## 

[Передумови:](#_e4h0t0bbpc4r)

[Варіант 1](#_n4fewdsmu9fb)

[Варіант 2](#_ds7hyy3zj7v1)

[Варіант 3](#_z44e67ezmgou)

[Варіант 4](#_zap2dua6a37m)

[Варіант 5](#_2kn72wott042)

[Варіант 6](#_msfj2144mpjx)

[Варіант 7](#_ydbn8zz5fb1d)

[Варіант 8](#_7z0kcu5thhk2)

[Варіант 9](#_n2kt577ih0af)

## 

## Передумови:

1 Поставити та налаштувати Spark

2 На основі координат адрес Лондона наданих в даних: (<https://drive.google.com/drive/folders/1jtnV2uqNDg5JJE3h3CxJrE7KCrPyJiX4?usp=sharing> згенерувати 10ГБ+ даних, які б описували поїздки таксі та характеризувалися :

* водієм
* клієнтом
* стартова точка
* кінцева точка
* час початку поїздки
* час кінця поїздки
* вартість поїздки(**запропонуйте мені формулу, яка б враховувала час пік та нічний час**)
* оцінка водія (опціонально)
* відгук на водія ( за категоріями ) - опціонально
* текстовий відкук на водія (опціонально )
* оцінка на клієнта (опціонально)
* відгук на клієнта( за категоріями ) - опціонально

Увага! ОПЦІОНАЛЬНО = може бути присутнім не у всіх поїздках ( але в частині таки має бути) )))

водіїв 2к+

Клієнтів 4к+

3 на основі біг дата бази, що ви створили в п 2 отримати наступні “звіти” (використовуйте map reduce) та перевірити перевірки, що у вас потрібна кількість клієнтів, водіїв, оцінок і тд ( ви продумуєте які перевірки покажуть релевантність генерації ваших даних)). Продумайте, *що робити з текстовими коментами ? Враховуйте, що я можу попросити щось з ними зробити прямо на здачі лаби ;))*

## Варіант 1

* топ 100 водіїв за рейтингом

## Варіант 2

* кого з водіїв ми маємо зняти з нашої системи ( рейтинг менше 3,5)

## Варіант 3

* в який проміжок часу здійснюється найбільше поїздок

## Варіант 4

* топ 50 клієнтів за рейтингом

## Варіант 5

* топ 100 водів, що заробили найбільше

## Варіант 6

* топ 50 - водіїв, які переважно їздять вночі

## Варіант 7

* за що (яка категорія) найчастіше водіїв хвалять

## Варіант 8

* на що (найчастіше) скаржаться клієнти (категорія)

## Варіант 9

* Топ 10 найдовших текстових коментарів