

Mata kuliah : Big Data
Materi : Spark
Nama : Nurhaliza Anindya Putri
NIM : 2241720016
No. Presensi : 14
Kelas : TI-3D

1. Instalasi Apache Spark

- Silakan gunakan Cluster Hadoop dari hasil kuis sebelumnya di VBox kelompok Anda.
- Lakukan instalasi Apache Spark.
- Unduh versi terbaru Spark dari situs resmi atau gunakan wget dari dalam namenode vbox Anda:
 - wget
<https://downloads.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz>
 - (Ubah versi sesuai kebutuhan)

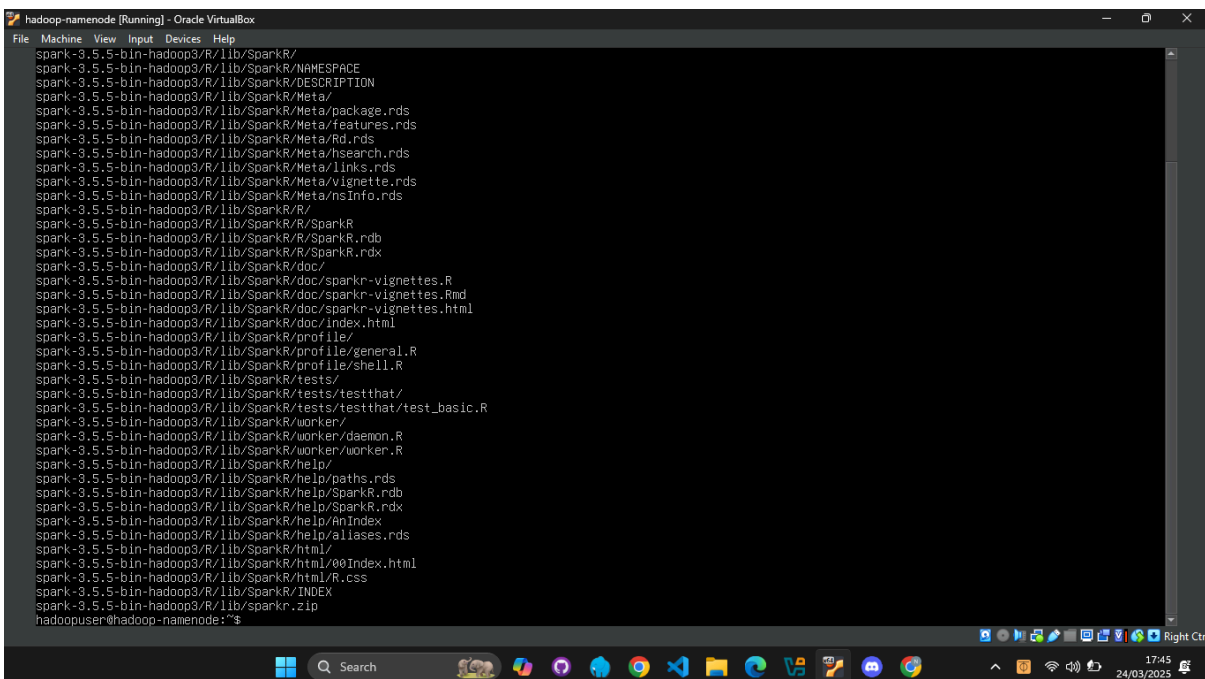
```
hadoopuser@hadoop-namenode:~$ wget https://downloads.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz
--2025-03-24 10:25:10-- https://downloads.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz
Resolving downloads.apache.org (downloads.apache.org)... 2a01:4f8:10a:39da::2, 2a01:4f9:3a:2c57::2, 88.99.208.237, ...
Connecting to downloads.apache.org (downloads.apache.org)|2a01:4f8:10a:39da::2|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 400724056 (382M) [application/x-gzip]
Saving to: 'spark-3.5.5-bin-hadoop3.tgz'

spark-3.5.5-bin-hadoop3.tgz      100%[=====] 382.16M  414KB/s   in 18m 44s

2025-03-24 10:43:55 (348 KB/s) - 'spark-3.5.5-bin-hadoop3.tgz' saved [400724056/400724056]

hadoopuser@hadoop-namenode:~$
```

- Ekstrak dan pindahkan direktori
 - tar -xvzf spark-3.4.1-bin-hadoop3.tgz



```
hadoop-namenode [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
spark-3.5.5-bin-hadoop3/R/lib/SparkR/
spark-3.5.5-bin-hadoop3/R/lib/SparkR/NAMESPACE
spark-3.5.5-bin-hadoop3/R/lib/SparkR/DESCRIPTION
spark-3.5.5-bin-hadoop3/R/lib/SparkR/Meta/
spark-3.5.5-bin-hadoop3/R/lib/SparkR/Meta/package.rds
spark-3.5.5-bin-hadoop3/R/lib/SparkR/Meta/features.rds
spark-3.5.5-bin-hadoop3/R/lib/SparkR/Meta/Rd.rds
spark-3.5.5-bin-hadoop3/R/lib/SparkR/Meta/hsearch.rds
spark-3.5.5-bin-hadoop3/R/lib/SparkR/Meta/links.rds
spark-3.5.5-bin-hadoop3/R/lib/SparkR/Meta/vignette.rds
spark-3.5.5-bin-hadoop3/R/lib/SparkR/Meta/nsInfo.rds
spark-3.5.5-bin-hadoop3/R/lib/SparkR/R/
spark-3.5.5-bin-hadoop3/R/lib/SparkR/R/SparkR
spark-3.5.5-bin-hadoop3/R/lib/SparkR/R/SparkR.rdb
spark-3.5.5-bin-hadoop3/R/lib/SparkR/R/SparkR.rdx
spark-3.5.5-bin-hadoop3/R/lib/SparkR/doc/
spark-3.5.5-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.R
spark-3.5.5-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.Rmd
spark-3.5.5-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.html
spark-3.5.5-bin-hadoop3/R/lib/SparkR/doc/index.html
spark-3.5.5-bin-hadoop3/R/lib/SparkR/profile/
spark-3.5.5-bin-hadoop3/R/lib/SparkR/profile/general.R
spark-3.5.5-bin-hadoop3/R/lib/SparkR/profile/shell.R
spark-3.5.5-bin-hadoop3/R/lib/SparkR/tests/
spark-3.5.5-bin-hadoop3/R/lib/SparkR/tests/testthat/
spark-3.5.5-bin-hadoop3/R/lib/SparkR/tests/testthat/test_basic.R
spark-3.5.5-bin-hadoop3/R/lib/SparkR/worker/
spark-3.5.5-bin-hadoop3/R/lib/SparkR/worker/daemon.R
spark-3.5.5-bin-hadoop3/R/lib/SparkR/worker/worker.R
spark-3.5.5-bin-hadoop3/R/lib/SparkR/help/
spark-3.5.5-bin-hadoop3/R/lib/SparkR/help/paths.rds
spark-3.5.5-bin-hadoop3/R/lib/SparkR/help/SparkR.rdb
spark-3.5.5-bin-hadoop3/R/lib/SparkR/help/SparkR.rdx
spark-3.5.5-bin-hadoop3/R/lib/SparkR/help/anIndex
spark-3.5.5-bin-hadoop3/R/lib/SparkR/help/alluses.rds
spark-3.5.5-bin-hadoop3/R/lib/SparkR/html/
spark-3.5.5-bin-hadoop3/R/lib/SparkR/html/00Index.html
spark-3.5.5-bin-hadoop3/R/lib/SparkR/html/R.css
spark-3.5.5-bin-hadoop3/R/lib/SparkR/INDEX
spark-3.5.5-bin-hadoop3/R/lib/sparkr.zip
hadoopuser@hadoop-namenode:~$
```

- sudo mv spark-3.5.5-bin-hadoop3 /opt/spark

```
hadoopuser@hadoop-namenode:~$ sudo mv spark-3.5.5-bin-hadoop3 /opt/spark
[sudo] password for hadoopuser:
hadoopuser@hadoop-namenode:~$
```

2. Konfigurasi Apache Spark

- Konfigurasi environment variables. Edit `.bashrc` atau `.profile` :
 - `nano ~/.bashrc`
- Tambahkan baris berikut:
 - `export SPARK_HOME=/opt/spark`
 - `export PATH=$SPARK_HOME/bin:$SPARK_HOME/sbin:$PATH`
 - `export`
`LD_LIBRARY_PATH=$HADOOP_HOME/lib/native:$LD_LIBRARY_PATH`
 - `export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop`
 - `export SPARK_MASTER_HOST=<IP_MASTER_NODE>`

```
# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

export SPARK_HOME=/opt/spark
export PATH=$SPARK_HOME/bin:$SPARK_HOME/sbin:$PATH
export LD_LIBRARY_PATH=$HADOOP_HOME/lib/native:$LD_LIBRARY_PATH
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export SPARK_MASTER_HOST=192.168.1.25

hadoopuser@hadoop-namenode:~$
```

- Kemudian jalankan:
 - `source ~/.bashrc`

```
hadoopuser@hadoop-namenode:~$ source ~/.bashrc
hadoopuser@hadoop-namenode:~$
```

- Konfigurasi `spark-env.sh`
- Salin template dan edit:
 - `cp /opt/spark/conf/spark-env.sh.template /opt/spark/conf/spark-env.sh`

```
hadoopuser@hadoop-namenode:~$ cp /opt/spark/conf/spark-env.sh.template /opt/spark/conf/spark-env.sh
hadoopuser@hadoop-namenode:~$
```

- `nano /opt/spark/conf/spark-env.sh`
- Tambahkan:
 - `export JAVA_HOME=$(readlink -f /usr/bin/java | sed "s:bin/java::")`
 - `export SPARK_MASTER_HOST=<IP_MASTER_NODE>`
 - `export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop`
 - `export SPARK_WORKER_CORES=2`
 - `export SPARK_WORKER_MEMORY=4g`
 - `export SPARK_DRIVER_MEMORY=2g`
 - `export SPARK_EXECUTOR_MEMORY=2g`

```
# Options for beeline
# - SPARK_BEELINE_OPTS, to set config properties only for the beeline cli (e.g. "-Dx=y")
# - SPARK_BEELINE_MEMORY, Memory for beeline (e.g. 1000M, 2G) (Default: 1G)

export JAVA_HOME=$(readlink -f /usr/bin/java | sed "s:bin/java::")
export SPARK_MASTER_HOST=192.168.1.25
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export SPARK_WORKER_CORES=2
export SPARK_WORKER_MEMORY=4g
export SPARK_DRIVER_MEMORY=2g
export SPARK_EXECUTOR_MEMORY=2g

hadoopuser@hadoop-namenode:~$ _
```

3. Menjalankan Apache Spark di Cluster Hadoop

- Jalankan Spark Master di namenode (Master Node), jalankan:
 - start-master.sh

```
hadoopuser@hadoop-namenode:~$ start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs/spark-hadoopuser-org.apache.spark.deploy.master.Master-1-hadoop-namenode.out
hadoopuser@hadoop-namenode:~$ jps
3268 ResourceManager
4233 Jps
4186 Master
3099 SecondaryNameNode
2862 NameNode
hadoopuser@hadoop-namenode:~$
```

- Buka di browser: http://<IP_MASTER>:8080

Spark Master at spark://192.168.1.25:7077

URL: spark://192.168.1.25:7077

Alive Workers: 0

Cores in use: 0 Total, 0 Used

Memory in use: 0.0 B Total, 0.0 B Used

Resources in use:

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

▼ Workers (0)

Worker Id	Address	State	Cores	Memory	Resources
-----------	---------	-------	-------	--------	-----------

▼ 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
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

- Di setiap Worker Node (data node-pastikan spark sudah setup), jalankan:
 - start-worker.sh spark://<IP_MASTER>:7077

```
hadoopuser@hadoop-namenode:~$ start-worker.sh spark://192.168.1.25:7077
starting org.apache.spark.deploy.worker.Worker, logging to /opt/spark/logs/spark-hadoopuser-org.apache.spark.deploy.worker.Worker-1-hadoop-namenode.out
hadoopuser@hadoop-namenode:~$ jps
3268 ResourceManager
4186 Master
3099 SecondaryNameNode
4267 Worker
2862 NameNode
4302 Jps
hadoopuser@hadoop-namenode:~$ _
```

4. Uji Apache Spark

- Cek apakah Spark bekerja dengan baik:
 - spark-shell --master spark://<IP_MASTER>:7077

- Atau jalankan contoh aplikasi:
 - `/opt/spark/bin/run-example SparkPi 10`
- Jika terdapat masalah atau error, silahkan diskusikan dengan kelompok Anda untuk mencari solusinya.