Разработка распределенной системы обработки данных

Дымникова Наталья, 344 гр.
Научный руководитель: Коновалов М.В.

Введение

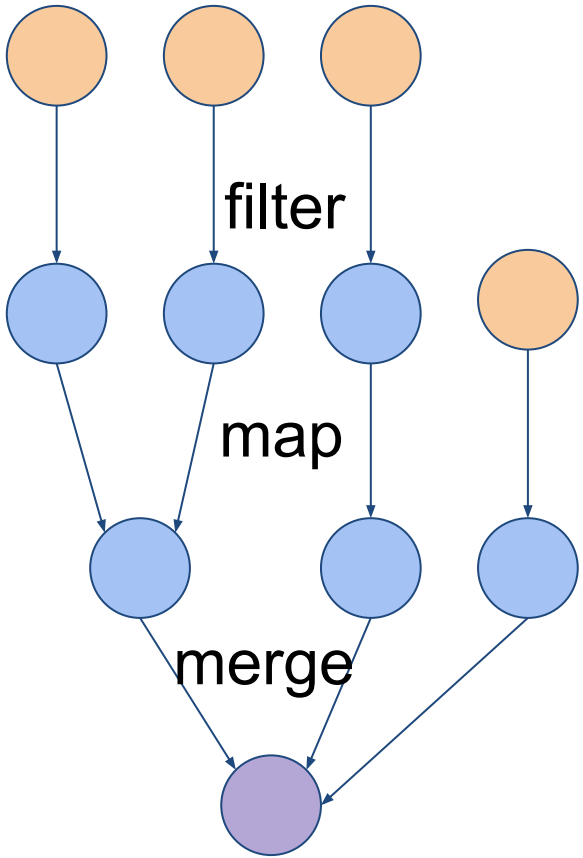
- Обработка данных требует больших ресурсов
- Решение — кластеризация

Постановка задачи

- Разработка распределенной системы обработки данных на кластере
- Создание интерфейса построения запросов к системе
- Реализация модуля нахождения наилучшей конфигурации запроса

Идея

- Потоки данных
  - java.Stream
  - rx.Observable

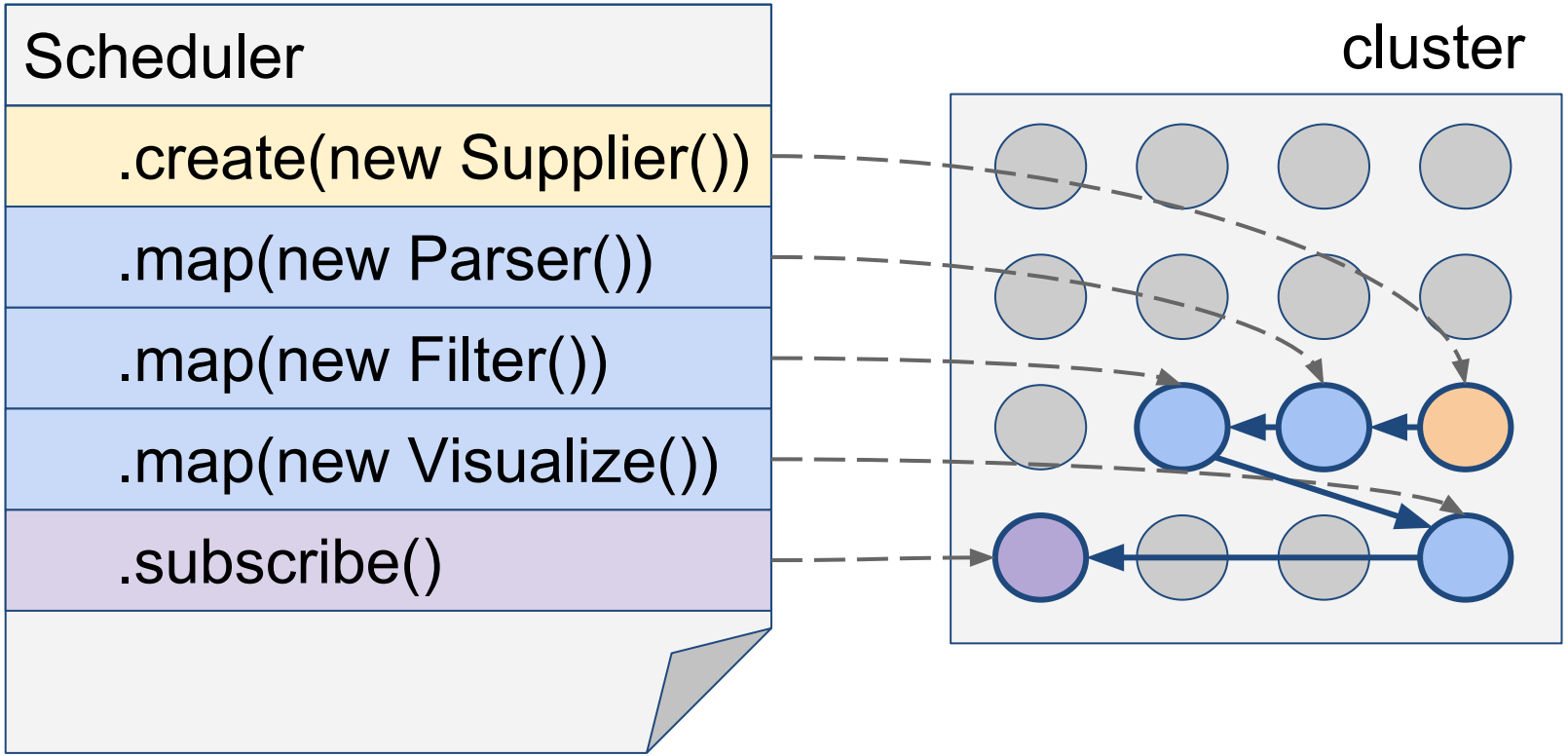


Архитектура

- Akka
- Scheduler

```
.createObservable(new Supplier())
.map(new Parser())
.map(new Filter())
.map(new Visualize())
.subscribe()
```

Scheduler


Scheduler
.create(new Supplier())
.map(new Parser())
.map(new Filter())
.map(new Visualize())
.subscribe()
cluster

Compute Pool

- Запрос разбивается на этапы
- На каждом узле ComputePool
- Контроль выполнения запроса

Конфигурация запроса

- Find first
- Round-robin
- Less loaded member

Особенности решения

- Pure Java
- Управление кластером
- Нет дублирования данных
- Нет немедленной индексации
- Не требует покупки коммерческих продуктов

Результат

- Разработана система распределенной обработки данных
- Создан интерфейс построения запросов к системе
- Реализован модуль нахождения различных конфигураций запроса
- Система протестирована и используется во внутреннем проекте компании EMC