



PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

Pós-graduação *Lato Sensu* em Analytics e Business Intelligence

RELATÓRIO TÉCNICO

ANÁLISE DOS DADOS DE DESPESAS PÚBLICAS DO GOVERNO DE MINAS GERAIS

Lucas Lukasavicus Silva

São Paulo

2023

Sumário

Resumo.....	3
1. Introdução.....	4
1.1. Contexto.....	4
1.2 Objetivos	4
2. Revisão Bibliográfica	6
3. Metodologia	9
3.1. Modelagem de Dados	9
3.1.1. Desafio	9
3.1.2. Escopo.....	9
3.1.3. Modelo de Dados.....	10
3.1.4. Solução.....	28
3.2. Engenharia de Dados.....	29
3.3. Ciência de Dados.....	35
3.3.1 Modelo de Forecast.....	36
3.3.2 Modelo de Regras de Associação.....	37
3.3.3 Modelo de Detecção de Anomalia	39
3.4. Visualização de Dados.....	41
4. Homologação e Resultados	51
4.1. Homologação	51
4.2. Resultados	56
5. Conclusões e Próximos Passos.....	62
5.1. Conclusões	62
5.2. Próximos Passos.....	62
6. Anexos.....	62
7. Referencias.....	64

Resumo

O presente trabalho tem como objetivo demonstrar as habilidades adquiridas ao longo do curso de *Business Intelligence e Analytics* da Pontifícia Universidade Católica (PUC) de Minas Gerais. Para isso, foi proposto o desenvolvimento de uma solução que implementasse atividades de *Discovery* de dados, a construção de um Painel de Controle com níveis estratégico, tático e operacional e a aplicação de Análises Avançadas (modelos de *machine learning*). Com esse objetivo selecionamos o tema de Gestão de Despesas Públicas do estado de Minas Gerais. Esses dados foram coletados, tratados, ingeridos e armazenados em diferentes tecnologias de armazenamento utilizando um framework moderno de computação distribuída em nuvem. Utilizamos a ferramenta *PowerBI* para a construção do painel de visualização e implementamos três modelos de *machine learning*. Todo trabalho foi homologado para atestar a qualidade dos dados e o seu correto processamento e importantes conclusões e insights foram obtidos a partir dos dados. Realizamos também uma pesquisa preliminar e um levantamento bibliográfico das pesquisas na área de gestão de dados públicos e transparência governamental.

1. Introdução

1.1. Contexto

Com a arrecadação e o investimento público de recursos sendo as principais atividades das áreas econômicas de diferentes esferas do governo, cada vez mais se faz necessário ter uma boa gestão dos recursos. Para tal gestão ser efetiva, é necessário adotar métodos que fazem uso da tecnologia para tal, e por conta do volume crescente de dados, soluções que façam uso de frameworks e ferramentas de Big Data têm sido amplamente difundidas (1). Outro ponto relevante é a Lei de Responsabilidades Fiscais (16) que determina políticas para o controle dos gastos públicos, por conta disso, não só é um ponto importante de ganho, no sentido de maior gestão das contas públicas, como também a questão do atendimento à legislação.

O Governo Brasileiro pertence a uma seleta lista de membros-fundadores do *Open Government Partnership*, associação que tem por principais objetivos promover uma política de transparência de dados públicos através da divulgação de dados dos mais variados setores públicos (educação, saúde, economia etc.), difundir através da tecnologia maior abertura, transparência e responsabilidade dos órgãos públicos, suportar a participação cível através de programas específicos entre outros (11). Tendo em vista as políticas que constituem o Acordo, o Brasil no ano de 2012 criou o Portal de Dados Abertos do Governo Brasileiro, um website que possuía mais de 240 datasets e mais de 2100 recursos digitais. Os dados presentes no portal são de diferentes autarquias do governo (federal, estadual e municipal), e foram a partir dos dados do Governo de Minas Gerais que realizamos o desenvolvimento desse trabalho.

Esse trabalho está dividido da seguinte maneira: Na seção 1 temos a introdução e contextualização da pesquisa. A seção 2 contém a revisão bibliográfica realizada para embasar o trabalho. A seção 3 detalha a metodologia desse trabalho dividindo as áreas do conhecimento abordadas (engenharia, ciência e visualização de dados). Na quarta seção temos os registros de homologação (componente obrigatório para o desenvolvimento desse trabalho) e os resultados obtidos após as análises. Na quinta seção discorremos sobre a conclusão e os próximos passos que sugerimos para dar continuidade a essa pesquisa. As duas últimas seções apresentam os anexos e as referências utilizadas no trabalho.

1.2 Objetivos

Os objetivos do presente trabalho são:

- Demonstrar o fluxo de Engenharia de Dados *end-to-end*, aplicando métodos para a extração, transformação e carga dos dados.
- Aplicar modelos de Ciência de Dados robustos que agreguem valor às análises desenvolvidas ao longo da pesquisa;
- Desenvolver e apresentar visualizações que permitam responder às perguntas de eventuais partes interessadas (sociedade civil, controladores de gastos, pessoas de áreas relevantes do governo e demais interessados).
- Discorrer sobre o tema de controle de gastos públicos e da necessidade da implantação, com ajuda da tecnologia, de mecanismos de gestão.

A principal motivação do desenvolvimento desse trabalho é a promoção da transparência dos gastos públicos. Nesse sentido temos por objetivos estratégicos criar uma plataforma de dados automatizada, de alta-performance e completa com os dados que temos à disposição para poder habilitar diferentes públicos, como a sociedade, órgãos governamentais, auditores entre outros possam realizar o acompanhamento dos indicadores de gastos e investimentos públicos no estado de Minas Gerais avaliando o desempenho do governo quanto às políticas públicas prometidas e cumpridas através do empenho desses valores. Vale destacar que apesar de apresentar uma linguagem simplificada, esse trabalho por ser de uma área acadêmica bastante técnica, tem como público-alvo pessoas que tenham familiaridade com tecnologia, principalmente ferramentas de *Business Intelligence*.

2. Revisão Bibliográfica

Diversos trabalhos têm mostrado a importância da adoção de tecnologia por parte dos governos ((2),(3),(4)). A incorporação de ferramentas e metodologias tecnológicas tem ajudado governos a promover mais transparência que é um tema central dentro de novas políticas modernas. A transparência pode ser vista de acordo com (5) ocorrendo através de: uma postura proativa do governo, divulgação de dados e materiais do governo, reuniões e assembleias públicas ou vazamento de informação através de agentes internos. Um estudo em (6) também mostrou que países que implementam políticas públicas de transparência tem 3 vezes mais chance de fornecer dados quando algum órgão ou entidade requisita. Essa nova interface entre a sociedade-e-governo criada a partir do desenvolvimento de plataforma de acesso aos dados permite cada vez mais a participação ativa da sociedade no governo e assim o desenvolvimento democrático cada vez maior (8), além de promover ações concretas de anticorrupção (9).

Uma vez que tem se mostrado crítico o desenvolvimento dessas soluções (10), o processo de democracia atualmente permeia o desenvolvimento tecnológico, sendo assim, é de suma importância que governos trabalhem para a correta disponibilização dos dados públicos.

Como parte dessas iniciativas de tecnologia de dados para maior transparência nos governos, uma das mais proeminentes ações foi em 2011 a criação da *Open Government Partnership*. Essa associação que atualmente conta com mais de 70 países e 100 governos locais, já estabeleceu e implementou mais de 4000 compromissos que vão desde inclusão, ações de anticorrupção e integridade, justiça social, saúde e educação até clima e meio ambiente (11). Os compromissos assinados através dessa parceria, se concretizam planos de ação com duração de 3 anos. No Brasil, o órgão responsável pela interface com a OGP é a Controladoria Geral da União (CGU). A CGU nos mais de 10 anos de participação da parceria já desenvolveu 5 Planos de Ação, onde podemos destacar os pontos conforme a tabela a seguir:

Plano	Órgãos Envolvidos	Principais ações	Principais Desafios	Quantidade de Compromissos Cumpridos
1 (2011-2013)	5 (CGU, Ciência e tecnologia, Educação, Planejamento, Orçamento e gestão)	Pesquisas diagnósticas sobre conhecimento de tecnologias de acesso à informação, Elicitação de demandas, Cartilhas sobre acesso à informação, Reestruturação do Portal da Transparência, Capacitação técnica e Desenvolvimento de sistemas de informação etc.	Integridade pública, Gestão efetiva de recursos e Melhoria na prestação de serviços públicos	26 / 32
2 (2013-2016)	11 (AGU, CGU, MEC, MI, MS, MDA, SeGov, MD, MJC, MP, MDS, MC)	Enriquecimento dos dados divulgados, Prestação de contas, Painéis online, Desenvolvimento de sistemas nos setores agro, educação e saneamento, Dados educacionais	Gestão mais efetiva de recursos públicos, Aumento da integridade pública, Melhoria dos serviços públicos, aumento da responsabilidade	36 / 52

		abertos, Proposta de dados abertos, Reestruturação do portal de dados abertos, Sistema eletrônico para consultas públicas etc.	corporativa, criação de comunidades mais seguras	
3 (2016 – 2018)	(5) CGU, CC, MF, MPDG, MRE, MJC + Órgãos civis	Elaboração de planos e matrizes com metas tangíveis nas áreas de transparência, participação social e segurança. Disponibilização de dados do governo através de meios digitais, Consolidação de redes abertas de forma colaborativa para participação da sociedade civil.	Temas estruturantes do governo aberto, Garantia de direitos, Inovação e melhoria do serviço público, rumo a um estado aberto.	8 / 16 (89%) *
4 (2018-2021)	(5) CGU, CC, MF, MPDG, MRE, MJC + Órgãos civis	Desenvolvimento do Plano do Governo Aberto em Estados e Municípios com dados de Inovação e Ciência, Clima e Recursos hídricos, Transparência do processo legislativo, fundiário e implantação de sistemas específicos (como o de controle do processo de reparação de Mariana e municípios da região)	Estruturante, priorizado pelo governo e priorizado pela sociedade civil	6 / 17 (88%) *
5 (2021 – vigente)	(5) CGU, CC, MF, MPDG, MRE, MJC + Órgãos civis	Ampliação da plataforma do Governo Aberto nos temas de meio ambiente, maus-tratos contra animais, direitos humanos, vigilância sanitária, licenciamento ambiental, dados eleitorais e cadeia agropecuária, Criação de programas de combate a corrupção no setor público e transparência de imóveis públicos federais	Priorizado pelo governo, priorizado por outros poderes e priorizado pela sociedade civil	2 / 12 (72%) *

Siglas:

- CGU – Controladoria-Geral da União;
- MF - Ministério da Fazenda;
- MPDG – Ministério do Planejamento, Desenvolvimento e Gestão,
- MRE – Ministério das Relações Exteriores;
- MJC – Ministério da Justiça e Cidadania;
- AGU – Advocacia Geral da União;
- MEC – Ministério da Educação;
- MP – Ministério Público;
- MD – Ministério da Defesa
- MDS – Ministério do Desenvolvimento Social e Agrário- Previdência
- SeGov – Secretaria de Governo da Presidência;
- MI – Ministério da Integração Nacional
- MS – Ministério da Saúde;
- MC – Ministério da Ciência, Tecnologia e Inovação
- MDA - Secretaria Especial de Agricultura Familiar e Desenvolvimento Agrário

* Obs.: As porcentagens dadas se referem a completude do projeto como um todo, somando a conclusão parcial de cada iniciativa.

Nesse contexto, dado a quantidade de inovação que o Governo Brasileiro vem implementando nos últimos anos, cada vez mais se faz necessário a utilização de ambientes de dados que abarquem a quantidade de dados e as suas respectivas demandas de velocidade, integridade, segurança entre outras. Para isso foi desenvolvido o Portal de Dados Abertos (13) como parte do resultado da Lei de Acesso à Informação (LAI) de 2011 (15).

Esse portal web segue o padrão da Open Knowledge Foundation que determina regras sobre a disponibilização dos dados, acesso, formato, transparência e responsabilidade. Todos os dados divulgados através do portal seguem a Instrução Normativa da Infraestrutura Nacional de Dados Abertos(14). No Portal são disponibilizadas mais de 12.000 bases de dados de 235 organizações públicas e privadas das áreas de: Agricultura, Assistência e Desenvolvimento Social, Ciência e Tecnologia, Comercio, Serviços e Turismo, Cultura, Lazer e Esporte, Dados estratégicos, dados de segurança, economia e finanças, educação, energia, equipamentos públicos, governo e política, habitação, saneamento e urbanismo indústria, saúde, transporte, trabalho entre outros.

Em 2012, o Governo de Minas Gerais atendendo aos critérios da Lei de Acesso à Informação criou o seu próprio Portal da Transparência (16) com os dados relativos à despesa, receita, dívida pública e remuneração dos servidores públicos. Desde a época até os dias atuais, o portal da transparência do governo de Minas Gerais evoluiu adotando as melhores práticas de mercado, aderindo a padrões internacionais de dados como o OKF e o *Frictionless Data*, incluindo cada vez mais bases de dados e mais recentemente, em 2021, implementando ações de anonimização e segurança como parte da política defendida pela Lei Geral de Proteção de Dados. Atualmente conta com 28 bases de dados de 8 órgãos públicos do Estado em 4 principais categorias: Administração, Saúde, Segurança e Ordem Pública e Proteção Social. A presente pesquisa utilizou os dados de Despesa Pública da Controladoria Geral do Estado de Minas Gerais (17). Esse conjunto de dados tem 61 arquivos em formato CSV. As suas características, relacionamentos e descrições detalhadas estão presentes na próxima seção desse trabalho.

A integração de sistemas de computador em práticas de negócios modernas impactou significativamente a maneira como as organizações gerenciam e analisam dados para obter insights valiosos. A inteligência de negócios envolve a coleta, armazenamento e análise de dados de várias fontes, e os sistemas de computador desempenham um papel vital no suporte a esses processos ((23),(24),(25)).

Tendo em vista as aplicações modernas que podemos construir atualmente uma das maiores preocupações das empresas tem sido com relação aos dados ((26) (27)). Por conta disso é cada vez mais importante olhar de forma estratégico as áreas que lidam com os dados das empresas, se atentar para os processos de transformação de dados em informações e do uso dessas informações para geração de conhecimento e tomada de decisão.

As áreas-chave que mais trabalham com dados são as áreas de engenharia, ciência, visualização (analytics / BI) e operação (DataOps). As atividades comuns dessas áreas vão desde a implementação de pipelines de dados (com extração, transformação e carga de dados) para diferentes camadas de refinamento e consumo de dados (engenharia de dados), passando por criação de modelos prescritivos e preditivos (ciência de dados), desenvolvimento de painéis de gestão de informação, acompanhamento de indicadores, relatórios e análises customizadas (análise de dados), e a manutenção em si de toda a plataforma de dados, banco de dados, servidores, gestão de recursos de nuvem e etc. (operação de dados).

Esse trabalho visa implementar todo esse processo demonstrando uma aplicação de dados real a partir dos dados de despesa pública do governo de Minas Gerais obtidos através do Portal da Transparência do governo.

3. Metodologia

3.1. Modelagem de Dados

O processo de modelagem de dados passa por entender o problema, definir e limitar um escopo, desenhar uma solução inicial, validar a solução proposta e implementá-la. Sendo um processo iterativo e interativo. No presente trabalho, simulamos esse cenário com o entendimento do problema sendo definido pela instituição de ensino PUC Minas Gerais, a definição do escopo do problema como sendo o a coleta e tratamento de dados da despesa pública do governo de MG, aplicação de modelos de *machine learning* a esses dados e a exibição dos dados através da ferramenta de analytics PowerBI. A solução proposta foi implementada usando a tecnologia de computação em nuvem da *Amazon AWS* e *Microsoft PowerBI Cloud*.

3.1.1. Desafio

Para esse projeto foram propostos 8 itens obrigatórios divididos em 3 temas de acompanhamento do desenvolvimento do trabalho (30). Esses itens foram agrupados, nesse trabalho, nos seguintes tópicos:

1. **Projeto de Dados:** Integração e tratamento de dados (ETL) e estrutura e volume da base de dados – Durante o trabalho devemos usar um conjunto de dados com volumetria considerável (mínimo de 10 dimensões e 1 fato) com histórico de 24 meses e usar ferramentas capazes de realizar o processo de *ETL*. A modelagem e a construção das bases de dados devem permitir ser capaz de realizar o processo de exploração de dados de forma ágil e flexível.
2. **Visualização e Painel de Controle:** A construção do painel de controle (dashboard) deve permitir realizar diferentes tipos de análises usando gráficos diversos (com uma ou mais variáveis categóricas e numéricas), realizar filtros e operações de *drill-down*, *slice and dice* nos níveis estratégico, tático e operacional, integradas e complementares. A solução apresentada também deve ser performática (timeout para apresentação dos dados após filtros aplicados de no máximo 10 segundos).
3. **Análises Avançadas e Machine Learning:** Desenvolver modelos de Machine Learning que sejam capazes de responder a perguntas críticas de negócio e sejam incorporados ao Painel de Controle.
4. **Documentação e gestão de ativos:** Documentar todos os processos, etapas, ferramentas e transformações realizadas nos dados, desde a camada de dados até a camada de apresentação, agregando explicações quando necessário, apresentando uma visão geral do cenário e escopo tratados, mostrando os resultados e processos de homologação e análise e engajando stakeholders e demais interessados no processo de desenvolvimento da solução de Business Intelligence e Analytics proposta.

3.1.2. Escopo

Para atender os critérios definidos pela PUC nos propusemos a analisar os dados de despesa pública do governo de Minas Gerais. Esses dados podem ser encontrados no site do Portal da Transparência de MG.

Portal de Dados Abertos do Estado de Minas Gerais

Entrar

Conjuntos de dados Organizações Grupos Documentação Sobre Fale Conosco Pesquisar

Pesquisar dados

Ex: meio ambiente

Etiquetas populares Portal da Transparéncia coronavírus COVID-19

Portal de Dados Abertos do Estado de Minas Gerais estatísticas

28 conjuntos de 8 organizações 4 grupos dados

Transparéncia Ativa e Dados Abertos

É dever dos órgãos e entidades públicos promover, independentemente de recursos, a disponibilização em local de fácil acesso, no âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas em formatos abertos, estruturados e legíveis por máquinas.

Gestor Público

Se você é um gestor público do Estado de Minas Gerais gestor de dados de compartilhamento amplo, aqueles que não estão sujeitos a nenhuma restrição de uso, mas que não estão catalogados no Portal de Dados Abertos, entre em contato com o seu órgão ou entidade e faça o seu cadastro.

Citado(a)

Se você é um cidadão e acredita que determinada base de dados dos órgãos e entidades deveria estar catalogada no Portal de Dados Abertos, você pode fazer uma solicitação de abertura de bases de dados diretamente ao órgão gestor de dados utilizando o e-SIC.

Se quiser tirar dúvidas ou relatar problemas nas bases de dados publicadas entre em contato pelo Fale Conosco.

Portal de Dados Abertos do Estado de Minas Gerais

Conjuntos de dados Organizações Grupos Documentação Sobre Fale Conosco Pesquisar

Conjuntos de dados

▼ Organizações

Controleadoria Geral ... Fundação Hospitalar ... Secretaria de Estado ... Secretaria de Estado ... Gabinete Militar do ... Polícia Civil - POMA ... Secretaria de Estado ... Secretaria de Estado ...

Search datasets... 🔍

28 conjuntos de dados encontrados

Ordenar por: Relevância

Proposta Orçamentária e Lei Orçamentária

Dados Abertos sobre Proposta Orçamentária e Lei Orçamentária. Esse conjunto de dados, documentado de acordo com o padrão da metadados Friccioneis, corresponde ao modelo.

CSV **JSON**

Convênios de entrada de recursos

Dados Abertos sobre convênios de entrada de recursos celebrados pelo Estado com o governo federal, municipais e organizações da sociedade civil para execução de política pública.

HTML **CSV**

Recadastramento

Dados Abertos sobre recadas previstas e efetivamente arrecadadas pelo Estado, como tributos, operações de crédito, alienação de bens e outras fontes. Esse conjunto de dados...

Imagens 1 e 2 – Imagens do site do Portal da Transparéncia, de onde os dados foram coletados.

Despesa pública

Sugestões
0

Organização



Controladoria Geral do Estado - CGE

A CGE, órgão central do sistema de controle interno do Poder Executivo, tem como função assessor diretamente o Governo do RJ no desempenho de suas atribuições que usam asto.

Leia mais

Social

Twitter

Facebook

Licença

CC-BY-4.0

Conjunto de dados **Grupos** **Fluxo de Atividades**

Despesa pública

Dados Abertos sobre despesas empenhadas, liquidadas e pagas pelo Estado ano a ano.

Esse conjunto de dados, documentado de acordo com o padrão de metadados Frictionless, corresponde ao modelo dimensional que alimenta a consulta Despesa do Portal da Transparéncia do Estado de Minas Gerais.

E é composto pelas seguintes tabelas fato (e tabelas dimensões associadas):

- `ft_despesa`
- `ft_despesa_pcto`

Como participar

Ritmo como contribuir com a documentação deste conjunto de dados

A documentação deste conjunto de dados está sendo feita de forma aberta e colaborativa no GitHub. Existem duas alternativas para enviar sua contribuição:

- `Issues`: Para iniciar uma discussão sobre melhorias na documentação.
- `Pull requests`: Para sugerir uma alteração concreta na documentação.

Todas as contribuições são bem vindas. Alguns exemplos são:

- Indicação de expressões impróprias presentes na documentação;
- Sugestões para inclusão de descrições em campos específicos;
- Sugestões para clareza na organização das ideias;
- Correção de erros de ortografia e gramática.

Além disso, fique à vontade para utilizar os demais canais oficiais de atendimento do Poder Executivo Estadual:

- Faie Conosco:** Dúvidas;
- Manifestações de Cúvidos:** Denúncia, Reclamação, Crítica, Exigiu ou Sugestões;
- Pedidos de Acesso à Informação:** Acesso às informações dos órgãos e entidades estaduais que não estejam publicamente disponíveis;
- Pedido de abertura de bases de dados:** Solicitação de abertura de bases de dados dos órgãos e entidades que não estejam publicamente disponíveis.

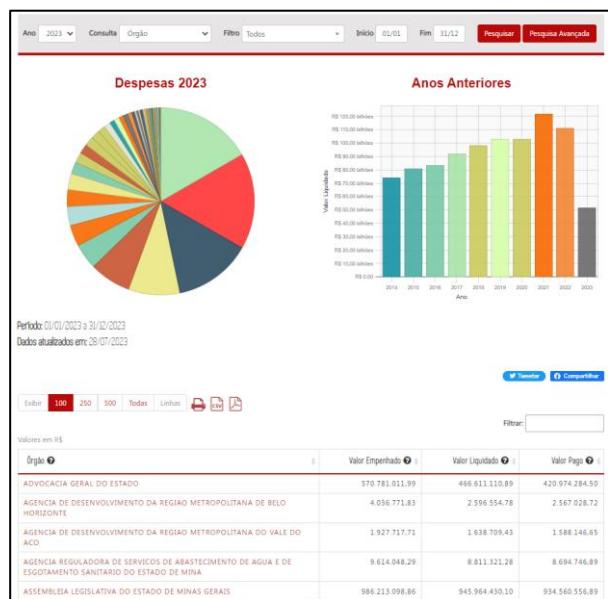
Controle de alterações

Documentam as principais alterações sofridas por este conjunto de dados.

Dados e recursos	
 Unidade Orçamentária	 Explorar
 Função	 Explorar
 Subfunção	 Explorar
 Calendário - Dias	 Explorar
 Tipo Documento	 Explorar
 Item de Despesa	 Explorar
 Programa	 Explorar
 Ação Orçamentária	 Explorar
 Empenho Despesa 2002	 Explorar
 Empenho Despesa 2003	 Explorar
 Empenho Despesa 2004	 Explorar
 Empenho Despesa 2005	 Explorar
 Empenho Despesa 2006	 Explorar
 Empenho Despesa 2007	 Explorar

Imagens 3 e 4 – Imagens do portal que mostra como os dados estão organizados e documentados.

Adicionar Filtro												
Grade	Gráfico	Mapa	44 records		1	-44						
_id	ANO	CODIGO...	ORGÃO...	CNPJ_C...	DOADO...	OBJETO...	QUANTI...	VIGÊNCIA...	VALOR	MOEDA	PROCE...	
1	2019	1501	SEPLAG	14.377.2	Instituto	Prestação	1	03 mese	300.000 BRL	1500.01	http://www...	
2	2019	1501	SEPLAG	14.377.2	Tropeia L.	Serviço	1	07 sema	310.000 BRL	1500.01	http://www...	
3	2019	1521	SEDESE	24.331.0	Improve	Prestação	1	28 sema	247.000.000 BRL	1250.01	http://www...	
4	2019	1501	SEPLAG	29.194.9	SEPLAG	Serviço	1	03 mese	15.180 BRL	1500.01	http://www...	
5	2019	1501	SEPLAG	65.461.1	Educa	Serviço	1	06 sema	5.500 BRL	1500.01	http://www...	
6	2019	1501	SEPLAG	27.685.5	Aracatu	Serviços	1	4 meses	150.000 BRL	1500.01	http://www...	
7	2019	1451	SEDESE	15.702.4	Macacá	Doação	1	30 junta	6.450 BRL	1500.01	http://www...	
8	2020	1260	SEE	13.891.7	Intendit	Doação	1	120 cen	81.440 BRL	1260.01	http://www...	
9	2020	2440	SEINPFA	30.024.0	Arco	Doação	1	60 sess	21.000 BRL	1260.01	http://www...	
10	2020	1300	SEINPFA	09.510.0	Fazitlo	Doação	1	30ثبتا	7.200 BRL	1300.01	http://www...	
11	2020	1481	SEDESE	25.078.4	Crafty Br.	Doação	1	24 vele	1.500 BRL	1400.01	http://www...	
12	2020	1501	SEPLAG	14.377.2	Instituto	Doação	1	02 dezo	290.000 BRL	1500.01	http://www...	
13	2020	1481	SEDESE	60.562.4	Junge C.	Doação	1	150cent	400.000 BRL	1480.01...	http://www...	
14	2020	1220	SEINPFA	22.923.5	Liga Ace	Doação	1	12(dize)	35.000 BRL	1220.01	http://www...	
15	2020	1300	SEINPFA	780.000	Tax do FI	Empresa	Doação	1	03 (trés)	26.253.000 BRL	1300.01...	http://www...
16	2020	1490	SEGOV	01.563.3	SEGOV	Doação	1	01(13)	81.440 BRL	1490.01	http://www...	
17	2020	1300	SEINPFA	35.755.8	Empresa	Doação	1	12(dize)	87.000 BRL	1300.01...	http://www...	
18	2020	1520	CGE	35.791.3	Empresa	Doação	1	12(dize)	24.525.000 BRL	1520.01...	http://www...	
19	2020	1300	SEINPFA	38.743.3	Empresa	Doação	1	90 nove	425.000 BRL	1300.01...	http://www...	
20	2020	1250	PMMO	26.798.7	Empresa	Doação	1	De azor	103.900 BRL	1250.01	http://www...	
21	2020	1300	SEINPFA	22.956.6	Empresa	Comenda	1	03 (Trés)	63.400 BRL	1300.01...	http://www...	
22	2020	1510	Polícia C.	04.346.5	Conselh	Comenda	1	Indetermin	116.054 BRL	1510.01	http://www...	
23	2020	1300	SEINPFA	02.846.0	CCR S.A	Doação	1	12(dize)	5.000 BRL	1300.01...	http://www...	



Imagens 5 e 6 – Imagens dos dados disponíveis no portal da transparência e de gráficos que a própria equipe de desenvolvimento do portal fez para demonstrar a capacidade dos dados.

Os dados foram então modelados da seguinte maneira.

3.1.3. Modelo de Dados

Nosso conjunto de dados original tem no total 61 tabelas, sendo elas:

- **38 tabelas dimensionais** (dm_acao, dm_categ_econ, dm_elemento_desp, dm_empenho_desp_2002, dm_empenho_desp_2003, dm_empenho_desp_2004, dm_empenho_desp_2005, dm_empenho_desp_2006, dm_empenho_desp_2007, dm_empenho_desp_2008, dm_empenho_desp_2009, dm_empenho_desp_2010, dm_empenho_desp_2011, dm_empenho_desp_2012, dm_empenho_desp_2013, dm_empenho_desp_2014, dm_empenho_desp_2015, dm_empenho_desp_2016, dm_empenho_desp_2017, dm_empenho_desp_2018, dm_empenho_desp_2019, dm_empenho_desp_2020, dm_empenho_desp_2021, dm_empenho_desp_2022, dm_empenho_desp_2023, dm_favorecido, dm_fonte, dm_funcao_desp, dm_grupo_desp, dm_item_desp, dm_modalidade_aplic, dm_procedencia, dm_programa, dm_situacao_op_desp, dm_subfuncao_desp, dm_tempo_diario, dm_tipo_documento, dm_unidade_orc)
- **23 tabelas fato** (fl_despesa_pgto, ft_despesa_2002, ft_despesa_2003, ft_despesa_2004, ft_despesa_2005, ft_despesa_2006, ft_despesa_2007, ft_despesa_2008, ft_despesa_2009, ft_despesa_2010, ft_despesa_2011, ft_despesa_2012, ft_despesa_2013, ft_despesa_2014, ft_despesa_2015, ft_despesa_2016, ft_despesa_2017, ft_despesa_2018, ft_despesa_2019, ft_despesa_2020, ft_despesa_2021, ft_despesa_2022, ft_despesa_2023);

Esse conjunto de tabelas-fato existe por conta da maneira como o governo decidiu disponibilizar os dados. Uma vez que a maioria das análises dos dados é feita de forma anual, os dados das tabelas-fato também seguem esse formato, assim, cada tabela-fato possui dados de um único ano. Como parte do processo de ETL decidimos alterar esse formato e criar uma única tabela-fato para o nosso conjunto de dados.

Nessa tabela-fato unificada além dos dados das dimensões também temos 3 importantes métricas: valor empenhado, valor liquidado e valor pago. Dada a importância dessas métricas, iremos explicá-las separadamente:

- **Valor Empenhado:** Valor reservado para realizar a compra ou contratação de um bem ou serviço, em geral esse valor vem de um planejamento de gastos (com exceções para gastos extraordinários). É a criação da obrigação de pagamento que deverá ser pago no futuro e é a primeira etapa da despesa pública.
- **Valor Liquidado:** Segundo estágio da despesa orçamentária, o valor liquidado é o processado por cada unidade executora quando essa recebe o objeto do empenho (a prestação do serviço, o bem, a benfeitoria, o material ou a obra). Tendo verificado os documentos comprobatórios do crédito, origem da despesa, objeto de empenho e valores a serem pagos então esse valor é liquidado podendo seguir para o fluxo de pagamento.
- **Valor Pago:** Valor é dito pago, quando o órgão público realiza a emissão do cheque, ordem bancária ou transferência em favor do credor (prestador do serviço ou comércio do bem), essa é a última etapa da despesa orçamentária.

A imagem a seguir mostra o Diagrama Entidade-Relacionamento do Modelo de Dados. Nesse diagrama é possível ver que a modelagem de dados seguiu o modelo ***Star-Schema*** (40).

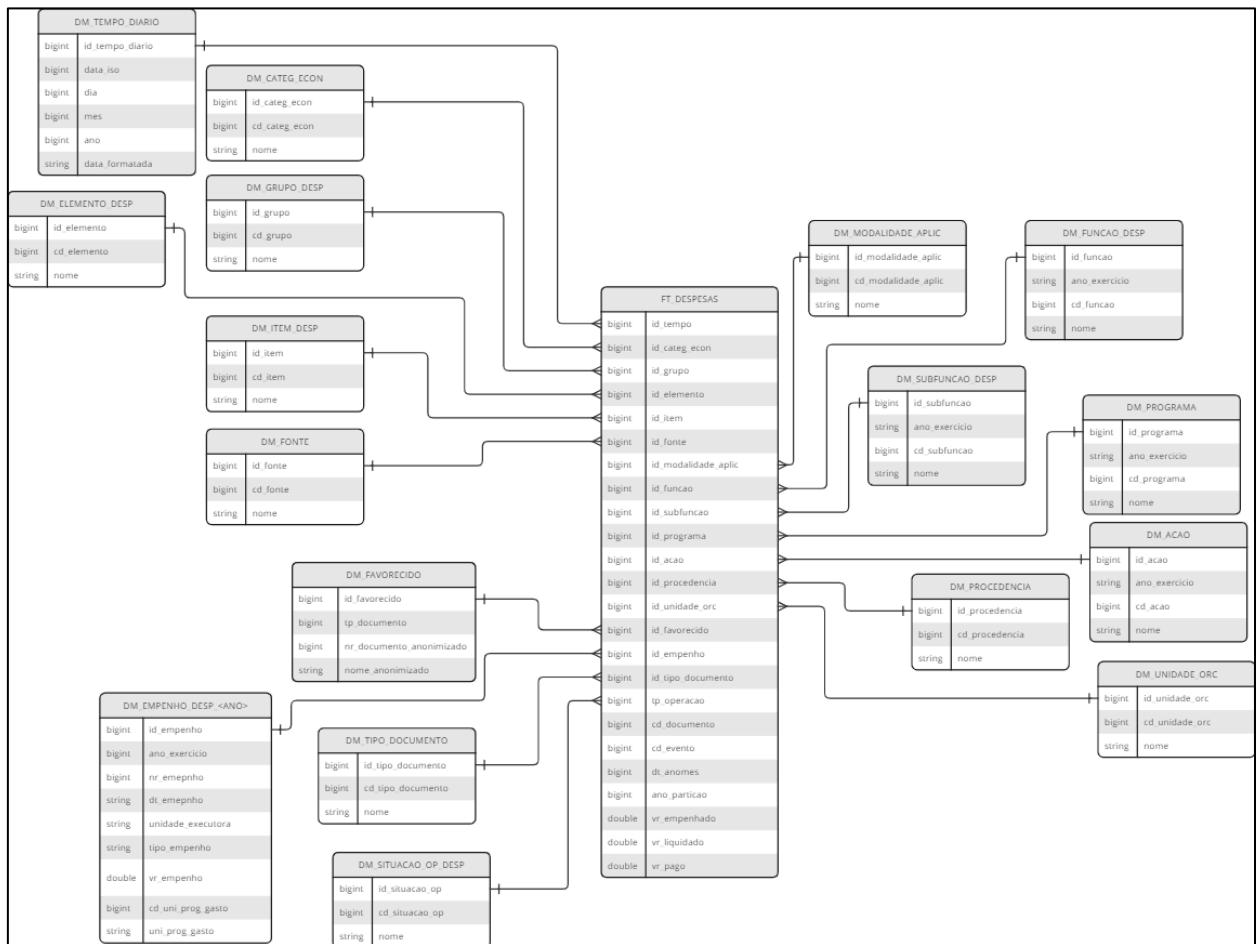


Imagen 7 – Imagem do Diagrama de Entidade-Relacionamento mostrando o esquema-estrela.

DM_ACAO

A dimensão de Ação descreve os possíveis valores para ações de gastos do Governo de MG.

Nome da Coluna	Tipo da Coluna	Descrição
ID_ACAO	Bigint	Identificador único do registro de uma ação (usado como chave estrangeira na tabela fato)
ANO_EXERCICIO	Bigint	Ano do exercício no qual aquela ação foi realizada
CD_ACAO	Bigint	Código único e interno a essa tabela que identifica uma ação
NOME	String	Nome da ação

id_acao	ano_exercicio	cd_acao	nome
36802	2014	1087	MAGISTRA - ESCOLA DE FORMACAO E DESENVOLVIMENTO PROFISSIONAL DE EDUCADORES - ENSINO MEDIO
36803	2014	2032	DESENVOLVIMENTO DO ENSINO MEDIO - ESCOLA ESTADUAL ORDEM E PROGRESSO
36804	2015	4128	FORMACAO DE PROFISSIONAIS EM NIVEL TECNICO, TECNOLOGOS E POS-GRADUACAO
36805	2016	4650	PLUG MINAS

Imagen 8 – Exemplo de dados da tabela DM_ACAO.

DM_CATEG_ECON

A dimensão de Categoria Econômica descreve os possíveis valores para as macrocategorias de gastos do Governo de MG.

Nome da Coluna	Tipo da Coluna	Descrição

ID_CATEG_ECON	Bigint	Identificador único do registro de uma categoria econômica (usado como chave estrangeira na tabela fato)
CD_CATEG_ECON	Bigint	Código único e interno a essa tabela que identifica uma categoria econômica
NOME	String	Nome da categoria econômica

id_categ_econ	cd_categ_econ	nome
20	3	DESPESAS CORRENTES
21	4	DESPESAS DE CAPITAL
22	9	A CLASSIFICAR
23	0	SEM CATEGORIA

Imagen 9 – Exemplo de dados da tabela DM_CATEG_ECON.

DM_ELEMENTO_DESP		
A dimensão de Elemento da Despesa descreve os possíveis valores para ações de gastos do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_ELEMENTO	Bigint	Identificador único do registro de um elemento de despesa (usado como chave estrangeira na tabela fato)
CD_ELEMENTO	Bigint	Código único e interno a essa tabela que identifica um elemento de despesa
NOME	String	Nome do elemento de despesa

id_elemento	cd_elemento	nome
545	83	DESPESAS DECORRENTES DE CONTRATO DE PARCERIA PUBLICO-PRIVADA - PPP, EXCETO SUBV
547	82	APORTE DE RECURSOS PELO PARCEIRO PUBLICO EM FAVOR DO PARCEIRO PRIVADO DECURRENT
546	40	SERVICOS DE TECNOLOGIA DA INFORMACAO E COMUNICACAO - PESSOA JURIDICA
450	3	PENSOES
451	16	OUTRAS DESPESAS VARIAVEIS - PESSOAL CIVIL
452	17	OUTRAS DESPESAS VARIAVEIS - PESSOAL MILITAR

Imagen 10 – Exemplo de dados da tabela DM_ELEMENTO_DESP.

DM_EMPENHO_DESP_<ANO>		
A dimensão de Empenho da Despesa descreve os possíveis valores para ações de gastos do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_EMPENHO	Bigint	Identificador único do registro de um empenho de despesa (usado como chave estrangeira na tabela fato)
ANO_EXERCICIO	Bigint	Ano do exercício no qual aquele empenho foi realizado
NR_EMPENHO	Bigint	Número do identificador do empenho, pode

		ser entendido como um
DT_EMPENHO	String	Data que o empenho foi realizado
UNIDADE_EXECUTORA	String	Unidade responsável pela execução do empenho da despesa
TIPO_EMPENHO	String	Tipo do empenho (estimado, ordinário, global ou sem descrição)
VR_EMPENHO	Double	Valor do empenho
CD_UNI_PROG_GASTO	Bigint	Código da unidade que programou o gasto
UNI_PROG_GASTO	String	Nome da unidade que programou o gasto

id_empenho	ano_exercicio	nr_empenh	dt_empenh	unidade_executora	tipo_empenho	vr_empenh	cd_uni_prog_gast	uni_prog_gasto
11249592	2016	64	2016-04-06	1510043 - 8º DRPC - TRES CORACOES	ESTIMADO	355.76	0	
11249593	2016	1672	2016-04-05	1260032 - SEE - SRE/POUSO ALEGRE	ESTIMADO	37.5	9322	ORIENTAR E ACOMPANHAR AS ESCOLAS
11249594	2016	2197	2016-04-05	1260022 - SEE - SRE/MONTES CLAROS	ESTIMADO	442.5	9223	REALIZAR ATIVIDADES DE INSPECÇÃO ESCOLAR
11249595	2016	547	2016-04-11	2330001 - IPREM	ORDINÁRIO	31.5	0	
11249596	2016	1150	2016-04-11	1260033 - SEE - SRE/PONTE NOVA	ESTIMADO	37.5	9332	ORIENTAR A ACOMPANHAR AS ESCOLAS
11249597	2016	218	2016-04-08	1510054 - 1º DRPC - TEOFILO OTONI	ESTIMADO	221.43	0	

Imagen 11 – Exemplo de dados da tabela DM_EMPENHO_2016.

Obs.: Temos 22 tabelas chamadas DM_EMPENHO_DESP por ano (de 2002 até 2023 - DM_EMPENHO_DESP_2002, DM_EMPENHO_DESP_2003, ... DM_EMPENHO_DESP_2023). Além dessas tabelas também temos uma tabela chamada DM_EMPENHOS_DESP (que consolida os dados de todas essas tabelas). Todas essas tabelas têm o mesmo formato supracitado.

DM_FAVORECIDO		
A dimensão de Favorecido descreve os possíveis valores para os prestadores de bens e serviços para o Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_FAVORECIDO	Bigint	Identificador único do registro de um favorecido (usado como chave estrangeira na tabela fato)
TP_DOCUMENTO	Bigint	Tipo do documento podendo ser 1 – pessoa física, 2 – pessoa jurídica, 3 – programas do governo estadual, 4 – programas do governo federal
NR_DOCUMENTO_ANONIMIZADO	Bigint	Número do cpf, cnpj e código interno do programa estadual e federal
NOME_ANONIMIZADO	String	Nome do favorecido (prestador do serviço ou vendedor do bem)

id_favorecido	tp_documento	nr_documento_anonimizado	nome_anonimizado
1490129	1		PABLO DAVIDSON DA ROCHA RODRIGUES
1490143	2	2430377000106	IVETE MARIA CORDEIRO LACERDA
1490144	2	2431136000181	VALTER NUNES MACHADO ME
1208615	3	99999958885	PREMIO LOTERICOS
918130	3	99999912036	WILLIAM CONNEL MCFARLAND
1426775	4	9999997989820	PREMIOS LOTERICOS
922253	4	9999990106256	INSTITUTO NACIONAL DE SALUD LIMA PERU

Imagen 12 – Exemplo de dados da tabela DM_FAVERECIDO.

DM_FONTE		
A dimensão de Fonte descreve os possíveis valores para as fontes de arrecadação do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_FONTE	Bigint	Identificador único do registro de uma fonte de receita (usado como chave estrangeira na tabela fato)
CD_FONTE	Bigint	Código único e interno a essa tabela que identifica uma fonte de receita
NOME	String	Nome da fonte de receita

id_fonte	cd_fonte	nome
486	91	TAXA DE EXPEDIENTE - ADMINISTRACAO INDIRETA
487	29	TAXA DE EXPEDIENTE - ADMINISTRACAO DIRETA
389	22	RECURSOS DO SISTEMA UNICO DE SAUDE - SUS
390	39	MULTAS PECUNIARIAS E JUROS DE MORA FIXADOS EM SENTENCIAS JUDICIAIS
391	76	TAXA DE ADMINISTRACAO DO FUNPEMG
392	36	TRANSFERENCIAS DE RECURSOS DA UNIAO VINCULADOS A EDUCACAO
393	34	NOTIFICACAO DE INFRACAO DE TRANSITO

Imagen 13 – Exemplo de dados da tabela DM_FONTE.

DM_FUNCAO_DESP		
A dimensão de Função descreve os possíveis valores para os destinos dos gastos do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_FUNCAO	Bigint	Identificador único do registro de uma função de despesa (usado como chave estrangeira na tabela fato)
ANO_EXERCICIO	Bigint	Ano do exercício no qual aquela função envolvida no empenho de alguma despesa
CD_FUNCAO	Bigint	Código único e interno a essa tabela que identifica uma função de despesa
NOME	String	Nome da função de despesa

id_funcao	ano_exercicio	cd_funcao	nome
1809	2017	18	GESTAO AMBIENTAL
1810	2017	20	AGRICULTURA
1811	2017	2	JUDICIARIA
1812	2017	3	ESSENCEIAL A JUSTICA
1813	2017	4	ADMINISTRACAO
1814	2017	8	ASSISTENCIA SOCIAL
1815	2017	16	HABITACAO

Imagen 14 – Exemplo de dados da tabela DM_FUNCAO_DESP.

DM_GRUPO_DESP

A dimensão de Grupo descreve os possíveis valores para os grupos de gastos do Governo de MG, outra subcategoria das despesas públicas.

Nome da Coluna	Tipo da Coluna	Descrição
ID_GRUPO	Bigint	Identificador único do registro de um grupo de despesa (usado como chave estrangeira na tabela fato)
CD_GRUPO	Bigint	Código único e interno a essa tabela que identifica um grupo de despesa
NOME	String	Nome do grupo de despesa

id_grupo	cd_grupo	nome
39	1	PESSOAL E ENCARGOS SOCIAIS
40	4	INVESTIMENTOS
41	0	SEM GRUPO DE DESPESA
42	3	OUTRAS DESPESAS CORRENTES
43	5	INVERSOES FINANCEIRAS
44	0	DESPESAS DE CAPITAL
45	9	RESERVA DE CONTINGENCIA

Imagen 15 – Exemplo de dados da tabela DM_GRUPO_DESP.

DM_ITEM_DESP

A dimensão de Item de Despesa descreve os possíveis valores para itens de gastos do Governo de MG, registro mais granular da despesa.

Nome da Coluna	Tipo da Coluna	Descrição
ID_ITEM	Bigint	Identificador único do registro de um item de despesa (usado como chave estrangeira na tabela fato)
CD_ITEM	Bigint	Código único e interno a essa tabela que identifica um item de despesa
NOME	String	Nome do item de despesa

id_item	cd_item	nome
2812	0	DESPESAS DE CAPITAL
2813	9	EQUIPAMENTOS HOSPITALARES, ODONTOLOGICOS E DE LABORATORIO
2814	13	MATERIAL ESPORTIVO E RECREATIVO
2815	19	INSTRUMENTOS MUSICIAIS E ARTISTICOS
2816	3	EXECUCAO DE OBRAS POR CONTRATO DE BENS PATRIMONIAVEIS

Imagen 15 – Exemplo de dados da tabela DM_ITEM_DESP.

DM_MODALIDADE_APPLIC		
A dimensão de Modalidade de Aplicação descreve os possíveis valores para as formas de aplicação dos recursos orçamentários do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_MODALIDADE_APPLIC	Bigint	Identificador único do registro de uma modalidade de aplicação (usado como chave estrangeira na tabela fato)
CD_MODALIDADE_APPLIC	Bigint	Código único e interno a essa tabela que identifica uma modalidade de aplicação
NOME	String	Nome da modalidade de aplicação

id_modalidade_aplic	cd_modalidade_aplic	nome
145	95	APLICACAO DIRETA A CONTA DE RECURSOS DE QUE TRATAM OS EE 1 E 2 DO ART.
146	67	EXECUCAO DE CONTRATO DE PARCERIA PUBLICO-PRIVADA - PPP
120	0	JUROS E ENCARGOS DA DIVIDA
121	70	TRANSFERENCIAS A INSTITUICOES MULTIGOVERNAMENTAIS
122	0	RESERVA DE CONTINGENCIA
123	91	APLICACAO DIRETA DECORRENTE DE OPERACOES ENTRE ORGAOS, FUNDOS E ENTIDA

Imagen 16 – Exemplo de dados da tabela DM_MODALIDADE_APPLIC.

DM_PROCEDENCIA		
A dimensão de Procedência descreve os possíveis valores para a origem da verba pública do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_PROCEDENCIA	Bigint	Identificador único do registro de uma procedência (usado como chave estrangeira na tabela fato)
CD_PROCEDENCIA	Bigint	Código único e interno a essa tabela que identifica uma procedência
NOME	String	Nome da procedência da verba

id_procedencia	cd_procedencia	nome
112	0	RECURSOS DECORRENTES DA DESVINCULACAO DE RECEITAS CONFORME A EC 93/2016
86	4	RECURSOS CONSTITUCIONALMENTE DESTINADOS PARA FOMENTO AO ENSINO
87	5	RECURSOS RECEBIDOS DA CONFIP

Imagen 17 – Exemplo de dados da tabela DM_PROCEDENCIA.

DM_PROGRAMA		
A dimensão de Programa descreve os possíveis valores para macro categorias de empenho da despesa pública do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_PROGRAMA	Bigint	Identificador único do registro de uma função de despesa (usado como chave estrangeira na tabela fato)
ANO_EXERCICIO	Bigint	Ano do exercício no qual aquele programa foi relacionado ao empenho de algum gasto público
CD_PROGRAMA	Bigint	Código único e interno a essa tabela que identifica uma função de despesa
NOME	String	Nome da função de despesa

id_programa	ano_exercicio	cd_programa	nome
9480	2000	476	FORMACAO PARA O EXERCICIO DA CIDADANIA
9481	2000	474	ASSISTENCIA JUDICIAL
9482	2001	476	FORMACAO PARA O EXERCICIO DA CIDADANIA
9483	2000	647	PROTECAO A FLORA E A FAUNA
9484	1995	13	ORGANIZACAO AGRARIA
9485	2001	363	ASSISTENCIA PREVIDENCIARIA
9486	2000	603	APOIO HABITACIONAL

Imagen 18 – Exemplo de dados da tabela DM_PROGRAMA.

DM_SITUACAO_OP_DESP		
A dimensão de Situação Operacional da Despesa descreve os possíveis valores para os status de uma despesa do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_SITUACAO_OP	Bigint	Identificador único do registro de uma situação operacional de despesa (usado como chave estrangeira na tabela fato)
CD_SITUACAO_OP	Bigint	Código único e interno a essa tabela que identifica uma situação operacional de despesa
NOME	String	Nome da situação operacional da despesa

id_situacao_op	cd_situacao_op	nome
0	23	Acatada pelo banco
1	1	Pendente de transmissão aos bancos
2	2	Acatada pelo banco
3	3	Sujeita a compensação bancária
4	4	Cancelada
5	5	Cancelada _ TED
6	6	Cancelada _ Recibo
7	7	Cancelada automaticamente após última transmissão bancária
8	8	Anulada automaticamente após última transmissão bancária
9	9	Cancelada sem cancelamento do IRRF retido

Imagen 19 – Exemplo de dados da tabela DM_SITUACAO_OP_DESP.

DM_SUBFUNCAO_DESP		
A dimensão de Subfunção da Despesa descreve os possíveis valores para subcategorias das funções de despesas do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_SUBFUNCAO	Bigint	Identificador único do registro de uma subfunção de despesa (usado como chave estrangeira na tabela fato)
ANO_EXERCICIO	Bigint	Ano do exercício no qual aquela subfunção envolvida no empenho de alguma despesa
CD_SUBFUNCAO	Bigint	Código único e interno a essa tabela que identifica uma subfunção de despesa
NOME	String	Nome da subfunção de despesa

id_subfuncao	ano_exercicio	cd_subfuncao	nome
2357	2010	781	TRANSPORTE AEREO
2358	2011	812	DESPORTO COMUNITARIO
2359	2011	813	LAZER
2360	2011	846	OUTROS ENCARGOS ESPECIAIS
2361	2011	0	SEM DESCRICAO
2362	2011	183	INFORMACAO E INTELIGENCIA
2363	2011	127	ORDENAMENTO TERRITORIAL
2364	2011	542	CONTROLE AMBIENTAL
2365	2011	125	NORMATIZACAO E FISCALIZACAO
2366	2011	695	COMERCIO EXTERIOR

Imagen 20 – Exemplo de dados da tabela DM_SUBFUNCAO_DESP.

DM_TEMPO_DIARIO		
A dimensão de Tempo Diário descreve os possíveis valores para os dias, meses e anos para cada empenho de despesa Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição

ID_TEMPO	Bigint	Identificador único do registro de um tempo diário de despesa (usado como chave estrangeira na tabela fato)
DATA_ISO	Bigint	Data formatada no padrão ISO (YYYYMMDD)
DIA	Bigint	Valor do dia para a data
MES	Bigint	Valor do mês para a data
ANO	Bigint	Valor do ano para a data
DATA_FORMATADA	String	Data formatada no padrão YYYY-MM-DD

id_tempo	data_iso	dia	mes	ano	data_formatada
884	19080924	24	9	1908	1908-09-24
885	19080925	25	9	1908	1908-09-25
886	19080926	26	9	1908	1908-09-26
887	19080927	27	9	1908	1908-09-27

Imagen 21 – Exemplo de dados da tabela DM_TEMPO_DIARIO.

DM_TIPO_DOCUMENTO		
A dimensão de Tipo de Documento descreve os possíveis valores para os tipos de documentos de empenho de despesa do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_TIPO_DOCUMENTO	Bigint	Identificador único do registro de uma função de despesa (usado como chave estrangeira na tabela fato)
CD_TIPO_DOCUMENTO	Bigint	Código único e interno a essa tabela que identifica uma função de despesa
NOME	String	Nome da função de despesa

id_tipo_documento	cd_tipo_documento	nome
1	0	EMPENHO
485	21	ANULACAO DE DESPESA ORCAMENTARIA
486	22	RETENCAO - DETALHE
487	23	VALOR GLOBAL DA OP DE RETENCAO - DETALHE
488	19	PAGAMENTO RESTO A PAGAR NAO PROCESSADO
489	66	ANULACAO SALDO LIQUIDADO - RETENCAO
490	68	ANULACAO SALDO LIQUIDADO RP - MULTA
491	74	APROPRIACAO DESPESA EXTRA ORCAMENTARIA - RETENCAO

Imagen 22 – Exemplo de dados da tabela DM_TIPO_DOCUMENTO.

DM_UNIDADE_ORC		
A dimensão de Unidade Orçamentária descreve os possíveis valores para áreas de destino de gastos do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_UNIDADE_ORC	Bigint	Identificador único do registro de uma unidade orçamentária (usado como

			chave estrangeira na tabela fato)
ANO_EXERCICIO	Bigint	Ano do exercício no qual aquela unidade orçamentária foi envolvida no empenho de alguma despesa	
CD_TIPO_DOCUMENTO	Bigint	Código do tipo de documento que foi usado pela unidade orçamentária no empenho da despesa	
ID_GRUPO_ADMINISTRACAO	Bigint	Id do grupo da administração ao qual essa unidade orçamentária faz parte	
GRUPO_ADMINISTRACAO	String	Nome do grupo da administração ao qual essa unidade orçamentária faz parte	
ID_ADMINISTRACAO	Bigint	Id da administração ao qual essa unidade orçamentária faz parte	
ADMINISTRACAO	String	Nome da administração ao qual essa unidade orçamentária faz parte	
NOME	String	Nome da unidade orçamentária	
SIGLA	String	Sigla da unidade orçamentária	

id_unidade_orc	ano_exercicio	cd_unidade_orc	id_grupo_administracao	grupo_administracao	id_administracao	administracao	nome	sigla
1962	2014	4331	4	FUNDOS	4	FUNDOS	FUNDO DE DESENVOLVIMENTO METROPOLITANO	FDM
1963	2015	1011	1	ADMINISTRACAO DIRETA	1	ADMINISTRACAO DIRETA	ASSEMBLEIA LEGISLATIVA DO ESTADO DE MINAS GERAIS	ALEMG
1964	2015	1301	1	ADMINISTRACAO DIRETA	1	ADMINISTRACAO DIRETA	SECRETARIA DE ESTADO DE TRANSPORTES E OBRAS PUBLICAS	SETOP
1965	2015	1441	1	ADMINISTRACAO DIRETA	1	ADMINISTRACAO DIRETA	DEFENSORIA PÚBLICA DO ESTADO DE MINAS GERAIS	DEF PUB
1966	2015	4601	4	FUNDOS	4	FUNDOS	FUNDO ESTADUAL DOS DIREITOS DO IDOSO	FEI
1967	2015	9999	1	ADMINISTRACAO DIRETA	1	ADMINISTRACAO DIRETA	EMG - ADMINISTRACAO DIRETA	EMG - ADM. DIRETA
1968	2015	4481	4	FUNDOS	4	FUNDOS	FUNDO DE PARCERIAS PÚBLICO-PRIVADAS DO ESTADO DE MINAS GERAIS	FUNDO PPP
1969	2015	4521	4	FUNDOS	4	FUNDOS	FUNDO DE UNIVERSALIZAÇÃO DO ACESSO A SERVIÇOS DE TELECOMUNICAÇÃO EM MINAS GERAIS	FUNDOMIC

Imagen 23 – Exemplo de dados da tabela DM_UNIDADE_ORC.

FT_DESPESA_<ANO>		
A dimensão de Ação descreve os possíveis valores para ações de gastos do Governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ID_TEMPO	Bigint	Id da dimensão DIM_TEMPO_DIARIO (chave-estrangeira)
ID_CATEG_ECON	Bigint	Id da dimensão DIM_CATEG_ECON (chave-estrangeira)
ID_GRUPO	Bigint	Id da dimensão DIM_GRUPO (chave-estrangeira)
ID_ELEMENTO	Bigint	Id da dimensão DIM_ELEM_DESP (chave-estrangeira)
ID_ITEM	Bigint	Id da dimensão DIM_ITEM_DESP (chave-estrangeira)
ID_FONTE	Bigint	Id da dimensão DIM_FONTE (chave-estrangeira)
ID_MODALIDADE_APPLIC	Bigint	Id da dimensão DIM_MODALIDADE_APPLIC (chave-estrangeira)
ID_FUNCAO	Bigint	Id da dimensão DIM_FUNCAO (chave-estrangeira)
ID_SUBFUNCAO	Bigint	Id da dimensão DIM_SUBFUNCAO (chave-estrangeira)
ID_PROGRAMA	Bigint	Id da dimensão DIM_PROGRAMA (chave-estrangeira)
ID_ACAO	Bigint	Id da dimensão DIM_ACAO (chave-estrangeira)
ID_PROCEDENCIA	Bigint	Id da dimensão DIM_PROCEDENCIA (chave-

		estrangeira)
ID_UNIDADE_ORC	Bigint	Id da dimensão DIM_UNIDADE_ORC (chave-estrangeira)
ID_FAVORECIDO	Bigint	Id da dimensão DIM_FAVORECIDO (chave-estrangeira)
ID_EMPENHO	Bigint	Id da dimensão DIM_EMPENHO_DESP (chave-estrangeira)
ID_TIPO_DOCUMENTO	Bigint	Id da dimensão DIM_TIPO_DOCUMENTO (chave-estrangeira)
TP_OPERACAO	Bigint	Id da dimensão DIM_SITUACAO_OP_TIPO (chave-estrangeira)
CD_DOCUMENTO	Bigint	
CD_EVENTO	Bigint	Código do evento da despesa. Cada evento pode conter uma ou mais despesas.
DT_ANOMES	Bigint	Data da despesa no formato ano + mês
ANO_PARTICAO	Bigint	Ano da despesa
VR_EMPENHADO	double	Valor empenhado para essa despesa
VR_LIQUIDADO	double	Valor liquidado para essa despesa
VR_PAGO	double	Valor pago para essa despesa

id_tempo	id_categ_econ	id_grupo	id_elemento	id_item	id_fonte	id_modalidade_aplic	id_funcao	id_subfuncao	id_programa	id_acao	id_procedencia	id_unidade_orc	id_favorecido	id_empenho	id_tipo_documento
47104	20	42	491	2604	58	130	1895	3773	14008	55362	93	4026	1342860	12626607	1
47104	20	42	491	2604	58	130	1895	3773	14008	55362	93	4026	1372949	12626581	1
47104	20	42	491	2604	58	130	1895	3773	14008	55362	93	4026	1738874	12626174	1
47104	20	42	491	2604	58	130	1895	3773	14008	55362	93	4026	1707207	12626814	1
47104	20	42	491	2604	58	130	1895	3773	14008	55362	93	4026	1919776	12626827	1
47104	20	42	491	2604	58	130	1895	3773	14008	55362	93	4026	1735389	12627398	1

Imagen 24 – Exemplo de dados da tabela FT_DESPESA_2020.

Além dessas tabelas iniciais ainda construímos para as análises um conjunto de outras **23 tabelas** (ft_despesa_last24m, ft_despesa, dm_empenhos_desp, tb_anomaly_detection, tb_association_rules, tb_prediction_mean, vw_adv_analytics_anomaly_detection, vw_adv_analytics_anomaly_detection_agents, vw_adv_analytics_anomaly_detection_metrics, vw_adv_analytics_forecast, vw_agg_categ_econ, vw_agg_elemento, vw_agg_fact, vw_agg_fonte, vw_agg_funcao, vw_agg_grouped_favorecidos, vw_agg_grupo, vw_agg_modalidade_aplic, vw_agg_procedencia, vw_agg_subfuncao, vw_agg_tipo_documento, vw_agg_unidade_orc, vw_dw);

Obs.: As tabelas *ft_despesa* e *dm_empenhos_desp* não serão detalhadas porque são somente uniões das tabelas *ft_despesa* e *dm_empenhos_desp* para vários anos, contendo os mesmos campos que as tabelas que as originam.

FT_DESPESA_LAST24M		
Tabela Fato constituída da união das tabelas <i>ft_despesa_2021</i> e <i>ft_despesa_2022</i> . Principal tabela do painel de controle.		
Nome da Coluna	Tipo da Coluna	Descrição
ID	String	Concatenação entre todos os campos das tabelas de origem, ou seja: id_tempo, id_categ_econ, id_grupo, id_elemento, id_item, id_fonte, id_modalidade_aplic, id_funcao, id_subfuncao, id_programa, id_acao, id_procedencia, id_unidade_orc, id_favorecido, id_empenho, id_tipo_documento, tp_operacao, cd_documento, cd_evento, dt_anomes, ano_particao.
Essa tabela possui todos os campos da tabela <i>ft_despesa</i> , com a adição do campo de ID		

id	id_tempo	id_catog_econ	id_grupo	id_elemento	id_item	id_fonte	id_modalidade_aplic	id_funcao	id_subfuncao	id_programa	id_acao	ida
50273-20-42-502-3012-408-130-1931-3853-14170-55805-93-4104-1457645-13174445-1-2-13825-502001-202111-2021	50273	20	42	502	3012	488	130	1931	3853	14170	55805	93
50273-20-42-502-3012-408-130-1931-3853-14170-55806-93-4104-1457645-13176539-1-2-13826-502001-202111-2021	50273	20	42	502	3012	488	130	1931	3853	14170	55805	93
50273-20-42-502-3012-408-130-1931-3853-14170-55805-93-4104-1622908-13176558-1-2-13827-502001-202111-2021	50273	20	42	502	3012	488	130	1931	3853	14170	55805	93
50273-20-42-502-3012-408-130-1931-3853-14170-55806-93-4104-1091960-13174444-1-2-13828-502001-202111-2021	50273	20	42	502	3012	488	130	1931	3853	14170	55805	93
50273-20-42-502-3012-408-130-1931-3853-14170-55805-93-4104-1925395-13176556-1-2-13830-502001-202111-2021	50273	20	42	502	3012	488	130	1931	3853	14170	55805	93

Imagen 25 – Exemplo de dados da tabela FT_DESPESA_LAST24M.

TB_ANOMALY_DETECTON		
Tabela gerada a partir do modelo de Detecção de Anomalias. Esse modelo usa como entrada os dados da tabela ft_despesa_last24m.		
Nome da Coluna	Tipo da Coluna	Descrição
ID	String	Id único do registro
ANOMALY	string	Flag binária com valores ‘True’ (caso o registro seja considerado um outlier) ou ‘False’ (em caso contrário)

id	anomaly
50283-20-42-520-2920-58-136-1924-3910-14174-55918-93-4113-1368950-12956563-510-2-8013-503003-202112-2021	False
50283-20-42-520-2920-58-136-1924-3910-14163-55916-109-4113-1369099-13181512-532-2-11437-701015-202112-2021	False
50283-20-42-520-2920-58-136-1924-3910-14163-55742-93-4113-1536523-13169437-510-2-7990-503003-202112-2021	False
50283-20-42-520-2920-58-136-1924-3910-14163-55742-93-4113-1032787-13165342-510-2-7962-503003-202112-2021	False
50283-20-42-520-2920-58-136-1924-3910-14163-55742-93-4113-1032787-13170025-510-2-7960-503003-202112-2021	False
50283-20-42-520-2920-58-136-1924-3910-14163-55742-93-4113-1144718-13170021-510-2-8010-503003-202112-2021	False
50283-20-42-520-2920-58-136-1924-3910-14163-55742-93-4113-1144718-13170021-510-2-8010-503003-202112-2021	False

Imagen 26 – Exemplo de dados da tabela TB_ANOMALY_DETECTON.

TB_ASSOCIATION_RULES		
Essa tabela contém os resultados do modelo de regras de associação que relacionam itens que são comumente comprados ou contratados juntos nos gastos do governo de MG.		
Nome da Coluna	Tipo da Coluna	Descrição
ANTECEDENTS	String	Descreve qual(is) item(ns) precede(m) a relação de associação. São os itens cuja presença é usada para prever a ocorrência de outros itens (as consequências) no conjunto de dados
CONSEQUENTS	String	Esses são os itens ou conjuntos de itens que aparecem no lado direito da regra de associação. São os itens cuja ocorrência está sendo prevista ou associada à presença dos antecedentes.
ANTECEDENT SUPPORT	Float	Medida da proporção de transações no conjunto de dados que contém o conjunto de itens antecedente. Indica a frequência com que o antecedente aparece nos dados.
CONSEQUENT SUPPORT	Float	Medida da proporção de transações que contêm o conjunto de itens consequente. Indica a frequência com que o consequente aparece nos dados.
SUPPORT	Float	O suporte mede a proporção de transações no conjunto de dados que contém os conjuntos de itens antecedentes e consequentes. Quantifica a frequência com que a associação entre o antecedente e o consequente ocorre em conjunto.
CONFIDENCE	Float	A confiança é a probabilidade condicional de encontrar o conjunto de itens consequente em uma transação, dado que o conjunto de itens antecedente está presente. É calculado como o suporte do conjunto de itens combinado dividido pelo suporte do

		conjunto de itens antecedente.
LIFT	Float	Medida da razão entre o suporte observado e o suporte esperado se o antecedente e o consequente forem independentes. Ele quantifica o quanto mais provável é o consequente ser comprado quando o antecedente também é comprado, em comparação com quando eles ocorrem independentemente.
LEVERAGE	Float	Calcula a diferença entre o suporte observado da combinação antecedente-consequente e o que seria esperado se fossem independentes. Ajuda a identificar o grau de associação entre o antecedente e o consequente.
CONVICTION	Float	É uma medida da dependência entre o antecedente e o consequente. Indica o quanto a regra de associação é dependente do antecedente ser verdadeiro.
ZHANGS_METRIC	Float	A métrica de Zhang é outra medida da força da associação entre o antecedente e o consequente. É uma combinação de levantamento e alavancagem e é útil para classificar regras com base em seu significado.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
['CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔNICA']	['ADQUIRIR AMPLIFICADOR DE ÁUDIO']	0.023855492	0.022322094	0.022085028	0.92578375	39.866505	0.021531053	13.161241	0.99874175
['ADQUIRIR AMPLIFICADOR DE ÁUDIO']	['CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔNICA']	0.023222094	0.023855492	0.022085028	0.95103514	39.866505	0.021531053	19.935623	0.99809414
['EXECUTAR O PROJETO FORTALECIMENTO DAS APRENDIZAGENS']	['ADQUIRIR AMPLIFICADOR DE ÁUDIO']	0.024290478	0.023222094	0.022168973	0.912661	39.301407	0.021604897	11.183756	0.9988174
['ADQUIRIR AMPLIFICADOR DE ÁUDIO']	['EXECUTAR O PROJETO FORTALECIMENTO DAS APRENDIZAGENS']	0.023222094	0.024290478	0.022168973	0.95465004	39.301407	0.021604897	21.515102	0.9977249

Imagen 27 – Exemplo de dados da tabela TB_ASSOCIATION_RULES.

TB_PREDICTION_MEAN

Essa tabela contém os dados do resultado do modelo de séries temporais que usamos para estimar o valor a ser pago nos próximos 15 meses com base nos valores pagos historicamente pelo governo de MG.

Nome da Coluna	Tipo da Coluna	Descrição
SUM_VR_PAGO	Float	Valor estimado a ser pago
DT_ANOMES	String	Data na qual esse pagamento irá ocorrer.

sum_vr_pago	dt_anomes
1.09300593E10	202301
9.5358413E9	202302
1.03088128E10	202303
9.8802688E9	202304
1.01178583E10	202305
9.9861361E9	202306
1.00591647E10	202307

Imagen 28 – Exemplo de dados da tabela TB_PREDICTION_MEAN.

VW_ADV_ANALYTICS_ANOMALY_DETECTION_AGENTS

Essa view descreve os principais agentes das transações e os classifica os gastos nos quais estão envolvidos entre outlier ou não-outlier. Essa visão também é baseada na tabela tb_anomaly_detection.

Nome da Coluna	Tipo da Coluna	Descrição
UNIDADE_ORC_NOME	String	Nome da unidade orçamentária
FAVORECIDO_NOME_ANONIMIZADO	String	Nome do favorecido (anonimizado)
VR_EMPENHADO	Float	Valor empenhado
VR_LIQUIDADO	Float	Valor Liquidado
VR_PAGO	Float	Valor Pago
ANOMALY	String	Flag que indica se o registro foi ou não considerado como outlier

unidade_orc_nome	favorecido_nome_anonimizado	vr_empenhado	vr_liquidado	vr_pago	anomaly
SECRETARIA DE ESTADO DE EDUCACAO	SERVICO NACIONAL DE APRENDIZAGEM INDUSTRIAL - DEPARTAMENTO REGIONAL DE	0.0	34398.0	0.0	False
SECRETARIA DE ESTADO DE GOVERNO	PM TABULEIRO	0.0	0.0	225000.0	False
SECRETARIA DE ESTADO DE GOVERNO	PM CONCEICAO DA BARRA DE MINAS	0.0	0.0	225000.0	False
FUNDO ESTADUAL DE SAUDE	SOCIEDADE BENEFICIENTE SAO CAMILO	0.0	0.0	134010.16	False
INSTITUTO DE PREVIDENCIA DOS SERVIDORES MILITARES DO ESTADO DE MINAS GERAIS	COOPERATIVA MEDICA DE ESPECIALIDADES LTDA	104164.97	0.0	0.0	False
SECRETARIA DE ESTADO DE EDUCACAO	PM FORMIGA	2564010.01	0.0	0.0	False

Imagen 29 – Exemplo de dados da tabela
VW_ADV_ANALYTICS_ANOMALY_DETECTION_AGENTS.

VW_ADV_ANALYTICS_ANOMALY_DETECTION_METRICS

Essa tabela resume os valores encontrados na tabela de *tb_anomaly_detection* para ser usada como mostrador macro dos números analisados na detecção de outliers. Possui somente 2 registros que summarizam os conjuntos de dados anômalos e dados não-anômalos.

Nome da Coluna	Tipo da Coluna	Descrição
ANOMALY	String	Flag que indica se a métrica summarizada pertence ao grupo de dados anômalos ou não
MIN_VR_EMPENHADO	Float	Menor valor empenhado para aquele grupo
MAX_VR_EMPENHADO	Float	Maior valor empenhado para aquele grupo
AVG_VR_EMPENHADO	Float	Valor médio empenhado para aquele grupo
MIN_VR_LIQUIDADO	Float	Menor valor liquidado para aquele grupo
MAX_VR_LIQUIDADO	Float	Maior valor liquidado para aquele grupo
AVG_VR_LIQUIDADO	Float	Valor médio liquidado para aquele grupo
MIN_VR_PAGO	Float	Menor valor pago para aquele grupo
MAX_VR_PAGO	Float	Maior valor pago para aquele grupo
AVG_VR_PAGO	Float	Valor médio pago para aquele grupo

anomaly	min_vr_empenhado	max_vr_empenhado	avg_vr_empenhado	min_vr_liquidado	max_vr_liquidado	avg_vr_liquidado	min_vr_pago	max_vr_pago	avg_vr_pago
False	-2.56802336256E9	3.0E9	431067.70467124233	-1.1690391898E8	1.1E9	431351.7519325066	-1.1690391898E8	1.1E9	392046.35155014513
True	-103500.0	64964.39	1515.9145879937737	-6.5567E8	31646.33	-1017.8857807942778	-1.0E9	30920.21	484.4788749685671

Imagen 30 – Exemplo de dados da tabela
VW_ADV_ANALYTICS_ANOMALY_DETECTION_METRICS.

VW_ADV_ANALYTICS_FORECAST

Essa visão contém os dados de saída do modelo de previsão de gastos (*tb_prediction_mean*) combinados com os dados de despesa (*ft_despesa*) para criar em uma única tabela os dados de gastos realizados e previstos.

Nome da Coluna	Tipo da Coluna	Descrição
DT_ANOMES	String	Ano e mês do gasto
SUM_VR_PAGO	Float	Valor pago ou previsto para ser pago
STATUS	String	Status do pagamento (realizado ou forecast)

dt_anomes	sum_vr_pago	status
202301	1.0930059264E10	FORECAST
202302	9.53584128E9	FORECAST
202303	1.03088128E10	FORECAST
202008	7.083481798740002E9	REALIZADO
202006	6.850080372050013E9	REALIZADO
201306	5.347647832220005E9	REALIZADO

Imagen 31 – Exemplo de dados da tabela VW_ADV_ANALYTICS_FORECAST.

VW_AGG_<DIM>

A dimensão de Ação descreve os possíveis valores para ações de gastos do Governo de MG.

Nome da Coluna	Tipo da Coluna	Descrição
ANO_PARTICAO	Bigint	Ano da participação da tabela
ID_<DIM>	Bigint	Id único dentro da dimensão específica
COUNTING	Bigint	Número de registros por ano e por id
S_VR_EMPENHADO	Float	Soma dos valores empenhados por ano e por id
S_VR_LIQUIDADO	Float	Soma dos valores liquidados por ano e por id
S_VR_PAGO	Float	Soma dos valores pagos por ano e por id

ano_particao	id_categ_econ	counting	s_vr_empenhado	s_vr_liiquidado	s_vr_pago
2002	20	2432127	1.750547334567003E10	1.71115608078E10	1.6707606344770008E10
2002	21	85471	1.8539680922300003E9	1.68789214274E9	1.38973052019E9
2003	20	2312789	1.8802965852729973E10	1.843380725112001E10	1.794110520165998E10
2003	21	64514	1.46049355755E9	1.2233110768300002E9	1.1610737406599996E9
2004	20	2339880	1.9766973463219982E10	1.92676311213996E10	1.8677622082299923E10
2004	21	77783	1.9596346338300002E9	1.5344345677900007E9	1.4102298307000005E9

Imagen 32 – Exemplo de dados da tabela VW_AGG_CATEG_ECON.

As tabelas cujo prefixo são **vw_agg**, são tabelas criadas para contabilizar métricas relevantes de cada dimensão do modelo. Sumarizamos nessas tabelas o número de linhas, a soma dos valores empenhados, liquidados e pagos por ano e pelo valor único do id de cada dimensão. Ou seja, na dimensão **categ_econ**, temos esses valores sumarizados por ano e pelo campo **id_categ_econ**, na dimensão **programa** temos esses valore sumarizados por ano e pelo campo **id_programa**.

```

CREATE OR REPLACE VIEW vw_agg_categ_econ AS
(
  SELECT
    ano_particao, id_categ_econ,
    count(*) as counting,
    sum(vr_empenhado) as s_vr_empenhado,
    sum(vr_liiquidado) as s_vr_liiquidado,
    sum(vr_pago) as s_vr_pago
    FROM ft_despesa GROUP BY ano_particao, id_categ_econ
    ORDER BY ano_particao, id_categ_econ
)
  
```

Imagen 33 – Exemplo de código SQL que gera a tabela vw_agg_categ_econ.

VW_DW

Visão consolidada da tabela fato. Essa visão já traz todos os nomes e demais colunas importantes das tabelas dimensões decodificado (sem necessidade de *join*).

Campo original na tabela de origem	Nome da Coluna	Tipo da Coluna
FT_DESPESA_LAST24M.ID	ID	Bigint
DM_TEMPO_DIARIO.DIA	DIA	Bigint
DM_TEMPO_DIARIO.MES	MES	Bigint
DM_TEMPO_DIARIO.ANO	ANO	Bigint
DM_CATEG_ECON.NOME	CATEG_ECON_NOME	String
DM_GRUPO_DESP.NOME	GRUPO_NOME	String
DM_ELEMENTO_DESP.NOME	ELEMENTO_DESP_NOME	String
DM_ITEM_DESP.NOME	ITEM_DESP_NOME	String
DM_FONTE.NOME	FONTE_NOME	String
DM_MODALIDADE_APPLIC.NOME	MODALIDADE_APPLIC_NOME	String
DM_FUNCAO_DESP.NOME	FUNCAO_DESP_NOME	String
DM_SUBFUNCAO_DESP.NOME	SUBFUNCAO_DESP_NOME	String
DM_PROGRAMA.NOME	PROGRAMA_NOME	String
DM_PROCEDENCIA.NOME	PROCEDENCIA_NOME	String
DM_UNIDADE_ORC.GRUPO_ADMINISTRACAO	UNIDADE_ORC_GRUPO_ADMINISTRACAO	String
DM_UNIDADE_ORC.ADMINISTRACAO	UNIDADE_ORC_ADMINISTRACAO	String
DM_UNIDADE_ORC.NOME	UNIDADE_ORC_NOME	String
DM_UNIDADE_ORC.SIGLA	UNIDADE_ORC_SIGLA	String
DM_FAVERECIDO.NOME_ANONIMIZADO	FAVERECIDO_NOME_ANONIMIZADO	String
DM_EMPENHOS_DESP.DT_EMPENHO	EMPENHOS_DESP_DT_EMPENHO	String
DM_EMPENHOS_DESP.UNIDADE_EXECUTORA	EMPENHOS_DESP_UNIDADE_EXECUTORA	String
DM_EMPENHOS_DESP.TIPO_EMPENHO	EMPENHOS_DESP_TIPO_EMPENHO	String
DM_EMPENHOS_DESP.VR_EMPENHO	EMPENHOS_DESP_VR_EMPENHO	
DM_EMPENHOS_DESP.UNI_PROG_GAS	EMPENHOS_DESP_UNI_PROG_GAS	String
DM_TIPO_DOCUMENTO.NOME	TIPO_DOCUMENTO_NOME	String
DM_SITUACAO_OP_DESP.NOME	SITUACAO_OP_DESP_NOME	String
FT_DESPESA_LAST24M	VR_EMPENHADO	Double
FT_DESPESA_LAST24M	VR_LIQUIDADO	Double
FT_DESPESA_LAST24M	VR_PAGO	Double

id	dia	mes	ano	categ_econ_nome	grupo_nome	elemento_desp_nome	item_desp_nome
43963-20-42-534-2873-506-130-1955-3950-14356-56360-93-4219-1872780-13218793-510-2-1024-503003-202203-2022	25	3	2022	DESPESAS CORRENTES	OUTRAS DESPESAS CORRENTES	OUTROS SERVICOS DE TERCEROS - PESSOA JURIDICA	TARIFA DE ENERGIA ELETTRICA
43963-20-42-466-2560-58-130-1948-3992-14318-56247-93-4259-1975101-13330755-532-2-7707-701003-202203-2022	25	3	2022	DESPESAS CORRENTES	OUTRAS DESPESAS CORRENTES	INDENIZACOES E RESTITUICOES	REEMBOLSO DE DESPESAS MEDICO-HOSPITALARES
43963-20-42-533-2798-484-130-1955-3992-14417-565537-93-4272-1817854-13336774-1-2-1207-502001-202203-2022	25	3	2022	DESPESAS CORRENTES	OUTRAS DESPESAS CORRENTES	OUTROS SERVICOS DE TERCEROS - PESSOA FISICA	ESTAGIARIOS
43963-20-42-534-2548-2-130-1955-3959-1493-56636-93-4219-1230245-532-2-1762-701008-202203-2022	25	3	2022	DESPESAS CORRENTES	OUTRAS DESPESAS CORRENTES	OUTROS SERVICOS DE TERCEROS - PESSOA JURIDICA	CURSOS DE FORMACAO E CAPACITACAO PARA O CIDADAO
43963-20-42-476-2791-58-130-1953-3948-14390-56583-97-4209-2017152-13337416-510-2-3361-503003-202203-2022	25	3	2022	DESPESAS CORRENTES	OUTRAS DESPESAS CORRENTES	SENTENCIAS JUDICIAIS	OUTRAS SENTENCIAS JUDICIAIS

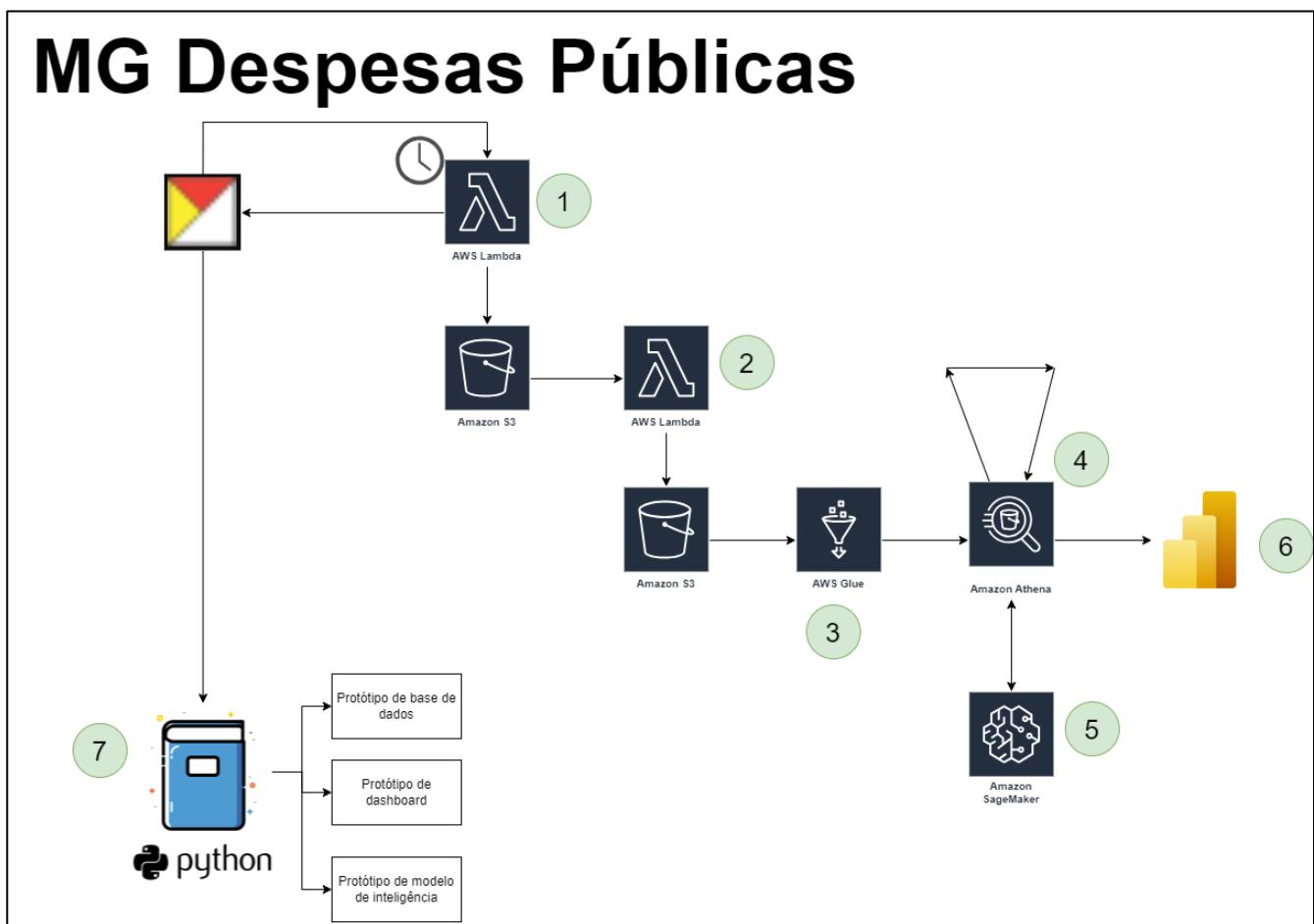
Imagen 34 - Exemplo de dados da tabela VW_DW.

Decidimos criar essa tabela no modelo de dados por conta da velocidade da ferramenta AWS Athena no processamento. Por ser uma ferramenta de computação distribuída e por termos somente 1 informação útil por tabela dimensional (normalmente o campo “nome”) cruzamos as dimensões à tabela fato para obter essa visão de *data warehouse*.

3.1.4. Solução

Para endereçar os pontos do desafio proposto, implementamos um fluxo completamente automatizado de dados em nuvem. Nossa solução de dados tem como principais funcionalidades:

- Utilizar computação distribuída para realizar a extração e o processamento dos dados;
- Implementar um modelo moderno de arquitetura de dados (*Medallion Architecture*), como visto em empresas como Uber e Google;
- Utilizar ferramenta de catálogo de dados que seja resiliente à mudança de esquema de dados (problema recorrente em empresas de diversos setores, como financeiro, saúde e transporte);
- Desenvolvimento de modelos de *machine learning* para resolver questões de real interesse dos stakeholders do processo de gestão de gastos públicos como acompanhamento de metas de orçamento e previsão, detecção e análise de outliers etc.
- Implementação de um painel de controle com visões estratégicas, táticas, operacional e avançadas integradas com gráficos que relacionem as métricas e dimensões permitindo realizar operações de filtragem e *drill-down* e realizar análises para responder a questões de negócios relevantes.
- Desenvolver um relatório técnico com a documentação completa e extensiva do processo de desenvolvimento das *pipelines* de dados, modelos de ciência de dados, visualizações e análises, além dos registros de homologação que comprovem a qualidade das transformações e processos de dados.



1. Os dados são coletados por um agente serverless desenvolvido em python que busca os dados de despesa pública e realiza o download deles no mesmo formato presente no Portal da Transparência;
2. Um segundo agente realiza uma transformação inicial nesses dados e coloca os dados já transformados numa segunda camada, onde poderão passar por transformações mais relacionadas ao consumo dos dados;
3. Utilizando um serviço específico da AWS (o AWS Glue) realizamos então a carga dos dados presentes nos arquivos em tabelas relacionais para o serviço de consulta e mecanismo de busca de dados da AWS (o AWS Athena);
4. Com o AWS Athena, realizamos outras transformações nos dados (que serão detalhadas posteriormente), para enriquecer os dados, cruzar tabelas e gerar métricas;
5. Utilizando o ambiente de machine learning da AWS (o AWS SageMaker) criamos e treinamos modelos a partir dos dados armazenados no Athena e com a saída dos modelos criamos novos dados;
6. Com todos os dados, desenvolvemos um Painel multidimensional com visualizações para o público estratégico, tático, operacional e de análises avançadas.
7. Usamos *jupyter notebooks* para o desenvolvimento dos protótipos de bases de dados, dashboards e modelos de inteligência artificial por conta da flexibilidade do python e sua vasta coleção de bibliotecas.

3.2. Engenharia de Dados

Essa seção descreve os processos de Engenharia de Dados aplicados para a realização da ingestão, extração e carga dos dados dessa pesquisa.

Para o desenvolvimento do trabalho implementamos uma arquitetura de dados baseada em 3 camadas: *Bronze*, *Silver* e *Gold*. Também chamada de *Medallion Architecture*, essa arquitetura promove maior organização, independência e autonomia dos processos que as alimentam e mantém. Utilizando tecnologias modernas de armazenamento e processamento, seja em nuvem (Amazon AWS, Microsoft Azure, Google Cloud Platform e etc.), seja em servidores locais (Databricks, Cloudera e etc.), podemos desenvolver complexos sistemas de integração e governança de dados que servem como base para aplicações modernas de Business Intelligence e Analytics.

A nossa implementação da Arquitetura de Dados segue o diagrama esquemático abaixo:

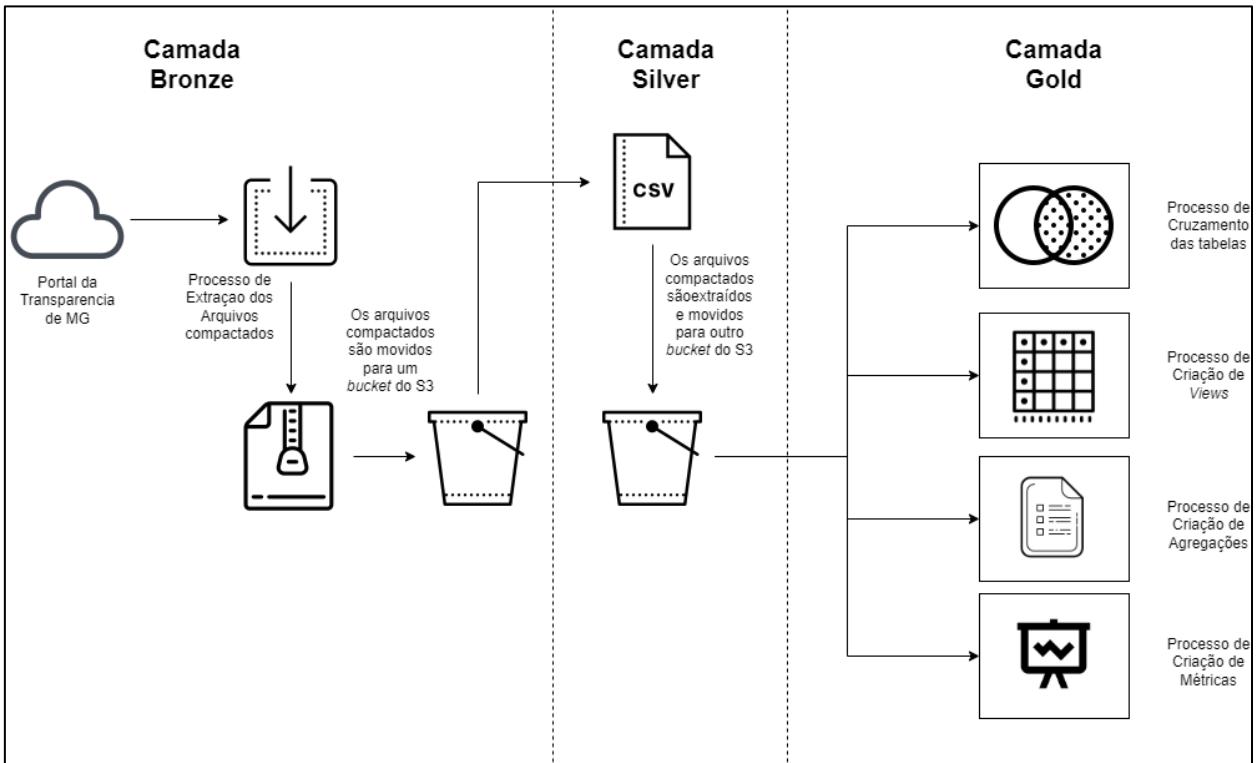


Imagen 36 – Diagrama da Arquitetura de Engenharia de Dados do processo.

Para cada camada realizamos um conjunto de processos de ETL. ETL pode ser definido como o processo complexo utilizado para coletar dados de diversas fontes (em diversos formatos, consistência, volumetria, qualidade e tempestividade), transformá-los (em diversas operações de agregação, filtragem, cruzamento e qualidade) e carregá-los (em bases performáticas, seja SQL ou No-SQL) para dar suporte à descoberta, à geração de relatórios, à análise e à tomada de decisões (31).

Processos de ETL por camada:

- **Camada Bronze:** Nessa camada os dados são coletados diretamente do site através de um *webcrawler*. O Webcrawler acessa a página principal dos dados de Despesa da Controladoria Geral do Estado de MG e realiza o download dos mesmos em formato compactado (.gz). Todos os arquivos baixados são então armazenados em um *bucket* do Amazon S3 (*simple storage service*) (18, 19, 20). Os dados baixados permanecem em formato *raw* ('cru' ou sem transformação), pois o propósito dessa primeira camada é servir como uma camada de recepção dos dados. Esse processo é realizado por um script em *python* que é executado em ambiente *serverless* (ou seja, não é necessário utilizar um servidor ou cluster para realizar a sua computação), oposto a isso, executamos tanto esse processo quanto o da *silver* a partir de outro serviço da Amazon o *AWS Lambda*. Esse serviço permite a criação de um ambiente de execução de código (Python, Javascript, Java entre outras opções) e a execução de código-fonte (21). Posterior ao download dos arquivos, esse processo ainda gera um arquivo de configuração com os hashes (identificadores únicos, baseados no conteúdo do arquivo) dos arquivos. Esse arquivo é lido toda vez que esse processo é executado, permitindo que o próprio processo identifique arquivos que já foram baixados anteriormente e cujo conteúdo não foi alterado (pois se houvessem alterações, o hash iria mudar), dessa forma o download do arquivo não é feito, assim temos um processo mais econômico com relação a latência e acesso ao portal do governo de Minas Gerais além de ser idempotente. O final do processo então temos 61 arquivos compactados e um arquivo (.json) com as configurações do processamento.
- **Camada Silver:** Nessa camada os arquivos são descompactados do formato original para o formato CSV (para cada um dos arquivos compactados é gerado um arquivo csv equivalente). Essa etapa apesar de ser mais simples, demanda muito mais de

processamento, por isso também criamos o script em python que realiza esse processamento usando o serviço AWS Lambda.

- **Camada Gold:** A última camada dessa arquitetura também é a mais complexa, pois é nessa camada que as transformações para o consumo dos dados são feitas. Nessa camada utilizamos o serviço de catálogo de dados da Amazon o *AWS Glue* (22), esse serviço serve, dentre outras coisas, para construir tabelas relacionais a partir de arquivos. O AWS Glue cria uma tabela, com um esquema inferido (tipos de dados são inferidos de acordo com uma amostra de dados para cada coluna) para cada arquivo, ou seja, ao final dessa etapa, temos 61 tabelas construídas. Além dessas tabelas também realizamos a criação de outras tabelas auxiliares para as fases de Ciência de Dados e Visualização de Dados:
 - **União de tabelas:** Nesse processo juntamos todas as tabelas *dm_empenhos* em uma única tabela de dados; outra tabela gerada a partir desse processo é a *ft_despesa* e a *ft_despesa_last24m* que contém a união de todas as tabelas *ft_despesa_<ano>*, sendo que a *ft_despesa_last24m* é a união somente das tabelas *ft_despesa_2021* e *ft_despesa_2022*.
 - **Criação de Views:** Nessa transformação criamos views específicas como a view ***vw_dw***, que é a principal tabela que usamos. Nessa view, cruzamos todos as dimensões com a fato *ft_despesa_last24m*, com o objetivo de obter os nomes dos valores de cada dimensão ao invés da chave-estrangeira. Esse processo de cruzar os dados na origem acelera muito o consumo de dados posteriormente em ferramentas de BI como o Power BI e Tableau, que apesar de terem mecanismos para realizar essas transformações não são tão performáticas quanto ferramentas de computação distribuída e em nuvem.
 - **Agregações:** Tabelas agregadas são tabelas que passam por um processo de agregação no qual uma função de agregação (média, soma, contagem, primeiro, último, dentre outras) é aplicada, reduzindo o número de linhas da tabela e diminuindo a granularidade. Usamos essa transformação para criar as views *vw_agg<dim>* (*vw_agg_categ_econ*, *vw_agg_grupo*, *vw_agg_procedencia*, etc).
 - **Criação de Métricas:** Essa transformação é usada para criar métricas (contagem e soma dos valores monetários) em cima da tabela fato *ft_despesa*. Como essa tabela é muito grande (mais de 65 milhões de linhas, por ser a união de todas as tabelas fatos), é necessário criar métricas sobre essa tabela para gráficos de mais alto nível, onde o interessante é olhar o todo e não o detalhe.

Obs.: Outras tabelas são geradas (*tb_anomaly_detection*, *vw_adv_analytics_forecast*, dentre outras), porém, essas são geradas para e pelos processos de ciência de dados que iremos descrever posteriormente.

```

if(isinstance(file_location_or_file_pointer, str)):
    if(mode == "buffered"):
        with open(file_location_or_file_pointer, 'rb') as f:
            while True:
                data = f.read(buffer_size)
                if not data:
                    break
                method_obj.update(data)
    else:
        data = read_file(file_location_or_file_pointer)
        method_obj.update(data)
elif(isinstance(file_location_or_file_pointer, io.TextIOWrapper)):
    raise Exception("Not implemented yet")
return method_obj.hexdigest()

def download_files(base_folder):
    """
    download_files:
    -----
    Download all data files defined by the base_url and saves them into a folder.
    """
    # --- 1. get initial data's website page -----
    base_url = "https://dados.mg.gov.br/dataset/despesa"
    r = requests.get(base_url, verify=False)
    page = bs(r.text, features="html.parser")

    # --- 2. get resources from initial page -----
    items = page.find_all('li', {'class' : 'resource-item'})
    ids = [item['data-id'] for item in items]
    links = [f"https://dados.mg.gov.br/dataset/despesa/resource/{i}" for i in ids]

    # --- 3. For every resource: get resource download link -----
    resources_links = []
    errs_links = []
    for link in links:
        try:
            resource_page = requests.get(link)
            resource_link = bs(resource_page.text, features="html.parser")
            href = resource_link.find('a', {'class' : "resource-url-analytics"}).attrs['href']
            resources_links.append(href)
        except Exception as e:
            errs_links.append({'resource' : link, 'err' : e})

    # --- 4. Do the downloads -----
    now_str = datetime.now().strftime("%Y-%m-%d-%H:%M:%S")
    errs_gz = []
    resources = []
    for rl in resources_links:

```

Imagen 37 - Imagem do código que constrói a camada Bronze – Esse código “baixa” os arquivos compactados do site do governo de Minas Gerais.

```

def listdir(folder):
    _s3r = boto3.resource('s3')
    s3_bckt = _s3r.Bucket(BUCKET_NAME)

    objs = [object_summary.key for object_summary in s3_bckt.objects.filter(Prefix=f'{folder}/')]
    return objs

def unzip_and_save(bucket_orig, file_orig, bucket_dest, file_dest):
    print("Unzip and Save", bucket_orig, file_orig, bucket_dest, file_dest)
    _s3 = boto3.client('s3', use_ssl=False) # optional
    _s3.upload_fileobj(
        Fileobj=gzip.GzipFile(None, 'rb',
            fileobj=BytesIO(_s3.get_object(Bucket=bucket_orig, Key=file_orig)['Body'].read())),
        Bucket=bucket_dest,
        Key=file_dest)

def list_to_unzip():
    objects = listdir(ORIGIN_PREFIX_PATH)
    print('objects', objects)
    errs_unzip = []
    for obj in objects:
        try:
            complete_filename = obj.split('/')[1]
            filename = '.'.join(complete_filename.split('.')[0:-1])
            if(filename.endswith('.csv')):
                # print(filename)
                unzip_and_save(BUCKET_NAME, f'{ORIGIN_PREFIX_PATH}/{filename}.gz', BUCKET_NAME, f'{DEST_PREFIX_PATH}/{filename}')
                pass
        except Exception as e:
            errs_unzip.append({'file' : obj, 'err' : e})

    if(len(errs_unzip)):
        print('errs_unzip', errs_unzip)

def lambda_handler(event, context):
    list_to_unzip()

```

Imagen 38 - Imagem do código de contrói a camada Silver – Esse código descompacta os arquivos e os coloca em outro bucket para serem processados posteriormente.

```

CREATE TABLE dm_empenhos_desp AS (
    SELECT * FROM "dm_empenho_desp_2002_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2003_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2004_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2005_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2006_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2007_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2008_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2009_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2010_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2011_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2012_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2013_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2014_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2015_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2016_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2017_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2018_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2019_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2020_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2021_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2022_csv" UNION ALL
    SELECT * FROM "dm_empenho_desp_2023_csv"
)

CREATE TABLE ft_despesa AS (
    SELECT * FROM "ft_despesa_2002_csv" UNION ALL
    SELECT * FROM "ft_despesa_2003_csv" UNION ALL
    SELECT * FROM "ft_despesa_2004_csv" UNION ALL
    SELECT * FROM "ft_despesa_2005_csv" UNION ALL
    SELECT * FROM "ft_despesa_2006_csv" UNION ALL
    SELECT * FROM "ft_despesa_2007_csv" UNION ALL
    SELECT * FROM "ft_despesa_2008_csv" UNION ALL
    SELECT * FROM "ft_despesa_2009_csv" UNION ALL
    SELECT * FROM "ft_despesa_2010_csv" UNION ALL
    SELECT * FROM "ft_despesa_2011_csv" UNION ALL
    SELECT * FROM "ft_despesa_2012_csv" UNION ALL
    SELECT * FROM "ft_despesa_2013_csv" UNION ALL
    SELECT * FROM "ft_despesa_2014_csv" UNION ALL
    SELECT * FROM "ft_despesa_2015_csv" UNION ALL
    SELECT * FROM "ft_despesa_2016_csv" UNION ALL
    SELECT * FROM "ft_despesa_2017_csv" UNION ALL
    SELECT * FROM "ft_despesa_2018_csv" UNION ALL
    SELECT * FROM "ft_despesa_2019_csv" UNION ALL
    SELECT * FROM "ft_despesa_2020_csv" UNION ALL
    SELECT * FROM "ft_despesa_2021_csv" UNION ALL
    SELECT * FROM "ft_despesa_2022_csv" UNION ALL
    SELECT * FROM "ft_despesa_2023_csv"
)

CREATE TABLE ft_despesa_last24m AS (
    WITH T AS (
        SELECT * FROM "ft_despesa_2021_csv" UNION ALL
        SELECT * FROM "ft_despesa_2022_csv"
    )
    SELECT
        CONCAT(
            CAST(id_tempo as varchar), '-',
            CAST(id_categ_econ as varchar), '-',
            CAST(id_grupo as varchar), '-',
            CAST(id_elemento as varchar), '-',
            CAST(id_item as varchar), '-',
            CAST(id_fonte as varchar), '-',
            CAST(id_modalidade_aplic as varchar), '-',
            CAST(id_funcao as varchar), '-',
            CAST(id_subfuncao as varchar), '-',
            CAST(id_programa as varchar), '-'
        )
)

```

Imagens 39 e 40 – Imagens dos códigos de união das tabelas. Esses códigos e os próximos fazem parte da camada gold.

```

CREATE OR REPLACE VIEW "vw_dw" AS
(
  SELECT
    ft_despesa_last24m.id id,
    dm_tempo_diario_csv.dia dia
   , dm_tempo_diario_csv.mes mes
   , dm_tempo_diario_csv.ano ano
   , dm_categ_econ_csv.nome categ_econ_nome
   , dm_grupo_desp_csv.nome grupo_nome
   , dm_elemento_desp_csv.nome elemento_desp_nome
   , dm_item_desp_csv.nome item_desp_nome
   , dm_fonte_csv.nome fonte_nome
   , dm_modalidade_aplic_csv.nome modalidade_aplic_nome
   , dm_funcao_desp_csv.nome funcao_desp_nome
   , dm_subfuncao_desp_csv.nome subfuncao_desp_nome
   , dm_programa_csv.nome programa_nome
   , dm_procedencia_csv.nome procedencia_nome
   , dm_unidade_orc_csv.grupo_administracao unidade_orc_grupo_administracao
   , dm_unidade_orc_csv.administracao unidade_orc_administracao
   , dm_unidade_orc_csv.nome unidade_orc_nome
   , dm_unidade_orc_csv.sigla unidade_orc_sigla
   , dm_favorecido_csv.nome_anonimizado favorecido_nome_anonimizado
   , dm_empenhos_desp.dt_empenho empenhos_desp_dt_empenho
   , dm_empenhos_desp.unidade_executora empenhos_desp_unidade_executora
   , dm_empenhos_desp.tipo_empenho empenhos_desp_tipo_empenho
   , dm_empenhos_desp.vr_empenho empenhos_desp_vr_empenho
   , dm_empenhos_desp.uni_prog_gasto empenhos_desp_uni_prog_gasto
   , dm_tipo_documento_csv.nome tipo_documento_nome
   , dm_situacao_op_desp_csv.nome situacao_op_desp_nome
   , vr_empenhado
   , vr_liquidado
   , vr_pago
  FROM
    (((((((((ft_despesa_last24m
    INNER JOIN dm_categ_econ_csv ON (ft_despesa_last24m.id_categ_econ = dm_categ_econ_csv.id_categ_econ))
    INNER JOIN dm_elemento_desp_csv ON (ft_despesa_last24m.id_elemento = dm_elemento_desp_csv.id_elemento)
    INNER JOIN dm_empenhos_desp ON (ft_despesa_last24m.id_empenho = dm_empenhos_desp.id_empenho))
    INNER JOIN dm_favorecido_csv ON (ft_despesa_last24m.id_favorecido = dm_favorecido_csv.id_favorecido))
    INNER JOIN dm_fonte_csv ON (ft_despesa_last24m.id_fonte = dm_fonte_csv.id_fonte))
    INNER JOIN dm_funcao_desp_csv ON (ft_despesa_last24m.id_funcao = dm_funcao_desp_csv.id_funcao)
    INNER JOIN dm_grupo_desp_csv ON (ft_despesa_last24m.id_grupo = dm_grupo_desp_csv.id_grupo)
    INNER JOIN dm_item_desp_csv ON (ft_despesa_last24m.id_item = dm_item_desp_csv.id_item)
    INNER JOIN dm_modalidade_aplic_csv ON (ft_despesa_last24m.id_modalidade_aplic = dm_modalidade_aplic_csv.id_modalidade_aplic))
    INNER JOIN dm_procedencia_csv ON (ft_despesa_last24m.id_procedencia = dm_procedencia_csv.id_procedencia)
    INNER JOIN dm_programa_csv ON (ft_despesa_last24m.id_programa = dm_programa_csv.id_programa))
    INNER JOIN dm_situacao_op_desp_csv ON (ft_despesa_last24m.tp_operacao = dm_situacao_op_desp_csv.id_situacao_op))
    INNER JOIN dm_subfuncao_desp_csv ON (ft_despesa_last24m.id_subfuncao = dm_subfuncao_desp_csv.id_subfuncao))
    INNER JOIN dm_tempo_diario_csv ON (ft_despesa_last24m.id_tempo = dm_tempo_diario_csv.id_tempo))
    INNER JOIN dm_tipo_documento_csv ON (ft_despesa_last24m.id_tipo_documento = dm_tipo_documento_csv.id_tipo_documento))
    INNER JOIN dm_unidade_orc_csv ON (ft_despesa_last24m.id_unidade_orc = dm_unidade_orc_csv.id_unidade_orc))

```

Imagen 41 – Imagem da visão que criamos para cruzar todas as tabelas dimensionais e fato para obter uma tabela desnormalizada que é superior em performance quando carregada no PowerBI.

```

CREATE OR REPLACE VIEW vw_agg_categ_econ AS
(
  SELECT
    ano_particao, id_categ_econ,
    count(*) as counting,
    sum(vr_empenhado) as s_vr_empenhado,
    sum(vr_liquidado) as s_vr_liquidado,
    sum(vr_pago) as s_vr_pago
  FROM ft_despesa GROUP BY ano_particao, id_categ_econ
  ORDER BY ano_particao, id_categ_econ
)

CREATE OR REPLACE VIEW vw_agg_grupo AS
(
  SELECT
    ano_particao, id_grupo,
    count(*) as counting,
    sum(vr_empenhado) as s_vr_empenhado,
    sum(vr_liquidado) as s_vr_liquidado,
    sum(vr_pago) as s_vr_pago
  FROM ft_despesa GROUP BY ano_particao, id_grupo
  ORDER BY ano_particao, id_grupo
)

CREATE OR REPLACE VIEW vw_agg_procedencia AS
(
  SELECT
    ano_particao, id_procedencia,
    count(*) as counting,
    sum(vr_empenhado) as s_vr_empenhado,
    sum(vr_liquidado) as s_vr_liquidado,
    sum(vr_pago) as s_vr_pago
)

```

Imagen 42 – Imagem da criação das views agregadas que servem de base para o gráfico de “Dimensões vs Anos”.

Além dessas transformações, ainda são realizadas outras transformações no próprio PowerBI por simplicidade. As transformações do PowerBI também fazem parte do fluxo de Engenharia de Dados.

```

let
  Source = AmazonAthena.Databases("PBDSN2", null),
  AwsDataCatalog_Database = Source[[Name="AwsDataCatalog",Kind="Database"]][Data],
  "#dep-puc_Schema" = AwsDataCatalog_Database[[Name="dep-puc",Kind="Schema"]][Data],
  vw_dw_Table = "#dep-puc_Schema"[Name="vw_dw",Kind="Table"][[Data]],

  #"[Added Conditional Column] vr_pago_2021" = Table.AddColumn(vw_dw_Table, "vr_pago_2021", each if [ano] = 2021 then [vr_pago] else 0),
  #"[Added Conditional Column] vr_pago_2022" = Table.AddColumn(#"[Added Conditional Column] vr_pago_2021", "vr_pago_2022", each if [ano] = 2022 then [vr_pago] else 0),

  #"[Added Conditional Column] vr_empenhado_2021" = Table.AddColumn(#"[Added Conditional Column] vr_pago_2022", "vr_empenhado_2021", each if [ano] = 2021 then [vr_empenhado]),
  #"[Added Conditional Column] vr_empenhado_2022" = Table.AddColumn(#"[Added Conditional Column] vr_empenhado_2021", "vr_empenhado_2022", each if [ano] = 2022 then [vr_empenhado]),

  #"[Added Conditional Column] vr_liquidado_2021" = Table.AddColumn(#"[Added Conditional Column] vr_empenhado_2022", "vr_liquidado_2021", each if [ano] = 2021 then [vr_liquidado]),
  #"[Added Conditional Column] vr_liquidado_2022" = Table.AddColumn(#"[Added Conditional Column] vr_liquidado_2021", "vr_liquidado_2022", each if [ano] = 2022 then [vr_liquidado]),

  #"[Changed Type]" = Table.TransformColumnTypes(#"[Added Conditional Column] vr_liquidado_2022",{
    {"vr_pago_2021", type number}, {"vr_pago_2022", type number},
    {"vr_empenhado_2021", type number}, {"vr_empenhado_2022", type number},
    {"vr_liquidado_2021", type number}, {"vr_liquidado_2022", type number}
  })
in
  #"[Changed Type]"

```

Imagen 43 – Nessa imagem descrevemos um exemplo de transformação feito no PowerBI, 6 colunas são criadas dentro do PBI através de funções internas.

3.3. Ciência de Dados

Essa seção descreve os modelos e métodos aplicados para a geração das tabelas-resultado dos modelos.

A utilização de *machine learning* e inteligência artificial para a construção de painéis de controle mais eficientes e explicativos tem sido adotada em larga escala nos últimos tempos. Esses modelos podem ser usados nas etapas de **preparação de dados**, permitindo que análises padrão de distribuição dos dados (médias, medianas, desvio-padrão etc.), verificação de qualidade integridade de dados, enriquecimento de dados e *profiling*, sejam feitas de forma

semiautomática ou completamente automática. Também é possível usar essas técnicas para automatizar tarefas de **visualização de dados**, desde a parte de design e exploração de dados até a construção de gráficos em si. Outro importante uso, é durante o processo de **análise de dados** onde modelos são construídos para descobrir padrões, realizar tarefas de mineração de dados, classificar registros em diferentes categorias e realizar previsões a partir de dados históricos (32).

Modelos de *machine learning* são modelos matemáticos implementados utilizando alguma linguagem de programação que passaram por um processo de *treinamento* a partir de um conjunto de dados (categorizados ou não) para a realização de uma atividade que exigiria conhecimento humano (classificar objetos, prever a demanda futura, conversar usando linguagem natural, identificar objetos em imagens, traduzir textos, criar legendas de áudios etc.) para dados desconhecidos (33).

Como parte da solução proposta por esse trabalho, desenvolvemos 3 modelos de aprendizado de máquina para fornecer aos usuários do Painel de Visualização de Dados informações extremamente úteis no processo de tomada de decisão, como por exemplo, uma previsão dos eventuais custos que pode ser usada para o planejamento do orçamento.

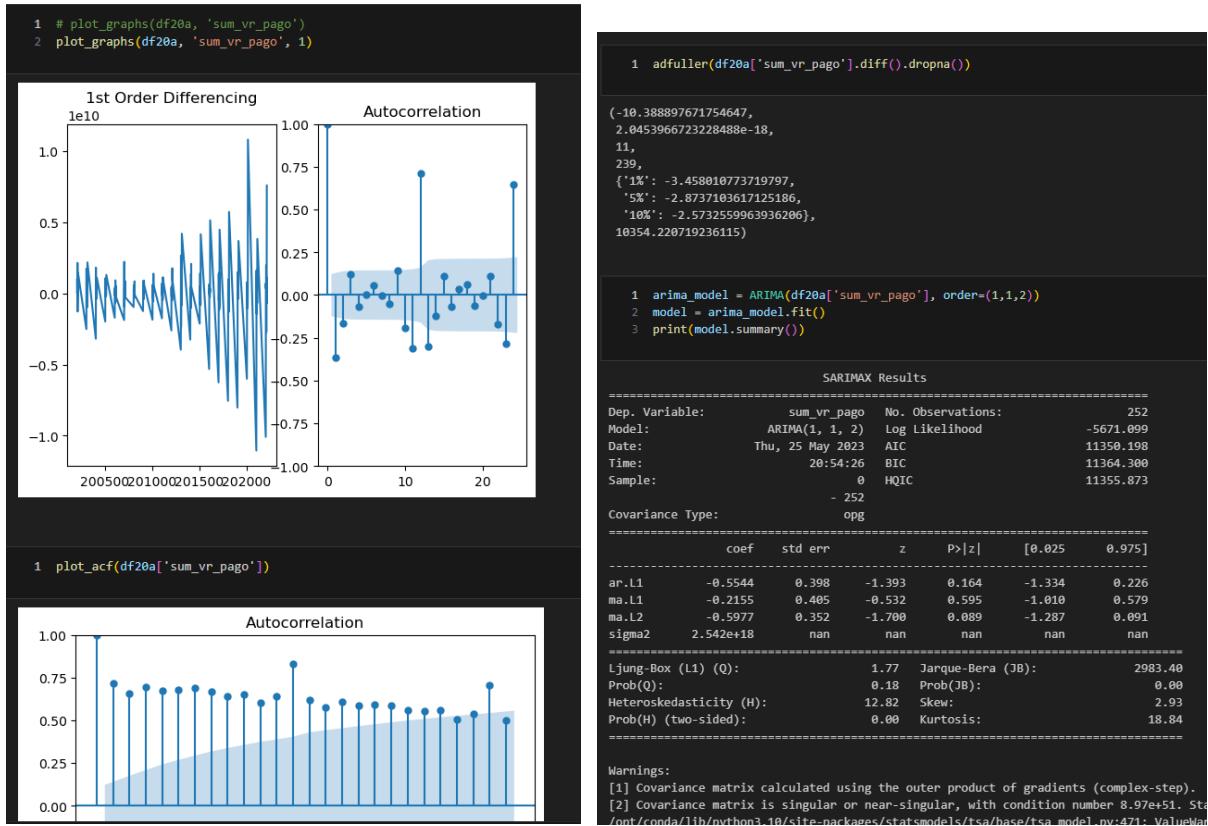
3.3.1 Modelo de Forecast

Modelo desenvolvido a partir do método estatístico ARIMA (AutoRegressive Integrated Moving Average), esse é um modelo popular de previsão de séries temporais usado para analisar e prever pontos de dados em uma sequência ordenada no tempo. É uma combinação de três componentes:

- Componente AutoRegressivo (AR): O componente AR envolve a modelagem da relação entre o ponto de dados atual e seus pontos de dados anteriores. Ele assume que o valor atual da série depende de seus valores passados. O parâmetro "p" representa o número de pontos de dados anteriores usados no modelo AR.
- Componente Integrado (I): O componente I refere-se à diferenciação dos dados da série temporal para torná-los estacionários. Estacionariedade significa que as propriedades estatísticas da série (como média e variância) permanecem constantes ao longo do tempo. A diferenciação envolve a subtração do ponto de dados anterior do ponto de dados atual. O parâmetro "d" representa o número de etapas de diferenciação necessárias para atingir a estacionariedade.
- Componente de Média Móvel (MA): O componente MA considera a relação entre o ponto de dados atual e os erros de previsão passados. Ele assume que o valor atual da série depende dos termos de erro das previsões anteriores. O parâmetro "q" representa o número de erros de previsão defasados usados no modelo MA.
- Ao combinar esses três componentes, o modelo ARIMA pode capturar vários padrões e tendências presentes nos dados da série temporal e fazer previsões para pontos de dados futuros. A ordem do modelo é representada como (p, d, q), onde "p", "d" e "q" são os parâmetros correspondentes aos componentes AR, I e MA, respectivamente.
- O modelo ARIMA é amplamente utilizado em áreas como finanças, economia, previsão do tempo e previsão de vendas, onde padrões históricos em dados de séries temporais podem fornecer informações valiosas para prever tendências futuras e tomar decisões informadas.

Desenvolvemos um modelo ARIMA (usando o pacote *statsmodels* para python (38)) com parâmetros **(1, 1, 2)** em suma esse modelo usa o valor lag-1 (a primeira diferença, ou a comparação entre um “*data point*” e o seu valor anterior imediato) como valor para a primeira componente do modelo (AR – autoregressive). Também temos 1 na segunda componente (I) que indica a diferença necessária para tornar a série estacionária, nesse caso somente a

primeira diferença já o faria. Por fim na componente MA (médias móveis) indica que os dois últimos erros de previsão são usados para fazer previsões para os dados da série temporal. O ajuste do modelo também foi realizado utilizando os gráficos das Funções de Autocorrelação e Função de Autocorrelação Parcial.



Imagens 44 e 45 – Imagens do modelo de forecast – A primeira imagem é o *plot* da diferença de primeira ordem do modelo e a sua autocorrelação e a segunda imagem é a saída do modelo ARIMA.

A saída desse modelo (em csv) forneceu os dados para a construção da tabela *tb_prediction_mean*, nessa tabela armazenamos os valores da média da previsão dos valores para a métrica *valor_pago* por mês para os próximos 15 meses. Esses dados foram lidos de forma automática pelo AWS Glue e foi criada essa tabela para armazená-los. Com esses dados construímos a visão *vw_adv_analytics_forecast* que combina os dados dessa tabela com os dados já presentes em na tabela-fato de despesas a *ft_despesa*.

```

CREATE OR REPLACE VIEW "vw_adv_analytics_forecast" AS
(
  SELECT
    dt_anomes
  , "sum"(vr_pago) sum_vr_pago
  , 'REALIZADO' status
  FROM
    "dep-puc"."ft_despesa"
  GROUP BY dt_anomes
  UNION ALL   SELECT
    dt_anomes
  , sum_vr_pago
  , 'FORECAST' status
  FROM
    "dep-puc"."tb_prediction_mean"
)

```

Imagen 46 – Imagem da criação da view *vw_adv_analytics_forecast*.

3.3.2 Modelo de Regras de Associação

Um modelo de regras de associação é uma técnica de mineração de dados usada para descobrir relacionamentos, padrões ou associações interessantes em grandes conjuntos de dados. É particularmente útil para analisar dados transacionais, como históricos de compras de clientes, onde o objetivo é encontrar co-ocorrências comuns ou relacionamentos entre itens (35).

O modelo de regras de associação funciona com base em duas medidas principais:

1. Suporte: O suporte refere-se à frequência ou à proporção de transações que contêm um conjunto de itens específico (uma coleção de um ou mais itens). Indica a frequência com que um conjunto de itens aparece no conjunto de dados.
2. Confiança: A confiança mede a probabilidade de um item B ser comprado quando o item A é comprado. É calculado como a razão entre o suporte do conjunto de itens {A, B} e o suporte do conjunto de itens {A}. Em outras palavras, a confiança quantifica a força da associação entre dois itens.

O modelo identifica regras de associação com limites mínimos de suporte e confiança. Esses limites permitem que o modelo filtre associações menos frequentes e menos significativas, concentrando-se nos padrões mais relevantes dos dados.

Por exemplo, em um ambiente de varejo, um modelo de regras de associação pode descobrir que os clientes que compram o produto A têm grande probabilidade de comprar o produto B também, com uma confiança de 80%. Essas informações podem ser valiosas para os varejistas otimizarem suas colocações de produtos, estratégias de vendas cruzadas e campanhas de marketing direcionadas.

O algoritmo mais conhecido para gerar regras de associação é o algoritmo Apriori. Ele explora com eficiência os conjuntos de itens nos dados para descobrir conjuntos de itens frequentes e gerar regras de associação relevantes.

As regras de associação são amplamente utilizadas em vários campos, como análise de cesta de compras, análise de comportamento do cliente, sistemas de recomendação e muito mais. Eles fornecem informações valiosas sobre as relações entre os itens e podem ajudar as empresas a tomar decisões informadas para aumentar a satisfação do cliente, aumentar as vendas e melhorar a eficiência geral.

Implementamos o modelo de regras de associação a partir da biblioteca mlxtend (39) do python. A partir dessa biblioteca criamos um modelo *apriori* com suporte mínimo de 0.22 para limitar o número de *itemsets* e número de regras (impedindo a criação de regras com pouca significância estatística). Também usamos como medida para limitar o número de regras a métrica *lift* para filtrarmos somente associações positivas.

```

1 groups = list(df_favors['grupo'].apply(lambda g : g.replace('[', '').replace(']', '').split(',') ))
2
3
4 te = TransactionEncoder()
5 te_ary = te.fit(groups).transform(groups)
6 df = pd.DataFrame(te_ary, columns=te.columns_)

1 frequent_itemsets = apriori(df, min_support=0.022, use_colnames=True)
2 print(frequent_itemsets)
3
4 rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
5 # print("\nAlgumas regras de associação geradas:\n", rules.head())
6 print("\nDimensões da matriz de regras:", rules.shape)
7 rules

support           itemsets
0 0.210426          ()
1 0.023069          ()
2 0.023222          ( ADQUIRIR AMPLIFICADOR DE AUDIO)
3 0.023611          ( ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDU...
4 0.040072          ( AUXÍLIOS E MONITORAMENTO)
5 0.022482          ( AVALIAÇÃO EXTERNA - SAEB)
6 0.023855          ( CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔ...
7 0.022337          ( EXECUTAR AVALIAÇÕES INTERNA)
8 0.024100          ( EXECUTAR O PROGRAMA DE MANUTENÇÃO PREDIAL)
9 0.024290          ( EXECUTAR O PROJETO FORTALECIMENTO DAS APREND...
10 0.023779         ( EXTERNA E DE CERTIFICAÇÃO)
11 0.024199         ( FUNDO EMERGENCIAL DE PREVENÇÃO AS CHUVAS)
12 0.022772         ( MOBILIÁRIO DE INCLUSÃO)
13 0.022902         ( REALIZAR AVALIAÇÃO INTERNA)
14 0.024252         ( REPASSAR RECURSOS PARA MANUTENÇÃO DE ESCOLAS)
15 0.147529         (APOIO A FORMAÇÃO PROFISSIONAL PRONATEC - FIC ...
16 0.052702         (DEPARTAMENTO DE PROGRAMAS DE BOLSAS E EVENTOS...
17 0.024321          (FHEMIG/ADC)
18 0.025557          (GASTOS GERAIS)
19 0.027267          (GERAL)

```

Imagen 47 – Imagem do modelo de regra de associações.

A saída do modelo é dada em um arquivo csv que é lido pelo AWS Glue para a criação da tabela *tb_association_rules*. Nessa tabela temos todas as métricas que o modelo nos dá como resultado, apesar de somente algumas delas exibirmos na visualização.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔ...	(ADQUIRIR AMPLIFICADOR DE AUDIO)	0.023855	0.023222	0.022085	0.925784	39.866506	0.021531	13.161240	0.998742
1	(ADQUIRIR AMPLIFICADOR DE AUDIO)	(CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔ...	0.023222	0.023855	0.022085	0.951035	39.866506	0.021531	19.935622	0.998094
2	(EXECUTAR O PROJETO FORTALECIMENTO DAS APREND...	(ADQUIRIR AMPLIFICADOR DE AUDIO)	0.024290	0.023222	0.022169	0.912661	39.301409	0.021605	11.183756	0.998817
3	(ADQUIRIR AMPLIFICADOR DE AUDIO)	(EXECUTAR O PROJETO FORTALECIMENTO DAS APREND...	0.023222	0.024290	0.022169	0.954650	39.301409	0.021605	21.515102	0.97725
4	(FUNDO EMERGENCIAL DE PREVENÇÃO AS CHUVAS)	(ADQUIRIR AMPLIFICADOR DE AUDIO)	0.024199	0.023222	0.022184	0.916746	39.477297	0.021622	11.732435	0.998840
5	(ADQUIRIR AMPLIFICADOR DE AUDIO)	(FUNDO EMERGENCIAL DE PREVENÇÃO AS CHUVAS)	0.023222	0.024199	0.022184	0.955307	39.477297	0.021622	21.833550	0.97841
6	(ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDU...	(CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔ...	0.023611	0.023855	0.022192	0.939884	39.399044	0.021629	16.237587	0.998187
7	(CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔ...	(ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDU...	0.023855	0.023611	0.022192	0.930262	39.399044	0.021629	14.000877	0.998437
8	(ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDU...	(EXECUTAR O PROGRAMA DE MANUTENÇÃO PREDIAL)	0.023611	0.024100	0.022093	0.935682	38.825468	0.021524	15.173043	0.997803
9	(EXECUTAR O PROGRAMA DE MANUTENÇÃO PREDIAL)	(ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDU...	0.024100	0.023611	0.022093	0.916719	38.825468	0.021524	11.724090	0.998302

Imagen 48 – Imagem dos resultados obtidos a partir da execução do modelo de regras de associação.

3.3.3 Modelo de Detecção de Anomalia

Modelo de detecção de outlier é uma técnica de análise de dados usada para identificar pontos de dados incomuns ou anômalos em um conjunto de dados. Outliers são pontos de dados que se desviam significativamente da maioria dos dados e podem indicar erros, eventos raros ou padrões interessantes que requerem uma investigação mais aprofundada. A detecção de outliers é comumente usada em vários campos, incluindo finanças, segurança cibernética, manufatura entre outras (36).

O objetivo principal de um modelo de detecção de outliers é separar os pontos de dados normais dos outliers. Ele faz isso aproveitando várias técnicas estatísticas e de aprendizado de máquina para quantificar o grau de anormalidade de cada ponto de dados. Depois que o modelo identifica possíveis discrepâncias, análises ou ações adicionais podem ser realizadas com base na natureza dos dados e no aplicativo específico.

Um algoritmo popular para detecção de outliers é o modelo *Isolation Forest* (37).

O modelo Isolation Forest é um algoritmo de aprendizado de máquina não supervisionado baseado em árvore projetado especificamente para detecção de outliers. Ele opera particionando recursivamente os dados em subconjuntos e isolando outliers em caminhos curtos pela árvore. A principal intuição por trás da Isolation Forest é que os outliers são mais facilmente separáveis e requerem menos partições para isolá-los em comparação com os pontos de dados normais. O processo de isolamento de outliers é o seguinte:

- Seleção aleatória: a Floresta de isolamento seleciona um recurso aleatório e um valor de divisão aleatório dentro do intervalo do recurso selecionado para cada nó na árvore.
- Particionamento recursivo: os pontos de dados são divididos recursivamente com base no recurso selecionado e no valor dividido, criando uma estrutura de árvore binária.
- Isolamento: Espera-se que os outliers que se desviam significativamente da maioria dos dados sejam isolados em caminhos mais curtos da árvore devido à sua singularidade. Os pontos de dados normais, por outro lado, exigem que mais partições sejam isoladas.
- Pontuação de outlier: O modelo atribui uma pontuação de anomalia a cada ponto de dados com base na profundidade média na qual ele é isolado em várias árvores. Pontuações mais baixas indicam maior probabilidade de ser um outlier.

O modelo Isolation Forest é computacionalmente eficiente e pode lidar bem com conjuntos de dados de alta dimensão. É particularmente eficaz na detecção de outliers globais que estão longe da maioria dos dados. No entanto, pode não funcionar tão bem na detecção de outliers locais que estão rodeados por pontos de dados semelhantes.

Em resumo, um modelo de detecção de outlier visa identificar pontos de dados incomuns, e o algoritmo Isolation Forest é uma técnica poderosa para realizar essa tarefa com eficiência e eficácia.

Para o nosso modelo usamos como parâmetro o valor 0.1 de contaminação, indicando que poderíamos encontrar até 10% de valores indicados como outliers.

```
1 query = '''  
2 | SELECT id, vr_empenhado, vr_liquidado, vr_pago FROM "dep-puc"."ft_despesa_last24m"  
3 |'''  
4 df_vals = pd.read_sql(query, conn)  
  
/tmp/ipykernel_211/1041627377.py:4: UserWarning: pandas only supports SQLAlchemy connectable (e  
df_vals = pd.read_sql(query, conn)  
  
1 df_vals.head()  
  


|   |                                                   | id | vr_empenhado | vr_liquidado | vr_pago |
|---|---------------------------------------------------|----|--------------|--------------|---------|
| 0 | 43949-20-42-502-3012-486-130-1961-3988-14363-5... |    | 0.0          | 913.5        | 0.0     |
| 1 | 43949-20-42-502-3012-486-130-1961-3988-14363-5... |    | 273.0        | 0.0          | 0.0     |
| 2 | 43949-20-42-502-3012-58-130-1961-3967-14341-56... |    | -352.5       | -352.5       | 0.0     |
| 3 | 43949-20-42-502-3012-58-130-1961-3967-14380-56... |    | -163.0       | -163.0       | 0.0     |
| 4 | 43949-20-42-502-3012-486-130-1961-3961-14310-5... |    | 0.0          | 756.0        | 0.0     |

  
1 ids = list(df_vals['id'])  
2 del df_vals['id']  
  
1 data = df_vals.values  
2 # identify outliers in the training dataset  
3 iso = IsolationForest(contamination=0.1)  
4 yhat = iso.fit_predict(data)  
5 # select all rows that are not outliers  
6 mask = yhat != -1
```

Imagen 49 – Imagen do modelo de detección de anomalia.

O modelo nos gera um arquivo csv que é lido pelo AWS Glue para a criação da tabela *tb_anomaly_detection*. Os dados dessa tabela são então usados por uma query para a construção de views específica: *vw_adv_analytics_anomaly_detection* e *vw_adv_analytics_anomaly_detection_agents*.

```

CREATE OR REPLACE VIEW "vw_adv_analytics_anomaly_detection" AS (
    | SELECT dw.*, ad.anomaly FROM "dep-puc"."vw_dw" as dw JOIN "dep-puc"."tb_anomaly_detection" as ad ON dw.id = ad.id
)
CREATE OR REPLACE VIEW "vw_adv_analytics_anomaly_detection_metrics" AS (
    | SELECT
        |     anomaly,
        |     MIN(vr_empenhado) as min_vr_empenhado, MAX(vr_empenhado) as max_vr_empenhado, AVG(vr_empenhado) avg_vr_empenhado,
        |     MIN(vr_liquidado) as min_vr_liquidado, MAX(vr_liquidado) as max_vr_liquidado, AVG(vr_liquidado) avg_vr_liquidado,
        |     MIN(vr_pago) as min_vr_pago, MAX(vr_pago) as max_vr_pago, AVG(vr_pago) avg_vr_pago
    |
    |     FROM "dep-puc"."vw_dw" as dw
    |     JOIN "dep-puc"."tb_anomaly_detection" as ad
    |     ON dw.id = ad.id
    |     GROUP BY anomaly
)
CREATE OR REPLACE VIEW "vw_adv_analytics_anomaly_detection_agents" AS (
    | SELECT unidade_orc_nome, favorecido_nome_anonimizado, vr_empenhado, vr_liquidado, vr_pago, ad.anomaly
    |     FROM "dep-puc"."vw_dw" as dw JOIN "dep-puc"."tb_anomaly_detection" as ad ON dw.id = ad.id WHERE ad.anomaly = 'False'
)

```

Imagen 50 – Imagem do código SQL utilizado para criar as tabelas e visões específicas a partir dos resultados intermediários dos modelos.

3.4. Visualização de Dados

Essa seção descreve o funcionamento do Painel de Visualização de Dados.

Painéis de Visualização de Dados são importantes ferramentas de Business Intelligence e Analytics pois permitem de forma prática e direta relacionar dados a fim de obter informações.

provem técnicas para realizar investigações mais profundas e habilitam responder questões relevantes de negócio e acompanhar indicadores-chave.

O PowerBI é uma das ferramentas mais utilizadas para esse fim (28), por conta disso, foi a ferramenta que optamos para implementar o painel de controle de dados contendo as visualizações e principais análises desse trabalho.

Para o desenvolvimento do Painel de Visualização de Dados desenvolvemos no PowerBI um conjunto de Dashboards com quatro visões para diferentes públicos-alvo, são elas:

1. **Visão Estratégica:** Na visão estratégica o objetivo é listar os números em uma visão mais ampla, sem muitos detalhes para realizar um acompanhamento macro e alertar caso algo urgente esteja errado. Nessa visão podemos analisar indicadores chave como os valores empenhados e pagos por ano, além de vermos quantas dimensões temos e a quantidade de registros com que estamos trabalhando em cada dimensão. Essa visão é principalmente usada por Presidentes e Diretores, sendo que no nosso caso poderia ser usado pelo Governador e seus assessores diretos.
2. **Visão Tática:** Na visão tática temos maior nível de detalhamento que na visão anterior, dessa forma, é interessante a sua utilização por Supervisores e Coordenadores, dentro do escopo do governo, esses dados poderiam ser vistos pelos líderes de cada secretaria do governo (uma vez que não temos os dados divididos pelos municípios não podemos delegar essa responsabilidade aos prefeitos). As análises nessa visão são mais detalhadas e permitem a realização de mais filtros.
3. **Visão Operacional:** Na visão operacional temos todos os dados disponíveis no nível mais granular possível, ou seja, com maior número de detalhes e menor agregação. Essa visão serve para um acompanhamento crítico e minucioso, cabendo análises de itens de despesa, lista de favorecidos em cada pagamento e detalhes de cada transação financeira. Essa visão pode ser usada pela sociedade e demais órgãos para a realização de auditoria, por exemplo.
4. **Visões avançadas:** Essas visões têm como objetivo prever valores futuros, criar regras baseado nos dados e classificar registros como outliers. Podem ser usadas em conjunto com as demais visões para a criação de novos insights como por exemplo, entender a previsão dos gastos do governo por área ou unidade orçamentária.

Observação: É válido destacar que os seguintes itens se aplicam para todos os gráficos como forma de melhorar a experiência do usuário na utilização do painel de controle:

- Usamos os mesmos símbolos para os ícones;
- Criamos submenus para facilitar a navegação;
- Temos um botão de “limpar os filtros” em cada painel;
- Todos os gráficos possuem *tooltip* com informações mais detalhadas sobre o gráfico em destaque;
- Todos os gráficos filtram os demais gráficos dentro do mesmo painel;



Imagen 51 – Imagem da capa do painel.

Capa: Nessa primeira tela exibimos o menu principal que dá acesso a todas os submenus que irão conter os links para cada painel do Dashboard. Além disso temos links externos para as páginas profissional do autor e do repositório de código-fonte. Note que todos os ícones usados ao longo do dashboard são os mesmos garantindo maior facilidade de navegação do painel e melhorando a experiência do usuário.

Imagen 52 – Imagem do submenu do Dashboard Estratégico.

Submenu de Dashboard Estratégico: Nesse submenu temos os painéis estratégicos. Temos 2 painéis estratégicos que permitem ter uma visão ampla dos dados que estamos trabalhando nesse dashboard. Esses painéis possuem filtros para manipulação dos dados e trabalham com mais de 20 anos de dados históricos, apresentando as visões de “Grandes Números” e “Dimensões vs Anos”.



Imagen 53 – Imagem do Dashboard de Grande Números.

Painel de Grandes Números: Esse painel apresenta o comportamento das principais métricas que temos à disposição no nosso conjunto de dados: valores empenhado, liquidado e pago, além de uma métrica adicional que criamos que é o número de registros por ano (número que serve para entendermos quantas despesas estão sendo criadas todos os anos).

Nesse painel temos o recurso de apresentação automática. O “player ano” quando pressionado altera a visão do painel aplicando o filtro de ano mudando de um ano para o outro (desde 2002 até 2022) permitindo ver a evolução e mudança dos valores das métricas de forma dinâmica. O filtro de ano também pode ser aplicado de forma estática, permitindo a seleção de mais de um ano inclusive para comparação.



Imagen 54 – Imagem do Dashboard de Dimensões x Anos.

Painel de Dimensões vs Anos: Nesse painel apresentamos as principais dimensões usadas na construção do dashboard e a quantidade de valores únicos de cada uma (na figura, os 18 quadrados à direita), isso é importante para entendermos como os *drill-downs* são construídos

nos demais gráficos do dashboard. Além disso também temos a apresentação dos valores empenhado, liquidado e pago distribuídos em 6 dimensões específicas olhando para os últimos 24 meses de dados (anos de 2021 e 2022). Note que os 6 gráficos à esquerda mudam de acordo com o filtro da métrica selecionada, na imagem a métrica selecionada é o Valor Empenhado.

The image shows a screenshot of the 'Dashboard Tático' submenu. On the left, there is a sidebar with a 'Dashboard Tático' icon and a descriptive text about its purpose. The main area contains three rounded rectangular boxes, each representing a different dashboard panel:

- Ano vs Ano:** Describes the 'Ano vs Ano' dashboard, which allows comparing two periods of major relevance (2021 and 2022) to verify the evolution of key metrics (values allocated, liquidated, and paid) across public expenditure areas. It includes a small screenshot of the dashboard interface.
- Composição:** Describes the 'Composição' dashboard, which provides insight into how public expenses of the Minas Gerais Government are composed over time, categorized by logical divisions (Favorecido / Fonte / Procedência / Categoría Económica). It includes a small screenshot of the dashboard interface.
- Fluxo:** Describes the 'Fluxo' dashboard, which is essential for understanding the allocation of values spent by the Government. It allows changing resource capture dimensions (Input values) and destination dimensions (Output values). It includes a small screenshot of the dashboard interface.

PUC Minas logo is visible in the bottom right corner.

Imagen 55 – Imagem do submenu de Dashboard Tático.

Submenu do Dashboard Tático: Nesse submenu apresentamos os três painéis que compõem a análise tática do dashboard. Nessa etapa de análise a ideia é permitir ao usuário maior controle sobre filtros e visões mais específicas por área ou assunto. Como os dados estão em um nível mais granular do que os dados estratégicos, nesse conjunto de painéis e no conjunto de painéis operacional olhamos somente para os dados dos últimos 24 meses, essa visão permite além de maior foco nos dados mais atuais maior velocidade nas análises. Os painéis apresentados nessa seção são: o painel de “Ano vs Ano” que permite ver o último (2022) ano versus o penúltimo ano (2021) comparando os valores por categoria de despesa, o painel de “Composição” que possui uma visão da composição do orçamento e dos gastos, bem como da maneira como estes são distribuídos ao longo do tempo. Por fim também temos a visão da movimentação dos valores pagos entre as origens e destinos a partir do painel de “Fluxo”.

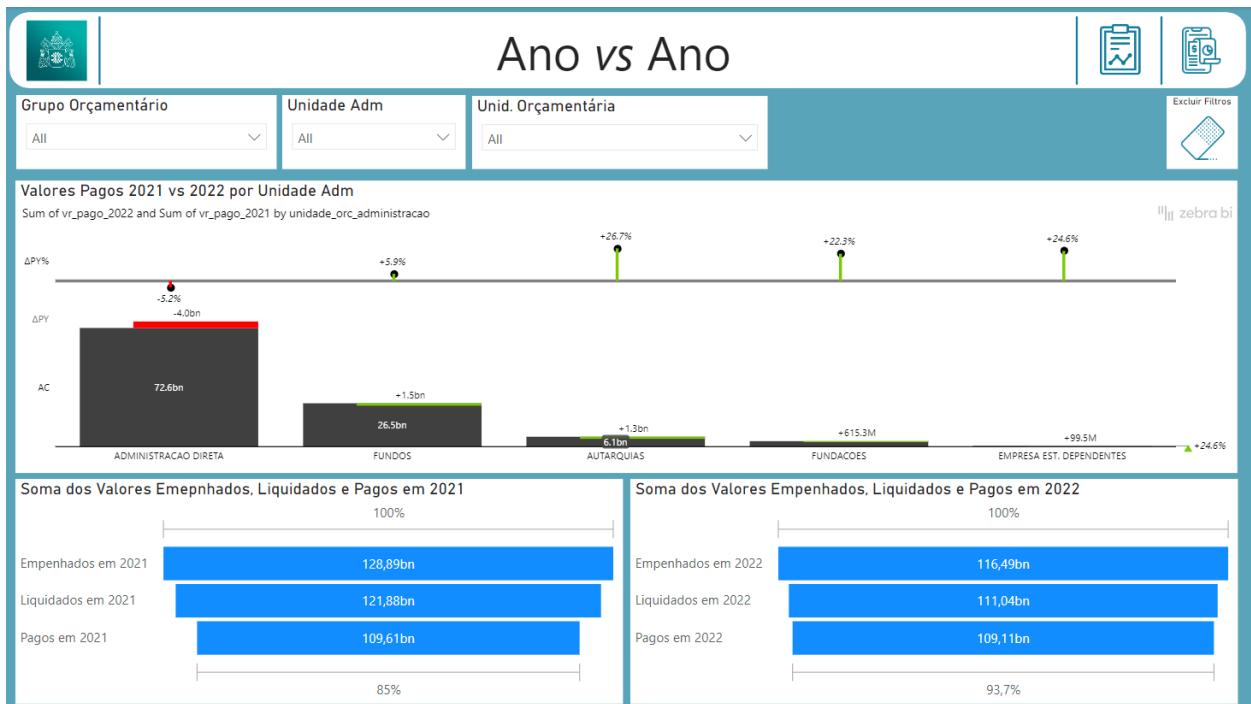


Imagen 56 – Imagem do Dashboard de Ano vs Ano.

Painel de Ano vs Ano: Através desse painel podemos ter uma visão comparativa entre os dois últimos anos, olhando os principais indicadores de valores empenhados, liquidados e pagos em formato de funil, mostrando dessa forma, como os valores que haviam sido planejados a princípio foram de fato usados em cada ano. Além disso também é possível entender a evolução (crescimento ou decrescimento) do valor pago por unidade administrativa (gráfico na porção superior do painel). Esse último painel por se tratar de um painel provido por uma extensão, permite que tenhamos mais visões comparando a evolução entre as unidades administrativas e extraíndo insights interessantes de como os valores são alocados por área em cada um dos anos.

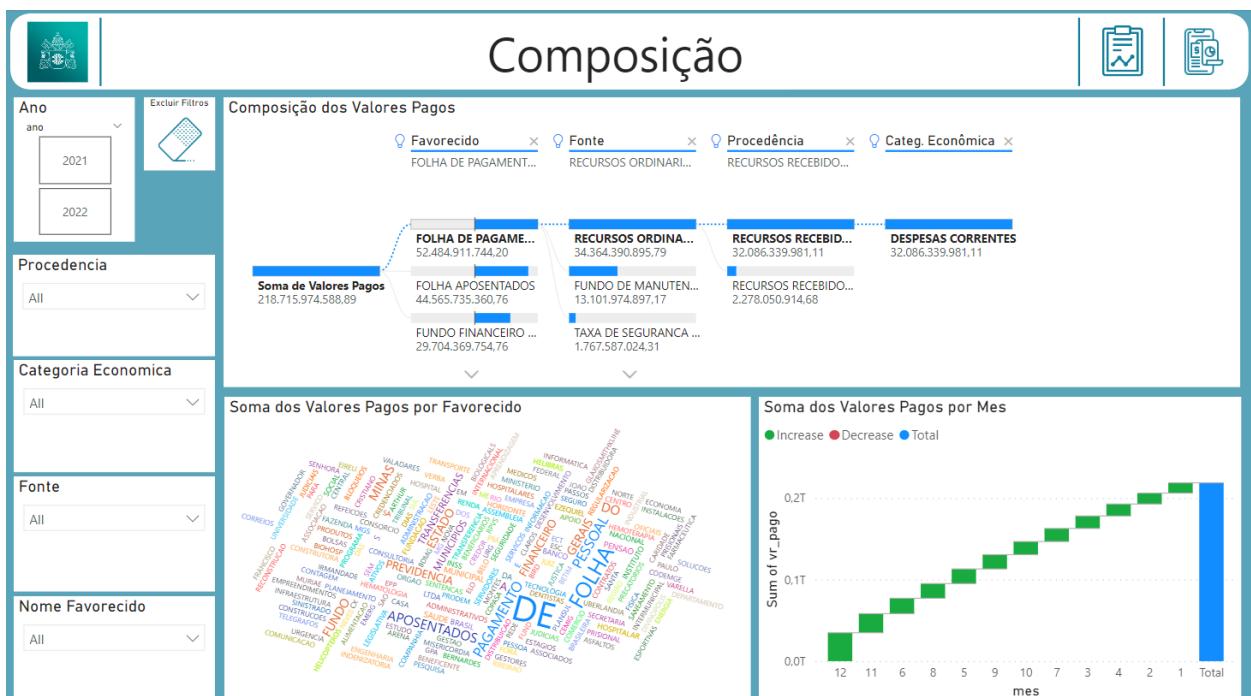


Imagen 57 – Imagem do Dashboard de Composição.

Painel de Composição: Esse painel permite visualizar como as receitas e despesas são compostas. Apesar dos dados terem mais referências às despesas, algumas dimensões como fonte e procedência indicam origens de receita. Essas dimensões e outras dimensões que apresentam onde os recursos são usados são cruzadas nesse painel para dar a ideia de como

estão compostas. Ademais, é possível ver também a composição dos valores pagos através do tempo no gráfico de valores pagos por mês. Essa visão permite termos noção da sazonalidade dos pagamentos, quando não selecionamos um ano específico, ou permite ver como dentro do ano é composta a dívida do estado de Minas Gerais. Vários filtros relativos aos dados que estão presentes nos gráficos são exibidos no canto esquerdo do painel.

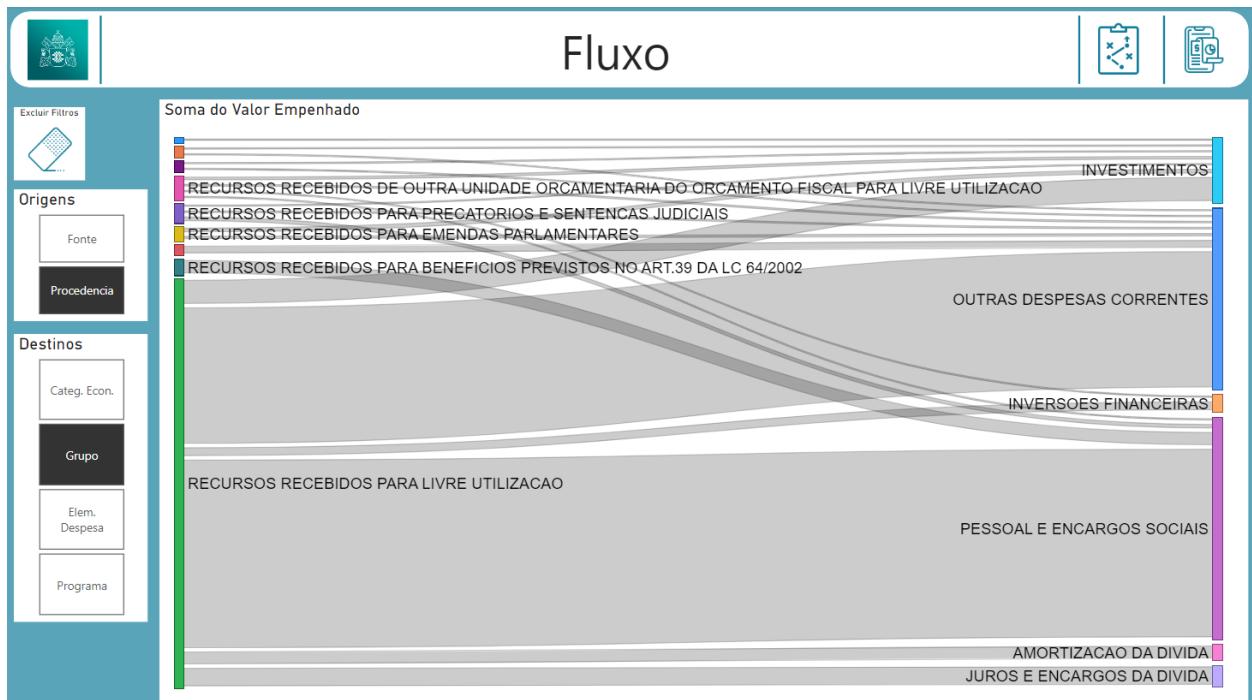


Imagen 58 – Imagem do Dashboard de Fluxo.

Painel de Fluxo: Importante para entender a movimentação de dinheiro dos últimos 2 anos, esse painel exibe em dois filtros as diferentes dimensões identificadas como origens (Fonte e Procedência) e destinos (Categoria Econômica, Grupo, Elemento de Despesa e Programa). As faixas mostram então o fluxo de valores empenhados entre cada relação origem/destino, sendo que a espessura da faixa é proporcional ao valor empenhado entre esse par.

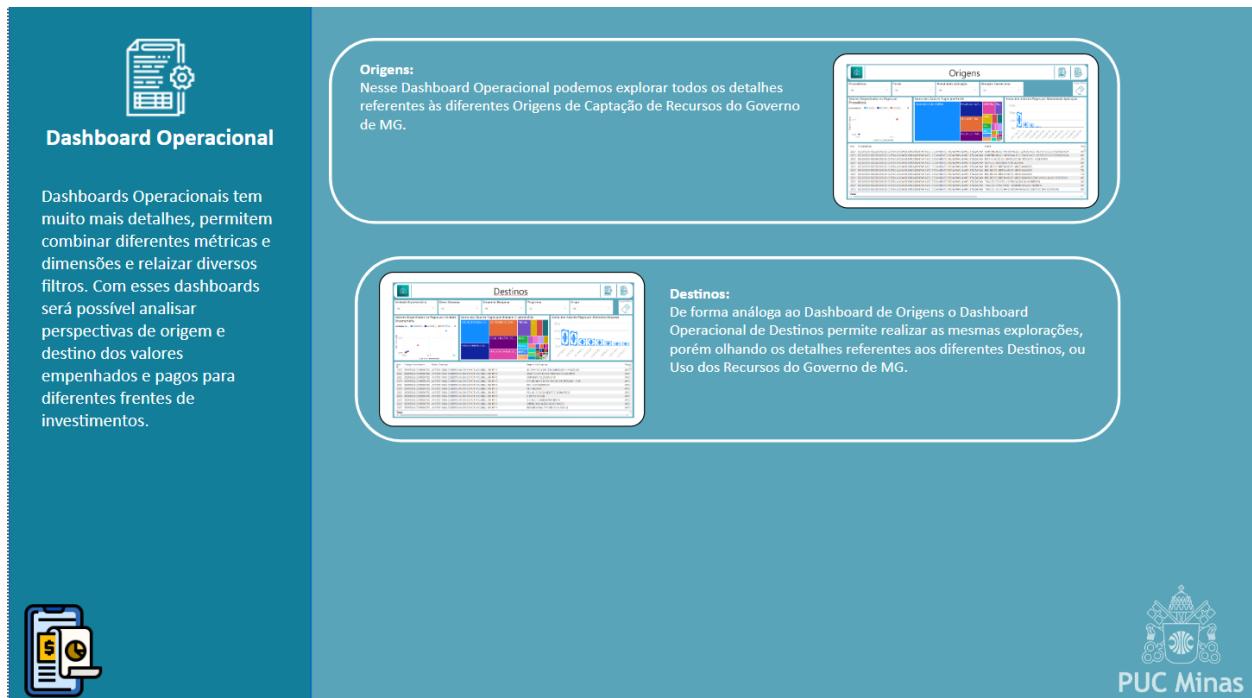


Imagen 59 – Imagem do submenu de Dashboard Operacional.

Submenu Operacional: Nesse submenu são apresentados os painéis operacionais. Esses são os que mais tem filtros e permitem realizar as análises mais minuciosas dado a alta

granularidade dos dados. Nesse conjunto temos dois painéis que tem por objetivo realizar uma análise profunda sobre as receitas e despesas do conjunto de dados com diferentes filtros e visões.

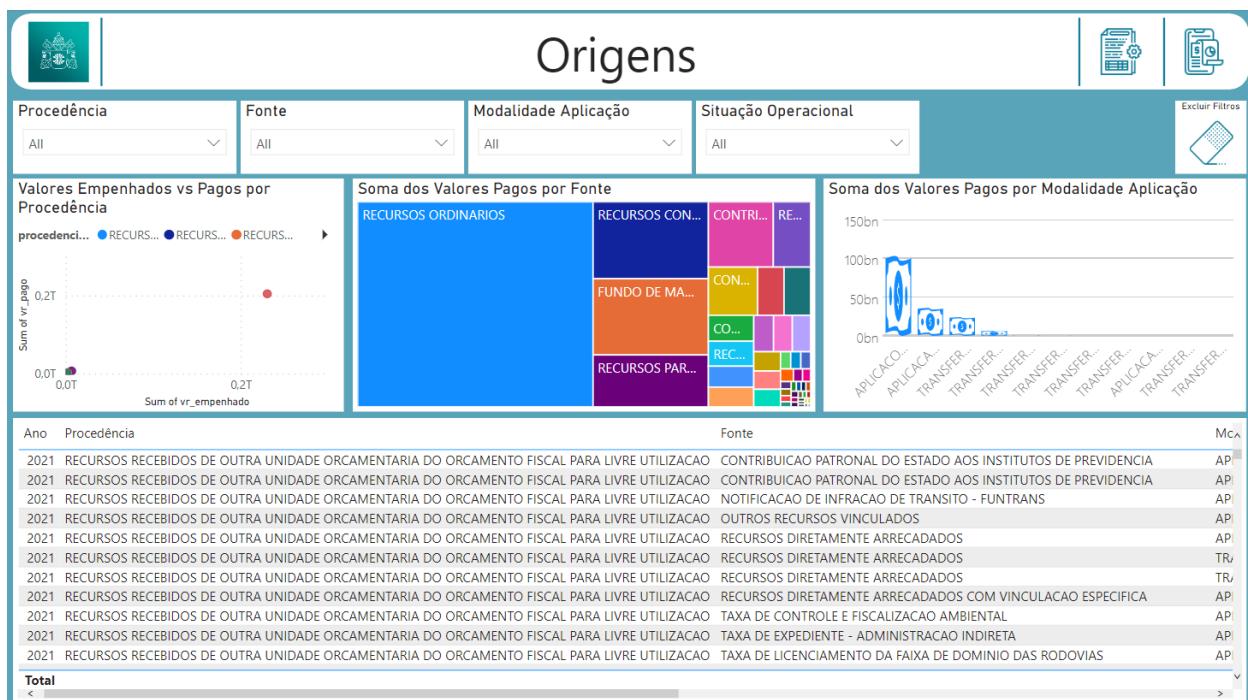


Imagen 60 – Imagem do Dashboard de Origens.

Painel de Origens: Nesse painel temos todos os detalhes referentes às origens de captação de recursos do governo de Minas Gerais, podendo realizar diversas operações de filtro e *drill-down*.



Imagen 61 – Imagem do Dashboard de Destinos.

Painel de Destinos: De forma análoga ao painel de origens, nesse painel temos todos os detalhes referentes aos destinos de recursos do governo de Minas Gerais, podendo realizar outras operações de filtro e *drill-down*.

Dashboard Análises Avançadas

Dashboards de Análises Avançadas mostram o resultado de modelos de *Machine Learning* para tarefas específicas, como Previsão de Gastos, Análise de Regras de Associação e Detecção de Anomalias.

Análises Avançadas - Forecast e Regras de Associação:
O Primeiro Dashboard de Análises Avançadas apresenta dois modelos. O primeiro diz respeito ao forecast dos valores pagos, trazendo uma visão do *Realizado* (valores já pagos) e *Forecast* (valores previstos pelo modelo de *Machine Learning*). O segundo modelo mostra um conjunto de Regras de Associação, útil para entender para cada Pedido como são associadas os diferentes Favorecidos.

Análises Avançadas - Detecção de Anomalias:
No segundo Dashboard de Análises Avançadas, mostramos as métricas do conjunto de registros que foram classificados como "Não-anomalias" vs "Anomalias". É possível também olhar para os registros considerados anomalias com mais detalhes através dos gráficos de rosca e entender o perfil das Unidades Orçamentárias e Favorecidos detectados como anomalias.

PUC Minas

Imagen 62 – Imagem do submenu do Dashboard de Análises Avançadas.

Submenu de Análises Avançadas: Esse submenu apresenta os painéis de análises avançadas. Para esse trabalho desenvolvemos três modelos avançados: análise de forecast para previsão dos valores pagos ao longo dos próximos meses, análise de regras de associação para entender quais itens são adquiridos em conjunto, permitindo melhorar o planejamento da demanda

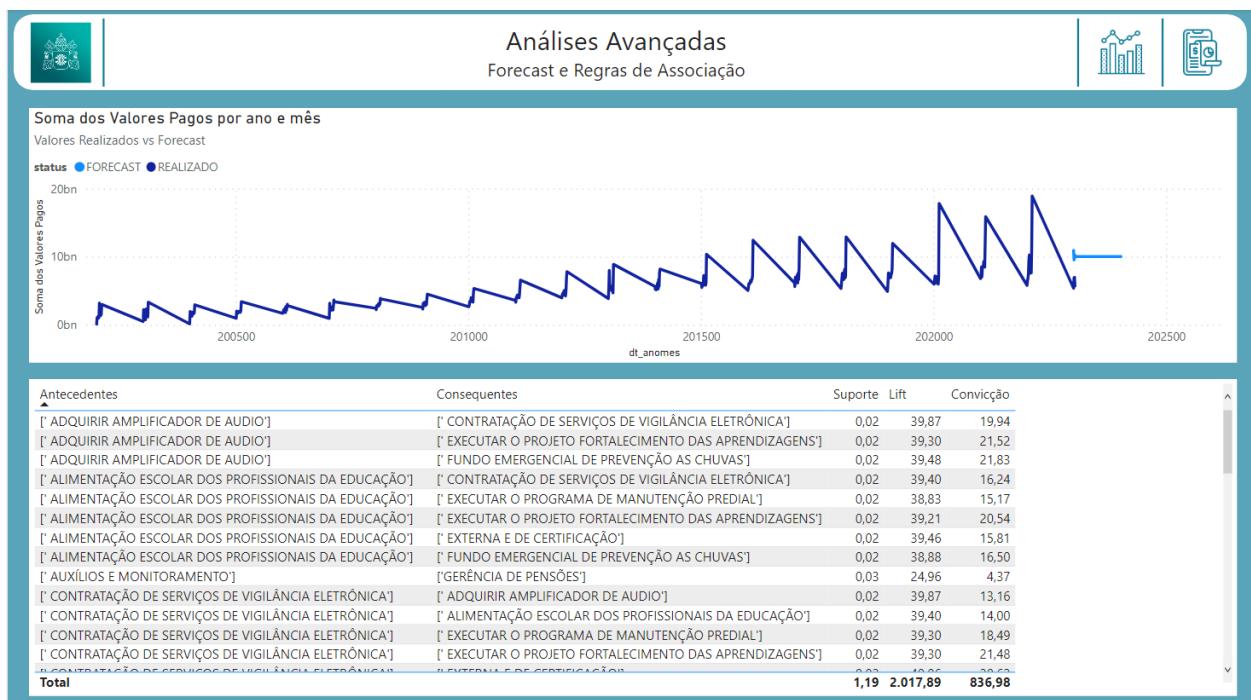


Imagen 63 – Imagem do Dashboard de Análises avançadas – Forecast e Regras de Associação.

Análises Avançadas – Forecast e Regras de Associação: Esse painel apresenta duas análises distintas: a primeira é a análise de forecast. Para chegarmos nesse gráfico, utilizamos um modelo de séries temporais conhecido como ARIMA. O output desse modelo é a previsão dos valores pagos ao longo dos próximos 15 meses com base no comportamento da série temporal (que tem dados de mais de 20 anos incluindo o período de pandemia, onde a

demandas ficou bastante instável). Com esse gráfico conseguimos visualizar que a demanda tende a ter uma média estável ao redor de 10 bilhões de reais, esse gráfico permite realizar um “zoom” nas janelas de dados e entender o comportamento dos valores pagos efetivos (ou realizados, que são os que realmente aconteceram) e os valores previstos (ou forecast, que foram obtidos como resultado do modelo). A segunda análise é a análise de regras de associação. Também foi obtida como saída de um modelo que aprendeu regras de associação entre os itens de gastos mais frequentes, permitindo que o governo crie estratégias de negociação com os fornecedores para diminuir os custos.

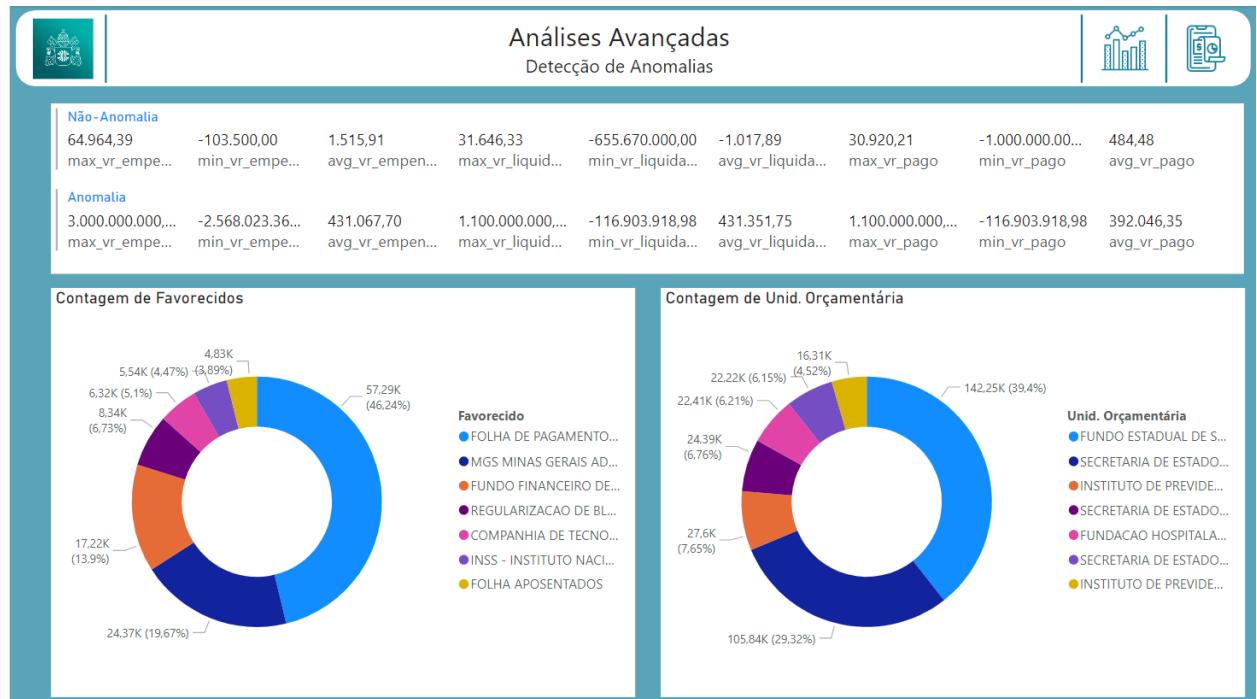


Imagen 64 – Imagem do Dashboard de Análises Avançadas – Detecção de Anomalias.

Análises Avançadas – Detecção de Anomalias: A partir desse painel podemos visualizar os outliers ou anomalias nos dados. O modelo usado para construir essa visualização considerou que pudesse haver até 10% de anomalias dentro da base de dados considerando os valores empenhados, liquidado e pago para diferentes fornecedores e unidades orçamentárias. Através dessa análise podemos ver as áreas que foram mais frequentemente classificadas como anomalias e permitir que o governo quando for realizar um processo de auditoria, que é de suma importância quando utilizamos recursos públicos para coibir abusos e desvios, foque primeiro nesse grupo destacado.

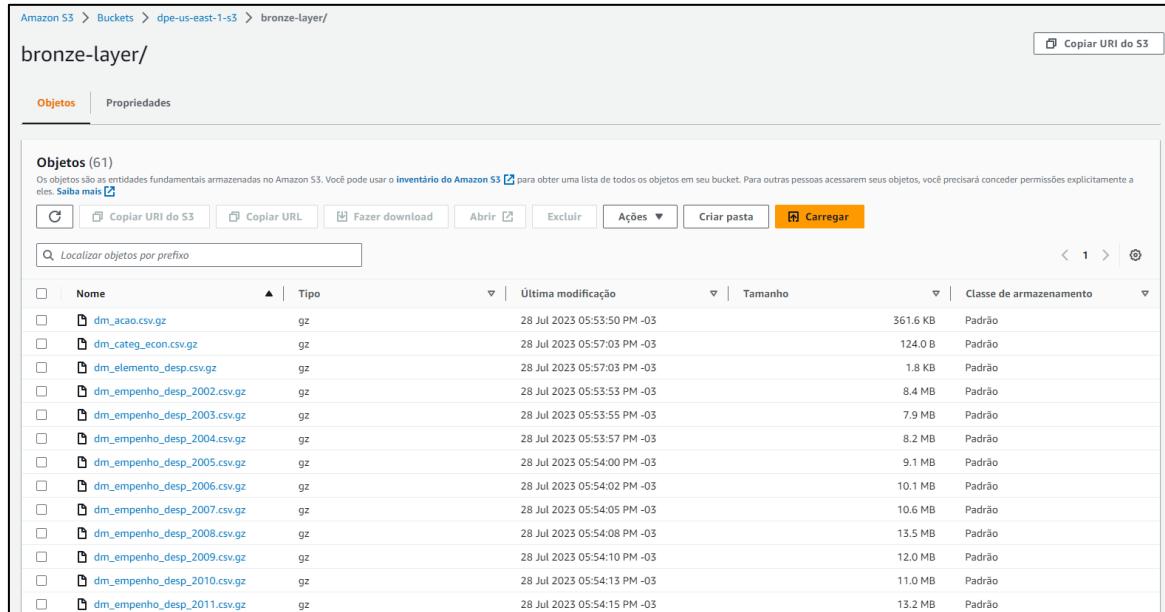
4. Homologação e Resultados

Essa seção descreve o processo de homologação e os resultados obtidos.

4.1. Homologação

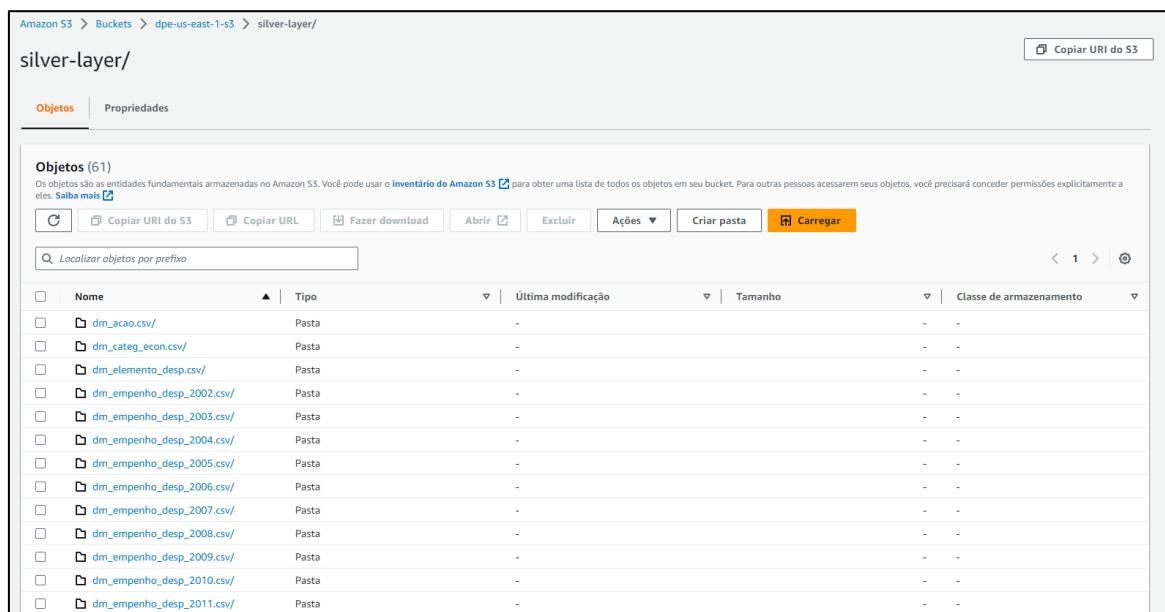
Evidências de homologação atestam a qualidade do processo de dados, das suas transformações e da consistência entre as camadas da arquitetura desenvolvida. Nessa seção mostramos evidencias que qualificam nosso processo de dados.

Evidência de Homologação #1: Análise da camada bronze vs camada silver:



The screenshot shows the Amazon S3 'Buckets' view for the 'dpe-us-east-1-s3' bucket. The 'bronze-layer/' folder is selected. The interface includes a breadcrumb navigation bar: 'Amazon S3 > Buckets > dpe-us-east-1-s3 > bronze-layer/'. A 'Copiar URI do S3' button is in the top right. Below it, there are tabs for 'Objetos' (selected) and 'Propriedades'. A search bar 'Localizar objetos por prefixo' is present. A toolbar below the search bar contains buttons for 'Copiar URI do S3', 'Copiar URL', 'Fazer download', 'Abrir', 'Excluir', 'Ações', 'Criar pasta', and 'Carregar'. A table lists 61 objects, all named 'dim_xxx.csv.gz' where 'xxx' is a date string from 2002 to 2011. The table columns are 'Nome', 'Tipo', 'Última modificação', 'Tamanho', and 'Classe de armazenamento'. All files are gzip compressed and have a size between 1.8 KB and 13.2 MB.

Imagen 65 – Imagem do *bucket S3* que representa a camada bronze contendo os arquivos no seu formato mais “cru” após terem sido “baixados” (61 arquivos compactados).

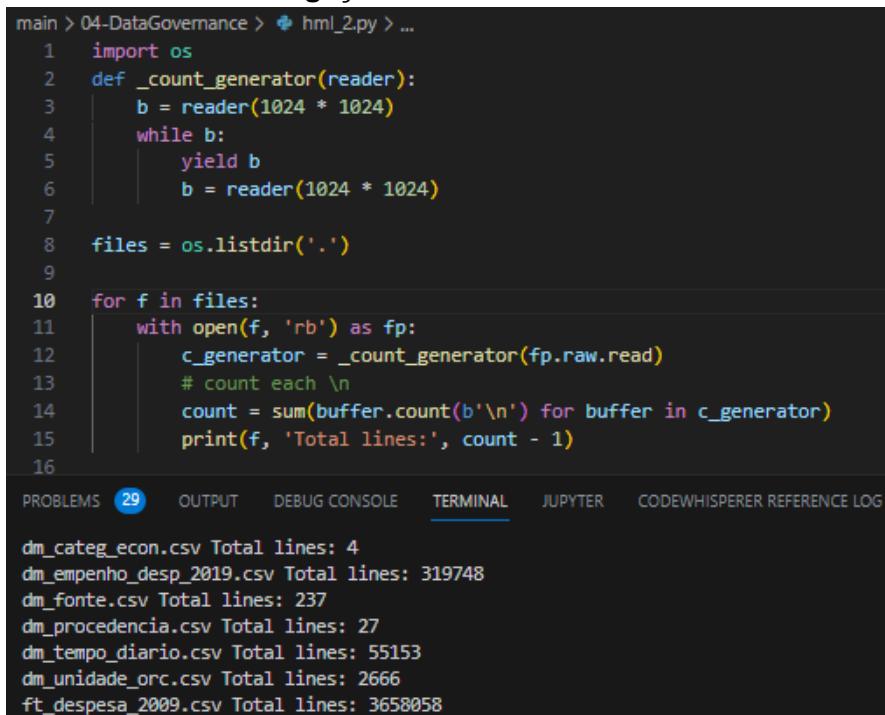


The screenshot shows the Amazon S3 'Buckets' view for the 'dpe-us-east-1-s3' bucket. The 'silver-layer/' folder is selected. The interface includes a breadcrumb navigation bar: 'Amazon S3 > Buckets > dpe-us-east-1-s3 > silver-layer/'. A 'Copiar URI do S3' button is in the top right. Below it, there are tabs for 'Objetos' (selected) and 'Propriedades'. A search bar 'Localizar objetos por prefixo' is present. A toolbar below the search bar contains buttons for 'Copiar URI do S3', 'Copiar URL', 'Fazer download', 'Abrir', 'Excluir', 'Ações', 'Criar pasta', and 'Carregar'. A table lists 61 objects, all named 'dim_xxx.csv/' where 'xxx' is a date string from 2002 to 2011. The table columns are 'Nome', 'Tipo', 'Última modificação', 'Tamanho', and 'Classe de armazenamento'. All entries show '-' in the 'Tamanho' column, indicating they are now ungzipped CSV files.

Imagen 66 – Imagem do *bucket S3* que representa a camada silver contendo os arquivos pós transformação inicial de descompactação (61 arquivos csvs).

Note que na camada em ambas as camadas temos 61 arquivos, mostrando que todos os arquivos baixados na camada bronze foram descompactados na camada silver.

Evidência de Homologação #2: Análise da camada silver vs camada gold.

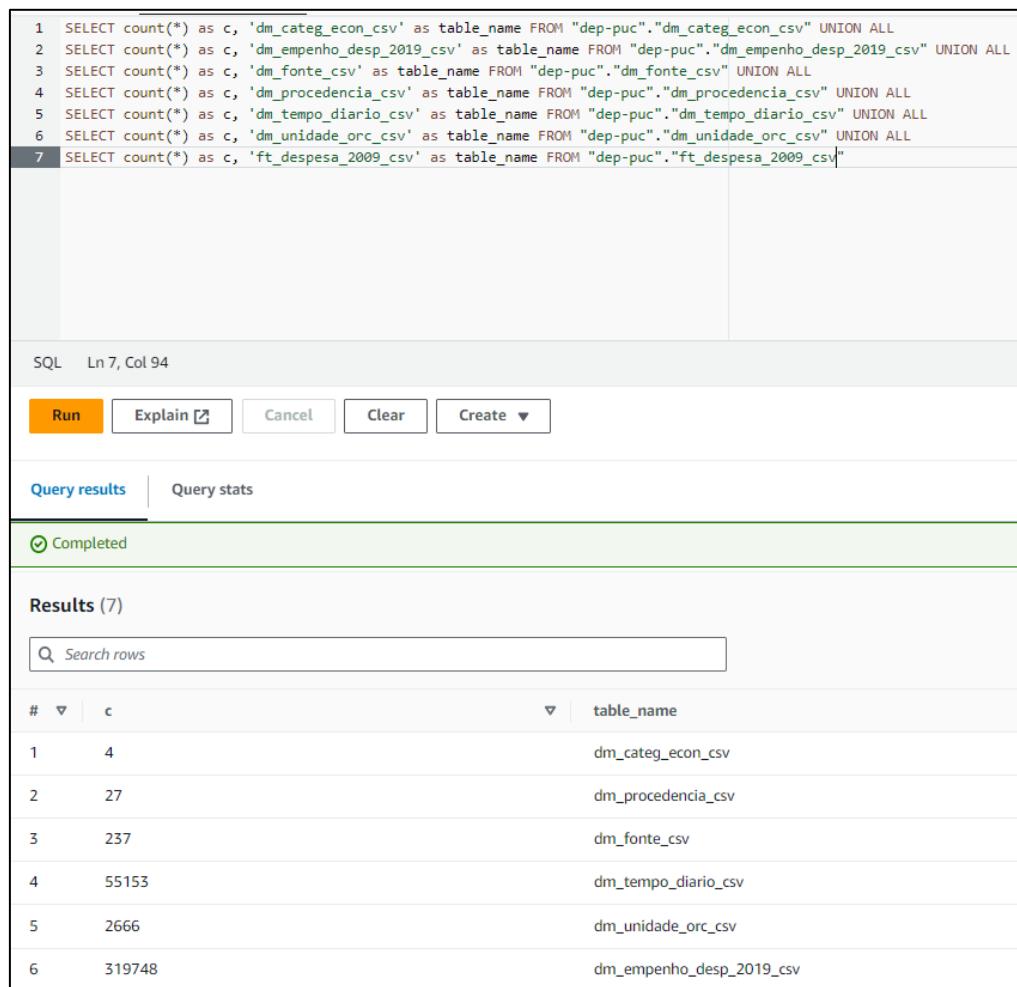


```
main > 04-DataGovernance > hml_2.py > ...
1 import os
2 def _count_generator(reader):
3     b = reader(1024 * 1024)
4     while b:
5         yield b
6         b = reader(1024 * 1024)
7
8 files = os.listdir('.')
9
10 for f in files:
11     with open(f, 'rb') as fp:
12         c_generator = _count_generator(fp.raw.read)
13         # count each \n
14         count = sum(buffer.count(b'\n') for buffer in c_generator)
15         print(f, 'Total lines:', count - 1)
16
```

PROBLEMS 29 OUTPUT DEBUG CONSOLE TERMINAL JUPYTER CODEWHISPERER REFERENCE LOG

```
dm_categ_econ.csv Total lines: 4
dm_empenho_desp_2019.csv Total lines: 319748
dm_fonte.csv Total lines: 237
dm_procedencia.csv Total lines: 27
dm_tempo_diario.csv Total lines: 55153
dm_unidade_orc.csv Total lines: 2666
ft_despesa_2009.csv Total lines: 3658058
```

Imagen 67 – Imagem do código e resultado da contagem de linhas de 7 arquivos da camada silver.



```
1 SELECT count(*) as c, 'dm_categ_econ_csv' as table_name FROM "dep-puc"."dm_categ_econ_csv" UNION ALL
2 SELECT count(*) as c, 'dm_empenho_desp_2019_csv' as table_name FROM "dep-puc"."dm_empenho_desp_2019_csv" UNION ALL
3 SELECT count(*) as c, 'dm_fonte_csv' as table_name FROM "dep-puc"."dm_fonte_csv" UNION ALL
4 SELECT count(*) as c, 'dm_procedencia_csv' as table_name FROM "dep-puc"."dm_procedencia_csv" UNION ALL
5 SELECT count(*) as c, 'dm_tempo_diario_csv' as table_name FROM "dep-puc"."dm_tempo_diario_csv" UNION ALL
6 SELECT count(*) as c, 'dm_unidade_orc_csv' as table_name FROM "dep-puc"."dm_unidade_orc_csv" UNION ALL
7 SELECT count(*) as c, 'ft_despesa_2009_csv' as table_name FROM "dep-puc"."ft_despesa_2009_csv"
```

SQL Ln 7, Col 94

Run Explain Cancel Clear Create ▾

Query results | Query stats

Completed

Results (7)

#	c	table_name
1	4	dm_categ_econ_csv
2	27	dm_procedencia_csv
3	237	dm_fonte_csv
4	55153	dm_tempo_diario_csv
5	2666	dm_unidade_orc_csv
6	319748	dm_empenho_desp_2019_csv

Imagen 68 – Imagem do código e resultado da consulta das linhas de 7 tabelas na camada gold.

Podemos ver pela contagem das linhas feitas pelo código em python considerando uma amostra aleatória de 7 arquivos da camada silver contra a query realizada na camada gold considerando as respectivas tabelas que temos o mesmo número de registros em ambas as camadas.

Evidência de Homologação #3: Análise dos processos dentro da camada gold.

The screenshot shows a database query editor with the following details:

- Query Text:**

```
1 v SELECT sum(c) as total FROM (
2     select count(*) as c from "ft_despesa_2002_csv" UNION ALL
3     select count(*) as c from "ft_despesa_2003_csv" UNION ALL
4     select count(*) as c from "ft_despesa_2004_csv" UNION ALL
5     select count(*) as c from "ft_despesa_2005_csv" UNION ALL
6     select count(*) as c from "ft_despesa_2006_csv" UNION ALL
7     select count(*) as c from "ft_despesa_2007_csv" UNION ALL
8     select count(*) as c from "ft_despesa_2008_csv" UNION ALL
9     select count(*) as c from "ft_despesa_2009_csv" UNION ALL
10    select count(*) as c from "ft_despesa_2010_csv" UNION ALL
11    select count(*) as c from "ft_despesa_2011_csv" UNION ALL
12    select count(*) as c from "ft_despesa_2012_csv" UNION ALL
13    select count(*) as c from "ft_despesa_2013_csv" UNION ALL
14    select count(*) as c from "ft_despesa_2014_csv" UNION ALL
15    select count(*) as c from "ft_despesa_2015_csv" UNION ALL
16    select count(*) as c from "ft_despesa_2016_csv" UNION ALL
17    select count(*) as c from "ft_despesa_2017_csv" UNION ALL
18    select count(*) as c from "ft_despesa_2018_csv" UNION ALL
19    select count(*) as c from "ft_despesa_2019_csv" UNION ALL
20    select count(*) as c from "ft_despesa_2020_csv" UNION ALL
21    select count(*) as c from "ft_despesa_2021_csv" UNION ALL
22    select count(*) as c from "ft_despesa_2022_csv" UNION ALL
23    select count(*) as c from "ft_despesa_2023_csv"
24 )
```
- Execution Status:** Completed (indicated by a green checkmark).
- Results:**

#	total
1	63536540

Imagen 69 – Imagem do código e do resultado da consulta que conta o número de linhas presente em todas as tabelas de despesa.

The screenshot shows a PostgreSQL query editor interface. At the top, a SQL command is entered:

```
1 SELECT count(*) as total from "ft_despesa"
```

Below the command, the status bar indicates "SQL Ln 1, Col 43".

Buttons for "Run again", "Explain", "Cancel", and "Close" are visible.

The "Query results" tab is selected, showing a green "Completed" status.

The results section displays the output of the query:

#	total
1	63536540

Imagen 70 – Imagem do código e do resultado da consulta com o número de linhas da tabela ft_despesa que reúne os dados de todas as tabelas ft_despesa_<ano>.

Evidência de Homologação #4: Análise da camada gold vs Visualização de Dados – Grandes Números



Imagen 71 – Imagem do dashboard de Grandes Números após terem sido filtrados os dados para os anos de 2018, 2019, 2020 e 2021.

```

1 SELECT
2     sum(counting) as contagem,
3     sum(vr_empenhado) as vr_empenhado,
4     sum(vr_liquidado) as vr_liquidado,
5     sum(vr_pago) as vr_pago
6     from "vw_agg_fact"
7 WHERE "ano_particao" in [2018, 2019, 2020, 2021]

```

SQL Ln 7, Col 53

Run again **Explain** **Cancel** **Clear** **Create**

Query results **Query stats**

Completed Time in queue: 147 ms Run time: 1.127 sec

Results (1)

Copy

#	contagem	vr_empenhado	vr_liquidado	vr_pago
1	8978360	4.4716658816585016E11	4.2611610212418964E11	3.8007993699291943E11

Imagen 72 – Imagem do código e resultado da consulta que verifica a contagem e as somas dos valores empenhado, liquidado e pago para os anos de 2018, 2019, 2020 e 2021 da tabela que alimenta o dashboard da imagem 71.

Evidência de Homologação #5: Análise das transformações da camada gold vs Visualização de Dados – Dimensões vs Ano

```

1 SELECT
2     COUNT(DISTINCT categ_econ_name) as categ_econ,
3     COUNT(DISTINCT elemento_desp_name) as elementos_desp,
4     COUNT(DISTINCT empenhos_desp_dt_empenho) as empenhos_desp_dt_empenho,
5     COUNT(DISTINCT empenhos_desp_unidade_executora) as empenhos_desp_unidade_executora,
6     COUNT(DISTINCT empenhos_desp_unl_prog_gasto) as empenhos_desp_unl_prog_gasto,
7     COUNT(DISTINCT empenhos_desp_tipo_empenho) as empenhos_desp_tipo_empenho,
8     COUNT(DISTINCT grupo_name) as grupo,
9     COUNT(DISTINCT procedencia_name) as procedencia,
10    COUNT(DISTINCT favorecido_nome_anonimizado) as favorecido_nome_anonimizado
11   from "vw_du"
12

```

SQL Ln 12, Col 1

Run again **Explain** **Cancel** **Clear** **Create**

Query results **Query stats**

Completed Time in queue: 175 ms Run time: 21.338 sec Data scanned: 150.03 MB

Results (1)

Copy **Download results**

#	categ_econ	elemento_desp	empenhos_desp_dt_empenho	empenhos_desp_unidade_executora	empenhos_desp_unl_prog_gast	empenhos_desp_tipo_empenho	grupo	procedencia	favorecido_nome_anonimizado
1	2	53	587	1046	2152	4	6	9	131038

Imagen 73 – Imagem da consulta e do código que analisam o número de valores únicos por dimensão do modelo.

```

1 SELECT
2     COUNT(DISTINCT unidade_orc_administracao) as unidade_orc_administracao,
3     COUNT(DISTINCT unidade_orc_grupo_administracao) as unidade_orc_grupo_administracao,
4     COUNT(DISTINCT programa_name) as programa,
5     COUNT(DISTINCT funcao_desp_name) as funcao_desp,
6     COUNT(DISTINCT item_desp_name) as item_desp,
7     COUNT(DISTINCT unidade_orc_sigla) as unidade_orc_sigla,
8     COUNT(DISTINCT fonte_name) as fonte,
9     COUNT(DISTINCT unidade_orc_name) as unidade_orc,
10    COUNT(DISTINCT modalidade_aplic_name) as modalidade_aplic
11   From "vw_du"
12

```

SQL Ln 12, Col 1

Run again **Explain** **Cancel** **Clear** **Create**

Query results **Query stats**

Completed Time in queue: 163 ms Run time: 15.386 sec Data scanned: 120.72 MB

Results (1)

Copy **Download results**

#	unidade_orc_administracao	unidade_orc_grupo_administracao	programa	funcao_desp	item_desp	unidade_orc_sigla	fonte	unidade_orc	modalidade_aplic
1	5	4	153	26	422	85	55	85	12

Imagen 74 – Imagem da consulta e do código que analisam o número de valores únicos por dimensão do modelo (continuação da consulta e resultados).

Categoria Econômica	Elemento Desp. Nome	Empenhos Desp. Data
2	53	587
Empenhos Desp. Unidade Executora	Empenho Desp. Unidade Programa Gasto	Empenhos Desp. Tipo
1046	2152	4
Grupo	Procedência	Favorecido
6	9	131,0...
Un. Orçamentária Administração	Un. Orçamentária Grupo Admnistração	Programa
5	4	153
Função Despesa	Item Despesa	Unidade Orçamentária
26	422	85
Fonte	Unidade Orçamentária	Modalidade de Aplicação
55	85	12

Imagen 75 – Imagem do dashboard de Dimensões vs Anos contendo as quantidades de valores únicos por dimensão.

Nessa análise podemos verificar os valores de cada dimensão dentro da camada *gold* vs os valores da visualização presente no dashboard de Dims vs Anos.

4.2. Resultados

Como resultados podemos destacar a criação de uma pipeline de dados automatizada em ambiente de nuvem que leva 13 minutos para construir as 3 camadas da nossa arquitetura baixando e transformando mais de 60 arquivos em tabelas diferentes.

```

Response
null

Function Logs
source/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"
s {"filename": "bronze-layer/ft_despesa_2008.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2009.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2010.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2011.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2012.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2013.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2014.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2015.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2016.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2017.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2018.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2019.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2020.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2021.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2022.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/ft_despesa_2023.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/dm_situacao_op_desp.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/dm_categ_econ.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/dm_elemento_desp.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}
s {"filename": "bronze-layer/datapackage.csv.gz", "link": "https://dados.mg.gov.br/dataset/despesa/resource/67fcfc88-66e4-4a65-b46e-4fbaf88c4496", "accessed_in": "2023-07-28-20:53:42"}}
None
END RequestId: 45f4efef13-0d68-4c2c-b206-28819d804361
REPORT RequestId: 45f4efef13-0d68-4c2c-b206-28819d804361 Duration: 259295.11 ms Billed Duration: 259296 ms Memory Size: 3008 MB Max Memory Used: 625 MB Init Duration: 814.68 ms

```

Imagen 76 - Imagem com a execução da criação da camada bronze pouco menos de 5 minutos.

```
* Execution results
> bronze-layer/ft_despesa_2014.csv.gz
ft_despesa_2014.csv
Unzip and Save dpe-us-east-1-s3 bronze-layer/ft_despesa_2014.csv.gz dpe-us-east-1-s3 silver-layer/ft_despesa_2014.csv/ft_despesa_2014.csv
> bronze-layer/ft_despesa_2015.csv.gz
ft_despesa_2015.csv
Unzip and Save dpe-us-east-1-s3 bronze-layer/ft_despesa_2015.csv.gz dpe-us-east-1-s3 silver-layer/ft_despesa_2015.csv/ft_despesa_2015.csv
> bronze-layer/ft_despesa_2016.csv.gz
ft_despesa_2016.csv
Unzip and Save dpe-us-east-1-s3 bronze-layer/ft_despesa_2016.csv.gz dpe-us-east-1-s3 silver-layer/ft_despesa_2016.csv/ft_despesa_2016.csv
> bronze-layer/ft_despesa_2017.csv.gz
ft_despesa_2017.csv
Unzip and Save dpe-us-east-1-s3 bronze-layer/ft_despesa_2017.csv.gz dpe-us-east-1-s3 silver-layer/ft_despesa_2017.csv/ft_despesa_2017.csv
> bronze-layer/ft_despesa_2018.csv.gz
ft_despesa_2018.csv
Unzip and Save dpe-us-east-1-s3 bronze-layer/ft_despesa_2018.csv.gz dpe-us-east-1-s3 silver-layer/ft_despesa_2018.csv/ft_despesa_2018.csv
> bronze-layer/ft_despesa_2019.csv.gz
ft_despesa_2019.csv
Unzip and Save dpe-us-east-1-s3 bronze-layer/ft_despesa_2019.csv.gz dpe-us-east-1-s3 silver-layer/ft_despesa_2019.csv/ft_despesa_2019.csv
> bronze-layer/ft_despesa_2020.csv.gz
ft_despesa_2020.csv
Unzip and Save dpe-us-east-1-s3 bronze-layer/ft_despesa_2020.csv.gz dpe-us-east-1-s3 silver-layer/ft_despesa_2020.csv/ft_despesa_2020.csv
> bronze-layer/ft_despesa_2021.csv.gz
ft_despesa_2021.csv
Unzip and Save dpe-us-east-1-s3 bronze-layer/ft_despesa_2021.csv.gz dpe-us-east-1-s3 silver-layer/ft_despesa_2021.csv/ft_despesa_2021.csv
> bronze-layer/ft_despesa_2022.csv.gz
ft_despesa_2022.csv
Unzip and Save dpe-us-east-1-s3 bronze-layer/ft_despesa_2022.csv.gz dpe-us-east-1-s3 silver-layer/ft_despesa_2022.csv/ft_despesa_2022.csv
> bronze-layer/ft_despesa_2023.csv.gz
ft_despesa_2023.csv
Unzip and Save dpe-us-east-1-s3 bronze-layer/ft_despesa_2023.csv.gz dpe-us-east-1-s3 silver-layer/ft_despesa_2023.csv/ft_despesa_2023.csv
> bronze-layer/hashcache.jccla-4a0d-81b9-0d12391ad3b7
END RequestId: ad35a4b1-cc1a-4a0d-81b9-0d12391ad3b7
REPORT RequestId: ad35a4b1-cc1a-4a0d-81b9-0d12391ad3b7 Duration: 231999.43 ms Billed Duration: 232000 ms Memory Size: 3008 MB Max Memory Used: 358 MB Init Duration: 377.23 ms
```

Imagen 77 - Imagen com a execução da criação da camada *silver* em menos de 4 minutos.

Timestamp	Mensagem
2023-05-03T15:27:19.068-03:00	[acd0e56f-0dcc-437f-a85a-6944f1a81909] BENCHMARK : Running Start Crawl for Crawler s3-dep-crawler-to-athena
2023-05-03T15:27:40.765-03:00	[acd0e56f-0dcc-437f-a85a-6944f1a81909] BENCHMARK : Classification complete, writing results to database dep-puc
2023-05-03T15:27:40.795-03:00	[acd0e56f-0dcc-437f-a85a-6944f1a81909] INFO : Crawler configured with Configuration {"Version":1.0,"Grouping":{"T
2023-05-03T15:28:05.355-03:00	[acd0e56f-0dcc-437f-a85a-6944f1a81909] INFO : Created table ft_despesa_2016_csv in database dep-puc
2023-05-03T15:28:06.708-03:00	[acd0e56f-0dcc-437f-a85a-6944f1a81909] BENCHMARK : Finished Writing to Catalog
2023-05-03T15:28:06.747-03:00	[acd0e56f-0dcc-437f-a85a-6944f1a81909] INFO : Run Summary For TABLE:
2023-05-03T15:28:06.748-03:00	[acd0e56f-0dcc-437f-a85a-6944f1a81909] INFO : ADD: 61
2023-05-03T15:29:14.344-03:00	[acd0e56f-0dcc-437f-a85a-6944f1a81909] BENCHMARK : Crawler has finished running and is in state READY

Imagen 78 - Imagem com a execução do primeiro passo da criação da camada *gold* em 2 minutos.

	Timestamp	Mensagem
▶	2023-05-25T10:57:43.285-03:00	[8e3834b4-799c-4f9e-8b47-05949c88ae91] BENCHMARK : Running Start Crawl for Crawler s3-analysis-dep-crawler-to-athena
▶	2023-05-25T10:57:54.462-03:00	[8e3834b4-799c-4f9e-8b47-05949c88ae91] ERROR : Not all read errors will be logged. com.amazonaws.services.s3.model.AmazonS3Exception: An error occurred (InternalError) when calling the HeadObject operation: Internal server error (Service: Amazon S3; Status Code: 500; Error Code: InternalError; Request ID: 8E3834B4799C4F9E8B4705949C88AE91; S3 Bucket: s3-analysis-dep-crawler-to-athena; S3 Key: 2023-05-25T10:57:43.285-03:00)
▶	2023-05-25T10:58:03.779-03:00	[8e3834b4-799c-4f9e-8b47-05949c88ae91] BENCHMARK : Classification complete, writing results to database dep-puc
▶	2023-05-25T10:58:03.805-03:00	[8e3834b4-799c-4f9e-8b47-05949c88ae91] INFO : Crawler configured with Configuration {"Version":1.0,"Grouping":{"TableGrouping":{}}}
▶	2023-05-25T10:58:24.452-03:00	[8e3834b4-799c-4f9e-8b47-05949c88ae91] WARN : Invalid table level found - 12. Table level must be less than or equal to 10
▶	2023-05-25T10:58:28.607-03:00	[8e3834b4-799c-4f9e-8b47-05949c88ae91] BENCHMARK : Finished writing to Catalog
▶	2023-05-25T10:59:37.118-03:00	[8e3834b4-799c-4f9e-8b47-05949c88ae91] BENCHMARK : Crawler has finished running and is in state READY

Imagen 79 - Imagen com a execución do segundo paso da creación da camada *gold* em 2 minutos.

Outro resultado positivo que podemos destacar foi o número e a variedade de análises que podemos fazer a partir do Dashboard construído para esse trabalho:



Imagen 80 – Imagem com 8 dos 9 painéis desenvolvidos no painel de controle.

Foram desenvolvidos 9 painéis com 50 gráficos de 17 tipos diferentes com 21 filtros, além dos 5 painéis de navegação e resumo:

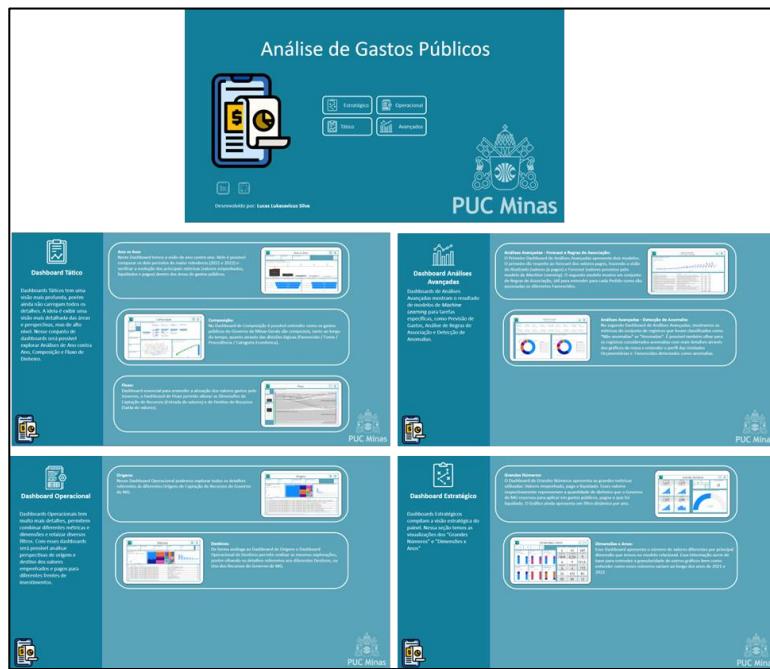


Imagen 81 – Imagem com as 5 telas desenvolvidas para o painel de controle.

Através dessas visualizações foi possível entender insights importantes como:

Insight 1: Apesar do número de registros durante a pandemia ter diminuído sinalizando menos registros feitos, o ticket médio desses gastos aumentou uma vez que os valores pagos não caíram na mesma proporção. O que indica que o governo continuou o ritmo de gastos, porém de forma mais concentrada.



Imagen 82 – Imagem dos números do Dashboard de Grandes Números.

Insight 2: A previsão dos gastos públicos oscila nos primeiros meses (janeiro, fevereiro e março de 2023), acompanhando o movimento instável e não-estacionário da série histórica, porém com o passar do tempo esse valor se concentra ao redor da média dos valores previstos por conta de o modelo considerar poucos períodos anteriores para a componente de médias móveis.

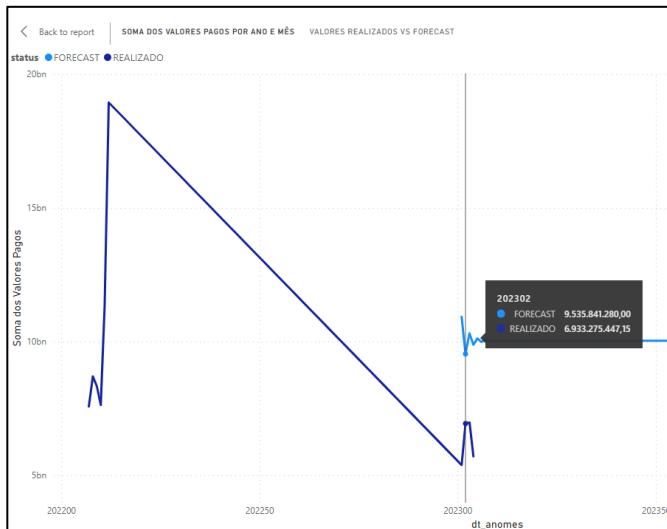


Imagen 83 – Imagem do gráfico de previsão de gastos do Dashboard de Análises Avançadas – Forecast e Regras de Associação.

Insight 3: Notamos que determinados equipamentos estão associados a determinados serviços (amplificador de áudio e serviços de vigilância eletrônica, por exemplo), nesse caso, essa análise permite entender que pode ser mais proveitoso (mais barato) contratar prestadores de serviço que possuam os equipamentos necessários para executar o serviço.

Antecedentes	Consequentes	Supor te	Lift	Convicção
['ADQUIRIR AMPLIFICADOR DE ÁUDIO']	['CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔNICA']	0,02	39,87	19,94
['ADQUIRIR AMPLIFICADOR DE ÁUDIO']	['EXECUTAR O PROJETO FORTALECIMENTO DAS APRENDIZAGENS']	0,02	39,30	21,52
['ADQUIRIR AMPLIFICADOR DE ÁUDIO']	['FUNDO EMERGENCIAL DE PREVENÇÃO AS CHUVAS']	0,02	39,48	21,83
['ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDUCAÇÃO']	['CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔNICA']	0,02	39,40	16,24
['ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDUCAÇÃO']	['EXECUTAR O PROGRAMA DE MANUTENÇÃO PREDIAL']	0,02	38,83	15,17
['ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDUCAÇÃO']	['EXECUTAR O PROJETO FORTALECIMENTO DAS APRENDIZAGENS']	0,02	39,21	20,54
['ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDUCAÇÃO']	['EXTERNA E DE CERTIFICAÇÃO']	0,02	39,46	15,81
['ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDUCAÇÃO']	['FUNDO EMERGENCIAL DE PREVENÇÃO AS CHUVAS']	0,02	38,88	16,50
['AUXÍLIOS E MONITORAMENTO']	['GERÊNCIA DE PENSÕES']	0,03	24,96	4,37
['CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔNICA']	['ADQUIRIR AMPLIFICADOR DE ÁUDIO']	0,02	39,87	13,16
['CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔNICA']	['ALIMENTAÇÃO ESCOLAR DOS PROFISSIONAIS DA EDUCAÇÃO']	0,02	39,40	14,00
['CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔNICA']	['EXECUTAR O PROGRAMA DE MANUTENÇÃO PREDIAL']	0,02	39,30	18,49
['CONTRATAÇÃO DE SERVIÇOS DE VIGILÂNCIA ELETRÔNICA']	['EXECUTAR O PROJETO FORTALECIMENTO DAS APRENDIZAGENS']	0,02	39,30	21,48

Imagen 84 – Imagem da tabela de resultados das regras de associação do Dashboard de Análises Avançadas – Forecast e Regras de Associação.

Insight 4: Dos registros detectados como anômalos temos valores altíssimos (3 Bilhões de reais empenhados) até valores muito negativos (normalmente proveniente de estorno quando o pagamento não é executado), além disso, podemos ver que o modelo de detecção de anomalias realmente destacou registros diferenciados uma vez que as médias dos conjuntos Anomalia e Não-Anomalia são muito diferentes.

Anomalias									
64.964,39	-103.500,00	1.515,91	31.646,33	-655.670.000,00	-1.017,89	30.920,21	-1.000.000,00...	484,48	
max_vr_empe...	min_vr_empe...	avg_vr_empen...	max_vr_liquid...	min_vr_liquida...	avg_vr_liquida...	max_vr_pago	min_vr_pago	avg_vr_pago	
3.000.000,000,...	-2.568.023,36...	431.067,70	1.100.000,000,...	-116.903.918,98	431.351,75	1.100.000,000,...	-116.903.918,98	392.046,35	
max_vr_empe...	min_vr_empe...	avg_vr_empen...	max_vr_liquid...	min_vr_liquida...	avg_vr_liquida...	max_vr_pago	min_vr_pago	avg_vr_pago	

Imagen 85 – Imagem dos Indicadores para os grupos de dados de Anomalia e Não-anomalia do Dashboard de Análises Avançadas – Detecção de Anomalia.

Insight 5: Os maiores gastos do governo são com pagamento a pessoas (folha de pagamento pessoal, aposentadoria, INSS entre outros). Os registros destacados pelos gráficos abaixo revelam que também são os favorecidos mais prováveis de serem alvo de anomalias, além do setor de tecnologia da informação. Outro ponto que chama atenção é o uso do dinheiro público em unidades orçamentárias como a área da saúde e da educação. Como parte do período apurado por esses gráficos (2021 e 2022) ainda estávamos em regime de pandemia faz sentido considerarmos esses valores, porém investigações mais aprofundadas são necessárias nessas áreas para garantir a destinação correta a essas áreas.

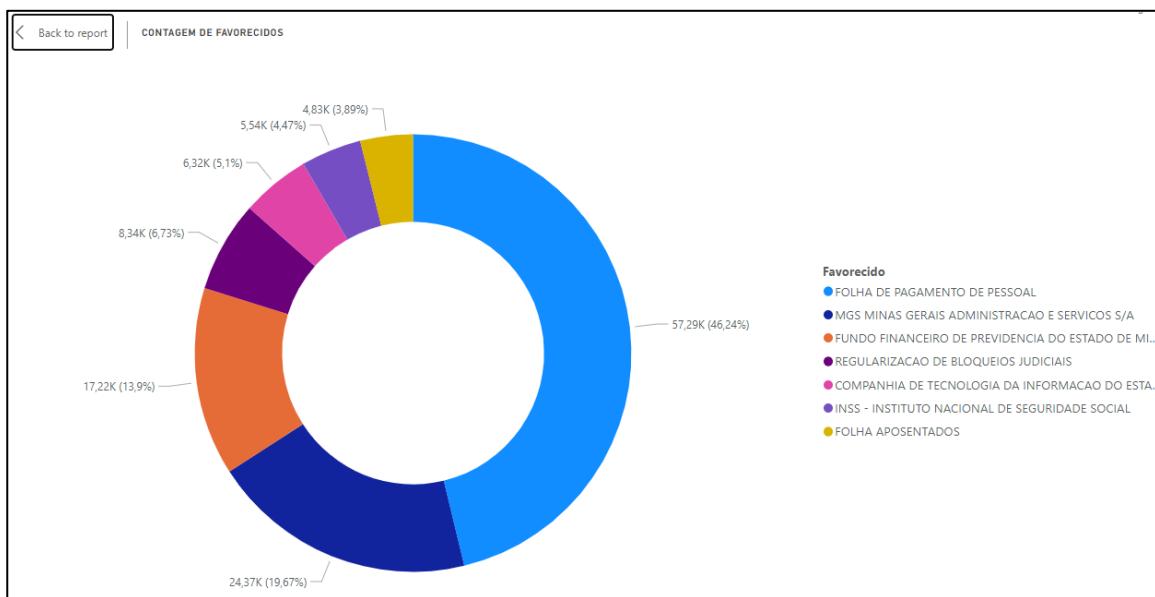


Imagen 86 – Imagem do gráfico de distribuição dos valores pagos por Favorecido dentro do grupo de registros detectados como anomalia do Dashboard de Análises Avançadas – Detecção de Anomalia.

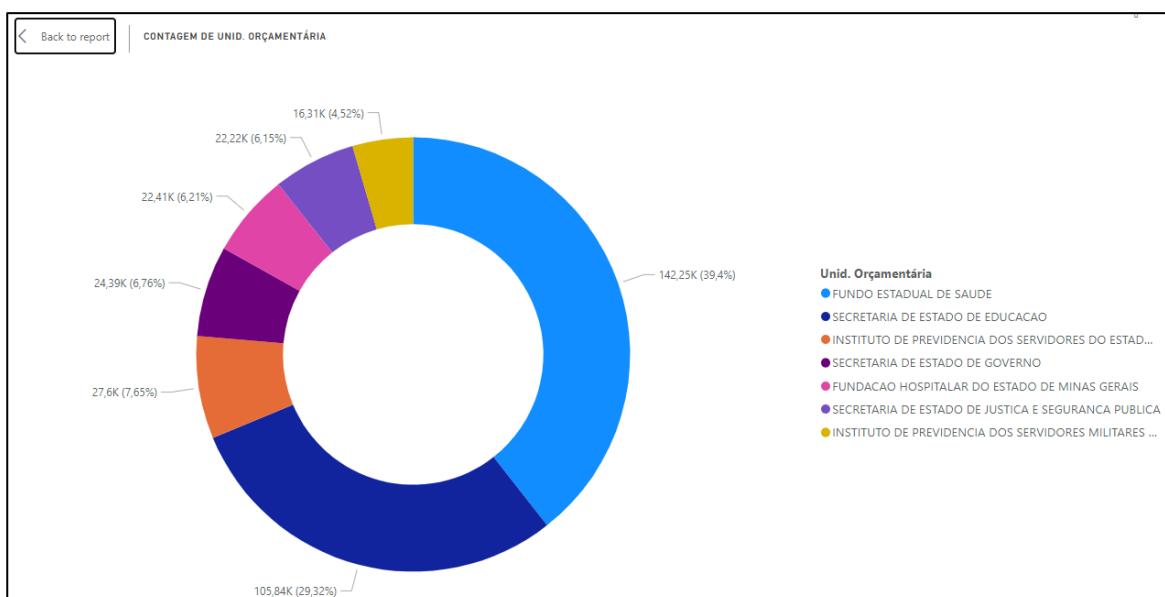


Imagen 87 – Imagem do gráfico de distribuição dos valores pagos por Unidade Orçamentária dentro do grupo de registros detectados como anomalia do Dashboard de Análises Avançadas – Detecção de Anomalia.

Insight 6: Gastos com Folha De Pagamento Pessoal representa um dos maiores gastos do governo, como ocorre em diversas empresas onde o setor mais onerado é o de Recursos Humanos.

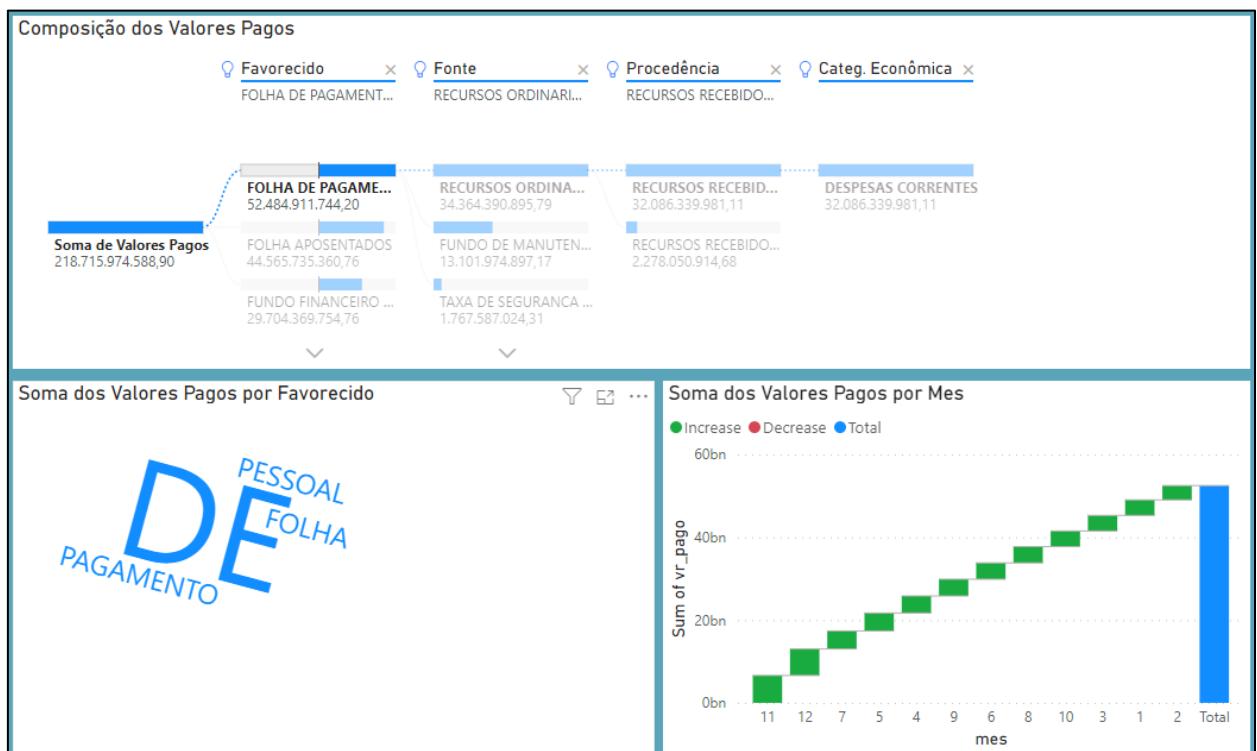


Imagen 88 – Imagem do Dashboard de Composição com filtro de Favorecido igual a “Folha de Pagamento Pessoal” aplicado.

Insight 7: “Recursos recebidos para livre utilização” é uma das maiores procedências que temos nos anos de 2021 e 2022. Essa verba é utilizada para os seguintes grupos: Investimentos, outras despesas correntes, Pessoa e encargos sociais, inversões financeiras, amortização da dívida e juros e encargos da dívida. Como é um valor muito alto (230 milhões de reais) vale questionar o ministério o destino desse dinheiro, ainda mais por conta de outras despesas correntes, ser um grupo ainda muito subjetivo.

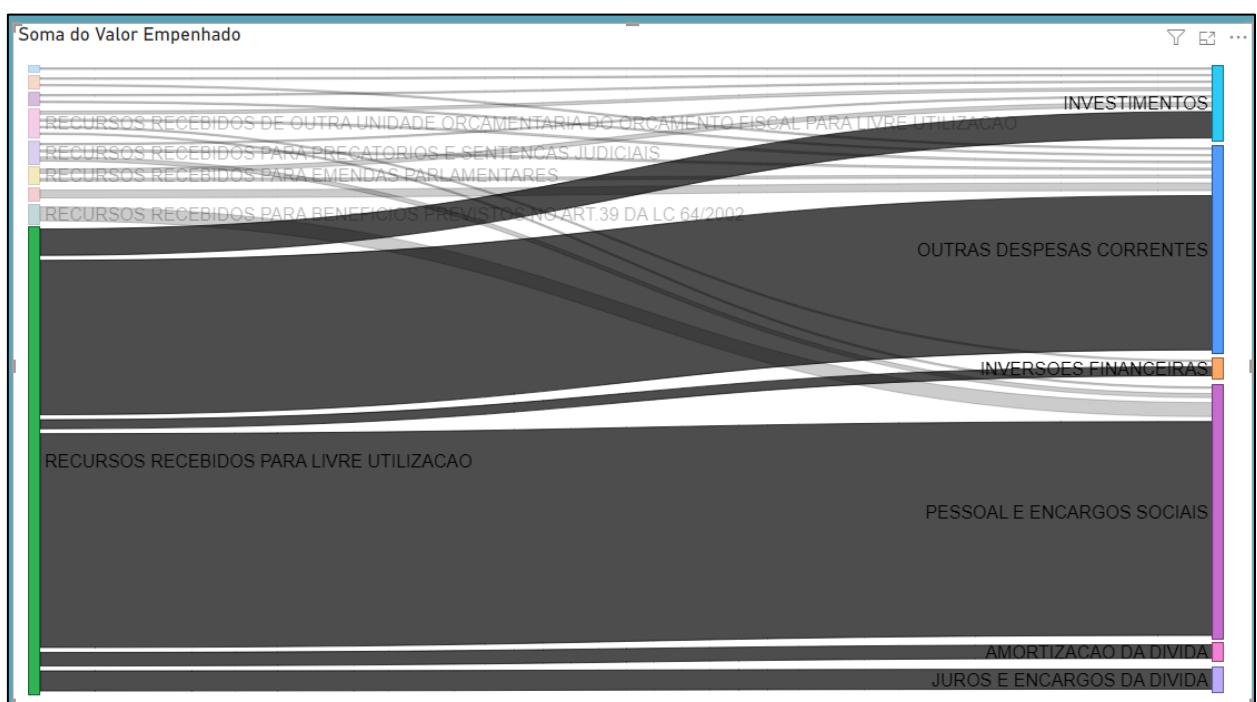


Imagen 89 – Imagem do Dashboar de Fluxo destacando o fluxo monetário entre Recursos recebidos para livre utilização e seus diferentes destinos.

5. Conclusões e Próximos Passos

Essa seção descreve as conclusões obtidas e os próximos passos que sugerimos seguir a partir do que já foi desenvolvido no presente trabalho.

5.1. Conclusões

Tendo desenvolvido todo o trabalho, pudemos entender a construção completa de uma arquitetura de dados *end-to-end*, observando as transformações feitas do ponto de vista de engenharia de dados, os modelos de ciência de dados e os painéis de visualização de dados. Entendemos a importância desses processos para a evolução de empresas em um ambiente cada vez mais moderno, versátil, dinâmico e crescente em volumetria de dados (40). Nas análises desenvolvidas nesse trabalho destacamos a importância desses processos também para promover mais transparência entre a administração pública por parte do governo e suas autarquias e a sociedade e como essa transparência gera interação e dessa forma mais democracia.

5.2. Próximos Passos

Essa pesquisa se revelou bastante interessante pela capacidade de entendermos o uso do dinheiro público nas ações do governo. Como próximos passos podemos sugerir:

1. Analisar as propostas do governo versus o destino dos valores;
2. Analisar de forma mais detalhada as receitas do governo, cruzando a base de dados usada aqui com outras disponíveis no portal da transparência;
3. Coletar os dados de municípios e verificar como esses dados podem ser cruzados e analisados com os dados do governo;
4. Expandir essa análise para cada estado da União;

6. Anexos

Nome do Item	Tipo do Item	Localização
Repositório do Github	Repositório de código contendo todos os códigos em python utilizados para o desenvolvimento do trabalho (tanto a parte de engenharia, quanto ciência de dados).	https://github.com/Lukasavicus/Lukasavicus-PUC-TCC
Painel de Controle	Dashboard criado em PowerBI como ferramenta de visualização de dados.	https://app.powerbi.com/view?r=eyJrJljojOTFjZGRkZjktNTc3MC00MGQxLTg3OWMtNjYzYzdiNTkzZGY1liwidCI6IjE0Y2JkNWE3LWVjOTQtNDZiYS1iMzE0LWNjMGZjOTcyYTE2MSIsImMiOjh9

Trabalho de Conclusão de Curso e Demais documentos (diagramas, especificações e imagens)		https://drive.google.com/drive/folders/1G0ZWB-SbkZFzevBHQqy5Kj9OyYqWNYfQ?usp=drive_link
--	--	---

7. Referencias

- (1) Kleppmann, M. (2023). Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems. Findaway World.
- (2) Kellyton dos Santos Brito, Marcos Antônio da Silva Costa, Vinicius Cardoso Garcia, and Silvio Romero de Lemos Meira. 2014. Brazilian government open data: implementation, challenges, and potential opportunities. In Proceedings of the 15th Annual International Conference on Digital Government Research (dg.o '14). Association for Computing Machinery, New York, NY, USA, 11–16. <https://doi.org/10.1145/2612733.2612770>
- (3) Fiscal Policy in Brazil through the Lens of an Estimated DSGE model, Working Paper Series, Disponível em: <https://www.bcb.gov.br/pec/wps/ingl/wps240.pdf>
- (4) LEI COMPLEMENTAR Nº 101, DE 4 DE MAIO DE 2000 – Lei de Responsabilidade Fiscal - Disponível em: https://www.planalto.gov.br/ccivil_03/leis/lcp/lcp101.htm
- (5) S.J. Piotrowski Governmental transparency in the path of administrative reform, SUNY Press, New York (2007)
- (6) Open Society Justice Initiative - Transparency and silence, Open Society Institute (2006) Available: http://www.justiceinitiative.org/db/resource2?res_id=103818
- (7) J. C. Bertot, P. T. Jaeger, and J. M. Grimes, "Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies," Government Information Quarterly, vol. 27, pp. 264-271, 2010
- (8) W. Wong and E. Welch, "Does E-Government Promote Accountability? A Comparative Analysis of Website Openness and Government Accountability," Governance, vol. 17, pp. 275-297, 2004.
- (9) T. B. Andersen, "E-Government as an anti-corruption strategy," Information Economics and Policy, vol. 21, pp. 201-210, 2009
- (10) D. Dada, "The Failure of E-Government in Developing Countries: A literature review," The Electronic Journal of Information Systems in Developing Countries, vol. 26, pp. 1-10, 2006.
- (11) Open Government Partnership - Open Government Declaration - Disponível em: <https://www.opengovpartnership.org/process/joining-ogp/open-government-declaration/>
- (12) Governo Aberto – 1o Plano de Ação do Brasil – publicado em: 12/12/2014 Disponível em: <https://www.gov.br/cgu/pt-br/governo-aberto/a-ogg/planos-de-acao/1o-plano-de-acao/balanco-primeiro-plano.pdf>
- (13) Governo Aberto – 2o Plano de Ação do Brasil – publicado em: 18/12/2014 - Disponível em: <https://www.gov.br/cgu/pt-br/governo-aberto/a-ogg/planos-de-acao/2o-plano-de-acao-brasileiro>
- (14) Governo Federal - Dados Abertos – Acessado em: 14/05/2023 – Disponível em: <https://dados.gov.br/home>
- (15) Governo Federal – Diário da União – Publicado em: 13/04/2012 – Disponível em: <https://www.gov.br/governodigital/pt-br/dados-abertos/InstrucaoNormativaINDA42012.pdf>

- (16) LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011. – Lei de Acesso à Informação –
Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm
- (17) Portal de Dados Abertos do Estado de Minas Gerais – Sobre o Portal – Disponível em:
<https://dados.mg.gov.br/about>
- (18) Portal de Dados Abertos do Estado de Minas Gerais – Dataset Despesa Pública –
Disponível em: <https://dados.mg.gov.br/dataset/despesa>
- (19) Amazon AWS – Serviço de Computação em Nuvem – Disponível em:
<https://aws.amazon.com/pt/>
- (20) Amazon AWS - O que é AWS? Como funciona Amazon Web Services – Disponível em:
<https://aws.amazon.com/pt/what-is-aws/>
- (21) Amazon AWS - Armazenamento S3 - Simple Storage Service – Disponível em:
<https://aws.amazon.com/pt/s3/>
- (22) Amazon AWS – O que é o AWS Lambda – Disponível em:
https://docs.aws.amazon.com/pt_br/lambda/latest/dg/welcome.html
- (23) Amazon AWS – O que é o AWS Glue – Disponível em:
https://docs.aws.amazon.com/pt_br/glue/latest/dg/what-is-glue.html
- (24) LITERATURE REVIEW OF BUSINESS INTELLIGENCE - Rasmey Heang and Raghul
Mohan School of Business and Engineering Halmstad University, Sweden
- (25) Clark, T. D., Jones, M. C., Armstrong, C. P. (2007). The Dynamic Structure of
Management Support Systems: Theory development, research focus and directions. MIS
Quarterly.
- (26) Neykova, M. ., & Zhelyazova, B. . (2020). THE ROLE OF BUSINESS INTELLIGENT
SYSTEMS IN MONITORING AND ANALYSIS OF UNIVERSITY DATA. Proceedings of
CBU in Natural Sciences and ICT, 1, 54-59. <https://doi.org/10.12955/pns.v1.121>
- (27) CDW Corporation - How the Modern Data Platform Fuels Success - Jesus Diaz,
Christopher Marcolis, Rex Washburn – Disponível em:
<https://www.cdw.com/content/cdw/en/articles/dataanalytics/how-the-modern-data-platform-fuels-success.html>
- (28) Harvard Business Review – Whats your data strategy – Thomas H. Davenport –
Publicado em: 06/2017 – Disponível em: <https://hbr.org/2017/05/whats-your-data-strategy>
- (29) Howson, C., Richardson, J., Sallam, R. &Kronz, A. (2019). Magic quadrant for analytics
and business intelligence platforms, Gartner, Retrieved
from <https://www.gartner.com/doc/3900992/magic-quadrant-analytics-business-intelligence>.
- (30) PUC Minas Gerais – Regulamento do Projeto Integrado – Disponível em:
<https://pucminas.instructure.com/courses/64608/pages/regulamento-do-projeto-integrado>
- (31) Oracle Cloud Infrastructure – O que é ETL – Disponível em:
<https://www.oracle.com/br/integration/what-is-etl/>

- (32) LinkedIn - How do you leverage AI and ML to enhance your BI capabilities and insights?
- Disponível em: <https://www.linkedin.com/advice/0/how-do-you-leverage-ai-ml-enhance-your-bi-capabilities>
- (33) Databricks – Machine Learning Models – Disponível em:
<https://www.databricks.com/glossary/machine-learning-models>
- (34) Shumway, R. H.; Stoffer, D. S. (2017). Time series analysis and its applications: With R examples. Springer.
- (35) Han, J., Tong, H., & Pei, J. (2023a). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.
- (36) Aggarwal, C. C. (2017). Outlier analysis. Springer.
- (37) Scikit-Learn – Sklearn.Emsemble.IsolationForest – Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html>
- (38) Statsmodels - statsmodels.tsa.arima.model.ARIMA – Disponível em:
<https://www.statsmodels.org/dev/generated/statsmodels.tsa.arima.model.ARIMA.html>
- (39) MLXtend - association_rules: Association rules generation from frequent itemsets -
Disponível em:
https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/
- (40) Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The definitive guide to dimensional modeling. Wiley.
- (41) Inmon, W. H. (2011). Building the data warehouse. Wiley.