JOBSHEET APACHE SPARK

BIG DATA

Dosen Pengampu:

Vipkas Al Hadid Firdaus,ST,. MT

Disusun Oleh:

Afinudy Ananda Pramudya	2241760096
Moch Fikri Setiawan	2241760105
Fida Cahya Sasmita	2241760001
Hertin Nurhayati	2241760025

PRODI D-lV SISTEM INFORMASI BISNIS
JURUSAN TEKNOLOGI INFORMASI
POLITEKNIK NEGERI MALANG
TAHUN 2025/2026

Instalasi Apache Spark

```
hadoopuser@hadoop-namenode:~$ spark-shell --version
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.5
      /_/

Using Scala version 2.12.18, OpenJDK 64-Bit Server VM, 1.8.0_442
Branch HEAD
Compiled by user ubuntu on 2025-02-23T20:30:46Z
Revision 7c29c664cdc9321205a98a14858aaf8daaa19db2
Url https://github.com/apache/spark
Type --help for more information.
hadoopuser@hadoop-namenode:~$ _
```
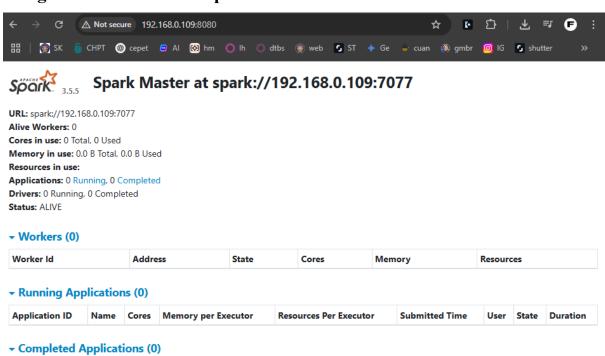
Konfigurasi Apache Spark

```
# 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=192.168.100.94
```

```
# 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}/l
# - SPARK_LOG_MAX_FILES   Max log files of Spark daemons can rotate to. Default
# - SPARK_PID_DIR         Where the pid file is stored. (Default: /tmp)
# - SPARK_IDENT_STRING    A string representing this instance of spark. (Default
# - SPARK_NICENESS        The scheduling priority for daemons. (Default: 0)
# - SPARK_NO_DAEMONIZE    Run the proposed command in the foreground. It will no
# Options for native BLAS, like Intel MKL, OpenBLAS, and so on.
# You might get better performance to enable these options if using native BLA
# - 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
# - SPARK_BEELINE_MEMORY, Memory for beeline (e.g. 1000M, 2G) (Default: 1G)

export JAVA_HOME=$(readlink -f /usr/bin/java | sed "s:bin/java::")
export SPARK_MASTER_HOST=192.168.100.94
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export SPARK_WORKER_CORES=2
export SPARK_WORKER_MEMORY=4g
export SPARK_DRIVER_MEMORY=2g
export SPARK_EXECUTOR_MEMORY=2g

[ Soft wrapping of o
```

Mengecek cluster di hadoop

Spark Master at spark://192.168.0.109:7077

URL: spark://192.168.0.109:7077
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