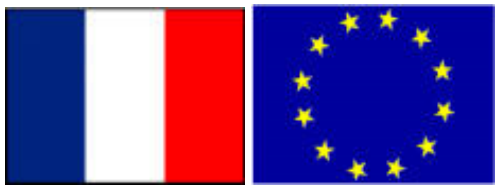


Big Data – TP1 Part 1

Using HDFS & Spark on the DCE clusters of CentraleSupelec

(Data Center for Education)

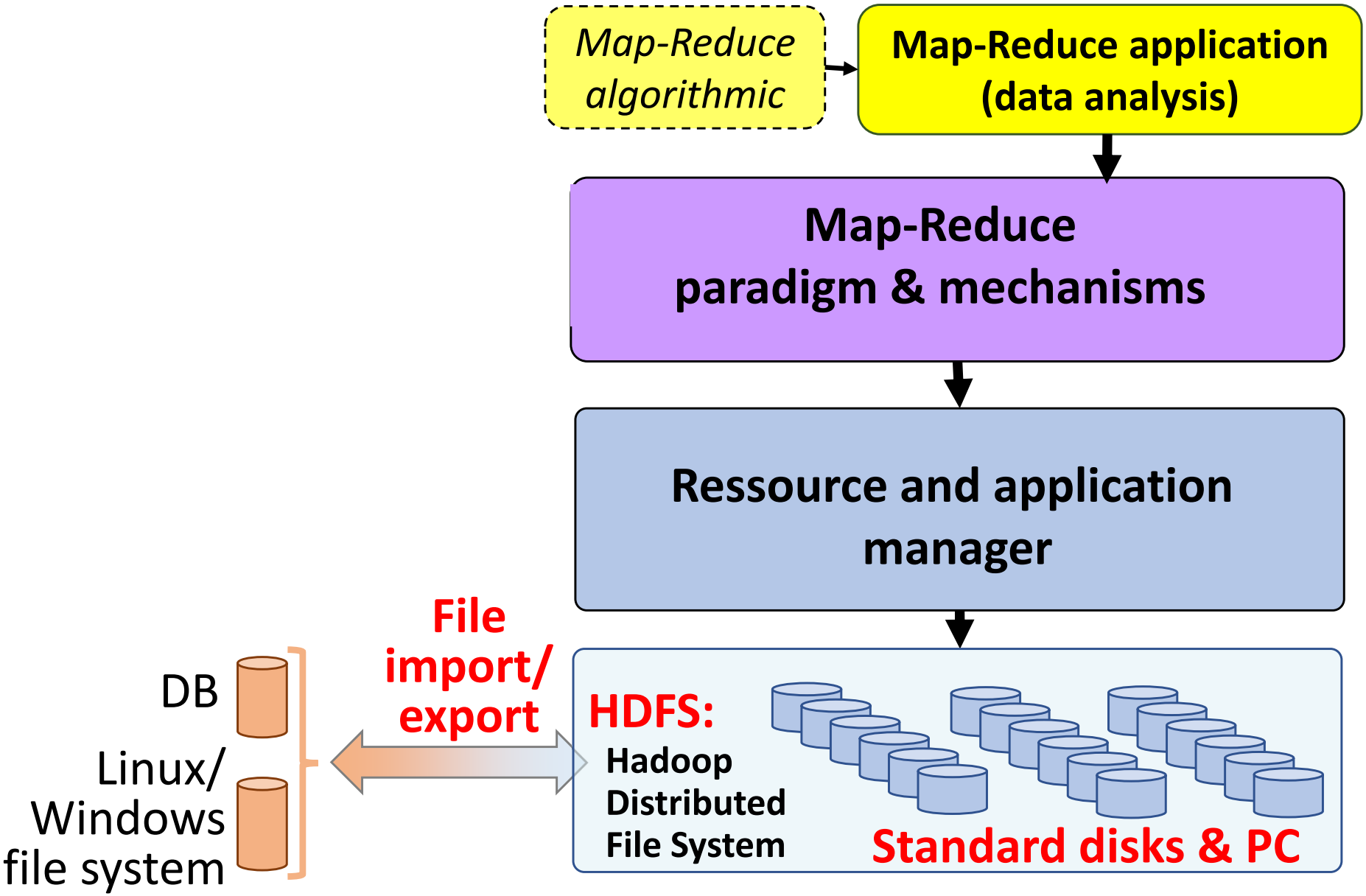
Stéphane Vialle **&** **Gianluca Quercini**

POLYTECH
PARIS-SACLAY

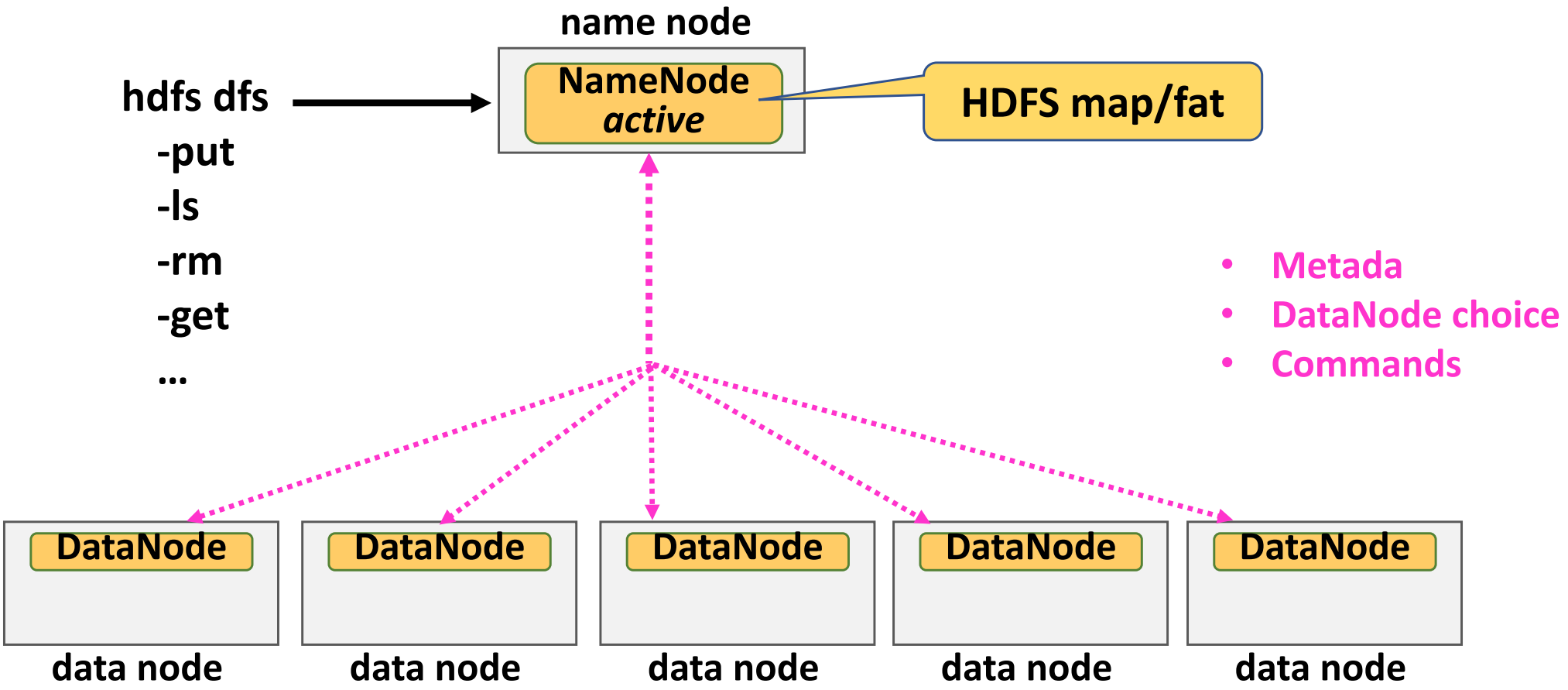
Using Spark cluster of CentraleSupelec DCE

- **HDFS principles & commands**
  → HDFS experiment
- **Spark principles & commands**
  → Spark experiment

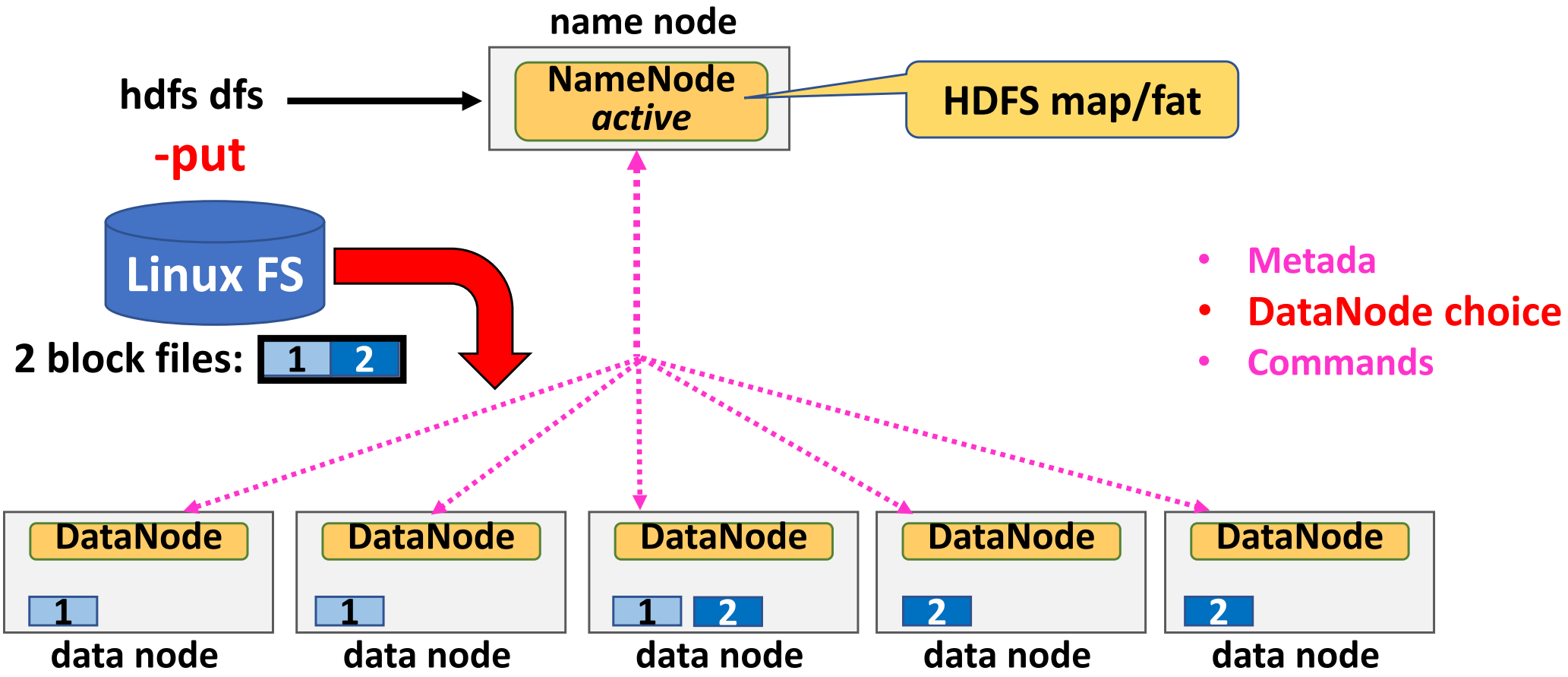
Hadoop software architecture



HDFS principles

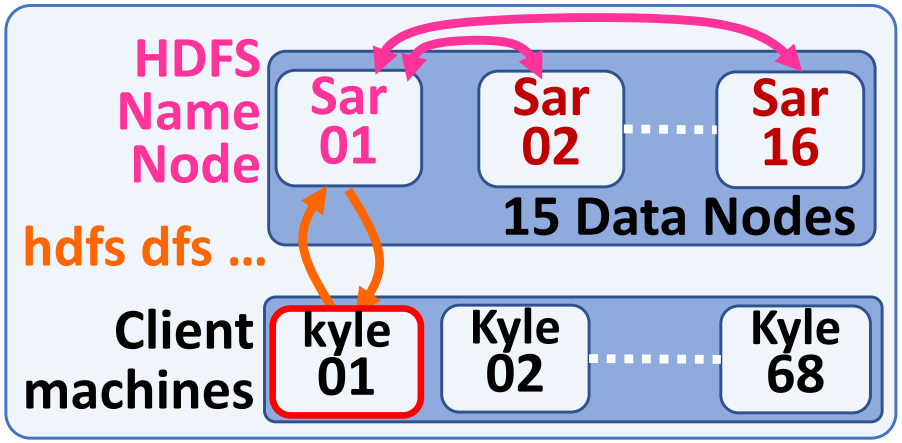


HDFS principles



- Files ares splitted into data blocks (64 or 128 Mbytes)
- Each block is replicated (default: x3) to ensure fault tolerance
- The NameNode chooses the Data Nodes storing blocks and replicas

HDFS commands on DCE



HDFS NameNode service: **sar01:9000**

or: sar**17**:9000

/data	*Read only*
/ppsbd1	
/ppsbd1_1/	*R/W for ppsbd1_1*
...	
/ppsbd1_20/	*R/W for ppsbd1_20*

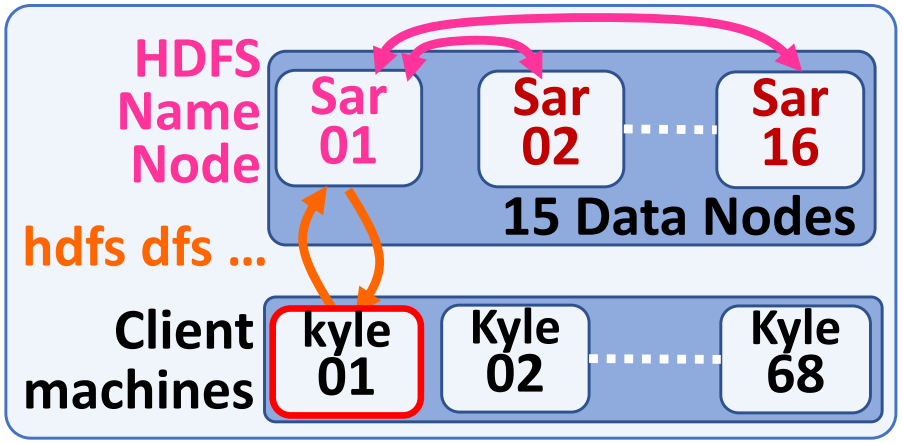
On a cluster node *(client machine)*:

hdfs dfs -ls -h hdfs://sar01:9000/data

hadoop fs -ls -h hdfs://sar01:9000/data (alternative syntax)

```
Found 2 items
drwxrwxr-x -   cpu_vialle   cpu_prof   0   2019-10-18   11:23
                                                hdfs://sar01:9000/data/sales
-rw-r--r--    3   cpu_vialle   cpu_prof   568.2 K   2019-10-04   13:58
      nb of replicas                          size          hdfs://sar01:9000/data/sherlock.txt
```

HDFS commands on DCE



HDFS NameNode service: **sar01:9000**

or: sar**17**:9000

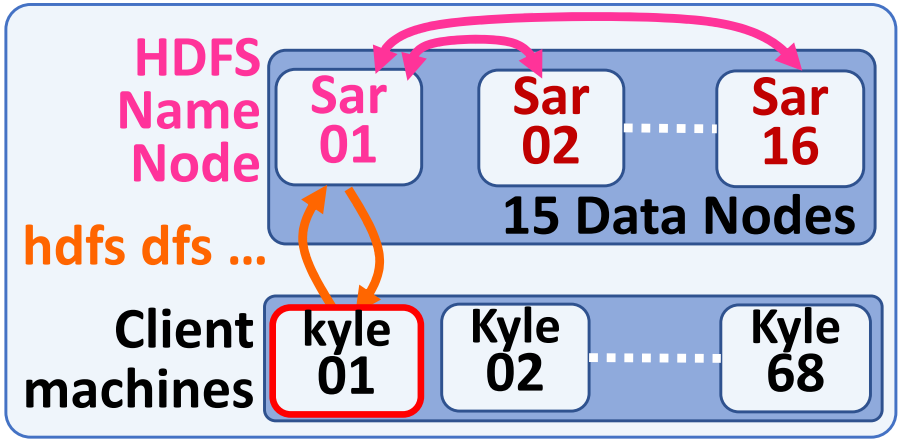
/data	*Read only*
/ppsbd1	
/ppsbd1_1/	*R/W for ppsbd1_1*
...	
/ppsbd1_20/	*R/W for ppsbd1_20*

On a cluster node *(client machine)*:

```
hdfs dfs -ls -h hdfs://sar01:9000/data/sales
```

```
Found 7 items
-rw-r--r--   3 cpu_vialle  cpu_prof    257.0 M  2023-09-14 10:24
                                  hdfs://sar01:9000/data/sales/customer_100.dat
......
-rw-r--r--   3 cpu_vialle  cpu_prof     37.4 G  2023-09-14 10:34
                                  hdfs://sar01:9000/data/sales/store_sales_1_4.400.dat
-rw-r--r--   3 cpu_vialle  cpu_prof     76.3 G  2023-09-14 10:45
                                  hdfs://sar01:9000/data/sales/store_sales_1_4.800.dat
```

HDFS commands on DCE



HDFS NameNode service: **sar01:9000**

or: sar**17**:9000

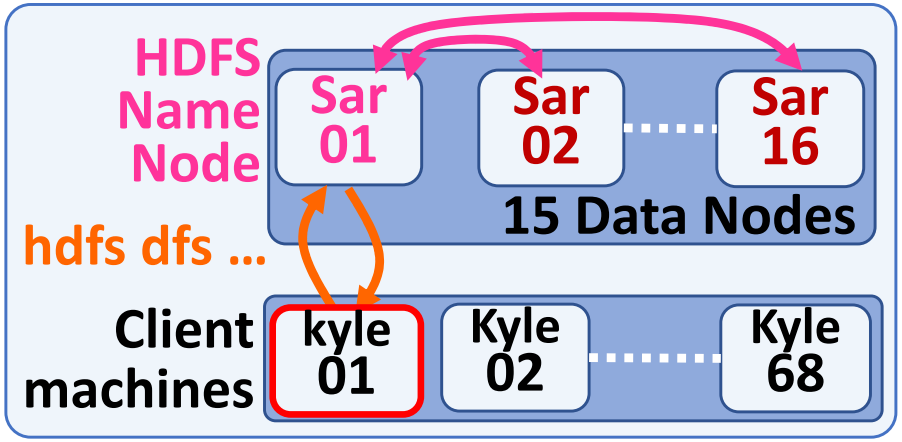
/data	*Read only*
/ppsbd1	
/ppsbd1_1/	*R/W for ppsbd1_1*
…	
/ppsbd1_20/	*R/W for ppsbd1_20*

On a cluster node *(client machine)*:

```
hdfs dfs -cat hdfs://sar01:9000/data/sherlock.txt | more
```

```
Project Gutenberg's The Adventures of Sherlock Holmes, by Arthur Conan Doyle
....
```

HDFS commands on DCE



HDFS NameNode service: **sar01:9000**

or: sar**17**:9000

/data	*Read only*
/ppsbd1	
/ppsbd1_1/	*R/W for ppsbd1_1*
…	
/ppsbd1_20/	*R/W for ppsbd1_20*

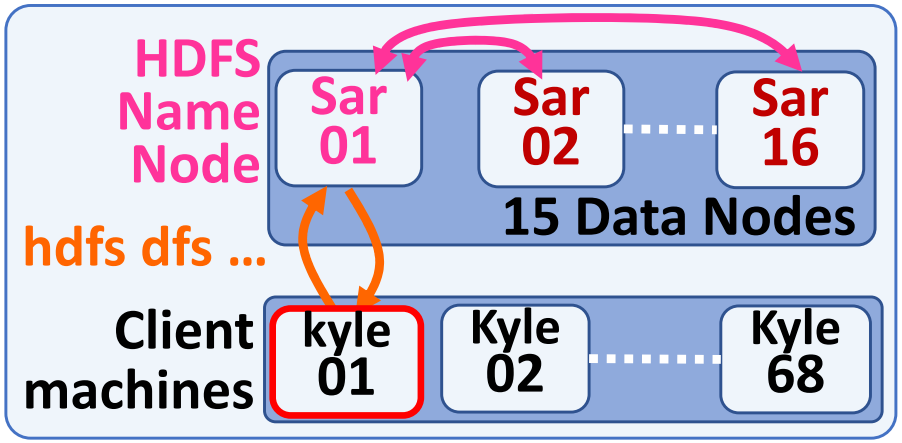
On a cluster node *(client machine)*:

```
hdfs dfs -ls -h hdfs://sar01:9000/ppsbd1/ppsbd1_1
```

empty

Use YOUR HDFS account

HDFS commands on DCE



HDFS NameNode service: **sar01:9000**

or: sar**17**:9000

/data	*Read only*
/ppsbd1	
/ppsbd1_1/	*R/W for ppsbd1_1*
...	
/ppsbd1_20/	*R/W for ppsbd1_20*

On a cluster node *(client machine)*:

```
cp ~cpu_vialle/DCE-Spark/RFC793-TCP.txt  .
hdfs dfs -put RFC793-TCP.txt hdfs://sar01:9000/ppsbd1/ppsbd1_1/
hdfs dfs -ls -h hdfs://sar01:9000/ppsbd1/ppsbd1_1/
```

```
Found 1 items
-rw-r--r--   3 ppsbd1_1 ppsbd1    173.8 K 2019-10-21 01:51
                                  hdfs://sar01:9000/ppsbd1/ppsbd1_1/RFC793-TCP.txt
```

Use YOUR HDFS account

POLYTECH PARIS-SACLAY

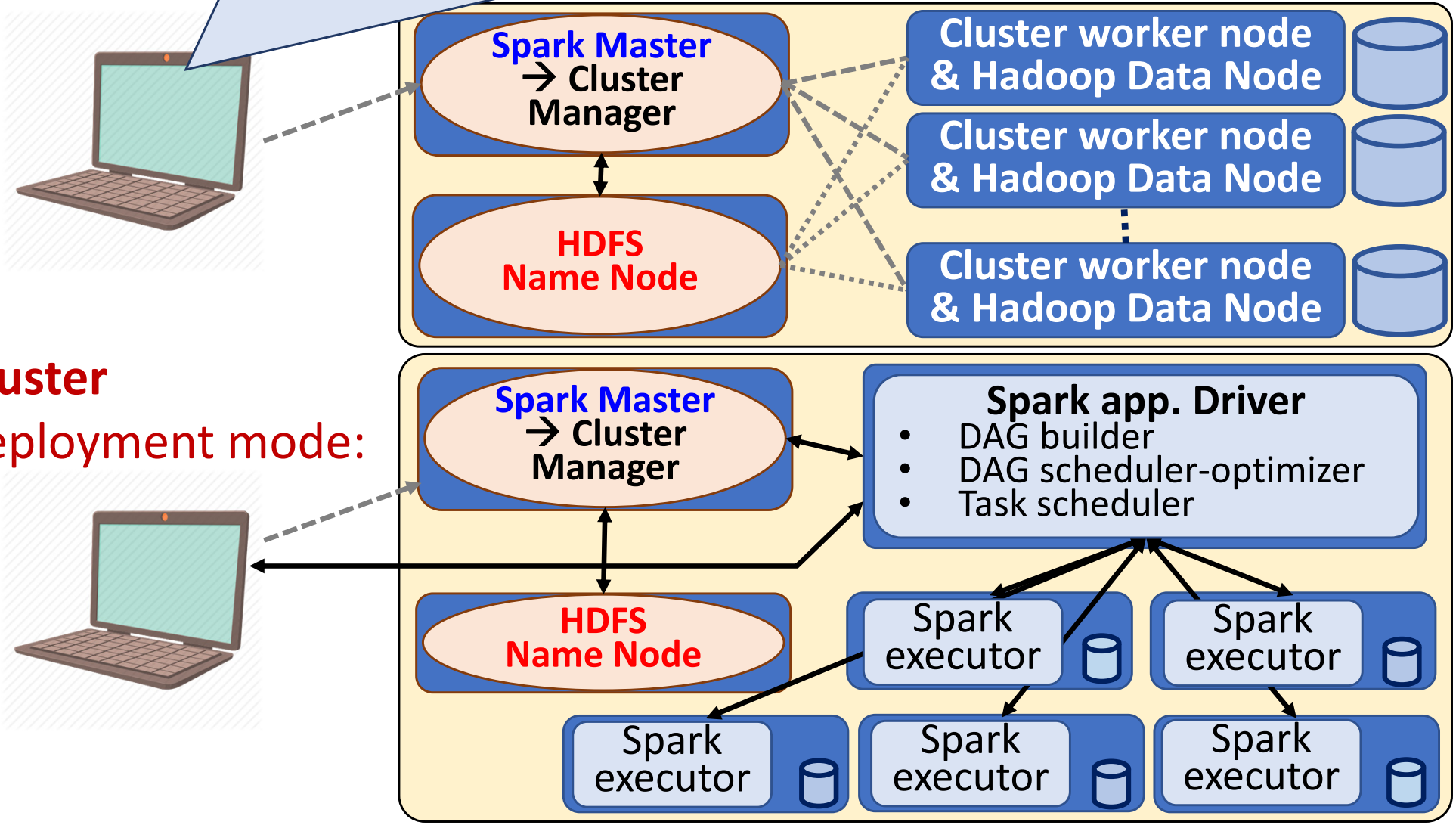
Using Spark cluster of CentraleSupelec DCE

- **HDFS principles & commands**
  → HDFS experiment
- **Spark principles & commands**
  → Spark experiment

Spark deployment on top of HDFS

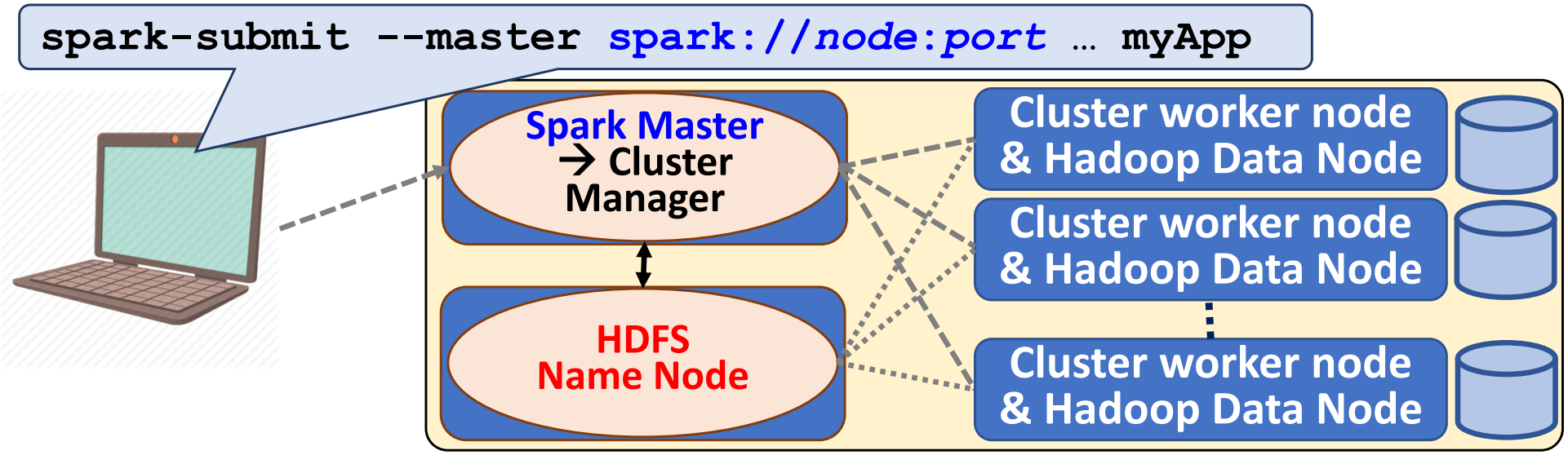
Spark Master as cluster manager: standalone mode

```
spark-submit --master spark://node:port … myApp
```



Spark deployment on top of HDFS

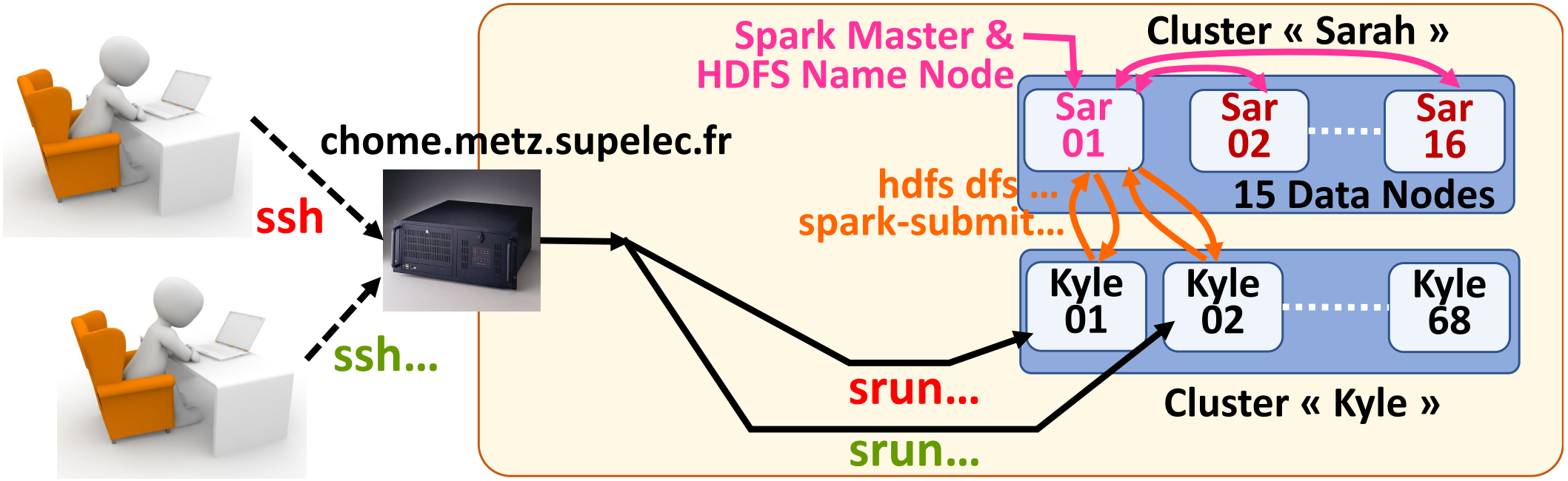
Spark Master as cluster manager: standalone mode



Strenght and weakness of standalone mode:

- Nothing more to install (included in Spark)
- Easy to configure
- Can run different jobs concurrently
- Can not share the cluster with non-Spark applications
- Limited scheduling mechanism (unique queue)
- Can not target data nodes hosting input data to launch Executors

Spark-HDFS configuration on DCE



Default configuration on DCE:

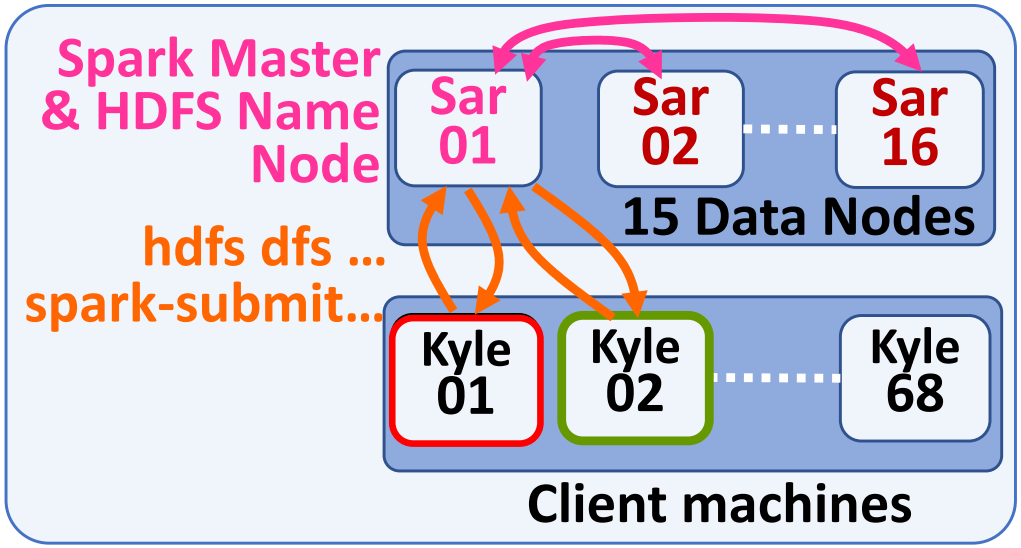
- (only) 15 cores per Spark application
- 15 Data nodes

→ One Spark application: 1Executor/Data Node & 1 core/Executor

- 16 (logical) cores per node

→ 16 Spark applications can run concurrently on the Spark cluster

Spark commands on DCE



HDFS NameNode service: **sar01:9000**

or: sar**17**:9000

/data	*Read only*
/ppsbd1	
/ppsbd1_1/	*R/W for ppsbd1_1*
…	
/ppsbd1_20/	*R/W for ppsbd1_20*

Spark Master service: **sar01:7077**

or: sar**17**:7077

On a cluster node *(client machine)*:

```
cp ~cpu_vialle/DCE-Spark/template_wc.py  ./wc.py
```

→ **Edit** and update the Python code

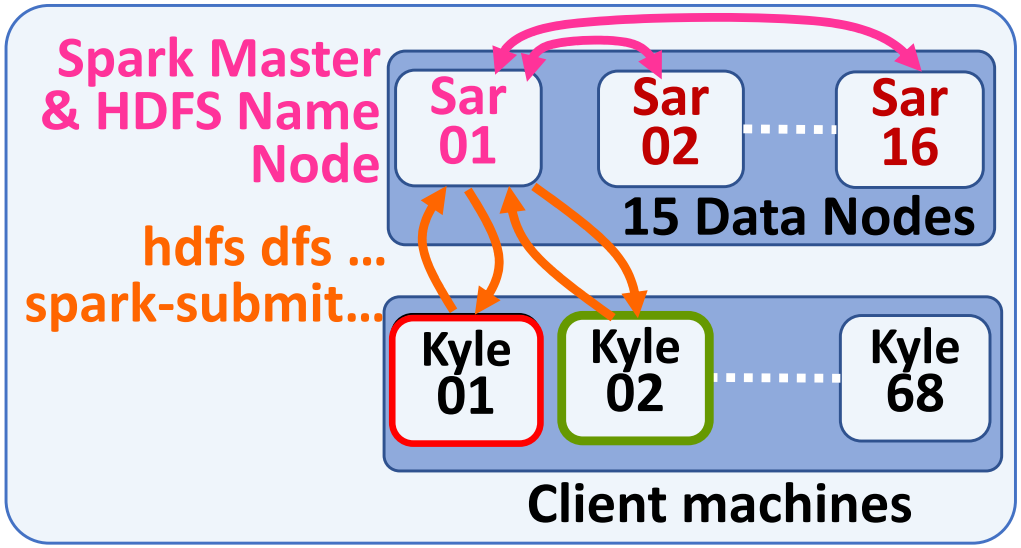
```
spark-submit  --master  spark://sar01:7077  wc.py
```

Use YOUR HDFS account

```
hdfs dfs -ls -h hdfs://sar01:9000/ppsbd1/ppsbd1_1/sherlock.out
```

```
hdfs dfs -cat hdfs://sar01:9000/ppsbd1/ppsbd1_1/sherlock.out/*
```

Spark commands on DCE



HDFS NameNode service: **sar01:9000**

or: sar**17**:9000

/data	*Read only*
/ppsbd1	
/ppsbd1_1/	*R/W for ppsbd1_1*
…	
/ppsbd1_20/	*R/W for ppsbd1_20*

Spark Master service: **sar01:7077**

or: sar**17**:7077

On a cluster node *(client machine)*:

Re-execute your Spark application

```
spark-submit --master spark://sar01:7077 wc.py
```

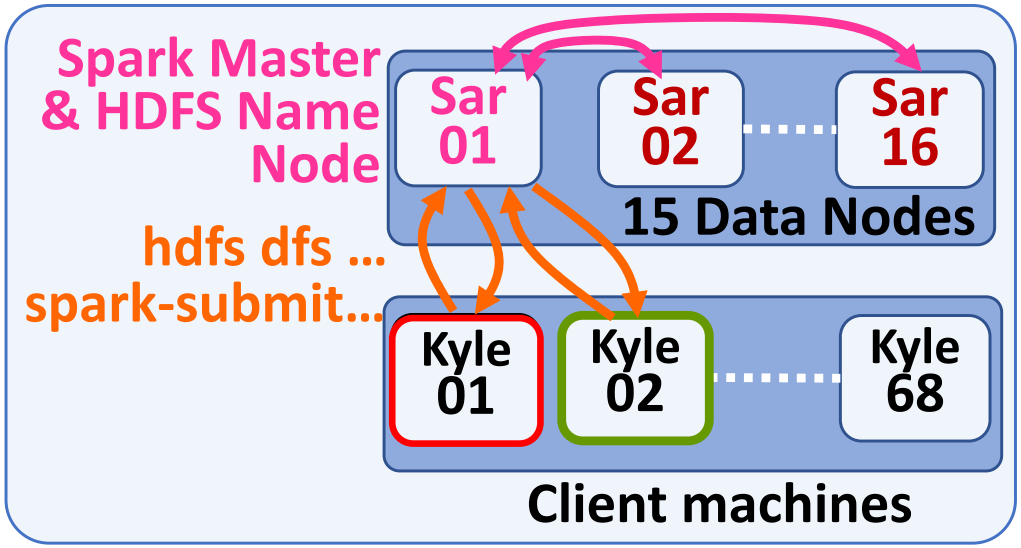
→ **ERROR** : output file already exists (and cannot overwrite)

→ Remove the output file which is a directory…

Use YOUR HDFS account

```
hdfs dfs -rm -r hdfs://sar01:9000/ppsbd1/ppsbd1_1/sherlock.out
```

Spark commands on DCE



HDFS NameNode service: **sar01:9000**

or: sar**17**:9000

/data	*Read only*
/ppsbd1	
/ppsbd1_1/	*R/W for ppsbd1_1*
…	
/ppsbd1_20/	*R/W for ppsbd1_20*

Spark Master service: **sar01:7077**

or: sar**17**:7077

Remarks (for this lab):

- Use the « **template-xx.py** » files to develop your code **xx.py** and execute your code with :

  spark-submit --master spark://sar01:7077 xx.py

- Write lambda with syntax: **(lambda a, b : (a[0] + b[0], a[1] + b[1]))**

  ~~(lambda (v1,n1), (v2,n2) : (v1+v2,n1+n2))~~

POLYTECH
PARIS-SACLAY

Big Data – TP1 Part 1

Using HDFS & Spark on the DCE clusters of CentraleSupelec

(Data Center for Education)

Questions ?