TEMA: Integração de dados no banco de dados estatístico de Goiás (BDE-GO) por meio dos processos de extração, transformação e carga (ETL)

Integração de dados do banco de dados estatístico de Goiás (BDE-GO) por meio dos processos de extração, transformação e carga (ETL)

Bernard Silva de Oliveira[1]

1. INTRODUÇÃO

O Banco de Dados Estatístico de Goiás (BDE-GO) é um sistema de informações estatísticas sobre o estado de Goiás e todos os seus municípios. As informações contidas são de diversas áreas como: econômica, física, social, financeira, política e administrativa, divididas em Regiões de Planejamento, Regiões do IBGE como: Mesorregião, Microrregião, Intermediárias e Imediata (IMB,2020).

O BDE Goiás é resultado do sistema IMP desenvolvido pela Fundação Seade do estado de São Paulo, sendo doada à Associação de Nacional das Instituições de Planejamento, Pesquisa e Estatística(Anipes), que permitiu o uso pela Secretaria de Estado da Administração de Goiás (SEAD-GO), antiga SEGPLAN, sob administração do Instituto Mauro Borges de Estatísticas e Estudos Socioeconômicos. As informações contidas no BDE-GO são agrupadas em 19 temas que derivam a 41 assuntos, compostos por mais de 800 variáveis de diferentes fontes como: IBGE, SANEAGO, TRE-GO, DETRANS- GO e entre outros (IMB, 2020).

O modelo lógico do BDE se baseia no modelo multidimensional, bastante utilizando em *DataWarehouse,* que são bancos de dados orientados a assunto, não voláteis e variável ao tempo. O *Data Warehouse* tem suas tabelas divididas em: dimensionais, que representam características do assuntoe valores que no qual se pretenda filtrar; e fato, com atributos mensuráveis dos assuntos(geralmente são dados numéricos). Além disso, o BDE tem uma característica de alta granularidade, que segundo Machado (2011), refere-se ao nível de detalhamento da informação no banco de dados. Conforme Turban et al. (2009), a inteligência de negócios (*Business Intelligence*) é um termo amplo que envolve arquitetura, ferramentas para análise (OLAP, *dashboards* etc.), banco de dados de aplicações e procedimentos metodológicos. Os principais objetivos da inteligência de negócios são interações com dados externos e/ou institucional com longo período (série histórica), proporcionando a manipulação desses, a fim de fornecer aos gestores, a capacidade de realizar a análise ou a tomada de decisão mais adequada (TURBAN et al., 2009).

A integração de dados é uma parte importante para aplicações de inteligência de negócio, pois esta vai além de uma simples ideia, envolvendo a coleta das referências (diferentes fontes) e os seus armazenamentos em um repositório ou container de dados. Algumas atividades importantes na integração de dados: identificar os sistemas de origens; compreender sistemas de origem; carregar dimensão conformada (dados) e outras atividades (Kimball e Caserta ,2009).

[1] Gerente de Dados e Estatísticas e Pesquisador do IMB. Mestre em Geografia (UFG). E-mail: bernard.oliveira@goias.gov.br.

TEMA: Integração de dados no banco de dados estatístico de Goiás (BDE-GO) por meio dos processos de extração, transformação e carga (ETL)

O ETL é um processo de extrema importância para integração, devido ao seu objetivo de carregar dados de fontes e estruturas distintas, sendo ele um banco de dados, planilhas eletrônicas e outras fontes para um banco de dados destino e/ou um *Data Warehouse* (Turban et al.,2009). A sigla ETL tem o significado, em português, de Extração, Transformação e Carga (*Load*).

Segundo Pall e Khaira (2013), as ferramentas do processo ETL são classificadas em: ETL codificadas manualmente, que são desenvolvidas por meio de linguagens de programações; e as ETL baseadas em ferramentas, em que o banco de dados e/ou aplicações de BI já possuem embutidostais processos. O Banco de Dados Estatístico de Goiás (BDE-GO) possui diferentes tipos de fontes de dados, este informe foi elaborado com o objetivo de mostrar aplicação do processo de ETL na atualização e manutenção e atualização de algumas variáveis no BDE - Goiás.

1.1 PROCESSO DE ETL (EXTRAÇÃO, TRANSFORMAÇÃO E CARGA)

1.2 - Extração de dados

Conforme Elias (2014), a extração é uma etapa na qual os dados serão adquiridos de diversas fontes e organizações distintas, e depois inseridas em um local que serão trabalhados para transformá-los em um único formato. A conversão de único formato é feita devido à diversificação das informações no banco de dados estatístico de Goiás.

A fonte de dados mais utilizado no BDE é do IBGE, sendo sua maioria neste banco de dados. O IBGE tem um sistema de recuperação automática de dados, conhecido como SIDRA, que reúne dadosde várias áreas do conhecimento. O SIDRA possui um API para que usuários que domine linguagens de programação possam extrair informações automáticas.

Para extrair as informações, primeiramente foi construído um mapa lógico dos campos da tabela de origem do Banco de Dados Estatístico de Goiás. O mapa lógico é um documento da base de metadados que são utilizados para garantir a qualidade e descrição exata do sistema de origem (Kimball e Caserta,2009). Os principais elementos para confecção do mapa lógico são: nome databela destino; nome da coluna de destino; tipo da tabela (fato ou dimensão) que, no caso deste informe, é a tabela fato; nome da tabela de origem. A Tabela 01 segue um exemplo do mapa lógico da variável **“Produção Agrícola - Cana-de-açúcar - Quantidade Produzida (t)”**.

TEMA: Integração de dados no banco de dados estatístico de Goiás (BDE-GO) por meio dos processos de extração, transformação e carga (ETL)

Tabela 01 - Mapa de dados lógicos para o BDE-GO

Dados destino				Dados fonte		
Nome Tabela	**Nome Coluna**	**Tipo de dado**	**Nome Banco**	**Fonte dados**	**Nome Tabela**	**Nome Coluna**
localidade	loc_nome	texto	MySQL	SIDRA-IBGE	1612	Município
variavel	var_cod	número	MySQL	SIDRA-IBGE	1612	D2N
variavel	d_ano	número	MySQL	SIDRA-IBGE	1612	Valor

Elaborado pelo próprio autor.

Após a extração dos dados do IBGE, estes são armazenados em um contêiner de base de dados temporário (*Data Store*), que serve de fundamentos para todo o processo de transformação antes da etapa de carga no banco de dados e/ou *Data WareHouse*.

1.3 - Transformação

Esta etapa nem sempre é obrigatória, pois dependendo da fonte de dados, a conversão e limpeza já são efetuadas pela fonte da base de dados original. Muitas das vezes, os dados apresentam grandes dificuldades em sua manipulação por exemplo: devido heterogeneidade dos formatos, nem sempre os campos são preenchidos e alguns até são duplicados. Na etapa de limpeza de dados são executadas as seguintes ações (Han e Kamber, 2006):

- Preenchimento dos valores ausentes;
- Suavização dos dados, eliminando dados anômalos (valores fora do padrão);
- Correção de inconsistência (duplicidade, erro de preenchimento).

No contexto deste trabalho, não houve necessidade de preencher valores ausentes devido àsinformações estarem completas que, é uma das características citadas por Kimball e Caserta (2009). Mas houve a correção de inconsistências como: a exclusão de informações de outro estado, neste caso os municípios do estado do Tocantins. Além disso, realizou-se a limpeza dos valores das colunas que caracterizam o nome do munícipio, retirados as unidades de federações.

TEMA: Integração de dados no banco de dados estatístico de Goiás (BDE-GO) por meio dos processos de extração, transformação e carga (ETL)

Na parte de conformidade de informação, verifica-se os conflitos de valores entre o nomedas variáveis do BDE-GO com o IBGE. Para resolver esse problema, são normalizados o nomes das variáveis do IBGE para o nome que está contido no BDE-GO, pois esse campo será utilizado para junção do dado que foi adquirido pelo IBGE com os dados do BDE-GO.

A última etapa foi juntar os dados do IBGE processados com os dados do BDE-GO por meio dos campos do Nome do Município e Nome da variável. A estrutura dos dados ficou da seguinte forma:

- ***loc_cod:*** código da localidade (município);
- ***cod_var:*** código da variável;
- ***d_ano:*** valor do ano da variável.

1.4 - Carga (*Load*)

O último processo do ETL, a carga, em inglês *Load*, resume-se à persistência dos dados na base consolidada (Elias, 2020). Neste trabalho, a carga foi preparada por um programa desenvolvido pela equipe de TI, da atual Secretaria de Estado de Administração do estado de Goiás (SEAD-GO), sendo que, o único parâmetro de entrada é um arquivo no formato *.csv*, na estrutura citada anteriormente do final do processo de transformação.

2. RESULTADOS

Com a implementação do ETL do tipo manual (desenvolvido por linguagem de programação), na atualização de dados no BDE-GO, obteve-se um ganho de produtividade de aproximadamente27% na atualização de variáveis com periocidade anual, sendo elas: Produção Agrícola Municipal (PAM) e a Produção Pecuária Municipal (PPM). Além disso, o ETL foi programado para executar quando tais pesquisas (PAM e PPM/IBGE) fossem divulgadas oficialmente.

3. CONSIDERAÇÕES FINAIS

A aplicação da técnica de ETL é fundamental na implementação de *Data Warehouse* e integração de dado*,* nesse caso, o Banco de Dados Estatístico de Goiás (BDE-GO)*,* pois este garante integridade nas informações como: ausência de duplicidade, sem valores anómalos (ruídos), valores ausentes, para ferramentas de análises como OLAP, *Dashboards*, Mineração de dados e outros, preparados para aplicação de inteligência de negócios.

TEMA: Integração de dados no banco de dados estatístico de Goiás (BDE-GO) por meio dos processos de extração, transformação e carga (ETL)

4. REFERÊNCIAS

ELIAS, D. **Entendendo o processo de ETL**. Disponível em: https://canaltech.com.br/business-intelligence/entendendo-o-processo-de-etl-22850/. Acesso em: 24 jul. 2020.

HAN, J.; KAMBER, M. **Data Mining: concepts and techniques**. Waltham: Elsevier, 2006.

IMB – INSTITUTO MAURO BORGES DE ESTATÍSTICAS E ESTUDOS SOCIOECONÔMICOS. **BDE-Goiás - Banco de Dados Estatísticos de Goiás**. Goiânia: IMB, 2018. Disponível em: https://www.imb.go.gov.br/bde/, Acesso em: junho 2020.

KIMBALL, R.; CASERTA, J. **The Data Warehouse ETL Toolkit: practical techniques for extracting, cleaning, conforming, and data delivering data**. Indianopolis: Wiley Publishing, 2009.

MACHADO, F. N. **Tecnologia e Projeto de Data Warehouse: uma visão multidimensional.** SãoPaulo: Érica, 2011.

PALL, A. S.; KHAIRA, J. S. **A comparative review of Extraction, Transformation and LoadingTools**. Database Systems Journal, Jalandhar, v. IV, n. 2, p. 42-51, 14 fev. 2013.

TURBAN, E. et al. **Business Intelligence: um enfoque gerencial para a inteligência do negócio**.Porto Alegre: Bookman, 2009.