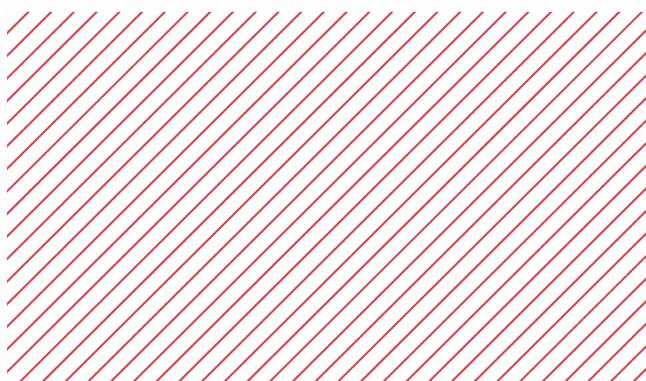


академия
больших
данных



HW01: Hadoop





Описание работы и критериев оценивания

В задании три блока:

- 1) Развёртывание локального кластера - 20 баллов
- 2) Работа с файловой системой - 40 баллов
- 3) Написание map reduce на Python - 40 баллов

Результаты ДЗ загрузить в репозиторий на Github и прислать ссылку на него в интерфейсе сдачи

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, 3 DN + NM, 1 RM, 1 History server ([инструкция](#))
- 2) Изучить настройки и состояние NM и RM в веб-интерфейсе
- 3) Сделать скриншоты NN и RM, добавить в репозиторий



Блок 2. Работа с HDFS

- 1) Выполните задания, записав выполненные команды последовательно в текстовый файл
- 2) Добавьте файл в репозиторий

Блок 2. Работа с HDFS

Все следующие задачи используют консольную утилиту “hdfs dfs”. Чтобы получить документацию / подсказку по HDFS-утилите или флагу, можно набрать:

- hdfs dfs -usage
- hdfs dfs -help

См. флаги “-mkdir” и “-touchz”

1. [2 балла] Создайте папку в корневой HDFS-папке
2. [2 балла] Создайте в созданной папке новую вложенную папку.
3. [3 балла] Что такое Trash в распределенной FS? Как сделать так, чтобы файлы удалялись сразу, минуя “Trash”?
4. [2 балла] Создайте пустой файл в подпапке из пункта 2.
5. [2 балла] Удалите созданный файл.
6. [2 балла] Удалите созданные папки.

См. флаги “-put”, “-cat”, “-tail”, “-cp”

1. [3 балла] Скопируйте любой в новую папку на HDFS
2. [3 балла] Выведите содержимое HDFS-файла на экран.
3. [3 балла] Выведите содержимое нескольких последних строчек HDFS-файла на экран.
4. [3 балла] Выведите содержимое нескольких первых строчек HDFS-файла на экран.
5. [3 балла] Переместите копию файла в HDFS на новую локацию.

Блок 2. Работа с HDFS

Полезные флаги:

- Для “hdfs dfs”, см. “-setrep -w”
- hdfs fsck /path -files - blocks -locations

Задачи:

2. [4 баллов] Изменить replication factor для файла. Как долго занимает время на увеличение / уменьшение числа реплик для файла?
3. [4 баллов] Найдите информацию по файлу, блокам и их расположениям с помощью “hdfs fsck”
4. [4 баллов] Получите информацию по любому блоку из п.2 с помощью “hdfs fsck -blockId”.
Обратите внимание на Generation Stamp (GS number).

Блок 3. Написание map reduce на Python

В данной задаче мы будем подсчитывать среднее значение (аналог `pumpry.mean`) и дисперсию (аналог `pumpry.var`) для сета из N сплитов данных с помощью map-reduce парадигмы. Маппер функция будет применяться нами к кортежам вида (ck, mk, vk) , где ck - размер `chunk_size`, mk -среднее данного `chunk` и vk -его дисперсия. Редюсер функция должна скомбинировать результаты среднего значения и дисперсии величины:

$$m_i = \frac{c_j m_j + c_k m_k}{c_j + c_k},$$
$$v_i = \frac{c_j v_j + c_k v_k}{c_j + c_k} + c_j c_k \left(\frac{m_j - m_k}{c_j + c_k} \right)^2$$

За правильное исполнение map-reduce части для подсчета среднего значения начисляется 20 баллов и также 20 баллов можно получить за map-reduce подсчета дисперсии указанной величины.

С документацией и примерами можно ознакомиться [здесь](#).

Блок 3. 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. В итоге у вас должно получиться 4 скрипта: 2 mapper и 2 reducer для каждой величины.
4. Проверьте правильность подсчета статистик методом map-reduce в сравнении со стандартным подходом
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.