Dev Potter:

Spark para Trouxas

Aurora Dev

A Primeira Centelha de Magia

Bem-vinda ao castelo do conhecimento, jovem feiticeira dos dados!Se você chegou até aqui, provavelmente já ouviu falar que o Spark é um feitiço poderoso capaz de processar milhões de linhas de dados em segundos — e que apenas os bruxos mais destemidos ousam domá-lo.

Mas não se preocupe. Neste livro, você vai aprender a usar o SparkSQL com Python passo a passo, sem precisar de uma varinha feita de grafite ou poções misteriosas.Cada capítulo será como uma aula em Hogwarts dos dados: prática, divertida e com feitiços que funcionam de verdade no seu computador.

Pegue sua varinha (ou melhor, seu notebook.ipynb) e vamos conjurar alguns códigos! ⚡

01

Lumos Initium

O feitiço que dá vida à sua sessão mágica

Antes de conjurar qualquer magia, precisamos invocar a chama inicial do Spark.Essa luz inicial é criada com o comando que todo mago dos dados precisa dominar: a SparkSession.

Ela é o seu portal de entrada — o momento em que o Spark desperta e começa a ouvir seus comandos.

Untitled-1

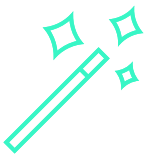
```python
from pyspark.sql import SparkSession

# Criando a sessão mágica ✨
spark = SparkSession.builder.appName(
    "Hogwarts dos Dados"
).getOrCreate()

# Conferindo se a magia foi invocada
print(spark)
```

✨ E pronto, você acabou de acender o Spark. A primeira magia foi feita com sucesso.

02

Wingardium

DataFrame



Fazendo Dados Flutuarem

Agora que o Spark está aceso, é hora de fazer seus dados levitarem!Os DataFrames são como criaturas mágicas que obedecem aos seus comandos: podem filtrar, voar, se transformar e até se multiplicar.

Um DataFrame é uma tabela com colunas e linhas, como no Excel, mas com poderes infinitamente maiores.

```
dados = [("Hermione", 98),
         ("Harry",    85),
         ("Ron",      76)]
colunas = ["Aluno", "Nota"]

df = spark.createDataFrame(dados, colunas)
df.show()
```

✨ Dica mágica extra ✨

O módulo .show() é o feitiço que você usará para mostrar as primeiras linhas do seu DataFrame.

03

Revelio SQL

Desvendando os Segredos Ocultos dos Dados

Agora que você tem seus dados flutuando em DataFrames, é hora de revelar seus segredos. Com o SparkSQL, você pode usar comandos tradicionais de SQL (como SELECT, WHERE e ORDER BY) de uma forma muito mais poderosa.

Basta registrar seu DataFrame como uma tabela temporária e depois lançar suas consultas SQL sobre ele. Veja um exemplo:

```
df.createOrReplaceTempView("alunos")

resultado = spark.sql("""
    SELECT
        Aluno,
        Nota
    FROM alunos
    WHERE Nota > 80
""")
resultado.show()
```

✨ E lá estão eles!
Apenas os alunos com notas acima de 80 — dignos de um feitiço bem lançado.

04

Expecto GroupBy

Unindo Forças dos Dados

Às vezes, o poder não está em um único dado, mas na união deles. O feitiço groupBy() é usado para reunir e combinar forças, calculando médias, totais, contagens e outras estatísticas encantadas.

Com groupBy() e agg(), você pode agrupar seus dados e conjurar feitiços de agregação. Veja:

```
from pyspark.sql import functions as F

dados = [("Grifinória", "Hermione", 98),
         ("Grifinória", "Harry",    85),
         ("Sonserina",  "Draco",    78)]
colunas = ["Casa", "Aluno", "Nota"]

df = spark.createDataFrame(dados, colunas)
df.groupBy(
    "Casa"
).agg(
    F.avg("Nota").alias("Média")
).show()
```

🔮 E assim, você vê a média de notas por casa. Parece que a Grifinória continua com mais pontos!

05

Accio CSV

Invocando Dados do Mundo Real

Chegou a hora de buscar dados fora de Hogwarts!O feitiço read.csv() serve para invocar arquivos do mundo trouxa direto para o seu ambiente Spark.

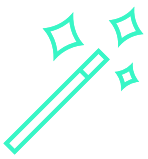
O Spark consegue ler arquivos .csv, .json, .parquet e muitos outros, com apenas um comando.

```
df = spark.read.csv(
    "notas_alunos.csv",
    header=True,
    inferSchema=True
)
df.show()
```

✨ Seus dados do mundo real agora estão sob seu controle.Um verdadeiro feitiço de invocação!

06

Obliviate Save



Guardando o Conhecimento

Todo feitiço precisa ser registrado, afinal, conhecimento esquecido é poder perdido. Com o comando write você pode salvar seus resultados e compartilhá-los com o mundo.

Observe com conjurar o comando:

```
df.write.mode(
    "overwrite"
).csv(
    "saida/resultados.csv",
    header=True
)
```

✨ E pronto!
Seus resultados foram guardados em pergaminhos que até os trouxas podem ler.

O Spark pode salvar DataFrames em diversos formatos mágicos: CSV, JSON, Parquet e outros.

Verifique qual formato de pergaminho se adapta melhor para as o uso de suas artes mágicas no futuro.

Conclusão – O Diploma de Bruxa dos Dados

Parabéns, jovem aprendiz!

Você passou por todas as aulas, conjurou feitiços poderosos e agora entende os mistérios do SparkSQL. Mas lembre-se: a magia dos dados nunca para.

Continue praticando, explore outros módulos do Spark — como MLlib, Streaming e GraphX — e mantenha viva sua curiosidade, afinal, todo bom bruxo dos dados sabe:

“Não é o poder do código que define quem você é, mas o jeito que você o usa.” 🪄💻

Acompanhe a bruxa que escreveu este grimório:

Renan Farias | LinkedIn

Obrigada e até a próxima