Massive Data Processing Techniques in Data Science

**Assignment 1: Spark Tutorial**

The Resistance War Against America

Trần Quốc Lĩnh – 51703124

*Khoa Công nghệ thông tin*

*Trường Đại học Tôn Đức Thắng*

|  |
| --- |
| **Tóm tắt**  Giải thích các phương pháp thao tác dữ liệu bằng các phương pháp thông thường và với map reduce (Spark). Trên cơ sở dữ liệu thực tế - dữ liệu về các cuộc không kích trong cuộc kháng chiến chống Mĩ cứu nước. |

**Mô tả dữ liệu**

Dữ liệu này ghi lại thông tin về các trận không kích trong cuộc kháng chiến chống Mĩ cứu nước, hay Vietnam war đối với Mĩ. Từ năm 1955 đến năm 1975.

Các tập tin đính kèm trong báo cáo này gồm: Bombing\_Operations.json.gz (lưu trữ trên google drive), Aircraft\_Glossary.json.gz (lưu trữ trên google drive), và một file colab Spark\_Tutorial.ipynb.

**Bombing\_Operations (Các cuộc ném bom)**

* AirCraft: Mẫu máy bay
* ContryFlyingMission: Quốc gia thực hiện
* MissionDate: Ngày thực hiện
* OperationSupported: Đơn vị hỗ trợ
* PeriodOfDay: Thời gian thực hiện
* TakeoffLocation: Sân bay
* TimeOnTarget: Thời gian tấn công
* WeaponType: Loại vũ khí
* WeaponsLoadedWeight: Sức chứa của vũ khí.

**Aircraft\_Glossary (Thuật ngữ quân sự)**

* AirCraft: Mẫu máy bay
* AirCraftName: Tên kỹ thuật
* AirCraftType: Mã kỹ thuật

**Dữ liệu mẫu**

Tải dữ liệu:

Bombing\_Operations = spark.read.json("Bombing\_Operations.json.gz")

Aircraft\_Glossary = spark.read.json("Aircraft\_Glossary.json.gz")

Xuất mẫu:

[in]:

Bombing\_Operations.take(3)

[Out]:

[Row(AirCraft='EC-47', ContryFlyingMission='UNITED STATES OF AMERICA', MissionDate='1971-06-05', OperationSupported=None, PeriodOfDay='D', TakeoffLocation='TAN SON NHUT', TargetCountry='CAMBODIA', TimeOnTarget=1005.0, WeaponType=None, WeaponsLoadedWeight=0),

Row(AirCraft='EC-47', ContryFlyingMission='UNITED STATES OF AMERICA', MissionDate='1972-12-26', OperationSupported=None, PeriodOfDay='D', TakeoffLocation='NAKHON PHANOM', TargetCountry='SOUTH VIETNAM', TimeOnTarget=530.0, WeaponType=None, WeaponsLoadedWeight=0),

Row(AirCraft='RF-4', ContryFlyingMission='UNITED STATES OF AMERICA', MissionDate='1973-07-28', OperationSupported=None, PeriodOfDay='D', TakeoffLocation='UDORN AB', TargetCountry='LAOS', TimeOnTarget=730.0, WeaponType=None, WeaponsLoadedWeight=0)]

[in]:

Aircraft\_Glossary.show()

[Out]:

+--------+--------------------+--------------------+

|AirCraft| AirCraftName| AirCraftType|

+--------+--------------------+--------------------+

| A-1|Douglas A-1 Skyra...| Fighter Jet|

| A-26|Douglas A-26 Invader| Light Bomber|

| A-37|Cessna A-37 Drago...|Light ground-atta...|

| A-4|McDonnell Douglas...| Fighter Jet|

| A-5|North American A-...| Bomber Jet|

| A-6|Grumman A-6 Intruder| Attack Aircraft|

| A-7| LTV A-7 Corsair II| Attack Aircraft|

| AC-119|Fairchild AC-119 ...|Military Transpor...|

| AC-123|Fairchild C-123 P...|Military Transpor...|

| AC-130|Lockheed AC-130 S...|Fixed wing ground...|

| AC-47|Douglas AC-47 Spooky|Ground attack air...|

| AH-1| Bell AH-1 HueyCobra| Helicopter|

| B-1| Rockwell B-1 Lancer|Heavy strategic b...|

| B-52| B-52 Stratofortress| Strategic bomber|

| B-57|Martin B-57 Canberra| Tactical Bomber|

| B-66|Douglas B-66 Dest...| Light Bomber|

| C-1| Grumman C-1A Trader| Transport|

| C-117| C-117D Skytrain| Transport|

| C-119|Fairchild C-119 F...|Military Transpor...|

| C-123|Fairchild C-123 P...|Military Transpor...|

+--------+--------------------+--------------------+

only showing top 20 rows

1. **Các quốc gia tham chiến và số lượng chiến dịch.**

Nhóm các dữ liệu có cùng ContryFlyingMission và đếm số lượng chiến dịch

[in]:

missions\_counts = Bombing\_Operations.groupBy("ContryFlyingMission")\

                                  .agg(count("\*").alias("MissionsCount"))\

                                  .sort(desc("MissionsCount"))

Câu lệnh trên tạo ra một bảng dữ liệu với 2 cột (ContryFlyingMission và MissionsCount) Hàm groupBy("ContryFlyingMission")thực hiện việc nhóm các dữ liệu giống nhau tại trường ContryFlyingMission. Hàm agg(count("\*") đếm và gộp lại dữ liệu. Hàm sort(desc()) sắp xếp các giá trị của dữ liệu theo thứ tự giảm dần của MissionsCount.

Hiển thị bảng đã tạo

[in]:

missions\_counts.show()

[Out]:

+--------------------+-------------+

| ContryFlyingMission|MissionsCount|

+--------------------+-------------+

|UNITED STATES OF ...| 3708997|

| VIETNAM (SOUTH)| 622013|

| LAOS| 32777|

| KOREA (SOUTH)| 24469|

| AUSTRALIA| 12519|

+--------------------+-------------+

Sử dụng lệnh SQL

[in]:

Bombing\_Operations.registerTempTable("Bombing\_Operations")

query = """

SELECT ContryFlyingMission, count(\*) as MissionsCount

FROM Bombing\_Operations

GROUP BY ContryFlyingMission

ORDER BY MissionsCount DESC

"""

missions\_counts = spark.sql(query)

Câu query trên có nghĩa như sau: Từ bảng Bombing\_Operations. Chọn trường ContryFlyingMission , đếm số lượng từ trường và hiển thị dưới label MissionsCount. Sau đó nhóm vào ContryFlyingMission rồi sắp xếp theo thứ tự giảm dần của MissionsCount.

[in]:

missions\_counts.show()

[Out]:

+--------------------+-------------+

| ContryFlyingMission|MissionsCount|

+--------------------+-------------+

|UNITED STATES OF ...| 3708997|

| VIETNAM (SOUTH)| 622013|

| LAOS| 32777|

| KOREA (SOUTH)| 24469|

| AUSTRALIA| 12519|

+--------------------+-------------+

Hiển thị dữ liệu trên dưới dạng Pandas

[in]:

missions\_count\_pd = missions\_counts.toPandas()

missions\_count\_pd.head()

[Out]:

|  | **ContryFlyingMission** | **MissionsCount** |
| --- | --- | --- |
| **0** | UNITED STATES OF AMERICA | 3708997 |
| **1** | VIETNAM (SOUTH) | 622013 |
| **2** | LAOS | 32777 |
| **3** | KOREA (SOUTH) | 24469 |
| **4** | AUSTRALIA | 12519 |

Vẽ biểu đồ cho dữ liệu trên ta được

[in]:

pl = missions\_count\_pd.plot(kind="bar",

                      x="ContryFlyingMission", y="MissionsCount",

                      figsize=(10, 7), log=True, alpha=0.5, color="olive")

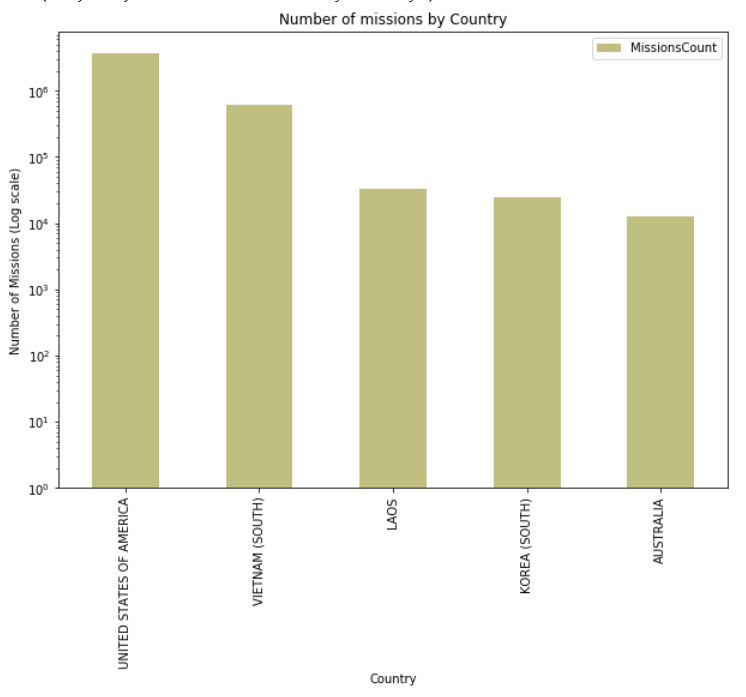
pl.set\_xlabel("Country")

pl.set\_ylabel("Number of Missions (Log scale)")

pl.set\_title("Number of missions by Country")

[Out]:

Text(0.5, 1.0, 'Number of missions by Country')



1. **Số lượng chiến dịch theo từng mốc thời gian của các quốc gia tham chiến.**

Chọn các dữ liệu cần thiết (MissionDate và ContryFlyingMission), trong đó MissionDate được chuyển định dạng sang date (một object trong python).

[in]:

missions\_countries = Bombing\_Operations.selectExpr(["to\_date(MissionDate) as MissionDate", "ContryFlyingMission"])

Nhóm các dữ liệu MissionDate và ContryFlyingMission để tính toán.

[in]:

missions\_by\_date = missions\_countries\

                    .groupBy(["MissionDate", "ContryFlyingMission"])\

                    .agg(count("\*").alias("MissionsCount"))\

                    .sort(asc("MissionDate")).toPandas()

Câu lệnh trên tạo ra một bảng dữ liệu với ba cột, lần lượt là ngày, quốc gia thực hiện và số phi vụ. Được sắp xếp theo thứ tự giảm dần của số phi vụ. Sau dó được chuyển sang định dạng Pandas.

Hiển thị dữ liệu:

[in]:

missions\_by\_date.head()

[Out]:

|  | **MissionDate** | **ContryFlyingMission** | **MissionsCount** |
| --- | --- | --- | --- |
| **0** | 1965-10-01 | UNITED STATES OF AMERICA | 447 |
| **1** | 1965-10-02 | UNITED STATES OF AMERICA | 652 |
| **2** | 1965-10-03 | UNITED STATES OF AMERICA | 608 |
| **3** | 1965-10-04 | UNITED STATES OF AMERICA | 532 |
| **4** | 1965-10-05 | UNITED STATES OF AMERICA | 697 |

Dữ liệu trên khi được biểu diễn dưới dạng biểu đồ:

[in]:

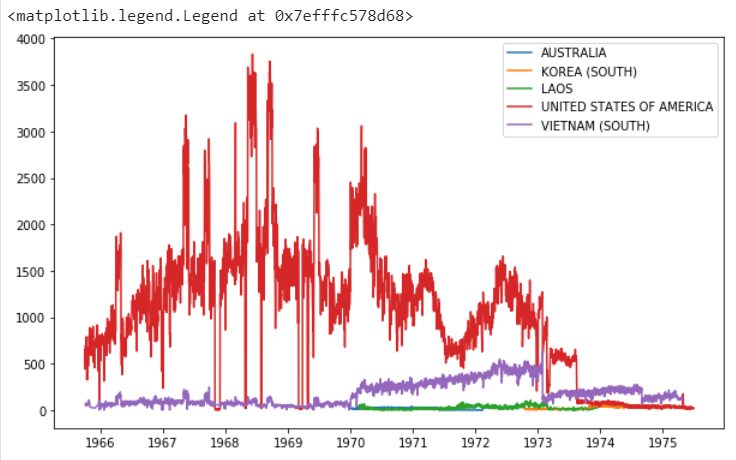
fig = plt.figure(figsize=(10, 6))

for country, missions in missions\_by\_date.groupby("ContryFlyingMission"):

    plt.plot(missions["MissionDate"], missions["MissionsCount"], label=country)

plt.legend(loc='best')

[Out]:



1. **Ai đánh bom khu vực này?**

Giả sử ta cần tìm thông tin về tất cả các chiến dịch vào ngày 29-06-1966 nhắm vào Hà Nội

[in]:

jun\_29\_operations = Bombing\_Operations.where("MissionDate = '1966-06-29' AND TargetCountry='NORTH VIETNAM'")

jun\_29\_operations.groupBy("ContryFlyingMission").agg(count("\*").alias("MissionsCount")).toPandas()

[Out]:

|  | **ContryFlyingMission** | **MissionsCount** |
| --- | --- | --- |
| **0** | VIETNAM (SOUTH) | 6 |
| **1** | UNITED STATES OF AMERICA | 389 |

Đếm số lượng chiến dịch

[in]:

%time jun\_29\_operations.count()

[Out]:

CPU times: user 1.54 ms, sys: 2.72 ms, total: 4.26 ms

Wall time: 14.8 s

395

Cache dữ liệu để tăng tốc cho lần query tiếp theo

[in]:

jun\_29\_operations.cache()

Hoặc lưu trữ vào file json…

[in]:

jun\_29\_operations.write.mode('overwrite').json("jun\_29\_operations.json")

..Rồi đọc lên

[in]:

jun\_29\_operations = spark.read.json("jun\_29\_operations.json")

1. **Mẫu chiến cơ được sử dụng nhiều nhất trong cuộc kháng chiến chống Mĩ cứu nước (Dựa vào số lượng chiến dịch).**

Xem thử 5 dòng đầu tiền của dữ liệu

[in]:

Aircraft\_Glossary.show(5)

[Out]:

+--------+--------------------+--------------------+

|AirCraft| AirCraftName| AirCraftType|

+--------+--------------------+--------------------+

| A-1|Douglas A-1 Skyra...| Fighter Jet|

| A-26|Douglas A-26 Invader| Light Bomber|

| A-37|Cessna A-37 Drago...|Light ground-atta...|

| A-4|McDonnell Douglas...| Fighter Jet|

| A-5|North American A-...| Bomber Jet|

+--------+--------------------+--------------------+

only showing top 5 rows

Liên kết AirCraft của Aircraft\_Glossary với AirCraft của Bombing\_Operations

[in]:

missions\_joined = Bombing\_Operations.join(Aircraft\_Glossary,

               Bombing\_Operations.AirCraft == Aircraft\_Glossary.AirCraft)

Nhóm dữ liệu và kết xuất kết quả phù hợp

[in]:

missions\_aircrafts.groupBy("AirCraftType")

.agg(count("\*").alias("MissionsCount"))\

                  .sort(desc("MissionsCount"))\

                  .show()

[Out]:

+--------------------+-------------+

| AirCraftType|MissionsCount|

+--------------------+-------------+

| Fighter Jet Bomber| 1073126|

| Fighter Jet| 882594|

| Jet Fighter Bomber| 451385|

| Attack Aircraft| 315246|

|Light ground-atta...| 267457|

| Fighter bomber jet| 242231|

|Military Transpor...| 228426|

| Utility Helicopter| 146653|

| Strategic bomber| 99100|

| Tactical Bomber| 82219|

|Observation Aircraft| 81820|

|Fixed wing ground...| 75058|

|Ground attack air...| 73843|

|Carrier-based Fig...| 58691|

| Training Aircraft| 48435|

| Light fighter| 39999|

| Light Bomber| 39262|

|Light Tactical Bo...| 34738|

| Light Utility Plane| 28582|

|Observation/ Ligh...| 24491|

+--------------------+-------------+

only showing top 20 rows

Sử dụng SQL

[in]:

Bombing\_Operations.registerTempTable("Bombing\_Operations")

Aircraft\_Glossary.registerTempTable("Aircraft\_Glossary")

query = """

SELECT AirCraftType, count(\*) MissionsCount

FROM Bombing\_Operations bo

JOIN Aircraft\_Glossary ag

ON bo.AirCraft = ag.AirCraft

GROUP BY AirCraftType

ORDER BY MissionsCount DESC

"""

spark.sql(query).show()

[Out]:

+--------------------+-------------+

| AirCraftType|MissionsCount|

+--------------------+-------------+

| Fighter Jet Bomber| 1073126|

| Fighter Jet| 882594|

| Jet Fighter Bomber| 451385|

| Attack Aircraft| 315246|

|Light ground-atta...| 267457|

| Fighter bomber jet| 242231|

|Military Transpor...| 228426|

| Utility Helicopter| 146653|

| Strategic bomber| 99100|

| Tactical Bomber| 82219|

|Observation Aircraft| 81820|

|Fixed wing ground...| 75058|

|Ground attack air...| 73843|

|Carrier-based Fig...| 58691|

| Training Aircraft| 48435|

| Light fighter| 39999|

| Light Bomber| 39262|

|Light Tactical Bo...| 34738|

| Light Utility Plane| 28582|

|Observation/ Ligh...| 24491|

+--------------------+-------------+

only showing top 20 rows