Задание 11

Цель – разработать систему запуска задач MapReduce. Не предполагается запуск внешних задач для отображения (map) и свёртки (reduce) (т. е. на вход принимаем только C++ функции или функторы, не нужно реализовывать запуск внешних скриптов).

На вход подаётся тестовый файл F, в котором каждая строка содержит адрес электронной почты. Для дальнейшей обработки не имеет значения, что именно содержат строки – самое важное, что обработка выполняется построчно.

В качестве дополнительных параметров также указывается количество потоков для запуска функций отображения (M) и свертки (R) соответственно.

На первом этапе необходимо выполнить разделение исходного файла на секции по числу потоков отображения M. При этом нужно следить, чтобы границы секций были выровнены по строке. То есть, каждая секция заканчивалась целой строкой и начиналась с начала новой строки. Таким образом мы не допустим, чтобы один из email-адресов оказался на границе секции и частично попал в две секции сразу. Чтение файла на этом этапе недопустимо, за исключением минимально возможного для выравнивания по строке.

На этапе работы отображения (map) следует запустить M потоков, каждый из которых обрабатывает свою секцию, полученную после разделения (split) исходного файла. Задача потока – построчно считать свою секцию и каждую строчку отправить в пользовательский функциональный объект. Результатом работы такого объекта будет новый список строк для стадии свертки (reduce). Полученные списки накапливаются в контейнере и затем сортируются. Для каждой секции получаем свой контейнер с отсортированными результатами.

Как только все потоки отображения будут завершены, необходимо запустить операцию смешивания (shuffle) и приготовить R контейнеров для будущей свертки. Общая задача на этапе смешивания – переместить строки из M контейнеров (результат этапа map) в R контейнеров (входные данные для этапа reduce). При этом сделать это нужно таким образом., чтобы одинаковые данные попали в один и тот же контейнер для свёртки. Важно, чтобы контейнеры для свёртки остались отсортированными. Необходимо понять, как реализовать объединение отсортированных последовательностей.

Как только shuffle будет завершён, должны будут запуститься R потоков для свертки (reduce). Каждый поток построчно отправляет данные из контейнера в пользовательский функциональный объект. Результатом работы такого объекта будет список строк, который должен быть сохранен в файл без какой-либо дальнейшей обработки.

К моменту завершения работы всех потоков свертки в файловой системе должны сформироваться R файлов с результатами.

Требования к реализации

Результатом работы должна стать система запуска задач MapReduce. С помощью этой системы нужно решить задачу определения минимально возможного префикса, однозначно идентифицирующего строку. Для этого потребуется написать два функциональных объекта – для отображения (map) и свертки (reduce).

Порядок запуска:

 # mapreduce <src> <mnum> <rnum>

, где:

- **src** – путь к файлу с исходными данными
- **mnum** – количество потоков для работы отображения
- **runm** – количество потоков для работы свертки

Проверка

Задание считается выполненным успешно, если после установки пакета и запуска с тестовыми данными вывод соответствует ожидаемому.