

Casas

March 4, 2025

```
[1]: +-----+
|      Casas      |
+-----+
| id_casa: INT    |
| cidade: STRING  |
| tipo: STRING    |
| area_m2: INT    |
| preco: DOUBLE   |
| quartos: INT    |
| casas_de_banho: INT |
+-----+
```

Cell In[1], line 1

```
+-----+
```

^

SyntaxError: invalid syntax

```
[2]: # Importar bibliotecas necessárias para o uso do PySpark
from pyspark.sql import SparkSession

# Criar sessão do Spark (ponto de entrada para utilizar o PySpark)
spark = SparkSession.builder \
    .appName("Casas_Analysis").config("spark.sql.catalogImplementation",
    ↪ "hive").enableHiveSupport().getOrCreate() # Criar ou obter uma sessão ativa
    ↪ do Spark

# Criar o banco de dados se não existir
spark.sql("CREATE DATABASE IF NOT EXISTS imobiliaria")

# Selecionar o banco de dados
spark.sql("USE imobiliaria")

# Criar a tabela permanente no Hive
spark.sql("""
```

```

CREATE TABLE IF NOT EXISTS casas (
    id_casa INT,
    cidade STRING,
    tipo STRING,
    area_m2 INT,
    preco DOUBLE,
    quartos INT,
    casas_de_banho INT
) USING PARQUET
""")

# Criar dados fictícios sobre casas
dados = [
    (1, 'Lisboa', 'Apartamento', 120, 500000, 3, 2),
    (2, 'Porto', 'Moradia', 200, 750000, 4, 3),
    (3, 'Coimbra', 'Apartamento', 90, 250000, 2, 1),
    (4, 'Faro', 'Moradia', 180, 600000, 4, 3),
    (5, 'Braga', 'Apartamento', 100, 300000, 3, 2)
]

# Definir os nomes das colunas para o DataFrame
colunas = ["id_casa", "cidade", "tipo", "area_m2", "preco", "quartos", "casas_de_banho"]

# Criar um DataFrame do Spark a partir dos dados
df = spark.createDataFrame(dados, colunas)

# Gravar os dados na tabela Hive
df.write.mode("overwrite").saveAsTable("imobiliaria.casas")

# Criar uma tabela temporária para permitir consultas SQL diretamente no DataFrame
df.createOrReplaceTempView("casas")

# Perguntas para os alunos:
# 1. Liste todas as casas do tipo "Apartamento".
# 2. Ordene as casas pelo preço em ordem decrescente.
# 3. Calcule a média de área por tipo de imóvel.
# 4. Identifique a cidade com a casa mais cara.
# 5. Filtre as casas que possuem pelo menos 3 quartos.

```

```

[3]: # a) Listar todas as casas do tipo "Apartamento"
spark.sql("""
    SELECT * FROM casas
    WHERE tipo = 'Apartamento'
""").show()

```

```

# b) Ordenar as casas pelo preço em ordem decrescente
spark.sql("""
    SELECT * FROM casas
    ORDER BY preco DESC
""").show()

# c) Calcular a média de área por tipo de imóvel
spark.sql("""
    SELECT tipo, AVG(area_m2) AS media_area
    FROM casas
    GROUP BY tipo
""").show()

# d) Identificar a cidade com a casa mais cara
spark.sql("""
    SELECT cidade, tipo, preco
    FROM casas
    ORDER BY preco DESC
    LIMIT 1
""").show()

# e) Filtrar as casas que possuem pelo menos 3 quartos
spark.sql("""
    SELECT * FROM casas
    WHERE quartos >= 3
""").show()

# Explicação:
# - `SparkSession.builder` cria uma sessão do Spark com suporte ao Hive.
# - `CREATE DATABASE IF NOT EXISTS` garante que o banco de dados "imobiliaria"
    ↳ exista.
# - `CREATE TABLE IF NOT EXISTS` cria uma tabela permanente no formato Parquet.
# - `createDataFrame(dados, colunas)` transforma uma lista de dados num
    ↳ DataFrame Spark.
# - `saveAsTable("imobiliaria.casas")` grava os dados de forma persistente.
# - `createOrReplaceTempView("casas")` cria uma tabela temporária para
    ↳ consultas SQL.
# - `sql()` permite executar consultas SQL diretamente no Spark.
# - `.show()` exibe os resultados no terminal.

# Com este código, os alunos podem aprender a criar bancos de dados, tabelas,
    ↳ manipular dados e executar consultas SQL com PySpark e Hive.

```

id_casa	cidade	tipo	area_m2	preco	quartos	casas_de_banho
1	Lisboa	Apartamento	120	500000	3	2

3	Coimbra	Apartamento	90	250000	2	1
5	Braga	Apartamento	100	300000	3	2

id_casa	cidade	tipo	area_m2	preco	quartos	casas_de_banho
2	Porto	Moradia	200	750000	4	3
4	Faro	Moradia	180	600000	4	3
1	Lisboa	Apartamento	120	500000	3	2
5	Braga	Apartamento	100	300000	3	2
3	Coimbra	Apartamento	90	250000	2	1

tipo	media_area
Apartamento	103.33333333333333
Moradia	190.0

cidade	tipo	preco
Porto	Moradia	750000

id_casa	cidade	tipo	area_m2	preco	quartos	casas_de_banho
1	Lisboa	Apartamento	120	500000	3	2
2	Porto	Moradia	200	750000	4	3
4	Faro	Moradia	180	600000	4	3
5	Braga	Apartamento	100	300000	3	2

[]: