MapReduce

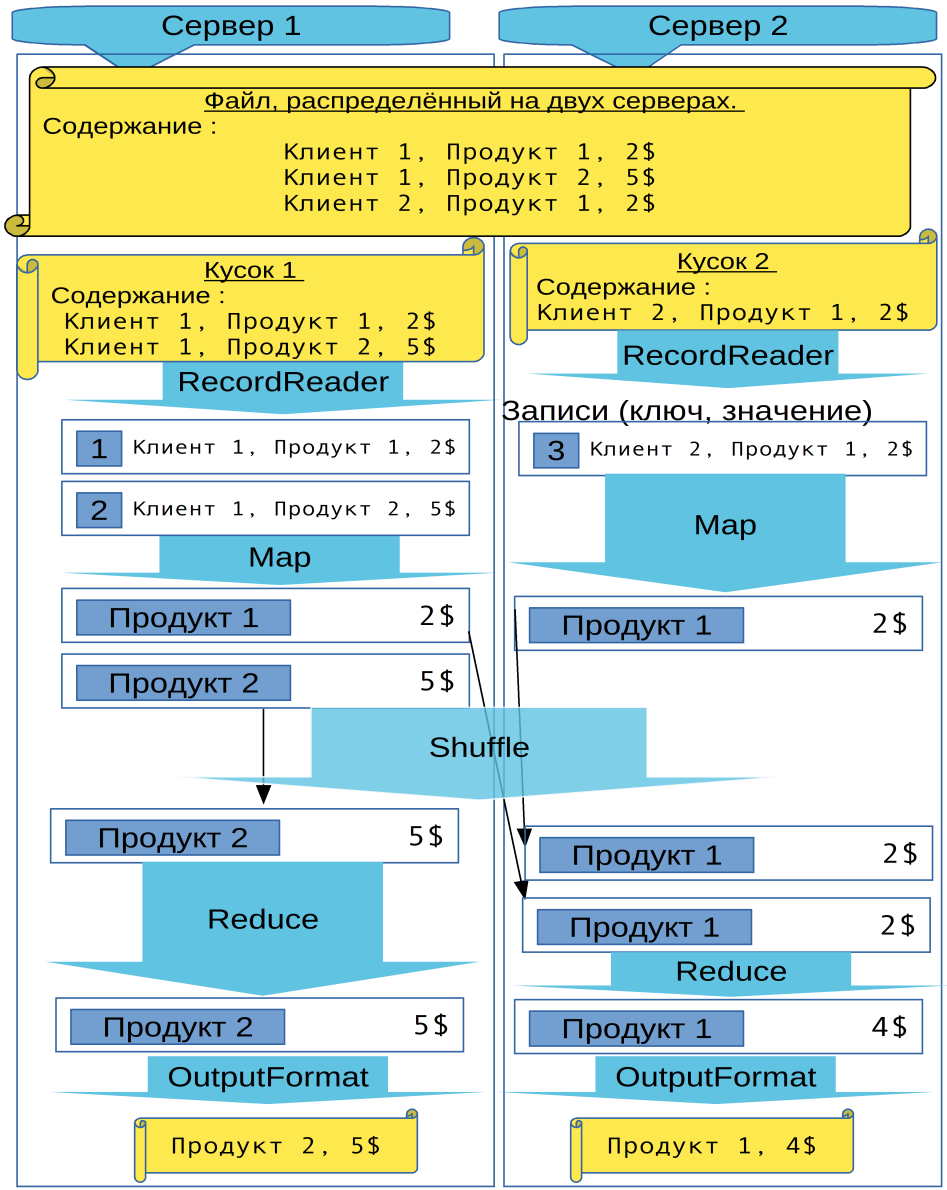
План семинара

1. Парадигма MapReduce
2. Разбор задачи wordcount (screencast)
3. Оптимизация и расширение возможностей

Материалы этого семинара

https://gitlab.com/pd2020-supplementary/8xx-GLOBAL/-/blob/master/practice/08-mapreduce_part1.md

Парадигма MapReduce



Парадигма MapReduce

Может быть воспроизведена **локально**:

```
cat input.txt | mapper.py | sort | reducer.py
```

Запуск **на кластере** (streaming):

```
yarn jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-
streaming.jar \
    -D mapreduce.job.reduces=${NUM_REDUCERS} \  # число reducer'ов
    -files mapper.py,reducer.py \  # копировать файлы на сегменты
    -mapper mapper.py \
    -reducer reducer.py \
    -input /data/wiki/en_articles_part \
    -output ${OUT_DIR}  # путь относительно домашней папки
```

См. также материалы семинара.

Wordcount на MapReduce

Задача: Для каждого слова подсчитать, сколько раз оно встречается в корпусе текстов.

1. **Входные данные**: Файл(ы) текстов
   - В Hadoop streaming — reader уже реализован
2. **Mapper**: Функция [файл текста] → [набор пар (слово, 1)]
3. **Shuffle + Sort**: Встроен в Hadoop
   - Любое отдельно взятый ключ (слово) попадает ровно на одну машину
4. **Reducer**: Функция [набор пар (слово, 1)] → [набор пар (слово, результат)]
   - Входные пары подаются в отсортированном порядке
   - Все пары с тем же ключом (словом) находятся на *этой* машине

Combiner

Оптимизация: выполнить операцию, аналогичную reduce, для подмножества ключей.

```
cat input.txt | mapper.py \
      | combiner.py | combiner.py | … \
      | sort | reducer.py
```

Счётчик

Можно определить «Глобальную переменную»:

- Запись: из любого mapper или reducer
- Чтение: после завершения задачи; агрегированная **сумма**

Использование в Streaming:

```
print("reporter:counter:Group,Name,{}".format(1), file=sys.stderr)
```

, где:

- `reporter:counter` — идентификатор user-defined счётчика
- `Group` — «группа» счётчика; `Name` — «имя» счётчика
- Последний элемент — число

Объединение задач MapReduce

Результаты MapReduce **сохраняются в HDFS**.
(как и результаты отдельных стадий MapReduce)

Их можно **подать на вход** следующей MapReduce задаче.

???

PROFIT

Полезные материалы

Документация Hadoop Streaming
https://hadoop.apache.org/docs/r1.2.1/streaming.html