

**Documentação Projeto Educação**

# São Paulo, 18 de março de 2021

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CONTROLE DE VERSÃO** | | |
| **Autor** | **Versão** | **Data** | **Descrição** |
| Talita Mesquita | 1.0 | 18/03/2021 | Criação do documento. |
| Talita Mesquita | 1.1 | 22/03/2021 | Inserção de informações |
| Felipe Zamberlam | 1.2. | 01/04/2021 | Alteração de Fontes |
| Talita Mesquita/Felipe Zamberlam | 1.3 | 07/04/2021 | Finalização |

# Introdução

O documento visa estabelecer e apresentar os aspectos técnicos sobre o projeto integrado ‘Educação – o Brasil e as metas do PNE’, que foi realizado a pedido da Blueshift do Brasil como forma de treinamento e objetivando colocar em prática os conhecimentos adquiridos durante o período de estudo do Blueshift Academy com o intuito de possibilitar a oportunidade de atuar com dados reais e favorecer o trabalho em equipe em projetos de engenharia de dados.

# Solicitação

O projeto realizado tem como objetivo analisar a educação no Brasil, tendo como parâmetro principal as metas educacionais estabelecidas pelo Plano Nacional de Educação (PNE) em junho de 2014 e que têm validade de 10 anos. Portanto, todas as metas estabelecidas devem ser atingidas até 2024, ou antes mesmo desta data de acordo com as próprias metas. O objetivo do projeto é analisar o quão perto do cumprimento de tais metas no prazo estipulado o país está. Ao todo, o PNE estabeleceu 20 metas que englobam todos os níveis de ensino (da educação básica ao ensino superior) e que buscam a equidade e a qualidade da educação em todo o país. Dessas 20 metas, 15 metas serão avaliadas no projeto, já que para 5 delas não há dados suficientes para fazer a análise ou mesmo apenas demonstrar aspectos gerais sobre o objetivo delas. O ensino básico regular no Brasil é essencialmente dividido em educação infantil, ensino fundamental e ensino médio. A educação infantil atende crianças de 0 a 5 anos e compreende creche e pré-escola. O ensino fundamental é dividido em anos iniciais, que compreende do 1º ao 5º ano, e anos finais, que compreende do 6º ao 9º ano – e atende crianças e adolescentes dos 6 aos 14 anos. O ensino médio é composto por 1º, 2º e 3º anos e atende adolescentes de 15 a 17 anos. Além disso existem outras modalidades de ensino, como EJA (Educação de Jovens e Adultos) para indivíduos que não puderam cursar o ensino fundamental/médio na idade prevista e ensino médio técnico e profissionalizante.

A entrega do projeto foi realizada em duas partes. Primeiramente foi realizado o processo de extração e transformação das bases de dados para conseguir tabelas organizadas com as informações relevantes ao atingimento das metas. Posteriormente, foram realizadas a análise e a apresentação dos dados encontrados para a compreensão de quão perto o Brasil está de atingir as metas no tempo indicado em cada uma delas.

A primeira parte da entrega do projeto consistiu na utilização de dois recursos: Azure Databricks e uma camada de armazenamento (*Blobs*) na nuvem Azure. Foram produzidos scripts em Python (que contém também comandos dbutils[[1]](#footnote-1) e %sh[[2]](#footnote-2)), usando notebooks do Databricks para fazer o processo de ETL. Ao total, foram produzidos 11 notebooks. Em 4 desses, a extração dos dados foi realizada via API, usando uma URL construída especificamente para a extração de cada tabela necessária, no formato JSON. Nos outros 7 notebooks a extração dos dados foi feita diretamente das URLs que os continham, em formato .zip[[3]](#footnote-3). Em todos os notebooks foi usado o conceito de Delta Lake para o armazenamento dos Dataframes em formato ‘delta’, com e sem transformações. Foram criados diretórios de armazenamento em pastas (bronze/prata/ouro) no armazenamento *Blob*, integrados ao HDFS do Databricks para acesso aos dados. Os Dataframes foram primeiramente salvos na camada bronze para armazenamento, sem transformações ou com pequenas transformações nos dados – somente para que todos tivessem o mesmo esquema. Em segundo lugar, foram feitas transformações em todos os Dataframes e sua unificação para salvar na camada prata. Finalmente, foram realizadas as transformações necessárias antes de utilizar a camada ouro como repositório final e da criação de tabelas no banco de dados do Databricks para alimentação do Power BI. Todas as tabelas seguindo o esquema necessário para análise na ferramenta.

A segunda parte da entrega do projeto consiste em carregar as tabelas estruturadas na plataforma Power BI e produzir dashboards com o intuito de visualizar os dados de maneira a verificar o atingimento, ou não, das metas dentro do prazo estabelecido. Todas as tabelas, com exceção de uma, foram extraídas diretamente do banco de dados do Databricks para alimentar o Power BI. A tabela de exceção foi salva no armazenamento *Blob* e depois foi feita a conexão com o Power BI devido a seu tamanho[[4]](#footnote-4).

As metas do PNE estabelecidas em 2014 são as seguintes:

1. Universalizar, até 2016, a educação infantil na pré-escola para as crianças de 4 (quatro) a 5 (cinco) anos de idade e ampliar a oferta de educação infantil em creches, de forma a atender, no mínimo, 50% (cinquenta por cento) das crianças de até 3 (três) anos até o final da vigência deste PNE.
2. Universalizar o ensino fundamental de 9 (nove) anos para toda a população de 6 (seis) a 14 (quatorze) anos e garantir que pelo menos 95% (noventa e cinco por cento) dos alunos concluam essa etapa na idade recomendada, até o último ano de vigência deste PNE.
3. Universalizar, até 2016, o atendimento escolar para toda a população de 15 (quinze) a 17 (dezessete) anos e elevar, até o final do período de vigência deste PNE, a taxa líquida de matrículas no ensino médio para 85% (oitenta e cinco por cento).
4. Universalizar, para a população de 4 (quatro) a 17 (dezessete) anos com deficiência, transtornos globais do desenvolvimento e altas habilidades ou superdotação, o acesso à educação básica e ao atendimento educacional especializado, preferencialmente na rede regular de ensino, com a garantia de sistema educacional inclusivo, de salas de recursos multifuncionais, classes, escolas ou serviços especializados, públicos ou conveniados.
5. Alfabetizar todas as crianças, no máximo, até o final do 3º (terceiro) ano do ensino fundamental.
6. Oferecer educação em tempo integral em, no mínimo, 50% (cinquenta por cento) das escolas públicas, de forma a atender, pelo menos, 25% (vinte e cinco por cento) dos(as) alunos(as) da educação básica.
7. Fomentar a qualidade da educação básica em todas as etapas e modalidades, com melhoria do fluxo escolar e da aprendizagem, de modo a atingir as seguintes médias nacionais para o IDEB (Índice de Desenvolvimento da Educação Básica): 6,0 nos anos iniciais do ensino fundamental; 5,5 nos anos finais do ensino fundamental; 5,2 no ensino médio.
8. Elevar a escolaridade média da população de 18 (dezoito) a 29 (vinte e nove) anos, de modo a alcançar, no mínimo, 12 (doze) anos de estudo no último ano de vigência deste Plano, para as populações do campo, da região de menor escolaridade no País e dos 25% (vinte e cinco por cento) mais pobres, e igualar a escolaridade média entre negros e não negros declarados à Fundação Instituto Brasileiro de Geografia e Estatística (IBGE).
9. Elevar a taxa de alfabetização da população com 15 (quinze) anos ou mais para 93,5% (noventa e três inteiros e cinco décimos por cento) até 2015 e, até o final da vigência deste PNE, erradicar o analfabetismo absoluto e reduzir em 50% (cinquenta por cento) a taxa de analfabetismo funcional.
10. Oferecer, no mínimo, 25% (vinte e cinco por cento) das matrículas de educação de jovens e adultos, nos ensinos fundamental e médio, na forma integrada à educação profissional.
11. Triplicar as matrículas da educação profissional técnica de nível médio, assegurando a qualidade da oferta e pelo menos 50% (cinquenta por cento) da expansão no segmento público.
12. Elevar a taxa bruta de matrícula na educação superior para 50% (cinquenta por cento) e a taxa líquida para 33% (trinta e três por cento) da população de 18 (dezoito) a 24 (vinte e quatro) anos, assegurada a qualidade da oferta e expansão para, pelo menos, 40% (quarenta por cento) das novas matrículas, no segmento público.
13. Elevar a qualidade da educação superior e ampliar a proporção de mestres e doutores do corpo docente em efetivo exercício no conjunto do sistema de educação superior para 75% (setenta e cinco por cento), sendo, do total, no mínimo, 35% (trinta e cinco por cento) doutores.
14. Elevar gradualmente o número de matrículas na pós-graduação stricto sensu, de modo a atingir a titulação anual de 60.000 (sessenta mil) mestres e 25.000 (vinte e cinco mil) doutores.
15. Garantir, em regime de colaboração entre a União, os Estados, o Distrito Federal e os Municípios, no prazo de 1 (um) ano de vigência deste PNE, política nacional de formação dos profissionais da educação de que tratam os incisos I, II e III do caput do art. 61 da Lei nº 9.394, de 20 de dezembro de 1996, assegurado que todos os professores e as professoras da educação básica possuam formação específica de nível superior, obtida em curso de licenciatura na área de conhecimento em que atuam.
16. Formar, em nível de pós-graduação, 50% (cinquenta por cento) dos professores da educação básica, até o último ano de vigência deste PNE, e garantir a todos(as) os(as) profissionais da educação básica formação continuada em sua área de atuação, considerando as necessidades, demandas e contextualizações dos sistemas de ensino.
17. Valorizar os(as) profissionais do magistério das redes públicas de educação básica de forma a equiparar seu rendimento médio ao dos(as) demais profissionais com escolaridade equivalente, até o final do sexto ano de vigência deste PNE.
18. Assegurar, no prazo de 2 (dois) anos, a existência de planos de Carreira para os(as) profissionais da educação básica e superior pública de todos os sistemas de ensino e, para o plano de Carreira dos(as) profissionais da educação básica pública, tomar como referência o piso salarial nacional profissional, definido em lei federal, nos termos do inciso VIII do art. 206 da Constituição Federal.
19. Assegurar condições, no prazo de 2 (dois) anos, para a efetivação da gestão democrática da educação, associada a critérios técnicos de mérito e desempenho e à consulta pública à comunidade escolar, no âmbito das escolas públicas, prevendo recursos e apoio técnico da União para tanto.
20. Ampliar o investimento público em educação pública de forma a atingir, no mínimo, o patamar de 7% (sete por cento) do Produto Interno Bruto (PIB) do País no 5º (quinto) ano de vigência desta Lei e, no mínimo, o equivalente a 10% (dez por cento) do PIB ao final do decênio.

O projeto visa analisar as metas estabelecidas, com exceção das metas 6, 14, 17, 18 e 19 que não foram analisadas por falta de dados.

# Premissas da solução

## Origem e especificação dos dados

Os dados foram retirados de plataformas governamentais das seguintes organizações que possuem política de acesso à informação e divulgam dados abertos para o público: IBGE (Instituto Brasileiro de Geografia e Estatística) e INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira).

Os dados gerais sobre a população do Brasil foram extraídos por meio da utilização da API do Banco SIDRA (Sistema IBGE de Recuperação Automática) via construção de uma URL que definiu a tabela a ser consumida pelo notebook Databricks em formato JSON. A tabelas retiradas do banco SIDRA, e suas respectivas URLs, foram as seguintes:

* Tabela 6407 - População residente, por sexo e grupos de idade: <http://api.sidra.ibge.gov.br/values/t/6407/n1/1/v/606/p/2017-2019/c2/all/c58/1140,104868,114535,2793,1144,1145,3299,3300,3301,3302/f/n?formato=json>.
* Tabela 7113 – Taxa de analfabetismo das pessoas de 15 anos ou mais de idade, por sexo e grupo de idade: <http://api.sidra.ibge.gov.br/values/t/7113/n1/1/v/10267/p/2017-2019/c2/all/c58/2795/f/n?formato=json>.
* Tabela 7127 – Número médio de anos de estudo das pessoas de 15 anos ou mais, por cor ou raça e grupo de idade: <http://api.sidra.ibge.gov.br/values/t/7127/n1/1/v/3593/p/2017-2019/c86/all/c58/2792,100052,108866/f/n?formato=json>.
* Tabela 7140 – Crianças de 0 a 5 anos de idade, por grupo de idade e frequência à creche ou escola: <http://api.sidra.ibge.gov.br/values/t/7140/n1/1/v/10278/p/2017-2019/c58/all/c12081/all/f/n?formato=json>.

Os dados sobre a educação básica, a educação superior, os investimentos públicos em educação e as notas do IDEB foram retirados de repositórios públicos em formato .zip, a partir das seguintes URLs:

* Censo escolar (educação básica). Os dados de cada ano foram retirados das seguintes URLs:
* <https://download.inep.gov.br/microdados/micro_censo_escolar_2017.zip> (2017)
* <https://download.inep.gov.br/microdados/microdados_educacao_basica_2018.zip> (2018)
* <https://download.inep.gov.br/microdados/microdados_educacao_basica_2019.zip> (2019)
* Censo (educação superior). Os dados de cada ano foram retirados das seguintes URLs
* <https://download.inep.gov.br/microdados/microdados_educacao_superior_2017.zip> (2017)
* <https://download.inep.gov.br/microdados/microdados_educacao_superior_2018.zip> (2018)
* <https://download.inep.gov.br/microdados/microdados_educacao_superior_2019.zip> (2019)
* Investimentos públicos em educação: <https://download.inep.gov.br/informacoes_estatisticas/investimentos_publicos_em_educacao/indicadores_financeiros_educacionais/investimento_pib_total.zip>.
* IDEB: <http://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2019/divulgacao_brasil_ideb_2019.zip>

## Ambiente de desenvolvimento

A empresa Blueshift disponibilizou todos os acessos aos recursos pagos usados durante o projeto. Os ambientes de desenvolvimento utilizados para realização desse projeto foram provisionados na nuvem Azure e foram:

* Databricks[[5]](#footnote-5):
* *Workspace*;
* *Notebooks*[[6]](#footnote-6);
* *Clusters*[[7]](#footnote-7);
* Banco de dados ‘*default*’.
* Armazenamento de *Blobs*.
* Power BI desktop.

A conta de armazenamento foi provisionada da seguinte maneira:

Nome: m03storage

Nome dos containeres de armazenamento: m03container e csvmatricula

Camada de acesso: Standard/Quente

Tipo de conta: StorageV2 (uso geral v2)

Replicação: LRS (armazenamento com redundância local)

A plataforma Databricks foi provisionada da seguinte maneira:

Tipo de preço: padrão

Nome: grupom03databricks

Para a utilização do *workspace* do Databricks e criação dos notebooks foram criados dois usuários: [talita.mesquita@blueshift.com.br](mailto:talita.mesquita@blueshift.com.br) e [felipe.takahasi@blueshift.com.br](mailto:felipe.takahasi@blueshift.com.br).

Além disso, para execução dos notebooks foram utilizados dois *clusters* Databricks:

**Projeto\_Educacao:**

Modo: nó unico

Versão Runtime: 7.5 (includes Apache Spark 3.0.1, Scala 2.12)

Tipo de nó (driver): Standard\_DS3\_v2 (14.0 GB Memory, 4 Cores, 0.75 DBU)

Configuração Spark: spark.databricks.delta.preview.enabled true / spark.databricks.cluster.profile singleNode / spark.master local[\*]

Variáveis de ambiente: PYSPARK\_PYTHON=/databricks/python3/bin/python3

**pi-cluster-dev:**

Modo: nó unico

Versão Runtime: 7.5 (includes Apache Spark 3.0.1, Scala 2.12)

Tipo de nó (driver): Standard\_DS3\_v2 (14.0 GB Memory, 4 Cores, 0.75 DBU)

Configuração Spark: spark.databricks.cluster.profile singleNode / spark.master local[\*]

Variáveis de ambiente: PYSPARK\_PYTHON=/databricks/python3/bin/python3

E para a alimentação do Power BI[[8]](#footnote-8) foi utilizado o seguinte *cluster* Databricks:

**Projeto\_Integrado:**

Modo: padrão

Versão Runtime: 7.5 (includes Apache Spark 3.0.1, Scala 2.12)

Tipo de nó (driver): Standard\_DS3\_v2 (14.0 GB Memory, 4 Cores, 0.75 DBU)

Tipo de nó (worker): Standard\_DS3\_v2 (14.0 GB Memory, 4 Cores, 0.75 DBU) – mínimo: 1 / máximo: 5

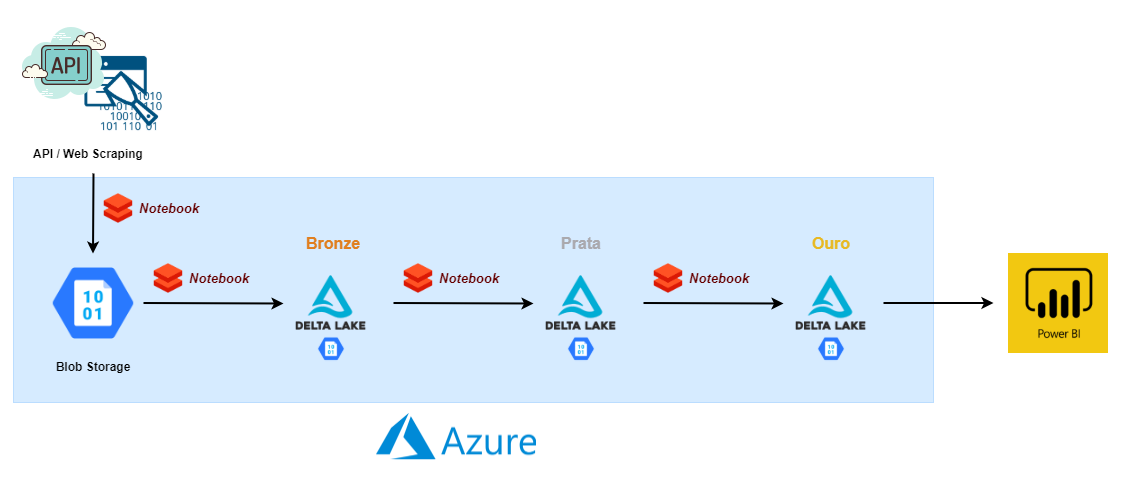
Configuração Spark: No Spark configuration set

Variáveis de ambiente: PYSPARK\_PYTHON=/databricks/python3/bin/python3

Todos os recursos foram provisionados usando o grupo de recursos **ProjetoIntegrado** na assinatura **Assinatura do Visual Studio Enterprise – MPN** e na localização **sul do Brasil**.

# Modelo da arquitetura sugerida

A figura abaixo apresenta o modelo da arquitetura sugerida, na qual os dados foram extraídos e transformados de modo a serem apresentados da melhor forma para visualização das metas do PNE.



1. Utilização da linguagem Python em todos os *notebooks* do Azure Databricks utilizados, além de comandos dbutils e %sh.
2. Extração dos dados diretamente da URL em formato .zip.
3. Extração de dados via API no formato JSON.
4. Criação da conexão entre o recurso de armazenamento e o Databricks.
5. Extração, transformação dos dados e armazenamento dos mesmos no Delta Lake, usando como diretório pastas (bronze/prata/ouro) criadas dentro do armazenamento de *Blobs* e integradas ao armazenamento do Databricks.
6. Criação da conexão do Databricks e do *Blob* ao Power BI.
7. Alimentação do Power BI para a elaboração do dashboard.

# Extração e transformações APIs – linguagem Python

Foram criados 4 *notebooks* em Python no Azure Databricks para fazer a extração e a transformação dos dados presentes em cada tabela do IBGE utilizada (6407, 7113, 7127 e 7140). A extração de todas as tabelas foi feita por meio da API do SIDRA, com a qual foram criadas URLs para a extração referentes a cada tabela.

Após a extração dos dados, as tabelas foram salvas na camada bronze do Delta Lake para posterior leitura e transformação destes. As camadas prata e ouro foram usadas como repositório dos dados já transformados. Ao salvar os dados na camada ouro foi criada uma tabela no banco de dados do Databricks, para a alimentação do Power BI.

Os respectivos *notebooks*, com todas as transformações realizadas podem ser visualizados no GitHub (<https://github.com/FZamberlam/BlueShiftAcademyPI>).

# Extração e transformação dos repositórios (URLs) – linguagem Python

Foram criados 7 *notebooks* em Python no Azure Databricks com o intuito de extrair arquivos das URLs no formato .zip e armazena-los primeiramente na camada bronze da maneira que foram extraídos e fazer, posteriormente, transformações nos dados. As camadas prata e ouro foram usadas como repositório dos dados já transformados. Ao salvar os dados na camada ouro foi criada uma tabela no banco de dados do Databricks, para a alimentação do Power BI.

Os arquivos .zip do censo escolar de educação básica foram primeiramente armazenados no *Blob* sem definição de Dataframes, visando melhorar a performance dos *notebooks* por conta do tamanho dos arquivos. Além disso, ao final das transformações, o Dataframe do censo escolar de educação básica referente às matrículas de alunos foi salvo primeiramente no *Blob* para posterior alimentação do Power BI.

Os respectivos *notebooks*, com todas as transformações realizadas podem ser visualizados no GitHub (<https://github.com/FZamberlam/BlueShiftAcademyPI>).

# Ingestão no Power BI

As tabelas foram importadas no Power BI por meio da conexão JDBC/ODBC do *cluster* Projeto\_Integrado do Azure Databricks. Para a criação da conexão das tabelas ao Power BI foi necessária a criação de um token pessoal (validade de 90 dias) no Databricks para autenticação desta conexão no Power BI. Além disso, ao importar as tabelas para o Power BI, as seguintes informações são necessárias: hostname[[9]](#footnote-9) do servidor e caminho HTTP[[10]](#footnote-10).

Após a criação da conexão do Azure Databricks ao Power BI, as tabelas foram importadas neste para a criação do dashboard. Foi criada também a conexão do *Blob* ao Power BI para a importação da tabela de matrículas do censo escolar de educação básica. Para tanto, foram, necessárias as seguintes informações: nome[[11]](#footnote-11) da conta e chave de acesso[[12]](#footnote-12).

## Modelagem tabelas

As tabelas foram modeladas da seguinte maneira:

### PNAD 6407

Foi executado script Python no Azure Databricks para a extração e a transformação dos dados da tabela PNAD 6407, que contém informações do IBGE sobre a quantidade de pessoas residentes no Brasil. A extração foi feita por meio do uso da API do Sidra (<http://api.sidra.ibge.gov.br>), buscando dados dos anos de 2017 a 2019. Foram efetuadas transformações visando conseguir a seguinte estrutura:

|  |
| --- |
| **populacao** |
| Ano (data) |
| Grupo\_de\_Idade (texto) |
| Sexo (texto) |
| Valor (número inteiro) |

### PNAD 7113

Foi executado script Python no Azure Databricks para a extração e a transformação dos dados da tabela PNAD 7113, que contém informações do IBGE sobre a taxa de analfabetismo de pessoas com 15 anos ou mais residentes no Brasil. A extração foi feita por meio do uso da API do Sidra (<http://api.sidra.ibge.gov.br>), buscando dados dos anos de 2017 a 2019. Foram efetuadas transformações visando conseguir a seguinte estrutura e tipos de dados:

|  |
| --- |
| **analfabetismo** |
| Ano (data) |
| Grupo\_de\_Idade (texto)  Sexo (texto) |
| Valor (número decimal) |

### PNAD 7127

Foi executado script Python no Azure Databricks para a extração e a transformação dos dados da tabela PNAD 7127, que contém informações do IBGE sobre o número médio de anos de estudo da população brasileira, levando em conta a cor/raça do residente no Brasil. A extração foi feita por meio do uso da API do Sidra (<http://api.sidra.ibge.gov.br>), buscando dados dos anos de 2017 a 2019. Foram efetuadas transformações visando conseguir a seguinte estrutura e tipos de dados:

|  |
| --- |
| **Anos\_estudo** |
| Ano (data) |
| Cor\_raca (texto) |
| Grupo\_de\_Idade (texto) |
| Valor (número decimal) |

### PNAD 7140

Foi executado script Python no Azure Databricks para a extração e a transformação dos dados da tabela PNAD 7140, que contém informações do IBGE sobre a frequentação ou não de escolas e creches por crianças com idades de 0 a 5 anos residentes no Brasil. A extração foi feita por meio do uso da API do Sidra (<http://api.sidra.ibge.gov.br>), buscando dados dos anos de 2017 a 2019. Foram efetuadas transformações visando conseguir a seguinte estrutura e tipos de dados:

|  |
| --- |
| **Criancas\_0\_a\_5** |
| Ano (data) |
| Grupo\_de\_Idade (texto) |
| Frequenta\_Escola\_Creche (texto) |
| Valor (número inteiro) |

### Censo Escolar – Educação Básica

Foi executado script Python no Azure Databricks para a extração e a transformação dos dados das tabelas de censo escolar da educação básica dos anos de 2017 a 2019 por meio da URL de acesso à informação (microdados) do INEP (<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar>). Foram efetuadas transformações visando conseguir duas tabelas com a seguinte estrutura e tipos de dados:

|  |
| --- |
| **Censo Escolar - Docentes** |
| Ano (data) |
| Idade\_Referencia (número inteiro) |
| Idade (número inteiro) |
| Sexo (texto) |
| Cor\_Raca (texto) |
| Nacionalidade (texto) |
| Zona\_Residencial (texto) |
| Escolaridade (texto) |
| Situacao\_Curso\_1 (texto) |
| Area\_Curso\_1 (texto) |
| Curso\_1 (texto) |
| Licenciatura\_Curso\_1 (texto) |
| Situacao\_Curso\_2 (texto) |
| Area\_Curso\_2 (texto) |
| Curso\_2 (texto) |
| Licenciatura\_Curso\_2 (texto) |
| Situacao\_Curso\_3 (texto) |
| Area\_Curso\_3 (texto) |
| Curso\_3 (texto) |
| Licenciatura\_Curso\_3 (texto) |
| Especializacao (texto) |
| Mestrado (texto) |
| Doutorado (texto) |
| Pos\_Graduacao (texto) |
| Funcao (texto) |

|  |
| --- |
| **Censo Escolar – Matrículas (csv)** |
| Ano (data) |
| Idade (número inteiro) |
| Sexo (texto) |
| Cor\_Raca (texto) |
| Nacionalidade (texto) |
| Necessidade\_Especial (texto) |
| Etapa\_Ensino (texto) |
| Especial\_Exclusiva (texto) |
| Ensino\_Regular (texto) |
| Ensino\_EJA (texto) |
| Ensino\_Profissionalizante (texto) |

### Censo - Educação Superior

Foi executado script Python no Azure Databricks para a extração e a transformação dos dados das tabelas do censo da educação superior dos anos de 2017 a 2019 por meio da URL de acesso à informação (microdados) do INEP (<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>). Foram efetuadas transformações visando conseguir duas tabelas com a seguinte estrutura e tipos de dados:

|  |
| --- |
| **Censo Superior - docentes** |
| Ano (data) |
| Categoria\_Administrativa (texto) |
| Organizacao\_Academica (texto) |
| Situacao (texto) |
| Escolaridade (texto) |
| Regime\_Trabalho (texto) |
| Cor\_Raca (texto) |
| Sexo (texto) |
| Idade (número inteiro) |
| Grupo\_de\_Idade (texto) |
| Nacionalidade (texto) |
| Tem\_Deficiencia (texto) |

|  |
| --- |
| **Censo Superior - alunos** |
| Ano (data) |
| Categoria\_Administrativa (texto) |
| Organizacao\_Academica (texto) |
| Turno (texto) |
| Grau\_Academico (texto) |
| Modalidade\_Ensino (texto) |
| Nivel\_Academico (texto) |
| Cor\_Raca (texto) |
| Sexo (texto) |
| Idade (número inteiro) |
| Grupo\_de\_Idade (texto) |
| Nacionalidade (texto) |
| Tem\_Deficiencia (texto) |
| Situacao (texto) |
| Ingressou\_por\_Vestibular (texto) |
| Ingressou\_por\_Enem (texto) |
| Apoio\_Social (texto) |

### IDEB

Foi executado script Python no Azure Databricks para a extração e a transformação dos dados da planilha do IDEB dos anos de 2017 a 2019 (com projeção de 2021) por meio da URL de acesso à informação do INEP (<https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>). Foram efetuadas transformações visando conseguir uma tabela com a seguinte estrutura e tipos de dados:

|  |
| --- |
| **IDEB** |
| Nivel\_Escolaridade (texto) |
| Rede (texto) |
| Ano (data) |
| Valor (número decimal) |

### Investimentos financeiros em educação

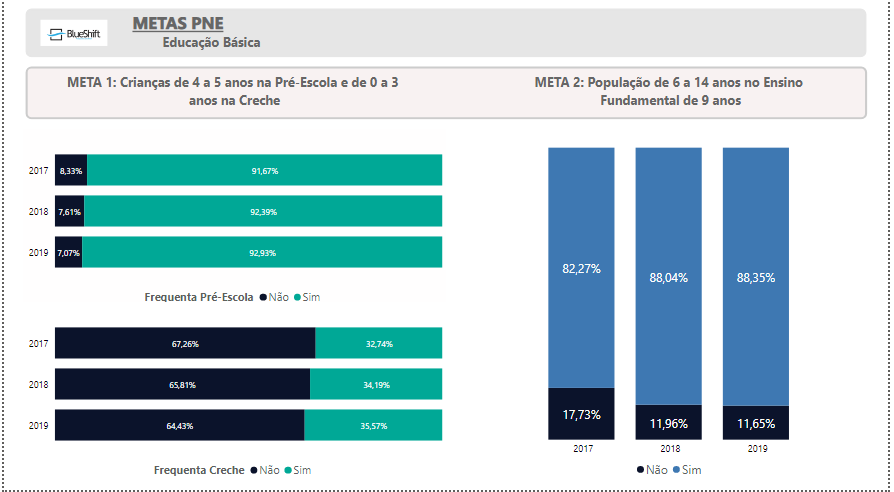
Foi executado script Python no Azure Databricks para a extração e a transformação dos dados da planilha de indicadores financeiros educacionais dos anos de 2017 a 2019[[13]](#footnote-13) por meio da URL de acesso à informação do INEP (<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-financeiros-educacionais>). A tabela de investimentos traz informações sobre o percentual de investimento público total em educação em relação ao PIB. Percentual este separado por nível educacional. Foram efetuadas transformações visando conseguir uma tabela com a seguinte estrutura e tipos de dados:

|  |
| --- |
| **Investimentos** |
| Ano (data) |
| Todos\_niveis (número decimal) |
| Ed.Basica (número decimal) |
| Ed.Infantil (número decimal) |
| E.F.Anos\_iniciais (número decimal) |
| E.F.Anos\_finais (número decimal) |
| Ensino\_Medio (número decimal) |
| Ed.Superior (número decimal) |

# Dashboards[[14]](#footnote-14)

Foram produzidos dashboards com o intuito de analisar os dados de 2017 a 2019 sobre a educação no Brasil, visando verificar quão perto de atingir as metas do PNE o país está. Os respectivos dashboards são apresentados abaixo:

1. Educação Básica\_1 – metas 1 e 2



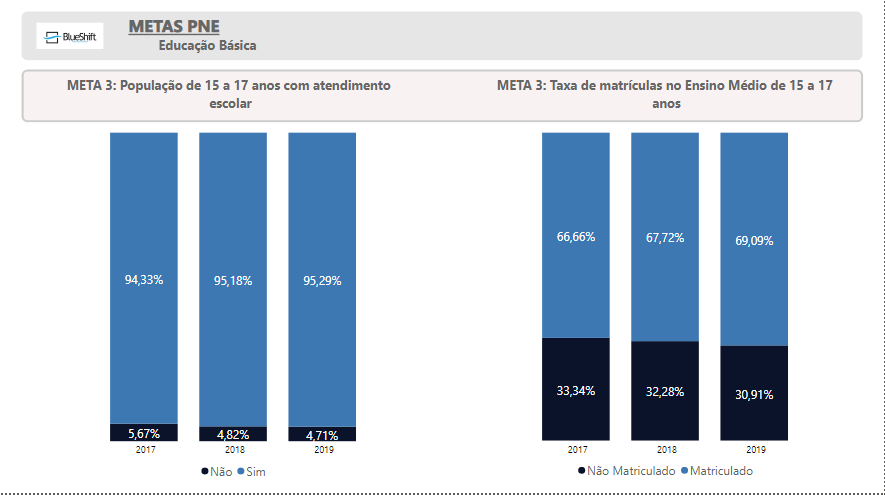
No dashboard acima são apresentadas as metas 1 e 2 referentes a educação básica no Brasil.

A meta 1 visa universalizar até 2016, a educação na pré-escola para crianças de 4 a 5 anos de idade e, ampliar até 2024 a oferta da educação em creches, atendendo no mínimo 50% das crianças de 0 a 3 anos de idade.

Para ambas as métricas da meta 1 foram utilizados filtros capturando os grupos de idade dessas crianças (0 a 1 ano e 2 a 3 anos para creche e 4 a 5 anos para pré-escola) e se frequentam ou não a escola/creche. A meta foi exibida em gráficos de barras 100% empilhadas a fim de exibir o percentual de adesão da educação e em ambos se percebe que estão distantes de alcançar a meta estabelecida pelo PNE.

Para a meta 2, onde estabelece que até 2024 seja universalizado o ensino fundamental de 9 anos para jovens de 6 a 14 anos, foram criadas duas medidas com dados da tabela de censo escolar (matriculas), identificando alunos que estão matriculados neste novo modelo e alunos que estão matriculados no modelo antigo ou até mesmo que não estão na escola. E nesta meta percebemos que houve uma diminuição considerável de alunos fora desta categoria de 2017 para 2018, porém ainda está distante de atingir os 100% necessários.

1. Educação Básica\_2 – meta 3

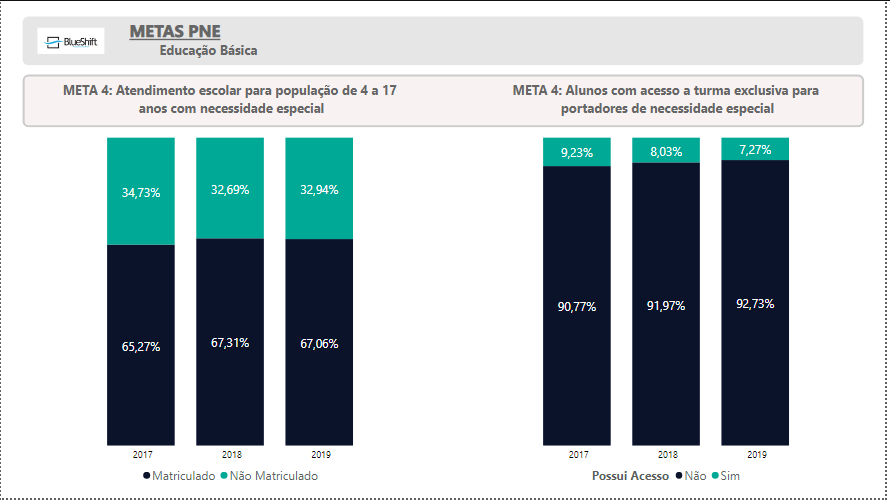


Neste dashboard é abordado a meta 3 onde o objetivo é universalizar até 2016 o atendimento escolar para jovens de 15 a 17 anos e atingir até 2024, a taxa de 85% de matrículas (desta mesma faixa etária) no ensino médio.

Para a primeira parte da meta foram criadas duas medidas com base na tabela do censo escolar (matriculas), uma filtrando todos os jovens desta faixa etária, matriculados em qualquer modalidade de ensino e outra com jovens não matriculados ou sem informação de qual etapa de ensino estão enquadrados. Nesta meta, considerando o prazo estipulado no PNE, está em atraso, porém próximo de ser alcançado com base na baixa porcentagem de alunos que ainda não possuem atendimento escolar.

Para a segunda parte também foram criadas duas medidas, também utilizando dados da tabela do censo escolar, porém desta vez considerando jovens de 15 a 17 matriculados no ensino médio (Normal, Integrado ou Magistério) e jovens que estão matriculados em etapas que não se enquadram como ensino médio ou até que não foram informados a etapa de ensino atual. Esta métrica por sua vez está distante da taxa estabelecida e analisando o crescimento de 2017 a 2019 percebe-se que a meta não será atingida até o prazo estipulado.

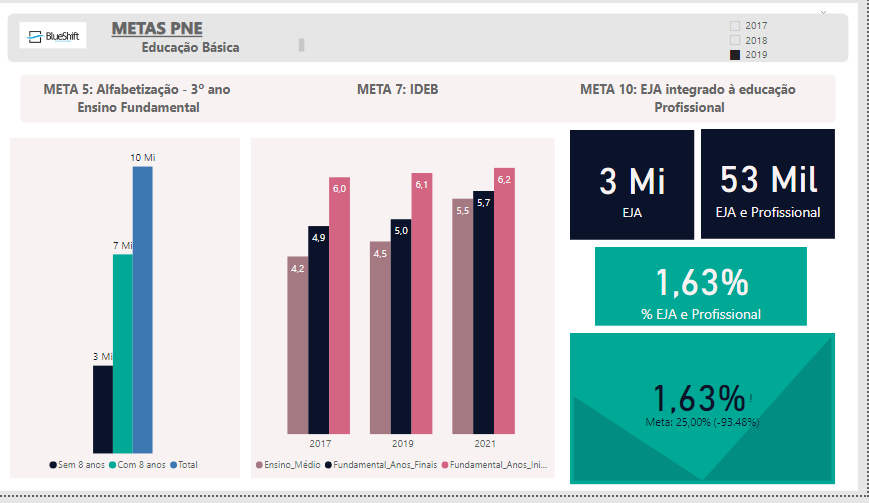
1. Educação Básica\_3 – meta 4



Neste dashboard é apresentado a meta 4, cujo objetivo é apresentar dados sobre a educação da população com alguma necessidade especial (deficiência, transtorno global do desenvolvimento ou superdotação), porém como não temos dados detalhados sobre essa parcela da população foram analisados de acordo com medidas criadas com dados disponíveis na tabela do censo escolar, a população de 4 a 17 anos que possui alguma necessidade especial e que estão enquadrados em alguma etapa de ensino e os que possui alguma necessidade especial mas não estão matriculados ou não tiveram os dados da etapa de ensino informados e posteriormente mais duas metas foram criadas com base nos alunos que estão matriculados, classificando-os se eles pertencem a turmas exclusivas para alunos com alguma necessidade especial ou não.

E nesta meta verifica-se que apesar de mais da metade dos alunos com alguma necessidade especial estarem na escola, há uma boa parcela sem o atendimento escolar e mesmo os que estão matriculados, grande parte não estão inclusos em turmas exclusivas para eles.

1. Educação Básica\_4 – metas 5, 7 e 10



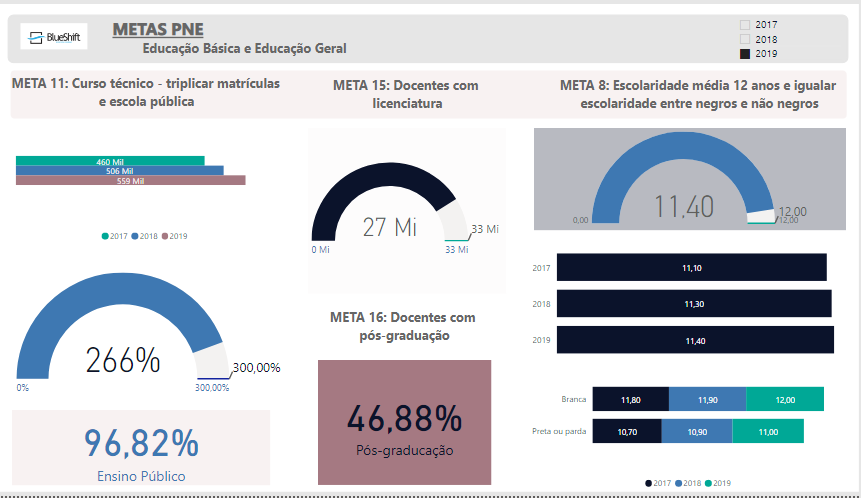
No dashboard acima são apresentadas as metas referentes à educação básica no Brasil, mostrando dados sobre as metas 5, 7 e 10.

A meta 5 visa alfabetizar todas as crianças, no máximo, até o 3º ano do ensino fundamental, ou seja, aos 8 anos de idade, até 2024. Por não termos dados sobre a alfabetização das crianças, trouxemos um panorama geral sobre quantas das crianças matriculadas no 3ª ano do ensino fundamental têm a idade correta para cursar este ano. Para tanto, fizemos 3 medidas com dados da tabela de censo escolar (matrículas): quantidade total de matriculados no 3º ano, quantidade de matriculados no 3º ano com 8 anos e quantidade de matriculados no 3º ano que não têm 8 anos. Pudemos verificar que a maior parte (70%) dos matriculados no terceiro ano têm a idade correta, mas ainda é preciso diminuir a porcentagem de alunos que não estão na idade correta para alfabetização.

A meta 7 visa atingir notas do IDEB específicas para cada etapa escolar até 2021, sendo elas: ensino fundamental – anos iniciais (6,0), ensino fundamental – anos finais (5,5) e ensino médio (5,2). De acordo com evolução mostrada a partir de dados do próprio IDEB, o ensino médio está próximo de atingir a meta. Os ensinos fundamentais ainda precisam crescer no mínimo 0,7 pontos para atingi-la, o que, levando em consideração o crescimento de 2017 para 2019, não é provável que aconteça. O visual contém a projeção ideal para 2021, para demonstrar visualmente essa constatação.

A meta 10 visa oferecer (no mínimo) 25% das matriculas da EJA de forma integrada à educação profissionalizante até 2024. Para análise desta meta foram produzidas medidas, a partir da tabela de censo escolar (matrículas) que mostram: o total de alunos matriculados na EJA, o total de alunos matriculados na EJA integrado à educação profissionalizante e a porcentagem de alunos matriculados na EJA integrada à educação profissionalizante em relação ao total de alunos matriculados nessa modalidade. Com isso conseguimos verificar que a meta está longe de ser atingida, já que apenas 1,63% dos alunos da modalidade está estudando de forma integrada à educação profissionalizante (levando em conta os números de 2019).

1. Educação Básica e Educação Geral – metas 8, 11, 15 e 16



O dashboard acima visa analisar metas relacionadas à educação básica e à educação de modo geral.

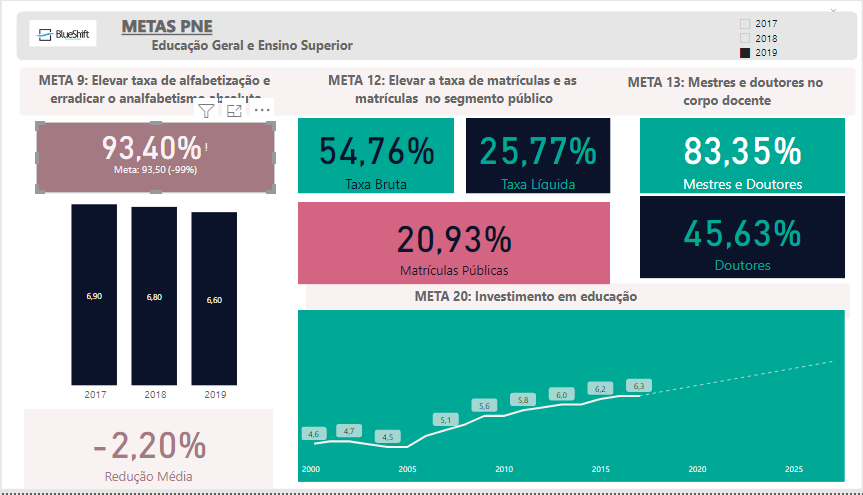
A meta 11 visa triplicar as matrículas da educação profissional técnica de nível médio e garantir que pelo menos 50% dessa expansão seja no segmento público. Para analisar a primeira parte da meta foram criadas medidas, com base na tabela de censo escolar (matrículas), visando entender o crescimento anual das matrículas de educação técnica e fazer uma projeção do crescimento atingido em 10 anos, com base no crescimento médio anual. Com isso conseguimos verificar que, apesar de uma projeção de crescimento de 266% em 10 anos, o crescimento anual não seria capaz de atingir o panorama de 300% em 10 anos, como planejado. Para a segunda parte da meta, decidimos verificar qual porcentagem o segmento público atinge em relação ao total de matrículas nessa modalidade de ensino e verificamos que foi de 96,82% em 2019 – o que pode ser considerado como um ganho.

A meta 15 pretende garantir até 2015 política de nacional de formação dos profissionais da educação básica, política essa que foi criada em 2017[[15]](#footnote-15). Além disso, assegurar que todos os professores tenham formação específica de nível superior (licenciatura) na área de conhecimento em que atuam até 2024. Como não temos dados sobre a área de atuação de cada professor, buscamos visualizar um panorama geral da quantidade de professores que possuem licenciatura em relação ao total de professores do corpo docente, usando a tabela de censo escolar (docentes). No ano de 2019 verificamos que apenas 27 milhões, dos 33 milhões de professores, possuem licenciatura.

A meta 16 visa formar 50% dos professores da educação básica em nível de pós-graduação e a garantia de formação continuada a eles até 2024. Não temos informações sobre a segunda parte da meta. Assim, para a primeira parte da meta verificamos a quantidade de professores com pós-graduação em relação ao total de professores, usando a tabela de censo escolar (docentes). Com isso vimos que no ano de 2019 a meta ainda não havia sido atingida, chegando ao patamar de 46,88%.

A meta 8 objetiva elevar a escolaridade média da população de 18 a 29 anos para 12 anos e igualar a escolaridade entre negros e não negros. Usamos a tabela do PNAD sobre anos de estudo e mostramos um gráfico com a escolaridade média da população nos anos de 2017 a 2019 e também um indicador para demonstrar que em 2019 a meta ainda não tinha sido atingida, já que a média de anos de estudo foi de 11,40. Além disso mostramos também a comparação de média de anos de estudo da população negra e não negra, que ainda possui uma grande disparidade no ano de 2019 com diferença de 1 ponto (negros – 11,00 e não negros 12,00).

1. Educação Geral e Ensino Superior – metas 9, 12 e 13



O dashboard acima mostra dados sobre a educação no Brasil de modo geral e sobre o ensino superior.

A meta 9 visa elevar a taxa de alfabetização da população com 15 anos ou mais para 93,5% até 2024, erradicar o analfabetismo absoluto e reduzir em 50% a taxa de analfabetismo funcional. Não temos dados sobre o analfabetismo funcional, portanto vamos nos ater às primeiras partes da meta. Para descobrir a alfabetização da população de 15 anos ou mais usamos a tabela sobre analfabetismo do PNAD. Com base nela, fizemos uma medida para ver a porcentagem de alfabetização da população partindo dos dados percentuais de analfabetismo. Com isso descobrimos que, em 2019, estávamos perto da meta já, com um valor de 93,4%. Para a segunda parte da meta verificamos o analfabetismo absoluto e conseguimos ver que anualmente há uma diminuição média anual de 2,20% na taxa de analfabetismo.

A meta 12 deseja elevar a taxa bruta de matrículas (50%) e a taxa líquida de matrículas (33%) da população de 18 a 24 anos na educação superior. E garantir que pelo menos 40% das novas matrículas seja na rede pública. Visando analisar a primeira parte da meta foram criadas duas medidas que mostram a taxa bruta de matrículas e a taxa líquida de matrículas, com base na tabela de censo do ensino superior (alunos). Com isso verificamos que, em 2019, a taxa bruta era de 54,76% - atingindo a meta, e a taxa líquida era de 25,77% - ainda abaixo da meta. Pudemos analisar, portanto, que apesar de termos atingido a taxa bruta de matrículas, muitas delas ainda não estão na faixa de idade desejada. Para a segunda parte da meta fizemos uma análise da proporção de matrículas no segmento público em relação ao total e verificamos que apenas 20,93% das matrículas são no segmento público, ficando bem abaixo da meta.

A meta 13 tem como objetivo elevar a qualidade da educação superior, tendo no mínimo 75% do corpo docente de mestres e doutores e 35% de doutores. Com base na tabela de censo escolar (docentes), essa meta foi alcançada já em 2019, com 83,35% de mestres e doutores e 45,63% de doutores no corpo docente. Para chegar à essa análise foi feita a proporção de mestres e doutores (e apenas doutores) em relação ao total de professores.

A meta 20 via aumentar o investimento público em educação, de forma a atingir pelo menos 7% do PIB (Produto Interno Bruto) até 2019 e 10% até o 2024. Para esta meta foi feita uma projeção a partir da tabela de investimentos em educação que contém dados até 2017. Com isso pudemos verificar que, de acordo com a projeção que usa os dados anteriores como base, as metas não seriam atingidas. Em 2019 teríamos 6,55% (6,91 se levarmos em conta a projeção superior) e em 2024 7,16% (7,8 se levarmos em conta a projeção superior). Ou seja, em ambos os casos a meta não seria atingida, apesar de estarmos mais próximos de atingi-la em 2019.

# Fontes

**Bases de dados**

<http://api.sidra.ibge.gov.br/>

<https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>

<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>

<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar>

<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-financeiros-educacionais>

**Pesquisa Documental**

**PNE**

<http://pne.mec.gov.br/images/pdf/pne_conhecendo_20_metas.pdf>

<http://pne.mec.gov.br/>

**Educação**

<http://portal.mec.gov.br/>

**IBGE**

<https://www.ibge.gov.br/>

**IDEB**

<http://ideb.inep.gov.br/>

**Taxa bruta x taxa liquida – escolar**

<https://gestrado.net.br/verbetes/taxa-de-matricula-bruta/>

<https://gestrado.net.br/verbetes/taxa-de-matricula-liquida/>

**Cálculo de juros compostos**

<https://www.todamateria.com.br/juros-compostos/>

**Paleta de cores**

<https://coolors.co/0b132b-3e78b2-00a896-a57982-d36582>

**Databricks**

**Documentação e Precificação**

https://databricks.com/product/azure-pricing

https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html

<https://docs.microsoft.com/pt-br/azure/databricks/>

**Pandas x Spark**

https://towardsdatascience.com/spark-vs-pandas-part-1-pandas-10d768b979f5

<https://towardsdatascience.com/spark-vs-pandas-part-2-spark-c57f8ea3a781>

**Databricks Community Edition**

https://blog.gft.com/br/2020/09/29/getting-started-with-databricks-2/

https://databricks.com/ce-termsofuse

<https://www.topcoder.com/thrive/articles/databricks-community-edition-a-beginners-guide-part-2>

**Gerenciamento do Sistema de Arquivos**

<https://www.mssqltips.com/sqlservertip/6700/azure-databricks-local-file-system-management/>

**Dimensionamento de Cluster**

https://www.youtube.com/watch?v=7xzomULtqwY

**DBFS e Databricks Utilities**

https://docs.databricks.com/data/databricks-file-system.html

https://docs.databricks.com/dev-tools/databricks-utils.html

<https://www.youtube.com/watch?v=qR18LMfzULQ>

**Referência Notebooks Databricks**

<https://docs.databricks.com/notebooks/notebooks-use.html>

<https://docs.databricks.com/notebooks/notebooks-manage.html>

**Arquivos .zip no Databricks**

https://docs.databricks.com/data/data-sources/zip-files.html

<https://docs.databricks.com/_static/notebooks/zip-files-python.html>

**Formatação de Notebooks com Markdown**

https://forums.databricks.com/static/markdown/help.html

**Leitura de arquivos de URLs no Databricks**

<https://www.youtube.com/watch?v=783lhm2eIKw>

**Databricks REST API**

<https://docs.databricks.com/dev-tools/api/latest/examples.html>

**Importar, ler e modificar dados no Databricks**

<https://docs.databricks.com/data/data.html>

**Conectando Databricks ao Azure Storage**

https://docs.databricks.com/data/data-sources/azure/azure-storage.html

<https://www.youtube.com/watch?v=qGfcjmvCyiU&ab_channel=ShareCloudTech>

https://www.youtube.com/watch?v=qGfcjmvCyiU

<https://www.youtube.com/watch?v=zwMksSEjNvU>

https://www.youtube.com/watch?v=yeNgrBxHmCc

**Versionamento no Github**

<https://docs.databricks.com/notebooks/github-version-control.html>

**Integração com Power BI**

https://docs.databricks.com/integrations/bi/power-bi.html

<https://docs.databricks.com/integrations/bi/jdbc-odbc-bi.html#authentication>

**Erro - ODBC: ERROR**

https://community.powerbi.com/t5/Desktop/Spark-datasource-ODBC-ERROR-HY000/m-p/939399

**Erro - OutOfMemoryError**

https://docs.microsoft.com/pt-br/azure/hdinsight/spark/apache-spark-troubleshoot-outofmemory

**Erro - spark.driver.maxResultSize**

https://spark.apache.org/docs/latest/configuration.html

<https://docs.microsoft.com/pt-br/azure/databricks/kb/jobs/job-fails-maxresultsize-exception>

**Delta Lake**

https://www.zdnet.com/article/a-standard-for-storing-big-data-apache-spark-creators-release-open-source-delta-lake/

https://www.upsolver.com/blog/databricks-delta-lake-data-lake-etl-comparison#:~:text=Delta%20Lake%20tables%20are%20a,read%20by%20a%20Delta%20cluster.&text=Data%20lake%20ETLs%20only%20make,used%20to%20query%20the%20data.

https://delta.io/

https://www.exasol.com/resource/delta-lake-format-integration-how-can-it-help-you/

https://docs.databricks.com/delta/quick-start.html

https://databricks.com/blog/2019/08/21/diving-into-delta-lake-unpacking-the-transaction-log.html

https://www.youtube.com/watch?v=BMO90DI82Dc&list=PLTPXxbhUt-YXlZZYWEioChooc9QZ\_Mf4g&index=3

https://databricks.com/session\_eu19/simplify-and-scale-data-engineering-pipelines-with-delta-lake

https://databricks.com/blog/2019/08/14/productionizing-machine-learning-with-delta-lake.html

<https://medium.com/@mayur.specindia/how-to-call-web-api-from-an-azure-data-bricks-notebook-to-a-delta-lake-table-60abb36dd6ac>

**Linux**

**Wget**

https://www.gnu.org/software/wget/manual/wget.html#URL-Format

**Remover arquivos com determinada extensão**

https://www.tecmint.com/delete-all-files-in-directory-except-one-few-file-extensions/

**Extrair arquivos sem recriar estrutura de pastas**

<https://unix.stackexchange.com/questions/59276/how-to-extract-only-a-specific-folder-from-a-zipped-archive-to-a-given-directory>

**Pandas**

**Documentação**

https://pandas.pydata.org

**Pandas Dataframe**

https://pandas.pydata.org/pandas-docs/stable/reference/frame.html

https://dzone.com/articles/pyspark-dataframe-tutorial-introduction-to-datafra

**Pandas to\_csv**

https://stackoverflow.com/questions/59331694/handling-bad-lines-in-a-python-read-csv-execution

**Tipo de Dados no Pandas**

https://pbpython.com/pandas\_dtypes.html

**Alteração de valores das colunas**

<https://www.youtube.com/watch?v=Q2nFweWoL2M>

**Pandas – ler do blob**

<https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/explore-data-blob>

**PySpark**

**Documentação e Tutoriais**

https://sparkbyexamples.com/pyspark-tutorial/

https://spark.apache.org/docs/2.1.0/api/python/pyspark.sql.html

**Tipos de Dados no PySpark**

<https://spark.apache.org/docs/latest/sql-ref-datatypes.html>

**PySpark Save Modes**

https://spark.apache.org/docs/latest/sql-data-sources-load-save-functions.html

**Imprimir uma coluna específica**

https://stackoverflow.com/questions/35913506/how-to-print-only-a-certain-column-of-dataframe-in-pyspark

**Alteração de valores de colunas**

https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.DataFrame.replace.html#pyspark.sql.DataFrame.replace

**Deletar Colunas**

https://stackoverflow.com/questions/29600673/how-to-delete-columns-in-pyspark-dataframe

**Construção de Variação de Idade**

https://stackoverflow.com/questions/44186769/how-to-do-age-banding-in-pyspark

**Retornar valores distintos de uma coluna**

https://forums.databricks.com/questions/9655/how-to-get-unique-values-of-a-column-in-pyspark-da.html

**Substituição de valores null**

https://stackoverflow.com/questions/42312042/how-to-replace-all-null-values-of-a-dataframe-in-pyspark

https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.DataFrame.fillna.html

https://sparkbyexamples.com/pyspark/pyspark-fillna-fill-replace-null-values/#pyspark-fillna-and-fill-syntax

**Adicionar Colunas**

https://kontext.tech/column/spark/455/tutorial-add-constant-column-to-pyspark-dataframe

**Contagem de valores nulos de uma tabela**

<https://www.datasciencemadesimple.com/count-of-missing-nanna-and-null-values-in-pyspark/>

**Conversão de Spark Dataframe para Pandas Dataframe**

https://stackoverflow.com/questions/50958721/convert-a-spark-dataframe-to-pandas-df

https://medium.com/hashmapinc/5-steps-to-converting-python-jobs-to-pyspark-4b9988ad027a

**Conversão de Tipo String para Date**

https://www.journaldev.com/23365/python-string-to-datetime-strptime

https://docs.python.org/3/library/datetime.html#strftime-strptime-behavior

**Conversão de Tipo String para Float**

<https://stackoverflow.com/questions/35684018/how-to-convert-dataframe-columns-from-string-to-float-double-in-pyspark-1-6>

**Python Arguments**

https://careerkarma.com/blog/python-positional-argument-follows-keyword-argument/

**Exportando Dataframe para um único CSV**

https://towardsdatascience.com/databricks-how-to-save-files-in-csv-on-your-local-computer-3d0c70e6a9ab

<https://analyticjeremy.github.io/Databricks_Examples/Write%20to%20a%20Single%20CSV%20File>

**Power BI**

**Conversão de Tipo BigInt para Date**

https://community.powerbi.com/t5/Power-Query/Convert-bigint-to-datetime/td-p/1295238

**Adicionando zeros antes dos dígitos**

<https://www.powerbi-pro.com/en/power-query-adding-leading-zeros/>

**Operadores**

https://docs.microsoft.com/pt-br/dax/dax-operator-reference

https://www.sqlbi.com/articles/the-in-operator-in-dax/

https://community.powerbi.com/t5/Desktop/DAX-equivalent-of-SQL-s-NOT-IN/m-p/586184

**Funções**

https://docs.microsoft.com/pt-br/dax/sumx-function-dax

https://community.powerbi.com/t5/Desktop/Count-of-rows-after-filtering-by-a-measure/m-p/64259

https://community.powerbi.com/t5/Desktop/Multiple-Filters-COUNTROWS/m-p/64156

https://docs.microsoft.com/pt-br/dax/countrows-function-dax

**Tutoriais Youtube**

<https://www.youtube.com/watch?v=a5rRMD-2O-k>

<https://www.youtube.com/watch?v=8on3t15QPzQ>

**Alterar interações entre visuais**

<https://docs.microsoft.com/pt-br/power-bi/create-reports/service-reports-visual-interactions>

**Tabela Calendário**

<https://community.powerbi.com/t5/Desktop/Create-a-List-of-Years/td-p/232860>

**Calcular média de crescimento**

<https://community.powerbi.com/t5/Desktop/calculate-average-of-a-measure/m-p/297565>

**Calcular % diferença anual**

<https://www.mssqltips.com/sqlservertip/6424/calculate-percentage-growth-over-time-with-power-bi/>

<https://www.youtube.com/watch?v=8on3t15QPzQ>

**Web Scraping**

**Selenium**

https://selenium-python.readthedocs.io/

https://www.youtube.com/watch?v=b5jt2bhSeXs

https://forums.databricks.com/questions/15480/how-to-add-webdriver-for-selenium-in-databricks.html

https://askubuntu.com/questions/20689/what-does-add-apt-repository-mean#:~:text=add%2Dapt%2Drepository%20adds%20a,in%20the%20official%20Ubuntu%20repositories.

https://www.selenium.dev/documentation/en/

https://forums.databricks.com/questions/19376/if-my-notebook-downloads-a-file-from-a-website-by.html

<https://stackoverflow.com/questions/58347261/extracting-table-data-using-selenium-and-python-into-pandas-dataframe>

**Chrome Driver para Selenium**

https://sites.google.com/a/chromium.org/chromedriver/downloads

<https://chromedriver.chromium.org/>

**Beautiful Soup**

https://docs.microsoft.com/en-us/answers/questions/78527/how-add-class-library-for-beautifulsoup-in-azure-d.html

https://imasters.com.br/back-end/aprendendo-sobre-web-scraping-em-python-utilizando-beautifulsoup

https://datascience.stackexchange.com/questions/11730/how-to-scrape-a-website-with-a-searchbar/11734

**Pandas**

https://www.youtube.com/watch?v=ooj84UP3r6M

**Permissões para Web Scraping**

<https://developers.google.com/search/docs/advanced/robots/robots_txt>

1. comando utilitário do Databricks para facilitar a execução de tarefas [↑](#footnote-ref-1)
2. comando mágico do Databricks para acessar o sistema de arquivos [↑](#footnote-ref-2)
3. Em um deles os arquivos foram armazenados primeiramente no *Blob* e depois extraídos do mesmo antes das transformações, por questões de performance. [↑](#footnote-ref-3)
4. O tamanho máximo padrão do *driver.MaxResultSize* (limita os trabalhos serializados de resultados das partições de cada ação do spark – serve para proteger o driver de erros de falta de memória) do spark é 1GB e não suportava a importação da tabela diretamente pelo Databricks ao Power BI [↑](#footnote-ref-4)
5. Para a criação inicial dos scripts em Python foi usado o Databricks Community Edition, plataforma sem custo que possibilita o acesso ao ambiente de notebooks Databricks com o provisionamento de um cluster com capacidade limitada (Driver: Community Optimized 15.3 GB Memory, 2 Cores, 1 DBU) e que não permite a reativação pós término. Essa edição do Databricks é hospedada na nuvem AWS. [↑](#footnote-ref-5)
6. Todos os notebooks foram integrados ao GitHub (<https://github.com/FZamberlam/BlueShiftAcademyPI>). [↑](#footnote-ref-6)
7. Todos os clusters foram provisionados habilitando sua finalização depois de 20 minutos de inatividade. [↑](#footnote-ref-7)
8. A importação das tabelas foi realizada usando conexão JDBC/ODBC (usando o conector do Databricks) com as seguintes informações:

   *Hostname* do servidor: adb-812110395505188.8.azuredatabricks.net

   Caminho http: sql/protocolv1/o/812110395505188/0327-214619-party52 [↑](#footnote-ref-8)
9. adb-812110395505188.8.azuredatabricks.net [↑](#footnote-ref-9)
10. sql/protocolv1/o/812110395505188/0327-214619-party52 [↑](#footnote-ref-10)
11. m03storage [↑](#footnote-ref-11)
12. 3Ts0JpdSfDvYx8JOerU9CrExx1Ankj3W1q5J3Cscu2hXlQR90cpMlVy6p2mBB0NpKndKWixIcp1poqcFrAJ+oQ== [↑](#footnote-ref-12)
13. A tabela contém dados de 2000 a 2017, que iremos usar para projetar possíveis dados para visualização até 2024 [↑](#footnote-ref-13)
14. O dashboard do Power BI pode ser visualizado no GitHub (<https://github.com/FZamberlam/BlueShiftAcademyPI>). [↑](#footnote-ref-14)
15. <https://bit.ly/3uxt4li> [↑](#footnote-ref-15)