

Nama : Fitria Nur Sholikhah

NIM : 224176004

Kelas : SIB-3E

Mata Kuliah: Big Data

---

## 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.

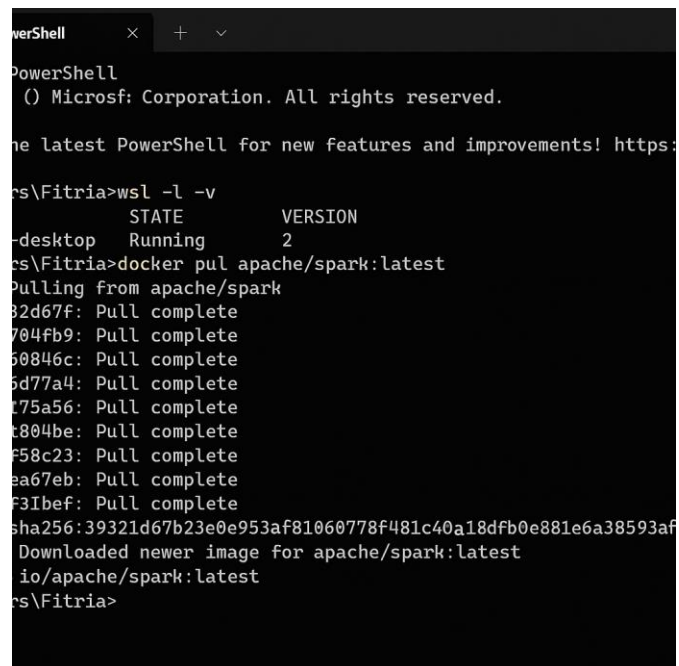
### Prasyarat

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

##### Hasil:



```
PowerShell
() Microsoft Corporation. All rights reserved.

Get the latest PowerShell for new features and improvements! https://aka.ms/PowerShellLatest

C:\Fitria>wsl -l -v
          STATE      VERSION
-----
-docker Running         2
C:\Fitria>docker pull apache/spark:latest
Pulling from apache/spark
32d67f: Pull complete
704fb9: Pull complete
50846c: Pull complete
5d77a4: Pull complete
175a56: Pull complete
t804be: Pull complete
f58c23: Pull complete
ea67eb: Pull complete
f31bef: Pull complete
sha256:39321d67b23e0e953af81060778f481c40a18dfb0e881e6a38593af
Downloaded newer image for apache/spark:latest
C:\Fitria>
```

## Menjalankan Master

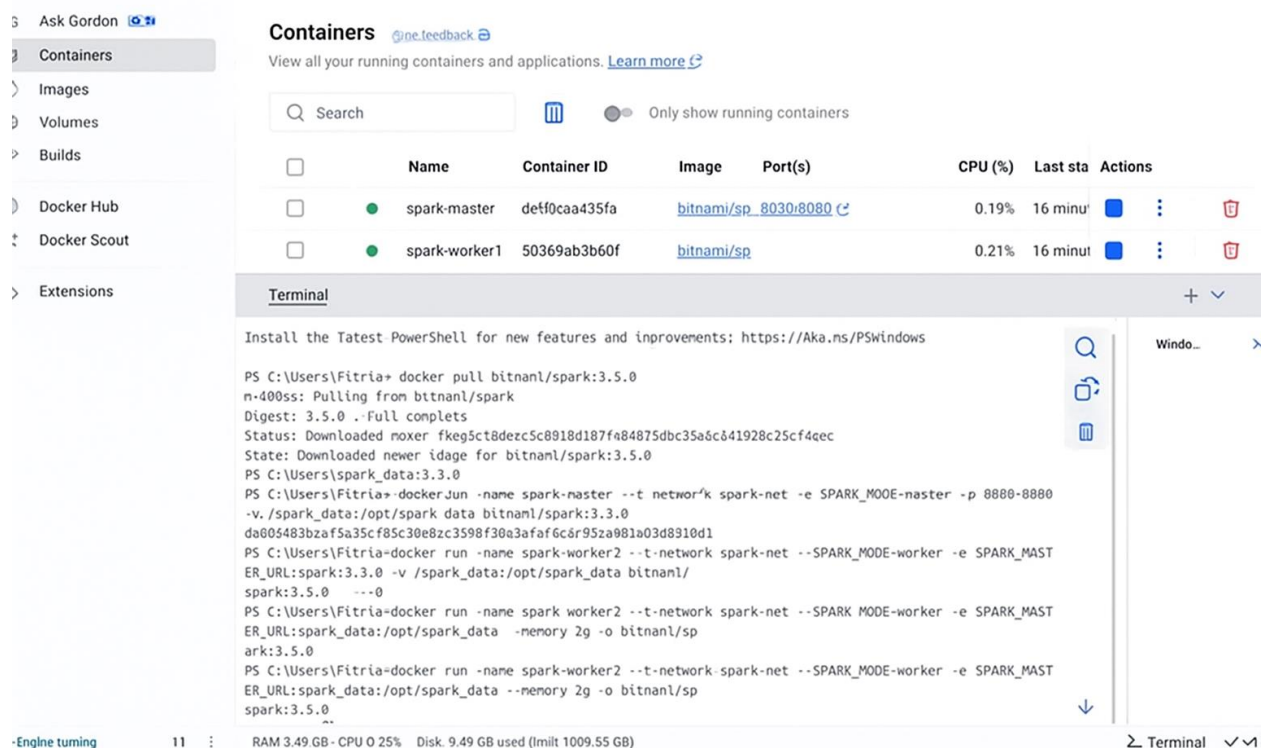
```
docker run -d --name spark-master --network spark-net -e SPARK_MODE=master -p 8080:8080 v
~/spark_data:/opt/spark_data bitnami/spark:3.5.0
```

## Menjalankan Worker

```
docker run -d --name spark-worker1 --network spark-net -e SPARK_MODE=worker -e
SPARK_MASTER_URL=spark://spark-master:7077 -v ~/spark_data:/opt/spark_data --memory 2g
--cpus 2 bitnami/spark:3.5.0
```

```
docker run -d --name spark-worker2 --network spark-net -e SPARK_MODE=worker -e
SPARK_MASTER_URL=spark://spark-master:7077 -v ~/spark_data:/opt/spark_data --memory 2g
--cpus 2 bitnami/spark:3.5.0
```

### Hasil:



The screenshot shows the Docker Desktop interface. On the left is a sidebar with navigation options: Ask Gordon, Containers (selected), Images, Volumes, Builds, Docker Hub, Docker Scout, and Extensions. The main area is titled 'Containers' and shows a table of running containers. Below the table is a terminal window displaying the commands used to pull the Spark image and run the master and worker containers.

Name	Container ID	Image	Port(s)	CPU (%)	Last sta	Actions
spark-master	def0caa435fa	bitnami/spark:3.5.0	8030:8080	0.19%	16 minu	[Stop] [Restart] [Delete]
spark-worker1	50369ab3b60f	bitnami/spark:3.5.0		0.21%	16 minut	[Stop] [Restart] [Delete]

```
PS C:\Users\Fitria> docker pull bitnami/spark:3.5.0
m-400ss: Pulling from bitnami/spark
Digest: 3.5.0 - Full complets
Status: Downloaded moxer fkeg5ct8dezc5c8918d187fa84875dbc35a6c841928c25cf4aec
State: Downloaded newer idage for bitnami/spark:3.5.0
PS C:\Users\spark_data> docker run -d --name spark-master --network spark-net -e SPARK_MODE=master -p 8080:8080
-v ~/spark_data:/opt/spark_data bitnami/spark:3.5.0
da005483b2af5a35cf85c30e8zc3598f30a3afaf6c8r952a081a03d8310d1
PS C:\Users\Fitria> docker run -d --name spark-worker1 --network spark-net -e SPARK_MASTER_URL=spark://spark-master:7077
-v ~/spark_data:/opt/spark_data bitnami/spark:3.5.0 --memory 2g --cpus 2
PS C:\Users\Fitria> docker run -d --name spark-worker2 --network spark-net -e SPARK_MASTER_URL=spark://spark-master:7077
-v ~/spark_data:/opt/spark_data bitnami/spark:3.5.0 --memory 2g --cpus 2
```

## Menjalankan Spark Shell lalu Menggunakan Jupyter Notebook dengan Spark

```
docker run -it -p 8888:8888 -p 4040:4040 --network spark-net --user root -e
SPARK_MASTER=spark://spark-master:7077 -e GRANT_SUDO=yes -v ~/spark_data:/opt/spark_data
jupyter/all-spark-notebook
```

### Hasil:

Ask Gordon

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
<input type="checkbox"/>	spark-master	deebcaa425fa	bitnami/spark	6030-8080	0.18%	17 min	<div></div> <div></div> <div></div>
<input type="checkbox"/>	spark-worker1	50569ab5b60f	bitnami/spark		0.18%	16 min	<div></div> <div></div> <div></div>

Terminal

STER\_URLS:spark://spark-master:7077 --v -v /spark\_data:/opt/spark\_data --memory 2g --cpus 2 bitnami/spark.3.5  
-B  
#914e77ff1a1vdeafcb8edc6d663718b7d8a31849dc11e436a  
PS C:\Users\JF1\1a>docker for -i1 -B 8888:3883 -p 4348:8158 network spark\_not --user root < GPARK\_HASTE  
n-spark://spark-master:7077 < POMAT\_SUPHoves -> /spark\_data/opt/spark\_data jupyter/all-spark-notebook  
unates latest from jupyter/all-spark-notebook:latest totally  
Pulling f5581: Pulling PS  
83997778ad084: Pulling PS  
0fct3566d58a: Pulling PS  
f83a118aaf588: Pulling PS  
0896a24cf5d: Pulling PS  
9f969858b952: Pulling PS  
af808823ef6: Pulling PS  
a87f6db84b5b: Pulling PS  
57482c78f76d: Pulling PS  
ae6bde39a518: Pulling PS  
85fac39a34c1: Pulling PS  
f7bacf8a377f1: Pulling PS  
f688afutcs15: Pulling PS  
PS C:\Users\JF1\1a>

Engine running | | RAM 9.41 GB, CPU 0.25% Disk 9.48 GB used (limit 1006.85 GB)

Ask Gordon

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
<input type="checkbox"/>	spark-master	deebcaa435fa	bitnami/spark	8080-8080	0.21%	17 min	<div></div> <div></div> <div></div>
<input type="checkbox"/>	spark-worker1	50359ab3b60f	bitnami/spark		0.25%	17 min	<div></div> <div></div> <div></div>

Terminal

15ab06de1b80: Pull complete  
68b6dc8dcfc: Pull complete  
ed7d1094f4e: Pull complete  
Digest: sha256:b03bae29d34779ac96d6eb4834af083891f7726c9a9a76f6b0d1f8678629  
Status: Downloaded newer image for jupyter/all-spark-notebook:latest  
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 jupyter passwordless sudo rights!  
Running hooks in: /usr/local/bin/before-notebook.d as uid: 0 gid: 0  
Sourcing shell scripts: /usr/local/bin/before-notebook.d/spark-config.sh  
Done running hooks in: /usr/local/bin/before-notebook.d  
Running as JupyterLab jupyter lab  
[I 2025-05-06 07:58:13.299 ServerApp] Package jupyterlab took 0.000s to import  
[I 2025-05-06 07:58:13.357 ServerApp] Package jupyter\_lab took 0.037s to import  
[I 2025-05-06 07:58:13.388 ServerApp] A \_jupyter\_server\_extension\_paths function was not found in jupyterlab. Instead, a \_jupyter\_server\_extensions\_paths function was found and will be used for now. This function name will be deprecated in future releases of Jupyter Server.  
[I 2025-05-06 07:58:13.382 ServerApp] Package jupyter\_server\_terminals took 0.000s to import  
[I 2025-05-06 07:58:13.383 ServerApp] Package jupyter\_server\_terminals took 0.022s to import

Engine running | | RAM 9.42 GB, CPU 0.00% Disk 9.48 GB used (limit 1006.85 GB)

Ask Gordon

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
<input type="checkbox"/>	spark-master	deebcaa435fa	bitnami/spark		0.24%	18 min	<div></div> <div></div> <div></div>
<input type="checkbox"/>	spark-worker1	50359ab3b60f	bitnami/spark		0.24%	17 min	<div></div> <div></div> <div></div>

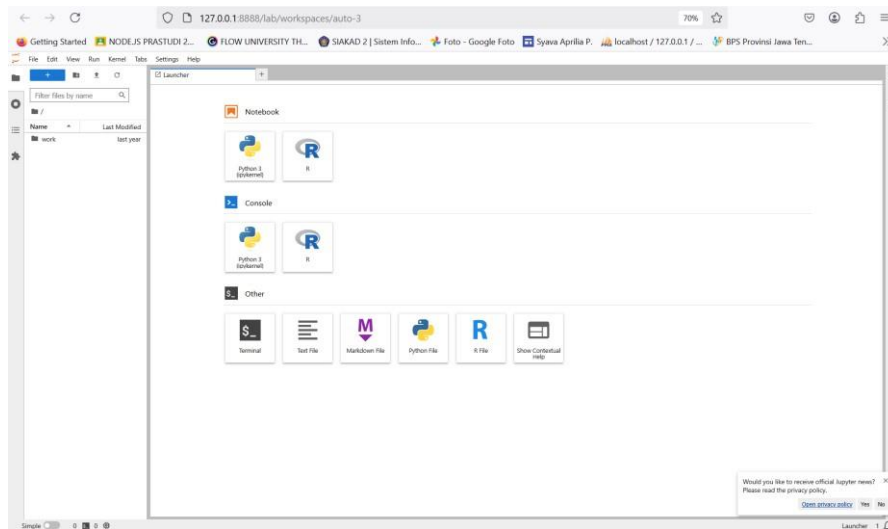
Terminal

IT 2025-05-06 07:50:42.411 serverpp1 Skipped non-installed server(s): bash-language-server, dockerfile-ls  
npuase-server-nodejs, javascript-typescript-languageserver, java-language-server, java-language-server, terin  
hi, python-language-server, python-lsp-server, n-languageserver, tal-language-server, kswikt, typescript-l  
anguage-server, whifled language-server, vscode-cts-languageserver-bin, vscode-kernel-languageserver-bin, vc  
code-jupyter-languageserver-bin, part-language-server  
0.005 - Debugger warning: It seems that trace modules are being used, which may  
0.005 - make the debugger-dist breakpoints. Please pass --no-debugger-module-leor  
0.005 - To python is visible from modules.  
0.005 - Next - debugging will pretend, See DISABLE\_FTL\_VALIDATION: FTL- to disable this validation.  
HTTPServerRequest(protocol: 'http', host=137.0.0.1:8080, method='GET', url: '/lab/api/settings/1744518  
8323431, werokan, RTTF, lial, contng, len 172, 18.8:41)  
Traceback (most recent call last):  
File: /opt/conda/lib/python3.9/site-packages/tornado/web.py, line 1786, in \_execute  
result = next result  
\*\*\*\*\* stream is closed  
tornado.tornado.StreamClosedError: stream is closed  
[W 2025-05-06 07:35:17.353 LabApp] Build is up to date  
IT 2025-05-06 07:58:47.881 LabApp Build up to date

Engine running | | RAM 5.46 GB, CPU 0.10% Disk 8.48 GB used (limit 1008.45 GB)

[illegible]

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



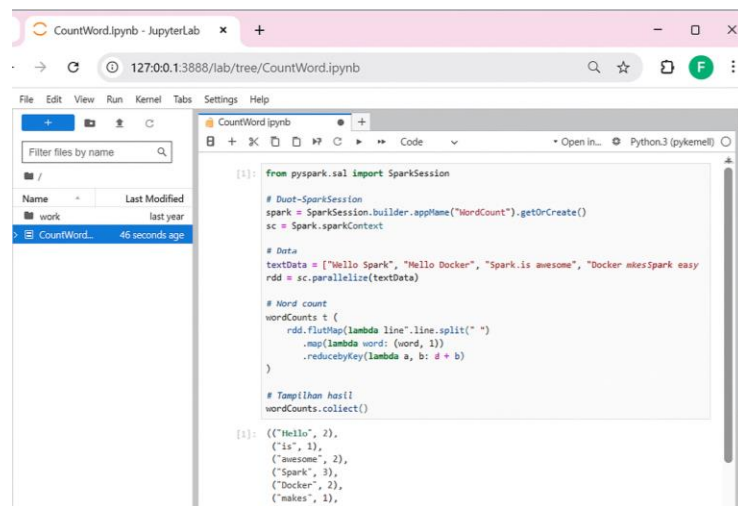
## Contoh Program Word Count dengan Spark di Docker

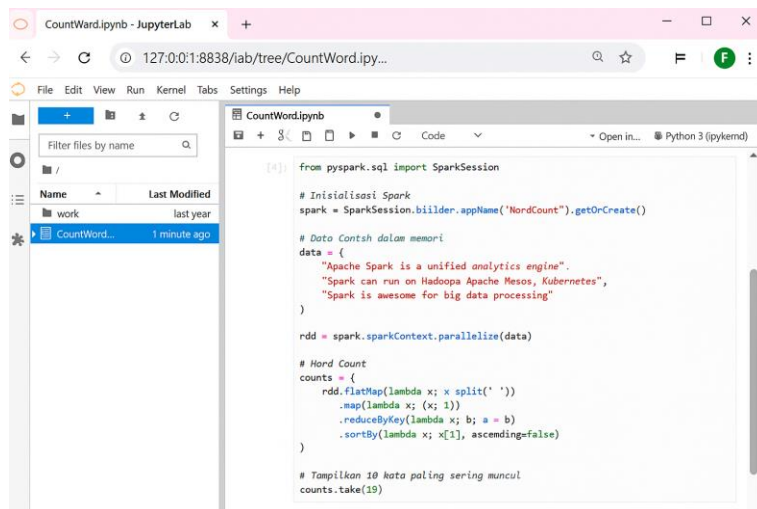
Berikut adalah contoh program Word Count (menghitung kemunculan kata) menggunakan Apache Spark yang bisa dijalankan di lingkungan Docker:

### Cara 3: Menggunakan Jupyter Notebook

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

#### Hasil:

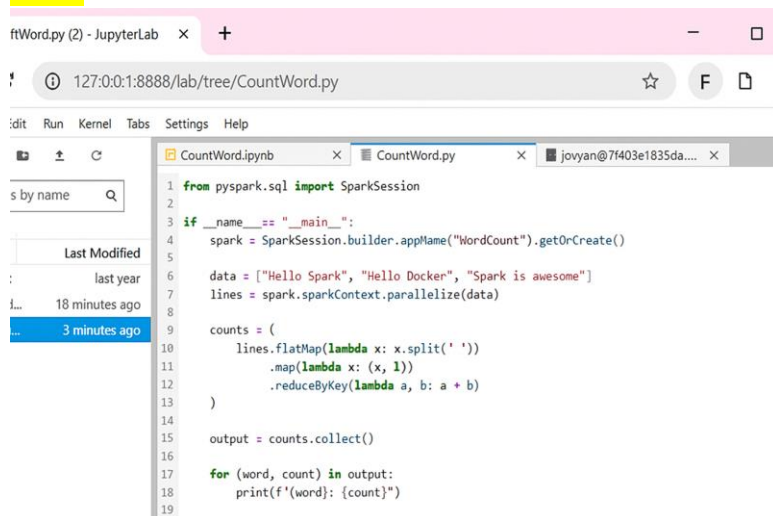




## Menjalankan Program sebagai Script

1. Buat file wordcount.py dengan isi berikut:

**Hasil:**



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

**Hasil:**

[illegible]