

LAPORAN
TUGAS PRAKTIKUM BIG DATA
(Apache Spark)



Disusun oleh:

Fitria Nur Sholikhah

2241760004

PROGRAM STUDI D-IV SISTEM INFORMASI BISNIS
JURUSAN TEKNOLOGI INFORMASI POLITEKNIK
NEGERI MALANG
2025

Tugas Praktikum

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](https://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>

1. Instalasi Apache Spark (2)

- Ekstrak dan pindahkan direktori tar -xvzf spark-3.4.1-bin-hadoop3.tgz sudo mv spark-3.4.1-bin-hadoop3 /opt/spark

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>](#)

Kemudian jalankan:

[source ~/.bashrc](#)

Hasil:

```

GNU nano 7.2 /home/hadoopuser/.bashrc
alias la='ls -A'
alias l='ls -CF'

# Add an "alert" alias for long running commands. Use like so:
# sleep 10; alert
alias alert='notify-send --urgency=low -i "${1} ${?} = 0" && echo terminal || echo error)' "${history}

# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# 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=192.168.109.158

hadoopuser@hadoop-datanode2:~$ _

```

2. Konfigurasi Apache Spark (2)

- Konfigurasi spark-env.sh
- Salin templat dan edit:

```
cp /opt/spark/conf/spark-env.sh.template /opt/spark/conf/spark-env.sh
```

```
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
```

Hasil:

```

GNU nano 7.2 /opt/spark/conf/spark-env.sh
# - SPARK_SHUFFLE_OPTS, to set config properties only for the external shuffle service (e.g. "-Dx=y")
# - SPARK_DAEMON_JAVA_OPTS, to set config properties for all daemons (e.g. "-Dx=y")
# - SPARK_DAEMON_CLASSPATH, to set the classpath for all daemons
# - SPARK_PUBLIC_DNS, to set the public dns name of the master or workers

# Options for launcher
# - SPARK_LAUNCHER_OPTS, to set config properties and Java options for the launcher (e.g. "-Dx=y")

# Generic options for the daemons used in the standalone deploy mode
# - SPARK_CONF_DIR      Alternate conf dir. (Default: ${SPARK_HOME}/conf)
# - SPARK_LOG_DIR       Where log files are stored. (Default: ${SPARK_HOME}/logs)
# - SPARK_LOG_MAX_FILES Max log files of Spark daemons can rotate to. Default is 5.
# - SPARK_PID_DIR       Where the pid file is stored. (Default: /tmp)
# - SPARK_IDENT_STRING  A string representing this instance of spark. (Default: $USER)
# - SPARK_NICENESS       The scheduling priority for daemons. (Default: 0)
# - SPARK_NO_DAEMONIZE  Run the proposed command in the foreground. It will not output a PID file.
# Options for native BLAS, like Intel MKL, OpenBLAS, and so on.
# You might get better performance to enable these options if using native BLAS (see SPARK-21305).
# - MKL_NUM_THREADS=1   Disable multi-threading of Intel MKL
# - OPENBLAS_NUM_THREADS=1 Disable multi-threading of OpenBLAS

# 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=$(readlink -f /usr/bin/java | sed "s:bin/java::")
export SPARK_MASTER_HOST=192.168.109.158
export HADOOP_CONF_DIR=$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-datanode2:~$

```

3. Menjalankan Apache Spark di Cluster Hadoop

- Jalankan Spark Master di namenode (Master Node), jalankan: `start-master.sh`
- Buka di browser: `http://<IP_MASTER>:8080`
- Di setiap Worker Node (data node-pastikan spark sudah setup), jalankan:
`start-worker.sh spark://<IP_MASTER>:7077`

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`

Hasil:

Not secure | 192.168.109.158:8080

Spark

3.5.5

Spark Master at spark://192.168.109.158:7077

URL: spark://192.168.109.158:7077

Alive Workers: 3

Cores in use: 6 Total, 0 Used

Memory in use: 12.0 GiB Total, 0.0 B Used

Resources in use:

Applications: 0 Running, 2 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory	Resources
worker-20250424080845-192.168.109.132-33705	192.168.109.132:33705	ALIVE	2 (0 Used)	4.0 GiB (0.0 B Used)	
worker-20250424080914-192.168.109.193-33905	192.168.109.193:33905	ALIVE	2 (0 Used)	4.0 GiB (0.0 B Used)	
worker-20250424080921-192.168.109.165-34531	192.168.109.165:34531	ALIVE	2 (0 Used)	4.0 GiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Completed Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20250424081106-0001	Spark shell	6	2.0 GiB		2025/04/24 08:11:06	hadoopuser	FINISHED	59 s
app-20250424081109-0000	Spark shell	6	2.0 GiB		2025/04/24 08:10:09	hadoopuser	FINISHED	38 s