Домашнее задание №3

Срок сдачи: 11 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 подается .csv файл (разделители – запятые) с финансовыми данными, доступный по адресу:

wasb://financedata@bigdatamsu.blob.core.windows.net/

Столбцы данного файла имеют следующие названия:

#SYMBOL,SYSTEM,MOMENT,ID_DEAL,PRICE_DEAL,VOLUME,OPEN_POS,DIRECTION

где #SYMBOL – название финансового инструмента;

MOMENT – время (дата);

PRICE_DEAL – цена.

Пример строки в файле:

SVH1,F,20110111100000080,255223067,30.46000,1,8714,S

Данный файл отсортирован по дате и времени.

Внимание! Так как данный файл с финансовыми данными очень большого размера, отладку программы следует производить на данных меньшего размера и только убедившись, что все работает, запускать программу на данных большого размера.

Указанный файл с финансовыми данными меньшего размера Вы можете скачать по адресу http://bigdata.cs.msu.ru/images/9/99/Finance_example.zip

Вам необходимо:

Привести данные к формату японских свечей.

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.

Входными параметрами программы являются:

- candle.width = 300000 #"ширина" свечи в числе миллисекунд;
- candle.securities = ".*" #шаблон инструментов – задается в виде регулярного выражения;
- 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.

Свечи "начинаются" в моменты времени, кратные "ширине".

На выходе необходимо получить директорию с файлами. Имя каждого файла равно SYMBOL (расширение csv). Каждый файл должен быть отсортирован по MOMENT.

Формат выходных данных (каждого файла) без шапки:

SYMBOL,MOMENT,OPEN,HIGH,LOW,CLOSE

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 Task3-Фамилия.rar (фамилия на англ.), содержащий следующие файлы:

1) Файл Candles.java с Вашим кодом;
2) Файл Candles.jar;
3) Файл Candles.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 полученную в результате директорию Candles.

Входные данные для дополнительного запуска:

- candle.width = 300000;
- candle.date.from = 20110111;
- candle.date.to = 20110112;
- candle.time.from = 1000;
- candle.time.to = 1015.

Необходимо провести дополнительный запуск на финансовых инструментах SVH1 и GDH1.