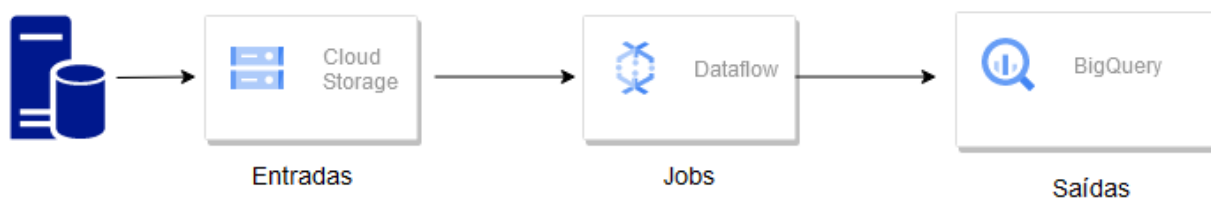


Apache Beam - Google Cloud Plataform (GCP) - DataFlow – BigQuery

Pipeline de Dados - com Big Query



Para a execução desse cenário algumas configurações foram necessárias na GCP (Google Cloud Plataforma) antes mesmo de rodar o script. Utilizei como informação uma planilha *.csv, arquivos de voos para essa atividade.

- 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;

curso-apachebeam-bucket

Local

southamerica-east1 (São Paulo)

Classe de armazenamento

Standard

Acesso público

Não público

Proteção

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	
<input type="checkbox"/>	ArquivoFinal/	—	Pasta	—	—	—	—
<input type="checkbox"/>	Entrada/	—	Pasta	—	—	—	—
<input type="checkbox"/>	Saida/	—	Pasta	—	—	—	—
<input type="checkbox"/>	Temp/	—	Pasta	—	—	—	—
<input type="checkbox"/>	Template/	—	Pasta	—	—	—	—

Coloquei o arquivo que iremos trabalhar na pasta de 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

Antes de executar o script no Python, se faz necessário criar alguns campos na tabela `dataflow_vooatrasados`. Esses campos serão os mesmos definidos no script.

curso-dataflow-beam-351817

- databigyquery
 - dataflow_vooatrasados**

 **Filtro** Insira o nome ou o valor da propriedade

Nome do campo	Tipo	Modo
airport	STRING	NULLABLE
lista_Qtd_Atrasos	INTEGER	NULLABLE
lista_Tempo_Atrasos	INTEGER	NULLABLE

```
table_schema = 'airport:STRING, lista_Qtd_Atrasos:INTEGER,
lista_Tempo_Atrasos:INTEGER'
tabela = 'curso-dataflow-beam-315923:curso_dataflow_voos.
curso_dataflow_voos_atraso'
```

Depois de executado o script, vamos verificar se nas pastas foram gerados os arquivos .

Pasta 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 > Template

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 armazenamen
<input type="checkbox"/>	 batch_job_df_gcs_voos	107,2 KB	application/octet-stream	31 de mai. de 2022 17:00:04	Standard
<input type="checkbox"/>	 batch_job_gcs_bigquerys_voos	224,2 KB	application/octet-stream	1 de jun. de 2022 15:22:58	Standard

```
pipeline_options = {
    'project': 'curso-dataflow-beam-351817' ,
    'runner': 'DataflowRunner',
    'region': 'southamerica-east1',
    'staging_location': 'gs://curso-apachebeam-bucket/Temp',
    'temp_location': 'gs://curso-apachebeam-bucket/Temp',
    'template_location': 'gs://curso-apachebeam-bucket/Template/batch_job_gcs_bigquerys_voos',
    'save_main_session' : True }
```

Gerado o arquivo batch_job_gcs_bigquery_voos

Pasta Temp/

Intervalos > curso-apachebeam-bucket > Temp

FAZER UPLOAD DE ARQUIVOS




CARREGAR PASTA

CRIAR PASTA

GERENCIAR RETENÇÕES

Filtrar apenas pelo prefixo do nome

Filtro Filtrar objetos e pastas

<input type="checkbox"/>	Nome	Tamanho
<input type="checkbox"/>	 beamapp-taraneh-0531195957-359064-k4eh5d9q.1654027197.359065/	—
<input type="checkbox"/>	 beamapp-taraneh-0601182248-581837-urz12b6z.1654107768.581838/	—
<input type="checkbox"/>	 bq_load/	—

Somente após a execução desse script podemos criar um job de execução no DataFlow.

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_gcs_bigquer

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

Criptografia

- ☒ Chave de criptografia gerenciada pelo Google
Nenhuma configuração necessária
- ☐ Chave de criptografia gerenciada pelo cliente (CMEK)
Gerenciar por meio do Google Cloud Key Management Service

✓ [MOSTRAR PARÂMETROS OPCIONAIS](#)

Parâmetros adicionais ?

+ [ADICIONAR PARÂMETRO](#)

EXECUTAR JOB

Nome	Tipo	Horário de término	Tempo decorrido	Horário de início	Status	Versão do SDK	ID
✓ dataflow_job_voos	Lote	1 de jun. de 2022 15:47:21	5 min 16 s	1 de jun. de 2022 15:42:05	Finalizado	2.39.0	2022-06-01.
✓ 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	2022-05-31.



Após a execução do Job , vamos verificar na ferramenta BigQuery se a tabela criada anteriormente foi preenchida.

Ver projetos fixos.

▼ curso-dataflow-beam-351817

▼ databigquery

dataflow_voosatrasados

dataflow_voosatrasados

CONSULTA COMPARTILHAR COPIAR

ESQUEMA DETALHES VISUALIZAR

Informações da tabela

ID da tabela	curso-dataflow-beam-351817.databigquery.dataflow_voosatrasados
Tamanho da tabela	105 B
Tamanho do armazenamento em longo prazo	0 B
Número de linhas	5
Criado	1 de jun. de 2022, 15:08:59 UTC-3
Última modificação	1 de jun. de 2022, 15:46:11 UTC-3
Validade da tabela	NUNCA
Local dos dados	southamerica-east1
Compilação padrão	[null]
Descrição	

Explorer

Digite para pesquisar

Ver projetos fixos.

curso-dataflow-beam-351817

databigyquery

dataflow_voosatrasados

dataflow... dos

dataflow_voosatrasados

CONSULTA

COMPARTILHAR

ESQUEMA

DETALHES

VISUALIZAR

Linha	airport	lista_Qtd_Atrasos	lista_Tempo_Atrasos
1	HNL	1	15
2	OGG	1	138
3	DFW	1	95
4	LAX	4	92
5	JFK	4	220

Dessa forma, a planilha foi preenchida com sucesso!! 😊