vantagens? Info Varejo. 07 fev. 2017. Disponível em: <https://confeb.liveuniversity.com/2017/02/21/como-manter-se-atualizado-sobre-mudancas-nas-leis-que-regem-seu-negocio/>. Acesso em: 30 abr. 2020.

KIMBALL, R. CASERTA, J. **The Data Warehouse Toolkit**: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley Publishing, Inc. 2004.

MACHADO, Felipe Nery Rodrigues. **Tecnologia e Projeto de Data Warehouse**. 6ª Ed. Érica, 2014, p. 303-312.

MACHADO, Felipe, ABREU, Mauricio. **Projeto de Banco de Dados**: Uma Visão Prática. 16ª Ed. Érica, 2011.

MEDEIROS, Higor. **Business Intelligence**: Conhecendo algumas ferramentas Open Source. 2015. Disponível em: <https://www.devmedia.com.br/business-intelligence-conhecendo-algumas-ferramentas-open-source/31963>

MUNIZ, Vander Emiro. **ETL**: Quais as ferramentas mais poderosas do mercado? DevMedia. 23 fev. 2018. Disponível em: <https://www.devmedia.com.br/etl-quais-as-ferramentas-mais-poderosas-do-mercado/6727>. Acesso em: 29 ago. 2018.

NETO, Trajano C. M. **Avaliação das Ferramentas ETL open-source Talend e Kettle para Projetos de Data Warehouse em Empresas de Pequeno Porte**. Lauro de Freitas, BA: 2012. Disponível em: <http://www.ambientelivre.com.br/downloads/doc_download/87-tcc-ferramentas-deetl-open-source-talend-e-kettle.html>. Acesso em: 30 set. 2018.

NASCIMENTO, Gabriel. **CFOP**: o que é, como aplicar e onde encontrar a tabela. Disponível em: <https://enotas.com.br/blog/cfop/>. Acesso em: 27 set. 2020.

NOVAIS, Ramon R. C. **Modelagem Dimensional**. São Paulo: 2012. Disponível em: <http://www.fatecsp.br/dti/tcc/tcc00071.pdf>. Acesso em: 25 set. 2018.

ÖZSU, M. Tamer. VALDURIEZ, Patrick. **Princípios de Sistemas de Banco de Dados Distribuídos**. Tradução da 2ª Edição Americana. Rio de Janeiro: Ed. Campus, 2001.

PENTAHO, Kettle. Project. Disponivel em: <http://community.pentaho.com/projects/data-integration/>. Acesso em: 29 set. 2018.

PORTAL TRIBUTÁRIO. **IPI**: Imposto sobre produtos industrializados. Disponível em: <http://www.portaltributario.com.br/tributos/ipi.html/>. Acesso em: 12 mai. 2020.

PRIETO, Beatriz. **Como manter-se atualizado sobre mudanças nas leis que regem seu negócio**. Live University. 21 fev. 2017. Disponível em: < https://www.infovarejo.com.br/classificacao-tributaria-no-cadastro-de-produtos/>. Acesso em: 10 mai. 2020.

SIGE CLOUD, Ajuda. **CEST**: Código Especificador da Substituição Tributária.

Disponível em: <https://suporte.sigecloud.com.br/hc/pt-br/articles/223083267-CEST-C%C3%B3digo-Especificador-da-Substitui%C3%A7%C3%A3o-Tribut%C3%A1ria>. Acesso em: 27 mai. 2020.

TAXWEB. **Impostos Incidentes Sob a NFS-e**: Saiba Quais São. 07 jul. 2019. Disponível em: <https://www.taxweb.com.br/impostos-incidentes-sob-nfs/>. Acesso em: 09 mai. 2020.

TANAKA, Asterio. **Tópicos Avançados de Banco de Dados (Business Intelligence)**: Integração de Dados e ETL. Disponível em: <http://www.uniriotec.br/~tanaka/SAIN/03-ETL-2015.1.pdf>. Acesso em: 31 out. 2016.

VARGAS, Marcelo F. **Construção De Data Warehouse para Pequenas e Médias Empresas Usando Software Livre**. Lages, SC: 2009. Disponível em: <https://revista.uniplac.net/ojs/index.php/tc_si/article/download/826/536>. Acesso em: 25 set. 2018.

ZORZO, André Luis. **ETL 2.0**: Uma proposta de extensão ao processo de extração, transformação e carga voltada à integração de dados estruturados e não estruturados. Florianópolis: 2009. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/184522/Projeto_Andre_Zorzo.pdf?sequence=-1>. Acesso em: 25 set. 2018.