



DIGITAL
TALENT
SCHOLARSHIP

DIGITAL TALENT SCHOLARSHIP 2019

Big Data Analytics



digitalent.kominfo.go.id



DIGITAL
TALENT
SCHOLARSHIP

Apache Spark

Instalasi Apache Spark Standalone dan Case Study

TERBUKA
UNTUK
DISABILITAS

BREK
YOUR
LIMITS



Requirement

- EC2 Instance
- Akses ke EC2 Instance.
 - (SSH Client, Putty, WinSCP, dll)
- Anaconda





DIGITAL
TALENT
SCHOLARSHIP

Instalasi Apache Spark Standalone

Big Data

Apache Spark


TERBUKA
UNTUK
DISABILITAS

BREAK
YOUR
LIMIT



Download Apache Spark

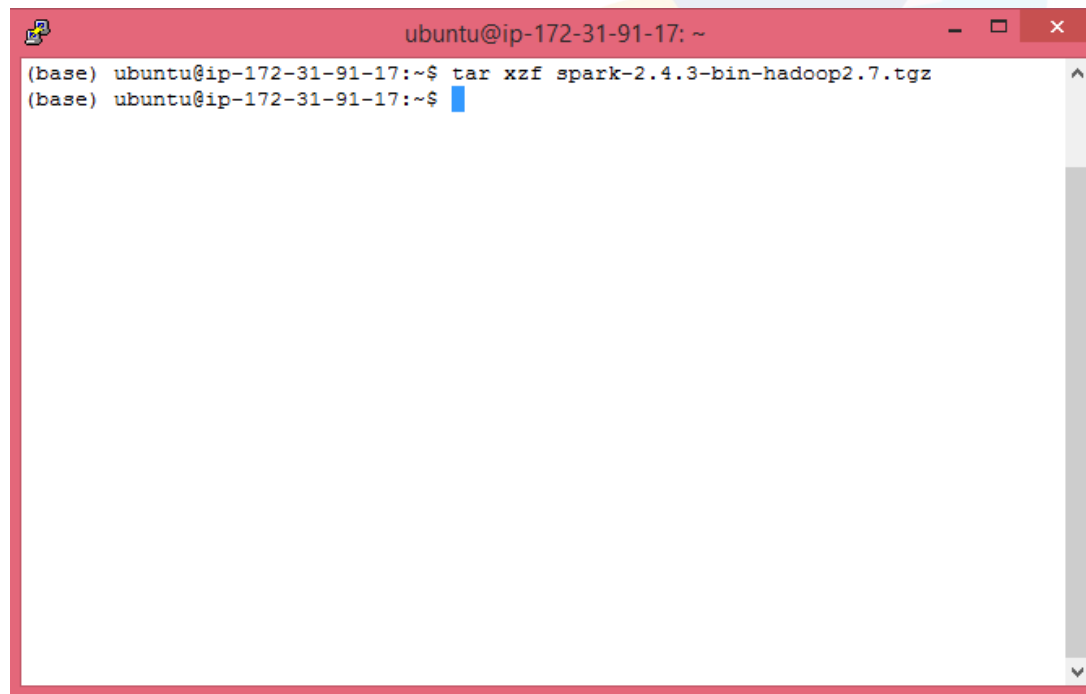
```
$ curl -O https://www-us.apache.org/dist/spark/spark-2.4.3/spark-2.4.3-bin-hadoop2.7.tgz
```



```
ubuntu@ip-172-31-91-17: ~  
(base) ubuntu@ip-172-31-91-17:~$ curl -O https://www-us.apache.org/dist/spark/spark-2.4.3/spark-2.4.3-bin-hadoop2.7.tgz  
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current  
                                 Dload  Upload   Total   Spent    Left   Speed  
100 219M  100 219M    0     0  37.5M      0  0:00:05  0:00:05 --:--:-- 38.2M  
(base) ubuntu@ip-172-31-91-17:~$
```

Extract Apache Spark

```
$ tar xzf spark-2.4.3-bin-hadoop2.7.tgz
```



```
ubuntu@ip-172-31-91-17: ~  
(base) ubuntu@ip-172-31-91-17:~$ tar xzf spark-2.4.3-bin-hadoop2.7.tgz  
(base) ubuntu@ip-172-31-91-17:~$
```

Variabel SPARK_HOME

Buka file `.bashrc` atau `.profile` dan tambahkan string berikut.

```
export SPARK_HOME=/home/ubuntu/spark-2.4.3-  
bin-hadoop2.7  
  
export  
PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin
```



Variabel SPARK_HOME (cont)

```
/home/ubuntu/.bashrc - master - Editor - WinSCP
elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
fi
fi

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
PATH=$PATH:$JAVA_HOME:$JAVA_HOME/jre/bin

# >>> conda initialize >>>
# !! Contents within this block are managed by 'conda init' !!
__conda_setup="$('/home/ubuntu/anaconda3/bin/conda' 'shell.bash' 'hook' 2> /dev/null)"
if [ $? -eq 0 ]; then
    eval "$__conda_setup"
else
    if [ -f "/home/ubuntu/anaconda3/etc/profile.d/conda.sh" ]; then
        . "/home/ubuntu/anaconda3/etc/profile.d/conda.sh"
    else
        export PATH="/home/ubuntu/anaconda3/bin:$PATH"
    fi
fi
unset __conda_setup
# <<< conda initialize <<<

export HADOOP_HOME=/home/ubuntu/hadoop-2.8.5
export LD_LIBRARY_PATH=$HADOOP_HOME/lib/native/:$LD_LIBRARY_PATH
export HADOOP_JAR=$HADOOP_HOME/share/hadoop/mapreduce
PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$HADOOP_JAR

export SPARK_HOME=/home/ubuntu/spark-2.4.3-bin-hadoop2.7
export PATH=$PATH:$SPARK_HOME/bin

Line: 1/145      Column: 1      Character: 35 (0x23)      Encoding: 1252 (ANSI - La
```


Variabel SPARK_HOME (cont)

```
$ source ~/.bashrc
```

Atau..

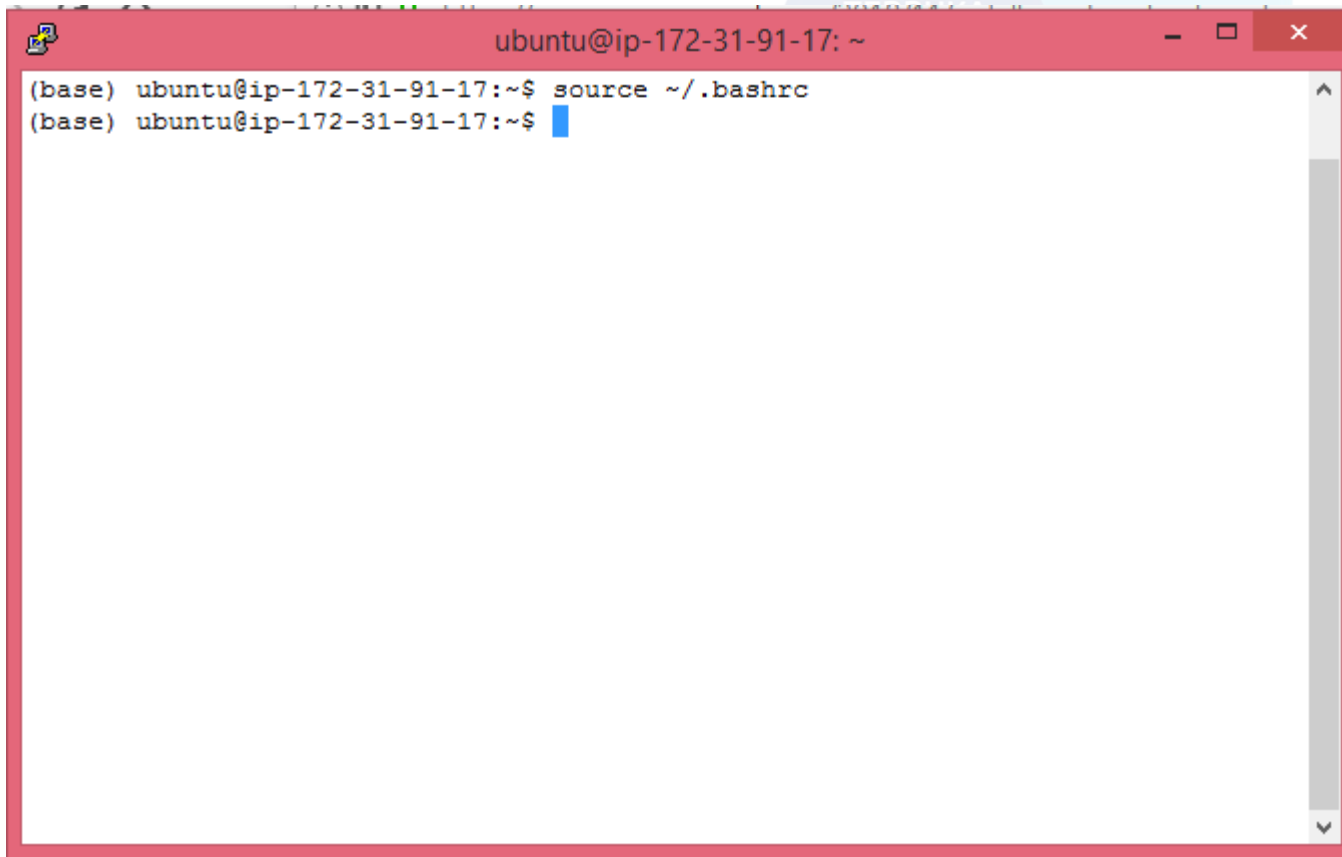
```
$ source ~/.profile
```

TERBUKA
UNTUK
DISABILITAS

BEYOND
YOUR
LIMITS!



Variabel SPARK_HOME (cont)



```
ubuntu@ip-172-31-91-17: ~  
(base) ubuntu@ip-172-31-91-17:~$ source ~/.bashrc  
(base) ubuntu@ip-172-31-91-17:~$
```

Periksa instalasi Spark

```
$ spark-shell
```

TERBUKA
UNTUK
DISABILITAS

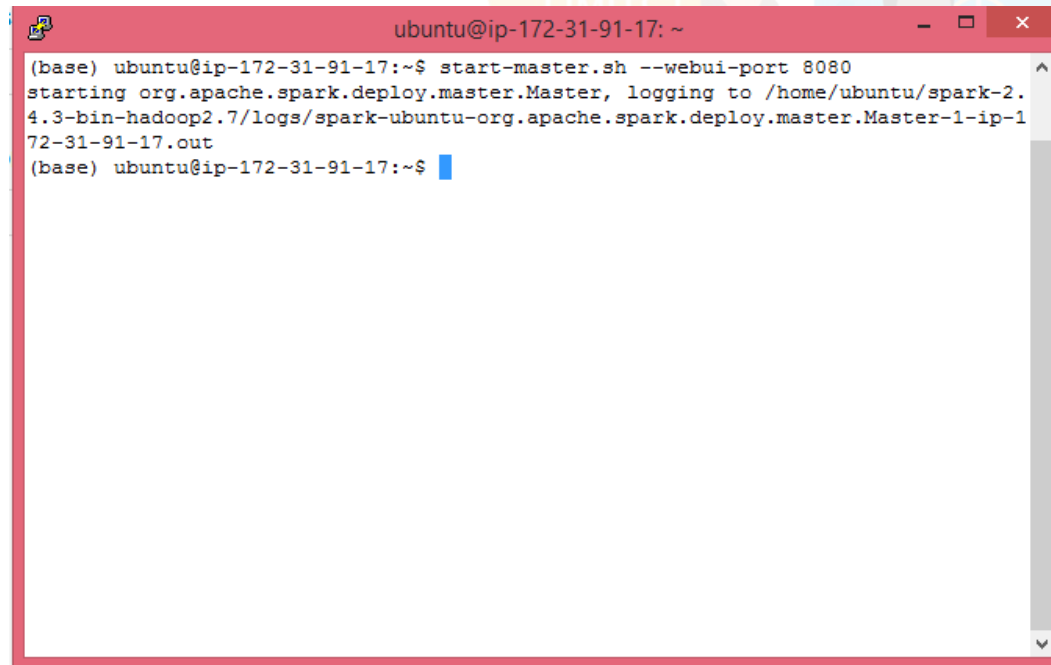


```
ubuntu@ip-172-31-91-17: ~  
(base) ubuntu@ip-172-31-91-17:~$ spark-shell  
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
Spark context Web UI available at http://ip-172-31-91-17.ec2.internal:4040  
Spark context available as 'sc' (master = local[*], app id = local-1561589034261).  
Spark session available as 'spark'.  
Welcome to  
  
          version 2.4.3  
  
Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_212)  
Type in expressions to have them evaluated.  
Type :help for more information.  
  
scala> 
```

Spark WebUI

Jalankan perintah berikut dan buka EC2 instance anda di browser dengan port 8080

```
$ start-master.sh --webui-port 8080
```



```
ubuntu@ip-172-31-91-17: ~  
(base) ubuntu@ip-172-31-91-17:~$ start-master.sh --webui-port 8080  
starting org.apache.spark.deploy.master.Master, logging to /home/ubuntu/spark-2.4.3-bin-hadoop2.7/logs/spark-ubuntu-org.apache.spark.deploy.master.Master-1-ip-172-31-91-17.out  
(base) ubuntu@ip-172-31-91-17:~$
```



DIGITAL
TALENT
SCHOLARSHIP

Spark WebUI

TERBUKA
UNTUK
DISABILITAS

Spark Master at spark://ip-172-31-91-17.ec2.internal:7077

URL: spark://ip-172-31-91-17.ec2.internal:7077
Alive Workers: 0
Cores in use: 0 Total, 0 Used
Memory in use: 0.0 B Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (0)

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

Running Applications (0)

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

Completed Applications (0)

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



Konfigurasi Hosts

Big Data

Apache Spark



Konfigurasi Hosts

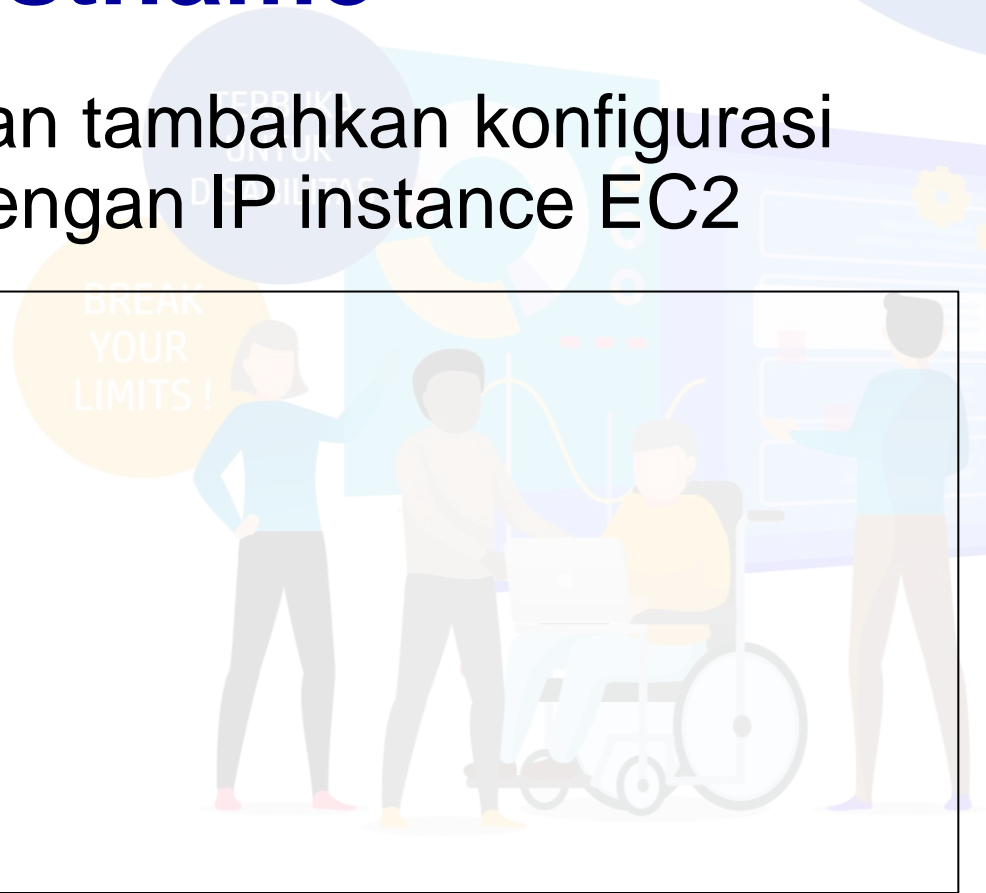
- Lewati bagian ini apabila anda sudah mengkonfigurasi cluster pada hadoop multi node cluster



Konfigurasi Hostname

- Buka file `/etc/hosts` dan tambahkan konfigurasi berikut. (sesuaikan dengan IP instance EC2 anda)

```
3.89.107.66 master  
54.167.237.4 slave1  
3.81.150.127 slave2
```



Konfigurasi Hostname

TERBUKA UNTUK



```
ubuntu@ip-172-31-91-17: ~  
GNU nano 2.5.3      File: /etc/hosts  
127.0.0.1 localhost  
3.89.107.66 master  
54.167.237.4 slave1  
3.81.150.127 slave2  
  
# The following lines are desirable for IPv6 capable hosts  
::1 ip6-localhost ip6-loopback  
fe00::0 ip6-localnet  
ff00::0 ip6-mcastprefix  
ff02::1 ip6-allnodes  
ff02::2 ip6-allrouters  
ff02::3 ip6-allhosts  
  
[ Read 13 lines ]  
^G Get Help  ^O Write Out ^W Where Is  ^K Cut Text  ^J Justify   ^C Cur Pos  
^X Exit      ^R Read File ^\ Replace   ^U Uncut Text ^T To Spell  ^_ Go To Line
```

Konfigurasi SSH

- Buka file `/.ssh/config` dan tambahkan konfigurasi berikut.

```
Host master
    HostName master
    User ubuntu
    IdentityFile ~/.ssh/digitalent.pem

Host slave1
    HostName slave1
    User ubuntu
    IdentityFile ~/.ssh/digitalent.pem
```



Konfigurasi SSH (cont)

```
Host slave2
    HostName slave2
    User ubuntu
    IdentityFile ~/.ssh/digitalent.pem
```

TERBUKA
UNTUK
DISABILITAS

BREAK
YOUR
LIMITS!



Konfigurasi SSH

- Coba login ke instance slave1 dan slave2 dari master.

```
$ssh slave1
```

```
$ssh slave2
```

```
ubuntu@ip-172-31-32-94: ~  
(base) ubuntu@ip-172-31-91-17:~$ ssh slave1  
The authenticity of host 'slave1 (54.167.237.4)' can't be established.  
ECDSA key fingerprint is SHA256:MH8Hm3PI0UccNcv0OpEwE9uZ4ovyOUjkHwpl4ZuEYWy.  
Are you sure you want to continue connecting (yes/no)? yes  
Warning: Permanently added 'slave1,54.167.237.4' (ECDSA) to the list of known ho  
sts.  
Welcome to Ubuntu 16.04.6 LTS (GNU/Linux 4.4.0-1083-aws x86_64)  
  
* Documentation:  https://help.ubuntu.com  
* Management:    https://landscape.canonical.com  
* Support:        https://ubuntu.com/advantage  
  
57 packages can be updated.  
32 updates are security updates.  
  
New release '18.04.2 LTS' available.  
Run 'do-release-upgrade' to upgrade to it.  
  
Last login: Wed Jun 26 22:12:31 2019 from 103.119.66.36  
(base) ubuntu@ip-172-31-32-94:~$
```

```
ubuntu@ip-172-31-47-60: ~  
(base) ubuntu@ip-172-31-91-17:~$ ssh slave2  
The authenticity of host 'slave2 (3.81.150.127)' can't be established.  
ECDSA key fingerprint is SHA256:Itphb1V+Gly0qFvS4lCdcx6Eit+S6rsHj70g3r9KrwI.  
Are you sure you want to continue connecting (yes/no)? yes  
Warning: Permanently added 'slave2,3.81.150.127' (ECDSA) to the list of known ho  
sts.  
Welcome to Ubuntu 16.04.6 LTS (GNU/Linux 4.4.0-1083-aws x86_64)  
  
* Documentation:  https://help.ubuntu.com  
* Management:    https://landscape.canonical.com  
* Support:        https://ubuntu.com/advantage  
  
57 packages can be updated.  
32 updates are security updates.  
  
New release '18.04.2 LTS' available.  
Run 'do-release-upgrade' to upgrade to it.  
  
Last login: Wed Jun 26 22:12:31 2019 from 103.119.66.36  
(base) ubuntu@ip-172-31-47-60:~$
```

Konfigurasi Cluster

Big Data

Apache Spark



Konfigurasi Cluster

- Lakukan konfigurasi seperti yang sudah anda lakukan ke masing-masing slave



Konfigurasi Master Node

- Salin file /home/ubuntu/spark-2.4.3-bin-hadoop2.7/conf/slaves.template pada master node (sesuaikan dengan ip anda).
- Rubah Namanya menjadi slaves
- kemudian tambahkan string berikut(gunakan WinSCP)

```
Slave1
```

```
slave2
```

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

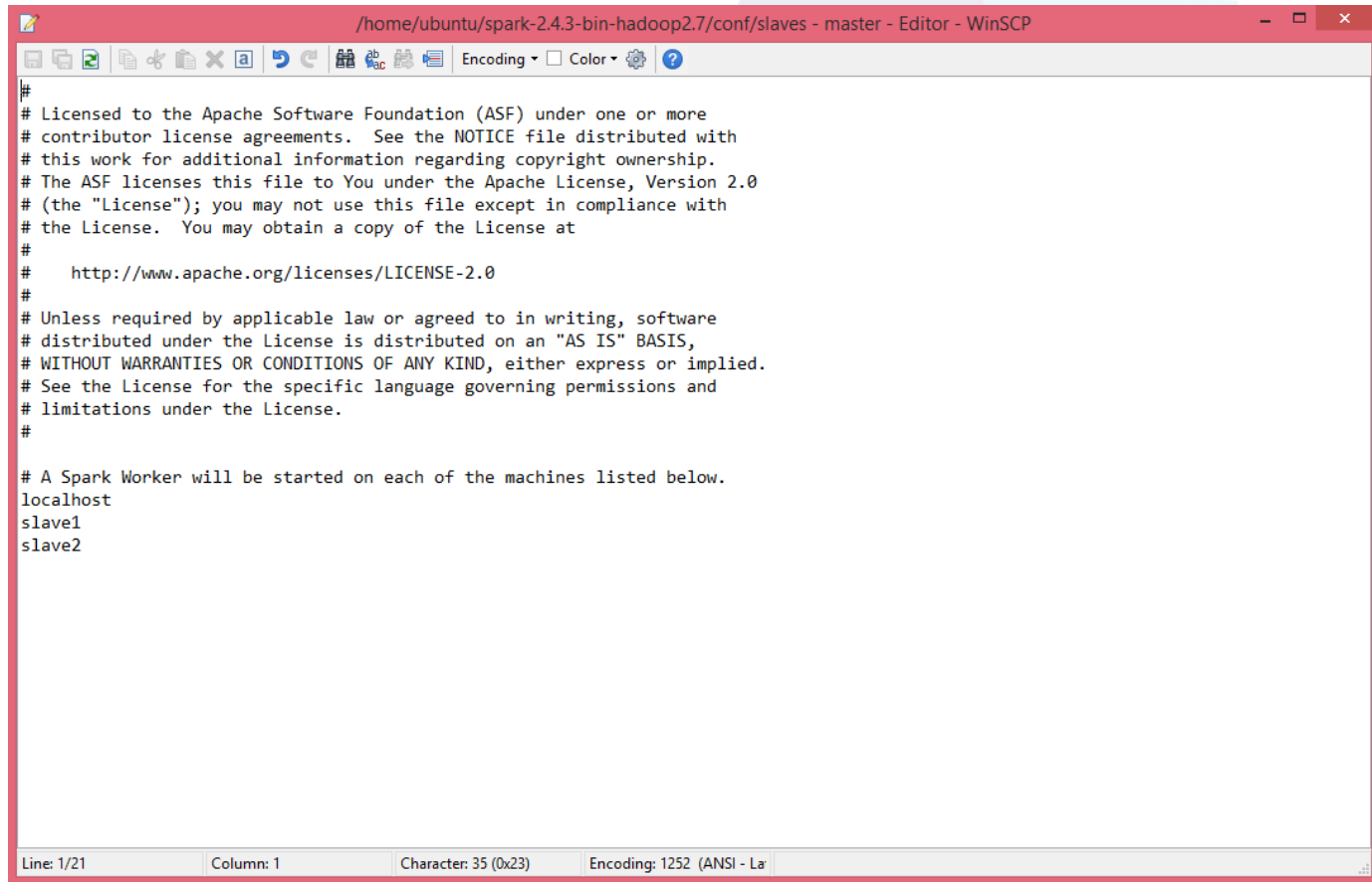
YOUR
DISABILITAS

YOUR
DISABILITAS

YOUR
DISABILITAS

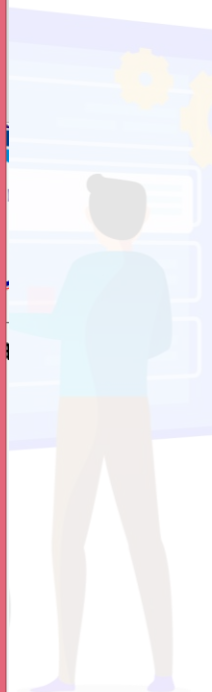
YOUR
DISABILITAS

Konfigurasi Master Node



```
#  
# 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.  
#  
  
# A Spark Worker will be started on each of the machines listed below.  
localhost  
slave1  
slave2
```

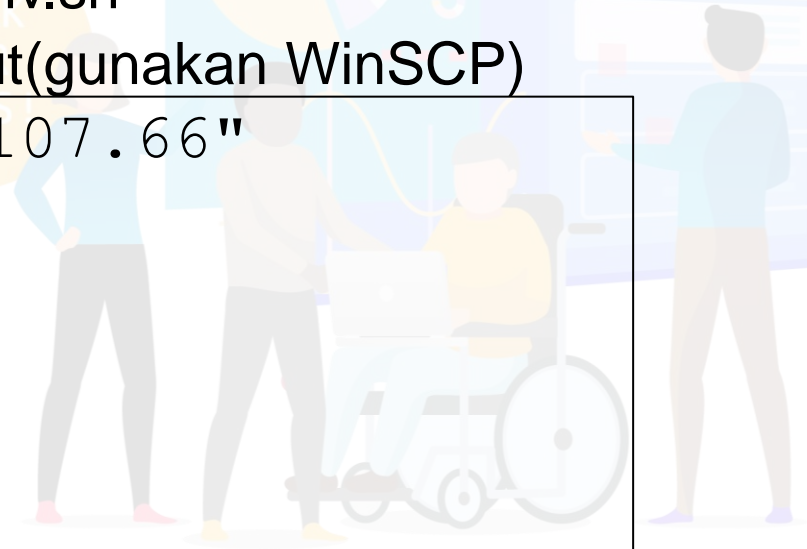
Line: 1/21 Column: 1 Character: 35 (0x23) Encoding: 1252 (ANSI - La)



Konfigurasi Slave Node

- Salin file /home/ubuntu/spark-2.4.3-bin-hadoop2.7/conf/spark-env.sh.template pada master node (sesuaikan dengan ip anda).
- Rubah Namanya menjadi spark-env.sh
- kemudian tambahkan string berikut(gunakan WinSCP)

```
SPARK_MASTER_HOST="3.89.107.66"  
SPARK_MASTER_PORT=7077
```

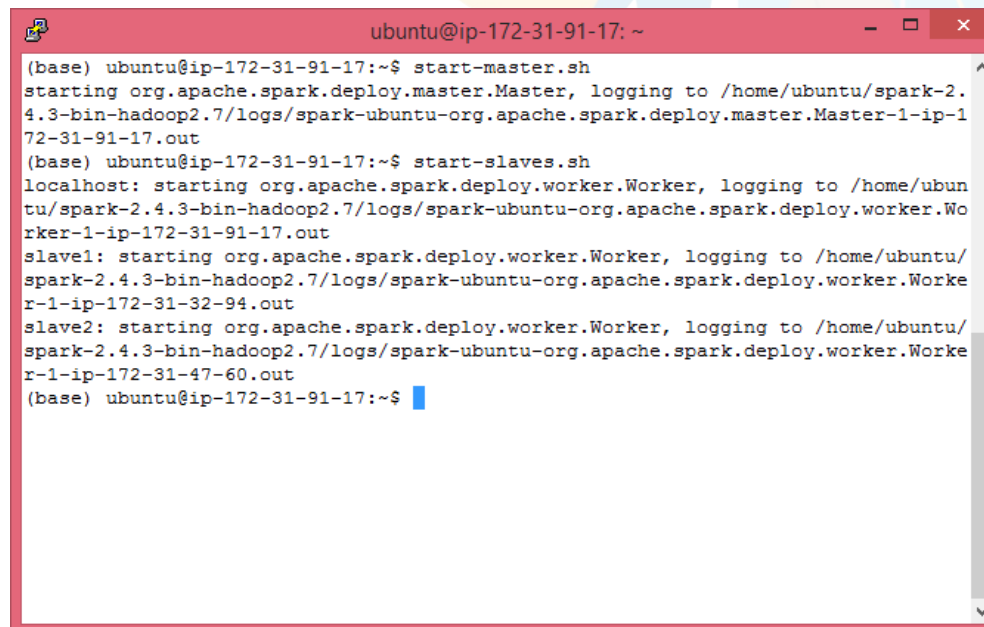


Start Cluster

- Jalankan 2 perintah berikut pada master untuk mulai menjalankan Cluster.

```
$ start-master.sh
```

```
$ start-slaves.sh
```



```
ubuntu@ip-172-31-91-17: ~  
(base) ubuntu@ip-172-31-91-17:~$ start-master.sh  
starting org.apache.spark.deploy.master.Master, logging to /home/ubuntu/spark-2.4.3-bin-hadoop2.7/logs/spark-ubuntu-org.apache.spark.deploy.master.Master-1-ip-172-31-91-17.out  
(base) ubuntu@ip-172-31-91-17:~$ start-slaves.sh  
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/ubuntu/spark-2.4.3-bin-hadoop2.7/logs/spark-ubuntu-org.apache.spark.deploy.worker.Worker-1-ip-172-31-91-17.out  
slave1: starting org.apache.spark.deploy.worker.Worker, logging to /home/ubuntu/spark-2.4.3-bin-hadoop2.7/logs/spark-ubuntu-org.apache.spark.deploy.worker.Worker-1-ip-172-31-32-94.out  
slave2: starting org.apache.spark.deploy.worker.Worker, logging to /home/ubuntu/spark-2.4.3-bin-hadoop2.7/logs/spark-ubuntu-org.apache.spark.deploy.worker.Worker-1-ip-172-31-47-60.out  
(base) ubuntu@ip-172-31-91-17:~$
```



DIGITAL
TALENT
SCHOLARSHIP

PySpark Wordcount (Python)

Big Data

Apache Spark

TERBUKA
UNTUK
DISABILITAS

BREAK
YOUR



PySpark Wordcount

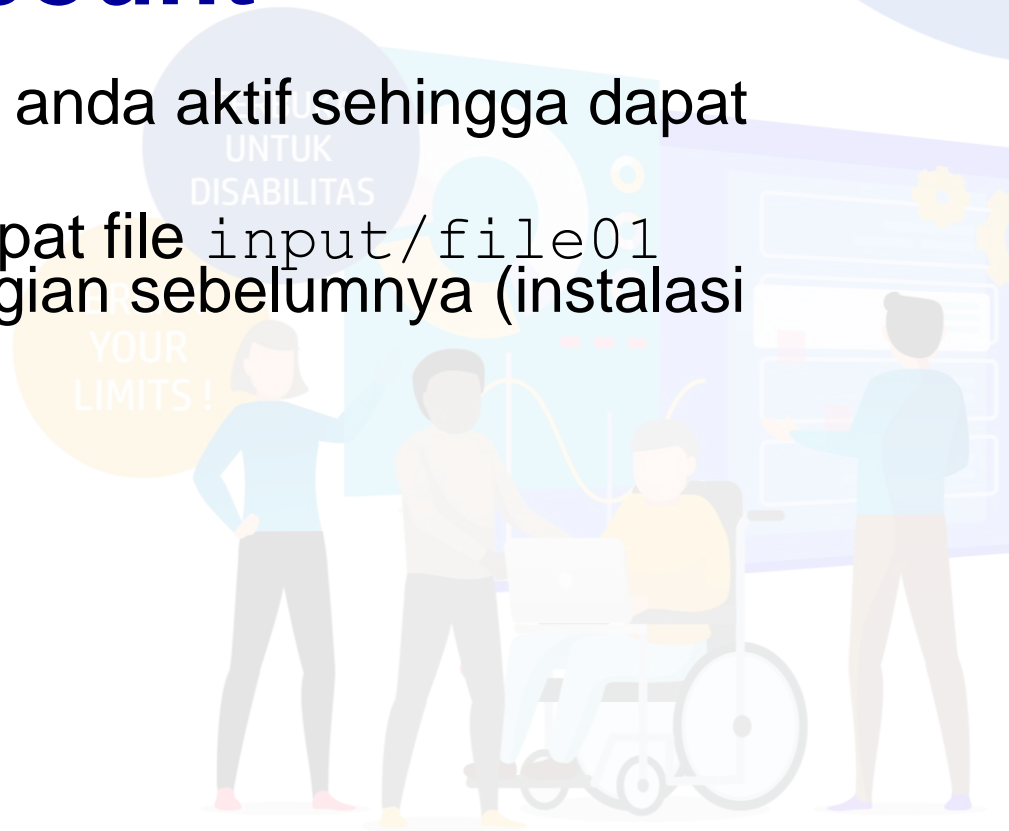
- Buat sebuah file dengan nama `wordcount.py` kemudian isi dengan string berikut. (ganti master dengan hdfs site)

```
from pyspark import SparkContext
sc = SparkContext(appName = "wordcount")
text_file =
sc.textFile("hdfs://master/user/ubuntu/input/file01")

counts = text_file.flatMap(lambda line:
line.split(" ")) \
                    .map(lambda word: (word, 1)) \
                    .reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://master/user/ubuntu/output/file01/pyspark")
```

PySpark Wordcount

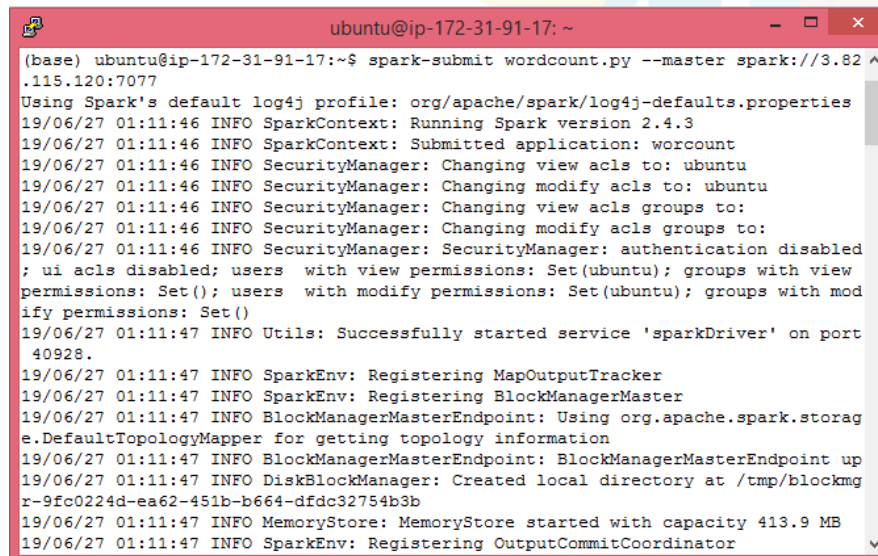
- Pastikan Hadoop Cluster anda aktif sehingga dapat mengakses hdfs
- Pastikan hdfs anda terdapat file `input/file01` yang digunakan pada bagian sebelumnya (instalasi hadoop single node)



PySpark Wordcount

- Jalankan perintah berikut untuk menjalankan wordcount pada cluster spark. (ganti ip dengan ip master anda)

```
$ spark-submit wordcount.py --master spark://3.82.115.120:7077
```

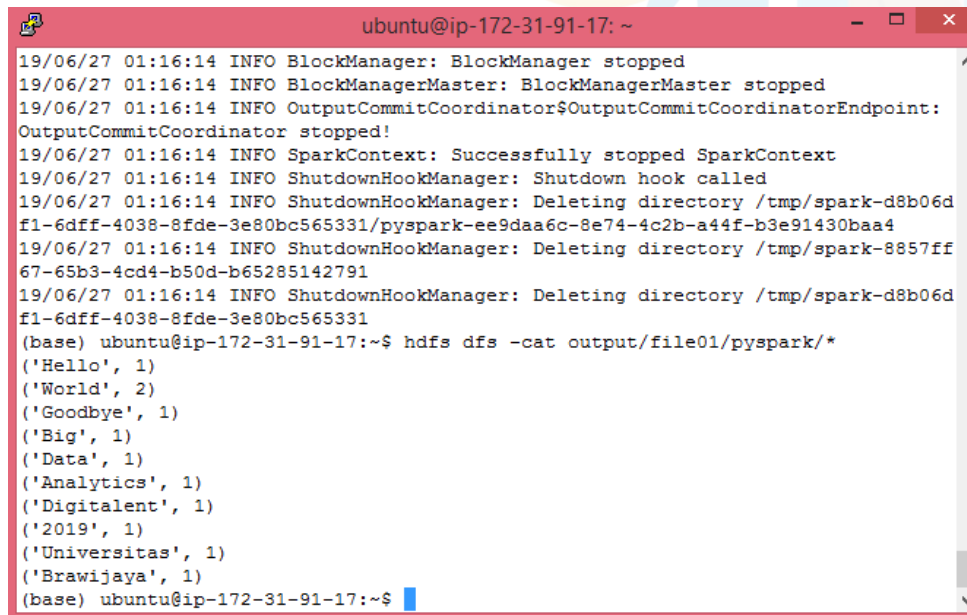


```
ubuntu@ip-172-31-91-17: ~  
(base) ubuntu@ip-172-31-91-17:~$ spark-submit wordcount.py --master spark://3.82.115.120:7077  
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties  
19/06/27 01:11:46 INFO SparkContext: Running Spark version 2.4.3  
19/06/27 01:11:46 INFO SparkContext: Submitted application: wordcount  
19/06/27 01:11:46 INFO SecurityManager: Changing view acls to: ubuntu  
19/06/27 01:11:46 INFO SecurityManager: Changing modify acls to: ubuntu  
19/06/27 01:11:46 INFO SecurityManager: Changing view acls groups to:  
19/06/27 01:11:46 INFO SecurityManager: Changing modify acls groups to:  
19/06/27 01:11:46 INFO SecurityManager: SecurityManager: authentication disabled  
; ui acls disabled; users with view permissions: Set(ubuntu); groups with view  
permissions: Set(); users with modify permissions: Set(ubuntu); groups with mod  
ify permissions: Set()  
19/06/27 01:11:47 INFO Utils: Successfully started service 'sparkDriver' on port  
40928.  
19/06/27 01:11:47 INFO SparkEnv: Registering MapOutputTracker  
19/06/27 01:11:47 INFO SparkEnv: Registering BlockManagerMaster  
19/06/27 01:11:47 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storag  
e.DefaultTopologyMapper for getting topology information  
19/06/27 01:11:47 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up  
19/06/27 01:11:47 INFO DiskBlockManager: Created local directory at /tmp/blockmg  
r-9fc0224d-ea62-451b-b664-dfdc32754b3b  
19/06/27 01:11:47 INFO MemoryStore: MemoryStore started with capacity 413.9 MB  
19/06/27 01:11:47 INFO SparkEnv: Registering OutputCommitCoordinator
```

PySpark Wordcount

- Jalankan perintah berikut untuk mencetak luaran program

```
$ hdfs dfs -cat output/file01/pyspark/*
```



```
ubuntu@ip-172-31-91-17: ~  
19/06/27 01:16:14 INFO BlockManager: BlockManager stopped  
19/06/27 01:16:14 INFO BlockManagerMaster: BlockManagerMaster stopped  
19/06/27 01:16:14 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:  
OutputCommitCoordinator stopped!  
19/06/27 01:16:14 INFO SparkContext: Successfully stopped SparkContext  
19/06/27 01:16:14 INFO ShutdownHookManager: Shutdown hook called  
19/06/27 01:16:14 INFO ShutdownHookManager: Deleting directory /tmp/spark-d8b06d  
f1-6dff-4038-8fde-3e80bc565331/pyspark-ee9daa6c-8e74-4c2b-a44f-b3e91430baa4  
19/06/27 01:16:14 INFO ShutdownHookManager: Deleting directory /tmp/spark-8857ff  
67-65b3-4cd4-b50d-b65285142791  
19/06/27 01:16:14 INFO ShutdownHookManager: Deleting directory /tmp/spark-d8b06d  
f1-6dff-4038-8fde-3e80bc565331  
(base) ubuntu@ip-172-31-91-17:~$ hdfs dfs -cat output/file01/pyspark/*  
(Hello', 1)  
(World', 2)  
(Goodbye', 1)  
(Big', 1)  
(Data', 1)  
(Analytics', 1)  
(Digitalent', 1)  
(2019', 1)  
(Universitas', 1)  
(Brawijaya', 1)  
(base) ubuntu@ip-172-31-91-17:~$
```


Case Study

Big Data

Apache Spark



Case Study

- Implementasikan case study pada instalasi hadop single node (penjumlahan ganjil genap) menggunakan PySpark

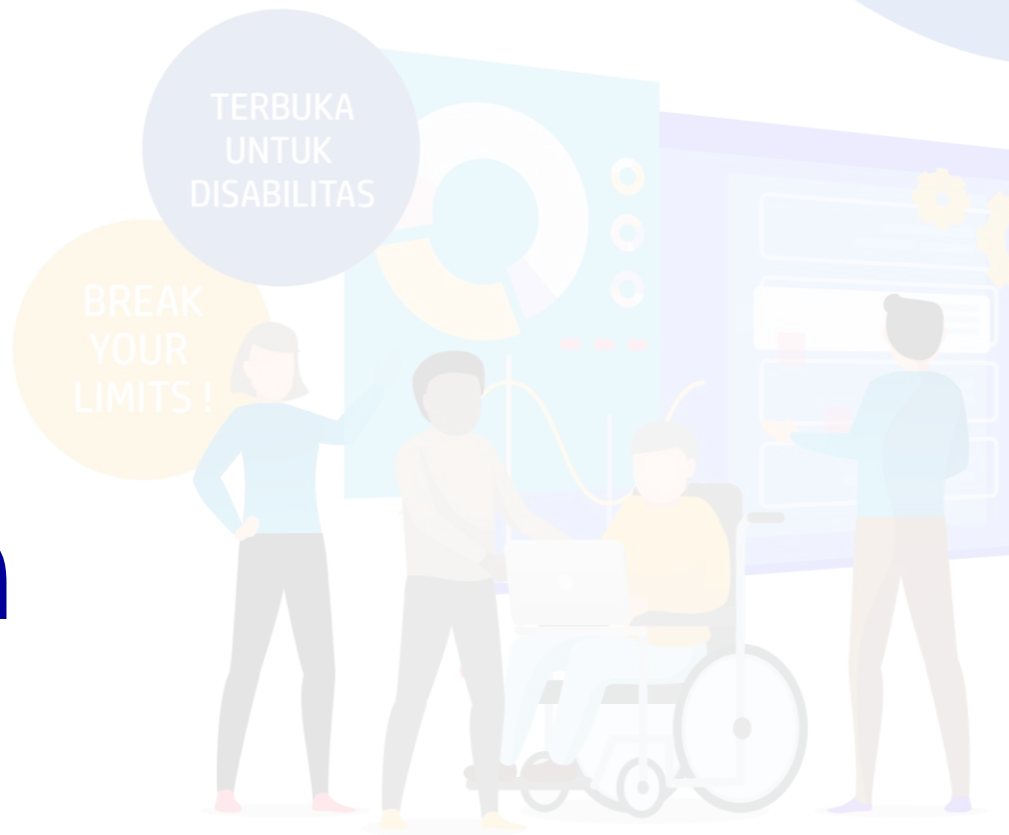
BREAK
YOUR
LIMITS!



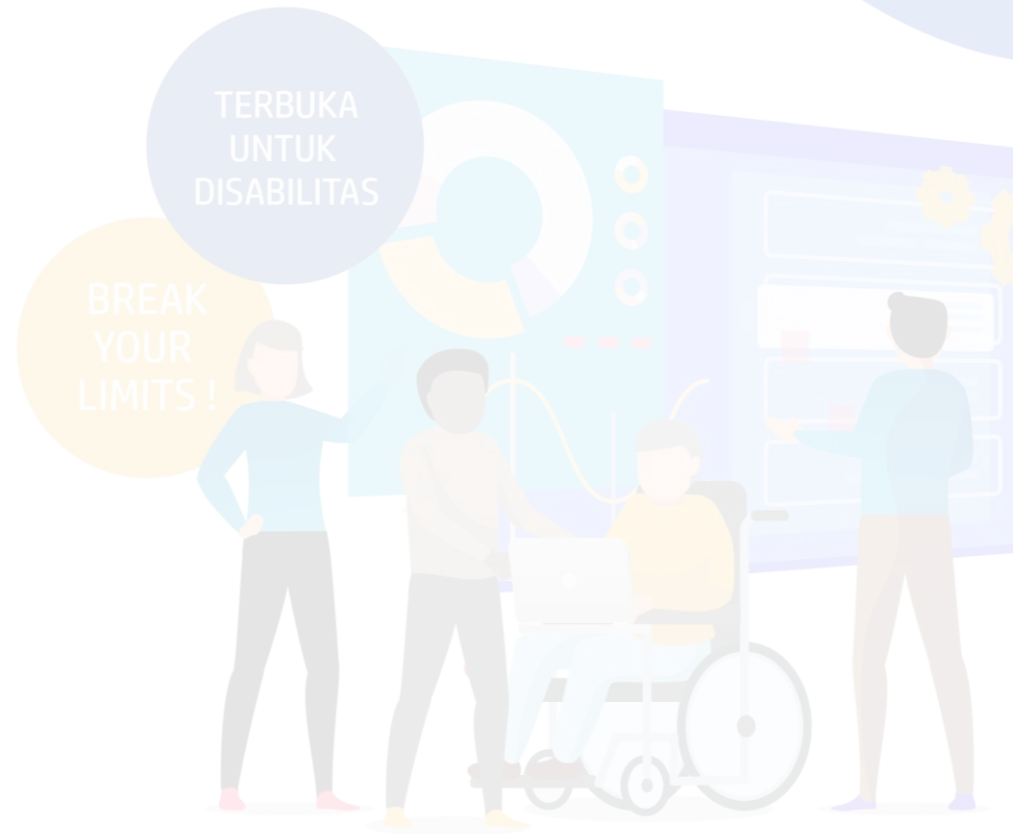
Pertanyaan

Big Data

Hadoop



Pertanyaan





DIGITAL
TALENT
SCHOLARSHIP

IKUTI KAMI



DIGITAL
TALENT
SCHOLARSHIP

- digitalent.kominfo
- digitalent.kominfo
- DTS_kominfo
- Digital Talent Scholarship 2019

Pusat Pengembangan Profesi dan Sertifikasi
Badan Penelitian dan Pengembangan SDM
Kementerian Komunikasi dan Informatika
Jl. Medan Merdeka Barat No. 9
(Gd. Belakang Lt. 4 - 5)
Jakarta Pusat, 10110



digitalent.kominfo.go.id