**4 DESCRIÇÃO DA MONTAGEM DO AMBIENTE**

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

4.1 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 pode-se perceber que o BI é uma metodologia, que possui regras, ordem e práticas para sua aplicação. Sendo assim, é necessário descrever cada uma das partes que vão compor o ambiente de inteligência, utilizando como base os autores Braghittoni (2017), Carvalhaes e Alves (2015), Inmon (2005) e Kimball e Ross (2013).

Esse ambiente divide-se 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.

4.2 Montagem do ambiente, parte 1 – Fontes de Dados (*Data Source*).

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 no link: <http://dados.gov.br/dataset/microdados-do-censo-escolar> e no próprio site do INEP no link: <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 se encontra as informações das turmas, das escolas, dos alunos e dos docentes envolvidos nos censos de cada ano, respectivamente.

Além dos dados principais, faz-se necessário o uso de tabelas auxiliares para auxiliar 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: Adaptado de INEP (2019).

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 base disponibilizada do ano de 2018 no link: <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: Adaptado de IBGE (2018).



Figura x – Tabela DTB de Distrito. Fonte: Adaptado de IBGE (2018).

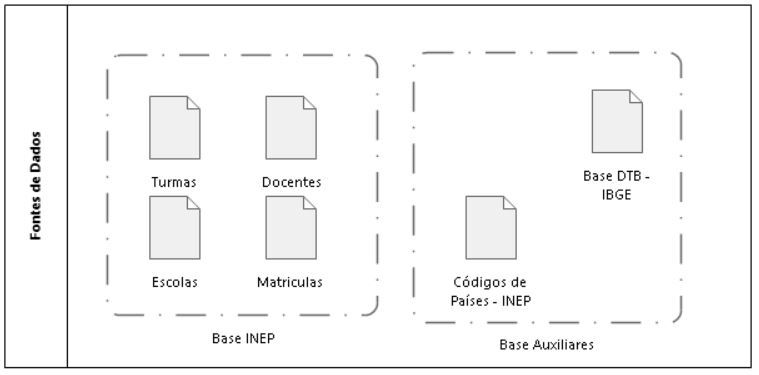


Figura x – Modelo de Fonte de Dados. Fonte: Autores (2019).

4.3 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 se faz necessária uma fase intermediária antes de carregar os dados no DW, para isso tem-se a *Staging Area* ou *Data Stage*. Com todos os dados já na máquina é iniciada a montagem dos processos de ETL para fazer a carga no Banco de Dados de *Staging*.

Será utilizado o *Pentaho Data Integrator* (PDI) versão 5.0.1 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 (2019).



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 esses dados estão em um formato diferente das outras bases.



Figura x – Visão geral da ETL de auxiliares. Fonte: Autores (2019).

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 podem-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 (2019).

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

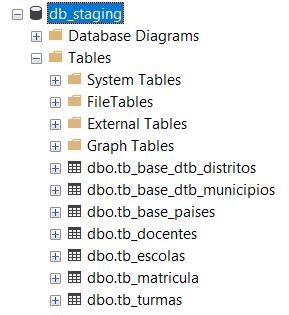


Figura x – Visão do Banco *Staging*. Fonte: Autores (2019).

4.4 Montagem do ambiente, parte 3 – O *Data Warehouse*

Após o banco *Staging* estar completamente carregado, será iniciado os processos para a formação do armazém de dados.

4.4.1 Abordagem Inmon x Kimball

Antes de começar o desenvolvimento das ETLs, deve-se pensar como será a estrutura e modelo do *Data Warehouse*, tendo como base a abordagem Inmon ou Kimball. Vale ressaltar que não há uma escolha certa ou errada, mas aquela atende melhor os requisitos e necessidades da organização.

Inmon utiliza a abordagem *top-down* em que o DW é um repositório de dados centralizado, sendo assim o componente mais importante da organização (PANOLY, 2019). Ele é o primeiro modelo criado logo após a extração de dados, e após sua finalização são criados todos os *Data Marts* necessários.



Figura x – Modelo Inmon. Fonte: Panoly.

Em contrapartida, Kimball utiliza a abordagem *bottom-up* em que são criados primeiro os *Data Marts* em cada área de interesse para depois se criar um grande *Data Warehouse* que é unicamente uma junção de todos esses *Marts* (PANOLY, 2019). Tal como Kimball (1997) afirma: “O *Data Warehouse* não é nada mais do que uma junção de diversos *Data Marts*”.



Figura x – Modelo Kimball. Fonte: Panoly.

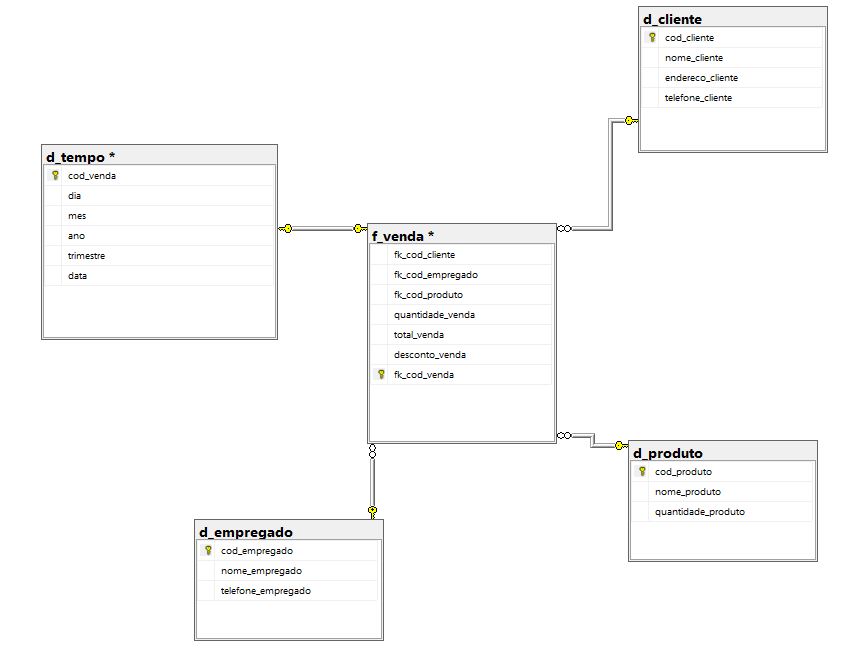
Para a realização desse trabalho foi escolhida a abordagem Inmon porque atende melhor os requisitos desse projeto, e também porque não será usado os *Data Marts*, assim sendo, será criado unicamente o *Data Warehouse* para armazenar os dados.

4.4.2 Modelos Estrela e Floco de Neve (*Star Schema and* *Snow-Flake* *Schema*)

Tendo definida a estrutura, inicia-se o desenvolvimento do modelo do DW. Para isso, possuímos duas maneiras de fazê-lo, utilizando o modelo do tipo Estrela (*Star Schema*) ou o modelo Floco de Neve (*Snow-Flake* *Schema*).

O modelo Estrela é o mais básico para a arquitetura do *Data Warehouse*. No seu desenho, a tabela fato assume o centro da arquitetura seguido pelas tabelas de dimensões, que em volta dela, definem a quantidade de pontas da Estrela (CARVALHAES e ALVES, 2015). Possui como vantagem uma visualização simplificada dos dados, além de mais agilidade nas análises.

Figura x – Exemplo de modelo Estrela. Fonte: Autores (2019).



O modelo Floco de Neve é um modelo específico que, partindo do modelo Estrela, as dimensões que possuem hierarquia são decompostas em outras tabelas (CARVALHAES e ALVES, 2015). Nesse modelo temos uma redução de redundância nas tabelas de dimensões e uma menor quantidade de memória utilizada. Podemos usar de exemplo uma dimensão chamada Data, em que ela poderá ser decomposta em outras tabelas como dia, mês, ano, trimestre, etc. Assim, essas “sub-dimensões” vão compor a dimensão principal.

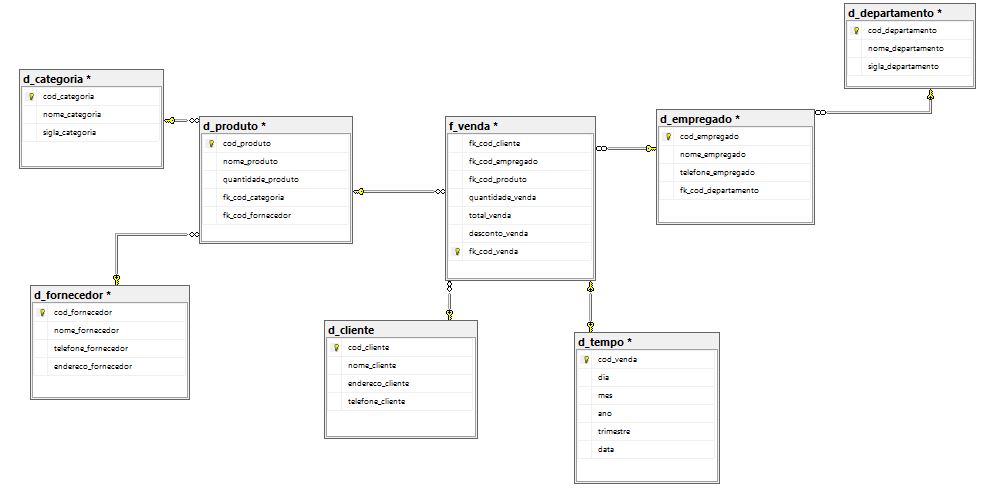


Figura x – Exemplo de modelo Floco de Neve. Fonte: Autores (2019).

Para o presente trabalho, será utilizado o modelo Floco de Neve, devido algumas dimensões apresentarem hierarquia nelas, além deste modelo se propor a ser mais didático, já que apresenta de forma detalhada todas as dimensões que estão em uso pela operação de BI.

4.4.3 Indicadores levantados para as análises

Ao desenvolver uma plataforma de BI, o objetivo é sempre responder a perguntas utilizando dados, que por sua vez se transformam em informação e auxílio na tomada de decisão. Para isso é necessário levantar perguntas que serão os indicadores da análise, com isso, teremos nossos fatos e dimensões. As perguntas levantadas pelo grupo são as seguintes:

1. Qual o total de alunos por cada Cor/Raça definida pelo Censo Escolar entre os anos da análise?
2. Qual o total de alunos que se declararam negros entre os anos da análise?
3. Qual o total de alunos estrangeiros que se declararam negros entre os anos da análise?
4. Qual o país que possui a maior quantidade de alunos estrangeiros negros no Brasil entre os anos da análise?
5. Qual é a diferença de declaração negra entre os anos da análise?
6. Qual o total de alunos negros por região, UF e município entre os anos da análise?
7. Qual é a diferença de alunos negros entre as regiões Nordeste e Sudeste nos anos da análise?
8. Qual é a quantidade de alunos negros no Distrito Federal entre os anos da análise?
9. Qual a quantidade de alunos negros que estudam em escolas sem água, energia, esgoto e alimentação entre os anos da análise?
10. Qual a quantidade de alunos negros por sexo entre os anos da análise?
11. Qual a quantidade de alunos negros nos módulos de ensino Presencial, Semipresencial e a Distância entre os anos da análise?
12. Qual a quantidade de alunos negros que moram em zona Urbana ou Rural entre os anos da análise?
13. Qual a quantidade de alunos negros que estudam em escolas Públicas e Privadas?
14. Qual a quantidade de alunos negros que estudam em escolas Urbanas e Rurais?
15. Qual a quantidade de alunos negros em cada etapa de ensino definida no censo entre os anos da análise?

Com as perguntas concluídas, pode-se agora levantar os fatos e dimensões da análise. Será utilizada apenas uma tabela fato, que é a tabela de matrículas (alunos) e as seguintes dimensões: Tempo (Ano), Cor/Raça, Nacionalidade, Mediação Didática Pedagógica, Etapa de ensino, Sexo, Localidade (Microrregião, Município, UF, Região, País, Mesorregião, Distrito) e Escola (Dependência, Localização, Categoria).

Com a fato e as dimensões já definidas, será criada as ETLs para a carga das informações no *Data Warehouse*.

4.4.4 Processo ETL para carga do *Data Warehouse*

Nessa parte de explicação das ETLs, separaremos por dimensões que possuem padrões de carga semelhantes, explicando os dados envolvidos e o processo.

4.4.4.1 Dimensões Categoria Escola Privada, Cor/Raça, Dependência, Etapa de Ensino, Localização, Mediação Didática, Nacionalidade, Sexo, Zona Residencial e Ano

Para a carga dessas dimensões, dada a sua quantidade pequena de informações, foi montado um conjunto de dados para a sua inserção.

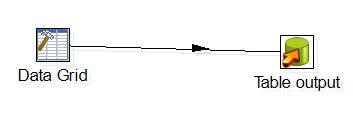


Figura x – Visão geral das primeiras ETLs Grid para o DW. Fonte: Autores (2019).

Os seguintes passos foram utilizados:

*Data Grid*: Neste *step* podemos criar uma tabela com um conjunto constante de dados, informando os nomes dos campos, seus tipos, e seus respectivos dados, conforme será mostrado para cada uma das dimensões.

*Table Output*: Conforme explicado na parte do *Staging*,esse passo 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.

Dados e definições de cada uma das dimensões:

Ano: Tempo é a dimensão que todo ambiente de BI precisa ter, no caso do presente trabalho, a nossa dimensão de tempo é formada pelos respectivos anos das análises do Inep. Além disso, foi adicionado um código para cada um dos dados, e um indicador de ano nulo, para caso o conjunto de dados não tiver essa informação.



Figura x – Dados inseridos na dimensão Ano. Fonte: Autores (2019).

Categoria de Escola Privada: Uma escola que é privada pode ser de quatros tipos distintos. Além do código que fora atribuído para cada e o seu indicador nulo, foi adicionado um indicador para quando a escola for pública, já que se uma escola tiver essa nomenclatura, ela não terá um tipo indicador de escola privada, ou seja, “não se aplica”.



Figura x – Dados inseridos na dimensão Categoria Escola Privada. Fonte: Autores (2019).

Cor/Raça: Cada aluno, nos questionários do Inep, se declara com um tipo de Cor/Raça, como também pode preferir pela não declaração. O Inep coloca cinco escolhas para os alunos, além de uma para a não declaração. Como nas outras dimensões, foi inserido também um indicador de Cor/Raça nula, para caso o conjunto da análise não tiver essa informação.



Figura x – Dados inseridos na dimensão Cor/Raça. Fonte: Autores (2019).

Dependência: Cada escola possui um tipo de dependência, indicando a qual esfera ela está associada, ou se é uma escola privada.



Figura x – Dados inseridos na dimensão Dependência. Fonte: Autores (2019).

Etapa de Ensino: Um aluno está associado a uma etapa de ensino, que é a série que ele estava cursando no momento do censo.



Figura x – Dados inseridos na dimensão Etapa Ensino. Fonte: Autores (2019).

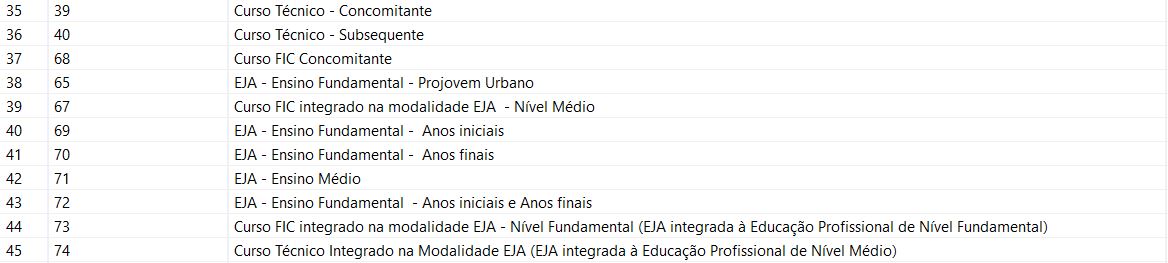


Figura x – Dados inseridos na dimensão Etapa Ensino (cont.). Fonte: Autores (2019).

Localização/Zona Residencial: Uma escola e um aluno podem estar localizados em dois tipos de zonas, Urbana e Rural.



Figura x – Dados inseridos na dimensão Localização e Zona Residencial. Fonte: Autores (2019).

Medicação Didático Pedagógica: A mediação didática informa a qual tipo de ensino o aluno está associado, sendo de três tipos: Presencial, Semipresencial e Ensino a Distância.

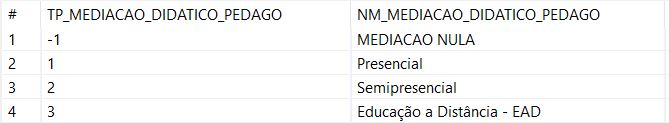


Figura x – Dados inseridos na dimensão Mediação. Fonte: Autores (2019).

Nacionalidade: O censo do Inep não abarca apenas alunos brasileiros, mas também estrangeiros e aqueles que nasceram fora do país ou se naturalizaram. Para cada um deles, temos seus respectivos códigos.

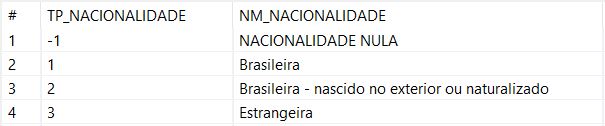


Figura x – Dados inseridos na dimensão Nacionalidade. Fonte: Autores (2019).

Sexo: Também se tem a dimensão que informa o sexo do aluno, conforme os termos utilizados nos questionários dos anos da respectiva análise.



Figura x – Dados inseridos na dimensão Sexo. Fonte: Autores (2019).