

OBJETIVO

Explorar insights a partir do conjunto de dados de informações das pessoas mais ricas do mundo. Este dataset é composto por uma variedade de atributos, incluindo a classificação, fortuna final, categoria, nome da pessoa, idade, país de origem, cidade, fonte de renda, setor de atuação, entre outros. Ao examinar esses dados, esperamos descobrir tendências, padrões e informações interessantes sobre as pessoas mais ricas do planeta.

DETALHAMENTO

Fonte dos dados

<https://github.com/dipucridigital/ciencia-de-dados-e-analytics/blob/main/mvp-engenharia-de-dados/aws-mvp-pipeline-simple-report.pdf>

CATÁLOGO DOS DADOS

Principais Características

rank: A classificação do bilionário em termos de riqueza.

finalWorth: O patrimônio líquido final do bilionário em dólares americanos.

category: A categoria ou setor em que o negócio do bilionário opera.

personName: O nome completo do bilionário.

age: A idade do bilionário.

country: O país em que o bilionário reside.

city: A cidade em que o bilionário reside.

source: A fonte da riqueza do bilionário.

industries: As indústrias associadas aos interesses comerciais do bilionário.

countryOfCitizenship: O país de cidadania do bilionário.

organization: O nome da organização ou empresa associada ao bilionário.

selfMade: Indica se o bilionário é autodidata (Verdadeiro/Falso).

status: "D" representa bilionários autodidatas (Fundadores/Empreendedores) e "U" indica riqueza herdada ou não adquirida.

gender: O gênero do bilionário.

birthDate: A data de nascimento do bilionário.

lastName: O sobrenome do bilionário.

firstName: O primeiro nome do bilionário.

title: O título ou honorífico do bilionário.

date: A data de coleta dos dados.

state: O estado em que o bilionário reside.

residenceStateRegion: A região ou estado de residência do bilionário.

birthYear: O ano de nascimento do bilionário.

birthMonth: O mês de nascimento do bilionário.

birthDay: O dia de nascimento do bilionário.

cpi_country: Índice de Preços ao Consumidor (IPC) para o país do bilionário.

cpi_change_country: Variação do IPC para o país do bilionário.

gdp_country: Produto Interno Bruto (PIB) para o país do bilionário.

gross_tertiary_education_enrollment: Matrícula na educação terciária no país do bilionário.

gross_primary_education_enrollment_country: Matrícula na educação primária no país do bilionário.

life_expectancy_country: Expectativa de vida no país do bilionário.

tax_revenue_country_country: Receita tributária no país do bilionário.

total_tax_rate_country: Taxa de imposto total no país do bilionário.


population_country: População do país do bilionário.

latitude_country: Coordenada de latitude do país do bilionário.

longitude_country: Coordenada de longitude do país do bilionário.

COLETA

Obtive o dataset do link acima e inseri manualmente no bucket da GCP

NIDULA ELGIRIYEWITHANA · UPDATED 2 DAYS AGO

21

New Notebook

Download (143 kB)

Billionaires Statistics Dataset (2023)

Exploring the Global Landscape of Success

Data Card

Code (5)

Discussion (0)

About Dataset

Description

This dataset contains statistics on the world's billionaires, including information about their businesses, industries, and personal details. It provides insights into the wealth distribution, business sectors, and demographics of billionaires worldwide.

Key Features

- rank:** The ranking of the billionaire in terms of wealth.
- finalWorth:** The final net worth of the billionaire in U.S. dollars.
- category:** The category or industry in which the billionaire's business operates.
- personName:** The full name of the billionaire.
- age:** The age of the billionaire.
- country:** The country in which the billionaire resides.

Usability

10.00

License

Other (specified in description)

Expected update frequency

Annually

Tags

Computer Science

Education

News

Data Visualization

Classification

Exploratory Data Analysis

Criacao do bucket

← → ↻

console.cloud.google.com/storage/create-bucket?authorizer=1&project=valid-tower-400521

🌐 ⭐ ⚙️ 🗂️ 📄 📁 📂 📅 📆 📇 📈 📉 📊 📋 📌 📍 📎 📏 📐 📑 📒 📓 📔 📕 📖 📗 📙 📚 📛 📜 📝 📞 📟 📠 📡 📢 📣 📤 📥 📦 📧 📨 📩 📪 📫 📬 📭 📮 📯 📰 📱 📲 📳 📴 📵 📶 📷 📸 📹 📺 📻 📼 📽 📾 📿

🏠

Status do período de teste gratuito: R\$ 1.669,43 de crédito e 91 dias restantes. Com uma conta completa, você tem acesso ilimitado a todos os recursos do Google Cloud Platform.

DISPENSAR

ATIVAR

☰

Google Cloud

My First Project ▾

Pesquise (/) recursos, documentos, produtos e muito mais

🔍 Pesquisa

🔍 1 ? ☰

☰

Cloud Storage

← Criar um bucket

Ajuda

📁 Buckets

📊 Monitoramento

⚙️ Configurações

🛒 Marketplace

📝 Notas de lançamento

✓ Nomeie seu bucket

Nome: bilionario-kaggle

✓ Escolha onde armazenar seus dados

Local: us (várias regiões nos Estados Unidos)

Tipo de local: Multi-region

• Escolha uma classe de armazenamento para seus dados

Uma classe de armazenamento define os custos de armazenamento, recuperação e operações, com diferenças mínimas de tempo de atividade. Escolha se você quer que os objetos sejam gerenciados automaticamente ou especifique uma classe de armazenamento padrão com base no tempo em que planeja armazenar seus dados e sua carga de trabalho ou caso de uso. [Learn more](#) 🔗

☐ Autoclass ?

Transfere de modo automático cada objeto para um armazenamento de acesso raro ou frequente com base na atividade no nível do objeto, a fim de otimizar o

📌 Importante

📄 Preços do local

As taxas de armazenamento variam dependendo da classe de armazenamento dos dados e da localização dos buckets. [Detalhes do preço](#) 🔗

Configuração atual: Multi-region / Standard

Item	Custo
us (várias regiões nos Estados Unidos)	\$0.026 por GB/mês
Com replicação padrão	\$0.020 por GB gravado

ESTIMAR SEU CUSTO MENSAL

Upload do arquivo para dentro do bucket

Google Cloud interface showing the Cloud Storage bucket details for "billionaire-kaggle". The bucket is located in "us (várias regiões nos Estados Unidos)" and contains a file named "Billionaires Statistics Dataset.csv" (661,9 KB, text/csv, created 30 de set. de 2023 03:04:10).

On the right, there are links for "Começar a usar o Cloud Storage", "Como receber informações sobre o bucket", "Como fazer upload de dados", "Como fazer download de dados", and "Casos de uso do Cloud Storage".

Criação do bucket temporário para o datafusion

Google Cloud interface showing the Cloud Storage Buckets list. The bucket "billionaire-kaggle-temp" is highlighted in red, indicating it is the temporary bucket created for datafusion.

Nome	Criado em	Tipo de local	Local	Classe de armazenamento
billionaire-kaggle	30 de set. de 2023 03:03:37	Multi-region	us	Standard
billionaire-kaggle-temp	30 de set. de 2023 03:08:11	Multi-region	us	Standard

Criacao da instancia no bigquery

Google Cloud interface showing the BigQuery "Criar conjunto de dados" (Create dataset) form. The dataset name is "billionairekaggledataset" and the location is set to "US (várias regiões nos Estados Unidos)".

The form includes fields for "ID do projeto" (valid-tower-400521), "Código do conjunto de dados" (billionairekaggledataset), "Tipo de local" (Multirregional), and "Expiração da tabela padrão" (Ativar expiração da tabela).

Importando a tabela que tinha sido adicionada no bucket para dentro da instancia do bigquery

Star

Google Cloud

My First Project

data fusion

Pesquisa

2

?

:

!

criar tabela

Origem

Criar tabela de

Google Cloud Storage

Selecione o arquivo do bucket do GCS ou use um padrão de URI

☒ billionaire-kaggle/Billionaires Statistics Dataset.csv

PROCURAR

Formato do arquivo

CSV

☐ Particionamento de dados de origem

Destino

Projeto *

valid-tower-400521

PROCURAR

Conjunto de dados *

billionairekaggledataset

Tabela *

É necessária uma tabela de destino.

MODELAGEM

Consistiu do próprio arquivo baixado do kaggle com os dados dos bilionários e diversas dimensões sobre isso.

CARGA

O ETL foi realizado através do DataFusion, limpando algumas colunas do dataset que não faziam sentido para o objetivo deste trabalho.

Criei instancia no data fusion

Status do período de teste gratuito: R\$ 1.669,43 de crédito e 91 dias restantes. Com uma conta completa, você tem acesso ilimitado a todos os recursos do Google Cloud Platform.

DISPENSAR

ATIVAR

Google Cloud

My First Project

data fusion

Pesquisa

2

?

:

!

Data Fusion

Detalhes da instância

ATUALIZAR

EXCLUIR

FAZER UPGRADE

Instâncias

Permissões

my-datafusion

ID da instância

my-datafusion

URL da instância

[View Instance](#)

Descrição

--

Edição

BASIC

Aceleradores

[ADICIONAR ACCELERADORES](#)

Região

us-west1

Zona

--

Criado

30 de set. de 2023 03:24:30

Última atualização

30 de set. de 2023 03:36:53

Registros do Stackdriver

Desativado

Stackdriver Monitoring

Desativado

Colocando o gcs p buscar do bucket que eu criei

Editando as configurações do gcs

Cloud Data Fusion | Studio OPERATIONS HUB SYSTEM ADMIN

GCS Properties 0.22.2
Reads objects from a path in a Google Cloud Storage bucket. Validate

Properties Documentation

Format: csv GET SCHEMA

Sample Size: 1000

Override: Field Name

Enable Quoted Values: ☐ False

Use First Row as Header: ☒ True

Output Schema

Field	Type	Nullable	Required
_rank	int	<input type="checkbox"/>	<input type="checkbox"/>
finalWorth	int	<input type="checkbox"/>	<input type="checkbox"/>
category	string	<input type="checkbox"/>	<input type="checkbox"/>
personName	string	<input type="checkbox"/>	<input type="checkbox"/>
age	string	<input checked="" type="checkbox"/>	<input type="checkbox"/>
country	string	<input checked="" type="checkbox"/>	<input type="checkbox"/>
city	string	<input checked="" type="checkbox"/>	<input type="checkbox"/>
source	string	<input type="checkbox"/>	<input type="checkbox"/>
industries	string	<input type="checkbox"/>	<input type="checkbox"/>
countryOfCitizenship	string	<input type="checkbox"/>	<input type="checkbox"/>
organization	string	<input checked="" type="checkbox"/>	<input type="checkbox"/>
selfMade	string	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Fazendo as transformações com o wrangle

Cloud Data Fusion | Studio OPERATIONS HUB SY

Wrangle

Cloud Storage Default - billionaire-kaggle/Billionaires Statis...
Billionaires Statistics Dataset.csv ...

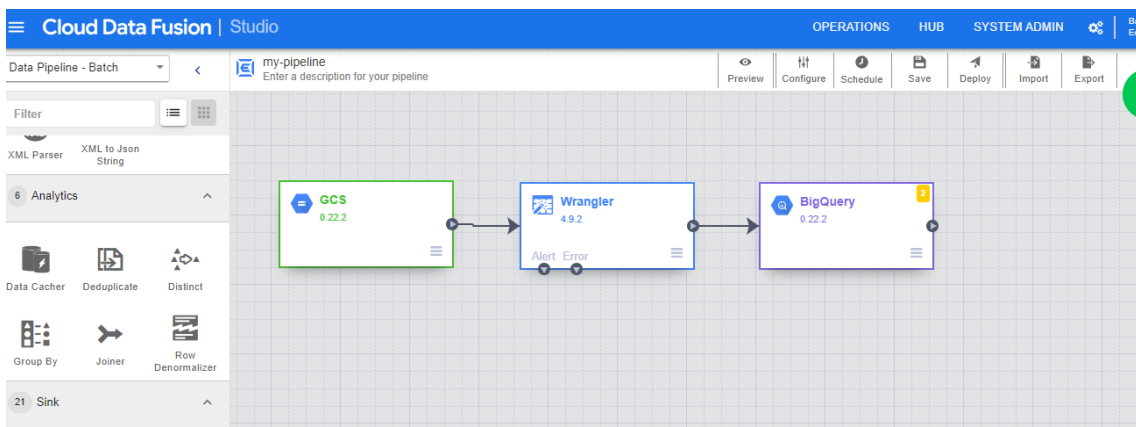
	String	String	String	String	String
	source	industries	countryOfCitizenship	organization	selfMade
	LVMH	Fashion & Retail	France	LVMH Moët Hennessy Louis Vuitton	FALSE
	Tesla, SpaceX	Automotive	United States	Tesla	TRUE
	Amazon	Technology	United States	Amazon	TRUE
	Oracle	Technology	United States	Oracle	TRUE
	Berkshire Hathaway	Finance & Investments	United States	Berkshire Hathaway Inc. (CI A)	TRUE
	Microsoft	Technology	United States	Bill & Melinda Gates Foundation	TRUE
	Bloomberg LP	Media & Entertainment	United States	Bloomberg	TRUE

Columns (35)

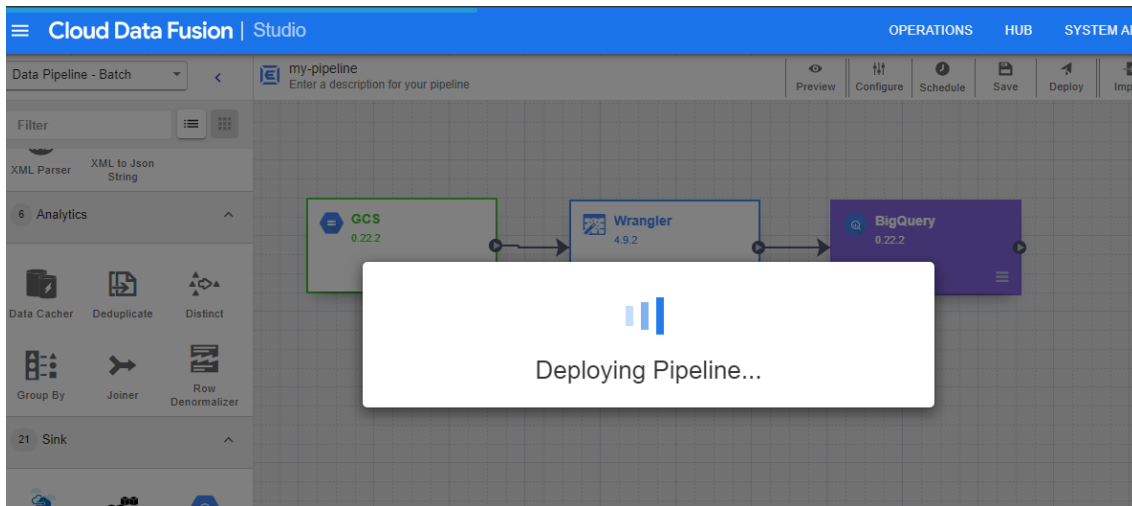
Search

- # Name
- 1 _rank
- 2 finalWor
- 3 category
- 4 personN
- 5 age
- 6 country
- 7 city
- 8 source
- 9 industrie

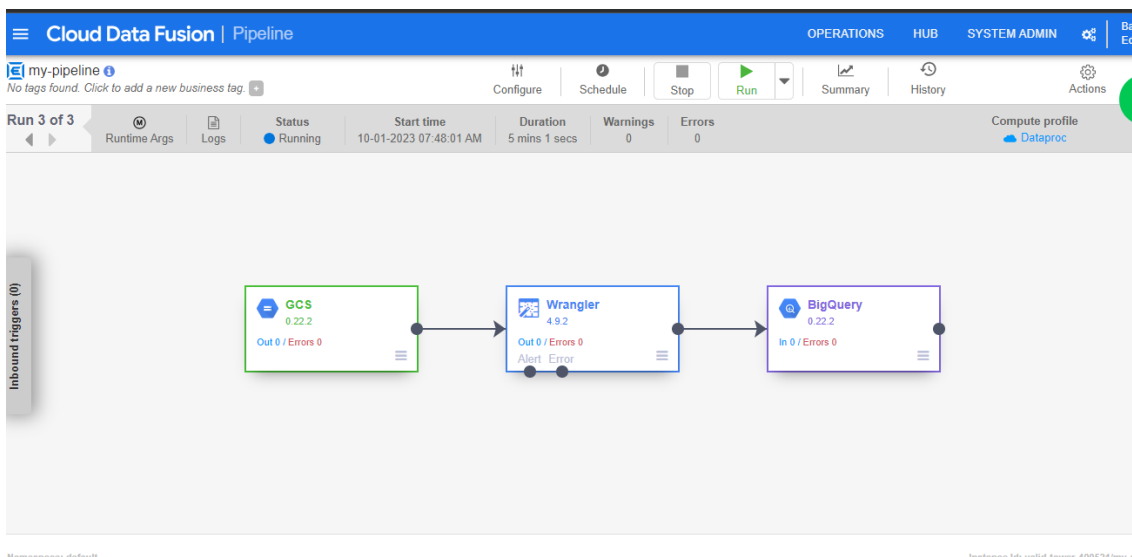
Desenho final do pipeline



Edite a caixinha do bigquery e dei o deploy



Pipeline criado



QUALIDADE DOS DADOS

O dataset possuía algumas informações faltantes (dados nulos) para alguns atributos, mas a exclusão de todas as informações faltantes de todos os atributos comprometeria muito a análise, dado que por hora um atributo faltante de um bilionário não possuía muita importância para a análise enquanto outros atributos importantes possuíam informação relevante e estavam presentes.

ANÁLISE E SOLUÇÃO DO PROBLEMA

Principais Insights da Análise:

1- Distribuição de Bilionários por País: Através da análise de agrupamento por país, consegui determinar quais nações têm a maior concentração de bilionários. Isso pode ser útil para entender as disparidades na distribuição de riqueza globalmente.

🔍	Sem título 4	EXECUTAR	SALVAR	COMPARTILHAR	PROGRAMAÇÃO	MAIS	Consulta concluída.
---	--------------	----------	--------	--------------	-------------	------	---------------------

```

1 SELECT country, COUNT(*) as num_bilionarios
2 FROM `valid-tower-400521.bilionairekaggledataset.tabela2`
3 GROUP BY country
4 ORDER BY num_bilionarios DESC;
5
6
7
8
9
10
11

```

Pressione Alt+F1 para abrir as opções de acessibilidade.

Resultados da consulta

SALVAR RESULTADOS EXPLORAR DADOS

<	INFORMAÇÕES DO JOB	RESULTADOS	JSON	DETALHES DA EXECUÇÃO	GRÁFICO	PRÉ-VISUALIZAÇÃO	GRÁFICO DE	>
Linha	country	num_bilionarios						
1	null	2525						
2	Hong Kong	65						
3	Taiwan	38						
4	Ireland	3						
5	Cayman Islands	3						
6	Bermuda	2						

2- Idade Média por Categoria: Descobri a idade média dos bilionários em diferentes categorias de riqueza. Essa análise revela as áreas de negócios ou setores que tendem a atrair pessoas mais jovens ou mais experientes.

🔍	Sem título 4	EXECUTAR	SALVAR	COMPARTILHAR	PROGRAMAÇÃO	MAIS	Consulta concluída.
---	--------------	----------	--------	--------------	-------------	------	---------------------

```

1 SELECT category, AVG(age) as idade_media
2 FROM `valid-tower-400521.bilionairekaggledataset.tabela2`
3 GROUP BY category
4 ORDER BY idade_media DESC;
5
6
7
8
9
10
11

```

Pressione Alt+F1 para abrir as opções de acessibilidade.

Resultados da consulta

SALVAR RESULTADOS EXPLORAR DADOS

<	INFORMAÇÕES DO JOB	RESULTADOS	JSON	DETALHES DA EXECUÇÃO	GRÁFICO	PRÉ-VISUALIZAÇÃO	GRÁFICO DE	>
Linha	category	idade_media						
1	Logistics	81.6						
2	Diversified	79.5						
3	Healthcare	73.0						
4	Automotive	73.0						
5	Metals & Mining	72.5						

Resultados por página: 50 1 - 18 de 18

3- Fonte de Renda Predominante: Identifiquei as 4 fontes de renda mais comuns entre os bilionários, o que pode fornecer informações sobre as indústrias mais lucrativas ou bem-sucedidas.

```

1 SELECT source, COUNT(*) as num_bilionarios
2 FROM `valid-tower-400521.bilionairekaggledataset.tabela2`
3 GROUP BY source
4 ORDER BY num_bilionarios DESC
5 LIMIT 4;
6
7
8
9
10
11

```

Pressione Alt+F1

Resultados da consulta

[SALVAR RESULTADOS](#)

	INFORMAÇÕES DO JOB	RESULTADOS	JSON	DETALHES DA EXECUÇÃO	GRÁFICO	PRÉ-VISUALIZAÇÃO
Linha	source	num_bilionarios				
1	null	2489				
2	Real estate	28				
3	Finance	13				

4- Calcular a média e a mediana da idade dos bilionários por gênero:

```

1 SELECT gender,
2        AVG(age) AS idade_media,
3        APPROX_QUANTILES(age, 2)[OFFSET(1)] AS idade_mediana
4 FROM `valid-tower-400521.bilionairekaggledataset.tabela2`
5 GROUP BY gender;
6
7
8
9
10
11

```

Pressione Alt+F1 para

Resultados da consulta

[SALVAR RESULTADOS](#)

	INFORMAÇÕES DO JOB	RESULTADOS	JSON	DETALHES DA EXECUÇÃO	GRÁFICO	PRÉ-VISUALIZAÇÃO
Linha	gender	idade_media	idade_mediana			
1	M	69.51694915254...	72			
2	F	58.82352941176...	58			
3	null	null	null			

5- Fortuna Média por Categoria: Calculei a fortuna média dos bilionários em diferentes categorias e as ordenamos em ordem alfabética. Isso nos ajuda a entender quais categorias tendem a ter maiores fortunas médias.

Sem título 4 **EXECUTAR** **SALVAR** **COMPARTILHAR** **PROGRAMAÇÃO** **MAIS**

```

1 SELECT category, AVG(finalWorth) as fortuna_media
2 FROM `valid-tower-400521.bilionairekaggledataset.tabela2`
3 GROUP BY category
4 ORDER BY category;
5
6
7
8
9
10

```

Pressione Alt+F1 para abrir

Resultados da consulta **SALVAR RESULTADOS** **EXPLC**

< INFORMAÇÕES DO JOB **RESULTADOS** JSON DETALHES DA EXECUÇÃO GRÁFICO **PRÉ-VISUALIZAÇÃO**

Linha	category	fortuna_media
1	null	null
2	Automotive	1250.0
3	Construction & Engineering	2600.0
4	Diversified	12000.0
5	Energy	2066.666666666...

6- Encontrar a idade média dos bilionários por país de cidadania.

Sem título 4 **EXECUTAR** **SALVAR** **COMPARTILHAR** **PROGRAMAÇÃO** **MAIS**

```

1 SELECT countryOfCitizenship, AVG(age) as idade_media
2 FROM `valid-tower-400521.bilionairekaggledataset.tabela2`
3 GROUP BY countryOfCitizenship
4 ORDER BY idade_media DESC;
5
6
7
8
9

```

Pressione Alt+F1 para abrir as opç

Resultados da consulta **SALVAR RESULTADOS** **EXPLORAR D**

< INFORMAÇÕES DO JOB **RESULTADOS** JSON DETALHES DA EXECUÇÃO GRÁFICO **PRÉ-VISUALIZAÇÃO**

Linha	countryOfCitizenship	idade_media
1	Colombia	84.0
2	St. Kitts and Nevis	83.0
3	Austria	82.5
4	Singapore	78.0
5	Japan	77.0

AUTOAVALIAÇÃO

Trabalhos futuros poderiam conter:

Análise Temporal: Um próximo passo valioso seria realizar uma análise temporal dos dados, se possível, para entender como a riqueza das pessoas mais ricas do mundo evoluiu ao longo do tempo. Isso exigiria um conjunto de dados atualizado.

Predição e Modelagem: Poderia ser interessante desenvolver modelos preditivos para prever tendências futuras na riqueza global com base em fatores como PIB, taxas de inflação e mudanças nas indústrias.

Visualizações Interativas: Criar visualizações interativas dos insights obtidos poderia tornar as descobertas mais acessíveis e envolventes para um público mais amplo.