Hadoop Map Reduce (Домашнее задание 1)

Семинар курса «Управление разно-структурированными большими данными»

http://synthesis.ipi.ac.ru/synthesis/student/BigData/seminar-hadoop/hadoop2014

alexey.vovchenko@gmail.com

Критерии оценки по курсу (РКТ, Магистратура)

Вопросы на зачетном занятии: Всем будет задано 2 простых вопроса по курсу. Каждый ответ – от 0 до 0.5 балла. 0 за совсем не правильный, 0.5 за полностью правильный.

Домашние задания: Каждое домашнее задание максимально оценивается в 1 балл. Всего два домашних задания.

Итоговая оценка: 2 + (кол-во набранных баллов). Округляется по правилам округления. Таким образом, для оценки 5 – достаточно сдать два домашних задания, и ответить на один из двух вопросов.

Посещение лекций и семинаров: Если вы пропустили какую-то лекцию или семинар, то на зачетном занятии будет задан дополнительный вопрос по этой теме. За не правильный ответ из итоговой оценки будет вычитаться 0.5 балла.

Список посещаемости будет выложен после последней лекции (9го декабря)

Hadoop Map Reduce (Домашнее задание 1)

Общие требования

- Задание выполняется в виде MapReduce приложения на Java.
- Срок выполнения задания – **01ое декабря**. За задание можно получить **1 балл**.
- Если задание будет прислано после 01го декабря, но до **23го декабря**, то за задание максимально можно будет получить **0.5 балла**.
- Выполненное задание должно быть прислано на почту в виде архива, содержащего проект Eclipse.
- Имя должно быть названо: <Family_Name>_MapReduce_Var<#>.rar (zip, gz, и.т.д.)
    - <Family_Name> - Ваша фамилия
    - <#> - номер варианта
    - Например: Vovchenko_MapReduce_Var7.rar
- Для получения варианта нужно написать мне на почту

Варианты заданий

1. Решение системы линейных уравнений

Уравнение в матричной форме.

Ax = b

A – матрица, ***x*** – вектор неизвестных, ***b*** – числовой вектор

Входные данные хранятся в файлах.

Формат представления данных в файлах – выбирается Вами, например csv.

Результат – пары: <ключ, значение>, где ключ - x_i, значение – посчитанное значение неизвестной

2. Произведение двух матриц

AB = C

A – первая исходная матрица, ***B*** – вторая исходная матрицы, ***C*** – матрица результат

Входные данные хранятся в файлах.

Формат представления данных в файлах – выбирается Вами, например csv.

Результат – пары: <ключ, значение>, где ключ – $c_{(i,j)}$, значение – посчитанное значение для элемента с индексами (i, j)

3. Построение инверсного индекса для списка документов

Входные данные хранятся в файлах.

Входные данные – набор текстов (аналогично задаче подсчета числа слов из семинарского практикума).

Результат – пары: <ключ, значение>, где ключ – слово, значение – Список пар <имя документа, индекс внутри документа> (Например - table: <text1,120>, <text1,233>, <text2,17>)

4. Реализация реляционных операций Select, Project

Дана таблица ***table (a, b, c, d, e)***

table – реляционная таблица с атрибутами - ***a, b, c, d, e*** (атрибуты ***a*** и ***b*** – одного типа)

Входные данные хранятся в файлах.

Формат представления данных в файлах – выбирается Вами, например csv.

Требуется выполнить в виде MapReduce программы запрос:

```
SELECT a, b, c
FROM table
WHERE a = b
```

В результате в файле должна быть таблица из атрибутов <***a, b, c***>, в том же формате, как и входные данные.

5. Реализация операции Join

Даны таблицы ***table1(a, b, c) u table2(c, d, e)***

table1 и ***table2*** – реляционные таблицы с атрибутами - ***a, b, c*** и ***c, d, e*** соответственно (атрибуты ***c*** в обоих таблицах должен быть одного типа)

Входные данные хранятся в файлах.

Формат представления данных в файлах – выбирается Вами, например csv.

Требуется выполнить в виде MapReduce программы запрос:

```
SELECT table1.a, table1.b, table1.c, table2.d, table2.e
FROM table1, table2
WHERE table1.c = table2.c
```

В результате в файле должна быть таблица из атрибутов <***a, b, c, d, e***>, в том же формате, как и входные данные.