



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
CURSO DE GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Gustavo Fukushima de Salles

**Extração, transformação e carga de dados sobre compras da Prefeitura de
Florianópolis usando Airflow**

Florianópolis
2025

Gustavo Fukushima de Salles

**Extração, transformação e carga de dados sobre compras da Prefeitura de
Florianópolis usando Airflow**

Trabalho de Conclusão de Curso do Curso de
Graduação em Ciências da Computação do Centro
Tecnológico da Universidade Federal de Santa
Catarina para a obtenção do título de bacharel em
Ciências da Computação.
Orientador: Prof. Renato Fileto, Dr.

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

de Salles, Gustavo Fukushima
Extração, transformação e carga de dados sobre compras da
Prefeitura de Florianópolis usando Airflow / Gustavo
Fukushima de Salles ; orientador, Renato Fileto, 2025.
64 p.

Trabalho de Conclusão de Curso (graduação) -
Universidade Federal de Santa Catarina, Centro Tecnológico,
Graduação em Ciências da Computação, Florianópolis, 2025.

Inclui referências.

1. Ciências da Computação. 2. Bancos de dados. 3.
Extração, transformação e carga. 4. Compras públicas. I.
Fileto, Renato. II. Universidade Federal de Santa
Catarina. Graduação em Ciências da Computação. III. Título.

Gustavo Fukushima de Salles

**Extração, transformação e carga de dados sobre compras da Prefeitura de
Florianópolis usando Airflow**

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de bacharel em Ciências da Computação e aprovado em sua forma final pelo Curso de Graduação em Ciências da Computação.

Florianópolis, 6 de dezembro de 2025.

Lúcia Helena Martins Pacheco, Dra.
Coordenadora do Curso

Banca Examinadora:

Prof. Renato Fileto, Dr.
Orientador

Profa. Carina Friedrich Dorneles, Dra.
Avaliadora
Universidade Federal de Santa Catarina

Prof. Ronaldo dos Santos Mello, Dr.
Avaliador
Universidade Federal de Santa Catarina

Este trabalho é dedicado aos meus pais, que me apoiaram durante todo o processo de formação e ao meu avô, *in memoriam*, que teria ficado muito orgulhoso.

AGRADECIMENTOS

Agradeço ao meu orientador, Prof. Dr. Renato Fileto, que me apresentou ao tema do trabalho e me guiou durante todo o processo de desenvolvimento e de escrita, sempre demonstrando paciência e compreensão ao corrigir meus erros e responder minhas dúvidas.

Agradeço aos amigos que conheci durante a Graduação, pela companhia fornecida não apenas nas aulas ou nos trabalhos em grupo, mas também nos encontros e nas confraternizações.

Também agradeço aos meus amigos Guilherme, Luiz Augusto e Ivo, pelas excelentes conversas que tivemos e pelo apoio emocional nas situações de dificuldade.

“Irracionalidade é a raiz quadrada de todo o mal.”
(HOFSTADTER, 1983)

RESUMO

Um fluxo de trabalho (*workflow*) especifica um processo envolvendo encadeamento de tarefas e fluxo de informações entre elas para alcançar um objetivo de negócio. A orquestração de um *workflow* automatiza e gerencia a execução de uma instância de processo de acordo com a especificação do *workflow*, usando um *software* de orquestração. O orquestrador invoca recursos humanos e/ou de TI apropriados para executar cada tarefa, coordenar a execução das mesmas de acordo com o encadeamento especificado e reportar eventuais problemas encontrados durante a execução. Este trabalho de conclusão de curso está no contexto do projeto de pesquisa CÉOS, executado pela Universidade Federal de Santa Catarina (UFSC), em parceria com o Ministério Público de Santa Catarina (MPSC). O projeto CÉOS tem o objetivo de desenvolver *workflows* para análise e processamento de dados em grande escala a fim de automatizar a extração de conhecimento e aumentar a eficiência das decisões tomadas em atividades do MPSC. Neste trabalho, é especificado e implementado um fluxo de trabalho para extração, transformação e carga (*extract, transform, load* - ETL) de dados referentes a licitações e compras feitas pela Prefeitura Municipal de Florianópolis. O fluxo é orquestrado pelo Airflow de acordo com grafos acíclicos direcionados (*directed acyclic graph* - DAG) especificados na plataforma. Os resultados são avaliados usando métricas referentes à quantidade e qualidade dos dados, a frequência de erros e o tempo de execução dos processos de ETL.

Palavras-chave: Extração, transformação e carga de dados. Dados abertos. Licitações.

ABSTRACT

A workflow specifies a process involving the chaining of tasks and the flow of information between them in order to achieve a business objective. The orchestration of a workflow automates and manages the execution of a process instance according to the workflow specification, using orchestration software. The orchestrator invokes appropriate human and/or IT resources to execute each task, coordinates their execution according to the specified chaining, and reports any issues encountered during execution. This undergraduate thesis is part of the CÉOS research project, carried out by the Federal University of Santa Catarina (UFSC) in partnership with the Public Prosecutor's Office of Santa Catarina (MPSC). The CÉOS project aims to develop workflows for the analysis and processing of large-scale data in order to automate knowledge extraction and improve the efficiency of decision-making processes within MPSC activities. In this work, a workflow is specified and implemented for extract, transform, and load (ETL) processes involving data related to bidding and procurement carried out by the Municipality of Florianópolis. The workflow is orchestrated by Airflow according to directed acyclic graphs (DAGs) specified in the platform. The results are evaluated using metrics related to data quantity and quality, error frequency, and ETL process execution time.

Keywords: Extraction, transformation and loading of data. Open data. Public procurement.

LISTA DE FIGURAS

Figura 1 – Diagrama mostrando os componentes do Airflow.	17
Figura 2 – Exemplo básico de <i>graph view</i> de uma DAG no Airflow.	18
Figura 3 – Visão em árvore do JSON de licitações.	21
Figura 4 – Visão em árvore do JSON de contratos.	26
Figura 5 – Diagrama de classe do esquema.	32
Figura 6 – Visualização das tarefas e dependências da DAG.	35
Figura 7 – Distribuição dos valores das licitações das principais unidades gestoras (1 e 22) no ano de 2024.	40
Figura 8 – Quantidade total de licitações e contratos entre 2014 e 2024.	41
Figura 9 – Variação anual de valores totais de licitações e contratos entre 2014 e 2024.	42
Figura 10 – Valores totais gastos com fornecedores de contratos entre 2014 e 2024. .	43
Figura 11 – Evolução anual de valores de contratos e empenhos vinculados entre 2014 e 2024.	45
Figura 12 – Exemplo de múltiplos registros de contratos com números idênticos. . .	46

LISTA DE TABELAS

Tabela 13 – Comparação de trabalhos relacionados	29
Tabela 14 – Resumo das modalidades de licitações entre 2014 e 2024	38
Tabela 15 – Resumo das modalidades de contratos entre 2014 e 2024	39
Tabela 16 – Resumo das finalidades de licitações entre 2014 e 2024	44

LISTA DE ABREVIATURAS E SIGLAS

API	<i>Application Programming Interface</i>
CSV	<i>Comma-separated Values</i>
DAG	<i>Directed Acyclic Graph</i>
ETL	<i>Extract, Transform, Load</i>
HTTP	<i>Hypertext Transfer Protocol</i>
IA	Inteligência Artificial
JSON	<i>JavaScript Object Notation</i>
LLM	<i>Large Language Model</i>
MPSC	Ministério Público de Santa Catarina
PDF	<i>Portable Document Format</i>
PMF	Prefeitura Municipal de Florianópolis
REST	<i>Representational State Transfer</i>
SQL	<i>Structured Query Language</i>
URL	<i>Uniform Resource Locator</i>
XML	<i>Extensible Markup Language</i>

SUMÁRIO

1	INTRODUÇÃO	13
1.1	OBJETIVOS	14
1.1.1	Objetivo Geral	14
1.1.2	Objetivos Específicos	14
1.2	METODOLOGIA	14
1.3	ORGANIZAÇÃO DO DOCUMENTO	15
2	FUNDAMENTOS	16
2.1	EXTRAÇÃO, TRANSFORMAÇÃO E CARGA (ETL) DE DADOS	16
2.2	APACHE AIRFLOW	16
2.3	TRANSPARÊNCIA E DADOS ABERTOS	18
2.3.1	Licitações e contratos administrativos	19
2.4	DADOS ABERTOS DE COMPRAS PÚBLICAS DA PMF	20
2.4.1	API de compras públicas da PMF	20
2.4.2	Descrição dos dados retornados	21
3	TRABALHOS RELACIONADOS	28
4	DESENVOLVIMENTO	30
4.1	CLIENTE PARA A API DE COMPRAS PÚBLICAS DA PMF	30
4.2	MÉTRICAS DA ANÁLISE EXPLORATÓRIA DOS DADOS COLETADOS EM JSON	31
4.3	ESQUEMA RELACIONAL PARA OS DADOS COLETADOS	32
4.4	PROCESSO DE ETL	33
4.4.1	Extração (<i>extract</i>)	33
4.4.2	Transformação (<i>transform</i>)	34
4.4.3	Carga (<i>load</i>)	34
4.4.4	Orquestração da DAG	35
4.4.5	Dificuldades	35
4.4.6	Considerações e integrações com IA	36
5	RESULTADOS	38
6	CONCLUSÕES E TRABALHOS FUTUROS	48
	REFERÊNCIAS	49
	APÊNDICE A – ARTIGO DO TCC	52
	APÊNDICE B – CÓDIGO FONTE	63

1 INTRODUÇÃO

O uso de plataformas de orquestração de *workflows* é imprescindível para gerenciar aplicações com uso intenso de dados, como *data warehouses*. (Hollingsworth, 1995) define *workflow* como a automação informatizada total ou parcial de um processo de negócio e afirma que um sistema de gerenciamento de *workflows* deve fornecer suporte a três tipos de funções: (i) as funções *build-time*, referentes à definição e modelagem do processo de *workflow*; (ii) as funções *run-time control*, que envolvem o sequenciamento da execução das atividades e (iii) as interações de *run-time* entre os usuários e as aplicações de TI que processam os passos das atividades.

Workflows são frequentemente utilizados para implementar processos de extração, transformação e carga (em inglês *extraction-transformation-loading* - ETL) de dados. (Vassiliadis; Simitsis; Skiadopoulos, 2002) fornecem uma descrição generalizada de processos ETL. Dados são obtidos de fontes como bancos de dados relacionais ou arquivos por rotinas de extração. Depois, os dados são propagados a uma ou mais *data staging areas*, onde são transformados a um formato homogêneo. Por fim, os dados formatados são inseridos em um banco de dados, muitas vezes no modelo multidimensional de um *data warehouse*.

O Apache Airflow é uma plataforma de orquestração de *workflows* de código aberto utilizada para implementar e manter *pipelines* de dados. No Airflow, *workflows* são definidos por grafos acíclicos direcionados, ou DAGs (do inglês *directed acyclic graph*). DAGs são *scripts* escritos na linguagem Python, utilizando construções básicas como operadores. DAGs definem as tarefas a serem executadas, as quais podem ser especificadas por operadores, e relações que indicam a ordem de execução e as dependências entre tarefas. Estes componentes formarão a *pipeline* do Airflow (Finnigan; Toner, 2021).

Este trabalho busca descrever o desenvolvimento de um fluxo de ETL e apresentar os resultados de uma análise exploratória feita sobre os dados obtidos neste fluxo. Primeiramente, são criados dois *scripts* em Python: (i) um *script* que faz uma requisição da API de dados públicos do Portal de Transparência da Prefeitura Municipal de Florianópolis buscando registros de contratos e licitações, incluindo número, nome e CPF/CNPJ do fornecedor, período de vigência e valor monetário, e (ii) um *script* que faz a requisição da API utilizada para mostrar os dados nas páginas individuais das licitações e contratos no *website* do Portal. Depois, os *scripts* são convertidos em tarefas para o DAG do Airflow utilizando o TaskFlow API. Finalmente, os dados obtidos de cada *script* são agregados pelos respectivos códigos de identificação e servem como base para a inserção de registros em um banco de dados PostgreSQL.

A análise exploratória e a visualização dos dados são feitas com a geração de tabelas e gráficos, que permitem analisar a estrutura e a distribuição dos dados, tais como os custos dos contratos para identificar irregularidades. Verifica-se, entre outros fatores, o número de registros de licitações e contratos por ano, a variação dos valores

totais negociados por ano, a distribuição de valores estimados de licitações por unidade gestora, a diferença entre os valores de um contrato e os empenhos vinculados a ele e os principais fornecedores de contratos. O desempenho do fluxo de ETL é avaliado com base nos tempos de execução das tarefas da DAG e na quantidade de dados tratados.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é orquestrar um fluxo de extração, transformação e carregamento de dados relacionados a compras feitas por órgãos públicos do município de Florianópolis, com o intuito de suprir dados para analisar os gastos com as compras no projeto CÉOS. Os resultados são avaliados através de uma análise exploratória de dados resultantes em tabelas e gráficos, que permitem visualizar suas distribuições. Também são discutidas possíveis irregularidades encontradas nos registros, tratados ou não dentro dos processos de ETL implementados e avaliados.

1.1.2 Objetivos Específicos

Os objetivos específicos são:

1. Estudar conceitos de orquestração de fluxos de ETL.
2. Desenvolver códigos para obtenção de dados utilizando métodos de web scraping e requisições em uma API pública.
3. Construir DAGs para implementar o fluxo de ETL, que padronizará os dados em um formato adequado.
4. Orquestrar processos de ETL usando as DAGs construídas, coletando métricas para avaliar sua capacidade, desempenho e custos computacionais.
5. Utilizar ferramentas de análise e visualização de dados para gerar tabelas e gráficos sobre as distribuições dos dados carregados e a análise dos custos computacionais dos processos de ETL.

1.2 METODOLOGIA

A metodologia adotada neste trabalho é estruturada em três etapas principais, contemplando desde o estudo preliminar até a implementação e análise dos dados extraídos.

A primeira etapa consiste em um levantamento teórico sobre os conceitos, ferramentas e práticas envolvidas na orquestração de *pipelines* de ETL (*extraction-transformation-loading*) utilizando o Apache Airflow. Foram reunidos artigos científicos, publicações técnicas e documentação especializada sobre a estrutura e implementação de DAGs no Airflow,

bem como sobre métodos de *web scraping*, a fim de possibilitar a extração de dados não estruturados diretamente de páginas da Web.

Na segunda etapa, foi realizada a implementação dos *scripts* responsáveis pela extração de dados por meio de *web scraping* e requisições a APIs REST. Os dados obtidos a partir dos *endpoints* públicos disponibilizados no Portal de Transparência da Prefeitura Municipal de Florianópolis estavam incompletos, o que tornou necessário consultar os *endpoints* internos do sistema para extrair os dados faltantes. Após este processo, realizou-se uma análise exploratória dos dados, consistindo na geração de gráficos para visualização de possíveis *outliers* entre os registros e na construção de um dicionário de dados e, subsequentemente, a modelagem de um esquema para bancos de dados relacionais.

A terceira e última etapa envolve a orquestração completa do fluxo ETL no ambiente do Airflow, seguida da análise dos dados carregados. Foram realizados testes de execução e ajustes nos fluxos com base na avaliação de métricas, como tempo de execução e ocorrência de erros, e a partir dos dados transformados, foram geradas tabelas e gráficos mais detalhados, permitindo a análise mais complexa.

1.3 ORGANIZAÇÃO DO DOCUMENTO

O trabalho está organizado da seguinte maneira: o Capítulo 2 apresenta os fundamentos de conceitos e tecnologias utilizados para desenvolver o trabalho. No Capítulo 3, faz-se uma comparação de trabalhos que abordam temas semelhantes a este. O Capítulo 4 descreve os processos de análise exploratória dos dados e implementação do fluxo ETL. No Capítulo 5, são discutidos os resultados obtidos a partir da análise. Finalmente, o Capítulo 6 apresenta conclusões e temas para trabalhos futuros.

2 FUNDAMENTOS

Neste capítulo são apresentados conceitos teóricos relacionados ao trabalho. Na Seção 2.1, apresenta-se o processo de extração, transformação e carga de dados com as definições de cada etapa do processo. A Seção 2.2 apresenta a plataforma de software Apache Airflow, utilizada para orquestrar fluxos de trabalho, além de seus componentes e sua interface gráfica. A Seção 2.3 apresenta o conceito de dados abertos e iniciativas de transparência governamental em nível federal e municipal; na Subseção 2.3.1, são definidos licitações e contratos administrativos de acordo com a Lei 14.133/2021. Por fim, a Seção 2.4 apresenta a API de compras públicas da Prefeitura Municipal de Florianópolis e define um dicionário de dados a partir dos resultados de consultas feitas.

2.1 EXTRAÇÃO, TRANSFORMAÇÃO E CARGA (ETL) DE DADOS

Processos de extração, transformação e carga de dados são responsáveis pela coleta, extração, limpeza, transformação integração e carga de dados de uma ou mais fontes para um sistema destino, tal como um *data warehouse*. De acordo com (Vassiliadis; Simitsis, 2009), qualquer programa que filtra registros, calcula novos valores e alimenta uma fonte de dados diferente da original pode ser considerado um programa de ETL.

(Vassiliadis; Simitsis, 2009) descrevem o fluxo de trabalho de um processo ETL como um grafo acíclico direcionado, no qual as tarefas executadas são análogas aos nodos do grafo e as relações entre entradas e saídas são análogas às arestas do grafo.

A primeira etapa do processo ETL é a extração. Nesta, dados são obtidos de fontes, que podem estar no formato de documentos de texto, planilhas, páginas da Web ou *streaming*. Geralmente, busca-se extrair apenas os dados que foram inseridos ou atualizados após a última execução do processo.

A segunda etapa é a transformação. Esta etapa envolve a normalização, formatação, padronização e enriquecimento de dados visando resolver problemas em nível de esquema (conflitos de nomeação de objetos ou diferenças na estruturação de dados entre fontes e o *data warehouse*) e em nível de registro (registros duplicados, contraditórios, ou com referências temporais diferentes).

A terceira e última etapa é a carga. Os dados transformados são carregados nas tabelas apropriadas do *data warehouse*, onde servirão como base para processos de análise, visualização e apoio a tomadas de decisões.

2.2 APACHE AIRFLOW

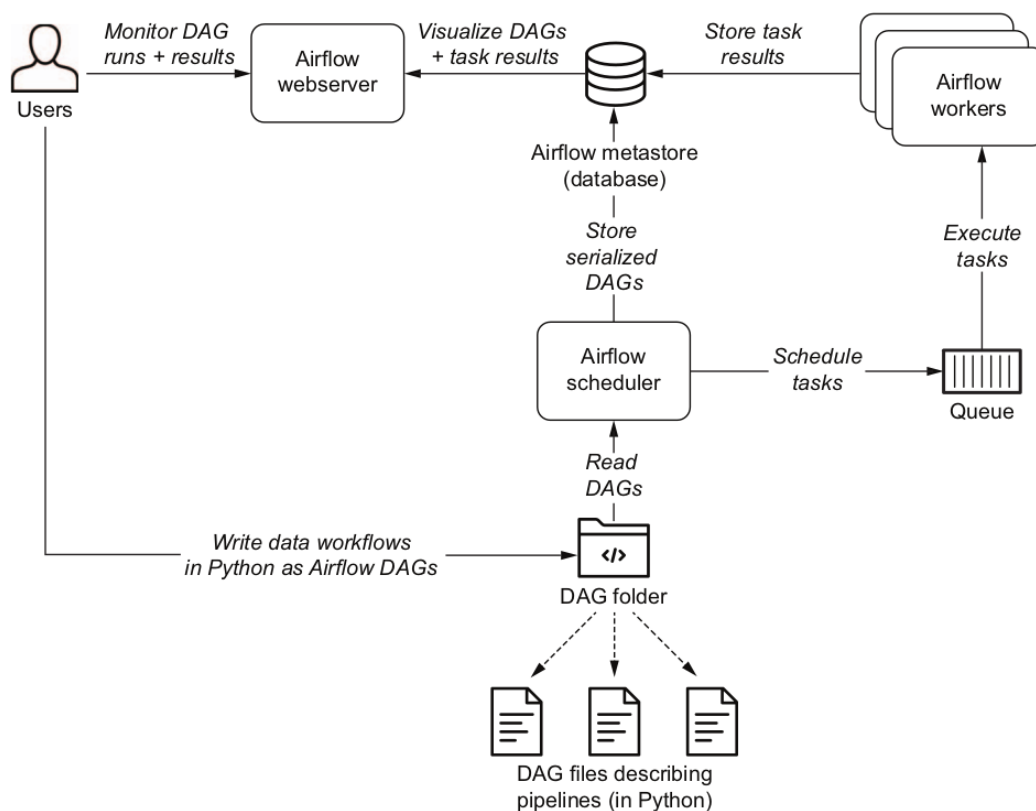
O Apache Airflow é uma plataforma de código aberto utilizada para orquestrar e monitorar fluxos de trabalho no formato de grafos acíclicos direcionados, ou DAGs. No Airflow, os DAGs são definidos por meio de arquivos, ou *scripts* na linguagem Python

que descrevem as tarefas a serem executadas e as dependências entre as tarefas, além de metadados com configurações específicas. O Airflow interpreta estes *scripts* e modela a estrutura das DAGs, executando as tarefas de acordo com o intervalo de tempo agendado nos metadados.

De acordo com (Ruiter; Harenslak, 2021), o uso de código em Python para definir DAGs garante flexibilidade ao desenvolvedor, pois qualquer função implementada em Python pode ser executada pelo Airflow. Assim, é possível importar funções de bibliotecas externas e criar fluxos de trabalho altamente escaláveis que estabelecem comunicações entre diversos serviços.

(Ruiter; Harenslak, 2021) afirmam que o Airflow pode ser dividido em três componentes: (i) o escalonador, que lê os *scripts* de DAGs desenvolvidos e obtém as tarefas, dependências e intervalos de execução, verifica se o intervalo de cada DAG passou e, se este for o caso, adiciona as respectivas tarefas à fila de execução, (ii) os trabalhadores, que obtêm tarefas escalonadas da fila e as executam em paralelo e (iii) o servidor *web*, que fornece uma interface com representações visuais das DAGs e permite que o usuário monitore os resultados das tarefas executadas.

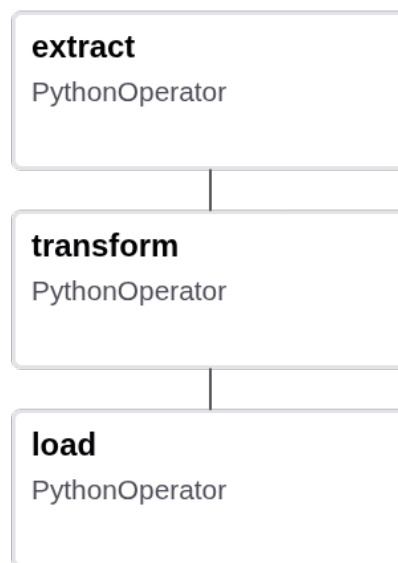
Figura 1 – Diagrama mostrando os componentes do Airflow.



Fonte: (Ruiter; Harenslak, 2021)

A interface *web* do Airflow oferece diferentes modos de visualização das DAGs. O modo *graph view* (visualização de grafo), exemplificado pela Figura 1, mostra um esquema contendo as tarefas e as respectivas dependências, enquanto o modo *tree view* (visualização de árvore) mostra o histórico dos resultados das execuções das DAGs, também permitindo verificar detalhes de tarefas específicas, com indicadores de status e *logs* de textos com informações referentes a possíveis erros de execução.

Figura 2 – Exemplo básico de *graph view* de uma DAG no Airflow.



Fonte: Elaborada pelo autor.

2.3 TRANSPARÊNCIA E DADOS ABERTOS

(Possamai; Souza, 2020) afirmam que dados abertos governamentais são dados públicos produzidos ou curados por órgãos estatais, padronizados em formatos abertos e legíveis por máquina, como JSON ou XML, e que podem ser livremente acessados e utilizados para quaisquer finalidades. A disponibilidade de dados abertos governamentais é importante para garantir a transparência com gastos públicos, incentivar a colaboração entre o cidadão e a Administração Pública e combater a corrupção.

Segundo (Possamai; Souza, 2020), a Lei 12.527/2011 regula o acesso às informações contidas em documentos de órgãos públicos e estabelece duas formas de transparência a serem implementadas: (i) a transparência passiva, que envolve o atendimento a solicitações de informação por meio físico ou eletrônico, e (ii) a transferência ativa, referente à divulgação de um rol mínimo de informações por parte dos órgãos independentemente de requerimento. A Lei 12.527/2011 faz referência explícita aos princípios dos dados abertos em seu texto.

De acordo com (Nazário; Silva; Rover, 2012), um dos principais projetos de visualização de dados abertos governamentais é o Portal de Transparência do Governo Federal, inicialmente lançado em novembro de 2004. Em concordância com a Lei Complementar 131/09, o Portal disponibiliza informações referentes a receitas, transferências e convênios recebidos, assim como despesas envolvendo processos licitatórios, contratos e gastos com diárias.

Em relação à cidade de Florianópolis, (Santos, M. T. S. et al., 2021) afirmam que a Prefeitura Municipal regulamentou o acesso à informação pública por meio do decreto 9988/12, posteriormente instituindo o Portal de Transparência de Florianópolis por meio da Lei Municipal 9.447/14. Analogamente ao Portal de Transparência do Governo Federal, o usuário possui acesso a receitas e despesas da Prefeitura e de seus órgãos.

2.3.1 Licitações e contratos administrativos

(TCU, 2023) define licitação como o processo por meio do qual a Administração Pública convoca, sob condições estabelecidas em ato próprio (edital de licitação), interessados para apresentação de propostas relativas ao fornecimento de bens, prestação de serviços ou execução de obras. O processo licitatório é regulamentado pela Lei 14.133/2021, cujo artigo 17 o divide em sete fases: (i) fase preparatória, (ii) divulgação do edital, (iii) apresentação de propostas e lances, (iv) julgamento, (v) habilitação, (vi) recursal e (vii) homologação.

O artigo 11 da Lei 14.133/2021 aponta como objetivos (i) garantir que seja selecionada a proposta apta a gerar o resultado de contratação mais vantajoso à Administração Pública, (ii) assegurar o tratamento isonômico e a justa competição entre os licitantes, (iii) evitar sobrepreço nas contratações e superfaturamento na execução dos contratos e (iv) incentivar a inovação e a sustentabilidade no desenvolvimento nacional.

Segundo a Lei 14.133/2021, existem cinco modalidades de licitação:

1. Pregão: aquisição de bens e serviços comuns definidos pelo edital, seguindo critérios de julgamento de menor preço ou maior desconto.
2. Concorrência: aquisição de bens e serviços especiais e de obras e serviços comuns e especiais de engenharia, utilizando critérios de julgamento de menor preço, maior desconto, melhor técnica ou conteúdo artístico, técnica e preço ou maior retorno econômico.
3. Concurso: escolha de trabalho técnico, científico ou artístico, utilizando o critério de julgamento de melhor técnica ou conteúdo artístico e concedendo prêmio ou remuneração ao vencedor.
4. Leilão: alienação de bens imóveis ou de bens móveis inservíveis ou legalmente apreendidos, seguindo o critério de julgamento de maior lance.

5. Diálogo competitivo: contratação de obras, serviços e compras por meio de diálogos entre a Administração Pública e licitantes previamente selecionados, visando uma solução que atenda a uma necessidade definida.

De acordo com (TCU, 2023), contratos administrativos são aqueles firmados entre órgãos ou entidades da Administração Pública e particulares, por meio do qual se estabelece acordo de vontades, para a formação de vínculo e a estipulação de obrigações recíprocas, que têm como objetivo principal atender a um interesse coletivo. Esses contratos se distinguem dos contratos regidos pelo direito privado por apresentarem cláusulas exorbitantes, isto é, prerrogativas conferidas à Administração Pública que visam assegurar a supremacia do interesse público sobre o privado. Entre essas prerrogativas estão, por exemplo, a possibilidade de alteração unilateral do contrato, a rescisão administrativa e a aplicação de sanções ao contratado, sempre observados os princípios da legalidade e da proporcionalidade.

Conforme dispõe o artigo 89 da Lei no. 14.133/2021, os contratos administrativos devem ser formalizados por escrito e conter cláusulas essenciais que estabeleçam as condições de execução, prazos, direitos, responsabilidades e garantias das partes. Assim, os contratos administrativos são instrumentos de execução de políticas públicas, por meio dos quais a Administração busca assegurar a prestação adequada de serviços, obras e fornecimentos de bens à sociedade, sob um regime jurídico pautado pela legalidade, impessoalidade e eficiência (TCU, 2023).

2.4 DADOS ABERTOS DE COMPRAS PÚBLICAS DA PMF

Esta seção é dividida em duas subseções. A Subseção 2.4.1 descreve a API de compras públicas disponível no *website* do Portal de Transparência da Prefeitura Municipal de Florianópolis, o método de consulta de dados e os parâmetros de pesquisa. A Subseção 2.4.2 apresenta um dicionário de dados, no qual os conteúdos das respostas da API são esquematizados em tabelas.

2.4.1 API de compras públicas da PMF

O Portal de Transparência da Prefeitura Municipal de Florianópolis disponibiliza informações relacionadas a compras públicas efetuadas pela Prefeitura e as unidades gestoras que a compõem, em concordância com o princípio de dados abertos. O *website* do Portal, além de mostrar listas com registros de compras, oferece uma API REST que permite ao usuário fazer requisições destes registros utilizando o formato JSON.

A API possui diversos *endpoints*, dos quais são utilizados os de consulta de licitações¹ e de contratos administrativos². O acesso a estes *endpoints* é livre, sem necessidade

¹ Disponível em: <https://transparencia.e-publica.net/epublica-portal/#/florianopolis/portal/dadosAbertos/licitacaoView>

² Disponível em: <https://transparencia.e-publica.net/epublica-portal/#/florianopolis/>

de chave para autenticação.

O único método fornecido para ambos os *endpoints* é o GET, que retorna dados de licitações ou contratos em formato JSON. Em ambos os *endpoints*, é possível filtrar os resultados pelas datas inicial e final do período de busca, o código da unidade gestora responsável e o limite máximo de registros a serem retornados. Apenas as datas inicial e final são parâmetros obrigatórios.

2.4.2 Descrição dos dados retornados

A estrutura dos registros retornados difere de acordo com os *endpoints*. A Figura 3 apresenta a estrutura dos registros retornados pelo *endpoint* Licitação. Estes registros contêm dados como o valor da licitação, datas de emissão e abertura, modalidade do processo licitatório, advogado responsável, lista de itens negociados e vencedores do processo, lista de *links* a documentos relacionados, lista de empenhos e lista de contratos.

A Figura 4 mostra a estrutura dos registros retornados pelo *endpoint* Contrato. Nestes registros, encontra-se dados como as datas de assinatura e início de vigência, valor total do contrato, nome do fornecedor, lista de itens e lista de empenhos.

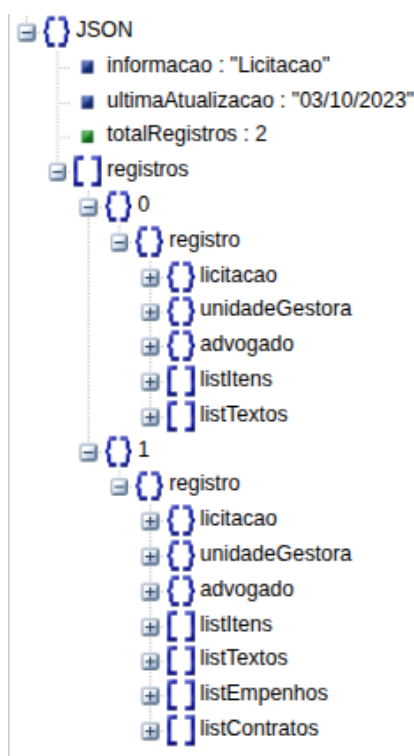


Figura 3 – Visão em árvore do JSON de licitações.

O dicionário de dados a seguir foi desenvolvido a partir do conteúdo retornado pelo método GET da API. Além do cabeçalho do arquivo JSON, que contém metadados, são retornados elementos de dados JSON para as entidades **Licitação** e **Contrato**

administrativo, bem como as entidades auxiliares, que representam objetos aninhados nos registros.

O *website* do Portal de Transparência da Prefeitura Municipal de Florianópolis também disponibiliza *endpoints* para a consulta de receitas, despesas orçamentárias, execuções de despesa e dados da gestão de pessoal; estes não foram utilizados no trabalho, devido aos dados possuírem pouca relação ao contexto dos processos licitatórios.

Os arquivos JSON seguem padrões semelhantes; ambos os *endpoints* consultados no trabalho retornam arquivos com as seguintes propriedades:

Cabeçalho do JSON

Nome do Campo	Tipo de Dado	Tamanho	Descrição
informacao	Texto	50	Tipo de informação retornada pela API ('Licitacao', 'Contrato', 'Receita', 'Despesa Orçamentária', 'Execução de Despesa').
ultimaAtualizacao	Data	-	Data da última atualização dos dados no formato "dd/MM/yyyy".
totalRegistros	Inteiro	-	Número total de registros retornados na consulta.
registros	Lista	-	Lista de registros (licitações ou contratos).

Licitação

Nome do Campo	Tipo de Dado	Tamanho	Descrição
numero	Texto	15	Número identificador da licitação.
modalidade	Texto	50	Modalidade da licitação (exemplo: 'Pregão', 'Tomada de Preço').
valorEstimado	Decimal	10,2	Valor estimado da licitação.
objetoResumido	Texto	500	Descrição resumida do objeto da licitação.
dataEmissao	Data	-	Data de emissão da licitação no formato "yyyy-MM-dd".
aberturaData	Data	-	Data de abertura da licitação.
finalidade	Texto	100	Finalidade da licitação (exemplo: 'Compras e Outros Serviços', 'Obras e Serviços de Engenharia').
formaJulgamento	Texto	50	Forma de julgamento da licitação (exemplo: 'Por item').
unidadeGestora	Objeto	-	Órgão responsável pela licitação.
advogado	Objeto	-	Representante jurídico vinculado ao processo licitatório.
listItens	Objeto	-	Lista de itens licitados.
listTextos	Objeto	-	Lista de documentos e arquivos complementares.
listEmpenhos (opcional)	Objeto	-	Lista de empenhos associados à licitação.
listContratos (opcional)	Objeto	-	Lista de contratos associados à licitação.

Unidade gestora

Nome do Campo	Tipo de Dado	Tamanho	Descrição
codigo	Inteiro	-	Código identificador da unidade gestora.
denominacao	Texto	100	Nome da unidade gestora responsável pela licitação ou contrato.

Advogado

Nome do Campo	Tipo de Dado	Tamanho	Descrição
pessoa.nome	Texto	100	Nome do advogado responsável pelo processo licitatório.

Item da licitação

Nome do Campo	Tipo de Dado	Tamanho	Descrição
numero	Texto	20	Número do item.
denominacao	Texto	100	Descrição do item.
quantidade	Decimal	-	Quantidade de unidades do item.
unidadeMedida	Texto	20	Unidade de medida do item (exemplo: 'Serviço' ou 'Unidade').
valorUnitarioEstimado	Decimal	10,2	Valor monetário de uma unidade do item.
situacao	Texto	20	Situação do item (exemplo: 'Homologado').
listVencedores	Objeto	-	Lista de vencedores do processo licitatório.

Vencedor

Nome do Campo	Tipo de Dado	Tamanho	Descrição
fornecedor	Texto	100	Nome da pessoa jurídica fornecedora do item.
quantidade	Decimal	-	Quantidade de unidades do item fornecido.
valorUnitario	Decimal	10,2	Valor monetário de uma unidade do item.
situacao	Texto	20	Situação da pessoa jurídica (exemplo: 'Vencedor').

Texto

Nome do Campo	Tipo de Dado	Tamanho	Descrição
denominacao	Texto	200	Título do documento, contendo números de identificação.
link	Texto	100	Endereço URL da página do documento.

Empenho

Nome do Campo	Tipo de Dado	Tamanho	Descrição
emissao	Data	-	Data de emissão do empenho.
numero	Texto	20	Número do empenho.
objetoResumido	Texto	255	Descrição resumida do empenho.
especie	Texto	20	Tipo do empenho (exemplo: 'Ordinário', 'Estimativo' ou 'Global').
categoria	Texto	20	Categoria do empenho (exemplo: 'Comum', 'Subvenção', 'Auxílio' etc.).
contrato	Texto	20	Número do contrato relacionado ao empenho.
licitacao	Texto	20	Número da licitação relacionada ao empenho.
recursoDiaria	Texto	20	Informação sobre recurso de diária (se aplicável).



Figura 4 – Visão em árvore do JSON de contratos.

No arquivo dos contratos, como mostra a Figura 4, o registro contém as seguintes propriedades:

Contrato

Nome do Campo	Tipo de Dado	Tamanho	Descrição
numero	Texto	20	Número do contrato.
assinatura	Data	-	Data da assinatura do contrato.
inicioVigencia	Data	-	Data de início da vigência do contrato.
vencimento	Data	-	Data de vencimento do contrato.
valorTotal	Decimal	10,2	Valor total do contrato.
objetoResumido	Texto	255	Descrição resumida do contrato.
unidadeGestora	Objeto	-	Órgão responsável pelo contrato.
fornecedor	Objeto	-	Fornecedor contratado.
listItens	Objeto	-	Lista de itens do contrato.
listEmpenhos (opcional)	Objeto	-	Lista de empenhos associados ao contrato.

Item do contrato

Nome do Campo	Tipo de Dado	Tamanho	Descrição
numero	Texto	20	Número do item.
denominacao	Texto	100	Descrição do item.
quantidade	Decimal	-	Quantidade de unidades do item.
unidadeMedida	Texto	20	Unidade de medida do item (exemplo: 'Serviço' ou 'Unidade').
valorUnitario	Texto	10	Valor monetário de uma unidade do item.
valorTotal	Texto	10	Valor monetário de todas as unidades do item.

Fornecedor

Nome do Campo	Tipo de Dado	Tamanho	Descrição
pessoa.nome	Texto	100	Nome do fornecedor contratado.

3 TRABALHOS RELACIONADOS

A análise de dados abertos governamentais é um tema cada vez mais pertinente na pesquisa acadêmica, especialmente considerando a disponibilidade de algoritmos de aprendizado de máquina e a capacidade de processar grandes quantidades de dados. Nesta seção, são apresentados trabalhos que envolvem a análise de dados de licitações e contratos administrativos e buscam facilitar o combate à corrupção nos processos licitatórios.

A busca bibliográfica foi feita pela plataforma Google Scholar utilizando as palavras-chave "ETL", "extração transformação e carga", "licitações", "portal de transparência" e "dados abertos". Também foram obtidos trabalhos a partir dos arquivos do Simpósio Brasileiro de Bancos de Dados de 2024 (SBBBD).

(Schmitz et al., 2024) propõem uma metodologia de aprendizado de máquina que utiliza modelos de mistura gaussiana (GMM) para identificar padrões de lances suspeitos de fraude em processos licitatórios, considerando que a escassez de casos confirmados de fraude torna o uso de métodos de aprendizado supervisionado inviável. O modelo avalia a similaridade entre casos de licitações possivelmente fraudulentas e casos conhecidos de fraude em diversos subespaços, criando um *ranking* eficaz de indicadores de risco.

(Santos, E. S. dos et al., 2024) utilizam um conjunto de dados referentes a licitações investigadas pela Operação Lava Jato para avaliar as capacidades de diferentes modelos de aprendizado de máquina em detectar casos de conluio. O trabalho propõe uma metodologia com extração de dados para enriquecimento do conjunto original e técnicas como validação cruzada e otimização de hiperparâmetros para treinar os modelos, alcançando resultados mais precisos.

(Santos, M. T. S. et al., 2021) apresentam uma ferramenta para visualizar dados obtidos do Portal de Transparência da Câmara Municipal de Florianópolis; o trabalho foca especificamente nos balancetes de vereadores. Os dados dos documentos em PDF são extraídos e coletados em arquivos CSV, que são lidos por uma aplicação escrita na linguagem PHP e adicionados a um banco de dados. Uma interface gráfica permite pesquisar o nome de um vereador e um ano, gerando gráficos dos gastos mensais durante este ano e comparando com gastos de outros anos.

(Jesus, 2021) apresenta um modelo preditivo que utiliza técnicas de mineração de dados e aprendizado de máquina para indicar riscos de irregularidades nas fases de divulgação do edital e apresentação de propostas do processo licitatório. O modelo analisa licitações do Estado de Goiás realizadas entre 2014 e 2019 e calcula o risco utilizando algoritmos de treinamento supervisionado com classificadores. Por fim, constrói-se um *ranking* de licitações potencialmente irregulares.

(Mello et al., 2024) propõem uma modelagem conceitual abrangente de dados para o domínio de licitações, também incluindo dados de fraudes associadas a licitações, denúncias e pessoas envolvidas no processo licitatório, com o objetivo de tornar o projeto

de um banco de dados mais robusto e fornecer apoio a sistemas de detecção de fraude. A modelagem inclui uma proposta de persistência poliglota, ou seja, o uso de diversos modelos de banco de dados, como relacional, orientado a grafos e orientado a documentos.

A Tabela 13 apresenta uma comparação entre as metodologias adotadas em cada trabalho e os conjuntos de dados utilizados.

Tabela 13 – Comparação de trabalhos relacionados

Trabalho	Metodologia	Conjunto de dados
(Schmitz et al., 2024)	Aprendizado de máquina com modelos de mistura gaussiana	Licitações referentes à Operação Patrola
(Santos, E. S. dos et al., 2024)	ETL com validação cruzada e otimização de hiperparâmetros	Licitações referentes à Operação Lava Jato
(Santos, M. T. S. et al., 2021)	Extração de arquivos CSV e interface gráfica	Balancetes de vereadores de Florianópolis
(Jesus, 2021)	Mineração de dados e treinamento supervisionado	Licitações do Estado de Goiás entre 2014 e 2019
(Mello et al., 2024)	Modelagem conceitual com persistência poliglota	-

Fonte: Elaborada pelo autor.

Um diferencial importante deste trabalho está na integração entre a extração automatizada de dados, a análise exploratória e a avaliação da consistência entre licitações, contratos e empenhos, utilizando fontes públicas e *endpoints* internos não documentados da Prefeitura Municipal de Florianópolis. Enquanto os estudos relacionados concentram-se principalmente em modelagem preditiva, identificação de fraudes ou visualização de informações já estruturadas, este trabalho aborda métodos de obter, integrar e validar dados administrativos municipais de forma confiável, preparando-os para análises posteriores. Ao propor um processo de ETL capaz de lidar com limitações das APIs públicas, inconsistências de registros e diferentes estruturas de dados, o trabalho contribui tanto para a melhoria da qualidade da informação governamental quanto para o desenvolvimento de *pipelines* que podem servir de base para aplicações futuras de auditoria, transparência e detecção de irregularidades.

4 DESENVOLVIMENTO

Este capítulo descreve o que foi desenvolvido no trabalho de conclusão de curso, incluindo codificação do cliente da API, análise exploratória de dados coletados, definição do esquema do banco de dados destino implementação do processo de ETL. Todo o código, dados e resultados obtidos neste trabalho estão disponíveis via GitHub¹.

4.1 CLIENTE PARA A API DE COMPRAS PÚBLICAS DA PMF

Ao consultar a API da prefeitura por meio dos *endpoints* descritos na Seção 2, observou-se que os registros retornados apresentavam informações incompletas. Especificamente, as licitações não possuíam contratos vinculados, e tanto licitações quanto contratos não exibiam os empenhos associados. Este problema indicou que os *endpoints* públicos disponibilizados pela API forneciam apenas dados resumidos, insuficientes para a análise completa pretendida.

Diante dessa limitação, foi necessário adotar uma estratégia alternativa para a extração das informações. O novo método baseou-se no uso dos *endpoints* internos do próprio portal de transparência, responsáveis por alimentar as tabelas exibidas na interface Web de licitações² e contratos³. Embora não documentados oficialmente, esses *endpoints* fornecem os dados completos que são utilizados pela aplicação para gerar as páginas detalhadas de cada licitação ou contrato.

O processo iniciou-se com o envio de requisições HTTP do tipo POST para o *endpoint* de listagem de licitações, com o objetivo de obter os identificadores internos (*IDs*) usados pelo sistema para acessar os registros individuais. A ferramenta Postman foi utilizada para testar e ajustar as requisições, variando parâmetros como intervalo de tempo e quantidade de registros retornados, e para exportar os resultados em arquivos no formato JSON, que serviram como base para as etapas seguintes. O mesmo procedimento foi aplicado ao *endpoint* responsável pela listagem dos contratos.

Em seguida, foi desenvolvido um *script* em Python para automatizar a extração dos dados completos. O programa utiliza os *IDs* coletados nas etapas anteriores como parâmetros em novas requisições POST, enviadas aos *endpoints* que retornam as informações detalhadas das licitações⁴ e dos contratos⁵. Essa abordagem possibilitou coletar não apenas os dados principais de cada licitação, mas também os contratos e os empenhos a

¹ Disponível em: <https://github.com/gustavodesalles/INE5454-trabalho>

² Disponível em: <https://transparencia.e-publica.net/epublica-portal/rest/florianopolis/compras/licitacao/listAll>

³ Disponível em: <https://transparencia.e-publica.net/epublica-portal/rest/florianopolis/compras/contrato/listAll>

⁴ Disponível em: <https://transparencia.e-publica.net/epublica-portal/rest/florianopolis/compras/licitacao/form>

⁵ Disponível em: <https://transparencia.e-publica.net/epublica-portal/rest/florianopolis/compras/contrato/form>

ela vinculados.

Todos os dados extraídos foram organizados e armazenados em arquivos JSON, que estão disponíveis publicamente no repositório do projeto no GitHub⁶.

4.2 MÉTRICAS DA ANÁLISE EXPLORATÓRIA DOS DADOS COLETADOS EM JSON

Esta etapa visa descrever as métricas utilizadas na análise exploratória do conjunto de dados descritos anteriormente com o objetivo de compreender a distribuição dos dados, identificar padrões, avaliar sua consistência e detectar possíveis inconsistências ou lacunas. Para fazer isso, criou-se *scripts* na linguagem Python, utilizando as bibliotecas Pandas e Matplotlib para estruturar os dados e gerar gráficos com base nos atributos dos registros.

Foi analisada a distribuição temporal das datas de assinatura de licitações e contratos ao longo do período abrangido pelos dados coletados. Foram gerados gráficos de séries temporais que representam a quantidade de licitações e contratos registrados por dia. Essa análise permite observar padrões sazonais, concentrações em determinados períodos e eventuais flutuações no volume de registros.

Com o intuito de compreender a composição dos dados, foram geradas distribuições das unidades gestoras vinculadas às licitações presentes no conjunto. Além disso, foi realizada uma análise estatística dos valores estimados das licitações, utilizando gráficos do tipo *boxplot* para identificar a dispersão, a mediana e possíveis valores discrepantes (*outliers*). Esta análise é fundamental para entender a variabilidade dos valores envolvidos nas compras públicas, além de fornecer indícios sobre padrões ou anomalias financeiras.

Outra etapa da análise consistiu na verificação da consistência dos dados financeiros entre os valores das licitações, contratos e respectivos empenhos. Foram comparados os valores dos contratos com os somatórios dos empenhos vinculados, a fim de identificar eventuais divergências e avaliar se os dados refletem corretamente os compromissos financeiros firmados pela administração pública.

Também foi realizada uma análise de integridade e qualidade dos dados, com o levantamento de campos nulos, ausentes ou inconsistentes. Foram identificados registros com ausência de informações relevantes, como dados de fornecedores, valores estimados, datas de assinatura dos contratos ou unidades gestoras. A identificação dessas irregularidades nos dados é fundamental tanto para a correta interpretação dos resultados quanto para o desenvolvimento de eventuais etapas de pré-processamento e limpeza dos dados.

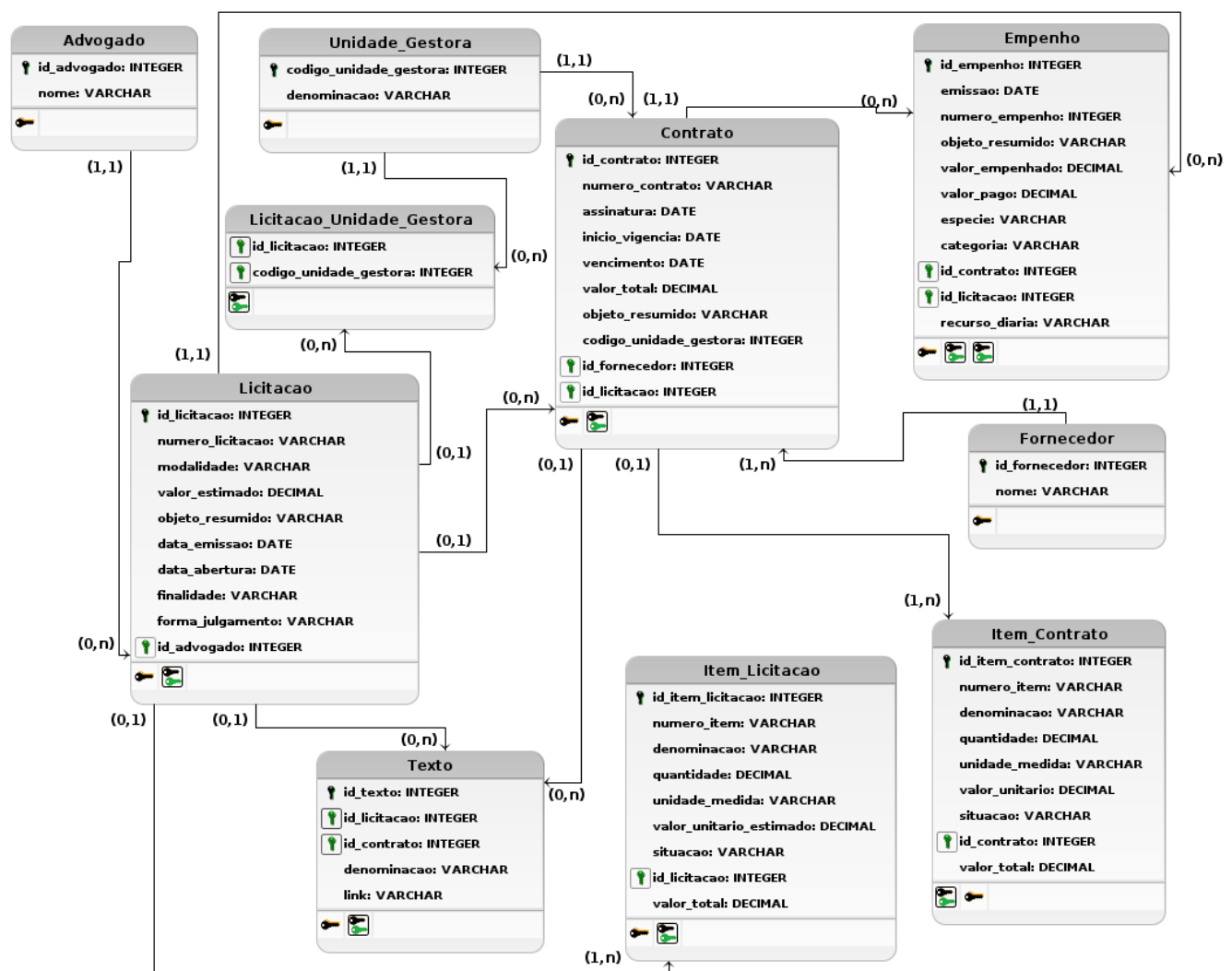
Com os resultados desta análise exploratória, obtém-se uma visão geral do comportamento dos dados e identifica-se limitações a serem consideradas nas etapas subsequentes do trabalho. Além disso, possíveis problemas de completude ou inconsistência nos dados podem indicar falhas nos sistemas de origem ou na própria disponibilização pública das informações.

⁶ <https://github.com/gustavodesalles/licitacoes-contratos-fln>

4.3 ESQUEMA RELACIONAL PARA OS DADOS COLETADOS

Nesta seção, é apresentado um esquema de dados modelado com base na API de compras públicas da Prefeitura Municipal de Florianópolis e no dicionário de dados descritos na Seção 2.4. O esquema visa representar a estrutura lógica dos dados de maneira clara e concisa para fins de análise. Um diagrama de classe é apresentado para mostrar a definição das tabelas em formato relacional e suas respectivas conexões.

Figura 5 – Diagrama de classe do esquema.



Fonte: Elaborada pelo autor.

Nota-se o uso de *surrogate keys* em todas as tabelas que possuem atributos denominados "id_<nome da tabela>", para facilitar a relação entre entidades e lidar com a possibilidade de uma licitação ou contrato ter sido registrado com um número irregular ou idêntico ao de outro registro no banco de dados.

A maioria das entidades corresponde às tabelas definidas no dicionário de dados: as entidades `Item_Licitacao` e `Item_Contrato` foram separadas, e no esquema de banco de dados do CÉOS, são análogas a `Lote` e `Cotacao`, respectivamente. Além disso, foi criada uma tabela para representar relações do tipo n:m (muitos-para-muitos); a tabela `Licitacao_Unidade_Gestora` modela a relação entre `Licitacao` e `Unidade_Gestora`, pois uma licitação pode ser vinculada a mais de uma unidade gestora.

4.4 PROCESSO DE ETL

Esta seção apresenta o planejamento e implementação do processo de extração, transformação e carga dos dados obtidos nas seções anteriores⁷. O Apache Airflow foi utilizado como ferramenta de orquestração da DAG para definir a sequência da execução de tarefas, garantindo reprodutibilidade, escalabilidade e tolerância a falhas no fluxo de dados. O processo consiste em três etapas:

4.4.1 Extração (*extract*)

A fase de extração foi implementada por meio de duas funções principais, definidas como tarefas do Airflow com o decorador `@task`:

- **`get_dados_api()`**: acessa a API pública do Portal de Transparência da Prefeitura Municipal de Florianópolis, recuperando registros em formato JSON referentes a licitações e contratos. Essa API retorna informações resumidas, como número do processo, modalidade, valores e datas principais.
- **`get_dados_internos()`**: complementa os dados obtidos na API através de requisições HTTP do tipo POST. Nessa etapa, são carregados registros mais detalhados, incluindo unidades gestoras responsáveis, fornecedores, itens licitados, vencedores, empenhos e documentos textuais associados. Para realizar essa extração, a função utiliza *payloads* em formato JSON armazenados localmente.

Além das duas funções previamente descritas, foi implementada a função **`get_dados_api_empenho()`**, que acessa o *endpoint* de execuções de despesas da API pública do Portal de Transparência⁸ com o objetivo de obter mais dados referentes aos empenhos e seus vínculos com licitações e contratos. Estes dados foram considerados impróprios para uso na fase de transformação; portanto, a função não foi integrada ao processo de ETL e foi mantida separada da função **`get_dados_api()`**.

⁷ Disponível em: <https://github.com/gustavodesalles/airflow-pmf>

⁸ Disponível em: <https://transparencia.e-publica.net/epublica-portal/#/florianopolis/portal/dadosAbertos/despesaView>

A estratégia de extrair dados de duas origens distintas foi adotada para mitigar limitações de cada fonte e assegurar maior completude e consistência nas informações obtidas.

4.4.2 Transformação (*transform*)

Uma vez coletados, os dados passam por tratamento e normalização, de forma a atender ao modelo de dados definido no banco PostgreSQL. Esse processo é realizado nas funções:

- **juntar_dados_licitacao()**: integra as informações provenientes da API e das consultas internas para cada licitação, consolidando-as em um registro único. Além disso, executa cálculos derivados, como o valor total de itens licitados, e relaciona as licitações com entidades auxiliares (advogados, fornecedores, unidades gestoras, vencedores, itens e textos/documentos).
- **juntar_dados_contrato()**: realiza procedimento análogo para os contratos administrativos. Nessa etapa, são inseridos dados de fornecedores, itens contratados, valores totais, documentos e empenhos vinculados, garantindo também o relacionamento entre contratos e licitações correspondentes.

Um aspecto relevante na etapa de transformação é a utilização de operações idempotentes, por meio da cláusula `ON CONFLICT` do PostgreSQL. Esse recurso assegura que, em caso de repetição de dados, os registros sejam atualizados em vez de duplicados, preservando a integridade e consistência do banco de dados.

4.4.3 Carga (*load*)

Na etapa final, os dados tratados são inseridos no banco PostgreSQL em um esquema relacional construído previamente. A DAG contempla tarefas específicas para garantir a infraestrutura necessária:

- **criar_tabelas**: executa o *script* SQL responsável por criar todas as tabelas do esquema relacional, caso ainda não existam.
- **inserir_unidades_gestoras**: realiza a inserção inicial das unidades gestoras de referência, fundamentais para relacionar as informações dos processos.

Após essa inicialização, as tarefas de carga inserem dados nas tabelas principais *Licitacao*, *Contrato*, *Item_Licitacao*, *Item_Contrato*, *Fornecedor*, *Advogado*, *Unidade Gestora*, *Texto* e *Empenho*. Cada tabela foi projetada para representar entidades-chave no contexto das compras públicas, permitindo consultas analíticas posteriores.

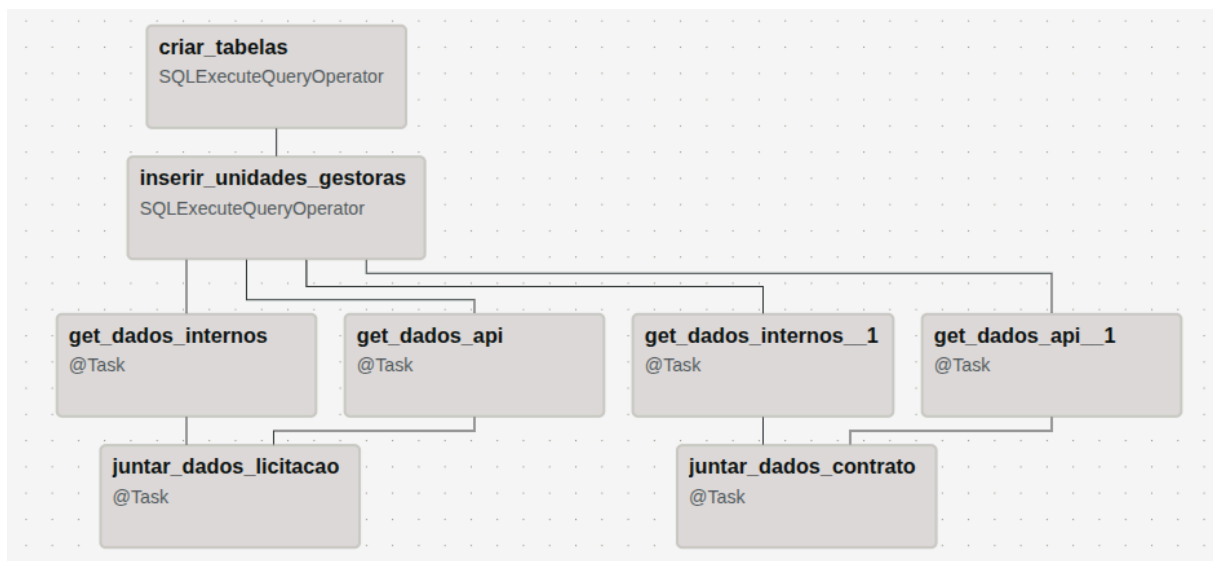
4.4.4 Orquestração da DAG

A DAG definida (`dag_compras_publicas_v2`) possui agendamento diário (`@daily`), de forma que novos dados sejam coletados automaticamente. A execução segue a seguinte ordem lógica:

1. Criação das tabelas (`criar_tabelas`);
2. Inserção das unidades gestoras (`inserir_unidades_gestoras`);
3. Extração de dados da API e das consultas internas, tanto para licitações quanto para contratos;
4. Consolidação e carga dos dados transformados em suas respectivas tabelas (`juntar_dados_licitacao` e `juntar_dados_contrato`).

A Figura 6 mostra o encadeamento das tarefas da DAG visualizado no modo *graph view* do Airflow, em ordem de execução de cima para baixo.

Figura 6 – Visualização das tarefas e dependências da DAG.



Fonte: Elaborada pelo autor.

4.4.5 Dificuldades

Durante o desenvolvimento da DAG no Airflow, uma das principais dificuldades enfrentadas foi a implementação do método de obtenção de dados a partir da API interna do Portal de Transparência da Prefeitura, correspondente à tarefa `get_dados_internos()`. As páginas do Portal são renderizadas dinamicamente, o que inviabilizou o uso de técnicas tradicionais de *web scraping* baseadas na extração de conteúdo estático.

Além disso, as requisições HTTP necessárias para recuperar os registros são do tipo POST, exigindo a identificação e reconstrução manual do *payload* adequado para cada solicitação. Esse processo envolveu a inserção de parâmetros específicos, como as datas de início e término do período de busca, bem como a quantidade de registros a serem retornados. A execução das requisições também apresentou instabilidades: em diversas ocasiões, o servidor respondia com o código HTTP 202 (*Accepted*), sem retornar os dados imediatamente, o que levou à implementação de um mecanismo de espera e reenvio da requisição até que o retorno com o código HTTP 200 (OK) fosse obtido.

Sobre a execução do Airflow, houve problemas com a configuração do ambiente. A plataforma foi instanciada em um *container* Docker, e foi necessário realizar ajustes no arquivo YAML para garantir que o ambiente localizasse o arquivo da DAG no diretório correto e estabelecesse a conexão com o banco de dados PostgreSQL corretamente.

Verificou-se que os dados obtidos apresentavam limitações significativas no que se refere aos empenhos vinculados às licitações e contratos. Como alternativa, foi explorado o *endpoint* de execução de despesas disponibilizado pela API pública da Prefeitura Municipal de Florianópolis, com o intuito de complementar as informações disponíveis. Contudo, essa abordagem mostrou-se pouco eficaz, uma vez que grande parte dos empenhos não apresentava vínculo explícito com licitações ou contratos. Além disso, identificou-se que o número de identificação dos empenhos não segue uma ordem clara (por exemplo, múltiplos empenhos com o número 1 e valores diferentes em uma mesma data, sem informações associadas de licitação ou contrato). Tais inconsistências inviabilizaram a verificação adequada de duplicidade antes da inserção no banco de dados, comprometendo a confiabilidade da integração.

4.4.6 Considerações e integrações com IA

O processo de ETL construído garante a disponibilização contínua de dados de licitações e contratos em uma base relacional robusta, apta para consultas exploratórias, geração de indicadores e análises futuras. O uso do Apache Airflow trouxe vantagens como controle de dependências, tolerância a falhas (com tentativas de reexecução em caso de erros) e rastreabilidade da execução por meio de sua interface *web*.

Para estruturar as etapas do processo de ETL, as funções inicialmente desenvolvidas em Python foram adaptadas e convertidas em tarefas (*tasks*) do Airflow com o auxílio da ferramenta de geração de código GitHub Copilot, que realizou a tradução do código para o formato compatível com o orquestrador. Como a API TaskFlow do Airflow permite utilizar funções em Python como tarefas simplesmente adicionando a anotação `@task` à definição da função, o código resultante não teve diferenças significativas das funções originais.

É possível alterar os parâmetros das datas inicial e final dos períodos nas consultas feitas durante a execução da DAG para obter dados de outros intervalos de tempo. Após

alguns testes com intervalos de diversos tamanhos, concluiu-se que a DAG consegue obter dados de um intervalo de até um mês com tempo de execução razoável, sem demoras excessivas.

Considerando a diversidade de fontes de dados referentes a compras públicas e as diferenças entre os registros de cada fonte, identificou-se o potencial na integração de modelos de linguagem de grande porte (LLM) como uma possível evolução do projeto. Esses modelos poderiam ser utilizados para interpretar e extrair informações de registros em diferentes formatos, adaptando-se a variações estruturais e semânticas nos dados provenientes de fontes distintas. Por meio de técnicas de compreensão contextual, um LLM seria capaz de reconhecer padrões, inferir campos ausentes e converter automaticamente descrições textuais em valores estruturados de acordo com os esquemas definidos no CÉOS.

Além disso, a incorporação de LLMs ao fluxo de ETL poderia ampliar as possibilidades de análise, permitindo a aplicação de métodos de enriquecimento semântico dos dados. Por exemplo, o modelo poderia classificar automaticamente as naturezas de despesas, identificar relações entre fornecedores e categorias de contratos ou mesmo detectar anomalias em valores e descrições, funcionando como uma camada adicional de validação inteligente. Porém, é necessário realizar estudos mais aprofundados sobre desempenho, custo computacional e acurácia dos resultados.

5 RESULTADOS

Nesta seção, apresentam-se os resultados obtidos a partir da análise exploratória dos dados de licitações e contratos administrativos da Prefeitura Municipal de Florianópolis. O conjunto de dados contempla registros entre 2014 e 2024, abrangendo informações como modalidades de licitação, valores contratados, datas de homologação, prazos de vigência e fornecedores participantes.

A análise exploratória teve como objetivo compreender o comportamento geral dos dados, identificar padrões e tendências ao longo do período, bem como levantar indícios de possíveis irregularidades ou inconsistências cadastrais. Para isso, foram elaborados gráficos e tabelas que sintetizam os aspectos mais relevantes. Os gráficos foram gerados com *scripts* na linguagem Python¹, utilizando a biblioteca Matplotlib.

Para complementar a análise, a Tabela 14 apresenta um resumo das modalidades de licitações registradas entre 2014 e 2024, considerando o valor total movimentado, a quantidade de processos licitatórios, o número de itens licitados e a quantidade de textos descritivos associados. Observa-se que as modalidades **Pregão Eletrônico** e **Dispensa** concentram os maiores valores totais, superando a marca de R\$ 15 bilhões cada. Modalidades tradicionais como **Pregão Presencial** e **Concorrência** também apresentam participação relevante, enquanto modalidades menos frequentes, como **Concurso** e **Leilão**, apresentam volumes reduzidos. Em particular, nota-se que há 1553 registros de licitações da modalidade **Pregão**, mas nenhum possui valor monetário, itens ou textos.

Tabela 14 – Resumo das modalidades de licitações entre 2014 e 2024

Modalidade	Valor total (em R\$)	No. licitações	No. itens	No. textos
Pregão Eletrônico	15507026111.65	2063	31258	2831
Dispensa	15476798797.17	1345	2227	1598
Convite	6486691903.41	87	43	43
Concorrência	1199136337.31	302	271	403
Pregão Presencial	487201833.11	540	3688	885
Tomada de Preço	316345806.41	521	402	313
Inexigibilidade	228545024.38	953	609	1761
Outros	213741098.32	248	556	161
Leilão	49320025.48	3	162	55
Credenciamento	29404786.46	29	154	64
Concurso	1516107.35	1	304	1
Pregão	0	1553	0	0

Fonte: Elaborada pelo autor.

A Tabela 15 relaciona as modalidades de licitações ao número de contratos adminis-

¹ Disponível em: <https://github.com/gustavodesalles/INE5454-trabalho>

trativos formalizados e a quantidade de empenhos vinculados aos processos licitatórios do mesmo período. Devido aos *endpoints* diferentes utilizados para obter os registros dos contratos e empenhos, houve uma quantidade menor de registros retornados em comparação com a consulta da API pública.

Nota-se a predominância do Pregão Eletrônico, que representa mais de 4 mil contratos e quase 10 mil empenhos, reforçando sua relevância no processo de compras públicas. Modalidades como Inexigibilidade e Dispensa também aparecem com destaque, especialmente no volume de empenhos, sugerindo uso recorrente para aquisições de natureza específica ou emergencial. A modalidade Leilão, apesar de conter somente três licitações, representa 118 contratos, enquanto Pregão e Concurso não possuem nenhum contrato ou empenho.

Tabela 15 – Resumo das modalidades de contratos entre 2014 e 2024

Modalidade	No. contratos	% contratos	No. empenhos	% empenhos
Pregão Eletrônico	4220	66.66%	9924	61.68%
Dispensa	797	12.59%	1944	12.08%
Inexigibilidade	492	7.77%	846	5.26%
Pregão Presencial	396	6.26%	589	3.66%
Outros	378	5.97%	450	2.80%
Tomada de Preço	226	3.57%	162	1.01%
Concorrência	168	2.65%	1949	12.11%
Leilão	118	1.86%	0	0.00%
Convite	36	0.57%	0	0.00%
Credenciamento	0	0.00%	223	1.39%
Pregão	0	0.00%	0	0.00%
Concurso	0	0.00%	0	0.00%
Total	6331	100%	16087	100%

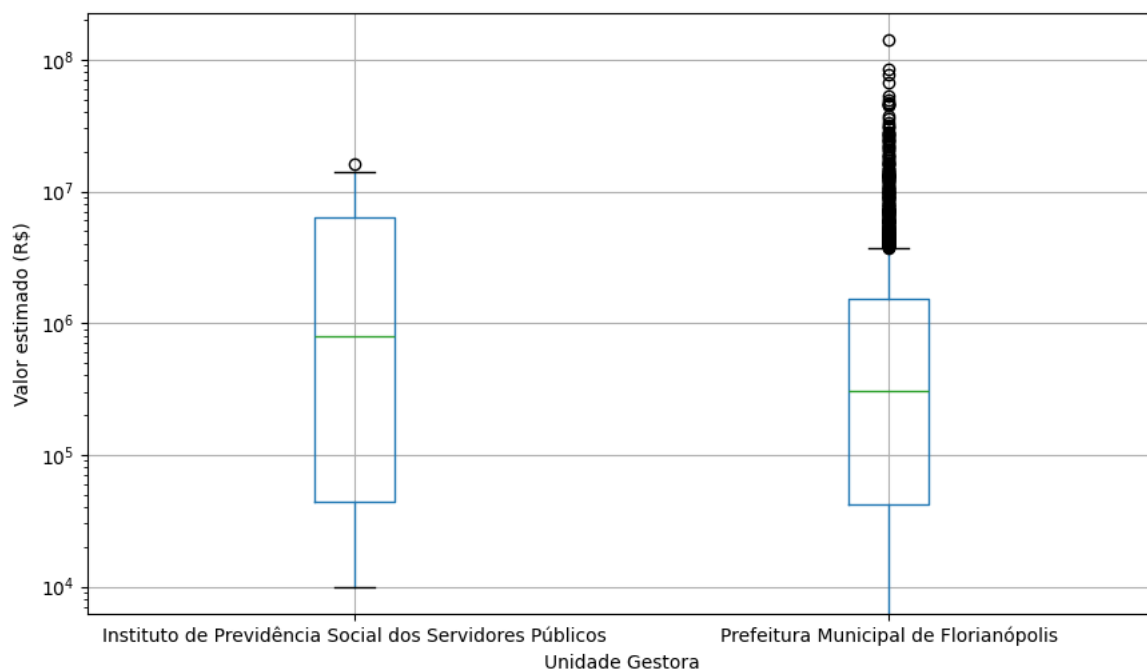
Fonte: Elaborada pelo autor.

Esses dados quantitativos fornecem uma visão inicial da distribuição das modalidades ao longo do período analisado, permitindo identificar aquelas que concentram maior movimentação financeira e operacional. Na sequência, são apresentados gráficos que exploram visualmente essas tendências e outros aspectos relevantes das licitações e contratos.

DISTRIBUIÇÃO DE VALORES ESTIMADOS DE LICITAÇÕES POR UNIDADE GESTORA

Em geral, nota-se uma grande quantidade de registros de licitações cuja unidade gestora é a Prefeitura Municipal de Florianópolis (código 1) ou o Instituto de Previdência Social dos Servidores Públicos (código 22) que não possuem valores estimados informados. Isto resulta em um *boxplot* com muitos *outliers*, como exemplificado pela Figura 7.

Figura 7 – Distribuição dos valores das licitações das principais unidades gestoras (1 e 22) no ano de 2024.



Fonte: Elaborada pelo autor.

Em relação aos contratos, verifica-se uma maior diversidade de unidades gestoras envolvidas, quando comparada às licitações. Essa diferença ocorre mesmo considerando que uma licitação pode estar associada a múltiplas unidades gestoras, enquanto um contrato está vinculado a apenas uma. Além disso, constatou-se uma quantidade significativamente menor de contratos sem o valor total informado, o que denota maior completude dos dados nessa categoria.

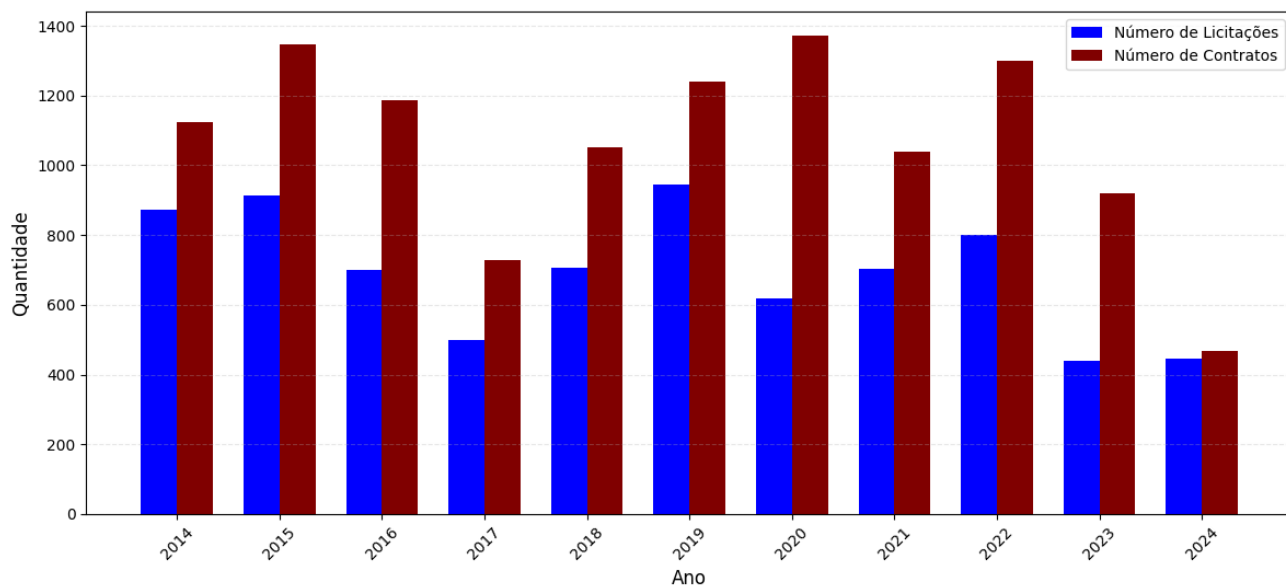
Entre todas as licitações analisadas, apenas quatro registros não apresentam unidade gestora informada, sendo todos referentes ao ano de 2024. No caso dos contratos, todos os registros contêm a informação da unidade gestora responsável.

QUANTIDADE ANUAL DE LICITAÇÕES E CONTRATOS

A Figura 8 compara a quantidade anual de licitações e contratos registrados entre 2014 e 2024. Observa-se que, na maior parte dos anos, o número de contratos supera significativamente o número de licitações, indicando que nem todas as licitações correspondem a um único contrato ou que há contratos decorrentes de processos anteriores.

De 2014 para 2015, há um crescimento tanto nas licitações quanto nos contratos, seguido de quedas expressivas em ambos os indicadores em 2016 e 2017. Considerando o fato de 2016 ter sido um ano de eleição municipal, esta redução está possivelmente relaci-

Figura 8 – Quantidade total de licitações e contratos entre 2014 e 2024.



Fonte: Elaborada pelo autor.

onada a alterações normativas, restrições orçamentárias ou mudanças nos procedimentos administrativos.

O ano de 2020 apresenta um comportamento atípico: o número de contratos foi mais que o dobro do número de licitações. Isto pode indicar a formalização de múltiplos contratos derivados de um número relativamente reduzido de processos licitatórios, possivelmente como resposta a demandas emergenciais, como as decorrentes da pandemia de COVID-19.

Por fim, o ano de 2024 exibe a menor diferença absoluta entre as quantidades de licitações e contratos em todo o período analisado. Esse alinhamento pode refletir maior proporcionalidade entre a quantidade de processos licitatórios e a execução contratual, sugerindo ajustes na gestão ou no planejamento das aquisições públicas.

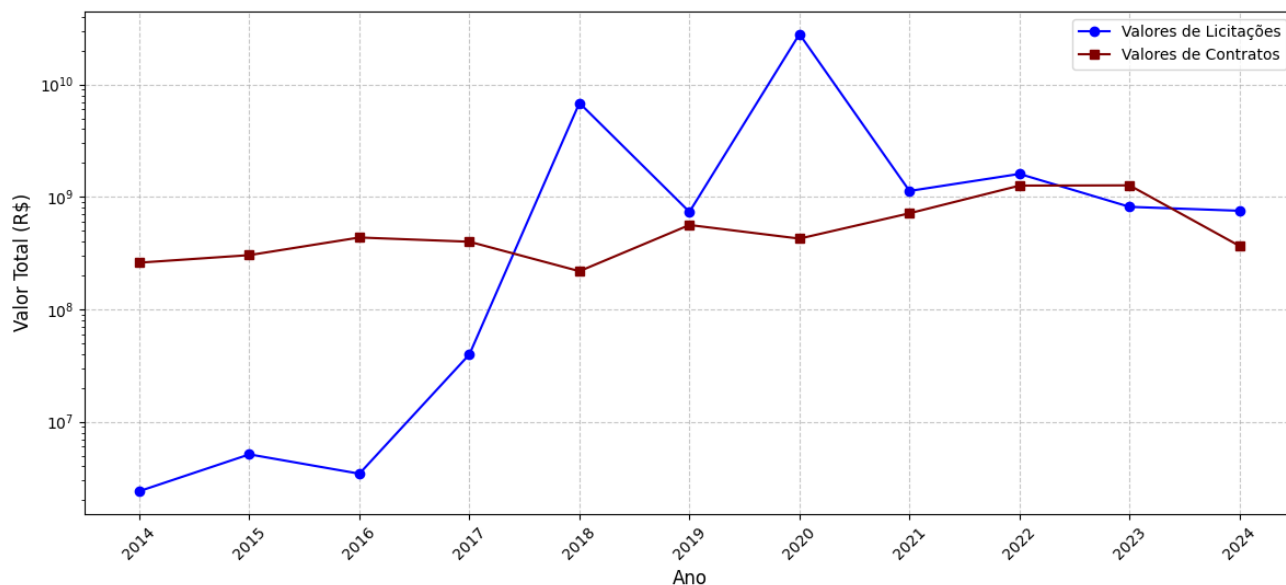
Observa-se que houve uma queda do número de registros no ano de 2017 em comparação com os anos anteriores. O ano de 2020 apresenta uma grande discrepância entre os números, sendo a quantidade de contratos maior que o dobro da quantidade de licitações. Em contrapartida, 2024 possui a menor diferença entre licitações e contratos.

VARIAÇÃO ANUAL DE VALORES TOTAIS DE LICITAÇÕES E CONTRATOS

A Figura 9 apresenta a variação anual dos valores totais de licitações e contratos registrados entre 2014 e 2024, representados em escala logarítmica para melhor visualização das discrepâncias de ordem de grandeza.

No período inicial, de 2014 a 2017, observa-se que os valores de contratos se man-

Figura 9 – Variação anual de valores totais de licitações e contratos entre 2014 e 2024.



Fonte: Elaborada pelo autor.

têm consistentemente superiores aos valores de licitações. Esse comportamento pode ser explicado pelo fato de que muitos contratos são firmados a partir de licitações de anos anteriores ou por contratações diretas, previstas na legislação, que não necessariamente guardam proporcionalidade imediata com os valores licitados no mesmo exercício.

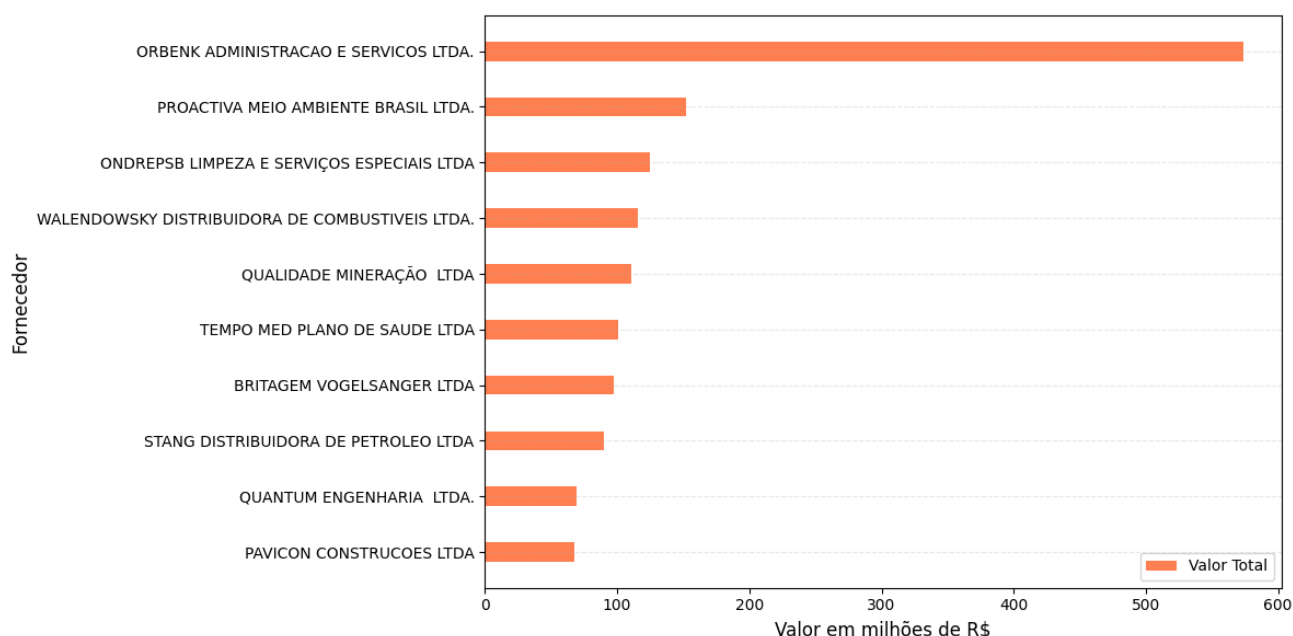
Em 2018, há um crescimento abrupto nos valores de licitações, ultrapassando em larga escala os valores de contratos. Essa discrepância pode estar associada a grandes editais lançados naquele ano, com expressiva previsão orçamentária, mas que não necessariamente se converteram no mesmo volume de contratos firmados de imediato.

Após um aparente equilíbrio no ano de 2019 entre os valores de licitações e contratos, em 2020 observa-se novamente uma disparidade, com os valores licitados alcançando seu ponto mais alto em toda a série, superando em várias ordens de magnitude os valores contratados. Esse comportamento pode estar relacionado a processos emergenciais desencadeados pela pandemia de COVID-19, em que foram anunciados grandes volumes de recursos licitados, mas cuja execução contratual pode ter ocorrido em montantes menores ou em fases posteriores.

A partir de 2021, há uma redução gradual da diferença entre licitações e contratos, atingindo um equilíbrio entre os valores de 2022 em diante. De 2023 até 2024, nota-se uma queda em ambos os índices, ainda que um pouco mais acentuada para os contratos. Essa retração pode refletir tanto um cenário de restrições orçamentárias quanto o processo de adaptação administrativa à plena vigência da nova Lei de Licitações e Contratos (Lei no. 14.133/2021), que se tornou a única lei vigente para licitações e contratações públicas em 2024.

PRINCIPAIS FORNECEDORES DE CONTRATOS

Figura 10 – Valores totais gastos com fornecedores de contratos entre 2014 e 2024.



Fonte: Elaborada pelo autor.

A Figura 10 apresenta os dez fornecedores com maior valor financeiro total de contratos administrativos firmados com a Prefeitura Municipal de Florianópolis durante o período analisado, expressos em milhões de reais.

A análise evidencia uma forte concentração de recursos em um número reduzido de fornecedores. Em especial, destaca-se a empresa Orbenk Administração e Serviços Ltda., cujo montante contratado ultrapassa R\$ 550 milhões, valor significativamente superior ao das demais. Essa predominância, ao lado de empresas como Ondrepsb Limpeza e Serviços Especiais Ltda., indica que os serviços de vigilância patrimonial e limpeza urbana representam uma parcela expressiva das despesas contratuais do município.

Entre os demais fornecedores com maiores volumes de recursos, predominam aqueles ligados ao setor de obras públicas, como saneamento básico (Proactiva Meio Ambiente Brasil Ltda.), pavimentação de ruas e estradas (Qualidade Mineração Ltda., Britagem Vogelsanger Ltda., Pavicon Construções Ltda.) e iluminação pública (Quantum Engenharia Ltda.), além de empresas de distribuição de combustíveis (Walendowsky Distribuidora de Combustíveis Ltda. e Stang Distribuidora de Petróleo Ltda.). A única exceção nesse conjunto é a Tempo Med Plano de Saúde Ltda., responsável pela oferta de planos de saúde aos servidores municipais.

Os dados da Tabela 16 complementam essa interpretação ao mostrar a distribuição das licitações segundo sua finalidade entre 2014 e 2024. Nota-se que a maior parte

Tabela 16 – Resumo das finalidades de licitações entre 2014 e 2024

Finalidade das licitações	Qtde.	Valor total
Outras contratações	3733	37.770.121.626,49
<i>None</i> (sem valor informado)	2841	0
Contratação de Serviços	378	352.536.424,34
Aquisição de Bens	331	409.253.713,35
Obras e Serviços de Engenharia	261	1.373.265.401,01
Compras e Outros Serviços	68	7.517.640,14
Concessão de Direitos	13	5.053.489,57
Contratação de Tecnologia da Informação e Comunicação	11	13.102.944,05
Concessão de Direito Real de Uso de Bem Público	3	5.112.806,61
Serviços de Publicidade	2	12.300.760,00
Contratação de Aluguéis	2	948.000,00
Termo de Parceria	1	0,01
Alienação de Bens	1	46.515.025,48

Fonte: Elaborada pelo autor.

das licitações encontra-se na categoria **Outras contratações**, que concentra mais de R\$ 37 bilhões em valores totais, representando a quase totalidade dos recursos. Esse agrupamento genérico dificulta a análise detalhada, mas sugere que boa parte das despesas se concentra em contratações sem uma classificação específica. Além disso, observa-se um número significativo de licitações (2.841 registros) sem finalidade informada, todas sem valor total registrado, o que indica lacunas importantes nos dados.

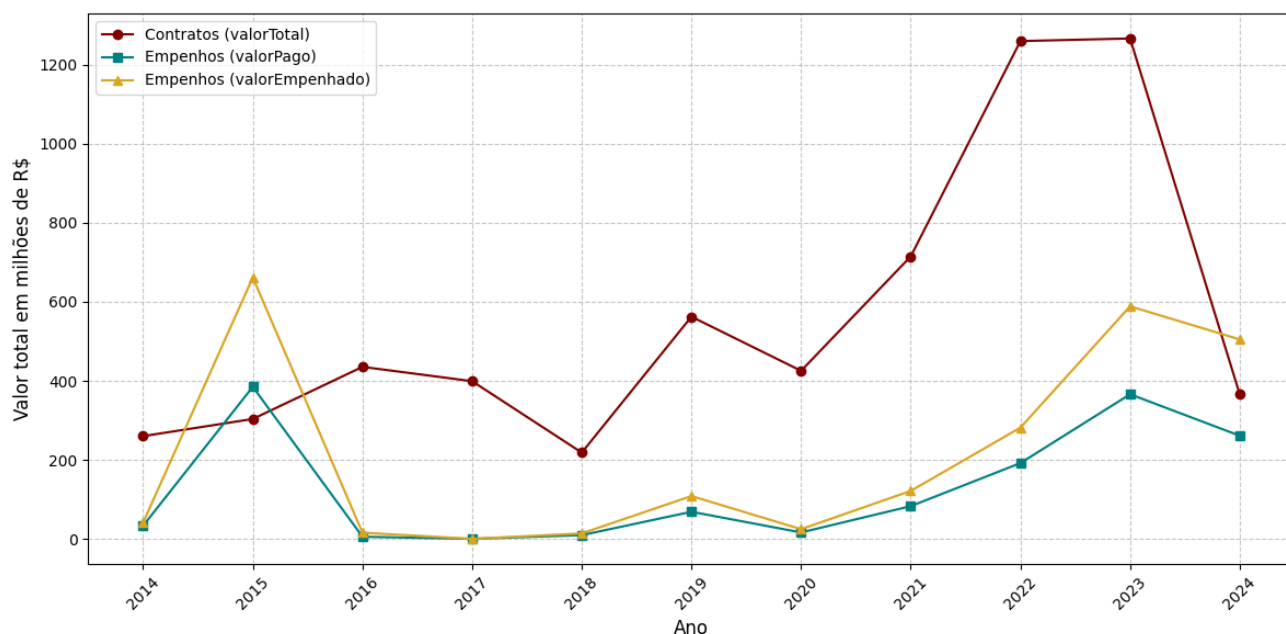
Em seguida, destacam-se as licitações voltadas a Obras e Serviços de Engenharia, que somam aproximadamente R\$ 1,37 bilhão, refletindo o peso desse setor nas despesas públicas, seguido por categorias como Aquisição de Bens (R\$ 409 milhões) e Contratação de Serviços (R\$ 352 milhões). Finalidades mais específicas, como Contratação de TIC (R\$ 13 milhões), Serviços de Publicidade (R\$ 12 milhões) e Concessão de Direitos (R\$ 5 milhões), aparecem com valores relativamente pequenos, mas ainda assim indicam a diversidade de áreas atendidas pelas licitações. Casos isolados, como a Alienação de Bens (R\$ 46 milhões), também se destacam, embora não representem o núcleo das despesas.

De modo geral, os resultados apontam que a maior parte dos recursos públicos contratados está concentrada em categorias genéricas e em um conjunto limitado de fornecedores, com destaque expressivo para um único ator econômico. Esse cenário sugere uma possível dependência da administração em relação a determinados prestadores e evidencia problemas de transparência decorrentes da classificação insuficiente das finalidades de licitação. Assim, emergem questões relevantes para discussão sobre concorrência, diversificação da base de fornecedores e gestão de riscos nas contratações públicas.

COMPARAÇÃO ENTRE VALORES DE CONTRATOS E EMPENHOS

O gráfico apresentado ilustra a evolução anual dos valores totais de contratos e de empenhos (nas dimensões `valorPago` e `valorEmpenhado`) entre os anos de 2014 e 2024. Os valores foram agregados em milhões de reais e permitem observar a relação entre os compromissos firmados por meio de contratos e a execução financeira representada pelos empenhos. Nota-se que os contratos apresentam valores sistematicamente superiores aos empenhos; isto é esperado, já que o contrato estabelece um teto de gastos, enquanto os empenhos refletem a efetiva alocação de recursos. Contudo, a discrepância entre as curvas é bastante acentuada em determinados períodos, especialmente a partir de 2021, quando os contratos atingem valores recordes, mas os empenhos crescem em proporção menor, evidenciando a distância entre planejamento e execução.

Figura 11 – Evolução anual de valores de contratos e empenhos vinculados entre 2014 e 2024.



Fonte: Elaborada pelo autor.

Também se observam variações bruscas ao longo da série. Em 2015, houve um aumento expressivo dos valores empenhados e pagos, mas essa tendência não se manteve nos anos seguintes. Em 2016, os valores registrados caíram de forma significativa e, em 2017, não há qualquer registro de empenhos, embora os contratos tenham permanecido em patamares elevados. Já entre 2019 e 2023 observa-se um crescimento contínuo dos contratos, com destaque para 2022 e 2023, que superam a marca de 1,2 bilhão de reais. Em contraste, os empenhos, apesar de acompanharem a tendência de crescimento, permanecem muito abaixo desses valores. Em 2024, verifica-se uma queda expressiva nos contratos, ao

passo que os empenhos se mantêm em níveis relativamente altos, possivelmente refletindo a execução de compromissos assumidos em exercícios anteriores.

Para aprofundar essa análise, foi desenvolvido um método específico para comparar o valor registrado em cada contrato com a soma dos valores de seus respectivos empenhos, com o objetivo de verificar a consistência entre as etapas do processo de execução orçamentária. Essa verificação permite identificar divergências que poderiam comprometer a confiabilidade dos dados ou revelar problemas de gestão administrativa. Caso seja encontrada diferença, o método imprime o número do contrato, seguido do valor total informado no registro da API pública da Prefeitura Municipal de Florianópolis e da soma dos valores dos empenhos.

Como os registros de contratos obtidos da API pública não continham dados de empenhos, foi necessário recorrer à API interna utilizada pelo próprio sistema para carregar as páginas, a qual contém os empenhos associados a cada contrato. Essa fonte alternativa possibilitou a execução da análise proposta, revelando uma grande quantidade de discrepâncias nos contratos firmados em 2015. Essas inconsistências podem ter diferentes origens, como erros de lançamento, mudanças nos procedimentos de registro adotados à época ou falhas de integração entre sistemas, e indicam a necessidade de investigações adicionais.

Figura 12 – Exemplo de múltiplos registros de contratos com números idênticos.

```
17/SMA/2024: 40000.0 - 22096962.02
17/SMA/2024: 80000.0 - 22096962.02
202/SMSOP/2024: 550000.0 - 722525.82
209/SMA/2024: 3098.79 - 204106.96
209/SMA/2024: 2277.0 - 204106.96
209/SMA/2024: 10329.3 - 204106.96
209/SMA/2024: 63342.0 - 204106.96
112/SMTI/2024: 2264000.0 - 2512438.08
250/FMAS/2024: 33148.92 - 37734.19
17/SMA/2024: 444000.0 - 22096962.02
142/SMLCP/2024: 11960.0 - 22771.84
17/SMA/2024: 109263.76 - 22096962.02
370/SMC/2024: 1549265.88 - 1711628.37
```

Fonte: Elaborada pelo autor.

Outro achado importante foi a existência de contratos que compartilham o mesmo número, mas que estão associados a unidades gestoras distintas, como demonstra a Figura 12. Esse comportamento decorre da modelagem adotada para o objeto "Contrato", que, ao contrário do objeto "Licitação", pode ser vinculado a uma única unidade gestora.

Como consequência, quando um contrato envolve mais de uma unidade, ele passa a ser registrado em duplicidade, gerando aparente inconsistência nos dados.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou o desenvolvimento de um processo de ETL (*Extract, Transform and Load*) utilizando o Apache Airflow para orquestração de *workflows*, aplicado à coleta e integração de dados de licitações e contratos da Prefeitura Municipal de Florianópolis.

A solução implementada permitiu automatizar as etapas de extração, transformação e carga, consolidando dados provenientes de fontes distintas em um banco de dados relacional PostgreSQL. A DAG desenvolvida realizou com sucesso a integração entre dados públicos via API e registros internos obtidos por requisições específicas, a normalização e tratamento dos dados para garantir a consistência, e a criação de uma base relacional estruturada capaz de suportar consultas analíticas.

Como continuidade deste trabalho, sugerem-se as seguintes linhas de evolução: (i) ampliar o pipeline para suportar maiores volumes de dados e múltiplas fontes governamentais; (ii) integrar o banco de dados com ferramentas de *business intelligence* para geração de *dashboards* e relatórios; (iii) explorar o uso de algoritmos de aprendizado de máquina para detecção de anomalias em licitações e contratos; (iv) aprofundar a coleta e integração de informações relacionadas a empenhos, permitindo maior detalhamento na análise do ciclo completo das despesas públicas; e (v) implementar mecanismos para revisar e atualizar registros antigos quando houver alterações ou complementações publicadas em períodos posteriores.

REFERÊNCIAS

CONTAS DA UNIÃO, Tribunal de. **Licitações & Contratos: Orientações e Jurisprudência do TCU / Tribunal de Contas da União**. 5. ed. Brasília: Secretaria-Geral da Presidência, 2023.

FINNIGAN, Leanne; TONER, Emily. Building and maintaining metadata aggregation workflows using apache airflow. **Temple University Libraries**, 2021.

HOLLINGSWORTH, David. Workflow management coalition: The workflow reference model, 1995.

JESUS, Mauricio Barros de. Modelo preditivo de risco de irregularidades em compras públicas no Estado de Goiás, 2021.

MELLO, Ronaldo et al. Uma Proposta de Modelagem de Dados no Domínio de Fraudes em Licitações Públicas. In: ANAIS Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados. Florianópolis/SC: SBC, 2024. p. 220–226. DOI: 10.5753/sbbd_estendido.2024.243665. Disponível em: https://sol.sbc.org.br/index.php/sbbd_estendido/article/view/30797.

NAZÁRIO, Débora Cabral; SILVA, Paulo Fernando da; ROVER, Aires José. Avaliação da qualidade da informação disponibilizada no Portal da Transparência do Governo Federal. **Revista Democracia Digital e Governo Eletrônico**, n. 6, 2012.

POSSAMAI, Ana Júlia; SOUZA, Vitoria Gonzatti de. Transparência e dados abertos governamentais: possibilidades e desafios a partir da lei de acesso à informação. **Administração Pública e Gestão Social**, 2020.

RUITER, Julian de; HARENSLAK, Bas. **Data Pipelines with Apache Airflow**. [S.l.]: Simon e Schuster, 2021.

SANTOS, Everton Schneider dos et al. Performance Variability of Machine Learning Models using Limited Data for Collusion Detection: A Case Study of the Brazilian Car Wash Operation. In: ANAIS do XXXIX Simpósio Brasileiro de Bancos de Dados. Florianópolis/SC: SBC, 2024. p. 431–443. DOI: 10.5753/sbbd.2024.240845. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/30711>.

SANTOS, Maria Teresa Silva et al. Ferramenta de visualização de dados abertos do portal de transparência da câmara municipal da cidade de florianópolis. In: SBC. WORKSHOP de Computação Aplicada em Governo Eletrônico (WCGE). [S.l.: s.n.], 2021. p. 71–82.

SCHMITZ, Fernando et al. Detecting Fraud in Public Procurement: A GMM-Based Approach to Analyzing Tender Data. In: ANAIS do XXXIX Simpósio Brasileiro de Bancos de Dados. Florianópolis/SC: SBC, 2024. p. 207–219. DOI: 10.5753/sbbd.2024.240649. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/30694>.

VASSILIADIS, Panos; SIMITSIS, Alkis. Extraction, Transformation, and Loading. **Encyclopedia of database systems**, Citeseer, v. 10, p. 14, 2009.

VASSILIADIS, Panos; SIMITSIS, Alkis; SKIADOPOULOS, Spiros. Conceptual modeling for ETL processes. In: PROCEEDINGS of the 5th ACM international workshop on Data Warehousing and OLAP. [S.l.: s.n.], 2002. p. 14–21.

Apêndices

APÊNDICE A – ARTIGO DO TCC

Extração, transformação e carga de dados sobre compras da Prefeitura de Florianópolis usando Airflow

Gustavo Fukushima de Salles¹, Renato Fileto²

¹ Departamento de Informática e Estatística - INE
Universidade Federal de Santa Catarina (UFSC) – Florianópolis, SC – Brasil

`gustavo.fukushima@grad.ufsc.br, r.fileto@ufsc.br`

Abstract. *A workflow defines a process composed of interconnected tasks and the flow of information between them. Workflow orchestration automates and manages execution through software that coordinates tasks, triggers resources, and reports failures. This work is part of the CÉOS project, developed by UFSC in partnership with MPSC, aimed at creating workflows for large-scale data analysis and processing. An Extract, Transform, and Load (ETL) workflow was specified and implemented to process procurement and bidding data from the Florianópolis City Hall, orchestrated by Apache Airflow using Directed Acyclic Graphs (DAGs). The results are evaluated regarding data quality and quantity, error frequency, and process execution time.*

Resumo. *Um fluxo de trabalho (workflow) define um processo composto por tarefas encadeadas e pelo fluxo de informações entre elas. A orquestração de workflows automatiza e gerencia sua execução por meio de um software que coordena tarefas, aciona recursos e reporta falhas. Este trabalho integra o projeto CÉOS, desenvolvido pela UFSC em parceria com o MPSC, voltado à criação de workflows para análise e processamento de dados em larga escala. Foi especificado e implementado um fluxo de extração, transformação e carga (Extract, Transform, Load — ETL) de dados sobre licitações e compras da Prefeitura Municipal de Florianópolis, orquestrado pelo Apache Airflow com grafos acíclicos direcionados (Directed Acyclic Graphs — DAGs). Os resultados são avaliados quanto à qualidade e quantidade dos dados, frequência de erros e tempo de execução dos processos.*

1. Introdução

O uso de plataformas de orquestração de *workflows* é essencial para gerenciar aplicações com grande volume de dados, como *data warehouses*. Um *workflow* consiste na automatização total ou parcial de um processo de negócio, envolvendo a modelagem das atividades, o controle de execução e a interação entre usuários e aplicações de TI.

Tais *workflows* são amplamente aplicados em processos de extração, transformação e carga de dados (*extract-transform-load*, ou ETL), que visam consolidar dados de diferentes fontes em um formato homogêneo e estruturado. O Apache Airflow destaca-se como uma plataforma de orquestração de código aberto amplamente utilizada para esse propósito. Nele, *workflows* são representados por grafos acíclicos direcionados (*directed acyclic graphs*, ou DAGs), definidos em Python, que descrevem as tarefas e suas dependências.

Este trabalho integra o projeto CÉOS, desenvolvido pela Universidade Federal de Santa Catarina (UFSC) em parceria com o Ministério Público de Santa Catarina (MPSC). O projeto tem como objetivo automatizar a extração e análise de dados públicos para apoiar atividades de fiscalização e tomada de decisão. Neste contexto, o presente estudo descreve o desenvolvimento e a avaliação de um fluxo de ETL para dados de licitações e compras realizadas pela Prefeitura Municipal de Florianópolis (PMF), utilizando o Apache Airflow como orquestrador.

1.1. Objetivo

O objetivo deste trabalho é orquestrar um fluxo de ETL para coleta, transformação e carregamento de dados referentes a compras públicas do município de Florianópolis, a fim de subsidiar análises de gastos no âmbito do projeto CÉOS.

1.2. Metodologia

A metodologia adota três etapas principais. Na primeira, foi realizado um levantamento teórico sobre orquestração de ETL e uso do Apache Airflow, abrangendo estudos acadêmicos e documentação técnica.

Na segunda etapa, implementaram-se scripts de extração de dados via *web scraping* e chamadas a APIs REST do Portal da Transparência da PMF. Devido à incompletude dos dados disponíveis nos endpoints públicos, foram utilizados também os conteúdos das páginas do Portal para complementar as informações. Os dados coletados passaram por análise exploratória, identificação de *outliers* e modelagem de um esquema relacional.

Por fim, a terceira etapa compreende a orquestração do fluxo completo de ETL no Airflow e a avaliação dos resultados com base em métricas como tempo de execução, volume de dados processados e ocorrência de erros. As informações tratadas foram utilizadas para gerar visualizações que apoiam a análise de gastos públicos.

2. Trabalhos Relacionados

A análise de dados abertos governamentais tem se tornado um tema de destaque na pesquisa em computação, especialmente com o avanço das técnicas de aprendizado de máquina e o aumento da disponibilidade de dados públicos. Diversos estudos exploram a aplicação dessas técnicas à detecção de irregularidades em licitações e contratos administrativos, com o objetivo de apoiar ações de transparência e combate à corrupção.

[Schmitz et al. 2024] propõem uma metodologia não supervisionada baseada em modelos de mistura gaussiana (*Gaussian Mixture Models*, ou GMM) para identificar padrões de fraude em licitações, enquanto [dos Santos et al. 2024] utilizam dados da Operação Lava Jato para comparar modelos de aprendizado supervisionado e técnicas de otimização de hiperparâmetros. Ambos os trabalhos demonstram o potencial do uso de aprendizado de máquina na detecção de conluios e comportamentos suspeitos em processos licitatórios.

Outros estudos focam na integração e visualização de dados públicos. [Santos et al. 2021] desenvolvem uma ferramenta para extração e visualização de dados do Portal da Transparência da Câmara Municipal de Florianópolis, permitindo a análise

comparativa de gastos públicos. Já [Jesus 2021] propõem um modelo preditivo para avaliar riscos de irregularidades em licitações do Estado de Goiás, empregando técnicas de mineração de dados e classificadores supervisionados.

Complementarmente, [Mello et al. 2024] apresentam uma modelagem conceitual abrangente para o domínio de licitações, incorporando dados sobre denúncias, fraudes e agentes envolvidos, com suporte a persistência poliglota.

A Tabela 1 resume as principais características metodológicas e os conjuntos de dados empregados em cada um desses trabalhos.

Tabela 1. Comparação de trabalhos relacionados

Trabalho	Metodologia	Conjunto de dados
[Schmitz et al. 2024]	Aprendizado de máquina com modelos de mistura gaussiana	Licitações referentes à Operação Patrola
[dos Santos et al. 2024]	ETL com validação cruzada e otimização de hiperparâmetros	Licitações referentes à Operação Lava Jato
[Santos et al. 2021]	Extração de arquivos CSV e interface gráfica	Balancetes de vereadores de Florianópolis
[Jesus 2021]	Mineração de dados e treinamento supervisionado	Licitações do Estado de Goiás entre 2014 e 2019
[Mello et al. 2024]	Modelagem conceitual com persistência poliglota	-

Em conjunto, esses trabalhos evidenciam o esforço crescente em explorar dados abertos para aprimorar a transparência pública. Contudo, poucos abordam a orquestração automatizada de fluxos de ETL voltados à coleta, padronização e análise sistemática desses dados, lacuna que o presente estudo busca preencher.

3. Fontes de Dados

O processo de ETL desenvolvido neste trabalho baseou-se em duas fontes principais de dados: a API pública de compras da Prefeitura Municipal de Florianópolis (PMF) e os *endpoints* internos do Portal da Transparência, ambos voltados à disponibilização de informações sobre licitações e contratos administrativos.

3.1. API pública da PMF

O Portal da Transparência da PMF disponibiliza uma API REST que fornece registros de licitações e contratos em formato JSON, de acordo com a política de dados abertos do município. Os endpoints principais de licitações¹ e contratos² permitem consultas via método GET, filtradas por período, unidade gestora e limite de resultados.

¹Disponível em: <https://transparencia.e-publica.net/epublica-portal/#/florianopolis/portal/dadosAbertos/licitacaoView>

²Disponível em: <https://transparencia.e-publica.net/epublica-portal/#/florianopolis/portal/dadosAbertos/contratoView>

Os dados retornados contêm informações essenciais, como valores, modalidades, fornecedores e documentos associados. No entanto, observou-se que os registros fornecidos pela API pública são incompletos, sem vínculos entre licitações, contratos e empenhos, o que inviabiliza análises integradas sobre a execução das despesas públicas.

3.2. Endpoints internos do Portal de Transparência

Para contornar essa limitação, optou-se pelo uso de endpoints internos do Portal da Transparência, responsáveis por alimentar dinamicamente as tabelas exibidas na interface web. Esses endpoints, embora não documentados oficialmente, oferecem acesso a dados completos e estruturados, incluindo licitações, contratos e empenhos vinculados.

O processo iniciou-se com o envio de requisições HTTP do tipo POST aos endpoints de listagem, obtendo identificadores internos (*IDs*) de cada registro. Em seguida, novos POSTs foram realizados aos endpoints de detalhe, retornando informações completas sobre licitações³ e contratos⁴.

A automatização foi inicialmente implementada como um *script* em Python, possibilitando a extração sistemática e integrada das informações, que foram posteriormente transformadas e armazenadas para uso na análise dos resultados.

4. Processo de ETL

O processo de extração, transformação e carga (ETL) foi implementado com o uso do Apache Airflow, que orquestra as tarefas por meio de uma DAG dedicada, garantindo reprodutibilidade, escalabilidade e tolerância a falhas. O código está disponível em repositório público⁵.

A Figura 1 mostra o encadeamento das tarefas da DAG visualizado no modo *graph view* do Airflow, em ordem de execução de cima para baixo.

4.1. Extração

A etapa de extração combina dados obtidos da API pública e dos endpoints internos do Portal da Transparência da PMF, permitindo maior completude e consistência das informações. As funções principais, `get_dados_api()` e `get_dados_internos()`, recuperam, respectivamente, registros resumidos e detalhados de licitações e contratos em formato JSON. As requisições POST foram adaptadas para obter informações complementares, como unidades gestoras, fornecedores, itens, vencedores e documentos associados. Uma terceira função, `get_dados_api_empenho()`, foi testada para extrair dados de execuções orçamentárias, mas não integrou o pipeline por inconsistências nos vínculos entre empenhos, licitações e contratos.

4.2. Transformação

Após a coleta, os dados passam por tratamento e normalização para adequação ao modelo relacional no PostgreSQL. As funções `juntar_dados_licitacao()` e

³Disponível em: <https://transparencia.e-publica.net/epublica-portal/rest/florianopolis/compras/licitacao/form>

⁴Disponível em: <https://transparencia.e-publica.net/epublica-portal/rest/florianopolis/compras/contrato/form>

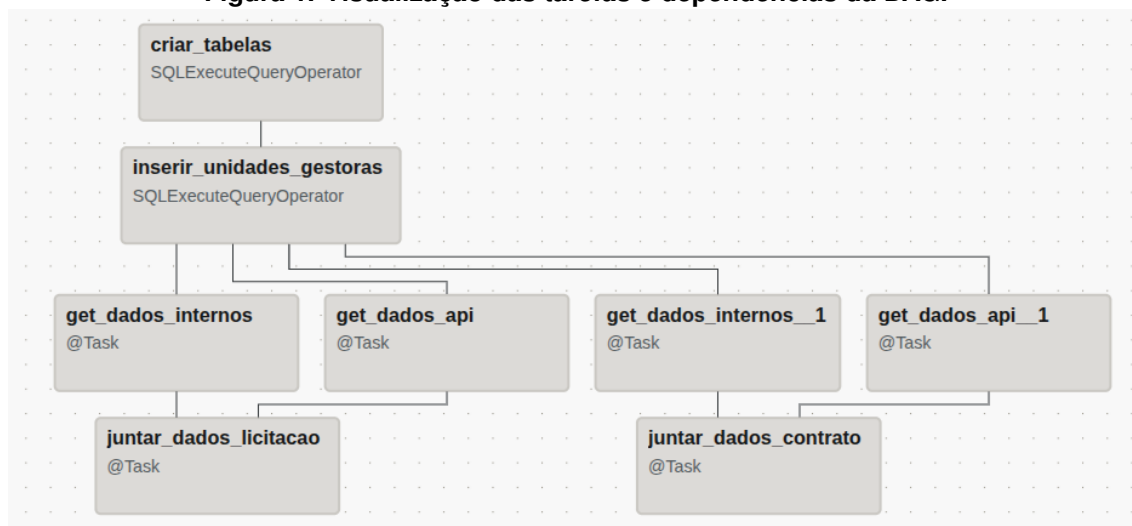
⁵Disponível em: <https://github.com/gustavodesalles/airflow-pmf>

`juntar_dados_contrato()` consolidam as informações obtidas em registros únicos, relacionando-os com entidades auxiliares e executando cálculos derivados, como o valor total licitado. A inserção é realizada de forma idempotente, utilizando a cláusula `ON CONFLICT` do PostgreSQL, o que evita duplicações e assegura a integridade dos dados.

4.3. Carga

Na fase final, as tarefas de carga criam as tabelas necessárias e inserem os dados consolidados nas entidades principais: `Licitacao`, `Contrato`, `Fornecedor`, `Item`, `Unidade Gestora`, `Advogado`, `Texto` e `Empenho`. A DAG (`dag_compras_publicas_v2`) é agendada para execução diária, automatizando a coleta e atualização contínua do banco de dados.

Figura 1. Visualização das tarefas e dependências da DAG.



4.4. Desafios e considerações

A principal dificuldade envolveu a extração via *endpoints* internos, que exigiram reconstrução manual de *payloads* POST e tratamento de respostas assíncronas (HTTP 202). Além disso, foram identificadas inconsistências nos dados dos empenhos vinculados às licitações e aos contratos.

Como perspectiva de evolução, prevê-se integrar modelos de linguagem de grande porte (LLMs) ao fluxo de ETL, a fim de interpretar registros em diferentes formatos e enriquecer semanticamente os dados, classificando naturezas de despesas, identificando relações entre entidades e detectando anomalias de forma automática.

5. Resultados

A análise exploratória dos dados de licitações e contratos administrativos da Prefeitura Municipal de Florianópolis (PMF), entre 2014 e 2024, permitiu identificar padrões relevantes de comportamento e inconsistências cadastrais. O conjunto de dados abrange informações sobre modalidades, valores contratados, prazos, unidades gestoras e fornecedores.

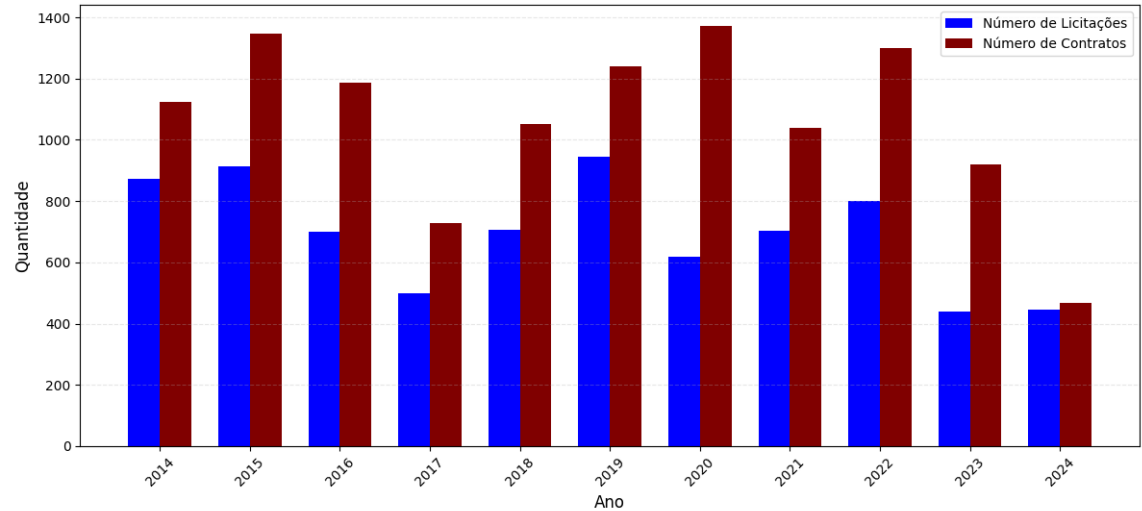
Tabela 2. Resumo das modalidades de licitações entre 2014 e 2024

Modalidade	Valor total (em R\$)	No. licitações	No. itens	No. textos
Pregão Eletrônico	15507026111.65	2063	31258	2831
Dispensa	15476798797.17	1345	2227	1598
Convite	6486691903.41	87	43	43
Concorrência	1199136337.31	302	271	403
Pregão Presencial	487201833.11	540	3688	885
Tomada de Preço	316345806.41	521	402	313
Inexigibilidade	228545024.38	953	609	1761
Outros	213741098.32	248	556	161
Leilão	49320025.48	3	162	55
Credenciamento	29404786.46	29	154	64
Concurso	1516107.35	1	304	1
Pregão	0	1553	0	0

As modalidades Pregão Eletrônico e Dispensa concentraram os maiores valores totais, superando R\$ 15 bilhões cada, seguidas por Convite e Concorrência. Observou-se, contudo, que 1.553 registros da modalidade Pregão não possuem valores, itens ou textos descritivos, indicando falhas na base de dados.

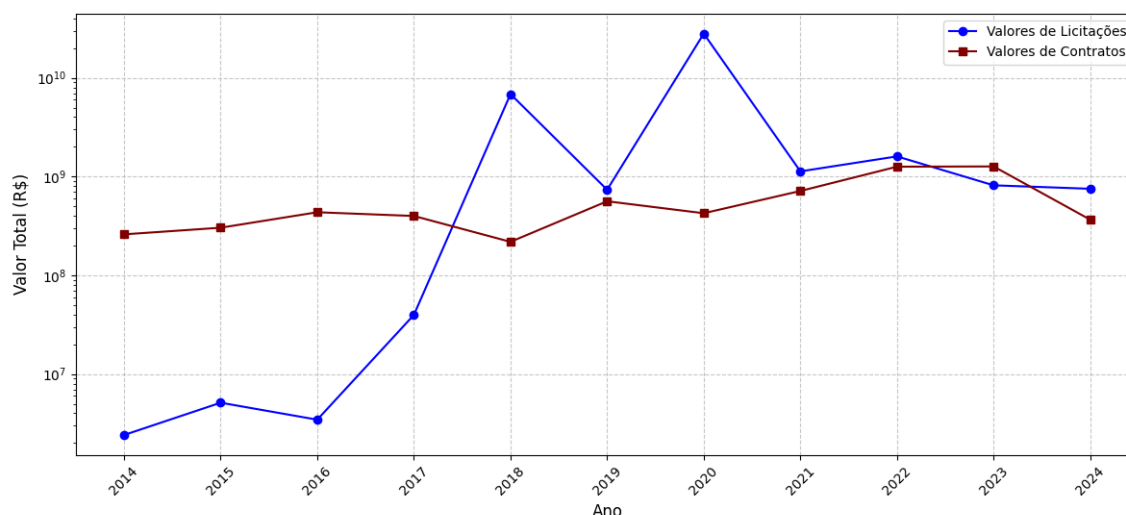
A comparação entre licitações e contratos revelou oscilações significativas ao longo da série histórica. Como mostra a Figura 2, houve queda acentuada em 2016–2017, possivelmente associada ao ciclo eleitoral municipal, e uma discrepância marcante em 2020, quando o número de contratos mais que dobrou o de licitações, indício de respostas emergenciais durante a pandemia de COVID-19. Em 2024, verificou-se maior equilíbrio entre ambos, sugerindo aprimoramento nos processos de planejamento e execução das compras públicas.

Figura 2. Quantidade total de licitações e contratos entre 2014 e 2024.



A Figura 3 apresenta a variação anual dos valores totais de licitações e contratos registrados entre 2014 e 2024, representados em escala logarítmica para melhor visualização das discrepâncias de ordem de grandeza.

Figura 3. Variação anual de valores totais de licitações e contratos entre 2014 e 2024.



No período inicial, de 2014 a 2017, observa-se que os valores de contratos se mantêm consistentemente superiores aos valores de licitações. Esse comportamento pode ser explicado pelo fato de que muitos contratos são firmados a partir de licitações de anos anteriores ou por contratações diretas, previstas na legislação, que não necessariamente guardam proporcionalidade imediata com os valores licitados no mesmo exercício.

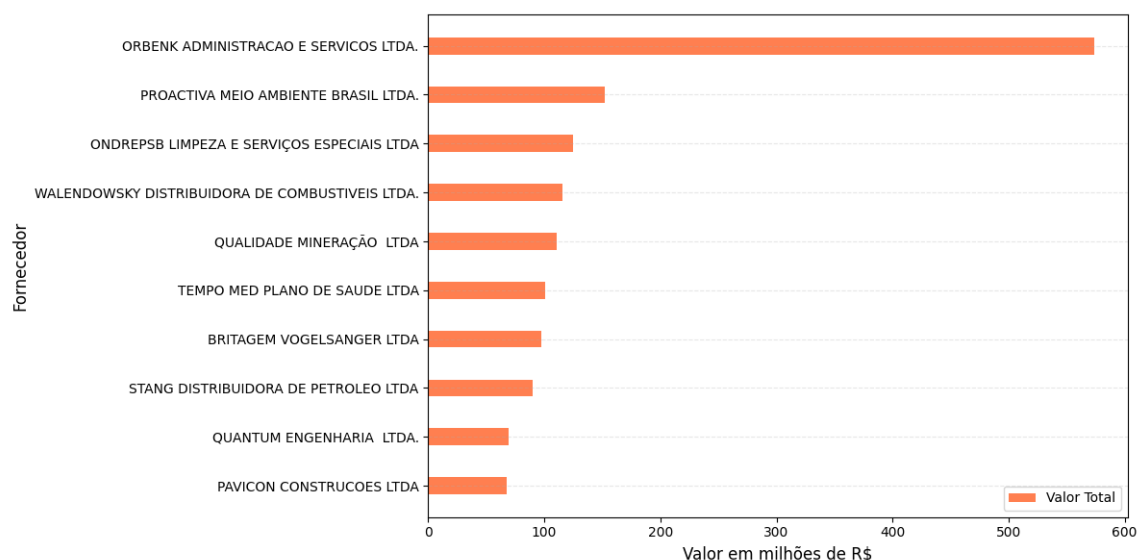
Em 2018, há um crescimento abrupto nos valores de licitações, ultrapassando em larga escala os valores de contratos. Essa discrepância pode estar associada a grandes editais lançados naquele ano, com expressiva previsão orçamentária, mas que não necessariamente se converteram no mesmo volume de contratos firmados de imediato.

Após um aparente equilíbrio no ano de 2019 entre os valores de licitações e contratos, em 2020 observa-se novamente uma disparidade, com os valores licitados alcançando seu ponto mais alto em toda a série, superando em várias ordens de magnitude os valores contratados. Esse comportamento pode estar relacionado a processos emergenciais desencadeados pela pandemia de COVID-19, em que foram anunciados grandes volumes de recursos licitados, mas cuja execução contratual pode ter ocorrido em montantes menores ou em fases posteriores.

A partir de 2021, há uma redução gradual da diferença entre licitações e contratos, atingindo um equilíbrio entre os valores de 2022 em diante. De 2023 até 2024, nota-se uma queda em ambos os índices, ainda que um pouco mais acentuada para os contratos. Essa retração pode refletir tanto um cenário de restrições orçamentárias quanto o processo de adaptação administrativa à plena vigência da nova Lei de Licitações e Contratos (Lei no. 14.133/2021), que se tornou a única lei vigente para licitações e contratações públicas em 2024.

A Figura 4 apresenta os dez fornecedores com maior valor financeiro total de contratos administrativos firmados com a Prefeitura durante o período analisado, expressos em milhões de reais.

Figura 4. Valores totais gastos com fornecedores de contratos entre 2014 e 2024.



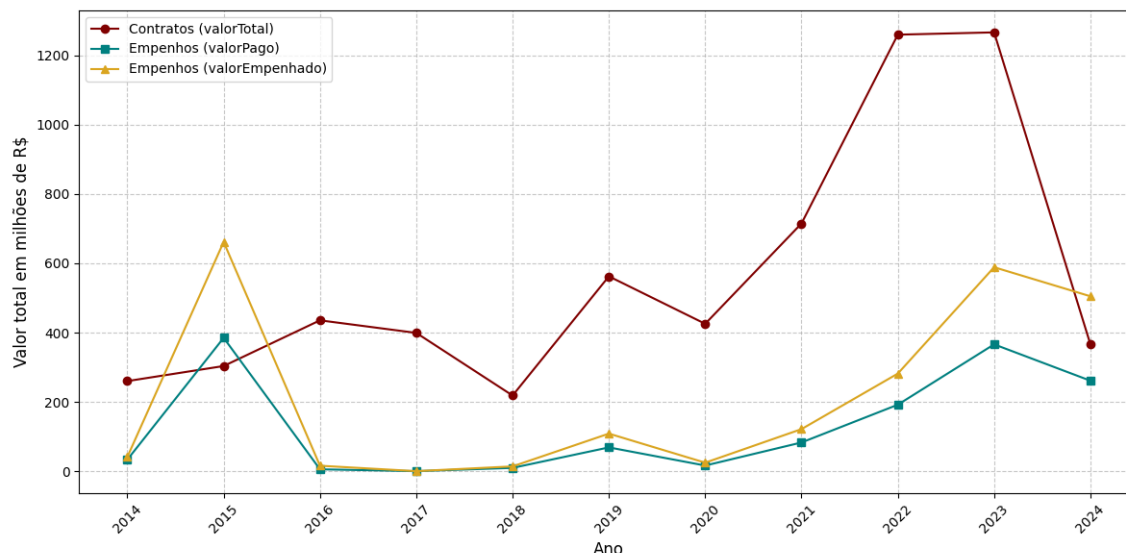
A análise evidencia uma forte concentração de recursos em um número reduzido de fornecedores. Em especial, destaca-se a empresa Orbenk Administração e Serviços Ltda., cujo montante contratado ultrapassa R\$ 550 milhões, valor significativamente superior ao das demais. Essa predominância, ao lado de empresas como Ondrepsb Limpeza e Serviços Especiais Ltda., indica que os serviços de vigilância patrimonial e limpeza urbana representam uma parcela expressiva das despesas contratuais do município.

Entre os demais fornecedores com maiores volumes de recursos, predominam aqueles ligados ao setor de obras públicas, como saneamento básico (Proactiva Meio Ambiente Brasil Ltda.), pavimentação de ruas e estradas (Qualidade Mineração Ltda., Britagem Vogelsanger Ltda., Pavicon Construções Ltda.) e iluminação pública (Quantum Engenharia Ltda.), além de empresas de distribuição de combustíveis (Walendowsky Distribuidora de Combustíveis Ltda. e Stang Distribuidora de Petróleo Ltda.). A única exceção nesse conjunto é a Tempo Med Plano de Saúde Ltda., responsável pela oferta de planos de saúde aos servidores municipais.

Por fim, a comparação entre valores de contratos e empenhos, representada na Figura 5, indicou que os primeiros são sistematicamente superiores, mas com discrepâncias acentuadas em 2015 e no período de 2021 a 2023. Também foram identificadas duplicidades de contratos com o mesmo número vinculados a diferentes unidades gestoras, resultado de limitações na modelagem do objeto *Contrato* na API pública da PMF.

Em síntese, os resultados evidenciam avanços na extração, integração e análise dos dados públicos, mas também revelam inconsistências estruturais que comprometem a confiabilidade e dificultam análises detalhadas sobre a eficiência e a execução orçamentária das contratações públicas.

Figura 5. Evolução anual de valores de contratos e empenhos vinculados entre 2014 e 2024.



6. Conclusões

O processo de ETL apresentado neste trabalho permitiu automatizar as etapas de extração, transformação e carga, consolidando dados provenientes de fontes distintas em um banco de dados relacional PostgreSQL. A DAG desenvolvida realizou com sucesso a integração entre dados públicos via API e registros internos obtidos por requisições específicas, a normalização e tratamento dos dados para garantir a consistência, e a criação de uma base relacional estruturada capaz de suportar consultas analíticas.

Como continuidade deste trabalho, sugerem-se as seguintes linhas de evolução: (i) ampliar o pipeline para suportar maiores volumes de dados e múltiplas fontes governamentais; (ii) integrar o banco de dados com ferramentas de *business intelligence* para geração de *dashboards* e relatórios; (iii) explorar o uso de algoritmos de aprendizado de máquina para detecção de anomalias em licitações e contratos; (iv) aprofundar a coleta e integração de informações relacionadas a empenhos, permitindo maior detalhamento na análise do ciclo completo das despesas públicas; e (v) implementar mecanismos para revisar e atualizar registros antigos quando houver alterações ou complementações publicadas em períodos posteriores.

Referências

- dos Santos, E. S., dos Santos, M. M., Castro, M., and Carvalho, J. T. (2024). Performance variability of machine learning models using limited data for collusion detection: A case study of the brazilian car wash operation. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 431–443, Porto Alegre, RS, Brasil. SBC.
- Jesus, M. B. d. (2021). Modelo preditivo de risco de irregularidades em compras públicas no estado de goiás.
- Mello, R., Bornia, L. G., da Silva, H., Pierotti, F., and Cardeal, M. (2024). Uma proposta de modelagem de dados no domínio de fraudes em licitações públicas. In *Anais Es-*

tendidos do XXXIX Simpósio Brasileiro de Bancos de Dados, pages 220–226, Porto Alegre, RS, Brasil. SBC.

Santos, M. T. S., Gasparini, I., Frigo, L. B., and Dalle Mülle, L. d. O. (2021). Ferramenta de visualização de dados abertos do portal de transparência da câmara municipal da cidade de florianópolis. In *Workshop de Computação Aplicada em Governo Eletrônico (WCGE)*, pages 71–82. SBC.

Schmitz, F., Ferrão, L., dos Santos, M. M., Castro, M., and Carvalho, J. T. (2024). Detecting fraud in public procurement: A gmm-based approach to analyzing tender data. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 207–219, Porto Alegre, RS, Brasil. SBC.

APÊNDICE B – CÓDIGO FONTE

Os códigos desenvolvidos para obter os dados das APIs do Portal de Transparência da Prefeitura Municipal de Florianópolis e gerar os gráficos apresentados neste trabalho estão disponíveis nos repositórios em <https://github.com/gustavodesalles/INE5454-trabalho> e <https://github.com/gustavodesalles/licitacoes-contratos-fln>.

O código do *workflow* implementado no Airflow está disponível em <https://github.com/gustavodesalles/airflow-pmf>.