

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Engenharia de Dados

Marcela Simão Gomes da Silva

**IMPLEMENTAÇÃO DE UM DATA LAKE NA AZURE: UM ESTUDO PRÁTICO EM
ENGENHARIA DE DADOS**

Belo Horizonte

2024

Marcela Simão Gomes da Silva

**IMPLEMENTAÇÃO DE UM DATA LAKE NA AZURE: UM ESTUDO PRÁTICO EM
ENGENHARIA DE DADOS**

Trabalho de Conclusão de Curso apresentado
ao Curso de Especialização em Engenharia de
Dados como requisito parcial à obtenção do
título de especialista.

Belo Horizonte
2024

SUMÁRIO

1.	INTRODUÇÃO	4
1.1.	CONTEXTUALIZAÇÃO	4
1.2.	O PROBLEMA PROPOSTO	4
1.3.	OBJETIVOS	4
2.	MODELAGEM CONCEITUAL E DEFINIÇÃO DAS TECNOLOGIAS/FERRAMENTAS/ARQUITETURA .	5
2.1.	ARQUITETURA	5
2.2.	CRIAÇÃO DO AMBIENTE NO PORTAL AZURE	6
2.3.	DADOS DE ORIGEM	9
2.3.1.	CADASTRO ÚNICO	9
2.3.2.	BENEFÍCIOS AO CIDADÃO.....	10
2.3.3.	MUNICÍPIOS.....	12
3.	INGESTÃO DE DADOS	12
4.	VISUALIZAÇÃO DE DADOS	16
4.1.	PAINEL.....	20
5.	LINKS	21
	REFERÊNCIAS.....	22

1. INTRODUÇÃO

1.1. CONTEXTUALIZAÇÃO

A Lei Nº 12.527, DE 18 DE NOVEMBRO DE 2011¹ e o decreto Nº 7.724, DE 16 DE MAIO DE 2012², visa garantir o acesso a informações dos órgãos públicos pelos cidadãos brasileiros. Desde então é obrigatório a divulgação dos dados públicos e muitos deles são divulgados via portal da transparência (dados abertos).

Assim, dados abertos são uma metodologia para a publicação de dados do governo em formatos reutilizáveis, visando o aumento da transparência e maior participação política por parte do cidadão, além de gerar diversas aplicações desenvolvidas colaborativamente pela sociedade.

Com a vasta volumetria de dados disponíveis nos sites de dados públicos, como o Portal da Transparência e o gov.br, a tarefa de analisar todo o histórico torna-se complexa. Uma delas é relacionada a problemática dos custos relacionados aos benefícios sociais no Brasil, por exemplo: o Bolsa Família e o Auxílio Emergencial, sendo uma questão crucial que demanda uma abordagem analítica.

1.2. O PROBLEMA PROPOSTO

A crescente demanda por programas sociais impõe desafios significativos na gestão eficiente dos recursos públicos, destacando a necessidade de estratégias de engenharia de dados para lidar com a dimensão e complexidade dessas informações. Além disso, a dificuldade de integração e interpretação dos dados, ressalta a importância de ferramentas avançadas de processamento e análise, visando extrair insights significativos para embasar políticas públicas mais eficazes.

Nesse contexto, a engenharia de dados surge como um instrumento fundamental para enfrentar os desafios inerentes a captura e tratamento dos dados relacionados à gestão de benefícios sociais, promovendo uma abordagem mais eficiente e transparente na alocação de recursos pelos municípios brasileiros.

1.3. OBJETIVOS

O objetivo do presente trabalho é capturar dados públicos relacionados a benefícios concedidos aos cidadãos brasileiros, cadastro único, dados do IBGE, armazená-los em um data lake e por fim disponibilizá-los em forma de painel para análise dos dados de maneira unificada.

De acordo com a IBM³, data lake é um repositório centralizado para gerenciar volumes de dados extremamente grandes. Ele serve como uma base para coleta e análise de dados estruturados, semiestruturados e não-estruturados em seus formatos nativos, para orientar novos insights.

A preferência recai sobre essa alternativa em virtude da significativa volumetria de dados, alcançando mais de 4 bilhões de linhas de histórico, as quais são disponibilizadas pelos órgãos públicos de diferentes lugares e com diferentes formatações.

2. MODELAGEM CONCEITUAL E DEFINIÇÃO DAS TECNOLOGIAS/FERRAMENTAS/ARQUITETURA

Esse tópico tem por objetivo consolidar os temas de arquitetura, ferramentas utilizadas e dados obtidos.

2.1. ARQUITETURA

A arquitetura do projeto está descrita abaixo:

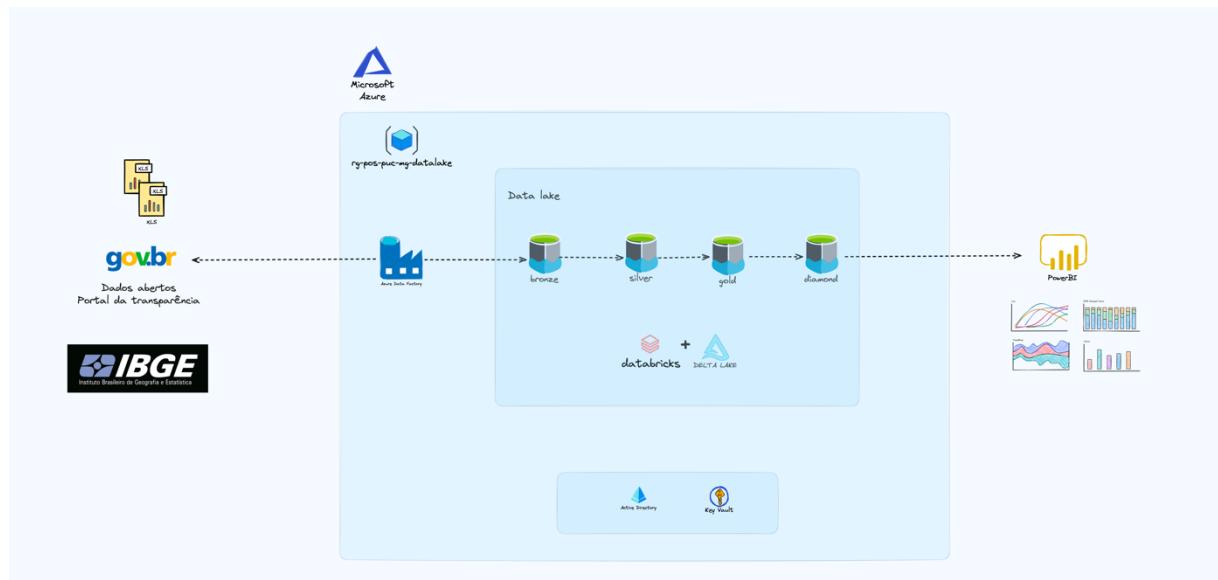


Figura 1: Arquitetura do projeto

O projeto será desenvolvido na Azure, plataforma de computação em nuvem da Microsoft, oferecendo uma ampla gama de serviços e soluções para hospedar, gerenciar e desenvolver aplicativos na nuvem.

Será utilizado os seguintes elementos:

Resource group: Organizador e gerenciador dos recursos associados a uma solução em nuvem.

Data factory: Integrador e orquestrador de pipelines para ingestão, validação, tratamento e armazenamento dos dados entre a origem e as camadas do data lake.

Storage account: Serviço para armazenamento dos dados de todas as camadas do data lake.

Databricks: Plataforma de análise baseada em Spark e otimizada para serviços de nuvem. Será o principal componente de processamento dos dados.

Key vault: Serviço criptográfico em nuvem para gerenciamento de chaves e segredos.

Power BI: Coleção de software, aplicativos e conectores que juntos transformam fontes de dados não relacionadas em informações visuais que será utilizado a construção do painel.

Para as camadas do data lake, será utilizado o conceito de arquitetura medalhão, que de acordo com a Databricks⁴, se refere ao design de dados usado para organizar logicamente os dados do data lake, que visa melhorar de forma incremental e progressiva a estrutura e a qualidade dos dados à medida que fluem pelas camadas da arquitetura (Bronze ⇒ Silver ⇒ Gold ⇒ Diamond). Também são conhecidas como arquitetura “multi-hop”.

Conceito das camadas:

Bronze: camada de recebimento do dado da origem, o mais próximo possível da formatação da fonte original.

Silver: aplicação de padronização de formatos, convenção de nomes de campo, caso seja necessário e apresenta o histórico dos dados.

Gold: aplicação da visão mais atualizada dos dados e aplicação de regra de negócio, quando necessário.

Diamond: definição de indicadores, geração de dados prontos para visualização e espaço de trabalho de analistas e exploração dos dados.

Em todas as camadas os dados serão armazenados em formato delta, no qual é um projeto de código aberto da Databricks que fornece funcionalidades avançadas para armazenamento e gerenciamento de dados em data lake.

2.2. CRIAÇÃO DO AMBIENTE NO PORTAL AZURE

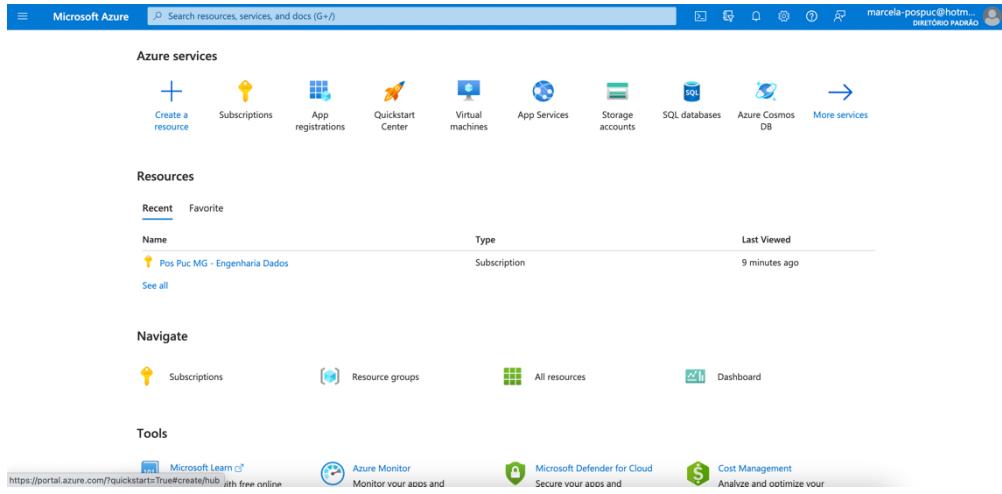


Figura 2 – Portal Azure

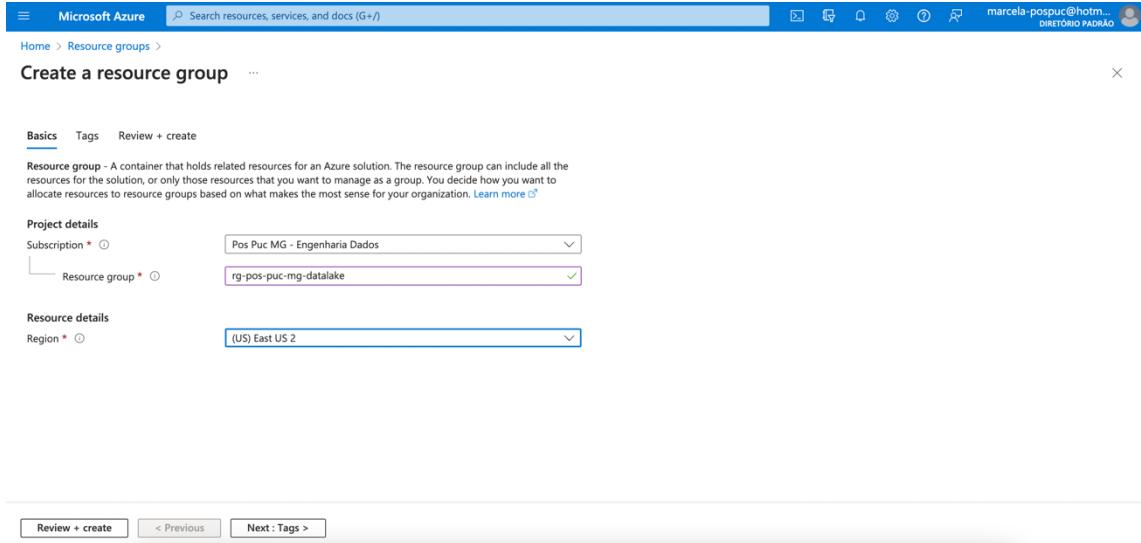


Figura 3 – Criação do resource group

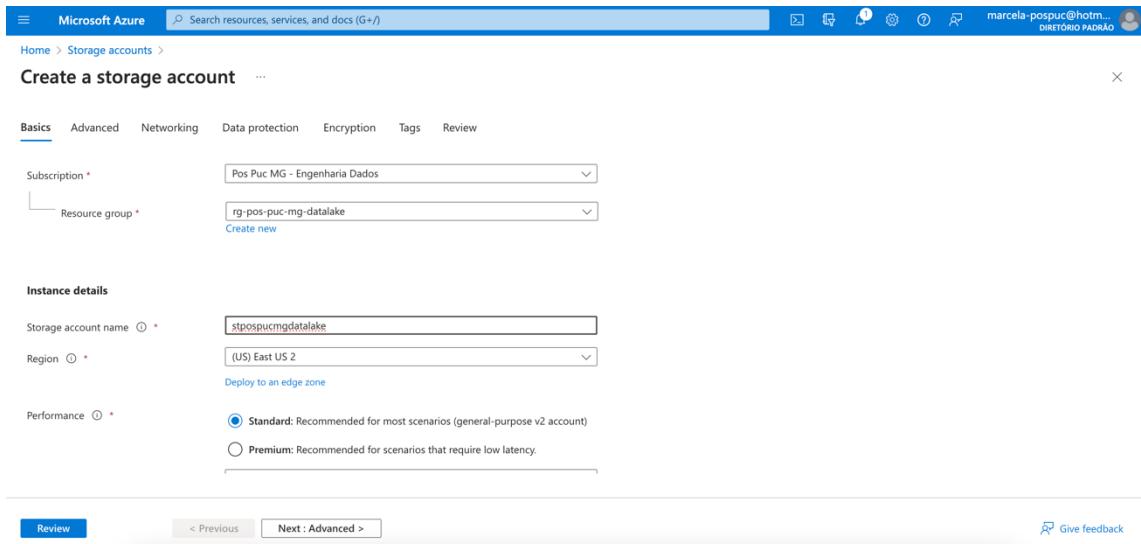


Figura 4 – Criação do storage account

The screenshot shows the Microsoft Azure Storage Explorer interface. On the left, a sidebar lists navigation options: Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage browser, Data storage (Containers, File shares, Queues, Tables), Security + networking (Networking, Access keys), and Help. The main area displays a table of containers:

Name	Last modified	Anonymous access level	Lease state
Slogs	19/01/2024, 20:07:27	Private	Available
bronze	19/01/2024, 20:08:52	Private	Available
diamond	19/01/2024, 20:09:06	Private	Available
gold	19/01/2024, 20:08:59	Private	Available
silver	19/01/2024, 20:08:29	Private	Available

Figura 5 – Estrutura dos contêineres do data lake

The screenshot shows the Microsoft Azure Data Factory Studio interface. On the left, a sidebar lists navigation options: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings (Networking, Managed identities, Properties, Locks), Getting started (Quick start), Monitoring (Alerts, Metrics), and Automation (Virtual Network Peerings, Encryption, Networking, Properties, Locks, Diagnostic settings). The main area displays the 'Essentials' section for the 'adf-pos-puc-mg-datalake' data factory:

Resource group (move)	Status	Type
rg-pos-puc-mg-datalake	Succeeded	Data factory (V2)
Subscription (move)	Pos Puc MG - Engenharia Dados	Getting started : Quick start
Subscription ID	7595a811-761c-4d8a-a8a6-3f8c58013f28	

A central icon represents the data factory, and below it is the text "Azure Data Factory Studio". A blue button labeled "Launch studio" is visible.

Figura 6 – Criação do Azure Data Factory

The screenshot shows the Microsoft Azure Databricks Service interface. On the left, a sidebar lists navigation options: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings (Virtual Network Peerings, Encryption, Networking, Properties, Locks), Monitoring (Diagnostic settings), and Automation (CLI / PS, Tasks (In progress)). The main area displays the "Essentials" section for the 'adb-pos-puc-mg-datalake' workspace:

Status	Managed Resource Group
Active	mrgr-pos-puc-mg-datalake
Resource group	rg-pos-puc-mg-datalake
Location	East US 2
Subscription	Pos Puc MG - Engenharia Dados
Subscription ID	7595a811-761c-4d8a-a8a6-3f8c58013f28
Tags (edit)	Add tags

A central icon represents the workspace, and below it is the text "Launch Workspace". Below the workspace are four buttons: Documentation, Getting Started, Import Data from File, and Import Data from Azure Storage.

Figura 7 – Criação do Azure Databricks

The screenshot shows the Microsoft Azure Key Vault 'Overview' page for a vault named 'kv-pos-puc-mg-datalake'. The main pane displays basic details such as the resource group ('rg-pos-puc-mg-datalake'), location ('East US 2'), and subscription ('Pos Puc MG - Engenharia Dados'). The vault's URI is listed as 'https://kv-pos-puc-mg-datalake.vault.azure.net/'. A prominent warning at the top right states: 'Upcoming TLS 1.0, 1.1 deprecation: Please enable support for TLS 1.2 on clients (applications/platform) to avoid any service impact. Learn more here.' Below this, there are tabs for 'Activity log', 'Tags', 'Diagnose and solve problems', 'Access policies', and 'Events'. On the left, a sidebar lists 'Objects' (Keys, Secrets, Certificates), 'Settings' (Access configuration, Networking, Microsoft Defender for Cloud), and 'Properties'. At the bottom, there are links for 'Get started', 'Properties', 'Monitoring', 'Tools + SDKs', and 'Tutorials'. A 'Manage keys and secrets used by apps and services' section is also present.

Figura 8 – Criação do Key Vault

2.3. DADOS DE ORIGEM

2.3.1. CADASTRO ÚNICO

De acordo com o portal do Ministério do desenvolvimento social⁵, o cadastro único é um grande mapa das famílias de baixa renda no Brasil. Ele mostra ao governo quem essas famílias são, como elas vivem e do que elas precisam para melhorar suas vidas.

Famílias inscritas no Cadastro Único - MI Social

Link do dataset: <https://dados.gov.br/dados/conjuntos-dados/familias-inscritas-no-cadastro-unico---mi-social>

Neste conjunto de dados temos as seguintes variáveis:

COLUNA	DESCRIÇÃO
codigo_ibge	Código IBGE do município.
anomes	Ano/mês de referência do dado.
cadun_qtd_familias_cadastradas_i	Quantidade total de famílias inscritas no Cadastro Único

Famílias por faixa de renda no Cadastro Único - MI Social

Link dataset: <https://dados.gov.br/dados/conjuntos-dados/familias-por-faixa-de-renda-no-cadastro-unico---mi-social>

Neste conjunto de dados temos as seguintes variáveis:

COLUNA	DESCRIÇÃO
codigo_ibge	Código IBGE do município.
anomes	Ano/mês de referência do dado.
cadun_qtd_familias_cadastradas_pobreza_pbf_i	Quantidade de famílias em situação de pobreza, segundo a faixa do Programa Bolsa Família, inscritas no Cadastro Único

cadun_qtd_familias_cadastradas_baixa_renda_i	Quantidade de famílias de baixa renda inscritas no Cadastro Único
cadun_qtd_familias_cadastradas_rfpc_ate_meio_sm_i	Quantidade de famílias com renda per capita mensal até meio salário-mínimo (Pobreza + Baixa renda) inscritas no Cadastro Único
cadun_qtd_familias_cadastradas_rfpc_acima_meio_sm_i	Quantidade de famílias com renda per capita mensal acima de meio salário-mínimo inscritas no Cadastro Único

2.3.2. BENEFÍCIOS AO CIDADÃO

Recursos publicados no Portal da Transparência do Governo Federal que são repassados diretamente a cidadãos.

Link dataset: <https://portaldatransparencia.gov.br/download-de-dados>

Auxílio Brasil:

COLUNA	DESCRIÇÃO
mês competência	Ano/Mês a que se refere a parcela, no formato AAAAMM
mês referência	Mês da folha de pagamento, no formato AAAAMM
uf	Sigla da Unidade Federativa do beneficiário do Auxílio Brasil
código município siafi	Código do município do beneficiário do Auxílio Brasil no SIAFI (Sistema Integrado de Administração Financeira)
nome município	Nome do município do beneficiário do Auxílio Brasil
cpf favorecido	Número no Cadastro de Pessoas Físicas (CPF) do beneficiário do Auxílio Brasil, caso possua
nis favorecido	Número de Identificação Social (NIS) do beneficiário do Auxílio Brasil, caso possua
nome favorecido	Nome do beneficiário do Auxílio Brasil
data saque	Data em que foi realizado o saque
valor parcela	Valor da parcela do benefício

Auxílio emergencial:

COLUNA	DESCRIÇÃO
mês disponibilização	Ano/Mês a que se refere a parcela, no formato AAAAMM.
uf	Sigla da Unidade Federativa do beneficiário do Auxílio Emergencial.
código município ibge	Código, no IBGE (Instituto Brasileiro de Geografia e Estatística), do município do beneficiário do Auxílio Emergencial.
nome município	Nome do município do beneficiário do Auxílio Emergencial.
nis beneficiário	Número de Identificação Social (NIS) do beneficiário do Auxílio Emergencial, caso possua.
cpf beneficiário	Número no Cadastro de Pessoas Físicas (CPF) do beneficiário do Auxílio Emergencial, caso possua.
nome beneficiário	Nome do beneficiário do Auxílio Emergencial.
nis responsável	Número de Identificação Social (NIS) do responsável pelo beneficiário do Auxílio Emergencial, caso possua.
cpf responsável	Número no Cadastro de Pessoas Físicas (CPF) do responsável pelo beneficiário do Auxílio Emergencial, caso possua.
nome responsável	Nome do responsável pelo beneficiário do Auxílio Emergencial, caso possua.
enquadramento	Identifica se o beneficiário é do grupo Bolsa Família, Inscrito no Cadastro Único (CadÚnico) ou Não Inscrito no Cadastro Único (ExtraCad).
parcela	Número sequencial da parcela disponibilizada.
observação	Indica alterações na parcela disponibilizada como, por exemplo, se foi devolvida ou está retida.
valor benefício	Valor disponibilizado na parcela.

Bolsa Família – Pagamentos:

COLUNA	DESCRÍÇÃO
ano/mês referência	Ano/Mês da folha de pagamento
ano/mês competência	Ano/Mês a que se refere a parcela
uf	Sigla da Unidade Federativa do beneficiário do Bolsa Família
código município siafi	Código, no SIAFI (Sistema Integrado de Administração Financeira), do município do beneficiário do Bolsa Família
nome município siafi	Nome do município do beneficiário do Bolsa Família
cpf beneficiário	Número no Cadastro de Pessoas Físicas (CPF) do beneficiário do Bolsa Família, caso possua.
nis beneficiário	NIS do beneficiário do Bolsa Família Criado pela Caixa Econômica Federal o NIS significa Número de Identificação Social e é ganho quando o cidadão brasileiro ingressa em algum Programa Social, seja o Bolsa Família, FGTS, emitiu sua Carteira de Trabalho, tornou-se contribuinte do INSS ou iniciou sua vida como trabalhador de iniciativa privada ou pública. Fonte: Caixa Econômica Federal
nome beneficiário	Nome do beneficiário do Bolsa Família
valor parcela	Valor da parcela do benefício

Bolsa Família – Saques:

COLUNA	DESCRÍÇÃO
ano/mês referência	Ano/Mês da folha de pagamento
ano/mês competência	Ano/Mês a que se refere a parcela
uf	Sigla da Unidade Federativa do beneficiário do Bolsa Família
código município siafi	Código, no SIAFI (Sistema Integrado de Administração Financeira), do município do beneficiário do Bolsa Família
nome município siafi	Nome do município do beneficiário do Bolsa Família
cpf beneficiário	Número no Cadastro de Pessoas Físicas (CPF) do beneficiário do Bolsa Família, caso possua.
nis beneficiário	NIS do beneficiário do Bolsa Família Criado pela Caixa Econômica Federal o NIS significa Número de Identificação Social e é ganho quando o cidadão brasileiro ingressa em algum Programa Social, seja o Bolsa Família, FGTS, emitiu sua Carteira de Trabalho, tornou-se contribuinte do INSS ou iniciou sua vida como trabalhador de iniciativa privada ou pública. Fonte: Caixa Econômica Federal
nome beneficiário	Nome do beneficiário do Bolsa Família
data saque	Data em que foi realizado o saque
valor parcela	Valor da parcela do benefício

Novo Bolsa Família:

COLUNA	DESCRÍÇÃO
mês competência	Ano/Mês a que se refere a parcela, no formato AAAAMM
mês referência	Mês da folha de pagamento, no formato AAAAMM
uf	Sigla da Unidade Federativa do beneficiário do Novo Bolsa Família
código município siafi	Código do município do beneficiário do Novo Bolsa Família no SIAFI (Sistema Integrado de Administração Financeira)
nome município	Nome do município do beneficiário do Novo Bolsa Família

cpf favorecido	Número no Cadastro de Pessoas Físicas (CPF) do beneficiário do Novo Bolsa Família, caso possua
nis favorecido	Número de Identificação Social (NIS) do beneficiário do Novo Bolsa Família, caso possua
nome favorecido	Nome do beneficiário do Novo Bolsa Família
data disponibilização	Data em que foi disponibilizada a parcela
valor parcela	Valor da parcela do benefício

2.3.3. MUNICÍPIOS

Informações do IBGE relacionados a municípios do brasil

Link dataset: <https://www.ibge.gov.br/explica/codigos-dos-municipios.php>

COLUNA	DESCRIÇÃO
uf	Abreviatura do estado onde o município está localizado
nome_uf	O nome do estado onde está localizado o município
regiao_geografica_intermediaria	O código da região geográfica intermediária onde o município está localizado
nome_regiao_geografica_intermediaria	O nome da região geográfica intermediária onde o município está localizado
regiao_geografica_imediata	O código da região geográfica imediata onde o município está localizado
nome_regiao_geografica_imediata	O nome da região geográfica imediata onde o município está localizado
mesorregiao_geografica	O código da mesorregião onde está localizado o município
nome_mesorregiao	O nome da mesorregião onde está localizado o município
microrregiao_geografica	O código da microrregião onde está localizado o município
nome_microrregiao	O nome da microrregião onde está localizado o município
municipio	O código do município
codigo_municipio_completo	O código completo do município, incluindo os códigos de estado e região
nome_municipio	O nome do município
distrito	O código do distrito dentro do município (se aplicável)
codigo_distrito_completo	O código completo do distrito, incluindo os códigos de estado, região e município (se aplicável)
nome_distrito	O nome do distrito dentro do município (se aplicável)

3. INGESTÃO DE DADOS

A etapa de ingestão de dados é efetuada por intermédio do Azure Data Factory, como descrito na arquitetura anteriormente. Este processo foi organizado dentro de uma pipeline de dados, visando facilitar a transição dos dados desde sua captação na fonte até a camada "gold" do Data Lake. Esse fluxo estruturado garante a promoção eficiente dos dados, mantendo a integridade e a qualidade ao longo do ciclo de vida no ambiente do Data Lake.

Microsoft Azure | Data Factory > adf-pos-puc-mg-datalake | Search factory and documentation | marcela-pospuc@hotmail.com | DIRETÓRIO Padrão | Preview experience Off

Factory Resources

- Pipelines (8)
 - Ingestao (8)
 - GOVBR (2)
 - Cadastro Unico (3)
 - pip_ing_govbr_cadunico_faixarenda_full
 - pip_ing_govbr_cadunico_familias_insc_full
 - IBGE (1)
 - pip_ing_ibge_municpios_full
 - Portal Transparencia (5)
 - pip_ing_portal_transparencia_aux_br_full
 - pip_ing_portal_transparencia_aux_emerg_full
 - pip_ing_portal_transparencia_bolsa_fam_pgto_full
 - pip_ing_portal_transparencia_bolsa_saq_full
 - pip_ing_portal_transparencia_novo_bolsa_fam_full
- Change Data Capture (preview) (0)
- Datasets (4)

Figura 9 – Organização das pipelines do Azure Data Factory

Demonstração da pipeline “pip_ing_govbr_cadunico_faixarenda_full”

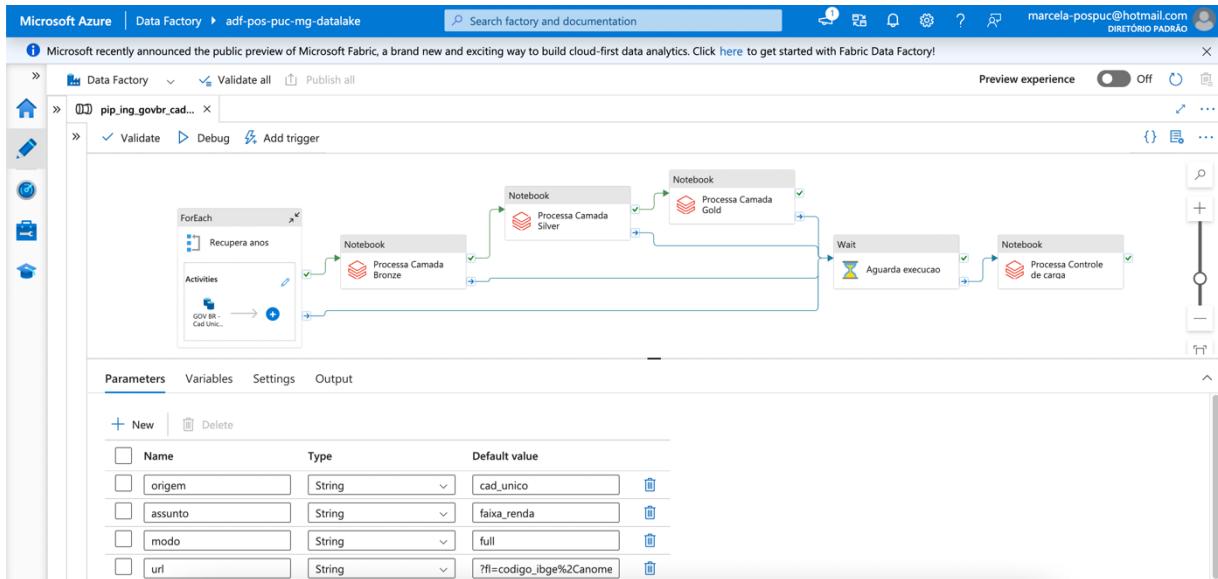


Figura 10 – Exemplo de pipeline no Azure Data Factory -“pip_ing_govbr_cadunico_faixarenda_full”

A pipeline recebe 4 parâmetros para o processamento dos dados:

Origem – sistema ou origem dos dados (ex: cad_unico | ibge | portal_transparencia)

Assunto – assunto relacionado aos dados (ex: faixa_renda | familias_insc | municípios

| auxilio_brasil | auxilio_emergencial | bolsa_familia_pagamentos|
bolsa_familia_saques) | novo_bolsa_familia

Modo – modo da ingestão (ex: full | delta)

url - url relativa ao processo de captura de dados (ex: ?fl=codigo_ibge%2Canomes_s%20cadun_qtd_familias_atualizadas_i%20cadun_qtd_familias_atualizadas_pobreza_pbf_i%20cadun_qtd_familias_atualizadas_baixa_renda_i%20cadun_qtd_familias_atualizadas_rfpc_ate_meio_sm_i%20cadun_qtd_familias_atualizadas_rfpc_acima_meio_sm_i%20cadun_qtd_familias_atualizadas_renda_zero_i%20cadun_taxa_atualizacao_cadastral_d%20cadun_taxa_atualizacao_cadastral_rfpc_ate_meio_sm_d&fq=cadun_qtd_familias_cadastradas_i%3A*&q=%3A*&rows=100000&sort=anomes_s%20desc%2C%20codigo_ibge%20asc&wt=csv&fq=ano_mes_s:)

A primeira caixinha “Recupera anos” realiza a captura dos dados da origem na forma recursiva de acordo com os anos de histórico que existe para o dataset. A conexão é feita por meio de uma requisição http (get).

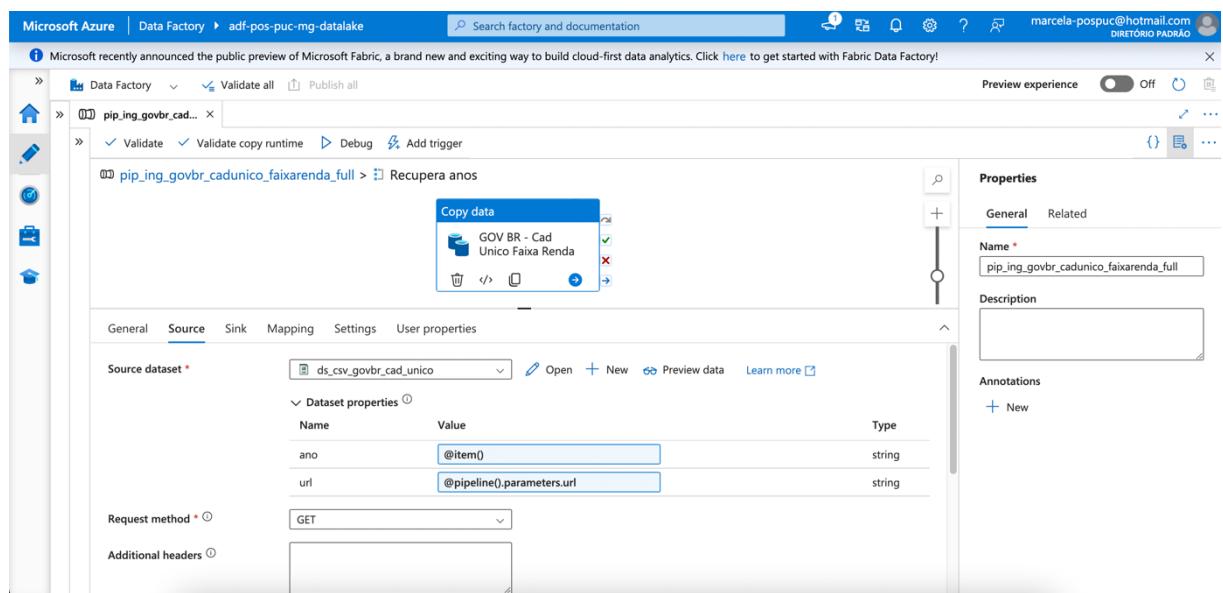


Figura 11 – Configuração do copy data aba origem

O copy data recupera o arquivo da origem e salva em formato parquet, que é um formato de arquivo de coluna eficiente para armazenamento e processamento de dados em big data, no data lake, na camada “bronze”.

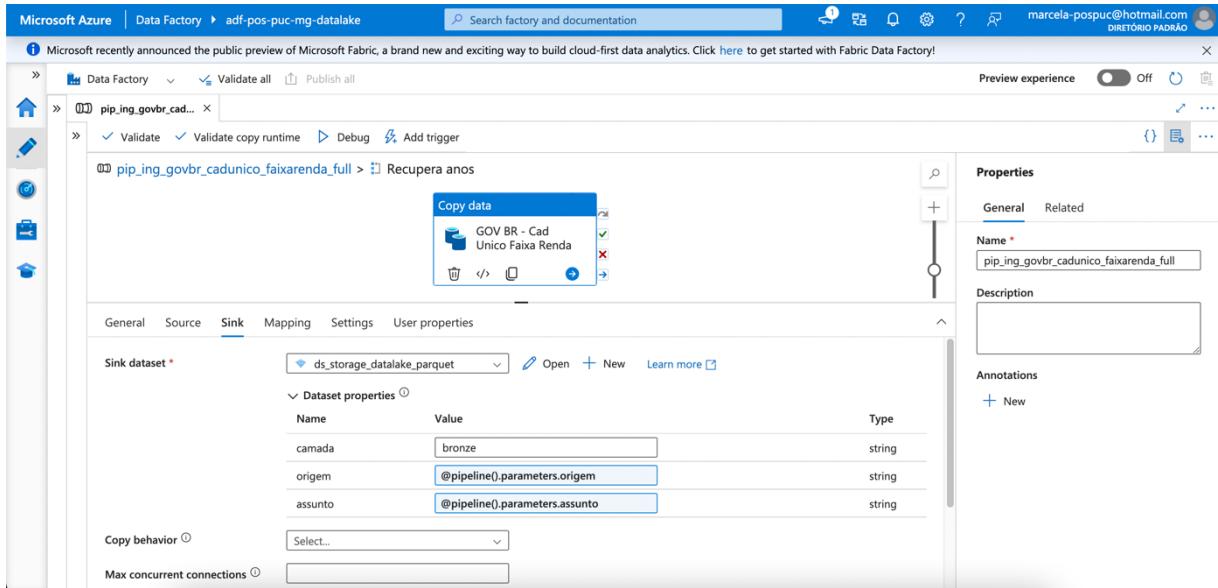


Figura 12 – Configuração do copy data aba destino

Depois de recuperar todas as informações da origem, inicia-se o processo de conversão para o formato delta e passagem entre as camadas do data lake.

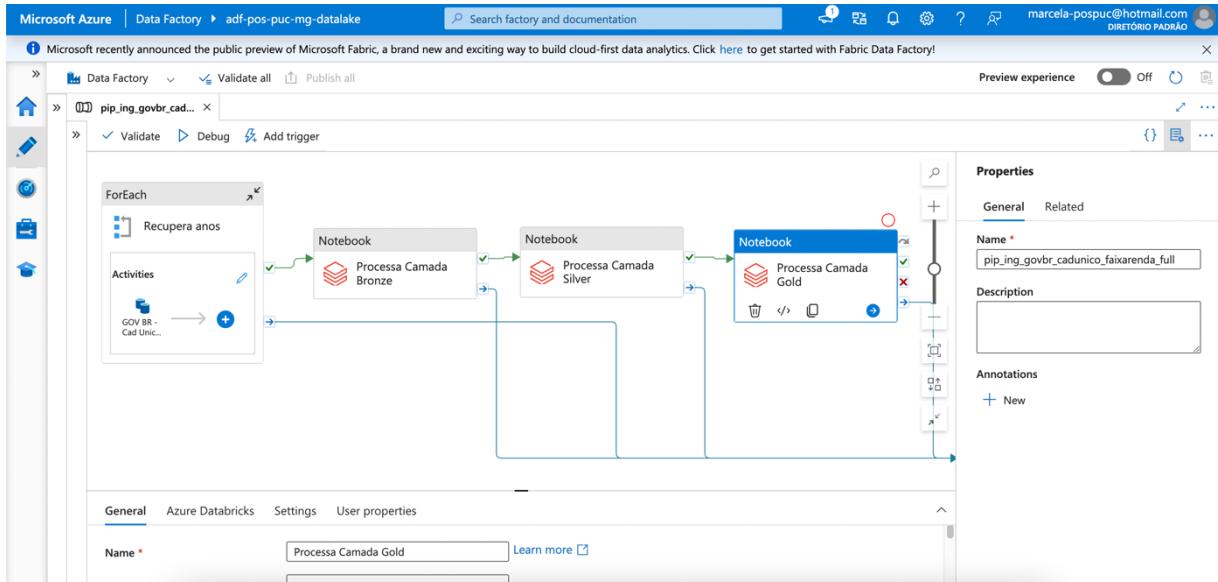


Figura 13 – Processamento das camadas do data lake

O processamento é feito por meio de um código pyspark, que é a interface Python para o Apache Spark, um poderoso framework de processamento de dados distribuído em larga escala, na Ferramenta Databricks, que está armazenado no projeto git hub com a nomenclatura de “nt_py_tabela_controle”.

Por fim todos as execuções são armazenadas em uma tabela de controle, com o nome da pipeline e o tempo de execução das caixinhas do Azure Data Factory, o código desse também é pyspark e estará no git hub com o nome “nt_py_controle_carga”

No término da execução da pipeline espera-se que tenha a tabela nas 3 camadas do data lake e esteja disponível para modelagem dos dados na camada de análise “Diamond”.

codigo_ibge	anomes_s	cadun_qtd_familias_atualizadas_i	cadun_qtd_familias_atualizadas_pobreza_pbf_i	cadun_qtd_familias_atualizadas_baixa_renda_i	cadun_qtd_familias
110001	202112	2399			1707
110002	202112	8921			6828
110003	202112	372			305
110004	202112	8882			6223
110005	202112	1370			1074
110006	202112	1206			782
110007	202112	818			616
110008	202112	2577			1992
110009	202112	2753			1764
110010	202112	5060			4117
110011	202112	4757			3380
110012	202112	9957			6470
110013	202112	3276			2228
110014	202112	2720			2016
110015	202112	3703			2513

Figura 14 – Exemplo das tabelas criadas nas camadas do data lake

4. VISUALIZAÇÃO DE DADOS

Para a visualização dos dados, foram criadas tabelas na camada “Diamond” do data lake modelando as regras de negócios. A ideia é que o painel construído no Power BI, se alimente dessa camada.

Foi construído uma visão de municípios, com o cruzamento da base do IBGE e do censo divulgado por eles:

```

1 %sql
2 create or replace viewdatalake.diamond.vw_municipios as
3 select
4   m.uf as cod_uf,
5   p.uf as sigla_uf,
6   m.nome_uf as uf,
7   m.codigo_municipio_completo,
8   p.cod_municipio,
9   p.nome_municipio,
10  s.cod_siafi,
11  p.`populacao` as populacao
12  | from datalake.gold.tb_ibge_base_populacao
13  left join (select distinct nome_uf, uf, codigo_municipio_completo, nome_municipio from datalake.gold.tb_ibge_municipios) m on m.
14  codigo_municipio_completo = p.cod_uf||p.cod_municipio
15  left join datalake.gold.tb_municipios_siafi_ibge s on s.cod_mun_ibge=p.cod_uf||p.cod_municipio

```

Figura 15 – Query de criação da tabela com as regras de negocio referente a municipio

New result table: ON								
Table		+						
	A _c cod_uf	A _c sigla_uf	A _c uf	A _c codigo_municipio_completo	A _c cod_municipio	A _c nome_municipio	A _c cod_siafi	A _c populacao
1	11	RO	Rondônia	1100015	00015	Alta Floresta D'Oeste	33	21558
2	11	RO	Rondônia	1100023	00023	Aríquemes	7	100896
3	11	RO	Rondônia	1100031	00031	Cabixi	37	5107
4	11	RO	Rondônia	1100049	00049	Cacoal	9	92202
5	11	RO	Rondônia	1100056	00056	Cerejeiras	27	15237
6	11	RO	Rondônia	1100064	00064	Colorado do Oeste	23	15747
7	11	RO	Rondônia	1100072	00072	Corumbiara	981	7503
8	11	RO	Rondônia	1100080	00080	Costa Marques	21	12633
9	11	RO	Rondônia	1100098	00098	Espigão D'Oeste	25	29722
10	11	RO	Rondônia	1100106	00106	Guajará-Mirim	1	39396
11	11	RO	Rondônia	1100114	00114	Jaru	15	52090
12	11	RO	Rondônia	1100122	00122	Ji-Paraná	5	136825
13	11	RO	Rondônia	1100130	00130	Machadinho D'Oeste	39	30626
14	11	RO	Rondônia	1100148	00148	Nova Brasilândia D'Oeste	41	17355
15	11	RO	Rondônia	1100155	00155	Ouro Preto do Oeste	17	36753

5,575 rows | 1.50 seconds runtime

Figura 16 – Dados da tabela criada para o painel referente a municípios

Para a base de benefícios foi unificado as tabelas criadas de cada dataset originalmente divulgado.

```

1 create or replace tabledatalake.gold.tb_beneficios_brasil as
2 select
3   ab.uf as cod_uf,
4   ab.uf as sigla_uf,
5   m.nome_uf as uf,
6   m.codigo_municipio_completo,
7   m.municipio as cod_municipio,
8   m.nome_municipio,
9   ae.uf as uf,
10  ab.mes_referencia,
11  ab.mes_competencia,
12  datalake.gold.encrypt(ab.cpf_favorecido) as cpfBeneficiario,
13  datalake.gold.encrypt(ab.nis_favorecido) as nisBeneficiario,
14  datalake.gold.encrypt(ab.nome_favorecido) as nomeBeneficiario,
15  null as pf,
16  cast(replace(ab.valor_parcela, ',', '.') as double) valorBeneficio,
17  null as dataSaque,
18  'Auxilio Brasil' as nomeBeneficio
19  from datalake.gold.tb_portal_transparencia_auxilio_brasil ab
20  left join datalake.gold.tb_municipios_siafi_ibge s on s.cod_siafi=cast(ab.CODIGO_MUNICIPIO_SIAFI)
21  left join (select distinct nome_uf, uf, codigo_municipio_completo, municipio, nome_municipio from datalake.gold.tb_ibge_municipios) m on m.codigo_municipio_completo = s.cod_mun_ibge
22
23  union all
24
25  select
26  m.uf as cod_uf,
27  ae.uf as sigla_uf,
28  m.nome_uf as uf,
29  m.codigo_municipio_completo,
30  m.municipio as cod_municipio,
31  m.nome_municipio,
32  s.cod_siafi,
33  ab.mes_referencia as mesReferencia,
34  null as mesCompetencia,
35  datalake.gold.encrypt(ab.cpfBeneficiario) cpfBeneficiario,
36  datalake.gold.encrypt(ab.nisBeneficiario) nisBeneficiario,
37  datalake.gold.encrypt(ab.nomeBeneficiario) nomeBeneficiario,
38  ae.parcela,
39  cast(replace(ab.valorBeneficio, ',', '.') as double) valorBeneficio,
40  null as dataSaque,

```

Figura 17– Query de criação da tabela de benefícios

Essa base unificada de benefícios tem a volumetria total de mais de 4 bilhões de linhas:

```
1   select count(*) from datalake.diamond.vw_beneficios_brasil  
2   |
```

► (6) trabalhos Spark

Tabela ▾ +	
1	1 ² 3 count(1) 4007837044

Figura 18 – Resultado da quantidade de linhas da tabela de benefícios

Estão divididas em:

```
select count(*) from datalake.gold.tb_portal_transparencia_bolsa_familia_pagamentos
```

Volumetria: 1.466.060.695

```
select count(*) from datalake.gold.tb_portal_transparencia_novo_bolsa_familia
```

Volumetria: 183.064.060

```
select count(*) from datalake.gold.tb_portal_transparencia_auxilio_emergencial
```

Volumetria: 781.655.982

```
select count(*) from datalake.gold.tb_portal_transparencia_auxilio_brasil
```

Volumetria: 293.956.157

Para ganhar performance no painel, foi construído indicadores já summarizados para o painel:

Quantidade de pessoas que obtiveram benefícios aos longos dos anos

```

1  create or replace tabledatalake.diamond.tb_beneficios_br_ano as
2
3  select left(ab.mes_referencia, 4) as ano,
4  count(distinct nis_favorecido) qtde_beneficios,
5  'Auxilio Brasil' as nome_beneficio
6  from datalake.gold.tb_portal_transparencia_auxilio_brasil ab
7  group by left(ab.mes_referencia, 4)
8
9
10 union all
11
12 select left(mes_disponibilizacao, 4) as ano,
13 count(distinct nis_beneficiario ) qtde_beneficios,
14 'Auxilio Emergencial' as nome_beneficio
15 from datalake.gold.tb_portal_transparencia_auxilio_emergencial ae
16 group by left(mes_disponibilizacao, 4)
17
18 union all
19
20
21
22 select left(mes_referencia, 4) as ano,
23 count(distinct nis_favorecido) qtde_beneficios,|
24 'Novo Bolsa Familia' as nome_beneficio
25 from datalake.gold.tb_portal_transparencia_novo_bolsa_familia nb
26 group by left(mes_referencia, 4)
27
28 union all
29
30 select left(mes_referencia, 4) as ano,
31 count(distinct nis_favorecido) qtde_beneficios,
32 'Bolsa Familia' as nome_beneficio
33 from datalake.gold.tb_portal_transparencia_bolsa_familia_pagamentos bp
34 group by left(mes_referencia, 4)

```

Figura 19 – Query de construção do indicador

Valor por município dispendido ao longo dos anos por benefícios:

```

1  create or replace tabledatalake.diamond.tb_beneficios_nv_bolsa_fam_ano as
2  select
3  TRANSLATE(m.nome_uf , 'àáââéêííôôôôûûçÃÃÃÉÉÉÍÍôôôôûûç', 'aaaaeeeiiooouccAAAAEEEII000UUC') as estado,
4  nb.uf as uf,
5  m.codigo_municipio_completo,
6  m.nome_municipio,
7  left(nb.mes_referencia, 4) ano,
8  sum(cast(replace(nb.valor_parcela, ',', '.') as double)) valor_beneficio,
9  DENSE_RANK () over (partition by left(nb.mes_referencia, 4) order by sum(cast(replace(nb.valor_parcela, ',', '.') as double)) desc) as DenseRank,
10 'Novo Bolsa Familia' as nome_beneficio
11 from datalake.gold.tb_portal_transparencia_novo_bolsa_familia nb
12 left join datalake.gold.tb_municipios_siafi_ibge s on s.cod_siafi=int(nb.CODIGO_MUNICIPIO_SIAFI)
13 left join (select distinct nome_uf, uf, codigo_municipio_completo, municipio, nome_municipio from datalake.gold.tb_ibge_municipios) m on m.codigo_municipio_completo = s.cod_mun_ibge
14 group by TRANSLATE(m.nome_uf , 'àáââéêííôôôôûûçÃÃÃÉÉÉÍÍôôôôûûç', 'aaaaeeeiiooouccAAAAEEEII000UUC'),
15 nb.uf,
16 m.codigo_municipio_completo,
17 m.nome_municipio,
18 left(nb.mes_referencia, 4)

```

Figura 20 – Query de construção do indicador

Todas as queries estão em anexo no git hub.

4.1. PAINEL

A conexão do Power BI foi feita por meio do conector Databricks Azure:

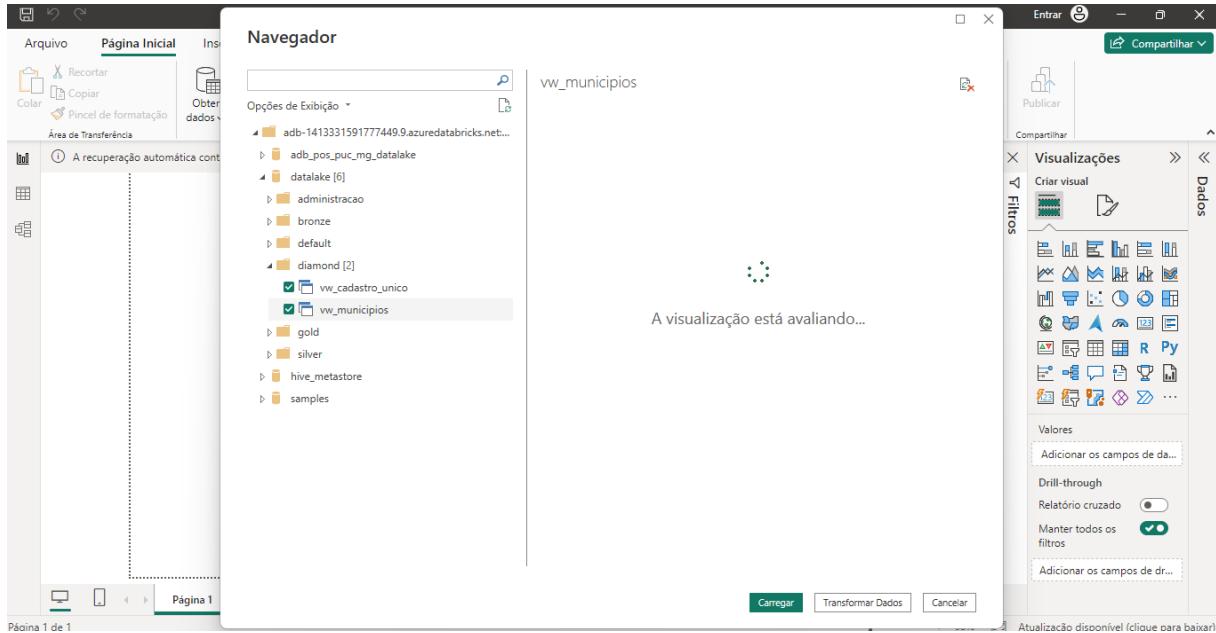


Figura 21 – Conexão Power BI com Azure Databricks

O painel contém duas páginas, a primeira é um resumo da população brasileira e como ela esta distribuída pelos estados, pelos benefícios e como é a evolução do cadastro único

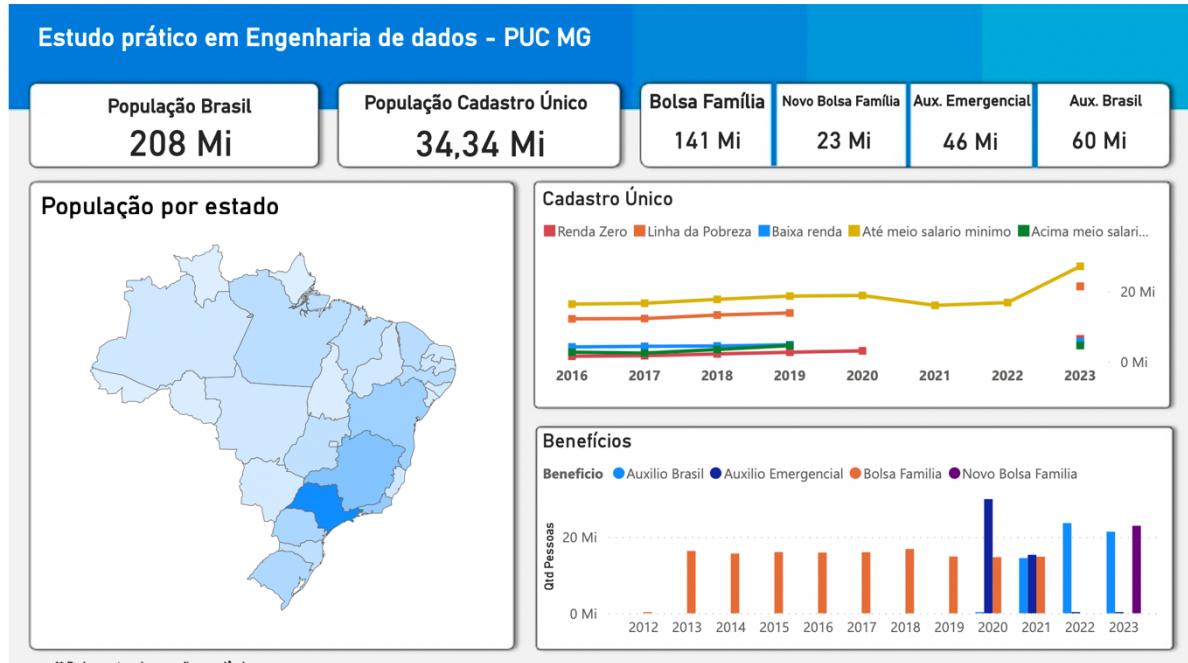


Figura 22 – Aba Visão Geral do painel

A segunda página demonstra como está distribuído os valores do benefícios pelos estados do brasil e os top 5 municípios e também é possível fazer filtros de: estado, beneficio e ano que se quer fazer a análise.

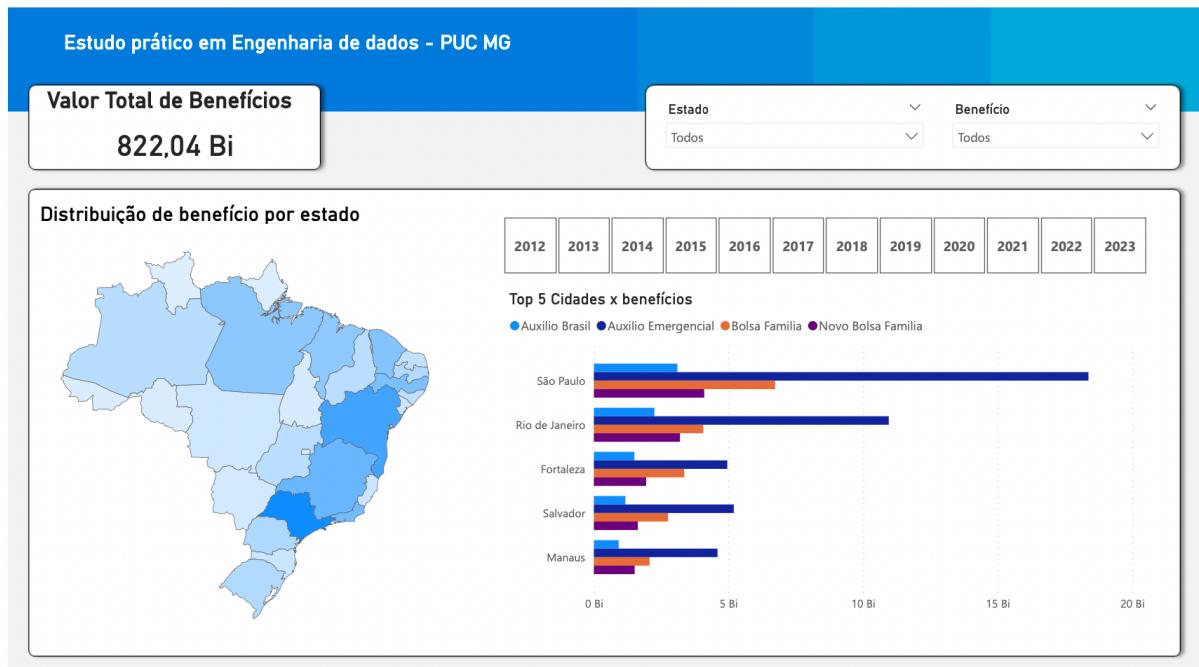


Figura 23 – Aba Visão Custos dos benefícios

5. LINKS

Link para o vídeo: <https://youtu.be/br6LnEo2bGg>

Link para o repositório: <https://github.com/simaomarcela/pos-puc-mg-eng-dados-2024>

REFERÊNCIAS

1. [Lei nº 12.527, de 18 de novembro de 2011. "Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências." Brasília: Presidência da República, 2011. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 03 de janeiro de 2024]
2. [Decreto nº 7.724, de 16 de maio de 2012. "Regulamenta a Lei nº 12.527, que dispõe sobre o acesso a informações." Brasília: Presidência da República, 2012. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/decreto/d7724.htm. Acesso em: 06 de janeiro de 2024]
3. [IBM Brasil. "IBM Data Lake." Disponível em: <https://www.ibm.com/br-pt/data-lake?utm_content=SRCWW&p1=Search&p4=43700078892965260&p5=e&gclid=CjwKCAiAk9itBhASEiwA1my_66So1IJrSKyjWsG0OTNi_2-o4hWUnsMGBTz1LqDlszz2zvIL4mrRERoCIL4QAvD_BwE&gclsrc=aw.ds>. Acesso em: 25 de janeiro de 2024]
4. [Databricks. "Medallion Architecture." Disponível em: <<https://www.databricks.com/br/glossary/medallion-architecture>>. Acesso em: 28 de janeiro de 2024]
5. [Ministério do Desenvolvimento Social (MDS). "Cadastro Único." Brasília: Governo Federal, s/d. Disponível em: <https://www.gov.br/mds/pt-br/acoes-e-programas/cadastro-unico>. Acesso em: 21 de janeiro de 2024]