МГУ имени М.В. Ломоносова
Объединенная компания «Афиши» и «Рамблера»

Применение NoSQL для построения рекомендательных сервисов реального времени

Павел Клеменков
p.klemenkov@rambler-co.ru

НОВОСТИ

Главное

У вас есть непрочитанные новости

Поиск по новостям

1 ЧАС НАЗАД

Удальцова вывели из квартиры в сопровождении спецназа

Интерфакс и еще 125 источников

Бойцы спецназа МВД вывели оппозиционера Сергея Удальцова из квартиры и сопровождают в микроавтобус, передает корреспондент "Интерфакса".

4 ЧАСА НАЗАД

Астрономы обнаружили похожую на Землю планету

YTPO.ru и еще 1 источник

1 ЧАС НАЗАД

Лукашенко обвинил российского олигарха во взятке

BFM.RU
и еще 27 источников

5 ЧАСОВ НАЗАД

Геннадий Гудков нашел способ вернуться

YTPO.ru
и еще 5 источников

4 ЧАСА НАЗАД

На московские вокзалы отправят православных миссионеров

Москва 24
и еще 7 источников

СЕГОДНЯ

Москвичам обещают насолить как следует

kommersant.ru
и еще 1 источник

2 ЧАСА НАЗАД

В бунтующую исправительную колонию в Петербурге введен спецназ

Lenta.ru
и еще 13 источников

4 ЧАСА НАЗАД

«Нужен игрок более высокого уровня, чем Кержаков!»

Советский спорт
и еще 2 источника

2 ЧАСА НАЗАД

В продуктовых магазинах разрешат работать гастарбайтерам

YTPO.ru
и еще 8 источников

2 ЧАСА НАЗАД

Следователи опять ищут виновных в ДТП на Ленинском

РБК.Daily
и еще 36 источников

Ещё новости

Прочитано
Россия — Азербайджан
Евгений Касперский
Анастасия Завгородняя
Аркадий Мамонтов
Убийство россиянки в Норвегии
Главное
Политика
Мир
Экономика
Общество
Происшествия
Спорт
Авто
Наука и техника
Культура
Образ жизни
Шоу-бизнес
Как бы новости
Москва

НОВОСТИ

YTPO.ru

Вблизи Солнечной системы астрономы обнаружили планету похожую на землю

4 ЧАСА НАЗАД, 09:44

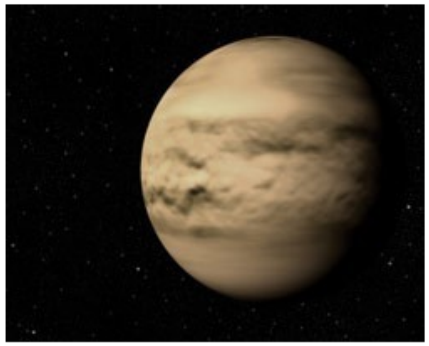


Фото: thinkstock.com

Европа Мир Наука Наука и техника

15 8 13 21

Команда европейских астрономов обнаружила похожую на Землю планету вблизи Солнечной системы. Об этом сообщают зарубежные СМИ. Планета соответствует расположению и размеру Земли, однако она намного горячее и ее поверхность, вероятно, покрыта лавой.

Планета находится в звездной системе Альфа Центавра В, расположенной в 40 трлн км от Солнечной системы. Астрономы предполагают, что рядом могут быть другие планеты, которые, возможно, не будут ни слишком горячими, ни слишком холодными. Ранее в этом месяце астрофизики Йельского университета (США) и Университета Тулузы (Франция) обнародовали информацию, что поверхность открытой в 2004 г. планеты «55 Рака», по их данным, покрыта алмазами.

Читать на сайте YTPO.ru

Комментарии к новости 198

Поиск по новостям

1 ЧАС НАЗАД
Удальцова вывели из квартиры в сопровождении спецназа
Интерфакс
и еще 125 источников

5 ЧАСОВ НАЗАД
Геннадий Гудков нашел способ вернуться
YTPO.ru
и еще 5 источников

1 ЧАС НАЗАД
Лукашенко обвинил российского олигарха во взятке
BFM.RU
и еще 27 источников

СЕГОДНЯ
Москвичам обещают насолить как следует
kommersant.ru
и еще 1 источник

4 ЧАСА НАЗАД
На московские вокзалы отправят православных миссионеров
Москва 24
и еще 7 источников

4 ЧАСА НАЗАД
«Нужен игрок более высокого уровня, чем Кержаков!»
Советский спорт
и еще 2 источника

Прочитано

Россия — Азербайджан

Евгений Касперский

Анастасия Завгородняя

Аркадий Мамонтов

Убийство россиянки в Норвегии

Главное

Политика

Мир

Экономика

Общество

Происшествия

Спорт

Авто

Наука и техника

Культура

Образ жизни

Шоу-бизнес

Как бы новости

Москва

Мера Жаккарда

$$J = \frac{|A \cap B|}{|A \cup B|}$$

Мера Жаккарда

$$J = \frac{|A \cap B|}{|A \cup B|}$$

$$O(min(|A|,|B|))$$

$$O(|A|+|B|)$$

MinHash

$$A = \{ url_1, url_2, \ldots, url_k \}$$

$$h : A \rightarrow \mathbb{N}$$

$$h_{min}(A) = min_{x \in A}(h(x))$$

$$J(A, B) = P[h_{min}(A) = h_{min}(B)]$$

Особенности MinHash

- Сигнатура множества вычисляется один раз
- Размер сигнатуры не меняется
- Одна хеш-функция имеет высокую дисперсию, поэтому нужно использовать среднее нескольких хеш-функций

Реализация. Прототип

- Один процесс
- Все логи загружаются в память
- Невозможно масштабировать вертикально (память)
- Невозможно масштабировать горизонтально (модель вычислений)

Реализация. Hadoop

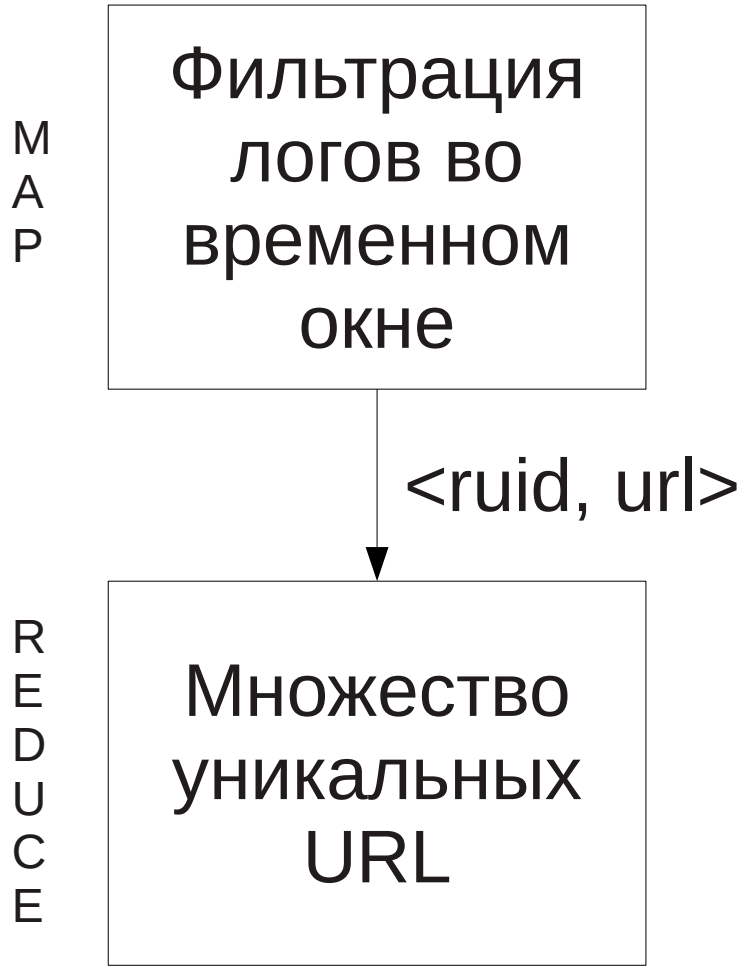
- Высокая скорость поточной обработки больших блоков данных
- Нет ограничений на размер временного окна
- Горизонтальное масштабирование (вычислительная модель MapReduce)

Реализация. Hadoop. Кластеризация

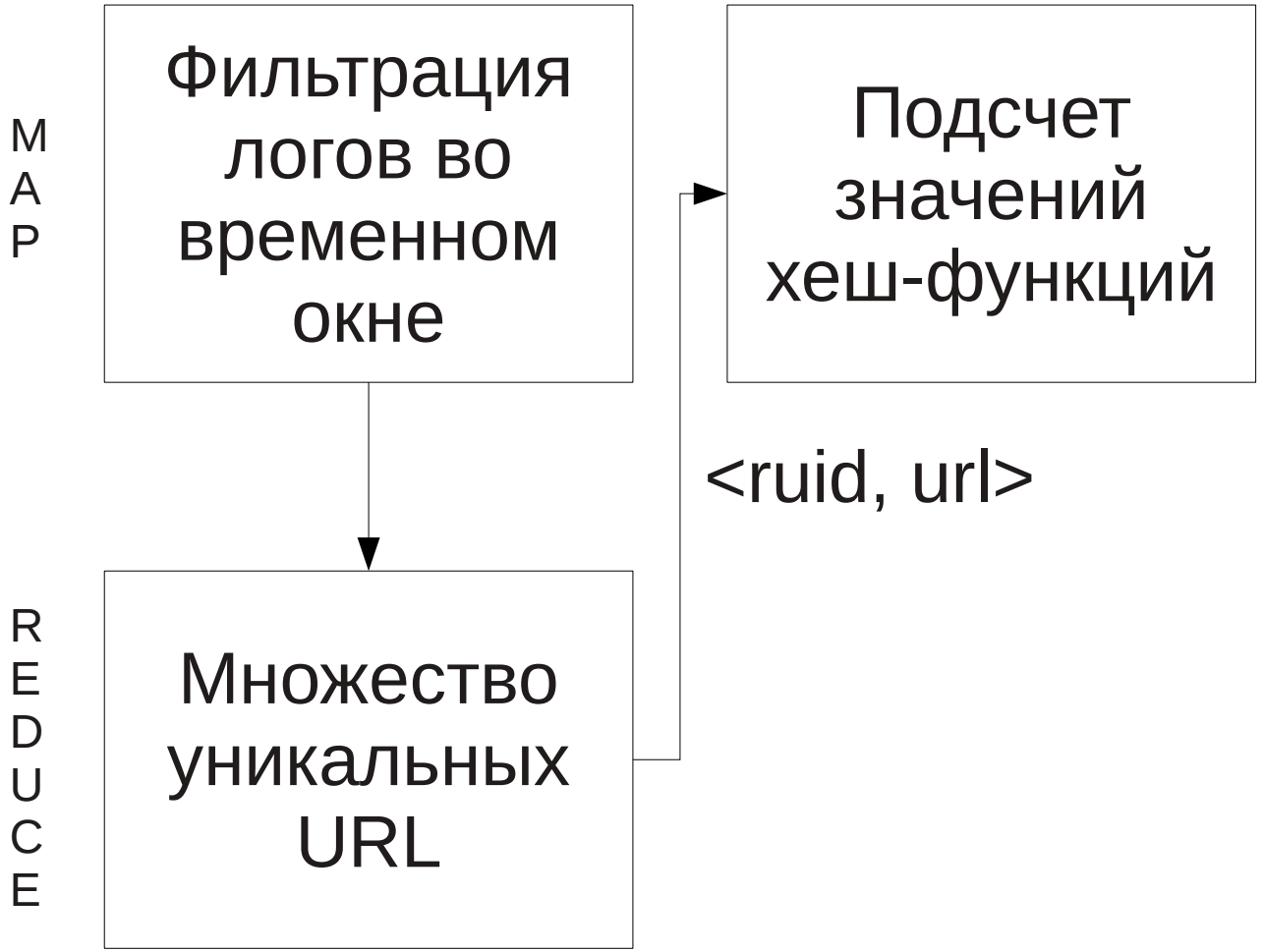
MAP

Фильтрация логов во временном окне

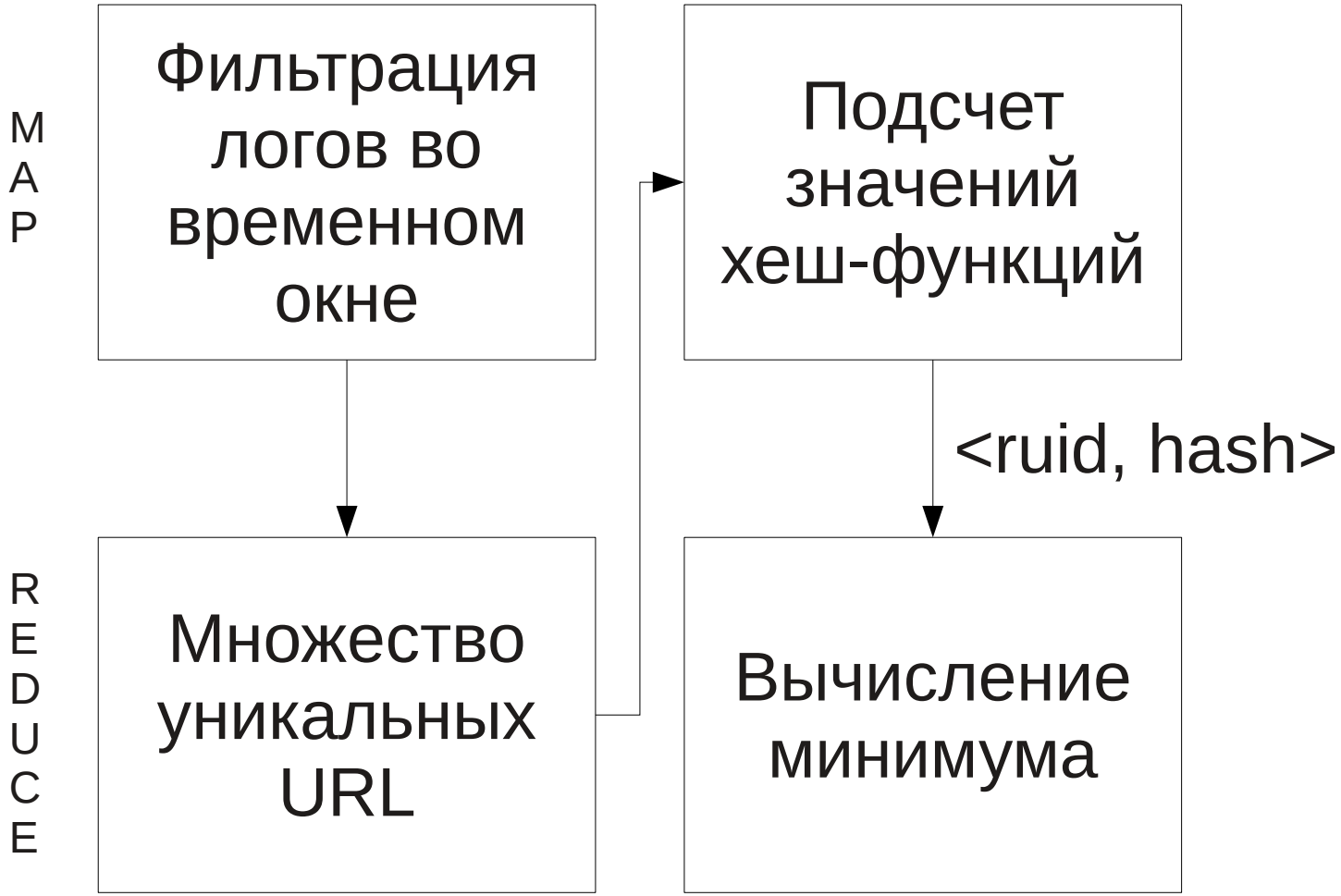
Реализация. Hadoop. Кластеризация



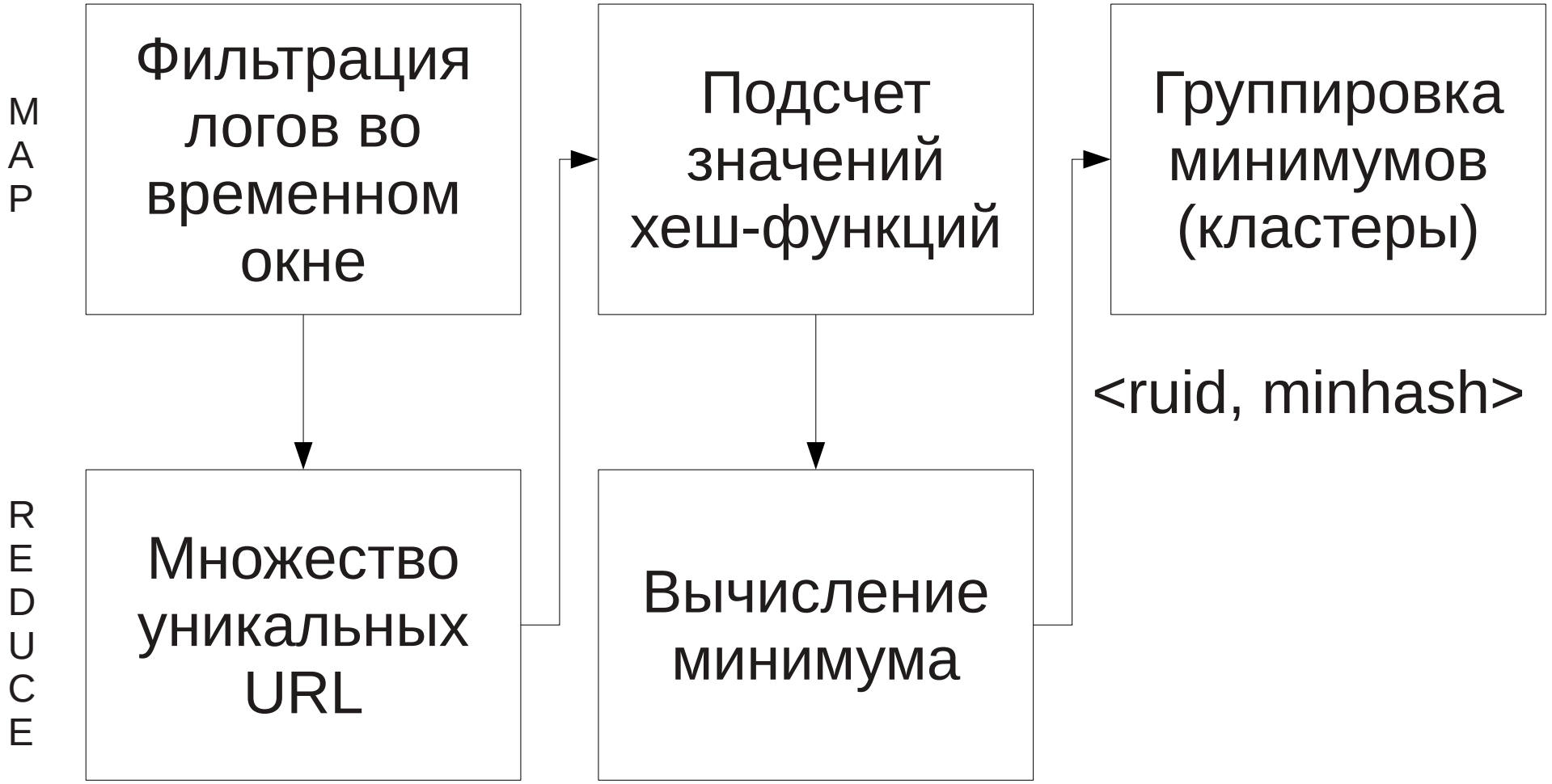
Реализация. Hadoop. Кластеризация



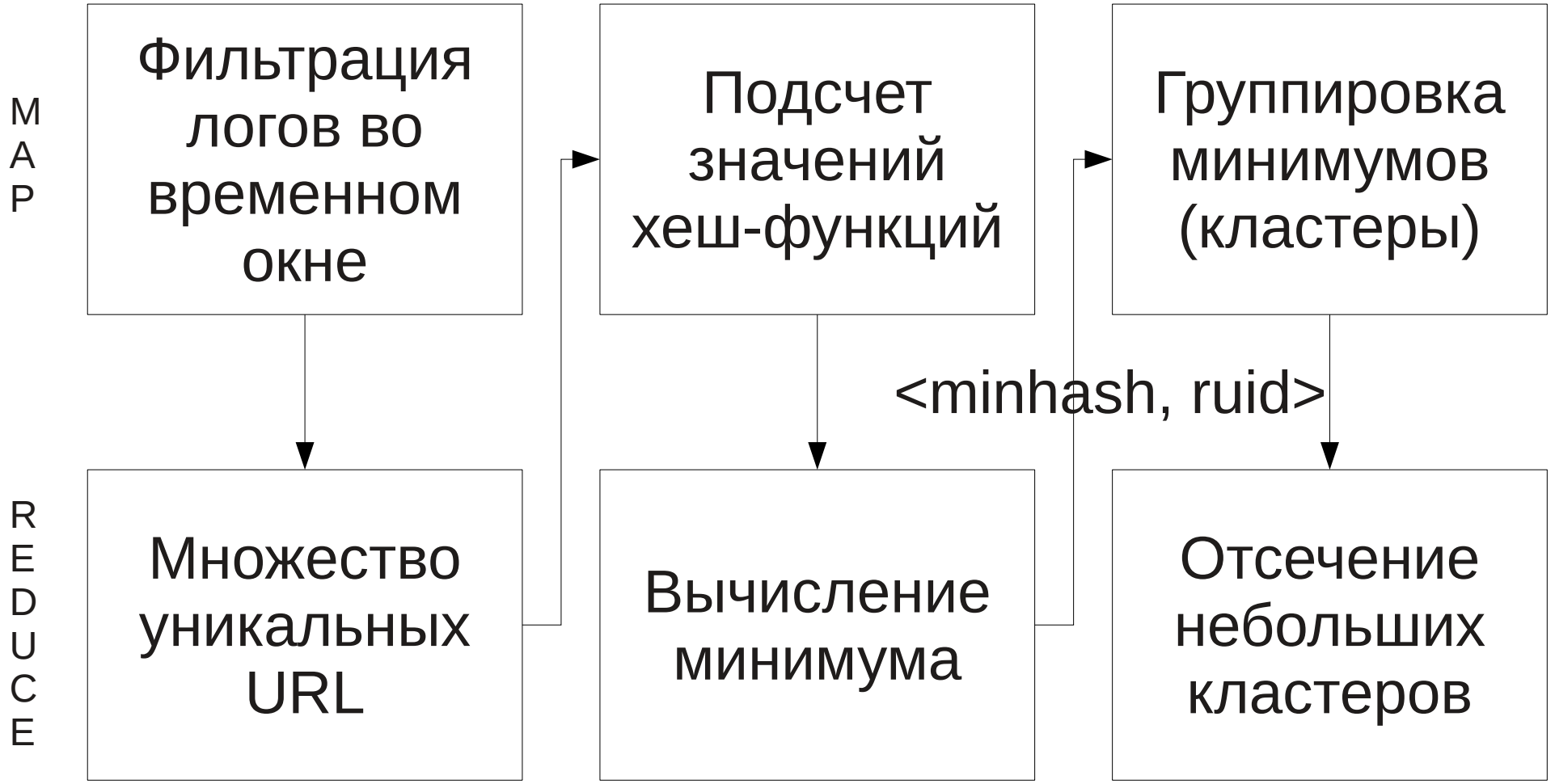
Реализация. Hadoop. Кластеризация



Реализация. Hadoop. Кластеризация



Реализация. Hadoop. Кластеризация

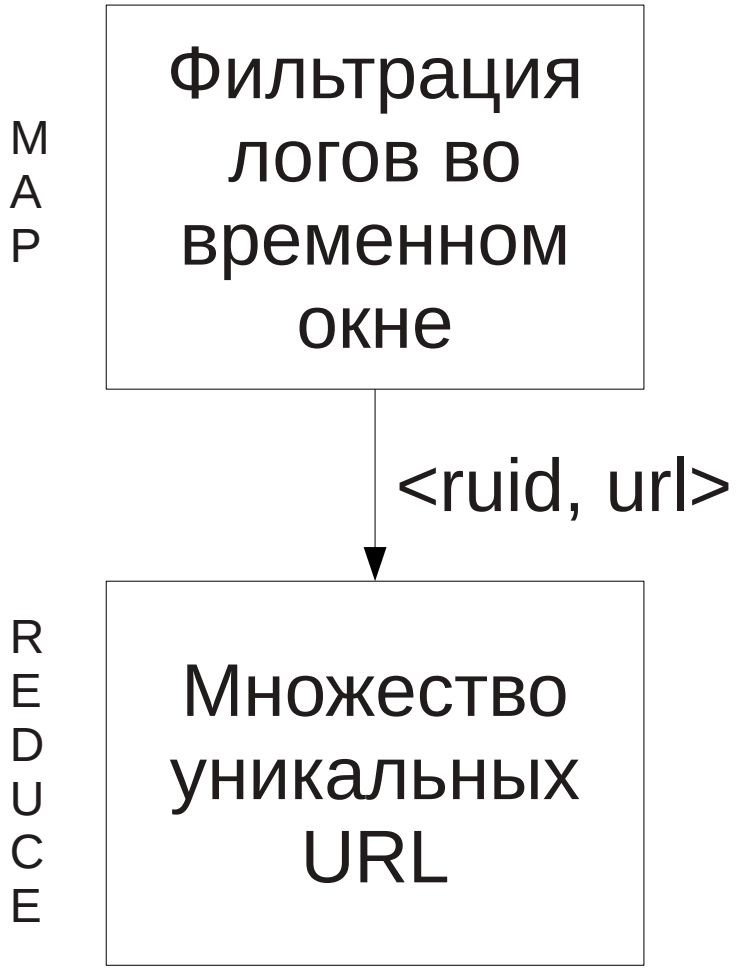


Реализация. Hadoop. Рекомендации

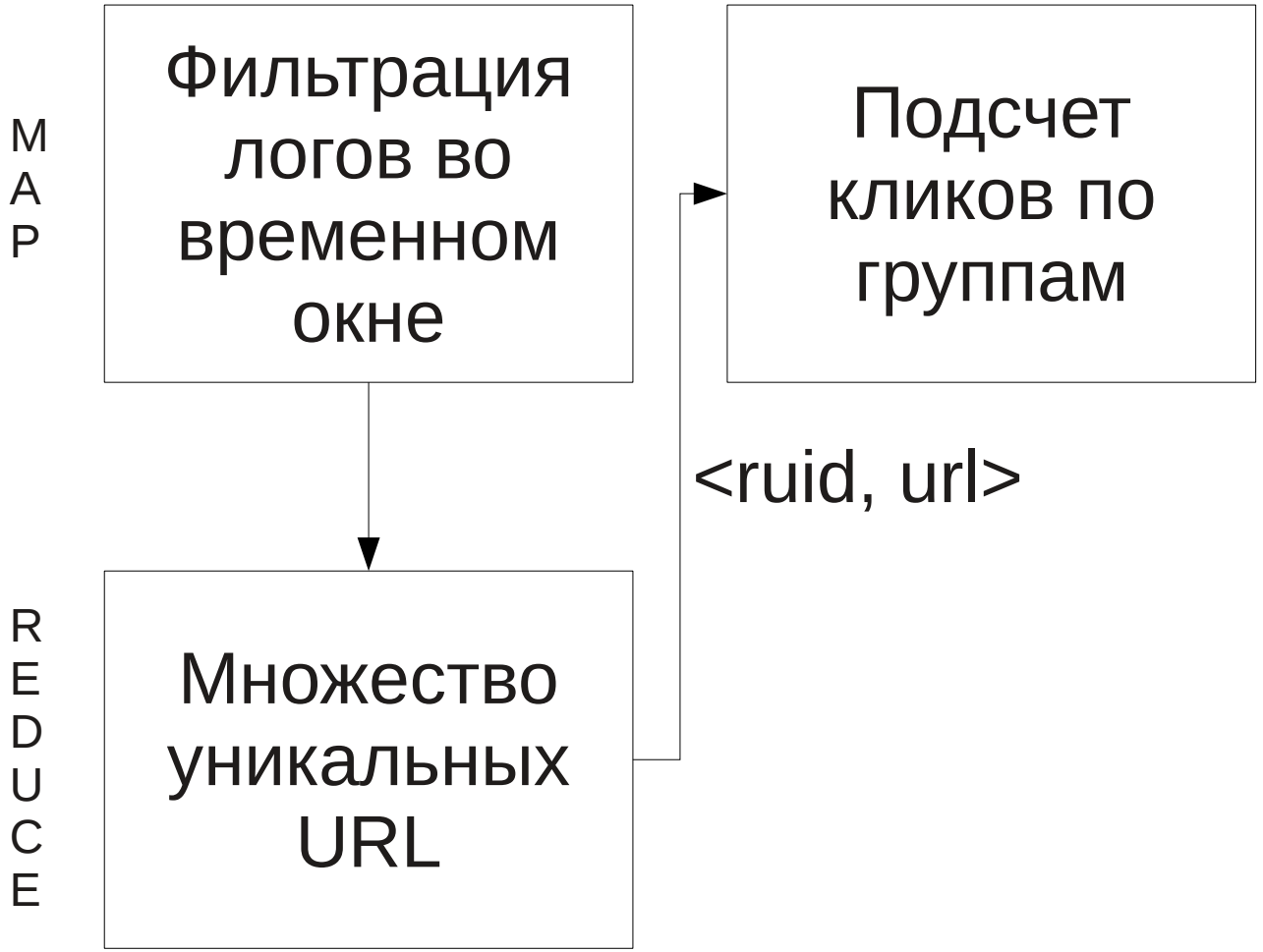
MAP

Фильтрация логов во временном окне

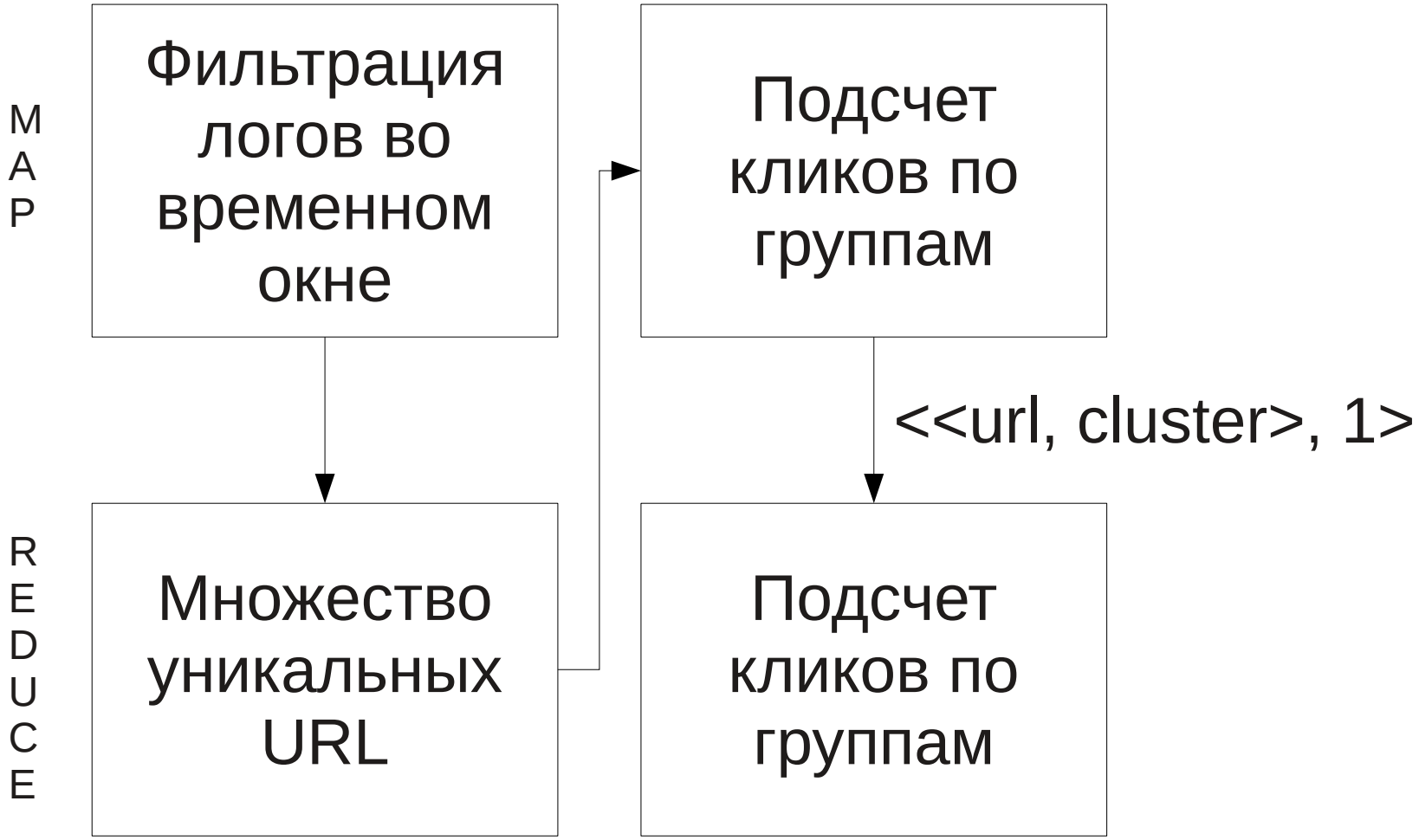
Реализация. Hadoop. Рекомендации



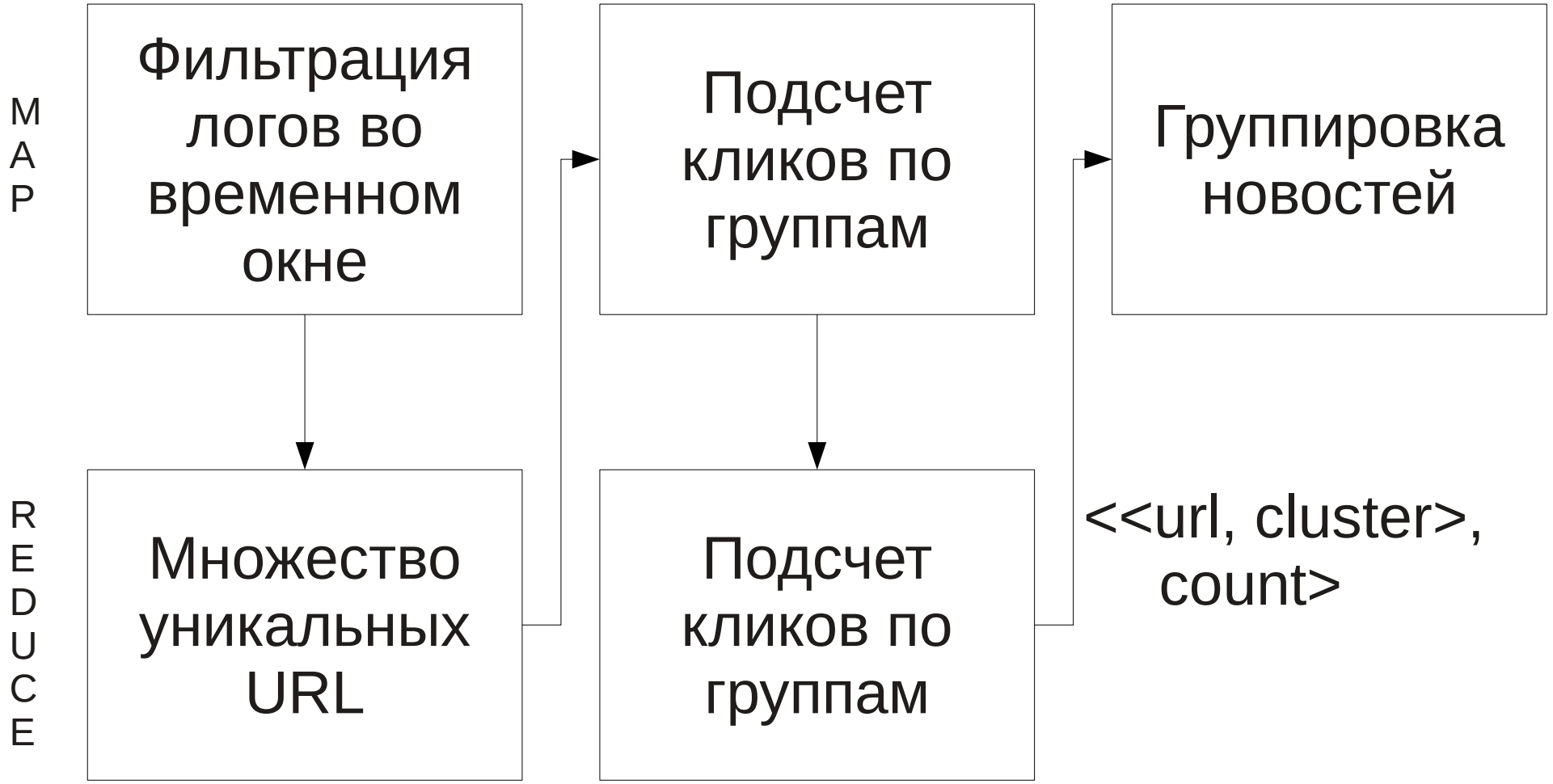
Реализация. Hadoop. Рекомендации



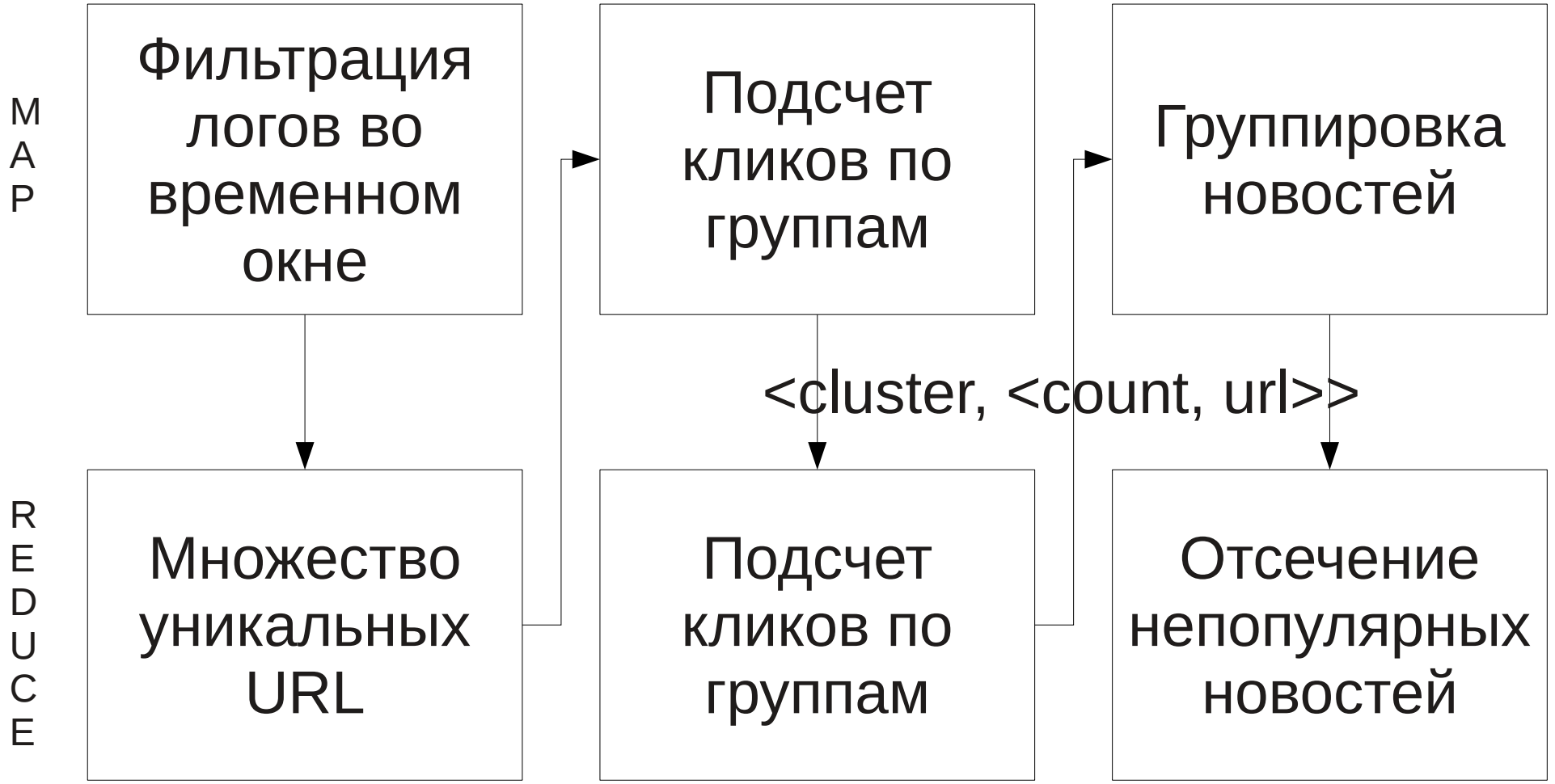
Реализация. Hadoop. Рекомендации



Реализация. Hadoop. Рекомендации



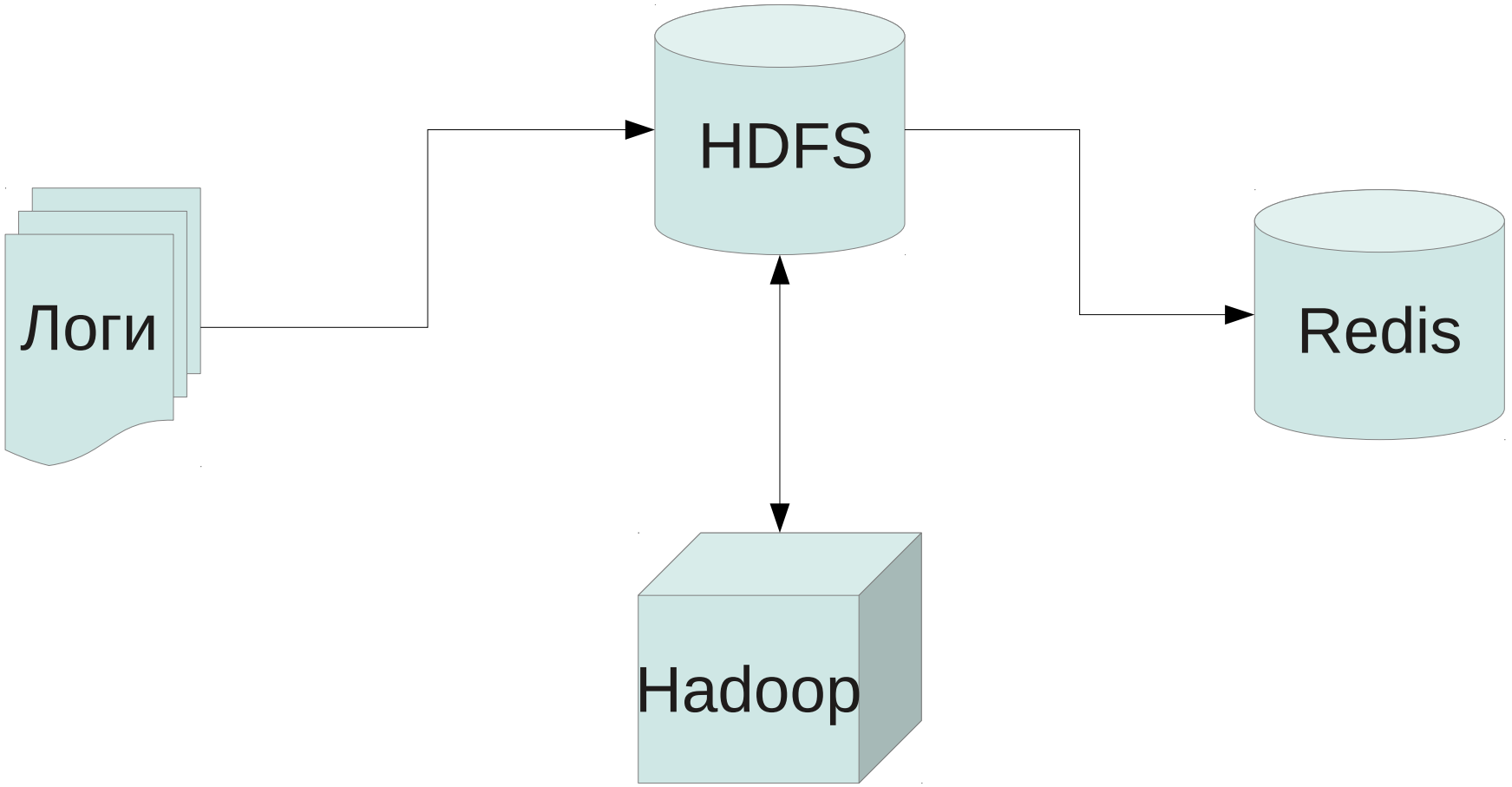
Реализация. Hadoop. Рекомендации



Реализация. Hadoop. Производительность

- Hadoop кластер из 8 узлов
- Временное окно кластеризации – 5 суток (8 ГБ логов)
- Время кластеризации – 7 минут
- Временное окно рекомендаций – 5 часов
- Время генерации рекомендаций – 3.5-4 минуты

Реализация. Hadoop. Архитектура



Реализация. Hadoop. Проблемы

- Загрузка логов в HDFS и их обработка – несвязанные задачи
- Обрабатывать данные больше чем из одного источника – проблематично
- Генерация рекомендаций и их использование – архитектурно разные задачи

Реализация. Hadoop. Проблемы

- Пакетный потоковый режим далек от реального времени
- Оптимален на больших объемах. Выполнение “пустого” задания ~ 40 сек

Pro NoSQL

- NoSQL – Not Only SQL
- Впервые термин использован в 1998 г.
- Популярность набирает с 2009 г. (Jon Oskarsson, Last.fm)
- Основной двигатель – высоконагруженный Веб

Особенности NoSQL

- Нет излишнему усложнению
- Высокая пропускная способность
- Неограниченное горизонтальное масштабирование
- Производительность важнее консистентности (зависит от задачи)
- Многие решения реализуют MapReduce

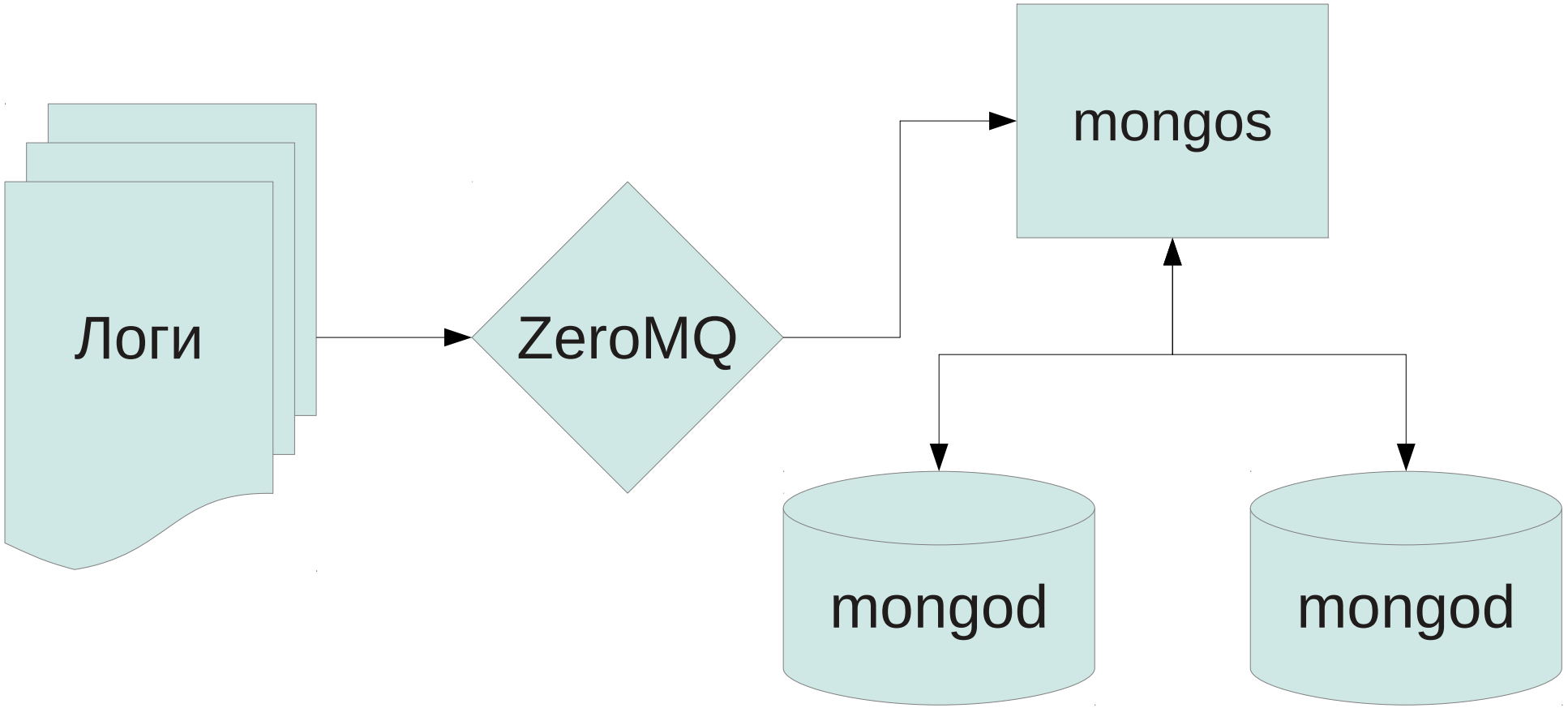
CouchDB

- JSON
- REST API
- B-деревья только на добавление
- Отсутствует язык запросов
- Отсутствуют ссылки на документы
- Избыточный трафик

MongoDB

- BSON
- Коллекции
- Мощный язык запросов
- Ссылки на документы (в том числе несуществующие)
- Индексы
- Шардинг

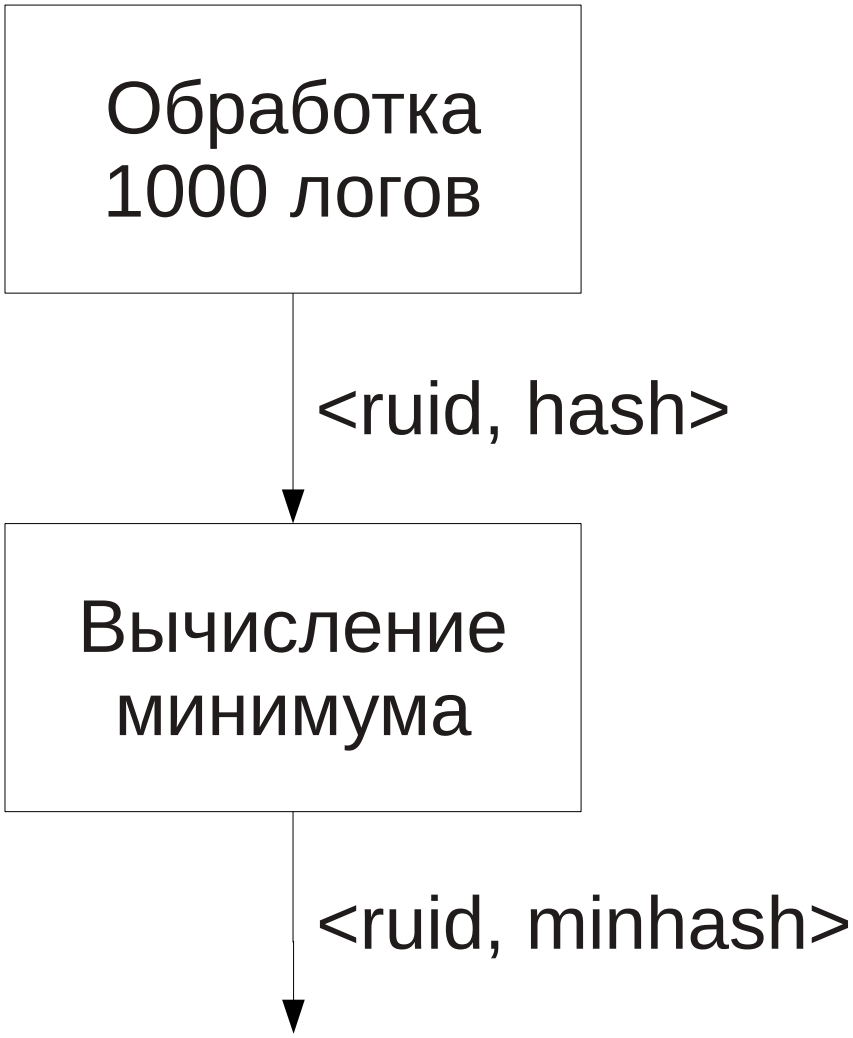
Реализация. MongoDB. Архитектура



Реализация. MongoDB. ZeroMQ

- Агрегация логов со всех фронтэндов
- Фильтрация логов
- Подсчет значений хеш-функций
- Запись обработанных логов в MongoDB
- 2500 оп/сек – скорость записи логов в базу (2 mongod с шардингом)
- Запуск MapReduce-задач

Реализация. MongoDB. Кластеризация



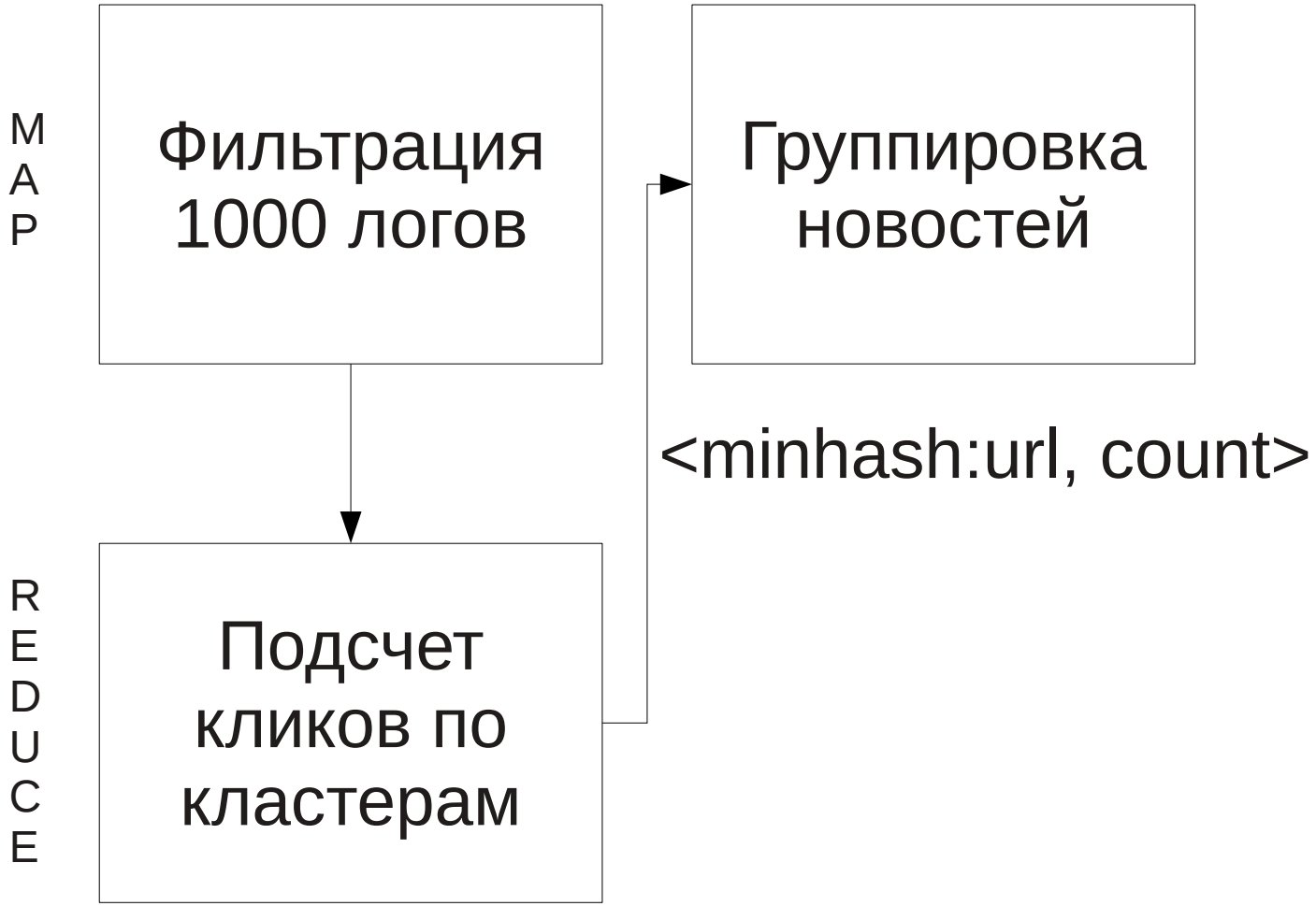
Реализация. MongoDB. Кластеризация

- Не отсекаются малочисленные кластеры
- Кластеризация суточного лога ~ 190 с
- Кластеризация 1000 записей ~ 3 с (без индекса)
- Кластеризация 1000 записей ~ 400-500 мс (с индексом)
- Индекс для выборки по временной метке

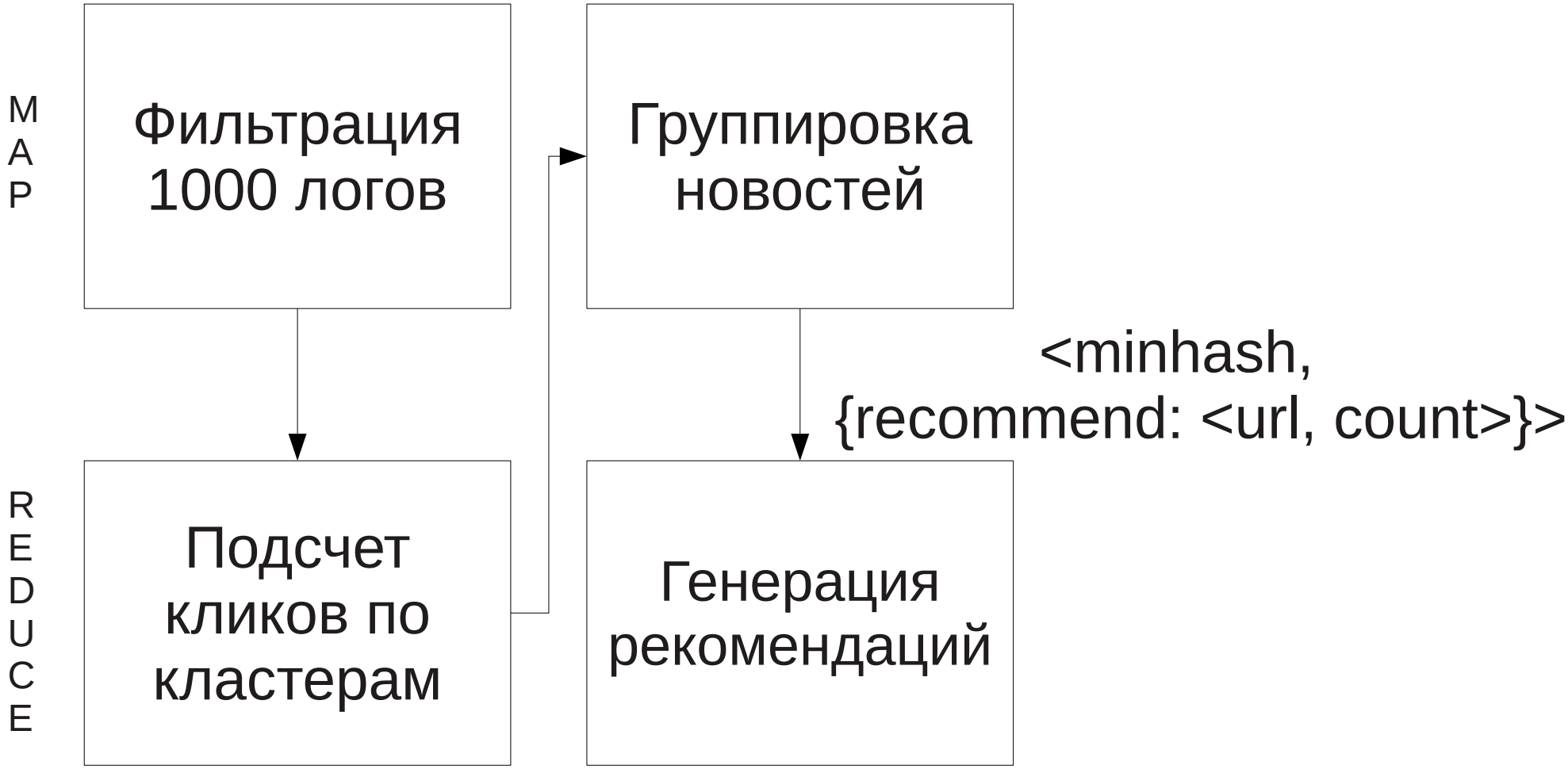
Реализация. MongoDB. Рекомендации

Реализация. MongoDB. Рекомендации

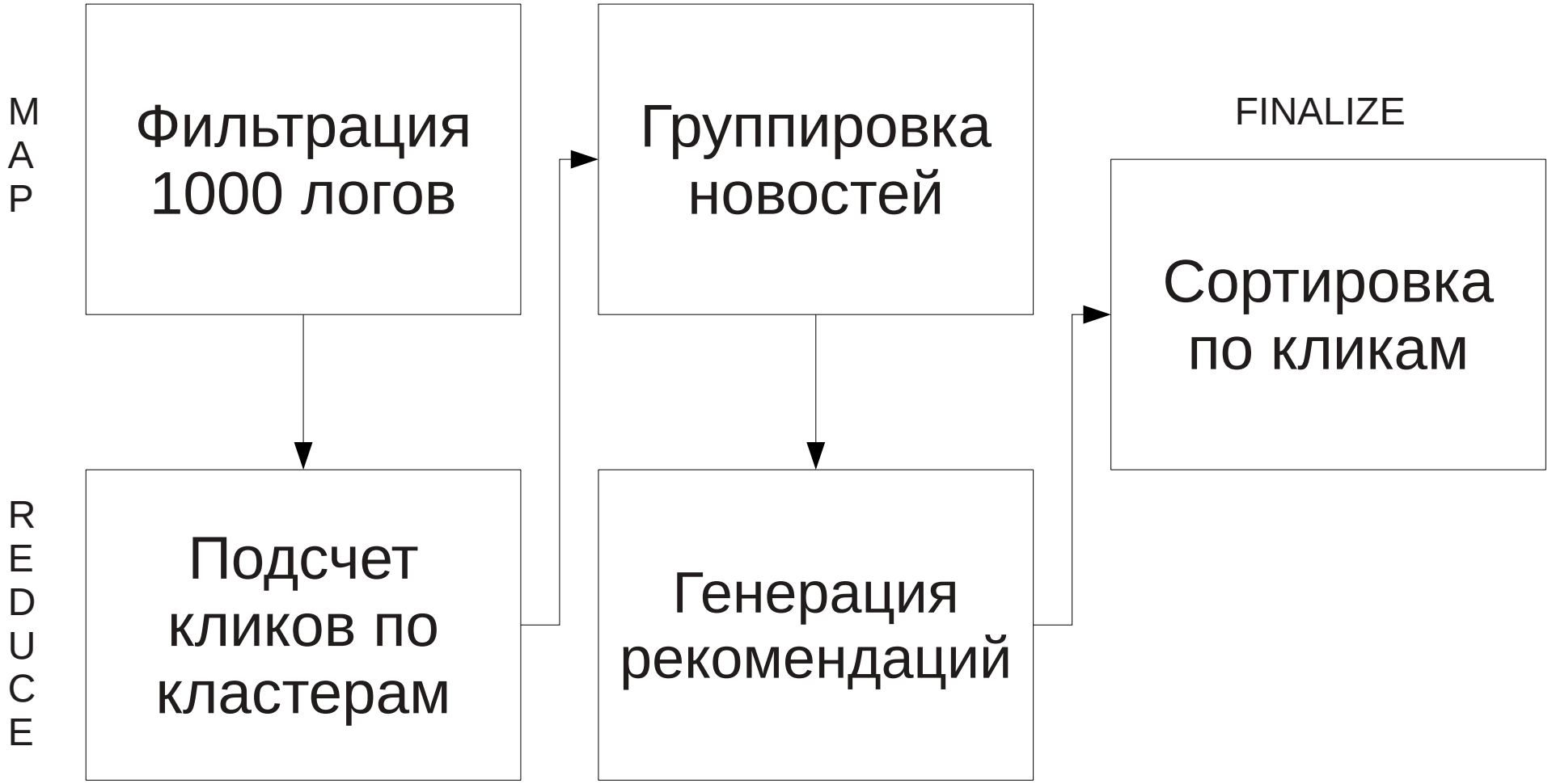
Реализация. MongoDB. Рекомендации



Реализация. MongoDB. Рекомендации



Реализация. MongoDB. Рекомендации



Реализация. MongoDB. Рекомендации

- Выборка логов по временной метке (индекс)
- Дополнительная фильтрация логов (JavaScript)
- Данные о кластерах из другой коллекции (DBRef)
- Сортировка рекомендаций по популярности
- Генерация рекомендаций по 1000 записей ~ 350 мс

Реализация. MongoDB. Проблемы

- Ротирование логов (избыточность данных)
- Сложные структуры (результаты map и reduce)
- Сложный синтаксис языка запросов
- Странный шардинг MongoDB
- JavaScript в коде на Python

Заключение

- Пересчет рекомендаций ~ 2-3 с на 1000 логов
- Качество рекомендаций (MinHash не подходит)
- Упрощение сбора повседневной статистики
- Рекомендации на графах (Covisitation, Neo4j)

Благодарность

- Добров Б.В., НИВЦ МГУ имени М.В. Ломоносова
- Кузнецов С.Д., ИСП РАН

Вопросы?