Map/Reduce using R, Hadoop and an EMR cluster

@anettebgo
@sivmhollup

bouvet

Agenda

What does all those words mean?

Amazon EMR example

More about R

Map/Reduce with R - Example 1

Map/Reduce with R - Example 2

Summary

R

A programming language

Hadoop

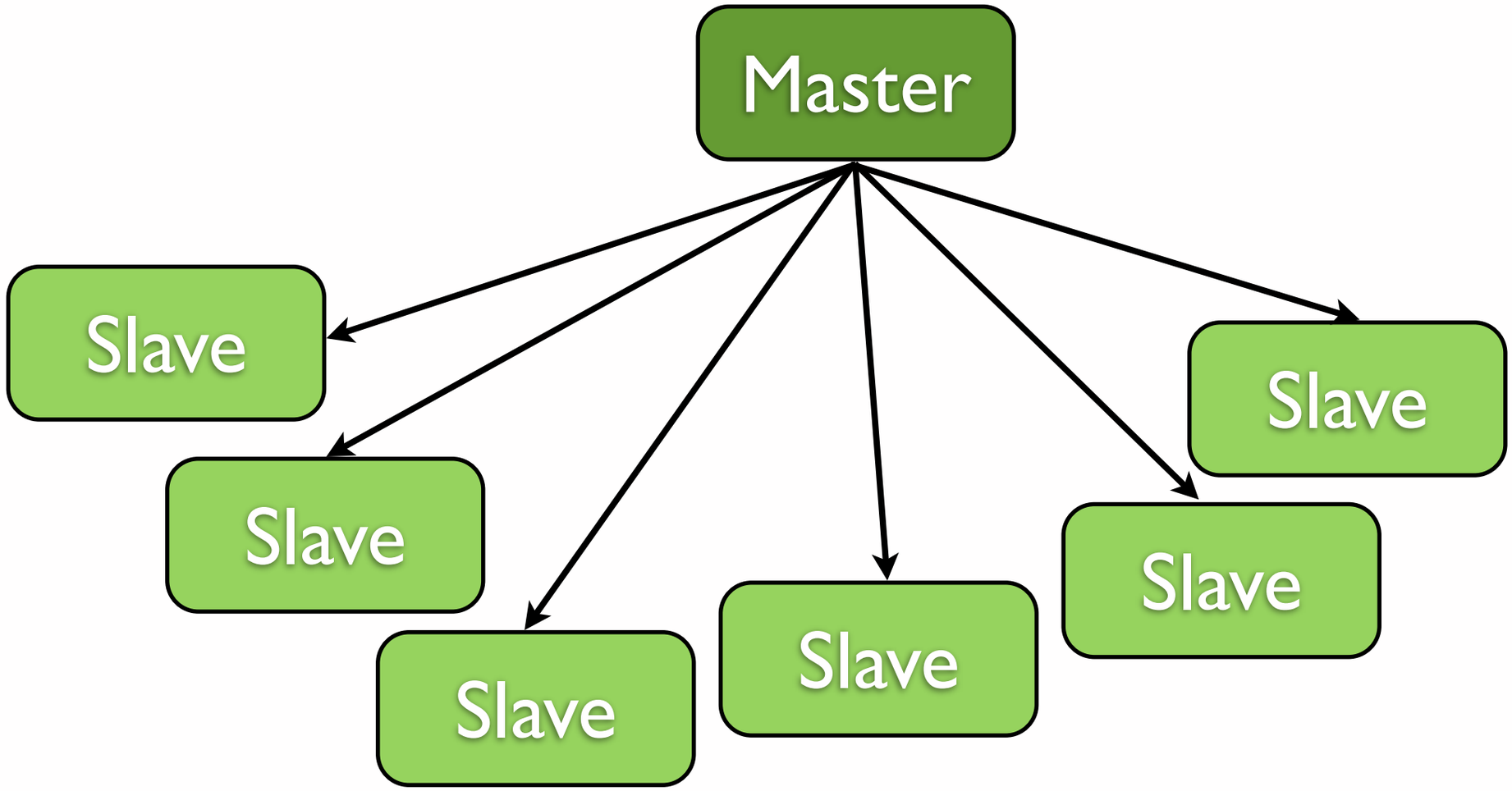
A framework to run map/reduce algorithms

EMR

Elastic Map/Reduce

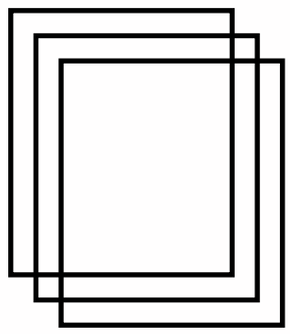
A service from Amazon to easily set up and tear down clusters with the Hadoop framework on them.

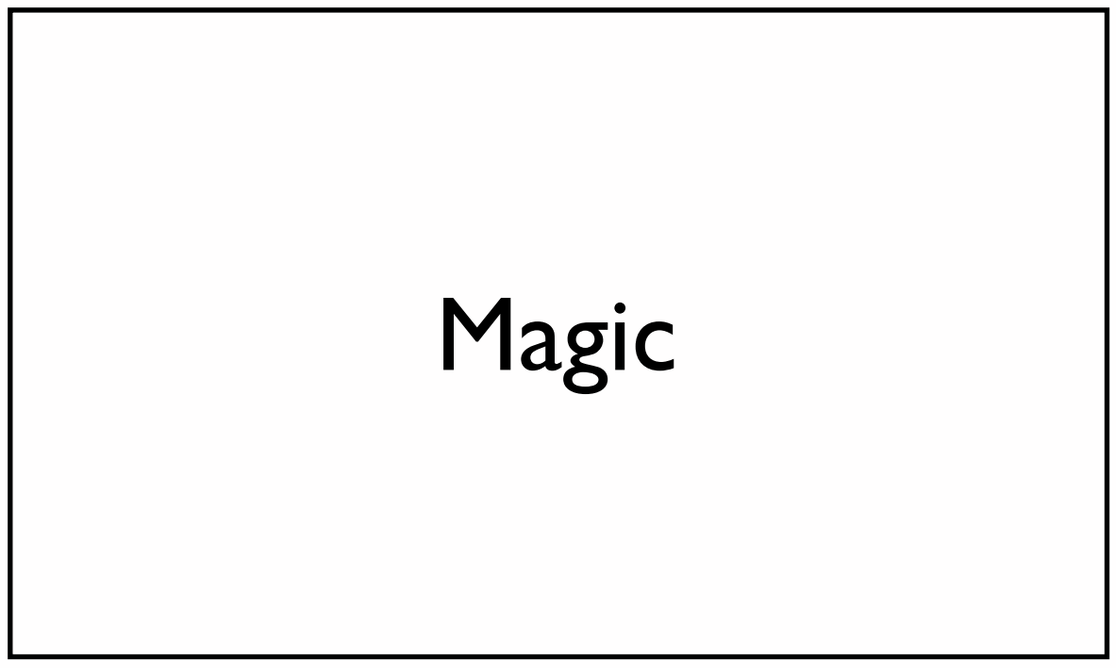
Cluster

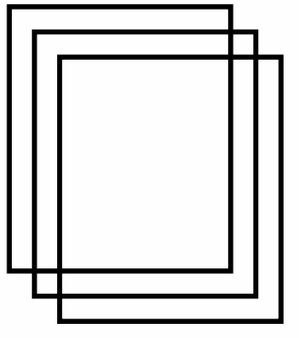


Map/Reduce

Input

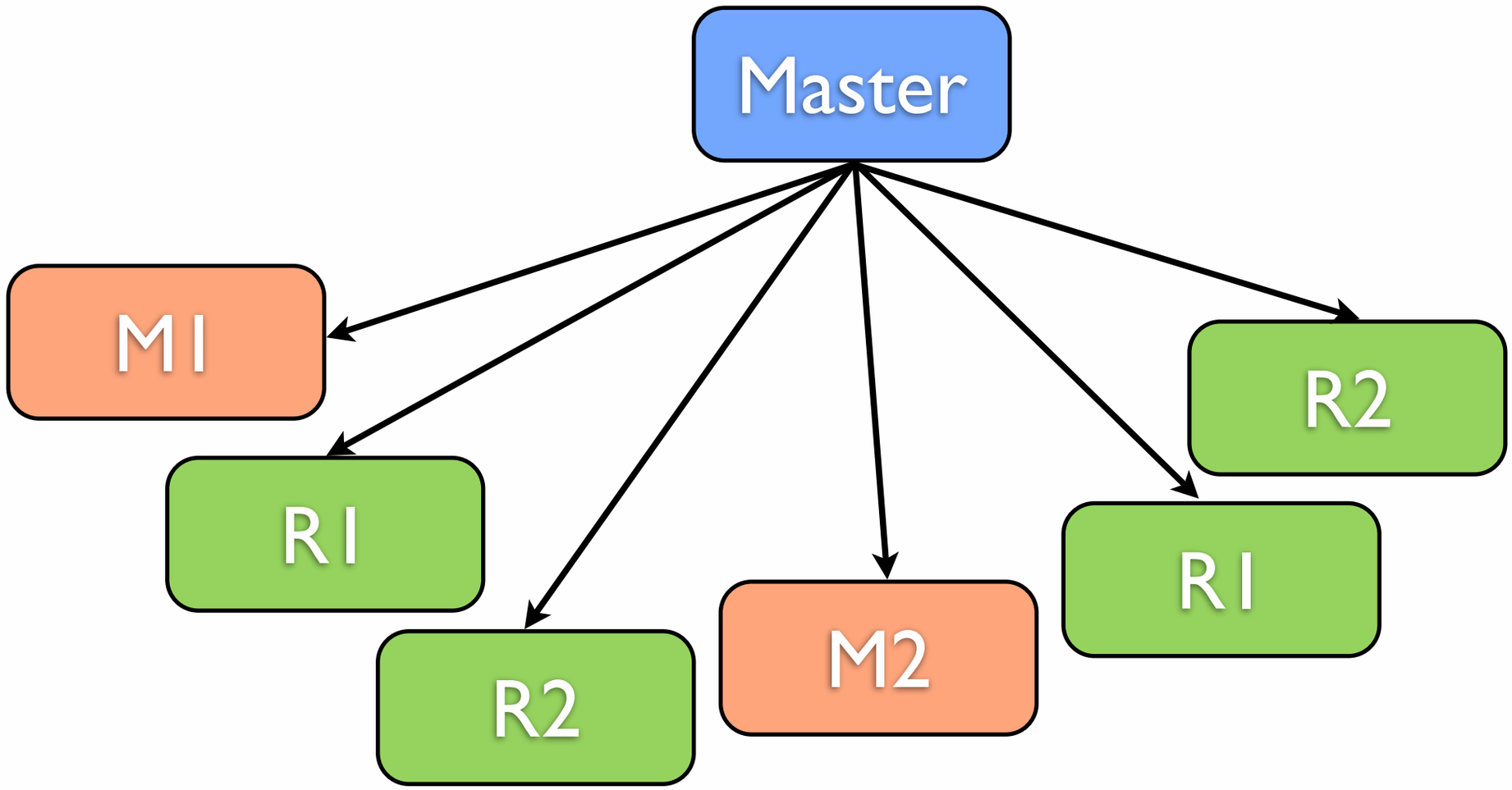






Output

Map/Reduce Cluster



Map/Reduce

- an *input reader*
- a ***Map*** function
- a *partition* function
- a *compare* function
- a ***Reduce*** function
- an *output writer*

Map/Reduce

Framework reads input (one or more data files) and passes chunks to mappers

Each Mapper creates a map* of the input

Framework sorts the map based on the keys

Framework allocates reducers to each key

**Reduce is called once per unique key, producing zero or more outputs

Framework writes output to permanent storage

* Map = set of key/value pairs

** Reduce = collapse map into result

Word Count Example

Word Count Example

Input

```
bla bla bla
and so
forth and
more
```

Word Count Example

Input

```
bla bla bla
and so
forth and
more
```

Mapped

```
bla    1
bla    1
bla    1
and    1
so     1
forth  1
and    1
more 1
```

Word Count Example

Input

```
bla bla bla
and so
forth and
more
```

Mapped

```
bla    1
bla    1
bla    1
and    1
so     1
forth  1
and    1
more 1
```

Sorted

```
and    1
and    1
bla    1
bla    1
bla    1
so     1
forth  1
more   1
```

Word Count Example

Input	Mapped	Sorted	Reduced
bla bla bla and so forth and more	bla 1 bla 1 bla 1 and 1 so 1 forth 1 and 1 more 1	and 1 and 1 bla 1 bla 1 bla 1 bla 1 so 1 forth 1 more 1	and 2 bla 3 forth 1 more 1 so 1

Word Count Example

```
cat data.txt | ./mapper.R | sort | ./reducer.R
```

Input	Mapped	Sorted	Reduced
bla bla bla and so forth and more	bla 1 bla 1 bla 1 and 1 so 1 forth 1 and 1 more 1	and 1 and 1 bla 1 bla 1 bla 1 bla 1 so 1 forth 1 more 1	and 2 bla 3 forth 1 more 1 so 1

Testing the account

- Log into your amazon account and go to AWS management console (top right)
- Click Services (top left), then S3 and create a bucket (region doesn't matter)
- **Write down** the name of your bucket.

Testing the account

- Click on your name in the top right corner > security credentials > access keys. If there isn't at least one access key here, create one.

Testing the account

- Click service (top left) then Elastic Map Reduce
- Click create cluster, then Configure sample application (top right)
- Choose word count, and make sure you fill in the Ouput location with your bucket name!
- Click Create Cluster at the bottom
-

This may take a while...
let's hear more about
R.

R in 5 minutes

Sort of

How do you work in R?

Command line interpreter

RStudio

Variable assignment

aVar = 23

aVar <- 23

Everything is a vector

Basic operations on vectors

Data into R

read.table()

Pretty plots

functions

sapply(), apply(), lapply()

R documentation

Example 1 - Locally

- Check out the code from git
- Run example one locally using the run.sh script
- You may need to install some missing packages:

```
> R CMD INSTALL /R_packages/HadoopStreaming_0.2.tar.gz

> R CMD INSTALL /R_packages/getopt_1.17.tar.gz
```

Example 1 - AWS

- Edit bootstrapR.sh with the name of your bucket, then upload it to the bucket
- Create an **ex1 folder** and upload the mapper.R reducer.R and data.txt here
- Create a folder called **R_packages** and upload content of R_Packages here

Running the example

- EMR > Create cluster
- Bootstrap Actions > Custom action. Give the bootstrapR.sh location
- Under steps, add a new Streaming step. Give it the location of mapper.R, reducer.R, data.txt, and a folder name where you want the result to show up. Click create!

All done?

Example 2

Do people try to trick their way around the toll-free import limit (currently 200 NOK) by having goods from a single larger purchase sent in multiple parcels?

Example 2 - bank data

```
1  2013-09-01 Goods Online   amazon blomster  NOK   595
2  2013-09-01 Goods Online   karis blomster   AUD   417
3  2013-09-01 Goods InStore  petters blomster GBP   206
```

Example 2 - post data

```
7 2013-09-01 amazon 12 gaten  Norge     alex smith    8 littlestreet
8 2013-09-01 karis  2 street  Australia alex singh    6 storegaten
9 2013-09-01 karis  7 gaten   Norge     petter smith  2 street
```

Example 2

- How many times do norwegians order goods from abroad?
- Does it match the number of parcels they get delivered?

Extensions

- Update wordcount to ignore case & punctuation
- How many parcels are delivered in total from Australia?
- What is the amount in NOK of services paid for in Germany?

Summary

Map/Reduce is for BIG DATA

Hadoop can be used with a range of languages

Amazon console is rubbish. Use boto!

Summary

Big Data is Dirty Data

Interpretation is important

Datasparsemkeit

Questions?