
ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

Especialização – Engenharia de Dados & Big Data

eEDB -011 Ingestão de Dados

EX01 – Ingestão de Dados via Programação Visual (ETL)

Grupo 01:

André Maggio

Nº PECE: 97421

Daniel Gatti Richart

Nº PECE: 97204

Denis Oliveira Duarte

Nº PECE: 96942

Diane de Paula

Nº PECE: 97103

Eder Paulo

Nº USP: 14665300

Thais Bortolotti

Nº USP: 10769427

Profa. Leandro Mendes Ferreira

São Paulo
(2023)

Descrição:

Nesta atividade vocês devem usar uma ferramenta de programação visual, sendo a sugerida em Cloud o AWS Glue ETL. Caso queiram fazer com ferramenta local (em suas próprias máquinas) sugiro o Apache Hop, Pentado PDI(Kettle) ou Talend, mas podem usar a ferramenta que quiserem.

A fonte de dados em anexo no ZIP. Abaixo uma descrição das fontes de dados e do que a tabela final que deve ser gerada:

Resumo: Os arquivos são relacionados a bancos comerciais do Brasil. As fontes são o Bacen e o Glassdor. Nesses arquivos contém reclamações dos bancos registrados no Bacen, Classificação dos Bancos e Avaliações dos funcionários dos bancos no Glassdor.

Diretório Bancos: Arquivos com nome, cnpj e classificação do banco

Diretório Empregados: Arquivos contendo um resumo das avaliações dos funcionários dos bancos (esse arquivo foi gerado via scrap e relacionado via lógica fuzzy com o CNPJ do banco)

Reclamações: Arquivos de reclamações registradas no Bacen

A tabela final entregue deve conter os seguintes dados:

Nome do Banco

CNPJ

Classificação do Banco

Quantidade de Clientes do Bancos

Índice de reclamações

Quantidade de reclamações

Índice de satisfação dos funcionários dos bancos

Índice de satisfação com salários dos funcionários dos bancos.

Solução:

1. Extract

1.1 Arquivos de input:

- Bancos:
 - EnquadramentoInicia_v2.tsv
- Empregados:
 - Glassdoor_consolidado_join_match_less_v2.csv
 - Glassdoor_consolidado_join_match_v2.csv
- Reclamações:
 - 2022_tri_02_nao_ha_dados.csv
 - 2022_tri_03.csv
 - 2022_tri_04.csv
 - 2021_tri_01.csv
 - 2021_tri_02.csv
 - 2021_tri_03.csv
 - 2021_tri_04.csv
 - 2022_tri_01.csv

1.2 Tratamento local dos dados nos arquivos de input:

- Padronização dos separadores
- Adição da informação de CNPJ no arquivo "Glassdoor_consolidado_join_match_v2.csv"
- Remoção de caracteres especiais dos nomes das colunas e de todos os caracteres especiais nos dados devido a erro de codificação UTF-8
- Todos os arquivos tiveram que ser reescritos em outros arquivos devido a problema de codificação UTF-8

1.3 Criação do bucket S3:

- Pastas com dados base:
 - \banco
 - \empregados
 - \reclamações
- Pastas intermediárias de Jobs de Joins
 - Join-Banco-Empregados (job Importar Bancos)
 - Join-Bancos-Reclamacoes (job Importar Reclamações - Bancos)
- Pasta de resultados
 - Result-Empregados-Reclamacoes (não populada for não executar o job (Juncao-Empregados-Reclamacoes)

Objetos (6)

Os objetos são as entidades fundamentais armazenadas no Amazon S3. Você pode acessar seus objetos, você precisará conceder permissões explicitamente a eles:









 Copiar URI do S3

 Copiar URL

 Fa

 Carregar

 Localizar objetos por prefixo

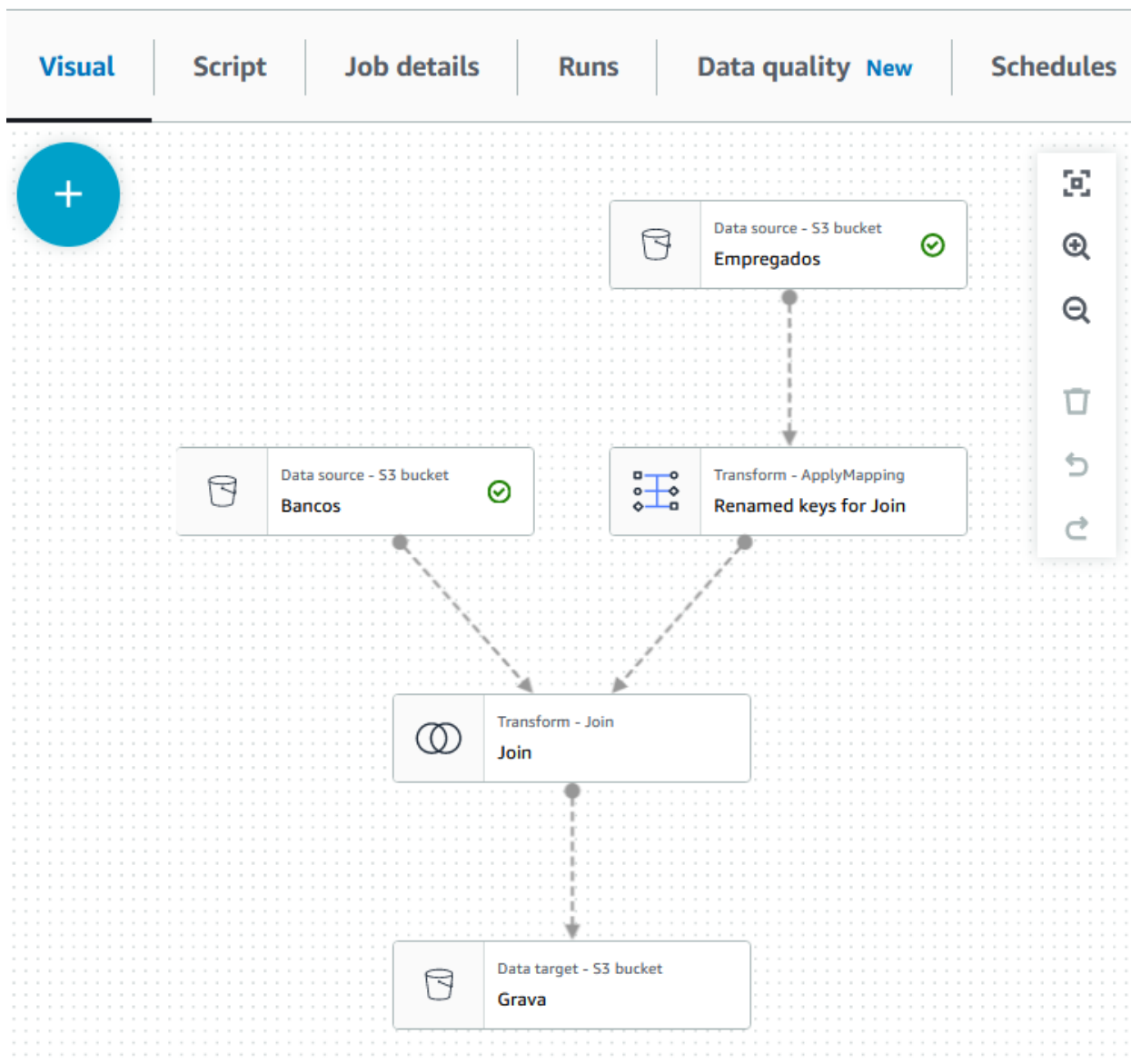
<input type="checkbox"/>	Nome ▲	Tipo
<input type="checkbox"/>	 banco/	Pasta
<input type="checkbox"/>	 empregados/	Pasta
<input type="checkbox"/>	 Join-Banco- Empregados/	Pasta
<input type="checkbox"/>	 Join-Bancos- Reclamacoes/	Pasta
<input type="checkbox"/>	 reclamacao/	Pasta
<input type="checkbox"/>	 Result-Empregados- Reclamacoes/	Pasta

2. Transform

Importar Bancos: Job para fazer join entre Bancos e Empregados
Saída em JSON com partition de CNPJ


Importar Bancos

Last m...



Importar Bancos

Last modified on 09/08/2023, 16:19:33

 Try new UI

Actions ▾

Save

Run



Unsaved job found

We found an unsaved job, do you wish to restore it?

Restore

Visual

Script

Job details

Runs

Data quality **New**

Schedules

Version Control

Job runs (1/13) Info

Last updated (UTC)
August 9, 2023 at 20:58:32



View details

Stop job run

Table View

Card View

Filter job runs by property

< 1 > ⚙

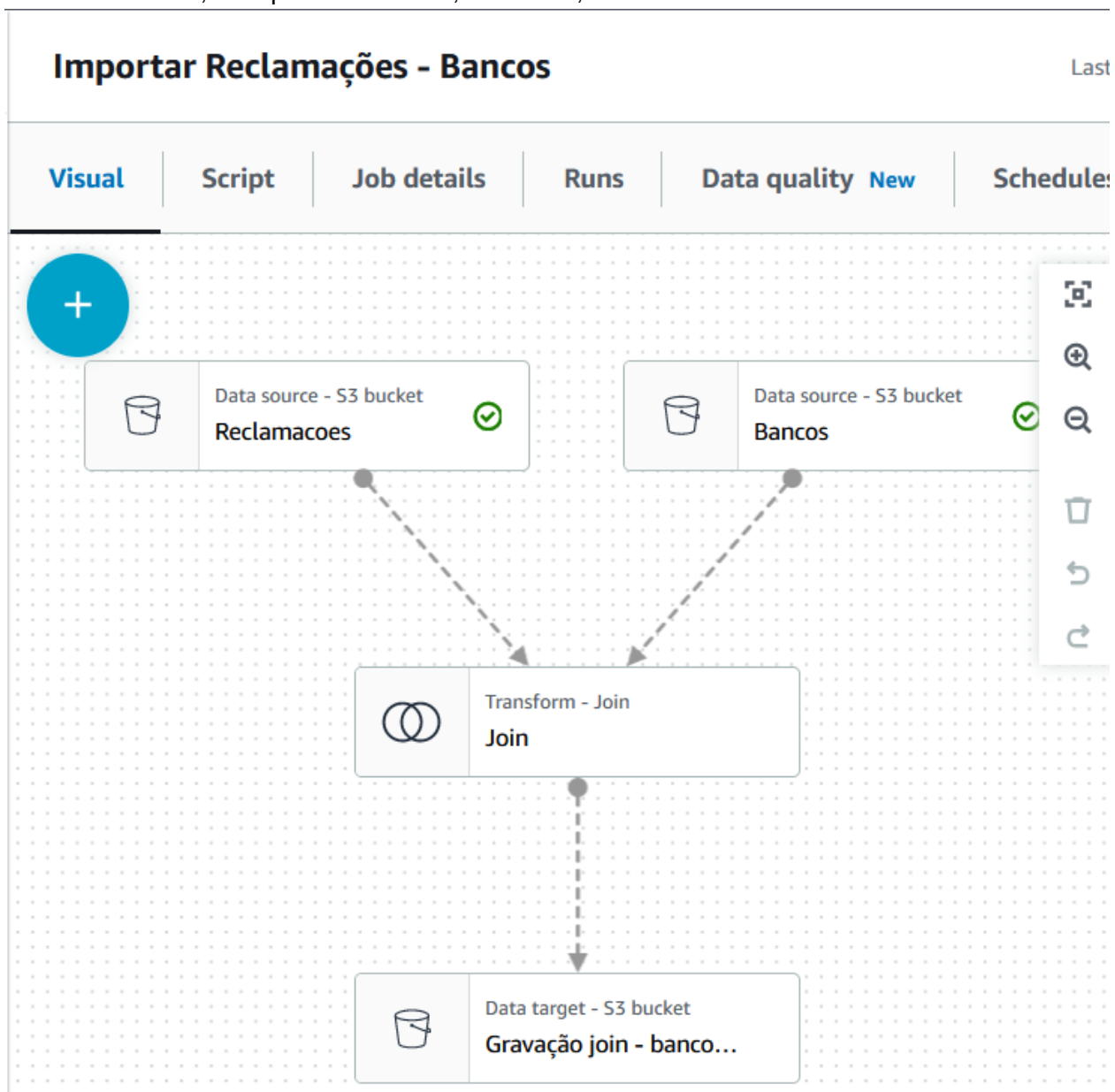
	Run status ▾	Retry ▾	Start time ▾	End time ▾	Duration ▾	Capacity ▾	Worker type ▾
<input checked="" type="radio"/>	✔ Succeeded	0	08/09/2023 16:19:37	08/09/2023 16:21:02	1 m 18 s	10 DPUs	G.1X
<input type="radio"/>	✔ Succeeded	0	08/09/2023 16:00:42	08/09/2023 16:02:43	1 m 54 s	10 DPUs	G.1X
<input type="radio"/>	✔ Succeeded	0	08/09/2023 14:42:16	08/09/2023 14:44:34	2 m 10 s	10 DPUs	G.1X
<input type="radio"/>	✔ Succeeded	0	08/08/2023 15:05:09	08/08/2023 15:07:46	2 m 9 s	10 DPUs	G.1X
<input type="radio"/>	✔ Succeeded	0	08/08/2023 14:56:47	08/08/2023 14:58:49	1 m 54 s	10 DPUs	G.1X
<input type="radio"/>	✔ Succeeded	0	08/07/2023 22:45:04	08/07/2023 22:46:01	50 s	10 DPUs	G.1X

08/09/2023 16:19:37



Job Importar Reclamações - Bancos: job para fazer a junção de bancos com reclamações:

Saída em JSON, com partition de Ano, Trimestre, CNPJ e Nome



Visual

Script

Job details

Runs

Data quality New

Schedules

Version Control

Job runs (1/8) Info

Last updated (UTC)
August 9, 2023 at 21:00:14

↺

View details

Stop job run

Table View

Card View

Q

Filter job runs by property

<

1

>

⚙

	Run status	Retry	Start time	End time	Duration	Capacity	Worker type	Gl
●	✔ Succeeded	0	08/09/2023 15:53:20	08/09/2023 15:55:29	2 m 2 s	10 DPUs	G.1X	4.0
○	✖ Failed	0	08/09/2023 15:44:27	08/09/2023 15:46:41	2 m 7 s	10 DPUs	G.1X	4.0
○	✖ Failed	0	08/09/2023 15:35:36	08/09/2023 15:37:22	1 m 39 s	10 DPUs	G.1X	4.0
○	✖ Failed	0	08/09/2023 15:27:10	08/09/2023 15:28:44	1 m 26 s	10 DPUs	G.1X	4.0
○	✖ Failed	0	08/09/2023 15:11:25	08/09/2023 15:12:53	1 m 21 s	10 DPUs	G.1X	4.0
○	✖ Failed	0	08/09/2023 14:59:54	08/09/2023 15:02:19	2 m 18 s	10 DPUs	G.1X	4.0

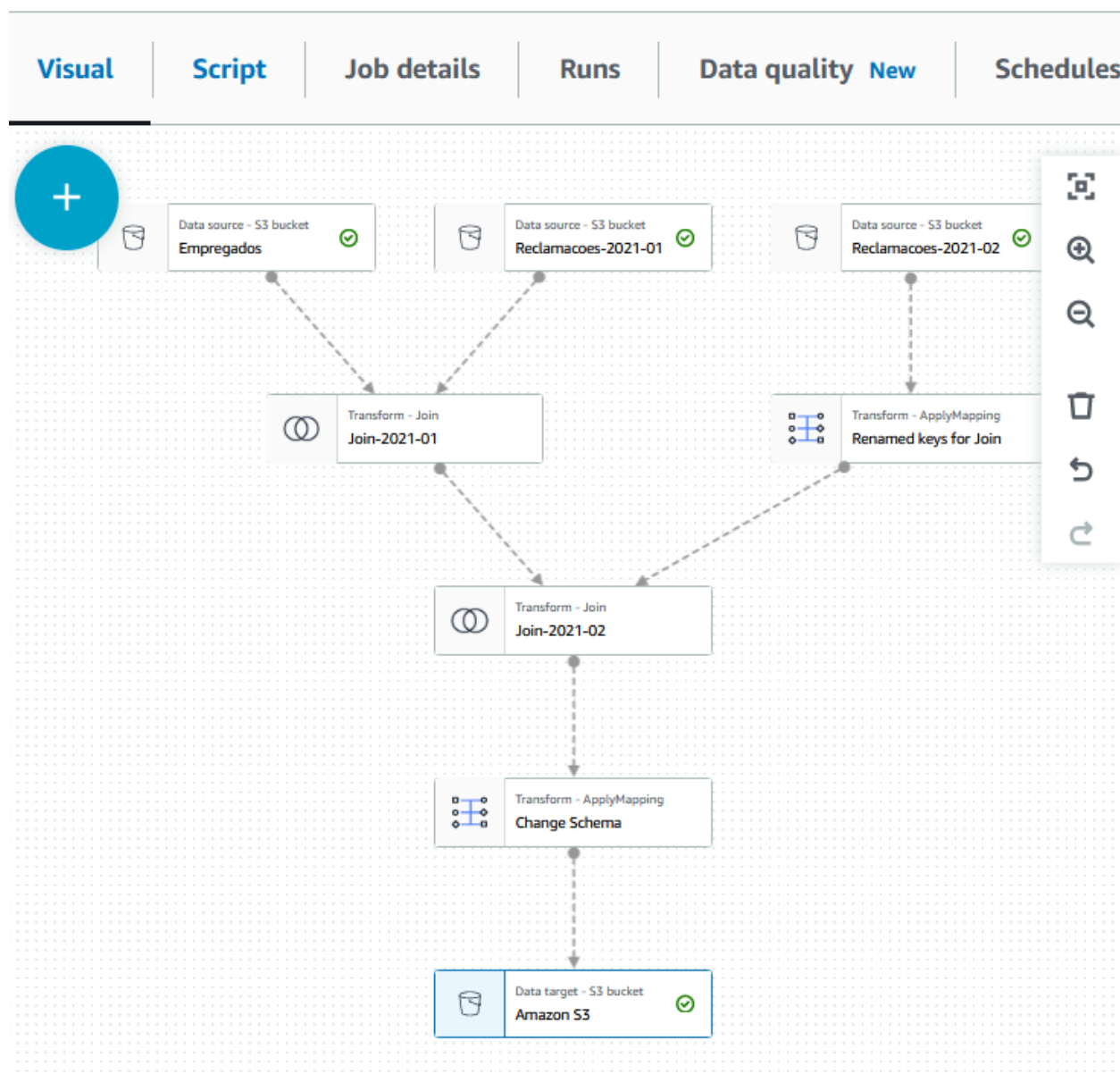
08/09/2023 15:53:20

×

Job Juncao-Empregados-Reclamacoes: Job para trazer resultado das com índices de empregados e das reclamações por trimestre (obs.: job não finalizado devido a tentativa de junção com vários itens de reclamações por trimestre o que torna o job muito grande e talvez sua execução muito longa).
Saída em JSON com partition pelo CNPJ

Juncao-Empregados-Reclamacoes

Last 1



3. Load




Apenas os jobs de join foram executados para apresentar resultados


Join-Bancos-Reclamacoes/



Objetos | Propriedades

Objetos (2)

Os objetos são as entidades fundamentais armazenadas no Amazon S3. Visando seus objetos, você precisará conceder permissões explicitamente a eles. Saiba mais

  Copiar URI do S3  Copiar URL

 Localizar objetos por prefixo

<input type="checkbox"/>	Nome ▲	Tipo
<input type="checkbox"/>	 Ano=2021/	Pasta
<input type="checkbox"/>	 Ano=2022/	Pasta

Join-Banco-Empregados/

Objetos

Propriedades

Objetos (33)

Os objetos são as entidades fundamentais armazenadas no Amazon S3. Você pode acessar seus objetos, você precisará conceder permissões explicitamente a eles



Copiar URI do S3



Copiar URL



Fa

 Localizar objetos por prefixo



Nome



Tipo



CNPJ=0/

Pasta



CNPJ=1181521/

Pasta



CNPJ=1522368/

Pasta



CNPJ=2801938/

Pasta



CNPJ=30306294/

Pasta



CNPJ=33058660/

Pasta



CNPJ=33172537/

Pasta

Referências:

<https://aws.amazon.com/pt/blogs/aws-brasil/resiliencia-do-data-lake-e-solucoes-analiticas-na-aws/>

<https://aws.amazon.com/pt/glue/>