Map/Reduce using R, Hadoop and an EMR cluster

@anettebgo

R

A programming language

Hadoop

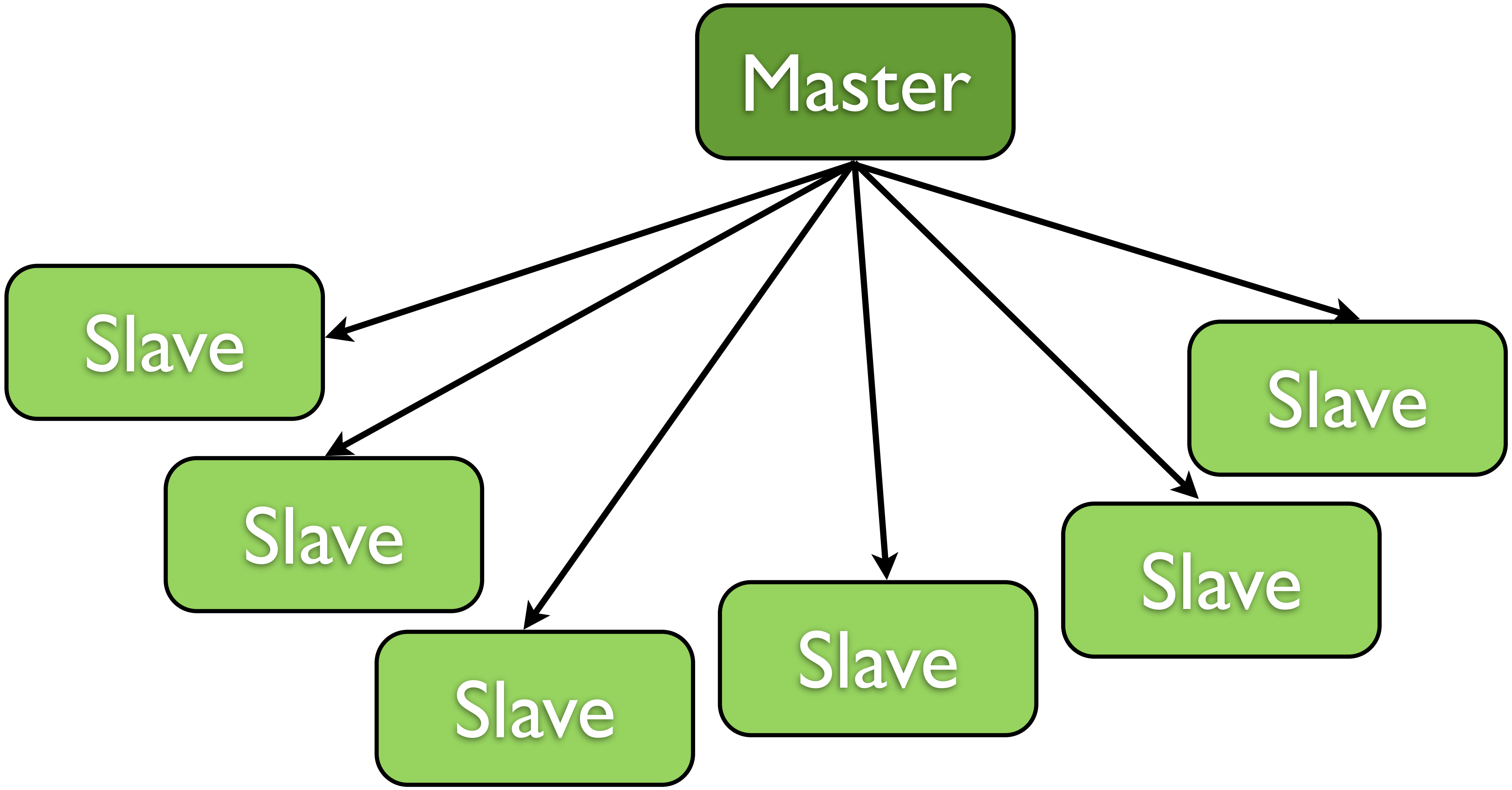
A framework to run map/reduce algorithms

EMR

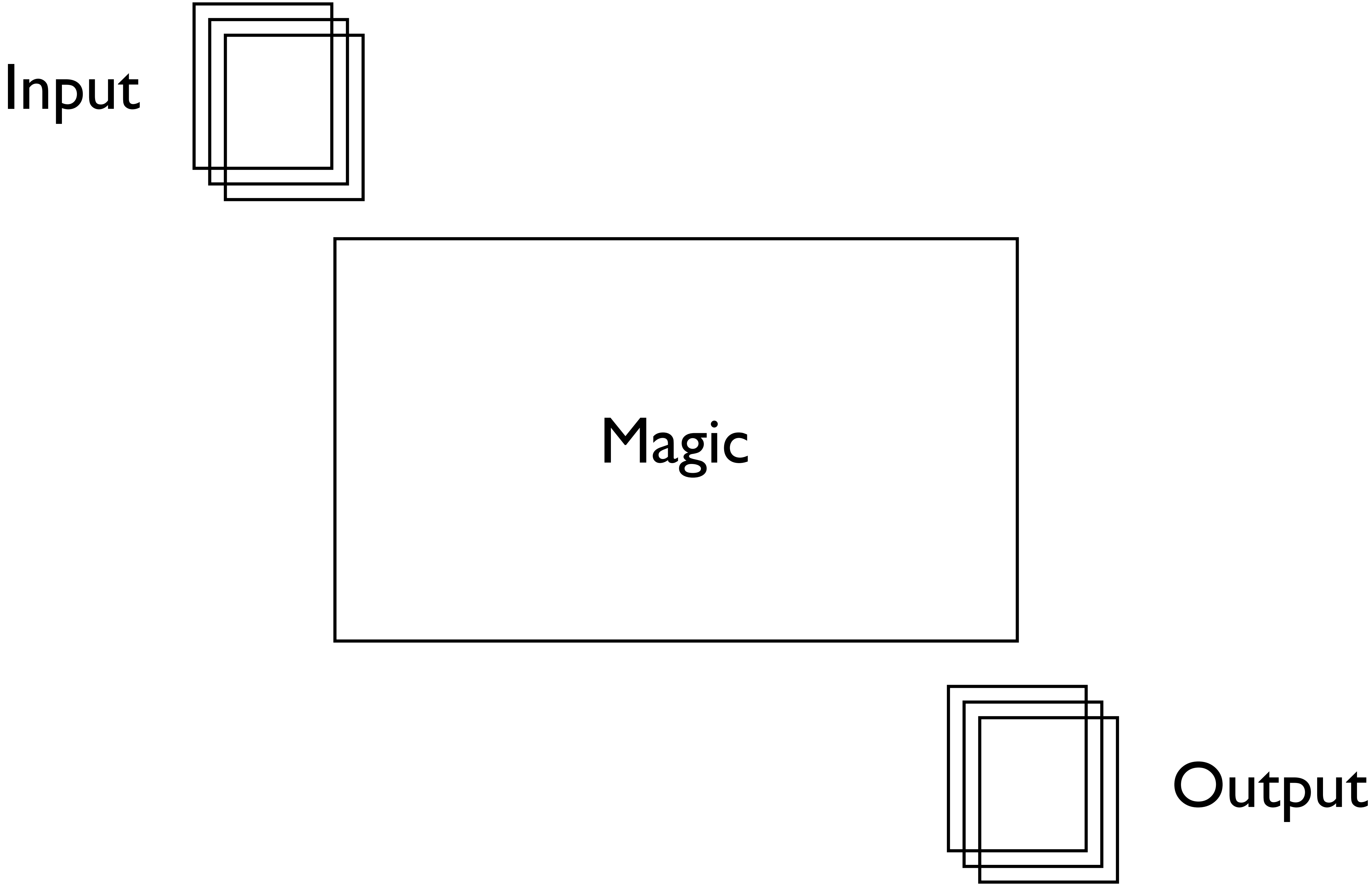
Elastic Map/Reduce

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

Cluster



Map/Reduce



Input

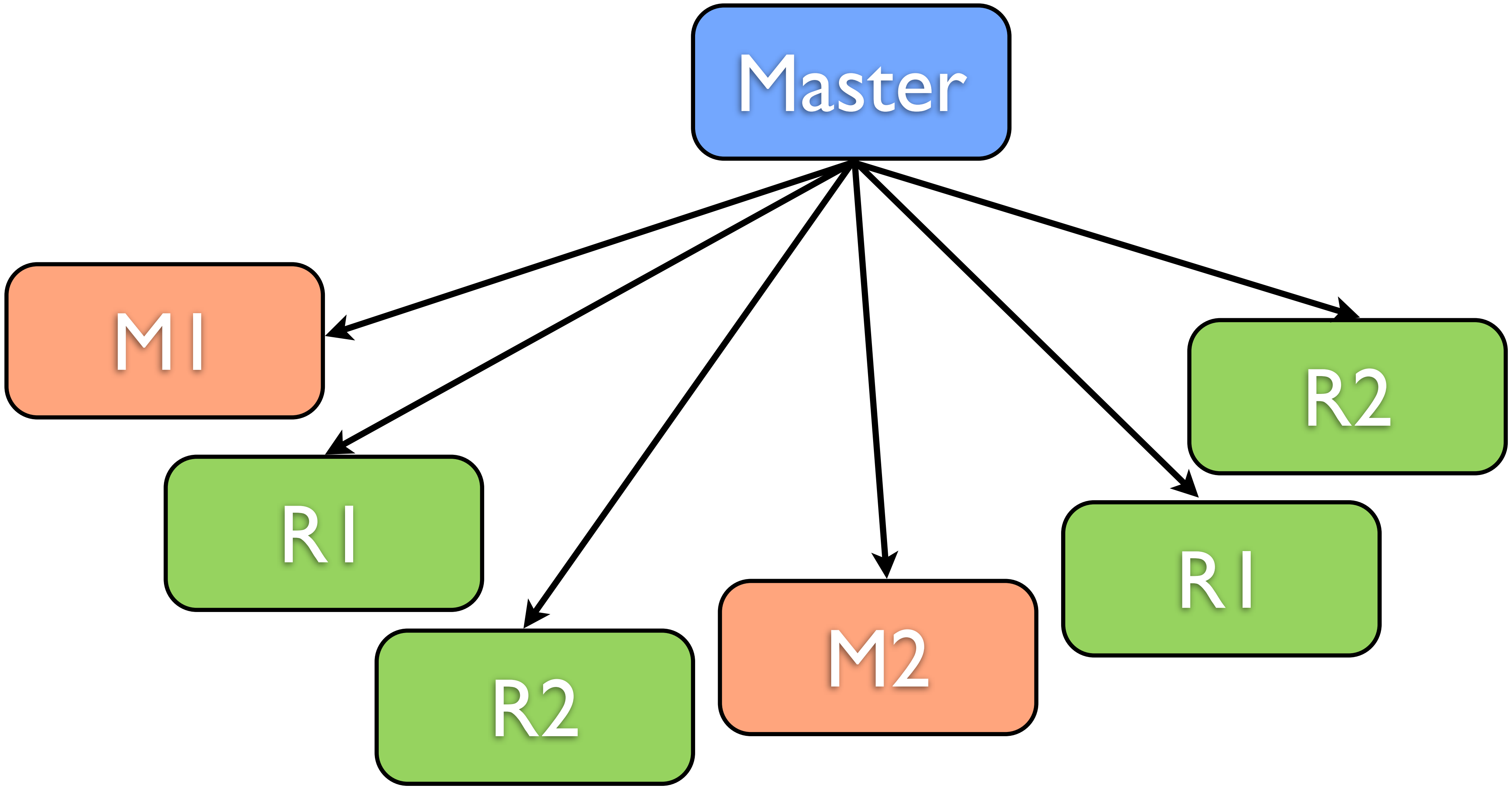
Magic

Output

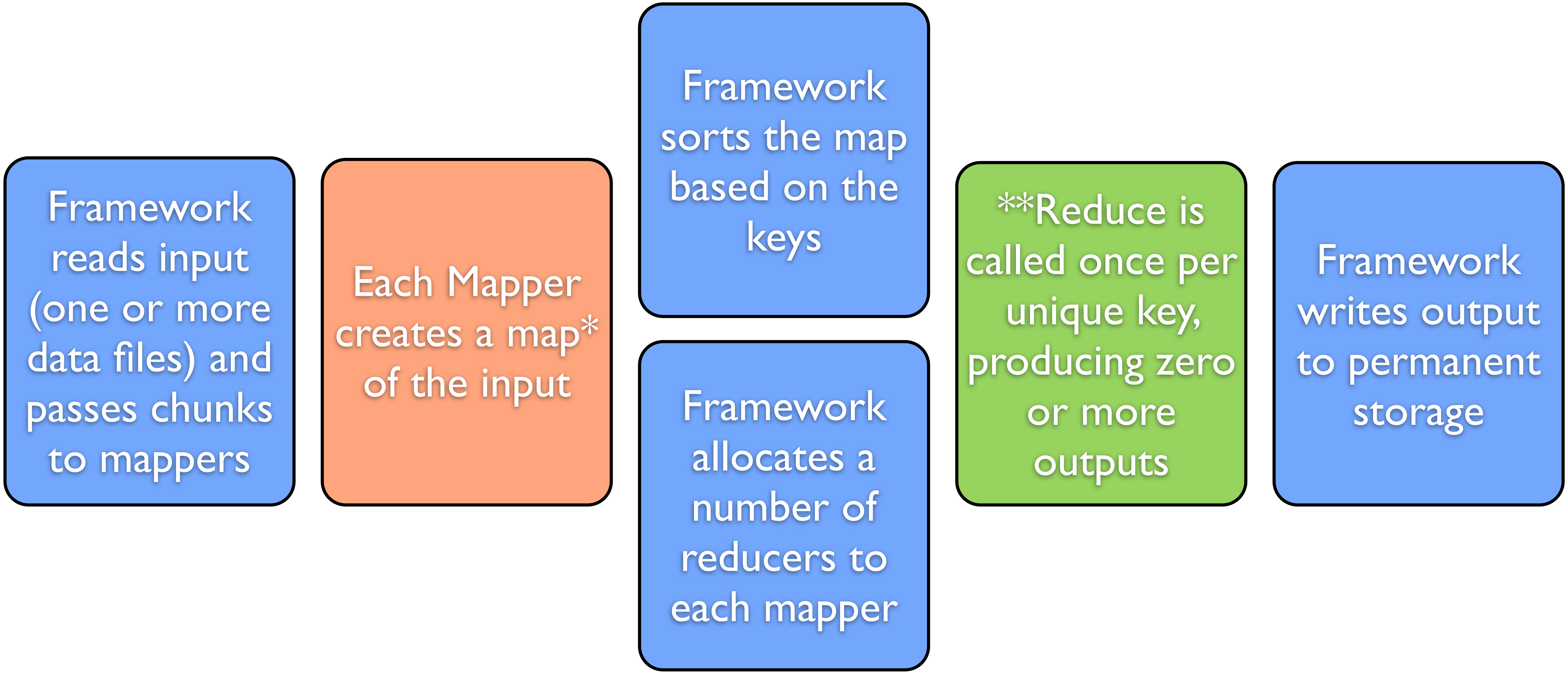
Map/Reduce

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

Map/Reduce Cluster



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 a number of reducers to each mapper

**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

In Our Case..

Input

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

Mapped

bla	1
bla	1
bla	1
and	1
so	1
forth	1
and	1
so	1
forth	1

Reduced

and	2
bla	3
forth	1
more	1
so	1

Let’s Do It!

- Check out the code from git
- Run it locally using the run.sh / run.cmd scripts
- You may need to install some missing packages - getopt and hadoop:

```
> R

...

> install.packages(HadoopStreaming)

> install.packages(getopt)
```

Let’s Do It!

```
> R

...

> install.packages(“HadoopStreaming”)

> install.packages(“getopt”)
```

Setting up

- Log into your AWS account, set up a S3 bucket (top left, S3, create a new bucket)
- Edit your boostrapR.sh to have the name of your bucket
- Upload the mapper.R, reducer.R, data.txt, bootstrapR.sh and the folder with the R packages.

Testing the account

- Go to EMR (Top left, EMR), click on new job
- Click on “new job”, and run the python word count (all defaults are ok, remember to input your bucket name for the logging though).
- You may get an error about missing keys/credentials. Go top right corner, click on your user name -> Security Credentials ->Access Keys and create a new set of

Running the example

- EMR -> New Job
- Streaming Job
- Give the data.txt as input location, mapper.R, reducer.R, and an output location
- Defaults all the way to the bootstrap step - here we need to run the bootstrapR.sh as a custom action

Questions?