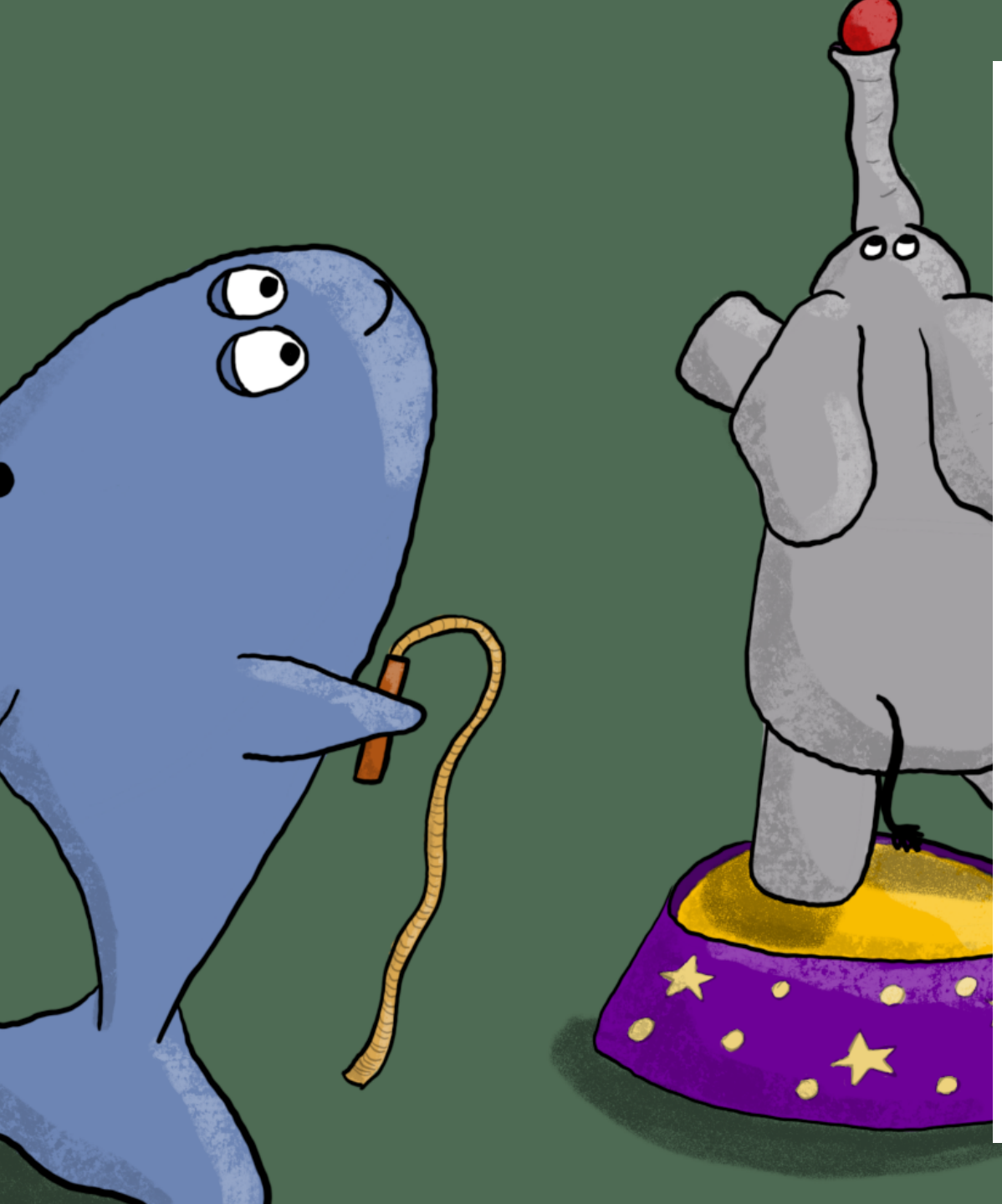


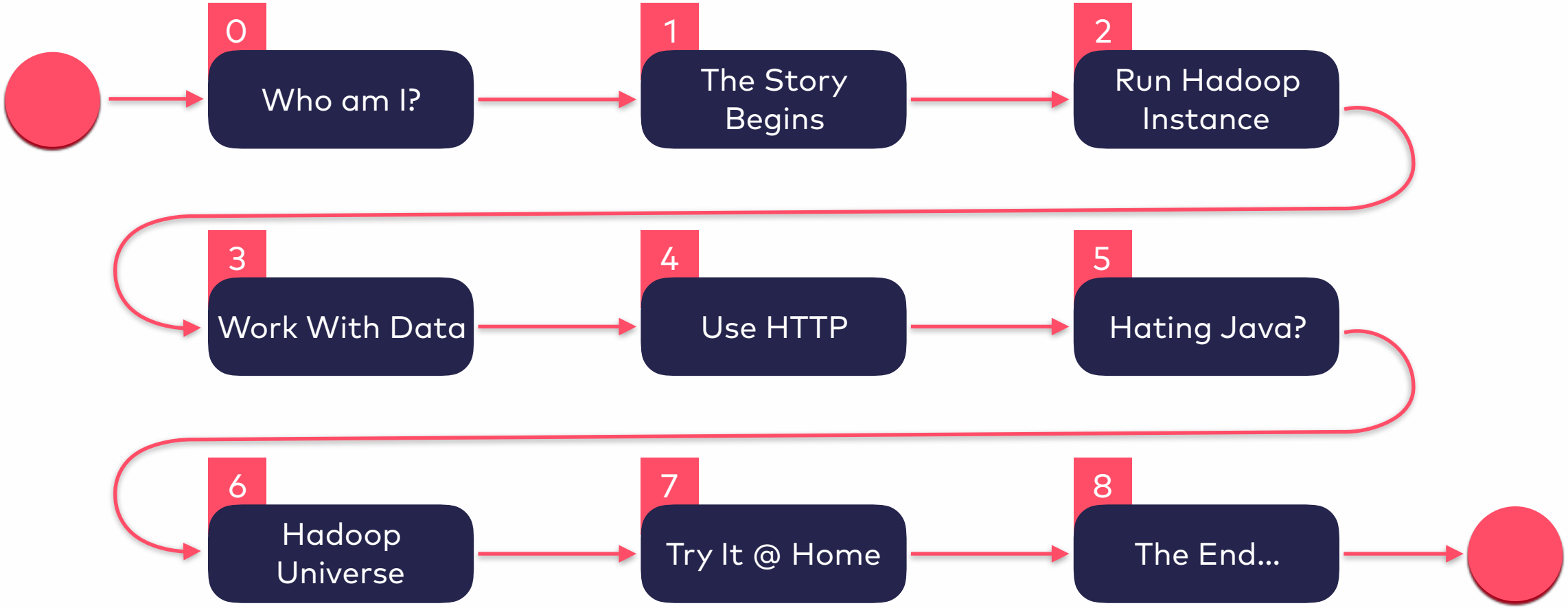
2019-01-14

Data Engineering Meetup, Munich

Hadoop - Taming the Elephant (With a Whale)

INNOQ

Our Journey


0
Who am I?
1
The Story Begins
2
Run Hadoop Instance
3
Work With Data
4
Use HTTP
5
Hating Java?
6
Hadoop Universe
7
Try It @ Home
8
The End...

0

Who am I?

Lisa Maria Moritz

lisa.moritz@innoq.com

+49 176 64 63 00 28

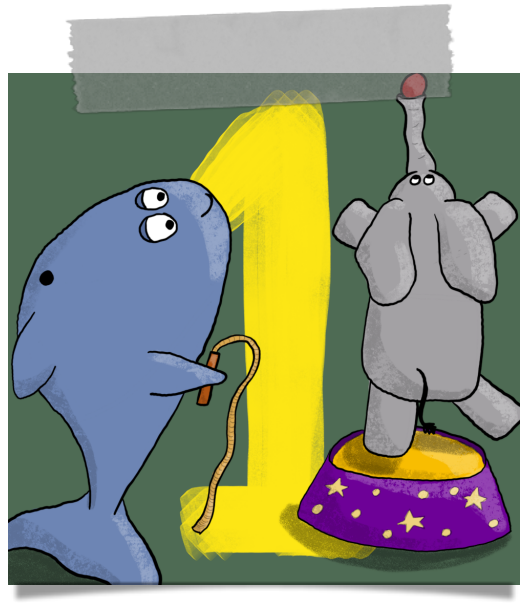
Teapot4181

Consultant since

September 2018

INNOQ

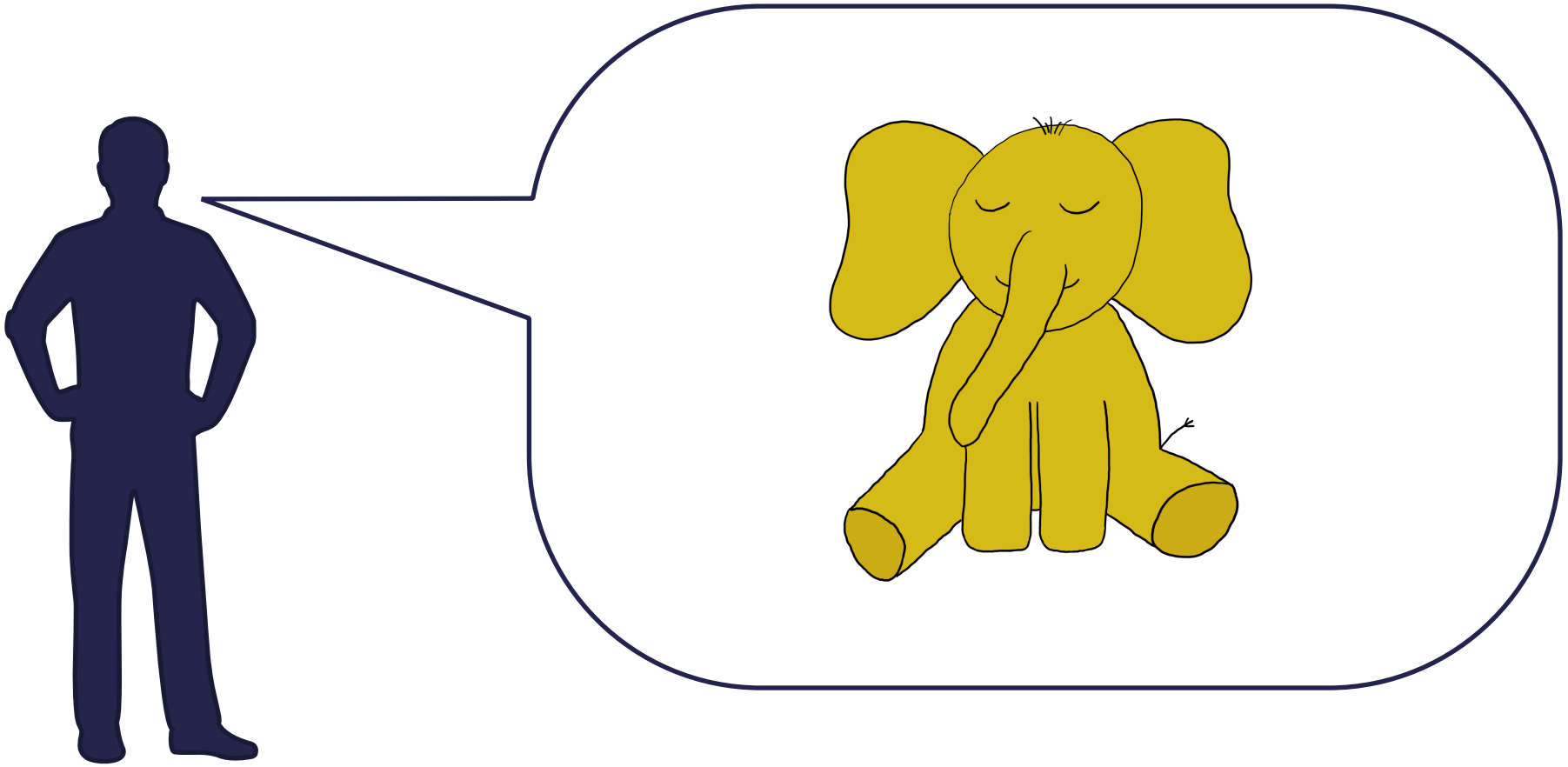
www.innoq.com



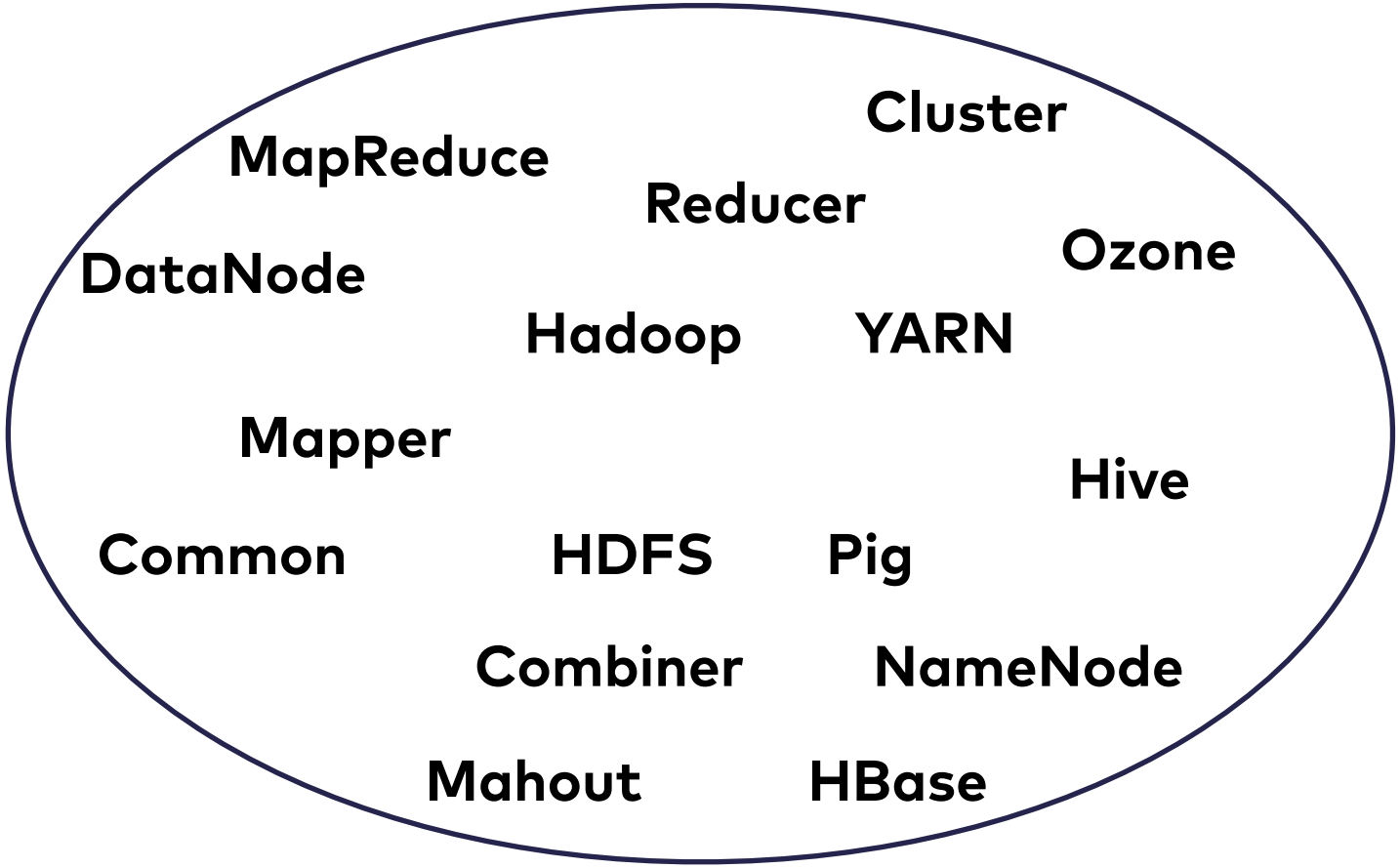
The Story Begins

1

How the Story Began...

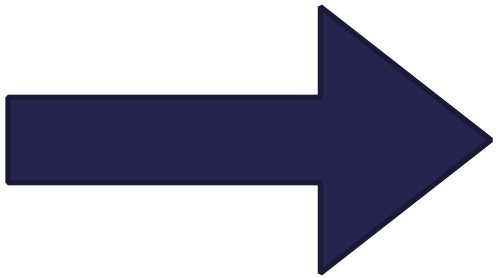


More Than One Elephant

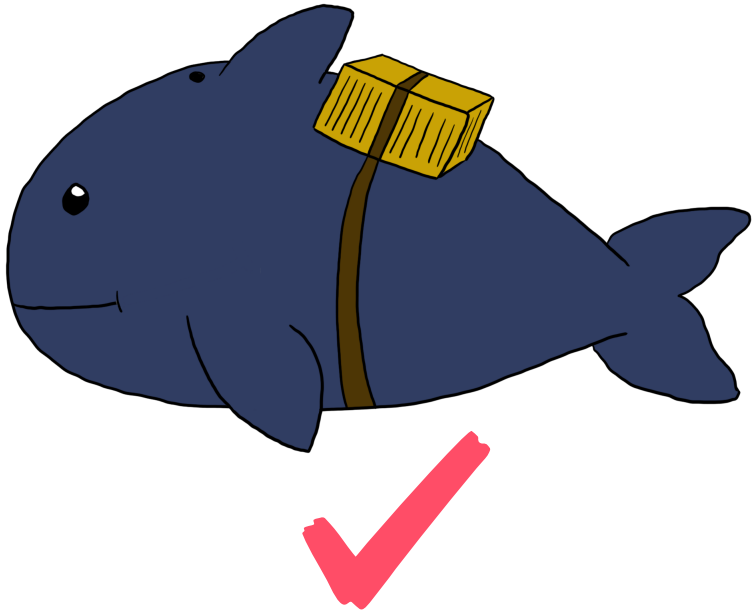


Break Into Smaller Pieces

Cluster



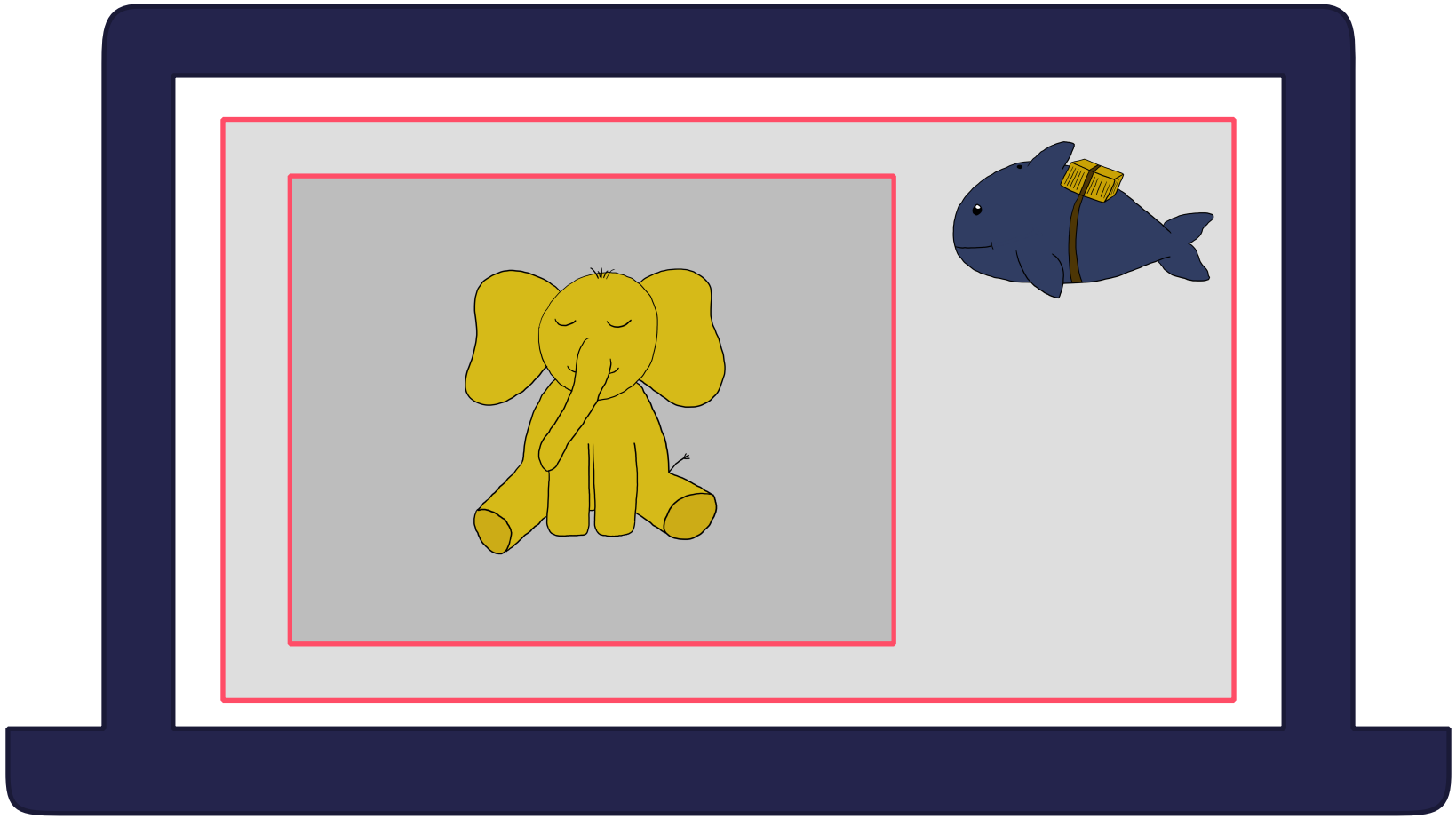
X



```
sequenceiq/hadoop-docker:2.7.0
```

Run Hadoop

📖 Goal - Part 2



📖 Some Docker Basics

Show all running containers

```
docker ps
```

Get bash in running container

```
docker exec -it <container-id> /bin/bash
```

Stop running container

```
docker container stop <container-id>
```

Start stopped container

```
docker container start <container-id>
```

🛠 Start Image

PLEASE NOTICE
It is not common to run Hadoop in Docker

```
docker run -it \
   -p 50070:50070 \
   -p 8088:8088 \
   -p 50075:50075 \
   sequenceiq/hadoop-docker:2.7.0 \
   /etc/bootstrap.sh -bash
```

- Run image in interactive mode
- Forward Web-UI port
- Forward job-tracking-port
- Forward result-download-port
- Name and version of Docker image
- Run this command

/etc/bootstrap.sh?

- Included script within sequenceiq Docker-image
- Important to know:
  - Starts services
  - Has two argument-options:
    - -d : Detached, Background-job
    - -bash: Open bash right after container start

🛠️ Run Example Job 1/2

Folder containing e.g. Examples

```
$HADOOP_PREFIX/share/hadoop/mapreduce
```

MapReduce Example Java App

```
hadoop-mapreduce-examples-2.7.0.jar
```

Running an Example MapReduce-Job

```
bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-
examples-2.7.0.jar grep input output 'dfs[a-z.]+'
```

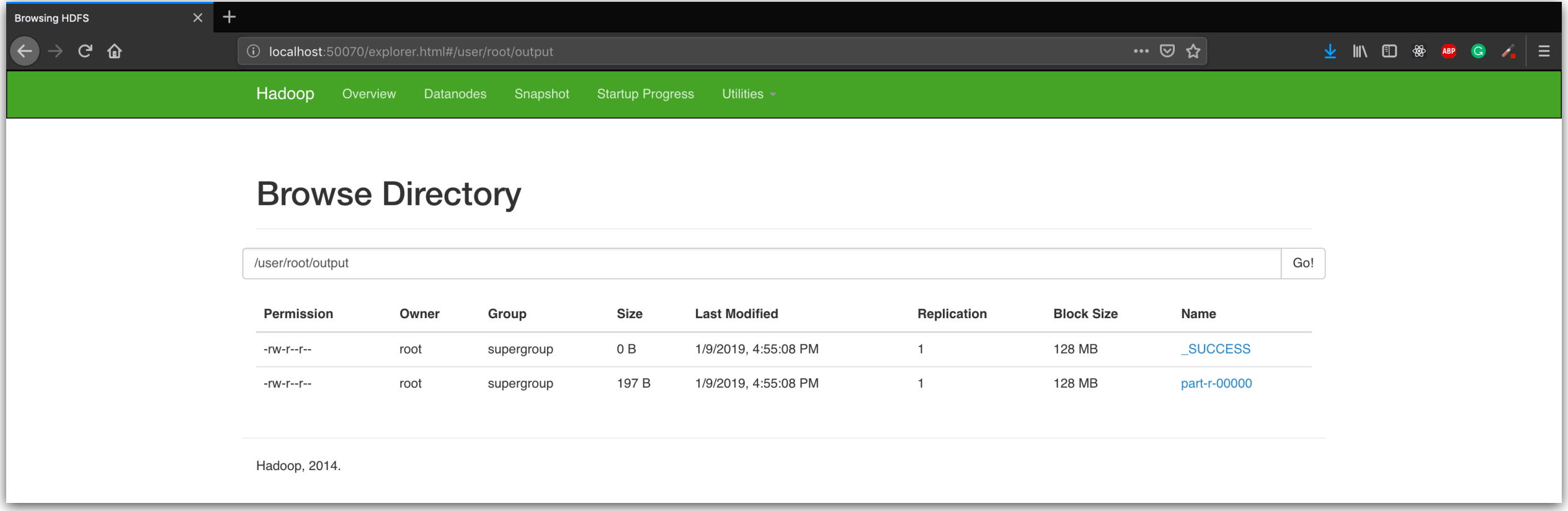
🛠 CLI - Check

```
bin/hdfs dfs -cat output/*
```

```
bash-4.1# bin/hdfs dfs -cat output/*
6       dfs.audit.logger
4       dfs.class
3       dfs.server.namenode.
2       dfs.period
2       dfs.audit.log.maxfilesize
2       dfs.audit.log.maxbackupindex
1       dfsmetrics.log
1       dfsadmin
1       dfs.servers
1       dfs.replication
1       dfs.file
```

2

🛠 Web UI - Check

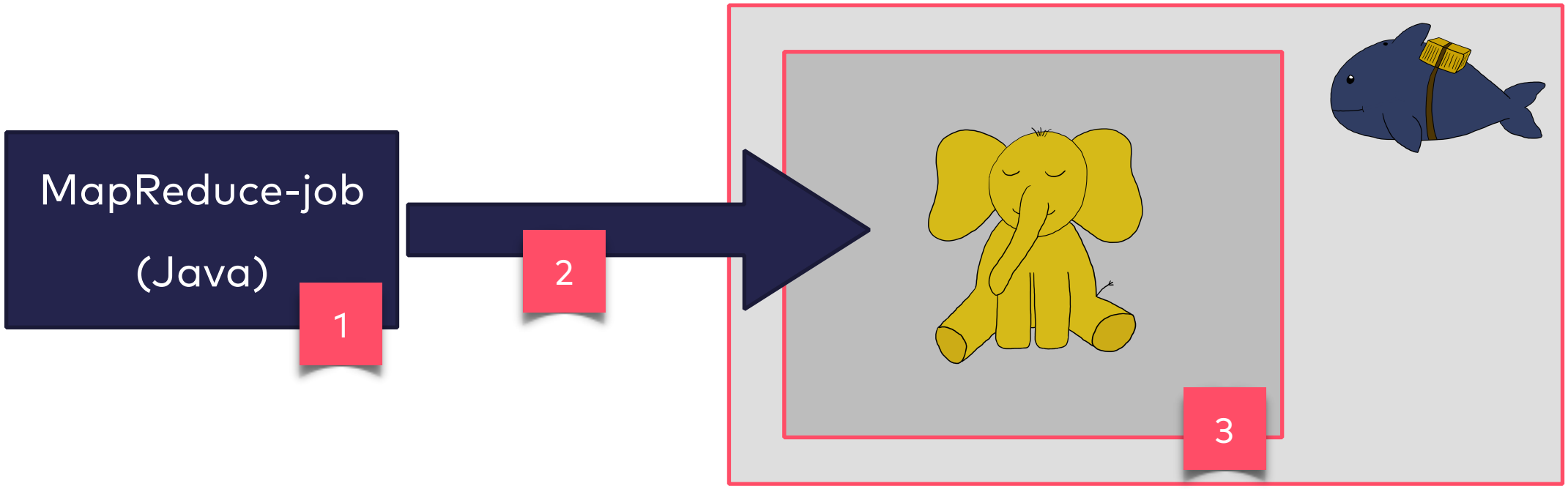


Running Hadoop-Instance

Working with Data

3

📖 Goal - Part 3



 MapReduce?

MapReduce

Mapper

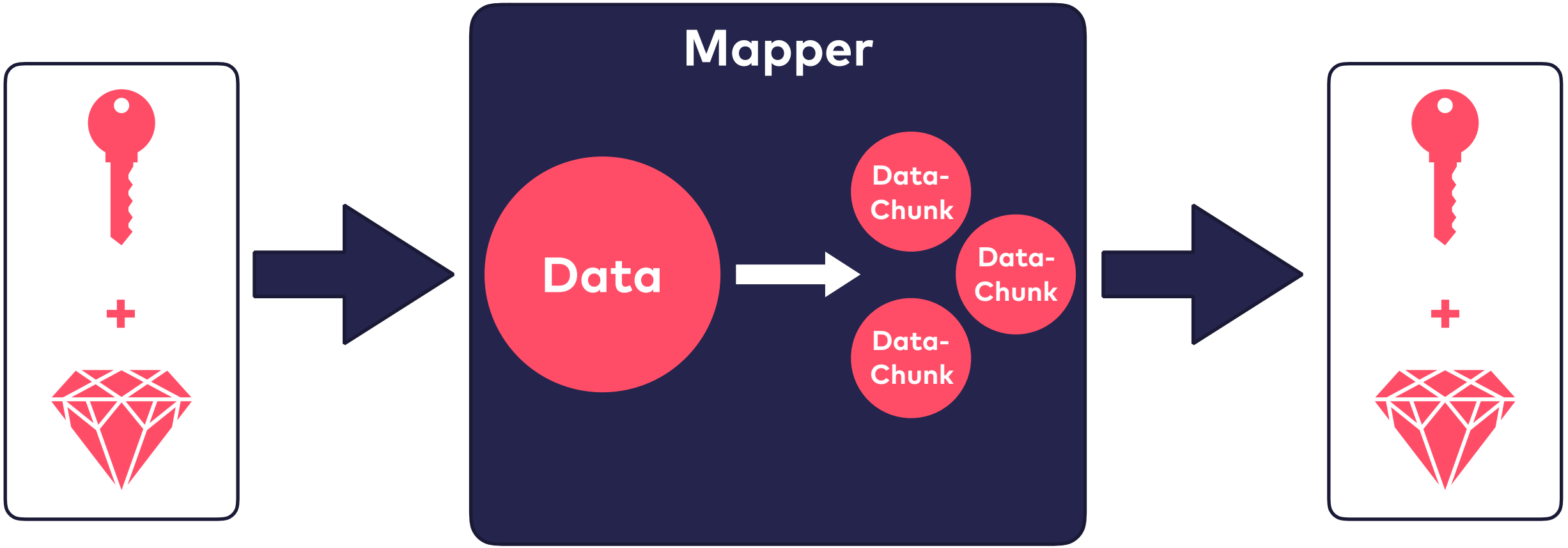
Reducer

Combiner

 MapReduce: Example

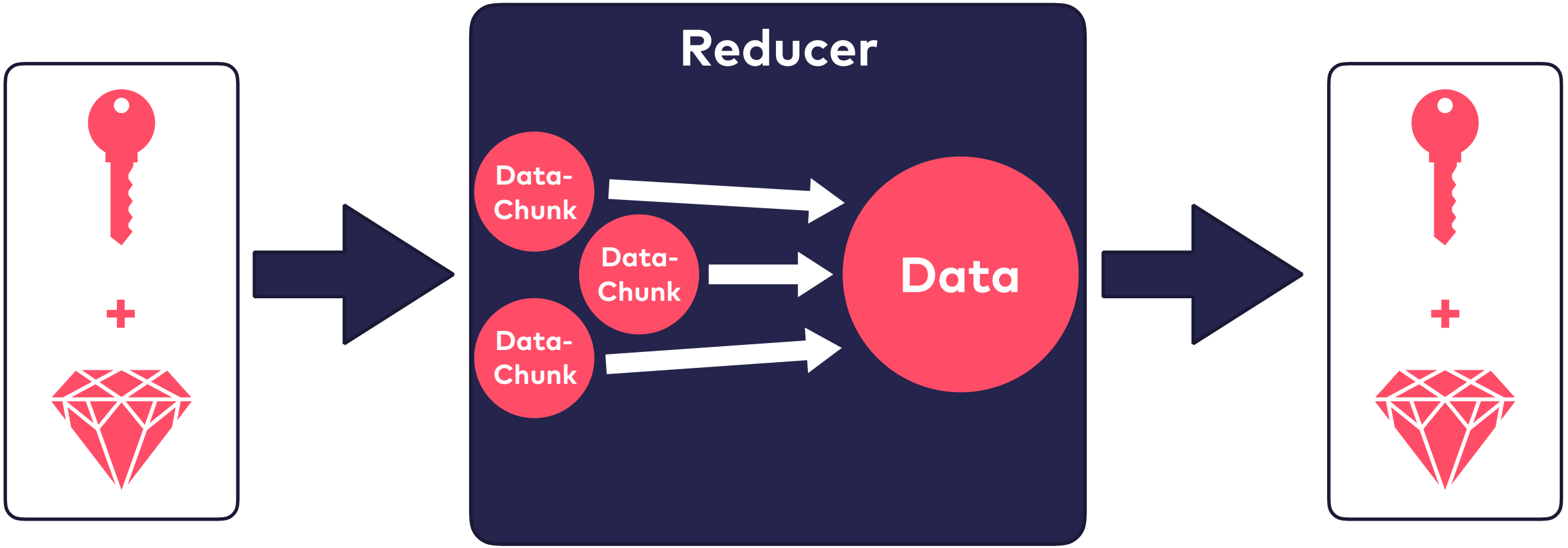
ID	Animal	Name
1	Dog	Berta
2	Cat	Miezi
3	Cat	Fluffy
4	Dog	Baxter
...		

📖 Mapper



📖 Mapper: Example

📖 Reducer



📖 Reducer: Example

 Combiner

PLEASE NOTICE
Combiners don't replace Reducers.

- Optional
- Minimize network traffic
  - Combine Mapper's output and send chunks of combined data to Reducer
- Reducer can often be used as Combiner

🛠️ We Need Data

Columns

- A Vendor
- A Category
- A Item
- A Item Description
- A Price
- A Origin
- A Destination
- A Rating
- A Remarks

- **Kaggle:**
  Dark Net Marketplace Data (Agora 2014-2015)
- **30.98 MB**
- **csv-file**

Original Dataset: https://www.kaggle.com/philipjames11/dark-net-marketplace-drug-data-agora-20142015

Slightly adapted version: https://github.com/Teapot-418/hadoop-taming-the-elephant/blob/master/darknet-data.csv

HDFS

🛠️ Get Data Into Hadoop

Copy dataset into container

```
docker cp darknet-data.csv \
   ab0978def5ae:/tmp/
```

Put dataset into Hadoop

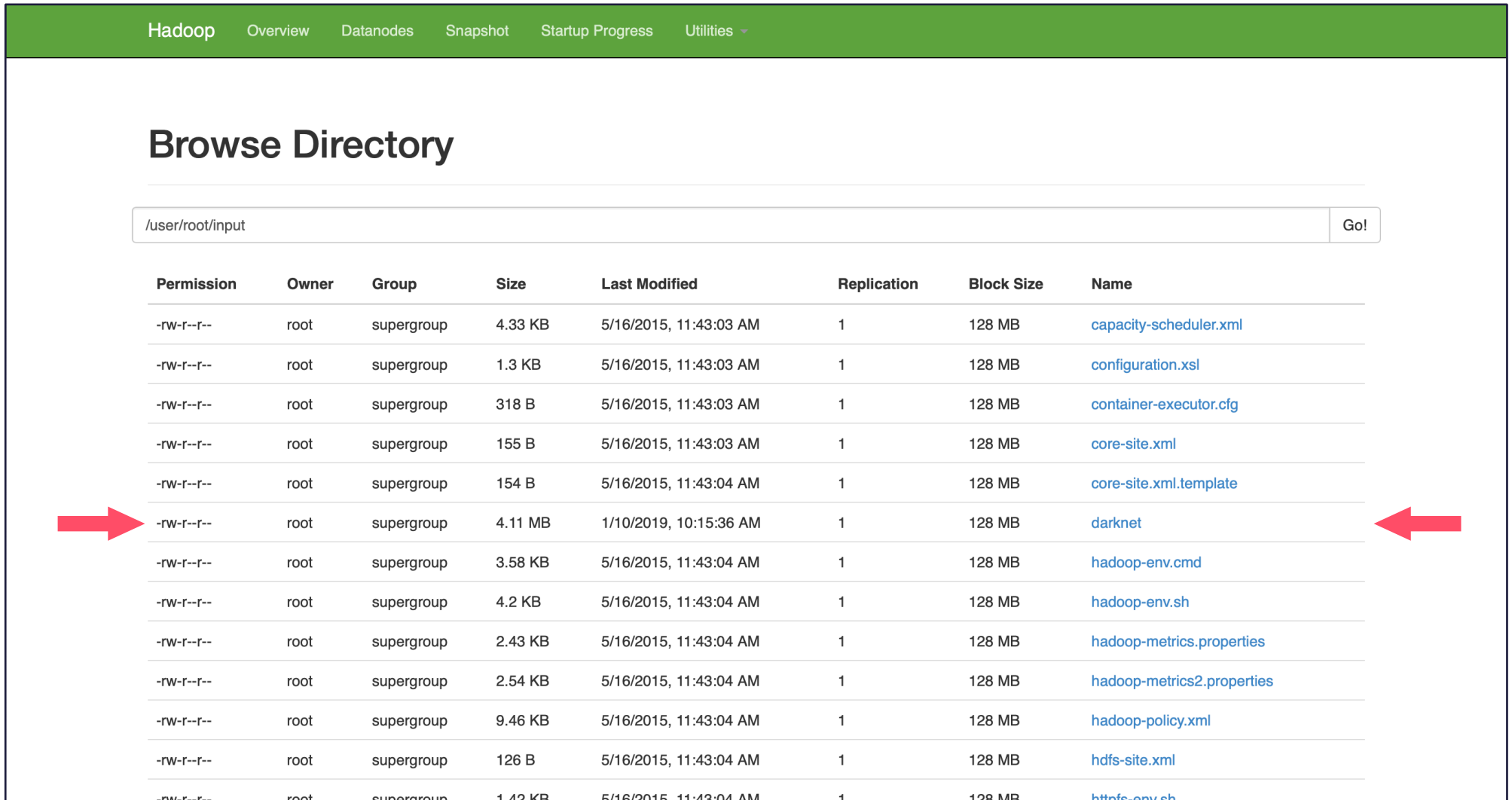
```
$HADOOP_PREFIX/bin/hdfs dfs \
   -put /tmp/darknet-data.csv \
   /user/root/input/darknet
```

🛠 CLI - Check

```
/usr/local/hadoop/bin/hdfs dfs -ls /user/root/input
```

```
bash-4.1# /usr/local/hadoop/bin/hdfs dfs -ls /user/root/input
Found 32 items
-rw-r--r--   1 root supergroup       4436 2015-05-16 05:43 /user/root/input/capacity-scheduler.xml
-rw-r--r--   1 root supergroup       1335 2015-05-16 05:43 /user/root/input/configuration.xsl
-rw-r--r--   1 root supergroup        318 2015-05-16 05:43 /user/root/input/container-executor.cfg
-rw-r--r--   1 root supergroup        155 2015-05-16 05:43 /user/root/input/core-site.xml
-rw-r--r--   1 root supergroup        154 2015-05-16 05:43 /user/root/input/core-site.xml.template
-rw-r--r--   1 root supergroup    4311104 2019-01-10 04:15 /user/root/input/darknet
-rw-r--r--   1 root supergroup       3670 2015-05-16 05:43 /user/root/input/hadoop-env.cmd
-rw-r--r--   1 root supergroup       4302 2015-05-16 05:43 /user/root/input/hadoop-env.sh
-rw-r--r--   1 root supergroup       2490 2015-05-16 05:43 /user/root/input/hadoop-metrics.properties
-rw-r--r--   1 root supergroup       2598 2015-05-16 05:43 /user/root/input/hadoop-metrics2.properties
-rw-r--r--   1 root supergroup       9683 2015-05-16 05:43 /user/root/input/hadoop-policy.xml
-rw-r--r--   1 root supergroup        126 2015-05-16 05:43 /user/root/input/hdfs-site.xml
-rw-r--r--   1 root supergroup       1449 2015-05-16 05:43 /user/root/input/httpfs-env.sh
-rw-r--r--   1 root supergroup       1657 2015-05-16 05:43 /user/root/input/httpfs-log4j.properties
-rw-r--r--   1 root supergroup         21 2015-05-16 05:43 /user/root/input/httpfs-signature.secret
-rw-r--r--   1 root supergroup        620 2015-05-16 05:43 /user/root/input/httpfs-site.xml
-rw-r--r--   1 root supergroup       3518 2015-05-16 05:43 /user/root/input/kms-acls.xml
-rw-r--r--   1 root supergroup       1527 2015-05-16 05:43 /user/root/input/kms-env.sh
-rw-r--r--   1 root supergroup       1631 2015-05-16 05:43 /user/root/input/kms-log4j.properties
```

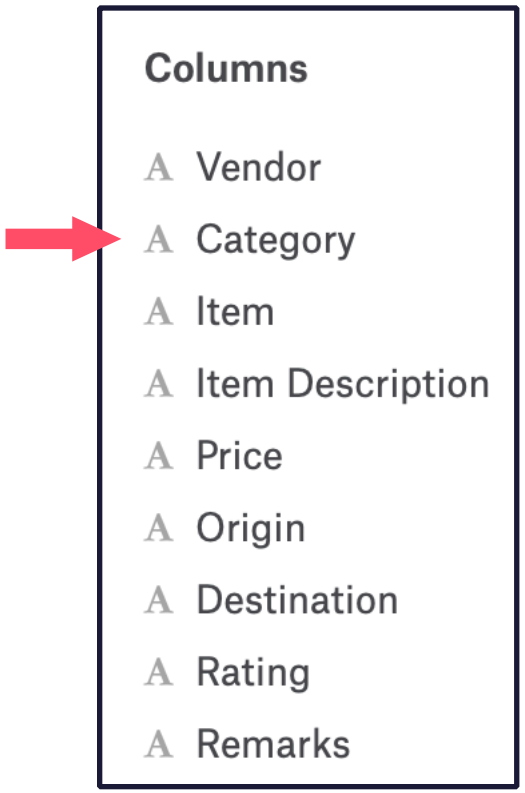
🛠 Web UI - Check



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	4.33 KB	5/16/2015, 11:43:03 AM	1	128 MB	capacity-scheduler.xml
-rw-r--r--	root	supergroup	1.3 KB	5/16/2015, 11:43:03 AM	1	128 MB	configuration.xsl
-rw-r--r--	root	supergroup	318 B	5/16/2015, 11:43:03 AM	1	128 MB	container-executor.cfg
-rw-r--r--	root	supergroup	155 B	5/16/2015, 11:43:03 AM	1	128 MB	core-site.xml
-rw-r--r--	root	supergroup	154 B	5/16/2015, 11:43:04 AM	1	128 MB	core-site.xml.template
-rw-r--r--	root	supergroup	4.11 MB	1/10/2019, 10:15:36 AM	1	128 MB	darknet
-rw-r--r--	root	supergroup	3.58 KB	5/16/2015, 11:43:04 AM	1	128 MB	hadoop-env.cmd
-rw-r--r--	root	supergroup	4.2 KB	5/16/2015, 11:43:04 AM	1	128 MB	hadoop-env.sh
-rw-r--r--	root	supergroup	2.43 KB	5/16/2015, 11:43:04 AM	1	128 MB	hadoop-metrics.properties
-rw-r--r--	root	supergroup	2.54 KB	5/16/2015, 11:43:04 AM	1	128 MB	hadoop-metrics2.properties
-rw-r--r--	root	supergroup	9.46 KB	5/16/2015, 11:43:04 AM	1	128 MB	hadoop-policy.xml
-rw-r--r--	root	supergroup	126 B	5/16/2015, 11:43:04 AM	1	128 MB	hdfs-site.xml

🛠 Maven Dependencies

🛠 What We Want



Category
Drugs/Cannabis/Weed
Drugs/Ecstasy/Pills
Services/Other
...

Main-Category	Quantity
Drugs	418
Services	42
...	

🛠️ Create Mapper

Extend Hadoop's Mapper

```
Mapper< [input-key], [input-value],
   [output-key], [output-value] >
```

Override "map"-function

```
public void map(
   [input-key] key,
   [input-value] value,
   Context context
)
```

🛠 Create Mapper

PLEASE NOTICE
Hadoop uses special value-types.

```
public class MainCategoryMapper extends
    Mapper<LongWritable, Text, Text, LongWritable> {
    @Override
    public void map(LongWritable key, Text value, Context context) throws […] {
        String line = value.toString();
        String[] lineData = line.split(",");
        String[] categories = lineData[1].split("/");
        if(categories.length > 0) {
            context.write(new Text(categories[0]), key);
        }
    }
}
```

Create output

🛠 Create Reducer

Extend Hadoop's Reducer

```
Reducer< [input-key], [input-value],
   [output-key], [output-value] >
```

Override "reduce"-function

```
public void reduce(
   [input-key] key,
   Iterable<[input-value]> values,
   Context context
)
```

🛠 Create Reducer

```
public class CategoryCountReducer extends
     Reducer<Text, LongWritable, Text, IntWritable> {
    @Override
    public void reduce(Text key, Iterable<LongWritable> values, Context context)
           throws […] {
        int count = 0;
        for (LongWritable value : values) {
            count++;
        }
        context.write(key, new IntWritable(count));
    }
}
```

Create output

🛠️ Create Entrypoint

A

Create Job

B

Define Location of input and output

C

Define Mapper and Reducer

```
A  Job job = Job.getInstance();
   job.setJobName("Main category count");

B  FileInputFormat.addInputPath(job, new Path(inputPath));
   FileOutputFormat.setOutputPath(job, new Path((outputPath)));

C  job.setMapperClass(MainCategoryMapper.class);
   job.setReducerClass(CategoryCountReducer.class);

D  job.setMapOutputKeyClass(Text.class);
   job.setMapOutputValueClass(LongWritable.class);

E  job.setOutputKeyClass(Text.class);
   job.setOutputValueClass(IntWritable.class);

F  job.setJarByClass(MainCategoryCount.class);
   System.exit(job.waitForCompletion(true) ? 0 : 1);
```

D

Define Mapper-output (key & value)

E

Define Reducer-output (key & value)

F

Start and wait for completion

🛠 Run it

Build executable jar

```
mvn clean package
```

Find jar into /target

Copy jar into container

```
docker cp darknet-mapreduce.jar\
ab0978def5ae:/tmp/darknet-mapreduce.jar
```

Execute

```
$HADOOP_PREFIX/bin/hadoop \
   jar /tmp/darknet-mapreduce.jar
```

🛠 Run it

```
[…] INFO mapreduce.Job: The url to track the job: http://ab0978def5ae:8088/
proxy/application_1547046677403_0012/
[…] INFO mapreduce.Job: Running job: job_1547046677403_0012
[…] INFO mapreduce.Job: Job job_1547046677403_0012 running in uber mode : false
[…] INFO mapreduce.Job:  map 0% reduce 0%
[…] INFO mapreduce.Job:  map 100% reduce 0%
[…] INFO mapreduce.Job:  map 100% reduce 100%
[…] INFO mapreduce.Job: Job job_1547046677403_0012 completed successfully
```

🛠 CLI - Check

```
$HADOOP_PREFIX/bin/hdfs dfs -cat output/darknet/main-category-count/*
```

```
Chemicals       1
Counterfeits    202
Data    38
Drug paraphernalia      2
Drugs   11834
Electronics     41
Forgeries       101
Info    15
Information     13
Jewelry 23
Other   97
Services        179
Tobacco 6
Weapons 83
```

Main-Category	Quantity
Chemicals	1
Counterfeits	202
Drug paraphernalia	2
Drugs	11834
…	

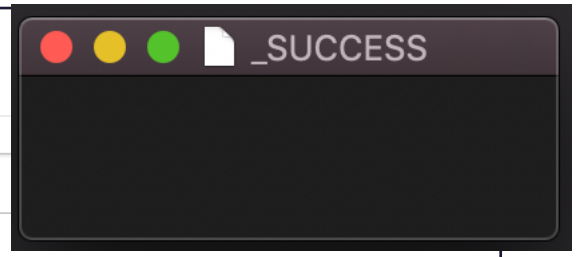
🛠 Web UI - Check

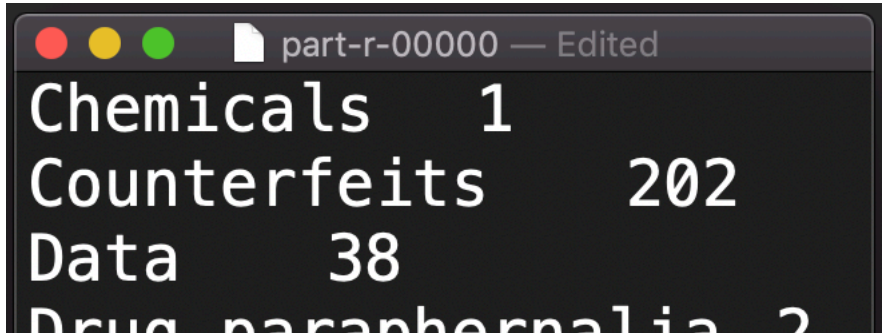
Browse Directory

/user/root/output/darknet/main-category-count

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	0 B	1/10/2019, 2:13:26 PM	1	128 MB	_SUCCESS
-rw-r--r--	root	supergroup	176 B	1/10/2019, 2:13:26 PM	1	128 MB	part-r-00000

Hadoop, 2014.



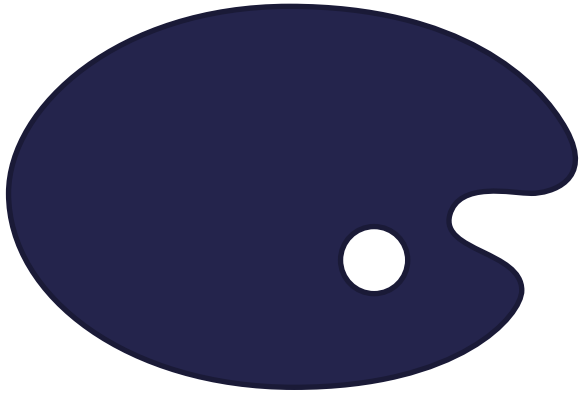


🛠 Remove Results

```
$HADOOP_PREFIX/bin/hdfs dfs -rm -r <folder-name>
```

Be Creative

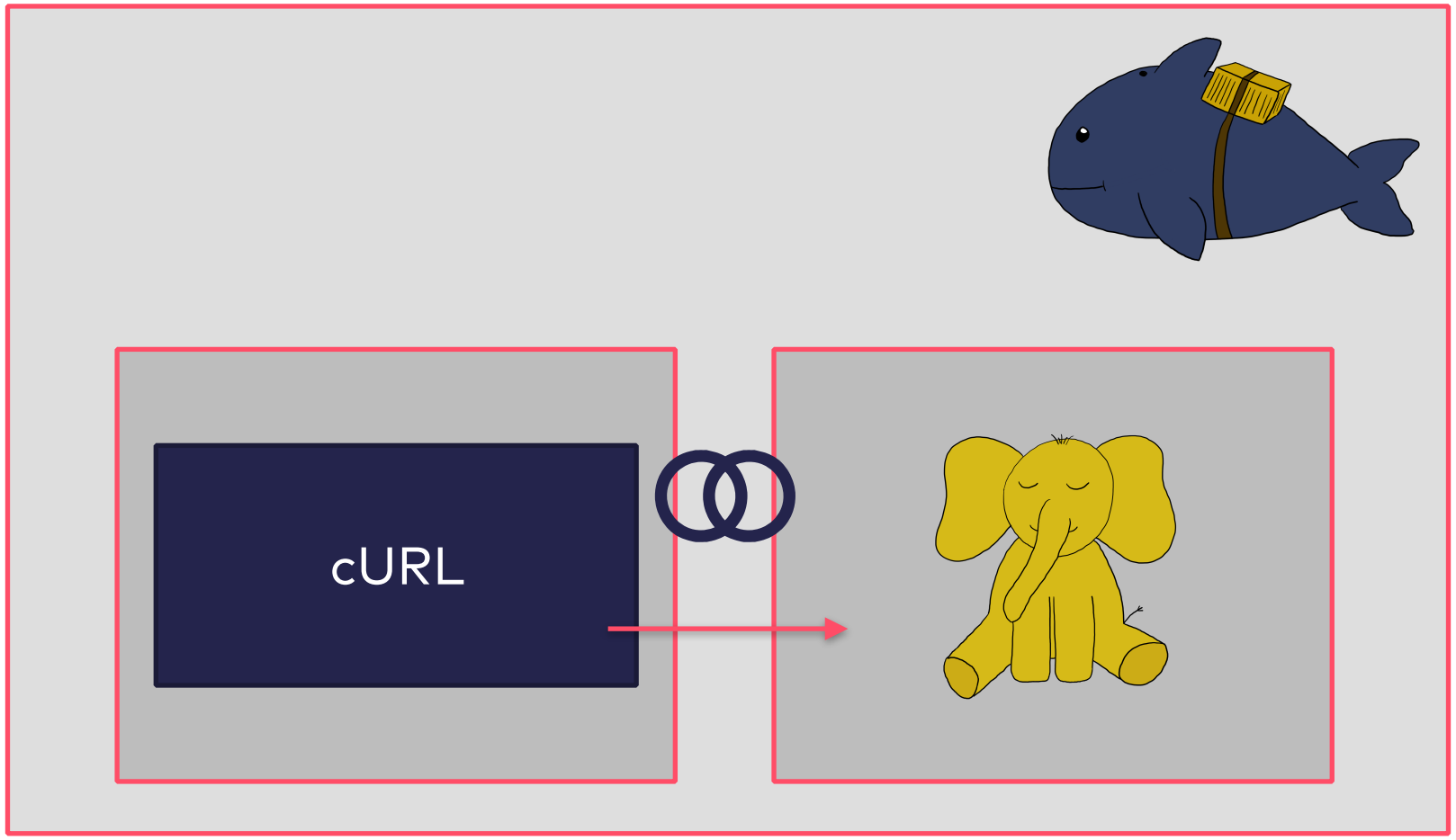
- Work with this dataset and try to do other things like:
  - Counting detailed categories
  - List the maximum price per main-category
  - ...



Run First MapReduce-job

Displaying Results (Http)

📖 Goal - Part 4



📖 HttpFS?

Hadoop HDFS over HTTP

🛠️ Linking Docker-Containers

```
docker run \
   -it \
   --link 7680676ced68:hadoop \
   –name curl_container \
   ubuntu:latest
```

Start in interactive mode directly

Link to container and give internal name

```
/ # cat /etc/hosts
127.0.0.1       localhost
::1     localhost ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
172.17.0.2      hadoop 7680676ced68 compassionate_robinson
```

🛠️ Navigate through HDFS

The Base URL

```
http://hadoop:50070/webhdfs/v1/user
```

Browsing Directories

```
curl -i -L \
“http://hadoop:50070/webhdfs/v1/user/
root/output/?op=LISTSTATUS"
```

Open Files

```
curl -i -L \
“http://hadoop:50070/webhdfs/v1/user/
root/output/darknet/category-count/
part-00000?op=OPEN”
```

🛠 Navigate through HDFS

Browsing Directories


```
{
    "FileStatuses": {
        "FileStatus": [{
            "accessTime": 0,
            "blockSize": 0,
            "childrenNum": 1,
            "fileId": 16435,
            "group": "supergroup",
            "length": 0,
            "modificationTime": 1547137495047,
            "owner": "root",
            "pathSuffix": "darknet",
            "permission": "755",
            "replication": 0,
            "storagePolicy": 0,
            "type": "DIRECTORY"
        }]
    }
}
```


Open Files

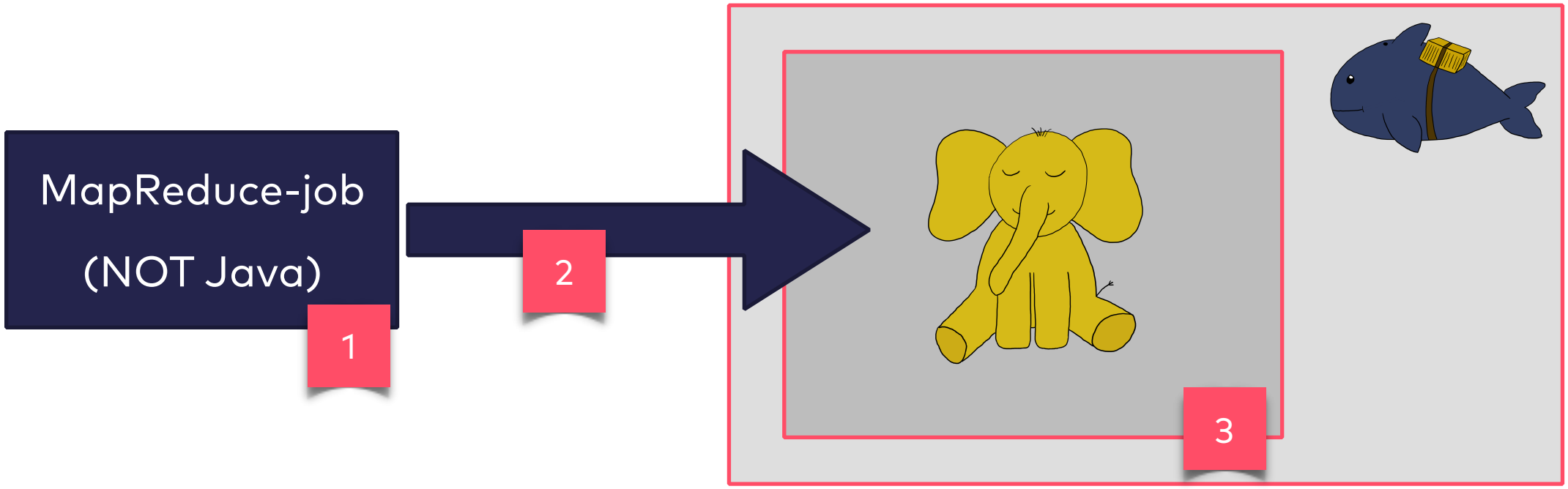
```
Pragma: no-cache
Content-Type: application/octet-stream
Location: http://7680676ced68:50075/webl
dress=7680676ced68:9000&offset=0
Content-Length: 0
Server: Jetty(6.1.26)

HTTP/1.1 200 OK
Access-Control-Allow-Methods: GET
Access-Control-Allow-Origin: *
Content-Type: application/octet-stream
Connection: close
Content-Length: 1905

Counterfeits/Accessories        1
Counterfeits/Clothing   8
Counterfeits/Electronics        10
Counterfeits/Money      5
Counterfeits/Watches    36
```

Fetch Data Via HttpFS

I don't like Java 😕

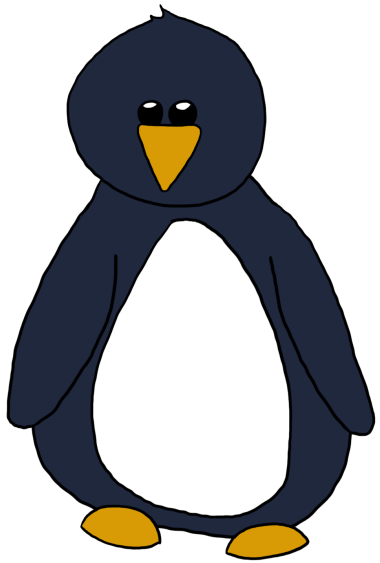
5

📖 Goal - Part 5



📖 Hadoop Streaming API

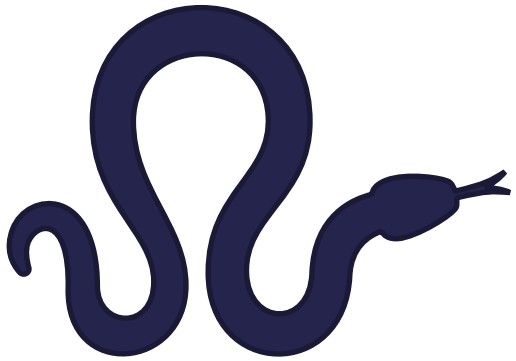
stdin



stdout

Reducer receives keys in order

🛠️ Example - Introduction



Columns

- A Vendor
- A Category
- A Item
- A Item Description
- A Price
- A Origin
- A Destination
- A Rating
- A Remarks

Category
Drugs/Cannabis/Weed
Drugs/Ecstasy/Pills
Services/Other
…

Category	Quanti ty
Drugs/Cannabis/Weed	418
Drugs/Ecstasy/Pills	42
…	

🛠 Mapper

```
#!/usr/bin/env python
import sys

id = 0

for line in sys.stdin:
  val = line.strip()
  data = val.split(',')
  category = data[1]
  print(category + '\t' + str(id))
  id = id + 1
```

Handle stdin line by line

Handover data to Reducer

🛠 Reducer

```
previous_key = None
count = 0

for line in sys.stdin:
    (key, val) = line.strip().split('\t')
    if previous_key is None:
        previous_key = key
    if key == previous_key:
        count = count + 1
    else:
        print(key + '\t' + str(count))
        count = 1
    previous_key = key
```

Get key and value from line

Print result for key, if last

Remember key

🛠️ Copy into Container

Copy Mapper into container

```
docker cp mapper.py ab0978def5ae:/tmp/
```

Copy Reducer into container

```
docker cp reducer.py ab0978def5ae:/tmp/
```

Make files executable

```
chmod +x mapper.py
chmod +x reducer.py
```

🛠️ Run it - Base Command

```
$HADOOP_PREFIX/bin/hadoop jar \
  $HADOOP_PREFIX/share/hadoop/tools/lib/hadoop-streaming-2.7.0.jar
```

🛠️ Run it - Arguments

```
-files /tmp/mapper.py,/tmp/reducer.py \
-input input/darknet \
-output output/darknet/category-count \
-mapper /tmp/mapper.py \
-reducer /tmp/reducer.py
```

- Ships the listed files to the cluster
- Input-file for MapReduce-job
- Output-path for MapReduce-job
- Set Mapper
- Set Reducer

🛠 CLI - Check

```
$HADOOP_PREFIX/bin/hdfs dfs -cat output/darknet/
category-count/*
```

```
Counterfeits/Accessories        1
Counterfeits/Clothing   8
Counterfeits/Electronics        10
Counterfeits/Money      5
Counterfeits/Watches    36
Data/Accounts   143
Data/Pirated    30
Data/Software   3
Drug paraphernalia/Pipes        5
Drugs/Barbiturates      2
Drugs/Benzos    4
Drugs/Cannabis/Concentrates     634
Drugs/Cannabis/Edibles  589
Drugs/Cannabis/Hash     24
```

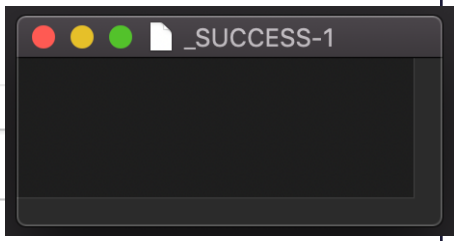
🛠 Web UI - Check

Browse Directory

/user/root/output/darknet/category-count

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	0 B	1/10/2019, 5:25:10 PM	1	128 MB	_SUCCESS
-rw-r--r--	root	supergroup	1.86 KB	1/10/2019, 5:25:10 PM	1	128 MB	part-00000

Hadoop, 2014.



part-00000

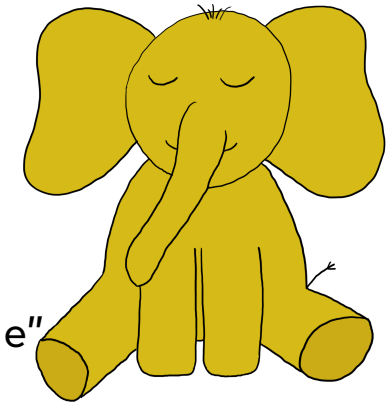
```
Counterfeits/Accessories        1
Counterfeits/Clothing   8
Counterfeits/Electronics        10
Counterfeits/Money      5
Counterfeits/Watches    36
Data/Accounts   143
Data/Pirated    30
Data/Software   3
Drug paraphernalia/Pipes        5
Drugs/Barbiturates      2
Drugs/Benzos    4
Drugs/Cannabis/Concentrates     634
Drugs/Cannabis/Edibles  589
Drugs/Cannabis/Hash     24
Drugs/Cannabis/Seeds    270
Drugs/Cannabis/Shake/trim       6
Drugs/Cannabis/Synthetics       8
```

Use Hadoop Without Java

What About the Rest?

📖 Naming

"The name my kid gave a stuffed yellow elephant. Short, relatively easy to spell and pronounce, meaningless, and not used elsewhere: those are my naming criteria. Kids are good at generating such. Googol is a kid's term."
(Doug Cutting)



Excerpt From: Tom White. "Hadoop: The Definitive Guide"

📖 Core of Hadoop

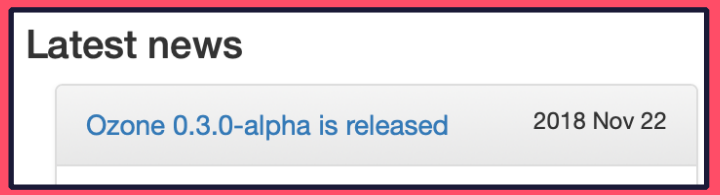
Hadoop Common

MapReduce

HDFS

YARN

Ozone



6

📖 HDFS

6

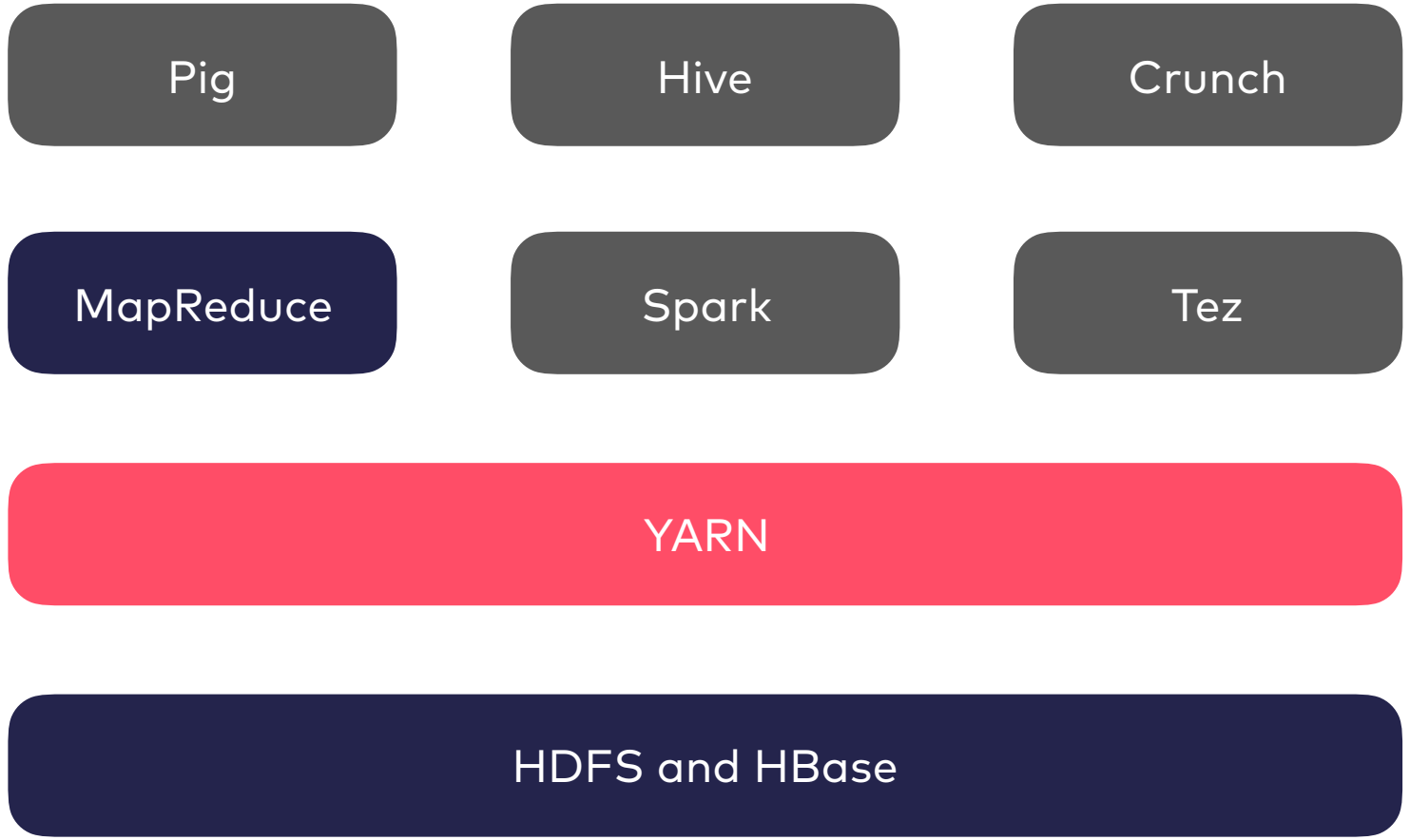
📖 YARN

Processing Frameworks

APPLICATION

COMPUTE

STORE



Small Introduction Into Hadoop Theory

I Want To Try It!

@ Home

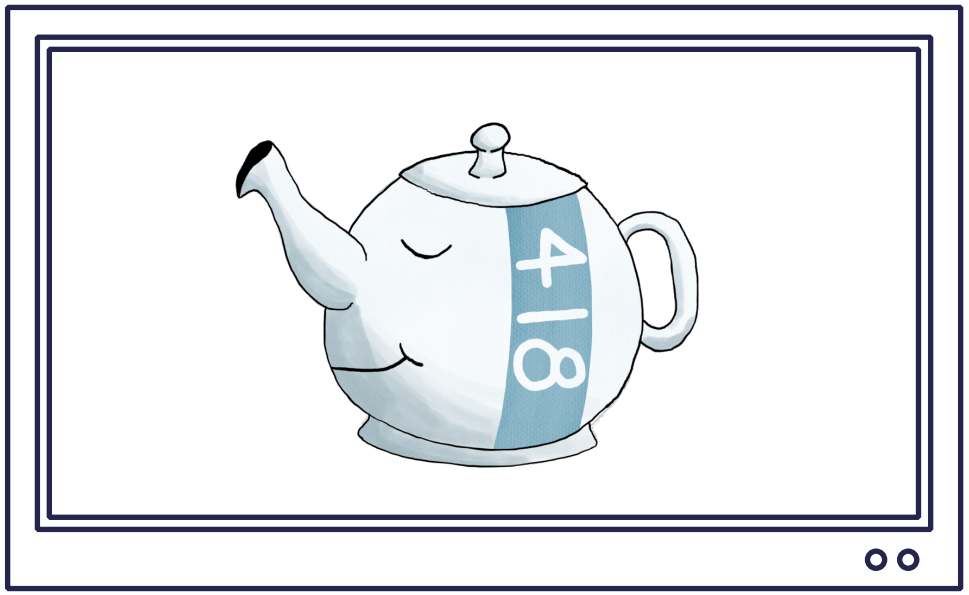
- More Information About Hadoop
  - http://hadoop.apache.org
  - Tom White - “Hadoop - The Definitive Guide” (4th)
  - Alex Holmes - “Hadoop In Practice” (2nd)

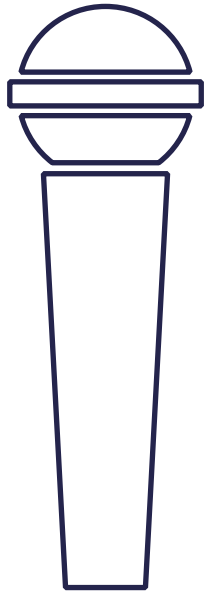
- More Information Regarding This Talk
  - www.teapot418.de
  - https://github.com/Teapot-418/hadoop-taming-the-elephant

The End...

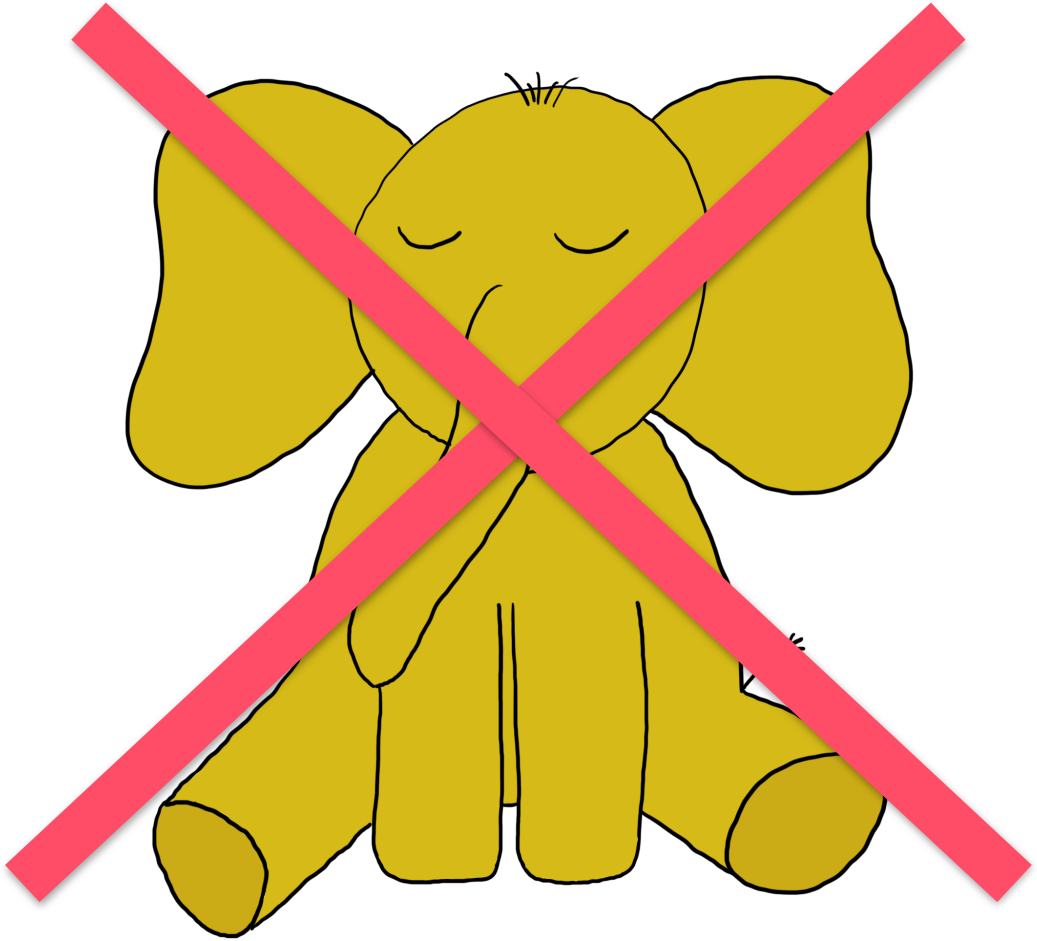
Share Discoveries

MY GOAL
Breaking the ice and giving a short introduction

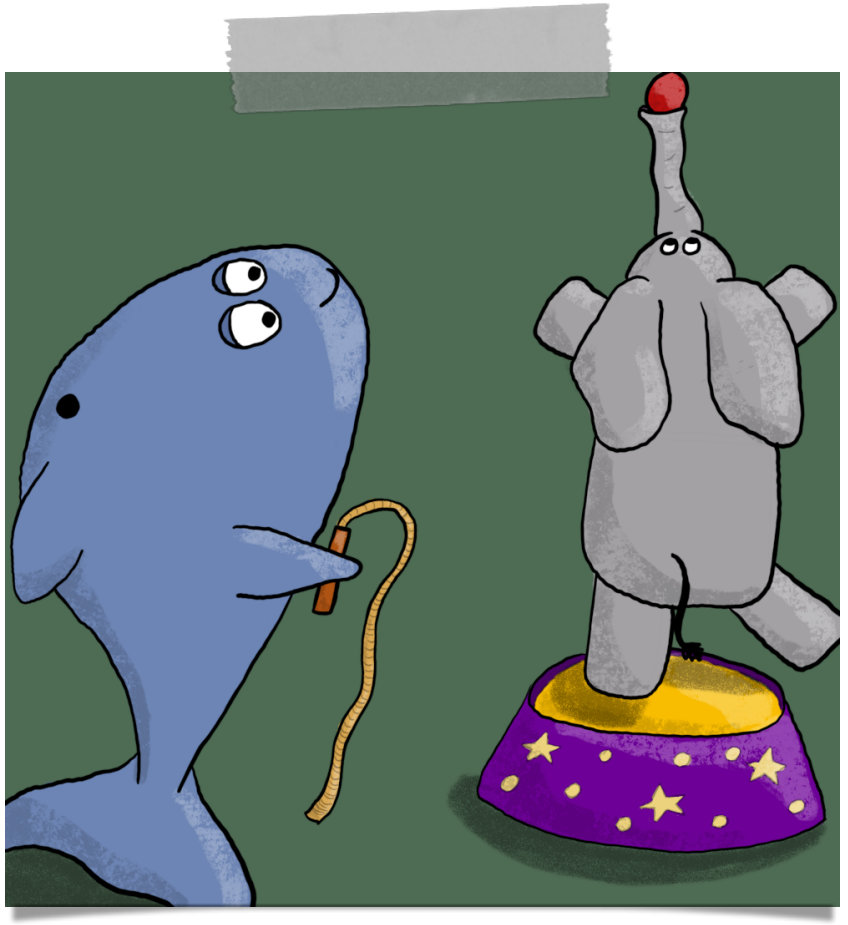




What's Up Now?



Thank you!



INNOQ
www.innoq.com

Lisa Maria Moritz
lisa.moritz@innoq.com

+49 176 64 63 00 28
Teapot4181

innoQ Deutschland GmbH

Krischerstr. 100
40789 Monheim am Rhein
Germany
+49 2173 3366-0

Ohlauer Str. 43
10999 Berlin
Germany
+49 2173 3366-0

Ludwigstr. 180E
63067 Offenbach
Germany
+49 2173 3366-0

Kreuzstr. 16
80331 München
Germany
+49 2173 3366-0

innoQ Schweiz GmbH

Gewerbestr. 11
CH-6330 Cham
Switzerland
+41 41 743 0116