UNIVERSIDADE FEDERAL DE SANTA MARIA

COLÉGIO POLITÉCNICO DA UFSM

CURSO DE SISTEMAS PARA INTERNET

Elis Regina Scherer

UTILIZAÇÃO DE RECURSOS DE BI PARA ANÁLISE DE DADOS PÚBLICOS

Santa Maria, RS

2021

Elis Regina Scherer

UTILIZAÇÃO DE RECURSOS DE BI PARA ANÁLISE DE DADOS PÚBLICOS

Trabalho de Graduação apresentado ao Curso de Sistemas para Internet, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção de grau para: **Tecnólogo em Sistemas para Internet.**

Orientador: Prof. Dr. Daniel Lichtnow.

Santa Maria, RS

2021

Elis Regina Scherer

UTILIZAÇÃO DE RECURSOS DE BI PARA ANÁLISE COM DADOS PÚBLICOS

Trabalho de Graduação apresentado ao Curso de Sistemas para Internet, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção de grau para: **Tecnólogo em Sistemas para Internet.**

Aprovado em

_____
Daniel Lichtnow, Dr. (UFSM)
(Presidente/Orientador)

_____
Eduardo Casagrande Stabel. Dr. (UFSM)

_____
Fernando Emilio Puntel. Me. (UFSM)

Santa Maria, RS

2021

AGRADECIMENTOS

Agradeço a mim mesma pela capacidade e perseverança de recomeçar a cada tropeço, esta força de vontade que me proporcionou chegar até aqui. Agradeço a meu namorado, amigo, mentor, fonte de inspiração Raul por me apoiar incondicionalmente nesse processo de aprendizado e formação de carreira, sendo suporte para que eu melhore cada dia mais, não perdendo minha identidade.

Agradeço aos professores que foram a base para construir meu conhecimento, em especial meu orientador Professor Daniel, pela insistência, competência e dedicação em seus ensinamentos sobre banco de dados, ao Professor Rafael pela incrível didática que contribui tanto para meus conhecimentos em Java e se tornou grande inspiração profissional e ao Professor Eduardo que me inspirou muito a aprofundar meus conhecimentos em matemática e estatística, esse é só o início desse passo.

Agradeço aos meus colegas pelo grande suporte psicológico e emocional nesta caminhada, pelas longas conversas e cumplicidade Pedro, Laviny, Wesley, Christian em especial ao Lucas, espero que essa amizade perdure para além da universidade, vocês são incríveis, agradeço e torço por vocês.

RESUMO

UTILIZAÇÃO DE RECURSOS DE BI PARA ANÁLISE COM DADOS PÚBLICOS

AUTORA: Elis Regina Scherer

ORIENTADOR: Prof. Dr. Daniel Lichtnow.

O objetivo geral é utilizar ferramentas de Business Intelligence - BI para explorar o dados públicos referentes à concessão de bolsas do ProUni dos últimos 15 anos.Neste sentido, o trabalho apresenta um estudo introdutório referente à utilização de dados públicos e ferramentas de BI. O estudo foi embasado com fundamentos de bancos de dados relacionais e modelos dimensionais, com foco nas consultas características de cada um e seu projeto. Apresenta ainda aspectos relacionados à utilidade de *Data Warehouse* no apoio à tomada de decisão, por fim, é demonstrada a utilização da ferramenta Power BI e sua destacando sua capacidade de reunir dados de diferentes formatos, além de sua compatibilidade com diferentes linguagens de programação.

Palavras chave: Dados Públicos, *Data Warehouse,* Power BI.

ABSTRACT

USE OF BI RESOURCES FOR PUBLIC DATA ANALYSIS

AUTHOR: Elis Regina Scherer

ADVISOR: Prof. Dr. Daniel Lichtnow.

The general objective is to use Business Intelligence - BI tools to explore public data related to the granting of grants from ProUni in the last 15 years. In this sense, the work presents an introductory study regarding the use of public data and BI tools. The study was based on fundamentals of relational databases and dimensional models, focusing on queries characteristic of each one and its project. It also presents aspects related to the usefulness of Data Warehouse to support decision making. Finally, the use of the Power BI tool is demonstrated, highlighting its ability to gather data from different formats, in addition to its compatibility with different programming languages.

Keywords: Public Data, Data Warehouse, Power BI.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SGBD	Sistema de Gerenciamento de Banco de Dados
OLAP	*Online Analytical Processing*
OLTP	*Online Transaction Processing).*
BI	*Business Intelligence*
DW	*Data Warehouse*
CSV	*Comma Separated Value*
CGU	Controladoria Geral da União
SQL	*Structured Query Language*
JPA	*Java Persistence API*
INDA	Infraestrutura Nacional de Dados Abertos
PROUNI	Programa Universidade para todos
ETL	Extração, Transformação e Carga
IBGE	Instituto Brasileiro de Geografia e Estatística
MEC	Ministério da Educação
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
FIES	Fundo de Financiamento Estudantil
LAI	Lei de Acesso à Informação
PDA	Plano de Dados Abertos
DIPES	Secretaria de Planejamento, Orçamento e Gestão
SESu	Secretaria de Educação Superior
EGDI	Índice de Desenvolvimento de Governo Eletrônico
SLTI	Secretaria de Logística e Tecnologia da Informação
EAD	Educação à Distância
KDE	*Kernel Density Estimate*

Sumário

1 INTRODUÇÃO

Com o crescente uso e evolução da *Web* nos últimos anos, nos deparamos constantemente com um grande volume de dados, a informação originada destes dados vem ganhando importância constante. Muitos dados de uso interno de organizações públicas e privadas começaram a ser disponibilizados para o público em geral, dando origem aos dados abertos (GOMES, 2016).

Há alguns anos o governo brasileiro vem tornando sua administração mais transparente, por meio de publicações de informações na *web.* De acordo com a Controladoria Geral da União (CGU), as páginas de transparência devem promover a visibilidade de gastos públicos. Os dados abertos devem ser livres ao uso, reutilização e redistribuição, compartilhando sua autoria e licença, (VICTORINO et.al, 2017).

Mais importante que apenas publicar dados, é disponibilizá-los de forma que possuam algum tipo de ligação, comumente os dados são disponibilizados em arquivos, muitos em formato proprietário ou armazenados em diferentes bases sem nenhuma integração, esses fatores dificultam a utilização e análise destes dados, (GOMES, 2016).

Os dados disponibilizados pelo governo podem ter valor econômico para a iniciativa privada, e tendem a ser uma alternativa viável a ser explorada. Esses podem oferecer valor para além do seu objetivo específico de coleta. Com as possibilidades existentes nas ferramentas de BI é possível facilitar a integração de dados de diferentes locais em um único repositório, para apoio à tomada de decisão nas organizações, (SILVA, 2017).

1.1. OBJETIVOS

1.1.1 Objetivo Geral

Este trabalho objetiva analisar a utilização de ferramentas que são frequentemente associadas à área de *Business Intelligence* (Inteligência de Negócios em português) para análise serão utilizados dados de domínio público referentes à ocupação de vagas do ProUni (Programa Universidade para Todos) de 2005 a 2019.

1.1.2 Objetivos Específicos

A partir do objetivo geral, foram definidos os seguintes objetivos específicos:

- Estudar as tecnologias e ferramentas de *Business Intelligence*;
- Compreender os dados publicados sobre ocupação de vagas do ProUni;
- Aplicar a ferramenta Power BI na análise dos dados explorando os recursos da ferramenta e a sua integração com outras ferramentas/softwares para análise de dados especialmente Python.

2 FUNDAMENTAÇÃO TEORICA

2.1. BANCO DE DADOS

O homem ao longo de sua existência, teve a necessidade de registrar eventos importantes e informações que pudesse utilizar futuramente. As técnicas foram de pinturas pré-históricas, hieróglifos, papiros, entre outras. A partir do século XV os registros passaram a ser armazenado no papel, apesar de sua indiscutível utilidade, utilizar os dados destes registros era trabalhoso. Com a era computacional o registro passou a ser feito com fita de papel perfurado, posteriormente pelo cartão perfurado. Os programas de banco de dados inicialmente eram softwares simples para manipular os dados do arquivo, com o passar do tempo passaram a ficar mais complexos, evoluindo para os sistemas de banco de dados que utilizamos hoje (PEREIRA, 2014).

Um banco de dados em sua definição é uma coleção de dados relacionados. Os dados são fatos que podem ser gravados e que possuem um significado implícito (ELSMARI, 2005). Para construção de um banco de dados, três partes são fundamentais: uma fonte de informação, um público que demonstre interesse nos dados do banco e uma interação com o mundo real, isto é, um banco de dados deve representar uma porção do mundo real e refletir qualquer alteração dele, a fonte de informação é de onde os dados serão extraídos, (PEREIRA, 2014).

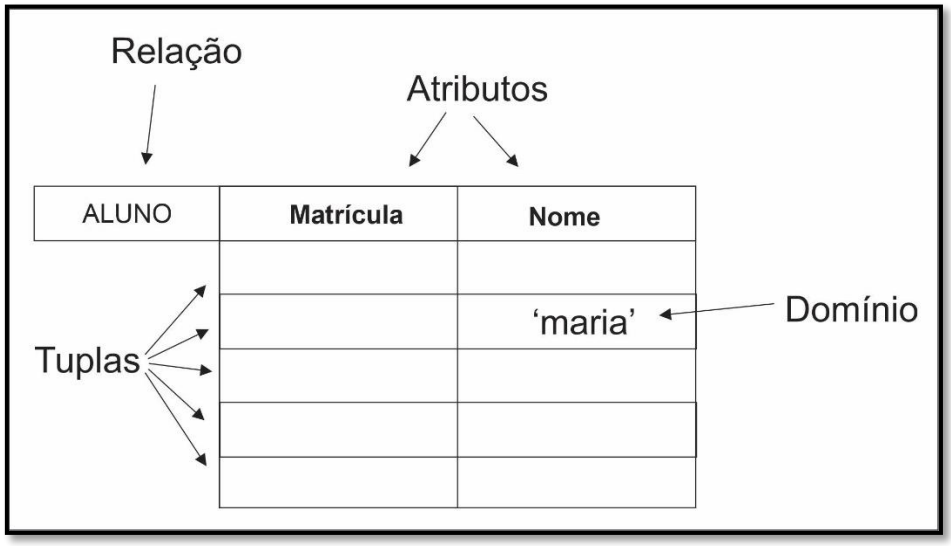
Um banco de dados tem tamanho e complexidade variável, pode ser gerado e mantido manualmente ou de forma automatizada. Sua construção utiliza um processo de armazenagem de dados de forma apropriada para manipulação. A manipulação deve incluir recuperação de dados, atualização para refletir as mudanças do mundo real e gerar relatórios, (ELSMARI, 2005).

2.2. BANCO DE DADOS RELACIONAL

De acordo com o autor SILBERSCHATZ (1999) o sistema de gerenciamento de banco de dados (SGBD) é um conjunto de dados associado a um conjunto de programas que possibilita o acesso a esses dados. Bancos de dados relacionais são caracterizados por coleções de tabelas. Cada linha em uma tabela, representa um fato que corresponde a uma entidade ou

relacionamento no mundo real (ELSMARI, 2005). O nome e as colunas são usados para interpretação dos valores de cada linha. Uma linha é denominada tupla, o cabeçalho da coluna é um atributo, a tabela é chamada de relação, o tipo dado que aparece em cada coluna é um domínio, como ilustra a Figura 1. Um domínio é um conjunto de valores atômicos, ou seja, indivisíveis, cada coluna possui um domínio de valores possíveis.

Figura 1 - Modelo Relacional.



Fonte: Autor.

Banco de dados relacionais são geralmente focados no nível operacional de uma organização, voltando-se à execução operacional, com consultas OLTP (*Online Transaction Processing).* Neste caso, os dados são sujeitos à modificações realizadas por usuário. Os bancos de dados relacionais atuam sob essas modificações por meio do controle de transações, permitindo assim a leitura, inserção, modificação e exclusão de dados. As consultas OLTP são caracterizadas por muitas transações de curta duração, que acessam ou produzem um pequeno volume de dados, diferente das consultas OLPA (*Online Analytical Processing*) que carregam grande volume de dados. Suponhamos que uma consulta OLTP retorne o total de vendas de uma empresa dos últimos 10 anos, o custo desta transação degrada a performance do processo, (ELSMARI, 2005).

Em aplicações empresariais, normalmente, é importante que o banco de dados garanta as chamadas propriedades ACID (Atomicidade, Consistência, Isolamento e Durabilidade). A atomicidade garante que uma transação seja realizada por completo, caso contrário o estado dos dados retorne ao estado anterior a execução da transação, a consistência garante que a transação

crie um estado válido dos dados e em caso de falha retome o estado anterior dos dados sem alteração, o isolamento garante que uma transação em andamento ainda invalidada permaneça de forma isolada de quaisquer outras operações e finalmente a durabilidade deve garantir que os dados validados sejam registrados pelo sistema de forma a continuar disponíveis de forma integra mesmo que falhas venham a ocorrer.

2.3. BUSINESS INTELLIGENCE – BI

Ramesh (2019), define os dados como principal ingrediente para qualquer iniciativa de BI, ciência de dados e analise. Os dados são base para que tecnologias de decisão produzam informações e conhecimento. Hoje os dados são considerados um bem valioso em uma organização, utilizados como diferencial competitivo, forma de analisar clientes e processos comerciais.

Empresas públicas e privadas, frequentemente necessitam responder rapidamente as situações inesperadas, além de contornar essas situações é necessário inovar em soluções frente às mudanças constantes do cenário organizacional. Essas circunstancias exigem agilidade na tomada de decisão. A tomada de decisão envolve conhecimento e dados relevantes, o processamento dessas informações deve ser em tempo real. O *Business Intelligence* é um suporte informatizado a nível gerencial para apoiar a tomada de decisão, (TURBAN, 2009).

Conforme Imhoff (2003), *Business Intelligence* é a capacidade de uma empresa estudar comportamentos e ações anteriores, a fim de identificar e compreender o que a levou a situação atual, além de prever ou mudar ações futuras.

Integrar a tecnologia em processos organizacionais é um elemento essencial, somado a um conjunto de estratégias organizacionais. Essa solução proporciona meios de agilizar processos de informação computadorizados, entretanto, necessita profissionais capacitados e ambientes tecnológicos adequados, resultando em um investimento consideravelmente alto. Estes recursos precisam estar de acordo com o negócio e o que a organização almeja, (GOMES, 2011).

Os sistemas de banco de dados convencionais, com modelos relacionais, não são projetados para gerar e armazenar informações com finalidade de estratégia de negócio. Dentro de uma organização a informação está descentralizada em diferentes departamentos, o que torna

os dados vagos e sem valor para apoiar a tomada de decisão. Com isso introduziu-se um novo conceito no mercado, o *Data Warehouse* (CAVALCANTI et.al, 2005).

2.4. DATA WAREHOUSE

O *Data Warehouse* (DW), em sua tradução literal para o português armazém de dados, é uma coleção de dados utilizados como suporte à tomada de decisão. Este repositório detém dados atuais e históricos de possível interesse à gerência da organização (TURBAN, 2009).

Segundo Turban (2009), um *Data Warehouse* deve orientar seus dados por assunto além de ser integrado, ou seja, integrar dados de diferentes fontes de forma consistente. Uma de suas características mais importantes é não apresentar volatidade, isso significa que os dados inseridos não podem ser alterados. Um repositório não-volátil é fundamental, pois um *data warehouse* deve manter históricos para realizar comparações e analises que auxiliam na tomada de decisão.

Kimball (2002) define um *Data Warehouse*, como um banco dedicado usado para auxiliar a tomada de decisão, para isto ele deve ser uma cópia de dados transacionais estruturados para consultas dedicadas à análise, denominadas processamento analítico online (Online Analytical Processing - OLAP). Esta estrutura de consulta torna possível realizar o tratamento de dados, oriundos de diferentes fontes em tempo real, além de utilizar ferramentas de visualização de dados.

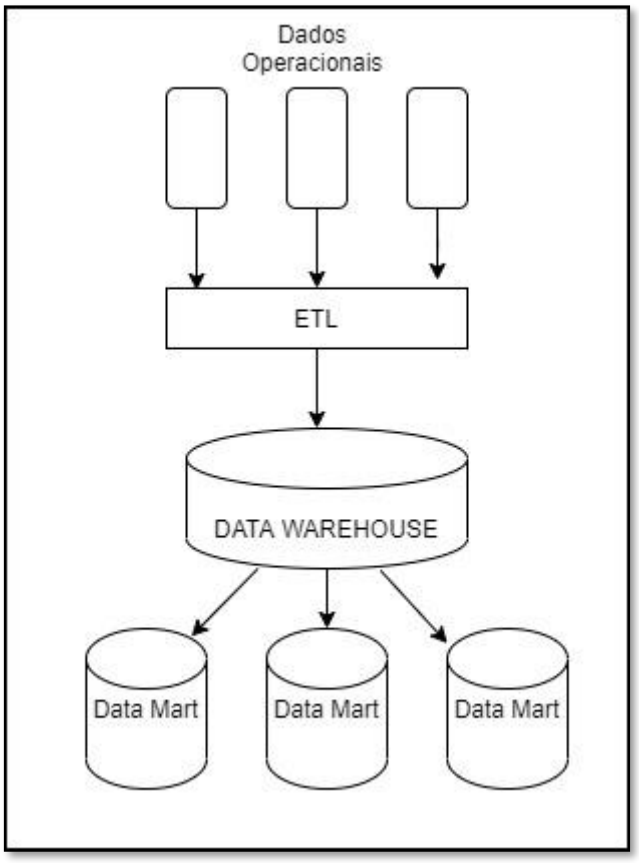
Entende-se que um armazém de dados reúne todos os bancos de dados de uma empresa, um *data mart*, no entanto é um subconjunto de um *data warehouse*, ou seja, ele é composto de uma área específica de uma empresa. Um exemplo disto é a empresa possuir diferentes setores, como financeiro, vendas, produção e marketing. Cada um destes setores possui seus próprios dados e esses constituem o *data mart*. A junção de todos estes *data mart* compõem o *data warehouse* da empresa (TURBAN, 2009).

Existem diferentes abordagens de uso e aplicação de *data mart*, pois, um *data mart* dependente deve ser composto a partir de um *data warehouse*, garantindo a apresentação de dados consistentes e de qualidade. Eles suportam um único modelo de dados em toda a empresa, sendo construído após o levantamento do *data warehouse*. Um *data mart* independente é uma espécie de *data warehouse* em menor escala, produzido para uma unidade estratégica, não

possuindo associação com outros *data marts*, nem dispondo de fonte de dados alimentada por um *data warehouse* (TURBAN, 2009).

A abordagem de Inmon (2002) é baseada no modelo Entidade Relacionamento. Esta abordagem permite o armazenamento de todos os acontecimentos de uma empresa, para isso, Inmon utiliza 4 níveis de distribuição da informação, sendo elas operacional, atômico, departamentais e níveis individuais. O primeiro nível suporta operações transacionais diárias como um banco relacional comum, já os 3 últimos níveis são voltados ao *data warehouse*. Em sua concepção, um *data warehouse* tem sua própria existência física, é orientado para o armazenamento, rastreabilidade e escalabilidade, enquanto o *data mart* é do tipo dependente em sua concepção, tornando sua modelagem bastante técnica, por ser mais complexa e detalhada, denominada *top-down* (de cima para baixo) (YESSAD et al. ,2016), como mostra a Figura 2.

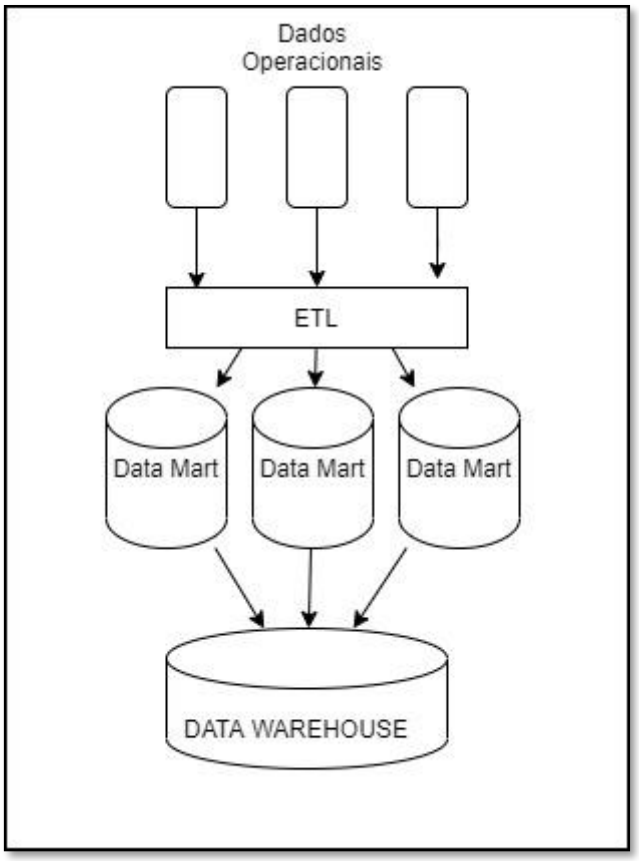
Figura 2 - Modelo Inmon.



Fonte: Autor.

Kimball se opõem aos princípios de isolamento de Inmon, e se baseia no conceito dimensional. Sua abordagem envolve fortemente os usuários finais. Ele visualiza um *data warehouse* como um conjunto de *data marts*, onde cada *data mart* é orientado por assunto ou departamento. Essa abordagem implica em um modelo *down-up* (de baixo para cima) como apresenta a Figura 3, onde o *data warehouse* será formado por data marts, tornando o processo de modelagem mais rápido, menos complexo e de fácil entendimento para o usuário final, (YESSAD et al. ,2016).

Figura 3 - Modelo Kimball.

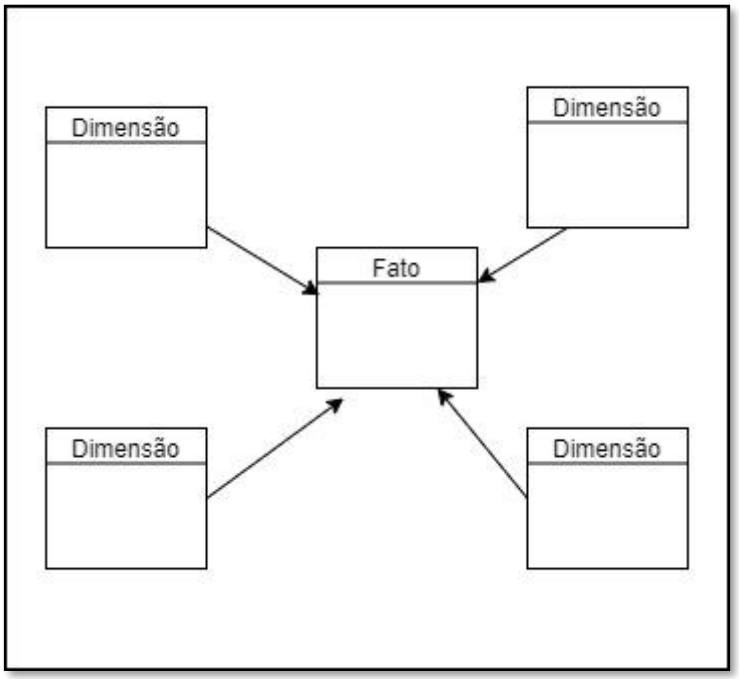


Fonte: Autor.

A modelagem dimensional é baseada em recuperação de dados, capaz de suportar grande volume de consultas (TURBAN, 2009). Essa modelagem possui 2 elementos fundamentais: as tabelas fatos e tabelas dimensões. Um fato é uma transação ou evento, utilizado para analisar um processo da empresa, enquanto uma dimensão são elementos que fazem parte de uma tabela fato, sendo que as dimensões determinam o contexto de um assunto ou negócio (JARDIM et al. ,2015).

A tabela fato é uma tabela central, cercada por diversas tabelas dimensões como mostra a Figura 4. A tabela central deve conter atributos necessários para realizar análise de decisão, atributos descritivos, vinculados às tabelas dimensões. Estes atributos consistem em medidas de desempenho e métricas.

Figura 4 - Star Schema.



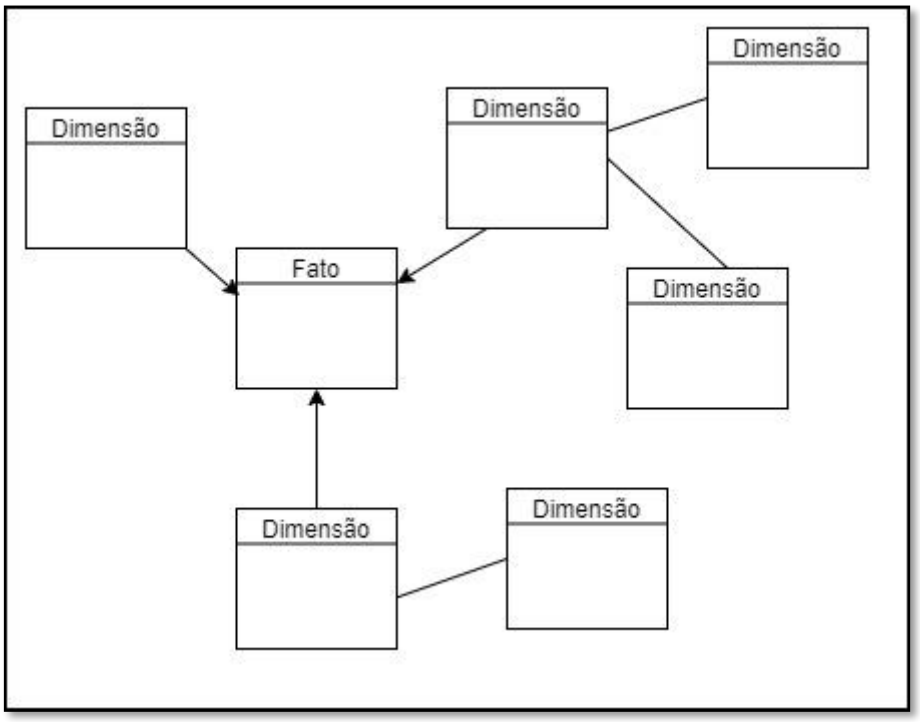
Fonte: Autor.

Segundo Nery (2013), para que a abordagem dimensional cumpra as finalidades de um *data warehouse*, existem alguns modelos de dados disponíveis, que são: *Star Schema* (esquema estrela) e *Snowflake* (floco de neve).

O *Star Schema* é um modelo de dados dimensional padrão, onde há apenas uma tabela fato e chaves simples nas tabelas de dimensões. Cada dimensão é representada por apenas uma tabela. Este modelo não é apenas um diagrama, mas a representação de um processo e do relacionamento dos participantes deste ao decorrer do tempo. Contudo devido a não normalização dos dados as tabelas dimensões tendem a ficar grandes e banco de dados menos flexível, (KIMBALL, 2002)

O *Snowflake* é bem semelhante ao modelo *Star Schema*, exceto pela particularidade de possuir duas ou mais dimensões interligadas, gerando uma hierarquia entre elas (NERY ,2013) como ilustra a Figura 6.

Figura 5 - Modelo Snowflake.



Fonte: Autor.

Além da modelagem de dados, outro processo na implementação de um *data warehouse* é definir a granularidade dos dados. À granularidade estabelece o nível de detalhamento dos dados, onde Kimball (2002) defende que lidar com dados atômicos com níveis de granularidade mais baixos, apresenta mais clareza à tomada de decisão. Dados atômicos fornecem flexibilidade analítica, contudo se a definição da granularidade for de nível superior, o detalhamento das dimensões será limitado, o modelo menos granular é vulnerável à solicitações inesperadas.

Uma base de dados com baixa granularidade (*Drill Down*) resulta em uma quantidade maior de dados armazenado no *data warehouse*, podendo afetar o desempenho das consultas, gerando um tempo de resposta maior. O alto nível de granularidade (*Roll Up*) implica num volume menor de dados e pode não admitir solicitações *Drill Up*, como exemplifica a Figura 6 (TURBAN, 2009).

Figura 6 – Ilustração de Granularidades de Dados.

Fonte: Autor.

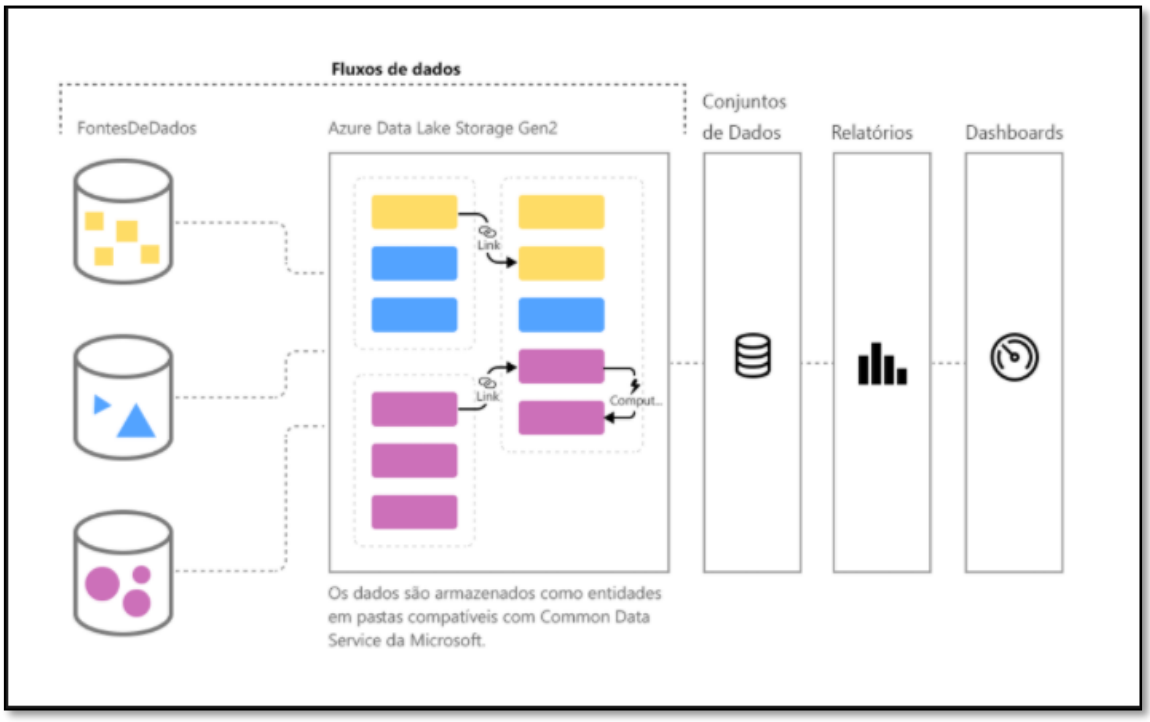
O carregamento de dados em um *data warehouse* é feito a partir do processo ETL – Extração, Transformação e Carga. Esse processo consiste na leitura de dados de um ou mais bancos, converter estes dados extraídos e colocar estes dados no data warehouse (TURBAN, 2009). A ferramentas de ETL transporta dados entre fontes e alvos, documentam e mudam conforme movimentam esses dados.

As ferramentas disponíveis no mercado vêm ganhando inúmeros avanços desde 1990, sendo direcionadas ao utilizador, apresentando uma interface mais intuitiva com mais recursos visuais na apresentação de dados. Uma boa ferramenta ETL deve ser capaz de se comunicar com diversas bases de dados e ler diferentes formatos, um recente exemplo no mercado tem sido o Power BI (FERREIRA, 2010).

2.5. POWER BI

O Power BI é um sistema proprietário disponibilizado pela empresa Microsoft Corporation. Este sistema visa ajudar as organizações a unificar dados de diferentes fontes e tratá-los para analise, através do seu fluxo de dados ilustrado na Figura 7. Os fluxos de dados são usados para transformar, integrar e enriquecer a base de dados. Os dados são armazenados no *Common Data Service*, esse serviço fornece uma linguagem de dados compartilhada para uso de aplicativos e de negócio, incluindo um conjunto de esquema de dados extensíveis e padronizados (POWER BI, 2020).

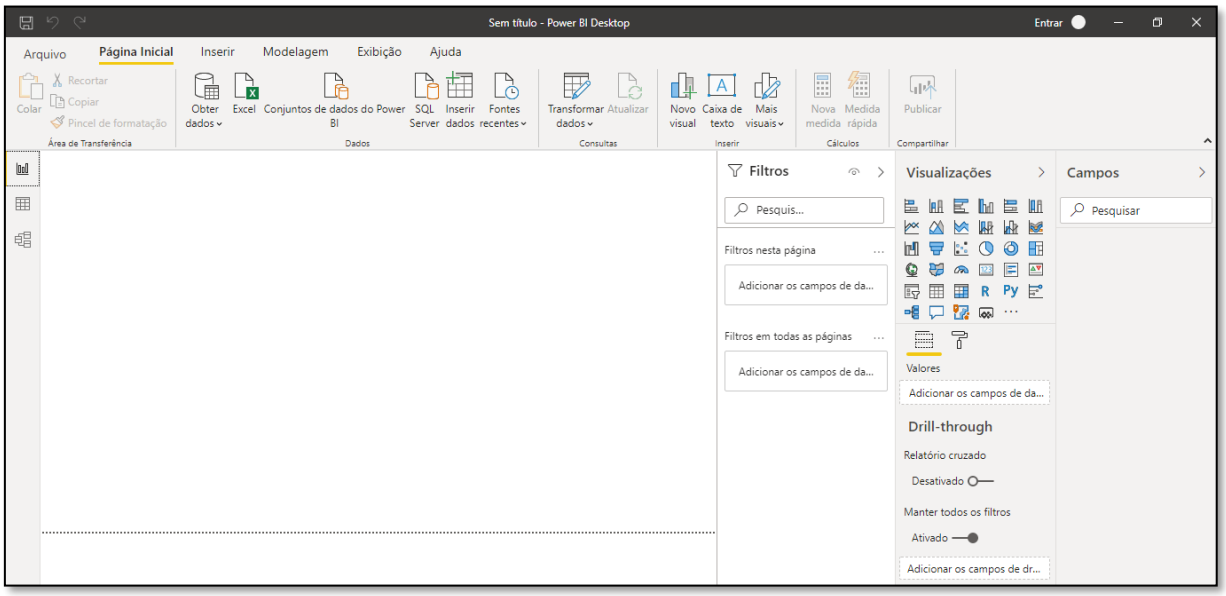
Figura 7 - Fluxo de Dados Power BI.



Fonte: Microsoft Corporation.

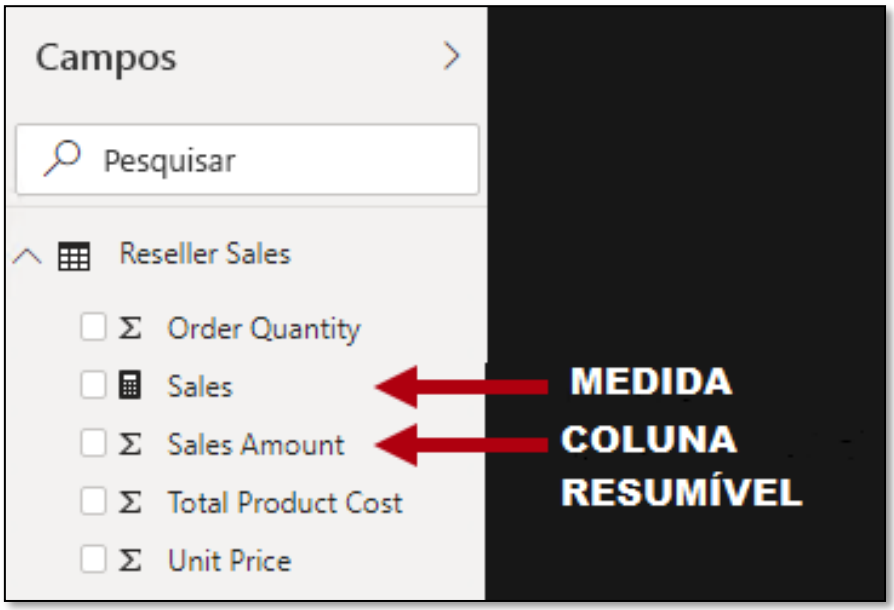
A modelagem dos dados é altamente relevante para o desenvolvimento de modelos otimizados no Power BI. Cada elemento visual no relatório do sistema é resultado de uma consulta enviada ao conjunto de dados do Power BI, na Figura 8 podemos ver a interface do programa, com padrão amistoso ao usuário. As consultas filtram, agrupam e resumem os dados do modelo. A modelagem de dados *Star Schema* é vantajosa nesse sentido, pois as tabelas de dimensões são compatíveis com a filtragem e agrupamento, sendo tabelas com poucas linhas normalmente, já as tabelas de fatos são compatíveis com o resumo, contendo um número maior de linhas no modelo. Uma medida no modelo dimensional é uma coluna da tabela fato, correspondente aos valores a serem resumidos, no Power BI essa medida é expressa por uma formula a DAX (Expressão de Análise de Dados), essa formula agrega as funções SUM, MIN, MAX e AVERAGE, como mostra a Figura 9. É possível trabalhar com design *SnowFlake* fazendo uso das tabelas normalizadas ligadas à tabela fato, ou integrando (desnormalizando) as de origem a uma única tabela do modelo, essa integração é mais benéfica para gerar relatórios, a decisão ideal vai depender do volume de dados (MICROSOFT, 2020).

Figura 8 - Interface Power BI.



Fonte: Autor.

Figura 9 - DAX Power BI.



Fonte: Microsoft Corporation.

2.5.1 Power BI integrado a linguagens de programação

Além dos recursos de medidas para realizar consultas com Power BI, o sistema suporta integração de consultas a partir de linguagens de programação, como linguagem R e a linguagem Python.

A linguagem R é considerada uma linguagem estatística muito utilizada no meio acadêmico, possui várias bibliotecas e código aberto. Um de seus grandes pontos fortes é a facilidade de produzir gráficos de qualidade, incluindo símbolos matemáticos e formuladas quando necessário, (R, 2020).

Com o Power BI é possível visualizar os dados usando o R. Para utilizar este recurso e executar scripts em R é necessário instalar a linguagem R no computador para posteriormente realizar a conexão com o sistema Power BI. Para adquirir o instalador da linguagem basta acessar site oficial (https://www.r-project.org/) baixar os arquivos correspondentes as configurações do seu sistema e realizar a instalação. Após instalar é preciso habilitar o uso da linguagem no software Power BI, especificando o local de instalação da linguagem R no seu computador, ao habilitar as opções de scripts a interface fornece um editor como mostra a Figura 10.

Figura 10 - Interface integrada ao R.

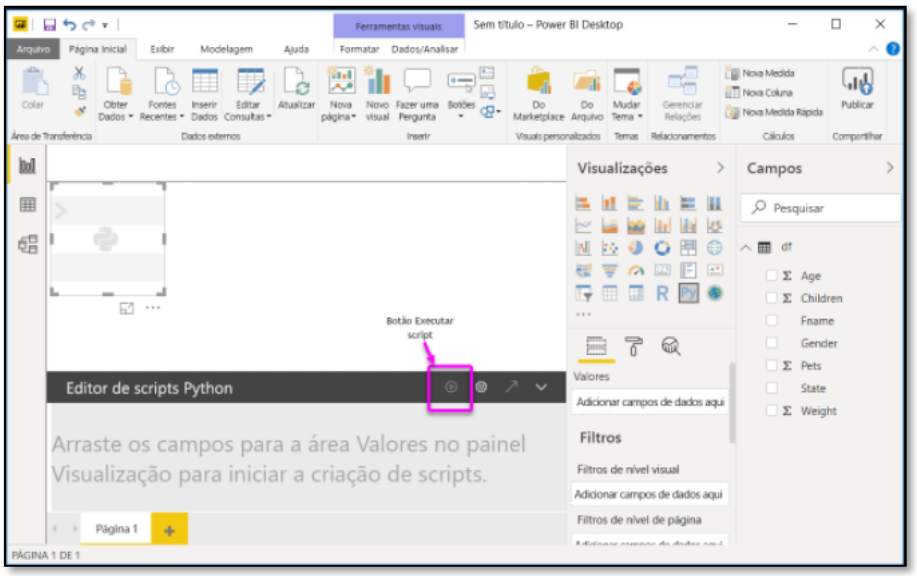
Fonte: Microsoft Corporation.

Python é uma linguagem de alto nível, orientada a objeto, com ampla aplicação, utilizada em construção de sistemas web com frameworks, análise de dados, mineração de dados, inteligência artificial, construção de aplicativos e construção de sistemas desktop. (Python documentação)

Em 2018 a Microsoft Corporation lançou a inclusão da linguagem Python no Power BI por meio de scripts. Para utilização integrada ao Power BI, assim como a linguagem R, é

necessário realizar a instalação da linguagem Python na máquina de trabalho de forma a parte ao Power BI, pode-se obter o instalador da linguagem Python através de seu site oficial (https://www.python.org/) escolhendo a versão desejada. Após a instalação é necessário habilitar a integração no Power BI e indicar o diretório de onde a linguagem foi instalada, na Figura 11 é possível ver interface para utilização dos scripts. Com o Python habilitado, é preciso adicionar duas bibliotecas importantes para utilização de funções a *Matplot* e o *Pandas*.

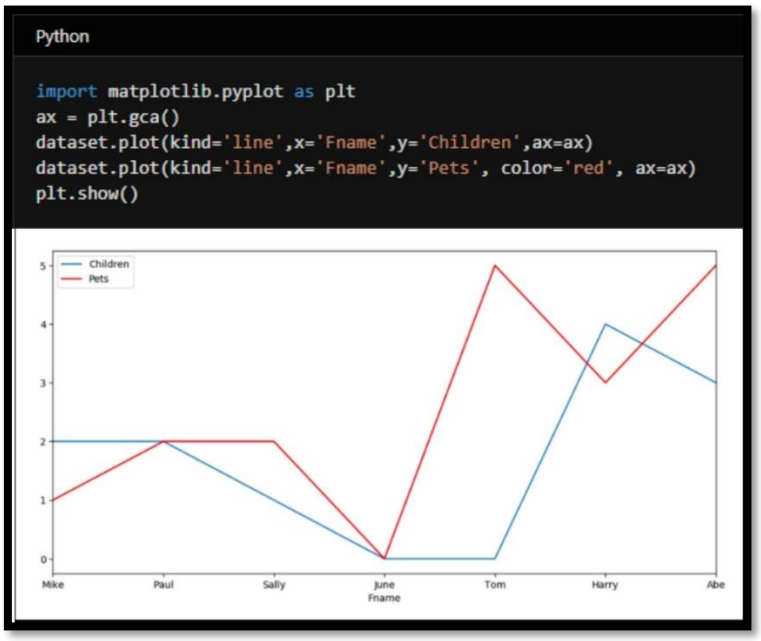
Figura 11- Interface Power BI integrado ao Python.



Fonte: Microsoft Corporation.

O *Matplotlib* é uma biblioteca de plotagem para Python e suas extensões matemáticas numéricas. Ela fornece uma API orientada a objeto para incorporar gráficos a aplicativos como exemplifica a Figura 12 O *Pandas* é uma biblioteca bastante usada para análise de dados em Python, ela dispõe de estruturas e operações para manipular tabelas numéricas e series temporais, suportando trabalhar com dados tabulares e matrizes.

Figura 12 – Recursos Matiplot.



Fonte: Microsoft Corporation.

3 REVISÃO DE TRABALHOS RELACIONADOS

A fim de realizar a análise dos dados públicos referentes ao ProUni, foi feita uma pesquisa buscando identificar e relacionar pesquisas e estudos, que realizaram análise destes mesmos dados, assim como ferramentas utilizadas no processo e fontes de dados relacionados na análise.

3.1. SELEÇÃO DE TRABALHOS

Para realizar a pesquisa de trabalhos relacionados foram realizadas buscas em buscadores da Google. O Google Acadêmico[1] é uma ferramenta que possibilita a localização de artigos, teses, dissertações, livros e outras publicações voltadas a pesquisas. Além da utilização do Google Acadêmico para identificar outros artigos ou ferramentas que não possuem vinculação a academia, utilizou-se o buscador Google.

Para efetuar a pesquisa foram utilizados termos genéricos derivados do objetivo do trabalho, com o intuito de encontrar artigos e publicações brasileiras, uma vez que o ProUni um programa nacional. Os termos utilizados para buscas foram: Análise de dados públicos do ProUni, Análise de dados abertos ProUni, Análise de dados abertos. Foram selecionados 15 trabalhos dos quais 4 foram considerados para utilizar neste trabalho, analisando a forma de análise e tratamento de dados, assim como a ferramenta utilizada neste processo.

3.2. ANÁLISE DE TRABALHOS RELACIONADOS

3.1.1 Um estudo focado ao PROUNI através da análise de dados abertos: período de 2005 até 2016.

O trabalho de Filho (2018), apresenta uma análise qualitativa dos dados do ProUni com gráficos interativos, a fim de verificar o perfil de estudantes que estão sendo contemplados pelo programa. O autor investigou a natureza dos dados, suas características, causas e relações, efetuando a análise e correlacionando com dados abertos do IBGE – Instituto Brasileiro de Geografia e Estatística. Para realizar os gráficos de análise foi utilizada o software *Tableau Desktop* na sua versão gratuita *(Public Edition)*.

[1] https://scholar.google.com.br/

O *Tableau* é uma plataforma de análise visual de dados, voltada a *Business Intelligence*, a fim de facilitar ao usuário a exploração e gerenciamento dos dados e compartilhamento de informações. Sua tecnologia base e patenteada é o VizQL, que expressa os dados visualmente traduzindo ações de arrastar e soltar em consultas de dados por meio de uma interface (Tableau, 2020). Neste software a pasta de trabalho é dividida em planilhas, estas planilhas podem ser usadas em conjuntos para montar painéis.

O autor trabalhou com os arquivos do ProUni referente aos anos de 2005 à 2016 em formato CSV, modificando a nomenclatura das colunas para melhor visualização, após a adequação dos dados foram criadas 5 planilhas, cada uma apresentando resumo do cálculo referente ao item escolhido para analise, sendo este “Beneficiário por Estado”, “Modalidade de Ensino”, “Instituições com mais Beneficiários”, “Nuvem Cursos” e “Total”.

Além da filtragem de dados, o trabalho apresenta um comparativo entre dados do ProUni e os dados nacionais do IBGE referentes ao ano de 2016, ambos os dados apresentam os mesmos critérios de classificação quanto às características do perfil do candidato. Apresentando a participação da população em relação a sua Cor (Figura 13).

Figura 13-Participação da População em Relação a Cor

Cor	IBGE	PROUNI
Branco	44,2%	44,68%
Pardo	46,6%	39,65%
Preto	8,2%	12,58%
Amarelos, indígenas e não declarados	Menos de 1%	3%

Fonte: Silva

O autor define que dado o volume de dados provenientes dos 12 arquivos CSV referentes ao ProUni, seria impossível realizar a análise sem uma ferramenta. O Tableau em sua versão gratuita demonstrou eficiência e recursos suficientes, além de uma rápida curva de aprendizagem. Através dos relatórios da ferramenta identificou-se que o número de beneficiários do ProUni vem tendo aumento desde sua criação, o maior número de bolsas é do tipo integral, além de serem destinadas ao turno noturno, atendendo população de menor renda.

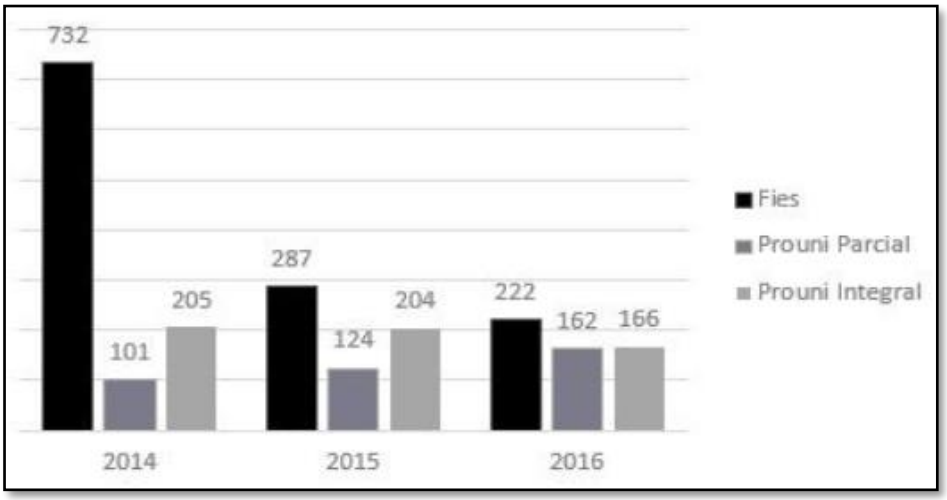
3.1.2 Uma análise temporal de dados aberto do ensino superior utilizando o software de visualização Tableau.

O trabalho de Pereira et al. (2018) apresenta a utilização de ferramentas de *Business Intelligence* para análise de dados abertos, a ferramenta utilizada pelo autor foi o Tableau, os dados utilizados foram coletados do portal do INEP- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, vinculado ao Ministério da Educação (MEC).

Para realizar análise o autor utilizou os dados do ensino superior referente aos anos de 2014 a 2016. A análise foi focada em apresentar as informações públicas de forma clara, possibilitando identificar tendências referentes ao ingresso no ensino superior em categorias públicas e privadas, analisando também programas de financiamento estudantil como FIES e bolsas concedidas pelo ProUni.

Através da utilização do Tableau o autor identificou uma queda das bolsas do tipo integral ofertadas pelo ProUni de 2014 a 2016, enquanto as bolsas do tipo parcial (50%) tiveram um pequeno aumento neste mesmo período (Figura 14)

Figura 14 - Bolsas e Financiamento ofertados



Fonte: PEREIRA, et al. (2018)

De acordo com o estudo realizado, o Brasil teve um aumento de 2,7% alunos matriculados em cursos de nível superior, em 2014 o total foi de 7.839.765 matrículas, já em

2016 foram 8.052.245, é possível ver a quantidade de alunos matriculados por região e por modalidade na Figura 15.

Figura 15- Total de alunos matriculados por categoria e região.

Modalidade	Região	2014	2015	2016
EAD		1.342 mil	1.393 mil	1.494 mil
Presencial	Centro-Oeste	613 mil	618 mil	606 mil
	Nordeste	1.400 mil	1.434 mil	1444 mil
	Norte	451 mil	473 mil	473 mil
	Sudeste	3.100 mil	3.092 mil	3023 mil
	Sul	996 mil	1.021 mil	1009 mil

Fonte: PEREIRA, et al. (2018)

Quanto a soma de alunos matriculados foi analisada a relação, ocupação de gênero por curso (Figura 16), de acordo com os dados o sexo feminino lidera a inserção ao ensino superior com mais de 55% para pouco mais de 44% de matrículas do sexo masculino, onde o curso de Pedagogia tem predominância de matriculas do sexo feminino e o curso de Direto com maior procura pelo sexo masculino.

Figura 16 - Matriculas por Gênero e Curso.

Ano	Rede	Total	Sem	Grad.	Espec.	Mest.	Dout.
2014	Pública	32.570	0	883	3.486	10.052	18.149
	Privada	41.088	1	265	12.058	20.536	8.228
2015	Pública	30.934	0	893	2.198	8.828	19.015
	Privada	43.789	1	106	12.421	22.185	9.076
2016	Pública	31.407	0	644	2.108	8.337	20.318
	Privada	42.376	1	62	11.229	21.476	9.608

Fonte: PEREIRA, et al. (2018)

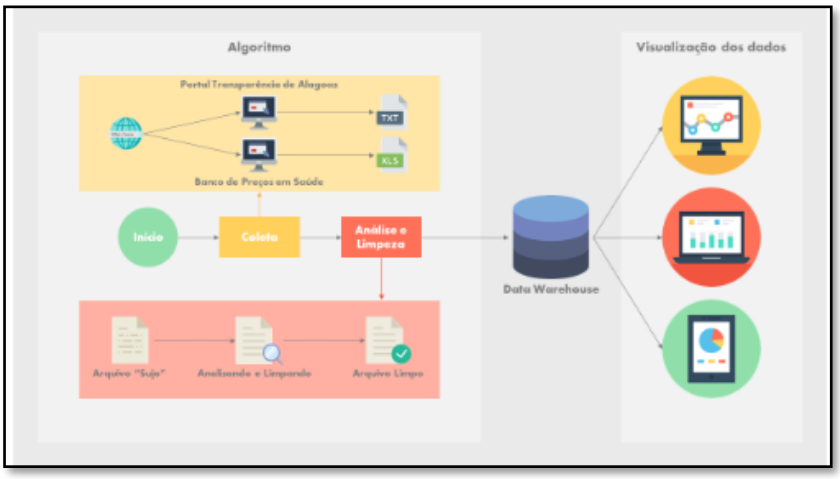
Diante dos resultados o autor de destaca que a utilização do software Tableau contribuiu significativamente para análise, por facilitar a conversão de dados e a visualização gráfica, uma vez que existem poucas informações disponíveis sobre a evolução temporal dos dados do ensino superior no Brasil e o intuito do trabalho foi demonstrar soluções de Inteligência de Negócio em um contexto de transparência publica, utilizando dados abertos.

3.1.3 Um Estudo dos Dados Governamentais Abertos do Estado de Alagoas – COSTA, et al.

O estudo de Costa, et al. (2019) é projeto para monitoramento de dados referentes a gastos públicos para o Estado de Alagoas. O autor defende o desenvolvimento de um *Data Warehouse*, para centralizar os dados públicos do estado, já que a infraestrutura de dados abertos do país não possui ferramentas que integrem estes dados, os dados são disponibilizados de forma setorizada por ministérios do governo. Com os dados centralizados, futuramente estima-se ser possível identificar pontos de ineficiência de administração de recursos públicos.

A metodologia do estudo foi a utilização de uma análise exploratória de dados, buscando identificar características e padrões presentes por meio de visualizações. Para o desenvolvimento foi aplicada técnicas de mineração de dados utilizando a linguagem Python, a linguagem também foi utilizada para realizar a limpeza dos dados, estes foram armazenados em um banco de dados implementado em MySQL, por fim para os relatórios foi utilizado o software Power BI, é possível ver o fluxo de trabalho na Figura 17.

Figura 17 - Fluxo do Estudo



Fonte: COSTA, et al.(2019).

Os relatórios para visualização, foram focados na comparação entre o preço dos itens comprados pelo Governo de Alagoas e os preços correntes do Banco de Preços em Saúde. Para gerar os relatórios está sendo utilizado o software Power BI. Até a publicação deste artigo, os algoritmos estão em fase de teste e ainda não há resultados apresentados.

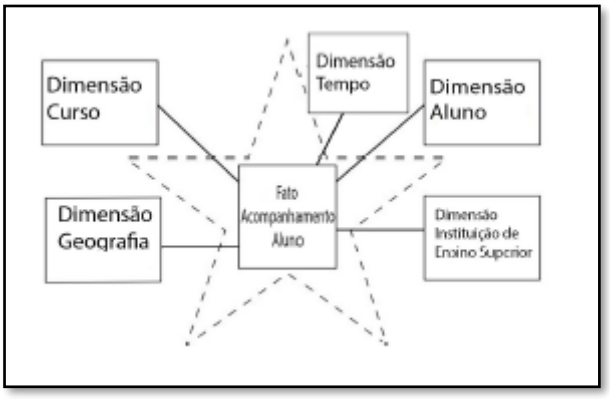
3.1.4 A análise de Dados Abertos sobre o Ensino Superior Brasileiro

De acordo com os autores Magalhães e Cardoso (2016), o intuito de seu trabalho foi realizar análise de dados públicos disponibilizados pelo portal do INEP. Sua análise foi realizada com consultas OLAP e os dados foram inseridos em um banco relacional MySQL,

este banco foi modelado com abordagem dimensional e para os relatórios foi utilizado o sistema *OpenSource – Pentaho*.

O objeto de análise disponibilizado pelo INEP que os autores utilizaram foi censo de ensino superior. Em seu trabalho foram utilizados micro dados do período de 1995 à 2014. Inicialmente foi realizado o processo de ETL que consistiu em baixar os arquivos, após realizada uma modelagem de dados relacional e a partir desta um modelo dimensional o qual recebeu os dados (Figura 18).

Figura 18 - Modelo Dimensional do Banco de Dados



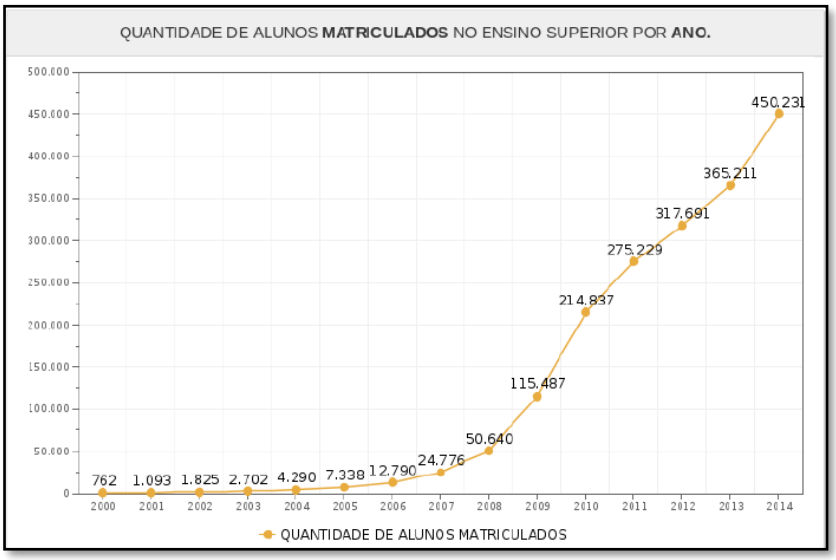
Fonte: MAGALHÃES, CARDOSO (2016)

A partir do modelo foram definidas consultas e analises de possíveis demandas, foram elencadas: Quantidade de alunos no ensino superior no decorrer dos anos; Sexo dos alunos do ensino superior por grupo de faixa etária; Quantidade de alunos matriculados por determinada grande área de conhecimento; Quantidade de alunos matriculados no Ensino Superior por curso; Níveis de titulação de docentes; Raça/cor dos alunos do Ensino Superior por Ano de Ingresso; Análise anual dos alunos do Ensino Superior formados por dada região; Situação acadêmica dos alunos no Ensino Superior por região.

Com a devida implementação a última etapa foi o projeto de aplicação de BI, para esta etapa foi utilizado o *Software Pentaho*. Com a ferramenta foi realizada as analises previstas pelos autores como por exemplo:

Nesta análise verificou-se um considerável aumento quantitativo de alunos matriculados após o ano de 2008 (Figura 19), o autor salienta possíveis motivos para o fenômeno, como, aumento de investimentos na educação e programas de financiamento estudantil.

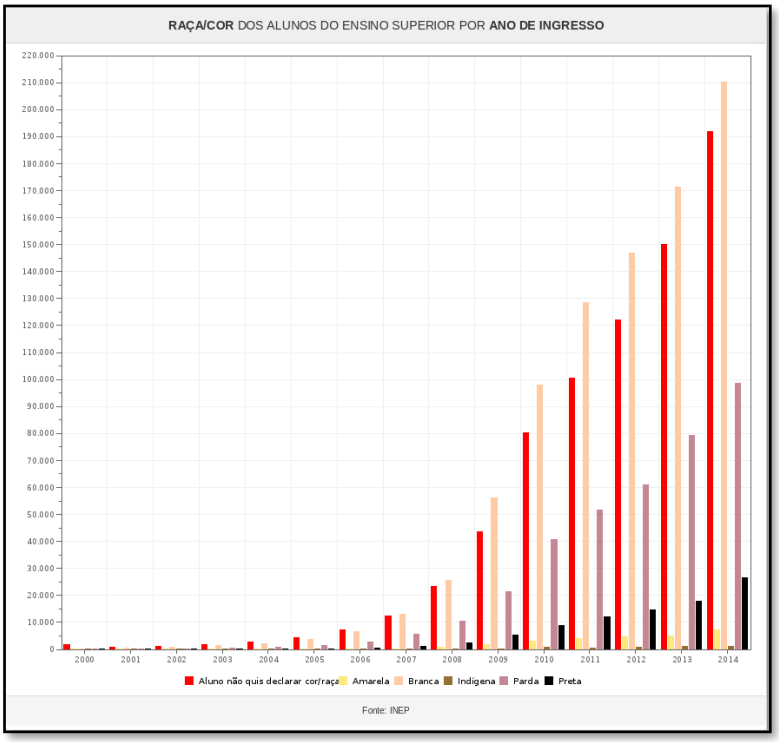
Figura 19- Alunos matriculados por ano



Fonte: MAGALHÃES, CARDOSO (2016)

O autor salienta o gradativo aumento da quantidade de negros e pardos nas instituições de ensino superior, apontando para o fenômeno os programas de cotas, contudo prevalece a quantidade de brancos e alunos que não quisera, declarar sua raça/cor. Já a raça/cor indígena se manteve constante ao longo de 2010 e 2014, Figura 20.

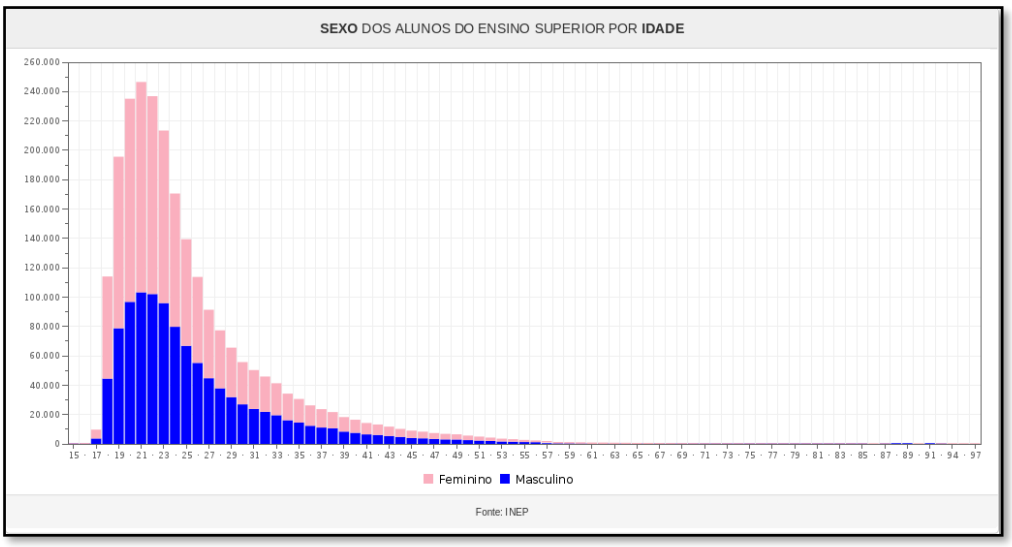
Figura 20- Raça/cor estudantes por ano de ingresso



Fonte: MAGALHÃES, CARDOSO (2016)

Os autores verificaram que a quantidade de alunos do sexo feminino prevalece no ensino superior, de maneira que a quantidade de mulheres na faixa de 21 anos são o dobro se comparado a quantidade de homens que cursam ensino superior, ressaltando que o ingresso de ambos ocorre em sua maioria na faixa etária entre 18 e 25 anos, representando mais de 75 % da amostra (Figura 21).

Figura 21 - Ingresso por sexo e faixa etária



Fonte: MAGALHÃES, CARDOSO (2016)

Com base em todas as etapas de seu trabalho os autores concluíram que não é uma tarefa fácil processar o volume massivo e crescente de dados gerados pelo governo brasileiro. Sua abordagem demonstra a eficácia da utilização de softwares livres neste tipo de processo e espera ajudar com insights para o estabelecimento de políticas públicas mais eficazes, além do concluído no trabalho os autores propõem para trabalhos futuros a utilização de banco de dados NoSQL, aplicação de mineração de dados e a inclusão de dados oriundos de outras áreas governamentais.

4 ESTUDO DE CASO PROPOSTO

4.1. DADOS PUBLICOS

Os principais objetivos da política brasileira de dados abertos é a transparência, participação social, além do melhoramento de serviços e integridade pública. O Portal de Dados Abertos, foi uma ação do existindo Ministério do Planejamento (atualmente Ministério da Economia), a plataforma foi desenvolvida pela sociedade e para ela, com ampla participação e totalmente aberta. Como estratégia da política de dados abertos, o governo incentiva os estados e municípios para implantação de políticas locais de dados abertos (GOVERNO DIGITAL, 2020).

No Portal Brasileiro de Dados abertos é possível o acesso e busca por dados públicos do país, cada órgão é responsável pela catalogação de seus dados que serão publicados na internet. Essa catalogação é realizada por pessoas de cada órgão que participam da INDA - Infraestrutura Nacional de Dados Abertos. A INDA institui a, de acordo com a sua normativa (Instrução normativa n° 4, 12 de abril de 2012), institui o Plano de Ação Nacional sobre o Governo Aberto, que estabelece o compromisso de o governo implantar a Infraestrutura Nacional de Dados Abertos, considerando que o direito à informação é um fundamento básico a democracia e que para o cidadão exercer esse direito plenamente lhe deve ser facilitado o acesso, considerando a utilização de meios eletrônicos (INDA, 2012).

O Decreto s/n° de 15 de setembro de 2011 atribuído na normativa foi revogado pelo Decreto n° 10.160 de 19 de dezembro de 2019, gerando algumas mudanças na política de dados abertos e sua disponibilização, antes conduzidas pelo Ministério da Economia, agora serão geridas pela Controladoria-Geral da União (CGU) (DECRETO 10.160, 2019). Desde 2012 à Lei de Acesso à Informação (LAI), criou mecanismos que facilitem o acesso a informações públicas sem necessidade de apresentar algum motivo, vigente para os três Poderes da União, estados e municípios, em 2019, no entanto houve alteração nas regras de aplicação da LAI, ampliando o número de autoridades que podem impor sigilo ultrassecreto a dados e documentos do governo, este tipo de sigilo prevê a restrição das informações por 25 anos (DECRETO 9.716, 2019).

Apesar dos avanços na distribuição de informações, ainda não há um local que concentre todos os Planos de Dados de Abertos do Governo (PDAs), sendo assim é necessário buscar os de dados referente a cada Órgão do Governo em seus respectivos sítios.

4.2. O PROGRAMA PROUNI

O Programa Universidade para Todos (ProUni) foi criado em 2004. A finalidade deste programa é subsidiar bolsas de estudos parciais e integrais de cursos de graduação em instituições de ensino superior privadas, com o intuito de viabilizar a capacitação e educação superior a jovens que se enquadrem no programa. Para participar os egressos do ensino médio da rede pública ou particular, devem ter renda familiar per capita máxima de até três salários-mínimos. As instituições, por sua vez, que aderem ao programa recebem isenção de tributos (MEC, 2020).

Os dados referentes ao preenchimento de vagas do ProUni são produzidos pela Diretoria de Políticas e Programas de Graduação (DIPES) da Secretaria de Educação Superior (SESu) do Ministério da Educação (MEC). O processo de criação do conjunto de dados é feito por "Barramento de Dados" e mantido pelo Escritório de Gestão de Dados e Informações Estratégicas (EGDIE/SE/MEC), os são modelados de forma multidimensional, utilizando mecanismos de ETL e servidor web. Os dados são disponibilizados em sua maioria em formato CSV, com frequência de atualização anual, apresentam granularidade geográfica municipal, cobrindo todos os municípios brasileiros que tenham beneficiários do Programa ProUni, a granularidade temporal dos dados também é anual (DADOS ABERTOS, 2020).

Cada arquivo CSV disponibilizado corresponde a 1 ano do programa, cada linha do arquivo representa uma linha da tabela e as colunas são separadas por virgula. A Secretaria de Logística e Tecnologia da Informação (SLTI), optou por esse formato, pela facilidade de manipulação das planilhas, mesmo por aqueles sem conhecimentos técnicos de desenvolvimento de software, além da facilidade de consumo de seu conteúdo por aplicações. Os campos dos arquivos consistem em bolsas concedidas pelo ProUni por ano, contendo a região; UF; município; instituição de educação superior; curso; modalidade de ensino (presencial ou EAD); turno e tipo de bolsa. Detalhamento do perfil dos beneficiários do ProUni por sexo; raça/cor; data de nascimento e pessoas com deficiência (PDA MEC, 2016).

4.3. FLUXO DE TRABALHO

Serão utilizados dados disponibilizados pelo Ministério da Educação (MEC), referente ao preenchimento de vagas do ProUni em todo território nacional referente aos anos de 2005 a 2019. Os dados se encontram em formado CSV, esse formato é caracterizado por apresentar dados separados por ponto e vírgula, disponíveis em (http://dados.gov.br/dataset/mec-prouni). Cada arquivo possui os mesmos campos descritos na tabela 1.

Tabela 1 - Dicionário de Dados do PROUNI

Dicionário PROUNI			
NOME	**CAMPO**	**TIPO**	**DESCRIÇÃO**
Ano da concessão da bolsa	ANO_CONCESSAO_BOLSA	Numérico	Ano da concessão da bolsa ProUni (início da vigência).
Código do e-MEC da IES que concedeu a bolsa	CODIGO_EMEC_IES_BOLSA	Numérico	Código do e-MEC referente a IES que concedeu a bolsa ProUni.
Nome da IES	NOME_IES_BOLSA	Alfanumérico	Nome/Razão Social da Instituição de Ensino Superior que concedeu a bolsa ProUni.
Tipo da Bolsa	TIPO_BOLSA	Alfanumérico	Descrição do tipo da bolsa concedida ao beneficiário do ProUni (integral

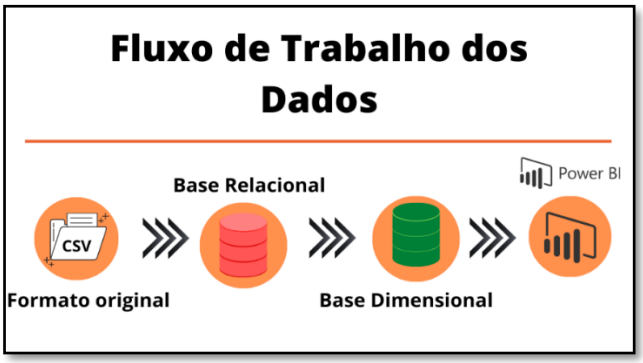
			– 100% ou parcial – 50%)
Modalidade de ensino	MODALIDADE_ENSINO_BOLSA	Alfanumérico	Descrição da modalidade de ensino da bolsa concedida ao beneficiário do ProUni (presencial ou ensino à distância – EAD).
Nome do Curso	NOME_CURSO_BOLSA	Alfanumérico	Nome do curso do beneficiário da bolsa ProUni.
Turno do Curso	NOME_TURNO_CURSO_BOLSA	Alfanumérico	Descrição do turno do curso do beneficiário da bolsa ProUni
CPF do beneficiário	CPF_BENEFICIARIO_BOLSA	Alfanumérico	CPF do beneficiário da bolsa ProUni
Sexo do beneficiário	SEXO_BENEFICIARIO_BOLSA	Alfanumérico	Sexo informado pelo beneficiário da bolsa ProUni
Raça/Cor	RACA_BENEFICIARIO_BOLSA	Alfanumérico	Raça/Cor informado pelo beneficiário da bolsa ProUni.
Data de nascimento do beneficiário	DT_NASCIMENTO_BENEFICIARIO	Data	Data de nascimento do beneficiário da bolsa ProUni.

Indicação se o beneficiário é portador de deficiência	BENEFICIARIO_DEFICIENTE_FISICO	Alfanumérico	Indicação se o beneficiário da bolsa ProUni é portador de algum tipo de deficiência (sim ou não).
Região	REGIAO_BENEFICIARIO_BOLSA	Alfanumérico	Nome da região de residência do beneficiário da bolsa ProUni
UF	SIGLA_UF_BENEFICIARIO_BOLSA	Alfanumérico	Sigla da UF de residência do beneficiário da bolsa ProUni.
Município	MUNICIPIO_BENEFICIARIO_BOLSA	Alfanumérico	Nome do Município de residência do beneficiário da bolsa ProUni.

Fonte: Ministério da Educação

Com a definição dos dados a utilizar, foram definidas as etapas necessárias e quais recursos de *Business Intelligence* seriam utilizados. A abordagem seguida para trabalhar os dados será a de Kimball (2002), criando um base operacional para receber os dados com tratamento prévio, e a partir dela, alimentar a base dimensional utilizada para análise posteriormente. Para realizar a análise foi utilizado o software Power BI, a fim de testar seus recursos e limitações incluindo a integração com linguagem Python na criação de relatórios (Figura 22).

Figura 22- Fluxo de Trabalho



Fonte: Autor

O primeiro passo foi padronizar os dados, os 15 arquivos apresentaram divergência nos valores correspondentes as colunas SEXO_BENEFICIARIO_BOLSA, onde alguns arquivos possuíam a definição de “Masculino” ou “Feminino”, enquanto outros possuíam apenas o indicativo “M” ou “F”; na MODALIDADE_ENSINO_BOLSA, as divergências eram entre a nomenclatura “Educação à Distância” e “EAD”; BENEFICIARIO_DEFICIENTE_FISICO, onde os valores possíveis eram “Sim” ou “Não”, alguns arquivos apresentavam “S” ou “N”, Foi realizada uma limpeza neste sentido deixando os valores padronizados escritos por extenso em todos os arquivos, além de adicionar o valor “NÃO INFORMADO” para os campos em brancos, a fim de prevenir problemas no processamento dos dados e na análise por campos com valores nulos.

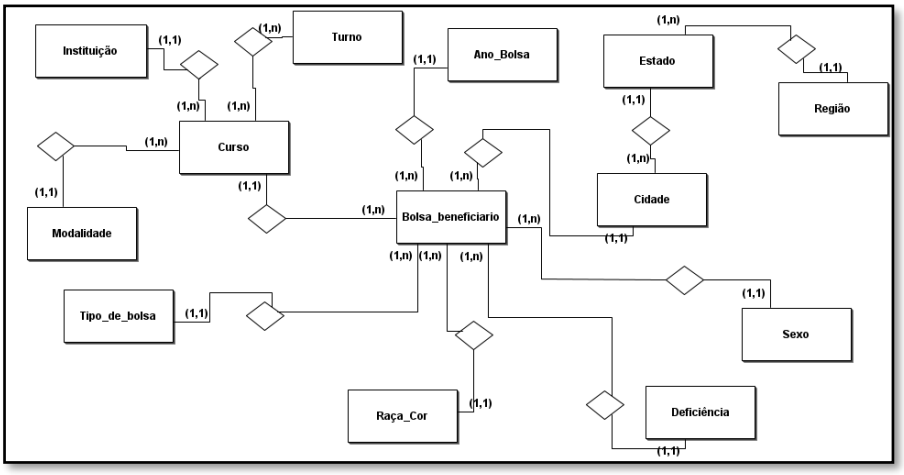
4.4. MODELAGEM DE DADOS

Seguindo a abordagem de Kimball (2002), foi implementada uma base de dados relacional (Figura 23) como uma base operacional, essa base será utilizada como suporte para criação e alimentação de dados da base de dados dimensional. Para implementação foi realizado um diagrama Entidade-Relacionamento, definindo as tabelas e seus relacionamentos, nesta base de dados será inserida cada linha dos arquivos CSV. Os valores no banco vão apresentar normalização evitando a redundância de dados.

Foi realizado o tratamento do campo referente a data de nascimento do beneficiário da bolsa, a partir do ano de nascimento e ano de ingresso do participante, foi definida sua idade, sendo assim na base relacional o campo referente a data de nascimento foi substituído pelo campo idade, esse novo valor foi considerado interessante para realizar possíveis análises,

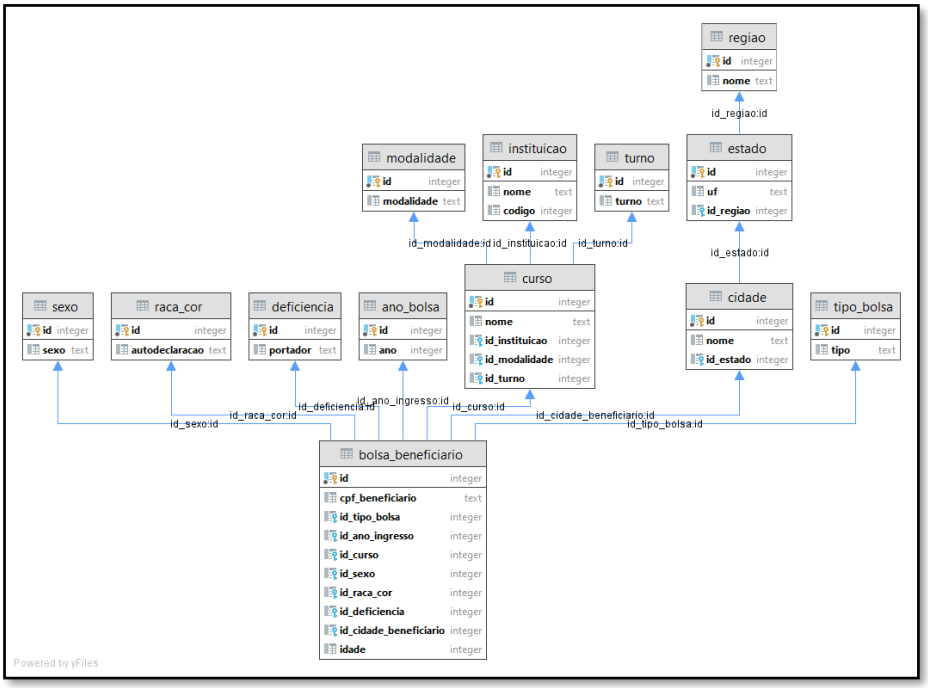
sendo uma métrica à mais a respeito do perfil dos candidatos contemplados com as bolsas (Figura 24).

Figura 23 - Modelo Conceitual da Base de Dados Relacional



Fonte: Autor

Figura 24 - Diagrama da Base de Dados



Fonte: Autor.

A partir do modelo conceitual Entidade-Relacionamento da base de dados relacional, foi feita a definição da tabela fato para a base dimensional, a estrutura da tabela fato definirá a granularidade dos dados, a granularidade irá afetar diretamente o volume de dados da tabela, para definir essa estrutura foi feita uma matriz de necessidade (Dimensão x Indicador), nessa matriz as linhas são compostas pelos Indicadores, já nas colunas da matriz são compostas pelas Dimensões (OLIVEIRA, 2016). Kimball (2002), propõem em sua abordagem uma matriz de

barramento, onde as colunas recebem as dimensões e as linhas os *Data Marts*, essa matriz não se mostrou a mais adequada para esta estrutura de dados, que consiste na criação de um *Data Mart* no contexto de dados públicos do Ministério da Educação.

Ao preencher a matriz (Figura 25) a primeira coisa a se analisar é a se a relação de hierarquia entre as dimensões, desta forma ficou definida a hierarquia entre as dimensões Curso e Instituição, e outra entre Cidade, Estado e Região. Com a presença de hierarquia, a melhor modelagem a ser utilizada é *SnowFlake*, cada indicador da matriz será uma tabela fato, suas relações são independentes entre si, mas suas relações com as dimensões são definidas pelas intersecções da matriz.

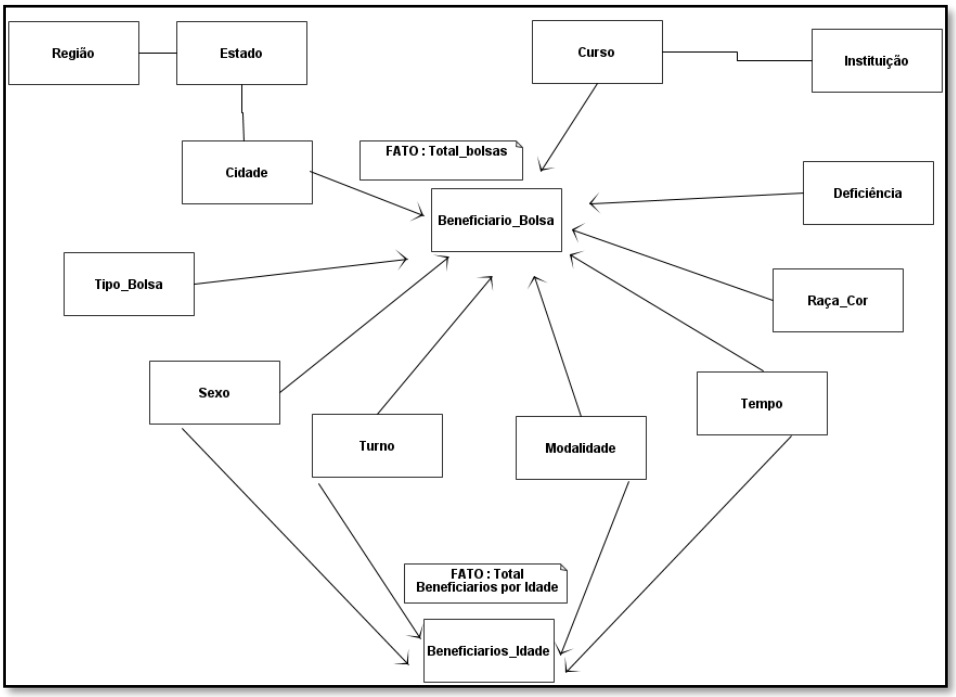
Figura 25 - Matriz de Necessidade

	turno	sexo	curso	instituição	cidade	região	estado	deficiencia	raça_cor	tipo_bolsa	ano
total_beneficiario	x	x	x	x	x	x	x	x	x	x	x
total_por_idade	x	x								x	x

Fonte: Autor.

A tabela fato referente ao total de beneficiários irá apresentar uma baixa granularidade, seu volume de dados será maior, enquanto a tabela referente ao total por idade apresentará um volume menor e uma granularidade maior, isso se deve a quantidade de intersecções, pois elas afetam diretamente o agrupamento de dados. A modelagem da base dimensional ficou como a representação conceitual apresentada na Figura 26.

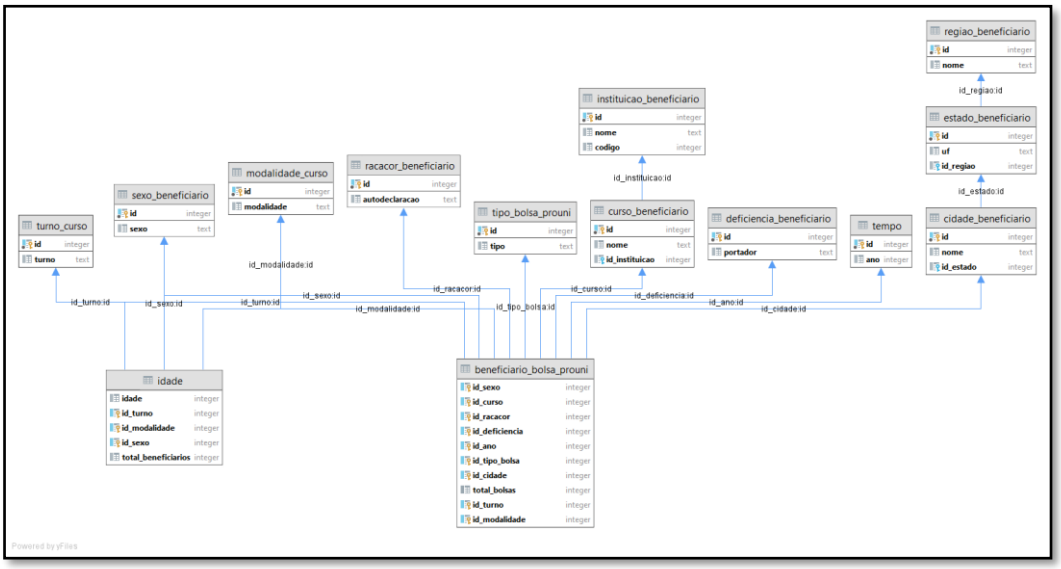
Figura 26 - Modelo Conceitual Base Dimensional



Fonte: Autor.

Este modelo possuí quatro dimensões que compartilham reação com as duas tabelas fatos, são elas: sexo, turno, modalidade e tempo. A dimensão sexo possui os valores masculino e feminino, a dimensão turno é referente ao turno das aulas do curso que concedeu bolsa, sendo integral, matutino, vespertino, noturno e curso à distância. Modalidade é referente à ensino presencial ou educação à distância, a dimensão tempo será a nossa referência temporal do *data mart*, possuindo o valor correspondente aos anos das concessões de bolsas, o detalhamento de dos campos de cada tabela pode ser observado na Figura 27.

Figura 27- Diagrama Base Dimensional



Fonte: Autor

4.5. PROCESSAMENTO DE DADOS

Para implementação das bases de dados foi utilizado o banco de dados PostgreSQL. Os arquivos CSV foram lidos e processados por um sistema desenvolvido na linguagem de programação Java com auxílio dos frameworks Spring Boot e Spring Data JPA. O paradigma utilizado foi orientação a objeto, nesse sistema as entidades do banco foram representadas por abstrações da linguagem utilizada, as classes. O banco utilizado foi o PostgreSQL, um banco relacional gratuito, apesar de transacional com a modelagem dimensional atende as necessidades de *Data Warehouse*.

Além da estrutura de abstração foi implementada a leitura dos arquivos CSV (Figura 28), esse processo consistiu em um código que lê cada linha do arquivo, essa alinha é quebrada por cada coluna do arquivo, essa coluna é identificada pelo ponto e vírgula, desta forma os dados são identificados e triados para o objeto da classe responsável, recebendo tratamento necessário e sendo salvo na base de dados. Para cada linha do arquivo é criado um objeto para receber os valores e invocar seus métodos de manipulação.

Figura 28 - Código de Leitura de Arquivo CSV

```
@Service
public class LeituraDocumentoService {
    @Autowired
    SalvarDadosService salvarDadosService;

    public void printMensage(String[] args) throws FileNotFoundException, UnsupportedEncodingException {
        System.out.println("entrou aqui caralha");
        int count = 0;
        BufferedReader arqIn = new BufferedReader(new InputStreamReader(new FileInputStream( name: "C:\\codificados prouni\\2016.csv")));

        Scanner leitor = new Scanner(arqIn);
        leitor.nextLine();
        while (leitor.hasNext()) {

            String linhaDoArquivo = new String();
            linhaDoArquivo = leitor.nextLine();
            String[] valoresEntreVirgula = linhaDoArquivo.split( regex: ";");
        salvarDadosService.saveline(valoresEntreVirgula[0],valoresEntreVirgula[1], valoresEntreVirgula[2],
                valoresEntreVirgula[3], valoresEntreVirgula[4], valoresEntreVirgula[5],
                valoresEntreVirgula[6], valoresEntreVirgula[7], valoresEntreVirgula[8],
                valoresEntreVirgula[9], valoresEntreVirgula[10], valoresEntreVirgula[11],
                valoresEntreVirgula[12], valoresEntreVirgula[13], valoresEntreVirgula[14]);
count ++;
        }
```

Fonte: Autor.

Para salvar os dados na base relacional foram utilizados métodos simples já implementados pelo framework Spring Data JPA. A tabela “bolsa_beneficiario” corresponde a cada linha de cada arquivo CSV que foi processado, importante ressaltar que a tabela “curso” e “instituição” houve o cuidado para que não apresentasse duplicidade nos valores, referente a tabela “curso” para evitar o uso de uma de ligação entre curso e turno, um mesmo curso com turnos distintos, foi inserido repetidamente com a indicação do turno especificado.

Tabela 2 - Volume de dados Base Relacional

Base Relacional	
Tabela	**Quantidade linhas**
bolsa_beneficiario	2.692.540
instituição	2.459
curso	49.572

Para inserção na base dimensional foram efetuadas buscas na base relacional, a inserção nas tabelas fatos necessitaram consultas SQL mais complexas, para isso foi utilizado do recurso de *query* nativa do *framework Spring Data JPA*, a *query* SQL foi encapsulada por um método em uma classe de implementação do repositório JPA, esse método retorna um *array* de *Object*, porque esse tipo genérico da linguagem Java, permite carregar qualquer tipo de dado seja ele do tipo texto, número ou booleano. Para realizar as inserções na tabela fato focada no total de bolsas, foi aplica a *query* (Figura 29).

Figura 29 - Query para Inserção em Tabela Fato com Foco no Total de Bolsas

```
Select bolsa_beneficiario.id_tipo_bolsa,curso.id_turno,curso.id_modalidade,bolsa_beneficiario.id_sexo,
       bolsa_beneficiario.id_raca_cor,bolsa_beneficiario.id_deficiencia,bolsa_beneficiario.id_curso,
       bolsa_beneficiario.id_cidade_beneficiario,estado.id AS estado,count(*)
       from modalidade, turno, instituicao,bolsa_beneficiario,curso, sexo,raca_cor,deficiencia, tipo_bolsa,
       cidade,estado where sexo.id= bolsa_beneficiario.id_sexo AND raca_cor.id= bolsa_beneficiario.id_raca_cor
       AND deficiencia.id = bolsa_beneficiario.id_deficiencia AND curso.id = bolsa_beneficiario.id_curso
       AND tipo_bolsa.id= bolsa_beneficiario.id_tipo_bolsa AND cidade.id = bolsa_beneficiario.id_cidade_beneficiario
       AND instituicao.id= curso.id_instituicao AND turno.id=curso.id_turno AND modalidade.id= curso.id_modalidade
       AND cidade.id_estado=estado.id AND bolsa_beneficiario.id_ano_ingresso=?
GROUP BY bolsa_beneficiario.id_tipo_bolsa,curso.id_turno,curso.id_modalidade,bolsa_beneficiario.id_sexo,
bolsa_beneficiario.id_raca_cor,bolsa_beneficiario.id_deficiencia,bolsa_beneficiario.id_curso,
bolsa_beneficiario.id_cidade_beneficiario,estado.id HAVING Count(*) > 0;
```

Fonte: Autor.

A consulta retorna o conjunto de dados com o total do agrupamento desde que ele apresente valor maior que zero, com esse conjunto de dados são realizadas buscas em casa entidade da base relacional, para posteriormente buscar na base dimensional o correspondente e assim inserir os identificadores e o total, compondo assim a tabela fato.

Para compor a tabela fato com ênfase em idade, foi utilizada a query da figura 30 executada na base relacional, a partir do conjunto de dados resultando foi aplicado o mesmo processo da composição da tabela fato focada em total de bolsas, foi feita a busca na base relacional, seu equivalente na base dimensional e assim agrupado o conjunto a ser inserido na tabela fato da base dimensional focada em total de beneficiários por idade.

Figura 30 - Query para Inserção em Tabela Fato com foco em Idades

```
select bolsa_beneficiario.idade,curso.id_turno,curso.id_modalidade,Count(*),bolsa_beneficiario.id_sexo
from sexo,bolsa_beneficiario,modalidade,turno, curso where bolsa_beneficiario.id_curso=curso.id AND
curso.id_turno=turno.id AND curso.id_modalidade=modalidade.id  AND bolsa_beneficiario.id_sexo=sexo.id
AND bolsa_beneficiario.id_ano_ingresso=?
GROUP BY bolsa_beneficiario.idade,curso.id_turno,curso.id_modalidade, bolsa_beneficiario.id_sexo;
```

Fonte: Autor.

Ao termino das inserções na base dimensional a tabela fato com maior granularidade focada no total de bolsas apresentou um total de 1.859.355 linhas referente ao período de 2005 à 2019, contudo esse grande volume de dados mostrou-se um problema em teste prévios para implementação da próxima etapa de trabalho, devido as configurações da máquina utilizada para realizar o procedimento, os recursos de hardware não foram suficientes para trabalhar com esse volume de dados no Power BI, apresentando baixo desempenho e erros de sistema. Dado esse empecilho foi realizado um backup em nuvem da base dimensional e foi definido que a análise no Power BI seria efetuada usando apenas 2 anos 20016 e 2017.

4.6. ANÁLISES COM POWER BI

Os dados foram carregados para o software Power BI através de sua compatibilidade de conexão com banco de dados PostgreSQL. O intuito do uso de software é utilizar seus recursos de análises com ênfase em sua integração com a linguagem Python. Segundo a documentação do software oferecido pela Microsoft, existem limitações conhecidas em relação a integração com Python, os dados a serem plotados com recursos visuais da linguagem não devem exceder 150.000 linhas, caso ocorra, a visualização apresentará inconsistência em relação aos dados existentes, além de existir um limite de tempo de cálculo de cinco minutos para operações (MICROSOFT, 2020).

Com as limitações existentes no uso da linguagem Python no software, foi decidido utilizar as operações de Python nos dados da tabela fato referente ao total de bolsas por idade que possui uma granularidade maior de dados e um volume menor, não impactando nas limitações do software. Os dados da tabela fato referente ao total de bolsas serão trabalhados com recursos de visualização e análise do próprio Power BI.

Tabela 3 - Descrição de volume tabelas FATOS

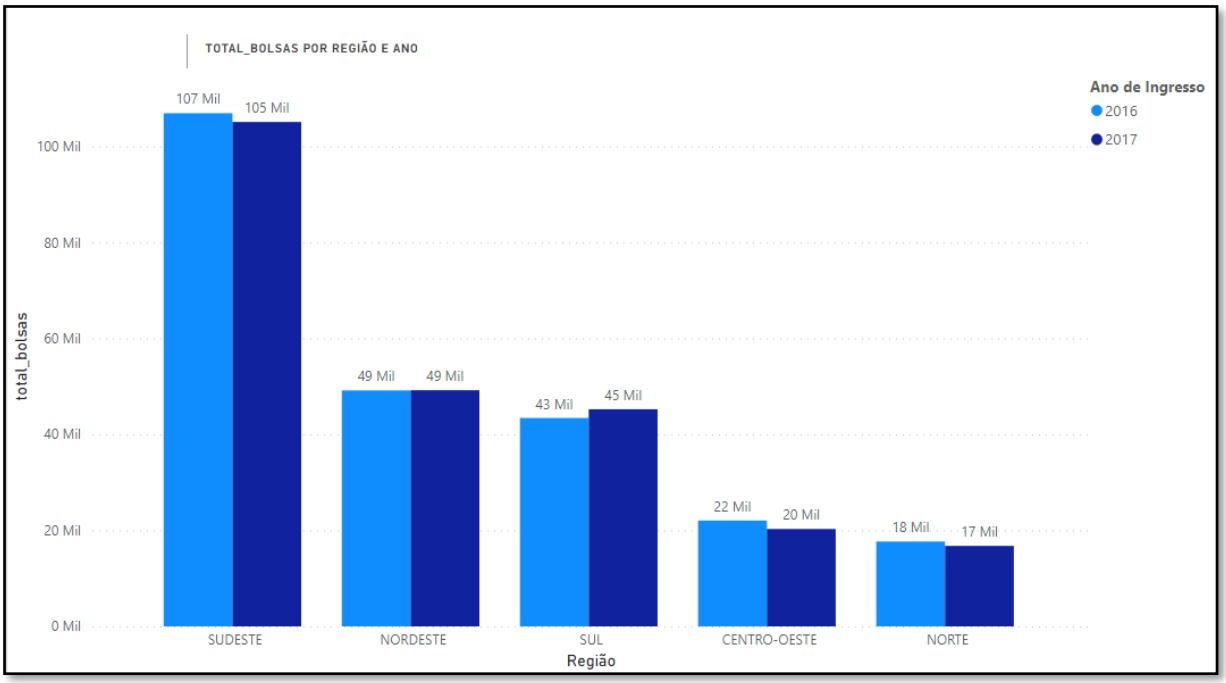
Base Dimensional usada para análise			
Tabela FATO	**FOCO TABELA**	**Granularidade**	**Linhas**
beneficiário_bolsa_prouni	Total de beneficiários	baixa	328.682
idade	Total de beneficiários por idade	alta	924

4.6.1 Visualizações nativas do Power BI

Utilizando a tabela fato com foco no total de bolsas que apresenta baixa granularidade (beneficiário_bolsa_prouni), foi utilizado recursos de visualização nativas do programa utilizado, devido ao grande volume de linhas. O foco da análise com esses recursos foi entender comportamentos da ocupação de vagas dos 2 anos selecionados e a ocupação de acordo com perfil dos candidatos.

A primeira visualização (Figura 31) foi a fim de comparar o total de bolsas concedidas por regiões do Brasil nos anos analisados 2016 e 2017, é possível ver uma queda na ocupação de vagas na maioria das regiões em relação ao ano de 2016, exceto na região Sul que apresentou um crescimento em 2017 se comparado com a ocupação no ano de 2016. Apesar dos dados disponibilizados mostrarem um declínio no ano de 2017, de acordo com dados estatísticos do INEP a rede de ensino privado teve uma expansão de 7,3%, esse valor considera os demais programas de financiamento estudantil e demais vagas (CENSO ENSINO SUPERIOR, 2018).

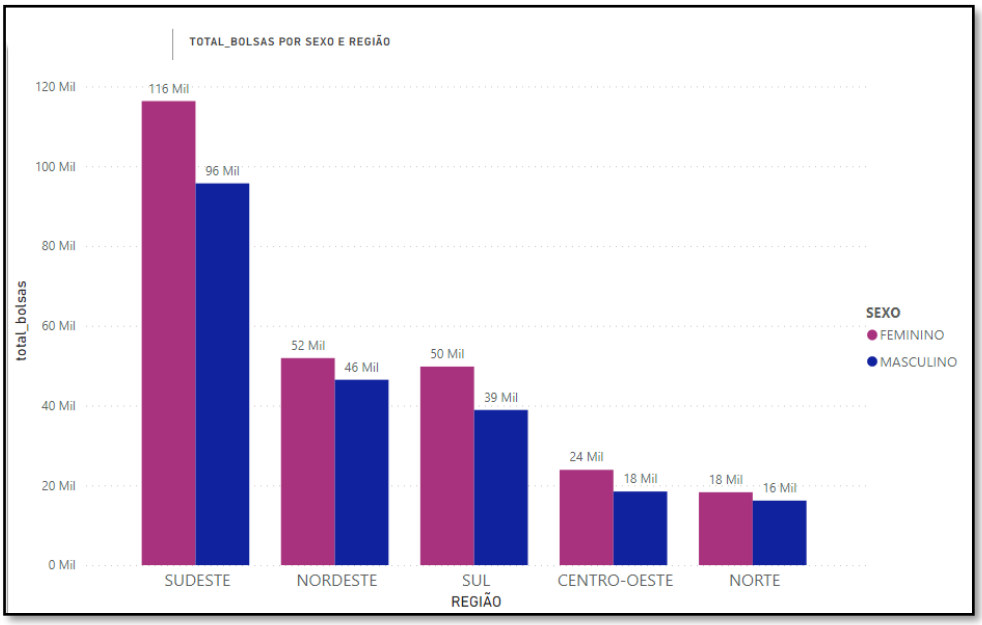
Figura 31- Total de Bolsas por Região ao Ano



Fonte: Autor.

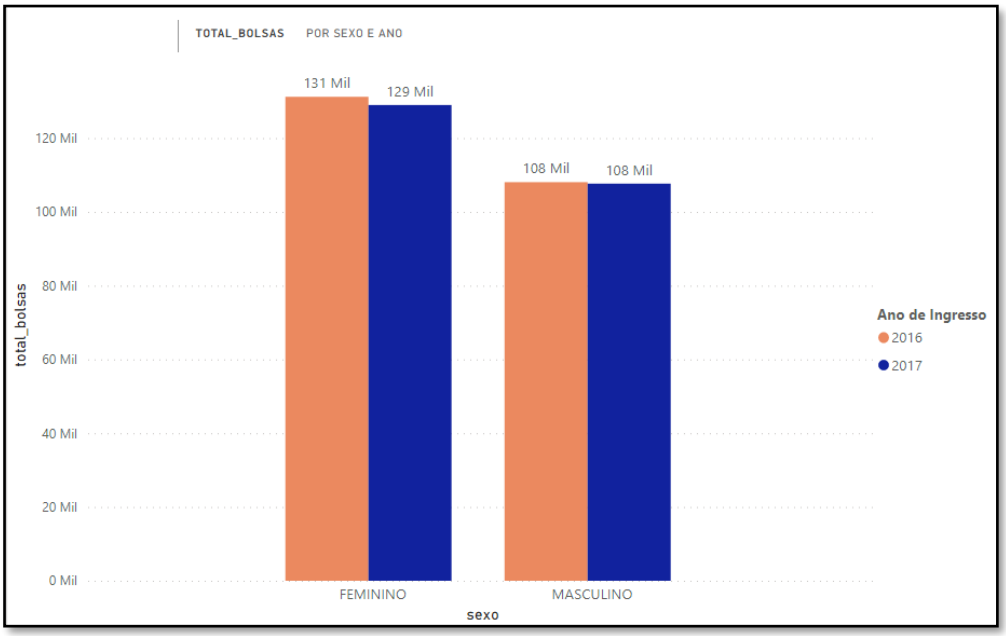
Outra visualização interessante foi analisar a distribuição de candidatos do sexo feminino e masculino por regiões, foi utilizado o total referente aos dois anos para gerar o gráfico, através dele (Figura 32) é possível ver a predominância de ocupação de bolsas pelo público feminino. Para analisar a distribuição por sexo de cada ano, a fim de ver se a predominância feminina é algo é comportamento comum, na figura 33 esse fato se mostra verdade dentro do período que foi analisado, ainda segundo dados do Censo de Ensino Superior (2018) o público feminino é maioria tanto na modalidade presencial quanto ensino a distância.

Figura 32 - Total de Bolsas por Sexo e Região



Fonte: Autor.

Figura 33 - Total de Bolsas por Sexo ao Ano

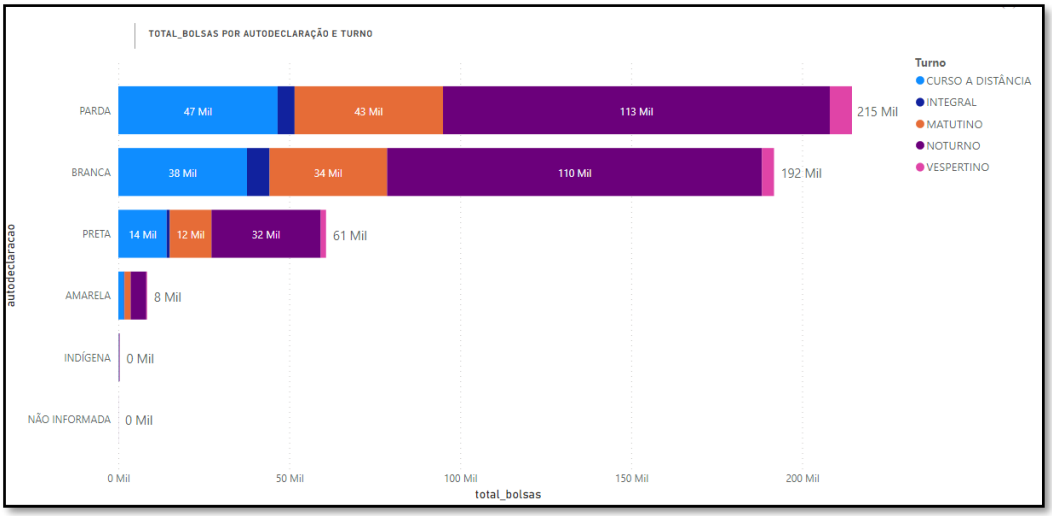


Fonte: Autor.

Sobre o perfil dos candidatos foi analisada a ocupação de bolsas por distinção de autodeclaração de raça/cor da pele, com foco no turno do curso do beneficiário (Figura 34). A partir do gráfico podemos concluir que a autodeclaração parda possui predominância na ocupação de vagas, enquanto a indígena possui menos de 1000 ocupantes nos 2 anos analisados, além disso independente da autodeclaração o turno com mais ocupantes é noturno, seguido do curso à distância. Nesse aspecto de turno predominante o resultado está de acordo com dados

do Censo de Ensino Superior (2018), que concluiu que no ensino presencial a predominância de estudantes é adepta de cursos noturnos.

Figura 34 - Total de Bolsas por Autodeclaração e Turno



Fonte: Autor.

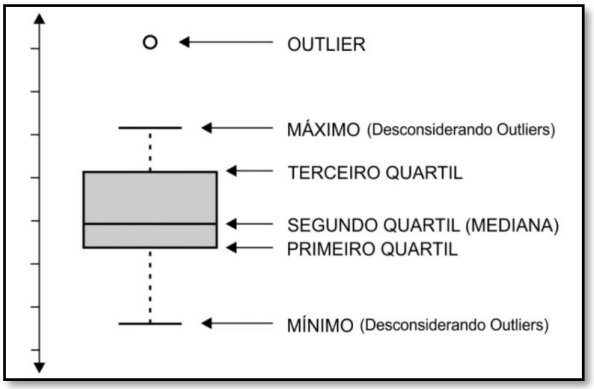
4.6.2 Visualizações com Recursos de Python

Para as visualizações utilizando os recursos da Linguagem Python integrada ao Power BI, a tabela fato utilizada foi a com foco no total de ocupação por idade (idade) que apresenta uma granularidade maior e volume de linhas menor. Para realizar as analises foram utilizadas as bibliotecas de Python, sendo elas *Pandas*[2], *Seaborn*[3] e *Matplotlib*[4], essas bibliotecas possuem funções e gráficos estatísticos mais específicos e complexos que os disponíveis de forma nativa no Power BI.

Um dos gráficos que foi explorado para análise foi o *boxplot* ou diagrama de caixa, ele permite visualizar a distribuição de valores discrepantes dos dados, além de ser uma representação gráfica comparativa. Essa representação se divide entre mínimo, máximo, primeiro quartil, mediana, terceiro quartil e *outlier* (Figura 35). A função que foi utilizada (Figura 36) buscou um comparativo entre a ocupação de vagas presenciais e à distância pela idade dos beneficiários das vagas.

[2] https://pandas.pydata.org/
[3] http://seaborn.pydata.org/index.html
[4] https://matplotlib.org/

Figura 35 - Interpretação de gráfico BOXPLOT



Fonte: Autor.

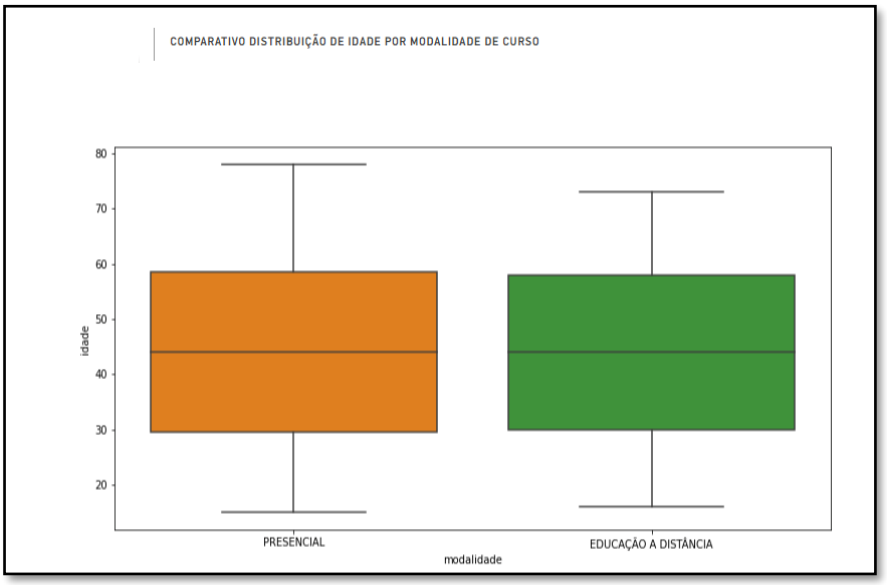
Figura 36 - Script de função Python (BOXPLOT)

```
# dataset = pandas.DataFrame(idade,modalidade)
import matplotlib.pyplot as plt
import seaborn as sn
plot =sn.boxplot(y='idade',x='modalidade',data=dataset, palette="Paired_r")
plt.show()
```

Fonte: Autor.

A partir da representação gerada (Figura 37), podemos observar que em ambas as modalidades há uma distribuição simétrica pela linha correspondente a mediana, ela encontra-se no centro do dos retângulos, indicando simetria. A modalidade presencial apresenta uma dispersão e uma amplitude maior que a modalidade Educação a Distância, a dispersão é definida pela altura do retângulo (primeiro e terceiro quartis), a amplitude é obtida através da diferença entre máximo e mínimo, observando a reta do máximo, identificamos que beneficiários acima de 75 anos optam por cursos presenciais, se focamos nas linhas de mínimo existe a mesma preferência por beneficiários mais jovens.

Figura 37- Distribuição de Idade por Modalidade de Curso



Fonte: Autor

Ainda analisando a variável de idade dos beneficiários, foi utilizando outro recurso da biblioteca *Seaborn* o gráfico de distribuição ou *Displot*, ele é composto por um histograma com a sobreposição de uma linha combinada a ele. O gráfico *displot* representa a variação na distribuição de dados por categorização ou contagem, a linha sob o gráfico é definida pelo argumento KDE (*Kernel Density Estimate*), utilizada para estimar a probabilidade de distribuição de variáveis continuas em valores de dados (SEABORN, 2021). A partir do *script* da figura 38, foi gerado o gráfico com intervalos de idades e contagem do total de beneficiários naquele intervalo (Figura 39).

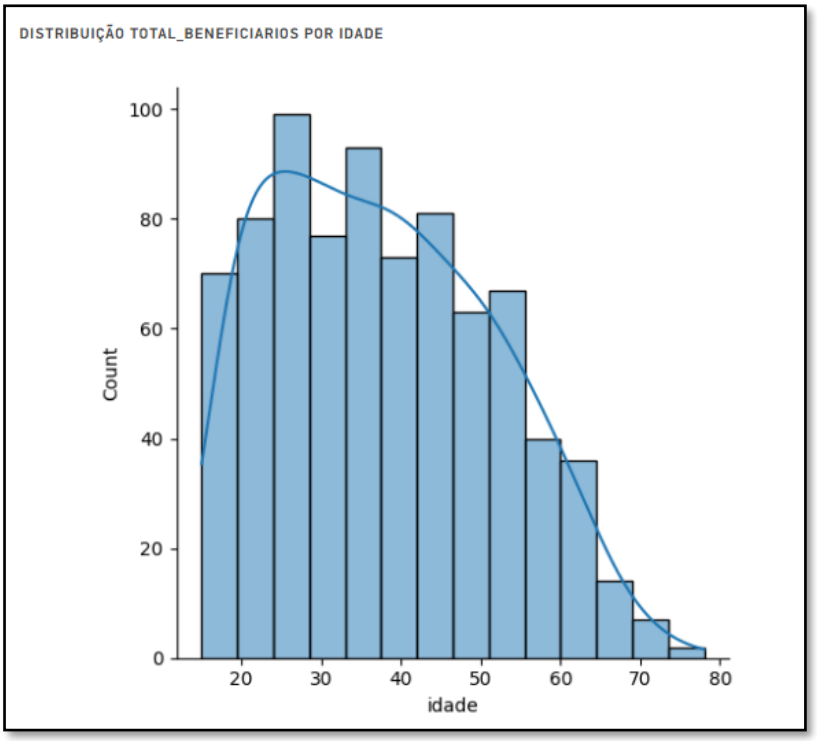
Figura 38 - Script Python (DISPLOT)

```
# dataset = pandas.DataFrame(idade, total_beneficiarios)
import matplotlib.pyplot as plt
import seaborn as sn
plot = sn.displot(dataset['idade'], kde=True)
plt.show()
```

Fonte: Autor.

Pelo caimento e curvatura da linha de densidade percebe-se uma assimetria do gráfico (Figura 39), isso significa que a distribuição de ocupação de bolsas por idade não é uma distribuição normal, de uma distribuição normal espera-se uma variação uniforme. A maior ocupação encontra-se entre 20 e 30 anos de idade, apresentando aumentos e quedas desiguais nos intervalos seguinte, apresentando declínio mais estável a partir dos 50 anos de idade.

Figura 39 - Total de Beneficiários por Idade



Fonte: Autor.

Baseado no gráfico de caixa (*Boxplot*) acrescido do conceito de densidade do KDE (*Kernel Density Estimate*), o gráfico de violino (*violin*) é usado para visualizar a distribuição dos dados e sua densidade de probabilidade (SEABORN, 2021). Com esse recurso buscou-se observar a distribuição total de beneficiários por sexo, essa análise foi verificada anteriormente por alguns gráficos simples de colunas, o intuito é agregar valor na interpretação dos dados com um gráfico com mais recursos de análise como violino.

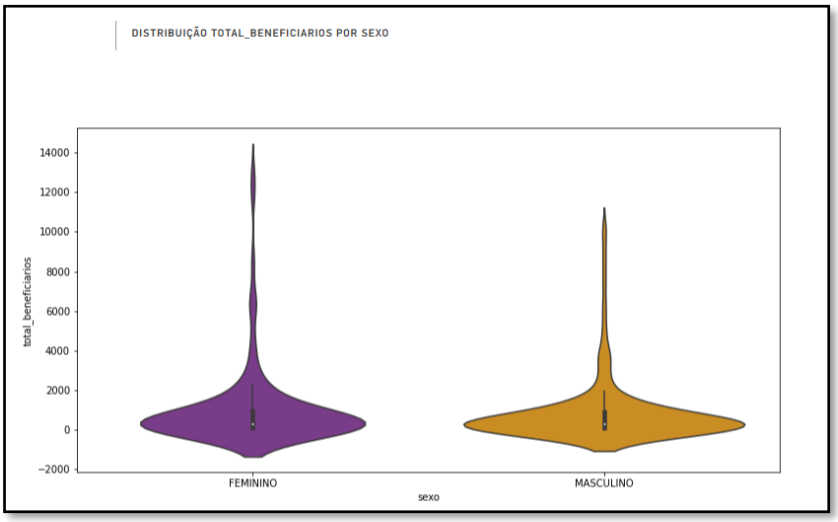
Com *scrypt* (Figura 40) Python foi obtido o gráfico da figura 41. Nesse gráfico podemos observar ponto referente a mediada no retângulo localizado do centro do elemento, o retângulo indica primeiro e segundo quartil assim como o gráfico de caixa, as linhas do eixo central do retângulo, representam o mínimo e máximo. As seções mais largas do gráfico do violino representam uma probabilidade mais alta de observações tomando um determinado valor, as seções mais finas correspondem a uma probabilidade mais baixa.

Figura 40 – *Scrypt* Python (VIOLIN)

```
# dataset = pandas.DataFrame(sexo, total_beneficiarios)
import matplotlib.pyplot as plt
import seaborn as sn
sn.violinplot(x=dataset["sexo"],y=dataset["total_beneficiarios"], data=dataset, palette="CMRmap")
plt.show()
```

Fonte: Autor.

Figura 41 - Total Beneficiários por Sexo

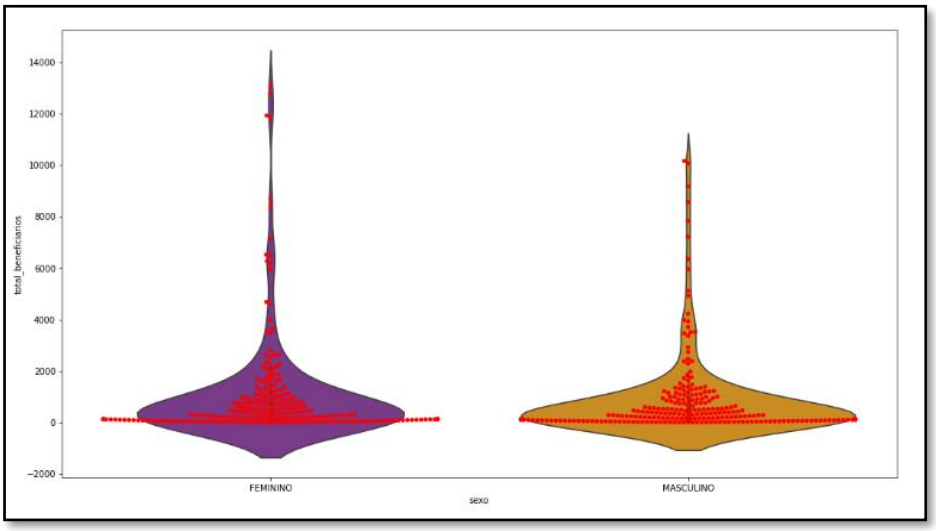


Fonte: Autor.

De forma geral a distribuição das pontas é semelhante para ambos, a mediana está próxima do primeiro quartil, situado antes do ápice da linha de densidade, é visível que o sexo feminino apresenta uma amplitude e frequência maior dada interpretação dos quartis do centro, o sexo masculino com intervalo menor entre os quartis, possui uma curva de densidade uniforme que termina de forma uniforme, enquanto a curva da distribuição feminina apresenta intervalos com protuberâncias na linha de densidade.

Para visualizar de forma mais ilustrada a distribuição pode-se utilizar um gráfico do tipo *swarmplot* também da biblioteca *Seaborn* sobreposto ao gráfico de violino (Figura 42). O gráfico *swarmplot* é um gráfico de dispersão para representar valores categóricos, seu diferencial é não realizar a sobreposição de pontos, ele deixa ainda mais evidente a distribuição de beneficiários por sexo mapeada pela estimativa de densidade de kernel (KDE) (SEABORN, 2021).

Figura 42 - Sobreposição de gráficos



Fonte: Autor

5 CONSIDERAÇÕES FINAIS

O *Business Intelligence* surgiu da necessidade de respostas rápidas a situações que envolvam a tomada de decisão. A modelagem de dados dimensional surgiu neste contexto como uma alterativa de bases de dados voltadas a relatórios, implementadas em bancos relacionais. A abordagem de Kimball (2002) se mostrou uma ferramenta poderosa para construção do *data mart* com os dados utilizados neste trabalho, a base de dados relacional criada primeiramente, foi fundamental para garantir a integridade dos dados na base dimensional criada na sequência, ainda que com o tipo de dados não se tenha uma base relacional exatamente com função de base de dados operacional e uma base dimensional garantindo a integridade de dados históricos, ambas se mostraram complementares no processo de trabalho dos dados.

A utilização de dados públicos é uma ótima alternativa para validar processos que envolvam grande volume de dados. O trabalho aplicado sobre os dados do PROUNI expressa a importância do processo de ETL, os dados passaram por alterações a fim de padroniza-los, transformar de campos para valores que pudessem ser mais interessantes para a análise. O processamento dos arquivos demandou bastante tempo e diversos ajustes no meio do percurso, assim como a inserção na base dimensional a partir da base relacional, isso foi necessário para que o valor das bases de dados mantivesse a integridade e refletisse o valor dos arquivos de origem dos dados.

O Power BI para realização da análise dos dados se mostrou uma ferramenta simples e eficiente, contudo houveram problemas no carregamento de dados, ao carregar a base referente aos 15 anos, o software apresentou-se lento e travou, entendendo-se ser um possível impacto do hardware utilizado para trabalhar, foi optado por se trabalhar com apenas 2 anos de dados. Infelizmente com a versão gratuita do software não é possível trabalhar com recursos em nuvem, sendo assim as configurações de hardware devem ser levadas em consideração assim como o volume de dados.

Os gráficos nativos conseguiram expressar resultados significativos do que foi analisado. A integração com a linguagem Python eleva o nível de recursos do software, ampliando a gama de gráficos estatísticos, suportando não só a linguagem como suas bibliotecas. A limitação de volume de dados suportada pelo Python no software o torna mais eficiente para trabalhar com amostras com menor volume de dados ao invés de *Data Warehouses*.

REFERÊNCIAS

CAVALCANTI, G.G. FELL, A.F.A. DORNELAS, J.S. **Data Warehouse: uma ferramenta de tecnologia de informação para as organizações**. XII SIMPEP Bauru - SP, 2005. Disponível em: <https://simpep.feb.unesp.br/anais/anais_12/copiar.php?arquivo=GOIS_GC_Data%20Warehouse.pdf>. Acesso em: Ago. de 2020.

CENSO ENSINO SUPERIOR, **Portal do Ministério da Educação**. 2018. Disponível em <https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2018/censo_da_educacao_superior_2017-notas_estatisticas2.pdf > Acesso em: Jul. 2021

COSTA, M. V. B. PEDROSA, T. Í. M. PIMENTA, E. A. **Um Estudo dos Dados Governamentais Abertos do Estado de Alagoas**. Encontro Nacional de Computação dos Institutos Federais – ENCOPIF, 2019. Disponível em https://sol.sbc.org.br/index.php/encompif/article/view/7197 > . Acesso em: Nov .2020.

DADOS ABERTOS. **Portal de Dados Abertos do Ministério da Educação**. Disponível em < http://dadosabertos.mec.gov.br/prouni?start=10> Acesso em: 24 Jun. de 2020.

DECRETO 10.160. DECRETO Nº 10.160, DE 9 DE DEZEMBRO DE 2019. **Comitê Interministerial de Governo Aberto.** Disponível em < http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2019/Decreto/D10160.htm#art13 > Acesso em: Nov. de 2020.

DECRETO 9.716. DECRETO Nº 9.716, DE 26 DE FEVEREIRO DE 2019. **Lei de Acesso à Informação.** Disponível em < http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2019/Decreto/D9716.htm#art1> Acesso em: Nov. 2020.

ELMASRI, R. NAVATHE, S. B. **Sistemas de Banco de Dados**. São Paulo : Pearson Education, sob o selo Addison Wesley, 2005.

FILHO, V. B. S; BRANDI, L. S. N. **Um estudo focado ao PROUNI através da análise de dados abertos: período de 2005 até 2016**. Revista: Prisma.com, 2018. Disponível em <https://ojs.letras.up.pt/ojs/index.php/prismacom/article/view/5204/5148> Acesso em: Nov. 2020.

FERREIRA, J. MIRANDA, M. ABELHA, A. MACHADO, J. **O Processo ETL em Sistemas Data Warehouse**. Universidade do Minho, Braga - Portugal, 2010. disponível em: < http://repositorium.sdum.uminho.pt/handle/1822/11435>. Acesso em: Jul. de 2020

GOMES, G. T; FERTIG, M. R. **Construção de uma base de conhecimento de dados governamentais abertos baseada em ontologia utilizando dados conectados**. Trabalho de Conclusão de Curso. Universidade do Sul de Santa Catarina, Florianópolis, 2016. Disponível em <https://riuni.unisul.br/handle/12345/2004>. Acesso em: Set. 2020.

GOMES,L.F.A.M., MORENO JR.,V.A. WOITOWICZ, B.B.C., LUCAS,S.M.F **Uma Abordagem Multicritério para a seleção de ferramentas de Business Intelligence**. Revista Eletrônica de Sistemas de Informação, v. 10, n. 2, artigo 5, 2011. Disponível em: <

http://www.spell.org.br/documentos/ver/5512/uma-abordagem-multicriterio-para-a-selecao-de-f--->. Acesso em : Jul. de 2020.

GOVERNO DIGITAL. Portal Brasileiro de Dados Abertos. Disponível em < https://www.gov.br/governodigital/pt-br/dados-abertos/portal-brasileiro-de-dados-abertos#:~:text=A%20pol%C3%ADtica%20brasileira%20de%20dados,o%20aumento%20da%20integridade%20p%C3%BAblica.> Acesso em: Nov. de 2020.

INDA. Instrução Normativa n°4, 12 de abril de 2012 . Disponível em < https://dados.gov.br/pagina/instrucao-normativa-da-inda> Acesso em: Nov. de 2020.

IMHOFF, C. GALEMMO, N.GEIGER, J.G. **Mastering Data Warehouse Design Relational and Dimensional Techniques**. Indiana: Wiley Publishing, 2003.

INMON, W.H. Bulding. **The Data Warehouse**. Canadá: John Wiley & Sons, Inc. 2002.

JARDIM, E.S. OLIVEIRA, M.V.A. MORAVIA, R.V. **Diferença Entre Banco de Dados Relacionais e Banco de Dados Dimensionais**. Revista Pensar Tecnologia, v.4, n.2, 2015. Disponível em < http://revistapensar.com.br/tecnologia/pasta_upload/artigos/a122.pdf> Acesso em : Ago. 2020.

KIMBALL, R. ROSS, M. **The Data Warehouse Toolkit, Second Edition**. Canadá: John Wiley & Sons, Inc. 2002.

MAGALHÃES, H. F. CARDOSO, L. A. Análise de Dados Abertos sobre o Ensino Superior Brasileiro. Monografia para Conclusão de Curso – Universidade de Brasília, Brasília 2016. Disponível em < https://bdm.unb.br/handle/10483/17719 > . Acesso em 03 fevereiro de 2021.

Society and Development, v. 9, n. 10, e1099108350, 2020. Disponível em <https://rsdjournal.org/index.php/rsd/article/view/8350/7414>. Acesso em: Nov. de 2020.

MICROSOFT. **Documentação de diretrizes do Power BI**. Disponível em : < https://docs.microsoft.com/pt-br/power-bi/guidance/star-schema >. Acesso em: Ago. de 2020.

MATPLOTLIB. Matplotlib: Vistualization with Pyhton. Disponível em :< https://matplotlib.org/ > . Acesso em: Jul. de 2021

NERY, Felipe Rodrigues Machado. **Tecnologia e Projeto de Data Warehouse**. 6ª Edição, São Paulo, SP, 2013.

OLIVEIRA, D. E.; OLIVEIRA, G. L. D. BI como deve ser - O guia definitivo. 2ª. ed.

Salvador: s.n., 2016.

PANDAS. Pandas Documentation . Disponível em < https://pandas.pydata.org/> Acesso em: Jul. 2021.

PDA MEC. Plano de Dados Abertos MEC 2016/2018, versão 2016. Disponível em< http://dadosabertos.mec.gov.br/images/pdf/dicionario-prouni-20161222.pdf > Acesso em: Nov. de 2020.

PEREIRA, A. W. **Banco de Dados**. [Digite o Local da Editora]: Editora Saraiva, 2014. 9788536518961. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788536518961/>. Acesso em: 14 Sep 2020.

PEREIRA P.F.O . CONCEIÇÃO K. R. LEMOS, R.R. FIUZA, P.J. GONÇALVES A.L. **Uma análise temporal de dados abertos do ensino superior utilizando o software de visualização Tabeau**. II Simpósio Ibero-Americano de Tecnologias Educacionais – SITE , 2018 - Universidade Federal de Santa Catarina – UFSC. Disponível em < https://publicacoes.rexlab.ufsc.br/old/index.php/sited/article/view/292 > Acesso em: Nov. 2020.

POSTEGRESQL. PostgreSQL**: The World's Most Advanced Open Source Relational Database**. Disponível em < https://www.postgresql.org/> Acesso em : Ago. 2020.

POWER BI. **Obtenha relatórios ricos e interativos com analise visual na ponta dos dedos**. Disponível em < https://powerbi.microsoft.com/pt-br/desktop/ >. Acesso em: Ago. de 2020.

PROUNI. **Portal do Ministério da Educação**. Disponível em : < http://portal.mec.gov.br/index.php?option=com_content&view=article&id=205&Itemid=298&msg=1&l=aW5kZXgucGhwP29wdGlvbj1jb21fY29udGVudCZ2aWV3PWJ1c2NhZ2VyYWwmSXRlbWlkPTE2NCZwYXJhbXNbc2VhcmNoX3JlbGV2YW5jZV09cHJvdW5pJmQ9cyZwYXJhbXNbZGVdPSZwYXJhbXNbYXRlXT0mcGFyYW1zW2NhdGlkXT0mcGFyYW1zW3NlYXJjaF9tZXRob2RdPWFsbCZwYXJhbXNbb3JkXT1wcg==#:~:text=O%20Programa%20Universidade%20para%20Todos,institui%C3%A7%C3%B5es%20privadas%20de%20educa%C3%A7%C3%A3o%20superior.>. Acesso em: Jun. de 2020.

PYTHON. **Python é uma linguagem de programação que permite trabalhar rapidamente e integrar sistemas de forma mais eficaz.** Disponível em <https://python.org.br/introducao/>. Acesso em: Ago. de 2020.

R. **What is R?**. Disponível em <https://www.r-project.org/about.html> acesso em : Set. de 2020.

RAMESH, S. **Business Intelligence e Análise de Dados para Gestão do Negócio**. [Digite o Local da Editora]: Grupo A, 2019. 9788582605202. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788582605202/. Acesso em: 10 Sep 2020.

SADALAGE, P. J.; FOWLER, M. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Indiana: Addison-Wesley, 2013.
SEABORN. Visualizing distributions of data. Disponível em < https://seaborn.pydata.org/tutorial/distributions.html> Acesso em: Ago. 2021.

SILBERSCHATZ, A. KORTH, H. F. SUDARSHAN, S. **Sistemas de Banco de Dados**. São Paulo: Makron Books, 1999.

SILVA, J. G. S. L; MEIRELLES, F. **O uso de bases de dados públicos por empresas em seus sistemas de Business Intelligence e seus benefícios par ao negócio**. International Conference on Information Resources Management (CONF- IRM), 2017. Disponível em < https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1020&context=confirm2017 >. Acesso em: Set. 2020.

TABLEAU. O que é o Tableau?. Disponível em < https://www.tableau.com/pt-br/why-tableau/what-is-tableau> Acesso em 25 novembro 2020.

TURBAN, E. SHARDA, R. ARONSON, J.E. KING, D. **Business Intelligence: Um Enfoque Gerencial**. Grupo A, 2009. 9788577804252. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788577804252/. Acesso em: 10 Aug 2020.

VICTORINO, M. C; SHIESSL, M; OLIVEIRA, E. C; ISHIKAWA, E; HOLANDA, M. T; HOKAMA, M. L. U**ma Proposta de Ecossistema de Big Data para a análise de dados abertos governamentais conectados**. Inf. & Soc.:Est., João Pessoa, v.27, n.1, p. 213-230, jan./abr. 2017. Disponível em < https://periodicos.ufpb.br/ojs/index.php/ies/article/download/29299/17505/> Acesso em: Set. de 2020.

YESSAD, L. LABIOD, A. **Comparative Study of Data Warehouses Modeling Approaches: Inmon, Kimball and Data Vault**. International Conference on System Reliability and Science, 2016. Disponível em < https://www.researchgate.net/publication/312486486_Comparative_study_of_data_warehouses_modeling_approaches_Inmon_Kimball_and_Data_Vault> Acesso: Jul. de 2020.