NAMA: HERTIN NURHAYATI

KELAS :SIB 3E NIM : 2241760025

Interaksi dengan Spark di Lingkungan Windows Menggunakan Docker

Dalam praktikum ini kita akan menjalankan Apache Spark di Windows menggunakan Docker dan mencoba membuat job sederhana dengan berbagai macam alternatif cara.

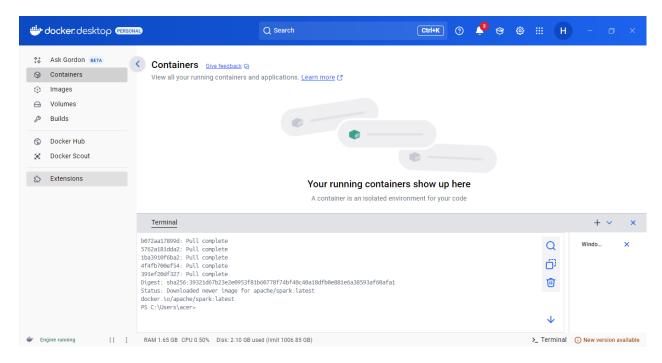
Prasyarat

- 1. Windows 10/11 (64-bit) dengan versi Pro, Enterprise, atau Education
- 2. Docker Desktop untuk Windows diinstal dan berjalan
- 3. WSL 2 (Windows Subsystem for Linux versi 2) diaktifkan

Langkah-langkah 1. Pull Image Spark Resmi

docker pull apache/spark:latest

```
f937e0a2086c: Pull complete
0f3083818c14: Pull complete
d3c7b6bd77aa: Pull complete
4d9bb71a5e54: Pull complete
b072aa17899d: Pull complete
5762a181dda2: Pull complete
1ba3910f6ba2: Pull complete
4d4fb700ef54: Pull complete
391ef20df327: Pull complete
Digest: sha256:39321d67b23e2e0953f81b60778f74bf40c40a18dfb0e881e6a38593af60afa1
Status: Downloaded newer image for apache/spark:latest
docker.io/apache/spark:latest
PS C:\Users\Acer>
```



2. Menjalankan Spark Master

Sebelumnya buat docker network sebagai berikut

PS C:\Users\Acer> docker network create spark-net 00568fcfd04abf0f4a9cf5d98aa568e7230e2dac924bee0bd70257c10ca3b3db

Hasil:

PS C:\Users\acer> docker network create spark-net 88c832cb4a110097733feb0d85cb2474c4b0b9d525f90f4acb1e893c287530ef PS C:\Users\acer>

U

Kemudian jalankan spark-master dalam network tersebut

Hasil:

PS C:\Users\acer> docker run -d -p 8080:8080 -p 7077:7077 --name spark-master --network spark-net -m 2g --cpus=2 apache/spark:latest /opt/spark/bin/spark-class org.apache.deploy.master.Master 2829950b488e710274778e4e59a0f488d873d43a9bbe89434644a3a48d39f64b
PS C:\Users\acer>

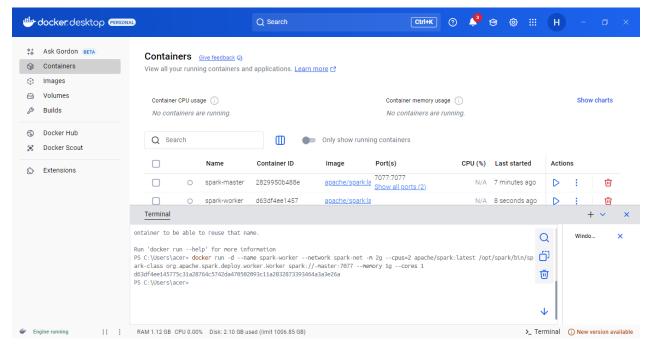
Kita alokasikan resource untuk memastikan tidak kekurangan resource dalam menjalankan job.

3. Menjalankan Spark Worker

PS C:\Users\Acer> docker run -d --name spark-worker --network spark-net -m 2g --cpus=2 apache/spark:latest /opt/spark/bin /spark-class org.apache.spark.deploy.worker.Worker spark://spark-master:7077 --memory 1g --cores 1 58c18a90d6c65b112cc2e72ed7cad42c991bc8bb2710f459e2bce6b53168f9a8

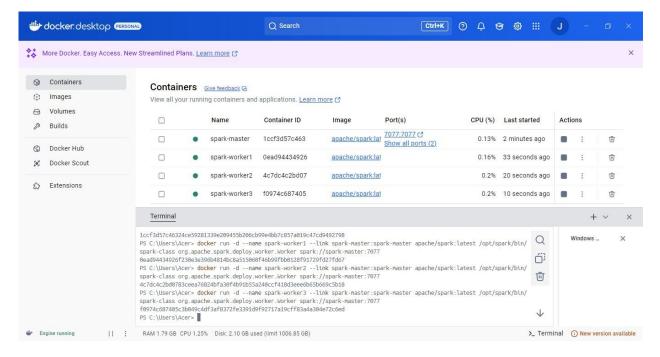


団



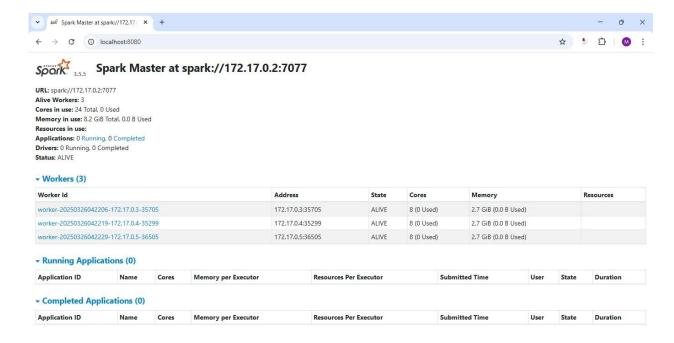
Kita perlu alokasikan resource misalnya 2G memori dan 2 core CPU

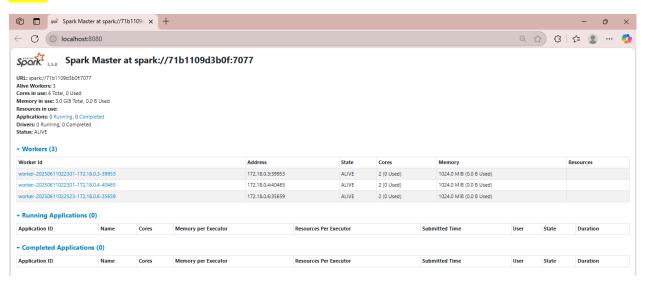
jalankan perintah di atas beberapa kali dengan nama yang berbeda untuk membuat beberapa worker. Misalnya spark-worker1, spark-worker2, dan seterusnya Contoh menggunakan 3 worker.



4. Mengakses Spark Web UI

http://localhost:8080





5. Menjalankan Spark Shell

PS C:\Users\Acer> docker run -it --rm --name spark-shell --network spark-net --link spark-master:spark-master apache/spark:latest /opt/spark/bin/spark-shell --master spark://spark-master:7077

Setting default log level to "WARN".

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

25/03/26 04:52:31 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cla sses where applicable

Spark context Web UI available at http://03f37d151387:4040

Spark context available as 'sc' (master = spark://spark-master:7077, app id = app-20250326045233-0000).

Spark session available as 'spark'.

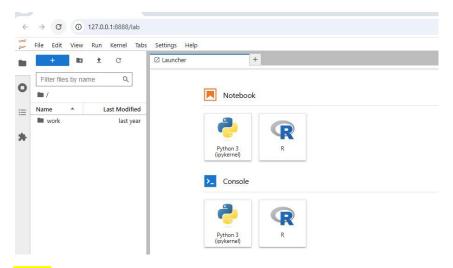
Welcome to

Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 11.0.26)

6. Menggunakan Jupyter Notebook dengan Spark

```
PS C:\Users\Acer> docker run -it -p 8888:8888 -p 4040:4040 --network spark-net jupyter/all-spark-notebook
Entered start.sh with args: jupyter lab
Running hooks in: /usr/local/bin/start-notebook.d as uid: 1000 gid: 100
Done running hooks in: /usr/local/bin/start-notebook.d
Running hooks in: /usr/local/bin/before-notebook.d as uid: 1000 gid: 100
Sourcing shell script: /usr/local/bin/before-notebook.d/spark-config.sh
Done running hooks in: /usr/local/bin/before-notebook.d
```

Setelah itu, akses Jupyter Notebook di: http://localhost:8888



Hasil:

Jalankan jupyter

```
PS C:\Users\acer> docker logs bold_lalande
>>
Entered start.sh with args: jupyter lab
Running hooks in: /usr/local/bin/start-notebook.d as uid: 0 gid: 0
Done running hooks in: /usr/local/bin/start-notebook.d
Granting jovyan passwordless sudo rights!
Running hooks in: /usr/local/bin/before-notebook.d as uid: 0 gid: 0
Sourcing shell script: /usr/local/bin/before-notebook.d/spark-config.sh
Done running hooks in: /usr/local/bin/before-notebook.d
Running as jovyan: jupyter lab

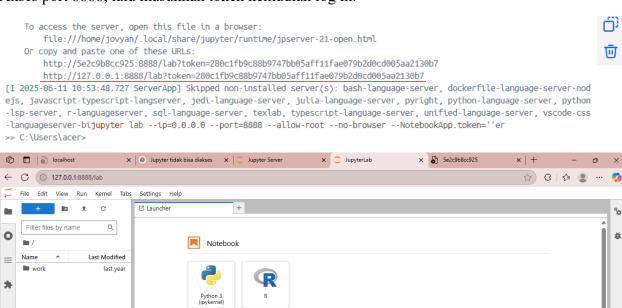
[I 2025-05-27 06:50:47.157 ServerApp] Package jupyterlab took 0.0000s to import
```

Akses port 8888, lalu masukkan token kemudian log in.

>_ Console

\$_ Other

\$_



Launcher 1 🗘

Untuk menghentikan container:

Simple 0 8 0 @

docker stop spark-master spark-worker

docker rm spark-master spark-worker

```
PS C:\Users\Acer> docker stop spark-master spark-worker1 spark-worker2 spark-master spark-worker1 spark-worker1 spark-worker2 PS C:\Users\Acer> docker rm spark-master spark-worker1 spark-worker2 spark-master spark-worker1 spark-worker1 spark-worker1 spark-worker1 spark-worker2 PS C:\Users\Acer> [
```

Contoh Program Word Count dengan Spark di Docker

Berikut adalah contoh program Word Count (menghitung kemunculan kata) menggunakan Apache Spark yang bisa dijalankan di lingkungan Docker:

Cara 1: Menggunakan Spark Shell

- 1. Jalankan Spark Shell di Docker seperti contoh di atas
- 2. Ketikkan kode berikut di Spark Shell:

```
// Buat RDD dari koleksi teks
val textData = List("Hello Spark", "Hello Docker", "Spark is awesome", "Docker makes
Spark easy") val rdd = sc.parallelize(textData)

// Lakukan word count val wordCounts = rdd.flatMap(line =>
line.split(" ")) // Split menjadi kata-kata .map(word =>
(word, 1)) // Map setiap kata ke tuple (word, 1)
.reduceByKey(_ + _) // Reduce dengan menjumlahkan

// Tampilkan hasil
wordCounts.collect().foreach(println)
```

scala> res1: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[3] at reduceByKey at <console>:24 scala> wordCounts.collect().foreach(println) Hello Spark Hello Docker Spark is awesome Docker makes Spark

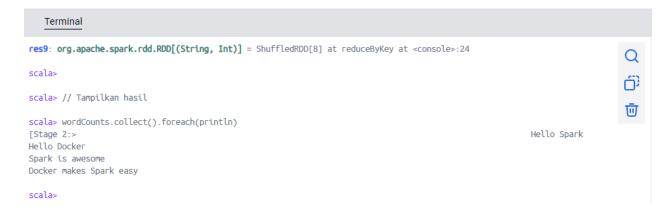
Untuk keluar dari spark-shell gunakan:

System.exit(0)

Hasil:

Untuk menjalankan Spark Shell:

Setelah diketikkan muncul output.



Cara 2: Menggunakan PySpark (Python)

1. Jalankan PySpark Shell di Docker:

Dalam command juga terdapat definisi network juga.

```
PS C:\Users\Acer> 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).

25/03/26 05:09:36 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cla sses where applicable

Welcome to
```

Terminal

```
PS C:\Users\acer> docker run -it --rm --name pyspark-shell --network spark-net --link spark-master:spark-master apache/spar
                                                                                                                             Q
k:latest /opt/spark/bin/pyspark --master spark://spark-master:7077
                                                                                                                             Ę)
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.
                                                                                                                             呵
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/06/11 11:56:24 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
Welcome to
Using Python version 3.8.10 (default, Feb 4 2025 15:02:54)
Spark context Web UI available at http://9757cba22ec3:4040
Spark context available as 'sc' (master = spark://spark-master:7077, app id = app-20250611115636-0000).
SparkSession available as 'spark'.
>>>
```

2. Ketikkan kode Python berikut:

Untuk keluar dari pyspark-shel menggunakan: exit()

Hasil:

```
Using Python version 3.8.10 (default, Feb 4 2025 15:02:54)

Spark context Web UI available at http://c9529144f0cc:4040

Spark context available as 'sc' (master = spark://spark-master:7077, app id = app-20250618053323-0000).

SparkSession available as 'spark'.

>>> from pyspark.sql import SparkSession

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

25/06/18 05:34:51 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.

>>> data = ["Hello Spark", "Hello Docker", "Spark is awesome"]

>>> 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()
```

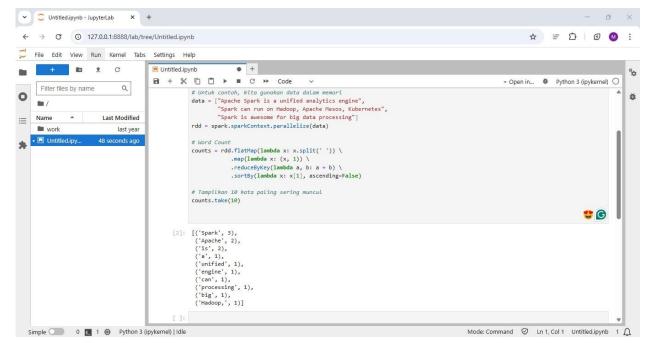
Cara 3: Menggunakan Jupyter Notebook

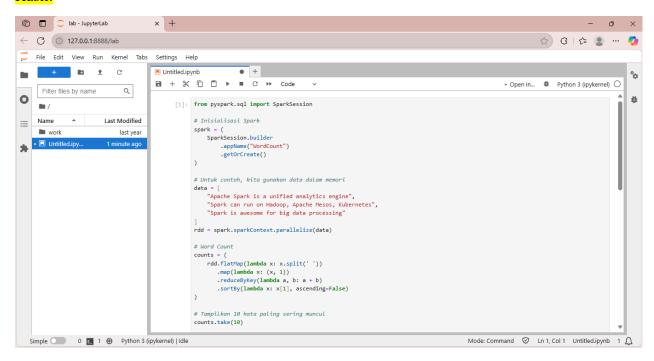
Jika Anda menggunakan Jupyter Notebook (seperti di container jupyter/all-spark-notebook):

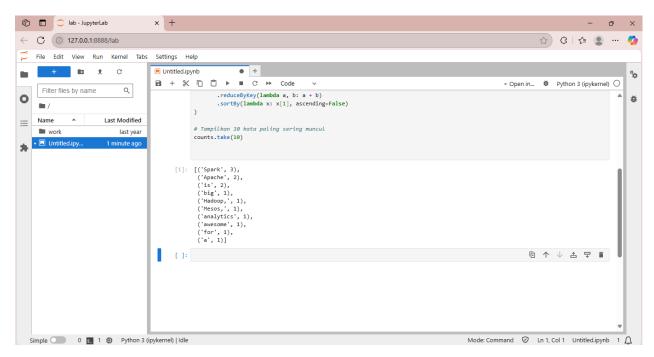
from pyspark.sql import SparkSession

Inisialisasi Spark

```
spark = SparkSession.builder \
  .appName("WordCount") \
  .getOrCreate()
# Baca file teks (jika ingin membaca dari file)
# text file =
spark.sparkContext.textFile("hdfs://.../input.txt")
# Untuk contoh, kita gunakan data 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)
```







Menjalankan Program sebagai Script

1. Buat file wordcount.py dengan isi berikut:

```
from pyspark.sql import SparkSession
if name == " main ":
  spark = SparkSession.builder.appName("WordCount").getOrCreate()
  # Untuk versi membaca file
  # lines = spark.read.text("input.txt").rdd.map(lambda r: r[0])
  # Untuk versi data contoh
  data = ["Hello Spark", "Hello Docker", "Spark is
             lines = spark.sparkContext.parallelize(data)
awesome"]
  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()
```

```
Kernel Tabs Settings Help
                       CountWord.ipynb
                                                 × ■ CountWord.py
                                                                             X ■ jovyan@7f403e183Sda.... X
           C
                        1 from pyspark.sql import SparkSession
              Q
s by name
                        3 if __name__ == "__main__":
                              spark = SparkSession.builder.appMame("WordCount").getOrCreate()
        Last Modified
                              data = ["Hello Spark", "Hello Docker", "Spark is awesome"]
            last year
                              lines = spark.sparkContext.parallelize(data)
      18 minutes ago
                              counts = (
       3 minutes ago
                                  lines.flatMap(lambda x: x.split(' '))
                       11
                                       .map(lambda x: (x, 1))
                       12
                                       .reduceByKey(lambda a, b: a + b)
                      13
14
15
16
17
18
19
                              output = counts.collect()
                              for (word, count) in output:
                                  print(f'(word): {count}")
```

2. Jalankan jika sudah memastikan skrip di atas benar: - Simpan file CountWord.py - Jalankan: spark-submit CountWord.py

