

LAPORAN TUGAS KELOMPOK PERTEMUAN 6

BIG DATA



Oleh :

| | |
|----------------------------------|-------------------|
| AHMAD FAZA ALFAN FASHLAH | 2241720186 |
| MOCHAMMAD ZAKARO AL FAJRI | 2241720175 |
| RIO BAGAS HERMAWAN | 2241720193 |
| SONY FEBRI HARI WIBOWO | 2241720202 |

D-IV TEKNIK INFORMATIKA
JURUSAN TEKNOLOGI INFORMASI
POLITEKNIK NEGERI MALANG
2025

Instalasi Apache Spark

Langkah 1: Instalasi Apache Spark

- a) Unduh versi terbaru Spark dalam namenode vbox

```
hadoopuser@hadoop-namenode:~$ wget https://downloads.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz
```

- b) Ekstrak lalu pindahkan ke direktori /opt/spark

```
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/aliases.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:~$
```

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

Langkah 2: Konfigurasi Apache Spark

- a) Konfigurasi environment variables

```

# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

# If not running interactively, don't do anything
case $- in
    *i*) ;;
    *) return;;
esac

# don't put duplicate lines or lines starting with space in the history.
# See bash(1) for more options
HISTCONTROL=ignoreboth

# append to the history file, don't overwrite it
shopt -s histappend

# for setting history length see HISTSIZE and HISTFILESIZE in bash(1)
HISTSIZE=1000
HISTFILESIZE=2000

# check the window size after each command and, if necessary,
# update the values of LINES and COLUMNS.
shopt -s checkwinsize

# If set, the pattern ".*" used in a pathname expansion context will
# match all files and zero or more directories and subdirectories.
shopt -s globstar

# make less more friendly for non-text input files, see lesspipe(1)
[ -x /usr/bin/lesspipe ] && eval "$(SHELL=/bin/sh lesspipe)"

# set variable identifying the chroot you work in (used in the prompt below)
[ Read 117 lines ]

```

lalu menambahkan baris:

```

# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

export SPARK_HOME=/opt/spark
export SPARK_MASTER_HOST=namenode
export PATH=$SPARK_HOME/bin:$SPARK_HOME/sbin:$PATH

export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export YARN_CONF_DIR=$HADOOP_HOME/etc/hadoop

# If not running interactively, don't do anything
case $- in
    *i*) ;;
    *) return;;
esac

# don't put duplicate lines or lines starting with space in the history.
# See bash(1) for more options
HISTCONTROL=ignoreboth

# append to the history file, don't overwrite it
shopt -s histappend

# for setting history length see HISTSIZE and HISTFILESIZE in bash(1)
HISTSIZE=1000
HISTFILESIZE=2000

# check the window size after each command and, if necessary,
# update the values of LINES and COLUMNS.
shopt -s checkwinsize

# If set, the pattern ".*" used in a pathname expansion context will

```

b) Jalankan kode sebelumnya

```

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

```

c) Konfigurasi spark-env.sh.

```
#!/usr/bin/env bash

SPARK_MASTER_HOST=192.168.72.244
SPARK_WORKER_CORES=2
SPARK_WORKER_MEMORY=2g
SPARK_EXECUTOR_MEMORY=2g
HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop

#
# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements. See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# The ASF licenses this file to You under the Apache License, Version 2.0
# (the "License"); you may not use this file except in compliance with
# the License. You may obtain a copy of the License at
#
#   http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#
# This file is sourced when running various Spark programs.
# Copy it as spark-env.sh and edit that to configure Spark for your site.
#
# Options read when launching programs locally with
# ./bin/run-example or ./bin/spark-submit
# - HADOOP_CONF_DIR, to point Spark towards Hadoop configuration files
# - SPARK_LOCAL_IP, to set the IP address Spark binds to on this node
# - SPARK_PUBLIC_DNS, to set the public dns name of the driver program
```

- d) Membuat daftar worker

```
hadoopuser@hadoop-namenode:/opt/spark/conf$ nano workers
```

- e) Membuat direktori /opt/ pada tiap datanode. Contoh datanode 1:

```
[sudo] password for hadoopuser:
hadoopuser@hadoop-datanode1:~$ sudo chown hadoopuser:hadoopuser /opt/
hadoopuser@hadoop-datanode1:~$ ls -ld /opt/
drwxr-xr-x 2 hadoopuser hadoopuser 4096 Aug 27 2024 /opt/
hadoopuser@hadoop-datanode1:~$ _
```

Langkah 3: Menjalankan Apache Spark di Cluster Hadoop

- a) Jalankan Spark Master dan Workers di namenode

- master:

```
hadoopuser@hadoop-namenode:/opt/spark/conf$ start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs
he.spark.deploy.master.Master-1-hadoop-namenode.out
hadoopuser@hadoop-namenode:/opt/spark/conf$ _
```

- Worker

```
hadoopuser@hadoop-namenode:/opt/spark/conf$ start-workers.sh
192.168.72.3: starting org.apache.spark.deploy.worker.Worker, logging to /opt
puser-org.apache.spark.deploy.worker.Worker-1-hadoop-datanode1.out
192.168.72.11: starting org.apache.spark.deploy.worker.Worker, logging to /opt
puser-org.apache.spark.deploy.worker.Worker-1-hadoop-datanode3.out
192.168.72.73: starting org.apache.spark.deploy.worker.Worker, logging to /opt
puser-org.apache.spark.deploy.worker.Worker-1-hadoop-datanode2.out
hadoopuser@hadoop-namenode:/opt/spark/conf$
```

- b) Buka di cluster hadoop

Spark

3.5.5

Spark Master at spark://192.168.72.244:7077

URL: spark://192.168.72.244:7077
Alive Workers: 3
Cores in use: 6 Total, 0 Used
Memory in use: 6.0 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-20250414082031-192.168.72.3-37029 | 192.168.72.3:37029 | ALIVE | 2 (0 Used) | 2.0 GiB (0.0 B Used) | |
| worker-20250414082050-192.168.72.73-34389 | 192.168.72.73:34389 | ALIVE | 2 (0 Used) | 2.0 GiB (0.0 B Used) | |
| worker-20250414082052-192.168.72.11-36535 | 192.168.72.11:36535 | ALIVE | 2 (0 Used) | 2.0 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 (0)

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