



# 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

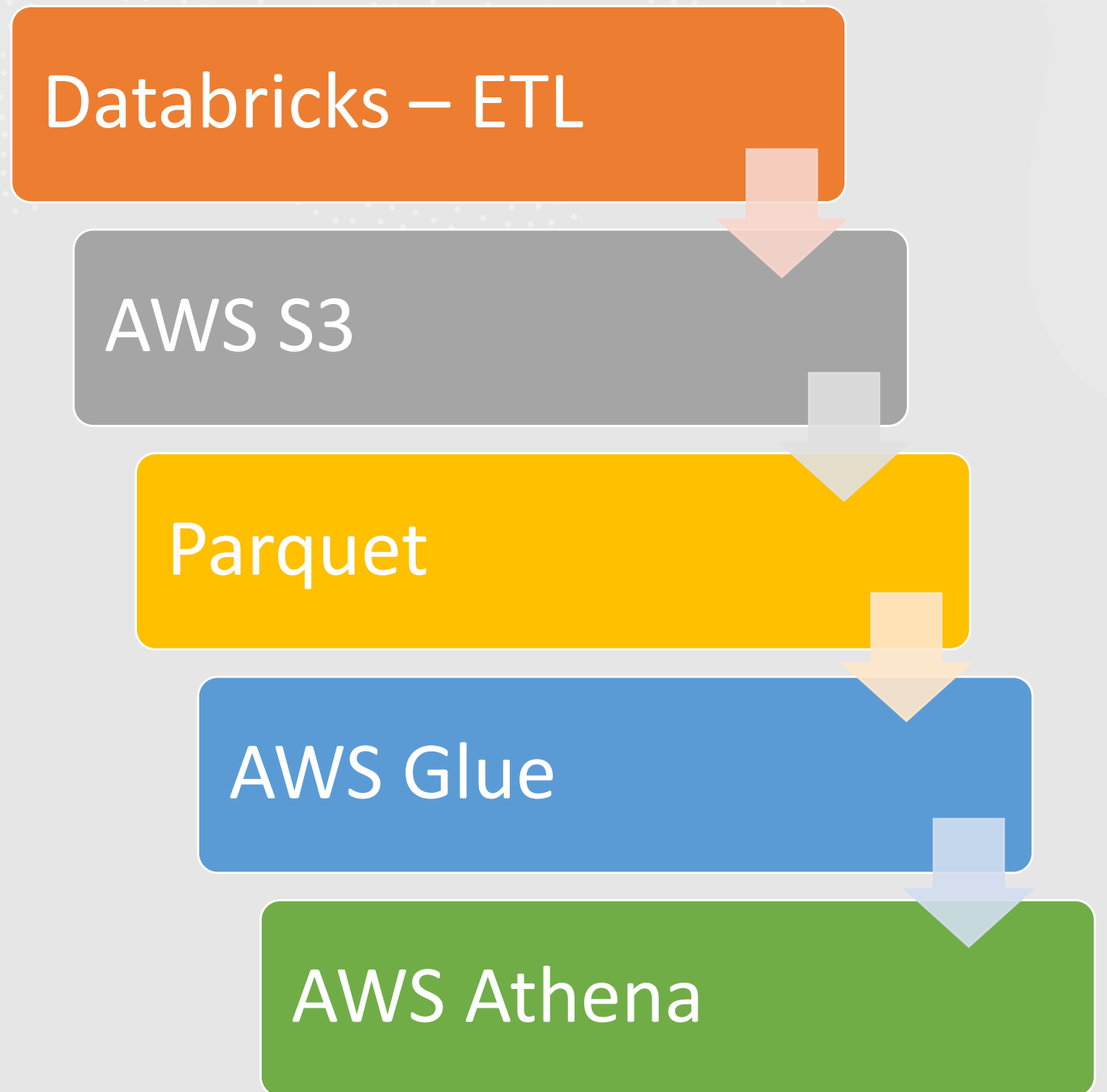
Databricks – ETL

AWS S3

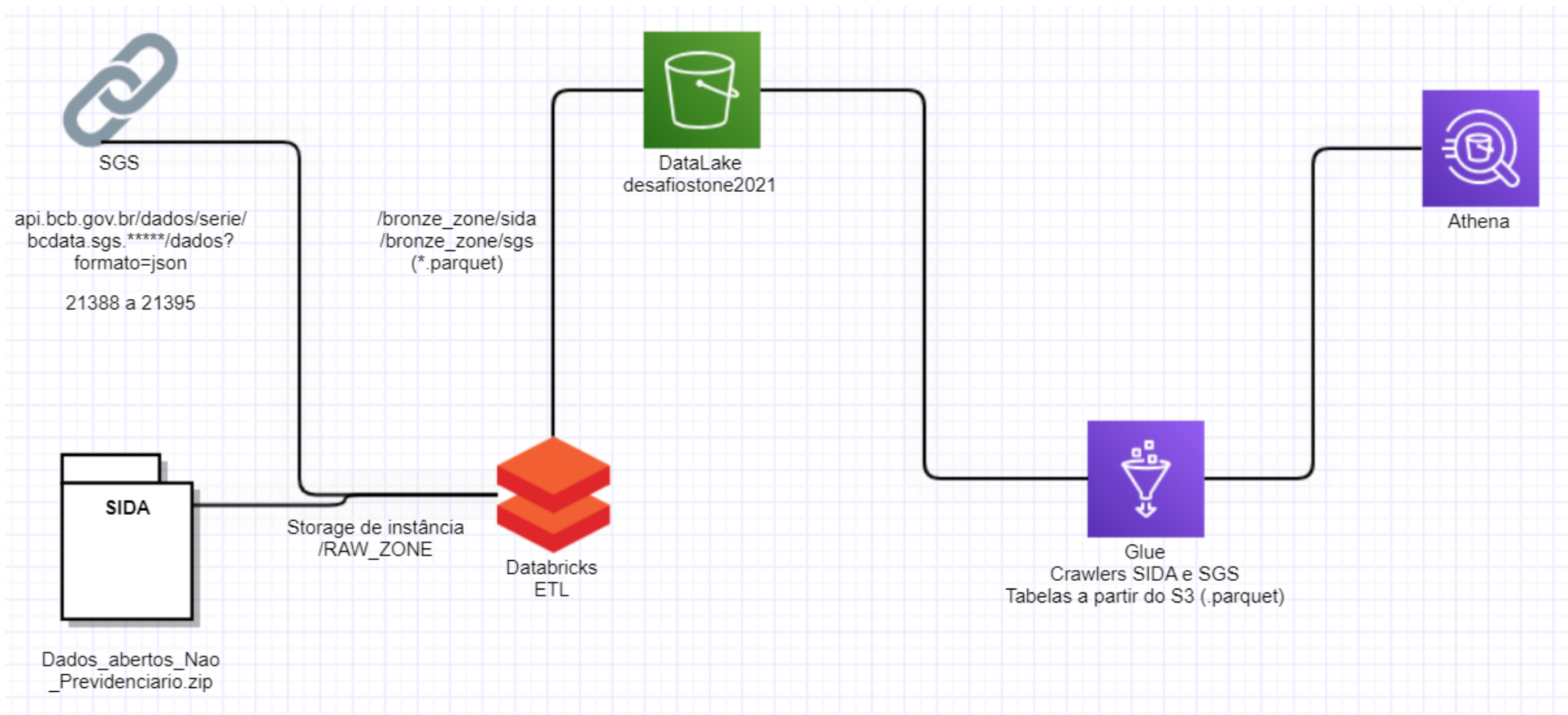
Parquet

AWS Glue

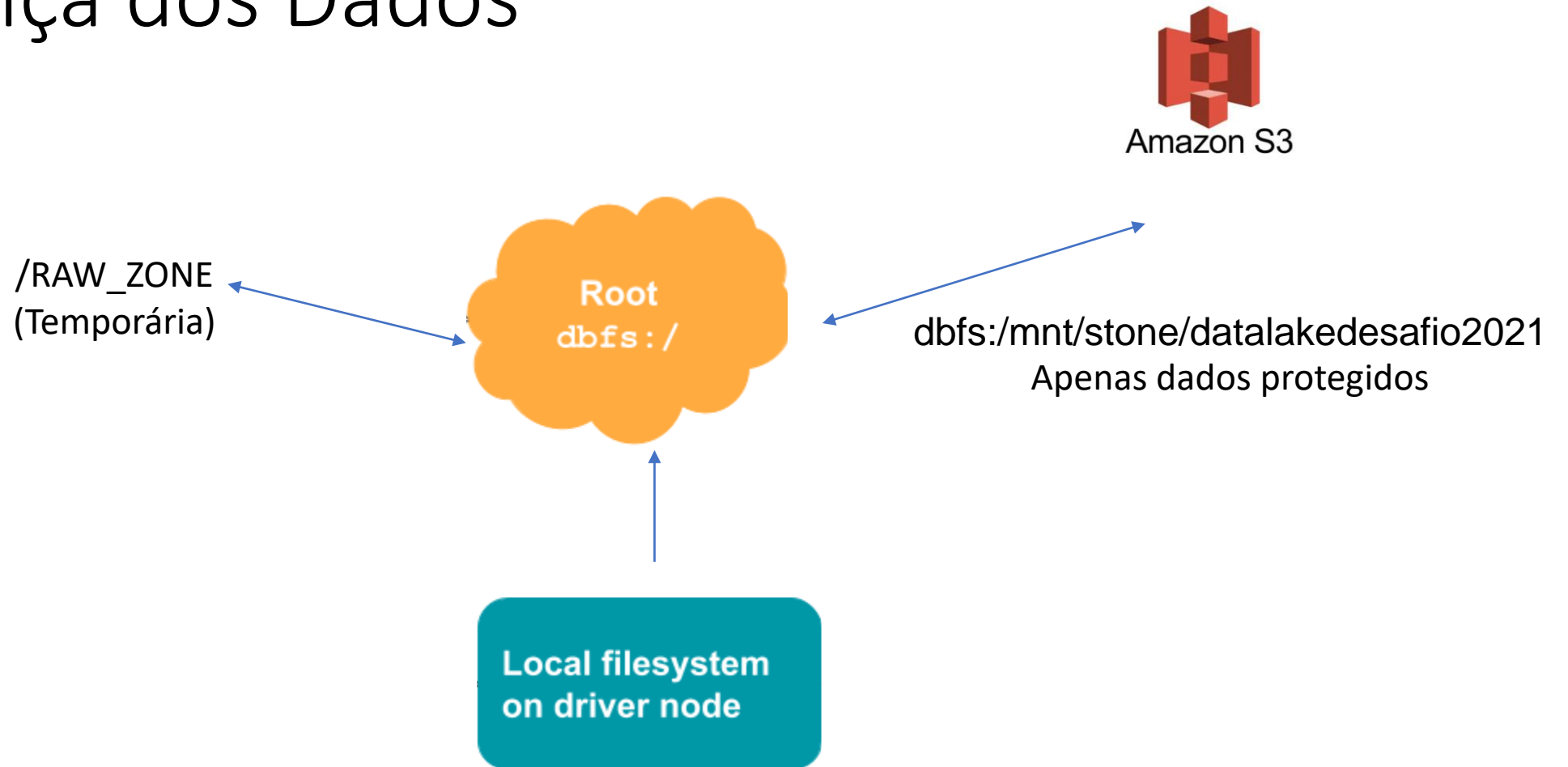
AWS Athena



# Solução Proposta



# Segurança dos Dados



## Características do Pipeline

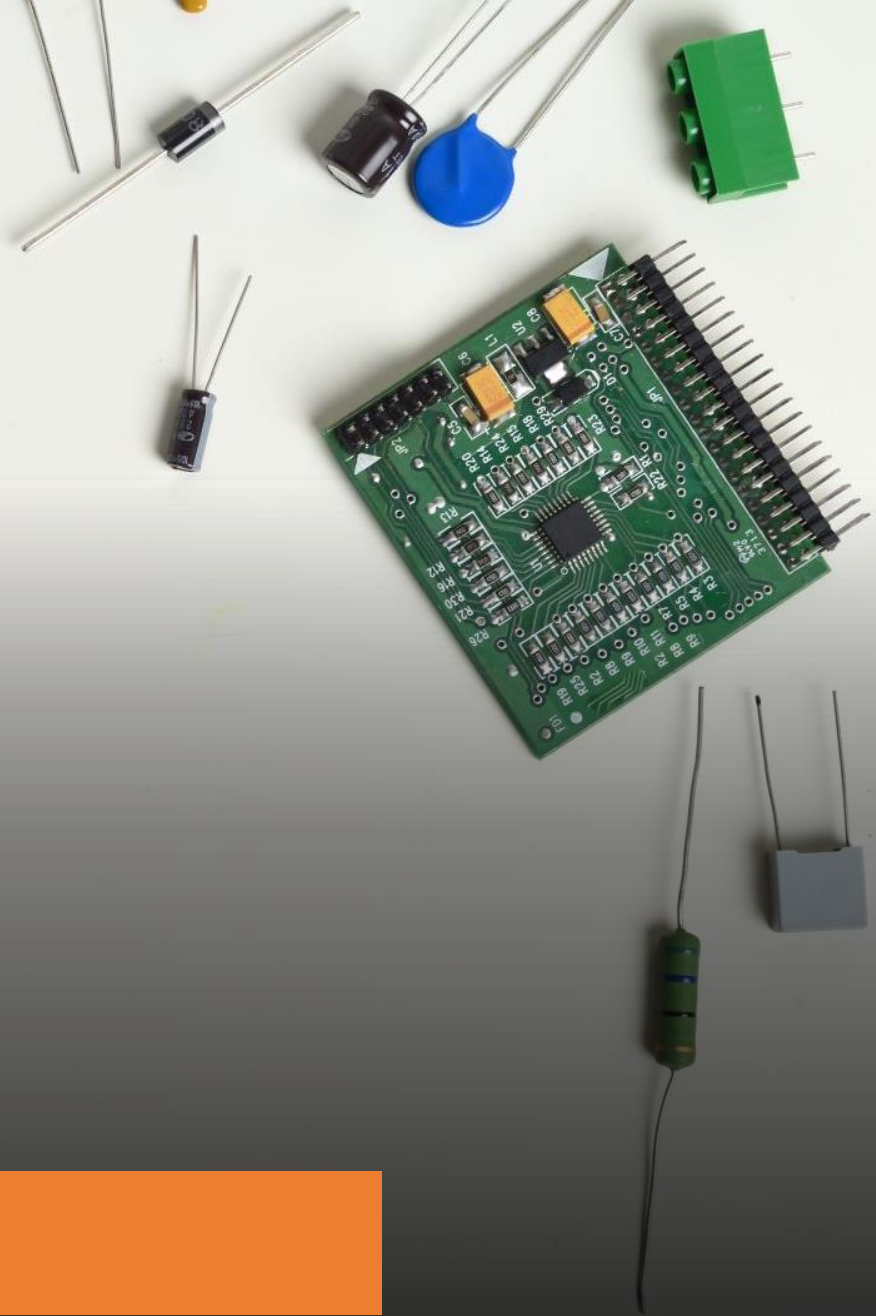
Totalmente automatizado com criação completa dos folders, arquivos e tabelas a cada execução

Crawlers + Tabelas Glue garantem disponibilidade

Consulta Athena Otimizada (parquets)



# Setups



# Usuário e permissões

**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)

# Configuração do Cluster

← → ↻ 🔒 community.cloud.databricks.com/?o=8557447580499845#create/cluster

## Create Cluster

### New Cluster

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

Cluster Name

Databricks Runtime Version ?  
 ▼

**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 ?

Bucket

## Amazon S3

### Buckets (1)

Os buckets são contêineres para dados armazenados no S3.

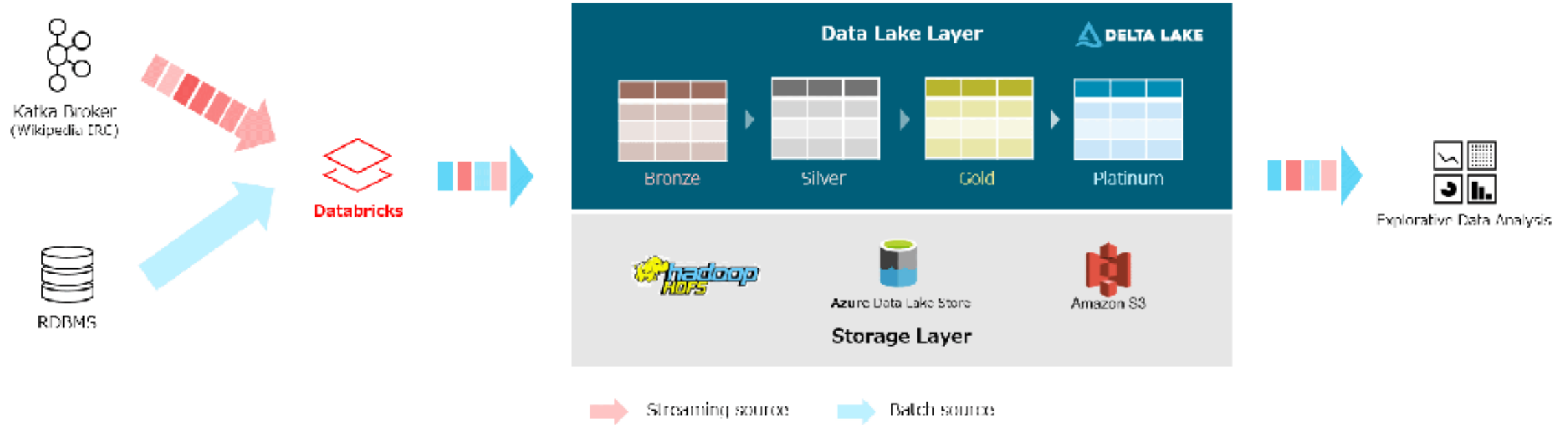
🔍 *Encontrar buckets por nome*

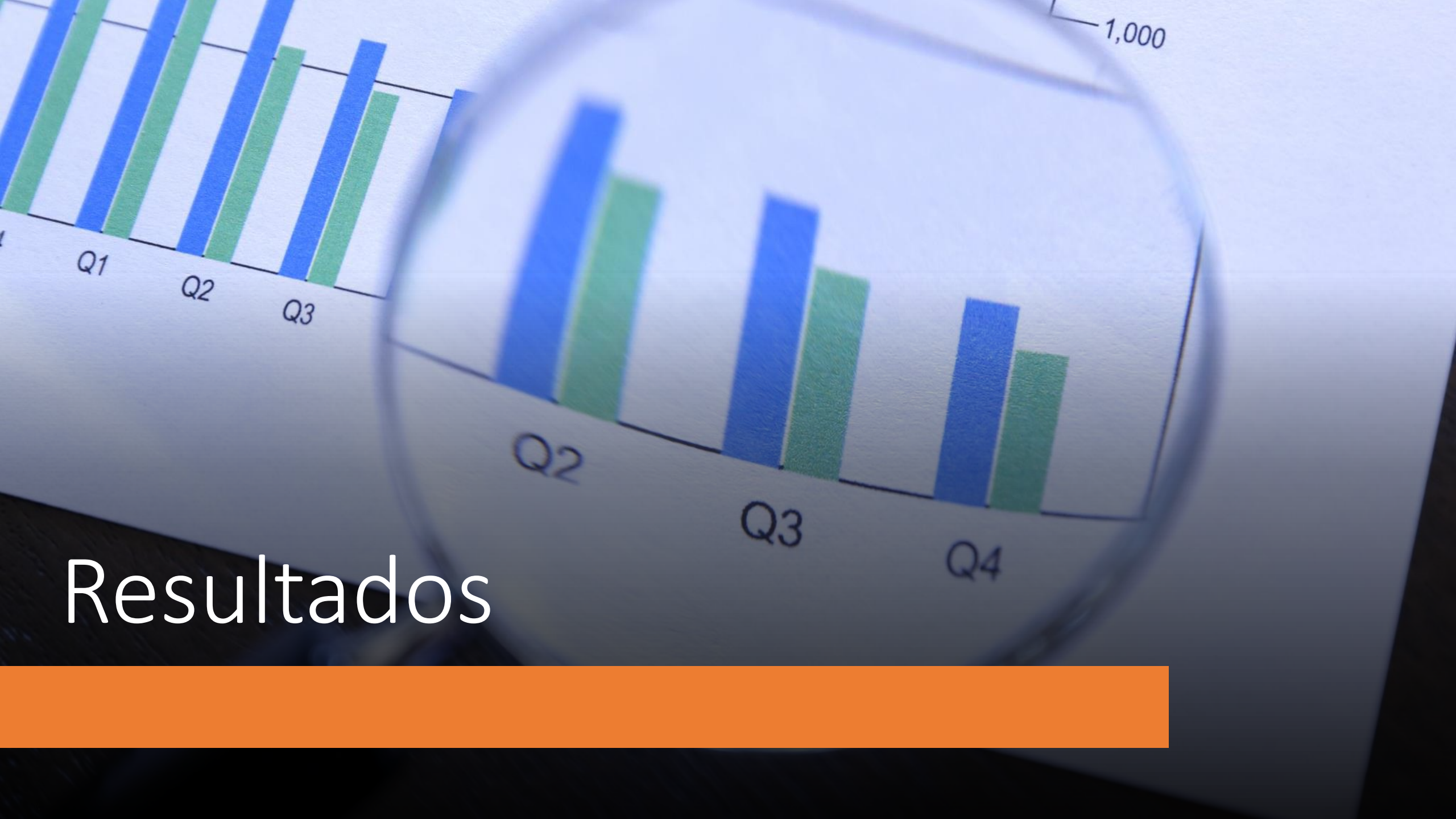
**Nome**



**datalakedesafio2021**

## Delta Architecture Workflow Demo














Resultados

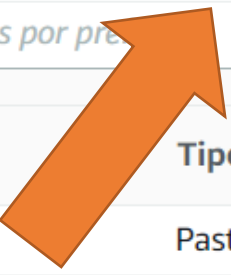
# Escrita no Data Lake

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	<a href="#">Logs</a>	47s
<input type="checkbox"/>	sida		Ready	<a href="#">Logs</a>	49s



# Tabelas a partir dos Dados (Glue)

## 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	desafio_bronze	→ s3://datalakedesafio2021/br... parquet	
<input type="checkbox"/> sida	desafio_bronze	→ s3://datalakedesafio2021/br... parquet	

# AWS Athena

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.

## Data source

AwsData

Connect data source

Filter tables and views

### Tables (2)

Create table

sgs

sida

### Views (0)

Create view

You have not created any views. To create a view, run a query and click "Create view from query"

## New query 1

```
1 SELECT sgs.*, round(avg(sida.valor_consolidado),2) as AVG_consolidada
2 FROM sgs
3 JOIN sida
4 ON sgs.q_dt_indicador=q_dt_inscricao where sgs.q_dt_indicador >= 20201
5 GROUP BY
6 q_dt_indicador,
7 SGS_21388,
8 SGS_21389,
9 SGS_21390,
10 SGS_21391,
11 SGS_21392,
12 SGS_21393,
13 SGS_21394,
14 SGS_21395
15 ORDER BY q_dt_indicador
```

Criar uma chave única para consultar a base de dívidas e outra chave temporal para cruzamento com a base de indicadores.

Run query

Save as

Create

(Run time: 2.27 seconds, Data scanned: 117.4 MB)

Format query

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

Athena engine version 1 [Release ver](#)

## Results

	q_dt_indicador	sgs_21388	sgs_21389	sgs_21390	sgs_21391	sgs_21392	sgs_21393	sgs_21394	sgs_21395	AVG_consolid:1395	AVG_consolid:
1	20201	-0.95	-0.14	-0.74	0.13	-0.89	-0.21	-0.5	0.17	141383.1	
2	20202	-0.18	-0.77	-0.13	-0.53	-0.06	-0.61	0.14	-0.29	47347.79	
3	20203	-0.05	-0.18	0.03	-0.17	0.44	0.06	0.14	0.29	576242.04	

# Chave temporal

DATA -> Ano + Conversão  
do mês em trimestre  
(quarter range)

Ex.: 20/6/2019 fica  
20192

# 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	-	-

