

Домашнее задание №2

Срок сдачи: 4 декабря, 23:59:59. Решения, присланные позже данного срока, не принимаются.

Формулировка задания:

Вам необходимо реализовать на Hadoop MapReduce Java API умножение больших квадратных матриц $C=AB$ с помощью алгоритмов, рассмотренных на лекции:

- Вариант 1. Однопроходный алгоритм с использованием групп полос;
- Вариант 2. Двухпроходный алгоритм с использованием блоков;
- Дополнительное задание – Поддержка неквадратных матриц.

Формат командной строки для запуска программы:

`hadoop jar mm.jar mm -conf config.xml <путь до матрицы A> <путь до матрицы B> <путь до матрицы C>`

где:

Каждый путь до матрицы – путь вида `/path_to_matrix`

(в HDFS для каждой матрицы должны создаваться следующие файл и директория:

`/path_to_matrix/data` – директория с файлами, содержащими элементы матрицы;

`/path_to_matrix/size` – файл, содержащий размеры матрицы).

Формат файла размеров матрицы (разделитель – табуляция):

`<число строк> <число столбцов>`

Формат хранения элементов матрицы (разделитель – табуляция):

`<тэг> <индекс i> <индекс j> <элемент Aij>`

Параметры программы должны содержаться в конфигурационном файле `config.xml`.

В программе используются следующие параметры:

Параметр	Значение по умолчанию	Комментарий
<code>mm.groups</code>	1	число групп
<code>mm.tags</code>	ABC	тэги матриц A, B, C. каждый тэг – буква латинского алфавита, все 3 тэга – разные
<code>mm.float-format</code>	<code>%.3f</code>	формат записи float

mapred.reduce.tasks	1	число редьюсеров
---------------------	---	------------------

Далее Вам необходимо отправить на почту bigdatamsu@gmail.com архив в формате Task2-Фамилия.zip. Архив должен содержать:

1. Директорию “prog”, в которой должны быть:
 - а. Файл pom.xml, в котором описывается вся структура Вашего проекта – .jar мы будем собирать на своей стороне;
 - б. Директорию с исходными файлами, необходимыми для сборки .jar (!) Очень важно, чтобы main class Вашей программы имел название **mm**;
 - в. Файл config.xml с параметрами задачи и оптимальными параметрами запуска (см. “Отчет”);
2. Файл readme.txt с описанием того, как Вы компилировали и запускали программу.

Обратите внимание, что Ваш архив должен соответствовать указанной выше структуре (т.е. readme.txt и prog должны находиться в корне архива).

В теме письма необходимо указать номер задания (в том же формате, что и архив: Task2-Фамилия.zip).

(!) Самое главное: задания будут взяты в обработку только если Вы отправляете своё решение с Вашей gse-почты, в противном случае задания приниматься не будут!

Если с этим имеются сложности – опишитесь, пожалуйста, на bigdata@cs.msu.ru.

Перед выполнением домашнего задания рекомендуем ознакомиться с примерами тестовых данных (приведены в данном задании ниже). Возможно, они смогут ответить на вопросы, которые могут у Вас возникнуть во время выполнения задания.

Отчет:

В тексте письма с Домашним заданием №2 необходимо указать ссылку на документ Google docs с Вашим отчетом по выполнению домашнего задания. В отчете необходимо:

1. Подробно описать принципы работы Вашей программы;
2. Оптимизировать физическое время обработки для неразрезанных матриц размером 800 x 800 по параметру mm.groups.
Построить графики зависимости от mm.groups физического времени обработки и релевантных счетчиков из набора:
MAP_INPUT_RECORDS
MAP_OUTPUT_RECORDS
MAP_INPUT_BYTES

MAP_OUTPUT_BYTES
REDUCE_INPUT_GROUPS
REDUCE_SHUFFLE_BYTES
REDUCE_INPUT_RECORDS
REDUCE_OUTPUT_RECORDS

Найденное оптимальное количество групп необходимо отразить в конфигурационном файле;

3. Подобрать оптимальное число reducer'ов (график зависимости физического времени исполнения и релевантных характеристик). Найденное оптимальное количество редьюсеров необходимо отразить в конфигурационном файле:

```
<configuration>
  <property>
    <name>mm.tags</name>
    <value>ABC</value>
  </property>
  <property>
    <name>mm.groups</name>
    <value>***</value>
  </property>
  <property>
    <name>mapred.reduce.tasks</name>
    <value>***</value>
  </property>
</configuration>
```

где *** – Ваши найденные оптимальные значения.

4. Сформулировать выводы по результатам проведенных экспериментов.

Перед отправкой домашнего задания на проверку, рекомендуем проверить Вашу программу на данных поднаборах тестовых данных:

- <https://m.cs.msu.ru/s/9iNen3KHBRyyJQd>
- <https://m.cs.msu.ru/s/K5Fze9Mwed2nZYQ>

Отметим, что сгенерировать большие матрицы для экспериментов Вы можете как самостоятельно (при этом приложив код по их генерации и его описание) и получить за это бонусные баллы, либо можете использовать большие матрицы из доступных по вышеуказанной ссылке тестов (тест №5), самостоятельно загрузив их на кластер.

Разбиение по вариантам:

[illegible]

	2
	1
	2