Entrega Final – Desafio XPE (Pipeline Postgres → Kafka → S3 → Spark)

Informações:

Repositorio: https://github.com/fdelfino93/desafioxpe
Bucket AWS: `ferdelfino-xpe` (regiao `us-east-2`)
Arquivos de evidencias (prints): `docs/img`

Visao geral
Pipeline ponta-a-ponta:
- Fonte: PostgreSQL (tabelas `public.dadostesouroipca` e `public.dadostesouropre`).
- Streaming: Kafka (Confluent Platform) com topics `postgres-dadostesouroipca` e `postgres-dadostesouropre`.
- Ingestao Data Lake: Kafka Connect S3 Sink (Avro converter + Schema Registry) gravando camada Bronze em S3.
- Processamento: Spark (master local[*] via notebook) lendo Bronze e gravando Silver (Parquet) e Gold (Parquet agregacoes).

Camadas:
- Bronze: `s3a://ferdelfino-xpe/bronze/ipca/` e `s3a://ferdelfino-xpe/bronze/pre/` (JSON).
- Silver: `s3a://ferdelfino-xpe/processed-data/ipca/silver/` e `.../pre/silver/` (Parquet).
- Gold: `s3a://ferdelfino-xpe/analytics/ipca/gold/` e `.../pre/gold/` (Parquet).

Para eu subir o ambiente (Docker)
```pwsh
cd "D:\Documents\Projeto\Prova\postgres"
docker-compose up -d

cd "D:\Documents\Projeto\Prova"
docker-compose up -d
docker ps --format "table {{.Names}}\t{{.Status}}\t{{.Ports}}"
```
Esperado: broker, zookeeper, schema-registry, connect, postgres em estado Up.

Criar tópicos Kafka
```pwsh
docker exec broker kafka-topics --create --bootstrap-server localhost:9092 --partitions 1 --replication-factor 1 --topic postgres-dadostesouroipca
docker exec broker kafka-topics --create --bootstrap-server localhost:9092 --partitions 1 --replication-factor 1 --topic postgres-dadostesouropre
```

Registrar conectores
Source (Postgres → Kafka):
```pwsh
cd "D:\Documents\Projeto\Prova\connectors\source"
curl -X POST -H "Content-Type: application/json" --data @connect_jdbc_postgres_ipca.config http://localhost:8083/connectors
curl -X POST -H "Content-Type: application/json" --data @connect_jdbc_postgres_pre.config http://localhost:8083/connectors
```

Sink (Kafka → S3 Bronze) – ja configurados para bucket `ferdelfino-xpe`, regiao `us-east-2`, Avro + Schema Registry:

Evidencias (prints)

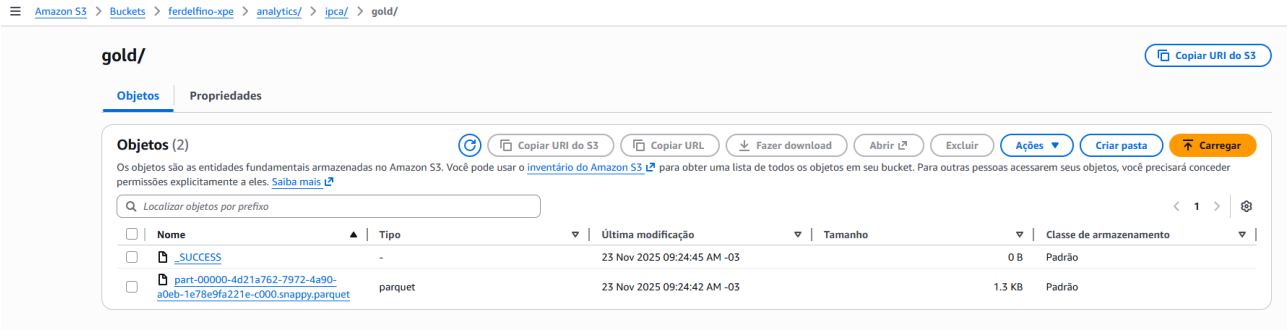
Conectores Kafka Connect RUNNING

Out[7]:

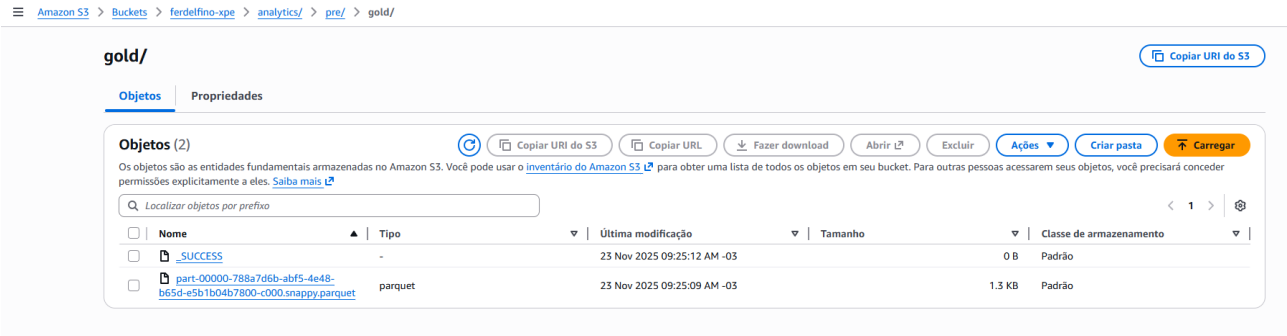
Tipo Titulo	Data Vencimento	Data Base	Taxa Compra Manha	Taxa Venda Manha	PU Compra Manha	PU Venda Manha	PU Base Manha	Data_Vencimento	Data_Base
Tesouro Prefixado	**01/01/2006**	**06/12/2005**	18.26	18.29	988.09	988.07	987.41	2006-01-01	2005-12-06
Tesouro IPCA+ com Juros Semestrais	**15/08/2006**	**06/12/2005**	12.88	12.90	1525.14	1524.96	1523.96	2006-08-15	2005-12-06
	15/05/2045	**06/12/2005**	9.07	9.17	1067.68	1056.28	1055.73	2045-05-15	2005-12-06
	15/08/2024	**06/12/2005**	9.00	9.08	1183.68	1175.01	1174.39	2024-08-15	2005-12-06
	15/05/2015	**06/12/2005**	8.98	9.04	1286.78	1281.80	1281.13	2015-05-15	2005-12-06
	15/05/2009	**06/12/2005**	11.20	11.24	1351.99	1350.48	1349.67	2009-05-15	2005-12-06
	15/05/2007	**06/12/2005**	11.90	11.92	1456.56	1456.20	1455.29	2007-05-15	2005-12-06
	15/08/2008	**06/12/2005**	11.43	11.47	1408.42	1407.18	1406.32	2008-08-15	2005-12-06
Tesouro IPCA+	**15/05/2015**	**06/12/2005**	8.68	8.74	714.11	710.42	710.05	2015-05-15	2005-12-06
	15/08/2024	**06/12/2005**	8.82	8.90	323.79	319.39	319.22	2024-08-15	2005-12-06
...	**...**	**...**	...	...	...	...	...	...	...
Tesouro Selic	**07/03/2015**	**02/08/2011**	0.00	0.04	4799.08	4792.19	4789.95	2015-03-07	2011-08-02
	07/03/2014	**02/08/2011**	0.00	0.03	4799.08	4795.36	4793.11	2014-03-07	2011-08-02
	07/03/2017	**02/08/2011**	0.00	0.04	4799.08	4788.39	4786.14	2017-03-07	2011-08-02
	07/03/2013	**02/08/2011**	0.00	0.02	4799.08	4797.56	4795.32	2013-03-07	2011-08-02
	07/03/2012	**02/08/2011**	0.00	0.01	4799.08	4798.80	4796.55	2012-03-07	2011-08-02
Tesouro Prefixado	**01/01/2014**	**02/08/2011**	12.82	12.88	747.49	746.53	746.17	2014-01-01	2011-08-02
	01/01/2013	**02/08/2011**	12.69	12.73	845.09	844.67	844.27	2013-01-01	2011-08-02
	01/01/2015	**02/08/2011**	12.85	12.91	661.63	660.43	660.11	2015-01-01	2011-08-02
	01/01/2012	**02/08/2011**	12.50	12.53	952.55	952.44	952.00	2012-01-01	2011-08-02
Tesouro IPCA+ com Juros Semestrais	**15/08/2024**	**02/08/2011**	6.29	6.37	2067.28	2053.44	2052.81	2024-08-15	2011-08-02

164293 rows × 7 columns

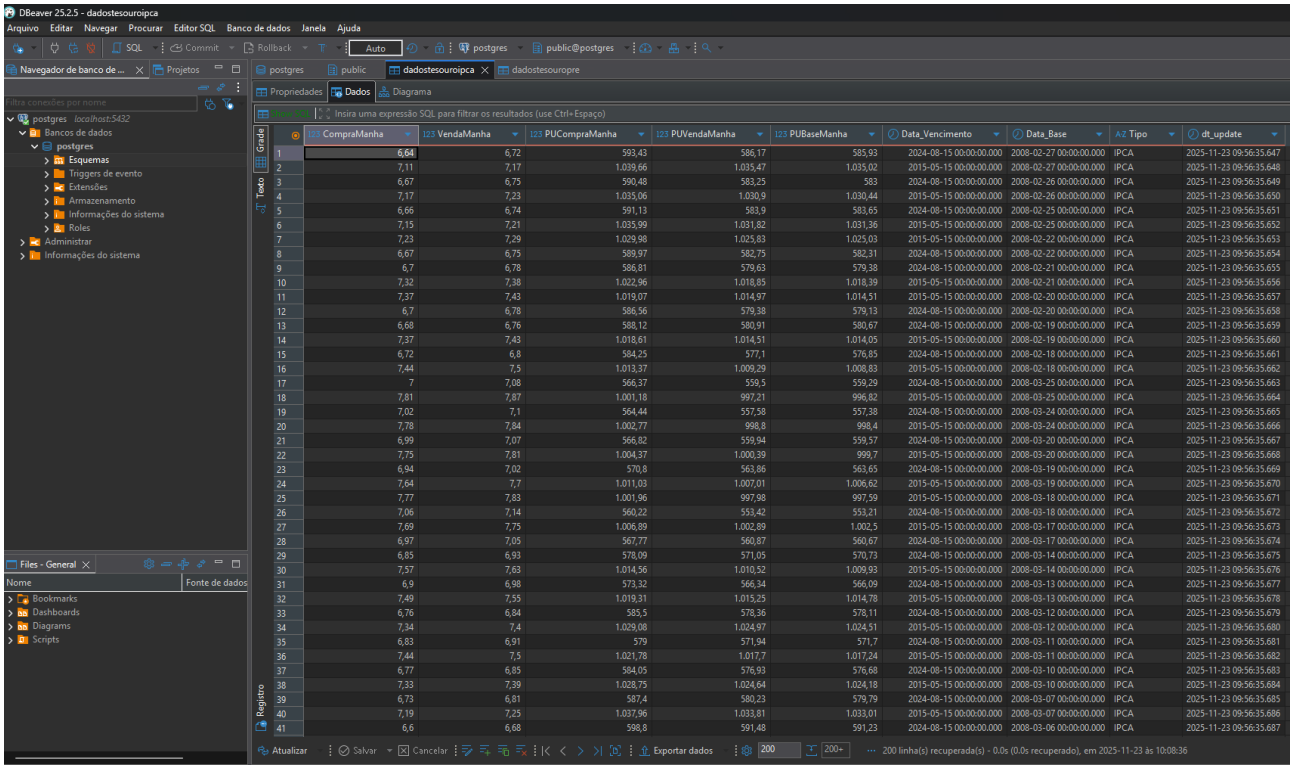
S3/Gold IPCA (analytics)



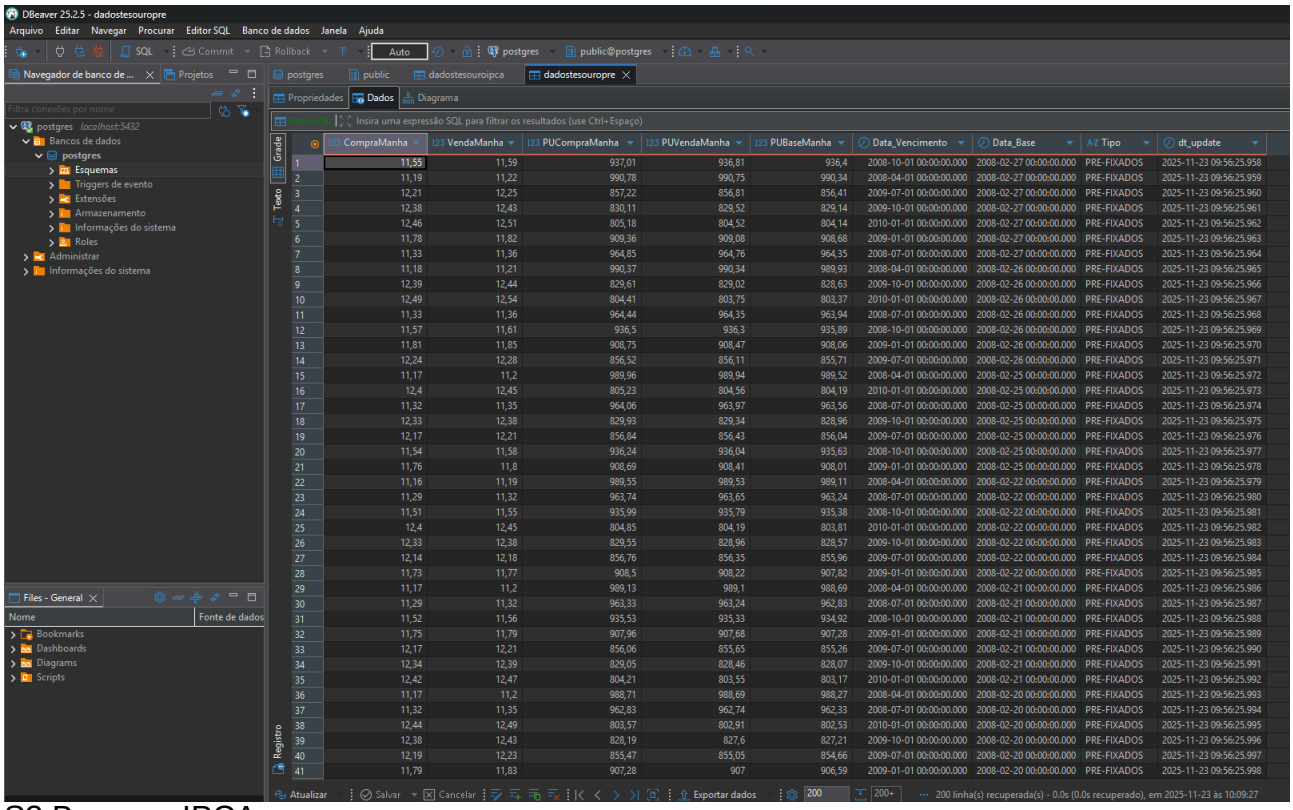
S3/Gold PRE (analytics)



Consulta Postgres dadostesouroipca

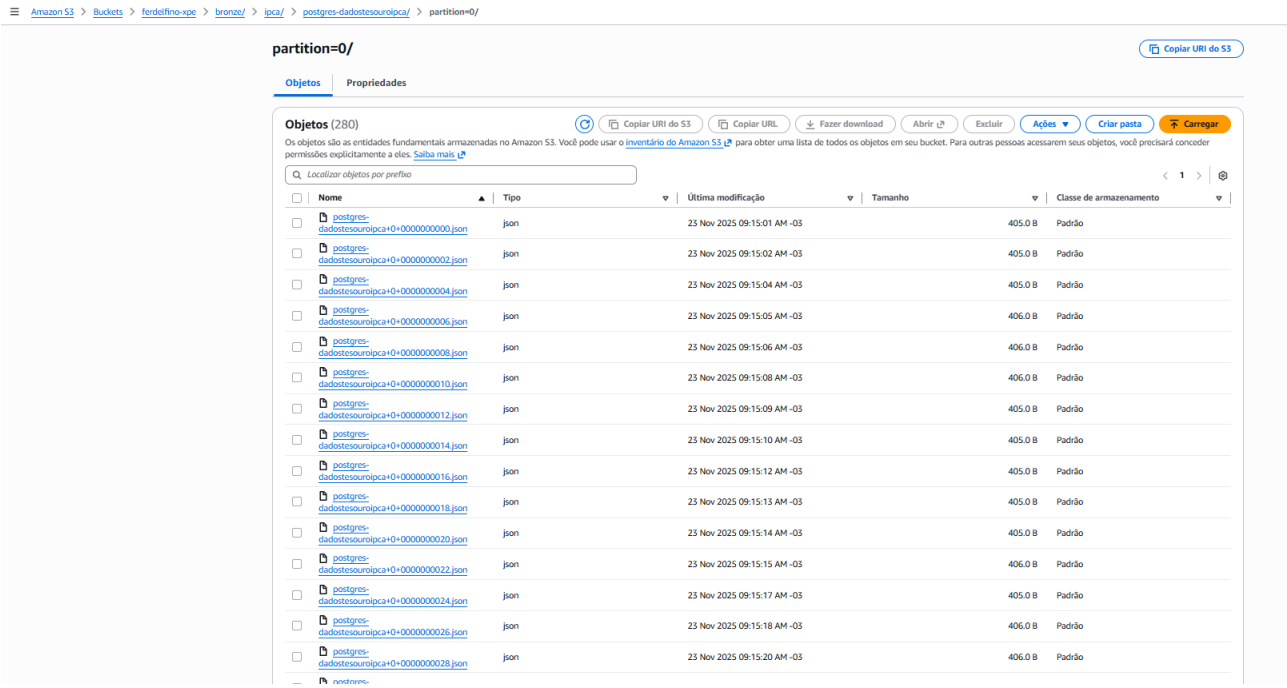


Consulta Postgres dadostesouropre



	123 CompraManha	123 VendaManha	123 PUCompraManha	123 PUVendaManha	123 PUBaseManha	Data_Vencimento	Data_Base	Tipo	dt_update
1	11,55	11,59	937,01	936,81	936,4	2008-10-01 00:00:00.000	2008-02-27 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.958
2	11,19	11,22	990,78	990,75	990,34	2008-04-01 00:00:00.000	2008-02-27 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.959
3	12,21	12,25	857,22	856,81	856,41	2009-07-01 00:00:00.000	2008-02-27 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.960
4	12,38	12,43	830,11	829,52	829,14	2009-10-01 00:00:00.000	2008-02-27 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.961
5	12,46	12,51	805,18	804,52	804,14	2010-01-01 00:00:00.000	2008-02-27 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.962
6	11,78	11,82	909,36	909,08	908,68	2009-01-01 00:00:00.000	2008-02-27 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.963
7	11,33	11,36	964,85	964,76	964,35	2008-07-01 00:00:00.000	2008-02-27 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.964
8	11,18	11,21	990,37	990,34	989,93	2008-04-01 00:00:00.000	2008-02-26 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.965
9	12,39	12,44	829,61	829,02	828,63	2009-10-01 00:00:00.000	2008-02-26 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.966
10	12,49	12,54	804,41	803,75	803,37	2010-01-01 00:00:00.000	2008-02-26 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.967
11	11,33	11,36	964,44	964,35	963,94	2008-07-01 00:00:00.000	2008-02-26 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.968
12	11,57	11,61	936,5	936,3	935,89	2008-10-01 00:00:00.000	2008-02-26 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.969
13	11,81	11,85	908,75	908,47	908,06	2009-01-01 00:00:00.000	2008-02-26 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.970
14	12,24	12,28	856,52	856,11	855,71	2009-07-01 00:00:00.000	2008-02-26 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.971
15	11,17	11,2	989,96	989,94	989,52	2008-04-01 00:00:00.000	2008-02-25 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.972
16	12,4	12,45	805,23	804,56	804,19	2010-01-01 00:00:00.000	2008-02-25 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.973
17	11,32	11,35	964,06	963,97	963,56	2008-07-01 00:00:00.000	2008-02-25 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.974
18	12,33	12,38	829,93	829,34	828,96	2009-10-01 00:00:00.000	2008-02-25 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.975
19	12,17	12,21	856,84	856,43	856,04	2009-07-01 00:00:00.000	2008-02-25 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.976
20	11,54	11,58	936,24	936,04	935,63	2008-10-01 00:00:00.000	2008-02-25 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.977
21	11,76	11,8	908,69	908,41	908,01	2009-01-01 00:00:00.000	2008-02-25 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.978
22	11,16	11,19	989,55	989,53	989,11	2008-04-01 00:00:00.000	2008-02-22 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.979
23	11,29	11,32	963,74	963,65	963,24	2008-07-01 00:00:00.000	2008-02-22 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.980
24	11,51	11,55	935,99	935,79	935,38	2008-10-01 00:00:00.000	2008-02-22 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.981
25	12,4	12,45	804,85	804,19	803,81	2010-01-01 00:00:00.000	2008-02-22 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.982
26	12,33	12,38	829,55	828,96	828,57	2009-10-01 00:00:00.000	2008-02-22 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.983
27	12,14	12,18	856,76	856,35	855,96	2009-07-01 00:00:00.000	2008-02-22 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.984
28	11,73	11,77	908,5	908,22	907,82	2009-01-01 00:00:00.000	2008-02-22 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.985
29	11,17	11,2	989,13	989,1	988,69	2008-04-01 00:00:00.000	2008-02-21 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.986
30	11,29	11,32	963,33	963,24	962,83	2008-07-01 00:00:00.000	2008-02-21 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.987
31	11,52	11,56	935,53	935,33	934,92	2008-10-01 00:00:00.000	2008-02-21 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.988
32	11,75	11,79	907,96	907,68	907,28	2009-01-01 00:00:00.000	2008-02-21 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.989
33	12,17	12,21	856,06	855,65	855,26	2009-07-01 00:00:00.000	2008-02-21 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.990
34	12,34	12,39	829,05	828,46	828,07	2009-10-01 00:00:00.000	2008-02-21 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.991
35	12,42	12,47	804,21	803,55	803,17	2010-01-01 00:00:00.000	2008-02-21 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.992
36	11,17	11,2	988,71	988,69	988,27	2008-04-01 00:00:00.000	2008-02-20 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.993
37	11,32	11,35	962,83	962,74	962,33	2008-07-01 00:00:00.000	2008-02-20 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.994
38	12,44	12,49	803,57	802,91	802,53	2010-01-01 00:00:00.000	2008-02-20 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.995
39	12,38	12,43	828,19	827,6	827,21	2009-10-01 00:00:00.000	2008-02-20 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.996
40	12,19	12,23	855,47	855,05	854,66	2009-07-01 00:00:00.000	2008-02-20 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.997
41	11,79	11,83	907,28	907	906,59	2009-01-01 00:00:00.000	2008-02-20 00:00:00.000	PRE-FIXADOS	2025-11-23 09:56:25.998

S3 Bronze - IPCA



Nome	Tipo	Última modificação	Tamanho	Classe de armazenamento
postgres-dadostesouroipca+0+0000000000.json	json	23 Nov 2025 09:15:01 AM -03	405.0 B	Padrão
postgres-dadostesouroipca+0+0000000002.json	json	23 Nov 2025 09:15:02 AM -03	405.0 B	Padrão
postgres-dadostesouroipca+0+0000000004.json	json	23 Nov 2025 09:15:04 AM -03	405.0 B	Padrão
postgres-dadostesouroipca+0+0000000006.json	json	23 Nov 2025 09:15:05 AM -03	406.0 B	Padrão
postgres-dadostesouroipca+0+0000000008.json	json	23 Nov 2025 09:15:06 AM -03	406.0 B	Padrão
postgres-dadostesouroipca+0+0000000010.json	json	23 Nov 2025 09:15:08 AM -03	406.0 B	Padrão
postgres-dadostesouroipca+0+0000000012.json	json	23 Nov 2025 09:15:09 AM -03	405.0 B	Padrão
postgres-dadostesouroipca+0+0000000014.json	json	23 Nov 2025 09:15:10 AM -03	405.0 B	Padrão
postgres-dadostesouroipca+0+0000000016.json	json	23 Nov 2025 09:15:12 AM -03	405.0 B	Padrão
postgres-dadostesouroipca+0+0000000018.json	json	23 Nov 2025 09:15:13 AM -03	405.0 B	Padrão
postgres-dadostesouroipca+0+0000000020.json	json	23 Nov 2025 09:15:14 AM -03	405.0 B	Padrão
postgres-dadostesouroipca+0+0000000022.json	json	23 Nov 2025 09:15:15 AM -03	406.0 B	Padrão
postgres-dadostesouroipca+0+0000000024.json	json	23 Nov 2025 09:15:17 AM -03	405.0 B	Padrão
postgres-dadostesouroipca+0+0000000026.json	json	23 Nov 2025 09:15:18 AM -03	406.0 B	Padrão
postgres-dadostesouroipca+0+0000000028.json	json	23 Nov 2025 09:15:20 AM -03	406.0 B	Padrão

S3 Bronze - PRE

Amazon S3 > Buckets > ferdelfino-xpe > bronze/ > pre/ > postgres-dadostesouropre/ > partition=0/

partition=0/

Copiar URI do S3

Objetos | Propriedades

Objetos (256)

Copiar URI do S3 | Copiar URL | Fazer download | Abrir | Excluir | Ações | Criar pasta | Carregar

Os objetos são as entidades fundamentais armazenadas no Amazon S3. Você pode usar o inventário do Amazon S3 para obter uma lista de todos os objetos em seu bucket. Para outras pessoas acessarem seus objetos, você precisará conceder permissões explicitamente a eles. Saiba mais

Localizar objetos por prefixo

Nome	Tipo	Última modificação	Tamanho	Classe de armazenamento
postgres-dadostesouropre+0+0000000000.json	json	23 Nov 2025 09:14:46 AM -03	424.0 B	Padrão
postgres-dadostesouropre+0+0000000002.json	json	23 Nov 2025 09:14:47 AM -03	424.0 B	Padrão
postgres-dadostesouropre+0+0000000004.json	json	23 Nov 2025 09:14:49 AM -03	424.0 B	Padrão
postgres-dadostesouropre+0+0000000006.json	json	23 Nov 2025 09:14:50 AM -03	422.0 B	Padrão
postgres-dadostesouropre+0+0000000008.json	json	23 Nov 2025 09:14:51 AM -03	423.0 B	Padrão
postgres-dadostesouropre+0+0000000010.json	json	23 Nov 2025 09:14:53 AM -03	424.0 B	Padrão
postgres-dadostesouropre+0+0000000012.json	json	23 Nov 2025 09:36:09 AM -03	423.0 B	Padrão
postgres-dadostesouropre+0+0000000014.json	json	23 Nov 2025 09:36:10 AM -03	424.0 B	Padrão
postgres-dadostesouropre+0+0000000016.json	json	23 Nov 2025 09:36:11 AM -03	423.0 B	Padrão
postgres-dadostesouropre+0+0000000018.json	json	23 Nov 2025 09:36:12 AM -03	424.0 B	Padrão
postgres-dadostesouropre+0+0000000020.json	json	23 Nov 2025 09:36:14 AM -03	424.0 B	Padrão
postgres-dadostesouropre+0+0000000022.json	json	23 Nov 2025 09:36:15 AM -03	421.0 B	Padrão
postgres-dadostesouropre+0+0000000024.json	json	23 Nov 2025 09:36:16 AM -03	424.0 B	Padrão
postgres-dadostesouropre+0+0000000026.json	json	23 Nov 2025 09:36:17 AM -03	423.0 B	Padrão
postgres-dadostesouropre+0+0000000028.json	json	23 Nov 2025 09:36:18 AM -03	422.0 B	Padrão
postgres-	json	23 Nov 2025 09:36:20 AM -03	422.0 B	Padrão

Containers Confluent/Connect/DB (docker ps)

docker desktop — Search for images, containers, volumes, extensions and more... Ctrl+K

Containers | Images | Volumes | Builds | Docker Scout | Extensions | Add Extensions

Containers

Give feedback

Container CPU usage: 7.04% / 800% (8 CPUs available)

Container memory usage: 3.24GB / 6.57GB

Show charts

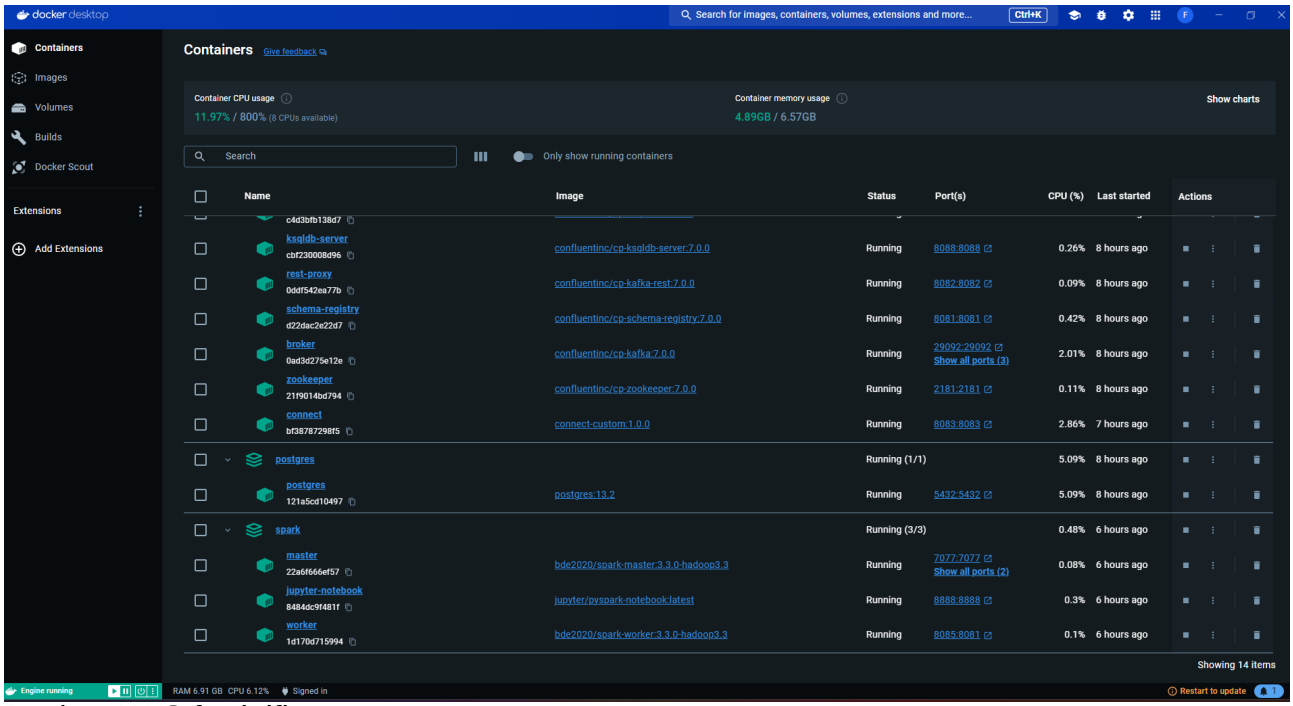
Search | Only show running containers

Name	Image	Status	Port(s)	CPU (%)	Last started	Actions
postgres		Running (1/1)		0.3%	44 minutes ago	
postgres 121a5cd10497	postgres:13.2	Running	5432:5432	0.3%	44 minutes ago	
prova		Running (7/7)		6.7%	36 minutes ago	
zookeeper 21f9014bd794	confluentinc/cp-zookeeper:7.0.0	Running	2181:2181	0.12%	36 minutes ago	
broker 0ad3d275e12e	confluentinc/cp-kafka:7.0.0	Running	29092:29092 Show all ports (3)	1.99%	36 minutes ago	
connect d3e41ea8323c	connect-custom:1.0.0	Running	8083:8083	3.47%	36 minutes ago	
schema-registry d22dac2e22d7	confluentinc/cp-schema-registry:7.0.0	Running	8081:8081	0.63%	36 minutes ago	
rest-proxy 0ddf542ea77b	confluentinc/cp-kafka-rest:7.0.0	Running	8082:8082	0.1%	36 minutes ago	
ksqldb-cli c4d3bfb138d7	confluentinc/cp-ksqldb-cli:7.0.0	Running		0%	36 minutes ago	
ksqldb-server cbf230008d96	confluentinc/cp-ksqldb-server:7.0.0	Running	8088:8088	0.39%	36 minutes ago	

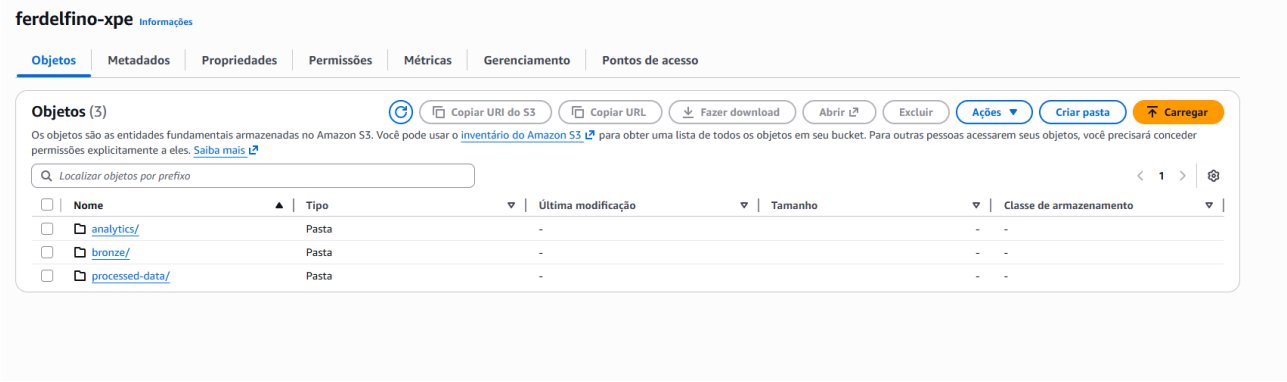
Showing 10 items

Engine running | RAM 6.93 GB CPU 0.50% | Signed in | New version available

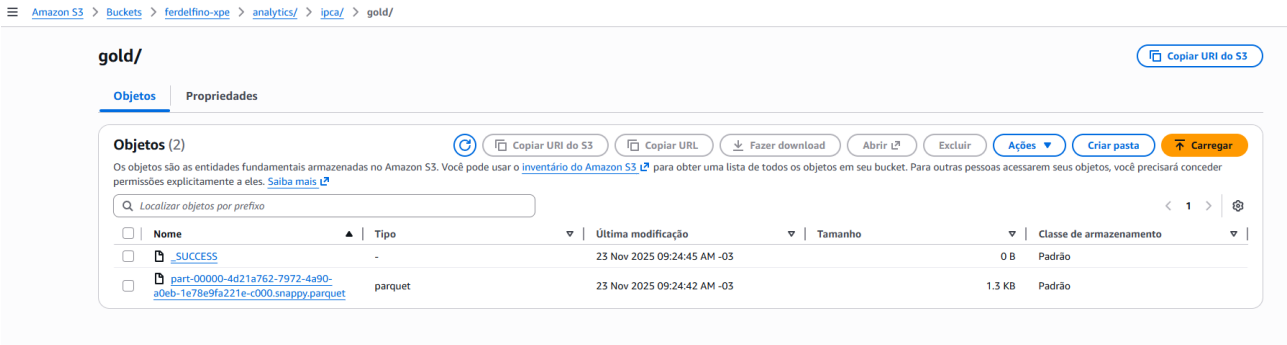
Containers Kafka/Spark/Connect (docker ps)



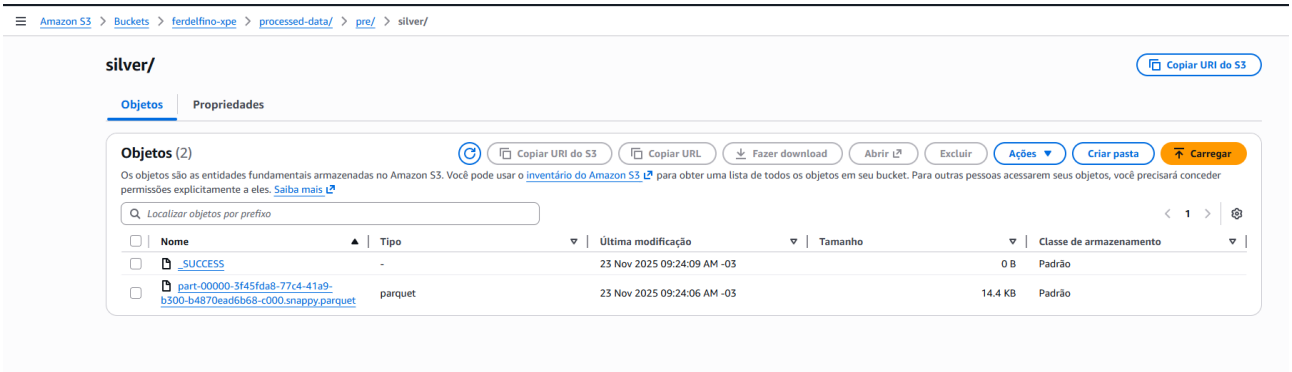
Bucket AWS ferdelfino-xpe



Spark Gold IPCA/PRE (visao)



Spark Silver IPCA/PRE (visao)



Jupyter Notebook executado (Spark)

Launcher | etl-spark.ipynb

Code

[21]:
```python
import os
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, from_unixtime, avg, count
from dotenv import load_dotenv

load_dotenv()  # carrega AWS_* do ambiente

spark = (
    SparkSession.builder
    .appName("S3 Access Example")
    .master("local[*]")
    .config("spark.jars", "/opt/spark/jars/hadoop-aws-3.3.4.jar,/opt/spark/jars/aws-java-sdk-bundle-1.12.262.jar")
    .config("spark.hadoop.fs.s3a.impl", "org.apache.hadoop.fs.s3a.S3AFileSystem")
    .config("spark.hadoop.fs.s3a.aws.credentials.provider", "com.amazonaws.auth.EnvironmentVariableCredentialsProvider")
    .config("spark.hadoop.fs.s3a.path.style.access", "true")
    .config("spark.hadoop.fs.s3a.connection.ssl.enabled", "true")
    .getOrCreate()
)

aws_access_key = os.getenv("AWS_ACCESS_KEY_ID")
aws_secret_key = os.getenv("AWS_SECRET_ACCESS_KEY")
aws_region = os.getenv("AWS_REGION", "us-east-2")

spark._jsc.hadoopConfiguration().set("fs.s3a.access.key", aws_access_key)
spark._jsc.hadoopConfiguration().set("fs.s3a.secret.key", aws_secret_key)
spark._jsc.hadoopConfiguration().set("fs.s3a.endpoint", f"s3.{aws_region}.amazonaws.com")

print("Spark session ready")
```

```
Spark session ready
```

[22]:
```python
bronze_ipca = "s3a://ferdelfino-xpe/bronze/ipca/"
bronze_pre = "s3a://ferdelfino-xpe/bronze/pre/"

silver_ipca = "s3a://ferdelfino-xpe/processed-data/ipca/silver/"
silver_pre = "s3a://ferdelfino-xpe/processed-data/pre/silver/"

gold_ipca = "s3a://ferdelfino-xpe/analytics/ipca/gold/"
gold_pre = "s3a://ferdelfino-xpe/analytics/pre/gold/"

for p in [bronze_ipca, bronze_pre, silver_ipca, silver_pre, gold_ipca, gold_pre]:
    print(p)
```

```
s3a://ferdelfino-xpe/bronze/ipca/
s3a://ferdelfino-xpe/bronze/pre/
s3a://ferdelfino-xpe/processed-data/ipca/silver/
```

Jupyter Notebook - contagens/paths

File Edit View Run Kernel Tabs Settings Help

Filter files by name

/ work /

Name	Last Modified
-.py	5 hours ago
etl-spark.ip...	40 seconds ago
etl-spark.ip...	6 hours ago

Launcher | etl-spark.ipynb

Code

[22]:
```
bronze_ipca = "s3a://ferdelfino-xpe/bronze/ipca/"
bronze_pre = "s3a://ferdelfino-xpe/bronze/pre/"

silver_ipca = "s3a://ferdelfino-xpe/processed-data/ipca/silver/"
silver_pre = "s3a://ferdelfino-xpe/processed-data/pre/silver/"

gold_ipca = "s3a://ferdelfino-xpe/analytics/ipca/gold/"
gold_pre = "s3a://ferdelfino-xpe/analytics/pre/gold/"

for p in [bronze_ipca, bronze_pre, silver_ipca, silver_pre, gold_ipca, gold_pre]:
    print(p)
```

```
s3a://ferdelfino-xpe/bronze/ipca/
s3a://ferdelfino-xpe/bronze/pre/
s3a://ferdelfino-xpe/processed-data/ipca/silver/
s3a://ferdelfino-xpe/processed-data/pre/silver/
s3a://ferdelfino-xpe/analytics/ipca/gold/
s3a://ferdelfino-xpe/analytics/pre/gold/
```

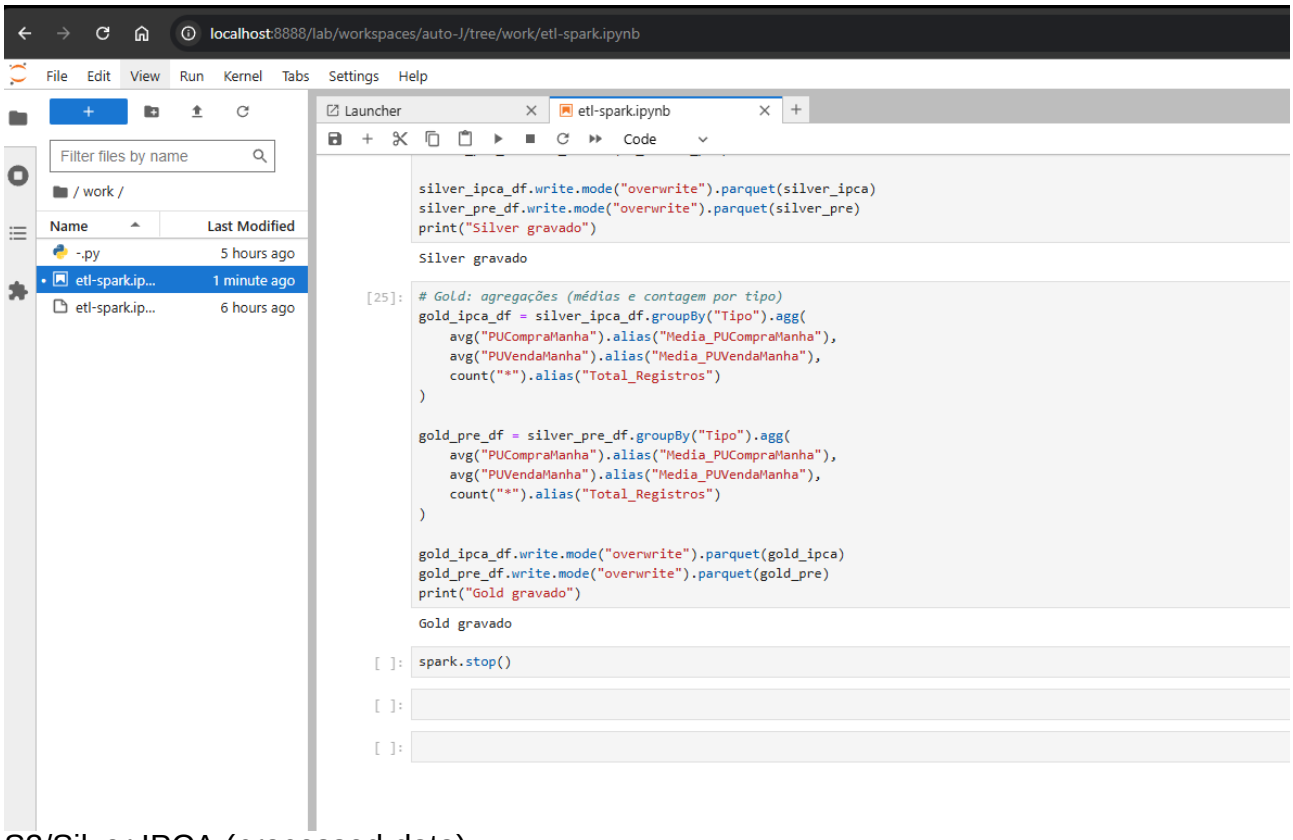
[23]:
```
# Ler Bronze
df_bronze_ipca = spark.read.json(bronze_ipca)
df_bronze_pre = spark.read.json(bronze_pre)
print("Bronze IPCA count:", df_bronze_ipca.count())
print("Bronze PRE count:", df_bronze_pre.count())
```

```
Bronze IPCA count: 560
Bronze PRE count: 512
```

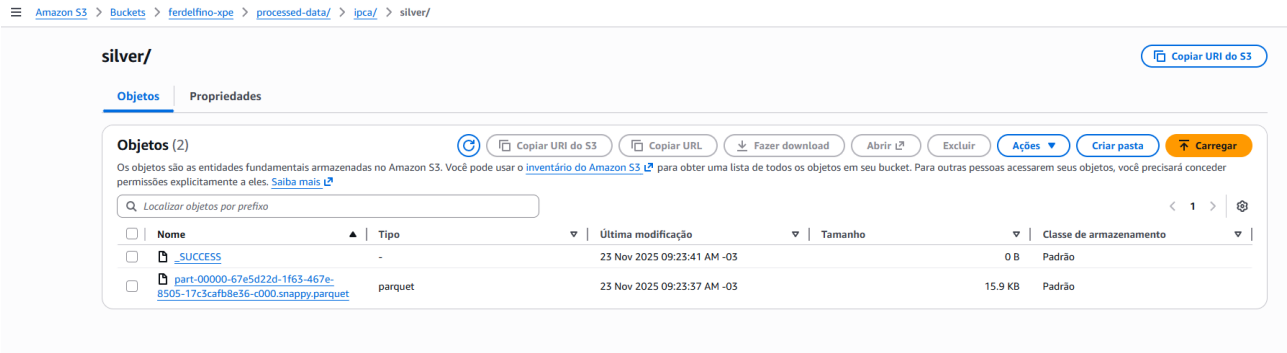
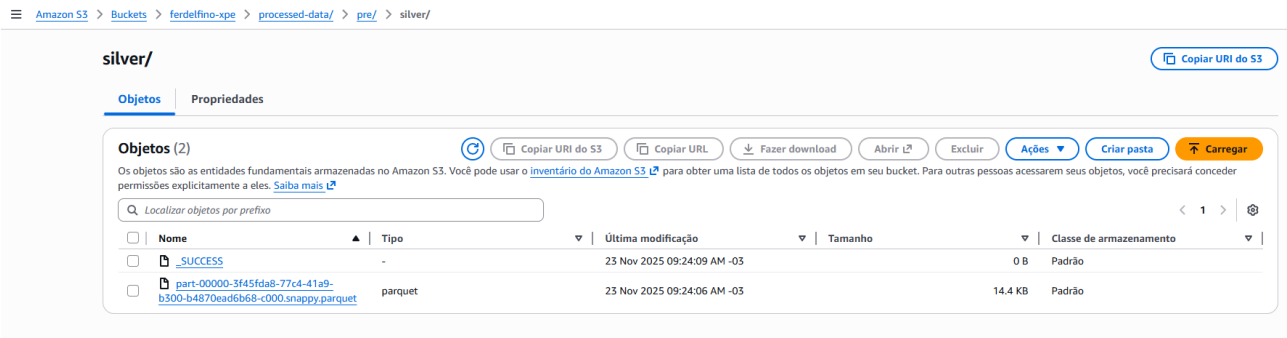
[24]:
```
# Silver: deduplicar, converter timestamps, tratar nulos
def to_silver(df):
    df_s = df.dropDuplicates()
    df_s = df_s.withColumn("Data_Vencimento", from_unixtime(col("Data_Vencimento")/1000, "yyyy-MM-dd")) \
               .withColumn("Data_Base", from_unixtime(col("Data_Base")/1000, "yyyy-MM-dd")) \
               .withColumn("dt_update", from_unixtime(col("dt_update")/1000, "yyyy-MM-dd HH:mm:ss"))
    df_s = df_s.fillna({"PUCompraManha": 0, "PUVendaManha": 0, "PUBaseManha": 0})
    return df_s

silver_ipca_df = to_silver(df_bronze_ipca)
silver_pre_df = to_silver(df_bronze_pre)

silver_ipca_df.write.mode("overwrite").parquet(silver_ipca)
silver_pre_df.write.mode("overwrite").parquet(silver_pre)
print("Silver gravado")
```

```
Silver gravado
```

Jupyter Notebook - Silver/Gold



S3/Silver IPCA (processed-data)



S3/Silver PRE (processed-data)