

Serviços Cloud – Big Query



Big Query

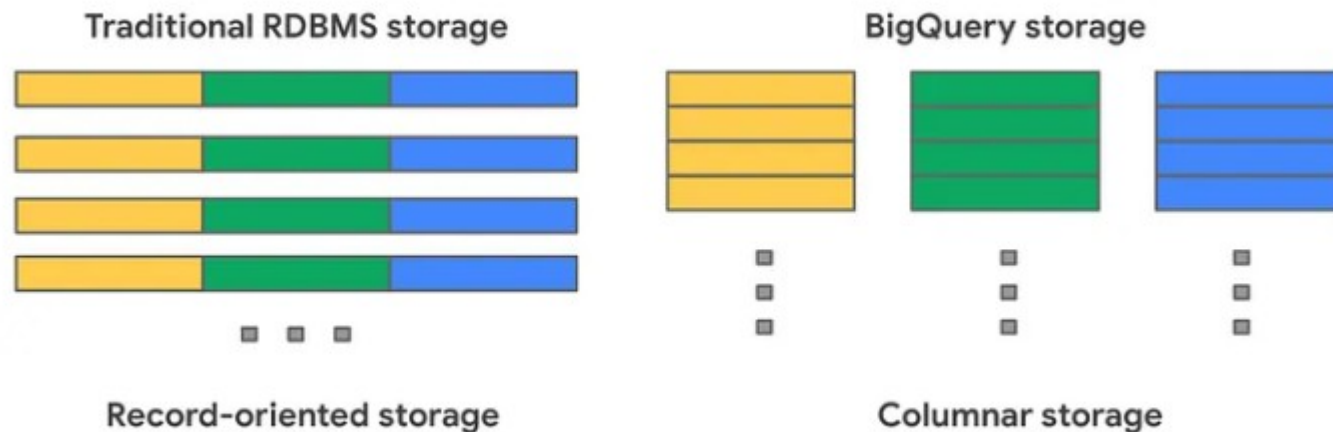
Big Query – características gerais

- Solução para Datawarehouse do google
- Banco de dados OLAP (Online Analytical Processing)
- Suporte para mutações (INSERT, UPDATE, MERGE, DELETE)
- Estrutura de armazenamento de dados corporativo
- Solução servless, escalável, utiliza a infra do google “automaticamente”
- Utiliza sintaxe SQL para manipulação dos dados
- Suporta dados na escala de PentaBytes
- API, interface gráfica, shell, linguagens de programação
- Gratuito até 10 GB de dados armazenados e 1 TB de consultas

Big Query

Big Query – Banco de dados colunar

- Armazena os dados de forma colunas, diminuindo o tempo de I/O no acesso ao dados
- Armazenamento colunas separa as colunas como se fossem tabelas, cada coluna é armazenada em um estrutura diferente
- Consultas de agregação só percorrem os dados da coluna no agrupamento
- Estrutura colunas tem melhor particionamento, indexação e compressão



Big Query

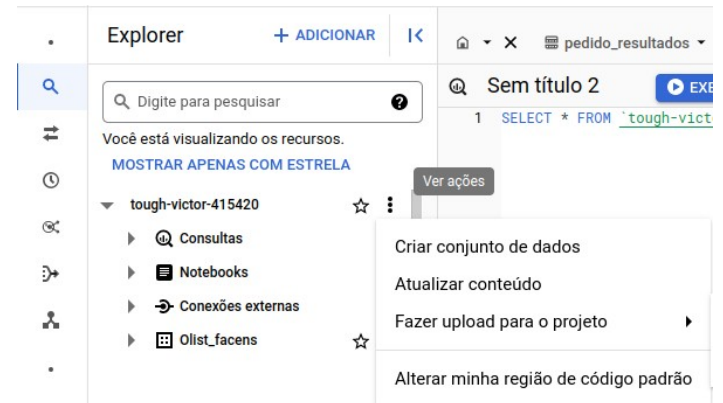
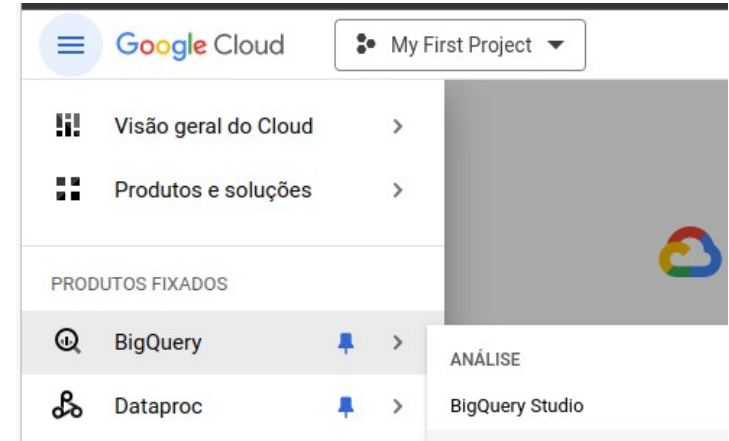
Big Query – Organização estrutural

- Projects
 - Raiz para os objetos
 - Multiplos datasets, tables, views, controles de acesso
- Datasets
 - Tables/Views ‘relacionadas’, do mesmo assunto
 - Região dos dados (Multi-regional ou regional)
- Tables
 - Registros com 1 ou mais colunas, fortemente tipados
- Views
 - Tabela virtual, definida por uma consulta SQL
- Jobs
 - Ações de carga, exportação, copia ou consulta de dados

Big Query

Big Query – Criação de Datasets e tables

- BigQuery e seus recursos
 - Menu serviços ou barra de pesquisa
 - BigQuery
 - BigQuery Studio
- Menu Explorer – Criação de datasets/Tables
 - Selecione o projeto desejado
 - Botão “Ver Ações” (3 pontos)
 - Criar conjunto de dados



Big Query

Big Query – Criação de Datasets e tables

- Criar conjunto de dados (datasets)
 - Informe o nome do dataset
 - Região ou Multiregional
 - Expiração de tabela padrão
 - Tabelas apagadas em x dias
 - Para datasets de dados temporários

Criar conjunto de dados

ID do projeto
tough-victor-415420 [MUDAR](#)

Código do conjunto de dados *
Olist_staging

Letras, números e sublinhados são permitidos

Tipo de local ?

☒ Região
Especifique uma região para colocar seus conjuntos de dados com outros serviços do Google Cloud.

☐ Multirregional
Permita que o BigQuery selecione uma região dentro de um grupo para atingir limites de cota mais altos.

Região *
us-central1 (Iowa) ▼

Expiração da tabela padrão

☒ Ativar expiração da tabela ?

Idade máxima padrão da tabela *
7 Dias

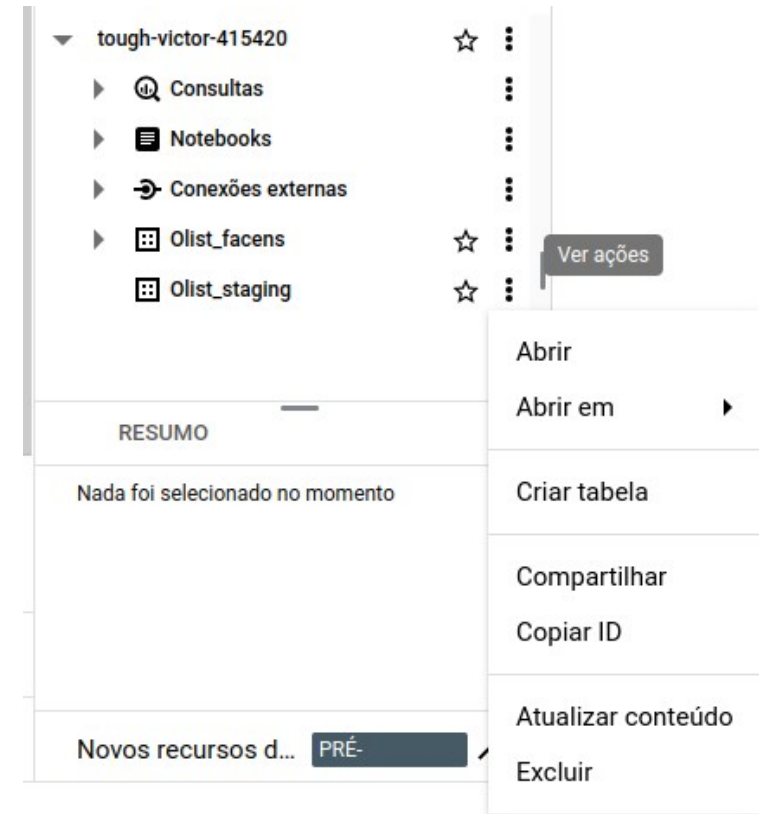
Opções avançadas ▼

[CRIAR CONJUNTO DE DADOS](#) [CANCELAR](#)

Big Query

Big Query – Criação de Datasets e tables

- Criar Tabelas (tables)
 - Selecione o dataset desejado
 - Botão “Ver Ações” (3 pontos)
 - Criar Tabela



Big Query

Big Query – Criação de Datasets e tables

- Criar Tabela - Origem
 - Origem – Cloud Storage
 - Escolher Caminho do arquivo
 - Formato arquivo – csv

Criar tabela

Origem

Criar tabela de

Google Cloud Storage

Selecione o arquivo do bucket do GCS ou [use um padrão de URI](#) *

☒ datalake_bigdata_facens_001/transient/departments/jobs/employees_2.csv

PROCURAR ?

Formato do arquivo

CSV

☐ Particionamento de dados de origem

Big Query

Big Query – Criação de Datasets e tables

- Criar Tabela - Destino
 - Projeto – Selecione seu projeto
 - Conjunto dados – Selecione o conjunto
 - Tabela – Informe o nome da tabela

Criar tabela



Destino

Projeto *

tough-victor-415420

PROCURAR

Conjunto de dados *

Olist_staging

Tabela *

employees

O tamanho máximo do nome é de 1.024 bytes UTF-8. Letras Unicode, marcas, números, conectores, traços e espaços são permitidos.

Tipo de tabela

Tabela nativa




Big Query

Big Query – Criação de Datasets e tables

- Esquema
 - Detectar automaticamente
- Particionamento
 - Sem particionamento

Esquema

☒ Detectar automaticamente

 O esquema será gerado automaticamente.

Configurações de particionamento e de cluster

Particionamento

Partição por tempo de processamento



Filtro de particionamento 

☐ Exigir cláusula WHERE para consultar dados

Tipo de particionamento 

☒ Por dia

☐ Por hora

☐ Por mês

☐ Por ano

Big Query

Sobre o particionamento de dados no Big Data

- Utilizado em muitas ferramentas de Big Data (talvez todas)
- Organiza os dados por uma coluna (sexo, status, etc)
- Menor volume de dados na leitura quando a partição é utilizada
- `df.write.format('delta').partitionBy("state").mode("overwrite").save("")`
- `spark.read.format('delta').load("/customers/state=AL")`

stackoverflow.questions_2018		
Creation_date	Title	Tags
2018-03-01	How do I??	Android
2018-03-01	When Should?	Linux
2018-03-02	This is great!	Linux
2018-03-03	Can this?	C++
2018-03-02	Help!!	Android
2018-03-01	What does?	Android
2018-03-02	When does?	Android
2018-03-02	Can you help?	Linux
2018-03-02	What now?	Android
2018-03-03	Just learned!	SQL
2018-03-01	How does!	SQL



stackoverflow.questions_2018_partitioned			
	Creation_date	Title	Tags
20180301	2018-03-01	How do I??	Android
	2018-03-01	When Should?	Linux
	2018-03-01	What does?	Android
	2018-03-01	How does!	SQL
20180302	Creation_date	Title	Tags
	2018-03-02	This is great!	Linux
	2018-03-02	Help!!	Android
	2018-03-02	When does?	Android
	2018-03-02	Can you help?	Linux
2018-03-02	What now?	Android	
20180303	Creation_date	Title	Tags
	2018-03-03	Can this?	C++
2018-03-03	2018-03-03	Just learned!	SQL

Big Query

Big Query – Laboratórios

- Laboratorio big query -
 - Criar dataset, tables, criar processo de carga e merge

Big Query

Big Query – Laboratórios

Laboratorio -

- Ler dados do mysql com spark
- Processar e salvar no storage
- Gravar e ler no big query

