Высшая Школа Экономики. ФКН

Контрольная работа.

"DS-09. Промышленное машинное обучение на Spark"

Начало выполнения задания: 19 апреля 2023 года, 17:00.
Жёсткий Дедлайн: **04 мая 2023 года, 23:59.**

Аннотация

Вам предлагается решить 10 задач с использованием Spark. Описание каждой задачи и формат входных данных указаны ниже. Решения задач необходимо оформить в виде Jupyter ноутбука. Задача считается решённой при выполнении следующих условий:

- Приведён правильный код для решения задачи
- Продемонстрирована корректность работы кода на примере из условия
- Было придумано два нетривиальных и различных примера входных данных и продемонстрирована корректность работы кода на данных примерах
- Примеры оформлены в виде `pyspark.sql.DataFrame`
- Решение использует только один **action** — `.toPandas` для отображения финального результата
- Решение использует только библиотеку `pyspark`
- В решении не используются SQL запросы (например, метод `.sql` или `.select` с SQL выражением) кроме как для операции **unpivot**

Порядок строк в ответе не важен, если явно не указано иное.

В качестве подсказки для каждой задачи указана ссылка на аналогичную задачу с решением через SQL запросы.

Решение в виде ноутбука с названием `[Exam][ФИО].ipynb` нужно отправить на почту nakhodnovms@my.msu.ru с темой письма `[HSE Spark 2022][Контрольная Работа][ФИО]`.

Содержание

1 Classes More Than 5 Students

Есть таблица курсов с колонками: студент и предмет. Перечислите все предметы, которым обучается не менее 5 человек.

INPUT **OUTPUT**

student	class
A	Math
B	English
C	Math
D	Biology
E	Math
F	Computer
G	Math
H	Math
I	Math

$\Longrightarrow$

class
Math

Рис. 1: Пример к задаче 1.

2 Second Highest Salary

Определите значение второй по величине зарплаты из таблицы с зарплатами сотрудников. Если второй по величине заработной платы нет, запрос должен вернуть строку `"Missing"`.

INPUT **OUTPUT**

id	salary
1	100
2	200
3	300

$\Longrightarrow$

SecondHighestSalary
200

id	salary
1	100

$\Longrightarrow$

SecondHighestSalary
Missing

Рис. 2: Пример к задаче 2.

3 Rising Temperature

В таблице содержится информация о температуре в определенный день. Найдите идентификаторы всех дат с более высокой температурой по сравнению с предыдущим ("вчерашним") днём.

INPUT **OUTPUT**

id	recordDate	temperature
1	2015-01-01	10
2	2015-01-02	25
3	2015-01-03	20
4	2015-01-04	30

$\Longrightarrow$

id
2
4

Рис. 3: Пример к задаче 3.

4 Duplicate Emails

Удалите из таблицы с адресами электронной почты повторяющиеся адреса, оставив только уникальные, которые соответствуют наименьшему идентификатору.

INPUT

id	email
1	john@example.com
2	bob@example.com
3	john@example.com

$\Longrightarrow$

OUTPUT

id	email
1	john@example.com
2	bob@example.com

Рис. 4: Пример к задаче 4.

5 Customers Who Never Order

По таблицам с информацией о клиентах и их заказах определите имена клиентов, которые ничего не заказывали.

INPUT

id	name
1	Joe
2	Henry
3	Sam
4	Max

id	customerId
1	3
2	1

$\Longrightarrow$

OUTPUT

Customers
Henry
Max

Рис. 5: Пример к задаче 5.

6 Employees Earning More Than Their Managers

Таблица содержит информацию о всех сотрудниках, включая их руководителей. У каждого сотрудника есть идентификатор, а также столбец с идентификатором их руководителя. Определите всех сотрудников, которые зарабатывают больше, чем их руководитель.

INPUT

id	name	salary	managerId
1	Joe	70000	3
2	Henry	80000	4
3	Sam	60000	Null
4	Max	90000	Null

$\Longrightarrow$

OUTPUT

Employee
Joe

Рис. 6: Пример к задаче 6.

7 Combine Two Tables

Объедините две таблицы так, чтобы определить поля `FirstName, LastName, City, State` для каждого человека в первой входной таблице.

INPUT

personId	lastName	firstName
1	Wang	Allen
2	Alice	Bob

addressId	personId	city	state
1	2	New York City	New York
2	3	Leetcode	California

$\Longrightarrow$

OUTPUT

firstName	lastName	city	state
Allen	Wang	Null	Null
Bob	Alice	New York City	New York

Рис. 7: Пример к задаче 7.

8 Not Boring Movies

В городе X открылся новый кинотеатр. Для каждого фильма в прокате известно его описание и рейтинг. Определите все фильмы с нечётными идентификаторами и описанием, которое не содержит подстроку `"boring"`. Выведите результат, отсортировав получившиеся фильмы по рейтингу.

INPUT

id	movie	description	rating
1	War	great 3D	8.9
2	Science	fiction	8.5
3	irish	boring	6.2
4	Ice song	Fantacy	8.6
5	House card	Interesting	9.1

$\Longrightarrow$

OUTPUT

id	movie	description	rating
5	House card	Interesting	9.1
1	War	great 3D	8.9

Рис. 8: Пример к задаче 8.

9 Swap Salary

Преобразуйте исходную таблицу, поменяв во всех записях пол сотрудников на противоположный (m $\Longleftrightarrow$ f).

INPUT

id	name	sex	salary
1	A	m	2500
2	B	f	1500
3	C	m	5500
4	D	f	500

$\Longrightarrow$

OUTPUT

id	name	sex	salary
1	A	f	2500
2	B	m	1500
3	C	f	5500
4	D	m	500

Рис. 9: Пример к задаче 9.

10 Big Countries

Будем называть страну *великой*, если ее площадь превышает 3 миллиона квадратных километров или население превышает 25 миллионов.

Выведите названия, население и площадь для всех *великих* стран из входной таблицы.

INPUT

name	continent	area	population	gdp
Afghanistan	Asia	652230	25500100	20343000000
Albania	Europe	28748	2831741	12960000000
Algeria	Africa	2381741	37100000	188681000000
Andorra	Europe	468	78115	3712000000
Angola	Africa	1246700	20609294	100990000000

$\Longrightarrow$

OUTPUT

name	population	area
Afghanistan	25500100	652230
Algeria	37100000	2381741

Рис. 10: Пример к задаче 10.