Hướng dẫn hoàn chỉnh về PySpark và SparkSQL.

Apache Spark là hệ thống tính toán dạng cluster cung cấp các thư viện và APIs toàn diện cho các nhà phát triển, và SparkSQL có thể được đại diện là một mô-đun trong Apache Spark để xử lý dữ liệu không cấu trúc với sự trợ giúp của DataFrame API.

chúng ta sẽ tìm hiểu cơ bản về cách chạy các tác vụ Spark với PySpark (Python API) và thực hiện các chức năng hữu ích bên trong. Nếu làm theo, bạn sẽ có thể nắm bắt được kiến thức cơ bản về PySparks và các chức năng phổ biến của nó.

```
# chúng tôi sử dụng thư viện findspark để xác định vị trí spark trên máy cục bộ của chúng
tôi
import findspark
findspark.init('C:/spark/spark-3.2.3-bin-hadoop2.7')


import pandas as pd
import numpy as np
from datetime import date, timedelta, datetime
import time

import pyspark
from pyspark.sql import SparkSession, SQLContext
from pyspark.context import SparkContext
from pyspark.sql.functions import *
from pyspark.sql.types import *
```

1. Khởi tạo phiên Spark

Chúng ta cần bắt đầu bằng việc khởi tạo phiên Spark. DataFrame có thể được tạo và đăng ký như các bảng. Hơn nữa, các bảng SQL có thể được thực thi, các bảng có thể được lưu vào bộ nhớ đệm, và các tệp định dạng parquet/json/csv/avro có thể được đọc.

```
sc = SparkSession.builder.appName("PysparkExample").config ("spark.sql.shuffle.partitions",
"50").config("spark.driver.maxResultSize","5g").config
("spark.sql.execution.arrow.enabled", "true").getOrCreate()


sc


<div>
    <p><b>SparkSession - in-memory</b></p>
```

SparkContext

```
<p><a href="http://172.16.2.39:4040">Spark UI</a></p>

<dl>
  <dt>Version</dt>
    <dd><code>v3.2.3</code></dd>
  <dt>Master</dt>
    <dd><code>local[*]</code></dd>
  <dt>AppName</dt>
    <dd><code>PysparkExample</code></dd>
</dl>


</div>
```

2. Tải dữ liệu

Spark rất tuyệt vời khi hỗ trợ đọc tất cả các loại dữ liệu khác nhau.

Các DataFrame có thể được tạo bằng cách đọc định dạng tệp txt, csv, json và parquet. Trong ví dụ của chúng tôi, chúng tôi sẽ sử dụng tệp định dạng .json. Bạn cũng có thể tìm và đọc các định dạng tệp văn bản, csv và parquet bằng cách sử dụng các hàm đọc tương ứng như được hiển thị bên dưới.

```
#JSON
dataframe = sc.read.json('./data/nyt2.json')

#TXT files
# dataframe_txt = sc.read.text('./data/text_data.txt')

#CSV files
# dataframe_csv = sc.read.csv('./data/csv_data.csv')

#PARQUET files
# dataframe_parquet = sc.read.load('./data/parquet_data.parquet')
```

Xem dữ liệu với show()

```
dataframe.show(10)
```

```
+--------------------+--------------------+--------------------+-----------------+---------
-----------+-------------+-----------------+-------------+----+-------------+-------------
-------+-------------+
|                 _id|  amazon_product_url|              author| bestsellers_date|
description|        price|   published_date|    publisher|rank|rank_last_week|
title|weeks_on_list|
+--------------------+--------------------+--------------------+-----------------+---------
-----------+-------------+-----------------+-------------+----+-------------+-------------
-------+-------------+
|{5b4aa4ead3089013...|http://www.amazon...|       Dean R Koontz|{{1211587200000}}|Odd
Thomas, who c...|   {null, 27}|{{1212883200000}}|       Bantam| {1}|           {0}|
ODD HOURS|          {1}|
|{5b4aa4ead3089013...|http://www.amazon...|     Stephenie Meyer|{{1211587200000}}|Aliens
have taken...|{25.99, null}|{{1212883200000}}|Little, Brown| {2}|           {1}|
THE HOST|          {3}|
|{5b4aa4ead3089013...|http://www.amazon...|        Emily Giffin|{{1211587200000}}|A woman's
happy m...|{24.95, null}|{{1212883200000}}| St. Martin's| {3}|           {2}|LOVE THE ONE
YOU'...|          {2}|
|{5b4aa4ead3089013...|http://www.amazon...|   Patricia Cornwell|{{1211587200000}}|A
Massachusetts s...|{22.95, null}|{{1212883200000}}|       Putnam| {4}|           {0}|
THE FRONT|          {1}|
|{5b4aa4ead3089013...|http://www.amazon...|     Chuck Palahniuk|{{1211587200000}}|An aging
porn que...|{24.95, null}|{{1212883200000}}|    Doubleday| {5}|           {0}|
SNUFF|          {1}|
|{5b4aa4ead3089013...|http://www.amazon...|James Patterson a...|{{1211587200000}}|A woman
finds an ...|{24.99, null}|{{1212883200000}}|Little, Brown| {6}|           {3}|SUNDAYS AT
TIFFANY'S|          {4}|
|{5b4aa4ead3089013...|http://www.amazon...|       John Sandford|{{1211587200000}}|The
Minneapolis d...|{26.95, null}|{{1212883200000}}|       Putnam| {7}|           {4}|
PHANTOM PREY|          {3}|
|{5b4aa4ead3089013...|http://www.amazon...|       Jimmy Buffett|{{1211587200000}}|A
Southern family...|{21.99, null}|{{1212883200000}}|Little, Brown| {8}|           {6}|
SWINE NOT?|          {2}|
|{5b4aa4ead3089013...|http://www.amazon...|    Elizabeth George|{{1211587200000}}|In
Cornwall, tryi...|{27.95, null}|{{1212883200000}}|       Harper| {9}|           {8}|
CARELESS IN RED|          {3}|
|{5b4aa4ead3089013...|http://www.amazon...|      David Baldacci|{{1211587200000}}|An
intelligence a...|{26.99, null}|{{1212883200000}}|Grand Central|{10}|           {7}|
THE WHOLE TRUTH|          {5}|
+--------------------+--------------------+--------------------+-----------------+---------
-----------+-------------+-----------------+-------------+----+-------------+-------------
-------+-------------+
only showing top 10 rows
```

Kiểm tra đơn giản dữ liệu.

Chúng tôi thường chỉ muốn lướt qua khung dữ liệu trước khi đi sâu hơn

```
# Trả về tên cột và kiểu dữ liệu của dataframe
dataframe.dtypes

# Hiển thị nội dung của dataframe
dataframe.show()

# Trả về n hàng đầu tiên
dataframe.head()

# Trả về hàng đầu tiên
dataframe.first()

# Return first n rows
dataframe.take(5)

# Tính toán số liệu thống kê tóm tắt
dataframe.describe().show()

# Trả về các cột của dataframe
dataframe.columns

# Đếm số hàng trong khung dataframe
dataframe.count()

# Đếm số hàng riêng biệt trong khung dữ liệu
dataframe.distinct().count()

# In các kế hoạch bao gồm vật lý và logic
dataframe.explain(4)
```

```
+--------------------+--------------------+--------------------+-----------------+---------
-----------+-------------+-----------------+--------------------+----+--------------+------
--------------+-------------+
|                 _id|  amazon_product_url|              author| bestsellers_date|
description|        price|   published_date|           publisher|rank|rank_last_week|
title|weeks_on_list|
+--------------------+--------------------+--------------------+-----------------+---------
-----------+-------------+-----------------+--------------------+----+--------------+------
--------------+-------------+
|{5b4aa4ead3089013...|http://www.amazon...|       Dean R Koontz|{{1211587200000}}|Odd
Thomas, who c...|   {null, 27}|{{1212883200000}}|              Bantam| {1}|           {0}|
ODD HOURS|          {1}|
|{5b4aa4ead3089013...|http://www.amazon...|     Stephenie Meyer|{{1211587200000}}|Aliens
have taken...|{25.99, null}|{{1212883200000}}|       Little, Brown| {2}|           {1}|
THE HOST|          {3}|
|{5b4aa4ead3089013...|http://www.amazon...|        Emily Giffin|{{1211587200000}}|A woman's
happy m...|{24.95, null}|{{1212883200000}}|        St. Martin's| {3}|           {2}|LOVE
THE ONE YOU'...|          {2}|
|{5b4aa4ead3089013...|http://www.amazon...|   Patricia Cornwell|{{1211587200000}}|A
Massachusetts s...|{22.95, null}|{{1212883200000}}|              Putnam| {4}|
{0}|           THE FRONT|          {1}|
|{5b4aa4ead3089013...|http://www.amazon...|     Chuck Palahniuk|{{1211587200000}}|An aging
porn que...|{24.95, null}|{{1212883200000}}|           Doubleday| {5}|           {0}|
SNUFF|          {1}|
|{5b4aa4ead3089013...|http://www.amazon...|James Patterson a...|{{1211587200000}}|A woman
finds an ...|{24.99, null}|{{1212883200000}}|       Little, Brown| {6}|
{3}|SUNDAYS AT TIFFANY’S|          {4}|
|{5b4aa4ead3089013...|http://www.amazon...|       John Sandford|{{1211587200000}}|The
Minneapolis d...|{26.95, null}|{{1212883200000}}|              Putnam| {7}|           {4}|
PHANTOM PREY|          {3}|
|{5b4aa4ead3089013...|http://www.amazon...|       Jimmy Buffett|{{1211587200000}}|A
Southern family...|{21.99, null}|{{1212883200000}}|       Little, Brown| {8}|
{6}|          SWINE NOT?|          {2}|
|{5b4aa4ead3089013...|http://www.amazon...|    Elizabeth George|{{1211587200000}}|In
```

```
Cornwall, tryi...|{27.95, null}|{{1212883200000}}|             Harper| {9}|          {8}|
CARELESS IN RED|          {3}|
|{5b4aa4ead3089013...|http://www.amazon...|      David Baldacci|{{1211587200000}}|An
intelligence a...|{26.99, null}|{{1212883200000}}|       Grand Central|{10}|          {7}|
THE WHOLE TRUTH|          {5}|
|{5b4aa4ead3089013...|http://www.amazon...|        Troy Denning|{{1211587200000}}|The New
Jedi orde...|   {null, 27}|{{1212883200000}}|  Del Rey/Ballantine|{11}|          {5}|
INVINCIBLE|          {2}|
|{5b4aa4ead3089013...|http://www.amazon...|          James Frey|{{1211587200000}}|A novel,
set in L...|{26.95, null}|{{1212883200000}}|             Harper|{12}|          {9}|BRIGHT
SHINY MORNING|          {2}|
|{5b4aa4ead3089013...|http://www.amazon...|         Garth Stein|{{1211587200000}}|A Lab-
terrier mix...|{23.95, null}|{{1212883200000}}|             Harper|{13}|          {0}|THE
ART OF RACING...|          {1}|
|{5b4aa4ead3089013...|http://www.amazon...|     Debbie Macomber|{{1211587200000}}|A widow
who owns ...|{24.95, null}|{{1212883200000}}|               Mira|{14}|          {10}|
TWENTY WISHES|          {4}|
|{5b4aa4ead3089013...|http://www.amazon...|         Jeff Shaara|{{1211587200000}}|A novel
about the...|   {null, 28}|{{1212883200000}}|          Ballantine|{15}|          {11}|
THE STEEL WAVE|          {2}|
|{5b4aa4ead3089013...|http://www.amazon...|    Phillip Margolin|{{1211587200000}}|
|    {null, 0}|{{1212883200000}}|HarperCollins Pub...|{16}|          {0}| EXECUTIVE
PRIVILEGE|          {0}|
|{5b4aa4ead3089013...|http://www.amazon...|       Jhumpa Lahiri|{{1211587200000}}|Stories
of the an...|    {null, 0}|{{1212883200000}}|               Knopf|{17}|          {0}|
UNACCUSTOMED EARTH|          {0}|
|{5b4aa4ead3089013...|http://www.amazon...|      Joseph O'Neill|{{1211587200000}}|A
Dutchman desert...|    {null, 0}|{{1212883200000}}|Knopf Publishing ...|{18}|
{0}|          NETHERLAND|          {0}|
|{5b4aa4ead3089013...|http://www.amazon...|        John Grisham|{{1211587200000}}|Political
and leg...|    {null, 0}|{{1212883200000}}|Doubleday Publishing|{19}|          {0}|
THE APPEAL|          {0}|
|{5b4aa4ead3089013...|http://www.amazon...|       James Rollins|{{1211587200000}}|
|    {null, 0}|{{1212883200000}}|Random House Publ...|{20}|          {0}|INDIANA JONES
AND...|          {0}|
+--------------------+--------------------+--------------------+-----------------+---------
-----------+-------------+-----------------+--------------------+----+--------------+------
--------------+-------------+
only showing top 20 rows

+-------+--------------------+---------------+--------------------+---------+--------------
----+
|summary|  amazon_product_url|         author|         description|publisher|
title|
+-------+--------------------+---------------+--------------------+---------+--------------
----+
|  count|               10195|          10195|               10195|    10195|
10195|
|   mean|                null|           null|                null|
null|1877.7142857142858|
| stddev|                null|           null|                null|     null|
370.9760613506458|
|    min|http://www.amazon...|        AJ Finn|                    |      ACE|  10TH
ANNIVERSARY|
|    max|https://www.amazo...|various authors|’Tis for the Rebe...|allantine|
ZOO|
+-------+--------------------+---------------+--------------------+---------+--------------
----+




---------------------------------------------------------------------------

TypeError                                 Traceback (most recent call last)
```

```
~\AppData\Local\Temp\ipykernel_13668\2535285985.py in <cell line: 29>()
     27
     28 # In các kế hoạch bao gồm vật lý và logic
---> 29 dataframe.explain(4)


C:/spark/spark-3.2.3-bin-hadoop2.7\python\pyspark\sql\dataframe.py in explain(self,
extended, mode)
    373             argtypes = [
    374                 str(type(arg)) for arg in [extended, mode] if arg is not None]
--> 375             raise TypeError(
    376                 "extended (optional) and mode (optional) should be a string "
    377                 "and bool; however, got [%s]." % ", ".join(argtypes))


TypeError: extended (optional) and mode (optional) should be a string and bool; however,
got [<class 'int'>].
```

3. Các hàm phổ biến hữu ích

[1] Xóa các giá trị trùng lặp

Có thể loại bỏ các giá trị trùng lặp trong một bảng bằng cách sử dụng hàm dropDuplicates().

```
dataframe_dropdup = dataframe.dropDuplicates()
dataframe_dropdup.show(10)
```

[2] Phép toán 'Select'

Có thể lấy các cột theo cột hoặc bằng cách lập chỉ mục (dataframe['author']).

```
#Hiển thị tất cả các mục trong cột tiêu đề
dataframe.select("author").show(10)

#Hiển thị các cột title, author, rank, price
dataframe.select("author", "title", "rank", "price").show(10)
```

[3] Phép toán 'When'

```
# Hiển thị tiêu đề và gán 0 hoặc 1 tùy thuộc vào tiêu đề
dataframe.select("title", when(dataframe.title != 'ODD HOURS', 1).otherwise(0)).show(10)
```

[4] Phép toán 'isin'

```
# Hiển thị các hàng có tác giả được chỉ định nếu trong các tùy chọn đã cho
dataframe [dataframe.author.isin("John Sandford", "Emily Giffin")].show(5)
```

[5] Phép toán 'Like'

```
# Hiển thị tác giả và tiêu đề là TRUE nếu tiêu đề có từ " THE " trong tiêu đề
dataframe.select("author", "title", dataframe.title.like("% THE %")).show(15)
```

[6] Phép toán 'Startswith' — 'Endswith'

StartsWith quét từ đầu từ/nội dung với các tiêu chí được chỉ định trong ngoặc. Song song, EndsWith xử lý từ/nội dung bắt đầu từ cuối. Cả hai chức năng đều phân biệt chữ hoa chữ thường.

```
dataframe.select("author", "title", dataframe.title.startswith("THE")).show(5)
dataframe.select("author", "title", dataframe.title.endswith("NT")).show(5)
```

[7] Phép toán 'Substring'

Trong các ví dụ sau, văn bản được trích xuất từ các số chỉ mục (1, 3), (3, 6) và (1, 6).

```
dataframe.select(dataframe.author.substr(1, 3).alias("title")).show(5)
dataframe.select(dataframe.author.substr(3, 6).alias("title")).show(5)
dataframe.select(dataframe.author.substr(1, 6).alias("title")).show(5)
```

[8] Thêm cột

```
from pyspark.sql import functions as F

# Lit() được yêu cầu trong khi chúng tôi đang tạo các cột có giá trị chính xác.
dataframe = dataframe.withColumn('new_column', F.lit('This is a new column'))

display(dataframe)
```

[9] Cập nhật cột

Đối với các hoạt động cập nhật của API DataFrame, hàm withColumnRenamed() được sử dụng với hai tham số.

```
# Cập nhật cột 'amazon_product_url' bằng 'URL'
dataframe = dataframe.withColumnRenamed('amazon_product_url', 'URL')

dataframe.show(5)
```

[10] Loại bỏ cột

Có thể xóa một cột theo hai cách: \

1. Thêm danh sách tên cột trong hàm drop()
2. Chỉ định cột bằng cách trỏ trong drop function

```
# 1.
dataframe_remove = dataframe.drop("publisher", "published_date").show(5)

# 2.
dataframe_remove2 =
dataframe.drop(dataframe.publisher).drop(dataframe.published_date).show(5)
```

[11] Phép toán 'GroupBy'

```
# Nhóm theo tác giả, đếm sách của các tác giả trong nhóm

dataframe.groupBy("author").count().show(10)
```

[12] Phép toán 'Filter'

```
# Lọc các mục tiêu đề
# Chỉ giữ các bản ghi có giá trị 'THE HOST'

dataframe.filter(dataframe["title"] == 'THE HOST').show(5)
```

[13] Xử lý các giá trị bị thiếu

```
# Thay thế giá trị null
dataframe.na.fill()
dataFrame.fillna()
dataFrameNaFunctions.fill()

# Trả về các hàng giới hạn khung dữ liệu mới có giá trị nulldataframe.na.drop()
dataFrame.dropna()
dataFrameNaFunctions.drop()

# Trả về khung dữ liệu mới thay thế một giá trị bằng một giá trị khác
dataframe.na.replace(5, 15)
dataFrame.replace()
dataFrameNaFunctions.replace()
```

[14] phân vùng lại

Có thể tăng hoặc giảm mức phân vùng hiện có trong RDD. \

Việc tăng có thể được hiện thực hóa bằng cách sử dụng chức năng **repartition(self, numPartitions)** để tạo ra một RDD mới thu được số lượng phân vùng bằng/cao hơn. \

Việc giảm có thể được xử lý bằng hàm **coalesce(self, numPartitions, shuffle=False)** dẫn đến RDD mới với số lượng phân vùng giảm xuống một số đã chỉ định

```
# Dataframe với 10 phân vùng
dataframe.repartition(10).rdd.getNumPartitions()

# Dataframe với 1 phân vùng
dataframe.coalesce(1).rdd.getNumPartitions()
```

[15] Chạy các lệnh SQL trong Spark

```
# Đăng ký bảng
dataframe.registerTempTable("df")

sc.sql("select * from df").show(3)


sc.sql("SELECT CASE WHEN description LIKE '%love%' THEN 'Love_Theme'
           WHEN description LIKE '%hate%' THEN 'Hate_Theme'
           WHEN description LIKE '%happy%' THEN 'Happiness_Theme' \
           WHEN description LIKE '%anger%' THEN 'Anger_Theme' \
           WHEN description LIKE '%horror%' THEN 'Horror_Theme' \
           WHEN description LIKE '%death%' THEN 'Criminal_Theme' \
           WHEN description LIKE '%detective%' THEN 'Mystery_Theme' ELSE 'Other_Themes' END
Themes from df").groupBy('Themes').count().show()
```

[16] Xuất dữ liệu

```
# Chuyển đổi khung dữ liệu thành RDD
rdd_convert = dataframe.rdd

# Chuyển đổi khung dữ liệu thành RDD của chuỗi
dataframe.toJSON().first()

# Lấy nội dung của df dưới dạng Pandas
dataFramedataframe.toPandas()
```

[17] Viết và lưu vào tệp

Bất kỳ loại nguồn dữ liệu nào được tải vào mã của chúng tôi dưới dạng khung dữ liệu đều có thể dễ dàng được chuyển đổi và lưu thành các loại khác bao gồm .parquet và .json

```
# Viết và lưu tệp ở định dạng .parquet
dataframe.select("author", "title", "rank", "description") \
        .write \
        .save("Rankings_Descriptions.parquet")


# Viết và lưu tệp ở định dạng .json
dataframe.select("author", "title") \
        .write \
        .save("Authors_Titles.json",format="json")
```

[18] Kết thúc phiên Spark

```
sc.stop()
```