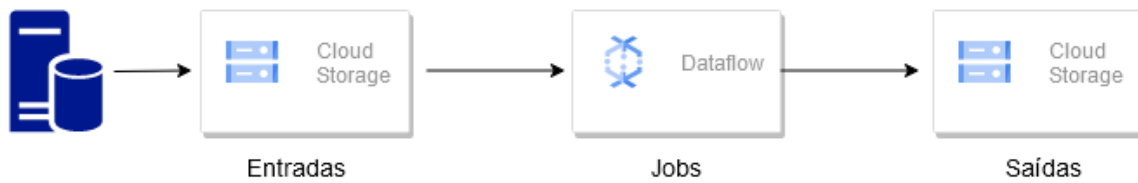


Pipeline de Dados - Engenharia de Dados



Apache Beam - Google Cloud Plataform (GCP) - DataFlow

Para a execução desse cenário algumas configurações foram necessárias na GCP (Google Cloud Plataforma) antes mesmo de rodar o script.

- Criação de Projeto;
- Criação de Bucket;
- Criação de conta de serviços (chave json);
- Habilitar as Api;

1-) Criei 04 pastas (Entrada, Saida, Temp, Template)

curso-apachebeam-bucket

Local

Classe de armazenamento

Acesso público

Proteção

southamerica-east1 (São Paulo)

Standard

Não público

Nenhum

OBJETOS

CONFIGURAÇÃO

PERMISSÕES

PROTEÇÃO

CICLO DE VIDA

Intervalos

>

curso-apachebeam-bucket

FAZER UPLOAD DE ARQUIVOS

CARREGAR PASTA

CRIAR PASTA

GERENCIAR RETENÇÕES

FAZER O DOWNLOAD

EXCLUIR

Filtrar apenas pelo prefixo do nome


Filtro

Filtrar objetos e pastas

<input type="checkbox"/>	Nome	Tamanho	Tipo	Criado ?	Classe de armazenamento	Última modificação	A
<input type="checkbox"/>	<div>ArquivoFinal/</div>	—	Pasta	—	—	—	—
<input type="checkbox"/>	<div>Entrada/</div>	—	Pasta	—	—	—	—
<input type="checkbox"/>	<div>Saida/</div>	—	Pasta	—	—	—	—
<input type="checkbox"/>	<div>Temp/</div>	—	Pasta	—	—	—	—
<input type="checkbox"/>	<div>Template/</div>	—	Pasta	—	—	—	—

Coloquei na pasta Entrada um **arquivo.csv**. Será a partir dele que iremos buscar a informação desejada. Nesse exemplo, utilizei uma planilha de voos. Criei a bucket na GCP chamada : **curso-apachebeam-bucket**.

Pasta Entrada

Intervalos > curso-apachebeam-bucket > Entrada 


FAZER UPLOAD DE ARQUIVOS


CARREGAR PASTA

CRIAR PASTA

GERENCIAR RE

Filtrar apenas pelo prefixo do nome ▼

 **Filtro** Filtrar objetos e pastas

<input type="checkbox"/>	Nome	Tamanho	Tipo
<input type="checkbox"/>	 voos_samples.csv	1,1 KB	application/vnd.ms-excel

Após a execução do script (**.ipynb**), presente também no github, geraram os seguintes arquivos;

Pasta Temporária

Intervalos > curso-apachebeam-bucket > Temp 


FAZER UPLOAD DE ARQUIVOS

CARREGAR PASTA

CRIAR


Filtrar apenas pelo prefixo do nome ▼

 **Filtro** Filtrar objetos

<input type="checkbox"/>	Nome	Tamanho	Ti
<input type="checkbox"/>	 beamapp-taraneh-0531195957-3...	—	Pe

Pasta Template

Será esses arquivos que usaremos no Dataflow do GCP.

Intervalos > curso-apachebeam-bucket > Template 


FAZER UPLOAD DE ARQUIVOS

CARREGAR PASTA

CRIAR P

Filtrar apenas pelo prefixo do nome ▼

 **Filtro** Filtrar objetos e

<input type="checkbox"/>	Nome	Tamanho	Tipo
<input type="checkbox"/>	 batch_job_df_gcs_voos	107,2 KB	appl

Criação do Dataflow (GCP)

Nome do job *

dataflow_job_voos

Precisa ser exclusivo entre o jobs em execução

Endpoint regional *

southamerica-east1 (São Paulo)

Escolha um endpoint regional de Dataflow para implantar as instâncias de worker e armazenar metadados de jobs. Se preferir, é possível implantar as instâncias de worker em qualquer região ou zona disponível do Google Cloud usando os parâmetros de região ou zona do worker. Os metadados de job sempre são armazenados no endpoint regional do Dataflow. [Saiba mais](#)

Modelo do Dataflow *

Modelo personalizado

Execute um modelo personalizado que você enviou para o Cloud Storage

Caminho do modelo *

☒ gs:// curso-apachebeam-bucket/Template/batch_job_df_gcs_voos **PROCURAR**

Caminho para seu arquivo modelo armazenado no Cloud Storage



Não foi possível analisar o arquivo de metadados deste modelo.

[VER DETALHES](#)

Se você estiver usando um modelo flexível, ative a respectiva opção de serviço.



Use o serviço de modelos flexíveis

Parâmetros obrigatórios

Local temporário *

gs://curso-apachebeam-bucket/Temp|

Prefixo do caminho e do nome de arquivo para gravar arquivos temporários. Por exemplo:
gs://your-bucket/temp

Após a execução do Dataflow

execução							
Nome	Tipo	Horário de término	Tempo decorrido	Horário de início	Status	Versão do SDK	
dataflow_job_voos	Lote	31 de mai. de 2022 18:05:18	5 min 33 s	31 de mai. de 2022 17:59:45	Finalizado	2.39.0	

Após finalizar o arquivo batch, para ter certeza que correu com sucesso, devemos visitar a pasta Saída e verificar se o arquivo está lá.


Pasta Saída

Intervalos > curso-apachebeam-bucket > Saída

FAZER UPLOAD DE ARQUIVOS CARREGAR PASTA CRIAR PASTA GERENCIAR RETENÇÕES FAZER O

Filtrar apenas pelo prefixo do nome

Filtro Filtrar objetos e pastas

<input type="checkbox"/>	Nome	Tamanho	Tipo	Criado ?	Classe de armazenament
<input type="checkbox"/>	 Voos_atrados_qtd.csv-00000-of-...	267 B	text/plain	31 de m...	Standard

```
('HNL', {'Qtd_Atrasos': [1], 'Tempo_Atrasos': [15]})
('OGG', {'Qtd_Atrasos': [1], 'Tempo_Atrasos': [138]})
('LAX', {'Qtd_Atrasos': [4], 'Tempo_Atrasos': [92]})
('DFW', {'Qtd_Atrasos': [1], 'Tempo_Atrasos': [95]})
('JFK', {'Qtd_Atrasos': [4], 'Tempo_Atrasos': [220]})
```