

**Untitled File 2026-01-16 12\_23\_33.py**

```
base = "dbfs:/Volumes/workspace/default/teste/raw_out/raw_out"
display(dbutils.fs.ls(base))

> See performance \(2\)
```

**Table**

path	name	size	modificationTime
dbfs:/Volumes/workspace/default/teste/raw_out/raw_out/overall_manifest.json	overall_manifest.json	3098	1768577576000
dbfs:/Volumes/workspace/default/teste/raw_out/raw_out/source=precodahora/	source=precodahor...	0	176857777324

2 rows | 1.22s runtime

Refreshed 14 minutes ago

---

**Untitled File 2026-01-16 12\_24\_15.py**

```
def list_files_recursive(path: str, suffix: str = ""):
    out = []
    for item in dbutils.fs.ls(path):
        if item.isDir():
            out.extend(list_files_recursive(item.path, suffix=suffix))
        else:
            if (not suffix) or item.path.endswith(suffix):
                out.append(item.path)
    return out

base = "dbfs:/Volumes/workspace/default/teste/raw_out/raw_out"
jsonl_files = list_files_recursive(base, suffix=".jsonl")

print("Found:", len(jsonl_files))
for p in jsonl_files[:10]:
    print(p)

> See performance \(14\)
```

Found: 4

```
dbfs:/Volumes/workspace/default/teste/raw_out/raw_out/source=precodahora/amp=DIESEL/dt=2026-01-16/run_id=7766e40b-9166-4271-9644-858a0b8e4049/data.jsonl
dbfs:/Volumes/workspace/default/teste/raw_out/raw_out/source=precodahora/amp=ETANOL/dt=2026-01-16/run_id=cad4d1ef-4b31-4757-8086-5661c9932974/data.jsonl
dbfs:/Volumes/workspace/default/teste/raw_out/raw_out/source=precodahora/amp=GASOLINA/dt=2026-01-16/run_id=d793dead-5959-4d67-adb1-a262bb2dafa/data.jsonl
dbfs:/Volumes/workspace/default/teste/raw_out/raw_out/source=precodahora/amp=GNV/dt=2026-01-16/run_id=7d49c175-93cc-49fd-9e8a-10e6af1b1ee/data.jsonl
```

---

**Untitled Notebook 2026-01-16 12:30:27**

```
df_raw = spark.read.json(jsonl_files)
display(df_raw)

> See performance \(1\)
```

**df\_raw:** pyspark.sql.connect.DataFrame = [amp: string, collected\_at\_utc: string ... 4 more fields]

**Table**

amp	collected_at_utc	query	raw
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"CENTRO","cep":43970000,"cnpi":33337122007130,"codMi...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"NOVA BRASILIA","cep":43810160,"cnpi":2805889001009,"cod...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"CONCEIÇÃO","cep":44066280,"cnpi":31421491000178,"cod...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"MATARIPE","cep":43900000,"cnpi":3427423308006,"cod...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"MATARIPE","cep":43900000,"cnpi":1125282001198,"codM...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"MATARIPE","cep":43900000,"cnpi":1125282001198,"codM...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"MATARIPE","cep":43900000,"cnpi":33453598013969,"cod...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"NOVA BRASILIA","cep":43810160,"cnpi":2805889001009,"cod...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"CONCEIÇÃO","cep":44066280,"cnpi":31421491000178,"cod...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"MATARIPE","cep":43900000,"cnpi":3427423308006,"cod...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"MATARIPE","cep":43900000,"cnpi":1125282001198,"codM...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"MATARIPE","cep":43900000,"cnpi":1125282001198,"codM...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"CENTRO","cep":44530000,"cnpi":40495341000108,"codMi...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"BATATAN","cep":44400000,"cnpi":4296047000103,"codMu...
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"CAIPE DE CIMA","cep":43900000,"cnpi":23314594002405,
DIESEL	2026-01-16T15:09:32+00:00	("horas":72,"latitude":-12.97111,"longitude":-38.51083,"ordenar":"preco.asc","pagina":1,"raio":...	{"estabelecimento":{"bairro":"DISTRITO INDUSTRIAL - DISTRITO INDUSTRIAL","cep":438...

100 rows | 4.20s runtime

Refreshed 14 minutes ago

---

**Untitled File 2026-01-16 12\_24\_15.py**

```
spark.sql("CREATE DATABASE IF NOT EXISTS precodahora")

(df_raw.write
 .format("delta")
 .mode("append")
 .saveAsTable("precodahora.bronze_raw"))

> See performance \(2\)
```

---

**Untitled Notebook 2026-01-16 12:30:27**

```
display(spark.sql("""
SELECT amp, count(*) AS rows
FROM precodahora.bronze_raw
GROUP BY amp
ORDER BY rows DESC
"""))

> See performance \(1\)
```

**Table**

amp	rows
GASOLINA	25
ETANOL	25
DIESEL	25
GNV	25

4 rows | 1.33s runtime

Refreshed 13 minutes ago

```

▶ ✓ 12:40 PM (4s) 6
from pyspark.sql import functions as F
df = spark.table("precodahora.bronze_raw")

df_silver = (
    df
    .select(
        F.col("run_id"),
        F.col("collected_at_utc"),
        F.col("source"),
        F.col("amp").alias("fuel_type"),

        # centro da busca (consulta)
        F.col("query.latitude").cast("double").alias("q_lat"),
        F.col("query.longitude").cast("double").alias("q_lng"),
        F.col("query.radius").cast("double").alias("q_radius_km"),
        F.col("query.hours").cast("int").alias("q_hours"),
        F.col("query.pagina").cast("int").alias("q_page"),

        # dados do produto (raw)
        F.col("raw_produto.codProduto").alias("product_code"),
        F.col("raw_produto.amp").alias("product_amp"),
        F.col("raw_produto.descricao").alias("product_desc"),
        F.col("raw_produto.unidade").alias("unit"),
        F.col("raw_produto.precoUnitario").cast("double").alias("price_unit"),
        F.col("raw_produto.precoLiquido").cast("double").alias("price_net"),
        F.col("raw_produto.precoBruto").cast("double").alias("price_gross"),
        F.col("raw_produto.desconto").cast("double").alias("discount"),
        F.to_timestamp(F.col("raw_produto.data")).alias("price_ts"),
    )
    .select(
        F.col("raw_estabelecimento.cnpj").alias("cnpj"),
        F.col("raw_estabelecimento.nomeEstabelecimento").alias("station_name"),
        F.col("raw_estabelecimento.municipio").alias("city"),
        F.col("raw_estabelecimento.uf").alias("uf"),
        F.col("raw_estabelecimento.bairro").alias("district"),
        F.col("raw_estabelecimento.endereco").alias("street"),
        F.col("raw_estabelecimento.endNumero").alias("number"),
        F.col("raw_estabelecimento.cep").alias("zip_code"),
        F.col("raw_estabelecimento.latitude").cast("double").alias("lat"),
        F.col("raw_estabelecimento.longitude").cast("double").alias("lng"),
        F.col("raw_estabelecimento.distancia").cast("double").alias("distance_km"),
    )
)

spark.sql("CREATE DATABASE IF NOT EXISTS precodahora")

(df_silver.write
    .format("delta")
    .mode("overwrite")
    .saveAsTable("precodahora.silver_prices"))

> ⓘ See performance (2)
> df: pyspark.sql.connect.DataFrame = [amp: string, collected_at_utc: string ... 4 more fields]
> df_silver: pyspark.sql.connect.DataFrame = [run_id: string, collected_at_utc string ... 27 more fields]

```

```

▶ ✓ 12:41 PM (1s) 7
display(spark.sql("""
SELECT fuel_type, MIN(price_unit) AS min_price, MAX(price_unit) AS max_price, COUNT(*) AS rows
FROM precodahora.silver_prices
GROUP BY fuel_type
ORDER BY fuel_type
"""))

> ⓘ See performance (1)

Table + 
-----+-----+-----+-----+
# fuel_type | 1.2 min_price | 1.2 max_price | 1.2 rows |
-----+-----+-----+-----+
1 DIESEL | 5.0532 | 6.19 | 25 |
2 ETANOL | 3.79 | 4.79 | 25 |
3 GASOLINA | 5.29056 | 6.79 | 25 |
4 GNV | 3.87 | 4.04 | 25 |
-----+-----+-----+-----+
4 rows | 1.16s runtime | Refreshed 11 minutes ago

```

```

▶ ✓ 12:42 PM (1s) 8
%sql
CREATE OR REPLACE VIEW precodahora.gold_latest AS
SELECT *
FROM (
    SELECT *,
        ROW_NUMBER() OVER (
            PARTITION BY fuel_type, cnpj, product_code
            ORDER BY price_ts DESC, collected_at_utc DESC
        ) AS rn
    FROM precodahora.silver_prices
) t
WHERE rn = 1;
> ⓘ See performance (1)
OK
Optimize

```

```

▶ ✓ 12:44 PM (1s) 9
%sql
CREATE OR REPLACE VIEW precodahora.gold_latest AS
SELECT
    run_id,
    collected_at_utc,
    source,
    fuel_type,
    q_lat, q_lng, q_radius_km, q_hours, q_page,
    product_code, product_amp, product_desc, unit,
    price_unit, price_net, price_gross, discount,
    price_ts,
    cnpj, station_name, city, uf, district, street, number, zip_code,
    lat, lng, distance_km
FROM (
    SELECT *
    ROW_NUMBER() OVER (
        PARTITION BY fuel_type, cnpj, product_code
        ORDER BY price_ts DESC, collected_at_utc DESC
    )
)
```

```

    ) AS rn
  FROM precodahora.silver_prices
) t
WHERE rn = 1;

> See performance \(1\)
OK

```

Optimize

[See performance \(1\)](#)

[\\_sqldf: pyspark.sql.connect.DataFrame](#) = [fuel\_type: string, rows: long]

Table	+
<a href="#">fuel_type</a>	<a href="#">rows</a>
1 DIESEL	25
2 ETANOL	25
3 GASOLINA	25
4 GNV	25

[↓](#) [▼](#) 4 rows | 1.33s runtime

Refreshed 8 minutes ago

1 This result is stored as `_sqldf` and can be used in other Python and SQL cells.

12:45 PM (1s)

```
%sql
SELECT fuel_type, station_name, district, city, uf, price_unit, price_ts, distance_km
FROM precodahora.gold_latest
WHERE city = 'SALVADOR'
ORDER BY fuel_type, price_unit ASC
LIMIT 40;
```

[See performance \(1\)](#)

fuel_type	station_name	district	city	uf	price_unit	price_ts	distance_km
1 DIESEL	EXTRELA TRANSPORTADOR REVENDEDOR RETALHISTA	PAU MIUDO	SALVADOR	BA	5.27	2026-01-16T13:57:00.000+00...	3
2 ETANOL	POSTO ZEUS	ACUPE DE BROTHAS	SALVADOR	BA	4.63	2026-01-16T14:53:13:000+00...	1
3 ETANOL	POSTO TRANSICAO	RETIRO	SALVADOR	BA	4.63	2026-01-16T14:53:34.000+00...	4
4 ETANOL	POSTO BONJOUR	JARDIM DAS MARGARIDAS	SALVADOR	BA	4.63	2026-01-16T14:50:59.000+00...	18
5 GASOLINA	POSTO ALELUIA	SAO CRISTOVAO	SALVADOR	BA	6.13	2026-01-16T14:47:14.000+00...	17
6 GASOLINA	POSTO 4	MATA ESCURA	SALVADOR	BA	6.13	2026-01-16T14:53:09.000+00...	7
7 GASOLINA	HIPER POSTO	CAMINHO DAS ARVORES	SALVADOR	BA	6.13	2026-01-16T14:50:58.000+00...	6
8 GNV	ENTREPOSTO AMARALINA	AMARALINA	SALVADOR	BA	3.87	2026-01-16T14:52:1.000+00...	1
9 GNV	POSTO ITAICE	BROTAS	SALVADOR	BA	3.89	2026-01-16T14:42:56.000+00...	3
10 GNV	SOGAS	ACUPE DE BROTHAS	SALVADOR	BA	3.89	2026-01-16T14:49:03.000+00...	2
11 GNV	POSTO CHAMINE	RIO VERMELHO	SALVADOR	BA	3.89	2026-01-16T14:43:21.000+00...	1
12 GNV	POSTO MACRO	PIRAJA	SALVADOR	BA	3.97	2026-01-16T12:28:30.000+00...	7
13 GNV	POSTO FLAMINGO	STIEP	SALVADOR	BA	3.99	2026-01-16T14:50:42.000+00...	1
14 GNV	POSTO SALVADOR 324	AGUAS CLARAS	SALVADOR	BA	3.99	2026-01-16T14:53:07.000+00...	13

[↓](#) [▼](#) 23 rows | 1.27s runtime

Refreshed 7 minutes ago

1 This result is stored as `_sqldf` and can be used in other Python and SQL cells.

12:45 PM (1s)

```
%sql
CREATE OR REPLACE VIEW precodahora.gold_timeseries AS
SELECT
  fuel_type,
  city,
  uf,
  MIN(price_unit) AS min_price,
  AVG(price_unit) AS avg_price,
  MAX(price_unit) AS max_price,
  COUNT(*) AS samples
FROM precodahora.silver_prices
GROUP BY fuel_type, city, uf, DATE_TRUNC('hour', price_ts);
```

[See performance \(1\)](#)

OK

[Optimize](#)

12:45 PM (1s)

```
%sql
SELECT *
FROM precodahora.gold_timeseries
ORDER BY hour_ts DESC
LIMIT 50;
```

[See performance \(1\)](#)

[\\_sqldf: pyspark.sql.connect.DataFrame](#) = [fuel\_type: string, city: string ... 6 more fields]

fuel_type	city	uf	hour_ts	min_price	avg_price	max_price	samples
1 ETANOL	ITAPARICA	BA	2026-01-16T14:00:00.000+00...	4.29	4.29	4.29	1
2 ETANOL	VALENCA	BA	2026-01-16T14:00:00.000+00...	4.2	4.2	4.2	1
3 ETANOL	SAO FELIPE	BA	2026-01-16T14:00:00.000+00...	4.79	4.79	4.79	1
4 GASOLINA	SALVADOR	BA	2026-01-16T14:00:00.000+00...	6.13	6.13	6.13	3
5 ETANOL	CATU	BA	2026-01-16T14:00:00.000+00...	4.33	4.33	4.33	1
6 ETANOL	SANTO ANTONIO DE JESUS	BA	2026-01-16T14:00:00.000+00...	4.26	4.26	4.26	1
7 GASOLINA	FEIRA DE SANTANA	BA	2026-01-16T14:00:00.000+00...	6.13	6.13	6.13	2
8 GASOLINA	NAZARE	BA	2026-01-16T14:00:00.000+00...	6.59	6.59	6.59	1
9 ETANOL	VERA CRUZ	BA	2026-01-16T14:00:00.000+00...	3.88	4.085	4.29	2
10 DIESEL	SAPEACU	BA	2026-01-16T14:00:00.000+00...	5.89	5.89	5.89	1
11 ETANOL	FEIRA DE SANTANA	BA	2026-01-16T14:00:00.000+00...	4.63	4.63	4.63	1
13 GNV	AREIA BRONQUES	BA	2026-01-16T14:00:00.000+00...	3.88	3.88	3.88	1
14 GNV	CAMACARI	BA	2026-01-16T14:00:00.000+00...	4.04	4.04	4.04	4
15 ETANOL	SALVADOR	BA	2026-01-16T14:00:00.000+00...	4.63	4.63	4.63	3

[Optimize](#)

 50 rows | 1.20s runtime

Refreshed 7 minutes ago

ⓘ This result is stored as `_sqldf` and can be used in other [Python](#) and [SQL](#) cells.

[Shift+Enter] to run and move to next cell  
[Ctrl+Shift+P] to open the command palette  
[Esc H] to see all keyboard shortcuts