

Objetivo

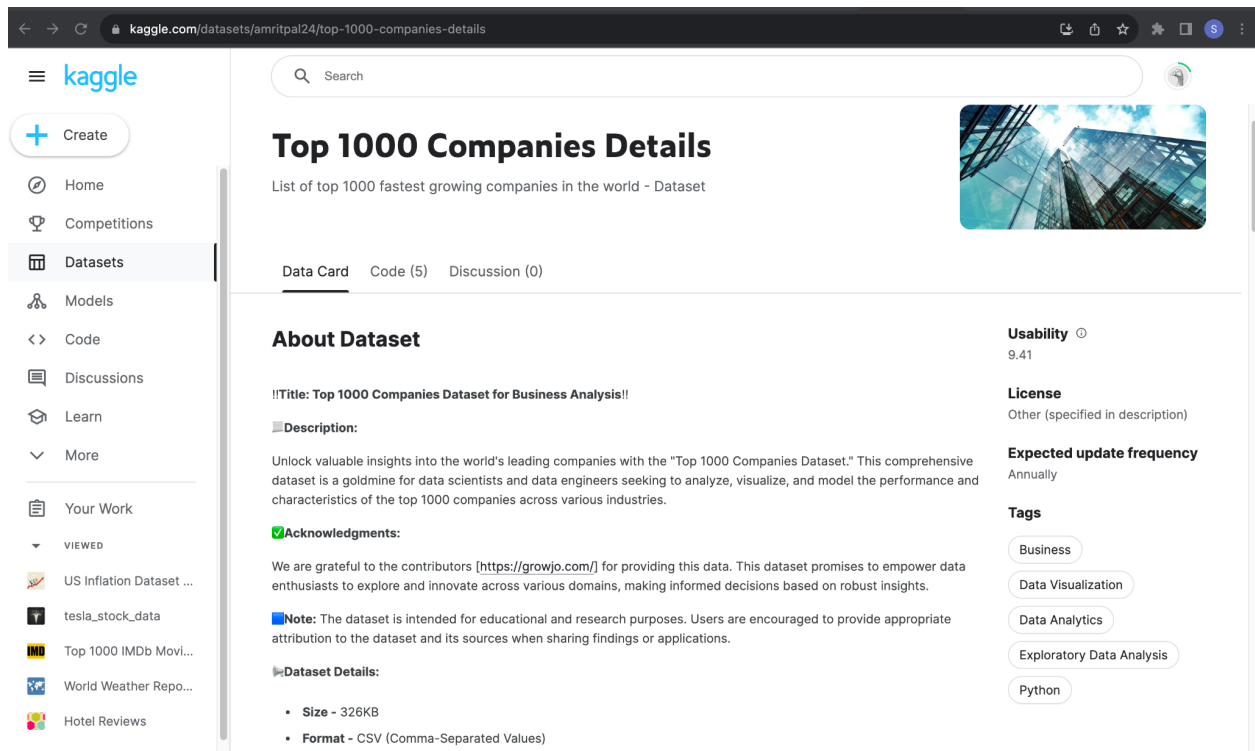
A partir de dados sobre as empresas que mais crescem no mundo, analisar:

1. Quantos colaboradores possui a empresa com o maior faturamento esperado?

Detalhamento

Busca pelos dados

<https://www.kaggle.com/datasets/amritpal24/top-1000-companies-details>



The screenshot shows the Kaggle dataset page for 'Top 1000 Companies Details'. The page layout includes a left sidebar with navigation links (Home, Competitions, Datasets, Models, Code, Discussions, Learn, More, Your Work, VIEWED) and a main content area. The main content area features a search bar, the dataset title 'Top 1000 Companies Details', a subtitle 'List of top 1000 fastest growing companies in the world - Dataset', and a thumbnail image of a modern building. Below the title, there are tabs for 'Data Card', 'Code (5)', and 'Discussion (0)'. The 'About Dataset' section contains the title '!!Title: Top 1000 Companies Dataset for Business Analysis!!', a description, acknowledgments, and a note. The 'Dataset Details' section lists the size (326KB) and format (CSV). On the right side, there are sections for 'Usability' (9.41), 'License' (Other), 'Expected update frequency' (Annually), and 'Tags' (Business, Data Visualization, Data Analytics, Exploratory Data Analysis, Python).

Top 1000 Companies Details
List of top 1000 fastest growing companies in the world - Dataset

About Dataset

!!Title: Top 1000 Companies Dataset for Business Analysis!!

Description:

Unlock valuable insights into the world's leading companies with the "Top 1000 Companies Dataset." This comprehensive dataset is a goldmine for data scientists and data engineers seeking to analyze, visualize, and model the performance and characteristics of the top 1000 companies across various industries.

Acknowledgments:

We are grateful to the contributors [https://growjo.com/] for providing this data. This dataset promises to empower data enthusiasts to explore and innovate across various domains, making informed decisions based on robust insights.

Note: The dataset is intended for educational and research purposes. Users are encouraged to provide appropriate attribution to the dataset and its sources when sharing findings or applications.

Dataset Details:

- **Size** - 326KB
- **Format** - CSV (Comma-Separated Values)

Usability 9.41

License
Other (specified in description)

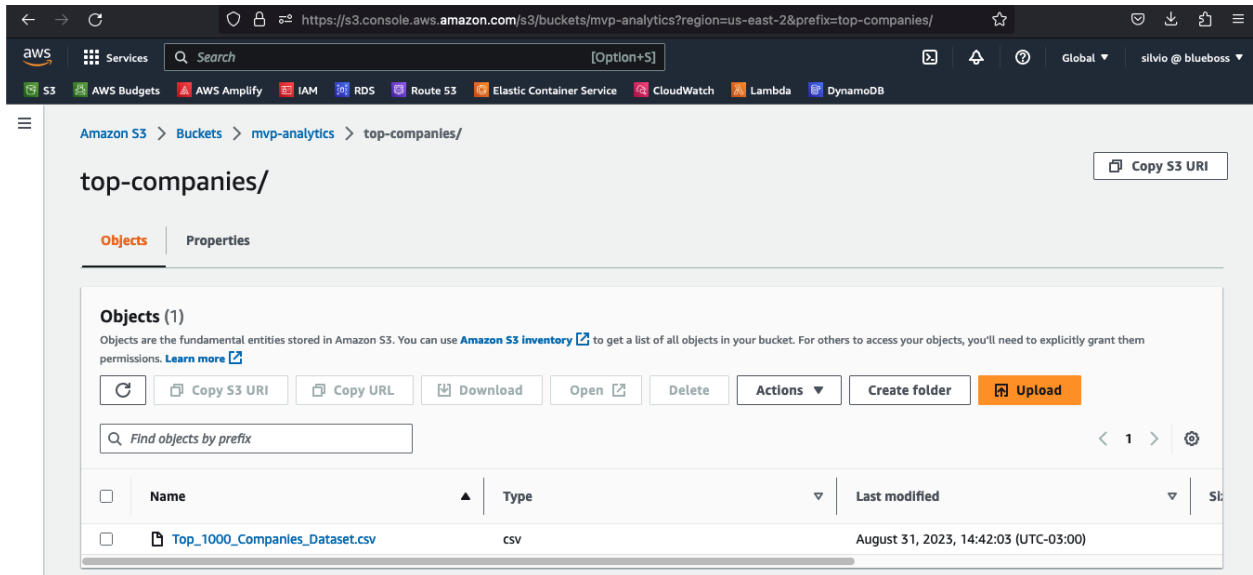
Expected update frequency
Annually

Tags

- Business
- Data Visualization
- Data Analytics
- Exploratory Data Analysis
- Python

Coleta

Dados baixados para a máquina local e inseridos manualmente em um bucket do S3.

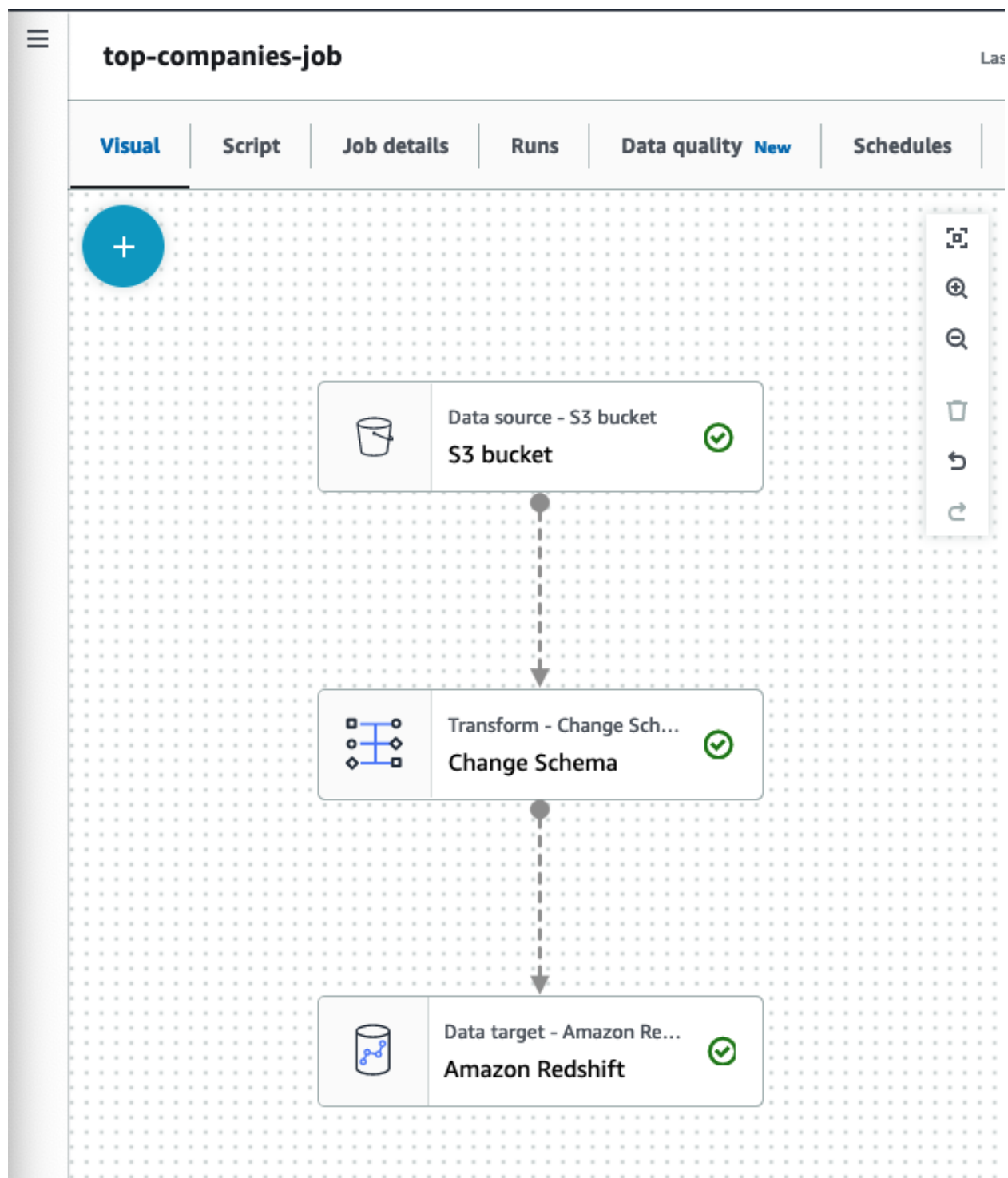


Modelagem

...

Carga

O ETL foi realizado utilizando o serviço AWS Glue. Através de sua interface visual foram criadas as seguintes etapas:



Na etapa 1, "Data source - S3 bucket", foram realizadas as configurações para extrair (*Extract*) os dados da fonte, no caso a pasta "top-companies" do bucket "mvp-analytics".

top-companies-job

Last modified on 8/31/2023, 3:08:51 PM Try new UI Actions Save Run

Visual Script Job details Runs Data quality New Schedules Version Control

Visual

+

Data source - S3 bucket S3 bucket

Transform - Change Schema Change Schema

Data target - Amazon Redshift Amazon Redshift

Data source properties - S3

Name S3 bucket

S3 source type Info

☒ S3 location Choose a file or folder in an S3 bucket.

☐ Data Catalog table

S3 URL s3://mvp-analytics/top-companies/ View Browse S3

☒ Recursive Read files in all subdirectories.

Data format CSV

Delimiter Comma (,)

Escape character - optional Enter a character to use for escaping

The character which immediately follows is used as-is, except for a small set of well-known escapes (\n, \r, \t, and \0)

Quote character Double quote (")

☒ First line of source file contains column headers

Na etapa 2, "Transform - Change Schema", realizamos a etapa de transformação (*Transform*) dos dados, convertendo os tipos dos campos "employees" e "estimated_revenues" para int.

top-companies-job

Last modified on 8/31/2023, 3:08:51 PM Try new UI Actions Save Run

Visual Script Job details Runs Data quality New Schedules Version Control

Visual

+

Data source - S3 bucket
S3 bucket

Transform - Change Sch...
Change Schema

Data target - Amazon Re...
Amazon Redshift

Data target properties - Amazon Redshift

Name
Amazon Redshift

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node

Change Schema
ApplyMapping - Transform

Redshift access type
☒ Direct data connection - recommended
☐ Glue Data Catalog tables

Redshift connection
Choose the AWS Glue connection for Amazon Redshift, or [create a new connection](#)

glue-redshift

Connection
View properties

Database
dev

Schema
Choose your Amazon Redshift schema.
public

Table
Search and enter the name of the source Amazon Redshift table.
top_companies

Handling of data and target table
☐ APPEND (insert) to target table

Por último, registramos a execução dos jobs após todas as configurações.

top-companies-job

Last modified on 8/31/2023, 3:08:51 PM

Try new UI

Actions

Save

Run

Visual

Script

Job details

Runs

Data quality New

Schedules

Version Control

Job runs (1/2) Info

Last updated (UTC)
August 31, 2023 at 20:11:59

View details

Stop job run

Table View

Card View

Filter job runs by property

Run status

Retries

Start time

End time

Duration

Capacity (DP...

Worker type

Glue ver

Succeeded

0

08/31/2023 15:09:03

08/31/2023 15:10:19

1 m 1 s

2 DPUs

G.1X

4.0

Succeeded

0

08/31/2023 14:50:56

08/31/2023 14:52:25

1 m 11 s

10 DPUs

G.1X

4.0

08/31/2023 15:09:03

Job name

Id

Run status

Glue version

top-companies-job

jr_92176f975777f4940c6599532db684a317

f850e7254d98af23c19b95e0447273

Succeeded

4.0

Retry attempt number

Start time

End time

Start-up time

Initial run

August 31, 2023 3:09:03 PM

August 31, 2023 3:10:19 PM

15 seconds

Execution time

Last modified on

Trigger name

Security configuration

1 minute 1 second

August 31, 2023 3:10:19 PM

-

-

Timeout

Max capacity

Number of workers

Worker type

3 minutes

2 DPUs

2

G.1X

Execution class

Log group name

Cloudwatch logs

Performance and debugging recommendations

Standard

/aws-glue/jobs

All logs

View in CloudWatch

Análise

Qualidade de dados

...

Solução do problema

The screenshot displays the Amazon Redshift Query Editor v2 interface. On the left, the 'Editor' sidebar shows a tree view of resources under 'Serverless: default-workgroup', including 'awsdatacatalog', 'dev', 'public', 'Tables' (2), 'sep', and 'top_companies' (highlighted). Below this, the 'top_companies' table schema is shown with columns: company_name (character varying(256)), employees (integer), and estimated_revenues (integer).

The main editor area shows a SQL query in 'Untitled 1':

```
1 select * from public.top_companies
2 where estimated_revenues = (select max(estimated_revenues) from public.top_companies)
```

The query is executed, and the results are displayed in the 'Result 1 (1)' pane. The results table has columns: company_name, employees, and estimated_revenues. The first row shows 'Alexandria Real Estate Equities' with 627 employees and 2010000000 in estimated revenues. The 'employees' value '627' is circled in red.

At the bottom right, the status bar indicates 'Elapsed time: 2500 ms' and 'Total rows: 1'.

company_name	employees	estimated_revenues
Alexandria Real Estate Equities	627	2010000000