

MapReduce

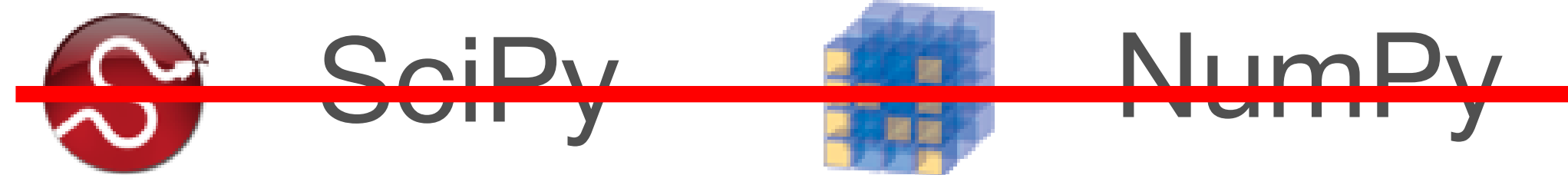
Streaming

MapReduce in Python

MapReduce in Python



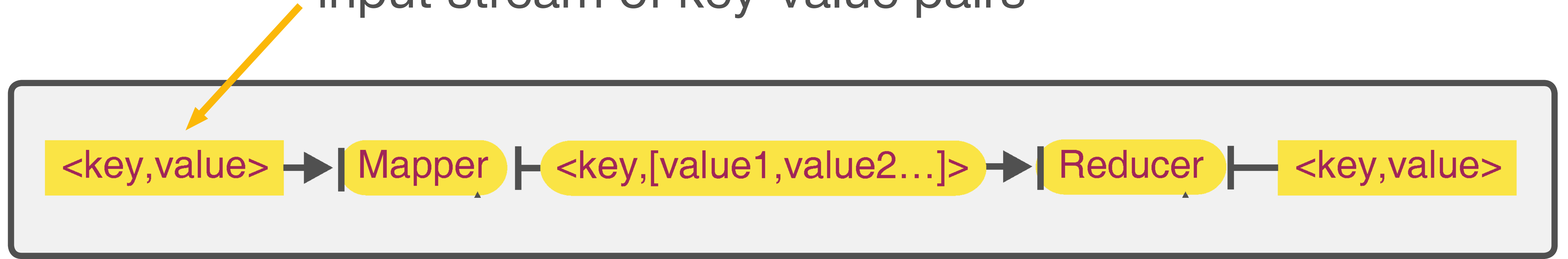
MapReduce in Python



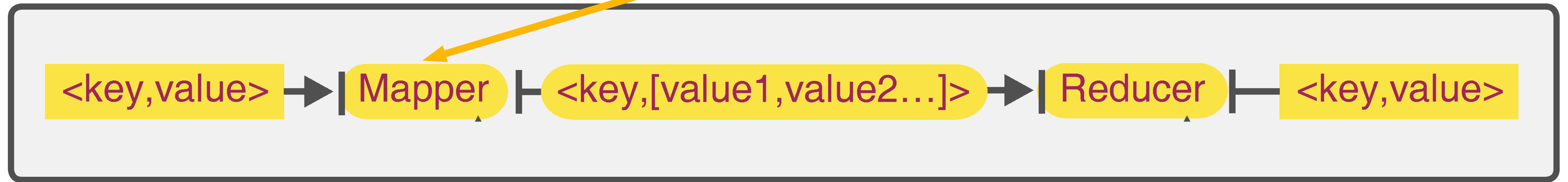
MapReduce in Python



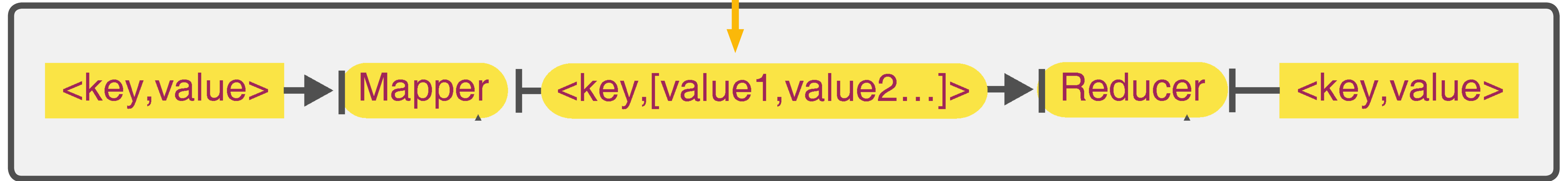
input stream of key-value pairs



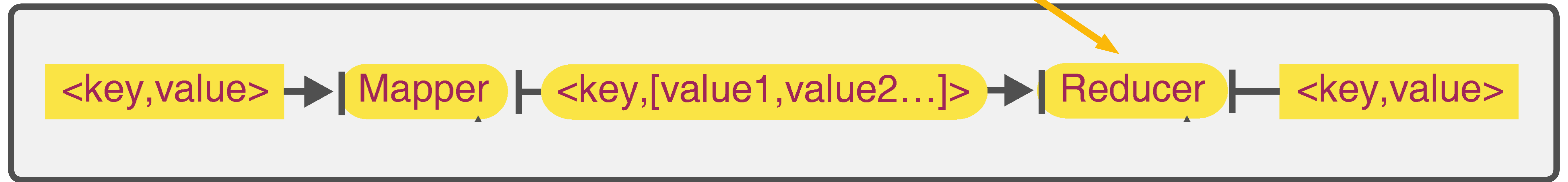
map: (k_in, v_in) --> [(k_interm, v_interm), ...]

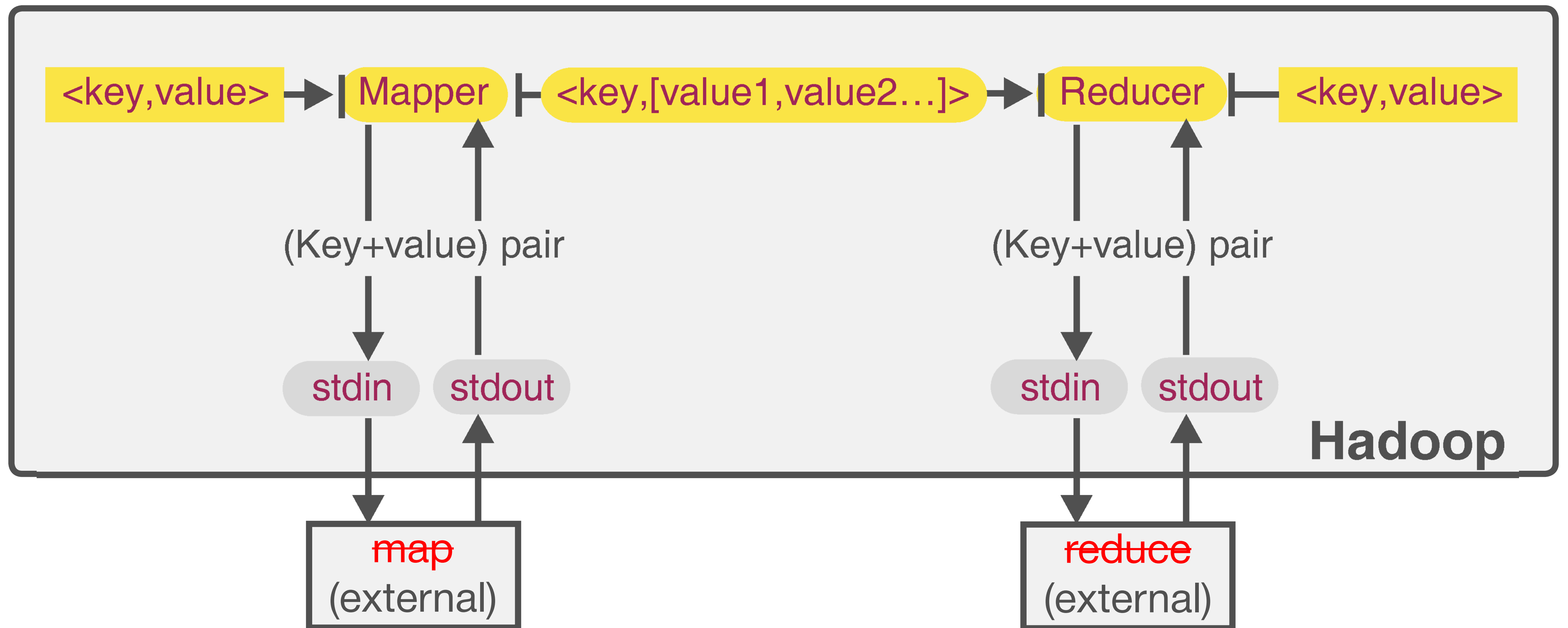


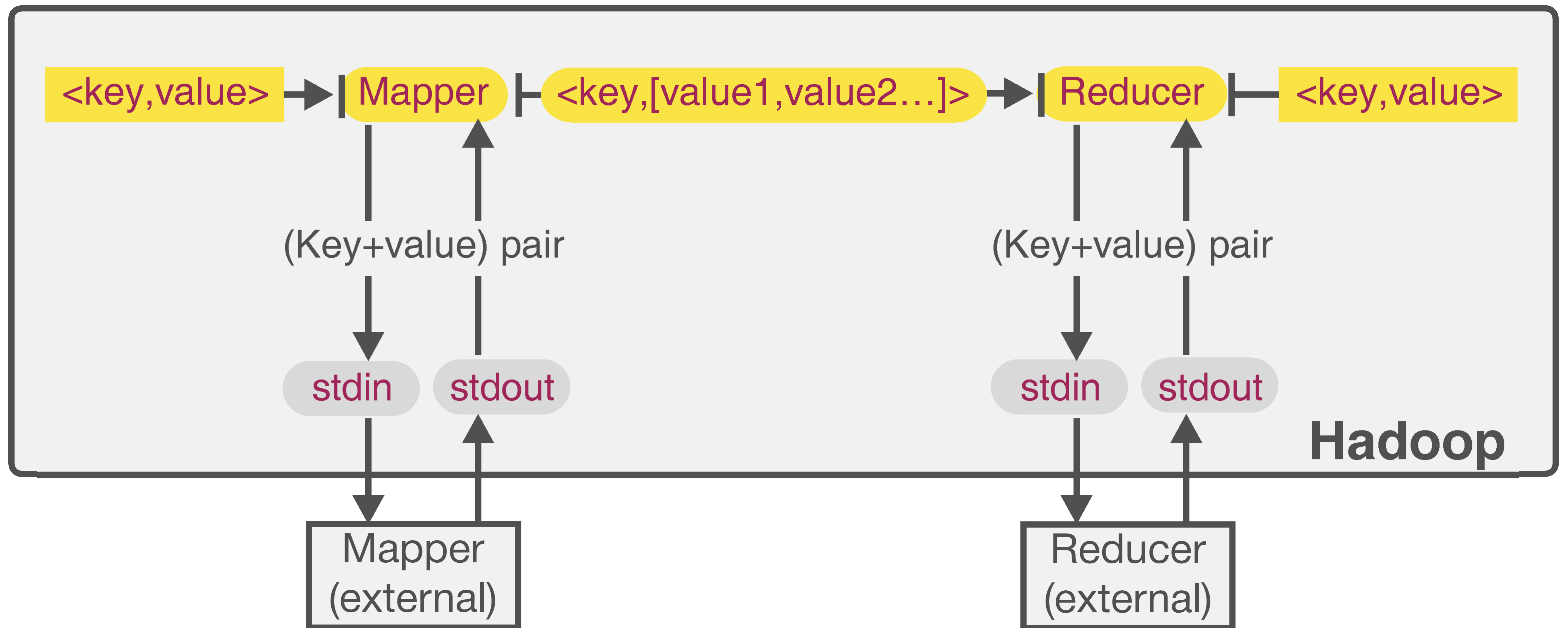
aggregate by key (Shuffle & Sort)

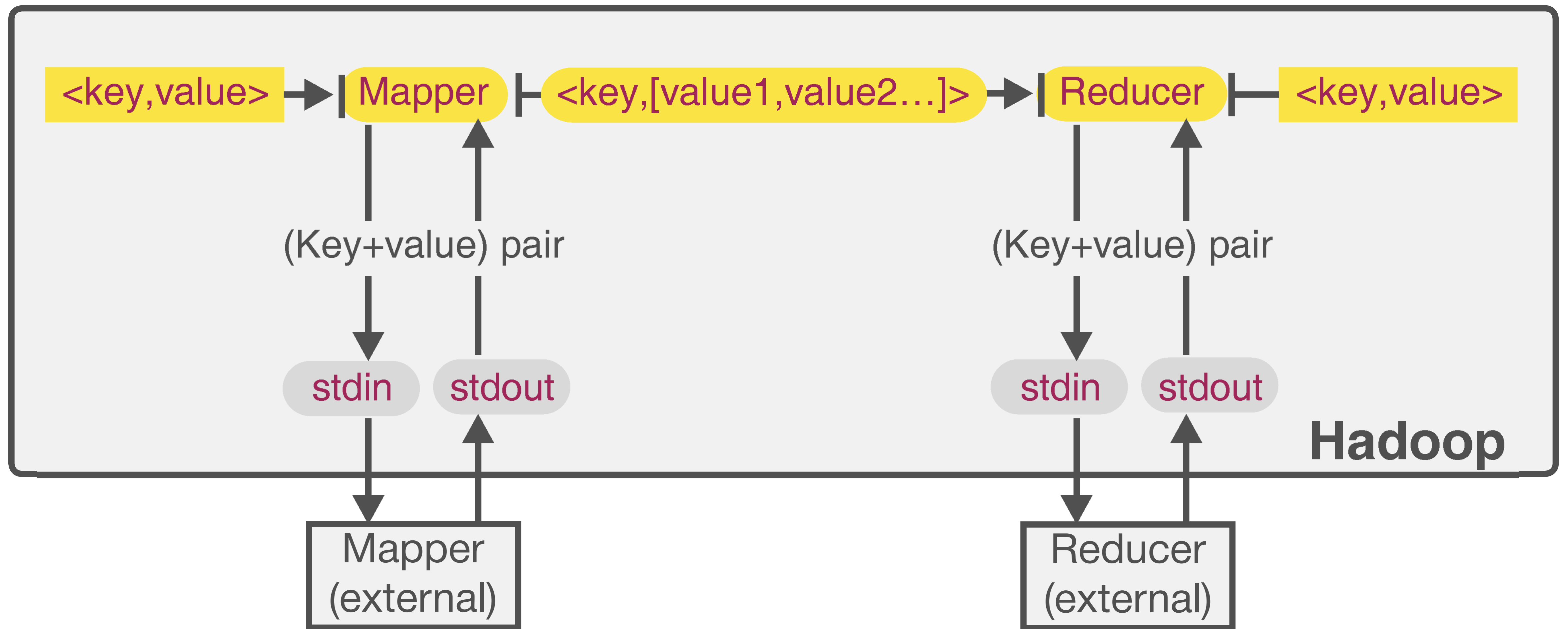


reduce: (k_interm, [(v_interm, ...)]) --> [(k_out, v_out), ...]

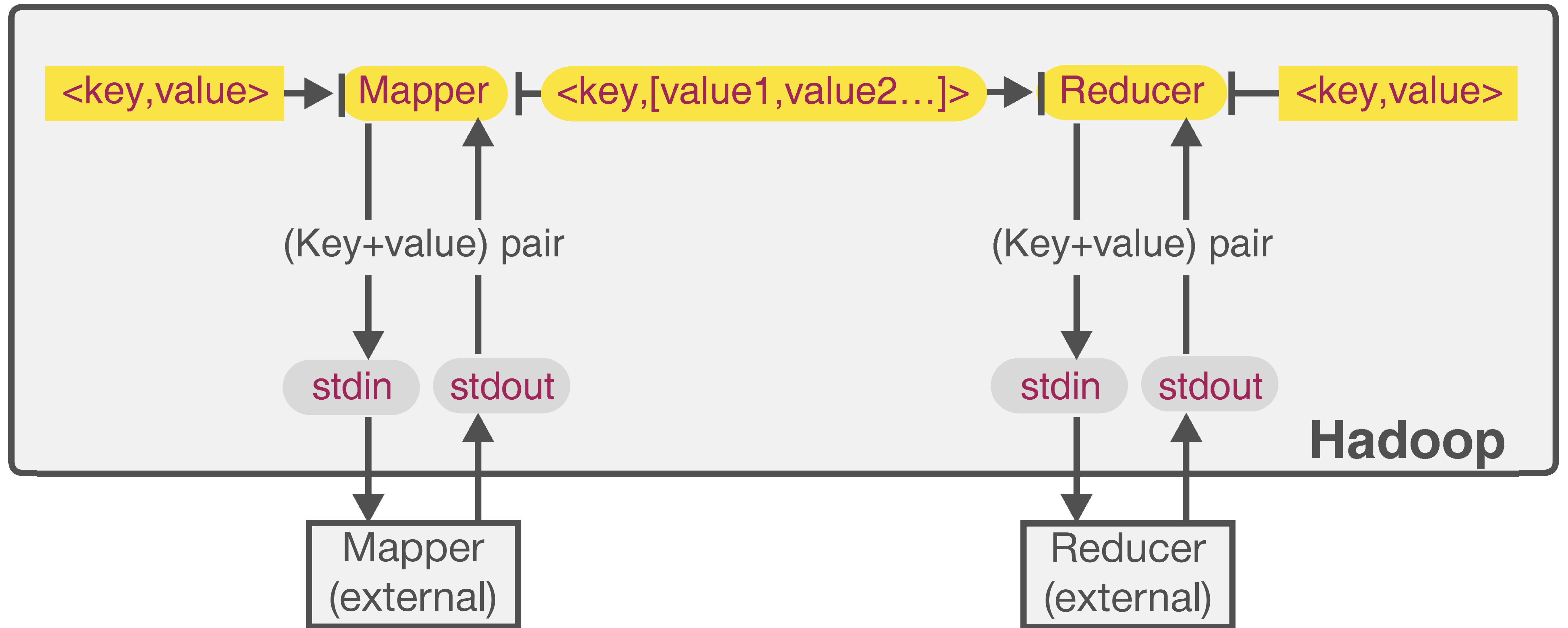




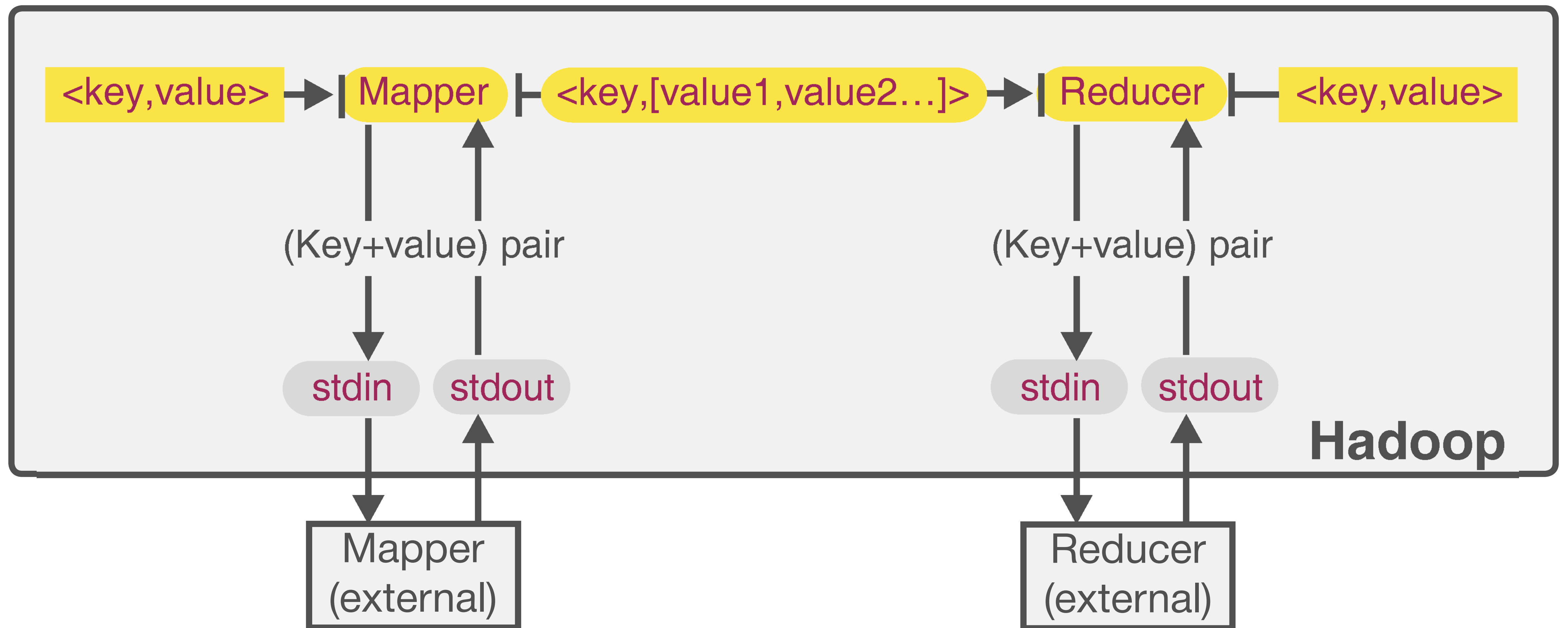




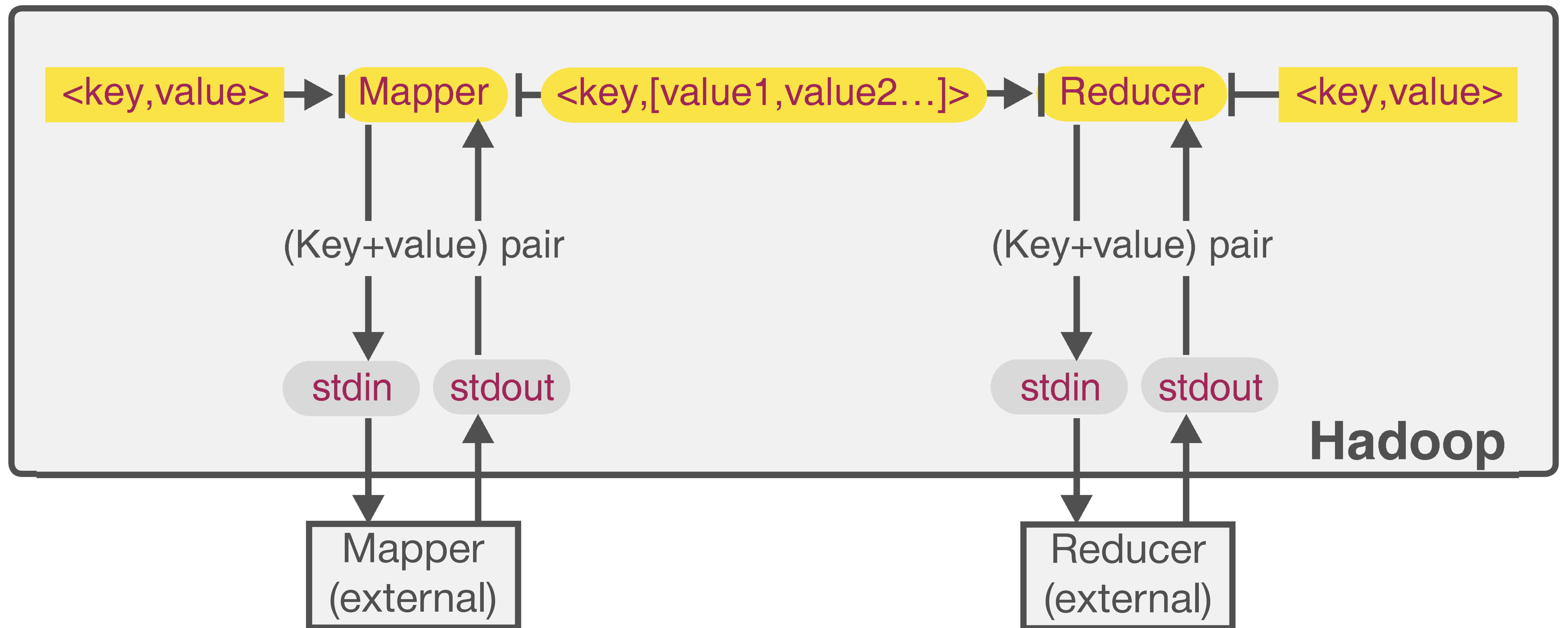
➤ define input format



- define input format
- process data

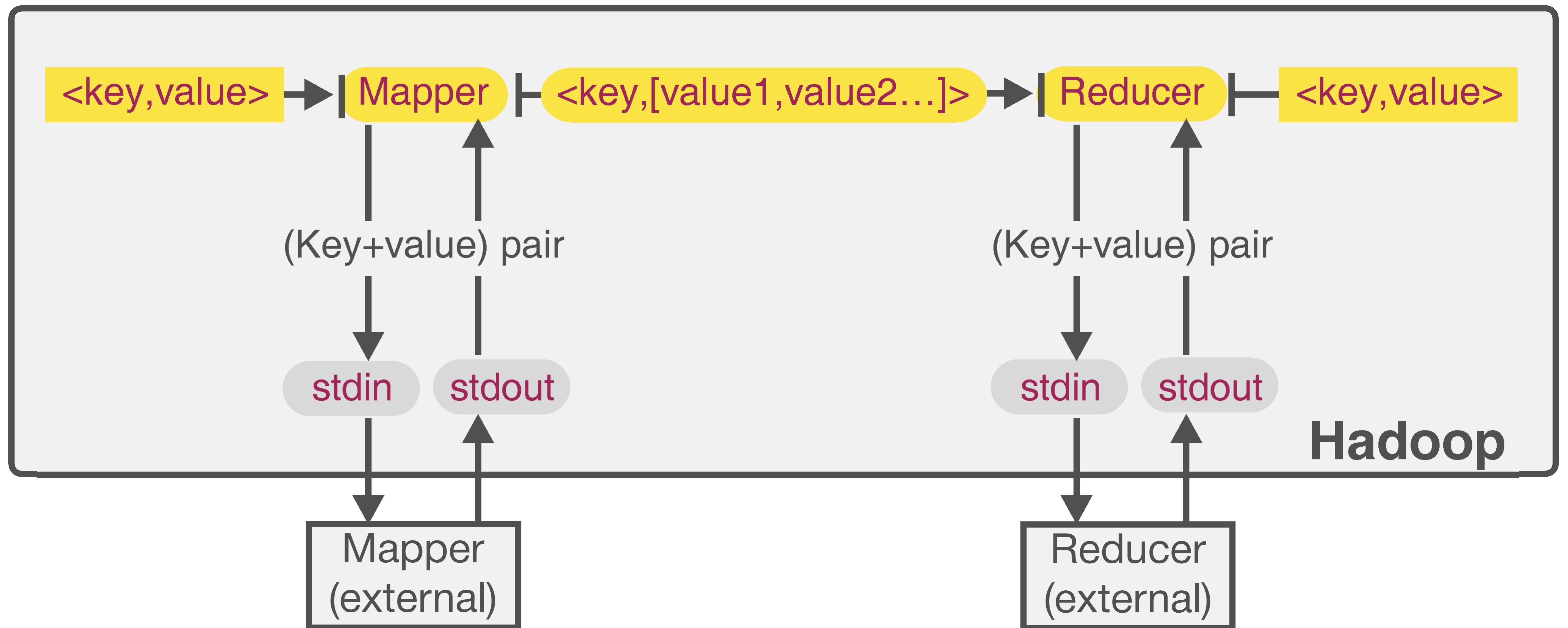


- define input format
- process data
- define output format



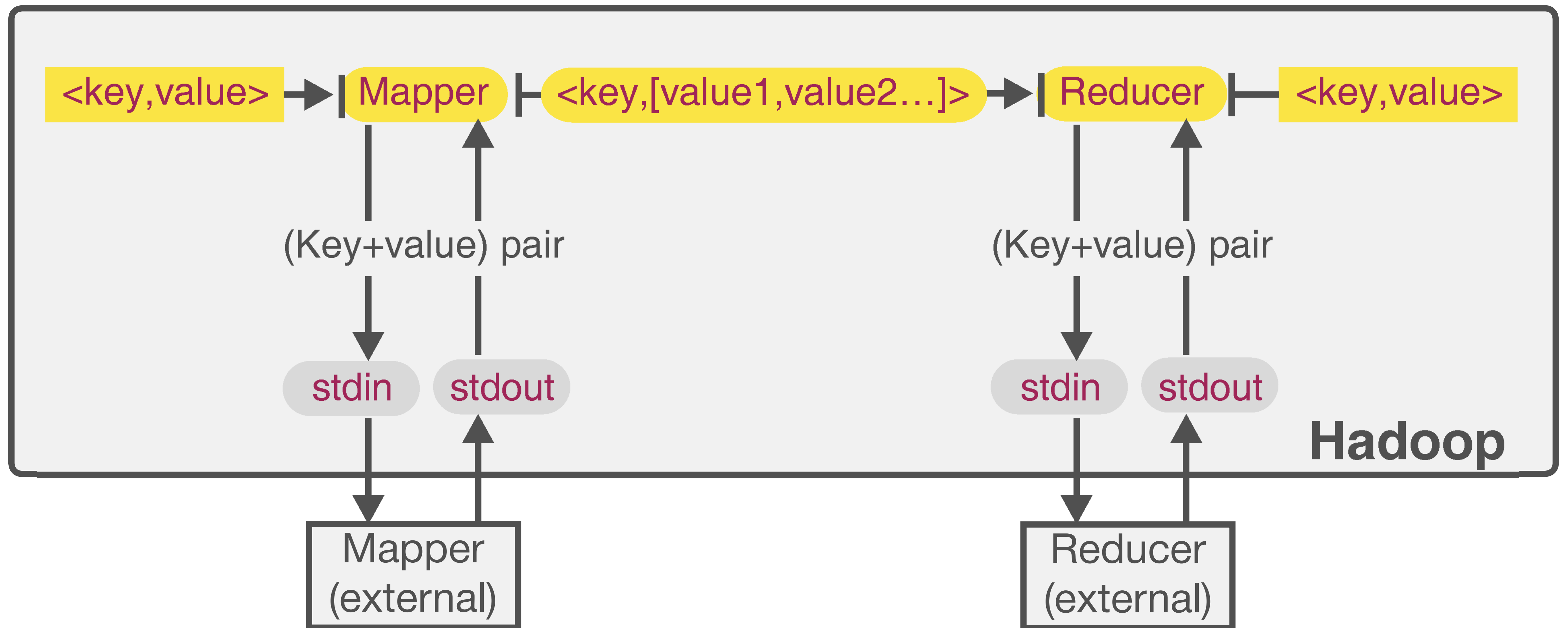
- define input format
- process data
- define output format

- define input format



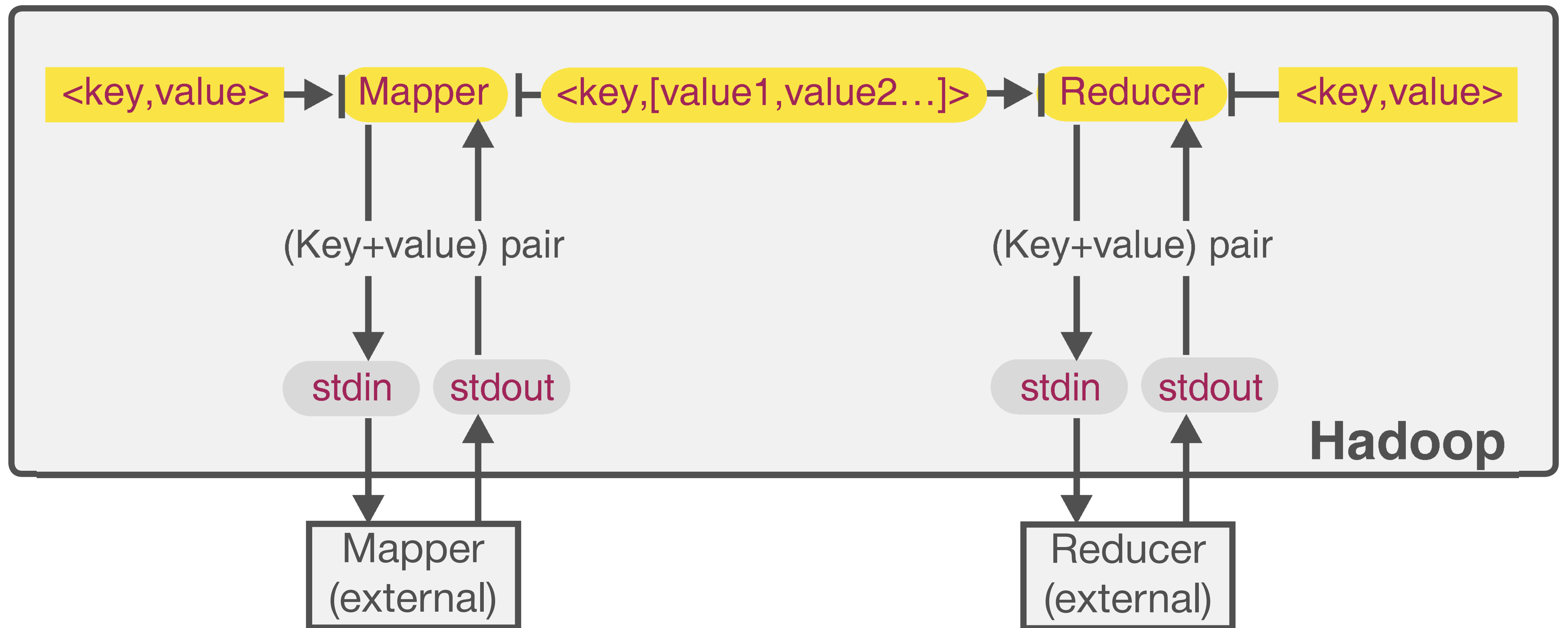
- define input format
- process data
- define output format

- define input format
- aggregate sorted data by key



- define input format
- process data
- define output format

- define input format
- aggregate sorted data by key
- process data



- define input format
- process data
- define output format

- define input format
- aggregate sorted data by key
- process data
- define output format



WIKIPEDIA
The Free Encyclopedia

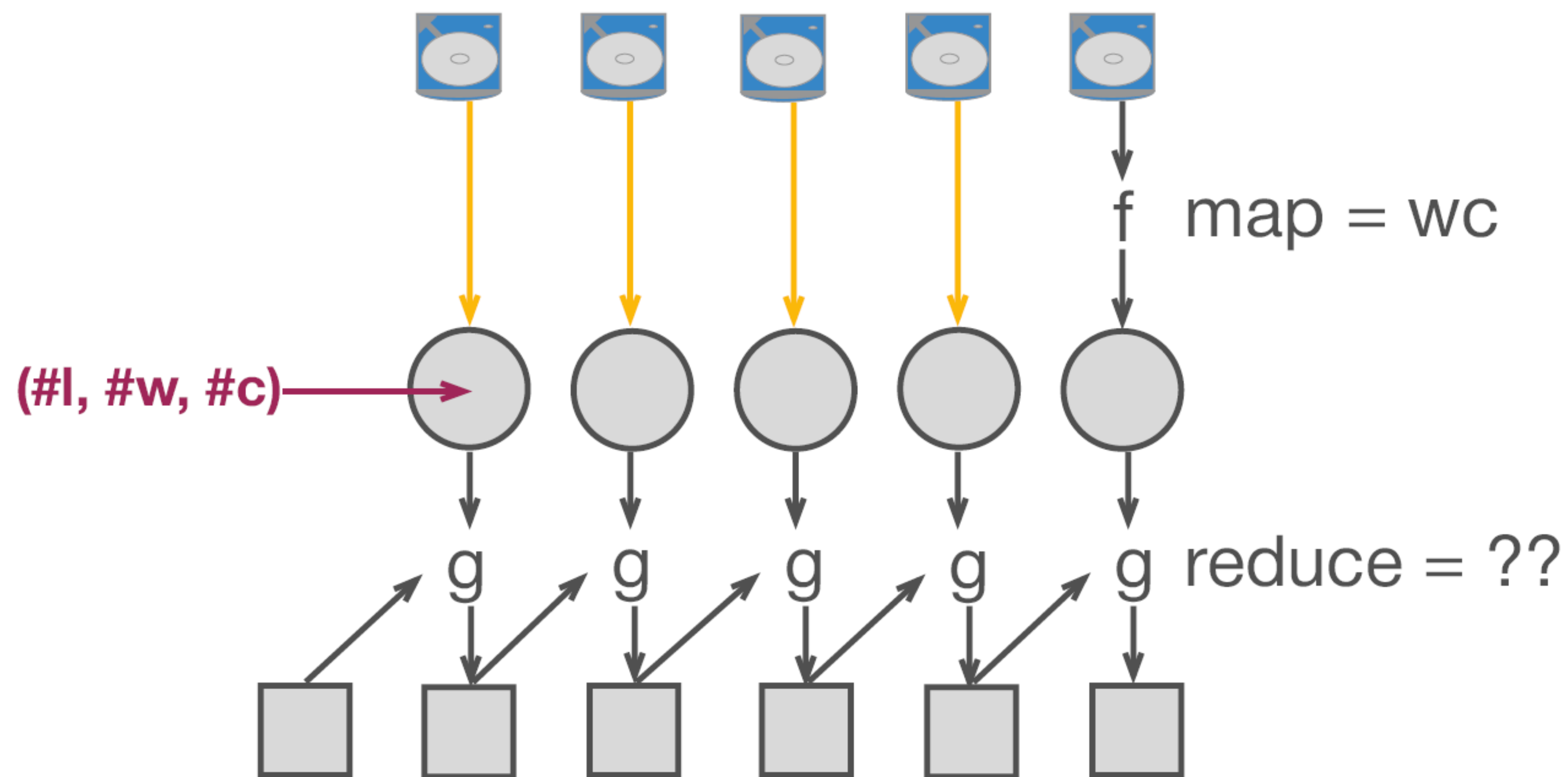


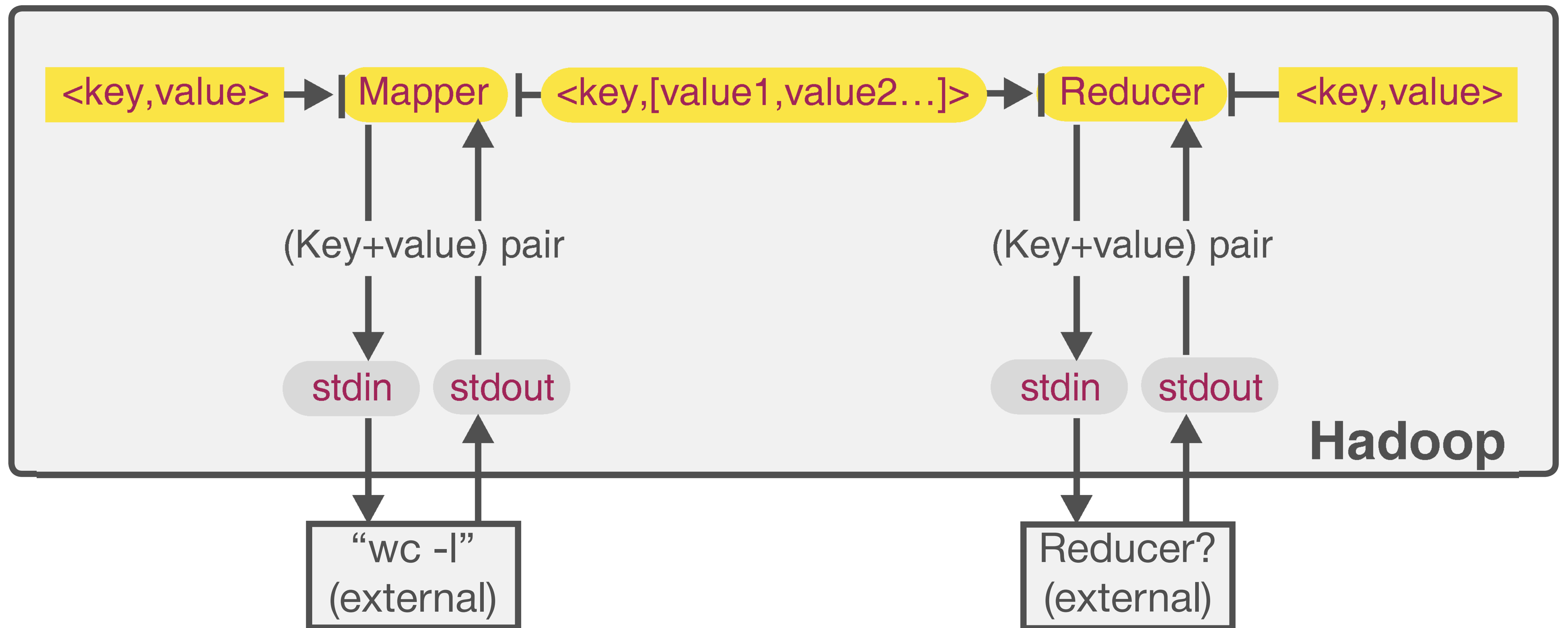
WIKIPEDIA
The Free Encyclopedia

<article id> <tab> <article content>

Line Count?

Distributed Shell: **wc**







```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```

/opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop-mapreduce/hadoop-streaming.jar



```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```



```
$ man locate
```

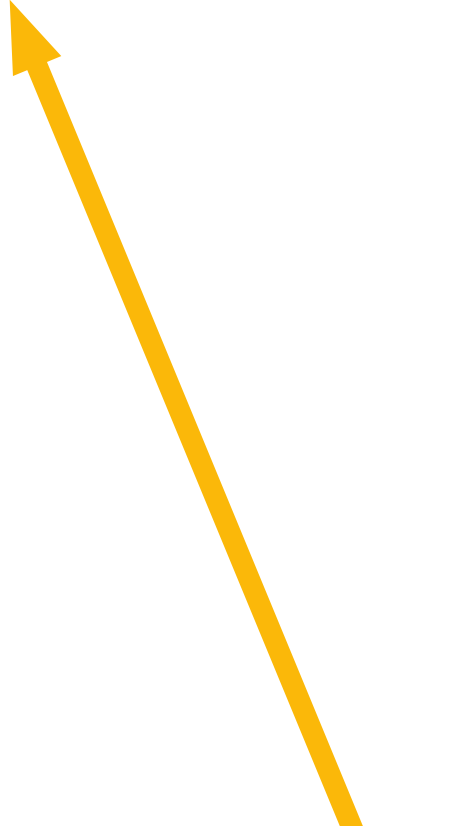
```
/opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop-mapreduce/hadoop-streaming.jar
```

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```



```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \
```



```
-mapper 'wc -l' \
```


```
-numReduceTasks 0 \
```

```
-input /data/wiki/en_articles \
```

```
-output wc_mr
```


```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```




```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```



```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```



```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```

ERROR streaming.StreamJob: Error Launching job : Output directory
hdfs://virtual-master.atp-fvt.org:8020/user/adral/**wc_mr already exists**
Streaming Command Failed!

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```

ERROR streaming.StreamJob: Error Launching job : Output directory
hdfs://virtual-master.atp-fvt.org:8020/user/adral/**wc_mr already exists**
Streaming Command Failed!

```
$ hdfs dfs -rm -r wc_mr
```



```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```

```
$ hdfs dfs -ls wc_mr
```

```
Found 3 items
```

-rw-r--r--	3	adral	adral	0	2017-03-21	14:48	wc_mr/_SUCCESS
-rw-r--r--	3	adral	adral	6	2017-03-21	14:48	wc_mr/part-00000
-rw-r--r--	3	adral	adral	6	2017-03-21	14:48	wc_mr/part-00001

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```

```
$ hdfs dfs -ls wc_mr
```

```
Found 3 items
```

-rw-r--r--	3	adral	adral	0	2017-03-21	14:48	wc_mr/_SUCCESS
-rw-r--r--	3	adral	adral	6	2017-03-21	14:48	wc_mr/part-00000
-rw-r--r--	3	adral	adral	6	2017-03-21	14:48	wc_mr/part-00001

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

```
1986
```

```
2114
```

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -numReduceTasks 0 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```

```
$ hdfs dfs -ls wc_mr
```

```
Found 3 items
```


-rw-r--r--	3	adral	adral	0	2017-03-21	14:48	wc_mr/_SUCCESS
-rw-r--r--	3	adral	adral	6	2017-03-21	14:48	wc_mr/part-00000
-rw-r--r--	3	adral	adral	6	2017-03-21	14:48	wc_mr/part-00001

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

```
1986
```


```
2114
```

$$1968 + 2114 = 4100$$

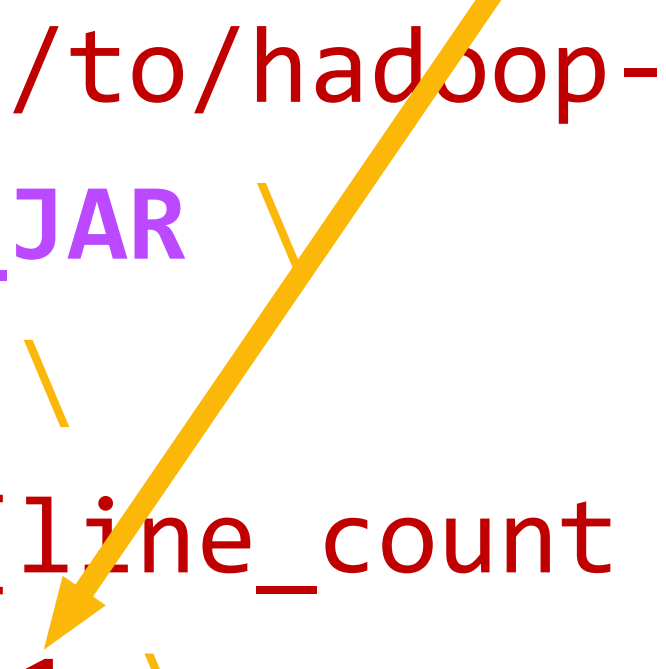


```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
yarn jar $HADOOP_STREAMING_JAR \
    -mapper 'wc -l' \
    -reducer "awk '{line_count += $1} END { print line_count }'" \
    -numReduceTasks 1 \
    -input /data/wiki/en_articles \
    -output wc_mr
```

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
yarn jar $HADOOP_STREAMING_JAR \
    -mapper 'wc -l' \
    -reducer "awk '{line_count += $1} END { print line_count }'" \
    -numReduceTasks 1 \
    -input /data/wiki/en_articles \
    -output wc_mr
```



```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
yarn jar $HADOOP_STREAMING_JAR \
    -mapper 'wc -l' \
    -reducer "awk '{line_count += \$1} END { print line_count }'" \
    -numReduceTasks 1 \
    -input /data/wiki/en_articles \
    -output wc_mr
```



```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -reducer "awk '{line_count += \$1} END { print line_count }'" \  
    -numReduceTasks 1 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```

```
$ hdfs dfs -ls wc_mr_with_reducer
```

```
Found 2 items
```

-rw-r--r--	3	adral	adral	wc_mr_with_reducer/_SUCCESS
-rw-r--r--	3	adral	adral	wc_mr_with_reducer/part-00000

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

```
4100
```

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \  
    -mapper 'wc -l' \  
    -reducer "awk '{line_count += \$1} END { print line_count }'" \  
    -numReduceTasks 1 \  
    -input /data/wiki/en_articles \  
    -output wc_mr
```

```
$ hdfs dfs -ls wc_mr_with_reducer
```

```
Found 2 items
```

-rw-r--r--	3	adral	adral	wc_mr_with_reducer/_SUCCESS
-rw-r--r--	3	adral	adral	wc_mr_with_reducer/part-00000


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

```
4100
```



HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"

yarn jar \$HADOOP_STREAMING_JAR \
 -mapper 'wc -l' \
 -reducer "awk '{line_count += \ \$1} END { print line_count }'" \
 -numReduceTasks 1 \
 -input /data/wiki/en_articles \
 -output wc_mr



reducer.sh

```
#!/usr/bin/env bash
```

```
awk '{line_count += $1} END { print line_count }'
```

reducer.sh

```
#!/usr/bin/env bash
```

```
awk '{line_count += $1} END { print line_count }'
```



reducer.sh

```
#!/usr/bin/env bash
```

```
awk '{line_count += $1} END { print line_count }'
```

```
HADOOP_STREAMING_JAR="/path/to/hadoop-streaming.jar"
```

```
yarn jar $HADOOP_STREAMING_JAR \
```

```
    -mapper 'wc -l' \
```



```
    -reducer './reducer.sh' \
```



```
    -file reducer.sh \
```

```
    -numReduceTasks 1 \
```

```
    -input /data/wiki/en_articles \
```

```
    -output wc_mr_with_reducer
```

