Streaming de Dados em Tempo Real: Aula 3

Prof. Felipe Timbó

UNI7

Ementa (dia 3)

- Resolução de problemas de mineração de dados com Python e Spark

Alterando nosso Setup...

Modificando a Instalação do Spark

1. Instalar o Java 8
   - ➢ `sudo apt-get update -y`
   - ➢ `sudo apt-get install openjdk-8-jdk -y`
2. Criar um ambiente Spark “spark-env.sh”
   - ➢ `cp /opt/spark/conf/spark-env.sh.template /opt/spark/conf/spark-env.sh`
   - ➢ `gedit /opt/spark/conf/spark-env.sh`
3. Adicionar ao arquivo “spark-env.sh” as seguintes linhas:

```
JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
PYSPARK_PYTHON=/usr/bin/python2.7
```

Modificando a Instalação do Spark

4. Configurar o arquivo “bashrc”:

```
gedit ~/.bashrc
```

5. **Alterar** no arquivo “bashrc” as seguintes linhas:

```
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
```

6. Salvar o arquivo (ctrl+s), fechá-lo e iniciar o bash

```
source ~/.bashrc
```

Modificando a Instalação do Spark

7. Configurar o arquivo “profile”:

```
gedit ~/.profile
```

8. **Alterar** no arquivo “profile” as seguintes linhas:

```
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
```

9. Salvar o arquivo (ctrl+s), fechá-lo e iniciar o profile

```
source ~/.profile
```

Modificando a Instalação do Spark

10. Abrir **novo terminal** e verificar a instalação com Spark shell:

- `pyspark`
- `nums = sc.parallelize([1, 2, 3, 4])`
- `nums.count()`
- `nums.collect()`

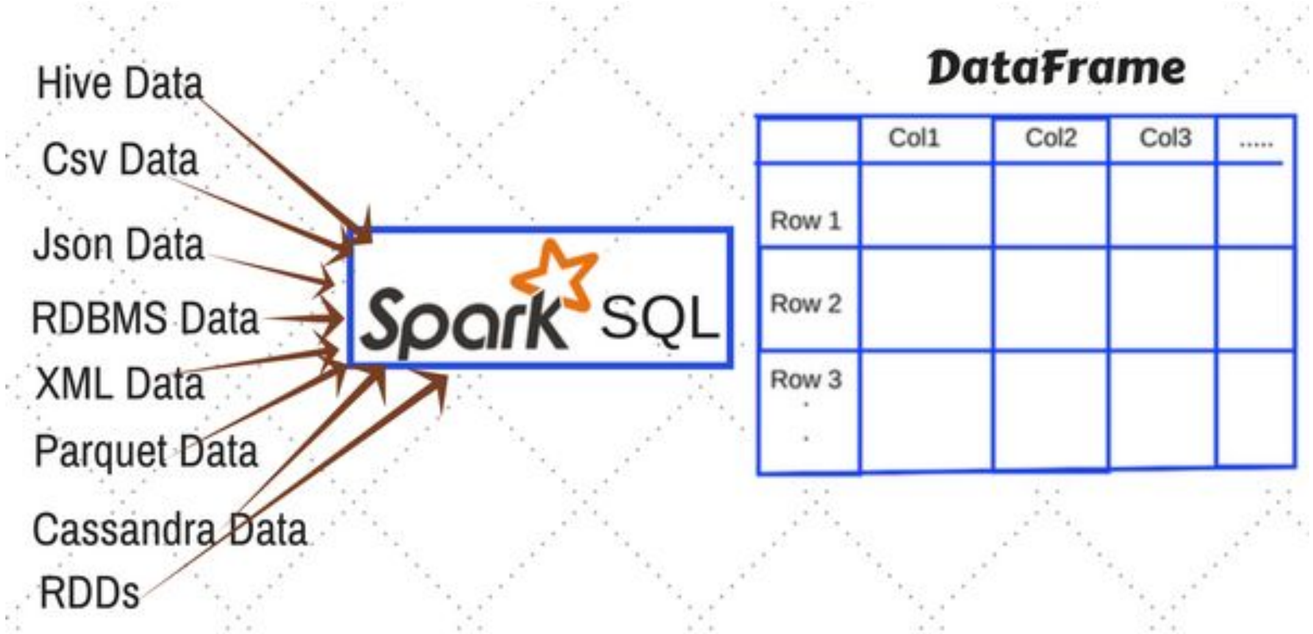
Lembre-se, para sair:

- `exit()`

Spark 2

Spark 2 - Dataframes

- Extensão do RDD
- Pode rodar consultas SQL
- Possui um esquema (maneira melhor de lidar)



Criando um SparkSession

Para criar um SparkSession:

- `pyspark`

- `from pyspark.sql import SparkSession`
- `spark = SparkSession.builder.appName("MyApp").getOrCreate()`

Spark 2 - Exemplo

Em outro terminal, baixar o conjunto de dados de pessoas

➢ `wget http://lia.ufc.br/~timbo/streaming/people.json`

Spark 2 - Exemplo

Criando um dataframe de pessoas no terminal Pyspark

➢ `df = spark.read.json("file:///home/posgrad/people.json")`

Brincando com Dataframes

- `df.show()`
- `df.printSchema()`
- `df.select(df.name).show()`
- `df.select(df['name'], df['age'] + 1).show()`
- `df.filter(df['age'] > 21).show()`
- `df.groupBy("age").count().show()`
- `df.orderBy("age", ascending = False).take(2)`
- `df.createOrReplaceTempView("pessoas")`
- `spark.sql("SELECT * FROM pessoas").show()`

Mais Exemplos

Em outro terminal, baixar o conjunto de dados de voos nos anos de 2014 e 2015:

- `wget http://lia.ufc.br/~timbo/streaming/voos-2014.csv`
- `wget http://lia.ufc.br/~timbo/streaming/voos-2015.csv`

Lendo os Dados

Criar um dataframe de pessoas no terminal Pyspark:

- `voos14 = spark.read.csv("file:///home/posgrad/voos-2014.csv")`
- `voos15 = spark.read.csv("file:///home/posgrad/voos-2015.csv")`
- `voos14.show()`
- `voos15.show()`

Ler os dados incluindo o cabeçalho e esquema:

- `voos14 = spark.read.option("inferSchema", "true").option("header",True).csv("file:///home/posgrad/voos-2014.csv")`
- `voos15 = spark.read.option("inferSchema", "true").option("header",True).csv("file:///home/posgrad/voos-2015.csv")`

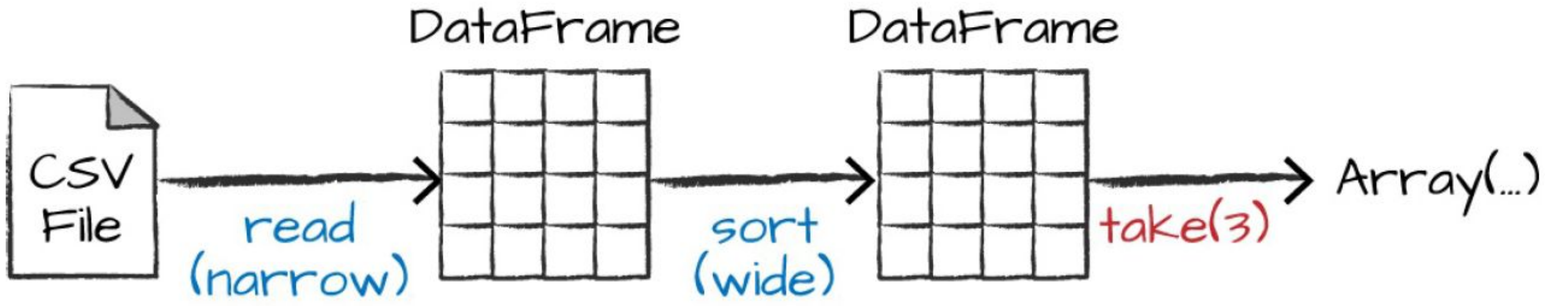
Obtendo Dataframes

Obter os 3 primeiros registros:

➢ `voos15.take(3)`

Ordenar e obter os 3 primeiros registros:

➢ `voos15.sort("count").take(2)`



Obtendo Dataframes

Obter a quantidade máxima de voos entre dois destinos

- `from pyspark.sql.functions import max`
- `voos15.select(max("count")).take(1)`

Obter o voo mais frequente em 2015

- `from pyspark.sql.functions import desc`
- `voos15.sort(desc("count")).take(1)`

Obter os 5 voos mais frequentes em 2015

- `voos15.sort(desc("count")).limit(5).show()`

Renomear uma coluna

- `voos15.sort(desc("count")).withColumnRenamed("DEST_COUNTRY_NAME", "destino").limit(5).show()`

Obtendo Dataframes

Obter a quantidade de voos em 2015

- `from pyspark.sql.functions import sum`
- `voos15.select(sum("count")).take(1)`

Obter a média de voos diários em 2015

- `voos15.select(sum("count")/365).take(1)`

Outras operações

União

- `voos_concat = voos14.union(voos15)`

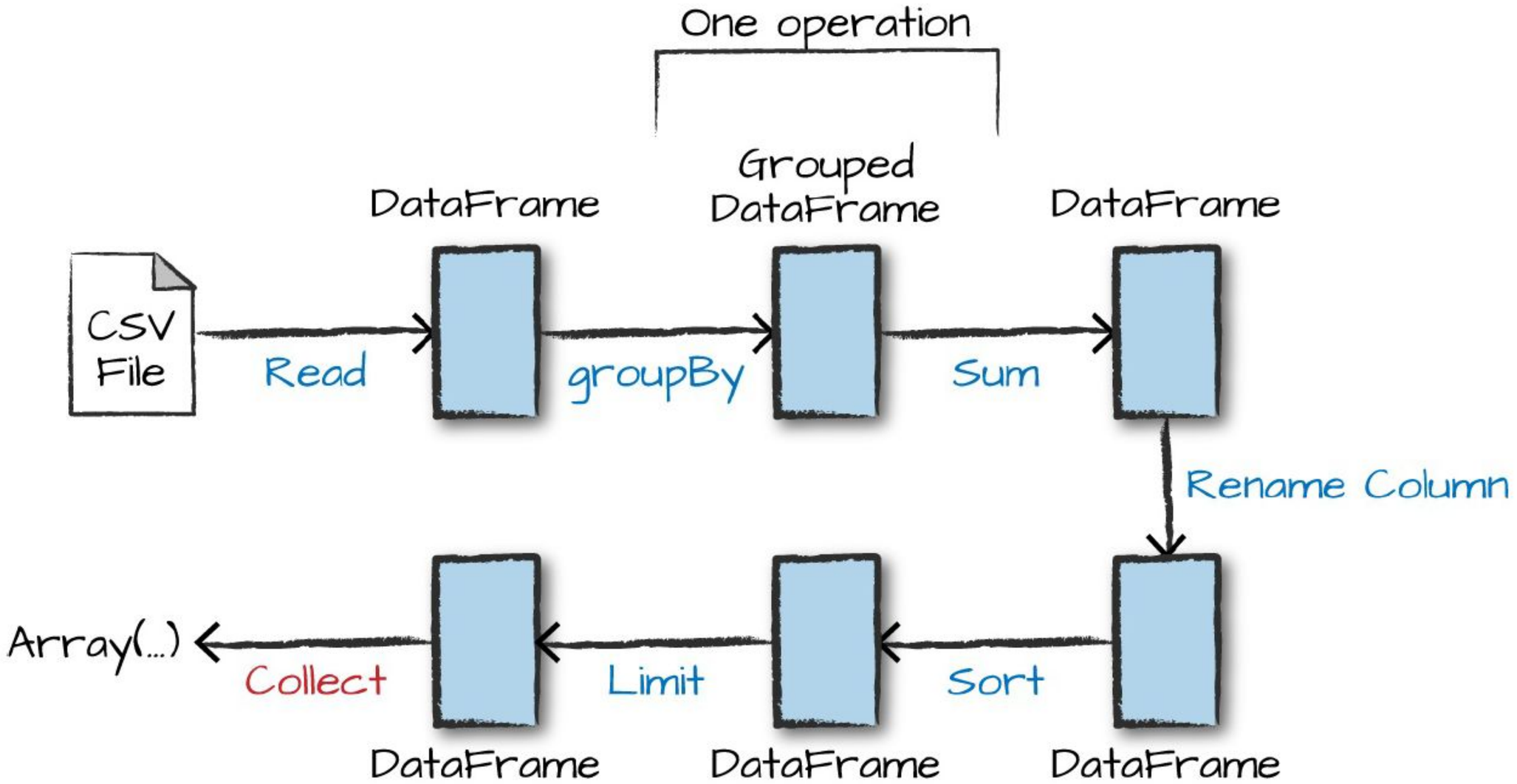
Interseção

- `voos_int = voos14.intersect(voos15)`

Exercício

Quais os 5 países destino que mais receberam voos em 2015?

Pipeline das operações



Exercício

Quais os 5 países destino que mais receberam voos em 2015? (em SQL)

Mais Exercícios

1. Qual a diferença no número total de voos entre 2014 e 2015?
2. Qual o voo mais frequente em 2014?
3. Qual a quantidade de voos total em 2014 entre os destinos "Bolivia" e "United States" ?
4. Qual a quantidade de voos total em 2014 e em 2015 entre os destinos "Germany" e "United States" ?
5. Qual a média de voos dentro dos EUA por dia considerando os anos de 2014 e 2015?

Em Script Python

Criar um arquivo Python no VSCode chamado voos.py

Escrever as seguintes linhas:

```
from pyspark.sql import SparkSession

if __name__ == "__main__":
    spark = SparkSession.builder.appName("App").getOrCreate()
    voos15 = spark.read.csv("file:///home/posgrad/voos-2015.csv")
    voos15.show()
    spark.stop()
```

Para rodar o script:

```
spark-submit voos.py
```

Mais Exercícios

Escreva um script Python para:

1. Saber se existe algum registro de voos entre Brasil e EUA e em caso positivo, qual a quantidade no ano de 2014?
2. Saber se existe algum registro de voos que não parte ou não chega nos EUA. Em caso positivo, quais esses voos?
3. Sumarizar os dados de 2014 e 2015 em um só Dataframe. (Certifique-se de que não contenha registros repetidos).