Running MapReduce in Docker

EECS 4415

Big Data Systems

Tilemachos Pechlivanoglou

tipech@eecs.yorku.ca

Last week's WordCount example

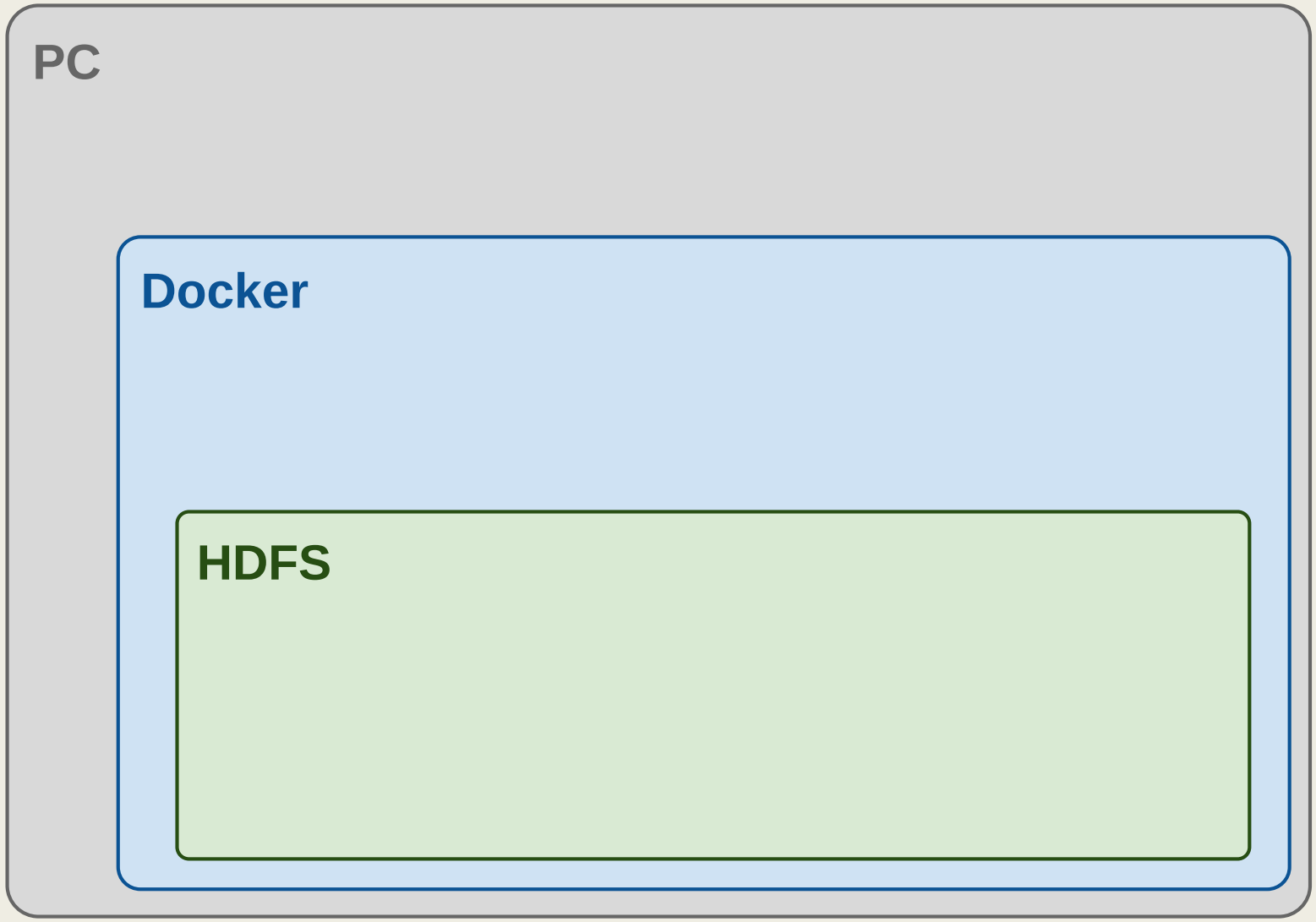
mapper.py

```
#!/usr/bin/python

import sys
import re

for line in sys.stdin:
    line = re.sub( r'^\W+|\W+$', '', line )
    words = re.split(r"\W+", line)

    for word in words:
        print( word.lower() + "\t1" )
```

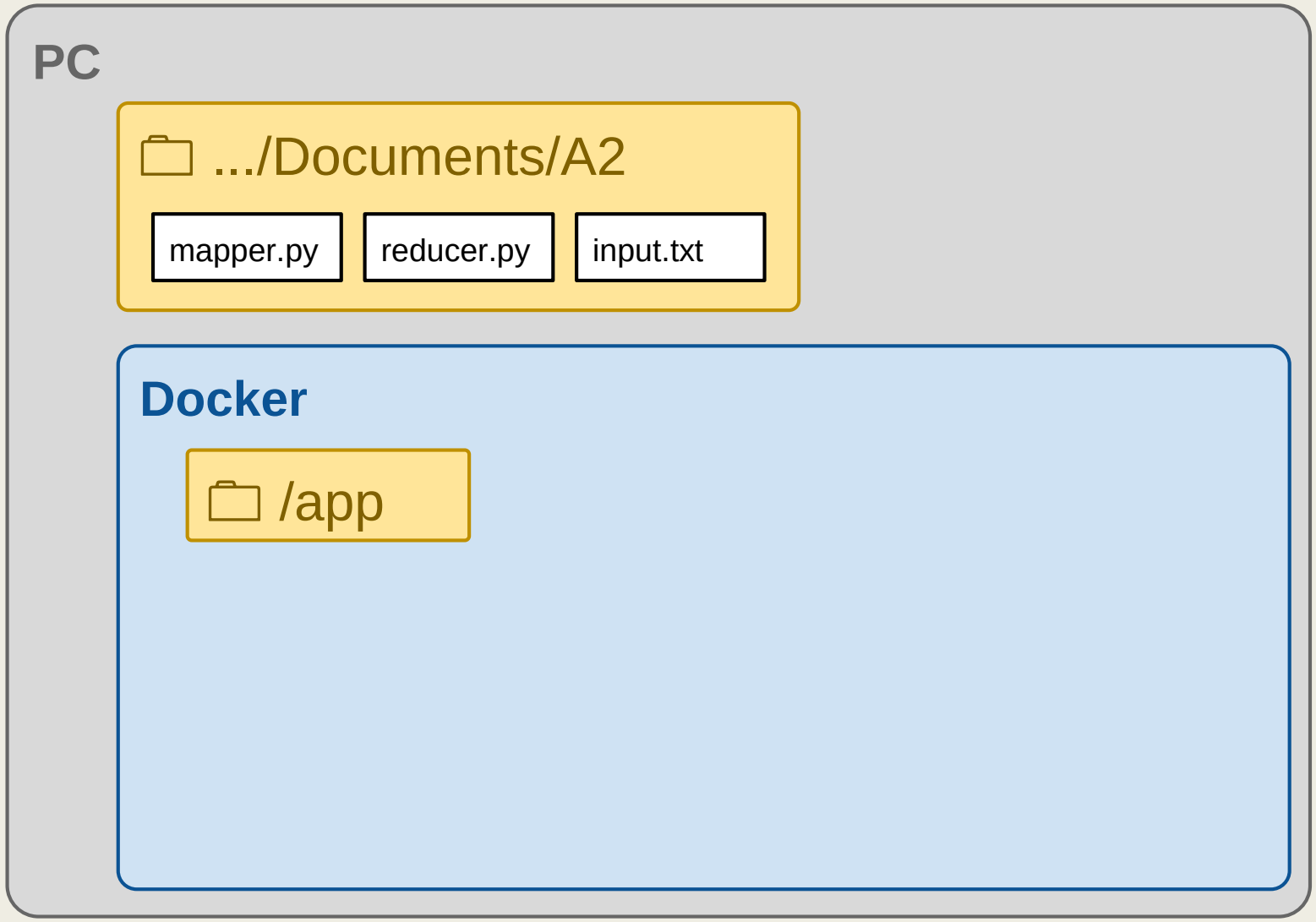
reducer.py

```
#!/usr/bin/python

import sys

previous = None
sum = 0

for line in sys.stdin:
    key, value = line.split( '\t' )

    if key != previous:
        if previous is not None:
            print str( sum ) + '\t' + previous
        previous = key
        sum = 0

    sum = sum + int( value )

print str( sum ) + '\t' + previous
```

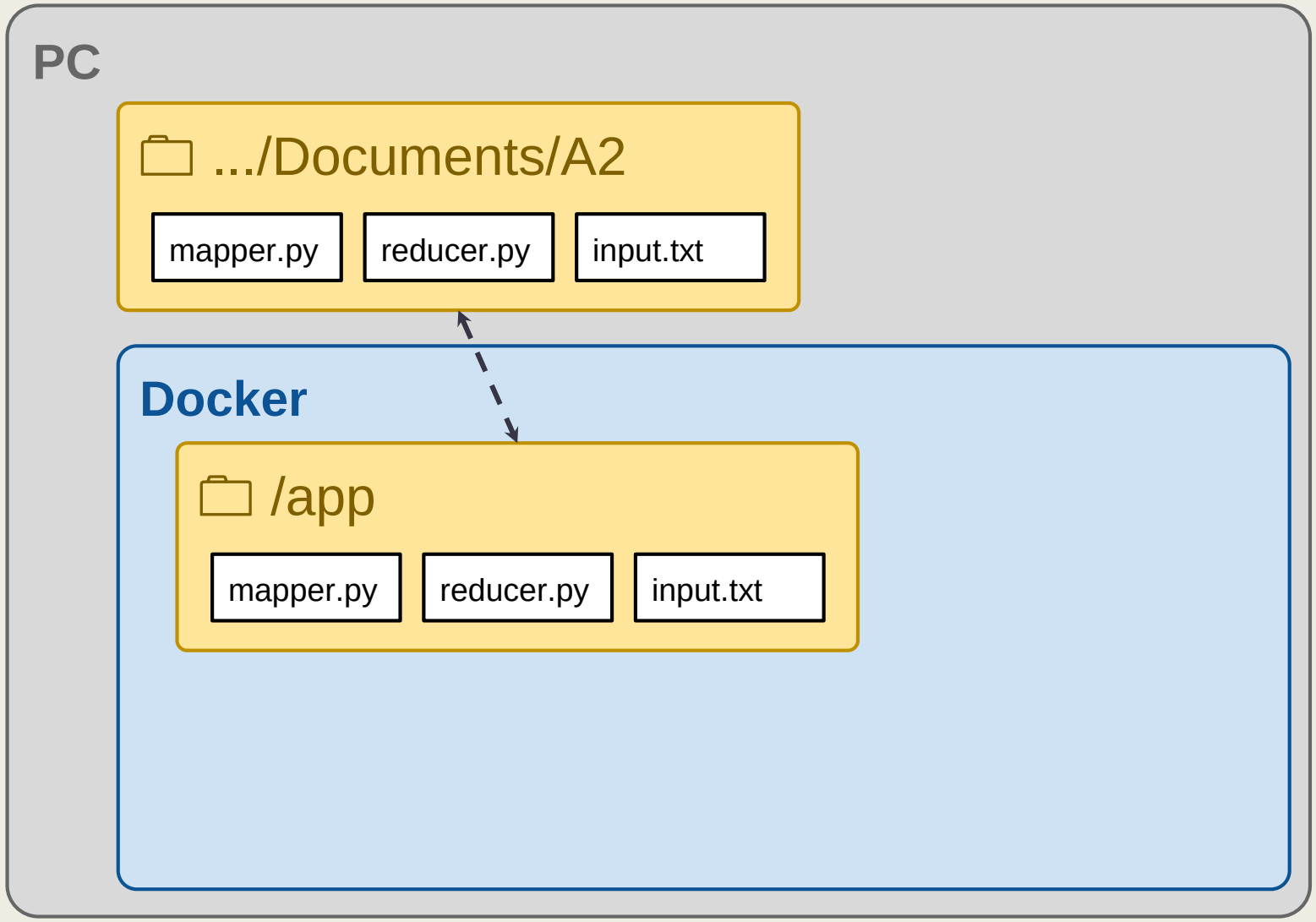
System representation



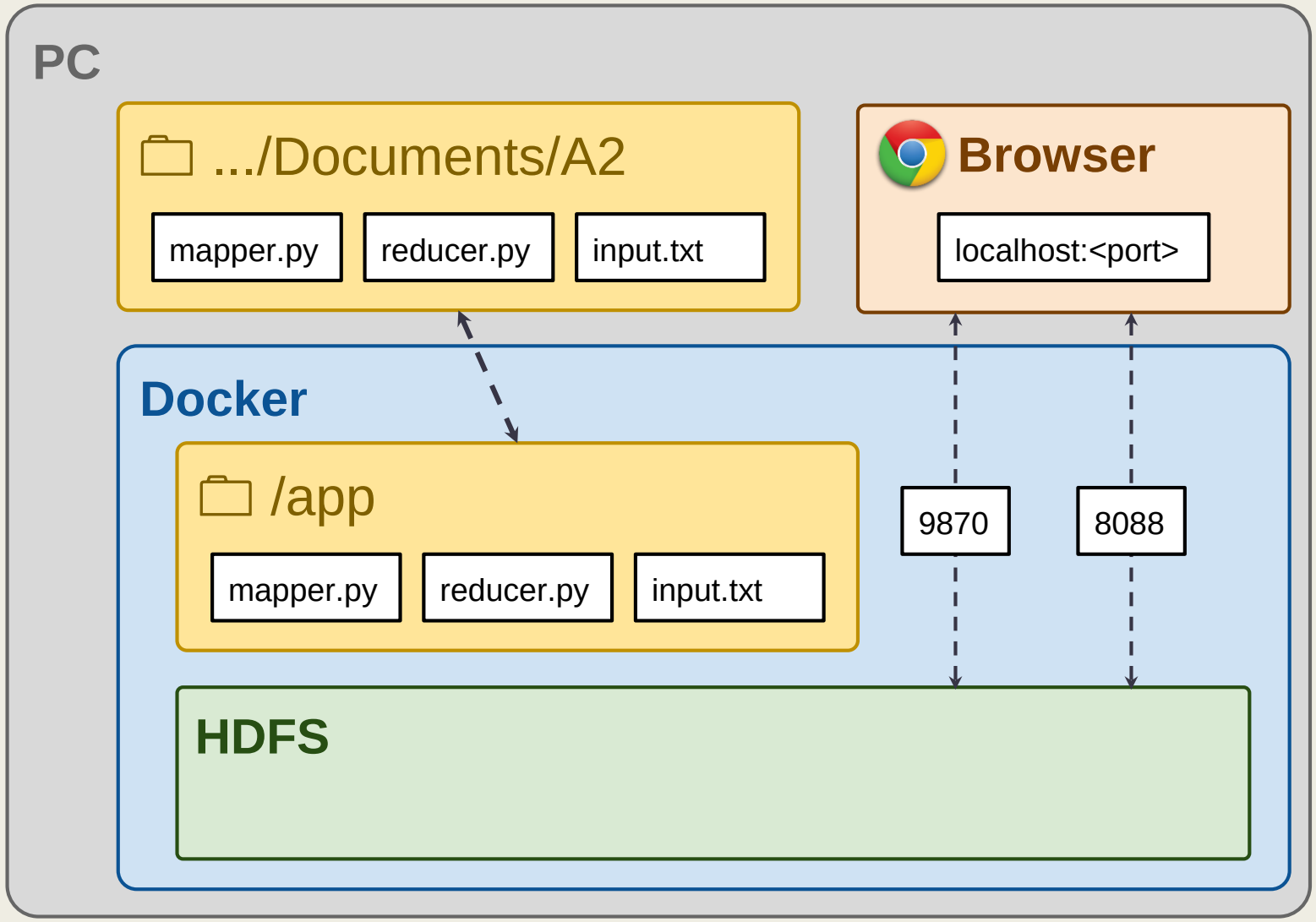
`docker -it run eecsyorku/eecs4415`



```
docker -it run -v $PWD:/app eecsyorku/eecs4415
```



```
docker -it run -p 9870:9870 -p 8088:8088
-v $PWD:/app eecsyorku/eecs4415
```

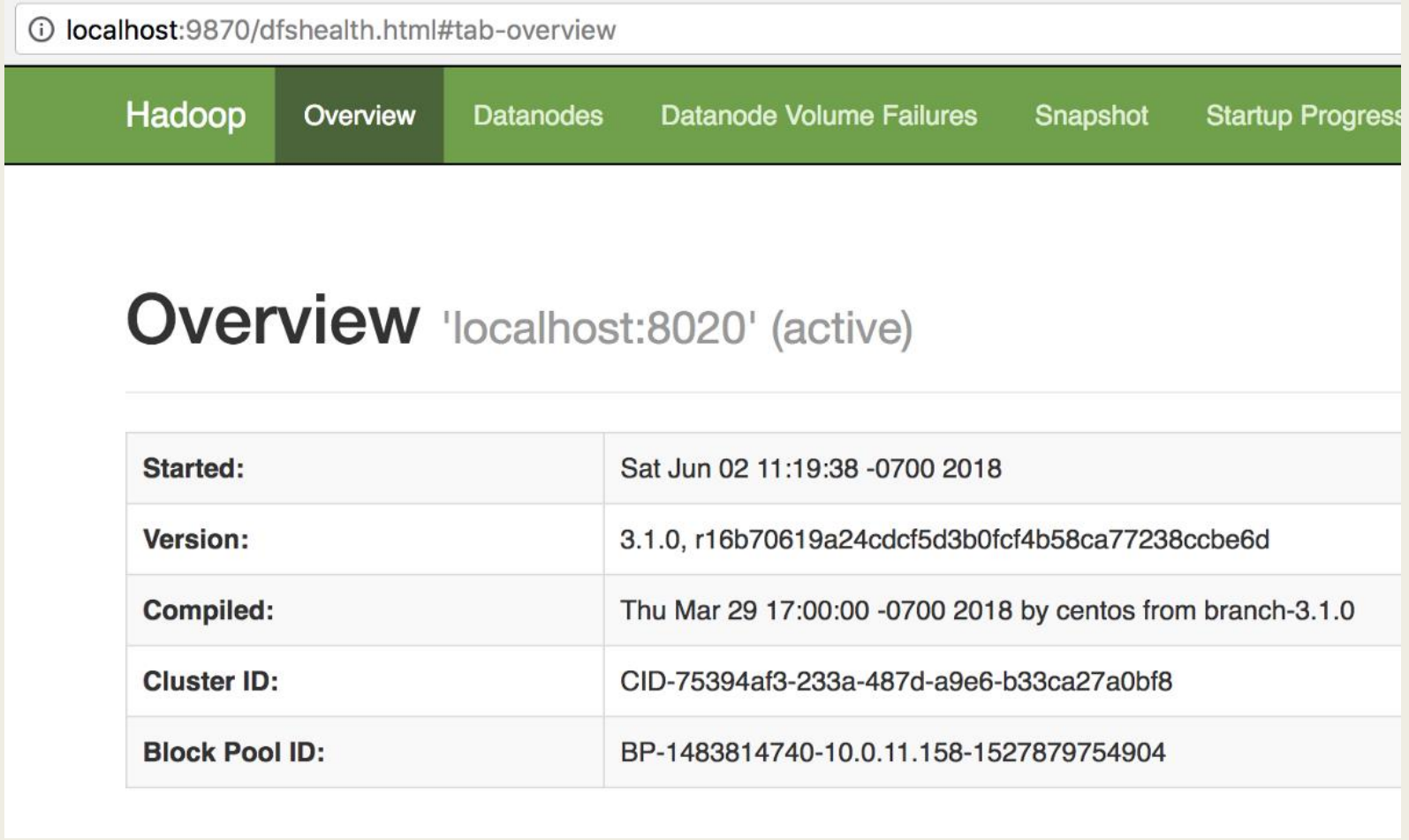


docker -it run -p 9870:9870 -p 8088:8088 -v $PWD:/app eecsyorku/eecs4415

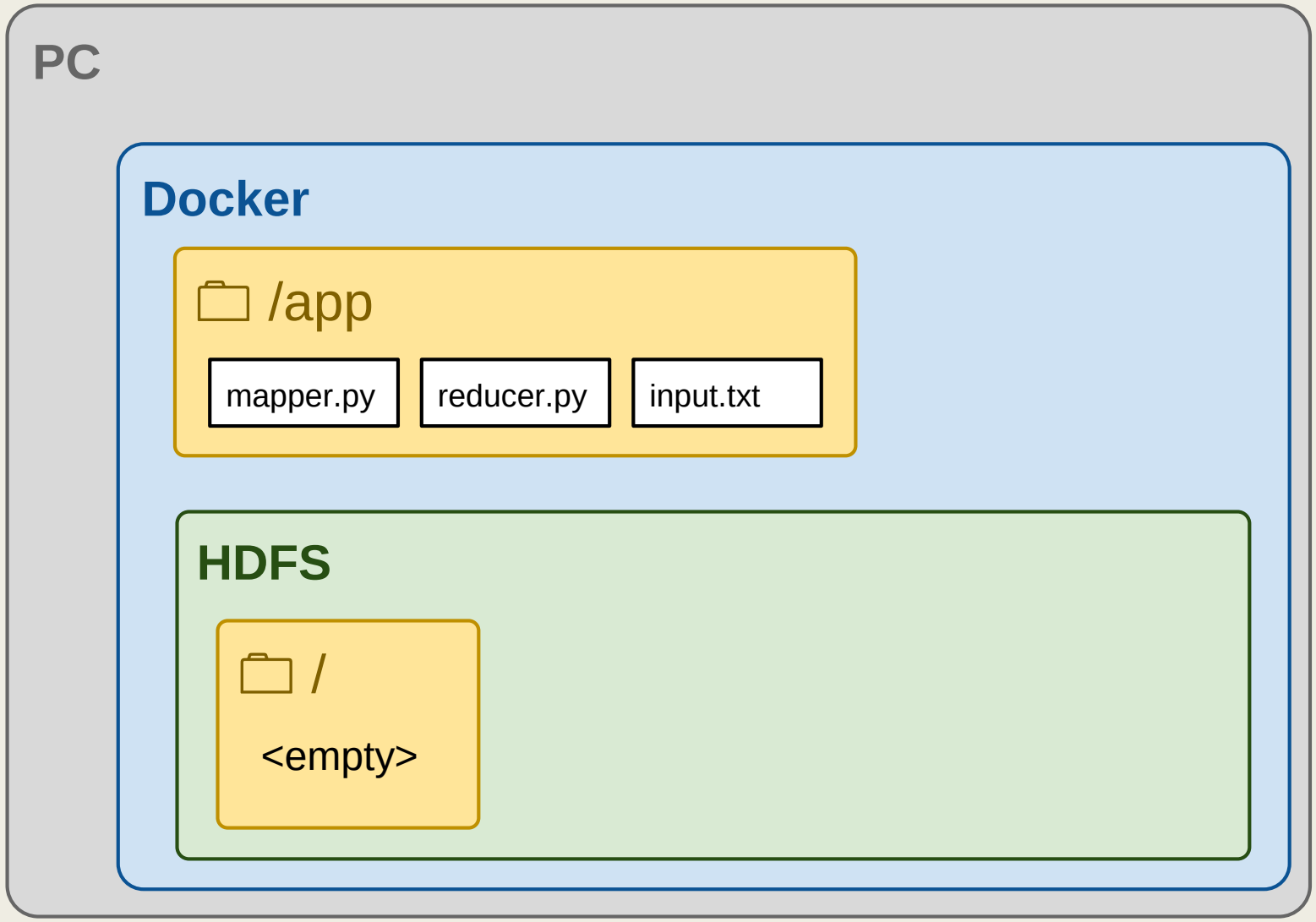
```
> ls
input.txt  mapper.py  reducer.py
```

```
> docker run -it -p 9870:9870 -p 8088:8088 -p 8042:8042 -v $PWD:/app eecsyorku/eecs4415
[ ok ] Starting OpenBSD Secure Shell server: sshd.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [8326269d4ee8]
Starting resourcemanager
Starting nodemanagers
root@8326269d4ee8:/app# ls
input.txt  mapper.py  reducer.py
root@8326269d4ee8:/app#
```

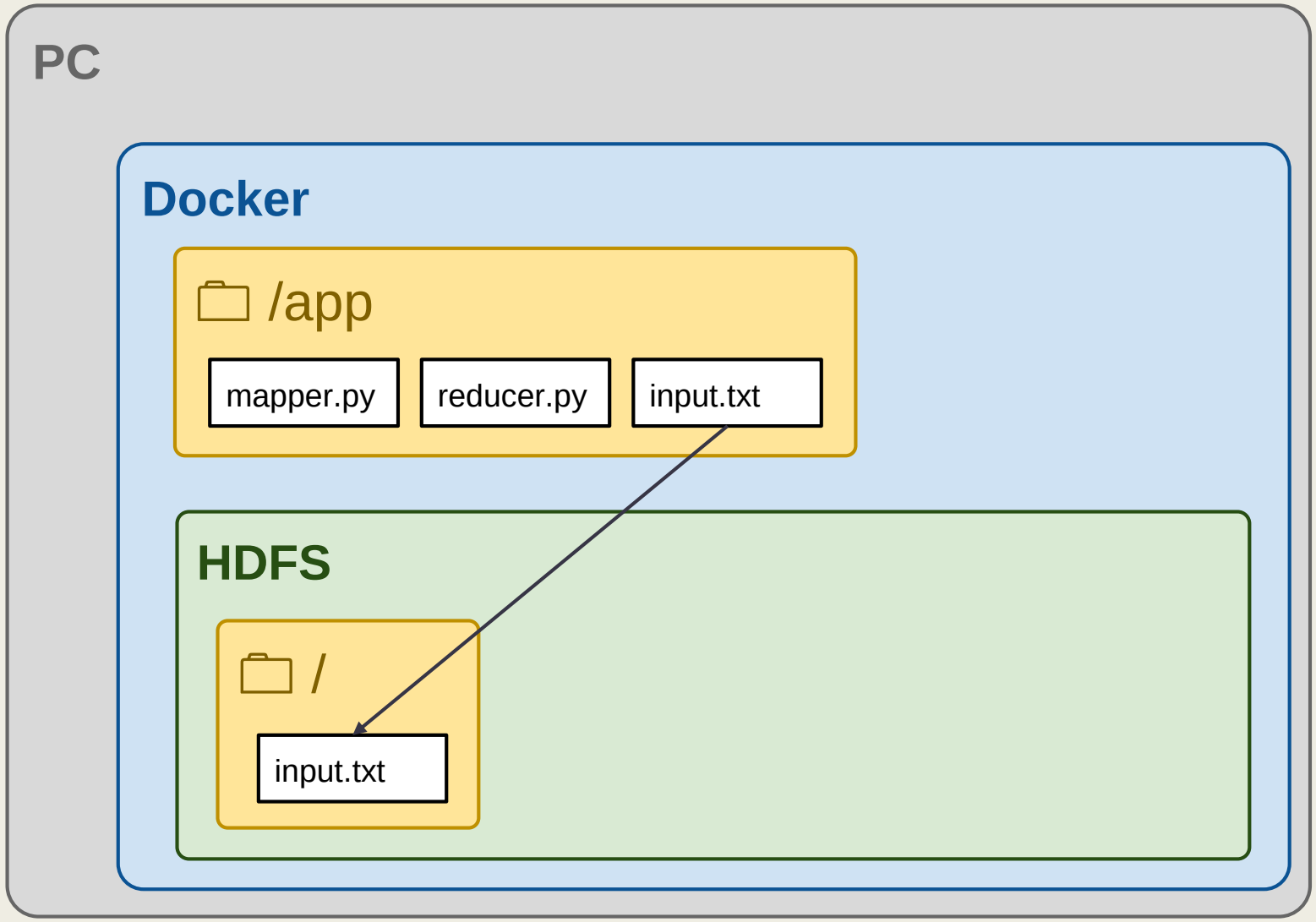
Web UI


localhost:9870/dfshealth.html#tab-overview
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Overview 'localhost:8020' (active)
Started: Sat Jun 02 11:19:38 -0700 2018
Version: 3.1.0, r16b70619a24cdcf5d3b0fcf4b58ca77238ccbe6d
Compiled: Thu Mar 29 17:00:00 -0700 2018 by centos from branch-3.1.0
Cluster ID: CID-75394af3-233a-487d-a9e6-b33ca27a0bf8
Block Pool ID: BP-1483814740-10.0.11.158-1527879754904

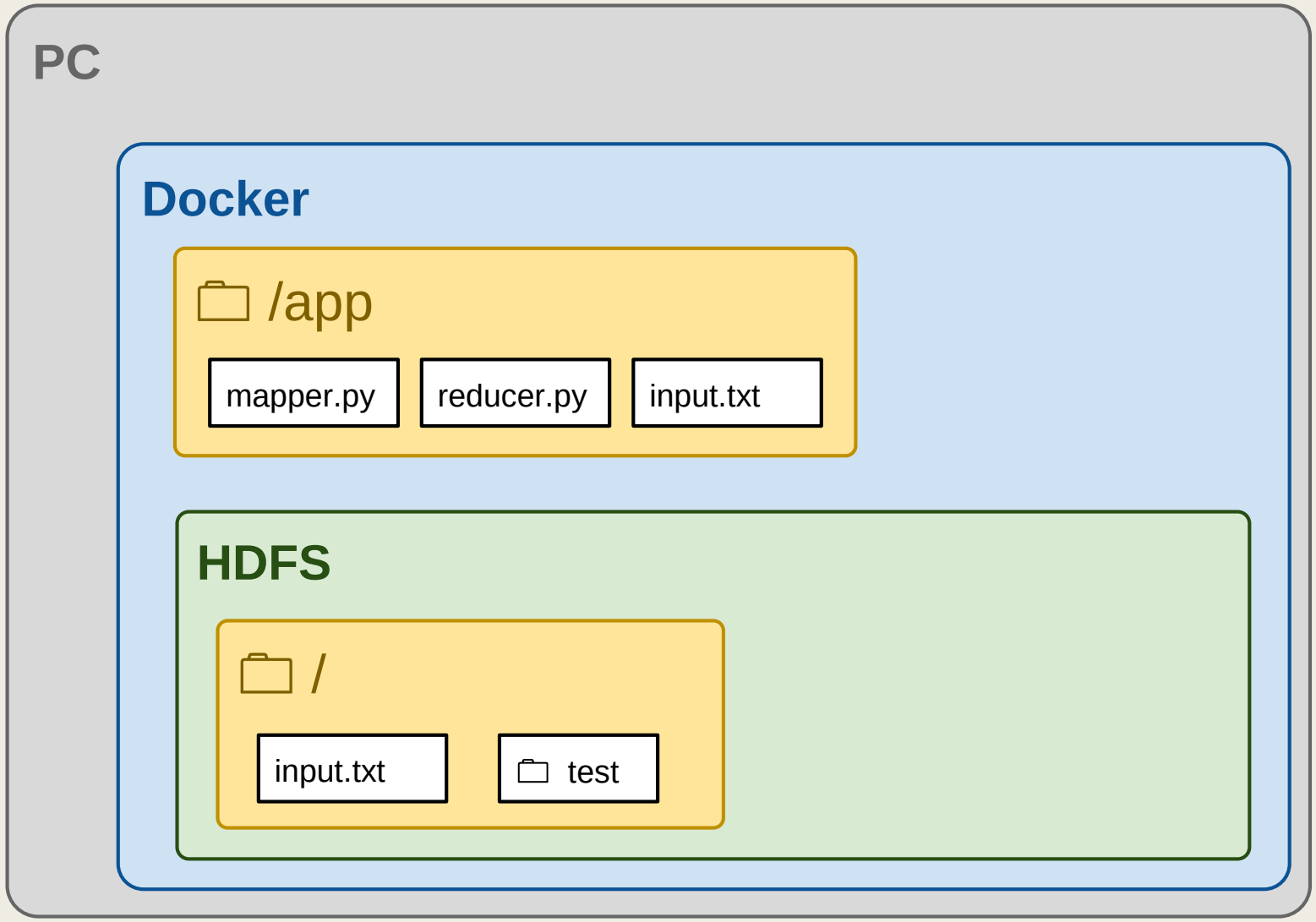
```
hdfs dfs -ls /
```



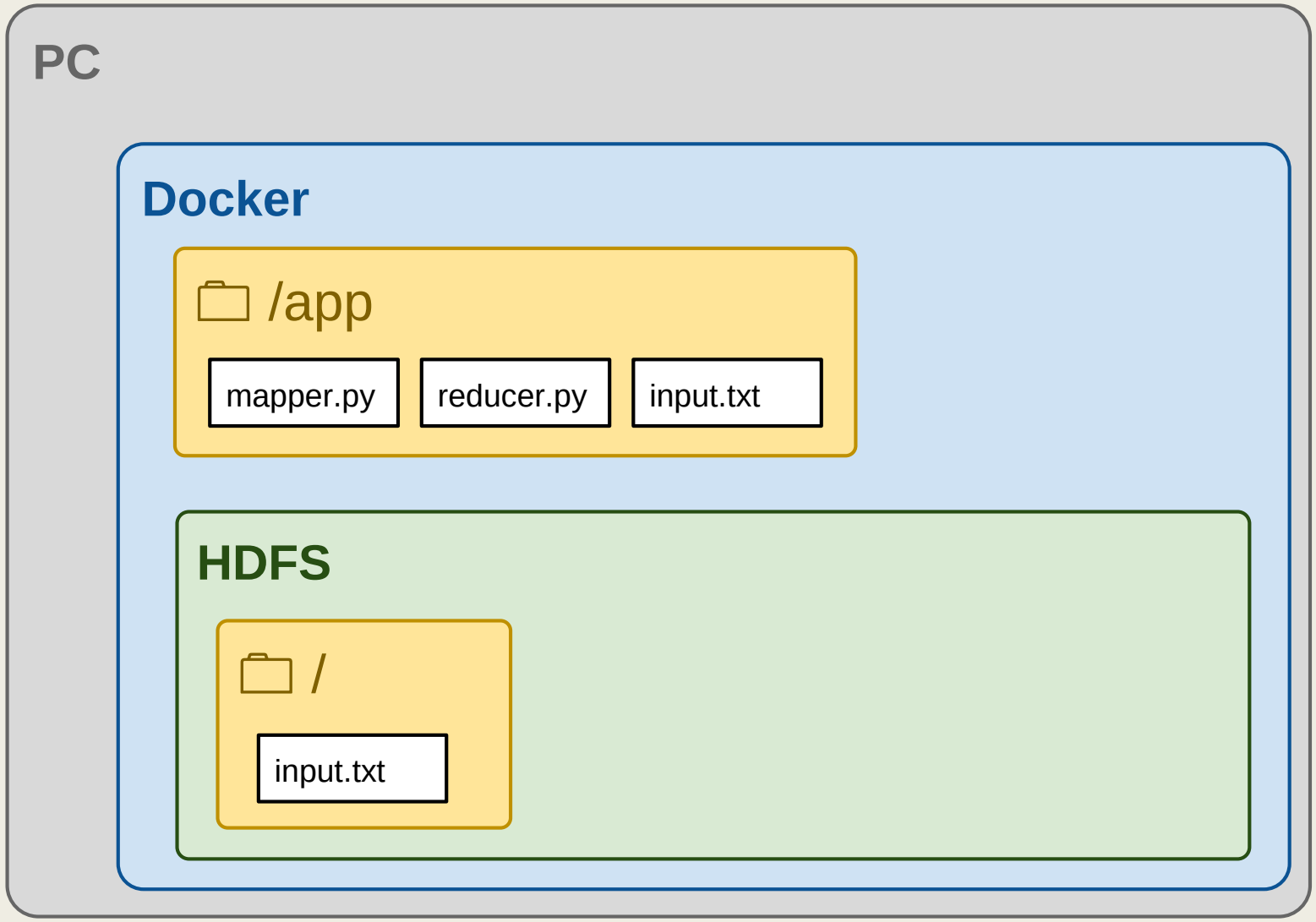
```
hdfs dfs -put ./input.txt /
```



```
hdfs dfs -mkdir /test
```



```
hdfs dfs -rm -r /test
```

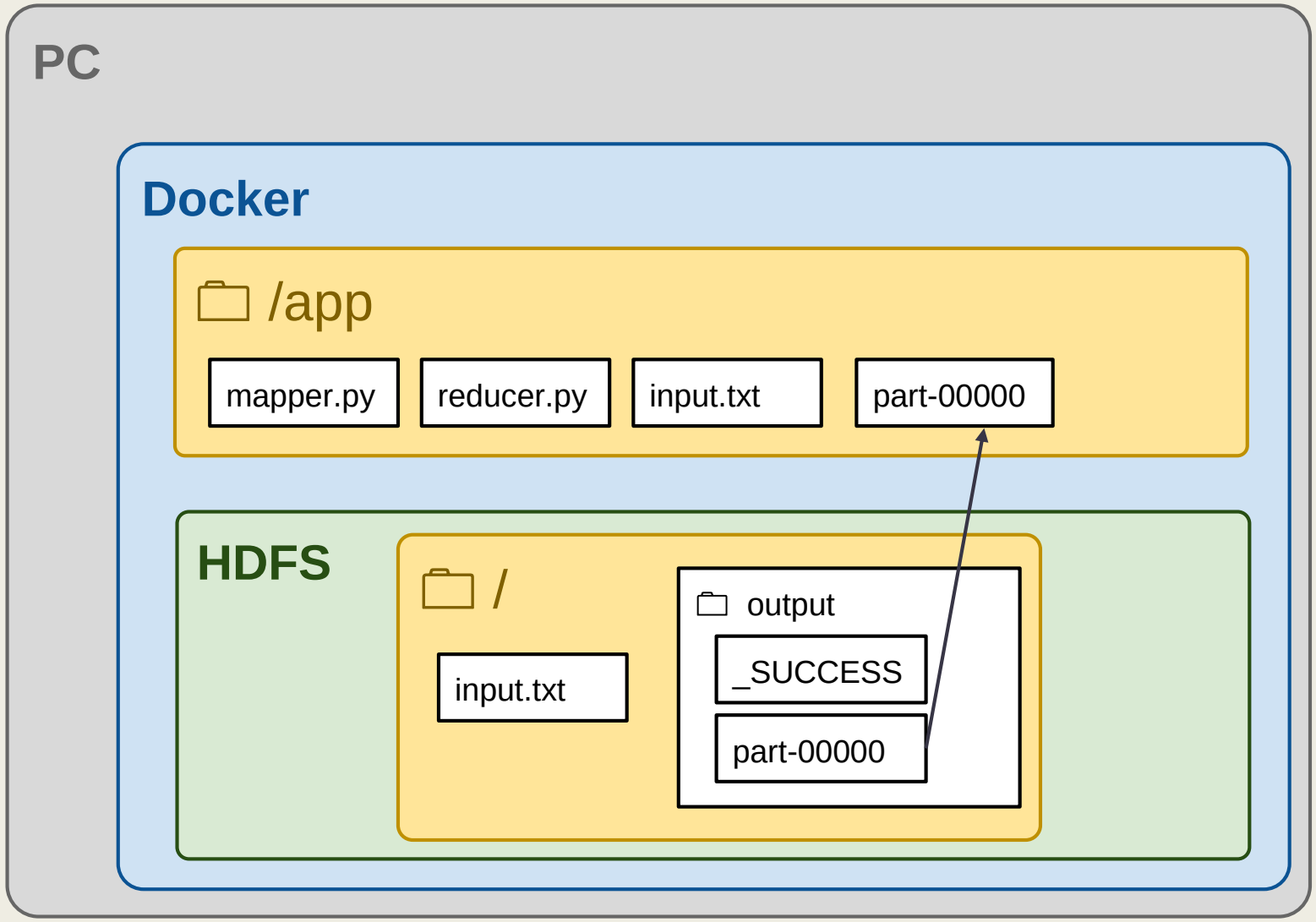


Running MapReduce

```
hadoop jar /usr/hadoop-3.0.0/share/hadoop/tools/lib/hadoop-streaming-3.0.0.jar \
-file ./mapper.py \
-mapper ./mapper.py \
-file ./reducer.py \
-reducer ./reducer.py \
-input /input.txt \
-output /output
```

```
> hadoop jar /usr/hadoop-3.0.0/share/hadoop/tools/lib/hadoop-streaming-3.0.0.jar \
> -file ./mapper.py \
> -mapper ./mapper.py \
> -file ./reducer.py \
> -reducer ./reducer.py \
> -input /input.txt \
> -output /output
2018-10-30 22:12:16,761 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [./mapper.py, ./reducer.py, /tmp/hadoop-unjar6527467371105808420/] [] /tmp/streamjob9036274102151078764.jar tmpDir=null
2018-10-30 22:12:17,371 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2018-10-30 22:12:17,514 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2018-10-30 22:12:17,713 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1540937348743_0003
2018-10-30 22:12:18,719 INFO mapred.FileInputFormat: Total input files to process : 1
2018-10-30 22:12:19,154 INFO mapreduce.JobSubmitter: number of splits:2
2018-10-30 22:12:19,179 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-
ublisher.enabled
2018-10-30 22:12:19,234 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1540937348743_0003
2018-10-30 22:12:19,235 INFO mapreduce.JobSubmitter: Executing with tokens: []
2018-10-30 22:12:19,341 INFO conf.Configuration: resource-types.xml not found
2018-10-30 22:12:19,341 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2018-10-30 22:12:19,378 INFO impl.YarnClientImpl: Submitted application application_1540937348743_0003
2018-10-30 22:12:19,401 INFO mapreduce.Job: The url to track the job: http://be521460fac0:8088/proxy/application_1540937348743_0003/
2018-10-30 22:12:19,402 INFO mapreduce.Job: Running job: job_1540937348743_0003
2018-10-30 22:12:23,454 INFO mapreduce.Job: Job job_1540937348743_0003 running in uber mode : false
2018-10-30 22:12:23,454 INFO mapreduce.Job:  map 0% reduce 0%
2018-10-30 22:12:28,497 INFO mapreduce.Job:  map 100% reduce 0%
2018-10-30 22:12:32,517 INFO mapreduce.Job:  map 100% reduce 100%
2018-10-30 22:12:32,521 INFO mapreduce.Job: Job job_1540937348743_0003 completed successfully
```

```
hdfs dfs -get /output/part*
```



Success!

Contents of part-00000:

```
yield   15
yielded 2
yielding        3
yields  3
yojo    17
yoke    3
yoked   4
yokes   1
yoking  1
yon     10
yonder  18
yore    2
york    5
york_   1
yorkshire       1
you     958
young   80
younger 2
youngest        1
youngish        1
your    257
yours   8
yourselbs       1
yourself        26
yourselves      7
youth   9
youthful        2
zag     1
zay     1
zeal    2
zealand 7
zealanders      1
zephyr  1
zeuglodon       1
zig     1
zip     1
zodiac  5
zogranda        1
zone    5
zoned   2
zones   3
zoology 2
zoroaster       1
```

Dealing with imports

- For Python's externally imported packages (nltk, sklearn):
  - *program will run properly outside Hadoop, but will fail without reason in it*
  - *they need to be loaded into HDFS somehow*

- To load them, compress as zip:
  - `zip -r nltkandyaml.zip nltk sklearn`
  - `mv ntlk_sklearn.zip /path/to/where/your/mapper/will/be/nltk_sklearn.mod`
  - `hadoop … -file ./nltk_sklearn.mod`

- And manually import:
  - `import zipimport`
  - `importer = zipimport.zipimporter('nltk_sklearn.mod')`
  - `sklearn = importer.load_module('sklearn')`
  - `nltk = importer.load_module('nltk')`

Thank you!

Links to check:

https://zettadatanet.wordpress.com/2015/04/04/a-hands-on-introduction-to-mapreduce-in-python/
https://afourtech.com/guide-docker-commands-examples/
https://medium.com/@rrfd/your-first-map-reduce-using-hadoop-with-python-and-osx-ca3b6f3dfe78
https://hadoop.apache.org/docs/r1.2.1/streaming.html