Задание 14

Условие

Разработать систему запуска задач MapReduce. Не предполагается запуск внешних задач для отображения и свертки.

На вход подается текстовый файл F, в котором каждая строка содержит электронный адрес. Для дальнейшей обработки не имеет значения что именно содержат строки - самое важное, что обработка выполняется построчно.

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

На первом этапе необходимо выполнить разделение исходного файла на секции по количеству потоков отображения M, при этом следить, чтобы границы секций были выровнены по строке, то есть каждая секция заканчивалась целой строкой и начиналась с начала строки. Чтение фала на этом этапе недопустимо, за исключением минимально возможного для выравнивания по строке.

На этапе работы отображения необходимо запустить M потоков, каждый из которых параметризуется параметрами своей секции, полученной после разделения (split) исходного файла.

Задача потока построчно читать свою секцию и каждую строку отправлять в пользовательский функциональный объект. Результатом работы такого объекта будет новый список строк для стадии свертки. Полученные списки накапливаются в контейнере и сортируются. Для каждой секции свой контейнер. Предполагаем, что памяти достаточно для хранения всех строк.

Как только все потоки отображения будут завершены, необходимо запустить M потоков смешивания (shuffle) и приготовить R контейнеров для будущей свертки. Общая задача на этапе смешивания переместить строки из M контейнеров в R контейнеров, при этом сделать это таким образом, чтобы одинаковые из них гарантированно попали в один и тот же контейнер для свертки. Важно, чтобы контейнеры для свертки остались отсортированные. Необходимо понять, как использовать объединение сортированных последовательностей.

Как только все потоки смешивания завершат свою работу, должны будут запуститься R потоков для свертки. Каждый поток построчно отправляет контейнер в пользовательский функциональный объект. Результатом работы такого объекта будет список строк, который немедленно должен быть сохранен в файл.

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

При помощи полученного инструментария решить задачу определения минимально возможного префикса однозначно идентифицирующего строку. Для это потребуется написать два функциональных объекта - для отображения и свертки.

Порядок запуска:

```
# yamr <src> <mnum> <rnum>
```

где

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

Требования к реализации

Результатом работы должен стать исполняемый файл `yamr` и находиться в пакете `yamr`.

Результат работы должен быть опубликован на bintray.

Проверка

Задание считается выполненным успешно, если после установки пакета и запуска с тестовыми данными существует ответ на вопрос - как по полученным результатам узнать минимально необходимый префикс.