HW01: Hadoop

Описание работы и критериев оценивания

В задании два блока:

1) Развертывание локального кластера - 50 баллов
2) Написание map reduce на Python - 50 баллов

Результаты ДЗ загрузить в репозиторий на Github

Бонусы и штрафы:

- **100%** за плагиат
- **30%** за посылку решения в течение недели после deadline

Блок 1. Развертывание локального кластера Hadoop

1) Развернуть локальный кластер в конфигурации 1 NN, 2 DN + 2 NM, 1 RM, 1 History server (опционально) (инструкция)
2) Изучить настройки и состояние NM и RM в веб-интерфейсе
3) Сделать скриншоты NN и RM, добавить в репозиторий

Блок 2. Написание map reduce на Python

В данной задаче мы будем подсчитывать среднее значение (аналог- numpy.mean) и дисперсию (аналог numpy.var) для сета из N сплитов данных с помощью map-reduce парадигмы.

1) Маппер функция должна генерить k кортежей вида (ck, mk, vk), где ck-размер chunk_size, mk-среднее данного chunk и vk - его дисперсия.
2) Редюсер функция должна скомбинировать полученные кортежи, вычислить результаты среднего значения и дисперсии величины и записать их в выходной файл.

За правильное исполнение map-reduce части для подсчета среднего значения начисляется 25 баллов и также 25 баллов можно получить за map-reduce подсчета дисперсии указанной величины.

С документацией и примерами можно ознакомиться здесь.

Блок 2. Написание map reduce на Python

1. Загрузите датасет по ценам на жилье Airbnb, доступный на kaggle.com: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

2. Подсчитайте среднее значение и дисперсию по признаку ”price” стандартными способами (”чистый код” или использование библиотек). Не учитывайте пропущенные значения при подсчете статистик.

3. Используя Python, реализуйте скрипт mapper.py и reducer.py для расчета каждой из двух величин.

4. Проверьте правильность подсчета статистик методом map-reduce в сравнении со стандартным подходом (могут быть минорные расхождения из-за особенностей чтения датасета, которые можно устранить предобраткой данных)

5. Результаты сравнения (то есть, подсчета двумя разными способами) для среднего значения и дисперсии запишите в файл .txt. В итоге, у вас должно получиться две пары значений (стандартного расчета и map-reduce)- одна пара для среднего, другая - для дисперсии.

6. Итоговый результат с выполненным заданием должен включать в себя сам код, а также результаты его работы, который необходимо разместить в репозитории.