

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS**  
**NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**  
**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**Bruno Marcos da Silva Miranda**

**O USO DA CIÊNCIA DE DADOS NA ANÁLISE DO GASTO PÚBLICO FEDERAL:  
UM ESTUDO APLICADO AOS DADOS DO SISTEMA INTEGRADO DE  
PLANEJAMENTO E ORÇAMENTO**

Belo Horizonte  
2020

**Bruno Marcos da Silva Miranda**

**O USO DA CIÊNCIA DE DADOS NA ANÁLISE DO GASTO PÚBLICO FEDERAL:  
UM ESTUDO APLICADO AOS DADOS DO SISTEMA INTEGRADO DE  
PLANEJAMENTO E ORÇAMENTO**

Trabalho de Conclusão de Curso apresentado  
ao Curso de Especialização em Ciência de  
Dados e Big Data como requisito parcial à  
obtenção do título de especialista.

Belo Horizonte

2020

## SUMÁRIO

<b>1. Introdução .....</b>	<b>4</b>
<b>1.1. Contextualização.....</b>	<b>5</b>
<b>1.2. A pergunta de estudo proposta.....</b>	<b>6</b>
<b>2. Coleta dos Dados .....</b>	<b>8</b>
<b>2.1. Estrutura dos dados do SIOP .....</b>	<b>9</b>
<b>2.2. Desenho da solução .....</b>	<b>11</b>
<b>2.3. Seleção de variáveis .....</b>	<b>12</b>
<b>2.4. Processo de obtenção.....</b>	<b>15</b>
<b>2.5. Modelagem do banco de dados.....</b>	<b>18</b>
<b>3. Processamento / Tratamento de Dados .....</b>	<b>20</b>
<b>3.1. Etapa da carga de dados.....</b>	<b>20</b>
<b>3.2. Etapa do tratamento de dados.....</b>	<b>24</b>
<b>3.3. Etapa da seleção do escopo .....</b>	<b>25</b>
<b>4. Análise e Exploração dos Dados .....</b>	<b>26</b>
<b>5. Apresentação e Análise dos Resultados.....</b>	<b>29</b>
<b>5.1. Análise da série histórica dos gastos em saneamento básico.....</b>	<b>31</b>
<b>5.2. Uma análise do Plansab sobre os gastos em saneamento básico .....</b>	<b>31</b>
<b>5.3. Análise dos gastos georreferenciados nas regiões Norte e Nordeste....</b>	<b>32</b>
<b>5.4. Análise dos gastos georreferenciados nas regiões Sul e Sudeste.....</b>	<b>33</b>
<b>5.5. Análise dos gastos georreferenciados na região Centro-Oeste.....</b>	<b>34</b>
<b>6. Conclusão .....</b>	<b>35</b>
<b>Referências Bibliográficas.....</b>	<b>37</b>
<b>APÊNDICE .....</b>	<b>40</b>

## 1. Introdução

Problemas de pesquisa que envolvam análise de efetivação de políticas públicas governamentais são, por sua natureza e complexidade, uma potencial fonte para a aplicação dos conceitos e técnicas da Ciência de Dados. Portanto, esse trabalho de conclusão de curso visa por meio de tais técnicas, demonstrar a sua efetivação aplicada à extração, tratamento, processamento e análise de dados oriundos do Sistema Integrado de Planejamento e Orçamento (SIOP) do Governo Federal.

A saber, segundo a Secretaria de Orçamento Federal (BRASIL, 2018) do Ministério da Economia, o sistema SIOP em sua origem possui as seguintes finalidades: 1) formular o planejamento estratégico nacional; 2) formular planos nacionais, setoriais e regionais de desenvolvimento econômico e social; 3) formular o plano plurianual, as diretrizes orçamentárias e os orçamentos anuais; 4) gerenciar o processo de planejamento e orçamento federal; 5) promover a articulação com os Estados, o Distrito Federal e os Municípios, visando a compatibilização de normas e tarefas afins aos diversos Sistemas, nos planos federal, estadual, distrital e municipal.

É, portanto, na terceira finalidade que se originam os dados interessantes ao estudo proposto, uma vez que se tratam das variáveis que compõem o orçamento federal brasileiro para a grande parte dos gastos públicos federais.

Por óbvio, que todo gasto federal deverá estar previsto na Lei Orçamentária Anual (LOA), fazendo, portanto, a sua contrapartida com os Orçamentos da União, o qual goza de transparência e possui acesso aberto a todo cidadão brasileiro (CÂMARA DOS DEPUTADOS, 2020).

A Lei Orçamentária Anual (LOA) estabelece os Orçamentos da União, por intermédio dos quais são estimadas as receitas e fixadas as despesas do governo federal. Na sua elaboração, cabe ao Congresso Nacional avaliar e ajustar a proposta do Poder Executivo, assim como faz com a Lei de Diretrizes Orçamentárias (LDO) e o Plano Plurianual (PPA). (CÂMARA DOS DEPUTADOS, 2020)

O referido trabalho contempla os seguintes objetivos gerais:

- I. Conhecer a estrutura de dados fornecida pelo SIOP;
- II. Analisar o potencial de uso dos dados em vista das perguntas de pesquisa;

- III. Implementar solução de modelagem e de extração dos dados;
- IV. Analisar os dados processados para a efetivação de um estudo de caso.

### 1.1. Contextualização

Diante aos objetivos colocados na seção introdutória, deseja-se observar que para fins de elaboração de um estudo de caso, ou seja, trabalhar com a aplicabilidade da arquitetura e técnicas de Ciência de Dados referidas, suportadas na sua parte de análise dos resultados pela Economia Ambiental, oportunizou-se para o estudo o tema: Políticas Públicas para o Desenvolvimento do Saneamento Básico no Brasil.

Fica de antemão esclarecido, que tais informações que norteiam e ou respondam às perguntas de pesquisa acima colocadas, não se encontram respondidas e ou processadas no sistema SIOP, a este sistema cabendo o justo (e não apenas) *input* e gestão dos dados que foram obtidos e processados para fins negociais neste trabalho.

Não obstante, evidencia-se também que todo um planejamento, desenho de solução, implementação e testes se farão necessários até que os dados estejam *off-line*, ou seja, em base de dados local extraídos a partir do SIOP para fins de processamento, onde, tais definições técnicas serão definidas nas seções posteriores do trabalho.

Por que o SIOP foi a principal base de dados escolhida para determinado problema de pesquisa? Com um total aproximado de 5.7 milhões (jan/2000 a dez/2019) de registros disponíveis, a definição do SIOP como base de dados proporciona o acompanhamento da evolução do histórico dos planos setoriais, programas e gastos do orçamento do Governo Federal em ações orçamentárias que tangem diversos setores de investimento do governo brasileiro, e dentre eles o Setor de Saneamento Básico, foco da análise do estudo de caso proposto nesse trabalho.

Para além deste critério negocial, tecnicamente os dados apresentam boa estrutura com relação aos formatos empregados e completude dos registros, uma base histórica dos últimos vinte anos dos gastos (planejamento e execução) públicos federais, ampla documentação técnica, API (*Application Programming Interface*) de

acesso, possibilidade de obtenção de dados abertos (ou para usuários registrados), bem como manual de uso e dicionário de dados atualizado.

## 1.2. A pergunta de estudo proposta

A pergunta proposta para o trabalho foi colocada dada a carência de estudos empíricos na área ambiental, sobretudo na temática do saneamento básico no Brasil, o qual hoje tanto se faz necessário para uma discussão precisa e clara sobre, por exemplo, o Novo Marco Legal do Saneamento Básico, o projeto de lei PL 4.162/2019 (SENADO FEDERAL, 2020a) e que está sendo discutido pela Câmara dos Deputados e pelo Senado Federal no presente ano de 2020 (SENADO FEDERAL, 2020b).

A saber, um último esforço nacional empregado na agenda para fins de traçar um panorama situacional, bem como prover o planejamento de políticas públicas de longo prazo, foi o Plano Nacional de Saneamento Básico (PLANSAB), o qual teve a sua elaboração prevista na **Lei nº 11.445/2007** (BRASIL, 2007), sendo elaborado pelo Ministério das Cidades durante a gestão de Dilma Rousseff (BRASIL, 2013).

Almejando ser o principal instrumento da política pública nacional de saneamento básico no Brasil nos próximos 20 anos, o documento norteador é coerente nesta direção e planejamento.

Destaca-se que a lógica adotada para a elaboração do Plansab é a de um planejamento que dá ênfase a uma visão estratégica de futuro. Nesse modelo, o futuro não é simplesmente uma realidade desenhada pela equipe de planejamento, abordagem esta usual no planejamento tradicional, que a adota a despeito de se saber que o planejador não dispõe da capacidade de influenciar todos os fatores determinantes desse futuro. O enfoque adotado, ao contrário, é o de procurar visualizar possíveis futuros, denominados de cenários, a partir das incertezas incidentes, com base em sólida análise da situação atual e pregressa. (BRASIL, 2013, p. 13)

Com um caráter efetivo de ações, o plano define ainda as medidas necessárias para que tais metas sejam alcançadas, dividindo tais medidas em estruturais e estruturantes. Para uma definição de medidas estruturais:

[...] correspondem aos tradicionais investimentos em obras, com intervenções físicas relevantes nos territórios, para a conformação das infraestruturas físicas de abastecimento de água potável, esgotamento sanitário, limpeza urbana e manejo de resíduos sólidos e drenagem e manejo das águas pluviais urbanas. São evidentemente necessárias para suprir o *déficit* de cobertura pelos serviços e a proteção da população quanto aos riscos epidemiológicos, sanitários e patrimoniais. (BRASIL, 2013, p. 15)

Em complemento às medidas estruturais citadas, as medidas estruturantes prestam, “[...] suporte político e gerencial para a sustentabilidade da prestação dos serviços. Encontram-se tanto na esfera do aperfeiçoamento da gestão, [...] quanto na da melhoria cotidiana e rotineira da infraestrutura física” (BRASIL, 2013, p. 15).

Cabe ressaltar que a temática do saneamento básico é transversal, podendo esta ser abordada no âmbito das esferas econômica, saúde pública, ambiental, social, dentre outras. Desse modo, se faz presente o desafio – para um estudo empírico, de se obter uma origem de dados sólidos, confiáveis e coerentes com os resultados que se almejam alcançar com a pesquisa, buscando sempre o seu suporte em evidências.

Para este estudo, os dados analisados possuem duas fontes principais, ambas do Governo Federal e que possuem mecanismos de coleta aberta e/ou registrada de acessos. A seguir, apresentam-se a relação de seus mantenedores / bases de dados que foram utilizadas no referido trabalho:

- i. Secretaria de Orçamento Federal, do Ministério da Economia, para os dados do orçamento federal providos por meio do SIOP (BRASIL, 2020);
- ii. Instituto Brasileiro de Geografia e Estatística, para os dados referentes à Divisão Territorial Brasileira – DTB, com arquivos de dados que proveem dados das municipalidades brasileira (IBGE, 2019).

Para fins de delimitação de escopo do estudo de caso que fundamenta o trabalho proposto, tem-se por objetivos específicos, em desdobramento dos objetivos gerais delimitados na introdução -, responder as seguintes perguntas de pesquisa:

- i. Qual foi a evolução dos gastos públicos federais na série histórica observada (2000-2019) para a agenda do saneamento básico no Brasil?
- ii. Qual foi o comportamento dos gastos públicos federais, aplicados para o investimento em saneamento básico no Brasil a partir de 2013, ano de lançamento do Plansab (BRASIL, 2013)? Houveram acréscimos significativos após a implementação do plano?

Devido ao fato do Plansab ter sido um plano de abrangência nacional, fica óbvio que sua cobertura deva contemplar todas as unidades federativas do Brasil, porém, para fins de análise deste trabalho, foi necessário visualizar de forma mais detida o

nível de completude dos registros do SIOP no que tangem as variáveis geoespacializadas, referentes aos gastos orçamentários do governo nos períodos selecionados. Fato que será explorado e evidenciado nas seções posteriores.

Desse modo, o objetivo é tentar captar o movimento (se houve) do fluxo de investimento em saneamento básico no Brasil depois da implementação do Plansab, ou seja, um cenário *ex ante* contra a visão *ex post* da política, uma típica visão de análise de custo *versus* benefício. Sendo assim, a pesquisa contemplou dois períodos de análise para fins comparativos, são estes:

- i. Início da série histórica dos gastos públicos federais providos pelo SIOP, ou seja, o ano de 2000, até o período anterior à implementação do Plansab, - a saber, o ano de 2013 -, formam o primeiro período;
- ii. Período a partir de 2013, ano de implementação do referido plano como política pública para o saneamento básico no Brasil (BRASIL, 2013). Cabendo a ressalva que, ter dividido a análise em dois períodos, possibilitou ao estudo aferir uma medida de comparação da eficácia da política pública a que se deseja avaliar.

## **2. Coleta dos Dados**

O processo de escolha e planejamento da coleta dos dados é sem sombra de dúvidas uma das partes mais delicadas em um projeto de Ciência de Dados. Seja devido à diversidade / dimensão de bases de dados hoje disponíveis, seja devido à complexidade em acessá-las e organizá-las numa estrutura eficaz, bem como torná-las úteis e orientadas às perguntas de pesquisa que se desejam responder.

Desse modo, optou-se por subdividir essa seção do trabalho em cinco subseções, para fins de esclarecimentos e detalhamentos mais precisos a cerca de cada etapa, são estas: 2.1) Estrutura dos dados do SIOP; 2.2) Desenho de solução; 2.3) Seleção de variáveis; 2.4) Processo de obtenção; e, 2.5) Modelagem do banco de dados.

Por óbvio, o SIOP não disponibiliza acesso direto ao seu banco de dados, ou mesmo em documentação o seu modelo de dados. Desse modo, conforme a maioria



dos portais de dados abertos, os usuários consumidores desses dados possuem a opção de conexão / extração via API de serviços remotos, ou ainda, via *download* de arquivos de dados. O SIOP disponibiliza ambas versões (SIOPDOC, 2020), onde para o projeto de arquitetura em questão, optou-se pela segunda forma de acesso.

## 2.1. Estrutura dos dados do SIOP

Basicamente, os dados exportados via planilha de dados a partir do SIOP foram carregados em uma entidade (tabela) do banco de dados criado para o projeto, o qual é o responsável por armazenar o espelhamento dos dados obtidos, suportar as etapas de processamento (limpeza, enriquecimento, etc.), bem como ser fonte de dados para as análises que foram realizadas na última fase do trabalho.

Para as consultas realizadas, um total de 15 variáveis foram selecionadas para o trabalho, as quais serão detalhadas em dicionário de dados nas seções que seguem o texto. Dentre as variáveis selecionadas, 3 destas merecem especial atenção, são estas: a) Ação orçamentária; b) Ano exercício; e, c) Pago + RAP Pago.

Uma Ação Orçamentária (ou um conjunto) basicamente são lançamentos de dispêndios financeiros governamentais, que por sua vez possuem o objetivo de financiar bens ou serviços que contribuam para atender um ou mais objetivos de um determinado programa no âmbito de alguma política pública implementada pelo governo brasileiro, segundo o Manual Técnico do Orçamento 2018 (MTO 2018) da Secretaria de Orçamento Federal do Ministério da Fazenda (BRASIL, 2018).

Ainda segundo o MTO 2018 (BRASIL, 2018) a variável Ano Exercício, representa o período ao qual se referem a previsão das receitas / despesas registradas na LOA. E por último, a variável Pago + RAP Pago é o montante total gasto em determinada ação orçamentária que vise atender à uma política pública.

Os registros de ações orçamentárias podem se repetir (e se repetem) diversas vezes no mesmo ano de exercício, ou seja, com diversos lançamentos para atender à diversas iniciativas / projetos ligados a uma ação orçamentária. Logo, para se obter o saldo de uma ação orçamentária é preciso somar para determinado ano de exercício, os valores monetários para cada ação correspondente que se deseja analisar.

O sistema SIOP não exporta a variável que representa um identificador único dos registros, dado a isso, não é possível garantir a unicidade destes. É importante salientar que por padrão os dados são exportados por meio de planilhas de dados em arquivos do tipo csv (*comma separated values*). Portanto, o que se recebe é uma matriz de dados não ordenada dos registros transferidos, como demonstra um exemplo da sua estrutura na figura 1.

21 colunas (variáveis)

5.683.085 milhões de linhas (observações)

cl1	cl2	cl3	...	...	...	...	...	...	...
l1									
l2									
l3									
...									
...									
...									
...									
...									
...									

Figura 1. Estrutura de dados matricial. Fonte: Elaborado pelo autor.

Isto posto, para cada ano selecionado de registros computados, o sistema permite exportar uma planilha de dados contendo aproximadamente 100.000 linhas. E caso o quantitativo ultrapasse esse valor, o sistema gera arquivos adicionais com o indicativo de qual parte do todo se está exportando no momento. Tomando por exemplo o ano de 2010, caso a consulta retorne 250.000 registros, a exportação seguirá a ordem: a) nome-arquivo.f1.csv, contendo 100.000 linhas; b) nome-arquivo.f2.csv, contendo 100.000 linhas; c) nome-arquivo.f3.csv, contendo 50.000 linhas, respectivamente.

De fato, caso se opte por enriquecer a base de dados em questão, medida conveniente é estender a normalização dessa base de acordo com o paradigma modelo relacional (CHEN, 2002), de modo a acrescentar entidades auxiliares, bem como os relacionamentos a fim de contemplar os novos conjuntos de dados desejados das demais fontes primárias necessárias a pesquisa.

Portanto, para o referido trabalho, fez-se necessário a contemplação da criação de um identificador único na entidade lógica da base de dados que foi desenvolvida

para o projeto, onde, a variável de identificador é um valor numérico inteiro sequencial gerado pelo próprio SGBD (sistema de gerenciamento de banco de dados), e de toda forma, também foram criados campos auxiliares na entidade para fins de registros do nome do arquivo de dados de onde o referido registro processado foi oriundo, bem como o Id da linha na planilha de dados processada na carga. Assim, se é possível garantir a unicidade dos registros, não incorrendo em erros possíveis de dupla contagem, por exemplo. Outro ganho substancial com essa medida, foi a possibilidade de rastrear o andamento da carga de dados que foi realizada e de também poder monitorar possíveis erros durante a execução do procedimento.

## 2.2. Desenho da solução

O desenho de solução proposto se baseou em gerar um banco de dados a partir dos registros obtidos do sistema SIOP. Desse modo, foi então desenvolvido o banco de dados DB\_GASTOS\_PUBLICOS (a ser detalhado posteriormente) para permitir a melhor exploração na fase de análise e de forma a possibilitar:

- i. Eliminar as possíveis redundâncias de registros;
- ii. Criar visões detalhadas de análise;
- iii. Validar e replicar os dados;
- iv. Realizar a junção de dados com outras bases auxiliares para fins de enriquecimento – por exemplo, DTB do IBGE;
- v. Exportar os dados trabalhados de maneira agregada.

Ademais, a solução adotada, focada na implementação do referido banco de dados, visa facilitar / automatizar tarefas de:

- i. Pré-processamento (exemplo: correção de formatação, alteração de tipos de variáveis, identificação de expurgos, etc.);
- ii. Pós-processamento (exemplo: normalização de registros, identificação de padrões, análise do nível de completude dos registros, etc.);
- iii. Enriquecimento das bases de dados com novas entidades e/ou variáveis;

- iv. Desenho de *datasets* estruturados e resumidos, focados à determinadas perguntas negociais, ou seja, na elaboração dos relatórios finalísticos do projeto do estudo de caso proposto.

É importante salientar que, para além das variáveis monetárias que contemplam o rastreio do orçamento público federal com as políticas nacionais da agenda do saneamento básico (dentre outras), o sistema SIOP ainda provê diversas variáveis de análise que permitem a compatibilização (junção) com dados de outras fontes governamentais, tais como: dados municipais e estaduais das bases de dados do IBGE (por exemplo, informações georreferenciadas); e/ou indicadores sociais como PIB e renda per capita (IBGE, 2020), IDH (PNUD BRASIL, 2020) e IDHM (PNUD BRASIL, 2016), indicadores de educação (INEPDATA, 2020), dentre outros.

Neste quesito, do enriquecimento de informações à base de dados original, é preciso salientar que o referido trabalho contemplou apenas a criação de variáveis (binárias ou categóricas) para fins de classificação dos registros com o objetivo de facilitar consultas na fase de análise, bem como monitorar o processo de carga dos dados para o banco de dados local. Apesar de não ter sido contemplado neste trabalho (devido a clara delimitação de escopo e propósito) o enriquecimento com outras bases de dados governamentais disponíveis, sua prática é altamente factível e recomendável como possíveis pontos de extensão ao referido estudo.

### 2.3. Seleção de variáveis

As variáveis oriundas da exportação de dados do sistema SIOP, que foram incorporadas ao banco de dados do projeto foram organizadas no quadro 1.

Quadro 1. Dicionário de dados das variáveis obtidas do SIOP.  
Fonte: Elaborado pelo autor com base no MTO 2018 (BRASIL, 2018).

Nome da variável	Descrição	Tipo
Ano Exercício	Ano da previsão das receitas e a fixação das despesas registradas na LOA.	String de caracteres
Ano Referência	Ano em que o recurso, inscrito e executado em restos a pagar, foi previsto no orçamento.	String de caracteres
Órgão	Representa o nível superior da classificação institucional.	String de caracteres

Unidade Orçamentária	Instituição que irá executar o orçamento, ou seja, que serão responsáveis pela realização das ações.	String de caracteres
Poder	Domínio: Executivo, Legislativo ou Judiciário.	String de caracteres
Função	Indica em qual área de despesa a ação governamental será executada.	String de caracteres
Programa	Instrumento de organização da ação governamental visando à concretização dos objetivos pretendidos.	String de caracteres
Ação orçamentária	Uma Ação Orçamentária (ou um conjunto destas) basicamente são operações das quais resultam em dispêndios financeiros governamentais, que por sua vez financiam produtos (bem ou serviços) que contribuem para atender ao(s) objetivo(s) de um programa no âmbito de alguma política pública implementada pelo governo brasileiro.	String de caracteres
Id Ação	Representa o identificador da ação orçamentária.	String de caracteres
Tipo de Ação	Domínio: Projeto, Atividade, Operações Especiais, Não Orçamentárias e Reserva de Contingência.	String de caracteres
Município	Código do IBGE do Município.	String de caracteres
UF	Identifica o Estado da Federação ao qual o orçamento está vinculado.	String de caracteres
Natureza de Despesa	Compreende o tipo de classificação da natureza do gasto.	String de caracteres
LOA	Compreende o valor da Lei Orçamentária do Ano Exercício analisado.	Monetário
Pago + RAP Pago	Representa a soma dos valores pagos no ano de exercício para determinada ação orçamentária.	Monetário

As variáveis que foram acrescentadas afim de enriquecer o conjunto oriundo da exportação de dados do sistema SIOP, foram organizadas no quadro 2.

Quadro 2. Dicionário de dados das variáveis enriquecidas.  
Fonte: Elaborado pelo autor.

Nome da variável	Descrição	Tipo
Id	Atributo <u>criado</u> para representar a chave primária artificial.	Inteiro
DataCarga	Se refere ao <i>datetime</i> da operação da carga de dados.	Data e hora

Escopo	Utilizado para marcar as ações alvo do estudo.	Binário
Expurgo	Utilizado para marcar as ações de registro inconsistente.	Binário
ArquivolImportacao	Registra o caminho/nome do arquivo de dados de origem.	String de caracteres
IndiceLinha	Registra o número de linha do arquivo de dados de origem.	Inteiro

As variáveis oriundas da importação de dados do IBGE, que foram incorporadas ao banco de dados do projeto foram organizadas no quadro 3.

Quadro 3. Dicionário de dados das variáveis obtidas do IBGE.

Fonte: Elaborado pelo autor com base no DTB 2019 (IBGE, 2019).

Nome da variável	Descrição	Tipo
Id	Atributo <u>criado</u> para representar a chave primária artificial.	Inteiro
capital	Atributo <u>criado</u> para efetuar a diferenciação dos 27 municípios brasileiros que são capitais de seus respectivos estados.	Binário
cod_uf	Representa o código da unidade federativa na base do IBGE.	Float
uf	Nome da unidade federativa no Brasil.	String de caracteres
sgl_uf	Sigla da unidade federativa no Brasil.	String de caracteres
cod_ibge	Representa o código do município na base do IBGE.	Float
nm_municipio	Representa o nome do município na base do IBGE.	String de caracteres
nm_municipio_upper	Nome do município em caixa alta.	String de caracteres
nm_municipio_lower	Nome do município em caixa baixa.	String de caracteres
cod_latitude	Representa a latitude do município na base do IBGE.	Float
cod_longitude	Representa a longitude do município na base do IBGE.	Float
sgl_regiao	Nome da região do estado no Brasil.	String de caracteres
regiao	Sigla do nome da região do estado no Brasil.	String de caracteres

## 2.4. Processo de obtenção

Uma vez de posse da credencial de acesso em mãos<sup>1</sup>, o acesso foi realizado na plataforma disponibilizada pela Secretaria de Orçamento Federal do Ministério da Economia (BRASIL, 2020). O processo pode ser verificado na figura 2.

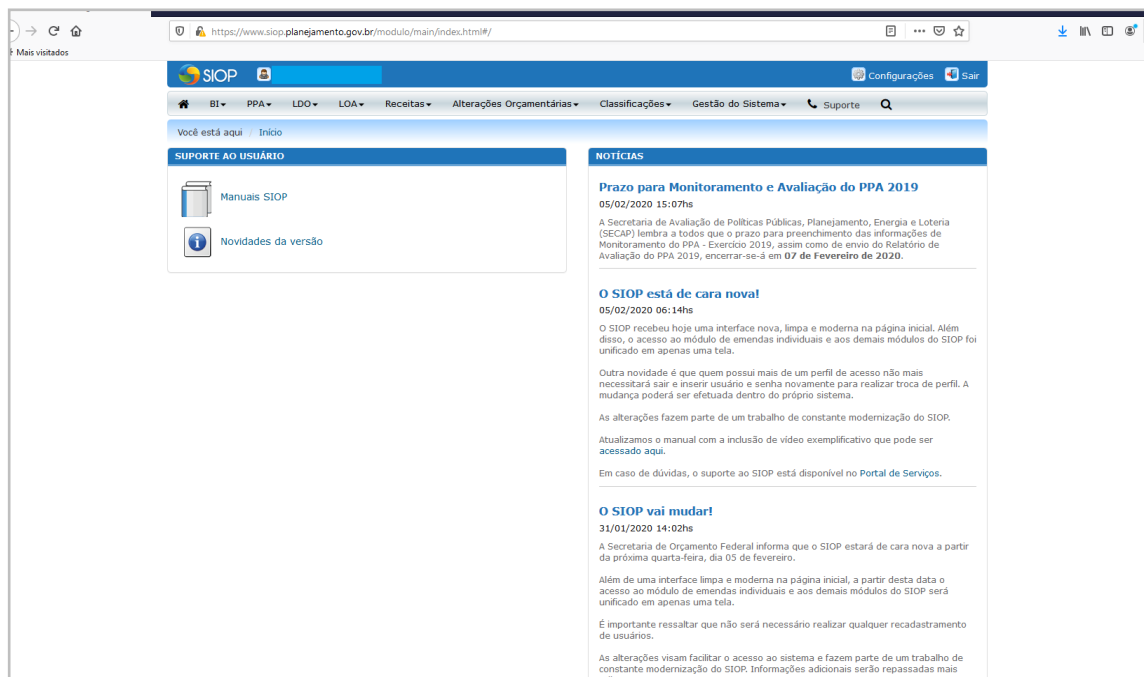


Figura 2. Tela inicial de *login* do sistema SIOP. Fonte: Adaptado de BRASIL (2020).

O sistema SIOP disponibiliza uma funcionalidade de busca e exportação dos dados, onde se é possível salvar sua estrutura no sistema como consultas nomeadas, o que facilita sua replicação num segundo momento pelo usuário (BRASIL, 2020).

Por meio da plataforma, então foram selecionadas a relação de variáveis a serem exploradas por meio do estudo de caso proposto, onde se aplicou o filtro para a variável Ano Exercício com os valores correspondentes ao período 2000 até o ano de 2019, como processo também pode ser verificado na figura 3.

<sup>1</sup> O processo de obtenção da credencial de acesso ao Sistema SIOP não será coberto pelo referido trabalho. Caberá a cada pesquisador explorar no site da plataforma e se informar a cerca dos seus pré-requisitos, exigências e afins para obtê-lo.

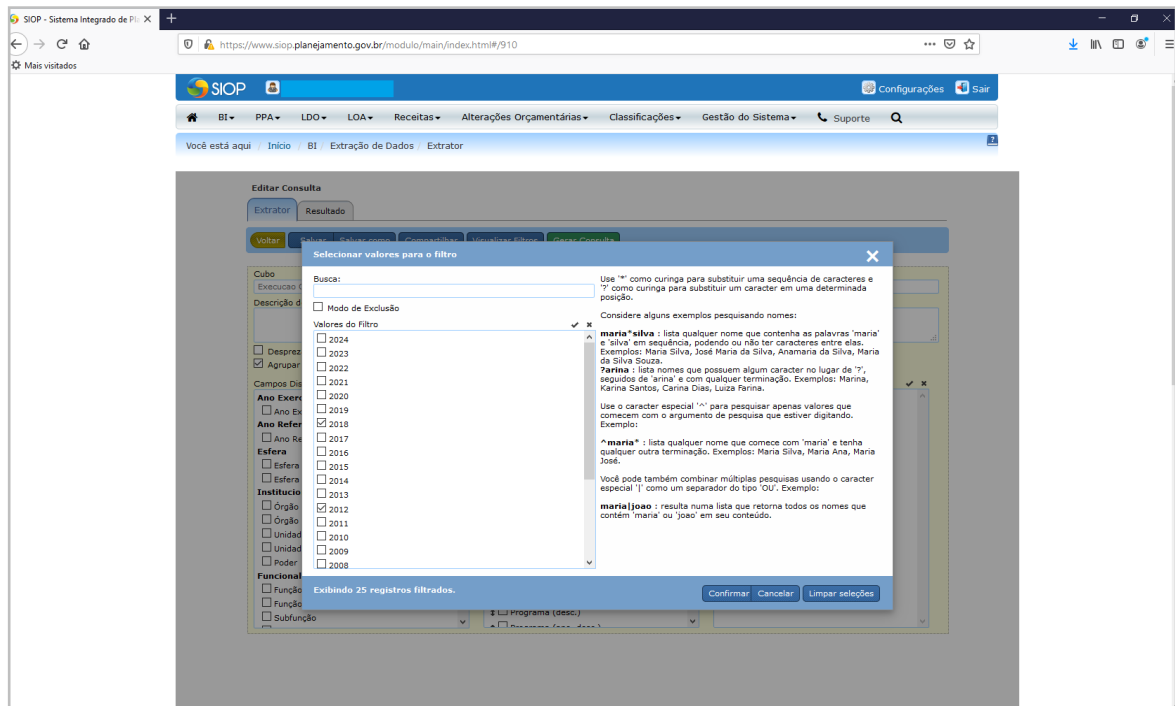


Figura 3. Seleção das variáveis de interesse e anos de exercício. Fonte: Adaptado de BRASIL (2020).

Como já destacado, o SIOPI particiona os arquivos de dados em lotes de até 100.000 registros, os quais em momento preliminar ao processo de busca e *download*, foram organizados em sistema de arquivos do sistema operacional do ambiente computacional, onde foi desenvolvido o projeto em questão - no caso aplicado foi o *Microsoft Windows*<sup>2</sup> -, por ano, conforme observado na figura 4.

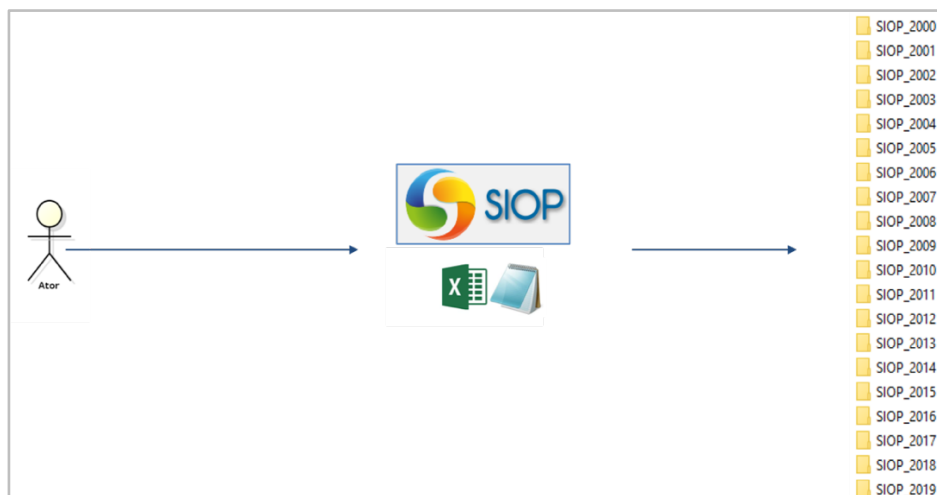


Figura 4. Processo de obtenção dos dados brutos. Fonte: Elaborado pelo autor.

<sup>2</sup> O referido sistema operacional utilizado poderia ser substituído por qualquer outro a escolha do pesquisador. No momento da pesquisa, o *Microsoft Windows* foi a opção disponível e factível para o trabalho.



Em seguida, um repositório de dados hierarquizados foi construído em um sistema de arquivos compartilhado em rede, o qual foi processado em estrutura de um banco de dados relacional (COUGO, 1997, p.33) em momento do processo de carga de dados.

De fato, parte considerável do tempo previsto em cronograma para o estudo de caso foi gasto no planejamento / desenho e implementação do processo de carga de dados, onde a decisão foi tomada por esse desenho de solução, ao se perceber a dificuldade (e inviabilidade, de certo modo) em se trabalhar com tantos arquivos de dados, agregando-os em um pacote estatístico posteriormente. Sobretudo, diante da necessidade de limpeza, formatação e deleção de registros, bem como acréscimo de variáveis, típicos da fase de modelagem da estrutura de dados almejada.

No quadro 4 é possível se ter uma dimensão do número de registros obtidos para o referido estudo de caso proposto por este trabalho.

Quadro 4. Quantitativo de registros obtidos para o estudo.

Fonte: Elaborado pelo autor a partir de dados obtidos do SIOP (2019).

<b>Tipo</b>	<b>Quantitativo</b>	<b>Descrição</b>
Anos / diretórios obtidos	19	Corresponde aos anos disponibilizados pelo SIOP para extração dos dados até o momento do delineamento de escopo e execução desse estudo.
Variáveis selecionadas	15	O SIOP disponibiliza mais de 60 variáveis na composição das consultas, donde 15 foram selecionadas para o presente estudo.
Arquivos de dados por ano	Entre 3 e 4	Cada arquivo plano de dados, contém até o limite máximo de 100.000 registros. Para os anos com maiores lançamentos, até 400.000 registros foram encontrados.
Total de arquivos de dados obtidos	63	Representa o quantitativo total de arquivos planos de dados que foram processados durante a fase de carga.
Total em <i>Gigabytes</i> de dados obtidos	1.7GB	Representa o total em <i>Gigabytes</i> em arquivos planos de dados que foram processados durante a fase de carga.
Total de registros na base de dados obtida	5.683.085	Corresponde aos lançamentos de ações orçamentárias disponibilizadas pelo SIOP para extração até o momento do delineamento de escopo e execução desse estudo.

Total de Ações Orçamentárias agrupadas obtidas	13.128	Corresponde ao número de ações orçamentárias agrupadas pelo Id, uma vez que cada ação pode ter diversos lançamentos no mesmo ano.
--	--------	---

## 2.5. Modelagem do banco de dados

Para a etapa de modelagem dos dados, o banco de dados DB\_GASTOS\_PUBLICOS foi criado com o intuito de armazenar as entidades de suporte negocial do estudo. É importante salientar, que o esforço de normalização do modelo segundo os preceitos do modelo entidade relacional (CHEN, 2002), não se aplicam diretamente, uma vez que se deseja o espelhamento, tratamento e replicação dos registros e não a construção de uma base de dados para fins de desenvolvimento de um sistema de informação visando uma aplicação corporativa típica (FOWLER, 2006, p.24).

Basicamente, o banco de dados desenhado para a solução (figura 5) conta com 5 tabelas, as quais se encontram descritas no quadro 5:

Quadro 5. Descrição das tabelas do banco de dados.

Fonte: Elaborado pelo autor.

Tabela	Função	Justificativa
AcaoSIOP	Espelhamento com os dados oriundos do SIOP, sem qualquer tratamento.	Dada a necessidade de se manter os dados brutos numa tabela a parte, para que o esforço de processamento da carga de dados não seja feito necessário novamente, caso algum erro ocorra durante o processo de limpeza, tratamento e enriquecimento dos dados.
AcaoTratada	Armazenamento dos dados após passar por fase de tratamento.	Se configura a necessidade de não alterar registros brutos oriundos da carga durante o processo de limpeza, expurgo e possíveis alterações. E caso hajam erros nesse processo, apenas uma deleção na entidade AcaoTratada e posterior cópia dos registros da entidade AcaoSIOP será feito.
AcaoAnalizada	Armazenamento das ações alvo dos gastos em saneamento básico.	É a estrutura de dados que irá armazenar as ações que sofrerão tratamento, bem como foram selecionadas para o escopo da análise, ou seja, será o repositório natural de onde os <i>datasets</i> serão carregados.
SelecaoEscopo	Registro e rastreamento da seleção de escopo da análise.	Por meio desta tabela é que será possível verificar quais ações de um total de 5.7 milhões de registros, serão de fato

		analisadas, por terem em sua característica principal, o gasto em saneamento básico no Brasil.
IBGEMunicipios	Agregar atributos georreferenciados ao banco de dados criado.	Por meio desta tabela é que será possível verificar a estrutura (nome, região, geolocalização) dos 5.570 municípios do Brasil.

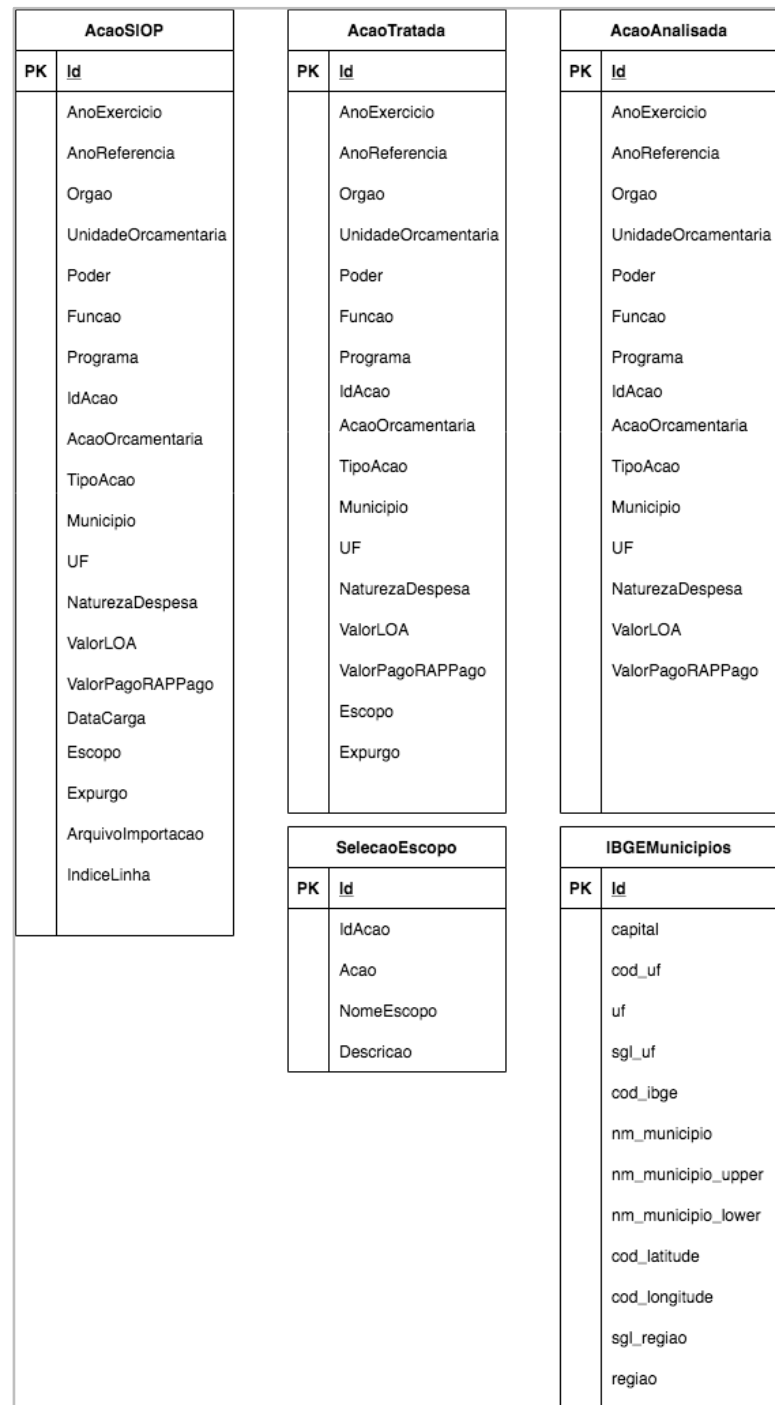


Figura 5. Modelo entidade relacional. Fonte: Elaborado pelo autor.

### 3. Processamento / Tratamento de Dados

A partir da organização dos dados brutos em uma estrutura de repositório hierarquizado, foi construído um aplicativo utilizando a linguagem de programação *Python 3.7* (PYTHON, 2020) para efetuar a carga completa dos arquivos de dados, processamento e tratamento de inconsistências.

Portanto, vencidas ambas etapas, foi realizada uma busca na base de dados construída afim de se recuperar as ações orçamentárias do contexto do saneamento básico, as quais foram definidas como foco do escopo no estudo de caso deste trabalho. Todo o processo será explorado a seguir nas subseções posteriores.

#### 3.1. Etapa da carga de dados

O processo de carga (figura 6) foi executado e disponibilizado o seu produto do processamento em infraestrutura de banco de dados local, fato que já permitiu uma exploração inicial dos dados. A aplicação codificada permitiu ainda, descobrir e mitigar as possíveis inconsistências oriundas do processo de exportação dos dados, dentre elas, a não unicidade dos registros no que diz respeito às ações orçamentárias, matéria-prima fundamental no entendimento e análise dos gastos públicos federais.

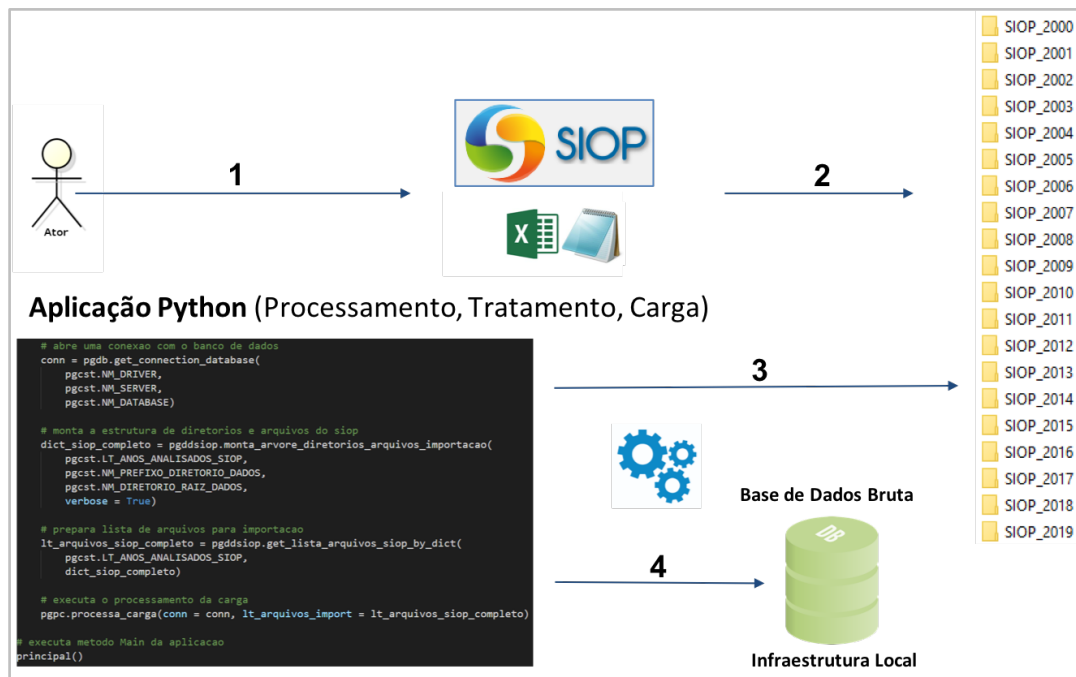
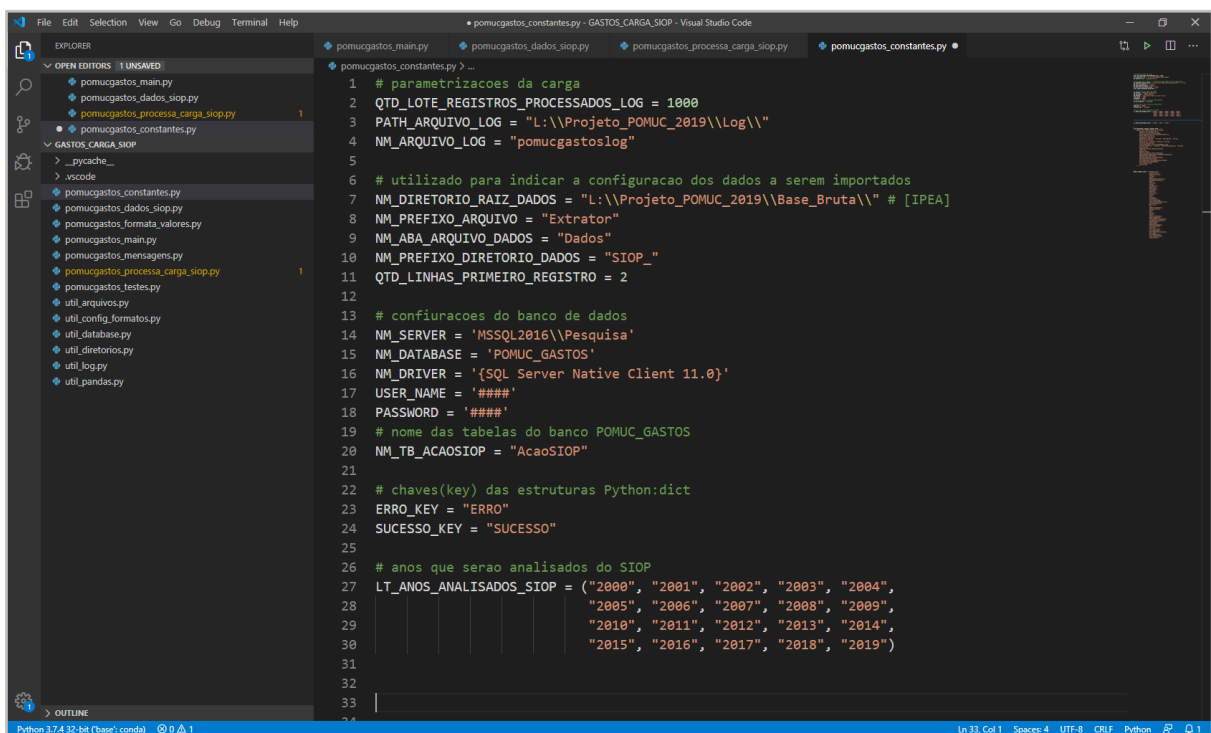


Figura 6. Diagrama de arquitetura do processo de carga. Fonte: Elaborado pelo autor.

A solução implementada permite a parametrização de diversas variáveis para o processo de carga, sendo estas:

- I. Log de operações no console e em arquivo texto previamente configurado;
- II. Local remoto da leitura dos arquivos de dados brutos;
- III. Dados do banco de destino;
- IV. Anos selecionados para a carga de dados.

A aplicação responsável pela tarefa da carga de dados, em sua estrutura, permite ainda que esta seja realizada de forma completa, parcial ou incremental, eliminando assim a necessidade de uma possível deleção completa e nova carga posterior, caso algum erro venha a ocorrer durante o processo. Na figura 7 é possível observar a tela de parametrização das variáveis indicadas.



```

1  # parametrizacoes da carga
2  QTD_LOTE_REGISTROS_PROCESSADOS_LOG = 1000
3  PATH_ARQUIVO_LOG = "L:\\Projeto_POMUC_2019\\Log\\"
4  NM_ARQUIVO_LOG = "pomucgastoslog"
5
6  # utilizado para indicar a configuracao dos dados a serem importados
7  NM_DIRETORIO_RAIZ_DADOS = "L:\\Projeto_POMUC_2019\\Base_Bruta\\" # [IPEA]
8  NM_PREFIXO_ARQUIVO = "Extrator"
9  NM_ABA_ARQUIVO_DADOS = "Dados"
10 NM_PREFIXO_DIRETORIO_DADOS = "SIOP_"
11 QTD_LINHAS_PRIMEIRO_REGISTRO = 2
12
13 # configuracoes do banco de dados
14 NM_SERVER = 'MSSQL2016\\Pesquisa'
15 NM_DATABASE = 'POMUC_GASTOS'
16 NM_DRIVER = '{SQL Server Native Client 11.0}'
17 USER_NAME = '####'
18 PASSWORD = '####'
19 # nome das tabelas do banco POMUC_GASTOS
20 NM_TB_ACAOSIOP = "AcaoSIOP"
21
22 # chaves(key) das estruturas Python:dict
23 ERRO_KEY = "ERRO"
24 SUCESSO_KEY = "SUCESSO"
25
26 # anos que serao analisados do SIOP
27 LT_ANOS_ANALISADOS_SIOP = ("2000", "2001", "2002", "2003", "2004",
28                             "2005", "2006", "2007", "2008", "2009",
29                             "2010", "2011", "2012", "2013", "2014",
30                             "2015", "2016", "2017", "2018", "2019")
31
32
33
34

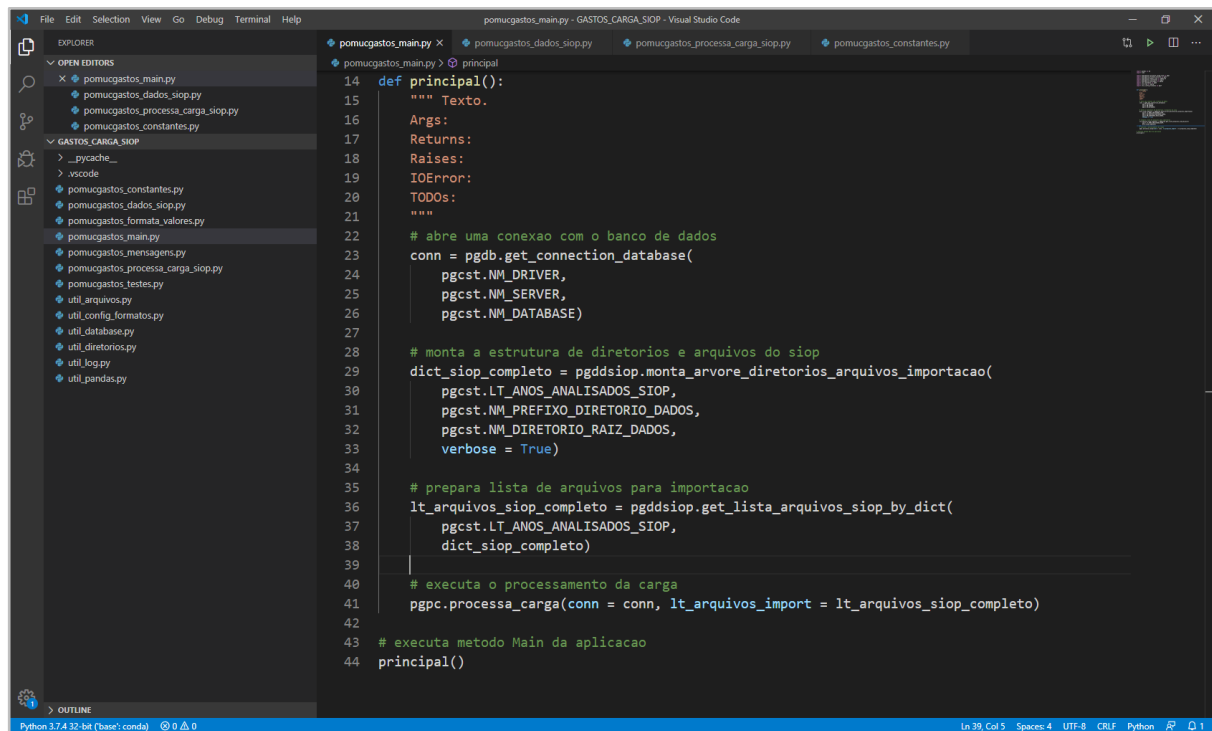
```

Figura 7. Classe de parametrização da aplicação. Fonte: Elaborado pelo autor.

Cabe a ressalva, que a solução foi projetada tomando por base uma execução de maneira *standalone* por meio de tecnologia de processamento paralelo, tal qual *Apache Spark* (SPARK, 2020) ou similares. Porém, também pode ser executado direto pelo console de uma IDE (*Integrated Development Interface*) de desenvolvimento *Python* (3.7 ou superior) ou em servidores que possuam a linguagem disponível previamente instalada e configurada. De todo modo, independente do meio de execução, o tempo de processamento dependerá necessariamente de três fatores básicos: a) quantidade de anos de exercício selecionados para a carga de dados; b)

tipo de execução (completa ou parcial); e, c) infraestrutura de banco de dados de destino (capacidade de processamento, alocação de memória, dentre outros).

Para se dirimir quaisquer dificuldades quanto à configuração de tecnologias não disponíveis no momento da execução do trabalho, a aplicação foi executada diretamente na console da ferramenta de desenvolvimento *Microsoft Visual Code* (MICROSOFT, 2020b), tendo como destino um servidor de banco de dados *Microsoft SQL Server 2016* (MICROSOFT, 2016), para o período de 2000-2019 de registros oriundos do SIOP. O tempo aferido de uma execução da carga completa para os dezenove anos contemplados, levou em média oito horas ininterruptas de processamento, tempo este que pode variar a depender do ambiente computacional. A classe principal da aplicação desenvolvida pode ser observada na figura 8.



```

14 def principal():
15     """ Texto.
16     Args:
17     Returns:
18     Raises:
19     IOError:
20     TODOS:
21     """
22     # abre uma conexao com o banco de dados
23     conn = pgdb.get_connection_database(
24         pgcst.NM_DRIVER,
25         pgcst.NM_SERVER,
26         pgcst.NM_DATABASE)
27
28     # monta a estrutura de diretorios e arquivos do siop
29     dict_siop_completo = pgddsiop.monta_arvore_diretorios_arquivos_importacao(
30         pgcst.LT_ANOS_ANALISADOS_SIOP,
31         pgcst.NM_PREFIXO_DIRETORIO_DADOS,
32         pgcst.NM_DIRETORIO_RAIZ_DADOS,
33         verbose = True)
34
35     # prepara lista de arquivos para importacao
36     lt_arquivos_siop_completo = pgddsiop.get_lista_arquivos_siop_by_dict(
37         pgcst.LT_ANOS_ANALISADOS_SIOP,
38         dict_siop_completo)
39
40     # executa o processamento da carga
41     pgpc.processa_carga(conn = conn, lt_arquivos_import = lt_arquivos_siop_completo)
42
43     # executa metodo Main da aplicacao
44     principal()
  
```

Figura 8. Código referente à classe principal da aplicação. Fonte: Elaborado pelo autor.

A aplicação de carga de dados recupera os arquivos de dados hierarquizados por ano de exercício de sua fonte original no sistema SIOP, a partir de um diretório raiz indicado para a aplicação. Desse modo, os arquivos foram colocados em fila de processamento ordenados pelo menor ano referenciado na lista, ou seja, para o estudo de caso em questão, a carga de dados começou no ano 2000 e finalizou o processo com o ano de 2019, seguindo a ordem crescente.

Além disso, o sistema de *log* permitiu descobrir os erros em tempo real de processamento, sabendo em qual diretório, arquivo de dados e linha o aplicativo processou, facilitando assim o monitoramento, rastreamento e correção de registros com defeito e/ou incompletos / corrompidos de alguma forma.

Desse modo, para cada arquivo de dados, o sistema de *log* contabilizou a quantidade de registros que foi processada e avaliou a completude / corretismo dos registros quanto ao seu preenchimento. Para os registros que apresentaram incoerências de formatos e ou preenchimento, esses registros foram indicados como expurgo e não foram inseridos no banco de dados da solução. Ao final da carga, o arquivo de *log* foi preenchido com um relatório de todo o processo e tomando por base todos os arquivos processados. Na figura 9 se é possível observar parte de um arquivo de *log* real originado e que fora comprimido para fins didáticos de apresentação.



```
[INFO]: início da carga de dados em 14:13:19.
[Arquivo]: L:\Projeto_POMUC_2019\Base_Bruta\SIOP_2003\Extrator1583333670471_2003_F1.xlsx com 100000 registros.
[Inserido]: 0 de 100000 registros.
[Inserido]: 99000 de 100000 registros.
[ERRO]: qtd registros em erro: 0
{'Arquivo processado': 'Extrator1583333670471_2003_F1.xls', 'A processar': '100000', 'Corretos': '100000',
'Erros': '0', 'Expurgo': '0.0%'}
[Arquivo]: L:\Projeto_POMUC_2019\Base_Bruta\SIOP_2003\Extrator1583333870825_2003_F2.xlsx com 99770 registros.
[Inserido]: 0 de 99770 registros.
[Inserido]: 99000 de 99770 registros.
[ERRO]: qtd registros em erro: 0
{'Arquivo processado': 'Extrator1583333870825_2003_F2.xls', 'A processar': '99770', 'Corretos': '99770', 'Erros':
'0', 'Expurgo': '0.0%'}
...
...

[{'Arquivo processado': 'Extrator1583333670471_2003_F1.xls', 'A processar': '100000', 'Corretos': '100000',
'Erros': '0', 'Expurgo': '0.0%'},
{'Arquivo processado': 'Extrator1583333870825_2003_F2.xls', 'A processar': '99770', 'Corretos': '99770', 'Erros':
'0', 'Expurgo': '0.0%'},
{'Arquivo processado': 'Extrator1583334441697_2011_F1.xls', 'A processar': '100000', 'Corretos': '100000',
'Erros': '0', 'Expurgo': '0.0%'},
{'Arquivo processado': 'Extrator1583334744308_2011_F2.xls', 'A processar': '100000', 'Corretos': '100000',
'Erros': '0', 'Expurgo': '0.0%'},
{'Arquivo processado': 'Extrator1583335004070_2011_f3.xls', 'A processar': '81656', 'Corretos': '81656', 'Erros':
'0', 'Expurgo': '0.0%'},
{'Arquivo processado': 'Extrator1583336343923_2016_f1.xls', 'A processar': '100000', 'Corretos': '100000',
'Erros': '0', 'Expurgo': '0.0%'},
{'Arquivo processado': 'Extrator1583336548510_2016_f2.xls', 'A processar': '100000', 'Corretos': '100000',
'Erros': '0', 'Expurgo': '0.0%'},
{'Arquivo processado': 'Extrator1583336741330_2016_f3.xls', 'A processar': '100000', 'Corretos': '100000',
'Erros': '0', 'Expurgo': '0.0%'},
{'Arquivo processado': 'Extrator1583336979243_2016_f4.xls', 'A processar': '65749', 'Corretos': '65749', 'Erros':
'0', 'Expurgo': '0.0%'}]
[INFO]: fim da carga de dados em 14:47:39.
```

Figura 9. Arquivo de *log* gerado pela aplicação. Fonte: Elaborado pelo autor.

Conforme descrito, a tabela AcaoSIOP armazena os dados em seu estado bruto. Dessa forma, após o processamento de 1.7 *gigabytes* (quadro 4) de arquivo plano de dados, o resultado obtido foram de aproximadamente 5.7 milhões de registros para análise, conforme explicitado na figura 10. Cabe a ressalva que as ações orçamentárias do contexto do saneamento básico, ainda precisaram ser encontradas e viabilizadas como escopo, antes da fase de análise dos dados.

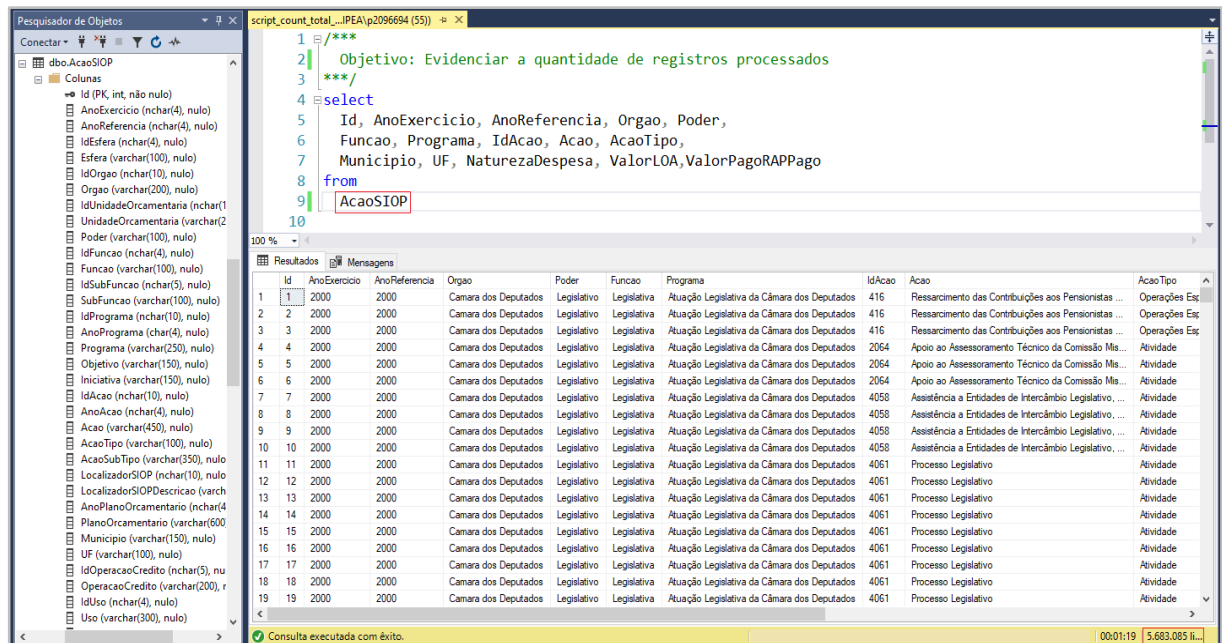


Figura 10. Aproximadamente 5.7 milhões de registros processados. Fonte: Elaborado pelo autor a partir de dados obtidos do SIOP (2019).

### 3.2. Etapa do tratamento de dados

Conforme já evidenciado, o sistema SIOP goza de corretismo em seus dados no que diz respeito ao formato e padronização dos tipos de campos definidos, embora aparentemente não se promova uma padronização forte quanto à uma nomenclatura de preenchimento de algumas variáveis. O que de fato, não representaria maiores problemas caso tal falta não incorresse justamente em uma de suas principais variáveis segundo o MTO 2018, a saber, a Ação Orçamentária (BRASIL, 2018).

Observa-se que essa possível falta de uma nomenclatura e/ou não observância no preenchimento dessa variável, acaba por causar o problema da não unicidade de registro, onde uma mesma ação orçamentária, com o mesmo identificador, tendo o mesmo propósito, acaba por ser identificada de maneira dúbia (figura 11).

Assim sendo, basicamente essa etapa de tratamento de dados consistiu em:

- i. Preencher como vazio, as variáveis que apresentaram erros de tipo de campo. Os erros foram sanados se utilizando da técnica de *missing data*, implementada por meio da biblioteca de manipulação de dados *pandas* (PANDAS, 2020), disponível para a linguagem de programação *Python* e de ampla utilização em projetos / estudos de Ciência de Dados;



- ii. Normalização do preenchimento da variável Ação Orçamentária, tomando como justo o seu último preenchimento encontrado na base de dados, ou seja, caso uma mesma ação orçamentária venha a ter registros para os anos de 2010, 2011 e 2019, por exemplo, este último ano terá o seu valor estabelecido também para os anos anteriores. Ressalta-se que tal decisão se deu metodologicamente, por acreditar que uma possível revisão possa ter sido realizada pelo gestor que operou o SIOF para fins de preenchimento desse campo, optando assim por editá-lo em último caso.

IdAcao	Acao
1	1P95 Apoio à Elaboração de Planos e Projetos de Saneamento em Municípios com População Superior a 50 mil Habitantes ou Integrantes de Regiões Metropolitanas ou de Regiões Integradas de Desenvolvimento.
2	1P95 Apoio à Elaboração de Planos e Projetos de Saneamento em Municípios com População Superior a 50 mil Habitantes ou Integrantes de Regiões Metropolitanas ou de Regiões Integradas de Desenvolvimento
3	1P95 Apoio à Elaboração de Projetos de Saneamento em Municípios de Regiões Metropolitanas, de Regiões Integradas de Desenvolvimento Econômico, Municípios com mais de 50 mil Habitantes ou Integrantes de Consórcios Públicos com mais de 150 mil Habitantes
4	3980 Projetos de Saneamento Básico Integrado (PAT/PROSANEAR)
5	3980 Projetos Integrados de Saneamento Básico
6	3997 Implantação de Serviços de Abastecimento de Água (Saúde e Saneamento no Piauí)
7	3997 Implantação dos Serviços de Abastecimento de Água - Saúde e Saneamento no Piauí - KFW
8	7656 Implantação, Ampliação ou Melhoria do Serviço de Saneamento em Áreas Rurais, em Áreas Especiais (Quilombos, Assentamentos e Reservas Extrativistas) e em Localidades com População Inferior a 2.500 Habitantes para Prevenção e Controle de Agravos
9	7656 Implantação, Ampliação ou Melhoria do Serviço de Saneamento em Localidades com População Inferior a 2.500 habitantes e Áreas Rurais
10	7656 Implantação, Ampliação ou Melhoria de Ações e Serviços Sustentáveis de Saneamento Básico em Pequenas Comunidades Rurais (Localidades de Pequeno Porte) ou em Comunidades Tradicionais (Remanescentes de Quilombos)
11	7656 Implantação, Ampliação ou Melhoria de Ações e Serviços Sustentáveis de Saneamento Básico em Pequenas Localidades, Comunidades Rurais, Tradicionais e Especiais para Prevenção e Controle de Doenças e Agravos
12	7656 Implantação, Ampliação ou Melhoria de Ações e Serviços Sustentáveis de Saneamento Básico em Comunidades Rurais, Tradicionais e Especiais
13	7656 Implantação, Ampliação ou Melhoria de Ações e Serviços Sustentáveis de Saneamento Básico em Comunidades Rurais, Tradicionais e Especiais para Prevenção e Controle de Doenças e Agravos
14	8871 Apoio à Elaboração de Estudos e Implementação de Projetos de Desenvolvimento Institucional e Operacional e à Estruturação da Prestação dos Serviços de Saneamento Básico e Revitalização dos Prestadores de Serviços Públicos de Saneamento
15	8871 Apoio à Elaboração e Monitoramento de Planos de Saneamento Regionais e Nacional
16	8871 Apoio à Elaboração, Implementação e Monitoramento de Planos de Saneamento Básico
17	8871 Apoio à Elaboração, Implementação e Monitoramento de Planos Nacional e Regionais de Saneamento Básico
18	8871 Apoio à Elaboração e Monitoramento de Planos de Saneamento Regionais e Nacional

Consulta executada com êxito. 18 linhas

Figura 11. Evidências de inconsistências no preenchimento. Fonte: Elaborado pelo autor a partir de dados obtidos do SIOF (2019).

### 3.3. Etapa da seleção do escopo

Uma vez concluída as etapas anteriores, se foi possível acessar a tabela AcaoTratada (quadro 5) e desse modo realizar as consultas e marcações das ações orçamentárias que serão alvo da análise dos gastos públicos federais em saneamento básico.

Basicamente, a estratégia aplicada foi por meio do uso da técnica de *text mining* (SILGE; ROBINSON, 2017), recuperar as ações orçamentárias (quadro 1) que possuam em alguma das suas variáveis o *token* descritivo “saneamento”. Este processo por sua vez, retornou 47 ocorrências de registros. Porém, observou-se que algumas ações contemplam o eixo do Saneamento Ambiental, e não do Saneamento Básico, estritamente, conforme o objetivo de pesquisa deste estudo. Dessa forma, 4 ações orçamentárias foram descartadas desse primeiro filtro por se tratarem da pauta do Saneamento Ambiental, restando assim, 43 ações direcionadas ao tema desejado.

Por fim, as ações orçamentárias selecionadas tiveram sua variável Escopo (quadro 2) preenchida, conforme pode ser observado na figura 12.

script\_consulta\_res... (PEA\p2096694 (74)) script\_mining\_acoe... (PEA\p2096694 (63))

```

1 1 /**
2 2 Objetivo: Evidenciar o processo de seleção de escopo realizado
3 3 /**/
4 4 select Id, IdAcao, Acao, NomeEscopo, Descricao from Escopo
5 5

```

	Id	IdAcao	Acao	NomeEscopo	Descricao
1	1	006H	Apoio a Empreendimentos de Saneamento Integrado em Assentamentos Precários em Municípios de Regiões Metropolitanas, de R...	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
2	2	006L	Apoio à Elaboração de Projetos de Saneamento em Municípios de Regiões Metropolitanas, de Regiões Integradas de Desenvolvim...	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
3	3	0582	Apoio a Projetos de Saneamento Integrado em Municípios com População de até 20 mil Habitantes na Região do Semi-Árido	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
4	4	0586	Apoio a Projetos de Ação Social em Saneamento (PASS)	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
5	5	0800	Apoio à Gestão dos Sistemas de Saneamento Básico em Municípios de até 30.000 Habitantes	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
6	6	1083	Melhoria das Condições Habitacionais, de Infra-Estrutura e de Saneamento Básico	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
7	7	10GC	Implantação e Melhoria de Serviços de Saneamento em Escolas Públicas Rurais - "Saneamento em Escolas"	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
8	8	10S5	Apoio a Empreendimentos de Saneamento Integrado em Municípios com População Superior a 50 mil Habitantes ou Municípios Int...	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
9	9	10T1	Apoio a Projetos de Ação Social em Saneamento (PASS)	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
10	10	10TA	Elaboração de Projetos de Saneamento nas Bacias Receptoras do São Francisco para Municípios com população abaixo de 50.00...	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
11	11	10TB	Elaboração de Projetos de Saneamento nas Bacias Receptoras da Integração com o Rio São Francisco em Municípios com mais d...	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
12	12	1P95	Apoio à Elaboração de Planos e Projetos de Saneamento em Municípios com População Superior a 50 mil Habitantes ou Integrante...	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
13	13	20AG	Apoio à Gestão dos Sistemas de Saneamento Básico em Municípios de até 50.000 Habitantes	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
14	14	20NW	Apoio a Estruturação e Implementação do Sistema Nacional de Informações em Saneamento Básico - SINISA	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
15	15	20Q8	Apoio à Implantação e Manutenção dos Sistemas de Saneamento Básico e Ações de Saúde Ambiental	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
16	16	20Z5	Apoio à Gestão e à Capacitação aplicados ao Saneamento	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
17	17	216F	Gestão da Política de Saneamento Básico	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
18	18	219R	Melhoria da Qualidade Regulatória do Setor de Saneamento	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
19	19	232	Dívidas Internas do Extinco Departamento Nacional de Obras de Saneamento - DNOS, Assumidas Pela União (Lei Nº 8. 029/90)	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
20	20	234	Dívidas Externas do Extinco Departamento Nacional de Obras de Saneamento - DNOS, Assumidas Pela União (Lei Nº 8. 029/90)	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO
21	21	2170	Implantação de Serviços de Saneamento Básico em Municípios com População Superior a 75 Mil Habitantes	Escopo 1	1-ESCOPO AÇÕES SANEAMENTO BASICO

Consulta executada com êxito. 00:00:00 43 linhas

Figura 12. Evidência das 43 ações orçamentárias selecionadas. Fonte: Elaborado pelo autor a partir de dados obtidos do SIOP (2019).

#### 4. Análise e Exploração dos Dados

Concluída a fase de processamento e carga dos dados, optou-se por criar uma camada de acesso aos dados que possibilite segmentar as consultas por assuntos estratégicos. A camada cria uma visão consolidada suportando informações específicas que se desejam obter, onde, para o referido trabalho tratam-se dos gastos públicos federais em políticas públicas para o desenvolvimento do saneamento básico.

Dessa forma, a criação de objetos *datasets* (visões) para análise dos dados, evita o trabalho “ad hoc” de manipulação, mitiga a questão da redundância das consultas, minimiza o retrabalho para construção dos relatórios, e ainda, promove a padronização dos formatos para se obter as respostas desejadas. Cada camada é mais focada à um mesmo domínio ou tema.

Ademais, a camada deve possibilitar a extração de informações por meio de diversas plataformas e *softwares* (figura 13), uma vez que se tomou a decisão por sua

implementação via tabela virtual (*View*), a qual é implementada em diversos bancos de dados relacionais comerciais, são facilmente consumidas via SQL ANSI (*Structured Query Language*) (GUIMARÃES, 2003), e possuem basicamente as funções: a) simplificar a percepção que cada usuário tem do banco de dados; b) permitir aos usuários acessarem dados por meio da exibição, sem conceder permissões para acessar diretamente a estrutura do banco ou de tabelas específicas (MICROSOFT, 2020a).

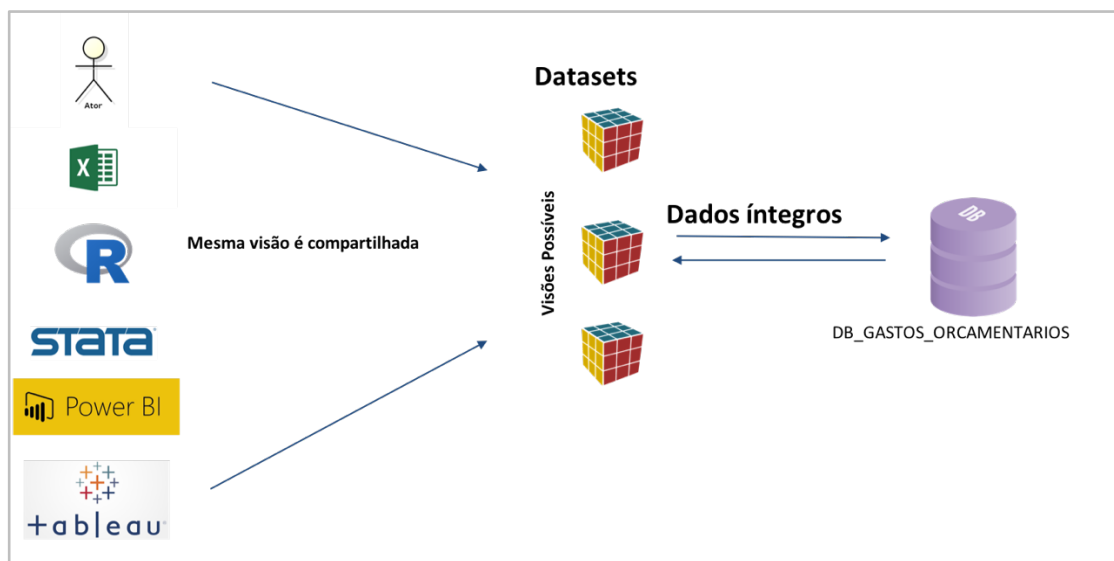


Figura 13. Diagrama de arquitetura dos *datasets* de análise. Fonte: Elaborado pelo autor.

A decisão técnica de optar pela elaboração de visões mais detalhadas e específicas para a(s) pergunta(s) a ser(em) analisada(s) levou ainda em consideração a complexidade de entendimento da base de dados do SIOP, a qual possui uma extensa lista de variáveis e seus domínios de preenchimentos possíveis a serem levados em consideração a fim de se obter os resultados de consulta almejados. Dessa forma, sua implementação forneceu ao estudo de caso os seguintes benefícios:

- I. Ganho de performance e confiabilidade;
- II. Facilidade para exportar os dados trabalhados;
- III. Possibilidade de validação e replicação da informação;
- IV. Possibilidade de junções de dados.

No exemplo demonstrado na figura 14, foi contemplada uma implementação utilizando a técnica descrita. Para o caso em específico - fora do escopo de análise desse trabalho -, se implementou um *dataset* que busca agregar os gastos governamentais para a função Gestão Ambiental, contidos na base de dados do SIOP para o período de janeiro de 2000 até dezembro 2019.

Desse modo, é importante notar na primeira parte da implementação a criação da tabela virtual, a qual possui regras de negócio envolvida, retornando, portanto, as variáveis Ano Exercício, Funcao, ValorLOATotal e ValorPagoRAPPagoTotal (quadro 1), em estrutura compatível com uma série temporal.

Já na segunda parte da implementação, agora sem maiores complexidades, um usuário sem qualquer conhecimento da estrutura de dados do SIOP, pode facilmente executá-la se utilizando de conectores para consultas SQL em uma planilha eletrônica, por exemplo, ou um pacote estatístico de sua preferência, ou ainda, ferramentas de *Business Intelligence* (BI) (TURBAN et al, 2009, p.27), e a partir daí realizar suas análises. Portanto, não é difícil de se perceber como produto final dessa etapa, a redução de complexidade e os ganhos de confiabilidade e agilidade para os consumidores finais da informação.

```

1  -- Criação do objeto View que fará a função de dataset
2  create view dataset_evolucao_gastos_gestaoambiental
3  as
4  select
5  AnoExercicio, Funcao,
6  FORMAT(sum(ValorLOA), 'c', 'pt-BR') as 'ValorLOATotal',
7  FORMAT(sum(ValorPagoRAPPago), 'c', 'pt-BR') as 'ValorPagoRAPPagoTotal'
8  from
9  AcaoSIOP
10 where
11 IdFuncao = 18 -- Gestão ambiental
12 group by
13 AnoExercicio, Funcao
14 order by
15 ValorPagoRAPPago asc
16
17 -- Exemplo de consulta após a criação do dataset
18 select * from dataset_evolucao_gastos_gestaoambiental order by ValorPagoRAPPagoTotal desc

```

Figura 14. Exemplo de implementação de *dataset*. Fonte: Elaborado pelo autor.

Para o referido estudo, os seguintes objetos de *datasets* foram criados para dar suporte às perguntas propostas, bem como análises gráficas:

Quadro 6. Relação de *datasets* implementados.

Fonte: Elaborado pelo autor.

Id	Dataset	Objetivo
01	dataset_serie_historica_gastos_saneamento_basico.sql	Demonstrar qual foi a evolução dos gastos públicos federais para a série histórica (2000 a 2019) disponível no SIOP para projetos de saneamento básico. Verificar se houveram acréscimos significativos após a implementação do Plansab.

02	dataset_gastos_saneamento_basico_plansab.sql	Demonstrar qual foi o orçamento federal aplicado para o investimento em saneamento básico no Brasil a partir de 2013, ano de lançamento do Plansab.
03	dataset_gastos_saneamento_basico_georreferenciados.sql	Demonstrar qual foi a distribuição geográfica dos gastos públicos federais em projetos de saneamento básico, para a série histórica (2000 a 2019) disponível no SIOP.

## 5. Apresentação e Análise dos Resultados

Ao iniciar essa última seção explicativa do trabalho, é importante salientar que o seu propósito é, sobretudo, o de evidenciar as abordagens analíticas provenientes de todo o trabalho anterior que contemplou os objetivos do estudo, os quais versam no conjunto de teorias, técnicas e ferramentas da Ciência de Dados para fins de elucidação de problemas que envolvam as complexas atividades de modelagem, manipulação, análise e interpretação de dados.

Cabe sempre ressaltar, que o foco prioritário desta seção de apresentação de resultados - apesar do forte viés de uma análise econômica -, não é prioritariamente sua interpretação finalística, e sim, demonstrar possibilidades e uma abordagem de análise dos dados resultantes de maneira resguardada por evidências empíricas, uma vez que se trata de um estudo de caso em que se utiliza de bases de dados reais.

Houve assim, um esforço de entendimento e interpretação dos importantes resultados que foram obtidos, mesmo quando da não possibilidade plena de sua interpretação, como é o caso, por exemplo, das variáveis de localização ausentes de preenchimento para a maior parte dos registros da série obtida do SIOP, a saber, UF e Município. E dessa forma, as subseções (5.3, 5.4 e 5.5) responsáveis por efetuarem uma análise georreferenciada teve que focar apenas numa demonstração de tendência. Sobretudo, para os registros que possuem essa possibilidade.

Para fins de resumo esquemático, a ser complementado no transcorrer da referida seção, observa-se no quadro 7 um descritivo dos gastos públicos federais analisados no trabalho.

Quadro 7. Resumo dos gastos públicos federais na agenda do saneamento básico no Brasil.

Fonte: Elaborado pelo autor a partir de obtidos do SIOP (2019).

Gastos totais	Valor	Observação
Total de gastos federais para a série histórica (2000-2019) na agenda de saneamento básico	R\$ 5.281.769.476,93	O montante não inclui gastos extra orçamentário e/ou investimento privados de qualquer natureza.
Total de gastos federais para a série histórica (2014-2019) reduzida na agenda de saneamento básico após a implantação do Plansab	R\$ 2.178.634.936,92	Em 6 anos, o Plansab representou o total de 41,24% dos gastos federais em saneamento básico no Brasil.
Total de gastos federais para a série histórica (2000-2019) com a agenda de saneamento básico e que <u>possuem</u> as variáveis georreferenciadas preenchidas	R\$ 2.254.644.948,82	Esse montante representa, portanto, um percentual de 41,24% dos gastos, onde se é possível rastrear em qual UF se deu o dispêndio.
Total de gastos federais na série histórica (2000-2019) com a agenda de saneamento básico e que <u>não possuem</u> as variáveis georreferenciadas preenchidas	R\$ 3.027.124.528,11	Esse montante representa, portanto, um percentual de 58,76% dos gastos, onde os registros não possibilitam como rastrear em qual UF se deu o dispêndio.

Informações sobre os dados finais estruturados em série temporal:

- Um conjunto de 43 ações orçamentárias agrupadas foram selecionadas, conforme já evidenciado anteriormente;
- Dentre as 43 ações orçamentárias agrupadas, originaram-se 1953 ações não agrupadas, ou seja, lançamentos dos gastos sob o identificador comum de uma mesma ação orçamentaria. Deste último quantitativo, 56 lançamentos não possuem a informação de UF, o qual representa um total de 2,8% dos registros sem dados georreferenciados;
- A média dos gastos, durante toda a série histórica disponível (2000-2019), foi de R\$ 186.512,99. Para seu valor mínimo foi auferido o montante de R\$ 1.400,00, e seu valor máximo foi de R\$ 4.100.000,00;
- Para os resultados apresentados a seguir (5.3, 5.4 e 5.5) utilizando dos recursos de georreferenciamento, é importante destacar que: a) na cor verde estão representados os valores abaixo da média histórica de R\$ 186.512,99; b) na cor laranja, os gastos em ações de até R\$ 2.000.000,00; e finalmente, c) na cor

vermelha, se encontram os gastos em ações com valores acima de R\$ 2.000.000,00 até o valor máximo registrado de R\$ 4.100.000,00.

### 5.1. Análise da série histórica dos gastos em saneamento básico

O Programa de Aceleração do Crescimento (PAC) foi criado no ano de 2007 durante o Governo Lula com o objetivo de incentivar o desenvolvimento / atividade econômica, baseado em investimento em infraestrutura no Brasil por meio da execução de grandes obras, sendo estas no segmento de saneamento básico, infraestrutura social e urbana, logística e energética (BRASIL, 2007); fato este que ajuda a explicar o incremento significativo nos gastos em saneamento básico no país a partir desse período, conforme pode ser observado no gráfico (figura 15).

- A média de gastos para a série foi de: R\$ 260.982,78.
- O valor mínimo da série foi de R\$ 00,00 no ano de 2000.
- O valor máximo da série foi de R\$ 91.776.991,69 no ano de 2009.

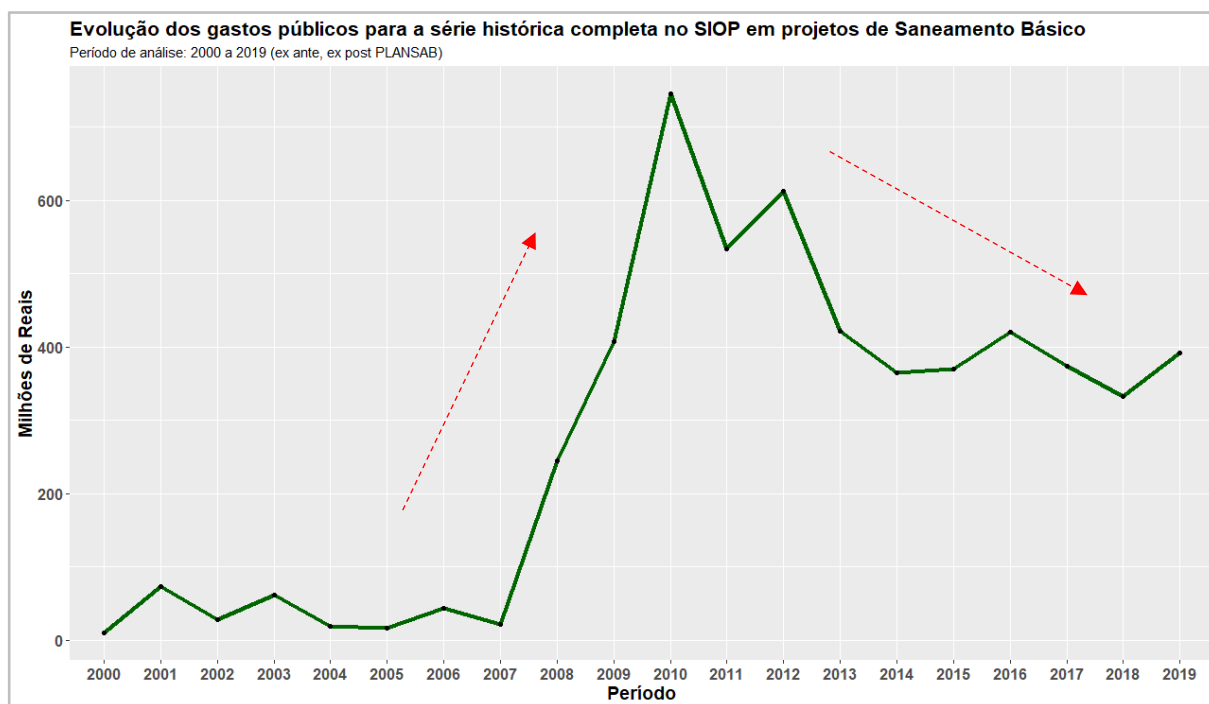


Figura 15. Série histórica completa para os gastos em saneamento básico. Fonte: Elaborado pelo autor a partir de dados obtidos do SIOF (2019).

### 5.2. Uma análise do Plansab sobre os gastos em saneamento básico

O PAC em sua segunda fase (a partir de 2011), sob o Governo Dilma Rousseff, manteve os mesmos objetivos definidos em sua criação (BRASIL, 2007), porém, o governo lançou ainda o Plano Nacional de Saneamento Básico no ano de 2013, por

meio do Ministério das Cidades (BRASIL, 2013). Fato este que impulsiona novamente os gastos públicos federais em saneamento básico no país a partir desse período, conforme pode ser observado no gráfico (figura 16), evidenciando, portanto, a eficácia do referido plano na promoção de políticas públicas voltadas à pauta de saneamento básico. Não obstante, tais investimentos e priorização da agenda demonstram uma forte interrupção nos gastos e execução do programa logo após o afastamento da mandatária e em seguida, seu processo de impeachment em 2016.

- A média de gastos para a série foi de: R\$ 323.657,20.
- O valor mínimo da série foi de R\$ 00,00 no ano de 2018.
- O valor máximo da série foi de R\$ 48.680.542,97 no ano de 2016.

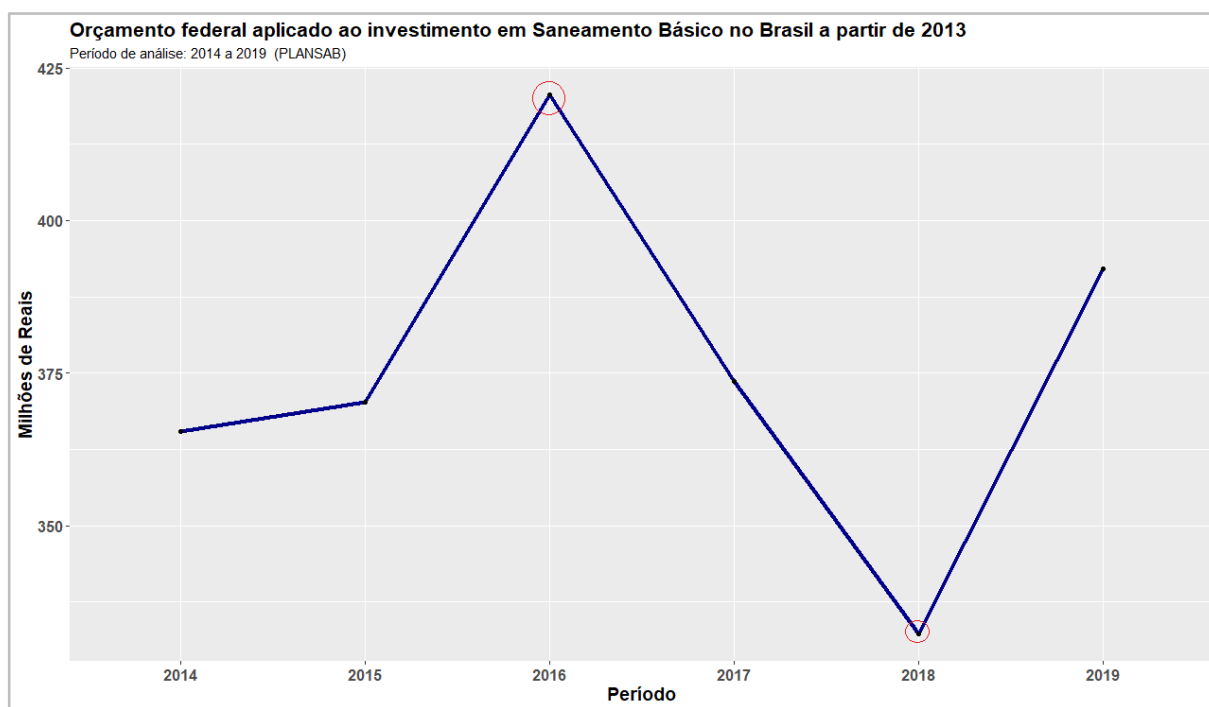


Figura 16. Série do período de vigência do Plansab sobre os gastos em saneamento básico. Fonte: Elaborado pelo autor a partir de dados obtidos do SIOP (2019).

### 5.3. Análise dos gastos georreferenciados nas regiões Norte e Nordeste

Para o caso das regiões Norte e Nordeste - analisadas conjuntamente, a média de gastos para a série histórica completa ficou acima do valor médio de R\$ 186.512,99, que considera todo o território nacional. É possível observar uma predominância de projetos na região Nordeste, frente a região Norte, bem como a maior quantidade de gastos de baixo valor investido, ou seja, abaixo da média geral dos gastos, indicados pela cor verde no mapa (figura 17). Ademais, apenas um gasto figurou acima dos R\$ 2.000.000,00, o qual foi realizado na região Nordeste.





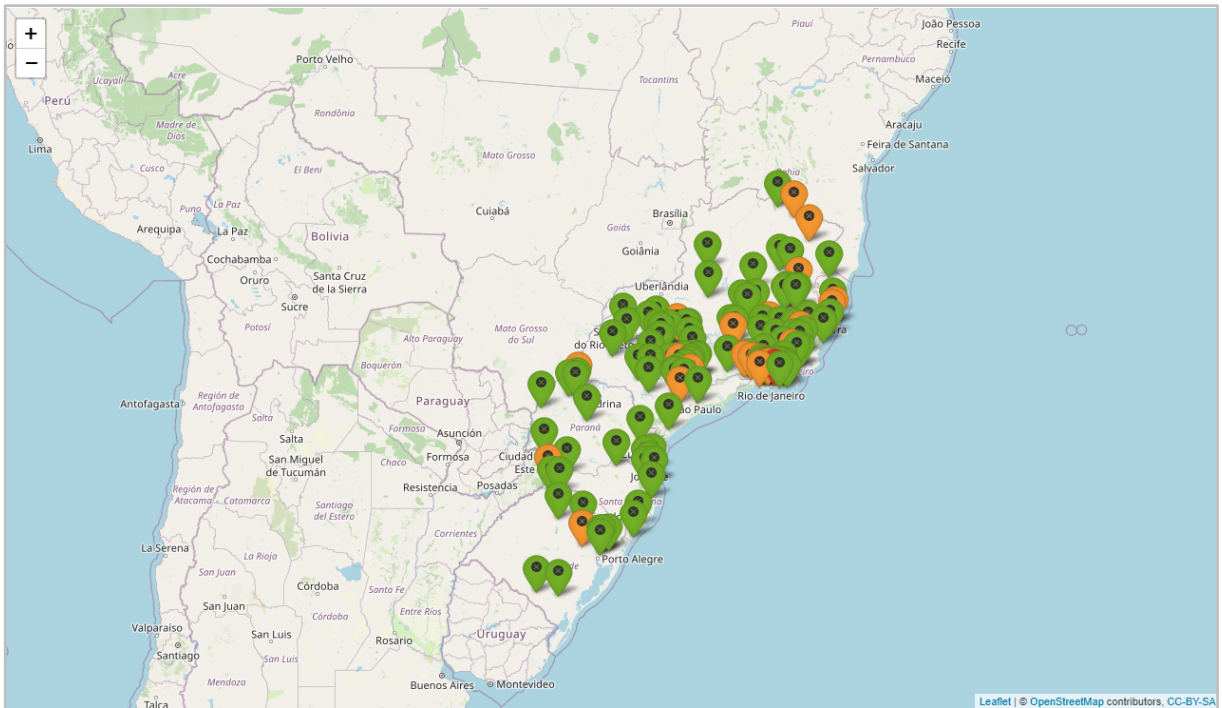


Figura 18. Distribuição geográfica dos gastos para as regiões Sul e Sudeste. Fonte: Elaborado pelo autor a partir de dados obtidos do SIOP (2019) e do IBGE (2019).

### 5.5. Análise dos gastos georreferenciados na região Centro-Oeste

Por fim, a região Centro-Oeste apresentou uma média de gastos bem acima do valor médio da série histórica, considerando todo o território nacional. É possível observar ainda a baixa distribuição de gastos para a região, bem como uma elevação média dos gastos frente às demais regiões analisadas (figura 19).

- A média de gastos para a série foi de: R\$ 273.502,70.
- O valor mínimo da série foi de R\$ 50.000,00.
- O valor máximo da série foi de R\$ 2.000.000,00.

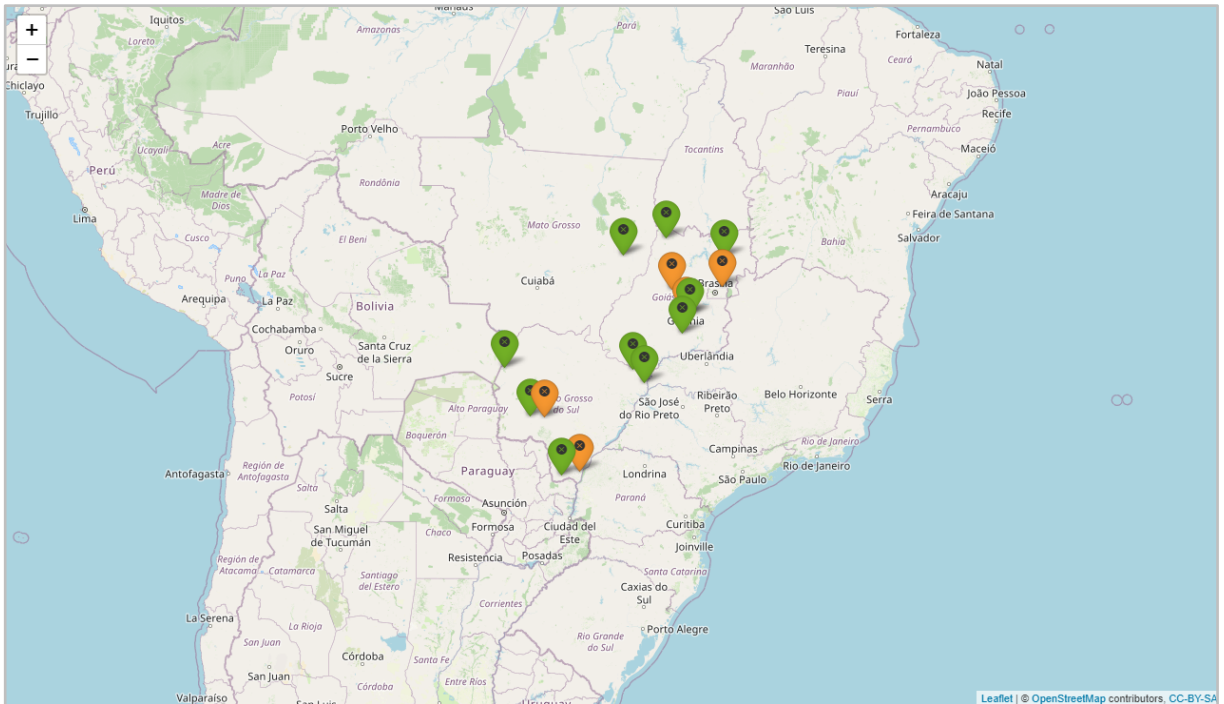


Figura 19. Distribuição geográfica dos gastos para a região Centro-Oeste. Fonte: Elaborado pelo autor a partir de dados obtidos do SIOP (2019) e do IBGE (2019).

## 6. Conclusão

É salutar a evidência teórica em qualquer tema que envolva a avaliação de políticas públicas. Desse modo, quando se optou por fazer desse Trabalho de Conclusão do Curso de Especialização em Ciência de Dados e Big Data, um estudo de caso aplicado, veio em conjunto a preocupação em elucidar conceitos teóricos, evidenciar a sistematização de cada fase do processo técnico executado, bem como realizar as análises baseadas em evidências. Sendo assim, tornou-se necessário o desenho de uma solução robusta de análise de dados, sua documentação e a utilização de fontes de dados reais aplicáveis ao tema abordado.

Uma vez definido o tema, bem como as questões a serem respondidas, um esforço de pesquisa e entendimento das bases de dados governamentais requeridas foi empregado a fim de selecionar as variáveis foco da análise de dados do estudo.

Inicialmente se fez uso da técnica de modelagem entidade relacional para a estruturação dos dados obtidos, uma vez que por meio desta se é possível realizar junções de dados diversos tornando um todo coeso e aplicável ao estudo de caso.

Posteriormente, o processo de carga de dados foi baseado no acesso formal aos repositórios. Comumente hoje utilizadas, as técnicas de *web scraping* possuem limitações tanto à padronização dos dados, quanto a forma de acesso, oferecendo certos riscos de inconsistências para a soluções técnicas. Ademais, optou-se por construir uma solução para fins de processamento automatizado, capaz de fornecer parametrizações e *log* das transações.

Em seguida, uma camada (*datasets*) de acesso aos dados foi criada para possibilitar a segmentação das consultas por questões estratégicas. Para essa camada, voltada para as questões específicas, contemplou-se: a) visão da série histórica completa; b) série segmentada pelo advento do Plansab; c) série histórica georreferenciada.

Finalizando a parte técnica do trabalho, na seção de apresentação dos resultados, cinco análises foram desenvolvidas objetivando a resposta para as questões colocadas no problema de pesquisa. Ainda, aplicou-se os conceitos de análise de dados georreferenciados com objetivo de fornecer maior robustez ao estudo por meio do indicativo da espacialização dos gastos públicos federais aplicados em saneamento básico no Brasil.

Necessariamente se é preciso fundamentar para todo estudo aplicado, um referencial teórico sólido e com evidências empíricas. De toda forma, é preciso garantir que os dados utilizados sejam íntegros e que não foram manipulados incorretamente para a obtenção de resultados obtidos. Fatos estes que desde o princípio foram premissas seguidas por esse autor. Desse modo, o estudo viabiliza: a) uma definição clara dos métodos e procedimentos utilizados para cada resultado obtido; b) viabilização da verificação independente e da replicabilidade dos resultados; c) disponibilização dos dados de pesquisa processados e utilizados.

De todo modo, esse estudo conclui que as técnicas de Ciência de Dados foram efetivas para a execução dos objetivos propostos, sendo estas de fundamental importância na seleção, tratativa, sistematização e entendimento dos dados governamentais oficiais disponíveis a tempo do referido trabalho.

Ademais, conclui-se também que os gastos públicos federais na agenda do saneamento básico no Brasil sofreram grande impacto positivo, ou seja, investiu-se

mais em obras de infraestrutura a partir do ano de 2007 com a implementação do PAC em sua fase 1, tendo como maior pico desses gastos o ano de 2009. De fato, outra ação de alavancagem nos investimentos da referida agenda foi a implementação do Plansab a partir do ano de 2013, tendo como o seu auge o ano de 2016.

## **Referências Bibliográficas**

BRASIL. Lei nº 11.445, de 5 de janeiro 2007. Estabelece diretrizes nacionais para o saneamento básico e no seu art. 52 determina a elaboração do Plano Nacional de Saneamento Básico (Plansab), sob a coordenação do Ministério das Cidades (MCidades), 2007.

BRASIL. Ministério do Planejamento, Desenvolvimento e Gestão, Secretaria de Orçamento Federal, Manual Técnico de Orçamento – MTO, 2018, 166p.

BRASIL. Plano Nacional de Saneamento Básico. Dezembro de 2013. Disponível em: <<https://www.mdr.gov.br/saneamento/proeesa/89-secretaria-nacional-de-saneamento/3137-plano-nacional-de-saneamento-basico-plansab>>. Acesso em: 26/01/2020.

BRASIL. Sistema Integrado de Planejamento e Orçamento - SIOP. Lei Orçamentária Anual (LOA) 2000 a 2019. Disponível em: <<https://www.siop.planejamento.gov.br/>> Acesso em: 26/01/2020.

BRASIL. SOBRE O PAC. Janeiro de 2007. Disponível em: <<http://pac.gov.br/sobre-o-pac>> Acesso em: 17/01/2020.

CÂMARA DOS DEPUTADOS. LOA - Lei Orçamentária Anual. Janeiro de 2020. Disponível em: <<https://www2.camara.leg.br/orcamento-da-uniao/leis-orcamentarias/loa>>. Acesso em: 26/01/2020.

CHEN, P. Entity-Relationship Modeling: Historical Events, Future Trends, and Lessons Learned. In: Broy M., Denert E. (eds) Software Pioneers. Berlin, Heidelberg: Springer. 2002.

COUGO, P. MODELAGEM CONCEITUAL e Projeto de Banco de Dados. Rio de Janeiro: Editora CAMPUS, 1997.

FOWLER, M. Padrões de Arquitetura de Aplicações Corporativas. Porto Alegre: Bookman. 2006.

GITHUB. Built for developers. Janeiro de 2020. Disponível em: <<https://github.com/>>. Acesso em: 20/06/2020.

GUIMARÃES, C. C. Fundamentos de Bancos de Dados: Modelagem, Projeto e Linguagem. Campinas: Editora da Unicamp. 2003.

IBGE. Divisão Territorial Brasileira – DTB. Janeiro de 2019. Disponível em: <<https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/23701-divisao-territorial-brasileira.html>>. Acesso em: 16/04/2020.

IBGE. Produto Interno Bruto – PIB. Abril de 2020. Disponível em: <<https://www.ibge.gov.br/explica/pib.php>>. Acesso em: 16/06/2020.

INEPDATA. Dados Abertos – InepData. Janeiro de 2020. Disponível em: <<http://inep.gov.br/web/guest/inep-data>>. Acesso em 16/04/2020.

MICROSOFT. CREATE VIEW (Transact-SQL). Abril de 2020a. Disponível em: <<https://docs.microsoft.com/pt-br/sql/t-sql/statements/create-view-transact-sql?view=sql-server-ver15>>. Acesso em 20/06/2020.

MICROSOFT. SQL Server 2016. Dezembro de 2016. Disponível em: <<https://www.microsoft.com/pt-br/sql-server/sql-server-2016>>. Acesso em: 20/06/2020.

MICROSOFT. Visual Studio 2019. Março de 2019. Disponível em: <<https://visualstudio.microsoft.com/pt-br/vs/>>. Acesso em: 20/06/2020.

MICROSOFT. Visual Studio Code - Getting Started. Fevereiro de 2020b. Disponível em: <<https://code.visualstudio.com/docs>>. Acesso em: 20/06/2020.

PANDAS. API reference. Janeiro de 2020. Disponível em: <<https://pandas.pydata.org/docs/reference/index.html>>. Acesso em 23/03/2020.

PNUD BRASIL. O IDHM do Brasil. Agosto de 2016. Disponível em: <<https://www.br.undp.org/content/brazil/pt/home/library/idh/o-idhm-do-brasil.html>>. Acesso em 26/06/2020.

PNUD BRASIL. O que é o IDH. Janeiro de 2020. Disponível em: <<https://www.br.undp.org/content/brazil/pt/home/idh0/conceitos/o-que-e-o-idh.html>>. Acesso em 26/06/2020.

PYTHON. Getting Started. Janeiro de 2020. Disponível em: <<https://www.python.org/about/>>. Acesso em: 20/06/2020.

R PROJECT. The R Project for Statistical Computing. Fevereiro de 2020. Disponível em: <<https://www.r-project.org/>>. Acesso em: 20/06/2020.

RSTUDIO. RStudio - Take control of your R code. Maio de 2020. Disponível em: <<https://blog.rstudio.com/2020/05/27/rstudio-1-3-release/>>. Acesso em: 20/06/2020.

SENADO FEDERAL. Projeto de Lei nº 4162, de 2019. Julho de 2020. Disponível em: <<https://www25.senado.leg.br/web/atividade/materias/-/materia/140534>>. Acesso em: 22/07/2020.

SENADO FEDERAL. Senado aprova novo marco legal do saneamento básico. Junho de 2020. Disponível em: <<https://www12.senado.leg.br/noticias/materias/2020/06/24/senado-aprova-novo-marco-legal-do-saneamento-basico>>. Acesso em: 22/07/2020.

SILGE, J.; ROBINSON, D. Text Mining with R. United States of America: O'Reilly Media, 2017.

SIOPDOC. Bem-vindo à Documentação de Usuário do SIOP. Janeiro de 2020. Disponível em: <<https://www1.siop.planejamento.gov.br/siopdoc/doku.php>>. Acesso em 26/01/2020.

SPARK, A. Spark Overview. Janeiro de 2020. Disponível em: <<https://spark.apache.org/docs/latest/>>. Acesso em 23/06/2020.

TURBAN, E.; SHARDA, R.; ARONSON, J. E.; KING, D. Business Intelligence: Um enfoque Gerencial para a Inteligência de Negócio. São Paulo: bookman, 2009.

## APÊNDICE

### I. Descritivo das tecnologias / ferramentas utilizadas no projeto

Quadro 8. Ferramentas utilizadas para desenvolvimento do estudo.

Fonte: Elaborado pelo autor baseado em MICROSOFT (2016), PYTHON (2020), R PROJECT (2020), RSTUDIO (2020), MICROSOFT (2020b), MICROSOFT (2019) e GITHUB (2020).

Nome	Descrição	Justificativa
Microsoft SQL Server 2016	Software de banco de dados proprietário do fabricante Microsoft.	Ademais, a plataforma é líder de mercado quanto à gestão de dados, amplamente difundido, confiável e robusto para a missão.
Python 3.7	Linguagem de programação open source de múltiplo propósito.	É uma linguagem de programação amplamente utilizada no meio técnico e científico, a qual possui diversos pacotes (bibliotecas) e facilidades para se trabalhar com manipulação de arquivos de dados, bem como análise de dados.
R 3.6	Linguagem de programação estatística Open Source.	É uma linguagem de programação amplamente utilizada no meio técnico e científico, a qual possui diversos pacotes (bibliotecas) e facilidades para se trabalhar com manipulação de arquivos de dados, bem como análise de dados.
RStudio 1.3	IDE de desenvolvimento Open Source para a linguagem R.	Trata-se de uma ferramenta robusta tecnicamente para desenvolvimento em R e não ocasiona custos financeiros ao projeto.
Microsoft Visual Code	IDE de desenvolvimento de propósito geral Open Source, fabricada e mantida pela Microsoft.	Trata-se de uma ferramenta robusta tecnicamente para desenvolvimento em Python e não ocasiona custos financeiros ao projeto.
Microsoft Visual Studio 2019	IDE de desenvolvimento de propósito geral Open Source, fabricada e mantida pela Microsoft.	A ferramenta em questão é padrão das soluções desenvolvidas baseadas nas tecnologias Microsoft, desse modo, o projeto banco de dados, se utilizou das facilidades e incremento da produtividade fornecidos pelo ecossistema do fabricante.
GitHub	Trata-se de uma plataforma de Open Source de hospedagem de código-fonte com controle de versão usando o Git.	A ferramenta se tornou pouco a pouco padrão de mercado, com ampla aceitação em projetos tanto no meio privado quanto no meio científico e acadêmico.



## II. Links para divulgação do trabalho

Conforme determinação das diretrizes para a elaboração do trabalho de conclusão de curso, no referido trabalho os seguintes recursos externos foram criados:

Quadro 9. Recursos disponíveis para divulgação do estudo.

Fonte: Elaborado pelo autor.

Recurso	Link
Vídeo da apresentação do trabalho	
Link do repositório GitHub com os códigos produzidos	
Link do <i>dataset</i> gerado a partir de extração do banco de dados	