UNIVERSIDADE PAULISTA – UNIP
INSTITUTO DE CIÊNCIAS EXATAS E TECNOLOGIA - ICET

CIÊNCIA DA COMPUTAÇÃO

DANIEL GADS MELO SOUSA
GABRIEL DE BRITO SILVA
MARCELO ANTÔNIO DA SILVA JÚNIOR
PEDRO HENRIQUE PEREIRA DE OLIVEIRA

ANÁLISE DO PANORAMA DA INSERÇÃO DO ALUNO NEGRO NA EDUCAÇÃO BÁSICA BRASILEIRA DE 2015 A 2018 UTILIZANDO A ABORDAGEM DE BUSINESS INTELLIGENCE

BRASÍLIA - DF
2019

DANIEL GADS MELO SOUSA

GABRIEL DE BRITO SILVA

MARCELO ANTÔNIO DA SILVA JÚNIOR

PEDRO HENRIQUE PEREIRA DE OLIVEIRA

ANÁLISE DO PANORAMA DA INSERÇÃO DO ALUNO NEGRO NA EDUCAÇÃO BÁSICA BRASILEIRA DE 2015 A 2018 UTILIZANDO A ABORDAGEM DE BUSINESS INTELLIGENCE

Trabalho de Conclusão de Curso para obtenção do título de Bacharel em Ciência da Computação da Universidade Paulista – UNIP, *campus* Brasília.

Aprovado em:

BANCA EXAMINADORA

_____

Prof. Msc. Claudio Bernardo

Universidade Paulista

_____

Prof. Msc. Josyane Lannes

Universidade Paulista

_____

Prof. Msc. Luiz Antônio

Universidade Paulista

AGRADECIMENTOS

Agradeço a Deus pela a vida e por todas as graças a mim concedidas.

À instituição Universidade Paulista – UNIP na pessoa da coordenadora Liliane Cordeiro por ter oferecido todas as oportunidades de fomentação do conhecimento para o curso de Ciência da Computação.

Ao orientador Claudio Bernardo pela paciência, incentivo e orientação no presente trabalho.

A professora Josyane Lannes por todo o auxílio na parte de Banco de Dados e pela sua incrível e inspiradora didática nas aulas da respectiva matéria.

A Caixa e ao FNDE por terem auxiliado no início da minha caminhada no mundo profissional, por meio do estágio. Em especial agradeço aos meus colegas Murillo Higor Fernandes Carvalhes e Débora Arnaud Lima Formiga pelos seus inestimáveis auxílios e incentivos referentes à área de *Business Intelligence.* Além da EPROJ, setor que me acolheu tão bem nos meus últimos momentos de estágio e que proporcionou a minha atuação em um setor de Análise de Dados.

Aos meus colegas de trabalho que tanto auxiliaram para a conclusão desse projeto.

Aos meus colegas de curso em que juntos compartilhamos conhecimento, dificuldades e momentos de alegria.

Daniel Gads Melo Sousa

Agradeço a Deus pela minha vida e pelas bênçãos, me proporcionando estar realizando este trabalho.

Agradeço a minha família pelo direcionamento, pelo apoio incondicional, pelas dicas e pelo suporte, especialmente a meus pais e irmãos.

Agradeço aos colegas de turma, em especial os integrantes deste grupo de trabalho, que sempre se ajudaram e auxiliaram neste projeto. Além de um grande amigo e ex-colega de turma Filipe Ribeiro, que sempre se disponibilizou e nos auxiliou neste projeto.

Agradeço aos profissionais da Caixa e FNDE que me impulsionaram na vida profissional, auxiliando sempre que puderam.

Gabriel de Brito Silva

Agradeço a Deus por todas as bênçãos e oportunidades que me propôs a conquistar.

À minha família, pelo carinho, pelo amor e pelo apoio essencial e de imediato sempre que foi necessário. Especialmente a meus pais por todo suporte me dado.

Ao Banco do Brasil por me proporcionar um aprendizado, e ter me acolhido ao longo do meu percurso acadêmico. Com um agradecimento especial ao caro colega Shalom Galvão Montoril, que dividiu seus conhecimentos me guiando na área de *Business Intelligence,* e também a todos que me acompanharam nessa jornada dentro da instituição.

Aos mestres que dispuseram de seu tempo e conhecimento para me guiar nessa jornada. E por fim agradeço aos meus colegas de curso, por todos os momentos de alegria, companheirismo e aprendizagem.

Marcelo Antônio da Silva Júnior

Agradeço aos meus companheiros de grupo, a minha família pelo apoio, aos professores do curso que me guiaram por esse caminho, também a todos os obstáculos que tive no meio do caminho, que antes não percebia, mas hoje percebo que me fizeram mais forte.

Pedro Henrique Pereira de Oliveira

“Sem dados você é apenas mais uma pessoa com uma opinião”.

(W. Edwards Deming)

LISTA DE ABREVIATURAS E SIGLAS

BI – *Business Intelligence* (Inteligência de Negócio)

DW – *Data Warehouse*

DM – *Data Mart*

ETL – *Extract, Transform and Load* (Extração, Transformação e Carga)

SQL – *Structured Query Language*

MEC – Ministério da Educação

IBM – *International Business Machines Corporation*

INEP – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

IBGE – Instituto Brasileiro de Geografia e Estatística

OLAP – *Online Analytical Processing* (Processamento Analítico *Online*)

OLTP – *Online Transaction Processing* (Processamento de Transações *Online*)

PDI – *Pentaho Data Integration*

DBMS – *Database Management Systems* (Sistema Gerenciador de Banco de Dados)

EIS – *Executive Information System* (Sistema de Informação Executiva)

ATM – *Automatic Teller Machine* (Máquina de Caixa Automático)

ERP – *Enterprise Resource Planning*

CSV – *Comma-separated Values*

UF – Unidade da Federação

XLSX – *Excel Microsoft Office Open XML Format Spreadsheet File*

LISTA DE FIGURAS

RESUMO

Este estudo teve como objetivo analisar o modelo de implantação do *Business Intelligence* e de que forma a aplicação do BI pode contribuir com informações relevantes sobre a inserção do aluno negro na educação básica brasileira. Para tanto, explicou-se conceitos, técnicas e características essenciais do *Business Intelligence*, além de uma rápida passagem sobre o contexto educacional brasileiro básico. Os dados utilizados para a análise são os microdados do censo escolar da educação básica brasileira do ano de 2015 a 2018, que foram coletados do site de dados abertos do governo brasileiro. A partir da análise desses dados percebeu-se como a localização, a imigração e a falta de qualidade de vida básica do ser humano influenciam no panorama do aluno negro na educação básica brasileira. É importante ressaltar que esse estudo é voltado para o conceito, implantação e utilização do *Business Intelligence* e como a utilização do próprio pode contribuir com informações relevantes. Enfim, por meio do estudo realizado, foi possível demonstrar o processo de *Business Intelligence* para analisar dados importantes sobre a atuação do aluno negro na educação básica brasileira.

Palavras-Chave: ETL, OLAP, *dashboard*, *Data Warehouse*, *Business Intelligence*, Análise de Dados, Educação Básica brasileira, alunos negros.

ABSTRACT

This study aimed to analyze the Business Intelligence deployment model and how the BI application can contribute with relevant information about student insertion in Brazilian basic education. To this end, the concepts, techniques and essential features of Business Intelligence, as well as a quick passage on the basic Brazilian educational context. The data used for analysis are the microdata of the Brazilian basic education school census from 2015 to 2018, which were collected on the Brazilian government's open data site. From the analysis of these data, we realize how the location, migration and lack of basic quality of life of human beings influence the panorama of black students in Brazilian basic education. Importantly, this study is focused on the concept, implementation and use of Business Intelligence and how the use of the individual can contribute relevant information. Finally, through the study, it was possible to demonstrate the Business Intelligence process to analyze important data about student performance in Brazilian basic education.

Keywords: ETL, OLAP, *dashboard*, *Data Warehouse*, *Business Intelligence*, *Data Analysis*, Brazilian elemetary education, black students.

SUMÁRIO

1 INTRODUÇÃO

Nessa seção será apresentado o trabalho junto dos seus objetivos gerais e específicos.

1.1 Justificativa

Devido a necessidade de uma análise da inserção do aluno negro na educação básica brasileira essa pesquisa se justifica através da aplicação da abordagem de *Business Intelligence* (BI) em contribuição a sua utilidade para análise de dados.

1.2 Problematização

Buscou-se reunir dados/informações com o propósito de responder ao seguinte problema de pesquisa: de que forma a aplicação da abordagem de *Business Intelligence* auxilia na análise do panorama da inserção do aluno negro na educação básica brasileira?

1.3 Delimitação do tema

Este projeto de pesquisa limitou-se a colher informações sobre de que forma a aplicação da abordagem de *Business Intelligence* auxilia uma análise da inserção do aluno negro na educação básica brasileira tendo como referência a base de microdados do Censo Escolar dos anos de 2015 a 2018 do INEP.

1.4 Objetivos

O objetivo geral do trabalho e os objetivos específicos em prol da conclusão do mesmo são descritos abaixo.

1.4.1 Objetivos gerais

O presente trabalho tem como objetivo geral apresentar de que forma a aplicação da abordagem de *Business Intelligence* auxilia na análise dos dados da inserção do aluno negro na educação básica brasileira no contexto educacional básico brasileiro, com a finalidade de comprovar a sua utilidade para análise dos dados a partir da base de microdados do Censo Escolar dos anos de 2015 a 2018 do INEP.

1.4.2 Objetivos específicos

- Levantar o estado da arte no que tange a *Business Intelligence*, sua metodologia e processos;
- Apresentar os indicadores sobre a inserção do aluno negro na educação brasileira e seus requisitos;
- Aplicar a abordagem de *Business Intelligence*, bem como os processos de ETL (*Extraction, Transformation and Loading*, em inglês e Extração, Transformação e Carga, em português) e a montagem do ambiente de *Data Warehouse* no case estudado;
- Desenvolver os resultados das análises através da solução Power BI;

1.5 Metodologia

Esse estudo tem por finalidade realizar uma pesquisa aplicada, uma vez que utilizará conhecimento aplicado à resolução de problemas. Segundo Silva e Menezes (2005, p. 20), “a pesquisa aplicada objetiva gerar conhecimentos para aplicação prática e dirigidos à solução de problemas específicos. Envolve verdades e interesses locais”.

Para um melhor tratamento dos objetivos e melhor apreciação desta pesquisa, observou-se que ela é classificada como pesquisa descritiva. Segundo Silva e Menezes (2005, p. 21):

> “A pesquisa descritiva visa descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis. Envolve o uso de técnicas padronizadas de coleta de dados: questionário e observação sistemática. Assume, em geral, a forma de levantamento”.

Detectou-se também a necessidade da pesquisa bibliográfica no momento em que se fez uso de materiais já elaborados: livros, artigos científicos, revistas, documentos eletrônicos e enciclopédias na busca e alocação de conhecimento sobre a abordagem de *Business Intelligence* para a análise dos dados no contexto educacional básico brasileiro, correlacionando tal conhecimento com abordagens já trabalhadas por outros autores.

Como procedimentos, pode-se citar a necessidade de pesquisa bibliográfica, isso porque será feito uso de material já publicado, constituído principalmente de livros, também se entende como um procedimento importante o estudo de caso como

procedimento técnico (SILVA e MENEZES, 2005, p. 21). Tem-se como base para o resultado da pesquisa um caso em específico que poderá ser expandido futuramente.

A abordagem do tratamento da coleta de dados do estudo de caso será quantitativa, pois requer o uso de recursos e técnicas de estatística, procurando traduzir em números os conhecimentos gerados pelo pesquisador (SILVA e MENEZES, 2005, p. 20). Existirão gráficos; obtém opinião dos entrevistados com questões fechadas, por exemplo: Qual sua área de atuação? (Saúde, Administração) Medir através de números. Por exemplo: Pesquisa de satisfação. 40% Insatisfeito, 10% Não opinaram, 50% Satisfeitos.

O problema foi direcionando a pesquisa para as áreas de análise da inserção do aluno negro na educação básica brasileira, sendo este com a aplicação da abordagem de *Business Intelligence* para a análise dos dados no contexto educacional básico brasileiro. Em que será feito ao apresentar os indicadores sobre a inserção do aluno negro na educação brasileira e seus requisitos afirmando que objetivo geral:

O presente trabalho tem como objetivo geral apresentar de que forma a aplicação da abordagem de *Business Intelligence* permite uma análise da inserção do aluno negro na educação básica brasileira utilizando dados do contexto educacional, com a finalidade de analisar a utilidade da abordagem de BI para análise de dados na base de microdados do Censo Escolar dos anos de 2015 a 2018 do INEP.

2 EMBASAMENTO TEÓRICO

Nessa seção serão apresentadas as fontes, trabalhos, artigos e autores utilizados para o embasamento desse trabalho.

2.1 *Business Intelligence*

Richard Millar Devens, em 1865, foi quem introduziu o termo "*Business Intelligence*" (Inteligência de Negócios) em seu livro *Cyclopædia of Commercial and Business Anecdotes*. Lá ele conta sobre um bancário, Sir Henry Furnese, que conseguia atuar antes da competição reunindo informações e conseguindo lucrar com elas (DEVENS, 1865).

Em 1958, Hans Peter, um cientista da computação da IBM, publicou um artigo que foi um marco no assunto, na qual descrevia o quão potencial o *Business Intelligence* (BI) seria com o uso da tecnologia. O artigo intitulado "*A Business Intelligence System*" descrevia:

> An automatic system is being developed to disseminate information to the various sections of any industrial, scientific or government organization. This intelligence system will utilize data-processing machines for auto-abstracting and auto-encoding of documents […] (LUHN, 1958, p. 314).

Em tradução livre: "Um sistema automático está sendo desenvolvido para disseminar informação para os setores industriais, científicos ou organizações governamentais. Esse sistema de inteligência vai utilizar máquinas de processamento de dados para abstrair e codificar automaticamente os documentos".

O termo "pai do BI" é relativo, pois depende da interpretação, muitos autores relacionam a Devens por ter cunhado o termo, mas quem descreveu o uso na área da computação e seu potencial foi Hans Peter Luhn. Se houvesse a necessidade de vincular o termo a alguém, seria à Hans Peter, pois o BI moderno que é conhecido atualmente é graças a ele.

Após a Segunda Guerra Mundial, houve a necessidade de simplificar e organizar o grande e rápido crescimento dos dados tecnológicos e científicos. Hoje, Luhn (Figura 1) é popularmente conhecido como o pai do *Business Intelligence*.

Figura 1 – Hans Peter Luhn

Fonte: (IEEE Spectrum, 2018)

A partir dos anos 60, surgiram novas formas de armazenamento como os DBMS (*Database Management Systems*), tendo uma evolução no modo de gerenciar grandes volumes de dados, no final da década de 70, nasce o modelo relacional no DBMS. A partir desse momento, todas as informações eram apresentadas e armazenadas em formato digital, fazendo com que fosse possível a concretização do Business Intelligence nas próximas décadas.

Também nessa mesma época, os CPD (Centros de Processamento de Dados), estavam se consolidando, se transformando no meio-termo da tecnologia da informação com os negócios. Os CPD eram focados totalmente em dados, diferente da TI que o centro focava em software, hardware e redes.

Com o surgimento do conceito de sistemas de informações executivas (EIS) há uma grande disseminação do assunto, sendo um dos maiores aliados aos sistemas de BI. Segundo Turban (2009, p. 27), "Esse conceito expandiu o suporte computadorizado aos gerentes e executivos de nível superior. Alguns dos recursos introduzidos foram sistemas de geração de relatórios dinâmicos multidimensionais [...]".

Na década de 80, alguns fabricantes de softwares voltados ao campo do BI começaram a ganhar terreno. Softwares como *MicroStrategy*, *Business Objects* e

Crystal Reports começaram a ser populares nas empresas que começaram a usar realmente computadores na época.

Em 1988 houve outro marco importante, com o intuito de simplificar as análises em BI a conferência internacional em Roma, organizada pelo *Multiway Data Analysis Consortium*, dá início à era moderna do Business Intelligence, sendo o termo popularizado pelo analista do *Gartner Group*, Howard Dresner, em 1989.

Na década de 90, se populariza o conceito de *Data Warehouse*, como um sistema dedicado a auxiliar o BI, também separando em momentos distintos os processamentos OLAP (*Online Analytical Processing*) e OLTP (*Online Transaction Processing*), sendo o OLAP usado ao lado do BI para a montagem de relatórios e posteriormente painéis em inúmeras visões diferentes, e o OLTP sendo utilizado geralmente para explorações estatísticas. Ao mesmo tempo, por consequência, o conceito de ETL (*Extraction, Transformation and Loading*) é incorporado ao *Data Warehouse,* com o intuito de fornecer dados relevantes e fornecer uma extração focada.

O sistema ERP (*Enterprise Resource Planning*) é um software voltado para a gestão das empresas, garantindo a entrada de dados essencial para que os sistemas de BI, sendo possível reportar aos gestores análises em pontos específicos.

É importante ter em mente sobre a construção e organização do BI que este processo é sem fim, sempre haverá novos requisitos a serem cumpridos mesmo tendo terminado o trabalho, sendo necessário refazer todas as etapas novamente.

2.2 Sistemas de Informação OLAP/OLTP

De acordo com Primak (2008), o objetivo de uma ferramenta OLAP é permitir a análise multidimensional dinâmica dos dados, apoiando os usuários finais em suas atividades, oferecendo várias perspectivas, onde o próprio usuário cria suas consultas dependendo de suas necessidades, fazendo cruzamento dos dados de formas diferenciadas, auxiliando na busca pelas respostas desejadas.

Para oferecer as várias perspectivas o método mais comumente usado é o MOLAP (*Multidimensional On Line Analytical Processing*) onde se usa um banco de dados multidimensional com tabelas que mais parecem um cubo, que por esse motivo originou a denominação “dados cúbicos”. Com essas tabelas de múltiplas dimensões é possível cruzar informações que antes não seria possível por uma pessoa, fazendo assim necessário o uso da ferramenta. Existem também outros métodos como o ROLAP (*Relational On Line Analytical Processing*) que utiliza um banco de dados relacional para seus dados e possui um tempo maior para resposta.

Para Thomsen (2002) o OLAP precisa dos requisitos abaixo para funcionar:

- Uma estrutura dimensional;
- Especificação eficiente das dimensões e cálculos;
- Separação da estrutura e representação;
- Flexibilidade;
- Velocidade suficiente para suportar as análises *ad hoc*;
- Suporte para multiusuários;

Segundo Prasad (2007) o processo OLAP se diferencia do OLTP (*Online Transaction Processing* ou Processamento de Transações em Tempo Real) que foca em processar transações repetitivas em alta quantidade e manipulação simples. Já o OLAP envolve uma análise de vários itens de dados em relacionamentos complexos, que busca padrões, tendências e exceções.

O foco principal do OLTP é transações online. Como o nome já diz, suas consultas são simples e curtas, portanto não precisa de tanto tempo de

processamento, também utiliza pouco espaço. O banco de dados é atualizado frequentemente, pode acontecer no momento da transação e também é normalizado. Um exemplo muito utilizado ao se explicar o OLTP é o ATM (do inglês, caixa automático) onde cada transação modifica a conta do usuário.

2.3 *Data Warehouse*

As aplicações de *Business Intelligence* necessitam de um repositório específico para buscar os dados que serão usados para a operação, sejam eles de que tipo for. O local onde serão centralizadas essas informações é chamado de *Data Warehouse* (DW). Segundo Inmon (2005, p. 29) "*Data Warehouse* (que no português significa literalmente armazém de dados) é um depósito de dados orientado por assunto, integrado, não volátil, variável com o tempo, para apoiar as decisões gerenciais". Como é verificado na definição de Inmon, a necessidade de um repositório específico para as informações é definida pelos seguintes itens:

- Orientado por assunto: Significa que os dados ali presentes têm contexto direto com as atividades da empresa, ou seja, as informações contidas são, unicamente, as necessárias para definir as operações.
- Integrado: Todas as atividades do DW devem estar conectadas ao ambiente operacional.
- Não volátil: Os dados que estão especificadamente no *Warehouse* não podem sofrer mudanças, tal como alterações e inclusões (já que é necessária uma informação concreta que não se altere para tomar decisões), podendo ser apenas consultados ou excluídos.
- Variável com o Tempo: Está ligado ao conceito da Não-Volatilidade, mas aqui com um foco maior no tempo dessas informações. Já que os dados dentro do DW não podem ser alterados, o horário das informações também não pode mudar. Por isso é necessária a certeza do tempo em que elas estão armazenadas.

Enquanto o *Data Warehouse* agrega todos os dados, definições e relacionamentos entre eles (PANOLY, 2019), o *Data Mart* (DM) é um pequeno DW dentro do contexto maior que se preocupa em adquirir apenas uma parte dessas informações para uma operação específica. Pode ser feita uma analogia com um comerciante que possui uma loja e um armazém, ao comprar novos produtos para o

seu comércio, ele, primeiramente, vai armazenar tudo o que for possível nesse armazém, para, posteriormente, pegar certas quantidades de determinados produtos e apresentá-los na sua loja para vendê-los. Isso é feito para que os relatórios gerados utilizem uma única base para adquirir as informações.

2.4 O Método ETL

As informações que serão utilizadas no processo de Business Intelligence são adquiridas por meio de um processo chamado ETL (*Extract, Transform and Load)*. Esse processo tem como função "transportar" e transformar os dados para todas as instâncias do BI, seja na aquisição dos dados das fontes, inserção desses dados no Banco de Dados e transformação dos dados nos tipos necessários para a realização da análise. Esse processo já foi conhecido como ETLM, em que o "M" significava *Maintenance* (Manutenção), mas esse último já caiu em desuso (BRAGHITTONI, 2017). Cada uma das suas ações será explicada adiante.

E - *Extract (Extração*): A primeira parte do processo de transporte é a extração periódica das informações para que sejam posteriormente carregadas. Elas podem ser extraídas de diversas fontes, sejam arquivos do tipo CSV, tipo texto (TXT), arquivos *mainframe*, sites e até arquivos em diferentes fontes de dados (CARVALHAES e ALVES, 2015). A extração deve estar preparada para reconhecer esses dados nas mais variadas fontes possíveis.

T - *Transform (*Transformação): Após os dados extraídos, eles precisam passar por uma verificação para depois serem "carregados" no *Data Warehouse.* Como Inmon (2005, p. 29) afirma que os dados dentro do DW não podem ser modificados (propriedade Não Volátil), é recomendado o uso de uma instância intermediária ao *Warehouse para* que eles sejam transformados segundo as regras de negócio. Essa instância pode ser outro banco de dados chamado *Staging Area* ou a própria memória do servidor/computador (CARVALHAES e ALVES, 2015).

L - *Load* (Carga): O último processo é da carga propriamente dita no *Data Warehouse* e nos seus *Data Marts.* No final desse processo os dados já estão prontos para o uso de alguma ferramenta de BI, transformando eles em algum tipo de visualização gráfica (*Dashboards).* Essa carga é feita de forma incremental, já que,

como mencionado anteriormente por Inmon, esses dados não podem sofrer mudanças (BRAGHITTONI, 2017).

2.5 Ferramentas

Para o desenvolvimento desse trabalho foram utilizadas algumas ferramentas que serão explicadas adiante.

2.5.1 Pentaho Data Integrator

Com o desejo de alcançar uma mudança positiva no mercado de análise de negócio dominada por grandes vendedores que oferecem produtos baseados em plataformas com custo elevado, foi-se criado o *Pentaho.* É uma ferramenta de gerenciamento de *Business Intelligence (*BI), desenvolvido para fins de recolher o máximo de dados diversificados e não estruturados a partir de diversas fontes e analisá-los para encontrar novos padrões, indicadores de tendências e base de dados para inovação.

Por ser uma tecnologia *Open Source,* o *Pentaho* possui uma versão funcional extremamente potente em que não há custo algum com licenças. É a ferramenta de código aberto mais utilizada do mundo, contendo um ambiente de desenvolvimento integrado.

O *Pentaho* possui diversas funções a fim de entender e analisar o negócio empresarial. Algumas delas permitem que o usuário possa acessar e preparar fontes de dados para análise, mineração e geração de relatórios *on-line,* via *web.* Também vem integrado com um ambiente de desenvolvimento, baseado no *Eclipse* (API), que visa à resolução de soluções mais complexas de BI.

2.5.2 Microsoft Power BI

Com o crescimento de negócios das empresas, uma grande massa de dados que entram nessas empresas também aumentou e por causa disso, ficou difícil organizar, analisar, monitorar e compartilhar essa quantidade de informações. Para resolver esse problema, foi necessário criar ferramentas que pudessem atender a esses requisitos.

Dentre várias ferramentas que foram criadas, têm-se o *Power BI.* É um pacote de ferramentas de análise de negócios que tem como objetivo a visualização, organização e análise de dados. Segundo a Microsoft (2019) o *Power BI* é uma ferramenta baseada na nuvem, tornando possível, ao usuário, a conexão de seus dados em tempo real, ou seja, em qualquer lugar que eles estejam, com rapidez, eficiência e compreensão. Além disso, ele permite a criação de painéis de simples visualização, fornecimento de relatórios interativos e o compartilhamento desses dados obtidos.

Sendo uma ferramenta de *Business Intelligence,* o *Power BI* não é apenas para a inserção de dados e criação de relatórios. Ela transforma os dados brutos em informações significativas e úteis a fim de analisar o negócio, permite uma fácil interpretação do grande volume de dados e entrega análises que ajudam na decisão de uma grande variedade de negócios, variando do operacional ao estratégico. Também é capaz de limpar os dados formatados de forma irregular, garantindo que tenham apenas aqueles interessados ao negócio e modela os dados quem vêm de diversas fontes para que se consiga fazer com que esses dados trabalhem de forma eficiente.

3 ESTUDO DE CASO: MEC E INEP

Nessa seção serão explicados os trabalhos semelhantes a este e uma explicação sobre o Ministério da Educação (MEC) e o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

3.1 Demandas e observações sobre os dados do INEP

Outras pesquisas semelhantes já foram feitas nessa área de estudo, tais como o trabalho apresentado por Oliveira (2018) quando apresenta alguns aspectos históricos e contemporâneos do negro e discorre de maneira abrangente sua inserção no sistema educacional brasileiro, trazendo aspectos históricos desde as épocas da Colônia, Império e Primeira República até os dias atuais. Já no artigo de Almeida e Sanchez (2016) é apresentada uma compreensão das legislações que regem a vida do negro na sua caminhada educacional para a visibilidade e valorização deles na construção do cotidiano escolar, passando pelo início da entrada do negro na educação formal brasileira até a atuação do movimento negro nessa práxis.

No artigo de Oliveira (2013), a autora procura discorrer sobre a atuação da lei 10.639 para os professores, diante do contexto da educação básica e da questão racial. Zandona (2008) discorre sobre a problemática da desigualdade racial dos negros no contexto do ensino médio brasileiro, abordando temas como o racismo e a questão socioeconômica para o entendimento desse fato.

Passos (2010) discorre sobre a população negra e as dificuldades enfrentadas por ela no contexto da Educação de Jovens e Adultos (EJA), passando pela construção dessas desigualdades e pelas políticas nacionais aplicadas para a promoção da igualdade. O texto de Fonseca *et al.* (2001) faz uma compilação de diversos artigos voltados para a atuação do negro sob várias perspectivas, como a educação das crianças negras no contexto da promulgação da Lei do Ventre Livre, de 1871; análise de projetos e iniciativas sobre as relações raciais voltadas as escolas da rede municipal de Belo Horizonte; análise do perfil dos estudantes negros ingressantes nos cursos de Direito; análise das questões de raça e gênero das graduandas negras da Unicamp.

No ano de 2015 foi divulgado pelo Ministério da Educação (MEC), um relatório com o “Título de Educação para Todos”, nele foi feito um estudo centrado em vários aspectos da educação. Segundo a própria pesquisa, foram resumidos em seis principais tópicos, são eles: Cuidados e Educação na Primeira Infância, Educação Primária Universal, Habilidades de Jovens e Adultos, Alfabetização de Adultos, Paridade e Igualdade de Gênero, Qualidade da Educação.

Em virtude disso, este trabalho de conclusão traz a contribuição de uma análise com especificação, focada na observação da inserção do aluno negro voltada ao recorte temporal de 2015 a 2018 no contexto da educação básica brasileira utilizando a base de microdados do Censo Escolar do INEP.

3.2 MEC E INEP

A história do Ministério da Educação é antiga, começando em 1930 quando foi criado no governo de Getúlio Vargas, inicialmente era chamado de Ministério dos Negócios da Educação e Saúde Pública. Como se pode ver pelo nome, a educação não era o único foco de atividade. Apenas em 1995, no governo de Fernando Henrique Cardoso, a educação ficou exclusiva ao ministério. A sigla MEC surgiu em 1953, quando se criou o Ministério da Educação e Cultura.

Até 1960 de acordo com o MEC (2015), ele era centralizado, um modelo que era seguido por todos os municípios e estados. A primeira Lei de Diretrizes e Bases da Educação (LDB), que fora aprovada em 1961, diminuiu a centralização do MEC, fazendo assim os órgãos municipais e estaduais ganharem mais autonomia.

Em 2015 a primeira versão da BNCC (Base Nacional Comum Curricular) é disponibilizada, uma pauta muito debatida e vista como importante para a educação. Somente em 2017 ela foi homologada para Ensino Básico e um ano depois, para o Ensino Médio.

> A Base é um documento normativo da maior importância, porque define o conjunto de aprendizagens essenciais que todos os alunos devem desenvolver ao longo da Educação Básica e do Ensino Médio, e orientar as propostas pedagógicas de todas as escolas públicas e privadas de Educação Infantil, Ensino Fundamental e Ensino Médio, em todo o Brasil (MEC, 2015).

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) é vinculado ao MEC e engloba várias áreas educacionais, desde ensino básico até a graduação. A respeito de sua origem, é preciso considerar que:

> Chamado inicialmente de Instituto Nacional de Pedagogia, o Inep foi criado, por lei, em 13 de janeiro de 1937, no Rio de Janeiro. Foi em 1938, entretanto, que o órgão iniciou, de fato, seus trabalhos. A publicação do Decreto-Lei nº 580 regulamentou a organização e a estrutura da instituição, além de modificar sua denominação para Instituto Nacional de Estudos Pedagógicos. [...] em 1952, assumiu a direção do Instituto o professor Anísio Teixeira, que passou a dar maior ênfase ao trabalho de pesquisa. Seu objetivo era estabelecer centros de pesquisa como um meio de fundar em bases científicas a reconstrução educacional do Brasil. A ideia foi concretizada com a criação do Centro Brasileiro de Pesquisas Educacionais (CBPE), com sede no Rio de Janeiro, e dos Centros Regionais, nas cidades de Recife, Salvador, Belo Horizonte, São Paulo e Porto Alegre. Tanto o CBPE como os Centros Regionais estavam vinculados à nova estrutura do Inep (INEP, 2015).

Na década de 70, com a sede sendo transferida para Brasília e o CBPE (*Centro Brasileiro de Pesquisas Educacionais*) sendo extinto, fez com que o modelo que fora idealizado por Anísio Teixeira fosse finalizado, que causou um reconhecimento ao INEP tanto nacionalmente quanto internacionalmente. Nos anos 80, passou por uma reforma institucional após um período de dificuldades, e tinha dois objetivos, que eram reorientação das políticas de apoio a pesquisas educacionais e o reforço do processo de disseminação de informações educacionais.

O INEP que conhecido hoje é devido à incorporação do Serviço de Estatística da Educação e Cultura (SEEC) à Secretaria de Avaliação e Informação Educacional (SEDIAE) que de acordo com o INEP (2015) a partir de 1997 um único órgão encarregado das avaliações, pesquisas e levantamentos estatísticos educacionais no âmbito do governo federal passou a existir através de uma integração da Secretaria de Avaliação e Informação Educacional (SEDIAE) ao INEP. Nesse mesmo ano, o INEP foi transformado em autarquia federal.

3.3 Como os dados são coletados e disponibilizados

Os dados oferecidos pelo MEC, INEP e outros órgãos do governo são dados abertos, o que significa que estão disponíveis para todos usarem e também redistribuírem como quiserem, sem restrição de patentes, licenças ou parecido. Cada órgão disponibiliza os seus dados de acordo com seu Plano de Dados Abertos (PDA) e é responsável pela catalogação do mesmo.

O Censo Escolar é o principal instrumento de coleta de informações da educação básica e a mais importante pesquisa estatística educacional brasileira. É coordenado pelo INEP e realizado em regime de colaboração entre as secretarias estaduais e municipais de educação e com a participação de todas as escolas públicas e privadas do país. (INEP, 2015)

A coleta de dados das escolas segundo o INEP (2015) tem caráter declaratório e é dividida em duas etapas. A primeira etapa consiste no preenchimento da matrícula inicial, quando ocorre a coleta de informações sobre os estabelecimentos de ensino, gestores, turmas, alunos e profissionais escolares em sala de aula. A etapa seguinte ocorre com o preenchimento de informações sobre a situação do aluno, e considera os dados sobre o rendimento e movimento escolar dos alunos, ao final do ano letivo. O Censo Escolar é regulamentado por instrumentos normativos, que instituem a obrigatoriedade, os prazos, os responsáveis e suas responsabilidades, bem como os procedimentos para realização de todo o processo de coleta de dados.

No caso do INEP eles podem ser acessados no seu próprio site, no link: http://inep.gov.br/web/guest/microdados, onde haverá uma página nominada “Microdados”, lá estão separados por categoria e ano. Além disso, para outros dados, tem-se o Portal Brasileiro de Dados Abertos que possui os dados do INEP e de diversos outros órgãos, no link: http://dados.gov.br/.

De acordo com o Portal Brasileiro de Dados Abertos (2017) em 2007 um grupo de 30 pessoas se reuniu na Califórnia - EUA, para definir os princípios dos Dados Abertos Governamentais. Eles criaram oito princípios, que são:

1. **Completos**: Todos os dados públicos são disponibilizados, que no caso, não são sujeitos a limitações válidas de privacidade, segurança ou controle de acesso, reguladas por estatutos.
2. **Primários**: Os dados são publicados do mesmo jeito que fora coletada na fonte, e não de forma agregada ou transformada.
3. **Atuais**: Os dados são disponibilizados o mais rápido possível para preservar o seu valor.
4. **Acessíveis**: Os dados são públicos para o maior público possível.
5. **Processáveis por máquina**: Os dados são estruturados para possibilitar o seu processamento automático.

6. **Acesso não discriminatório**: Os dados estão disponíveis para todos sem necessidade de identificação.
7. **Formatos não proprietários**: Os dados estão disponíveis sem que nenhum ente tenha exclusivamente um controle.
8. **Licenças livres**: Os dados não estão sujeitos a restrições como dito no parágrafo acima.

4 DESCRIÇÃO DA MONTAGEM DO AMBIENTE

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

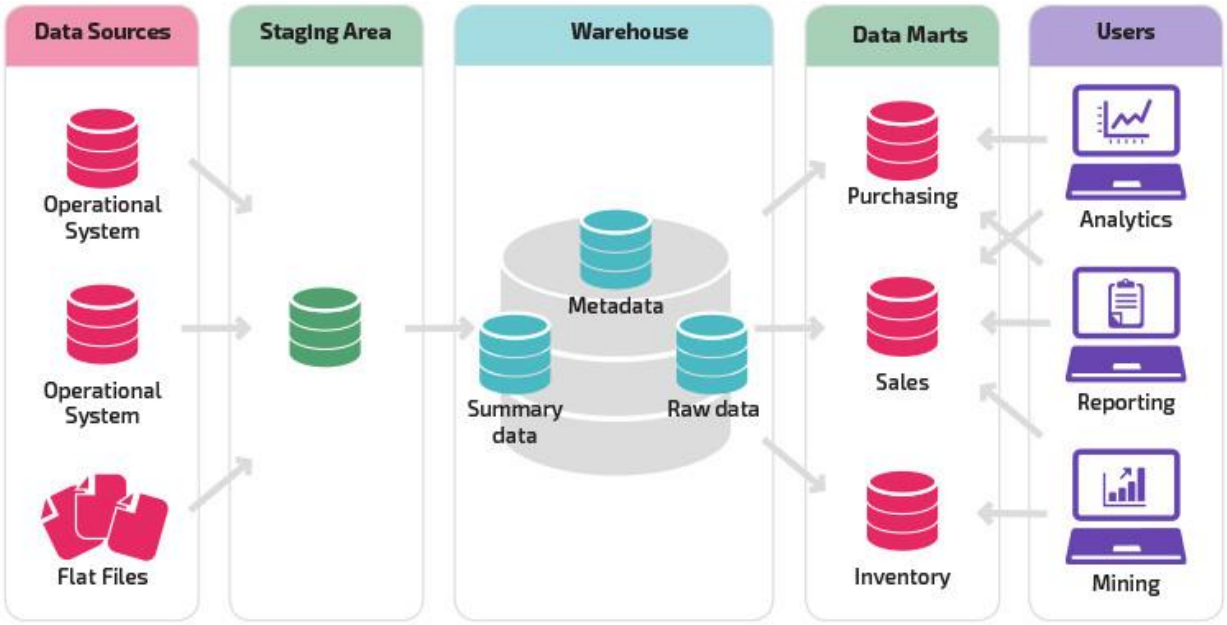
4.1 Introdução

Segundo Braghittoni (2017, p. 1): “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 abordagem 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 (Figura 2):

- Parte 1: Fontes de Dados (*Data Source*), em que é feita a definição da localização dos dados, seu formato e quais deles serão aproveitados para a análise.
- Parte 2: Área de *Staging* (*Staging Area*), passo intermediário e opcional onde os dados são gravados, na sua forma original, em tabelas para posterior transformação e inserção.
- Parte 3: *Data Warehouse* (DW), que adquire os dados do banco *Staging*, após serem feitas as transformações para que eles possam ser utilizados pela ferramenta de BI. Sua criação pode ser antes ou depois de um *Data Mart*.
- Parte 4: *Data Marts*, que são pequenos DW relacionados a um assunto específico. Sua criação pode ser antes ou depois de um *Data Warehouse* dependendo a abordagem escolhida. Tais abordagens serão explicadas adiante.
- Parte 5*:* Análise dos resultados, em que a ferramenta de BI escolhida acessa os dados de um *Data Warehouse* ou de um *Data Mart* para que sejam feitas as gerações das análises.

Figura 2 - Arquitetura do ambiente de BI



Fonte: Panoly (2019).

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

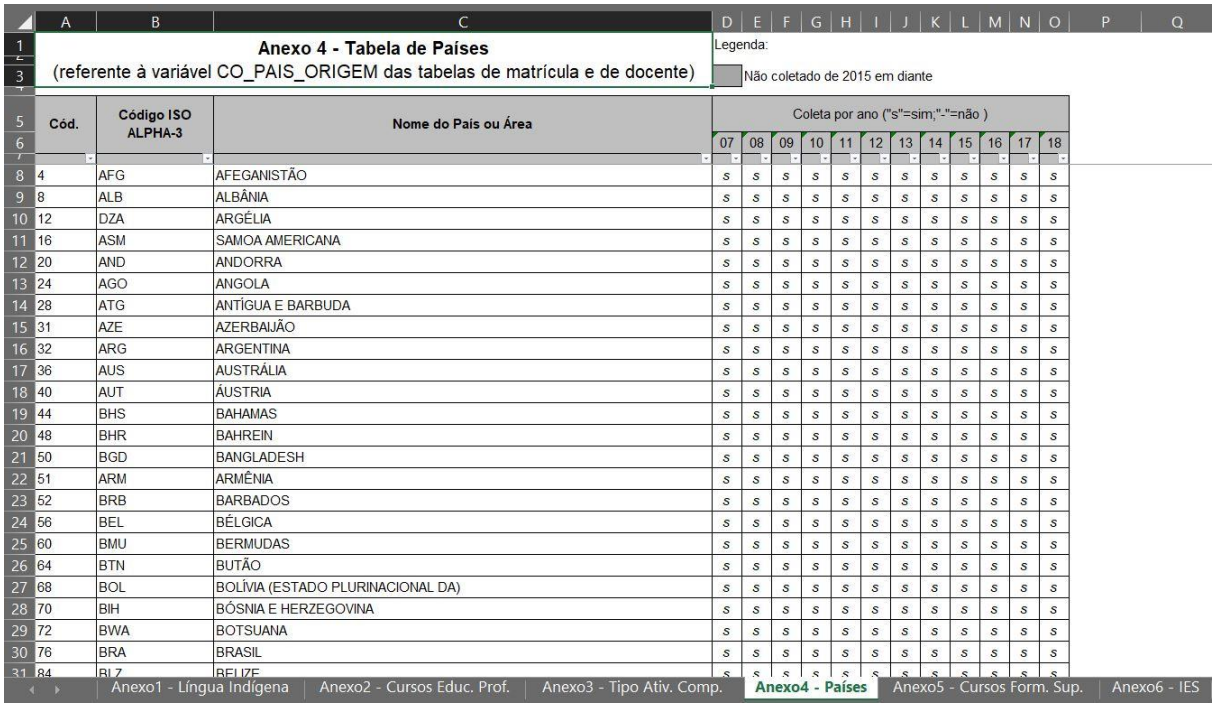
Como visto na Figura 2, o primeiro passo na aplicação dos processos de Business Intelligence é definir quais serão as bases de dados utilizadas para o processo e quais dados serão extraídos delas. No caso do presente trabalho, foram utilizadas as bases de microdados do censo escolar do INEP, disponíveis 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 (Figura 3) que contêm os códigos dos países descritos no censo, já que alunos estrangeiros também são envolvidos no censo escolar.

Figura 3 - Tabela de códigos dos países



Anexo 4 - Tabela de Países
(referente à variável CO_PAIS_ORIGEM das tabelas de matrícula e de docente)

Legenda:
Não coletado de 2015 em diante

Cód.	Código ISO ALPHA-3	Nome do País ou Área	Coleta por ano ("s"=sim;"-"=não)											
			07	08	09	10	11	12	13	14	15	16	17	18
4	AFG	AFEGANISTÃO	s	s	s	s	s	s	s	s	s	s	s	s
8	ALB	ALBÂNIA	s	s	s	s	s	s	s	s	s	s	s	s
12	DZA	ARGÉLIA	s	s	s	s	s	s	s	s	s	s	s	s
16	ASM	SAMOA AMERICANA	s	s	s	s	s	s	s	s	s	s	s	s
20	AND	ANDORRA	s	s	s	s	s	s	s	s	s	s	s	s
24	AGO	ANGOLA	s	s	s	s	s	s	s	s	s	s	s	s
28	ATG	ANTÍGUA E BARBUDA	s	s	s	s	s	s	s	s	s	s	s	s
31	AZE	AZERBAIJÃO	s	s	s	s	s	s	s	s	s	s	s	s
32	ARG	ARGENTINA	s	s	s	s	s	s	s	s	s	s	s	s
36	AUS	AUSTRÁLIA	s	s	s	s	s	s	s	s	s	s	s	s
40	AUT	ÁUSTRIA	s	s	s	s	s	s	s	s	s	s	s	s
44	BHS	BAHAMAS	s	s	s	s	s	s	s	s	s	s	s	s
48	BHR	BAHREIN	s	s	s	s	s	s	s	s	s	s	s	s
50	BGD	BANGLADESH	s	s	s	s	s	s	s	s	s	s	s	s
51	ARM	ARMÊNIA	s	s	s	s	s	s	s	s	s	s	s	s
52	BRB	BARBADOS	s	s	s	s	s	s	s	s	s	s	s	s
56	BEL	BÉLGICA	s	s	s	s	s	s	s	s	s	s	s	s
60	BMU	BERMUDAS	s	s	s	s	s	s	s	s	s	s	s	s
64	BTN	BUTÃO	s	s	s	s	s	s	s	s	s	s	s	s
68	BOL	BOLÍVIA (ESTADO PLURINACIONAL DA)	s	s	s	s	s	s	s	s	s	s	s	s
70	BIH	BÓSNIA E HERZEGOVINA	s	s	s	s	s	s	s	s	s	s	s	s
72	BWA	BOTSUANA	s	s	s	s	s	s	s	s	s	s	s	s
76	BRA	BRASIL	s	s	s	s	s	s	s	s	s	s	s	s
84	BLZ	BELIZE	s	s	s	s	s	s	s	s	s	s	s	s

Anexo1 - Língua Indígena | Anexo2 - Cursos Educ. Prof. | Anexo3 - Tipo Ativ. Comp. | Anexo4 - Países | Anexo5 - Cursos Form. Sup. | Anexo6 - IES

Fonte: Adaptado de INEP (2019).

Para as UF, Municípios, Distritos, Mesorregiões e Microrregiões, foram utilizadas as bases de códigos *Geodata* disponíveis no site *GitHub* no link*:* https://github.com/paulofreitas/geodata-br/tree/master/data/pt. O *GitHub* é um site para a criação de repositórios públicos e privados com o intuito de compartilhar informações e códigos e o repositório *Geodata* tem como propósito prover informações precisas e atualizadas acerca dos dados geográficos do Brasil. Essas informações são um compilado das informações disponíveis na SIDRA (Sistema IBGE de Recuperação Automática) formados pelo IBGE (Instituto Brasileiro de Geografia e Estatística).

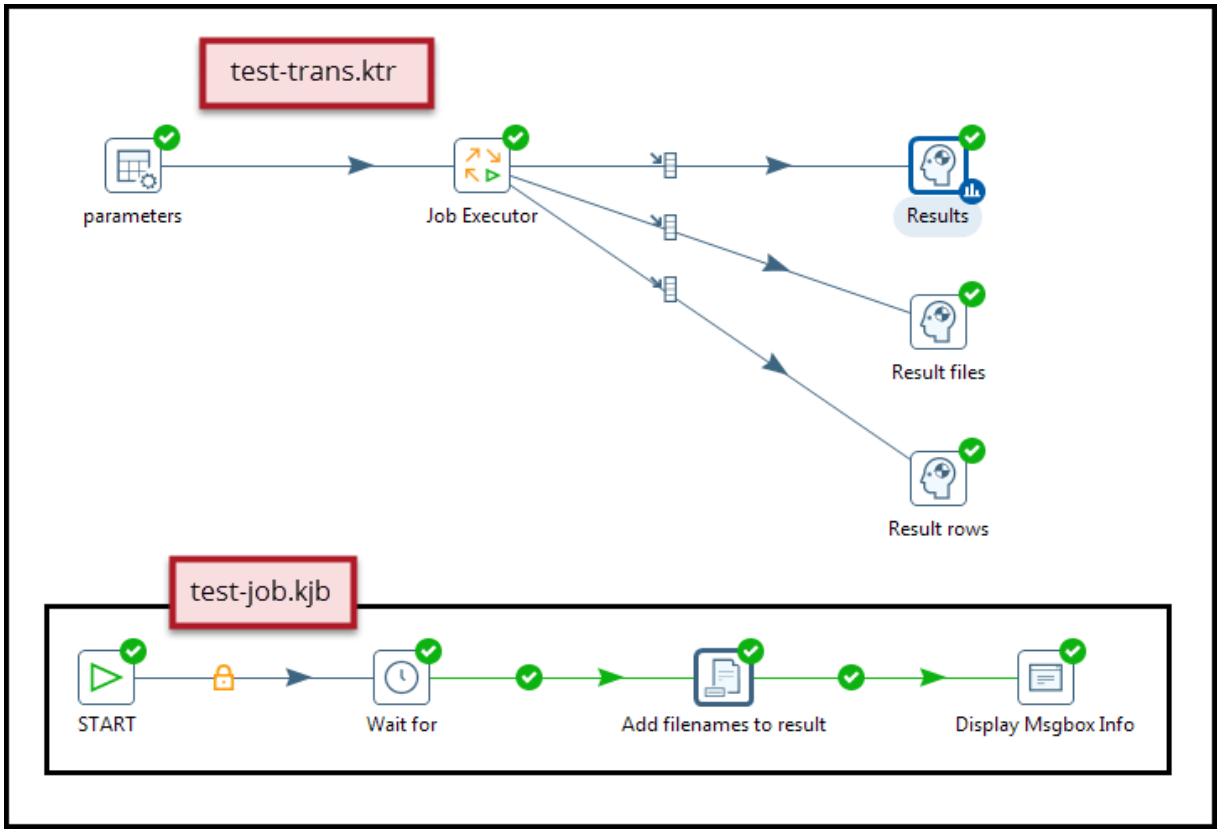
4.3 Montagem do ambiente – Área de *Staging*

Inmon (2005, p. 29) define em um dos seus postulados sobre *Data Warehouse* 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*, como é mostrado na Figura 2. 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*.

Foi 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. Como na imagem abaixo:

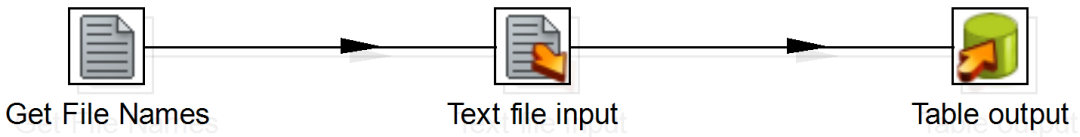
Figura 4 - Exemplo de Transformation e Job



Fonte: Pentaho (20-?).

A carga dos arquivos no BD dos arquivos principais (turmas, matrícula, escolas, docentes) e das bases de códigos das UF, Municípios, Distritos, Mesorregiões e Microrregiões são compostas por três passos, em que o PDI encontra os arquivos, prepara-os para a inserção e grava-os no BD, como pode ser visto na imagem abaixo:

Figura 5 - Visão da ETL das bases principais



Fonte: Autores (2019).

Os passos são descritos abaixo:

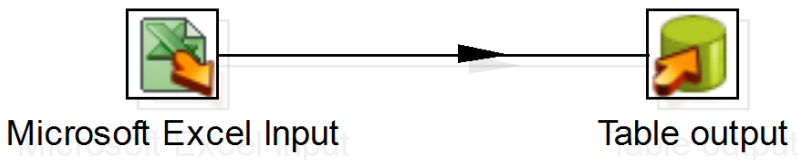
Get File Names: Esse *step* procura nomes de arquivos ou pastas. Ele é recomendado para quando se tem uma grande massa de dados em que todos precisam ser gravados. Os padrões dos nomes são adquiridos conforme uma expressão regular.

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 automaticamente para a mesma ser criada.

Já para a carga das tabelas contendo o código dos países, foi utilizado um padrão de carga diferente, já que o arquivo que possui esse dado está em um formato diferente das outras bases, como pode ser visto na figura 6:

Figura 6 - Visão geral da ETL de auxiliares



Fonte: Autores (2019).

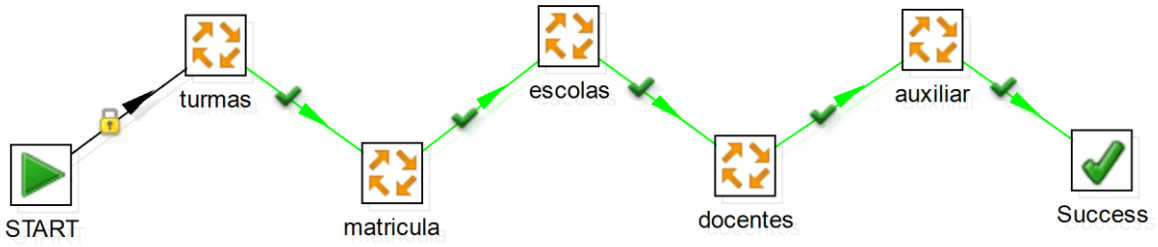
Os passos são descritos abaixo:

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: Como descrito nas cargas principais, 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 automaticamente para a mesma ser criada.

Após definir cada uma das ETLs, será usado uma *Transformation* para unir todos os outros *Jobs*, como pode ser visto na imagem abaixo:

Figura 7 - Visão geral da ETL Staging



Fonte: Autores (2019).

4.4 Montagem do ambiente – *Data Warehouse*

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

4.4.1 Fato e Dimensão

Em uma modelagem multidimensional temos dois tipos de tabelas principais: Fato e Dimensão.

Pela definição de Kimball (2013) dimensão é uma coleção de atributos semelhantes ao texto que estão altamente correlacionados entre si. Isso quer dizer que ela possui característica descritiva. Para se criar uma dimensão podem ser feitas

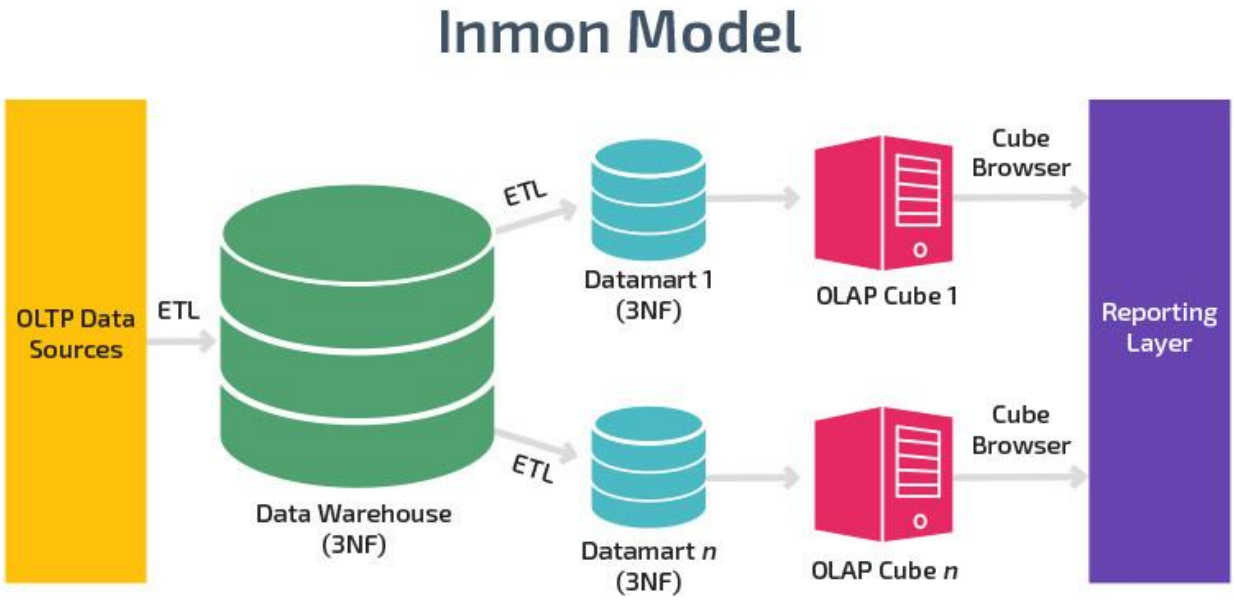
algumas perguntas como “quando”, “onde”, “quem”, e “o que”. Já na tabela fato, normalmente os dados são apenas números, categorizando-a em quantitativa, mas também se podem ter textos que estão classificando o fato em análise.

4.4.2 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 que 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. Seu diagrama é mostrado abaixo:

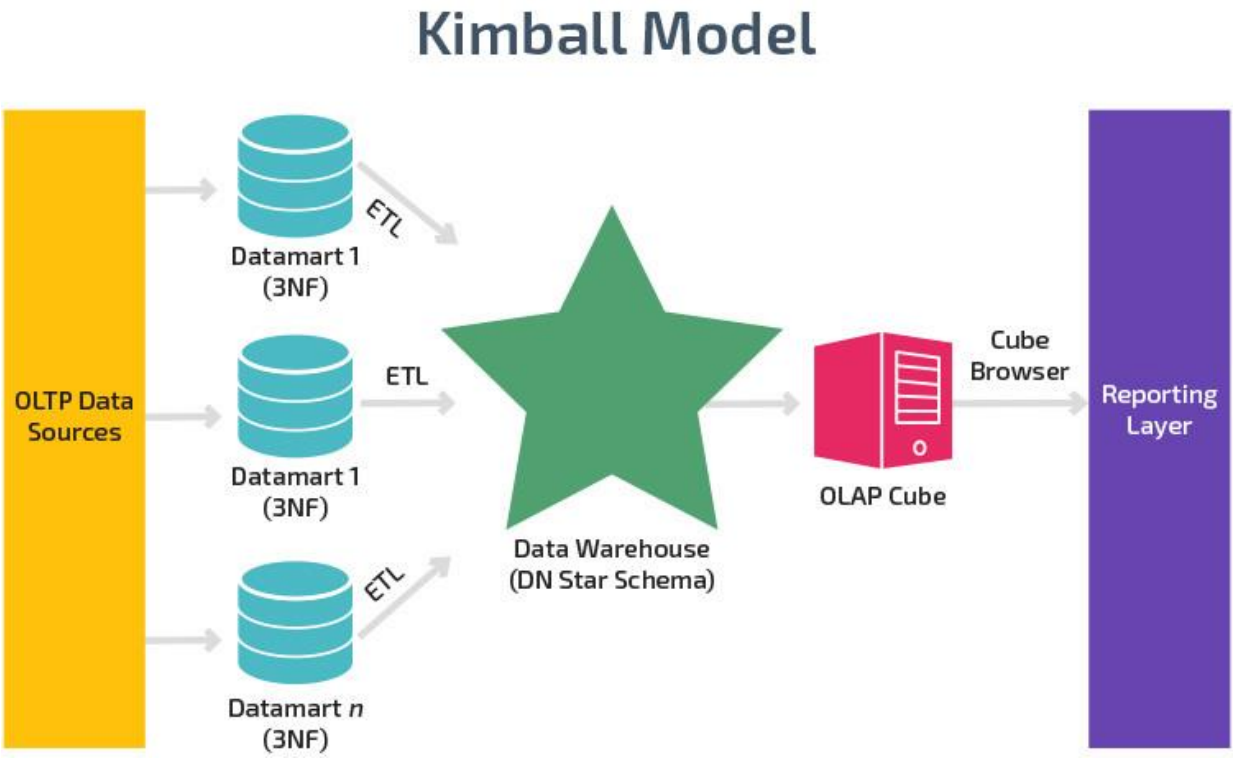
Figura 8 - Modelo Inmon



Fonte: Panoly (2019).

Em contrapartida, Kimball utiliza a abordagem *bottom-up* em que é feita primeiramente a criação de *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 (2013) afirma: “O *Data Warehouse* não é nada mais do que uma junção de diversos *Data Marts*”. Seu diagrama é mostrado abaixo:

Figura 9 - Modelo Kimball



Fonte: Panoly (2019).

Na Tabela 1 é feita uma comparação entre as abordagens Inmon e Kimball:

Tabela 1 - Comparação entre as abordagens do Inmon e Kimball

	Inmon	Kimball
Manutenção	Fácil	Difícil - muitas vezes redundante e sujeito a revisão
Tempo	Maior tempo para iniciar	Menor tempo para iniciar
Conhecimento preciso	Time especialista	Time generalista
Tempo de construção do Data Warehouse	Demorado	Rápido
Custo para implantar	Custos iniciais altos, com menores custos subsequentes de desenvolvimento do projeto	Custos iniciais pequenos, com cada projeto subsequente custando-o mais
Persistência dos dados	Alta taxa de mudança dos dados	Relativamente estável

Fonte: Autores (2019)

Para a realização desse trabalho foi escolhida a abordagem Inmon porque o projeto não fez uso de *Data Marts*, assim sendo, foi criado unicamente o *Data Warehouse* para armazenar os dados.

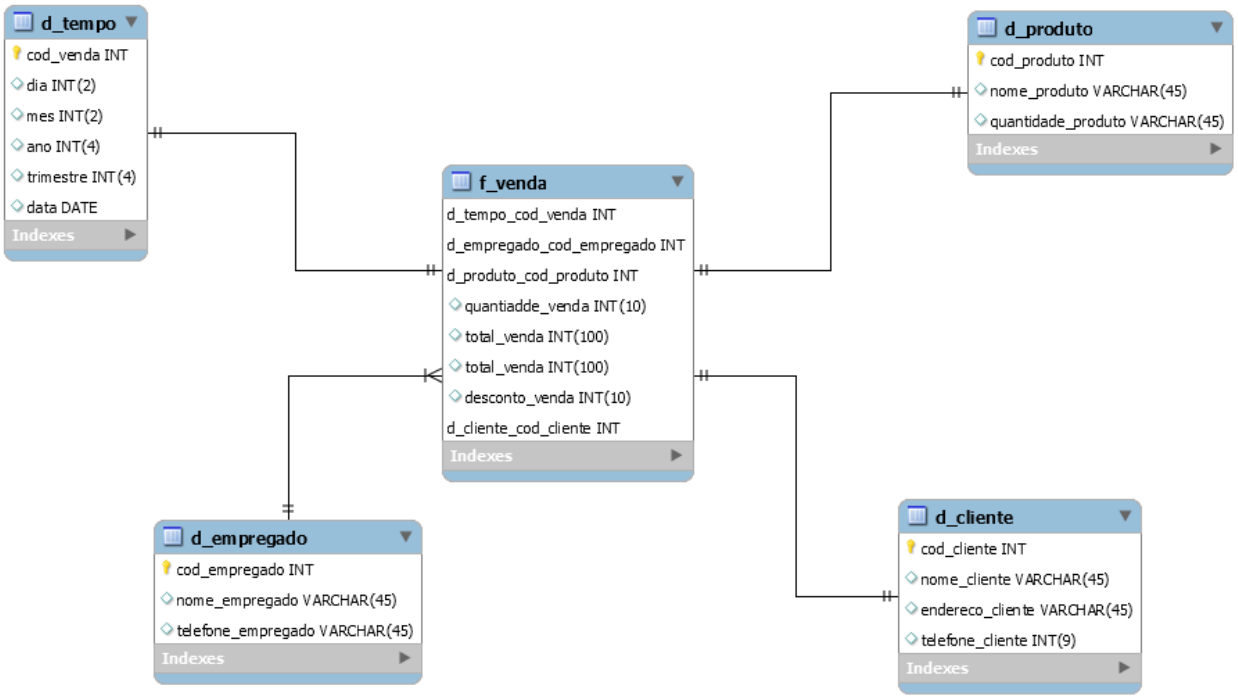
4.4.3 Modelos Estrela e Floco de Neve (*Star Schema and Snow-Flake Schema*)

Tendo definida a estrutura, inicia-se o desenvolvimento do modelo do DW.

Em um modelo de dados multidimensional pode ser utilizado dois tipos de modelos, que são: tipo Estrela (*Star Schema*) ou tipo Floco de Neve (*Snow-Flake Schema*).

O modelo Estrela é o mais básico e mais comum para a arquitetura do *Data Warehouse*. No seu desenho, a tabela fato (F_VENDA, Figura 11) 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 10 - Exemplo de modelo Estrela

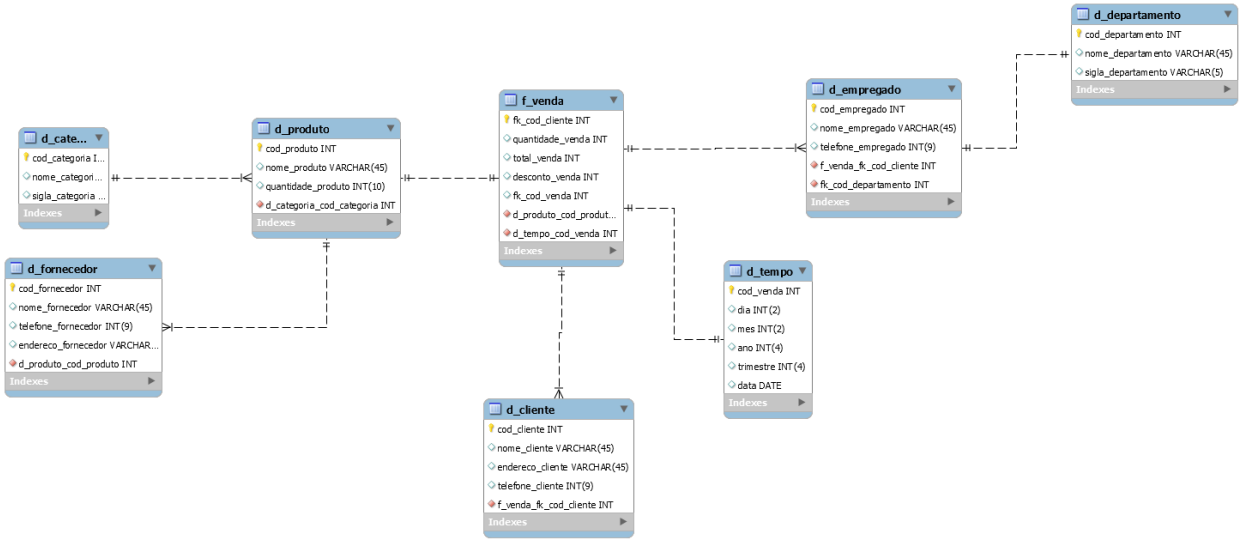


Fonte: Autores (2019).

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 tem-se uma redução de redundância nas tabelas de dimensões e uma menor quantidade de memória utilizada. Um exemplo seria a 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. Seu diagrama é mostrado abaixo:

Figura 11 - Exemplo de modelo Floco de Neve



Fonte: Autores (2019).

Na Tabela 2 é feita uma comparação entre os modelos Estrela e Floco de Neve:

Tabela 2 - Comparação entre Star Schema e Snow Flake Schema

	Star Schema	*Snow Flake*
Manutenção	Possui dados redundantes que dificultam manter ou alterar	Sem redundância, portanto, facilita manter e alterar
Facilidade de Uso	Consultas menos complexas e de fácil entendimento	As consultas são mais complexas o que torna difícil de entender
Performance nas consultas	Menos *foreign keys* o que torna a consulta mais rápida	Mais *foreign keys* o que torna a consulta mais lenta
Espaço	Não tem tabelas normalizadas o que aumenta o espaço	Tem tabelas normalizadas
Bom para:	Bom para *Data Mart* com relacionamentos simples (1:1 ou 1:N)	Bom para uso em *Data Warehouse* para simplificar as relações complexas (N:N)

Fonte: Autores (2019).

Para o presente trabalho, foi utilizado o modelo Floco de Neve. Devido algumas dimensões apresentarem hierarquia nelas, houve-se a necessidade de criar uma tabela adicional.

4.4.4 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, serão definidas as 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 a UF que possui a maior concentração de alunos estrangeiros negros?
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), Localidade Município (Município, UF, País), Localidade Distrito (Microrregião, Município, UF, Região, Mesorregião, Distrito) e Escola.

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

4.4.5 Processo ETL para carga do *Data Warehouse*

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

4.4.5.1 Definição dos indicadores nulos

Segundo Braghittoni (2017, p. 94) nenhuma coluna que esteja inserida no *Data Warehouse* pode aceitar valores nulos (*null*). O site *W3Schools* (2019) define valores *null* como um campo que não possui valor, deixado em branco no momento da gravação. Sendo assim, há a necessidade de criar valores genéricos para definir um valor que veio nulo. Em cada uma das dimensões explicadas adiante, será descrito os seus respectivos indicadores nulos.

4.4.5.2 Dimensão Tempo (Ano)

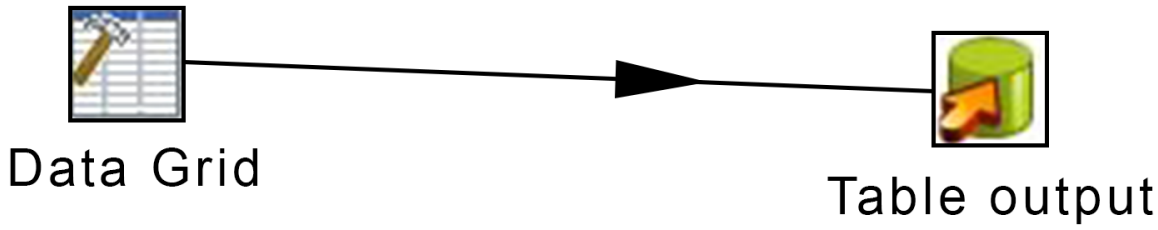
Inmon (2005, p. 29) define em um dos seus postulados sobre *Data Warehouse* a variabilidade com o tempo, ou seja, um DW e suas informações vivem com base no tempo. Com base dessa informação, Braghittoni (2017, p. 31) atesta que “Por mais que não exista nenhuma outra dimensão no seu DW, a dimensão temporal deve estar lá” e também (2017, p. 73) “Como postulado por Inmon, o DW é sempre variável com o tempo, ou seja, a dimensão DATA deve invariavelmente existir”.

Conforme definido pelos dois autores, todo projeto necessita obrigatoriamente de uma dimensão de tempo, em contrapartida, esse ‘tempo’ pode ser descrito de formas diferentes por cada projeto, com base nas necessidades das análises. Ele pode ser definido tanto como informações separadas (ano ou mês ou dia), uma data, formada por ano, mês, e dia, ou até uma informação mais complexa inserindo trimestre, dia da semana, hora, etc.

Para o presente trabalho, a dimensão de tempo será identificada pelos anos referentes a cada análise (2015 até 2018), um identificador para cada uma delas e seu indicador nulo que será explicado adiante.

Seu diagrama de carga é mostrado na imagem abaixo:

Figura 12 - Visão geral da ETL Ano



Fonte: Autores (2019).

Os seguintes passos foram utilizados:

Data Grid: Neste *step* será criada uma tabela com um conjunto constante de dados, informando os nomes dos campos, seus tipos, e seus respectivos dados. Aqui está sendo carregado cada um dos anos da análise e um indicador para cada um deles.

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 automaticamente para a mesma ser criada. Aqui os dados estão sendo gravados na tabela D_TEMPO do banco de dados.

Indicador nulo da dimensão:

Como explicado no início da seção, serão atribuídos valores para serem inseridos para caso o campo da informação, no momento da carga, for nulo. Para essa dimensão será usado o indicador ‘-1’ para caso em algum momento da carga essa informação estiver nula.

4.4.5.3 Dimensões Localidade

Como descrito na seção 4.4.4 acerca dos indicadores e das dimensões, será usado duas tabelas chamadas de Aluno e Matrícula, elas por sua vez utilizam informações geográficas na sua estrutura.

Em base dos microdados do INEP, as tabelas sobre Aluno utilizam as informações de município, UF, e país, por outro lado, a dimensão Escola faz uso das informações de distrito, município, UF, microrregião, mesorregião e região.

Tendo em vista que cada uma das tabelas apresenta uma combinação de informações diferentes, fez-se necessário dividir as informações de localidade, com cada uma sendo chamada pelo seu menor grão de informação. No caso das combinações de Aluno, a dimensão com as suas combinações será chamada de Localidade Município, e de Escola, chamada de Localidade Distrito, por este ser o menor nível de informação. Essas informações geográficas seguem a seguinte ordem (do maior para o menor):

1. País;
2. Região;
3. UF;
4. Mesorregião;
5. Microrregião;
6. Município;
7. Distrito;

As definições de cada uma das dimensões Localidade serão explicadas adiante.

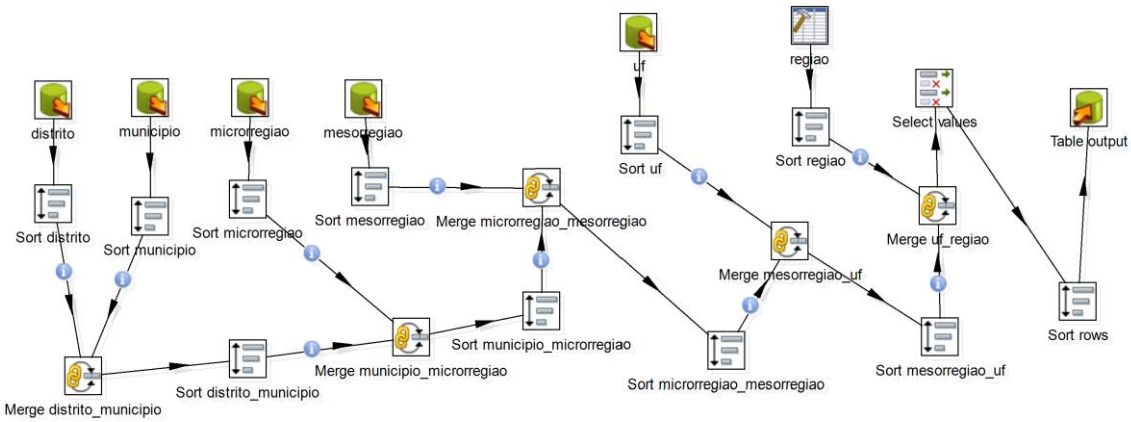
4.4.5.3.1 Dimensão Localidade Distrito

Como explicado anteriormente, uma das tabelas será a Localidade Distrito que irá apoiar as combinações da tabela Escola. Essa tabela de Localidade é formada

pela combinação das informações de distrito, município, microrregião, mesorregião, UF e região, junto de um identificador único para cada uma dessas combinações, além dos identificadores nulos.

Seu diagrama de carga é mostrado na imagem abaixo:

Figura 13 - Diagrama da ETL Localidade Distrito



Fonte: Autores (2019).

Os seguintes passos foram utilizados:

Table Input (na imagem acima com os nomes: distrito, municipio, microrregiao, mesorregiao, uf): Este *step* permite utilizar os dados já existentes em alguma tabela para fazer outras operações, como a inserção em outro banco, por exemplo. Os dados nesse passo são adquiridos por meio de um comando SQL, mas o *Pentaho* possui uma interface gráfica para selecionar esses dados, sem necessidade do comando, se assim o usuário preferir.

Data Grid: Neste *step* será criada uma tabela com um conjunto constante de dados, informando os nomes dos campos, seus tipos, e seus respectivos dados. Aqui está sendo carregado cada uma das regiões da análise.

Sort rows (na imagem acima com os nomes: Sort distrito, Sort municipio, Sort microrregiao, Sort mesorregião, Sort regiao, Sort distrito_municipio, Sort municipio_microrregiao, Sort microrregiao_mesorregiao, Sort mesorregiao_uf): Esse passo possibilita a ordenação de um conjunto de dados com base em uma coluna informada. Seu uso é semelhante ao comando *order by* do SQL. Nele podem ser

configuradas outras opções como ordenação ascendente ou descendente e diferenciação de maiúsculas e minúsculas.

Merge Join (na imagem acima com os nomes: Merge distrito_municipio, Merge municipio_microrregiao, Merge microrregiao_mesorregiao, Merge mesorregiao_uf, Merge uf_regiao): Com um funcionamento semelhante ao comando *JOIN* do SQL, esse *step* une dois fluxos de informação com base em uma coluna compartilhada (*key*), além de ser possível a configuração da forma de união (retornar apenas os dados que se relacionam ou também aqueles que não se relacionam). Requer o uso do *step Sort rows* antes deste para ordenação da coluna escolhida.

Select values: Este passo é utilizado para remover colunas, alterar o nome delas bem como seus tipos.

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 automaticamente para a mesma ser criada.

Este processo de carga inicia-se com a aquisição dos dados da tabela (*step Table input* com o nome ‘distritos’) que contêm as informações dos códigos de distritos (menor nível de informação, daí o nome da dimensão), dados estes inseridos previamente na seção 4.3 de montagem do *Staging*. Além desses códigos, são lidos também os códigos referentes aos municípios associados a cada distrito, na própria tabela de distritos e na tabela de municípios (*step Table input* com o nome ‘municipio’).

No momento que todas as informações são adquiridas, o próximo passo é ordená-las (*step Sort rows* com os nomes ‘Sort distrito’ e ‘Sort municipio’) de modo ascendente escolhendo a coluna que contêm os códigos dos municípios, esse passo de ordenação é necessário para o funcionamento correto do próximo passo.

Após as informações ordenadas, é feita a ‘união’ desses dois fluxos de informação (*step Merge Join* com o nome ‘Merge distrito_municipio’) utilizando como coluna de união os códigos de município em cada um dos fluxos. Por exemplo: na tabela de distritos, um distrito de nome Lua Nova (cód. 521295610) possui nas suas informações o código de município 5212956, fazendo a união desse código na tabela

de municípios, é encontrado esse código associado ao nome do município Matrinchã. Esse processo é feito para todos os distritos na tabela distritos no banco *Staging*.

A próxima tabela a ser acessada é a que contêm as informações dos códigos das Microrregiões (*step Table input* com o nome ‘microrregiao’). Como descrito no processo da tabela distrito, essa tabela foi carregada na seção 4.3 do banco de *Staging*. Em conjunto desses, no momento da carga da tabela anterior de municípios também foi adquirida as informações referentes aos códigos Microrregiões associadas a cada uma.

Tal como no processo anterior, após as informações serem adquiridas, elas são ordenadas (*step Sort rows* com os nomes ‘Sort microrregiao’ e ‘Sort distrito_municipio’) de modo ascendente, agora utilizando a coluna com os códigos das Microrregiões como referência.

Após isso é feita a ‘união’ (*step Merge Rows* com o nome ‘Merge município_microrregiao’) desses fluxos tal como o processo anterior, mas utilizando a coluna com o código das Microrregiões como forma de união. Por exemplo: continuando com o município anteriormente especificado (Matrinchã, cód. 5212956), ele possui em sua base o código de Microrregião 52002, que na tabela de Microrregião o código está associado ao nome Rio Vermelho.

O mesmo processo é aplicado a seguir, com as informações de Mesorregião sendo adquiridas (*step Table input* com o nome ‘mesorregiao’) e ordenadas (*step Sort rows* com os nomes ‘Sort mesorregiao’ e ‘Sort municipio_microrregiao’) pelo seu respectivo código junto com as informações de mesorregião adquiridas na tabela de microrregião, sendo feita a sua união (*step Merge Rows* com o nome ‘Merge microrregiao_ mesorregiao’) no final do processo.

Repetindo os processos anteriores, adquirem-se as informações dos códigos das UFs brasileiras (*step Table input* com o nome ‘uf’) e é feita a sua ordenação junto com o resultado da união anterior (*step Sort rows* com os nomes ‘Sort uf’ e ‘Sort microrregiao_mesorregiao’) e posteriormente a sua união (*step Merge Rows* com o nome ‘Merge mesorregião_uf’) com base nas informações dos códigos das UFs na ordenação anterior.

Por último, são adquiridas as informações sobre as regiões brasileiras (*step Data Grid* com o nome ‘regiao’), feita sua ordenação e da união anterior (*step Sort rows* com os nomes ‘Sort regiao’ e ‘Sort mesorregiao_uf’) e a união desses resultados com base na coluna de código das regiões (*step Merge Rows* com o nome ‘Merge uf_regiao’).

Finalmente, após todos os resultados serem retornados é usado o *step Select values* para remover as colunas redundantes geradas no momento das uniões, mantendo uma coluna para cada código e seus respectivos nomes, sendo ordenado logo após (*step Sort rows*) e inserido na sua dimensão de localidade (*step Table output*).

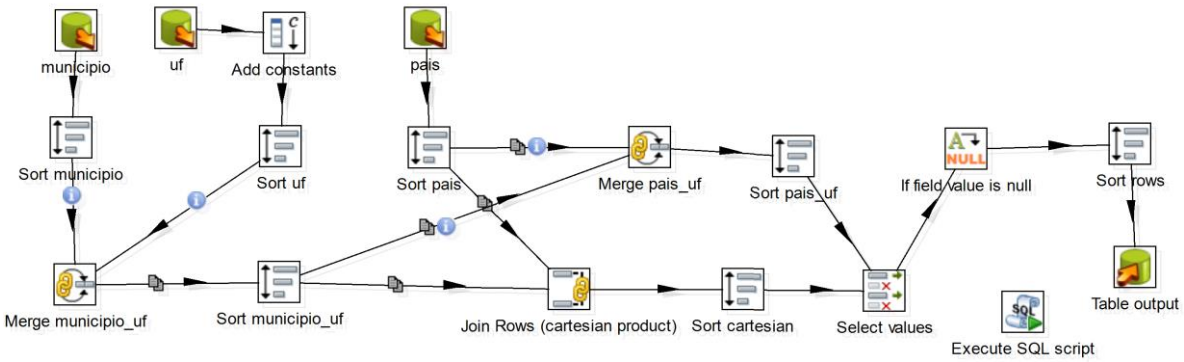
Como explicado no início da seção, serão atribuídos valores para serem inseridos para caso o campo da informação, no momento da carga, for nulo. Para essa dimensão será usado o indicador ‘-1’ para caso em algum momento da carga essa informação estiver nula.

4.4.5.3.2 Dimensão Localidade Município

Para a segunda tabela, tem-se a Localidade Município. A tabela em questão vai apoiar as combinações da fato Aluno, composta por município, UF e país. Além do identificador único para cada combinação e indicadores nulos.

Seu diagrama de carga é mostrado na imagem abaixo:

Figura 14 - Diagrama da ETL Localidade Município



Fonte: Autores (2019).

Os seguintes passos foram utilizados:

Table Input (na imagem acima com os nomes: municipio, uf, pais): Como descrito na carga anterior, este *step* permite utilizar os dados já existentes em alguma tabela para fazer outras operações, como a inserção em outro banco, por exemplo. Os dados nesse passo são adquiridos por meio de um comando SQL, mas o *Pentaho* possui uma interface gráfica para selecionar esses dados, sem necessidade do comando, se assim o usuário preferir.

Sort rows (na imagem acima com os nomes: Sort municipio, Sort uf, Sort pais, Sort municipio_uf, Sort pais_uf, Sort cartesian): Como dito na carga anterior, esse passo possibilita a ordenação de um conjunto de dados com base em uma coluna informada. Seu uso é semelhante ao comando *order by* do SQL. Nele podem ser configuradas outras opções como ordenação ascendente ou descendente e diferenciação de maiúsculas e minúsculas.

Merge Join (na imagem acima com os nomes: Merge município_uf, Merge pais_uf): Explicado na carga anterior, possui um funcionamento semelhante ao comando *JOIN* do SQL. Esse *step* une dois fluxos de informação com base em uma coluna compartilhada (*key*), além de ser possível a configuração da forma de união (retornar apenas os dados que se relacionam ou também aqueles que não se relacionam). Requer o uso do *step Sort rows* antes deste para ordenação da coluna escolhida.

Select values: Como explicado na carga anterior, este passo é utilizado para remover colunas, alterar o nome delas bem como seus tipos.

Execute SQL Script: Conforme explicado anteriormente, nesse *step* o *Pentaho* permite executar um ou mais comandos SQL para fazer alguma operação no BD, seja uma consulta (*SELECT*) ou uma inserção (*INSERT*). Além disso, é possível utilizar variáveis criadas no próprio PDI no código.

Add constants: Neste passo é criado um fluxo constante de dados para serem inseridos junto com outro fluxo. Nele é possível configurar o nome da coluna que vai gerar esses dados, bem como os próprios dados a serem gerados.

Join rows (cartesian product): Tem o funcionamento parecido com o *Merge Join*, mas no seu caso, ele é usado para multiplicar dois fluxos de informações criando todas as combinações possíveis entre eles, fazendo o chamado ‘produto cartesiano’.

If Field Value is Null: Aqui nesse passo o *Pentaho* permite inserir valores em campos que estão nulos. É possível escolher um valor padrão ou especificar valores para cada uma das colunas que chegam ao fluxo de dados.

Table Output: Conforme explicado na parte do *Staging*, esse *step* 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 automaticamente para a mesma ser criada.

Este processo de carga inicia-se com a aquisição dos dados dos códigos e nomes da tabela de Municípios (*step Table Input* com o nome ‘municipio’) carregadas na seção 4.3 de *Staging* junto com os códigos das UFs associadas a eles, além dos dados dos códigos e nomes das UFs brasileiras (*step Table Input* com o nome ‘uf’). Junto com essa carga, é adicionado o código ‘76’ que é referente ao Brasil para ser carregado junto (*step Add constants*).

Ao final dessas duas cargas, são feitas suas respectivas ordenações (*step Sort rows* com os nomes ‘Sort municipio’ e ‘Sort uf’). Após suas ordenações concluídas, é feita a união dos dois fluxos de informações (*step Merge Join* com o nome ‘Merge município_uf’) utilizando como coluna de união os códigos das UFs.

Junto da carga anterior, é feita a aquisição dos dados referentes aos códigos dos países (*step Table Input* com o nome ‘pais’) e sua ordenação ascendente pelo mesmo (*step Sort rows* com o nome ‘Sort pais’). Com os dois *steps* de ordenação prontos (‘Sort pais’ e ‘Sort município_uf’) é feita a cópia de seus dados para cada um dos passos seguintes: o primeiro (*step Merge Join* com o nome ‘Merge pais_uf’), faz a união dos dados dos municípios e UFs com o código ‘76’ que é referente ao país Brasil. O segundo (*step Join Rows (cartesian product)*), faz todas as combinações possíveis dos dados provenientes de municípios e UFs com os outros países, isso foi feito para ter as combinações dos alunos nascidos no exterior/naturalizados, que nasceram em outro país, mas que residem no Brasil. Ao final, os dois fluxos são mais uma vez ordenados (*step Sort rows* com o nome ‘Sort pais_uf’ e ‘Sort cartesian’).

Após a finalização das ordenações anteriores, são removidas as colunas redundantes resultantes das uniões (*step Select values*) e checados os seus valores nulos (*step If field is null*), que nessa situação, serão os alunos estrangeiros que, na base do INEP, não possuem registros de UF/Município de nascimento/endereço, diferente dos alunos nascidos no exterior/naturalizados que possuem um país estrangeiro e registro de UF/Município de nascimento. Ao final é feita sua ordenação ascendente pelo código do país (*step Sort rows*) e sua inserção na respectiva dimensão no *Data Warehouse* (*step Table output*).

Como passo independente, ou seja, que pode ser executado antes ou depois da inserção dos dados no *DW*, tem-se a inserção dos indicadores nulos (*step Execute SQL script*) na dimensão de combinações de municípios. Esses indicadores serão detalhados adiante.

Como citado anteriormente, essa dimensão possuirá os seguintes indicadores de informação nula:

-1: Caso todas as informações estiverem nulas. Esse indicador foi criado em duas combinações: Quando a informação de município, UF e país de origem for nula (-1, -1, -1); quando o aluno tem como país de origem o Brasil, mas não possui informações referentes ao seu município e sua UF (-1, -1, 76).

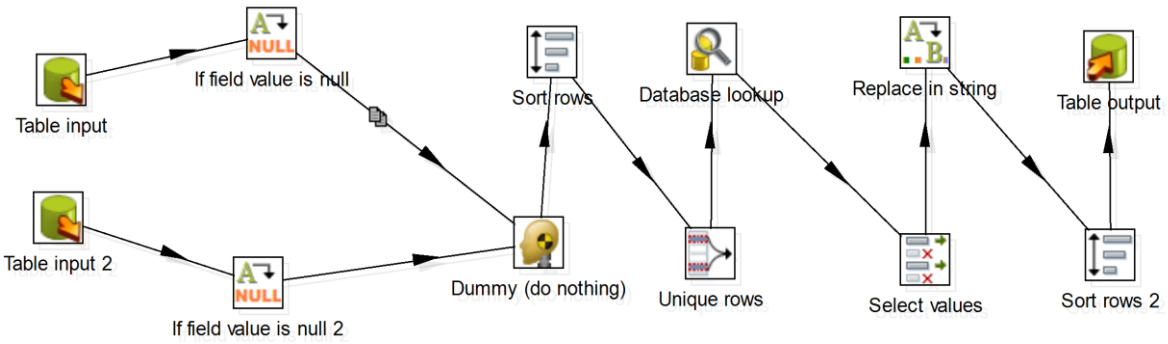
-2: Caso o aluno for estrangeiro. Esse indicador foi criado em uma combinação: Quando o aluno tiver como país de origem qualquer valor diferente de 76 (que faz referência ao Brasil) e suas informações de UF e município não estiverem disponíveis, por exemplo, se o aluno for natural dos Estados Unidos: (-2, -2, 840).

-3: Caso o aluno for naturalizado/nascido no exterior. Esse indicador foi criado em uma combinação: Quando o aluno for naturalizado/nascido no exterior e suas informações de município e UF não estiverem disponíveis (-3, -3, 76).

4.4.5.4 Dimensão Escola

Para a carga da dimensão das escolas, a ordem de ações consiste na extração dos dados dos dois tipos de escolas (públicas e privadas), transformação dos dados inserindo os indicadores de nulo dos dois tipos de escolas, remover as informações duplicadas e posterior inserção delas no *Data Warehouse*, como segue:

Figura 15 - Diagrama da ETL Escola



Fonte: Autores (2019).

Os passos estão descritos a seguir:

Table Input: Este *step* permite utilizar os dados já existentes em alguma tabela para fazer outras operações, como a inserção em outro banco, por exemplo. Os dados nesse passo são adquiridos por meio de um comando SQL, mas o *Pentaho* possui uma interface gráfica para selecionar esses dados, sem necessidade do comando, se assim o usuário preferir. A utilização de dois passos será explicada adiante.

If Field Value is Null: Aqui nesse passo o *Pentaho* permite inserir valores em campos que estão nulos. É possível escolher um valor padrão ou especificar valores para cada uma das colunas que chegam ao fluxo de dados. Seus dois usos nessa ETL serão explicados adiante.

Dummy (do nothing): Esse *step* não realiza ações, mas pode ser utilizado para unir diferentes fluxos de dados ou analisar os dados que estão sendo recebidos. Nessa ETL, ele está juntando os dois fluxos de dados para centralizar a inserção.

Sort rows: Como dito na carga anterior, esse passo possibilita a ordenação de um conjunto de dados com base em uma coluna informada. Seu uso é semelhante ao comando *order by* do SQL. Nele podem ser configuradas outras opções como ordenação ascendente ou descendente e diferenciação de maiúsculas e minúsculas.

Unique rows: Esse *step* é utilizado para eliminar dados que porventura venham duplicados no fluxo de carga, além disso, podem ser configurados contadores para a quantidade de duplicatas encontradas e redirecionamento delas para outro passo. Requer o uso do *step Sort rows* antes deste para ordenação da coluna escolhida.

Database lookup: Esse passo permite a comparação de um ou mais valores vindos de um fluxo de dados com uma ou mais colunas inseridas em uma tabela no banco de dados. Essa comparação, quando verdadeira, retorna uma coluna específica no banco, que pode ser renomeado. Além disso, podem ser feitas configurações para o uso de cache, que em dadas situações podem melhorar a performance da carga. Seu uso nessa ETL será explicado adiante.

Select values: Como explicado na carga anterior, este passo é utilizado para remover colunas, alterar o nome delas bem como seus tipos.

Replace In String: Esse passo permite que o PDI possa substituir um valor de algum campo por outro valor especificado. Esse campo pode ser especificado diretamente ou por meio de uma expressão regular, após, é informado o valor de procura e depois o novo valor, que também pode estar em outro campo.

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 automaticamente para a mesma ser criada.

Primeiramente, são feitas as aquisições dos dados referentes às escolas em duas partes, o primeiro (*step Table input*) procura as escolas públicas e o segundo (*step Table input 2*) procura as escolas privadas conforme a coluna TP_DEPENDENCIA de cada um deles. Caso a escola tiver os códigos 1, 2, 3 ela é considerada uma escola pública por ter dependência nas esferas federal, estadual ou municipal, respectivamente, ou ela possui o código 4 referente a uma escola privada.

Após a pesquisa dos dados tem-se a parte de transformação, em que os próximos passos substituem os valores nulos por dois tipos de indicadores. Na pesquisa de escolas públicas, é feita a substituição dos dados (*step If Field Value is Null*) acerca dos mantedores de escolas privadas e da categoria privada da mesma, já que por ser uma escola pública, esses indicadores não se aplicam a ela e sempre estarão nulos, além da substituição quando essa informação estiver vazia. No outro passo onde é feita a pesquisa de escolas privadas, é feita a substituição de todos os valores nulos (*step If Field Value is Null 2*). Os indicadores nulos dessa dimensão serão explicados adiante.

Com as substituições concluídas, o *Dummy* é utilizado como o *step* que recebe todo esse fluxo de dados para centralizá-los e enviar para o próximo *step* em que é feita a sua ordenação (*step Sort rows*), e após ser ordenado, é feita remoção das escolas duplicadas (*step Unique rows*) para manter uma lista única com todas as escolas envolvidas na análise.

Tendo os dados duplicados removidos, é feita a procura (*step Database lookup*) do código referente a cada combinação de região, distrito, microrregião, mesorregião, UF e município na tabela D_LOCALIDADE_DISTRITO, carregada anteriormente para apoiar essas combinações. Quando uma combinação é encontrada com sucesso, é retornado para o fluxo o código referente a essa combinação.

Com todas as combinações encontradas na dimensão de localidade distrito, é feita a substituição dos dados (*step Replace in string*) de algumas colunas com informações referentes às escolas pelo o seu significado segundo o dicionário de dados do INEP. Por exemplo, a coluna IN_AGUA_INEXISTENTE indica se a escola possui ou não abastecimento de água, no fluxo, os dados existentes nessa coluna são 0 para ‘Não’ e 1 para ‘Sim’, assim, esse *step* faz essa substituição do valor numérico pelo seu valor de significado. Ao finalizar, os dados são mais uma vez ordenados e inseridos na tabela da dimensão escolar.

Como citado anteriormente, essa dimensão possuirá os seguintes indicadores de informação nula:

-1: Caso alguma informação estiver nula.

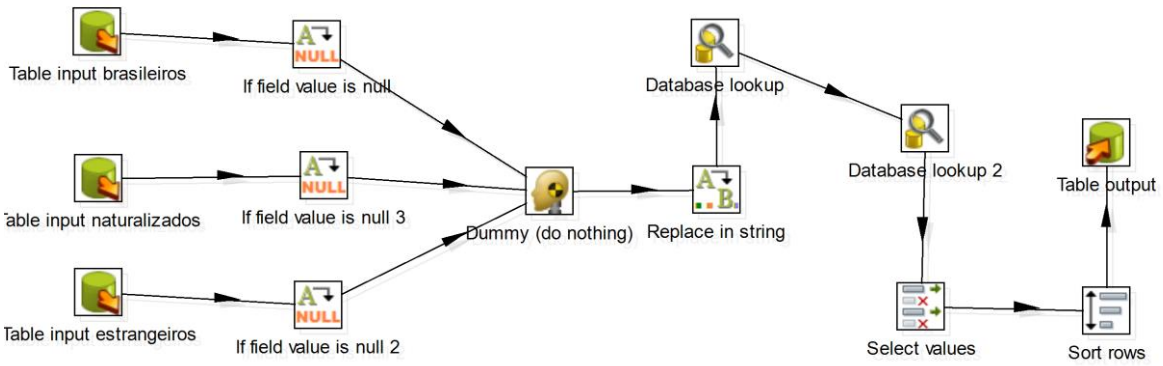
-2: Inserido nas colunas referentes aos mantedores privados e na categoria de escola privada, quando a escola for pública, já que essas duas informações não se aplicam a uma escola que é pública.

4.4.5.5 Fato Aluno

Agora na última tabela do *Data Warehouse*, tem-se a fato aluno que é gerada após todas as dimensões estiverem prontas no BD.

Na sua carga, o processo é semelhante à carga da dimensão escolas, com o diferencial da substituição dos anos da análise por seus respectivos códigos definidos na dimensão ano no momento da carga.

Figura 16 - Diagrama da ETL Aluno



Fonte: Autores (2019).

Os passos estão descritos a seguir:

*Table Input (*na imagem acima com os nomes: Table input brasileiros, Table input naturalizados, Table input estrangeiros): Este *step* permite utilizar os dados já existentes em alguma tabela para fazer outras operações, como a inserção em outro banco, por exemplo. Os dados nesse passo são adquiridos por meio de um comando SQL, mas o *Pentaho* possui uma interface gráfica para selecionar esses dados, sem necessidade do comando, se assim o usuário preferir. A utilização de três passos será explicada adiante.

If Field Value is Null: Aqui nesse passo o *Pentaho* permite inserir valores em campos que estão nulos. É possível escolher um valor padrão ou especificar valores para cada uma das colunas que chegam ao fluxo de dados. Seus três usos nessa ETL serão explicados adiante.

Dummy (do nothing): Esse *step* não realiza ações, mas pode ser utilizado para unir diferentes fluxos de dados ou analisar os dados que estão sendo recebidos. Nessa ETL, ele está juntando os três fluxos de dados para centralizar a inserção.

Replace In String: Esse passo permite que o PDI possa substituir um valor de algum campo por outro valor especificado. Esse campo pode ser especificado diretamente ou por meio de uma expressão regular, após, é informado o valor de procura e depois o novo valor, que também pode estar em outro campo.

Database lookup: Esse passo permite a comparação de um ou mais valores vindos de um fluxo de dados com uma ou mais colunas inseridas em uma tabela no

banco de dados. Essa comparação, quando verdadeira, retorna uma coluna específica no banco, que pode ser renomeado. Além disso, podem ser feitas configurações para o uso de cache, que em dadas situações podem melhorar a performance da carga. Seu uso nessa ETL será explicado adiante.

Select values: Como explicado na carga anterior, este passo é utilizado para remover colunas, alterar o nome delas bem como seus tipos.

Sort rows: Como dito na carga anterior, esse passo possibilita a ordenação de um conjunto de dados com base em uma coluna informada. Seu uso é semelhante ao comando *order by* do SQL. Nele podem ser configuradas outras opções como ordenação ascendente ou descendente e diferenciação de maiúsculas e minúsculas.

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 automaticamente para a mesma ser criada.

Primeiramente, são feitas três pesquisas de dados. A primeira (*step Table input* com o nome ‘Table input brasileiros’) procura os estudantes brasileiros, a segunda (*step Table input* com o nome ‘Table input naturalizados’) procura os estudantes naturalizados, a terceira (*step Table input* com o nome ‘Table input estrangeiros’) procura os estudantes estrangeiros, conforme a coluna CO_PAIS_ORIGEM de cada um deles. Caso o aluno possuir o código 76 nessa coluna, significa que ele tem nacionalidade brasileira/naturalizado (esse é o código do Brasil na tabela de países do INEP). Caso contrário, ele é definido como estrangeiro.

Após a pesquisa dos dados tem-se a parte de transformação, em que os próximos passos (*step If field value is null*) substituem os valores nulos por até três tipos de indicadores, esses indicadores serão explicados adiante.

Com as substituições concluídas, é feita a junção dos três fluxos de dados (*step Dummy (do nothing)*) e envio para o próximo passo (*step Replace in string*) que faz a substituição dos dados relativos aos anos da análise (2015, 2016, 2017 e 2018) pelos seus códigos definidos na tabela dimensão ano (1, 2, 3 e 4, respectivamente), além de indicadores referentes ao sexo do aluno, cor/raça, nacionalidade, entre outros.

Ao ser feitas as substituições, são feitas as comparações das combinações de município, UF e país de origem (tanto como nascimento e endereço), com o seu equivalente na dimensão D_LOCALIDADE_MUNICIPIO (*step Database lookup e Database lookup 2*), essa combinação ao ser verdadeira, retorna o código associado a ela existente na dimensão.

Concluindo esse passo, é feita a remoção das colunas com os códigos das UFs, municípios e país de origem (tanto como nascimento e endereço) para manter apenas o código referente à combinação dessas três informações para que possa ser ligada a dimensão D_LOCALIDADE_MUNICIPIO, após é feita a sua ordenação e finalmente a carga dessas informações na tabela fato dos alunos (*step Table output*).

Como citado anteriormente no *step If field value is null* essa dimensão possuirá os seguintes indicadores de informação nula:

-1: Caso alguma informação estiver nula.

-2: Indicador usado para representar uma informação nula nas colunas de UF e município de um aluno que é estrangeiro.

-3: Indicador usado para representar uma informação nula nas colunas de UF e município de um aluno que é naturalizado.

Nessa última carga são finalizadas todas as dimensões e a fato referente ao ambiente de BI do presente trabalho, com o banco de dados pronto para o próximo passo, a montagem das análises pela ferramenta escolhida.

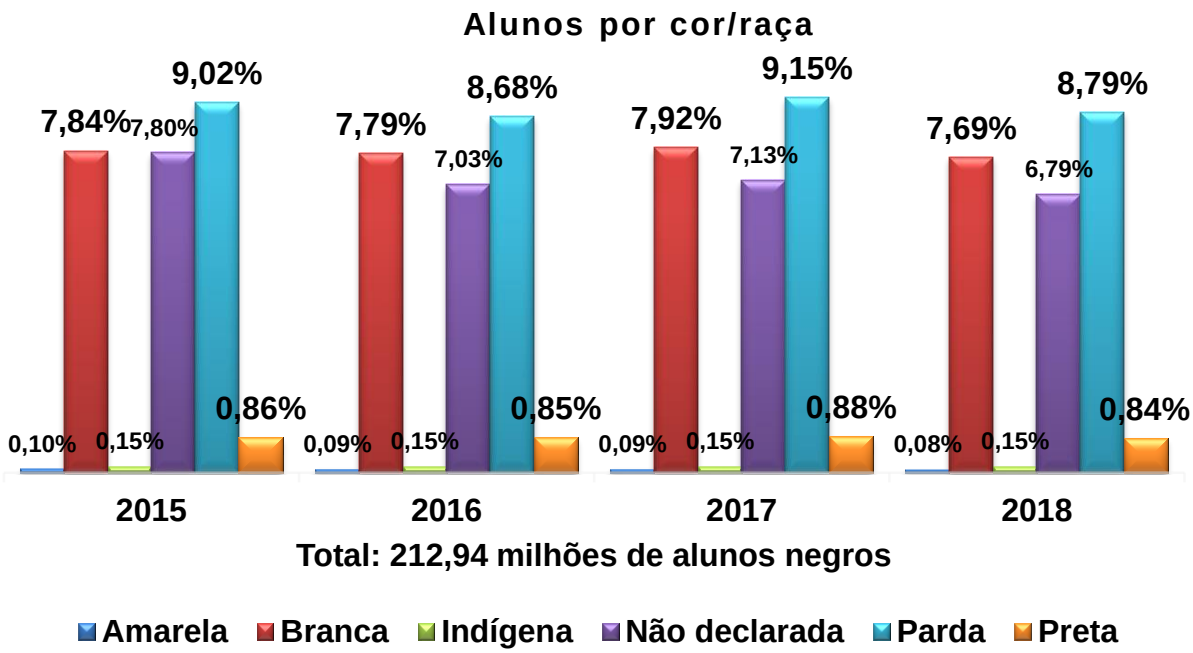
5 RESULTADOS DA ANÁLISE

Ao finalizar todas as ETLs para o *Data Warehouse* e geração dos indicadores pela ferramenta *Power BI* foi possível realizar a análise desses indicadores. Foram gerados quinze indicadores apresentados abaixo, a partir da leitura, interpretação de como eles poderiam ser úteis como indicadores e da bibliografia consultada constante no capítulo 3 deste trabalho.

Este trabalho não tem a pretensão de cobrir todo o assunto, mas pode ser útil para indicar caminhos para outras análises. É importante lembrar que essas análises são afetas a um recorte temporal, a saber, os anos de 2015 a 2018 da Educação Básica dos estados, municípios e distritos brasileiros. Os termos utilizados neste trabalho como “cor/raça”, são provenientes das bases de dados do INEP.

1. Qual o total de alunos por cada Cor/Raça definida pelo Censo Escolar entre os anos da análise?

Figura 17 - Contagem de cor/raça por ano



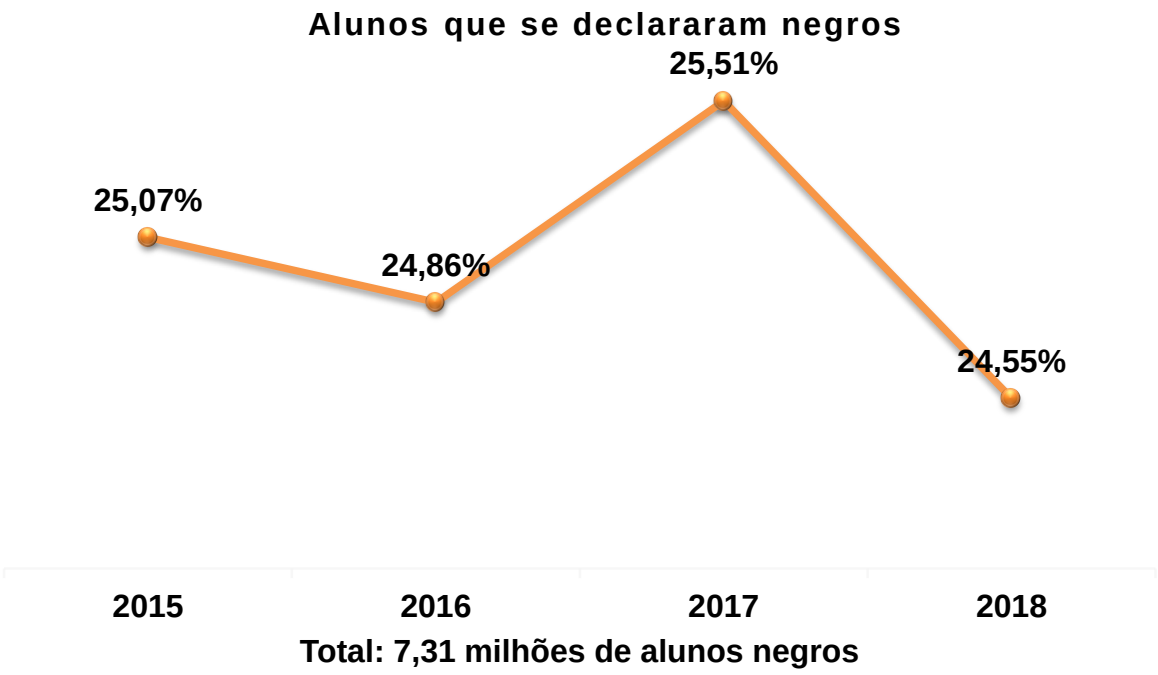
Fonte: Autores (2019).

Segundo o gráfico da figura acima a cor/raça de maior quantidade da base é dos alunos que se consideraram pardos, mantendo quase 10% de alunos entre todos os anos da análise, logo após tem-se a Cor/Raça branca, atingindo quase 8%, além dos

que preferiram não se declarar, com a sua menor quantidade em 2018 onde estiveram com 7,69%. Os negros se mantêm entre 0,85% e 0,88%, sendo a quarta maior Cor/Raça na base do INEP.

2. Qual o total de alunos que se declararam negros entre os anos da análise?

Figura 18 - Contagem de alunos negros por ano



Fonte: Autores (2019).

Segundo o gráfico para o segundo indicador, a quantidade de alunos negros na base do INEP não passou de 26%. Sua menor quantidade foi em 2018 onde se teve apenas 24,55% de alunos que se declararam negros e seu maior pico foi em 2017 tendo 25,51% de alunos com auto declaração negra.

3. Qual o total de alunos estrangeiros que se declararam negros entre os anos da análise?

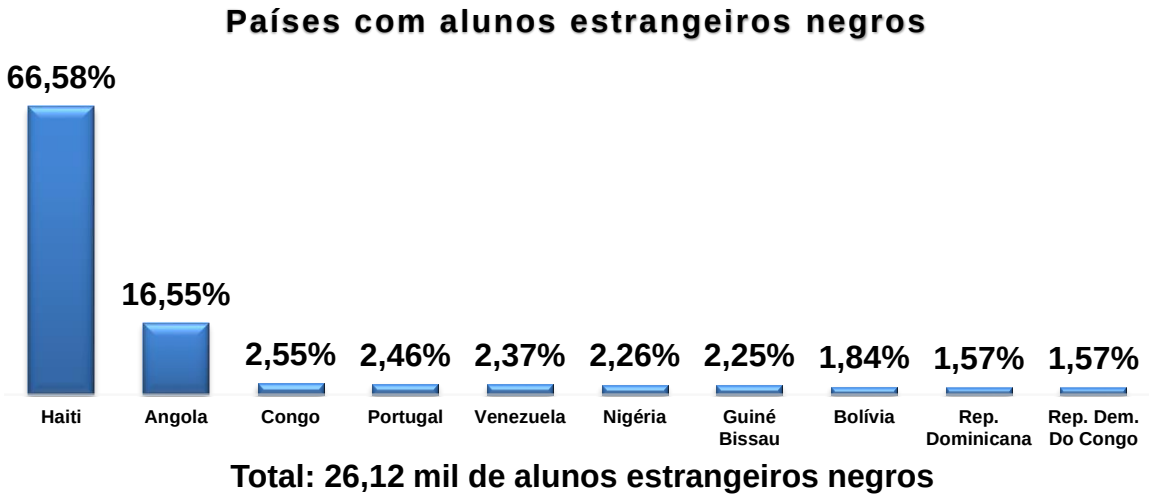
Figura 19 - Contagem de alunos estrangeiros negros por ano

Fonte: Autores (2019).

Para o gráfico referente ao terceiro indicador, é possível notar um aumento desde o início da análise em 2015, de 14% alunos, até o último ano onde chegou a marca de quase 37% de alunos estrangeiros segundo a base do INEP.

4. Qual o país que possui a maior quantidade de alunos estrangeiros negros no Brasil entre os anos da análise?

Figura 20 - Contagem de alunos estrangeiros negros por país (Top 10)



Fonte: Autores (2019).

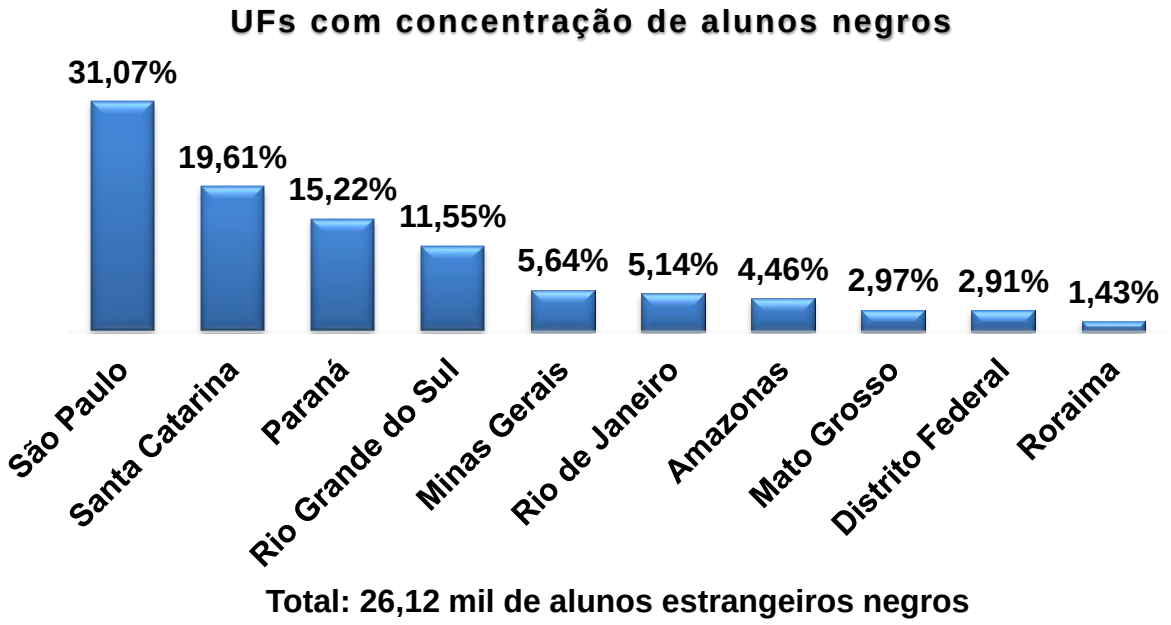
Para o quarto indicador, por questões de visualização, a análise foi reduzida para mostrar apenas os dez primeiros países que mais possuem alunos no Brasil, na

educação básica segundo a base do INEP. O Haiti se mostra o país com a maior quantidade, chegando a quase 67% de alunos, seguido por Angola e Congo. Portugal é o quarto país e os Estados Unidos estão abaixo desses 10 primeiros.

Não faz parte deste trabalho a correlação das missões de paz no Haiti com o fato deste país ser aquele com mais estrangeiros negros na Educação Básica, porém trata-se de um possível estudo futuro.

5. Qual a UF que possui a maior concentração de alunos estrangeiros negros?

Figura 21 - Contagem de alunos estrangeiros negros por UF (Top 10)



Fonte: Autores (2019).

Para o gráfico acima, por questões de visualização, o gráfico foi reduzido para mostrar apenas os 10 primeiros. A UF onde mais se concentram os alunos estrangeiros negros é o estado de São Paulo, seguido por Santa Catarina e Paraná. O Distrito Federal é o nono maior, chegando a quase 3% de registros.

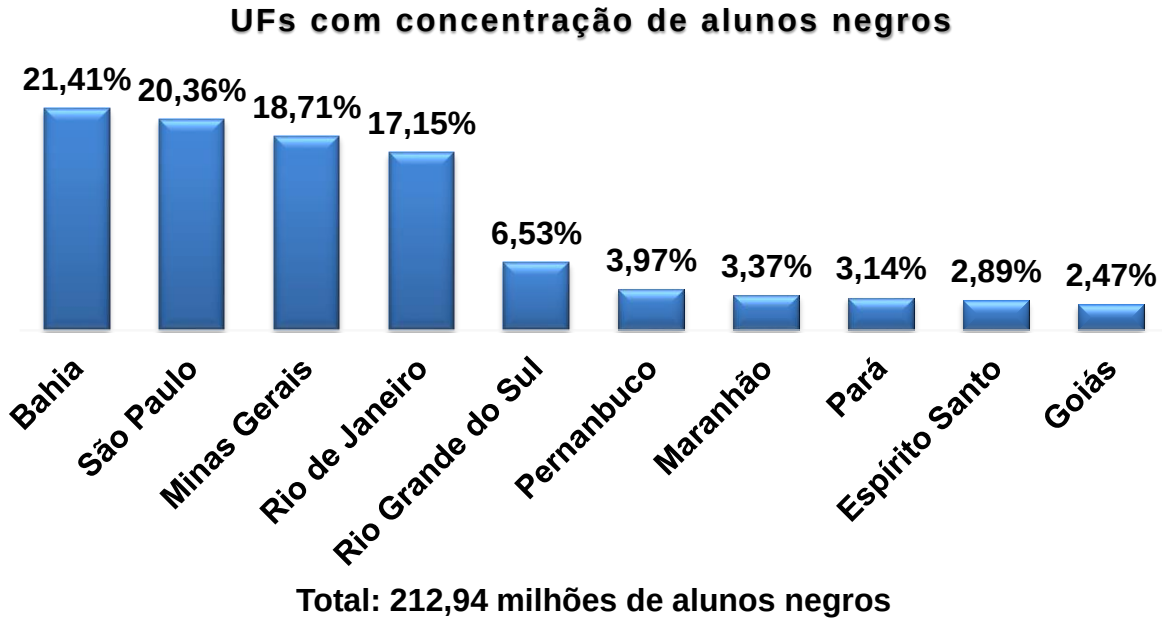
6. Qual o total de alunos negros por região, UF e município entre os anos da análise?

Figura 22 - Contagem de alunos negros por região

Fonte: Autores (2019).

Segundo o gráfico acima, o maior registro de alunos se concentra na região Sudeste onde tem-se quase 50% de dados, seguido logo após pelo Nordeste com quase 32% de registros, o Sul vem depois com quase 9% dos resultados, após o Norte com quase 6% dos resultados, e por último o Centro-Oeste com quase 5% dos resultados.

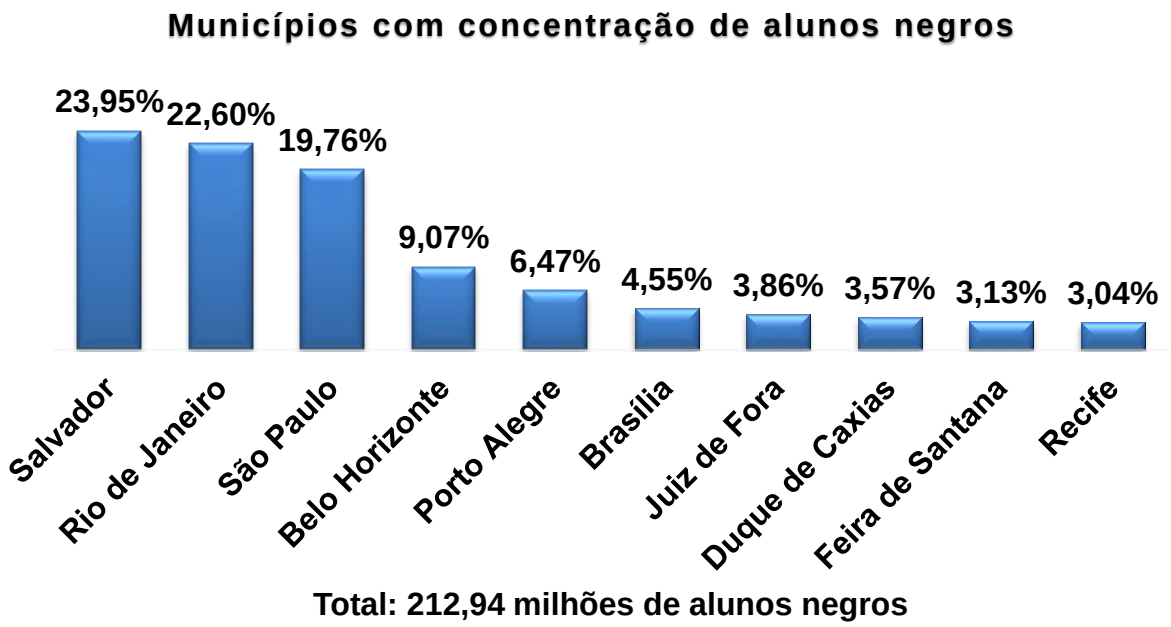
Figura 23 - Contagem de alunos negros por UF (Top 10)



Fonte: Autores (2019).

Segundo o gráfico acima (por questões de visualização foi reduzido para mostrar apenas os 10 primeiros), a maior quantidade de registros dos alunos negros se concentra no estado da Bahia, com quase 21,5% dos resultados. Logo após vem São Paulo com 20,36% e Minas Gerais com quase 19%, fechando os três primeiros. O Distrito Federal não fica entre os 10 primeiros nessa análise.

Figura 24 - Contagem de alunos negros por município (Top 10)



Fonte: Autores (2019).

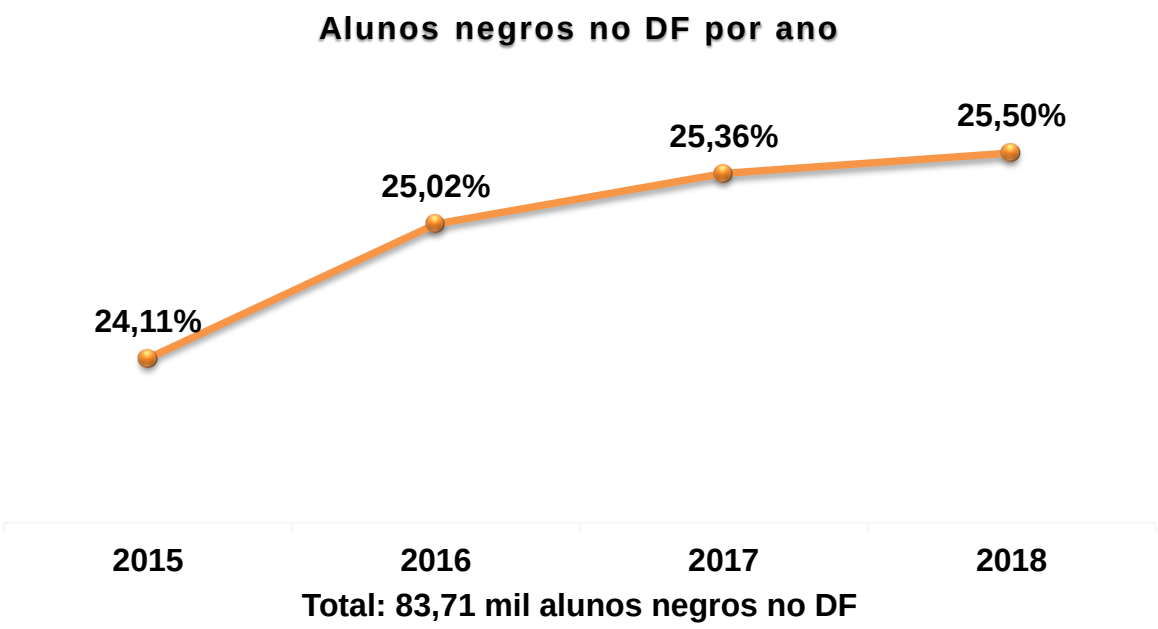
Segundo o gráfico acima (que por questões de visualização, foi reduzido para mostrar apenas os 10 primeiros), o município com a maior concentração de alunos negros é o município de Salvador na Bahia com quase 24% dos alunos, seguido pelo município do Rio de Janeiro com quase 23% e o município de São Paulo com 20% dos resultados. Brasília aparece na análise como o sexto maior município com quase 5% dos alunos (na base do INEP, Brasília, mesmo sendo oficialmente um distrito, possui um código de município para manter a padronização).

7. Qual é a diferença de alunos negros entre as regiões nordeste e sudeste nos anos da análise?

A figura 22 também responde este indicador, demonstrando a diferença de quase 18% entre a região sudeste e nordeste.

8. Qual é a quantidade de alunos negros no Distrito Federal entre os anos da análise?

Figura 25 - Contagem de alunos negros no DF por ano

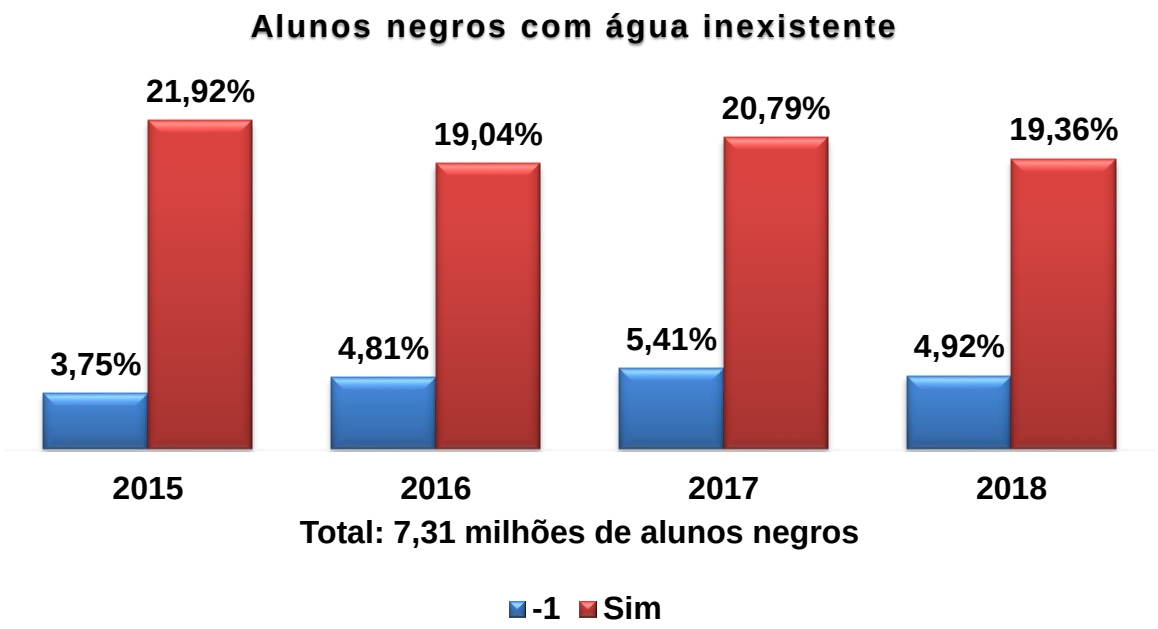


Fonte: Autores (2019).

Em análise ao gráfico, é possível observar que, em média, cerca de 25.675 estudantes negros estão anualmente matriculados em escolas de ensino básico. Segundo dados da Secretária de Estado da Educação do DF, cerca de 450 mil estudantes são atendidos pela Secretária de Educação do Distrito Federal, envolvendo Escolas de Educação Básica, Escolas Parque, Centros Interescolares de Línguas, Centros de Ensino Profissionalizante, além de um Centro de Ensino Médio Integrado (SECRETARIA DE ESTADO DA EDUCAÇÃO, 2018). Em dados levantados pela Codeplan (2014), a população negra do DF corresponde a 57%, porém esta realidade não é transmitida no gráfico, pois a avaliação é feita por auto declaração, sendo vários alunos não se declarando como negros ou nem sequer declaram uma cor/raça.

9. Qual a quantidade de alunos negros que estudam em escolas sem água, energia, esgoto e alimentação entre os anos da análise?

Figura 26 - Contagem de alunos negros com água inexistente nas escolas por ano



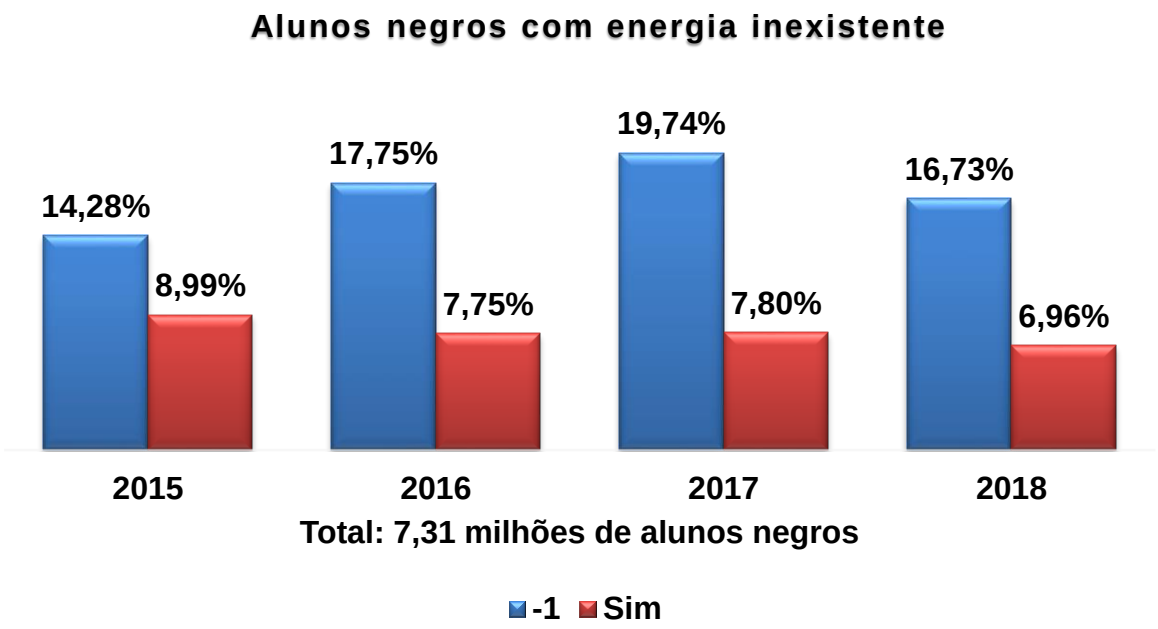
Fonte: Autores (2019).

Significado dos códigos da análise:

-1: Informação desconhecida.

A partir de uma análise do gráfico, é possível identificar que uma parcela de quase 22% dos alunos negros no ano de 2015 não tem acesso a água em suas escolas. Esta quantidade pode estar ligada a fatores geográficos e geopolíticos, pois segundo uma pesquisa realizada pelo jornal O Globo, casos como este, onde há falta de um dos princípios de saneamento básico, são extremados, afetando pincipalmente comunidades pobres e rurais (VASCONCELLOS, RIBEIRO e LINS, 2014). Também consta no gráfico a quantidade de dados inexistentes na base do INEP (representado por “-1”), onde em 2017 teve-se a maior quantidade de dados faltantes, 5,41%.

Figura 27 - Contagem de alunos negros com energia inexistente nas escolas por ano



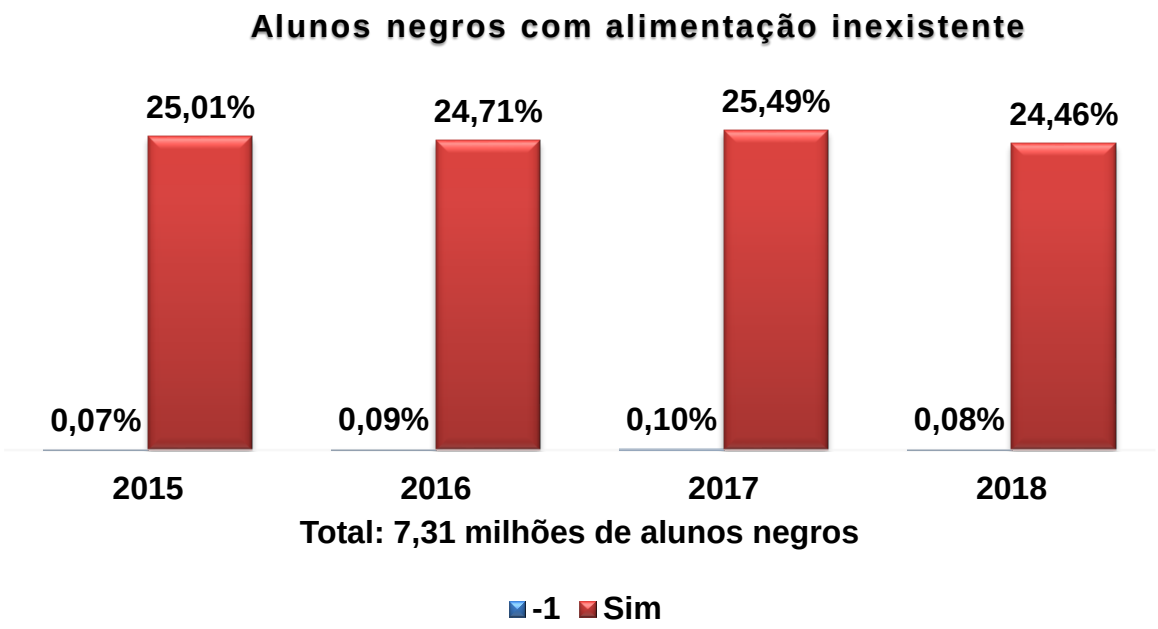
Fonte: Autores (2019).

Significado dos códigos da análise:

-1: Informação desconhecida.

Com base no gráfico, é possível detectar a quantidade de quase 9% de alunos negros no ano de 2015 que frequentam escolas com déficit de energia elétrica. Este número pode ser considerado baixo em relação à quantidade total de alunos negros, pois a partir de programas de desenvolvimento como o Programa Luz para Todos que asseguram a prioridade de desenvolvimento de medidas para o melhoramento do acesso à energia em escolas (PLANO DE DESENVOLVIMENTO DA EDUCAÇÃO, [200-?]). As informações inexistentes passam as informações existentes em todos os anos, chegando ao seu maior pico em 2017 a quase 20% dos registros.

Figura 28 - Contagem de alunos negros com alimentação inexistente nas escolas por ano



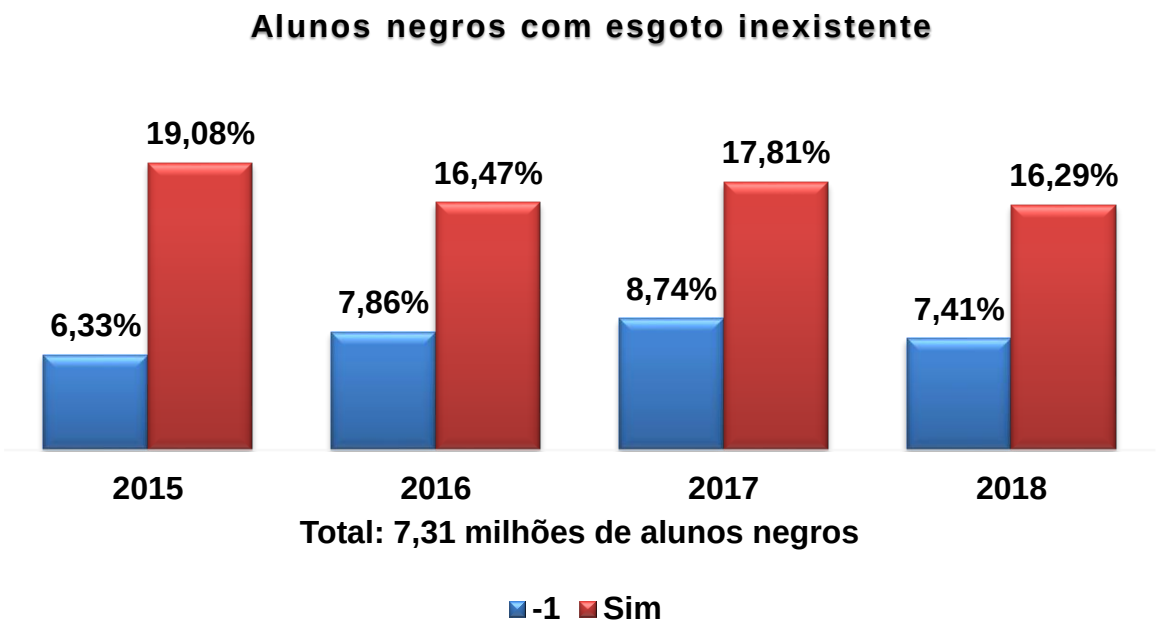
Fonte: Autores (2019).

Significado dos códigos da análise:

-1: Informação desconhecida.

Segundo o gráfico acima, tem-se a informação que em média, 1,6 milhões de alunos negros estudam em escolas sem alimentação, tendo o seu maior pico em 2017, onde teve-se quase 25,5% dos registros. As informações inexistentes não chegam a quase 1%, se mostrando a menor inexistência entre os outros gráficos.

Figura 29 - Contagem de alunos negros com esgoto inexistente nas escolas por ano



Fonte: Autores (2019).

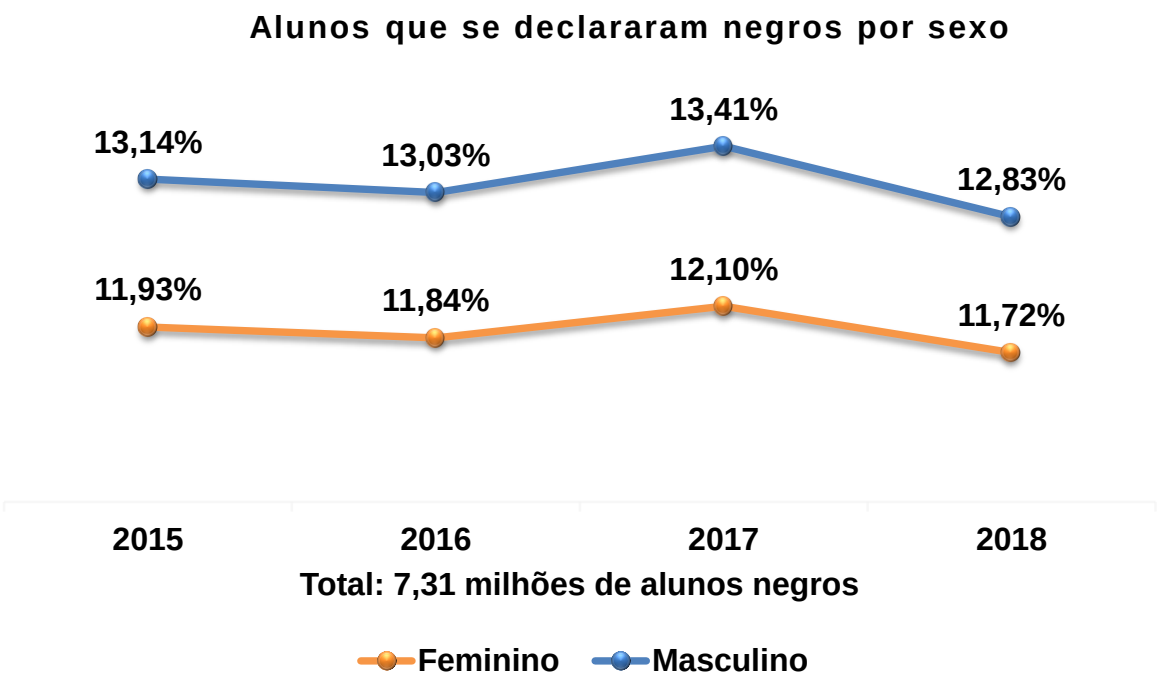
Significado dos códigos da análise:

-1: Informação desconhecida.

Segundo o gráfico acima, em média, 12,7 mil alunos estudam em escolas sem esgoto, com o seu maior pico em 2015 com quase 20% de alunos. Sobre as informações inexistentes, sua maior quantidade foi em 2017, onde teve-se quase 9% de registros faltantes.

10. Qual a quantidade de alunos negros por sexo entre os anos da análise?

Figura 30 - Contagem de alunos por sexo por ano

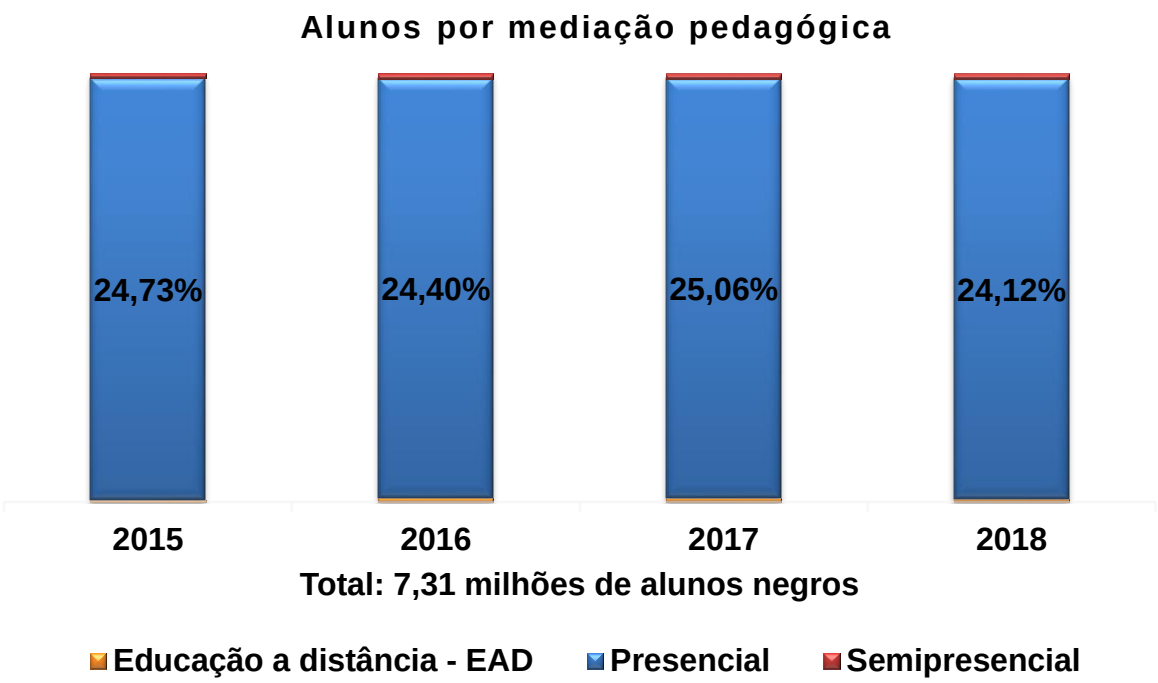


Fonte: Autores (2019).

Segundo o gráfico apresentado, percebe-se uma maior quantidade de alunos negros masculinos, segundo os dados da base do INEP. O maior pico foi em 2017, onde teve-se 13,41% de alunos, contra 12,10% de alunas no mesmo ano. Percebe-se também que nos dois sexos apresentados, essa quantidade foi variando, onde em 2015 teve-se uma quantidade maior que em 2016, que teve uma quantidade menor que em 2017, que teve uma quantidade maior em 2018.

11. Qual a quantidade de alunos negros nos módulos de ensino presencial, semipresencial e a distância entre os anos da análise?

Figura 31 - Contagem de alunos negros por mediação pedagógica por ano

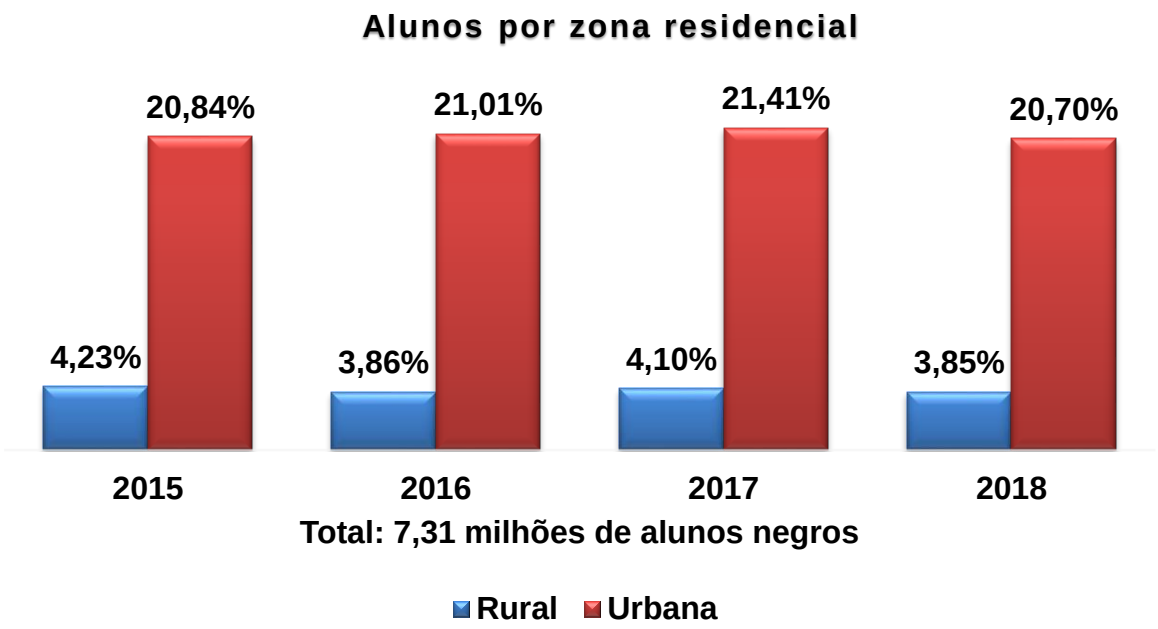


Fonte: Autores (2019).

Segundo o gráfico apresentado, percebe-se a maior inserção dos negros na educação presencial com quase 100% dos dados da base demonstrando isso.

12. Qual a quantidade de alunos negros que moram em zona urbana ou rural entre os anos da análise?

Figura 32 - Contagem de alunos negros por zona residencial por ano

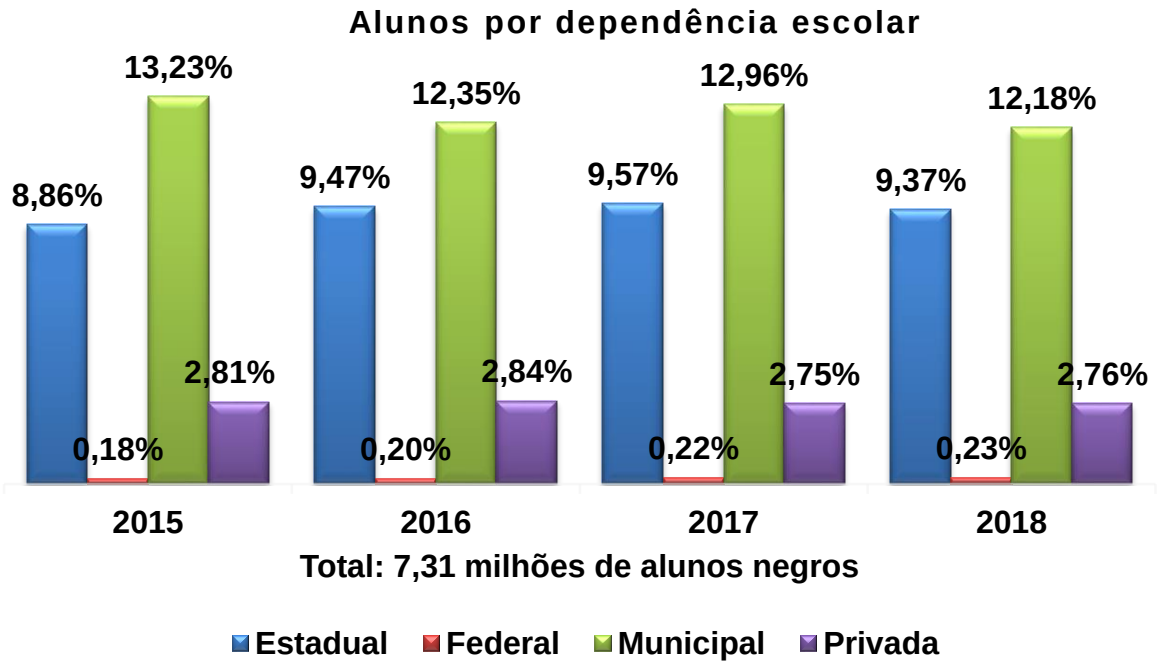


Fonte: Autores (2019).

Partindo de uma análise do gráfico, pode-se observar que grande parte da população de alunos negros se concentra em áreas urbanas, sendo mais exatos quase 83,96% desta amostra, e apenas 16,04% da população negra concentra-se em zonas rurais. Isso pode ser por conta de uma tendência nacional de concentração de população urbana. Segundo uma pesquisa do IBGE, cerca de 84,72% da população brasileira está reunida em centros urbanos e apenas 15,28% está localizada em zonas rurais (IBGE, 2015).

13. Qual a quantidade de alunos negros que estudam em escolas públicas e privadas?

Figura 33 - Contagem de alunos negros por dependência escolar por ano

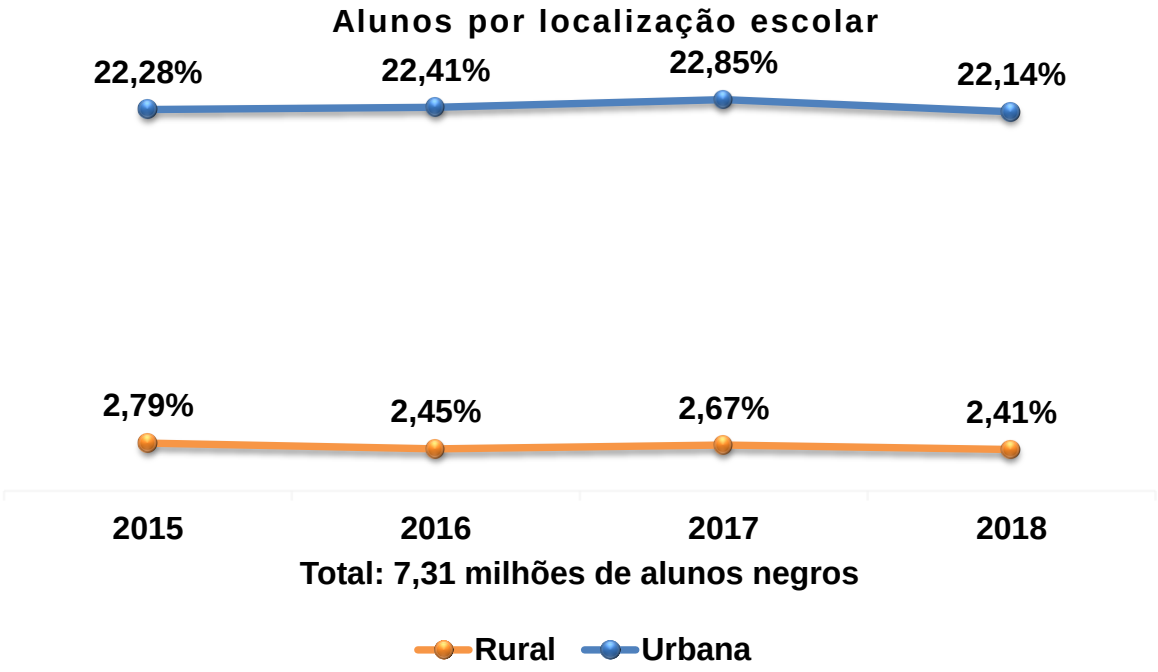


Fonte: Autores (2019).

Segundo o gráfico há uma maior concentração de alunos negros em escolas com dependência municipal, chegando até 13% no ano de 2015. Logo após tem-se a modalidade estadual, com o seu pico atingindo a quase 10% no ano de 2017. As escolas privadas não passaram de 3% de registros em todos os anos da análise.

14. Qual a quantidade de alunos negros que estudam em escolas urbanas e rurais?

Figura 34 - Contagem de alunos negros por localização escolar por ano



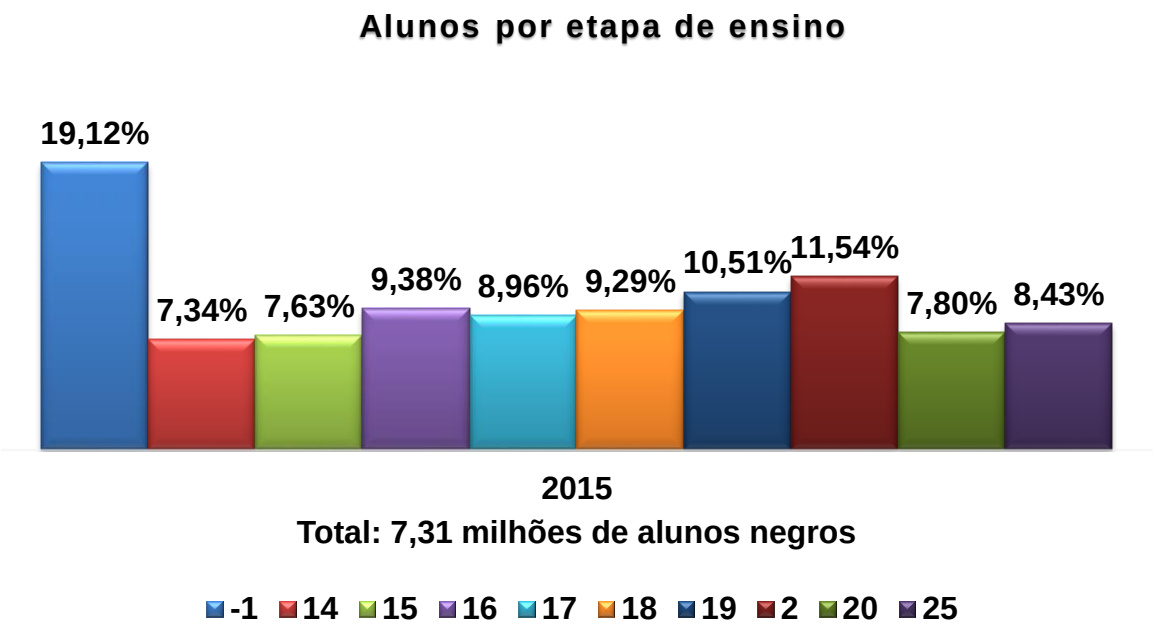
Fonte: Autores (2019).

Segundo os dados apresentados no gráfico, percebe-se uma maior quantidade de alunos negros envolvidos nas escolas de localização urbana, tendo seu maior pico em 2017, onde tem-se quase 23% de alunos nas escolas urbanas, comparado com o maior pico das escolas rurais em 2015, com quase 3% de alunos.

15. Qual a quantidade de alunos negros em cada etapa de ensino definida no censo entre os anos da análise?

Por questões de performance, não foram alterados os códigos referentes a cada uma das etapas de ensino, mas em cada um dos gráficos será descrito o significado dos mesmos segundo o dicionário de dados na base do INEP.

Figura 35 - Contagem de alunos negros por etapa de ensino em 2015 (Top 10)



Fonte: Autores (2019).

Significado dos códigos da análise:

-1: Informação desconhecida;

14: Ensino Fundamental de 9 anos - 1º Ano;

15: Ensino Fundamental de 9 anos - 2º Ano;

16: Ensino Fundamental de 9 anos - 3º Ano;

17: Ensino Fundamental de 9 anos - 4º Ano;

18: Ensino Fundamental de 9 anos - 5º Ano;

19: Ensino Fundamental de 9 anos - 6º Ano;

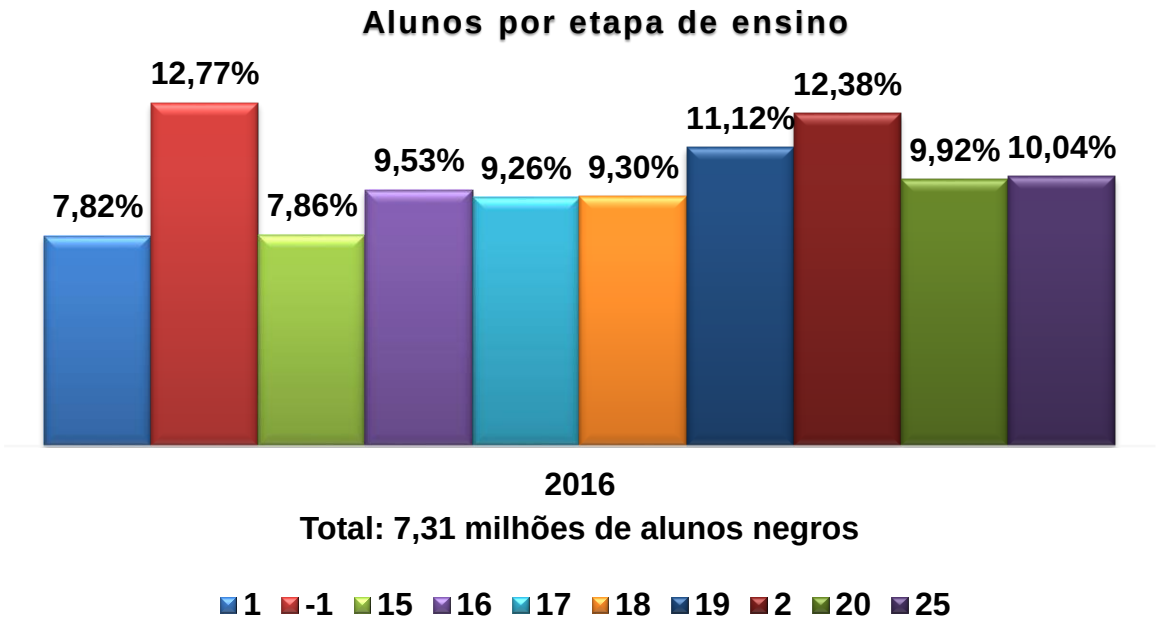
2: Educação Infantil - Pré-escola;

20: Ensino Fundamental de 9 anos - 7º Ano;

25: Ensino Médio - 1ª Série;

Segundo o gráfico acima (por questões de visualização foi reduzido para mostrar apenas os 10 primeiros), o dado em primeiro lugar da análise é o indicador de informação desconhecida, ou seja, na base, por algum motivo, essa informação estava vazia, chegando quase a 20% dos registros. Tirando o indicador nulo, o próximo dado de maior quantidade é o de Educação Infantil – Pré-escola com quase 12% dos registros. Logo após, a etapa de Ensino Fundamental de 9 anos - 6º Ano com quase 11%, e finalizando os três primeiros, a etapa Ensino Fundamental de 9 anos - 3º Ano com quase 10% dos registros.

Figura 36 - Contagem de alunos negros por etapa de ensino em 2016 (Top 10)



Fonte: Autores (2019).

Significado dos códigos da análise:

1: Educação Infantil – Creche;

-1: Informação desconhecida;

15: Ensino Fundamental de 9 anos - 2º Ano;

16: Ensino Fundamental de 9 anos - 3º Ano;

17: Ensino Fundamental de 9 anos - 4º Ano;

18: Ensino Fundamental de 9 anos - 5º Ano;

19: Ensino Fundamental de 9 anos - 6º Ano;

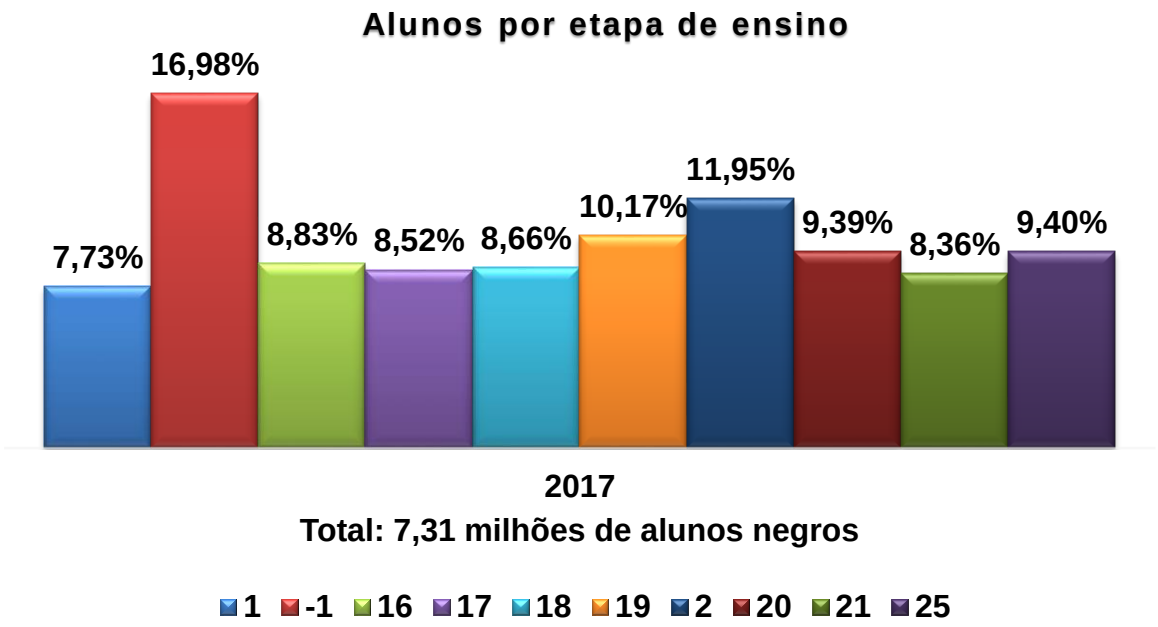
2: Educação Infantil - Pré-escola;

20: Ensino Fundamental de 9 anos - 7º Ano;

25: Ensino Médio - 1ª Série;

Segundo o gráfico acima (por questões de visualização foi reduzido para mostrar apenas os 10 primeiros), o dado em primeiro lugar da análise é o indicador de informação desconhecida, ou seja, na base, por algum motivo, essa informação estava vazia, chegando quase a 13% dos registros. Tirando o indicador nulo, o próximo dado de maior quantidade é o de Educação Infantil – Pré-escola com quase 12,4% dos registros. Logo após, a etapa de Ensino Fundamental de 9 anos - 6º Ano com quase 11,1%, e finalizando os três primeiros, a etapa Ensino Fundamental de 9 anos - 7º Ano com quase 10% dos registros.

Figura 37 - Contagem de alunos negros por etapas de ensino em 2017 (Top 10)



Fonte: Autores (2019).

Significado dos códigos da análise:

1: Educação Infantil – Creche;

-1: Informação desconhecida;

16: Ensino Fundamental de 9 anos - 3º Ano;

17: Ensino Fundamental de 9 anos - 4º Ano;

18: Ensino Fundamental de 9 anos - 5º Ano;

19: Ensino Fundamental de 9 anos - 6º Ano;

2: Educação Infantil - Pré-escola;

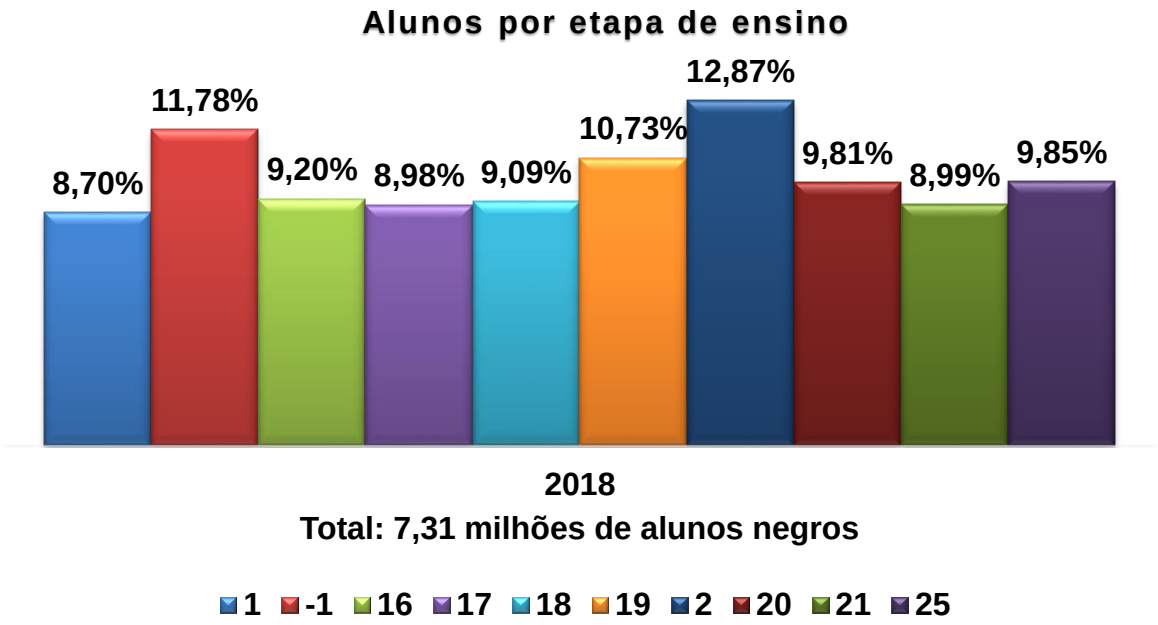
20: Ensino Fundamental de 9 anos - 7º Ano;

21: Ensino Fundamental de 9 anos - 8º Ano;

25: Ensino Médio - 1ª Série;

Segundo o gráfico acima (por questões de visualização foi reduzido para mostrar apenas os 10 primeiros), o dado em primeiro lugar da análise é o indicador de informação desconhecida, ou seja, na base, por algum motivo, essa informação estava vazia, chegando quase a 17% dos registros. Tirando o indicador nulo, o próximo dado de maior quantidade é o de Educação Infantil – Pré-escola com quase 12% dos registros. Logo após, a etapa de Ensino Fundamental de 9 anos - 6º Ano com 10,17% dos registros, e finalizando os três primeiros, a etapa Ensino Médio - 1ª Série com 9,40% dos registros.

Figura 38 - Contagem de alunos negros por etapa de ensino em 2018 (Top 10)



Fonte: Autores (2019).

Significado dos códigos da análise:

1: Educação Infantil – Creche;

-1: Informação desconhecida;

16: Ensino Fundamental de 9 anos - 3º Ano;

17: Ensino Fundamental de 9 anos - 4º Ano;

18: Ensino Fundamental de 9 anos - 5º Ano;

19: Ensino Fundamental de 9 anos - 6º Ano;

2: Educação Infantil - Pré-escola;

20: Ensino Fundamental de 9 anos - 7º Ano;

21: Ensino Fundamental de 9 anos - 8º Ano;

25: Ensino Médio - 1ª Série;

Segundo o gráfico acima (por questões de visualização foi reduzido para mostrar apenas os 10 primeiros), o dado em primeiro lugar da análise é a etapa Educação Infantil - Pré-escola, com quase 13% dos registros. O indicador nulo, diferentemente dos gráficos anteriores desse indicador, vem em segundo lugar com quase 12% dos registros. Tirando o indicador nulo, tem-se a etapa Ensino Fundamental de 9 anos - 6º Ano, com quase 11%. Por último, fechando os três primeiros (desconsiderando o indicador nulo), tem-se a etapa Ensino Médio - 1ª Série com quase 10% dos registros.

6 CONSIDERAÇÕES FINAIS

Este estudo possibilitou uma análise da inserção do aluno negro na educação básica brasileira demonstrando a utilidade e toda a abordagem de *Business Intelligence*.

Foram analisados quase 200 milhões de alunos envolvidos na base do INEP segundo o recorte temporal de 2015 a 2018, o que, em média seria 50 milhões de alunos para cada um dos anos da análise. Para os alunos negros, tem-se, em média, 1,8 milhões de registros em cada um dos anos, finalizando um total de 7 milhões de alunos negros.

De maneira geral, o *Business Intelligence* disponibiliza aos usuários formas claras de ver os dados, a partir de visualizações gráficas limpas e bem estruturadas. Desde o início teve-se o foco em demonstrar e implementar uma aplicação de BI tradicional por inteira. A implantação de um projeto de BI não é simples e neste caso não foi diferente, havendo uma longa fase de planejamento para a correta estruturação e carga dos dados.

Uma das maiores dificuldades no decorrer do trabalho foi a performance da carga dos dados, onde a cada erro constatado, era necessária a completa limpeza no *Data Warehouse*, procurar em qual parte da carga teve-se o erro e finalmente realizar a nova carga, custando muito ao computador. No começo, teve-se muitos problemas de falta de memória por causa das cargas feitas, esse problema foi contornado pelo uso das cargas incrementais, onde as cargas eram separadas por ano ou por região, isso facilitou, inclusive, a correção de erros que as cargas poderiam apresentar.

Este estudo ofereceu-nos a oportunidade de integrar todos os conhecimentos adquiridos no curso de Ciência da Computação, tais como a vivência das matérias de Lógica de Programação, Bancos de Dados e Tópicos de Atuação Profissional. O uso das ferramentas de BI foram essenciais para o fechamento do processo.

Por fim, com este estudo é possível entender o processo de Business Intelligence, bem como funcionalidades e características; os processos de ETL com o uso de uma ferramenta *Open Source*, o *Pentaho*; comparações entre as duas abordagens e modelos utilizados pelos autores da área. Mas como tudo que envolve

tecnologia, tende a evoluir, novas tecnologias para o BI surgirão. Também é possível, a partir deste estudo, contextualizar a inserção do aluno negro na educação básica brasileira de maneira geral nos anos de 2015 a 2018 com o auxílio das análises.

6.1 Limitações e Trabalhos Futuros

O trabalho possui certas limitações que abrem espaço para melhorias e trabalhos futuros, como o desenvolvimento de uma página *web*/*software mobile* ou *desktop* para a visualização dessas análises; implementação de um processo de *Machine Learning/Deep Learning* para a previsão, conforme os dados da base, de quantos alunos negros a base pode receber nos próximos anos; desenvolvimento de *Data Marts* para o foco das análises em mais indicadores; expandir o ambiente para unir os dados de todos os censos. O trabalho não cobre nenhum desses itens citados anteriormente, e como dito, abre espaço para uma grande melhoria.

6.2 Revisão dos objetivos alcançados

Sobre os objetivos específicos citados na seção 1.4.2, todos eles foram alcançados com sucesso, onde:

1. Foi possível levantar o estado da arte no que tange Business Intelligence, sua metodologia e processos no capítulo 2.
2. Foram apresentados indicadores sobre a atuação do aluno negro na educação brasileira que podem ser úteis em futuros trabalhos para essa área na seção 4.4.4.
3. Foi possível aplicar a metodologia de Business Intelligence, os processos de ETL e a montagem do ambiente de Data Warehouse no capítulo 4, onde foi criado todo um ambiente de *Business Intelligence* que possibilitou as análises.
4. Foram desenvolvidos os resultados das análises através da ferramenta de BI escolhida, onde foram gerados gráficos e visualizações dos dados.

REFERÊNCIAS

ALMEIDA, M. A. B. D.; SANCHEZ, L. REVEDUC. **Os negros na legislação educacional e educação formal no Brasil**, 2016. Disponivel em: <http://www.reveduc.ufscar.br/index.php/reveduc/article/view/1459/500>. Acesso em: 19 set. 2019.

ANTONELLI, R. A. TECAP. **Conhecendo o Business Intelligence (BI)**, Paraná, v. III, p. 79-85, 2009.

BRAGHITTONI, R. **Business Intelligence - Implementar do jeito certo e a custo zero**. 1ª. ed. São Paulo: Casa do Código, 2017.

CARVALHAES, M. H. F.; ALVES, A. L. **Estruturando o Business Intelligence através do processo de Data Warehouse**. Pontifícia Universidade Católica de Goiás. Goiânia, p. 11. 2015.

CODEPLAN. **A população negra do Distrito Federal**, 2014. Disponivel em: <http://www.codeplan.df.gov.br/wp-content/uploads/2018/02/Popula%C3%A7%C3%A3o-Negra-no-Distrito-Federal-Analisando-as-Regi%C3%B5es-Administrativas.pdf>. Acesso em: 03 dez. 2019.

DEVENS, R. M. **Cyclopædia of commercial and business anecdotes**. Estados Unidos: [s.n.], v. 1, 1865.

ÉPOCA NEGÓCIOS. **Power BI: muito além de Business Intelligence**, 2018. Disponivel em: <https://epocanegocios.globo.com/Publicidade/Microsoft/noticia/2018/03/power-bi-muito-alem-de-business-intelligence.html>. Acesso em: 04 set. 2019.

FONSECA, M. V. D. et al. Ação Educativa. **Negro e educação: Presença do negro no Sistema Educacional Brasileiro**, 2001. Disponivel em: <http://acaoeducativa.org.br/relacoesraciais/wp-content/uploads/2013/12/Negro-Educa%C3%A7%C3%A3o-1-INEP.pdf>. Acesso em: 19 set. 2019.

FREITAS, P. GitHub. **Geodata BR**, 2018. Disponivel em: <https://github.com/paulofreitas/geodata-br/tree/master/data/pt>. Acesso em: 26 set. 2019.

GUIMARÃES, L. Know Solution. **Saiba o que é Pentaho e por que escolhemos trabalhar com ele**, 2015. Disponivel em: <https://www.knowsolution.com.br/saiba-o-que-e-pentaho-e-por-que-escolhemos-trabalhar-com-ele/>. Acesso em: 04 set. 2019.

IBGE. **Estatísticas de gênero**, 2010. Disponivel em: <https://www.ibge.gov.br/apps/snig/v1/notas_metodologicas.html?loc=0>. Acesso em: 23 out. 2019.

IBGE. IBGE Educa. **População rural e urbana**, 2015. Disponivel em: <https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18313-populacao-rural-e-urbana.html>. Acesso em: 03 dez. 2019.

INEP. **História**, 2019. Disponivel em: <http://inep.gov.br/historia>. Acesso em: 19 set. 2019.

INMON, W. H. **Building the Data Warehouse**. 4ª. ed. Estados Unidos: John Wiley & Sons, 2005.

KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit: The definitive guide to dimensional modeling**. 3ª. ed. Estados Unidos: John Wiley & Sons, 2013.

KUMAR, R. **Machine Learning and cognition in enterprises: Business Intelligence transformed**. 1ª. ed. India: Apress, v. I, 2017.

LUHN, H. P. IBM Journal of Research and Development. **A Business Intelligence System**, Estados Unidos, v. II, n. 2, p. 314-319, Outubro 1958.

MICROSOFT. **O que é Power BI?**, 2019. Disponivel em: <https://docs.microsoft.com/pt-br/power-bi/power-bi-overview>. Acesso em: 04 set. 2019.

MICROSOFT. **Descrição do serviço Power BI**, 2019. Disponivel em: <https://docs.microsoft.com/pt-br/office365/servicedescriptions/power-bi-service-description>. Acesso em: 04 set. 2019.

MINISTÉRIO DA EDUCAÇÃO. Portal MEC. **Institucional**, 04 Setembro 2019. Disponivel em: <http://portal.mec.gov.br/institucional>. Acesso em: 19 set. 2019.

OLIVEIRA, F. D. Revista Educação e Políticas em Debate. **A Educação Básica e o tratamento da questão racial: As implicações da Lei 10.639 para a formação de professores**, Minas Gerais, v. 2, n. 1, p. 53-75, Janeiro/Julho 2013.

OLIVEIRA, I. D. Ministério Público do Estado do Rio de Janeiro. **O negro no Sistema Educacional Brasileiro: Alguns aspectos históricos**, 2018. Disponivel em: <https://www.mprj.mp.br/documents/20184/167086/apresentacao_iolanda_oliveira.pdf>. Acesso em: 19 set. 2019.

OLIVEIRA, V. Pentaho - Visão Geral. **BI na prática**. Disponivel em: <https://www.binapratica.com.br/visao-pentaho>. Acesso em: 04 set. 2019.

PANOLY. Data Warehouse Guide. **Data Mart vs. Data Warehouse**, 2019. Disponivel em: <https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/>. Acesso em: 18 out. 2019.

PASSOS, J. C. D. Secretaria da Educação do Paraná. **As desigualdades educacionais, a população negra e a educação de jovens e adultos**, 2010. Disponivel em: <http://www.educadores.diaadia.pr.gov.br/arquivos/File/pacto_nacional_em/artigos/desigualdades_educacionais_eja.pdf>. Acesso em: 19 set. 2019.

PLANO DE DESENVOLVIMENTO DA EDUCAÇÃO. **Programa Luz para Todos levará energia elétrica a todas as escolas públicas do país**, [200-?]. Disponivel em: <http://portal.mec.gov.br/arquivos/Bk_pde/luz.html>. Acesso em: 03 dez. 2019.

PORTAL BRASILEIRO DE DADOS ABERTOS. **Perguntas mais frequentes**, 2019. Disponivel em: <http://dados.gov.br/pagina/faq>. Acesso em: 19 set. 2019.

PORTAL BRASILEIRO DE DADOS ABERTOS. **O que são dados abertos?**, 2019. Disponivel em: <http://dados.gov.br/pagina/dados-abertos>. Acesso em: 19 set. 2019.

PRASAD, K. V. K. K. **Data Warehouse development tools - covering informatica, cognos, business objects and datastage with case studies**. 1ª. ed. Estados Unidos: Dreamtech Press, v. I, 2007.

PRIMAK, F. V. **Decisoes com B.I. - Business Intelligence**. 1ª. ed. Rio de Janeiro: Ciência Moderna, 2008.

SECRETARIA DE ESTADO DA EDUCAÇÃO. GDF. **Quantas escolas existem na rede pública de ensino do distrito federal?**, 2018. Disponivel em: <http://www.se.df.gov.br/unidades-escolares/>. Acesso em: 03 dez. 2019.

SILVA, E. L. D.; MENEZES, E. M. **Metodologia da pesquisa e elaboração da dissertação**. Universidade Federal de Santa Catarina. Florianópolis, p. 139. 2005.

THOMSEN, E. **OLAP solutions: Building multidimensional information systems**. 2ª. ed. Estados Unidos: John Wiley & Sons, 2002.

TURBAN, E.; SHARDA, R.; KING, D. **Business Intelligence: um enfoque gerencial para a inteligência de negócio**. 1ª. ed. Porto Alegre: Bookman, v. I, 2009.

VASCONCELLOS, F.; RIBEIRO, E.; LINS, L. O Globo. **Brasil tem 30% de suas escolas sem abastecimento de água**, 2014. Disponivel em: <https://oglobo.globo.com/sociedade/brasil-tem-30-de-suas-escolas-sem-abastecimento-de-agua-12315236>. Acesso em: 03 dez. 2019.

W3SCHOOLS. SQL NULL Values. **What is a NULL value?**, 2019. Disponivel em: <www.w3schools.com/sql/sql_null_values.asp>. Acesso em: 26 nov. 2019.

ZANDONA, E. P. ANPED. **Desigualdades raciais na trajetória escolar de alunos do negros do Ensino Médio**, 2008. Disponivel em: <http://www.anped.org.br/sites/default/files/gt21-4566-int.pdf>. Acesso em: 19 set. 2019.