Descrição da montagem do ambiente de BI

Nessa seção serão descritos todos os passos, técnicas e dados utilizados para a aplicação da metodologia de Business Intelligence no contexto do presente trabalho.

Introdução

Segundo Braghittoni (p.1, 2017): “O BI é um conjunto de conceitos e métodos para melhorar o processo de tomada de decisão, utilizando-se de sistemas fundamentados em fatos e dimensões”. Nesse caso podemos perceber que o BI é uma metodologia, que possui regras, ordem e práticas para sua aplicação. Sendo assim, precisamos descrever cada uma das partes que vão compor nosso ambiente de inteligência, utilizando como base os autores Braghittoni (2017), Carvalhaes e Alves (2015), Inmon (2005) e Kimball e Ross (2013).

Esse ambiente se divide em Fontes de Dados (*Data Source*), Área de Staging (*Staging Area*), *Data Warehouse* e *Data Marts*. Suas definições serão explicadas a frente.



Figura x – Arquitetura do ambiente de BI.

Fonte: Panoly.

Montagem do ambiente, parte 1 – Fontes de Dados.

O primeiro passo na aplicação dos processos de Business Intelligence é definir quais serão as bases de dados utilizadas para o processo e quais dados serão extraídos delas. No caso do presente trabalho, foram utilizadas as bases de micro dados do censo escolar do INEP, disponíveis no Portal Brasileiro de Dados Abertos (<http://dados.gov.br/dataset/microdados-do-censo-escolar>) e no próprio site do INEP (<http://inep.gov.br/web/guest/microdados>). Para a melhor delimitação do trabalho, foram utilizados os censos dos anos de 2015 a 2018.

Os arquivos estão em formato CSV (*Comma-separated Values*) que é um tipo de arquivo onde seus dados estão separados por algum delimitador, no caso das bases do INEP é utilizado o delimitador *Pipe* (|). Eles são divididos em Turmas, Escolas, Matriculas (Centro-Oeste, Nordeste, Norte, Sudeste e Sul), e Docentes (Centro-Oeste, Nordeste, Norte, Sudeste e Sul), onde temos as informações das turmas, das escolas, dos alunos e dos docentes envolvidos nos censos de cada ano, respectivamente.

Além dos dados principais, precisamos de tabelas auxiliares que vão auxiliar-nos na definição dos dados do INEP, já que são utilizados campos com os códigos dos Países, Unidades da Federação (UF), Municípios, Distritos, Mesorregiões e Microrregiões. Para o primeiro, o INEP disponibiliza em sua base, ao fazer download, uma tabela que contêm os códigos dos países descritos no censo, já que alunos estrangeiros também são envolvidos no censo escolar.



Figura x – Tabela de códigos dos países.

Fonte: Autores.

Para as UF, Municípios, Distritos, Mesorregiões e Microrregiões, foram utilizadas as bases de códigos do IBGE de Divisão Territorial Brasileira (DTB). Para o presente trabalho, foi escolhida a última disponibilizada do ano de 2018 (<ftp://geoftp.ibge.gov.br/organizacao_do_territorio/estrutura_territorial/divisao_territorial/2018/>). Diferente das bases do INEP, esses arquivos estão em formato XLSX (*Excel Microsoft Office Open XML Format Spreadsheet File*) que é o formato padrão em planilhas do programa Excel da *Microsoft*.



Figura x – Tabela DTB de Município.

Fonte: Autores.



Figura x – Tabela DTB de Distrito.

Fonte: Autores.

Assim finaliza-se a parte da aquisição de dados.

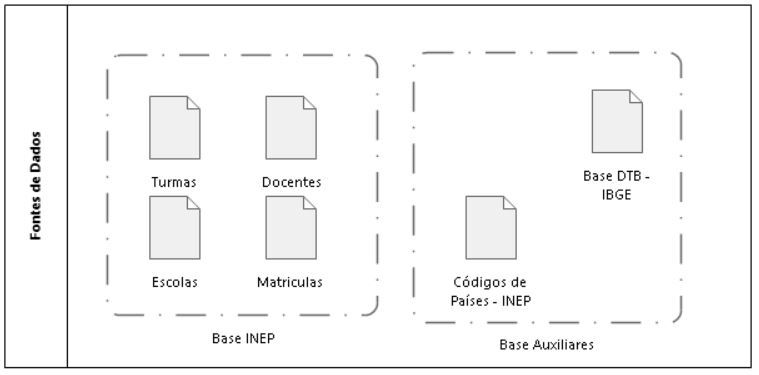


Figura x – Modelo de Fonte de Dados.

Fonte: Autores.

Montagem do ambiente, parte 2 – Área de *Staging*

Segundo um dos postulados de Inmon (p. 29, 2005) sobre *Data Warehouse*, ele define a não volatilidade, ou seja, os dados dentro do mesmo não podem sofrer alterações. Isso significa que vamos precisar de uma fase intermediária antes de carregar os dados no DW, para isso temos a *Staging Area* ou *Data Stage*. Com todos os dados já na máquina iniciaremos a montagem dos processos de ETL para fazer a carga no Banco de Dados de *Staging*.

Utilizaremos o *Pentaho Data Integrator* (PDI) para iniciar os processos de ETL, separando as cargas por assunto. O PDI utiliza duas nomenclaturas como *Job* e *Transformation*, o primeiro é a menor ação possível que o programa possa fazer como ler o arquivo ou fazer inserção, e o segundo é um conjunto de outros *Jobs* para fazer uma execução única e contínua.



Figura x – Exemplo de *Transformation* e *Job*.

Fonte: Pentaho.

A carga dos arquivos no BD dos arquivos principais (turmas, matrícula, escolas, docentes) é composta por três passos, em que o PDI encontra os arquivos, prepara-os para a inserção e grava-os no BD.

Figura x – Visão da ETL das bases principais.



Fonte: Autores.

Os passos são descritos abaixo:

*Get* *File Names*: Esse *step* procura nomes de arquivos ou pastas. Ele é recomendado para quando se tem uma grande massa de dados em que todos precisam ser gravados. Os padrões dos nomes são adquiridos conforme uma expressão regular.

*Text File Input*: Aqui o Pentaho prepara um ou mais arquivos de textos para a inserção, nele são configuradas diversas opções como os delimitadores do texto, linha de título, formato e colunas adicionais para serem adicionadas no momento da carga.

*Table Output*: Realiza a carga dos dados estruturados em uma tabela no banco de dados. A tabela não precisa ser criada com antecedência, já que o PDI prepara um comando SQL com os dados que serão inseridos.

Já para a carga das tabelas auxiliares (código dos países e base dtb), foi utilizado um padrão de carga diferente, já que os arquivos que possuem os dados estão em um formato diferente das outras bases.



Figura x – Visão geral da ETL de auxiliares.

Fonte: Autores.

Os passos são descritos abaixo:

*Microsoft Excel Input*: Esse *step* procura nomes de arquivos do tipo XLS (formato utilizado nas versões de 97 até 2003) e/ou XLSX (utilizado na versão de 2007 em diante). Nele pode-se configurar opções como, especificar de qual linha e/ou coluna deve-se iniciar a análise, se os títulos das colunas estão na primeira linha (*Header*), além de especificar campos adicionais no momento da carga.

*Table Output*: Realiza a carga dos dados estruturados em uma tabela no banco de dados. A tabela não precisa ser criada com antecedência, já que o PDI prepara um comando SQL com os dados que serão inseridos.



Figura x – Visão geral da ETL *Staging*.

Fonte: Autores.

Com todo o fluxo executado, o Banco de Dados *Staging* foi finalizado da seguinte forma:

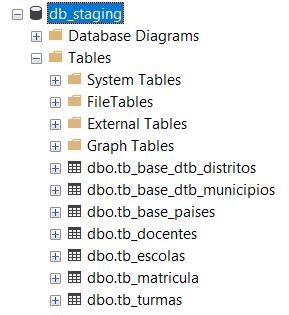


Figura x – Visão do Banco *Staging*.

Fonte: Autores.

Assim encerra-se a parte de carga do Banco de Dados *Staging*.