LAPORAN JOBSHEET 7

BIG DATA

Oleh:

ALBANI RAJATA MALIK

2241760080/05

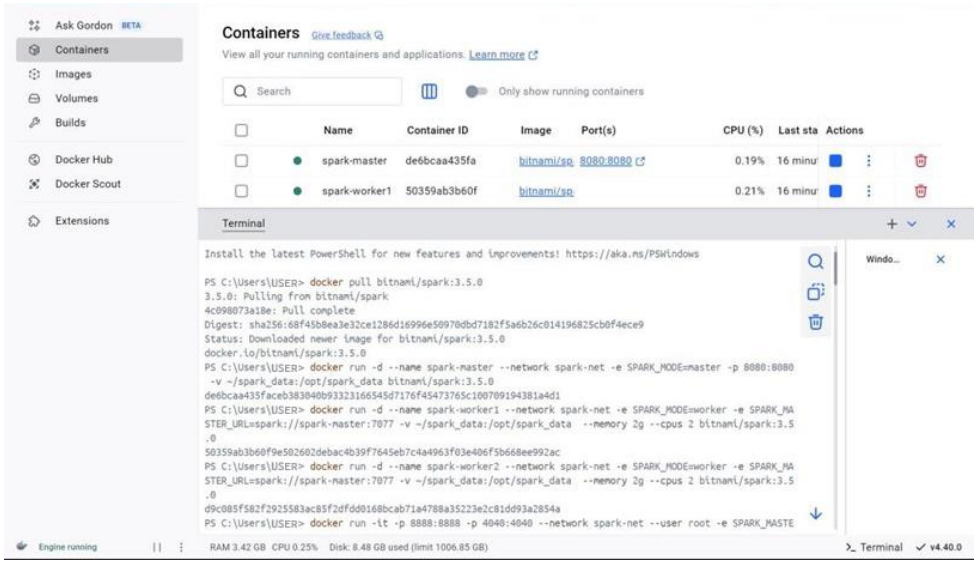
SIB 3E

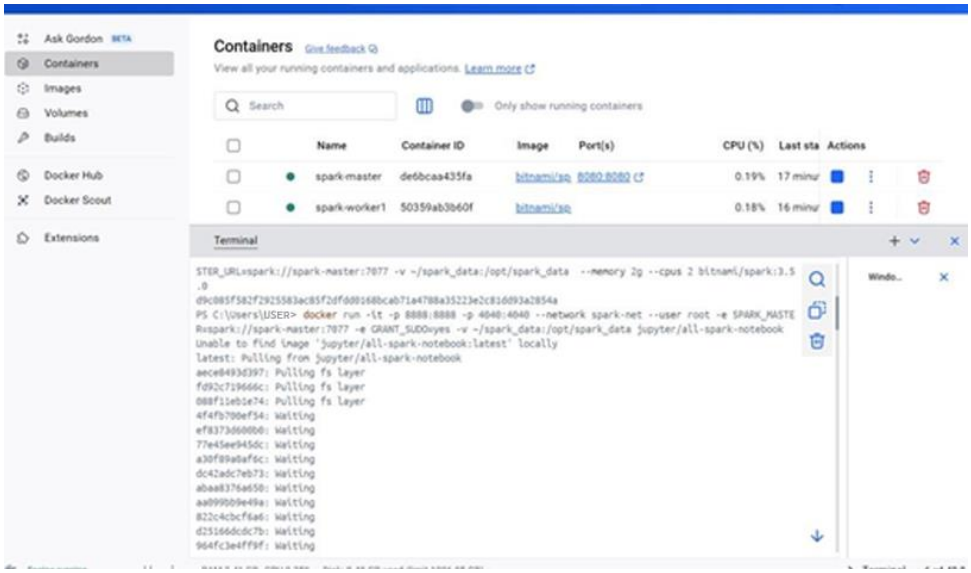
PROGAM STUDI D-IV SISTEM INFORMASI BISNIS

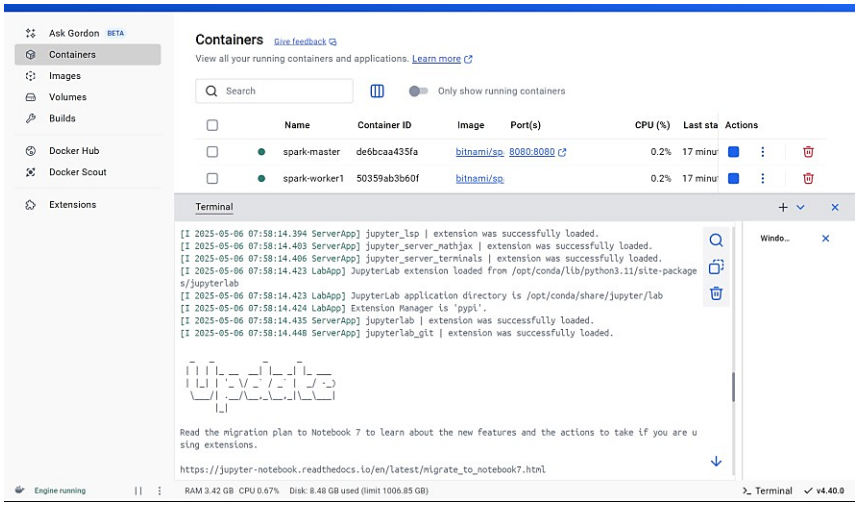
JURUSAN TEKNOLOGI INFORMASI

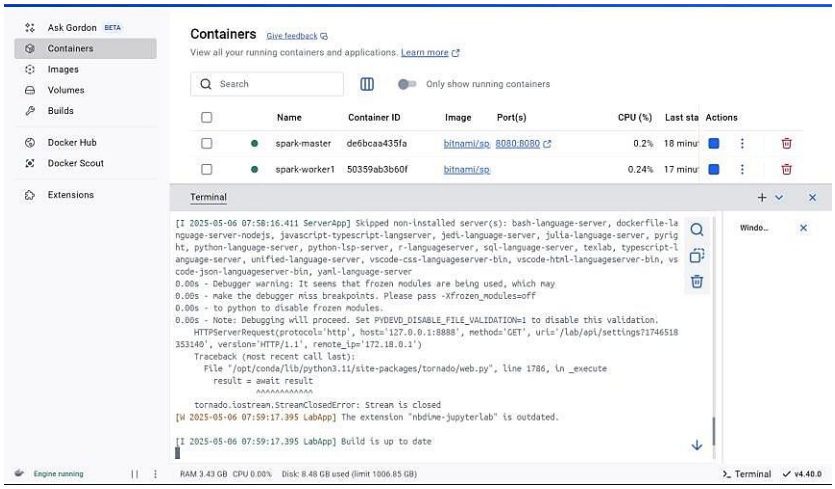
POLITEKNIK NEGERI MALANG

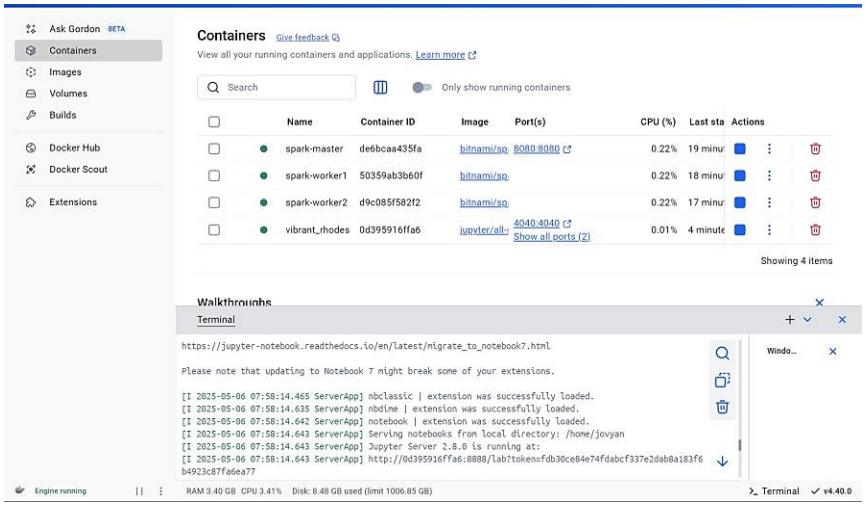
```
PS C:\Users\USER> wsl -l -v
  NAME              STATE           VERSION
* docker-desktop    Running         2
PS C:\Users\USER> docker pull apache/spark:latest
latest: Pulling from apache/spark
b100ab72b3e9: Pull complete
c74c7c34259b: Pull complete
053e658edcf9: Pull complete
2fd40b3fb58e: Pull complete
215ed5a63843: Pull complete
ead058ffaa09: Pull complete
4f4fb700ef54: Pull complete
d4ebd433e6b9: Pull complete
3eaf8f3f646b: Pull complete
357990f05276: Pull complete
8f0df7ed423b: Pull complete
Digest: sha256:2ecf9cab4a1d0df1052731636f6029f6e80489fbd9ad0d24aaa31061f8aa8eea
Status: Downloaded newer image for apache/spark:latest
docker.io/apache/spark:latest
PS C:\Users\USER>
```

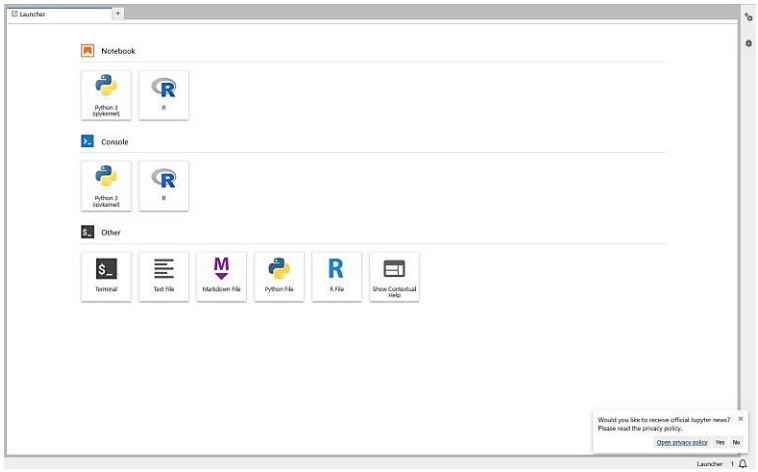


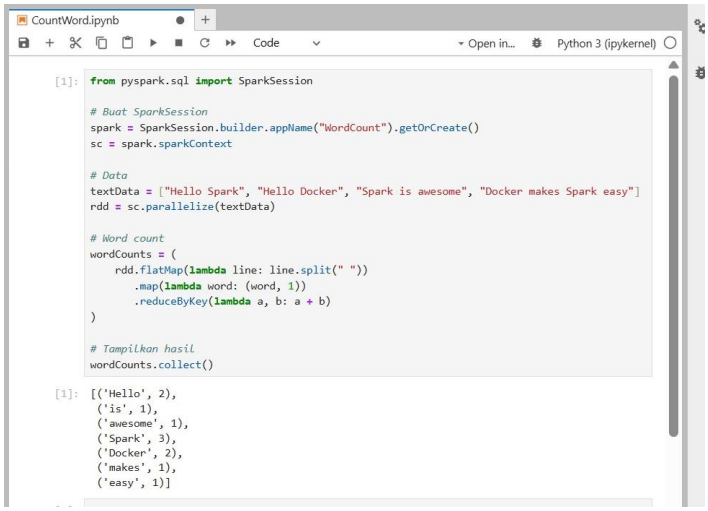


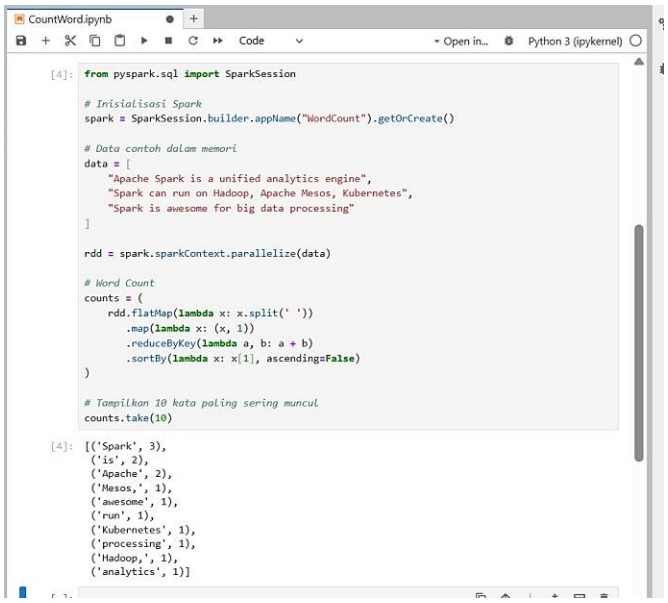

Ask Gordon BETA
Containers
Images
Volumes
Builds
Docker Hub
Docker Scout
Extensions
Containers
Give feedback
View all your running containers and applications. Learn more
Search
Only show running containers
Name
Container ID
Image
Port(s)
CPU (%)
Last sta
Actions
spark-master
de6bcaa435fa
bitnami/sp
8080:8080
0.2%
17 minu
spark-worker1
50359ab3b60f
bitnami/sp
0.2%
17 minu
Terminal
[I 2025-05-06 07:58:14.394 ServerApp] jupyter_lsp | extension was successfully loaded.
[I 2025-05-06 07:58:14.403 ServerApp] jupyter_server_nathjax | extension was successfully loaded.
[I 2025-05-06 07:58:14.406 ServerApp] jupyter_server_terminals | extension was successfully loaded.
[I 2025-05-06 07:58:14.423 LabApp] JupyterLab extension loaded from /opt/conda/lib/python3.11/site-packages/jupyterlab
[I 2025-05-06 07:58:14.423 LabApp] JupyterLab application directory is /opt/conda/share/jupyter/lab
[I 2025-05-06 07:58:14.424 LabApp] Extension Manager is 'pypi'.
[I 2025-05-06 07:58:14.435 ServerApp] jupyterlab | extension was successfully loaded.
[I 2025-05-06 07:58:14.448 ServerApp] jupyterlab_git | extension was successfully loaded.
Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions.
https://jupyter-notebook.readthedocs.io/en/latest/migrate_to_notebook7.html
Windo...
Engine running
RAM 3.42 GB CPU 0.67% Disk: 8.48 GB used (limit 1006.85 GB)
Terminal
v4.40.0

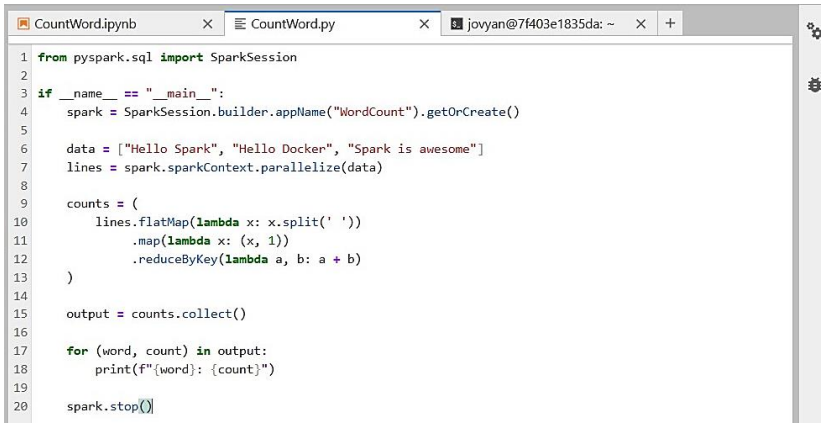

Ask Gordon BETA
Containers
Images
Volumes
Builds
Docker Hub
Docker Scout
Extensions
Containers
Give feedback
View all your running containers and applications. Learn more
Search
Only show running containers
Name
Container ID
Image
Port(s)
CPU (%)
Last sta
Actions
spark-master
de6bcaa435fa
bitnami/sp
8080:8080
0.2%
18 minu
spark-worker1
50359ab3b60f
bitnami/sp
0.24%
17 minu
Terminal
[I 2025-05-06 07:58:16.411 ServerApp] Skipped non-installed server(s): bash-language-server, dockerfile-language-server-nodejs, javascript-typescript-langserver, jedi-language-server, julia-language-server, pyright, python-language-server, python-lsp-server, r-languageserver, sql-language-server, texlab, typescript-language-server, unified-language-server, vscode-css-languageserver-bin, vscode-html-languageserver-bin, vscode-json-languageserver-bin, yaml-language-server
0.00s - Debugger warning: It seems that frozen modules are being used, which may
0.00s - make the debugger miss breakpoints. Please pass -Xfrozen_modules=off
0.00s - to python to disable frozen modules.
0.00s - Note: Debugging will proceed. Set PYDEVD_DISABLE_FILE_VALIDATION=1 to disable this validation.
HTTPServerRequest(protocol='http', host='127.0.0.1:8888', method='GET', uri='/lab/api/settings?1746518353140', version='HTTP/1.1', remote_ip='172.18.0.1')
Traceback (most recent call last):
File "/opt/conda/lib/python3.11/site-packages/tornado/web.py", line 1786, in _execute
result = await result
^^^^^^^^^^^^
tornado.iostream.StreamClosedError: Stream is closed
[W 2025-05-06 07:59:17.395 LabApp] The extension "nbdime-jupyterlab" is outdated.
[I 2025-05-06 07:59:17.395 LabApp] Build is up to date
Windo...
Engine running
RAM 3.43 GB CPU 0.00% Disk: 8.48 GB used (limit 1006.85 GB)
Terminal
v4.40.0

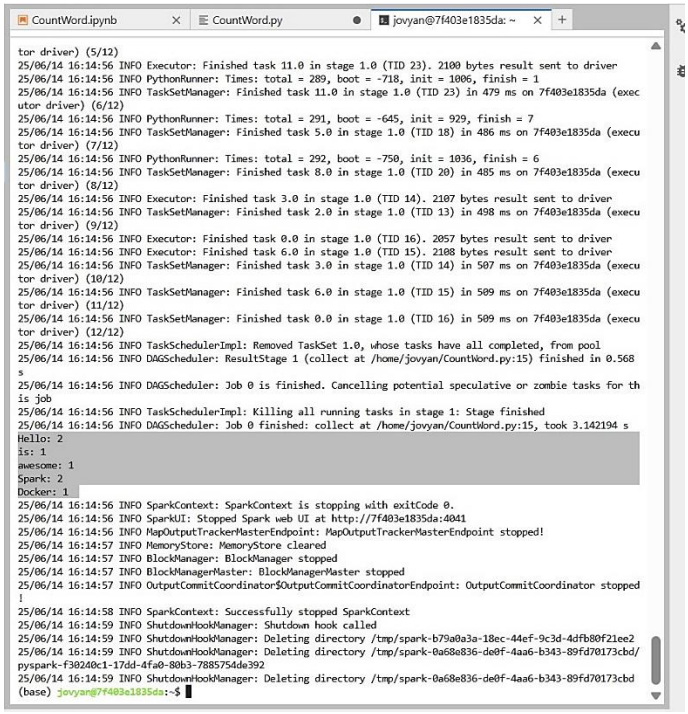

Ask Gordon BETA
Containers
Images
Volumes
Builds
Docker Hub
Docker Scout
Extensions
Containers
Give feedback
View all your running containers and applications. Learn more
Search
Only show running containers
Name
Container ID
Image
Port(s)
CPU (%)
Last sta
Actions
spark-master
de6bcaa435fa
bitnami/sp
8080:8080
0.22%
19 minu
spark-worker1
50359ab3b60f
bitnami/sp
0.22%
18 minu
spark-worker2
d9c085f582f2
bitnami/sp
0.22%
17 minu
vibrant_rhodes
0d395916ffa6
jupyter/all-
4040:4040
Show all ports (2)
0.01%
4 minute
Showing 4 items
Walkthroughs
Terminal
https://jupyter-notebook.readthedocs.io/en/latest/migrate_to_notebook7.html
Please note that updating to Notebook 7 might break some of your extensions.
[I 2025-05-06 07:58:14.465 ServerApp] nbclassic | extension was successfully loaded.
[I 2025-05-06 07:58:14.635 ServerApp] nbdime | extension was successfully loaded.
[I 2025-05-06 07:58:14.642 ServerApp] notebook | extension was successfully loaded.
[I 2025-05-06 07:58:14.643 ServerApp] Serving notebooks from local directory: /home/jovyan
[I 2025-05-06 07:58:14.643 ServerApp] Jupyter Server 2.8.0 is running at:
[I 2025-05-06 07:58:14.643 ServerApp] http://0d395916ffa6:8888/lab?token=fdb30ce84e74fdabcf337e2dab8a183f6b4923c87fa6ea77
Windo...
Engine running
RAM 3.40 GB CPU 3.41% Disk: 8.48 GB used (limit 1006.85 GB)
Terminal
v4.40.0


Launcher
Notebook
Python 3 (ipykernel)
R
Console
Python 3 (ipykernel)
R
Other
Terminal
Text File
Markdown File
Python File
R File
Show Contextual Help
Would you like to receive official Jupyter news?
Please read the privacy policy.
Open privacy policy
Yes
No
Launcher


CountWord.ipynb
Code
Open in...
Python 3 (ipykernel)
[1]:
from pyspark.sql import SparkSession
# Buat SparkSession
spark = SparkSession.builder.appName("WordCount").getOrCreate()
sc = spark.sparkContext
# Data
textData = ["Hello Spark", "Hello Docker", "Spark is awesome", "Docker makes Spark easy"]
rdd = sc.parallelize(textData)
# Word count
wordCounts = (
rdd.flatMap(lambda line: line.split(" "))
.map(lambda word: (word, 1))
.reduceByKey(lambda a, b: a + b)
)
# Tampilkan hasil
wordCounts.collect()
[1]:
[('Hello', 2),
('is', 1),
('awesome', 1),
('Spark', 3),
('Docker', 2),
('makes', 1),
('easy', 1)]


CountWord.ipynb
Code
Open in...
Python 3 (ipykernel)
[4]:
from pyspark.sql import SparkSession
# Inisialisasi Spark
spark = SparkSession.builder.appName("WordCount").getOrCreate()
# Data contoh dalam memori
data = [
"Apache Spark is a unified analytics engine",
"Spark can run on Hadoop, Apache Mesos, Kubernetes",
"Spark is awesome for big data processing"
]
rdd = spark.sparkContext.parallelize(data)
# Word Count
counts = (
rdd.flatMap(lambda x: x.split(' '))
.map(lambda x: (x, 1))
.reduceByKey(lambda a, b: a + b)
.sortBy(lambda x: x[1], ascending=False)
)
# Tampilkan 10 kata paling sering muncul
counts.take(10)
[4]:
[('Spark', 3),
('is', 2),
('Apache', 2),
('Mesos,', 1),
('awesome', 1),
('run', 1),
('Kubernetes', 1),
('processing', 1),
('Hadoop,', 1),
('analytics', 1)]


CountWord.ipynb
CountWord.py
jovyan@7f403e1835da: ~
from pyspark.sql import SparkSession
if __name__ == "__main__":
spark = SparkSession.builder.appName("WordCount").getOrCreate()
data = ["Hello Spark", "Hello Docker", "Spark is awesome"]
lines = spark.sparkContext.parallelize(data)
counts = (
lines.flatMap(lambda x: x.split(' '))
.map(lambda x: (x, 1))
.reduceByKey(lambda a, b: a + b)
)
output = counts.collect()
for (word, count) in output:
print(f"{word}: {count}")
spark.stop()


CountWord.ipynb
CountWord.py
jovyan@7f403e1835da: ~
tor driver) (5/12)
25/06/14 16:14:56 INFO Executor: Finished task 11.0 in stage 1.0 (TID 23). 2100 bytes result sent to driver
25/06/14 16:14:56 INFO PythonRunner: Times: total = 289, boot = -718, init = 1006, finish = 1
25/06/14 16:14:56 INFO TaskSetManager: Finished task 11.0 in stage 1.0 (TID 23) in 479 ms on 7f403e1835da (exec
utor driver) (6/12)
25/06/14 16:14:56 INFO PythonRunner: Times: total = 291, boot = -645, init = 929, finish = 7
25/06/14 16:14:56 INFO TaskSetManager: Finished task 5.0 in stage 1.0 (TID 18) in 486 ms on 7f403e1835da (execu
tor driver) (7/12)
25/06/14 16:14:56 INFO PythonRunner: Times: total = 292, boot = -750, init = 1036, finish = 6
25/06/14 16:14:56 INFO TaskSetManager: Finished task 8.0 in stage 1.0 (TID 20) in 485 ms on 7f403e1835da (execu
tor driver) (8/12)
25/06/14 16:14:56 INFO Executor: Finished task 3.0 in stage 1.0 (TID 14). 2107 bytes result sent to driver
25/06/14 16:14:56 INFO TaskSetManager: Finished task 2.0 in stage 1.0 (TID 13) in 498 ms on 7f403e1835da (execu
tor driver) (9/12)
25/06/14 16:14:56 INFO Executor: Finished task 0.0 in stage 1.0 (TID 16). 2057 bytes result sent to driver
25/06/14 16:14:56 INFO Executor: Finished task 6.0 in stage 1.0 (TID 15). 2108 bytes result sent to driver
25/06/14 16:14:56 INFO TaskSetManager: Finished task 3.0 in stage 1.0 (TID 14) in 507 ms on 7f403e1835da (execu
tor driver) (10/12)
25/06/14 16:14:56 INFO TaskSetManager: Finished task 6.0 in stage 1.0 (TID 15) in 509 ms on 7f403e1835da (execu
tor driver) (11/12)
25/06/14 16:14:56 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 16) in 509 ms on 7f403e1835da (execu
tor driver) (12/12)
25/06/14 16:14:56 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
25/06/14 16:14:56 INFO DAGScheduler: ResultStage 1 (collect at /home/jovyan/CountWord.py:15) finished in 0.568
s
25/06/14 16:14:56 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for th
is job
25/06/14 16:14:56 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
25/06/14 16:14:56 INFO DAGScheduler: Job 0 finished: collect at /home/jovyan/CountWord.py:15, took 3.142194 s
Hello: 2
is: 1
awesome: 1
Spark: 2
Docker: 1
25/06/14 16:14:56 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/06/14 16:14:56 INFO SparkUI: Stopped Spark web UI at http://7f403e1835da:4041
25/06/14 16:14:56 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/06/14 16:14:57 INFO MemoryStore: MemoryStore cleared
25/06/14 16:14:57 INFO BlockManager: BlockManager stopped
25/06/14 16:14:57 INFO BlockManagerMaster: BlockManagerMaster stopped
25/06/14 16:14:57 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped
!
25/06/14 16:14:58 INFO SparkContext: Successfully stopped SparkContext
25/06/14 16:14:59 INFO ShutdownHookManager: Shutdown hook called
25/06/14 16:14:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-b79a0a3a-18ec-44ef-9c3d-4dfb80f21ee2
25/06/14 16:14:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-0a68e836-de0f-4aa6-b343-89fd70173cbd/
pyspark-f30240c1-17dd-4fa0-80b3-7885754de392
25/06/14 16:14:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-0a68e836-de0f-4aa6-b343-89fd70173cbd
(base) jovyan@7f403e1835da:~$