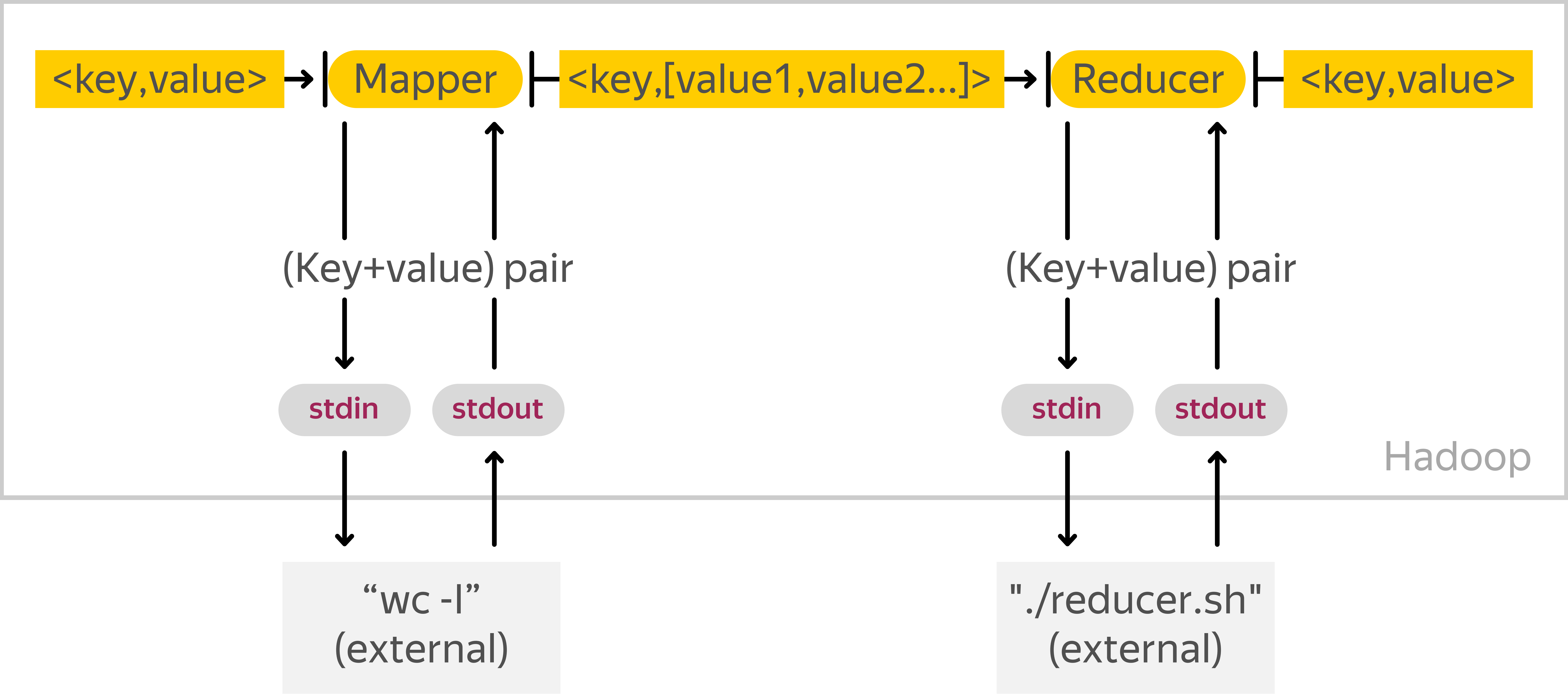
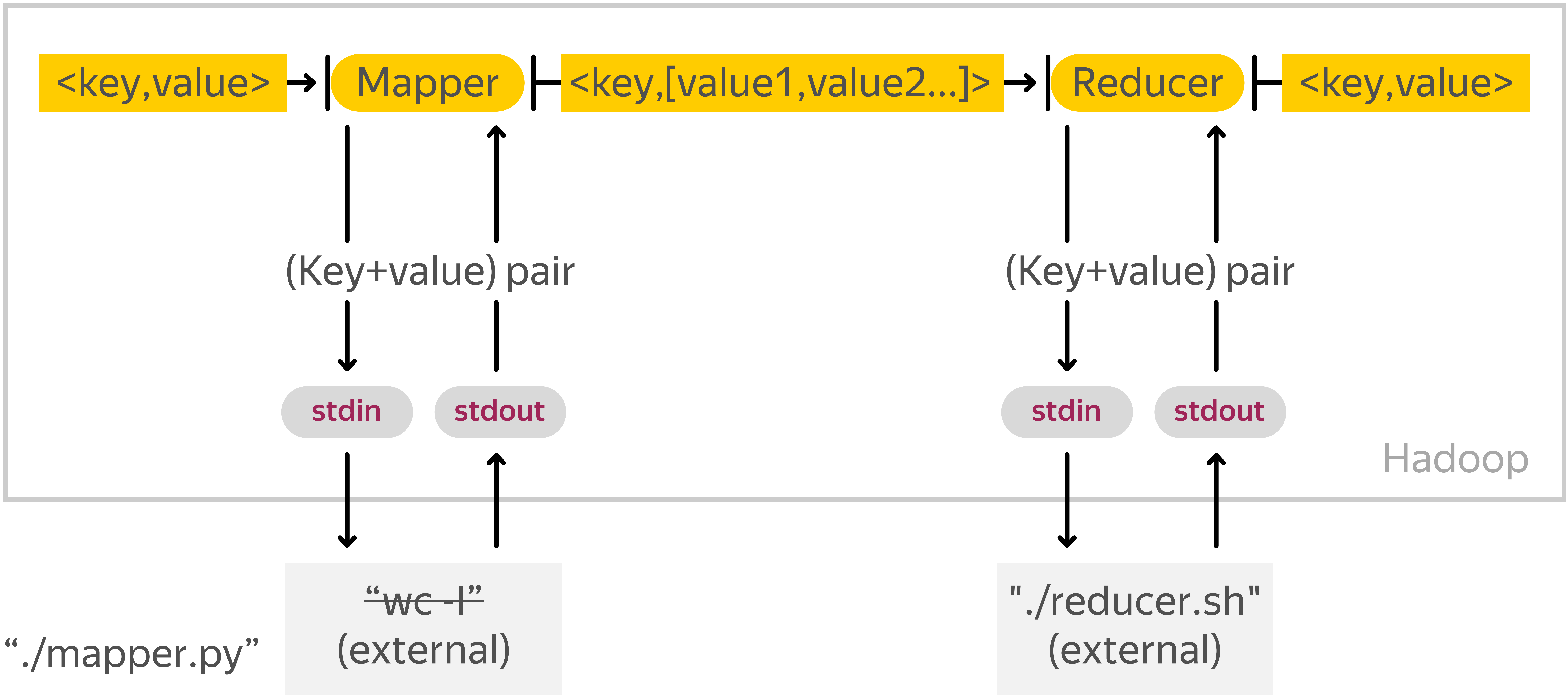
Yandex

MapReduce

Streaming in Python



<key,value>
Mapper
<key,[value1,value2...]>
Reducer
<key,value>
(Key+value) pair
(Key+value) pair
stdin
stdout
stdin
stdout
Hadoop
“wc -l”
(external)
"./reducer.sh"
(external)



<key,value> → Mapper <key,[value1,value2...]> → Reducer <key,value>

(Key+value) pair

stdin stdout

(Key+value) pair

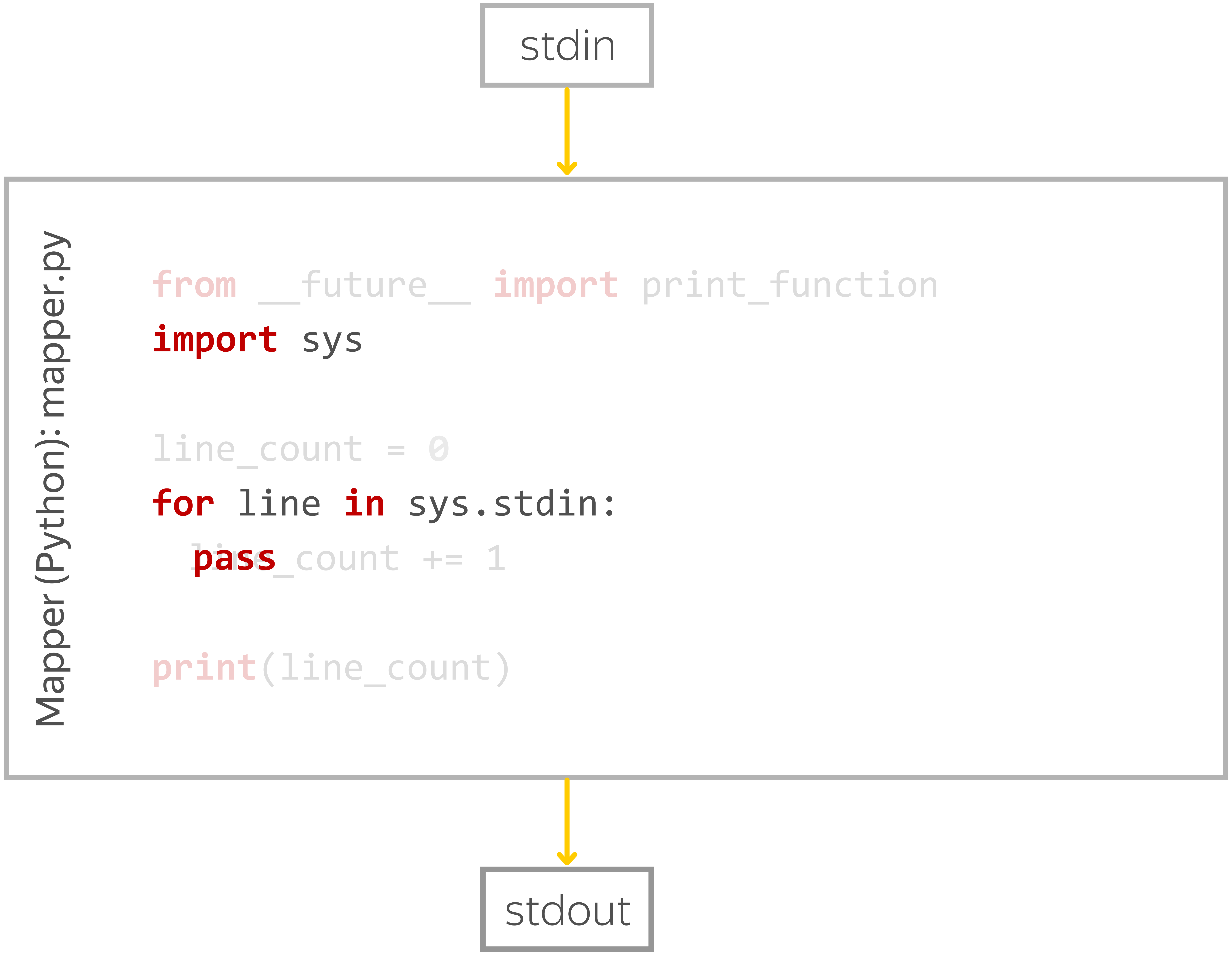
stdin stdout

Hadoop

“./mapper.py”

~~“wc -l”~~
(external)

"./reducer.sh"
(external)



stdin

Mapper (Python): mapper.py

```
from __future__ import print_function
import sys

line_count = 0
for line in sys.stdin:
  pass
  line_count += 1

print(line_count)
```

stdout

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
yarn jar $HADOOP_STREAMING_JAR \
             -files mapper.py, reducer.sh \
             -mapper 'python mapper.py' \
             -reducer './reducer.sh' \
             -numReduceTasks 1 \
             -input /data/wiki/en_articles \
             -output wc_mr_with_reducer
```

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
yarn jar $HADOOP_STREAMING_JAR \
             -files mapper.py, reducer.sh \
             -mapper 'python mapper.py' \
             -reducer './reducer.sh' \
             -numReduceTasks 1 \
             -input /data/wiki/en_articles \
             -output wc_mr_with_reducer
```

The general command line syntax is

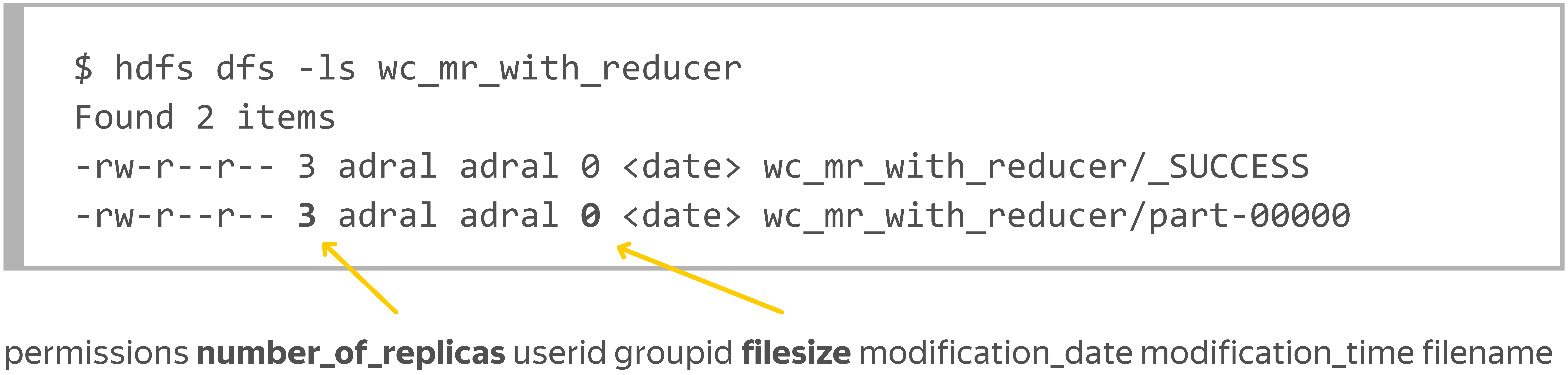
```
bin/hadoop command [genericOptions] [commandOptions]
-conf <configuration file>
-D <property=value>
-fs <local|namenode:port>
-jt <local|resourcemanager:port>
-files <comma separated list of files>
-libjars <comma separated list of jars>
-archives <comma separated list of archives>
```

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
yarn jar $HADOOP_STREAMING_JAR \
            -files mapper.py, reducer.sh \
            -mapper 'python mapper.py' \
            -reducer './reducer.sh' \
            -numReduceTasks 1 \
            -input /data/wiki/en_articles \
            -output wc_mr_with_reducer
```

```
$ hdfs dfs -ls wc_mr_with_reducer
Found 2 items
-rw-r--r-- 3 adral adral 0 <date> wc_mr_with_reducer/_SUCCESS
-rw-r--r-- 3 adral adral 0 <date> wc_mr_with_reducer/part-00000
```

? ?

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
yarn jar $HADOOP_STREAMING_JAR \
             -files mapper.py, reducer.sh \
             -mapper 'python mapper.py' \
             -reducer './reducer.sh' \
             -numReduceTasks 1 \
             -input /data/wiki/en_articles \
             -output wc_mr_with_reducer
```



```
$ hdfs dfs -ls wc_mr_with_reducer
Found 2 items
-rw-r--r-- 3 adral adral 0 <date> wc_mr_with_reducer/_SUCCESS
-rw-r--r-- 3 adral adral 0 <date> wc_mr_with_reducer/part-00000
```

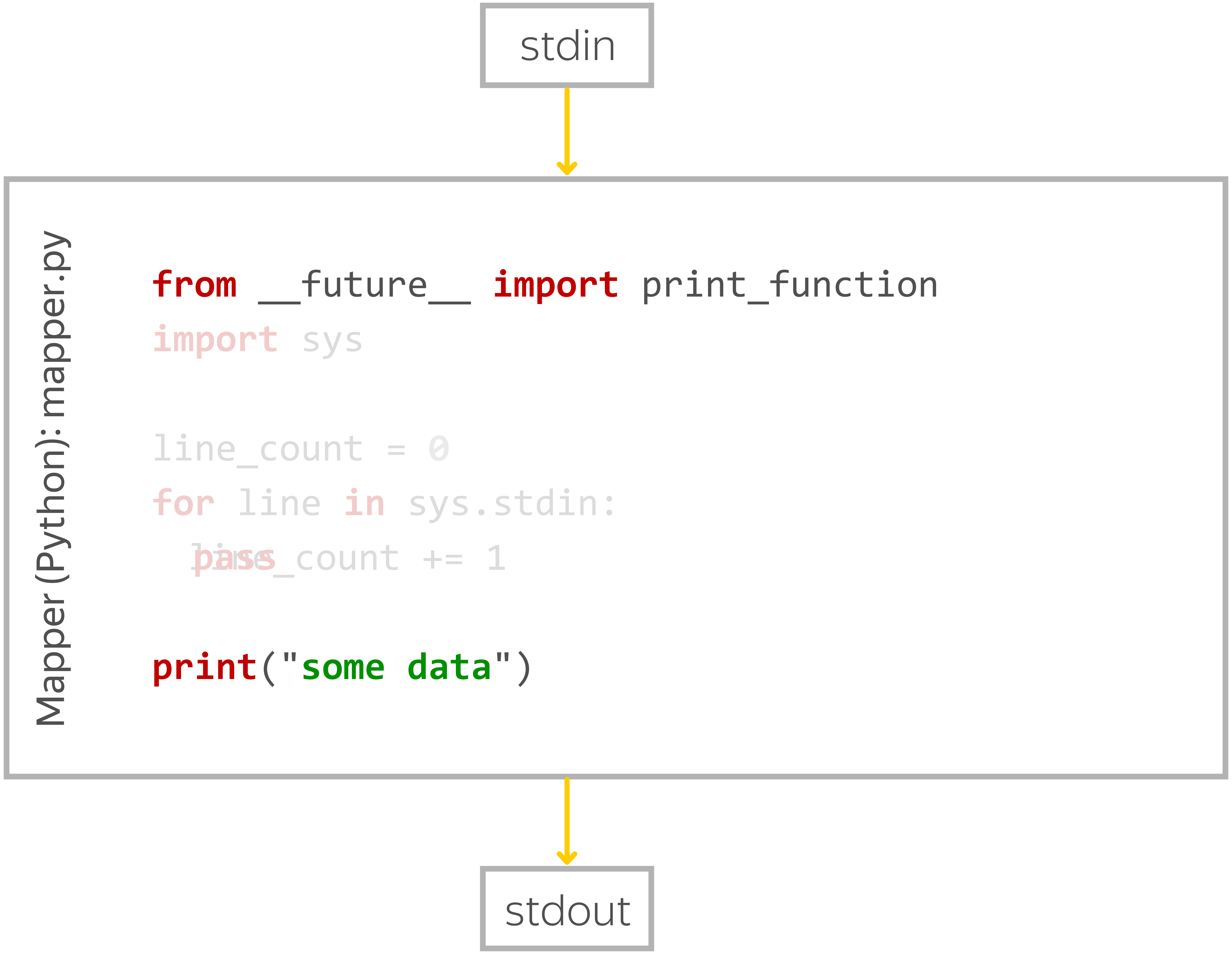
permissions **number_of_replicas** userid groupid **filesize** modification_date modification_time filename

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
yarn jar $HADOOP_STREAMING_JAR \
             -files mapper.py, reducer.sh \
             -mapper 'python mapper.py' \
             -reducer './reducer.sh' \
             -numReduceTasks 1 \
             -input /data/wiki/en_articles \
             -output wc_mr_with_reducer
```

```
$ hdfs dfs -ls wc_mr_with_reducer
Found 2 items
-rw-r--r-- 3 adral adral 0 <date> wc_mr_with_reducer/_SUCCESS
-rw-r--r-- 3 adral adral 0 <date> wc_mr_with_reducer/part-00000
```

permissions **number_of_replicas** userid groupid **filesize** modification_date modification_time filename

```
$ hdfs dfs -text wc_mr_with_reducer/*
--
```



stdin

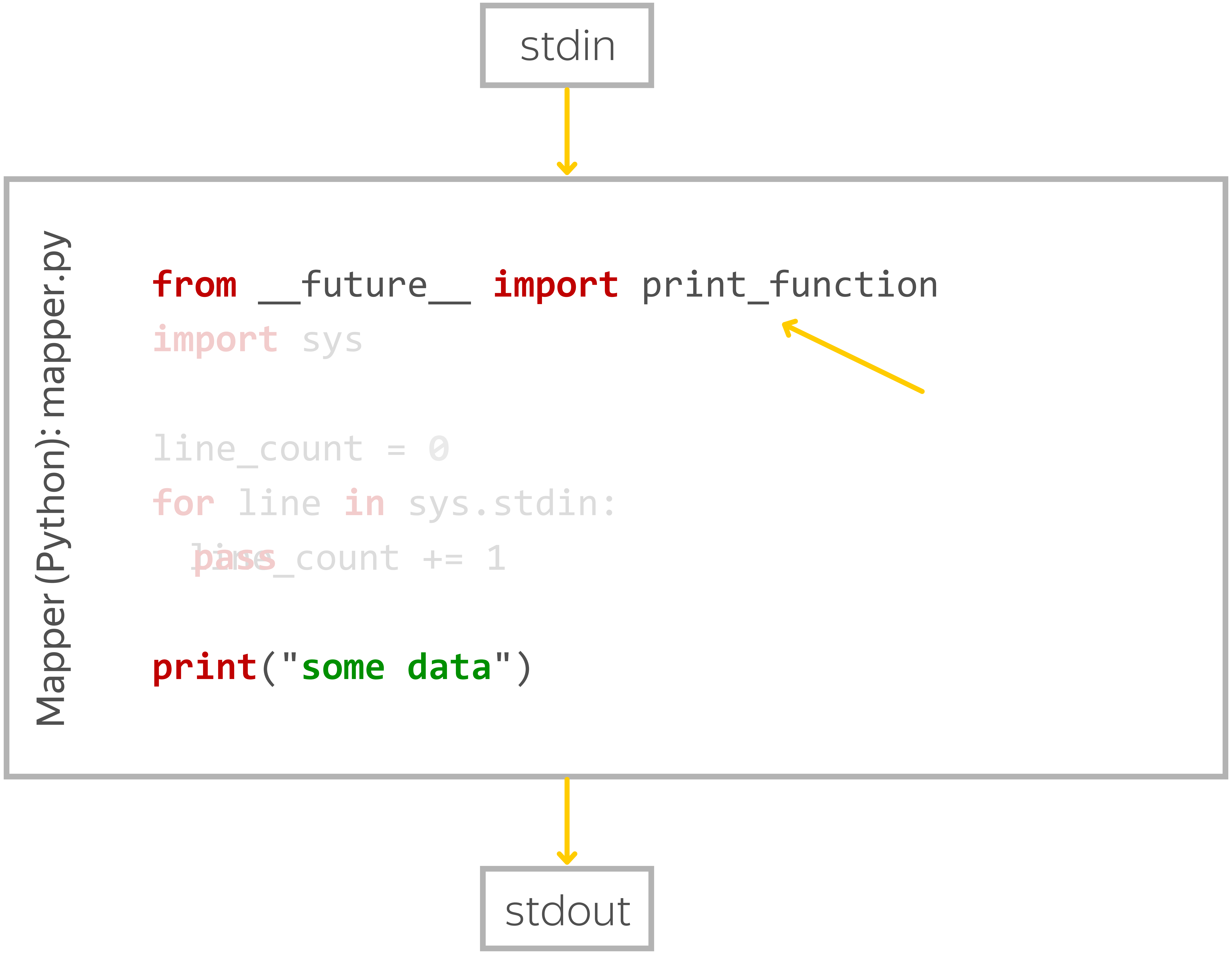
Mapper (Python): mapper.py

```python
from __future__ import print_function
import sys

line_count = 0
for line in sys.stdin:
  line_count += 1
  pass

print("some data")
```

stdout



stdin

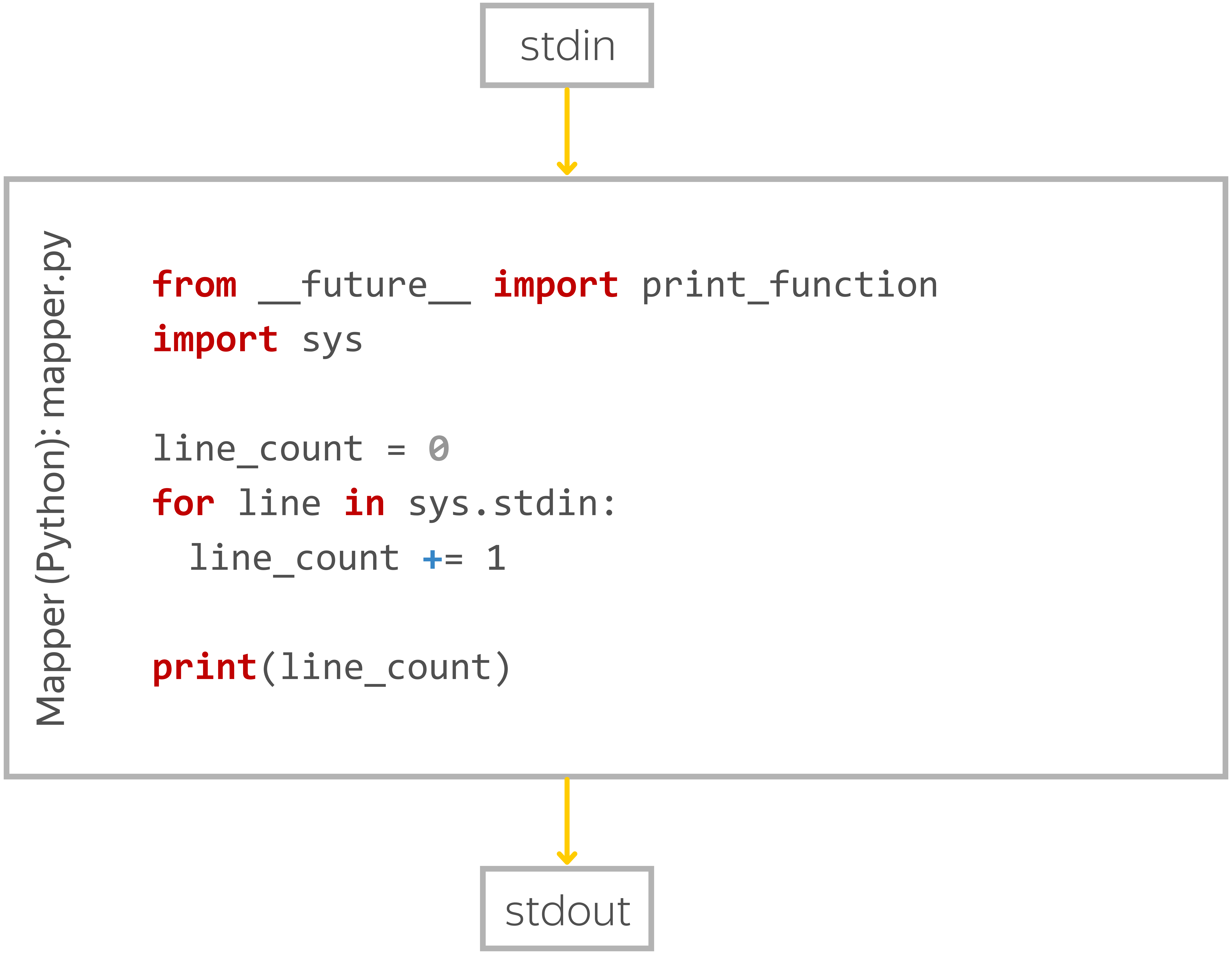
Mapper (Python): mapper.py

```
from __future__ import print_function
import sys

line_count = 0
for line in sys.stdin:
  line_count += 1
  pass

print("some data")
```

stdout



stdin

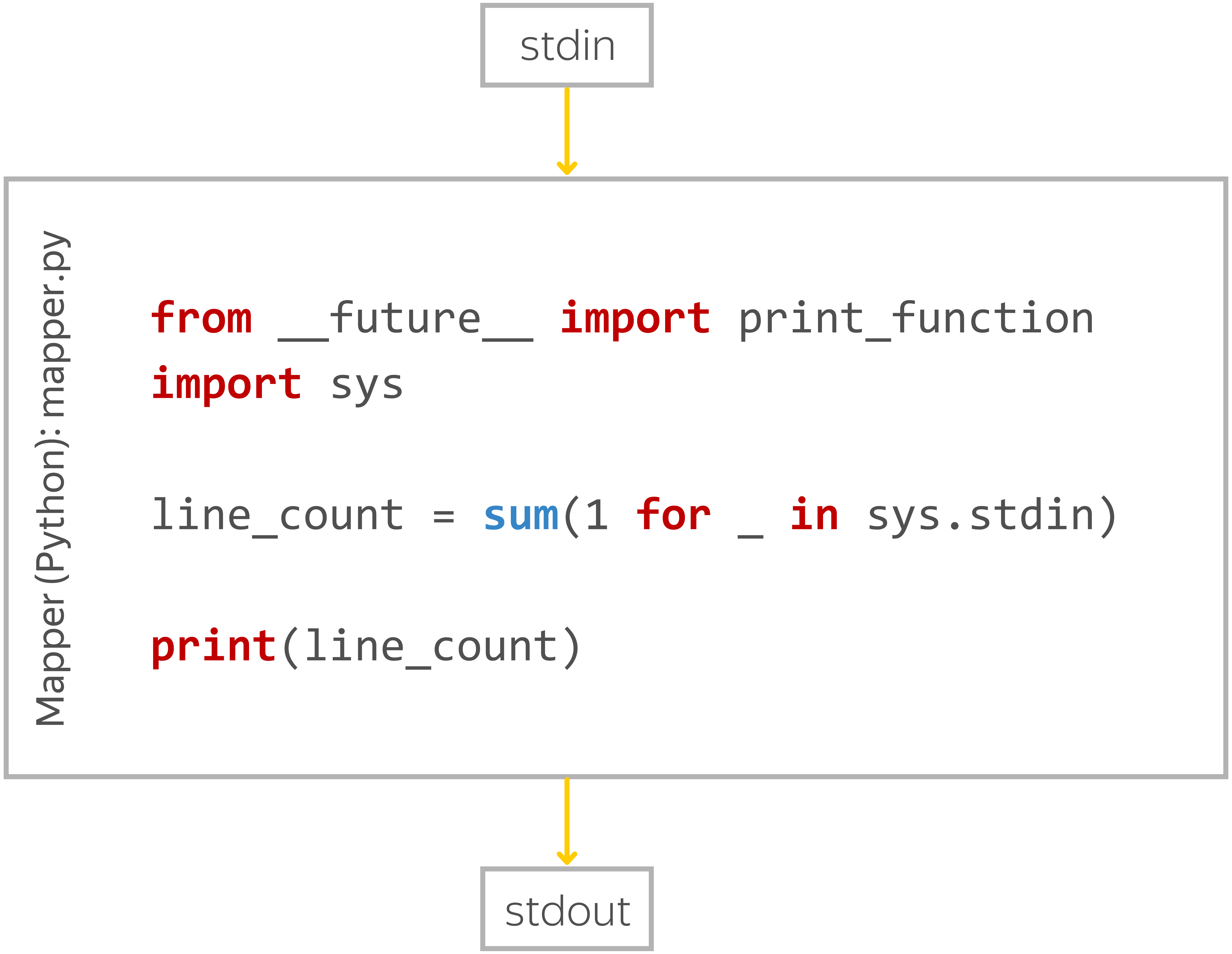
Mapper (Python): mapper.py

```python
from __future__ import print_function
import sys

line_count = 0
for line in sys.stdin:
  line_count += 1

print(line_count)
```

stdout



stdin

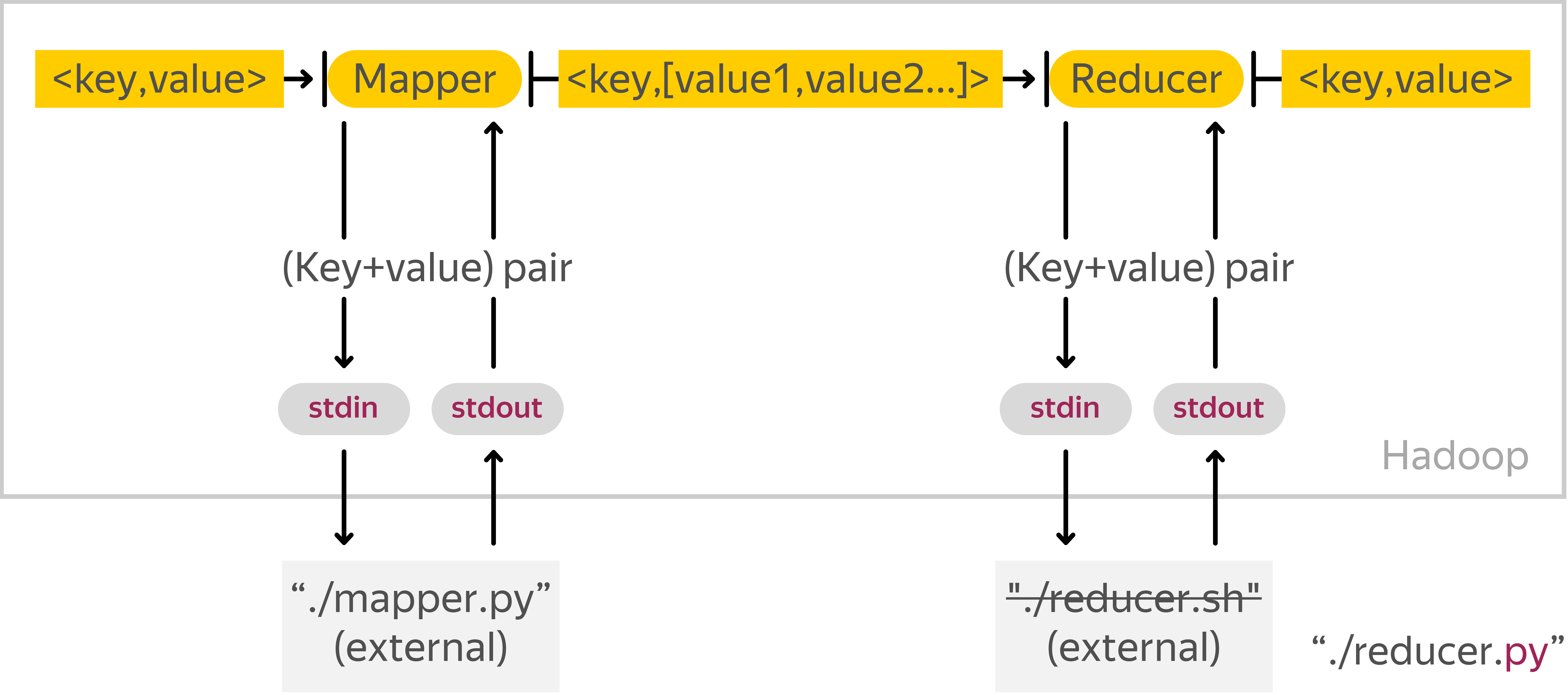
Mapper (Python): mapper.py

```python
from __future__ import print_function
import sys

line_count = sum(1 for _ in sys.stdin)

print(line_count)
```

stdout



<key,value>

Mapper

<key,[value1,value2...]>

Reducer

<key,value>

(Key+value) pair

stdin

stdout

(Key+value) pair

stdin

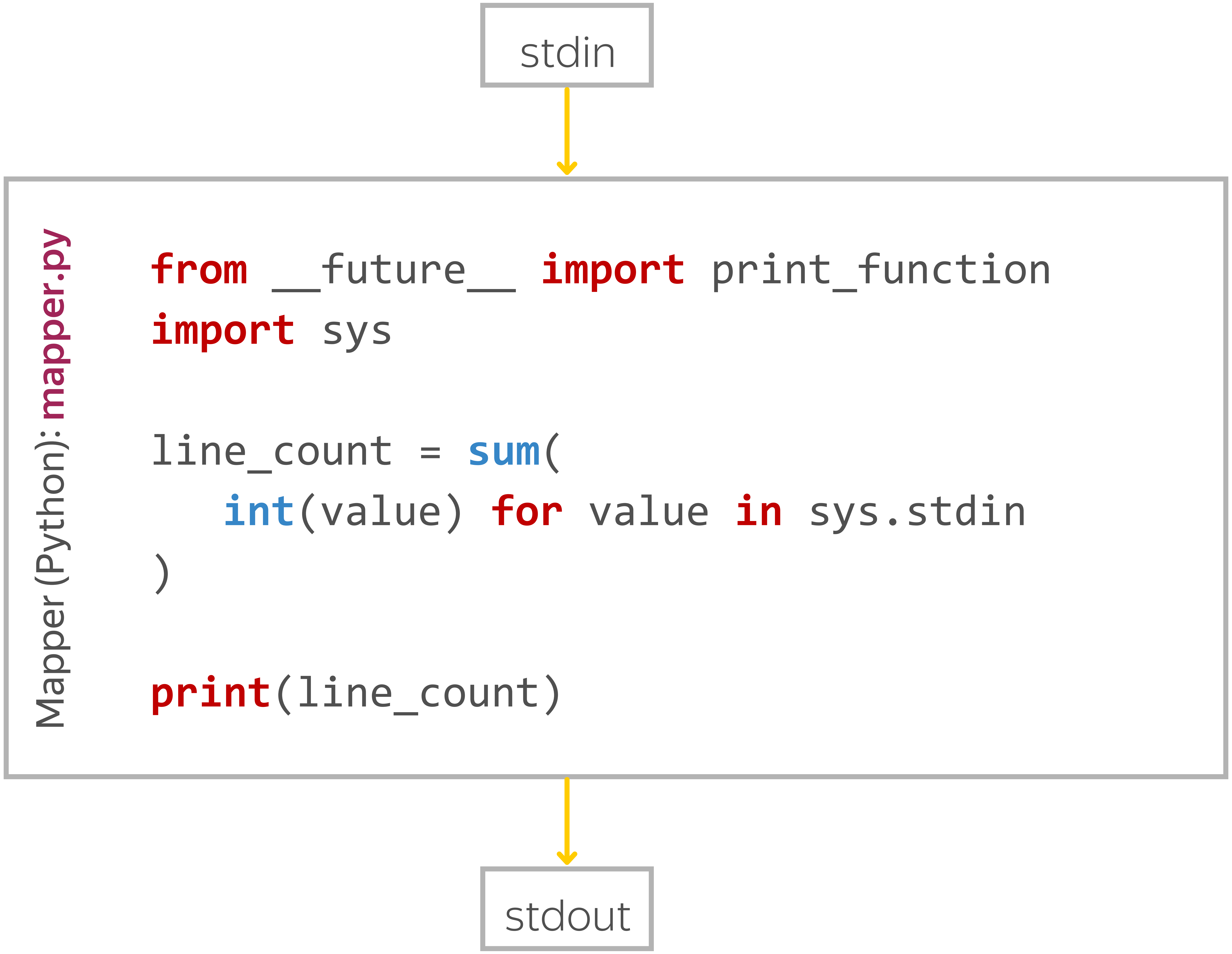
stdout

Hadoop

“./mapper.py”
(external)

~~"./reducer.sh"~~
(external)

“./reducer.py”



stdin

Mapper (Python): **mapper.py**

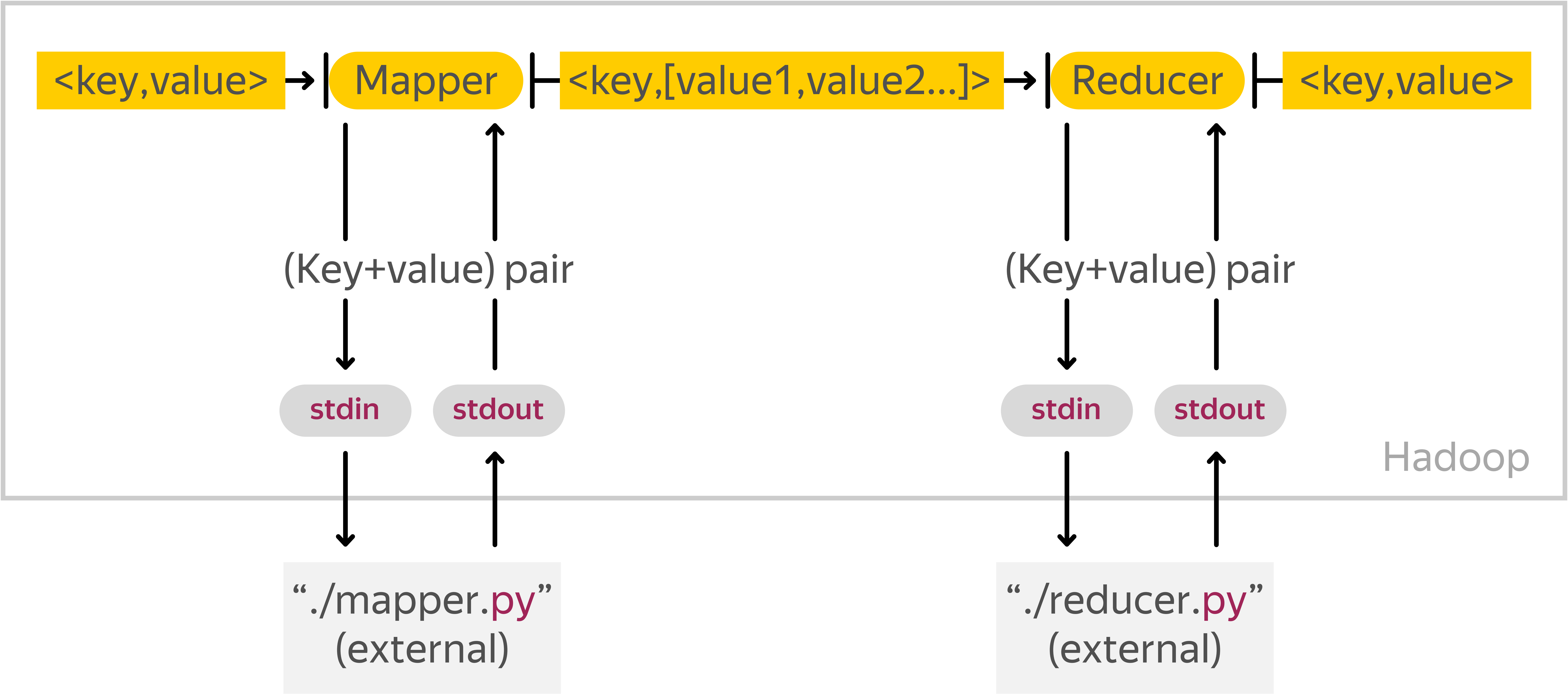
```python
from __future__ import print_function
import sys

line_count = sum(
   int(value) for value in sys.stdin
)

print(line_count)
```

stdout

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
yarn jar $HADOOP_STREAMING_JAR \
          -files mapper.py, reducer.py \
          -mapper 'python mapper.py' \
          -reducer './reducer.sh' \
          -numReduceTasks 1 \
          -input /data/wiki/en_articles \
          -output wc_mr_with_reducer
```



<key,value> → Mapper ⊢ <key,[value1,value2...]> → Reducer ⊢ <key,value>

(Key+value) pair

stdin

stdout

“./mapper.py”
(external)

(Key+value) pair

stdin

stdout

“./reducer.py”
(external)

Hadoop

BigDATAteam