Lưu ý:

* Tạo file dữ liệu theo định dạng yêu cầu của từng bài. (mặc định csv)

1. **Cho tập dữ liệu sau**

id,name,population

0,Warsaw,1 764 615

1,Villeneuve-Loubet,15 020

2,Vranje,83 524

3,Pittsburgh,1 775 634

Input

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

| id| name|population|

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

| 0| Warsaw| 1 764 615|

| 1|Villeneuve-Loubet| 15 020|

| 2| Vranje| 83 524|

| 3| Pittsburgh| 1 775 634|

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

Output

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

|population|

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

| 1775634|

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

Yêu cầu:

* Load dữ liệu từ csv file
* Query population lớn nhất

1. **Cho tập dữ liệu sau**

+---+-----+

| id|group|

+---+-----+

| 0| 0|

| 1| 1|

| 2| 0|

| 3| 1|

| 4| 0|

+---+-----+

Output:

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

|group| ids|

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

| 0|[0, 2, 4]|

| 1| [1, 3]|

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

Lưu ý: biểu diễn [0,2,4] là biễu diễn của mảng

1. Cho tập dữ liệu sau

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

| id|day|price|units|

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

|100| 1| 23| 10|

|100| 2| 45| 11|

|100| 3| 67| 12|

|100| 4| 78| 13|

|101| 1| 23| 10|

|101| 2| 45| 13|

|101| 3| 67| 14|

|101| 4| 78| 15|

|102| 1| 23| 10|

|102| 2| 45| 11|

|102| 3| 67| 16|

|102| 4| 78| 18|

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

Output:

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

| id|price\_1|price\_2|price\_3|price\_4|unit\_1|unit\_2|unit\_3|unit\_4|

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

|100| 23| 45| 67| 78| 10| 11| 12| 13|

|101| 23| 45| 67| 78| 10| 13| 14| 15|

|102| 23| 45| 67| 78| 10| 11| 16| 18|

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

1. **Cho tập dữ liệu sau**

time,department,items\_sold

1,IT,15

2,Support,81

3,Support,90

4,Support,25

5,IT,40

6,IT,24

7,Support,31

8,Support,1

9,HR,27

10,IT,75

Input:

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

|time|department|items\_sold|

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

| 1| IT| 15|

| 2| Support| 81|

| 3| Support| 90|

| 4| Support| 25|

| 5| IT| 40|

| 6| IT| 24|

| 7| Support| 31|

| 8| Support| 1|

| 9| HR| 27|

| 10| IT| 75|

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

Output:

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

|time|department|items\_sold|running\_total|

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

| 9| HR| 27| 27|

| 1| IT| 15| 15|

| 5| IT| 40| 55|

| 6| IT| 24| 79|

| 10| IT| 75| 154|

| 2| Support| 81| 81|

| 3| Support| 90| 171|

| 4| Support| 25| 196|

| 7| Support| 31| 227|

| 8| Support| 1| 228|

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

Yêu cầu:

* Load dữ liệu vào spark từ file csv
* Tính running\_total hay tổng tích lũy số item đã bán được đến thời điểm time.

1. **Cho tập dữ liệu sau**

time,department,items\_sold,running\_total

1,IT,15,15

2,Support,81,81

3,Support,90,171

4,Support,25,196

5,IT,40,55

6,IT,24,79

7,Support,31,227

8,Support,1,228

9,HR,27,27

10,IT,75,154

Input:

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

|time|department|items\_sold|running\_total|

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

| 1| IT| 15| 15|

| 2| Support| 81| 81|

| 3| Support| 90| 171|

| 4| Support| 25| 196|

| 5| IT| 40| 55|

| 6| IT| 24| 79|

| 7| Support| 31| 227|

| 8| Support| 1| 228|

| 9| HR| 27| 27|

| 10| IT| 75| 154|

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

Output:

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

|time|department|items\_sold|running\_total|diff|

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

| 9| HR| 27| 27| 27|

| 1| IT| 15| 15| 15|

| 5| IT| 40| 55| 40|

| 6| IT| 24| 79| 24|

| 10| IT| 75| 154| 75|

| 2| Support| 81| 81| 81|

| 3| Support| 90| 171| 90|

| 4| Support| 25| 196| 25|

| 7| Support| 31| 227| 31|

| 8| Support| 1| 228| 1|

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

Yêu cầu:

* Đọc dữ liệu vào spark từ file csv chứa dữ liệu.
* Dùng SparkSQL query độ lệch giữa các running\_total liên tiếp nhau theo thời gian của các department

1. **Cho tập dữ liệu sau**

id,name,department,salary

1,Hunter Fields,IT,15

2,Leonard Lewis,Support,81

3,Jason Dawson,Support,90

4,Andre Grant,Support,25

5,Earl Walton,IT,40

6,Alan Hanson,IT,24

7,Clyde Matthews,Support,31

8,Josephine Leonard,Support,1

9,Owen Boone,HR,27

10,Max McBride,IT,75

Input

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

| id| name|department|salary|

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

| 1| Hunter Fields| IT| 15|

| 2| Leonard Lewis| Support| 81|

| 3| Jason Dawson| Support| 90|

| 4| Andre Grant| Support| 25|

| 5| Earl Walton| IT| 40|

| 6| Alan Hanson| IT| 24|

| 7| Clyde Matthews| Support| 31|

| 8|Josephine Leonard| Support| 1|

| 9| Owen Boone| HR| 27|

| 10| Max McBride| IT| 75|

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

Output

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

| id| name|department|salary|diff|

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

| 9| Owen Boone| HR| 27| 0|

| 1| Hunter Fields| IT| 15| 60|

| 5| Earl Walton| IT| 40| 35|

| 6| Alan Hanson| IT| 24| 51|

| 10| Max McBride| IT| 75| 0|

| 2| Leonard Lewis| Support| 81| 9|

| 3| Jason Dawson| Support| 90| 0|

| 4| Andre Grant| Support| 25| 65|

| 7| Clyde Matthews| Support| 31| 59|

| 8|Josephine Leonard| Support| 1| 89|

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

Yêu cầu :

* Đọc dữ liệu vào spark từ file csv
* Tìm sự chênh lệch salary giữa nhân viên có lương cao nhất với các nhân viên còn lại trong từng phòng ban

1. **Cho tập dữ liệu sau**

Employee,Salary

Tony,50

Alan,45

Lee,60

David,35

Steve,65

Paul,48

Micky,62

George,80

Nigel,64

John,42

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

|Employee|Salary|

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

| Tony| 50|

| Alan| 45|

| Lee| 60|

| David| 35|

| Steve| 65|

| Paul| 48|

| Micky| 62|

| George| 80|

| Nigel| 64|

| John| 42|

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

Output:

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

|Employee|Salary|Percentage|

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

| George| 80| High|

| Steve| 65| High|

| Nigel| 64| High|

| Micky| 62| High|

| Lee| 60| Average|

| Tony| 50| Low|

| Paul| 48| Low|

| Alan| 45| Low|

| John| 42| Low|

| David| 35| Low|

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

Yêu cầu:

-Load dữ liệu từ file csv

- Thêm cột Percentage với giá trị được quy định như sau

* 30% số nhân viên đầu có lương cao nhất nhận giá trị “High”
* 40% tiếp theo nhận giá trị “Average”
* Phần còn lại nhận giá trị “Low”

1. **Cho tập dữ liệu sau**

id,title,genre,quantity

1,Hunter Fields,romance,15

2,Leonard Lewis,thriller,81

3,Jason Dawson,thriller,90

4,Andre Grant,thriller,25

5,Earl Walton,romance,40

6,Alan Hanson,romance,24

7,Clyde Matthews,thriller,31

8,Josephine Leonard,thriller,1

9,Owen Boone,sci-fi,27

10,Max McBride,romance,75

Input

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

| id| title| genre|quantity|

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

| 1| Hunter Fields| romance| 15|

| 2| Leonard Lewis|thriller| 81|

| 3| Jason Dawson|thriller| 90|

| 4| Andre Grant|thriller| 25|

| 5| Earl Walton| romance| 40|

| 6| Alan Hanson| romance| 24|

| 7| Clyde Matthews|thriller| 31|

| 8|Josephine Leonard|thriller| 1|

| 9| Owen Boone| sci-fi| 27|

| 10| Max McBride| romance| 75|

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

Output:

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

| id| title| genre|quantity|rank|

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

| 10| Max McBride| romance| 75| 1|

| 5| Earl Walton| romance| 40| 2|

| 3| Jason Dawson|thriller| 90| 1|

| 2|Leonard Lewis|thriller| 81| 2|

| 9| Owen Boone| sci-fi| 27| 1|

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

Yêu cầu:

* Load dữ liệu từ csv
* Lọc những tilte có số lượng quantity top 1 và top 2 với mỗi genre

1. **Cho tập dữ liệu sau**

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

|Qid| Question|AnswerText|ParticipantID|Assessment| GeoTag|

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

| 1|Question1Text| Yes| abcde1| 0|(x1,y1)|

| 2|Question2Text| No| abcde1| 0|(x1,y1)|

| 3|Question3Text| 3| abcde1| 0|(x1,y1)|

| 1|Question1Text| No| abcde2| 0|(x2,y2)|

| 2|Question2Text| Yes| abcde2| 0|(x2,y2)|

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

Qid,Question,AnswerText,ParticipantID,Assessment,GeoTag

1,Question1Text,Yes,abcde1,0,"(x1,y1)"

2,Question2Text,No,abcde1,0,"(x1,y1)"

3,Question3Text,3,abcde1,0,"(x1,y1)"

1,Question1Text,No,abcde2,0,"(x2,y2)"

2,Question2Text,Yes,abcde2,0,"(x2,y2)"

Output:

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

|ParticipantID|Assessment| GeoTag|Qid\_1|Qid\_2|Qid\_3|

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

| abcde1| 0|(x1,y1)| Yes| No| 3|

| abcde2| 0|(x2,y2)| No| Yes| null|

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

Yêu cầu:

* Load dữ liệu từ json
* Tạo báo cáo giống output

1. **Kafka**

- Cho topic: “topic1”

- Việc config để tạo ra object producer và consumer như bên dưới:

Properties props = ResourceUtils.loadProperties("kafka.producer.properties");

Producer<String, String> producer = new KafkaProducer<>(props);

Properties props = ResourceUtils.loadProperties("kafka.consumer.properties");

Consummer<String, String> producer = new KafkaConsumer<>(props);

-. Câu hỏi:

+ Ghi 5 số 1, 2, 3, 4, 5 vào topic1

+ Dùng consumer ở trên đăng ký topic1

+ Lấy ra message cũ nhất

+ Lấy ra message mới nhất

+ Lấy ra tất cả các message có trong topic1 hiện tại