



Desafio Stone 2021

Engenharia de Dados

O desafio

Procuradoria-Geral da Fazenda Nacional

Proteção de Dados (Lei nº 13.709/2018).

Salienta-se que os débitos inscritos em dívida ativa não estão cobertos por sigilo, conforme disposto no

Fica dica! Para uma melhor compreensão do teor dos arquivos, acesse o [dicionário de campos](#).

Conjunto de dados: "Devedores da União"


Descrição: conjunto de informações sobre débitos com a Fazenda Nacional ou o FGTS inscritos em Divi devedores, na condição de devedor principal, corresponsável ou solidário, atualizada trimestralmente.

Divida Ativa Geral (Sistema SIDA) Dívida FGTS Dívida Previdenciária (Sistema Dívida)

[download](#)

[download](#)

[download](#)

**BANCO CENTRAL DO BRASIL**

SGS - Sistema Gerenciador de Séries Temporais - v2.1
Módulo público

[Consultar](#) | [Minhas listas de séries](#) | [Configurações](#) | [Ajuda](#) | [Login](#)

[Início](#) → [Consultar séries](#) → Localizar séries

Pesquisa

Selecione a periodicidade

Todas

Selecione uma opção

Por tema

Por código

Por fonte

Abecip e BCB-Depec

Não há lista(s).
Para criar clique aqui

Séries mais pesquisadas

Séries desativadas

Pesquisa textual
(nome da série)

Localizar séries - Pesquisa por tema

+ Clique para visualizar Parâmetros de pesquisa

Total de séries localizadas: 8

Sel.	Cód.	Nome abreviado
<input type="checkbox"/>	21388	PTC - Grandes Empresas - Oferta esperada
<input type="checkbox"/>	21389	PTC - Grandes Empresas - Oferta observada
<input type="checkbox"/>	21390	PTC - MPME - Oferta esperada
<input type="checkbox"/>	21391	PTC - MPME - Oferta observada
<input type="checkbox"/>	21392	PTC - Consumo - Oferta esperada
<input type="checkbox"/>	21393	PTC - Consumo - Oferta observada
<input type="checkbox"/>	21394	PTC - Habitacional - Oferta esperada
<input type="checkbox"/>	21395	PTC - Habitacional - Oferta observada

Volumetria

SIDA 18.001.164 x 13

SGS 39 x 9

Abordagem Adotada

Framework Big Data

Arquitetura Delta (“Bronze”)

Proteção de dados pessoais

Tecnologias

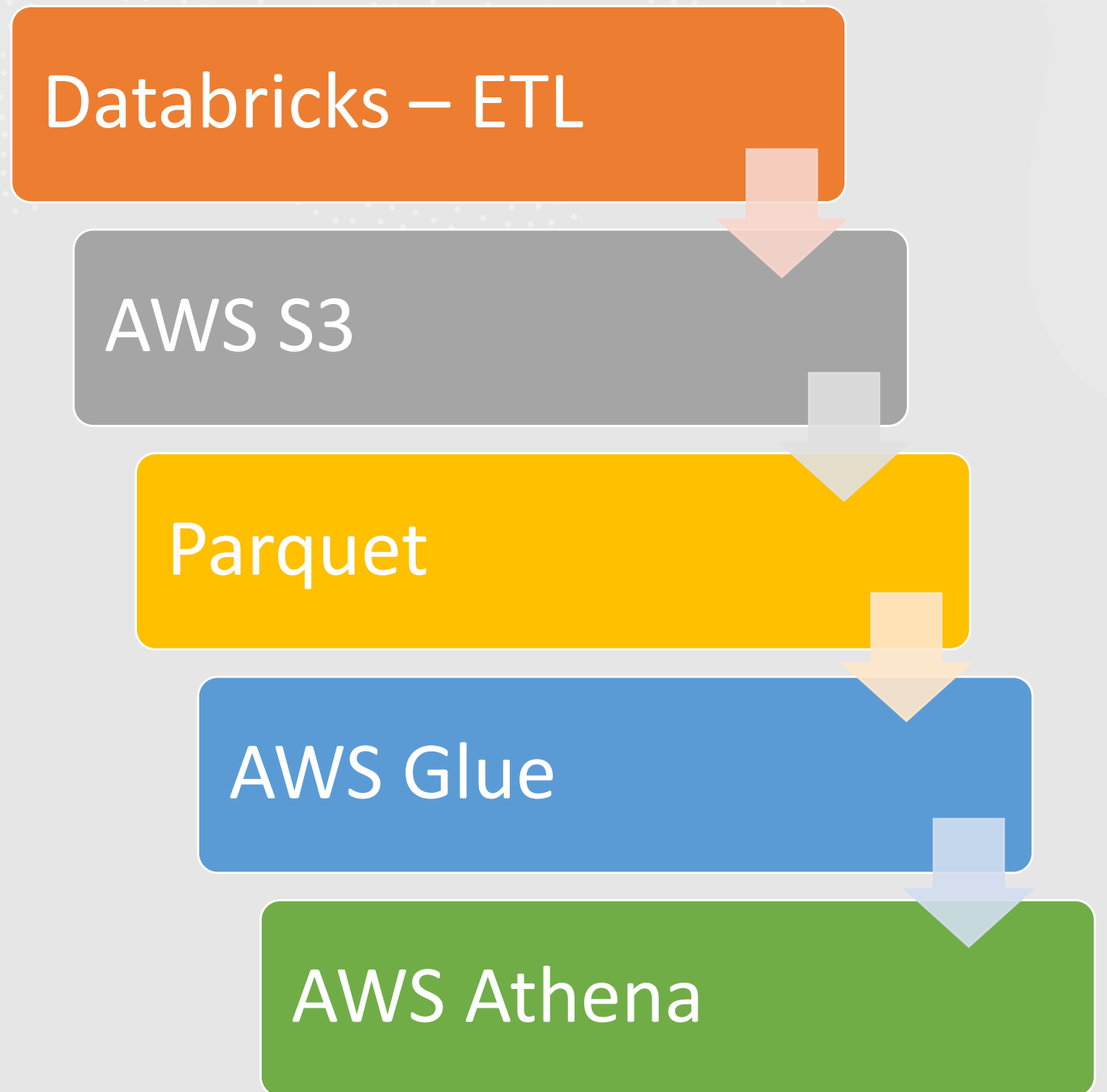
Databricks – ETL

AWS S3

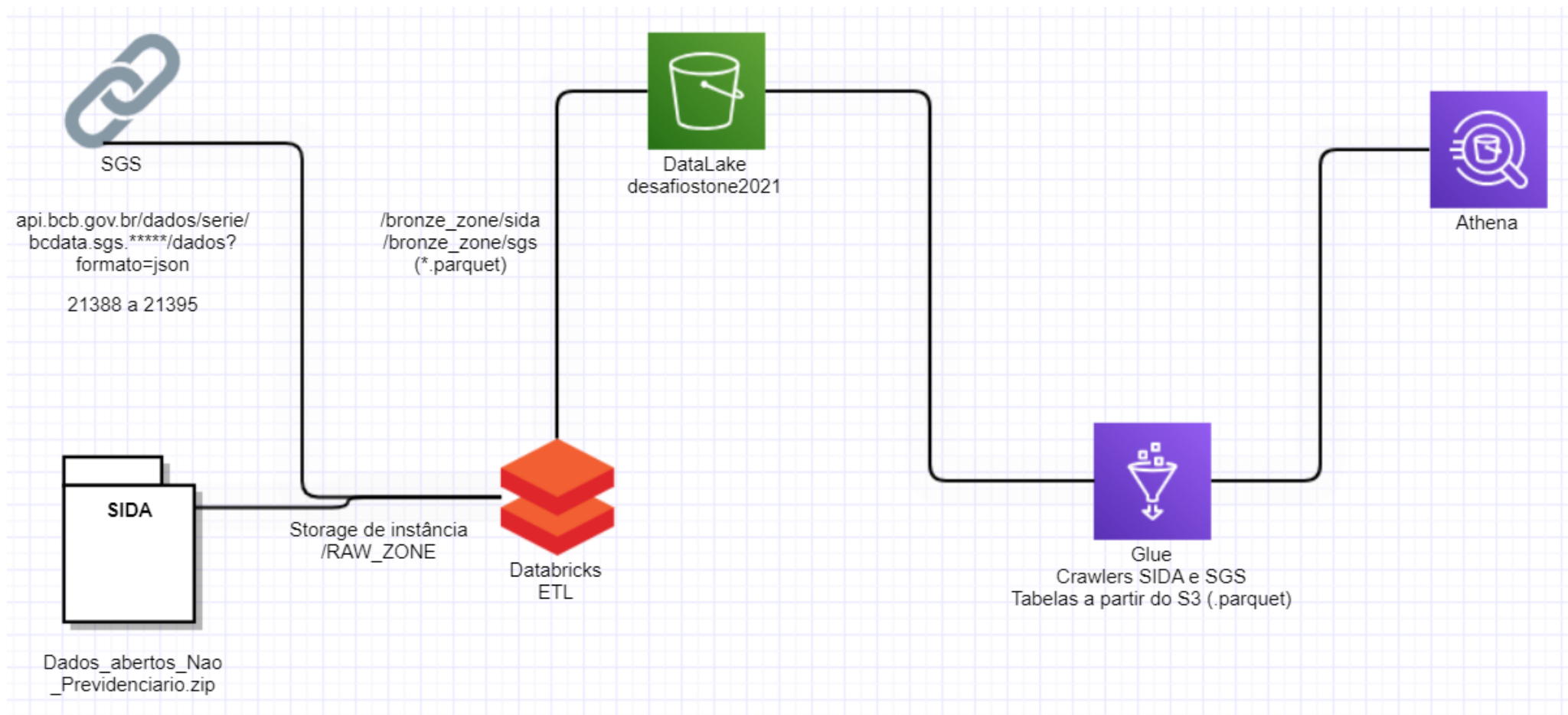
Parquet

AWS Glue

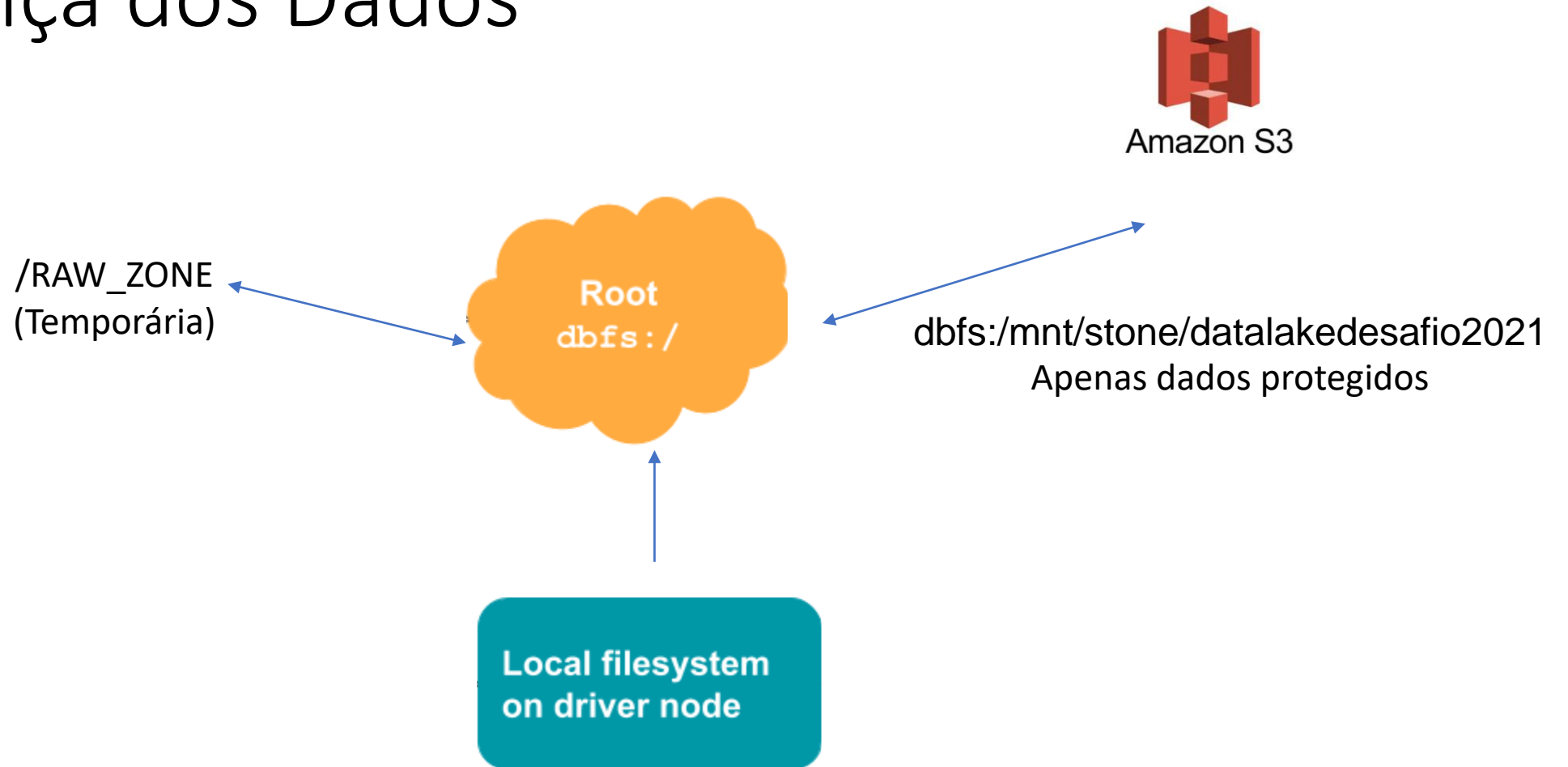
AWS Athena



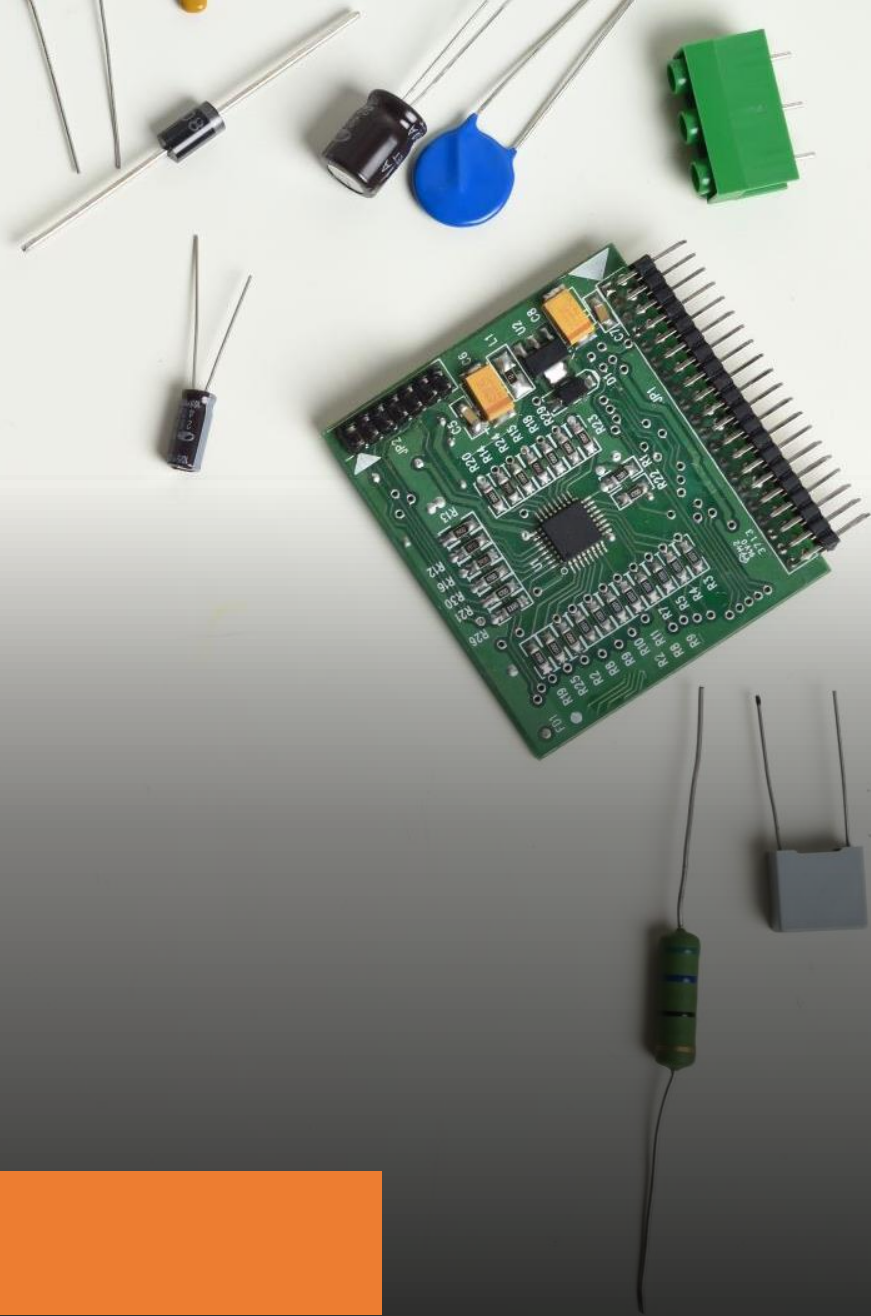
Solução Proposta



Segurança dos Dados



Setups



Permissões

Grupos (1)

Tags

Credenciais de segurança

▼ Permissions policies (1 política aplicada)

Adicionar permissões

	Nome da política ▼
Anexado a partir do grupo	
▶	 AmazonS3FullAccess
▶ Permissions boundary (not set)	

Usuário e
permissões

community.cloud.databricks.com/?o=8557447580499845#create/cluster

Create Cluster

New Cluster

[Cancel](#) [Create Cluster](#)

0 Workers: 0.0 GB Memory, 0 Cores, 0 DBU
1 Driver: 15.3 GB Memory, 2 Cores, 1 DBU ?

Cluster Name

DESAFIO

Databricks Runtime Version ?

Runtime: 8.0 (Scala 2.12, Spark 3.1.1) | v

Note Databricks Runtime 8.x uses Delta Lake as the default table format. [Learn more](#)

Instance

Free 15GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period. For [more configuration options](#), please [upgrade your Databricks subscription](#).

Instances [Spark](#)

Availability Zone ?

Configuração
do Cluster

Bucket


Amazon S3 > datalakedesafio2021 > bronze_zone/

bronze_zone/

Objetos

Propriedades

Objetos (0)

Os objetos são as entidades fundamentais armazenadas no Amazon S3. Você pode usar o [inventário](#) acessarem seus objetos, você precisará conceder permissões explicitamente a eles. [Saiba mais](#) 



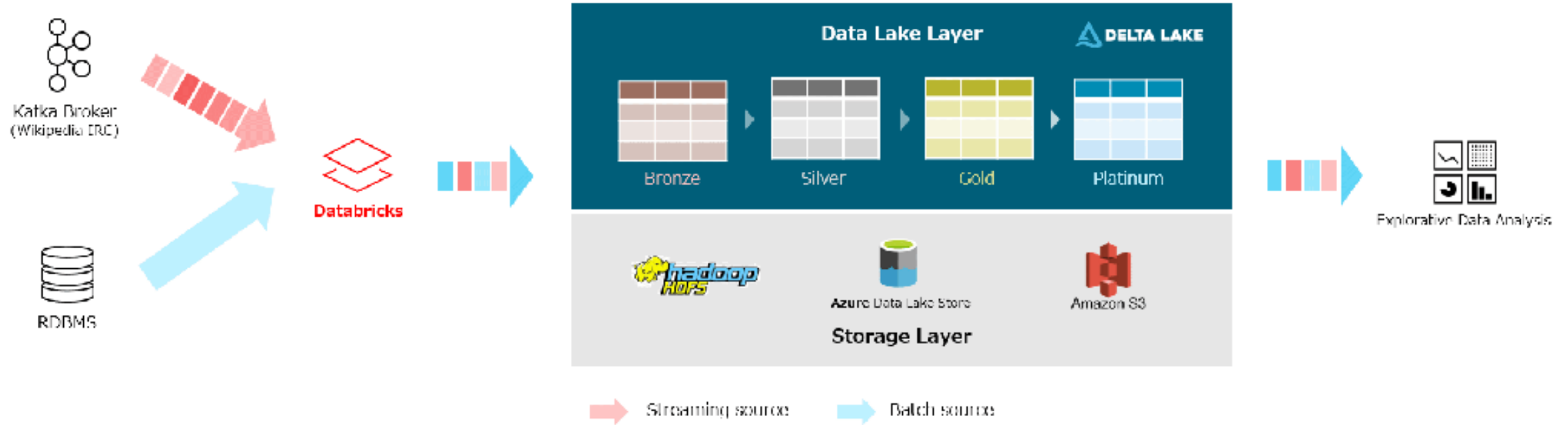
Excluir

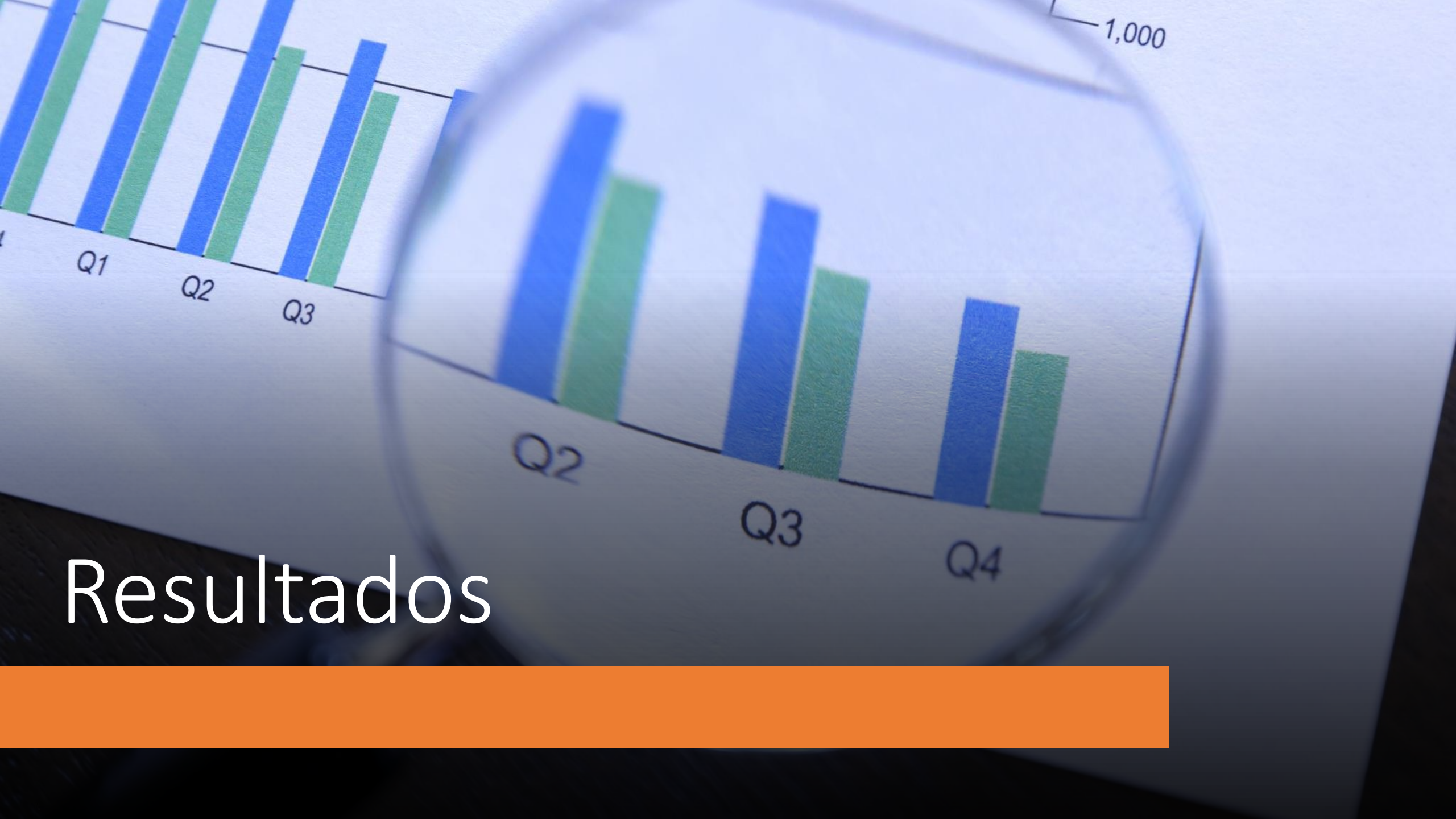
Ações ▼

Criar pasta

Carregar

Delta Architecture Workflow Demo














Resultados

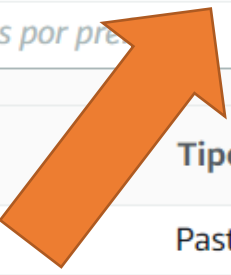
Escrita no Data Lake

Amazon S3 > datalakedesafio2021 > bronze_zone/

<input type="checkbox"/>	 _committed_1880068806606115596	-	28 Mar 2021 05:23:25 PM -03
<input type="checkbox"/>	 _started_1880068806606115596	-	28 Mar 2021 05:15:37 PM -03
<input type="checkbox"/>	 _SUCCESS	-	28 Mar 2021 05:23:27 PM -03
<input type="checkbox"/>	 part-00000-tid-1880068806606115596-e97dcc87-4e8b-4985-879b-35b60ad650a9-11-1-c000.snappy.parquet	parquet	28 Mar 2021 05:17:59 PM -03
<input type="checkbox"/>	 part-00001-tid-1880068806606115596-e97dcc87-4e8b-4985-879b-35b60ad650a9-12-1-c000.snappy.parquet	parquet	28 Mar 2021 05:17:55 PM -03
<input type="checkbox"/>	 part-00002-tid-1880068806606115596-e97dcc87-4e8b-4985-879b-35b60ad650a9-13-1-c000.snappy.parquet	parquet	28 Mar 2021 05:18:00 PM -03
<input type="checkbox"/>	 part-00003-tid-1880068806606115596-e97dcc87-4e8b-4985-879b-35b60ad650a9-14-1-c000.snappy.parquet	parquet	28 Mar 2021 05:17:57 PM -03

🔍 Localizar objetos por prefixo

<input type="checkbox"/>	Nome	Tipo	Últi
<input type="checkbox"/>	 sgs/	Pasta	-
<input type="checkbox"/>	 sida/	Pasta	-



Crawlers S3 (Glue)

Adicionar crawler

Executar crawler

Ação ▼

🔍 Filter by tags and attributes

<input type="checkbox"/>	Nome	Programação	Status	Logs	Último tempo de execução
<input type="checkbox"/>	sgs		Ready	Logs	47s
<input type="checkbox"/>	sida		Ready	Logs	49s

Tabelas a partir dos Dados

AWS Glue

Catálogo de dados

Bancos de dados

Tabelas

Conexões

Crawlers

Tabelas

Uma tabela é a definição de metadados que representa seus dados, incluindo o esquema. Uma tabela pode ser usada como origem

Adicionar tabelas ▼

Ação ▼

Filter by attributes or search by keyword

Salvar visualização

<input type="checkbox"/> Nome	Banco de dados	Local	Classificação
<input type="checkbox"/> sgs	sgs	s3://datalakedesafio2021/br...	parquet
<input type="checkbox"/> sida	sida	s3://datalakedesafio2021/br...	parquet

AWS Athena

Data source [Connect data source](#)

AwsDataCatalog

Database

sida

Filter tables and views...

▼ **Tables (1)** [Create table](#)

▼ sida

- id (bigint)
- cpf_cnpj (string)
- tipo_pessoa (string)
- tipo_devedor (string)
- nome_devedor (string)
- uf_unidade_responsavel (string)
- unidade_responsavel (string)
- numero_inscricao (string)
- tipo_situacao_inscricao (string)
- situacao_inscricao (string)
- receita_principal (string)
- data_inscricao (date)
- indicador_ajuizado (string)
- valor_consolidado (double)

New query 1 **New query 2** ✕ +

```
1 select * from sida limit 10;
```

[Run query](#) [Save as](#) [Create](#) (Run time: 1.33 seconds)

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

	id	cpf_cnpj
1	171798691840	0ae1a54a1c4a4fd1c4060f
2	171798691841	26c2095172716519ff36b5
3	171798691842	2c0eaab1e0b2160d7b6d4
4	171798691843	6fc303f7d821574ec4d59e
5	171798691844	5816cc0e9daa4093c6d18
6	171798691845	869935d5ef3d9815bc0d7f
7	171798691846	8850c341f5c3530181c

Data source [Connect data source](#)

AwsDataCatalog

Database

sgs

Filter tables and views...

▼ **Tables (1)** [Create table](#)

▼ sgs

- data (timestamp)
- sgs_21388 (double)
- sgs_21389 (double)
- sgs_21390 (double)
- sgs_21391 (double)
- sgs_21392 (double)
- sgs_21393 (double)
- sgs_21394 (double)
- sgs_21395 (double)

New query 1 **New query 2** ✕ +

```
1 select * from sgs limit 10;
```

[Run query](#) [Save as](#) [Create](#) (Run time: 1.33 seconds)

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

	data	sgs_21388	sgs_21389
1	2011-01-01 00:00:00.000	-0.05	-0.09
2	2011-04-01 00:00:00.000	0.05	0.2
3	2011-07-01 00:00:00.000	-0.77	-0.45

Objetivos

Sua tarefa implica inicialmente em:

- OK • Coletar todo o histórico disponível e armazenar ambas as bases no s3, respeitando as boas práticas de tipos de arquivos, particionamento, zonas de armazenamento comuns em um Datalake e anonimização dos dados exigidos pela LGPD.
- OK • Criar tabelas no Athena de forma que os cientistas de dados possam analisar o histórico e as correlações entre os dados de dívidas e os indicadores de crédito.
- OK • Criar uma chave única para consultar a base de dívidas e outra chave temporal para cruzamento com a base de indicadores.

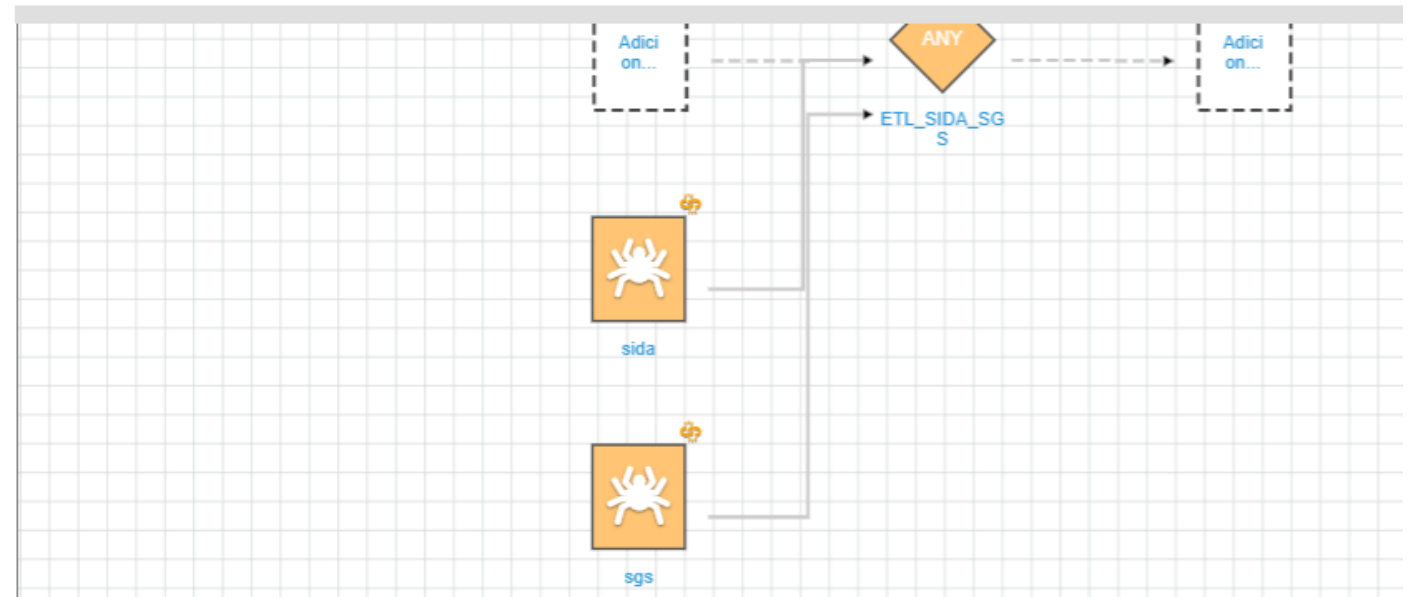
Melhorias futuras

- Adição do Delta Lake (ACID, TimeTravel)
- Job do ETL do Databricks
- Feature Engineering aprimorada (datas e valores)

Fluxos de trabalho (1)

Um fluxo de trabalho é uma orquestração usada para visualizar e gerenciar o relacionamento e a

Adicionar fluxo de trabalho	Ações ▾	↺	🔍 Filtrar fluxos de trabalho
	Nome ▾	Última execução ▾	Status da última execução
●	desafio2021	-	-



Demonstração do Pipeline

The screenshot displays the Stone_ETL (Python) interface. On the left is a dark sidebar with icons for Home, Workspace, Recents, and Data. The main area has a title bar 'Stone_ETL (Python)' and a toolbar with buttons for DESAFIO, File, Edit, View: Standard, Permissions, Stop Execution, and Clear. Below the toolbar, 'Cmd 1' contains the text 'DESAFIO STONE 2021 - ENGENHARIA DE DADOS' and 'Marlon Ferrari'. 'Cmd 2' contains the text '1. Configurações de Acesso ao S3'. There are expand/collapse icons on the left of each command box and add/edit icons on the right of the 'Cmd 2' box.

Stone_ETL (Python)

DESAFIO

File Edit View: Standard Permissions Stop Execution Clear

Cmd 1

DESAFIO STONE 2021 - ENGENHARIA DE DADOS

Marlon Ferrari

Cmd 2

1. Configurações de Acesso ao S3