ТЕХНОСФЕРА

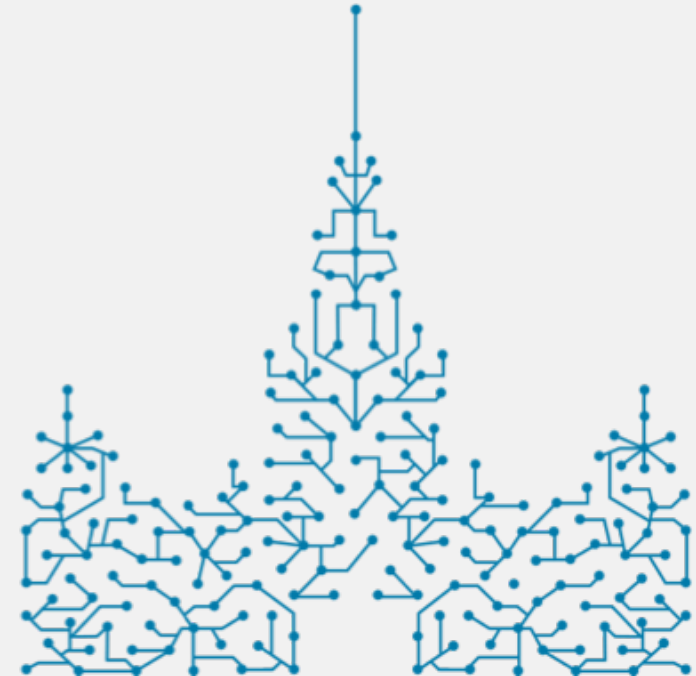
Практическое занятие №2

MapReduce Сортировка и Композитные ключи

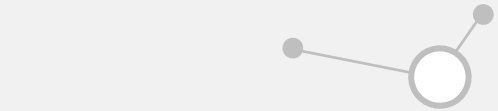
Кисель Ян

План занятия

- Логирование и отладка в MapReduce
- Устройство сортировки
- Вторичная сортировка
- ДЗ



Логирование и отладка в MapReduce



Логирование и счетчики

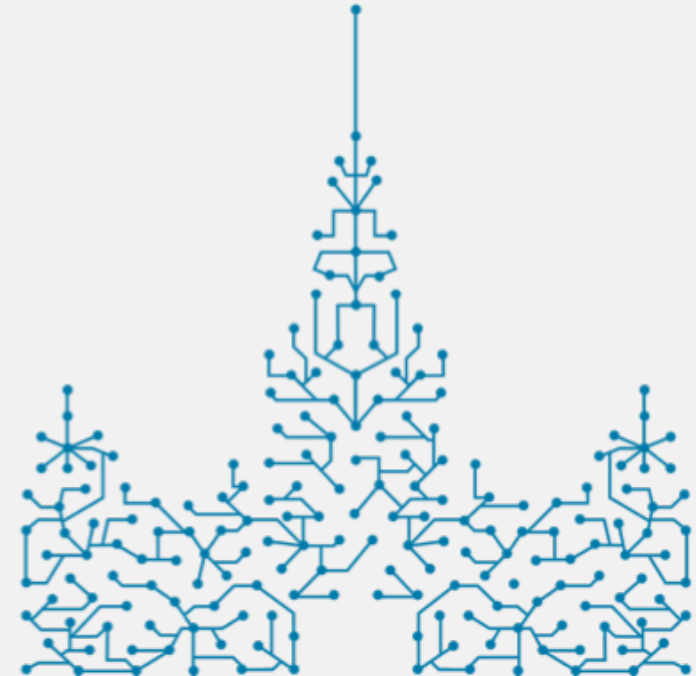
```
import org.slf4j.Logger;
import org.slf4j.LoggerFactory;

private final static Logger LOG =
  LoggerFactory.getLogger(MyClass.class);
  …
  LOG.error("Failed to send counters", e);
```

```
context.getCounter("COMMON_COUNTERS",
                   ”BadURLS").increment(1);
```

Логирование и счетчики

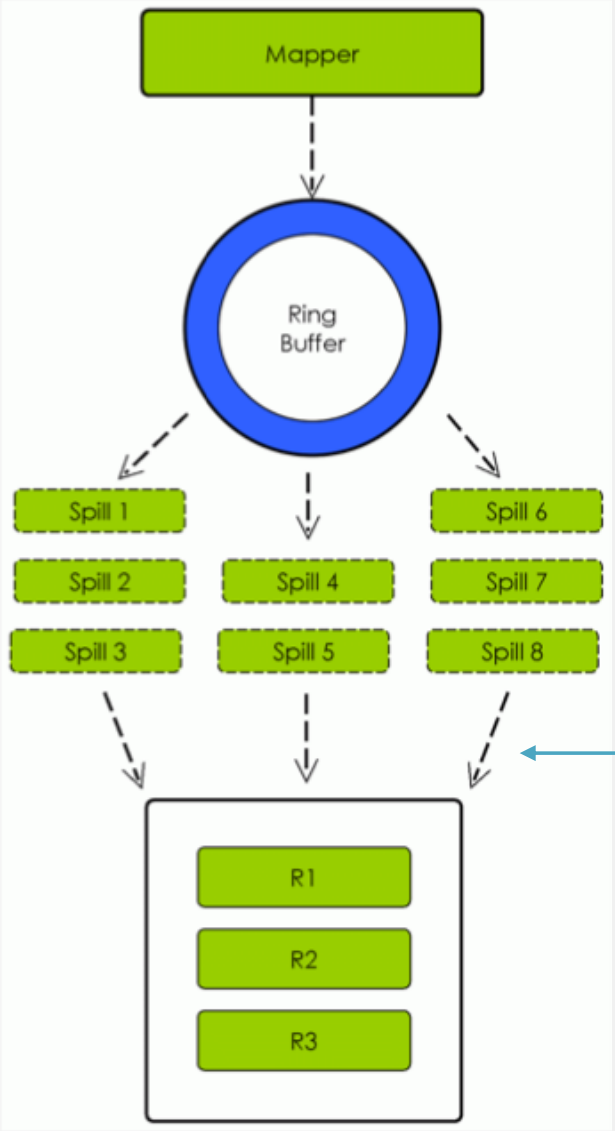
Не логируйте на каждую входную строку!

Не создавайте 100500 счетчиков



MapReduce и сортировка

InputSplit,
Map, CircularBuffer, Spill, Combine,
Sort, Shuffle, Copying,

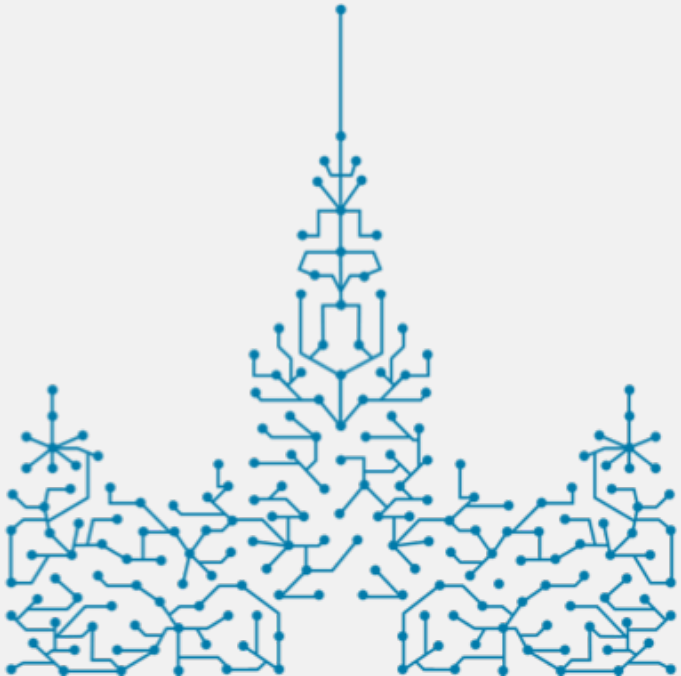


Shuffle окончательно объединяет Spill-ы в 1 для каждого recuder-а

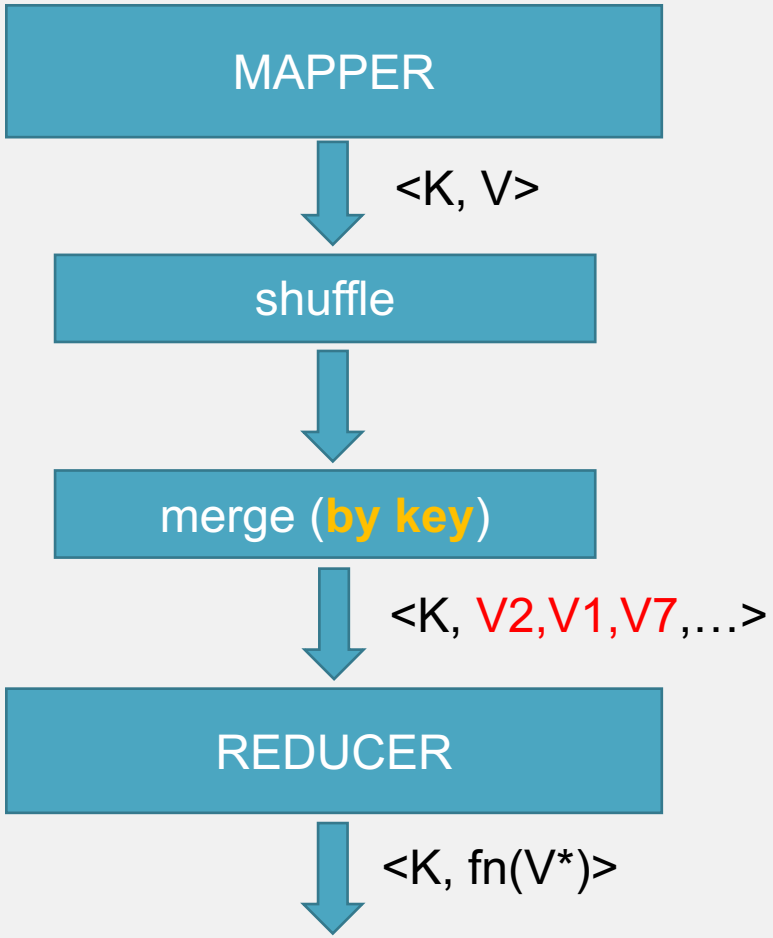
Подробней об устройстве hadoop-mapreduce

- http://ercoppa.github.io/HadoopInternals/
- http://trimc-hdfs.blogspot.ru/2015/03/hadoop-architecture.html

Вторичная сортировка

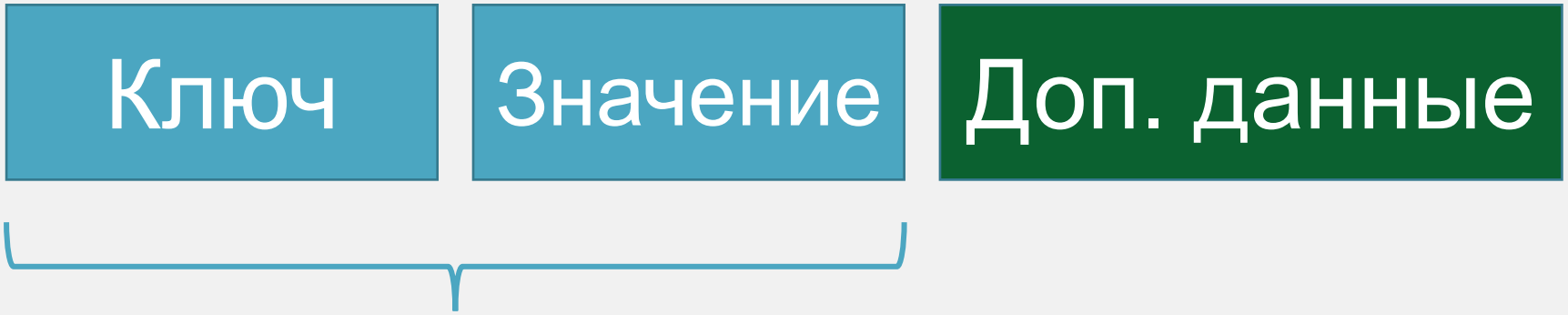


Стандартный flow



А если хотим порядок values?

1. MapReduce сортирует **только** по ключам
2. Значит часть значения должна стать ключом



3. Определить свои Comparator/Group и Partitioner

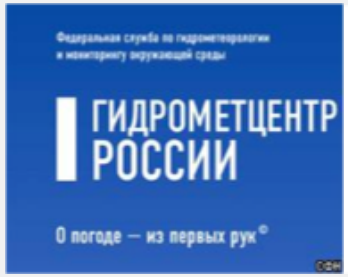
Задача

Задача – средствами framework найти температурный максимум по дням

Вход – данные Global Surface Summary Of Day Data

Выход – StationID, month/day, maxT(year)

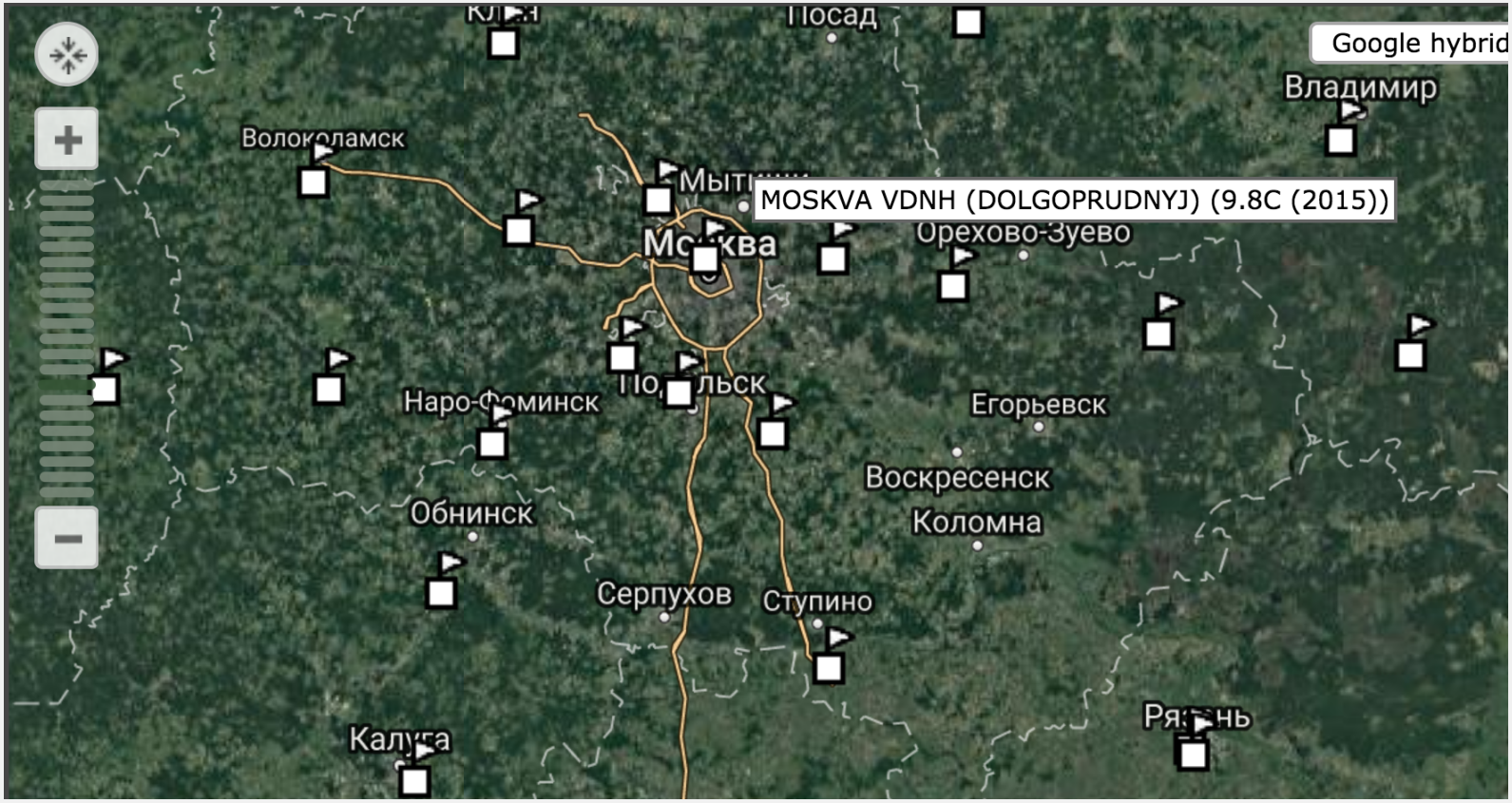
… посмотрим результат на Google Maps



В Казани погода установила новый температурный рекорд

5 марта воздух прогрелся до +6,2

Метеостанции на карте (этот день)



http://www.gpsvisualizer.com/

Что именно происходит

Mapper:

KEY={id-станции:день.месяц}|temp(℃)

composite(10000:15.05, -2)

VALUE =ГОД

2017

Reducer:

{ Composite(10000:15.05, -2), 2017 }, ⟸ Ответ **всегда** первый

{ Composite(10000:15.05, -3), 2001 },

…

{ Composite(10000:15.05, -4), 2018 },

Код

https://github.com/dkrotx/hadoop_sem2

См. run.sh

Обратите внимание

job.setPartitionerClass(MDStationPartitioner.class)

Свое разбиение по ключу

job.setSortComparatorClass(KeyComparator.class)

Свое упорядочивание для reduce (полное)

job.setGroupingComparatorClass(MDStationGrouper.class)

Своя группировка по ключам

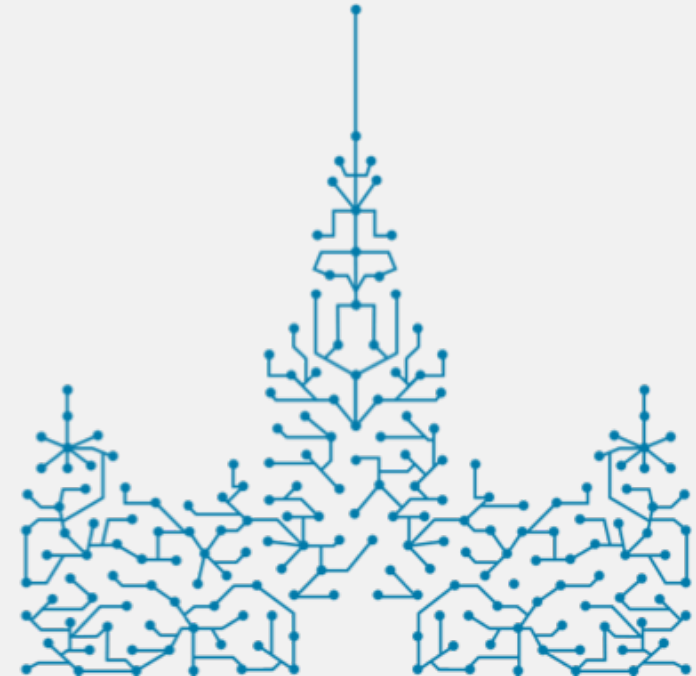
job.setMapOutputKeyClass(TextFloatPair.class)

Свой тип ключа с операциями

Неочевидный момент

Итерация по значениям в reduce() меняет значение ключа!

- Все ОК, ключ композитный
- Итерируемся в рамках группы



ДЗ #2
см. Блог