Big Data

Doker (Jobsheet 7)

Dosen :

M. Hasyim Ratsanjani, S.Kom., M.Kom.

Dibuat Oleh :

Nurhaliza Anindya Putri 2241720016

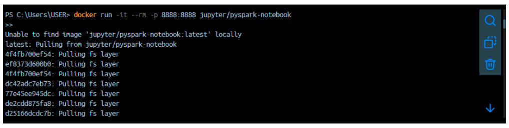
Kelas : TI – 3D

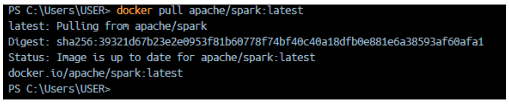
POLITEKNIK NEGERI MALANG

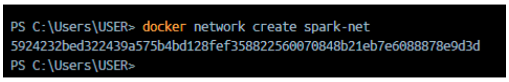
2025

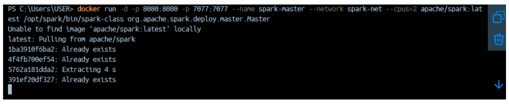
1. Implementasi Python Spark (di Dockers)

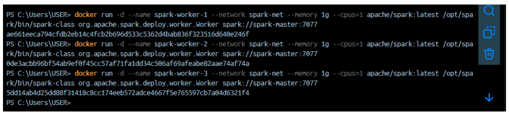
No	Gambar	Deskripi

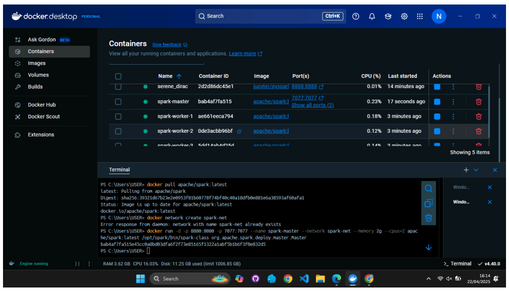
	
	`PS C:\Users\USER> docker run -it --rm -p 8888:8888 jupyter/pyspark-notebook`
	`>>`
	`Unable to find image 'jupyter/pyspark-notebook:latest' locally`
	`latest: Pulling from jupyter/pyspark-notebook`
	`4f4fb700ef54: Pulling fs layer`
	`ef8373d600b0: Pulling fs layer`
	`4f4fb700ef54: Pulling fs layer`
	`dc42adc7eb73: Pulling fs layer`
	`77e45ee945dc: Pulling fs layer`
	`de2cdd875fa8: Pulling fs layer`
	`d25166dcdc7b: Pulling fs layer`	

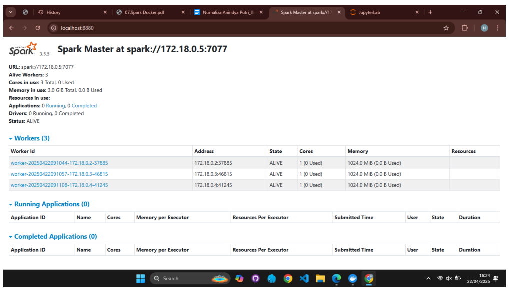
	
	`PS C:\Users\USER> docker pull apache/spark:latest`
	`latest: Pulling from apache/spark`
	`Digest: sha256:39321d67b23e2e0953f81b60778f74bf40c40a18dfb0e881e6a38593af60afa1`
	`Status: Image is up to date for apache/spark:latest`
	`docker.io/apache/spark:latest`
	`PS C:\Users\USER>`
		Pull Image Spark Resmi

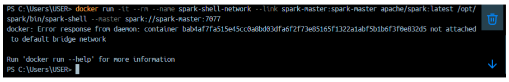
	
	`PS C:\Users\USER> docker network create spark-net`
	`5924232bed322439a575b4bd128fef358822560070848b21eb7e6088878e9d3d`
	`PS C:\Users\USER>`
		Buat docker network

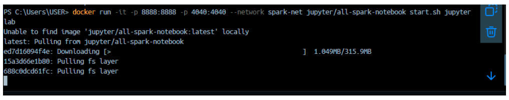
	
	`PS C:\Users\USER> docker run -d -p 8080:8080 -p 7077:7077 --name spark-master --network spark-net --cpus=2 apache/spark:lat`
	`est /opt/spark/bin/spark-class org.apache.spark.deploy.master.Master`
	`Unable to find image 'apache/spark:latest' locally`
	`latest: Pulling from apache/spark`
	`1ba3910f6ba2: Already exists`
	`4f4fb700ef54: Already exists`
	`5762a181dda2: Extracting 4 s`
	`391ef20df327: Already exists`
		Menjalankan Spark Master

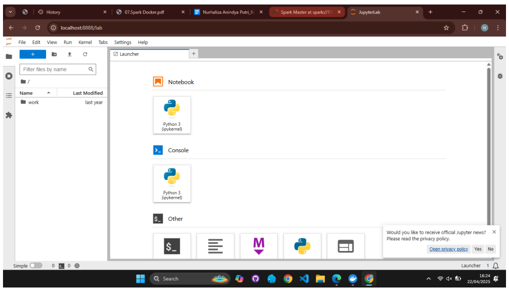
	
	`PS C:\Users\USER> docker run -d --name spark-worker-1 --network spark-net --memory 1g --cpus=1 apache/spark:latest /opt/spa`
	`rk/bin/spark-class org.apache.spark.deploy.worker.Worker spark://spark-master:7077`
	`ae661eeca794cfdb2eb14c4fcb2b696d533c5362d4bab836f323516d640e246f`
	`PS C:\Users\USER> docker run -d --name spark-worker-2 --network spark-net --memory 1g --cpus=1 apache/spark:latest /opt/spa`
	`rk/bin/spark-class org.apache.spark.deploy.worker.Worker spark://spark-master:7077`
	`0de3acbb96bf54ab9ef0f45cc57af71fa1dd34c506af69afeabe82aae74af74a`
	`PS C:\Users\USER> docker run -d --name spark-worker-3 --network spark-net --memory 1g --cpus=1 apache/spark:latest /opt/spa`
	`rk/bin/spark-class org.apache.spark.deploy.worker.Worker spark://spark-master:7077`
	`5dd14ab4d25dd88f31418c8cc174eeb572adce4667f5e765597cb7a04d6321f4`
	`PS C:\Users\USER>`
		Menjalankan Spark Worker

	
	 docker desktop PERSONAL
	Search Ctrl+K
	Ask Gordon BETA
	Containers
	Images
	Volumes
	Builds
	Docker Hub
	Docker Scout
	Extensions
	Give feedback
	Terminal
	Windo...
	Windo...
	Engine running
	Terminal
	v4.40.0
	Containers
	View all your running containers and applications. Learn more
	Name · Container ID · Image · Port(s) · CPU (%) · Last started · Actions
	serene_dirac · 2d2d86dc45e1 · jupyter/pyspar · 8888:8888 · 0.01% · 14 minutes ago
	spark-master · bab4af7fa515 · apache/spark:l · 7077:7077 Show all ports (2) · 0.23% · 17 seconds ago
	spark-worker-1 · ae661eeca794 · apache/spark:l · · 0.18% · 3 minutes ago
	spark-worker-2 · 0de3acbb96bf · apache/spark:l · · 0.12% · 3 minutes ago
	spark-worker-3 · 5dd14ab4d25d · apache/spark:l · · 0.14% · 3 minutes ago
	Showing 5 items
	`PS C:\Users\USER> docker pull apache/spark:latest`
	`latest: Pulling from apache/spark`
	`Digest: sha256:39321d67b23e2e0953f81b60778f74bf40c40a18dfb0e881e6a38593af60afa1`
	`Status: Image is up to date for apache/spark:latest`
	`docker.io/apache/spark:latest`
	`PS C:\Users\USER> docker network create spark-net`
	`Error response from daemon: network with name spark-net already exists`
	`PS C:\Users\USER> docker run -d -p 8880:8080 -p 7077:7077 --name spark-master --network spark-net --memory 2g --cpus=2 apac`
	`he/spark:latest /opt/spark/bin/spark-class org.apache.spark.deploy.master.Master`
	`bab4af7fa515e45cc0a8bd03dfa6f2f73e85165f1322a1abf5b1b6f3f0e832d5`
	`PS C:\Users\USER>`
	RAM 3.62 GB CPU 16.03% Disk: 11.25 GB used (limit 1006.85 GB)
	16:14 22/04/2025
		Hasil menggunakan 3 worker

	 History 07.Spark Docker.pdf Nurhaliza Anindya Putri_B Spark Master at spark://17 JupyterLab localhost:8880 3.5.5 **Spark Master at spark://172.18.0.5:7077** **URL:** spark://172.18.0.5:7077 **Alive Workers:** 3 **Cores in use:** 3 Total, 0 Used **Memory in use:** 3.0 GiB Total, 0.0 B Used **Resources in use:** **Applications:** 0 Running, 0 Completed **Drivers:** 0 Running, 0 Completed **Status:** ALIVE **Workers (3)** Worker Id · Address · State · Cores · Memory · Resources worker-20250422091044-172.18.0.2-37885 · 172.18.0.2:37885 · ALIVE · 1 (0 Used) · 1024.0 MiB (0.0 B Used) worker-20250422091057-172.18.0.3-46815 · 172.18.0.3:46815 · ALIVE · 1 (0 Used) · 1024.0 MiB (0.0 B Used) worker-20250422091108-172.18.0.4-41245 · 172.18.0.4:41245 · ALIVE · 1 (0 Used) · 1024.0 MiB (0.0 B Used) **Running Applications (0)** Application ID · Name · Cores · Memory per Executor · Resources Per Executor · Submitted Time · User · State · Duration **Completed Applications (0)** Application ID · Name · Cores · Memory per Executor · Resources Per Executor · Submitted Time · User · State · Duration 16:24 22/04/2025	Mengakses Spark Web UI

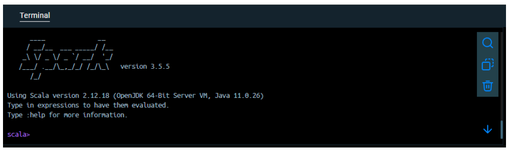
	
	`PS C:\Users\USER> docker run -it --rm --name spark-shell-network --link spark-master:spark-master apache/spark:latest /opt/`
	`spark/bin/spark-shell --master spark://spark-master:7077`
	`docker: Error response from daemon: container bab4af7fa515e45cc0a8bd03dfa6f2f73e85165f1322a1abf5b1b6f3f0e832d5 not attached`
	` to default bridge network`
	
	`Run 'docker run --help' for more information`
	`PS C:\Users\USER>`
		Menjalankan Spark Shell

	
	`PS C:\Users\USER> docker run -it -p 8888:8888 -p 4040:4040 --network spark-net jupyter/all-spark-notebook start.sh jupyter`
	`lab`
	`Unable to find image 'jupyter/all-spark-notebook:latest' locally`
	`latest: Pulling from jupyter/all-spark-notebook`
	`ed7d16094f4e: Downloading [>                                                  ]  1.049MB/315.9MB`
	`15a3d66e1b80: Pulling fs layer`
	`688c0dcd61fc: Pulling fs layer`
		Menggunakan Jupyter Notebook dengan Spark



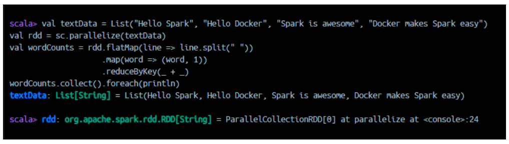
Would you like to receive official Jupyter news?
Please read the privacy policy.

akses Jupyter Notebook di: http://localhost:8888



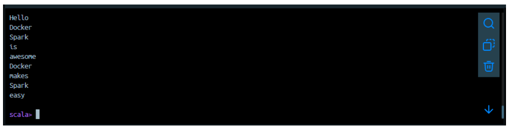
```
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.5
      /_/

Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 11.0.26)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Jalankan Spark Shell di Docker seperti contoh di atas



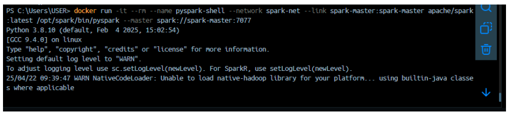
```
scala> val textData = List("Hello Spark", "Hello Docker", "Spark is awesome", "Docker makes Spark easy")
val rdd = sc.parallelize(textData)
val wordCounts = rdd.flatMap(line => line.split(" "))
                    .map(word => (word, 1))
                    .reduceByKey(_ + _)
wordCounts.collect().foreach(println)
textData: List[String] = List(Hello Spark, Hello Docker, Spark is awesome, Docker makes Spark easy)

scala> rdd: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[0] at parallelize at <console>:24
```

Ketikkan kode berikut di Spark Shell:



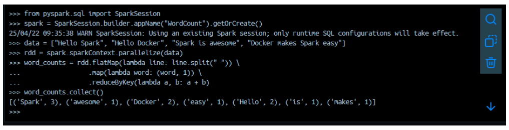
```
Hello
Docker
Spark
is
awesome
Docker
makes
Spark
easy

scala> 
```

Hasil



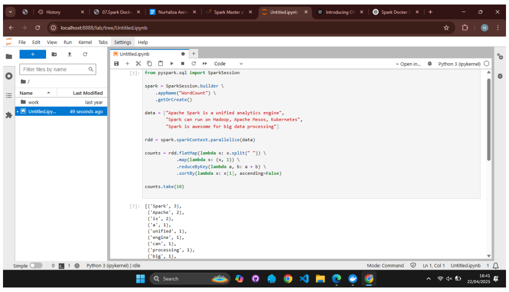
```
PS C:\Users\USER> docker run -it --rm --name pyspark-shell --network spark-net --link spark-master:spark-master apache/spark
:latest /opt/spark/bin/pyspark --master spark://spark-master:7077
Python 3.8.10 (default, Feb  4 2025, 15:02:54)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/04/22 09:39:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
```

Menggunakan PySpark (Python)



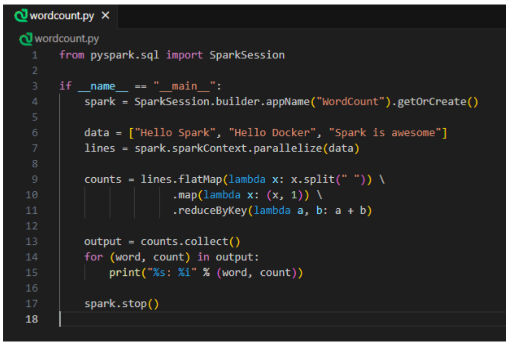
```
>>> from pyspark.sql import SparkSession
>>> spark = SparkSession.builder.appName("WordCount").getOrCreate()
25/04/22 09:35:38 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
>>> data = ["Hello Spark", "Hello Docker", "Spark is awesome", "Docker makes Spark easy"]
>>> rdd = spark.sparkContext.parallelize(data)
>>> word_counts = rdd.flatMap(lambda line: line.split(" ")) \
...                 .map(lambda word: (word, 1)) \
...                 .reduceByKey(lambda a, b: a + b)
>>> word_counts.collect()
[('Spark', 3), ('awesome', 1), ('Docker', 2), ('easy', 1), ('Hello', 2), ('is', 1), ('makes', 1)]
>>>
```

Ketikkan kode Python berikut:



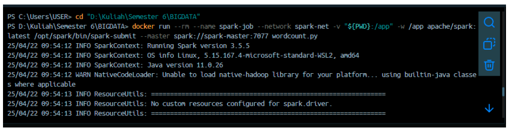
```
[3]: from pyspark.sql import SparkSession

     spark = SparkSession.builder \
         .appName("WordCount") \
         .getOrCreate()

     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)

     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)

     counts.take(10)

[3]: [('Spark', 3),
      ('Apache', 2),
      ('is', 2),
      ('a', 1),
      ('unified', 1),
      ('engine', 1),
      ('can', 1),
      ('processing', 1),
      ('big', 1),
```

Menggunakan Jupyter Notebook



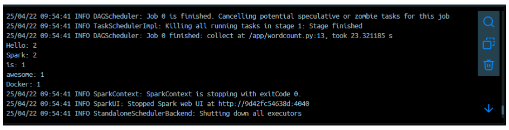
```python
from pyspark.sql import SparkSession

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("%s: %i" % (word, count))

    spark.stop()

```

Menjalankan Program sebagai Script
Buat file wordcount.py dengan isi berikut:



```
PS C:\Users\USER> cd "D:\Kuliah\Semester 6\BIGDATA"
PS D:\Kuliah\Semester 6\BIGDATA> docker run --rm --name spark-job --network spark-net -v "${PWD}:/app" -w /app apache/spark:
latest /opt/spark/bin/spark-submit --master spark://spark-master:7077 wordcount.py
25/04/22 09:54:12 INFO SparkContext: Running Spark version 3.5.5
25/04/22 09:54:12 INFO SparkContext: OS info Linux, 5.15.167.4-microsoft-standard-WSL2, amd64
25/04/22 09:54:12 INFO SparkContext: Java version 11.0.26
25/04/22 09:54:12 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
25/04/22 09:54:13 INFO ResourceUtils: ==============================================================
25/04/22 09:54:13 INFO ResourceUtils: No custom resources configured for spark.driver.
25/04/22 09:54:13 INFO ResourceUtils: ==============================================================
```

Jalankan script, jangan lupa juga mendifinisikan network spark-net



```
25/04/22 09:54:41 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/22 09:54:41 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
25/04/22 09:54:41 INFO DAGScheduler: Job 0 finished: collect at /app/wordcount.py:13, took 23.321185 s
Hello: 2
Spark: 2
is: 1
awesome: 1
Docker: 1
25/04/22 09:54:41 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/22 09:54:41 INFO SparkUI: Stopped Spark web UI at http://9d42fc54638d:4040
25/04/22 09:54:41 INFO StandaloneSchedulerBackend: Shutting down all executors
```

Program-program di atas akan menghasilkan output seperti: