Домашнее задание №4

Срок сдачи: 25 ноября, 23:59:59. Решения, присланные позже данного срока, не принимаются.

Обратите внимание, что в данном задании оценивается временная эффективность решения!

Формулировка задания:
На вход подаются источники данных, каждый источник данных содержит японские свечи по одному инструменту, упорядоченные по времени. Каждый источник данных может быть представлен единственным файлом или файлами, состоящими из нескольких частей, соответствующим результатам работы MapReduce задачи, формирующей свечи, при этом части файлов с одинаковыми номерами partitions должны содержать одинаковые диапазоны времени и в рамках partitions каждого источника должна обеспечиваться глобальная сортировка по времени.

Напомним, каждая свеча – это:

- MOMENT – время начала свечи;
- OPEN – цена первой сделки за свечу;
- HIGH – максимальная цена за свечу;
- LOW – минимальная цена за свечу;
- CLOSE – цена последней сделки за свечу.

Формат входных данных (каждого файла) без шапки:

SYMBOL,MOMENT,OPEN,HIGH,LOW,CLOSE

Вам необходимо:

Программа должна посчитать коэффициент корреляции Пирсона для каждой пары инструментов (всего n(n-1)/2) за выбранный промежуток времени и отсортировать пары инструментов согласно убыванию модуля коэффициента корреляции.

Формат строки вывода:
инстр1,инстр2\tкоэфф_корелл
Коэффициент корреляции считается для величин $(CLOSE_{i+1}-CLOSE_i)/CLOSE_i$, где i - номер свечи. В случае присутствия данных по одному инструменту и отсутствию данных по другому инструменту за некий момент времени, данные по этому моменту не включаются в расчет коэффициента.
Входными параметрами программы являются:

- candle.date.from = 19000101 #первый день периода времени (ГГГГММДД);

- candle.date.to = 20200101 #первый день после последнего дня периода (ГГГГММДД);
- candle.time.from = 1000 #время (ЧЧММ) начала первой свечи;
- candle.time.to = 1800 #время (ЧЧММ) после начала последней свечи;
- Название входной директории;
- Название выходной директории.

Необходимо сделать отчет о переданном по сети трафике и времени исполнения для двух и четырех рабочих A3-узлов в кластере, подобрать оптимальное число редьюсеров для обоих вариантов по времени исполнения. Включить в отчет таблицы по результатам проведенных экспериментов, сформулировать соответствующие выводы.

Далее Вам необходимо на почту курса bigdata@cs.msu.ru отправить архив в формате Task4-Фамилия.rar (фамилия на англ.), содержащий следующие файлы:

1) Файл Correlation.java с Вашим кодом (либо архив Correlation с исходными файлами);
2) Файл Correlation.jar;
3) Файл Correlation.pdf с выполненным отчетом;
4) Вспомогательные файлы для сборки (если используются);
5) Файл readme.txt с описанием того, как Вы компилировали и запускали программы;
6) Так как выходные данные слишком большого объема (и их выкачивание стоит дорого), для быстрой проверки корректности работы программы просим дополнительно запустить программу с входными параметрами, указанными ниже, и прислать полученный в результате файл Correlation.txt.

Входные данные для дополнительного запуска:

- candle.date.from = 20110111;
- candle.date.to = 20110112;
- candle.time.from = 1000;
- candle.time.to = 1020.

Необходимо провести дополнительный запуск на финансовых инструментах SVH1 и GDH1.