

CHALLENGE BANCO PAN

DATAQUALITY | AWS



BEATRIZ CARDOSO CUNHA

EVERSON DANIEL DA SILVA

GABRIEL SOLER BELMONTE

HARYEL COSTA ASSENÇÃO

WESLEY FERREIRA DE SOUZA

MBA EM ENGENHARIA DE SOFTWARE | 4aojo

O desafio Data Quality é destinado ao curso de Engenharia de Software e comprehende, utilizando como premissa obrigatória serviços AWS (quais ficam por conta do grupo) para criar mecanismos parametrizados de controle de qualidade dos dados. O Banco Pan fornecerá aos alunos uma base de dados modelo, contendo os problemas mais comuns observados na chegada dos dados do data lake, e o grupo deverá criar uma solução funcional para tratar os dados.

1. INGESTÃO DE DADOS

Primeiramente, foi criado o Armazenamento de Dados em S3 como repositório, conforme os passos a seguir.

Acessar a conta no console de gerenciamento da AWS

The screenshot shows the AWS Management Console homepage. At the top, there's a navigation bar with the AWS logo, a dropdown menu for 'Services', a search bar, and links for 'Help' and 'Home'. Below the navigation bar is the main content area. On the left, there's a sidebar titled 'Serviços da AWS' with sections for 'Serviços acessados recentemente' (listing AWS Glue DataBrew, S3, Amazon FSx, Key Management Service, AWS Glue, Cloud9, Database Migration Service, RDS, Lambda, Data Pipeline, Billing, and AWS Organizations) and a 'Todos os serviços' link. The main content area has several sections: 'Criar uma solução' (with links for 'Executar uma máquina virtual', 'Criar uma aplicação Web', 'Criar usando servidores virtuais', 'Registrar um domínio', 'Conectar um dispositivo IoT', 'Começar a migrar para a AWS', 'Iniciar um projeto de desenvolvimento', and 'Implantar um microserviço sem servidor'); 'Explorar a AWS' (with links for 'Executar contêineres sem servidor com o AWS Fargate', 'AWS Marketplace', 'Backup e restauração com Amazon S3', and 'Amazon Redshift'); and a sidebar on the right titled 'Mantenha-se conectado aos seus recursos da AWS em qualquer lugar' with information about the AWS mobile app. At the bottom, there are links for 'Feedback', 'English (US)', and '© 2008 - 2021, Amazon Web Services, Inc. ou suas afiliadas. All rights reserved.' followed by 'Privacy Policy' and 'Terms'.

Clicar no canto superior esquerdo em Services e em Armazenamento em S3

Criamos um Bucket para armazenar o dado no S3: Criar Bucket

The screenshot shows the AWS S3 console interface. On the left, a sidebar navigation includes 'Amazon S3' (selected), 'Buckets', 'Pontos de acesso', 'Pontos de acesso Lambda de objeto', 'Operações em lotes', 'Analizador de acesso para S3', 'Configurações de bloqueio do acesso público desta conta', 'Storage Lems' (expanded), 'Painéis', 'Configurações do AWS Organizations', and 'Recurso em destaque'. The main content area has a banner with 'Estamos melhorando o console do S3 para torná-lo mais rápido e fácil de usar. Se você tem comentários sobre a experiência atualizada, escolha Forneca comentários.' and a link 'Saiba mais...'. Below this, a section titled 'Amazon S3' contains a 'Snapshot da conta' button with a note about Storage Lens, a 'Visualizar painel do Storage Lens' button, and a 'Buckets (1) Informações' table. The table lists one bucket: 'Frapinacopan' (Nome), 'Leste dos EUA (Norte da Virgínia) us-east-1' (Região da AWS), 'Bucket e objetos não públicos' (Acesso), and '8 Aug 2021 10:20:15 AM -03' (Data de criação). At the bottom, there are links for 'AWS Marketplace para S3' and 'Feedback English (US)'.

CRIAR S3 BUCKET

Na tela de criação faremos a Configuração Geral dando nome ao Bucket seguindo o padrão exigido de letras minúsculas sem espaços. Deixar a região da AWS Leste dos EUA (norte da Virgínia) us-east-1. Deixar bloqueado todo o acesso público.

Nome do S3: grupoc-dq-pa

Criar bucket Info

Os buckets são contêineres para dados armazenados no S3. [Saiba mais](#)

Configuração geral

Nome do bucket
grupoc-dq-parl

O nome do bucket deve ser exclusivo e não deve conter espaços ou letras maiúsculas. [Consulte as regras de nomenclatura de bucket](#)

Região da AWS
Leste dos EUA (Norte da Virgínia) us-east-1

Copiar configurações do bucket existente - *opcional*
Somente as configurações de bucket na configuração a seguir são copiadas.

[Escolher bucket](#)

Deixaremos desativada a opção de versionamento e para este desafio não faremos a criptografia dos dados.

Clicar em Criar Bucket

Amazon S3 X

Amazon S3

Buckets

- Pontos de acesso
- Pontos de acesso Lambda de objeto
- Operações em lotes
- Analisador de acesso para S3

Configurações de bloqueio do acesso público desta conta

Storage Lens

- Painéis
- Configurações do AWS Organizations

Recurso em destaque 1

AWS Marketplace para S3

Snapshot da conta

O Storage Lens fornece visibilidade sobre o uso e as tendências de atividades. [Saiba mais](#)

[Visualizar painel do Storage Lens](#)

Buckets (5) Info

Os buckets são contêineres para dados armazenados no S3. [Saiba mais](#)

<input type="button" value="C"/>	<input type="button" value="Copiar ARN"/>	<input type="button" value="Vazio"/>	<input type="button" value="Excluir"/>	<input type="button" value="Criar bucket"/>
<input type="text" value="Encontrar buckets por nome"/> < 1 > ②				
Nome	Região da AWS	Acesso	Data de criação	
<input type="radio"/> aws-athena-query-results-us-east-1-700584023387	Leste dos EUA (Norte da Virgínia) us-east-1	Bucket e objetos não públicos	17 Aug 2021 09:13:41 AM -03	
<input type="radio"/> desafio-pan-g	América do Sul (São Paulo) sa-east-1	Bucket e objetos não públicos	12 Aug 2021 08:22:42 PM -03	
<input checked="" type="radio"/> grupoc-dq-parl	Leste dos EUA (Norte da Virgínia) us-east-1	Os objetos podem ser públicos	16 Aug 2021 12:17:29 PM -03	
<input type="radio"/> resultadoathenas	Leste dos EUA (Norte da Virgínia) us-east-1	Bucket e objetos não públicos	16 Aug 2021 08:28:31 PM -03	
<input type="radio"/> teste-haryel	Leste dos EUA (Norte da Virgínia) us-east-1	Os objetos podem ser públicos	16 Aug 2021 02:13:38 PM -03	

INGESTÃO DE DADOS NO S3

Clicamos em Visualizar Detalhes para fazer a ingestão do arquivo no S3

Adicionamos o arquivo enviado previamente pelo Banco Pan `base_exemplo_score.csv` para upload, clicando em Carregar:

The screenshot shows the 'Objetos' tab selected in the navigation bar. Below it, a message states 'Os objetos são as entidades fundamentais armazenadas no Amazon S3. Você pode usar o inventário do Amazon S3 para obter uma lista de todos os objetos em seu bucket. Para outras pessoas acessarem seus objetos, você precisará conceder permissões explicitamente a eles. Saiba mais'. A toolbar below includes buttons for Copy URI, Copy URL, Download, Open, Delete, Actions, and Create folder. A search bar labeled 'Localizar objetos por prefixo' is present. The main table header includes columns for Nome, Tipo, Última modificação, Tamanho, and Classe de armazenamento. A message 'Nenhum objeto' indicates no objects are found.

Em seguida em Adicionar Arquivo.

This screenshot shows the first step of the upload wizard. It features a large dashed blue box for dragging files. Below it, a table lists one file: 'Arquivos e pastas (1 Total, 310.7 KB)' with a 'base_exemplo_score.csv' entry. Buttons for Remove, Add files, and Add folder are available. The 'Destino' section shows the target as 's3://grupoc-dq-pan'. The 'Permissões' and 'Propriedades' sections are partially visible at the bottom.

PERMITIR USUÁRIOS

Para permitir que outros usuários acessem o S3 é preciso estar muito atento a seguir as Políticas do Bucket S3 ou de IAM para acesso. Em Lista de Controle de Acesso vamos escolher entre ACLs predefinidas e depois em Privado. O controle de acesso deve ser feito pelo IAM e Políticas do Bucket S3.

Finalmente clicar em Carregar

The screenshot shows the 'Listagem de controle de acesso (ACL)' (ACL Control List) configuration screen. It includes a note from AWS recommending the use of bucket policies or IAM policies instead of ACLs. Below this, there are two options: 'Escolher entre ACLs predefinidas' (Select predefined ACLs) and 'Especificar permissões individuais da ACL' (Specify individual ACL permissions). Under 'Predefinidas' (Predefined), 'Privado (recomendado)' (Private (recommended)) is selected. At the bottom right, there are 'Cancelar' (Cancel) and 'Carregar' (Upload) buttons.

A tela indicando que o upload foi bem-sucedido será assim:

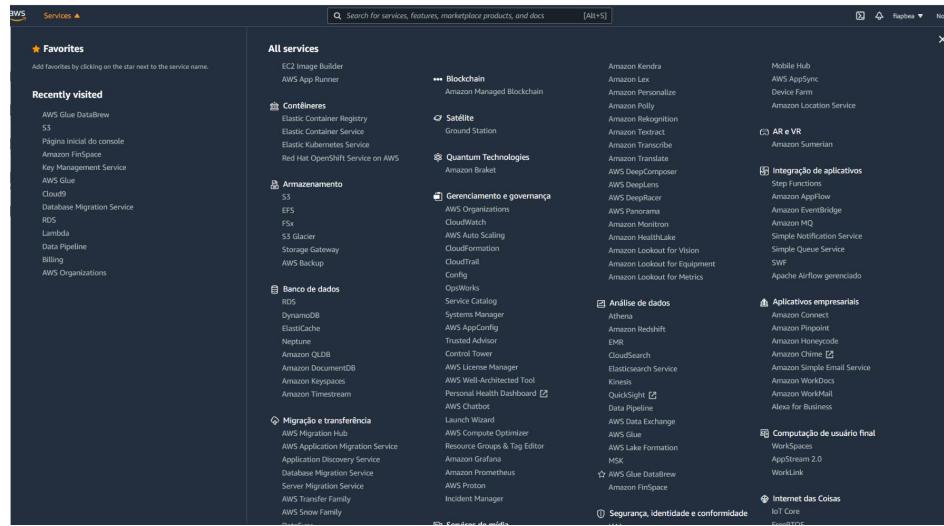
The screenshot shows the 'Upload: status' (Upload status) page. A green header bar indicates a successful upload: 'Upload bem-sucedido' (Upload successful). The main area displays the upload summary: 'Destino' (Destination) is 's3://grupoc-dq-pan', 'Status' (Status) is 'Bem-sucedida' (Successful), and 'Com falha' (With error) is '0 arquivos, 0 B (0%)' (0 files, 0 B (0%)). Below this, there are tabs for 'Arquivos e pastas' (Files and folders) and 'Configuração' (Configuration). The 'Arquivos e pastas' tab is active, showing a table with one file: 'base_exemplo_score.csv'. The file details are: Nome (Name) 'base_exemplo_score.csv', Pasta (Folder) '-', Tipo (Type) 'application/vnd.ms-excel', Tamanho (Size) '310.7 KB', Status (Status) 'Bem-sucedida' (Successful), and Erro (Error) '-'. There are navigation arrows at the bottom of the table.

Clicar em Fechar.

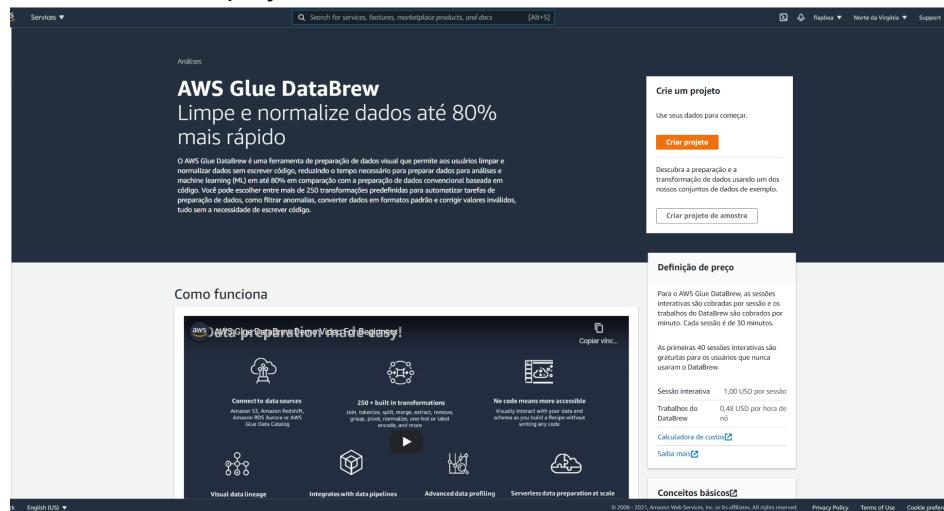
2. NORMALIZAÇÃO DOS DADOS COM O AWS GLUE DATABREW

Escolhemos o Glue DataBrew para o desafio por sua facilidade de preparação, normalização e análise primária dos dados, além do baixo custo em uma única ferramenta.

Clicar no canto superior esquerdo em Services e em Análise de Dados clicar em AWS Glue DataBrew.



Criar um projeto.



Em Detalhes do Projeto nomeamos de dq-grupoc e será automaticamente criada uma receita para as etapas de limpeza de dados vinculadas a este projeto.

Criar projeto

Detalhes do projeto

Nome do projeto

O nome do projeto deve conter de 1 a 255 caracteres. Os caracteres válidos são alfanuméricos (A-Z, a-z, 0-9), hífen (-), ponto (.) e espaço.

Detalhes da receita

As etapas de limpeza de dados no DataBrew são armazenadas como uma receita. Uma receita é conectada a um projeto por padrão. Uma receita existente sem projeto associado também pode ser aplicada a um projeto.

Receita anexada	Nome da receita
<input style="width: 150px; height: 25px;" type="button" value="Criar nova receita"/> ▾	<input type="text" value="dq-grupoc -recipe"/>
O nome da receita deve conter de 1 a 255 caracteres. Os caracteres válidos são alfanuméricos (A-Z, a-z, 0-9), hífen (-), ponto (.) e espaço.	

Importar etapas da receita
 Importe etapas de receita de uma receita existente para o seu projeto. A receita existente escolhida não será editada.

Em Selecionar um conjunto de dados marcamos Novo Conjunto de Dados e também lhe damos um nome sugestivo do projeto. Agora fazemos a conexão com o novo conjunto de dados, clicando em Upload de arquivo.

Criar projeto

Detalhes do projeto

Nome do projeto

O nome do projeto deve conter de 1 a 255 caracteres. Os caracteres válidos são alfanuméricos (A-Z, a-z e 0-9), hífens (-), pontos (.) e espaços.

Detalhes da descrição

As etapas de limpeza de dados no DataStudio são armazenadas como uma receta. Uma receta é conectada a um projeto por padrão. Uma receta existente sem projeto associado também pode ser aplicada a um projeto.

Receta anexada

Nome da receta

Criar nova receta

O nome da receta deve conter de 1 a 255 caracteres. Os caracteres válidos são alfanuméricos (A-Z, a-z, 0-9), hífens (-), pontos (.) e espaços.

Importar etapa de receta

Importar etapa de receta de uma receta existente para o seu projeto. A receta existente escolhida não será editada.

Selecionar um conjunto de dados

Selecione o conjunto de dados no qual deseja trabalhar

Meus conjuntos de dados

Meus conjuntos de dados importados

Arquivos de amostra

Escolher arquivos de exemplo para o conjunto de dados

Novo conjunto de dados

Importar novo conjunto de dados

Detalhes do novo conjunto de dados

Nome do conjunto de dados

O nome do conjunto de dados deve conter de 1 a 255 caracteres. Os caracteres válidos são alfanuméricos (A-Z, a-z, 0-9), hífens (-), pontos (.) e espaços.

Conectar-se ao novo conjunto de dados

Arquivo carregado com sucesso!

Insira sua origem: S3
Para selecionar uma pasta, todos os arquivos nella precisam conter o mesmo tipo de arquivo. Se houver respostas diferentes, elas serão misturadas.

s3://grupodc-project-base/exemplo_score.csv

O formato é: s3://bucket/prefixo

base_exemplo_score.csv está selecionado

53 Buckets > grupo-dg-pyan

Pesquisar objetos do S3 por ...

Definir parâmetros de conjunto de dados dinâmico

O que posso Escolher com os parâmetros?

Exemplo de Regra que pode ser adicionada ao caminho:

- Selecionar arquivos somente na pasta pai
- Selecionar arquivos que terminam com .csv somente na pasta pai

Escolha arquivos filtrados

Especificar o número de arquivos a serem incluídos

Inserimos a origem do S3 que foi recém criado; grupoc-dq-pan e o selecionamos o arquivo csv.

Em Configurações Adicionais selecionamos o tipo de arquivo csv e o delimitador de nossa tabela é o ponto-vírgula (;) e a primeira linha é cabeçalho.

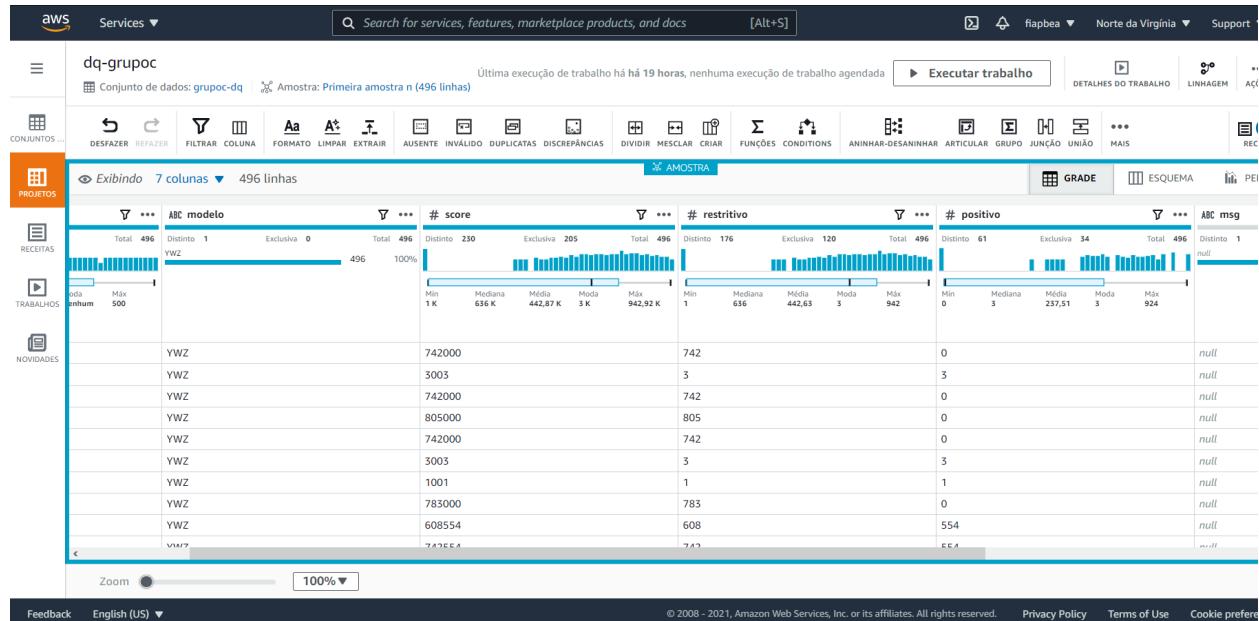
Em Permissões, se ainda não foi criada uma, devemos criar nova permissão IAM e dar um novo sufixo. Ao clicar em Criar Projeto será criada a nova função. Chamamos de project.

Tela do Projeto sendo criado.

USANDO O DATABREW

ANÁLISE DO DADO BRUTO

A tela do DataBrew nos fornece uma série de informações da qualidade do dado antes do tratamento. Sem a normalização teríamos métricas sem sentido e colunas sem acurácia.



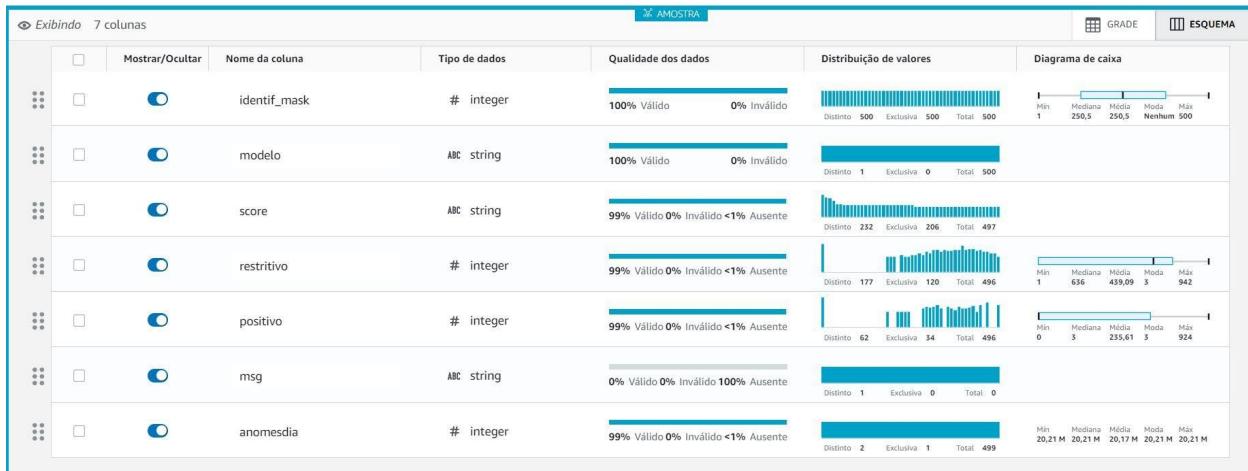
Na Coluna identif_mask não tem sentido adotar métricas de mín, máx, mediana, etc já que estamos adotando um id para cada cliente.

Na coluna score temos 3 nulas, ou seja, 1%.

Nas colunas restritivo e positivo temos 496 registros preenchidos, com perda em 4, correspondendo a 1%.

Na coluna anomesdia não temos o formato date e as métricas não fazem sentido.

Na aba ESQUEMA podemos ver um relatório da Qualidade dos Dados sem tratamento:



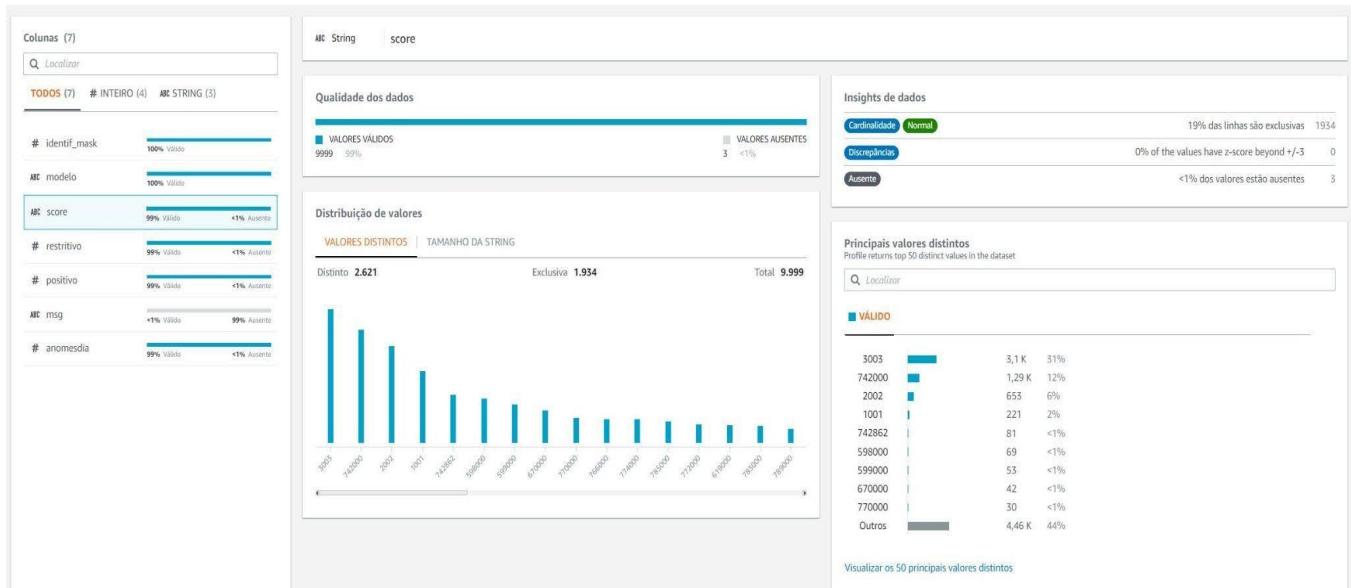
Em Visão Geral do Perfil de Dados temos o relatório bruto das colunas:

Visão geral do perfil de dados Estatísticas de coluna

Nome da coluna	identif_mask	modelo	score	restrictivo	positivo	msg	anomesdia
Tipo da coluna	# Inteiro	ABC String	ABC String	# Inteiro	# Inteiro	ABC String	# Inteiro
Qualidade dos dados	100% Válido 0% Inválido	100% Válido 0% Inválido	99% Válido 0% Inválido <1% Ausente	99% Válido 0% Inválido <1% Ausente	99% Válido 0% Inválido <1% Ausente	<1% Válido 0% Inválido 99% Ausente	99% Válido 0% Inválido <1% Ausente
Distribuição	Distincto 5140 Exclusiva 206 Total 10.002	Distincto 2 Exclusiva 0 Total 10.002	Distincto 2421 Exclusiva 1554 Total 3.899	Distincto 535 Exclusiva 29 Total 5.299	Distincto 307 Exclusiva 141 Total 3.982	Distincto 8 Exclusiva 5 Total 16	Distincto 3 Exclusiva 1 Total 10.001
Diagrama de caixa	Min 1 Mediana 250,5 Média 250,5 Moda 1 Máx 5145	Min 1 Mediana 250,5 Média 250,5 Moda 1 Máx 5145	Min 1 Mediana 496,25 Média 496,25 Moda 3 Máx 954	Min 0 Mediana 3 Média 231,08 Moda 3 Máx 980	Min 0 Mediana 3 Média 231,08 Moda 3 Máx 980	Min 20,21 M Mediana 20,21 M Média 20,21 M Moda 20,21 M Máx 20,21 M	Min 20,21 M Mediana 20,21 M Média 20,21 M Moda 20,21 M Máx 20,21 M
Total válido	10002 (100%)	10002 (100%)	9999 (100%)	9998 (100%)	9993 (100%)	16 (0%)	10001 (100%)
Total ausente	0 (0%)	0 (0%)	3 (0%)	4 (0%)	9 (0%)	9986 (100%)	1 (0%)
Valores distintos	5145 (51%)	2 (0%)	2621 (26%)	535 (5%)	367 (4%)	9 (0%)	3 (0%)
Valores exclusivos	288 (3%)	0 (0%)	1934 (19%)	89 (1%)	141 (1%)	5 (0%)	1 (0%)
Tamanho mínimo da string	-	3	3	-	-	1	-
Tamanho máximo da string	-	3	6	-	-	8	-
Min	1	-	-	1	0	-	20210331
Máx	5145	-	-	954	980	-	20210450
Mediana	2502,5	-	-	626	3	-	20210331
Média	2504,996400719856	-	-	436,23114622924584	231,07665365756029	-	20210379,079692032
Moda	1	-	-	3	3	-	20210331
Desvio padrão	1448,2598759615644	-	-	360,10843687422575	352,30912889985484	-	49,481897196106694
Discrepâncias	0 valores	-	-	0 valores	0 valores	-	0 valores
Distorção	0,0081568365145588	-	-	-0,2997702734462556	0,9579769770536442	-	0,05741742438178736
Soma	25054974	-	-	4361439	2309149	-	202124001176
Primeiro quartil (25%)	1252	-	-	3	0	-	20210331
Segundo quartil (50%)	2502,5	-	-	626	3	-	20210331
Terceiro quartil (75%)	3754	-	-	742	608	-	20210450
Intervalo interquartil	2502	-	-	739	608	-	99
Variância	2097456,668520206	-	-	129678,08630799896	124121,72230617453	-	2448,4581501260714
Curtose	-1,18915209224926	-	-	-1,7801556074530663	-0,9460225669438448	-	-1,9967026116364823

Em Estatísticas de coluna temos as seguintes métricas, antes do tratamento:

Coluna Score:



NORMALIZANDO OS DADOS

RECEITAS NO DATABREW

A primeira coluna para criar uma receita e comparar o tratamento do dado é a anomesdia-converter de String para Date.

1) Criar uma receita clicando em Adicionar etapa

anomedia

GRADE ESQUEMA PERFIL

Exclusiva Total 499

Distincto 2

Mín 20,21 M Mediana 20,21 M Média 20,17 M Moda 20,21 M Máx 20,21 M

20210331

20210331

20210331

20210331

20210331

20210331

20210331

20210331

20210331

20210331

20210331

20210331

20210331

20210331

20210331

20210331

Receita (0)

GrupoC-DataQuality-BancoPan-recipe

Versão em operação

Mais

Criar sua receita

Comece a aplicar etapas de transformação aos seus dados. Todas as etapas de preparação de dados serão rastreadas na receita.

Adicionar etapa

- 2) Em AÇÕES DE COLUNA selecionar Alterar Tipo
 - 3) Em Coluna de Fonte selecionar a coluna anomesdia
 - 4) Em Alterar para o Tipo selecionar Date
 - 5) Em Identificar valores inválidos selecionar Substituir por nulo
 - 6) Clicar em Inscrever-se

The screenshot shows two windows side-by-side. On the left is the 'Alterar tipo de dados' (Change data type) dialog for the 'anomesdia' dataset. It has fields for 'Coluna de fonte' (anomesdia) and 'Alterar o tipo para' (date). A note says: 'Quando uma coluna de string é convertida em uma coluna de data, a coluna de data resultante está no formato aaaa-mm-dd.' Below are options for handling invalid values: 'Excluir linhas' (Exclude rows), 'Substituir por nulo' (Replace by null) (selected), 'Substitua pelo valor de data personalizado' (Replace with custom date value), and a 'Visualizar alterações' (View changes) button. At the bottom are 'Cancelar' and 'Inscriver-se' buttons. On the right is a metrics report titled 'anomesdia' showing distribution of data types: Distinto 2, Exclusiva 1, Total 499. It includes a bar chart for '2021-03-31' (499, 99,8%) and 'null' (1, 0,2%). Below the report is a table with six rows, each containing the date '2021-03-31'.

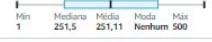
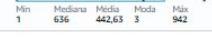
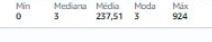
[adicionamos o arquivo json das receitas criadas:]

Depois de aplicadas as Receitas podemos usar os relatórios do Databrew para várias análises dos dados.

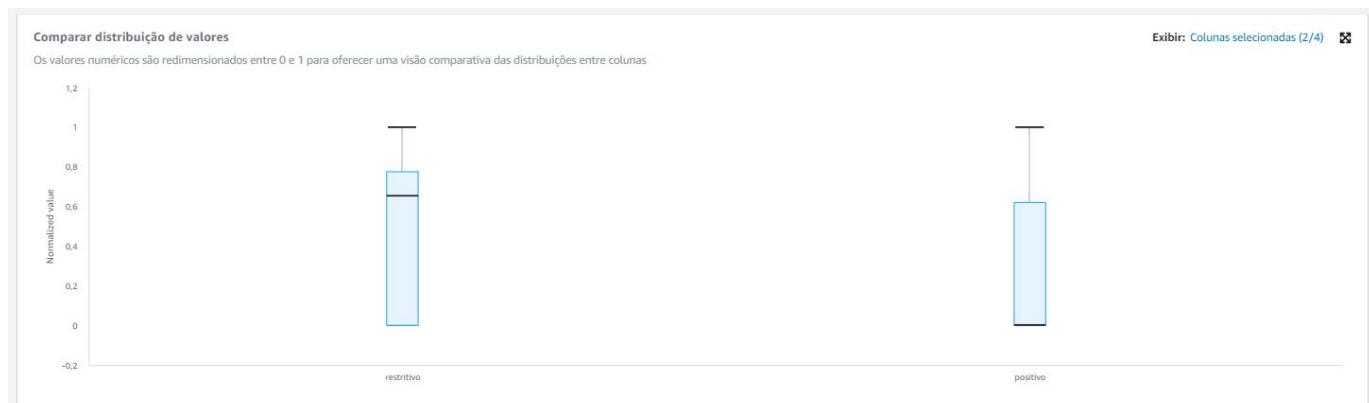
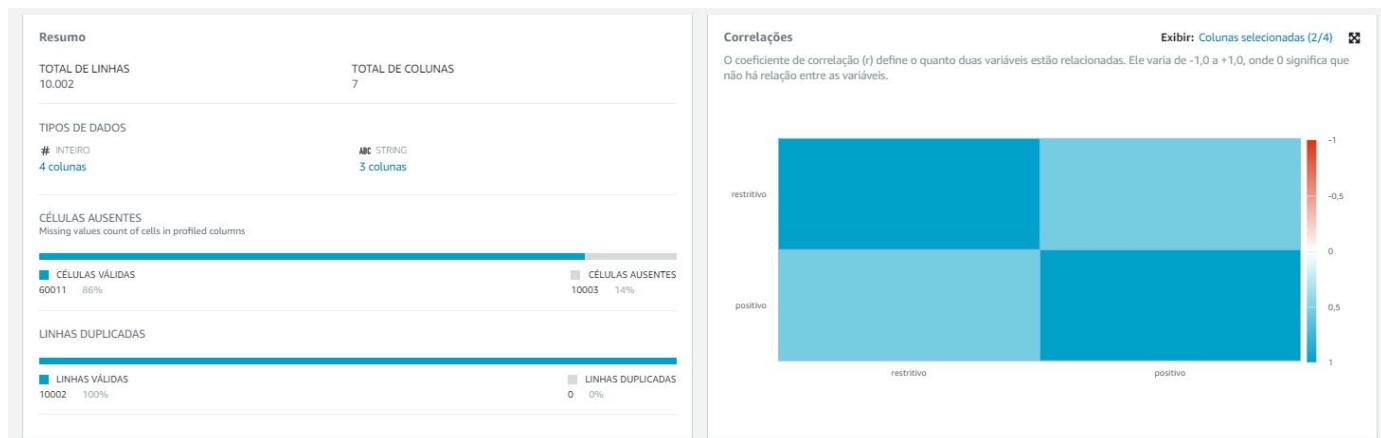
Clicando em ESQUEMA e em PERFIL podemos ver os resultados de métricas do próprio DataBrew. No Perfil precisamos CRIAR PERFIL DE DADOS, com um novo S3 de saída. Neste caso criamos o s3://grupoc-dq-pan/perfil, e o formato é json:

The screenshot shows the 'Executar trabalho: grupoc-dq profile job' dialog. It has sections for 'Locais de saída' (Output locations) and 'Exemplo de dados' (Example of data). In 'Locais de saída', it shows 'Tipo de arquivo' (JSON) and 'Local do S3' (s3://grupoc-dq-pan/perfil/). In 'Exemplo de dados', it shows 'Conjunto de dados completo' (selected) and 'Exemplo personalizado' (unchecked). A text input field contains '20000' with the label 'linhas'. Below it is a note: 'O valor deve ser maior que zero'. At the bottom are 'Cancelar' and 'Executar trabalho' buttons.

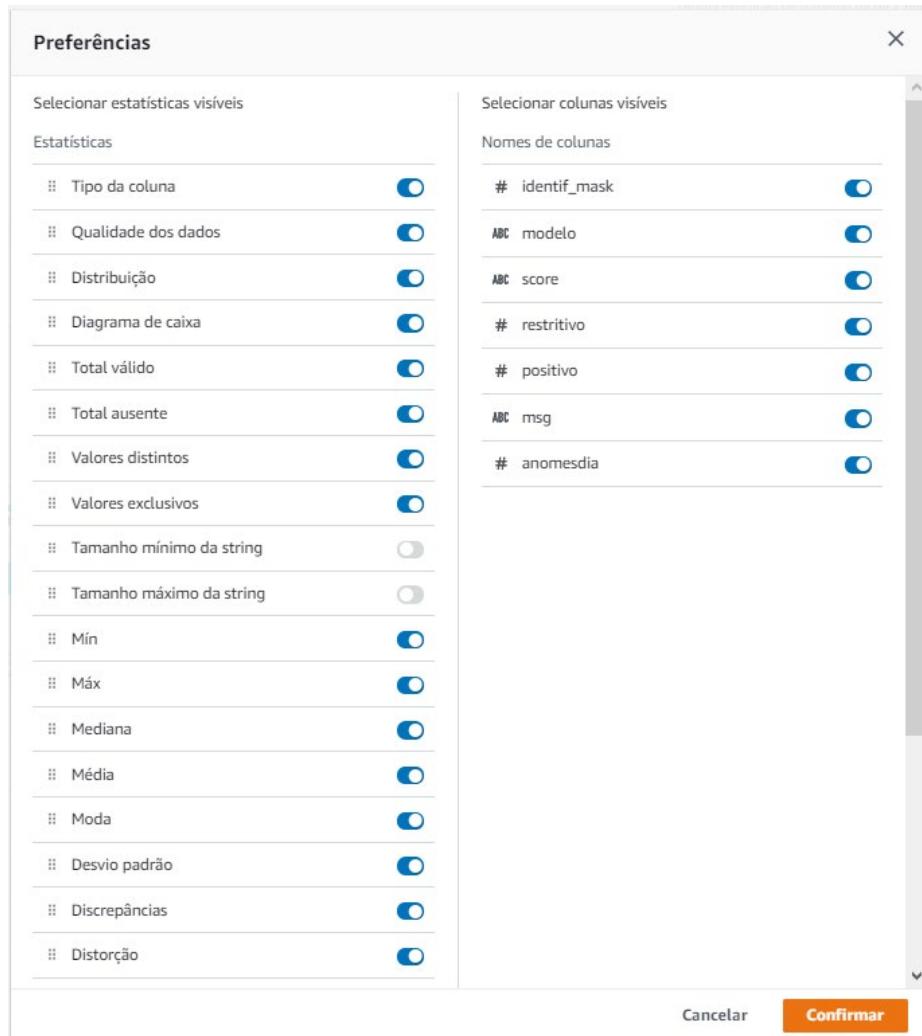
VISUALIZANDO O ESQUEMA DOS DADOS NO DATABREW

	<input type="checkbox"/> Mostrar/Ocultar	Nome da coluna	Tipo de dados	Qualidade dos dados	Distribuição de valores	Diagrama de caixa
1	<input checked="" type="checkbox"/>	identif_mask	# integer	100% Válido 0% Inválido	 Distinto: 496 Exclusiva: 0 Total: 496	
2	<input checked="" type="checkbox"/>	modelo	ABC string	100% Válido 0% Inválido	 Distinto: 1 Exclusiva: 0 Total: 496	
3	<input checked="" type="checkbox"/>	score	# integer	100% Válido 0% Inválido	 Distinto: 230 Exclusiva: 295 Total: 496	
4	<input checked="" type="checkbox"/>	restritivo	# integer	100% Válido 0% Inválido	 Distinto: 176 Exclusiva: 120 Total: 496	
5	<input checked="" type="checkbox"/>	positivo	# integer	100% Válido 0% Inválido	 Distinto: 61 Exclusiva: 34 Total: 496	
6	<input checked="" type="checkbox"/>	msg	ABC string	0% Válido 0% Inválido 100% Ausente	 Distinto: 1 Exclusiva: 0 Total: 0	
7	<input checked="" type="checkbox"/>	Data	Date Time	100% Válido 0% Inválido	 Distinto: 1 Exclusiva: 0 Total: 496	

VISUALIZANDO O PERFIL DOS DADOS NO DATABREW



As métricas estatísticas disponíveis no DataBrew são:



Resumo das Colunas #Strings

Resumo de colunas (7)			
<input style="width: 100px; height: 20px; border: 1px solid black; padding: 2px; margin-bottom: 5px;" type="button" value="Rolar até a coluna"/>			
Nome da coluna	modelo	score	msg
Tipo da coluna	ABC String	ABC String	ABC String
Qualidade dos dados <small>Validade baseada no tipo da coluna</small>	100% Válido 0% Inválido	99% Válido 0% Inválido <1% Ausente	<1% Válido 0% Inválido 99% Ausente
Distribuição	 Distinto 2 Exclusiva 0 Total 10.002	 Distinto 2.621 Exclusiva 1.934 Total 9.999	 Distinto 9 Exclusiva 5 Total 16
Total válido	10002 (100%)	9999 (100%)	16 (0%)
Total ausente	0 (0%)	3 (0%)	9986 (100%)
Valores distintos <small>Valores que ocorrem pelo menos uma vez</small>	2 (0%)	2621 (26%)	9 (0%)
Valores exclusivos <small>Valores que ocorrem somente uma vez</small>	0 (0%)	1934 (19%)	5 (0%)
Tamanho mínimo da string	3	3	1
Tamanho máximo da string	3	6	8

Resumo das Colunas #Integer

Nome da coluna	identif_mask	restrictivo	positivo
Tipo da coluna	# Inteiro	# Inteiro	# Inteiro
Qualidade dos dados Validade baseada no tipo da coluna	100% Válido 0% Inválido	99% Válido 0% Inválido <1% Ausente	99% Válido 0% Inválido <1% Ausente
Distribuição	 Distinto 5.145 Exclusiva 288 Total 10.002	 Distinto 535 Exclusiva 89 Total 9.998	 Distinto 367 Exclusiva 141 Total 9.993
Diagrama de caixa			
Total válido	10002 (100%)	9998 (100%)	9993 (100%)
Total ausente	0 (0%)	4 (0%)	9 (0%)
Valores distintos Valores que ocorrem pelo menos uma vez	5145 (51%)	535 (5%)	367 (4%)
Valores exclusivos Valores que ocorrem somente uma vez	288 (3%)	89 (1%)	141 (1%)
Mín	1	1	0
Máx	5145	954	980
Mediana	2502.5	626	3
Média	2504.996400719856	436.23114622924584	231.07665365756029
Mediana	2502.5	626	3
Média	2504.996400719856	436.23114622924584	231.07665365756029
Moda	1	3	3
Desvio padrão	1448.2598759615644	360.10843687422675	352.30912889985484
Discrepâncias	0 valores	0 valores	0 valores
Distorção	0.008151568363145588	-0.2997702734462556	0.9579769770536442
Soma	25054974	4361439	2309149
Primeiro quartil (25%)	1252	3	0
Segundo quartil (50%)	2502.5	626	3
Terceiro quartil (75%)	3754	742	608
Intervalo interquartis	2502	739	608
Variância	2097456.668320206	129678.08630799896	124121.72230617453
Curtose	-1.18915209224926	-1.7801356074530663	-0.9460223669438448

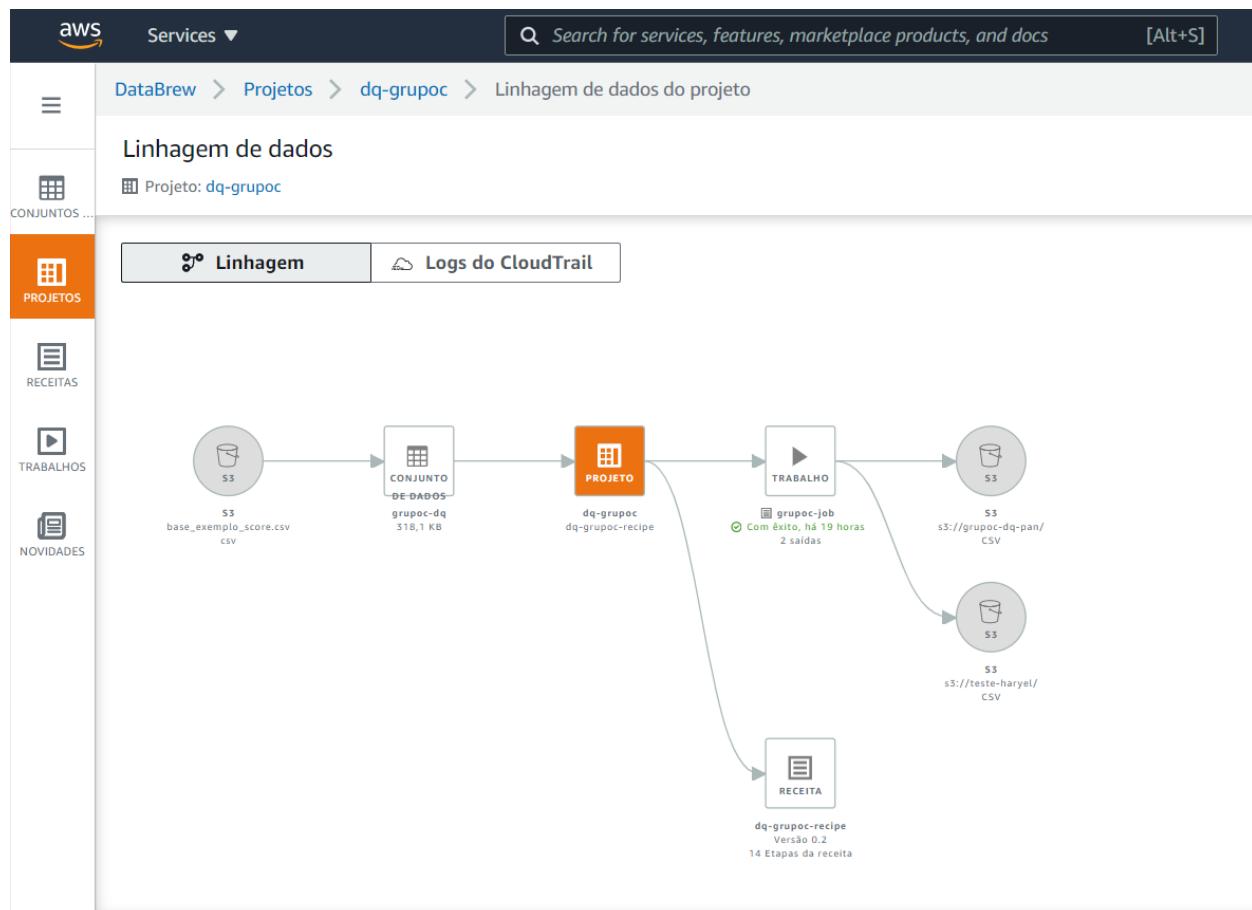
PREPARANDO A SAÍDA DOS DADOS

CRIANDO JOBS

Para que as receitas criadas sejam aplicadas a todas as alterações e inclusões de dados feitas na base precisamos criar Trabalho - Jobs

Configuramos duas saídas, conforme Linhagem de Dados, com dois Buckets de output; s3://grupoc-dq-pan/csv para futuros ETLs, os quais também podem ser criadas as saídas em json e parquet, além de csv, com compactação, pensando na redução de custos e escalabilidade.

O segundo Bucket é o s3://teste-haryel/csv para configurar a saída a ser usada no AWS ATHENA.



- 1) No menu à esquerda clicar em Trabalhos ou no botão à direita superior - Criar Trabalho



2) Na tela de configuração, em Detalhes do Trabalho daremos um nome a este Job, neste caso job-bancopan.

3) No Tipo de Trabalho deixaremos a configuração padrão para o Conjunto de dados associado e para a Receita associada.

4) Aqui configuramos a saída. Uma para deixar disponível para o Banco Pan para futuras ações (até de Machine Learning) e outra que usamos no nosso desafio para associar ao Athena e ao QuickSight. Mantendo o delimitador (;) e neste caso não faremos a compactação, pois o arquivo é pequeno.

Manteremos o formato csv, que receberá as alterações feitas. Configuramos o bucket de output em: s3://grupoc-dq-pan

Na segunda com csv em: s3://teste-haryel

É aqui onde se pode configurar outras saídas para os formatos json e parquet.

5) Em Configurações Avançadas de Trabalho se pode habilitar o CloudWatch para controle de logs.

6) Em Permissões precisamos associar uma função do IAM pré-estabelecida; a project.

7) Clicar em Criar e Executar Trabalho.

Configurações de saída do trabalho
A execução de um trabalho gera arquivos de saída em destinos de arquivos especificados.

Saída 1

Saída para Local de saída	Tipo de arquivo Formato de saída	Delimitador Separador CSV	Compactação Tipos disponíveis
Amazon S3	CSV	Ponto e vírgula (,)	None

Local do S3
O formato é: s3://bucket/pasta/

Resumo da configuração

Arazenamento de saída de arquivos
Substitua arquivos de saída para cada execução de trabalho
Partição personalizada por valores de coluna
Nenhum

Previsualização do caminho de saída
s3://grupoc-dq-pan/grupoc-job_17ago2021_timestam_p_part00000.csv

Os arquivos de saída serão particionados se forem muito grandes.

Saída 2

Saída para Local de saída	Tipo de arquivo Formato de saída	Delimitador Separador CSV	Compactação Tipos disponíveis
Amazon S3	CSV	Ponto e vírgula (,)	None

Local do S3
O formato é: s3://bucket/pasta/

Resumo da configuração

Arazenamento de saída de arquivos
Crie uma nova pasta para cada execução de trabalho
Partição personalizada por valores de coluna
Nenhum

Previsualização do caminho de saída
s3://teste-harvel/grupoc-job_17ago2021_timestam_p_part00000.csv

Os arquivos de saída serão particionados se forem muito grandes.

Configurações de trabalho avançadas
Configurações que controlam o processamento e a capacidade computacional usados nos trabalhos executados no projeto

Número máximo de unidades
Definir o número máximo de nós DataBrew que podem ser alocados quando um trabalho é executado

Tempo limite do trabalho (minutos)
Definir o tempo limite de um trabalho

CloudWatch Logs
 Habilitar Amazon CloudWatch logs para o trabalho
Habilite a criação de Amazon CloudWatch logs quando este trabalho for executado. [Saiba mais](#)

Número de novas tentativas
Número máximo de vezes para repetir o trabalho em caso de falha

Tags -opcional
Metadados que você pode definir e atribuir a recursos da AWS. Cada tag é um rótulo simples que consiste em uma chave definida pelo cliente (nome) e um valor opcional. O uso de tags pode facilitar o gerenciamento, a pesquisa e a filtragem de recursos por finalidade, proprietário, ambiente ou outros critérios.

Esse recurso não tem tags.
[Adicionar tag](#)
Você pode adicionar mais tags.

Permissões
DataBrew needs permission to connect to data on your behalf. Use an IAM role with the [política obrigatória](#) attached.

Nome da função
Escolha a função que tem acesso para se conectar aos seus dados. Recarregue para ver as atualizações mais recentes.

Ao clicar em "Salvar", você autoriza o DataBrew a adicionar as permissões necessárias para acessar todos os conjuntos de dados neste trabalho à função de serviço selecionada.

Podemos conferir o Trabalho acessando no menu lateral à esquerda:

3. CRIAÇÃO DAS MÉTRICAS COM O AWS ATHENA

- 1) Selecionar na seção Análise de Dados o Athena
- 2) NA console do Athena clicar em Get Started
- 3) Conectar-se a uma fonte de dados – Connect data source
- 4) Selecionar AWSDataCatalog
- 5) Criamos uma database: Gabriel
- 6) Create Table – criamos a tabela manualmente chamada score
- 7) E create view com view_metrics

Executamos as seguintes queries para criar as métricas:

Athena Query editor **Saved queries** History Data sources Workgroup: primary Settings Tutorial Help What's new

Saved queries

Search for query Delete Query New query

Name	Description	Query
<input type="radio"/> Create CloudFront Table	Sample Hive DDL statement to create a table pointing to Clou...	<code>CREATE EXTERNAL TABLE IF NOT EXISTS cloudFront_logs (Date...</code>
<input type="radio"/> Create Table	criar tabela a partir do csv	<code>CREATE EXTERNAL TABLE IF NOT EXISTS hancopen.score (`idem...</code>
<input type="radio"/> Flights Select Query	Sample query to get the top 10 airports with the most number...	<code>SELECT origin, count(*) AS total_departures FROM #flights_per...</code>
<input type="radio"/> Create ELB Table	Sample Hive DDL statement to create a table pointing to ELB ...	<code>CREATE EXTERNAL TABLE IF NOT EXISTS #elb_logs_src (request...</code>
<input type="radio"/> Create table with partitions	Sample Hive DDL statement to create a partitioned table pon...	<code>CREATE EXTERNAL TABLE IF NOT EXISTS #flights_parquet (yr INT...</code>
<input type="radio"/> ELB Select Query	Sample query to view peak load ELBs during a particular time...	<code>SELECT elb_name, count(1) FROM #elb_logs_src WHERE elb_respo...</code>
<input type="radio"/> Load flights table partitions	Sample query to load flights table partitions using MSCK REP...	<code>MSCK REPAIR TABLE #flights_parquet;</code>
<input type="radio"/> CloudFront Select Query	Sample query to view requests per operating system during a ...	<code>SELECT os, COUNT(*) count FROM cloudFront_logs WHERE date BE...</code>
<input type="radio"/> create-table	criação da tabela no Athena	<code>CREATE EXTERNAL TABLE IF NOT EXISTS desviopadrao.example_score...</code>

Beginning of List Previous Page Next Page



`create_database_create_table.sql`



`view_metricas.sql`

Arquivos:

MÉTRICAS CRIADAS:

- [modelo \(string\)](#)
- [momento \(date\)](#)
- [count \(bigint\)](#)
- [avgscore \(double\)](#)
- [avgrestritivo \(double\)](#)
- [avgpositivo \(double\)](#)
- [maiorscore \(int\)](#)
- [menorscore \(int\)](#)
- [maiorrestritivo \(int\)](#)
- [menorrestritivo \(int\)](#)
- [maiorpositivo \(int\)](#)
- [menorpositivo \(int\)](#)
- [amplitudescore \(int\)](#)
- [amplituderestritivo \(int\)](#)
- [amplitudepositivo \(int\)](#)
- [varianciascore \(double\)](#)
- [varianciarestritivo \(double\)](#)
- [variaciapositivo \(double\)](#)
- [desviopadraoscore \(double\)](#)
- [desviopadraorestritivo \(double\)](#)
- [desviopadraapositivo \(double\)](#)

4. VISUALIZAÇÃO DAS MÉTRICAS NO QUICKSIGHT

O primeiro passo é integrar o ATHENA com o QUICKSIGHT:

1. Clicar em Nova análise
2. Depois em Novo Conjunto de Dados

The image shows two screenshots of the QuickSight interface. The top screenshot displays the 'Analyses' page with several pre-existing analyses like 'base_exemplo_score.csv analysis', 'People Overview analysis', 'Web and Social Media Analysis', 'Business Review analysis', and 'Sales Pipeline analysis'. The bottom screenshot shows the 'Novo conjunto de dados' (New data set) page where various data sets are listed, including 'view_metrics', 'score', 'desafiopian_2021_08_14...', 'base-exemplo-score-nov...', 'exemplo_score', 'desafio', 'base_exemplo_score.csv', 'People Overview', 'Web and Social Media A...', 'Sales Pipeline', and 'Business Review'. Each data set is represented by a thumbnail icon and a name.

3. Escolher Athena

This screenshot shows the 'Criar um conjunto de dados' (Create a data set) page under the 'Fontes de dados' (Data sources) section. It lists various data sources: 'Fazer upload de um arquivo' (Upload a file), 'Salesforce', 'S3 Analytics', 'S3', 'Athena' (with a red arrow pointing to it), 'RDS', 'Redshift' (discovered automatically), 'Redshift' (manual connection), 'MySQL', 'PostgreSQL', 'ORACLE', 'SQL Server', 'Aurora', 'MariaDB', 'Presto', and 'Spark'. The 'Athena' option is highlighted with a red arrow.

4. Dar um nome para a fonte de dados. O nosso é “análise-teste1” e clicar em Criar.

Nova fonte de dados do Athena ×

Nome da fonte de dados
análise_teste1

Grupo de trabalho do Athena
[primary] ▼

Validar conexão O SSL está habilitado Criar fonte de dados

5. Escolher a tabela, selecionar view_metricas e Selecionar:

Escolher sua tabela ×

gabriel

Catálogo: contém conjuntos de bancos de dados.
AwsDataCatalog ▼

Banco de dados: contém conjuntos de tabelas.
gabriel ▼

Tabelas: contém os dados que você pode visualizar.

score
 view_metricas

Editar/pré-visualizar dados Usar SQL personalizado Selecionar

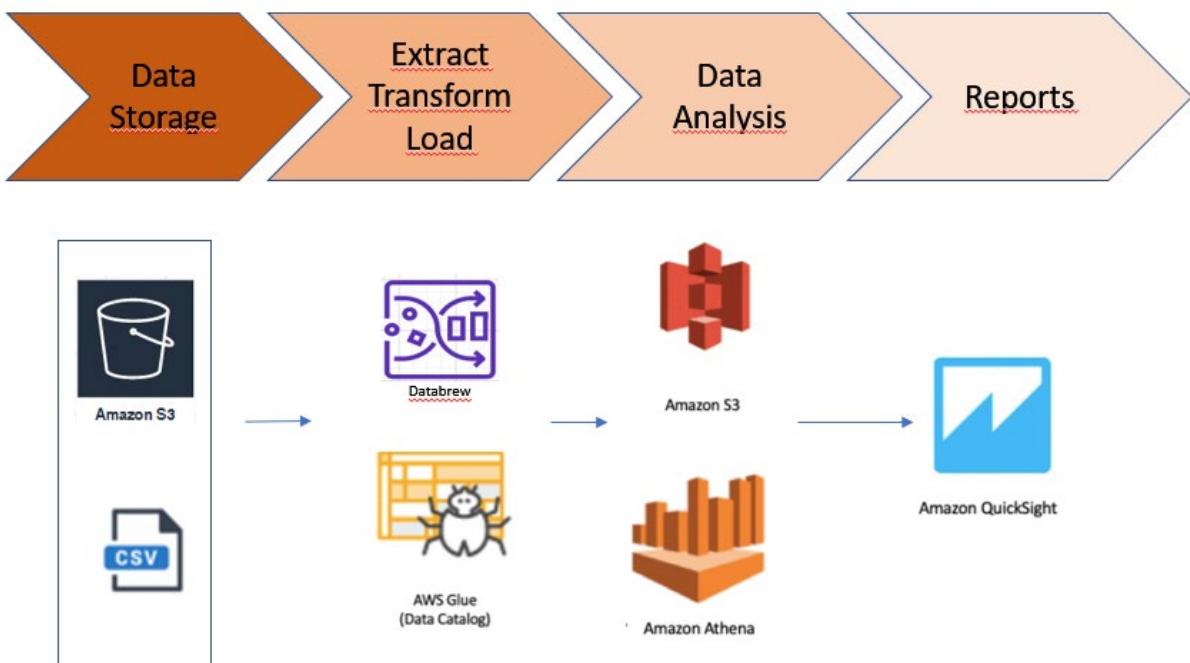




dashboard_grupoc.p
df

Arquivo:

5. ARQUITETURA DA SOLUÇÃO NA AWS



6. ACESSO CONTA AWS

E-MAIL: fiapbea@gmail.com

SENHA: AWSfiapgrupoc@2021