Running hadoop on Grid'5000

Vinicius Cogo
vielmo@lasige.di.fc.ul.pt

Marcelo Pasin
pasin@di.fc.ul.pt

Andrea Charão
andrea@inf.ufsm.br

Outline

1 - Introduction

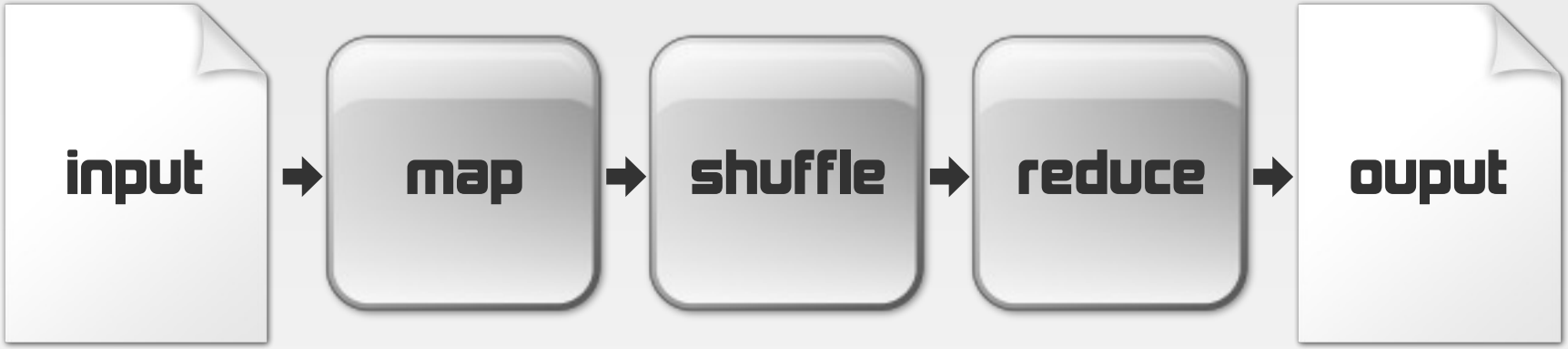
- **Main goal:** Introduce the development of MapReduce applications using the Hadoop framework.
- **Important:** Prepare a Hadoop environment.
- Grid'5000 Hadoop environment available.
- www.grid5000.fr/index.php/Run_Hadoop_On_Grid'5000

2 - MapReduce

- Programming model.
- Proposed by Google in 2004.
- Based on LISP ***map*** and ***reduce*** functions.
- Uses the parallelism to share the **data load**, instead of parallelizing **processing loads**.

2 - MapReduce

- MapReduce data flow example:



3 - Hadoop

- Set of sub-projects.

Pig	Chukwa	Hive	HBase
MapReduce	HDFS		ZooKeeper
Core		Avro	

- Yahoo!'s MapReduce implementation.
- Free and open-source framework.

3 - Hadoop

- **Split** = piece of input

“Lorem ipsum dolor sit amet”
“Lisbon 20”

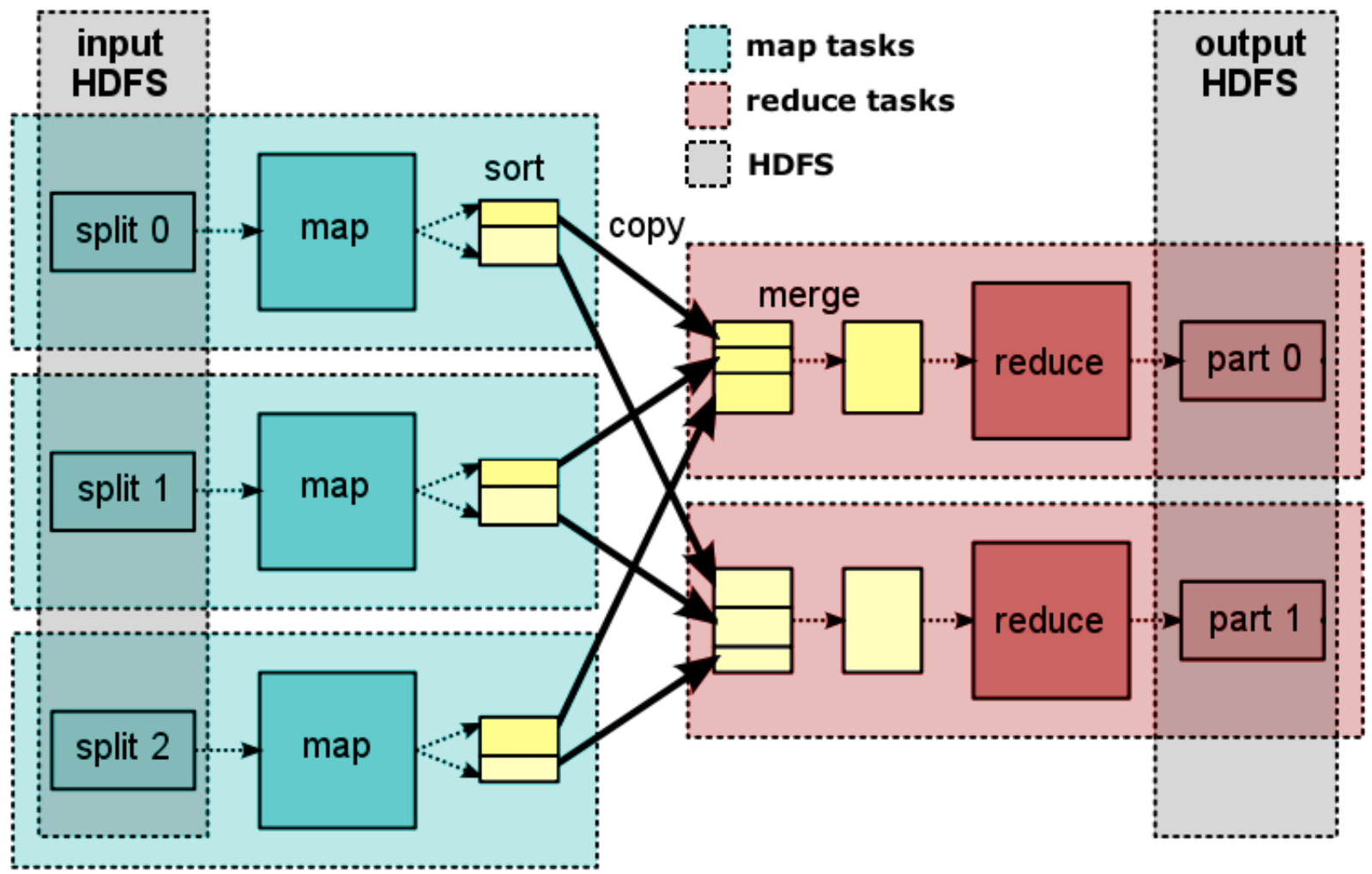
- **Information** = <**key**, **value**> pairs

<0, Lorem ipsum dolor sit amet>
<Lisbon, 20>

- **Task** = part of the work (mapTask or reduceTask)

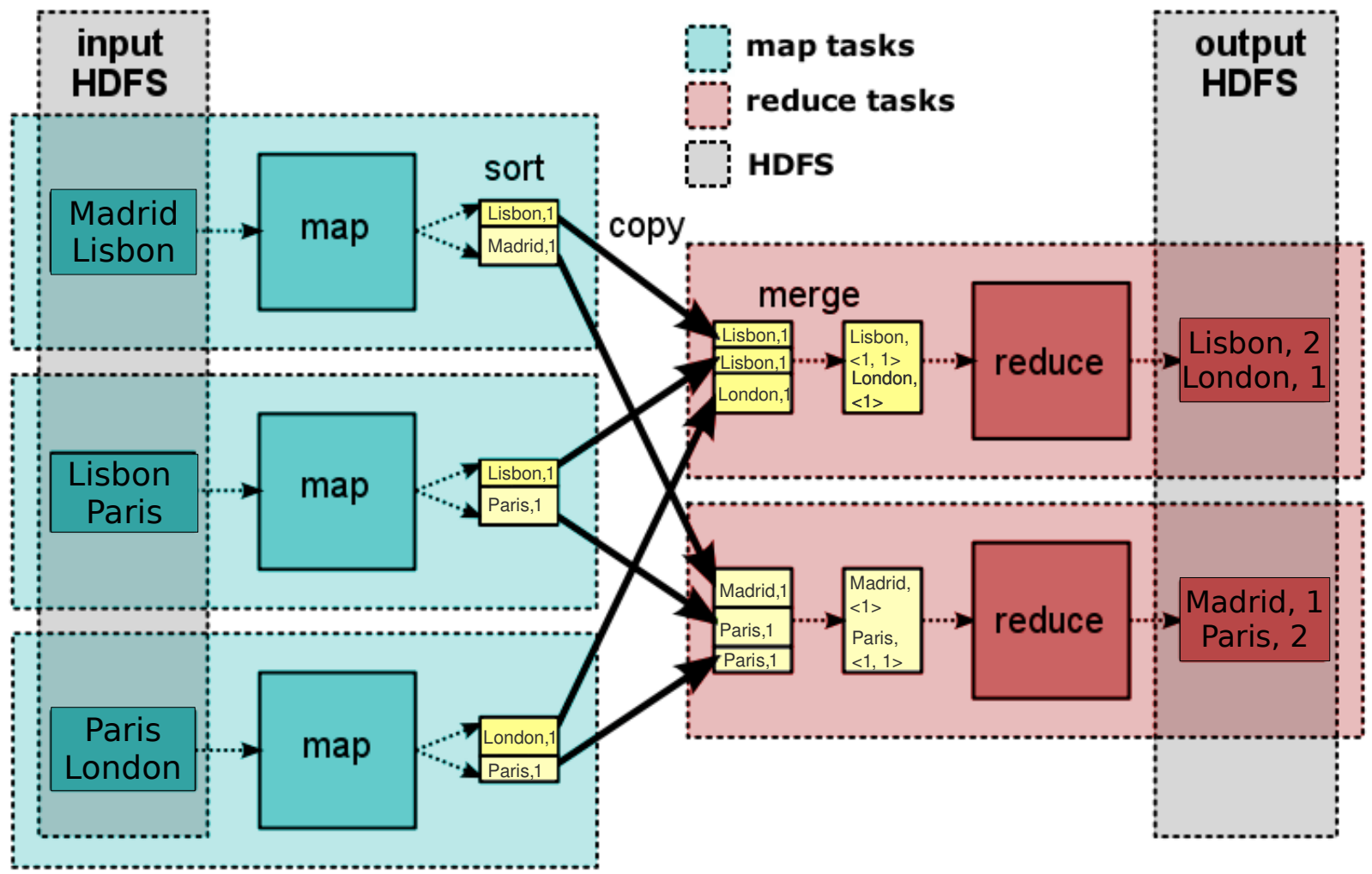
- **Job** = entire work

3 - Hadoop



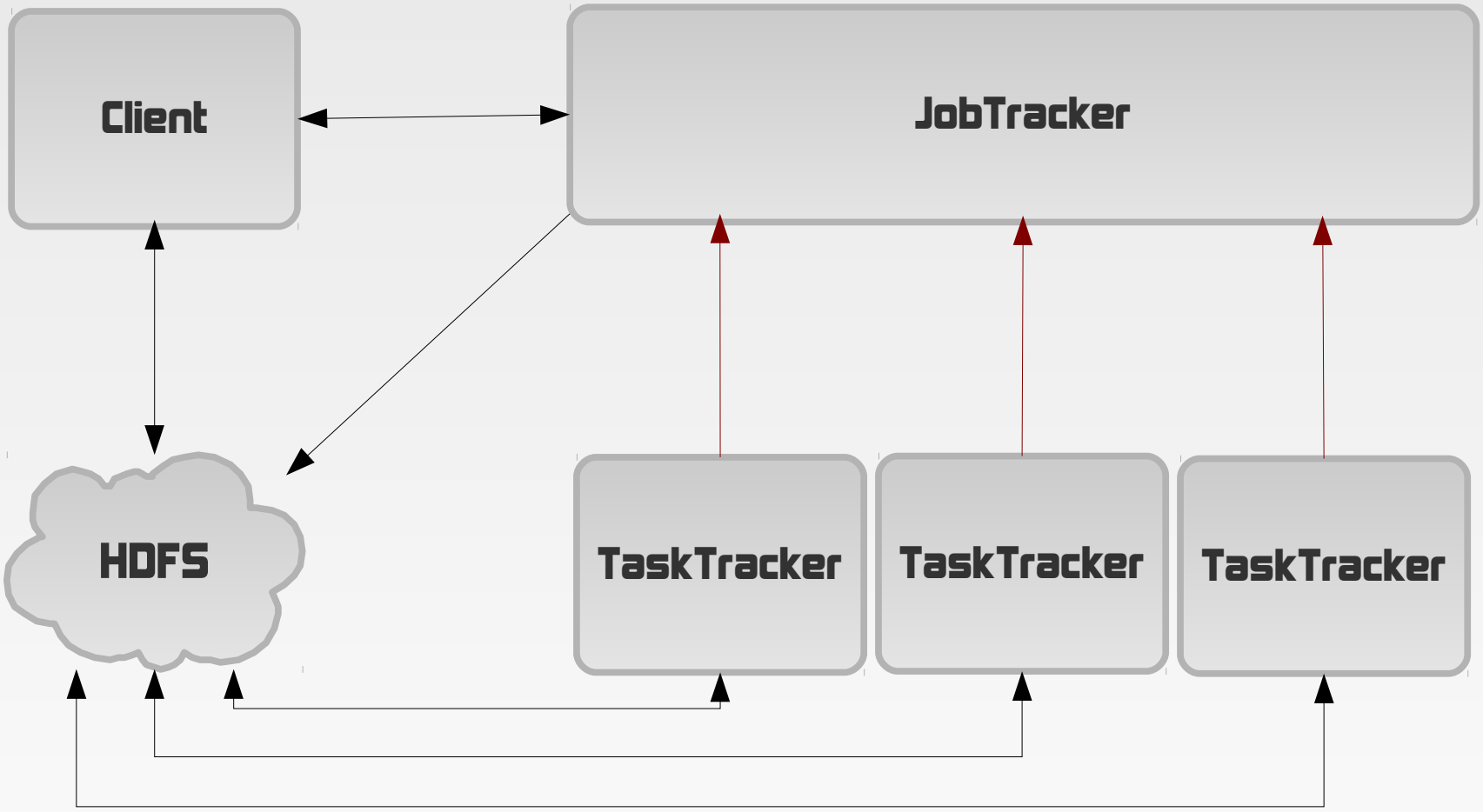
Input › Map | Shuffle | Reduce › Output

3 - Hadoop



Input › Map | Shuffle | Reduce › Output

3 - Hadoop

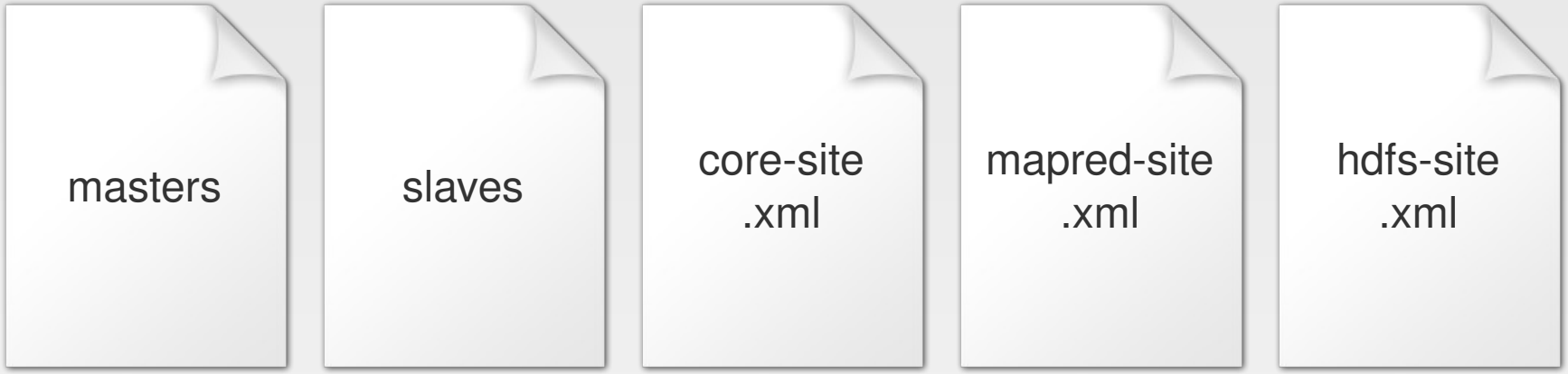

Client
JobTracker
HDFS
TaskTracker
TaskTracker
TaskTracker

4 - How to Install Hadoop?

- Install Java 1.6.XX.
- Configure SSH to works based on RSA or DSA key authentication method.
- Download a Hadoop version.
- Unzip the files in some folder, e. g. $PATH = /opt/hadoop/.
- Configure the **JAVA_HOME** property in **hadoop-env.sh** file, located in $PATH/conf/ folder.

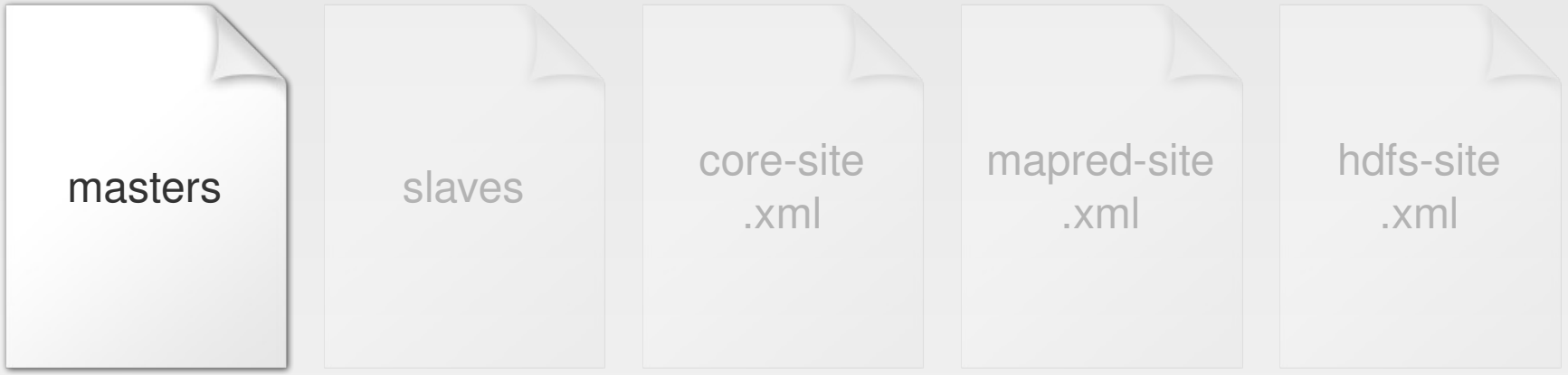
```
From:
# export JAVA_HOME=/usr/lib/j2sdk1.5-sun

To:
export JAVA_HOME=/usr/lib/jvm/java-6-sun
```

5 - How to Configure Hadoop?



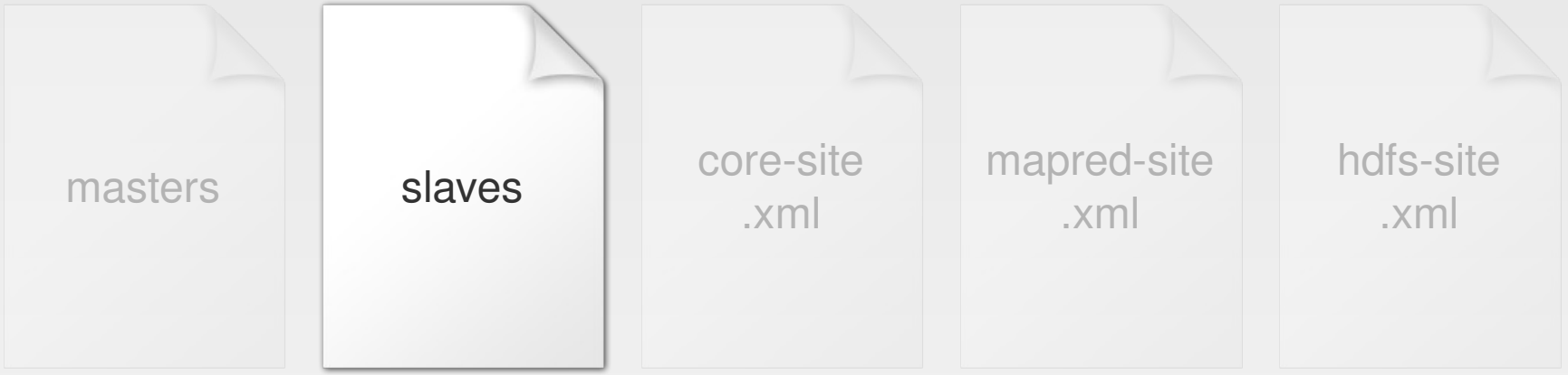

```
$PATH/conf/
```

5 - How to Configure Hadoop?



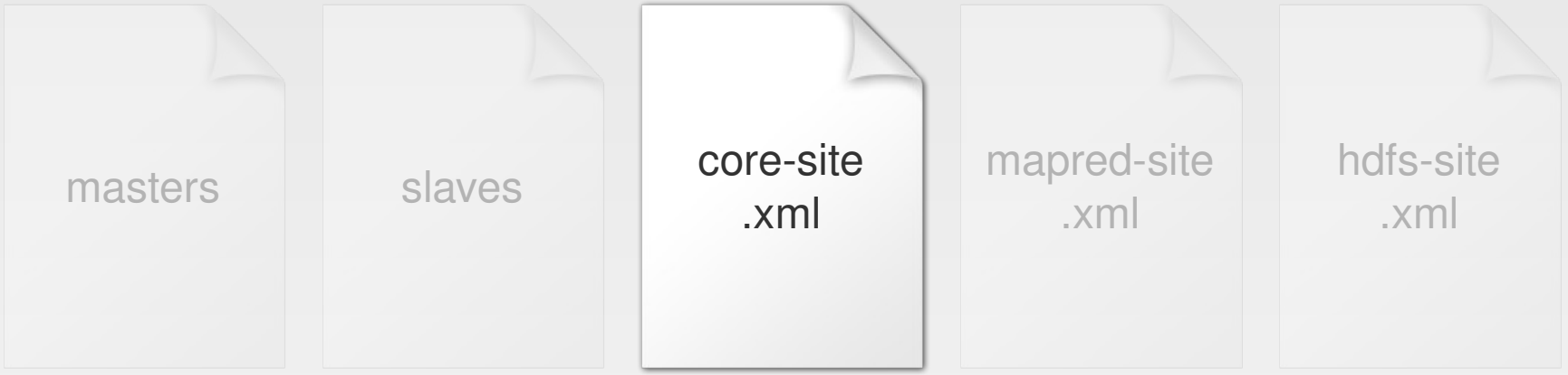

```
node01.site.grid5000.fr
```

5 - How to Configure Hadoop?



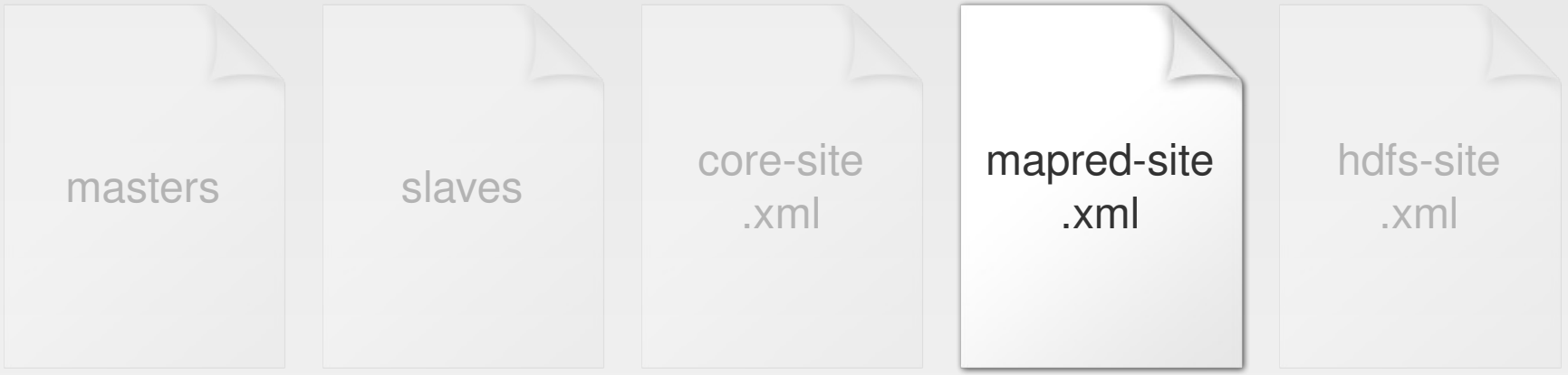

```
node01.site.grid5000.fr
node02.site.grid5000.fr
node03.site.grid5000.fr
...
nodeNN.site.grid5000.fr
```

5 - How to Configure Hadoop?



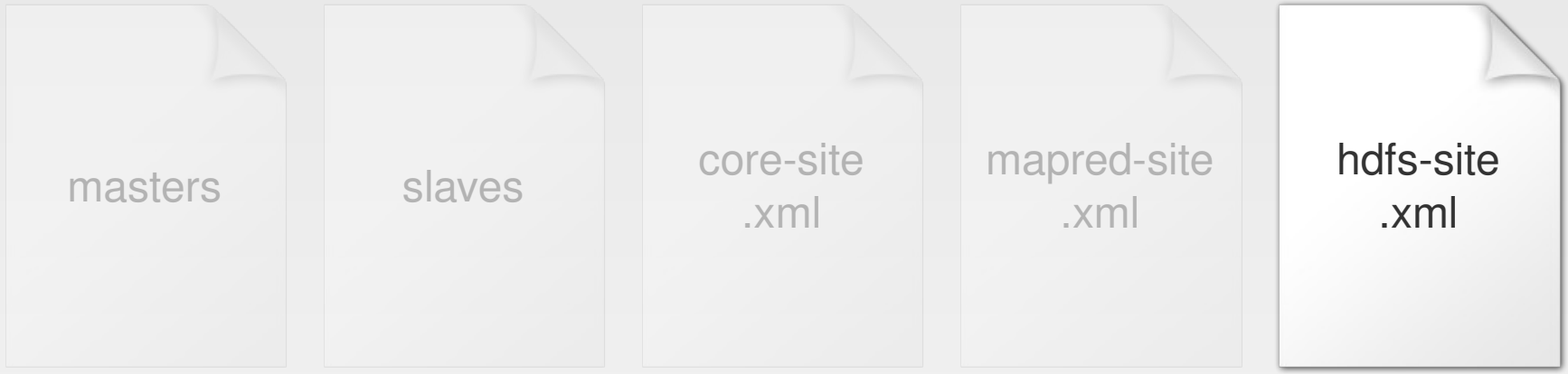

```
<?xml version="1.0"?>
<configuration>
     <property>
          <name>hadoop.tmp.dir</name>
          <value>/tmp/hadoop-\${user.name}</value>
     </property>
     <property>
          <name>fs.default.name</name>
          <value>hdfs://node01.site.grid5000.fr:54310</value>
     </property>
</configuration>
```

5 - How to Configure Hadoop?




```
<?xml version="1.0"?>
<configuration>
     <property>
          <name>mapred.job.tracker</name>
          <value>hdfs://node01.site.grid5000.fr:54311</value>
     </property>
</configuration>
```

5 - How to Configure Hadoop?




```
<?xml version="1.0"?>
<configuration>
     <property>
          <name>dfs.replication</name>
          <value>1</value>
     </property>
</configuration>
```

6 - How to Start Hadoop?

Connect to the Master node:

```
ssh user@node01.site.grid5000.fr
```

Stop all the current HDFS and Hadoop MapReduce instances:

```
$PATH/bin/stop-all.sh
```

Format the HDFS namenode:

```
$PATH/bin/hadoop namenode -format
```

Initialize the HDFS:

```
$PATH/bin/start-dfs.sh
```

Initialize the Hadoop MapReduce:

```
$PATH/bin/start-mapred.sh
```

7 - How to Run Hadoop Applications?

Example of a generic call:

```
$PATH/bin/hadoop jar file.jar [<parameters>]
```

Example of a real call:

```
$PATH/bin/hadoop jar                                    \
                     $PATH/hadoop-0.20.1-examples.jar   \
                     Pi 4 1000
```

7 - How to Run Hadoop Applications?

- Some important HDFS commands:

```
$PATH/bin/hadoop dfs -ls [folderName]

$PATH/bin/hadoop dfs -rm [fileName]

$PATH/bin/hadoop dfs -mkdir [folderName]

$PATH/bin/hadoop dfs -copyFromLocal  [fileLocal] [fileHDFS]

$PATH/bin/hadoop dfs -copyToLocal [fileHDFS] [fileLocal]
```

8 - How to Use Hadoop Environment on Grid'5000?

Allocate, with OAR, the quantity of nodes you will need for the Hadoop job:

```
oarsub -I -t deploy -l nodes=NUM_HOSTS,walltime=HH:MM:SS
```

Deploy the Hadoop environment and run the script at the nodes allocated for the job.

```
kadeploy3                                                          \
      -a ~vvielmocogo/hadoop/0.20.1/lenny-x64-nfs-hadoop.dsc3      \
      -f  $OAR_FILE_NODES                                          \
      -k ~/.ssh/id_rsa.pub                                         \
      -s ~vvielmocogo/hadoop/0.20.1/config.sh
```

9 - How to Develop Hadoop Applications?

- Examples are in the folder $PATH/src/examples/org/apache/hadoop/examples/

- What do you need to program?

```
public void map(Type key, Type value, Context context)
throws IOException, InterruptedException
{
    // map code
}


public void reduce(Type key, Type value, Context context)
throws IOException, InterruptedException
{
    // reduce code
}
```

9 - How to Develop Hadoop Applications?

- WordCount.java

```
public void map(Object key, Text value, Context context)
throws IOException, InterruptedException
{
      IntWritable one = new IntWritable(1);
      Text word = new Text();
      StringTokenizer itr = new StringTokenizer(value.toString());
      while (itr.hasMoreTokens()) {
            word.set(itr.nextToken());
            context.write(word, one);
      }
}

public void reduce(Text key, Iterable<IntWritable> values, Context context)
throws IOException, InterruptedException
{
      IntWritable result = new IntWritable();
      int sum = 0;
      for (IntWritable val : values) {
            sum += val.get();
      }
      result.set(sum);
      context.write(key, result);
}
```

9 - How to Develop Hadoop Applications?

- **Exercise 1:**

Copy to HDFS one file

```
$PATH/bin/hadoop dfs -copyFromLocal                    \
    ~vvielmocogo/hadoop0.20.1/gutenberg/               \
    gutenberg
```

- For each word in input, returns the number of the line that have the bigger quantity of words.

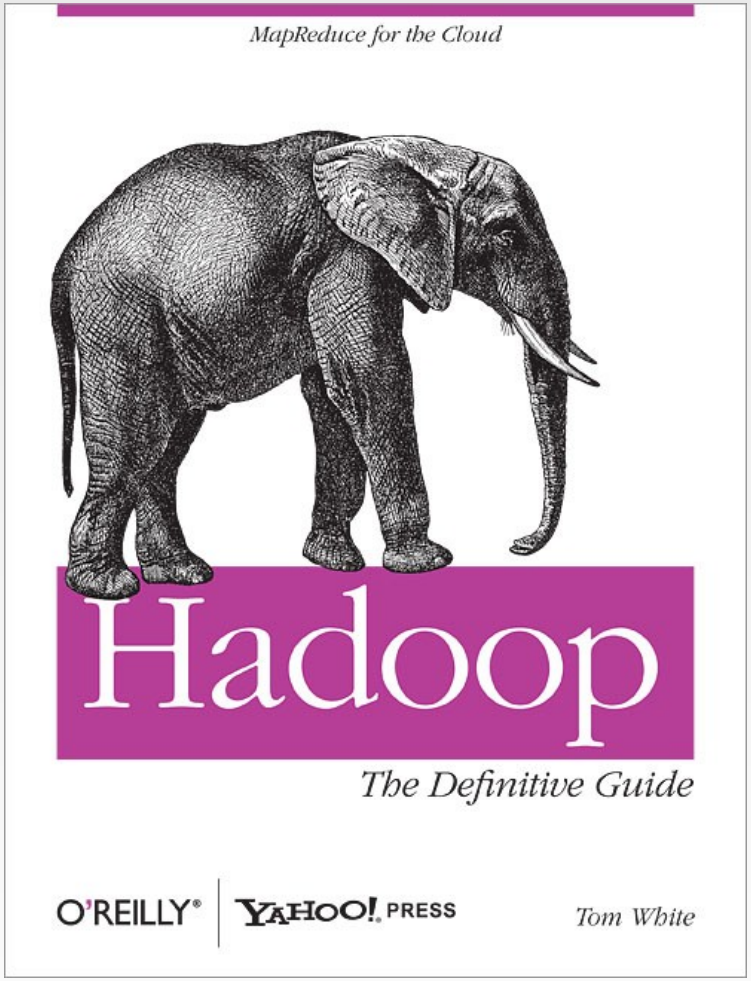
- **Exercise 2:**

For each word in input, returns the list of lines which contents the word, just like an index.

P.S.1: To add a new exercise in hadoop-examples JAR, do you have to create a new Java file in examples folder add it's class in ExampleDriver.java file.
P.S.2: To generate the hadoop-examples JAR, use:

```
ant -Doffline=true examples
```

10 - Read More



Grid'5000

FTH GRID

Extras

Grid'5000 Hadoop Environment:

- Deploy lenny-x64-nfs
- Extract the Hadoop files and configure $JAVA_HOME
- Create a new environment (tgz-g5k)
- Create the descriptor file
- Create the **script** to configure the environment:
  - Fills masters and slaves files based on $OAR_FILE_NODES
  - Fills others configuration XML files
  - Copy configuration files for all nodes
  - Startup Hadoop on master node