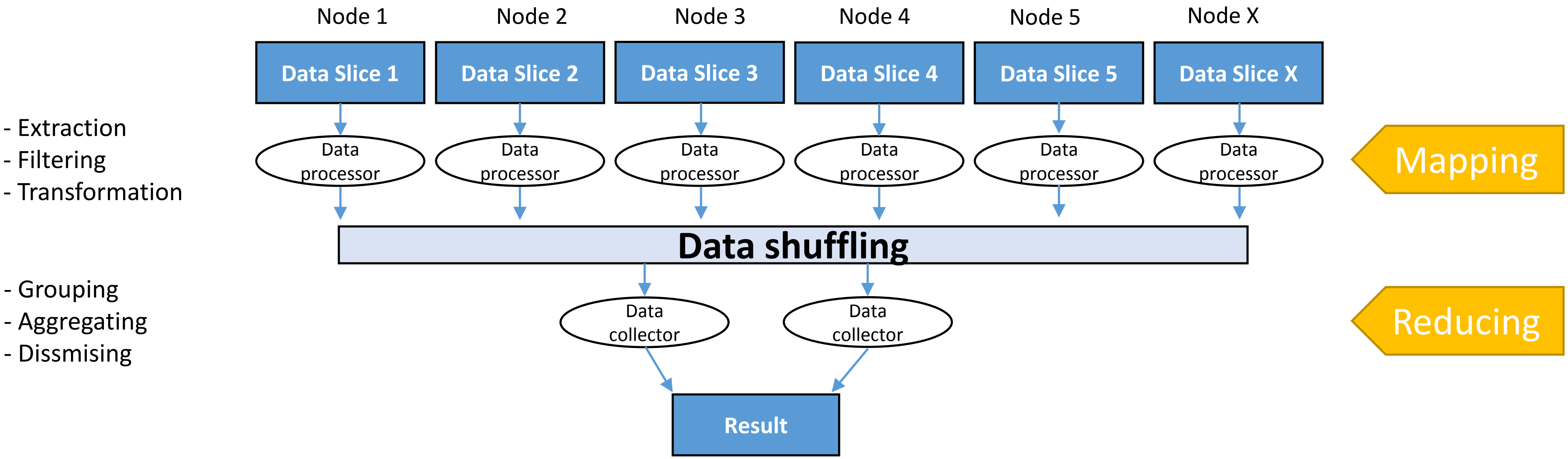
MapReduce

What is MapReduce? (1)

- A programing model for parallel processing of a distributed data on a cluster



Node 1 Node 2 Node 3 Node 4 Node 5 Node X

Data Slice 1 Data Slice 2 Data Slice 3 Data Slice 4 Data Slice 5 Data Slice X

- Extraction
- Filtering
- Transformation

Data processor Data processor Data processor Data processor Data processor Data processor

Mapping

Data shuffling

- Grouping
- Aggregating
- Dissmising

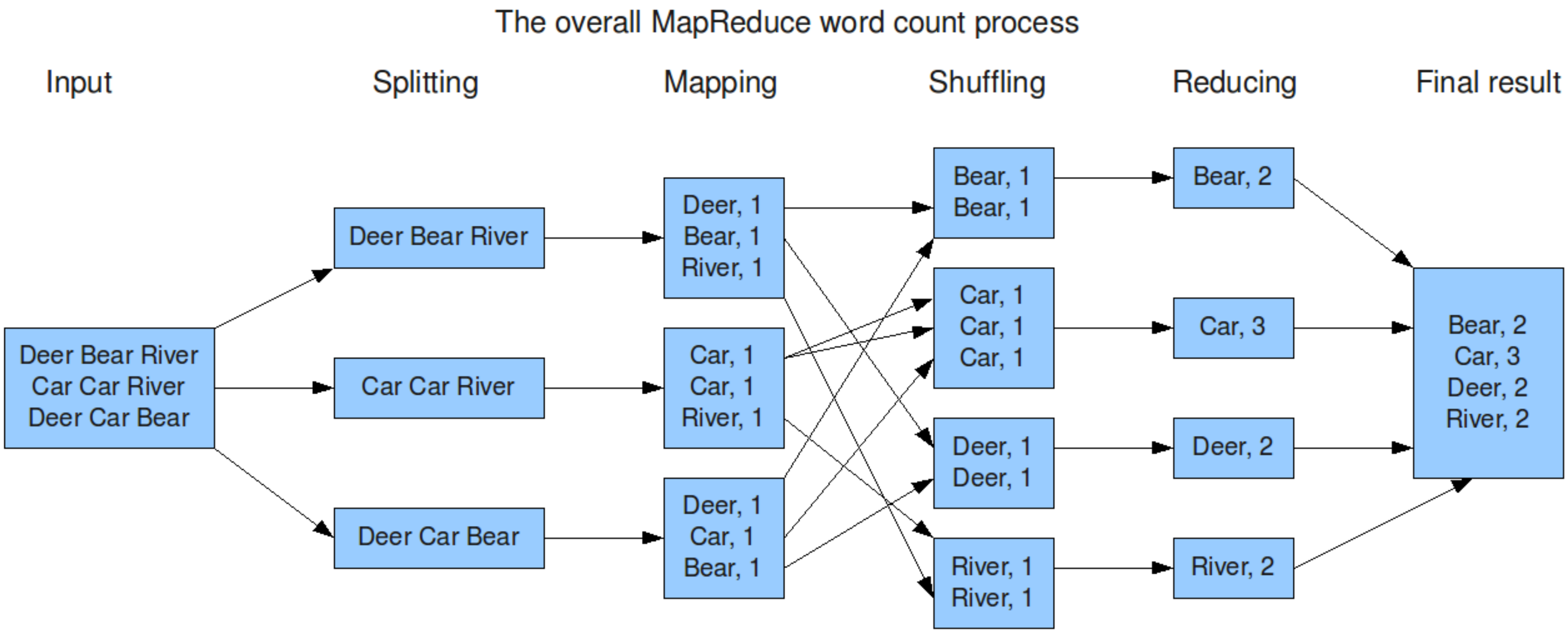
Data collector Data collector

Reducing

Result

- It is an ideal solution for processing data on HDFS

Example: The famous „world counting”

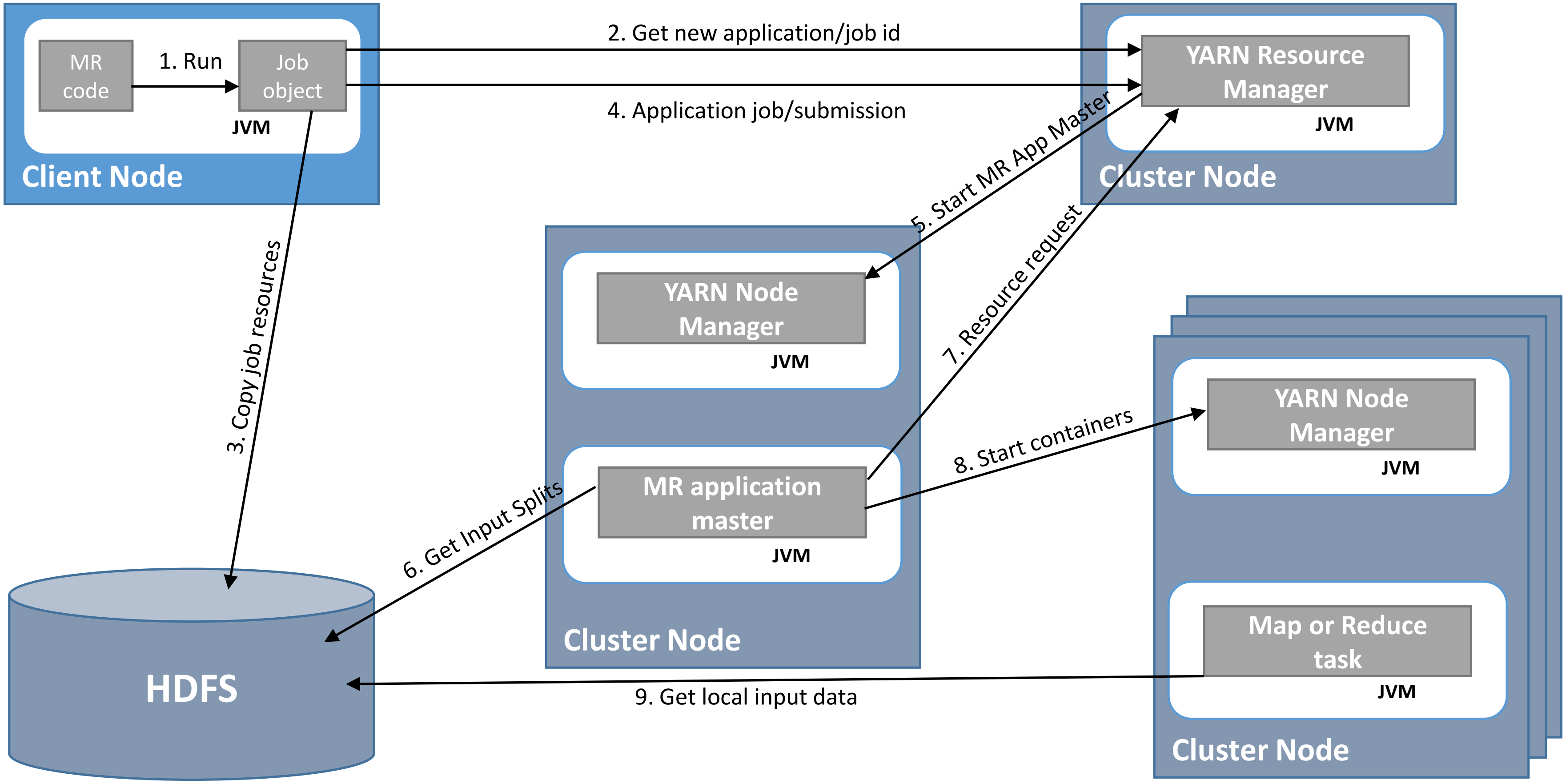


What is MapReduce? (2)

- 2 staged data processing
  - Map and Reduce
- Each stage emits key-value pairs as a result of its work
- Programing MapReduce
  - In Java
  - 3 classes
    - Map
    - Reduce (optional)
    - Job configuration (with a ‚main' function)

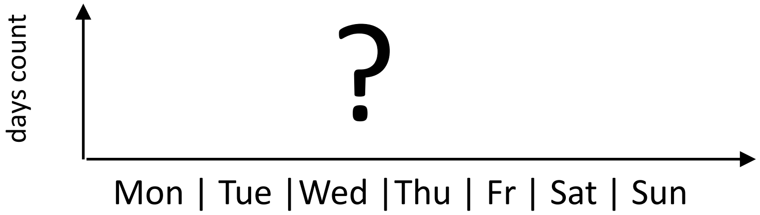
MapReduce on Hadoop

- In V2 controlled by YARN demons
  - ResourceManager, NodeManager



MR hands on (1)

- The problem
  - Q: „What follows two rainy days in the Geneva region?”
  - A: „Monday”
- The goal
  - Proof if the theory is true or false
- Solution
  - Lets take meteo data from GVA and build a histogram of days of a week followed by 2 or more bad weather days



MR hands on (2)

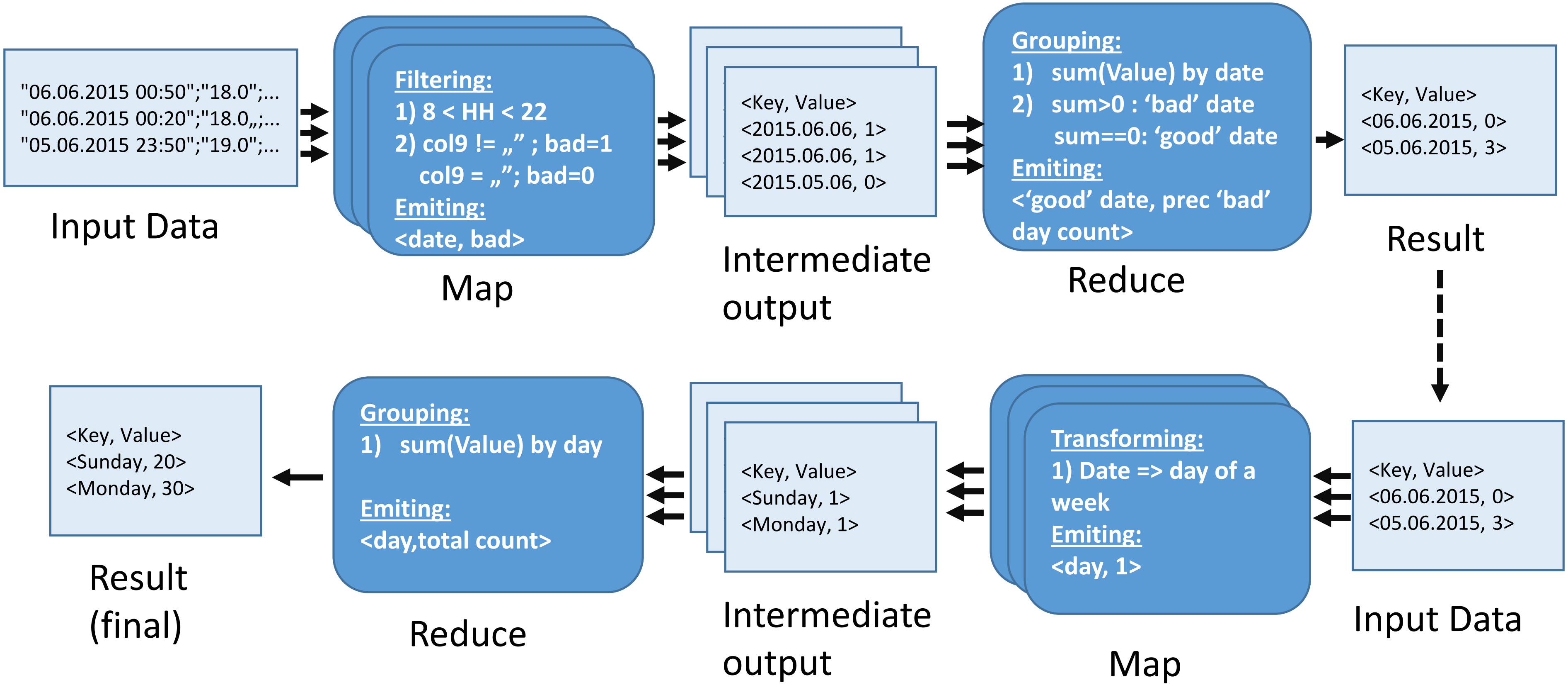
- The source data (http://rp5.co.uk)
  - Source: Last 3 years of weather data taken at GVA airport
  - CSV format

```
"Local time in Geneva (airport)";"T";"P0";"P";"U";"DD";"Ff";"ff10";"WW";"W'W'";"c";"VV";"Td";
"06.06.2015 00:50";"18.0";"730.4";"767.3";"100";"variable wind direction";"2";"";"";"";"No Significant Clouds";"10.0 and more";"18.0";
"06.06.2015 00:20";"18.0";"730.4";"767.3";"94";"variable wind direction";"1";"";"";"";"Few clouds (10-30%) 300 m, scattered clouds (40-50%) 3300 m";"10.0 and m
"05.06.2015 23:50";"19.0";"730.5";"767.3";"88";"Wind blowing from the west";"2";"";"";"";"Few clouds (10-30%) 300 m, broken clouds (60-90%) 5400 m";"10.0 an
"05.06.2015 23:20";"19.0";"729.9";"766.6";"83";"Wind blowing from the south-east";"4";"";"";"";"Few clouds (10-30%) 300 m, scattered clouds (40-50%) 2400 m,
"05.06.2015 22:50";"19.0";"729.9";"766.6";"94";"Wind blowing from the east-northeast";"5";"";"Light shower(s), rain";"";"Few clouds (10-30%) 1800 m, scattered
"05.06.2015 22:20";"20.0";"730.7";"767.3";"88";"Wind blowing from the north-west";"2";"";"Light shower(s), rain, in the vicinity thunderstorm";"";"Few clouds (10
"05.06.2015 21:50";"22.0";"730.2";"766.6";"73";"Wind blowing from the south";"7";"";"Thunderstorm";"";"Few clouds (10-30%) 1800 m, cumulonimbus clouds , sc
"05.06.2015 21:20";"23.0";"729.6";"765.8";"78";"Wind blowing from the west-southwest";"4";"";"Light shower(s), rain, in the vicinity thunderstorm";"";"Few cloud
"05.06.2015 20:50";"23.0";"728.8";"765.0";"65";"variable wind direction";"2";"";"In the vicinity thunderstorm";"";"Scattered clouds (40-50%) 1950 m, cumulonimb
"05.06.2015 20:20";"23.0";"728.2";"764.3";"74";"Wind blowing from the west-northwest";"4";"";"Light thunderstorm, rain";"";"Scattered clouds (40-50%) 1950 m,
"05.06.2015 19:50";"28.0";"728.0";"763.5";"45";"Wind blowing from the south-west";"5";"11";"Thunderstorm";"";"Scattered clouds (40-50%) 1950 m, cumulonimb
"05.06.2015 19:20";"28.0";"728.0";"763.5";"42";"Wind blowing from the north-northeast";"2";"";"In the vicinity thunderstorm";"";"Few clouds (10-30%) 1950 m,
```

- What is a bad weather day?:
  - Weather anomalies (col nr 9) between 8am and 10pm

MR hands on (3)

- Designing MapReduce flow



"06.06.2015 00:50";"18.0";...
"06.06.2015 00:20";"18.0„;...
"05.06.2015 23:50";"19.0";...

Input Data

Filtering:
1) 8 < HH < 22
2) col9 != „” ; bad=1
col9 = „”; bad=0
Emiting:
<date, bad>

Map

<Key, Value>
<2015.06.06, 1>
<2015.06.06, 1>
<2015.05.06, 0>

Intermediate output

Grouping:
1) sum(Value) by date
2) sum>0 : ‘bad’ date
sum==0: ‘good’ date
Emiting:
<‘good’ date, prec ‘bad’ day count>

Reduce

<Key, Value>
<06.06.2015, 0>
<05.06.2015, 3>

Result

<Key, Value>
<06.06.2015, 0>
<05.06.2015, 3>

Input Data

Transforming:
1) Date => day of a week
Emiting:
<day, 1>

Map

<Key, Value>
<Sunday, 1>
<Monday, 1>

Intermediate output

Grouping:
1) sum(Value) by day

Emiting:
<day,total count>

Reduce

<Key, Value>
<Sunday, 20>
<Monday, 30>

Result (final)

Hand on (4)

- Loading the data to HDFS

```
cd ~/tutorials;  hdfs dfs –put  data data;
```

- Getting script and code

```
mkdir myMR; cd myMR
wget https://cern.ch/test-zbaranow/script.txt  (and MRtutorial.zip);
```

- Compiling the MapReduce source code

```
unzip MRtutorial.zip
javac –classpath `hadoop classpath` *.java
```

- Packing into a jar file

```
jar –cvf GVA.jar *.class
```

- Submitting a MapReduce jobs

```
hadoop jar GVA.jar AggByDateJob data stage
hadoop jar GVA.jar AggByDayJob stage result
```

Things that have not covered

- Types of YARN schedulers
- Combiner – just after map reducer
- Writing own: input splitters, data serializes, partitioners etc.
- Hadoop streaming – map and reducer as an external executable
- Distributed cache – caching of arbitrary files caching

Summary

- MapReduce is a model for parallel data processing on Hadoop in a batch fashion
  - 2 staged
  - Job submission submission is not immediate
- Logic written in Java (but not only)
  - A developer skills required
  - Fully customizable
- Resource allocation controlled by YARN