Домашнее задание №1

Японские свечи

Принцип работы программы

Программа реализована на языке программирования Java с использованием фреймворка Hadoop. Вычисление свечей происходит в один map-reduce проход с использованием комбайнера. Первым этапом на этапе Map строки csv-файлов переводятся в пару ключ-значение для определения текущего состояния свечи. На этапе Reduce происходит объединение свечей с общим ключом и печать в файлы с группировкой по финансовому инструменту. Для уменьшения нагрузки на редьюсер перед этапом Reduce используется комбайнер, практически совпадающий по алгоритму с редьюсером за исключением печати значений в файлы.

Специальные классы

Для хранения и передачи ключа и значения в mapreduce реализованы классы CandleKey и CandleValue соответственно.

Ключ — CandleKey

Аттрибутами класса являются символ финансового инструмента SYMBOL и время начала свечи, которые при известном начале отсчета времени и ширине заданных в конфигурации однозначно определяют свечу.

Для CandleKey реализованны интерфейс WritableComparable, а так же переопределены методы toString для печати и hashCode для корректного разбиения по редьюсерам стандартным HashPartitioner.

Значение — CandleValue

Аттрибутами класса являются id первой и последней транзакции в свече для корректного нахождения цен открытия и закрытия, а также сами значения цен: открытия OPEN, закрытия CLOSE, минмальная LOW, максимальная HIGH.

Для CandleValue реализован интерфейс Writable, а так же переопределен метод toString для печати.

Этап Map

На этапе Map происходит преобразование строк исходных csv-файлов в пары ключ-значение. Для ключа берется названия финансового инструмента SYMBOL, а время

начали свечи вычисляется как модуль по ширине свечи при вычислении в милисекундах. В значение записываются все одинаковые значения: цена транзакции в данной строке.

Так же происходит фильтрация свечей:

- Название финансового инструмента проверяются по regexp.
- Время проверяется на вхождение в заданные рамки.
- Заголовок csv игнорируется.

Этап Combine

Эквивалентен этапу Reduce за исключением записи в файлы.

Этап Reduce

Этап reduce получает все значения транзакций для конкретной свечи в виде свечей с одинаковыми значениями (либо свечи вычисленные на подмножестве транзакций в Combine) и итеративно проходя по этим значениям собирает финальные значения:

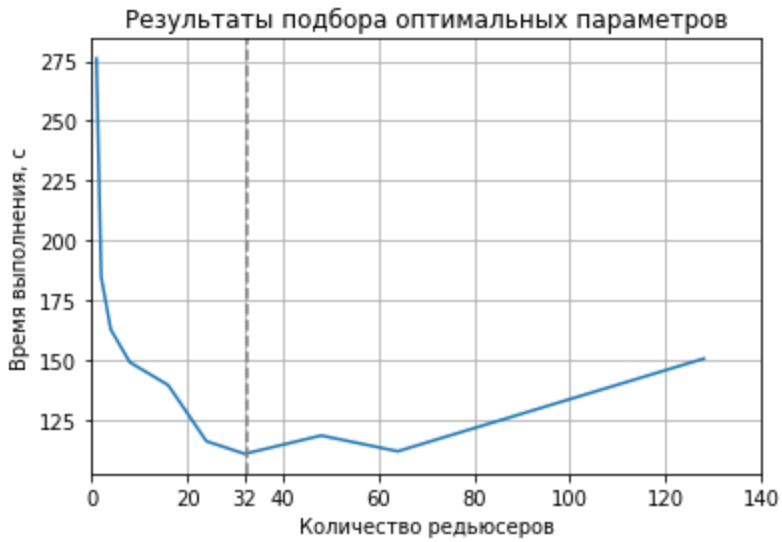
- При нахождении наиболее ранней цены, id первой транзакции и цена открытия обновляются
- При нахождении наиболее поздней цены, id последней транзакции и цена закрытия обновляются
- При нахождение более высокой цены, текущая максимальная цена обновляется
- При нахождение более низкой цены, текущая минимальная цена обновляется

Затем значение финальной полученной свечи пишется в файл с названием финансового инструмента в формате csv.

Подбор оптимальных параметров

Для подбора оптимальных параметров программа запускалась в Hue с разными параметрами количествоа редьюсеров. Перебор проводился по степеням двойки вплоть до 128 редьюсеров. Затем в окрестности наиболее оптимального значения было проведено еще два эксперимента.

Оптимальным количеством редьюсеров оказалось число 32.



Выводы

По итогам задание была реализована программа для решения поставленной задачи. Был проведен эксперимент по подбору оптимального количества редьюсеров. Оптимальным количеством оказалось 32 редьюсера. График привден в предыдущем разделе.