



DIGITAL
TALENT
SCHOLARSHIP



DIGITAL TALENT SCHOLARSHIP 2019

Big Data Analytics



EMR Lab: Hive, Hue, & Others

Oleh: Imam Cholissodin | imamcs@ub.ac.id, Putra Pandu Adikara, Sufia Adha Putri

Asisten: Guedho, Sukma, Anshori, Aang dan Gusti

Fakultas Ilmu Komputer (Filkom) Universitas Brawijaya (UB)

Pokok Pembahasan

- Menggunakan Hive pada EMR
- Menggunakan Hue pada EMR
- Tugas



TERBUKA
UNTUK
DISABILITAS

BREAK
YOUR
LIMITS!

Lab Hive



Big Data Query

```
data.head()
```

| | created_at | in_reply_to_status_id | id_str | retweeted | in_reply_to_user_id_str | coordinates | retweet_count | contributors | favorite_count |
|---|---------------------|-----------------------|----------------------|-----------|-------------------------|-------------|---------------|--------------|----------------|
| 0 | 2019-07-22 05:41:45 | NaN | 1153178298081374208 | False | NaN | NaN | 0 | NaN | 0 |
| 1 | 2019-07-22 05:41:47 | 1.153177e+18 | 1153178304603545600 | False | 731167584.0 | NaN | 0 | NaN | 0 |
| 2 | 2019-07-22 05:41:48 | NaN | 1.153178311515766785 | False | NaN | NaN | NaN | NaN | 0 |

Siapa user yang paling sering nge-tweet?

| | created_at | in_reply_to_status_id | id_str | retweeted | in_reply_to_user_id | source | user_id_str | user_screen_name | place | geo | text |
|---|---------------------|-----------------------|---------------------|-----------|---------------------|---------------------|--------------------|---------------------|------------|-----|---|
| 3 | 2019-07-22 05:41:49 | NaN | 1153178311515766785 | False | NaN | Twitter for iPad | 513783027 | fierceangel | NaN | NaN | Aaaakkk YukSay balik |
| 4 | 2019-07-22 05:41:49 | NaN | 115317831394421248 | False | 1.153177e+18 | Twitter for iPhone | 731167584.0 | 1003356794683416576 | CatsRule98 | NaN | @RonBrownstein That's why #ISlandWithHilhan |
| | | | | False | NaN | Twitter for Android | 1097334060 | JejakaShahid | NaN | NaN | RT @mathsanova: mana pakai faceapp lagi 🤔 |
| | | | | False | NaN | Twitter for Android | 1014117841 | nurlaneksa | NaN | NaN | RT @BadmintonTalk: #BtalkBWFRankPrediction MS ... |
| | | | | False | NaN | Twitter for Android | 701611356496113664 | sticvh | NaN | NaN | RT @Husen_Jafar: FaceApp haram? #RecehanDakwah... |

Tweet mana yang paling sering difavoritkan?

Aplikasi apa saja yang paling populer untuk nge-twit?

Total data : 5.186.814 baris (row)

Dataset

Object URL :

<https://hadoopbucket-digitalent.s3.amazonaws.com/tweet/tweet.csv>

EMR Lab

1. Buatlah security group pada EC2 anda dengan nama SSH, dan juga pastikan inbound dan outbound rule seperti berikut :

Create Security Group

Security group name: ssh

Description: ssh

VPC: vpc-ccc1acb6 (default)

Security group rules:

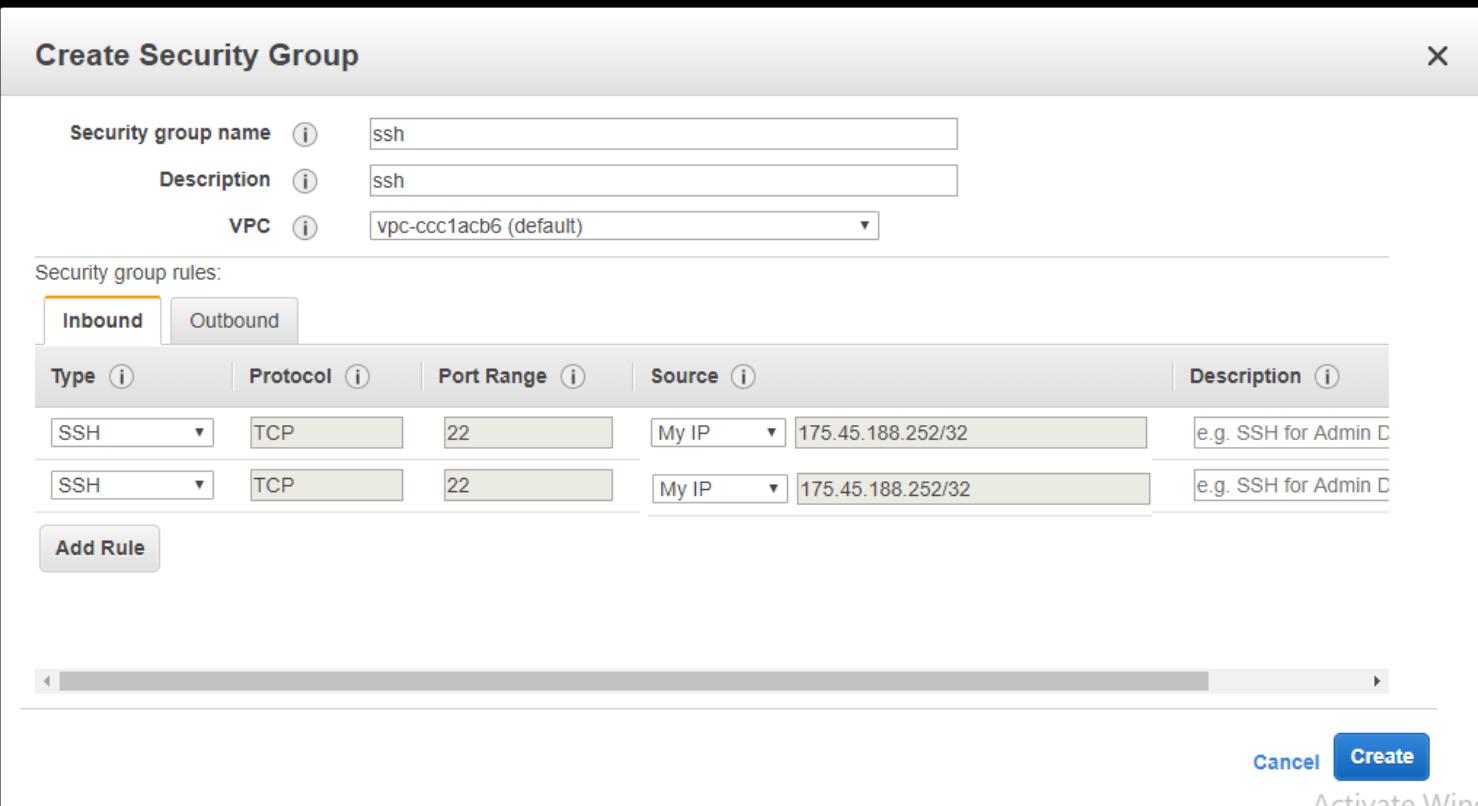
Inbound Outbound

| Type | Protocol | Port Range | Source | Description |
|------|----------|------------|----------------------------|----------------------|
| SSH | TCP | 22 | My IP 175.45.188.252/32 | e.g. SSH for Admin D |
| SSH | TCP | 22 | My IP 175.45.188.252/32 | e.g. SSH for Admin D |

Add Rule

Cancel Create

Activate Win



EMR Lab

2. Buka Service EMR pada Analytics

The screenshot shows the AWS Services Catalog interface. On the left, there's a sidebar with links like History, EMR, Console Home, EC2, Redshift, Support, and Billing. The main area has a search bar at the top with the placeholder "Find a service by name or feature (for example, EC2, S3 or VM, storage)". Below the search bar, services are categorized into four groups: Compute, Developer Tools, Analytics, Application Services, Storage, Management Tools, Artificial Intelligence, Messaging, and Business Productivity. The "Analytics" group is currently selected, indicated by a blue border around its title and a yellow arrow icon next to the "EMR" service. Other services in the Analytics group include Athena, CloudSearch, Elasticsearch Service, Kinesis, Data Pipeline, and QuickSight.

| Compute | Developer Tools | Analytics | Application Services |
|-----------------------|-----------------|-----------------------|----------------------|
| EC2 | CodeStar | Athena | Step Functions |
| EC2 Container Service | CodeCommit | CloudSearch | SWF |
| Lightsail | CodeBuild | Elasticsearch Service | API Gateway |
| Elastic Beanstalk | CodeDeploy | Kinesis | Elastic Transcoder |
| Lambda | CodePipeline | Data Pipeline | |
| Batch | X-Ray | QuickSight | |

| Storage | Management Tools | Artificial Intelligence | Messaging |
|-----------------|------------------|-------------------------|-----------------------------|
| S3 | CloudWatch | Lex | Simple Queue Service |
| EFS | CloudFormation | Polly | Simple Notification Service |
| Glacier | CloudTrail | Rekognition | SES |
| Storage Gateway | Config | Machine Learning | |
| | OpsWorks | | |

| Business Productivity |
|-----------------------|
| WorkDocs |

EMR Lab

3. Pada EMR Klik Create Cluster à go to advanced options

The screenshot shows the 'Create Cluster - Quick Options' interface. At the top, there's a navigation bar with 'Services', 'Resource Groups', and a user icon 'Sanja'. Below the title, a link 'Go to advanced options' is visible. The main section is titled 'General Configuration'.

General Configuration:

- Cluster name:** My cluster
- Logging:** Logging
- S3 folder:** s3://aws-logs-145550458401-us-east-1/elasticmapreduce/
- Launch mode:** Cluster Step execution

Software configuration:

- Release:** emr-5.6.0
- Applications:**
 - Core Hadoop: Hadoop 2.7.3 with Ganglia 3.7.2, Hive 2.1.1, Hue 3.12.0, Mahout 0.13.0, Pig 0.16.0, and Tez 0.8.4
 - HBase: HBase 1.3.0 with Ganglia 3.7.2, Hadoop 2.7.3, Hive 2.1.1, Hue 3.12.0, Phoenix 4.9.0, and ZooKeeper 3.4.10

EMR Lab

4. Sesuaikan software configuration seperti ini, lalu klik Next à

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Software Configuration

Release emr-5.6.0

| | | |
|--|--|--|
| <input checked="" type="checkbox"/> Hadoop 2.7.3 | <input type="checkbox"/> Zeppelin 0.7.1 | <input type="checkbox"/> Tez 0.8.4 |
| <input type="checkbox"/> Flink 1.2.1 | <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 1.3.0 |
| <input type="checkbox"/> Pig 0.16.0 | <input checked="" type="checkbox"/> Hive 2.1.1 | <input type="checkbox"/> Presto 0.170 |
| <input type="checkbox"/> ZooKeeper 3.4.10 | <input type="checkbox"/> Sqoop 1.4.6 | <input type="checkbox"/> Mahout 0.13.0 |
| <input type="checkbox"/> Hue 3.12.0 | <input type="checkbox"/> Phoenix 4.9.0 | <input type="checkbox"/> Oozie 4.3.0 |
| <input checked="" type="checkbox"/> Spark 2.1.1 | <input type="checkbox"/> HCatalog 2.1.1 | |

Edit software settings (optional) [i](#)

Enter configuration Load JSON from S3

```
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
```

Add steps (optional) [i](#)

Step type [Select a step](#) [Configure](#)

Auto-terminate cluster after the last step is completed

EMR Lab

5. Sesuaikan hardware configuration seperti ini, lalu klik Next à

The screenshot shows the AWS EMR console's hardware configuration step. It lists three task instance groups: Master, Core, and Task. Each group has one m3.xlarge instance type selected, with 8 vCPU, 15 GiB memory, and 80 SSD GB storage. The EBS Storage field is set to none. The Master group has 1 instance, while Core and Task have 0 instances. Purchasing options are set to Spot with a maximum bid price of \$0,07. The Auto Scaling and On-demand options are disabled. Buttons at the bottom include 'Cancel', 'Previous', and 'Next'.

| Node type | Instance type | Instance count | Purchasing option | Auto Scale |
|----------------------|--|----------------|--|---------------|
| Master Master - 1 | m3.xlarge 8 vCPU, 15 GiB memory, 80 SSD GB storage EBS Storage: none | 1 Instances | <input type="radio"/> On-demand <input checked="" type="radio"/> Spot Maximum bid price: \$ 0,07 | Not available |
| Core Core - 2 | m3.xlarge 8 vCPU, 15 GiB memory, 80 SSD GB storage EBS Storage: none | 0 Instances | <input checked="" type="radio"/> On-demand <input type="radio"/> Spot Maximum bid price: \$ | Not enabled |
| Task Task - 3 | m3.xlarge 8 vCPU, 15 GiB memory, 80 SSD GB storage EBS Storage: none | 0 Instances | <input checked="" type="radio"/> On-demand <input type="radio"/> Spot Maximum bid price: \$ | Not enabled |

+ Add task instance group

Cancel Previous Next

EMR Lab

6. Sesuaikan General Cluster setting seperti berikut, lalu klik Next à

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

General Options

Cluster name: emrlab

Logging i
S3 folder: s3://bigdatalabssk/emrlogs/ ...

Debugging i

Termination protection i

Scale down behavior: Terminate at instance hour ...

Tags i

| Key | Value (optional) |
|---------------------------|--------------------------|
| Name | EMR Lab <small>x</small> |
| Add a key to create a tag | |

Additional Options

EMRFS consistent view i

▶ Bootstrap Actions

Cancel **Previous** **Next**

EMR Lab

7. Pada Security pilih EC2 Key pair yang sudah dibuat sebelumnya, lalu klik EC2 Security Options

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Security Options

EC2 key pair [?](#)

Cluster visible to all IAM users in account [?](#)

Permissions

Default Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR_DefaultRole](#) [?](#)

EC2 Instance profile [EMR_EC2_DefaultRole](#) [?](#)

Auto Scaling role [EMR_AutoScaling_DefaultRole](#) [?](#)

[Encryption Options](#)

[EC2 Security Groups](#)

[Cancel](#) [Previous](#) **Create cluster**

EMR Lab

8. Pada Master klik pilihan edit pada Additional security group

The screenshot shows the AWS EMR Cluster Configuration interface at Step 4: Security. The 'Additional security groups' section for the Master role is highlighted with a red box and a cursor pointing to the 'Edit' icon. The 'EMR managed security groups' section for the Master role shows 'Default: sg-7d8b7e03 (ElasticMapReduce-master)'. The 'Additional security groups' section for the Master role shows 'No security groups selected' with an edit icon.

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

EC2 Key pair: digoastacert

Cluster visible to all IAM users in account

Permissions

Default Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role: [EMR_DefaultRole](#)

EC2 instance profile: [EMR_EC2_DefaultRole](#)

Auto Scaling role: [EMR_AutoScaling_DefaultRole](#)

Encryption Options

EC2 Security Groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, EMR managed security groups and additional security groups. EMR will automatically update the rules in the EMR managed security groups in order to launch a cluster. Learn more.

| Type | EMR managed security groups | Additional security groups |
|-------------|--|-----------------------------|
| Master | Default: sg-7d8b7e03 (ElasticMapReduce-master) | No security groups selected |
| Core & Task | Default: sg-c18b7ebf (ElasticMapReduce-slave) | No security groups selected |

Create a security group

Cancel Previous Create cluster

EMR Lab

9. Lalu, pilih ssh security group yang telah dibuat sebelumnya, dan klik Assign security group, Lalu klik Create Cluster



EMR Lab

10. Akhirnya Cluster telah dibuat

Amazon EMR

Add step Resize Clone Terminate AWS CLI export

Cluster list Starting

Cluster: emrlab

Connections: --

Master public DNS: --

Tags: Name = EMR Lab View All / Edit

Summary Configuration Details

ID: j-1KYL95NDBNWL Release label: emr-5.6.0

Creation date: 2017-06-30 11:09 (UTC-4) Hadoop distribution: Amazon 2.7.3

Elapsed time: 0 seconds Applications: Hive 2.1.1, Spark 2.1.1

Auto-terminate: No Log URI: s3://bigdatalabssk/emrlogs/

Termination On Change protection: EMRFS consistent view: Disabled

Network and Hardware Security and Access

Availability zone: -- Key name: bigdatacert

Subnet ID: subnet-6009e216 EC2 instance profile: EMR_EC2_DefaultRole

Master: Provisioning 1 m3.xlarge (Spot: 0.049) EMR role: EMR_DefaultRole

Core: -- Auto Scaling role: EMR_AutoScaling_DefaultRole

Task: -- Visible to all users: All Change

Security groups for sg-7d8b7e03 (ElasticMapReduce- Master: master) More

Security groups for sg-c18b7ebf (ElasticMapReduce- Core & Task: slave)

EMR Lab

11. Copy Master public DNS, dan buka sesion terminal SSH

Amazon EMR

Add step Resize Clone Terminate AWS CLI export

Cluster list Security configurations VPC subnets Events Help

Cluster: emrlab Waiting Cluster ready after last step completed.

Connections: Enable Web Connection – Spark History Server, Resource Manager ... (View All)

Master public DNS: 54.89.186.242 SSH

Tags: Name = EMR Lab View All / Edit

Summary

ID: j-21D9UFGIC8NJR
Creation date: 2017-06-30 13:17 (UTC-4)
Elapsed time: 24 minutes
Auto-terminate: No
Termination On Change protection:

Configuration Details

Release label: emr-5.6.0
Hadoop distribution: Amazon 2.7.3
Applications: Hive 2.1.1, Spark 2.1.1
Log URI: s3://bigdatalabssk/emrlogs/

EMRFS consistent Disabled view:

Network and Hardware

Availability zone: us-east-1a
Subnet ID: subnet-6009e216
Master: Running 1 m3.xlarge (Spot: 0.049)
Core: --
Task: --

Security and Access

Key name: bigdatacert
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All Change
Security groups for sg-7d8b7e03 (ElasticMapReduce-Master: master) More
Security groups for sg-c18b7ebf (ElasticMapReduce-Core & Task: slave)

Download dataset

11. wget https://hadoopbucket-digitalent.s3.amazonaws.com/tweet/tweet.csv

```
code — hadoop@ip-172-31-74-253:~ — ssh hadoop@ec2-18-208-143-2.comp...
```

```
[hadoop@ip-172-31-74-253 ~]$ wget https://hadoopbucket-digitalent.s3.amazonaws.com/tweet/tweet.csv
--2019-07-31 08:19:52--  https://hadoopbucket-digitalent.s3.amazonaws.com/tweet/tweet.csv
Resolving hadoopbucket-digitalent.s3.amazonaws.com (hadoopbucket-digitalent.s3.amazonaws.com)... 52.216.137.244
Connecting to hadoopbucket-digitalent.s3.amazonaws.com (hadoopbucket-digitalent.s3.amazonaws.com)|52.216.137.244|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1276813940 (1.2G) [text/csv]
Saving to: 'tweet.csv.1'
```

tweet.csv.1 27% 338.84M 68.0MB/s eta 13s

Copy data ke HDFS

11. hdfs dfs -put tweet.csv /user/hadoop/

```
code — hadoop@ip-172-31-74-253:~ — ssh hadoop@ec2-18-208-143-2.comp...
pbucket-digitalent.s3.amazonaws.com)... 52.217.38.204
Connecting to hadoopbucket-digitalent.s3.amazonaws.com (h
adoobucket-digitalent.s3.amazonaws.com)|52.217.38.204|:4
43... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1276813940 (1.2G) [text/csv]
Saving to: 'tweet.csv'

tweet.csv          100%[=====>]    1.19G  51.4
MB/s    in 21s

2019-07-31 07:20:20 (58.0 MB/s) - 'tweet.csv' saved [1276
813940/1276813940]

[hadoop@ip-172-31-74-253 ~]$ hdfs dfs -put tweet.csv /use
r/hadoop/tweet.csv
```

Menggunakan Hive

12. Ketik 'hive' di terminal

```
code — hadoop@ip-172-31-26-166:~ — ssh hadoop@ec2-3-80-112-187.com...

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
8 package(s) needed for security, out of 13 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEE MMMMMMM   MMMMMMM RRRRRRRRRRRRRRRR
E::::::: E M:::::M   M:::::M R:::::::::::R
EE:::::E EEEEEEEEEE:::E M:::::M   M:::::M R:::::RRRRRR:::::R
E:::::E     EEEE  M:::::M   M:::::M RR:::::R    R:::::R
E:::::E     M:::::M::::M   M::::M:::::M R:::R    R:::::R
E:::::EEEEEEEEEE   M:::::M M::::M M::::M M::::M R:::RRRRRR:::::R
E:::::::::::E   M:::::M M::::M::::M M:::::M R:::::::::::RR
E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M R:::RRRRRR:::::R
E:::::E     M:::::M   M:::::M   M:::::M R:::R    R:::::R
E:::::E     EEEE  M:::::M   MMM   M:::::M R:::R    R:::::R
EE:::::EEEEEEEEEE:::E M:::::M   M:::::M R:::R    R:::::R
E:::::::::::E:::::E M:::::M   M:::::M RR:::::R    R:::::R
EEEEEEEEEEEEEEEEEE MMMMMMM   MMMMMMM RRRRRRRR

[[hadoop@ip-172-31-26-166 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> ]
```

Membuat schema di EMR

13. Ketik 'create schema twitter;'

```
code — hadoop@ip-172-31-26-166:~ — ssh hadoop@ec2-3-80-112-187.com...
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEE MMMMMMM   MMMMMMM RRRRRRRRRRRRRRR
E::::::::::: E M:::::::M   M:::::::M R:::::::R:::::R
EE:::::EEEEEEE:::E M:::::::M   M:::::::M R:::::RRRRRR:::::R
  E::::E     EEEEE  M:::::::M   M:::::::M RR:::::R      R:::::R
  E::::E           M:::::M:::M  M:::M:::M:::::M   R:::R      R:::::R
  E:::::EEEEEEEEE  M:::::M M:::M M:::M M:::::M   R:::::RRRRRR:::::R
  E:::::::::::E    M:::::M M:::M:::M M:::::M   R:::::R:::::RR
  E:::::EEEEEEEEE  M:::::M   M:::::M   M:::::M   R:::::RRRRRR:::::R
  E:::::E           M:::::M   M:::::M   M:::::M   R:::::R
  E:::::E     EEEEE  M:::::::M   M:::::M   R:::::R      R:::::R
  E:::::E::::::: E M:::::::M   M:::::M   R:::::R      R:::::R
  E:::::E::::::: E M:::::::M   M:::::M   RR:::::R      R:::::R
  EEEEEEEEEEEEEEEEEE MMMMMMM   MMMMMMM RRRRRRRR      RRRRRR

[[hadoop@ip-172-31-26-166 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
[hive> create schema twitter
[> ;
OK
Time taken: 0.77 seconds]]]
```

Import data dari S3

14. Perintah : <https://pastebin.com/6P1D7Pym>

```
properties AS yml. yaml
[hive> CREATE EXTERNAL TABLE twitter(tweet (created_at STRING, in_reply_to_status
_id STRING, id_str STRING, retweeted BOOLEAN, in_reply_to_user_id_str STRING, co
ordinates STRING, retweet_count INT, contributors STRING, favorite_count INT, fav
orited BOOLEAN, in_reply_to_status_id_str STRING, source STRING, in_reply_to_use
r_id STRING, user_id_str STRING, user_screen_name STRING, place STRING, geo STRI
NG, text STRING) LOCATION 's3://hadoopbucket-digitalent/tweet/' TBLPROPERTIES (""
skip.header.line.count"="1");
OK
Time taken: 4.782 seconds
hive> ]
```

Cek banyak baris pada tabel tweet



14. Perintah : select count(*) from tweet;

```
[● ● ● code — hadoop@ip-172-31-26-166:~ — ssh hadoop@ec2-3-80-112-187.com...]  
[hive> select count(*) from tweet;  
Query ID = hadoop_20190731043722_455b0eff-8a29-46d6-bcd7-36d5dc38f7b1  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1564545374633  
_0003)  
  
Map 1: -/-      Reducer 2: 0/1  
Map 1: 0/1      Reducer 2: 0/1  
Map 1: 0/1      Reducer 2: 0/1  
Map 1: 0(+1)/1  Reducer 2: 0/1  
Map 1: 1/1      Reducer 2: 0(+1)/1  
Map 1: 1/1      Reducer 2: 1/1  
OK  
11262928  
Time taken: 31.412 seconds, Fetched: 1 row(s)
```

User paling sering nge-tweet

15. Perintah : select user_screen_name,count(*) as c from tweet group by user_screen_name order by c desc limit 10;

```
code — hadoop@ip-172-31-26-166:~ — ssh hadoop@ec2-3-80-112-187.com...
Map 1: 1/1      Reducer 2: 0(+1)/5      Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 1(+1)/5      Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 1(+2)/5      Reducer 3: 0/1
Map 1: 1/1      Reducer 2: 2(+1)/5      Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 2(+2)/5      Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 2(+3)/5      Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 2(+3)/5      Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 3(+2)/5      Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 4(+1)/5      Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 5/5      Reducer 3: 0(+1)/1
Map 1: 1/1      Reducer 2: 5/5      Reducer 3: 1/1
OK
NULL      6076108
2428
wordnuvola    2208
GooglePayIndia 2134
Weinbach      1855
WwuRadio      1192
pogosj1 1158
Christo73106853 1130
lililin79416363 933
MercadoLechugas 917
Time taken: 63.396 seconds, Fetched: 10 row(s)
hive> ■
```

TERBUKA
UNTUK
DISABILITAS

BREAK
YOUR
LIMITS!

Lab Hue



Tentang Hue

- Hue (*Hadoop User Experience*) adalah antarmuka berbasis Web untuk memudahkan dalam memonitoring atau memanage (create, delete, edit, etc) data HDFS pada Apache Hadoop dan beberapa fungsi lainnya, serta bisa diinstall di pc/notebook dengan versi hadoop manapun.



- Membuat “any users” untuk lebih fokus pada big data processing.

Tutorial Hue

Buat cluster pada layanan EMR, beri centang seperti pada gambar dibawah, kemudian next

The screenshot shows the 'Create Cluster - Advanced Options' page in the AWS EMR console. The 'Software Configuration' section lists various software packages. Several checkboxes are checked, indicated by arrows pointing to them:

- Hadoop 2.8.5
- Hive 2.3.5
- Hue 4.4.0
- Spark 2.4.3
- HBase 1.4.9
- Sqoop 1.4.7
- ZooKeeper 3.4.14
- Pig 0.17.0
- Oozie 5.1.0

Other checked boxes include Zeppelin 0.8.1, Tez 0.9.2, Presto 0.220, Phoenix 4.14.1, HCatalog 2.3.5, Livy 0.6.0, Flink 1.8.0, Mahout 0.13.0, and TensorFlow 1.13.1.

Tutorial Hue

Berikan mark Spot untuk master dan core node dan ubah instance type

The screenshot shows the 'Create Cluster' wizard in the AWS EMR console. The 'Node type' column lists 'Master', 'Core', and 'Task'. The 'Instance type' column for each lists 'm4.large' with details: 4 vCore, 8 GiB memory, EBS only storage, and EBS Storage: 32 GiB. The 'Instance count' column shows 1 instance for Master and 2 instances for Core. The 'Purchasing option' column for both Master and Core has 'Spot' selected, indicated by a blue circle. Arrows point from the text instructions to the 'Spot' radio buttons for the Master and Core rows.

| Node type | Instance type | Instance count | Purchasing option |
|-----------|---|----------------|---|
| Master | m4.large 4 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB | 1 Instances | <input checked="" type="radio"/> On-demand <small>i</small> <input checked="" type="radio"/> Spot <small>i</small> Use on-demand as max price |
| Core | m4.large 4 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB | 2 Instances | <input checked="" type="radio"/> On-demand <small>i</small> <input checked="" type="radio"/> Spot <small>i</small> Use on-demand as max price |
| Task | m3.xlarge 8 vCore, 15 GiB memory, 80 SSD GB storage EBS Storage: none | 0 Instances | <input checked="" type="radio"/> On-demand <small>i</small> <input type="radio"/> Spot <small>i</small> Use on-demand as max price |

+ Add task instance group

Tutorial Hue

Hingga muncul tampilan berikut, lanjutkan dengan klik Next

The screenshot shows the AWS EMR console interface for creating a new cluster. The URL in the browser is <https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#create-cluster>. The page is titled "Step 3: General Cluster Settings".

Cluster name: My cluster

Logging: S3 folder: s3://aws-logs-146080089880-us-east-1/elasticmapred

Debugging:

Termination protection:

Tags: Add a key to create a tag

Additional Options:

- EMRFS consistent view
- Custom AMI ID: None

Bootstrap Actions: (This section is collapsed)

At the bottom right, there are "Cancel", "Previous", and "Next" buttons. The "Next" button is highlighted in blue.

Tutorial Hue

Pilih EC2 key-pair yang pernah dibua, kemudian Create Cluster

The screenshot shows the 'Create Cluster - Advanced Options' page in the AWS EMR console. On the left, a sidebar lists steps: Step 1: Software and Steps, Step 2: Hardware, Step 3: General Cluster Settings, and Step 4: Security (which is selected). The main area is titled 'Security Options'. It shows an 'EC2 key pair' dropdown set to 'golobok', which is highlighted with a red arrow. Below it is a checkbox for 'Cluster visible to all IAM users in account'. Under 'Permissions', there are three tabs: 'Default' (selected), 'Custom', and 'Default IAM roles'. A note says: 'Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.' Below this are sections for 'EMR role', 'EC2 instance profile', and 'Auto Scaling role', each with a dropdown menu and an info icon. At the bottom, there are links for 'Authentication and encryption' and 'EC2 security groups'. At the very bottom are 'Cancel', 'Previous', and 'Create cluster' buttons.

Tutorial Hue

Tunggu status **Starting** menjadi **Waiting**. Kemudian pilih **Hue**

The screenshot shows the AWS EMR console interface. On the left, there's a sidebar with options like Clusters, Security configurations, VPC subnets, Events, Notebooks, Help, and What's new. The main area displays a cluster named "My cluster" which is currently in a "Waiting" state, indicated by a green label. Below the cluster name, there are tabs for Summary, Application history, Monitoring, Hardware, Configurations, Events, Steps, and Bootstrap actions. Under the "Summary" tab, there are sections for Connections, Master public DNS, and Tags. The "Connections" section includes a link to "Hue". In the "Configuration details" section, it lists the release label as "emr-5.25.0", Hadoop distribution as "Amazon 2.8.5", and applications like Hive 2.3.5, Pig 0.17.0, Hue 4.4.0, Spark 2.4.3, HBase 1.4.9, and ZooKeeper 3.4.14. It also shows the Log URI as "s3://aws-logs-146080089880-us-east-1/elasticmapreduce/". The "Network and hardware" section shows the availability zone as "us-east-1f", subnet ID as "subnet-09404b06", and a master instance type of "m4.large Bootstrapping". The "Security and access" section includes a key name "golobok", EC2 instance profile "EMR_EC2_DefaultRole", and EMR role "EMR_DefaultRole". At the bottom, there are links for Feedback, English (US), and footer information including copyright, privacy policy, and terms of use.

Tutorial Hue

Apabila pilihan **Hue** tidak tersedia, ikuti tutorial berikut:

- Klik **Enable Web Connection** dan ikuti langkah-langkahnya

The screenshot shows the AWS Management Console for an Amazon EMR cluster. At the top, there's a navigation bar with tabs: Summary, Application history, Monitoring, Hardware, Configurations, Events, Steps, and Bootstrap actions. Below the navigation bar, there are sections for Connections, Master public DNS, and Tags. A modal window titled "Enable Web Connection" is open, containing the "Setup Web Connection" guide. The guide explains that Hadoop, Ganglia, and other applications publish user interfaces as web sites hosted on the master node. It also states that to reach these interfaces, an SSH tunnel must be established using either dynamic or local port forwarding. Below this, a section titled "Step 1: Open an SSH Tunnel to the Amazon EMR Master Node" provides a link to "Learn more". Underneath, there are two tabs: "Windows" and "Mac / Linux". A numbered list of 12 steps outlines the process of setting up PuTTY for the Windows tab. A mouse cursor is visible over the "Mac / Linux" tab.

Connections: [Enable Web Connection](#) – Hue, Spark History Server, HBase, Resource Manager ... (View All)

Master public DNS: ec2-18-205-56-147.compute-1.amazonaws.com SSH

Tags:

Enable Web Connection

Setup Web Connection

Hadoop, Ganglia, and other applications publish user interfaces as web sites hosted on the master node. For security reasons, these web sites are only available on the master node's local web server.

To reach the web interfaces, you must establish an SSH tunnel with the master node using either dynamic or local port forwarding. If you establish an SSH tunnel using dynamic port forwarding, you must also configure a proxy server to view the web interfaces.

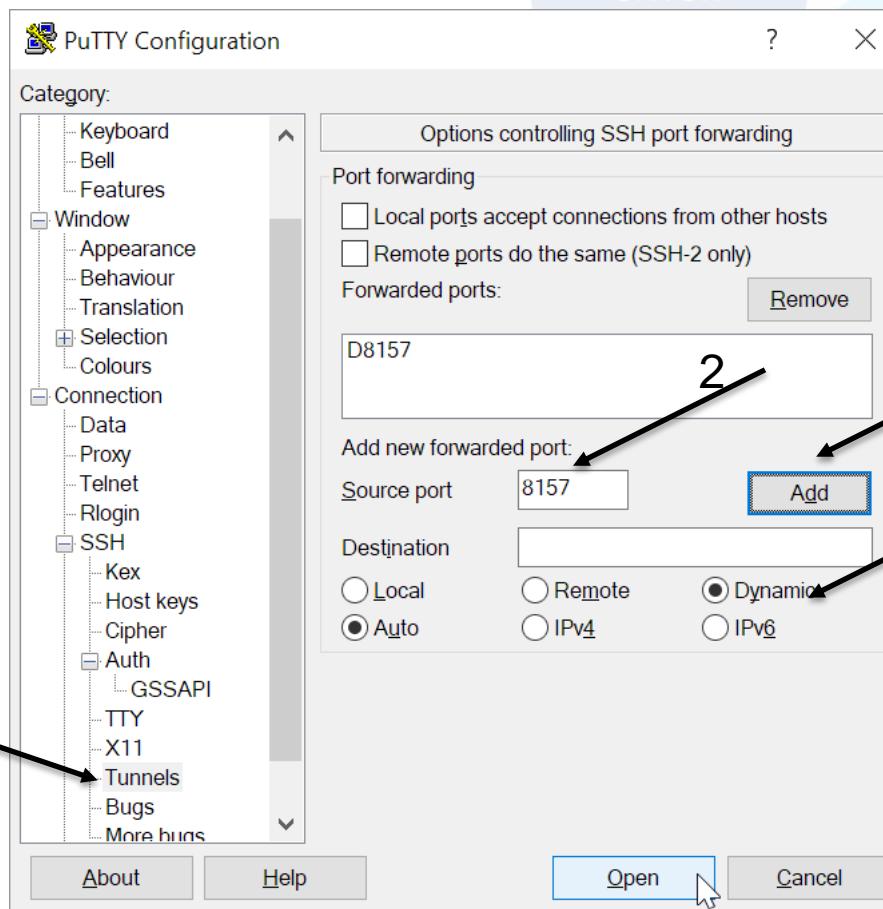
Step 1: Open an SSH Tunnel to the Amazon EMR Master Node - [Learn more](#)

Windows Mac / Linux

1. Download PuTTY.exe to your computer from:
<http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>
2. Start PuTTY.
3. In the Category list, click Session
4. In the Host Name field, type **hadoop@ec2-18-205-56-147.compute-1.amazonaws.com**
5. In the Category list, expand Connection > SSH > Auth
6. For Private key file for authentication, click Browse and select the private key file (**golobok.ppk**) used to launch the cluster.
7. In the Category list, expand Connection > SSH, and then click Tunnels.
8. In the Source port field, type **8157** (a randomly chosen, unused local port).
9. Select the Dynamic and Auto options.
10. Leave the Destination field empty and click Add.
11. Click Open.
12. Click Yes to dismiss the security alert.

Tutorial Hue

Buka putty dan masukkan konfigurasi seperti biasa,
Tambahkan pengaturan dibawah ini :

TERBUKA
UNTUK

Tutorial Hue

Buat file baru dengan nama “foxyproxy-settings.xml**”,
Buka dan paste script dibawah kemudian simpan.**

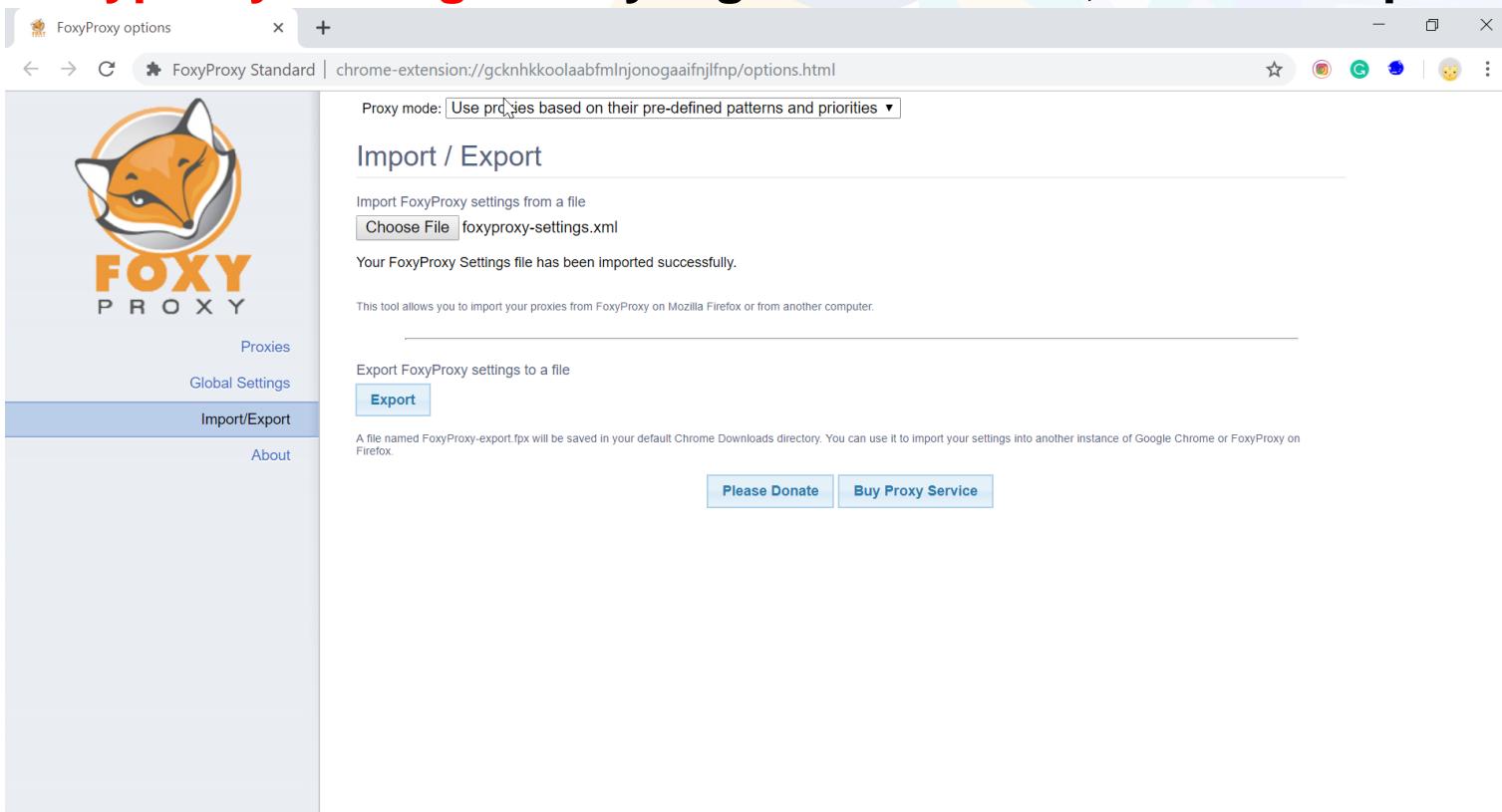
```
<?xml version="1.0" encoding="UTF-8"?>
<foxyproxy>
    <proxies>
        <proxy name="emr-socks-proxy" id="2322596116" notes="" fromSubscription="false" enabled="true"
mode="manual" selectedTabIndex="2" lastresort="false" animatedIcons="true" includeInCycle="true" color="#0055E5"
proxyDNS="true" noInternalIPs="false" autoconfMode="pac" clearCacheBeforeUse="false" disableCache="false"
clearCookiesBeforeUse="false" rejectCookies="false">
            <matches>
                <match enabled="true" name="*ec2*.amazonaws.com*" pattern="*ec2*.amazonaws.com*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*ec2*.compute*" pattern="*ec2*.compute*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="10.*" pattern="http://10.*" isRegEx="false" isBlackList="false"
isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*10*.amazonaws.com*" pattern="*10*.amazonaws.com*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*10*.compute*" pattern="*10*.compute*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*.compute.internal*" pattern="*.compute.internal*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*.ec2.internal*" pattern="*.ec2.internal*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
            </matches>
            <manualconf host="localhost" port="8157" socksversion="5" isSocks="true" username="" password=""
domain="" />
        </proxy>
    </proxies>
</foxyproxy>
```

Tutorial Hue

- Buka chrome, tambahkan ekstensi berikut :

<https://chrome.google.com/webstore/search/foxy%20proxy>

- Klik kanan ikon foxyproxy – options – Import/Export - Choose File - **foxyproxy-settings.xml** yang telah dibuat , kemudian pilih Add.



Tutorial Hue

Untuk pertama kali, buat username dan password yang akan digunakan untuk login kedalam Hue

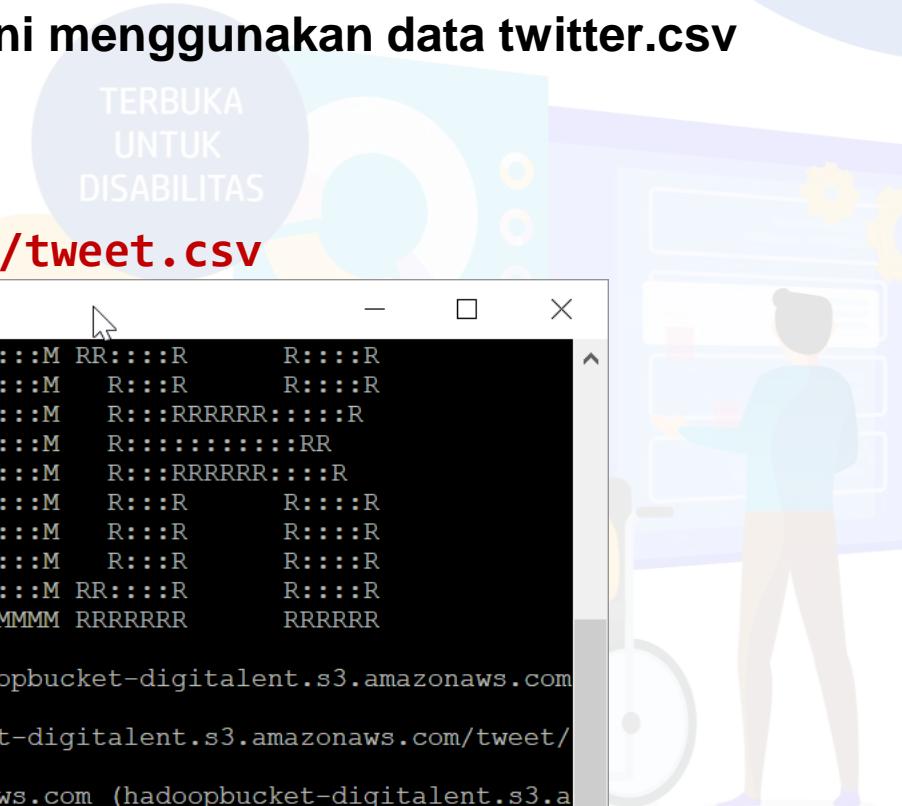
The screenshot shows a web browser window titled "Hue - Welcome to Hue". The address bar indicates the URL is "Not secure | ec2-18-207-240-52.compute-1.amazonaws.com:8888/hue/accounts/login?next=/". The main content is a "Create Account" form for Hue. At the top, there's a logo with the word "HUE" and the tagline "Query. Explore. Repeat.". Below this, a message states: "Since this is your first time logging in, pick any username and password. Be sure to remember these, as they will become your Hue superuser credentials." It also specifies password requirements: "The password must be at least 8 characters long, and must contain both uppercase and lowercase letters, at least one number, and at least one special character." There are two input fields: "Username" and "Password", followed by a large blue "Create Account" button. A small note at the bottom says: "Hue and the Hue logo are trademarks of Cloudera, Inc."

Tutorial Hue

Unduh data yang akan digunakan, disini menggunakan data `twitter.csv`

Ketikkan perintah berikut pada putty:

```
wget https://hadoopbucket-digitalent.s3.amazonaws.com/tweet/tweet.csv
```



```
hadoop@ip-172-31-61-7:~
```

```
E::::E      EEEEE M::::::M      M:::::::M RR:::R      R:::R
E::::E          M:::::M:::M    M:::M:::::M    R:::R    R:::R
E:::::EEEEEEEEE M:::::M M:::M M:::M M:::::M    R::::RRRRRR:::::R
E:::::::::::E   M:::::M M:::M:::M M:::M    R:::::::::::RR
E:::::EEEEEEEEE M:::::M    M:::::M    M:::::M    R::::RRRRRR:::::R
E:::::E        M:::::M    M:::::M    M:::::M    R:::R    R:::R
E:::::E        EEEEE  M:::::M      MMM    M:::::M    R:::R    R:::R
EE:::::EEEEEEEEE::::E M:::::M          M:::::M    R:::R    R:::R
E:::::::::::E     M:::::M          M:::::M    RR:::R    R:::R
EEEEEEEEEEEEEEEEEE  MMMMM    MMMMM RRRRRRR    RRRRRR

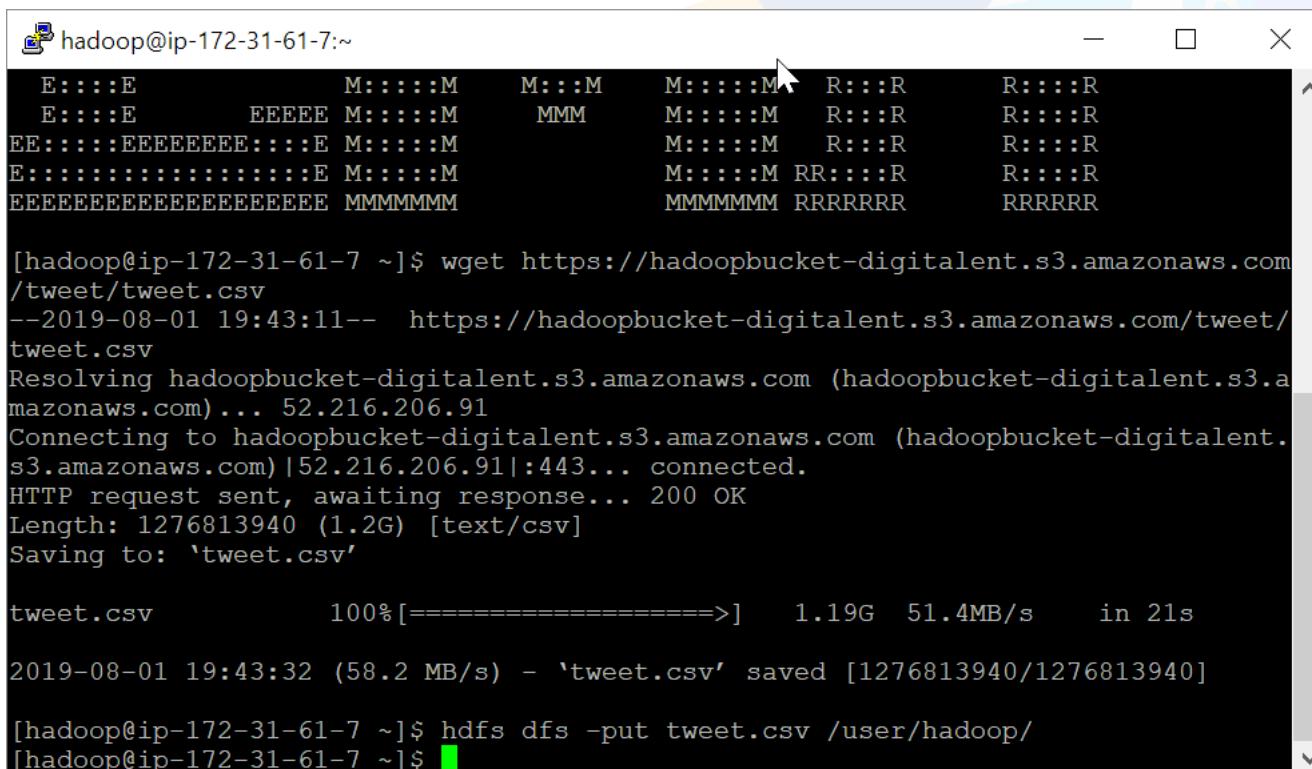
[hadoop@ip-172-31-61-7 ~]$ wget https://hadoopbucket-digitalent.s3.amazonaws.com/tweet/tweet.csv
--2019-08-01 19:43:11--  https://hadoopbucket-digitalent.s3.amazonaws.com/tweet/tweet.csv
Resolving hadoopbucket-digitalent.s3.amazonaws.com (hadoopbucket-digitalent.s3.amazonaws.com)... 52.216.206.91
Connecting to hadoopbucket-digitalent.s3.amazonaws.com (hadoopbucket-digitalent.s3.amazonaws.com)|52.216.206.91|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1276813940 (1.2G) [text/csv]
Saving to: 'tweet.csv'

tweet.csv      15% [==>] 191.14M 68.2MB/s
```

Tutorial Hue

Salin file twitter.csv kedalam /user/hadoop/:

Hdfs dfs -put tweet.csv /user/Hadoop/ TERBUKA



Tutorial Hue

Buka Hue kembali dan ketikkan perintah berikut:

Create schema twitter;

The screenshot shows the Hue Editor interface. In the top navigation bar, there are tabs for 'EMR - AWS Console' and 'Hue - Editor'. The URL in the browser is 'Not secure | ec2-18-205-56-147.compute-1.amazonaws.com:8888/hue/editor?editor=5'. The main area has a title 'TERBUKA UNTUK' with a blue circular graphic. On the left, there's a sidebar with icons for databases, tables, and jobs, and a section for 'Tables' which says '(0)'. The main query editor window has a title 'Hive' and a status bar showing '2.4s Database default Type text'. A query is being typed into the editor: 'create schema twitter;'. Below the editor, a log window shows the execution of the command: '2019-08-01 11:50:03 INFO : Starting task [Stage-0:DDL] in serial mode', 'INFO : Completed executing command(queryId=hive_20190801115003_9de6c0ce-0f35-490e-97a4-340fb128f2dc); Time taken: 0.305 seconds', and 'INFO : OK'. At the bottom of the editor window, a message says 'Success.'

Tutorial Hue

Create external table dengan perintah kueri dibawah:

<https://pastebin.com/6P1D7Pym>

The screenshot shows the Hue web interface running on AWS. The browser tab is titled "Hue - Editor". The URL is "ec2-18-205-56-147.compute-1.amazonaws.com:8888/hue/editor?editor=6". The interface has a sidebar on the left with icons for HDFS, HIVE, HBASE, and METASTORE. The main area is titled "Hive" and shows a query editor with the following code:

```
1 CREATE EXTERNAL TABLE twitter.tweet (created_at STRING,
2 in_reply_to_status_id STRING,
3 id_str STRING,
4 retweeted BOOLEAN,
5 in_reply_to_user_id_str STRING,
6 coordinates STRING,
7 retweet_count INT,
8 contributors STRING,
9 favorite_count INT,
10 favorited BOOLEAN,
11 in_reply_to_status_id_str STRING,
12 source STRING,
13 in_reply_to_user_id STRING,
14 user_id_str STRING,
15 user_screen_name STRING,
16 place STRING,
17 geo STRING,
18 text STRING)
19 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
20 LOCATION '/user/hadoop/' TBLPROPERTIES ("skip.header.line.count"="1");
```

Below the code, the logs show the execution results:

```
LOCATION '/user/hadoop/' TBLPROPERTIES ('skip.header.line.count' = '1')
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20190801195133_a77a91b5-9748-42cb-9
315-552c51764ae0); Time taken: 0.265 seconds
INFO : OK
```

Tutorial Hue

Kemudian load data kedalam tabel:

LOAD DATA INPATH 'hdfs:/user/hadoop/tweet.csv' INTO TABLE twitter(tweet;

The screenshot shows the Hue web interface running on a browser. The URL is <http://ec2-18-205-56-147.compute-1.amazonaws.com:8888/hue/editor?editor=7>. The interface includes a top navigation bar with tabs for EMR - AWS Console, Hue - Editor, and a search bar. On the left, there's a sidebar for the 'default' database showing tables and a 'No entries found' message. The main area is titled 'Hive' and contains a query editor with the following code:

```
4.7s Database default Type text
1 LOAD DATA INPATH 'hdfs:/user/hadoop/tweet.csv' INTO TABLE twitter(tweet;
```

Below the query editor, the output window shows the execution log:

```
hdfs:/user/hadoop/tweet.csv
INFO : Starting task [Stage-1:STATS] in serial mode
INFO : Completed executing command(queryId=hive_20190801195413_cd25d529-8f4d-4096-8
b27-0aa1147cfdf); Time taken: 1.855 seconds
INFO : OK
```

Tutorial Hue

Coba kueri:

select count(*) from tweet;

The screenshot shows the Hue web interface for Apache Hive. The top navigation bar includes tabs for 'EMR - AWS Console' and 'Hue - Editor'. The URL is 'ec2-18-205-56-147.compute-1.amazonaws.com:8888/hue/editor?editor=8#id=application_1564687472748_0003'. The main area is titled 'HUE' and shows a 'Hive' query editor. The query 'select count(*) from tweet;' has been run, taking 30.85s and returning 1 result. The results table shows a single row with value '_c0'. The right sidebar displays 'Jobs 1' and 'Tables' (No tables identified). A large watermark 'TERBUKA UNTUK' is visible in the background.

```
select count(*) from tweet;
```

30.85s Database twitter Type text

| _c0 |
|------------|
| 1 11262928 |

Tutorial Hue

Coba kueri:

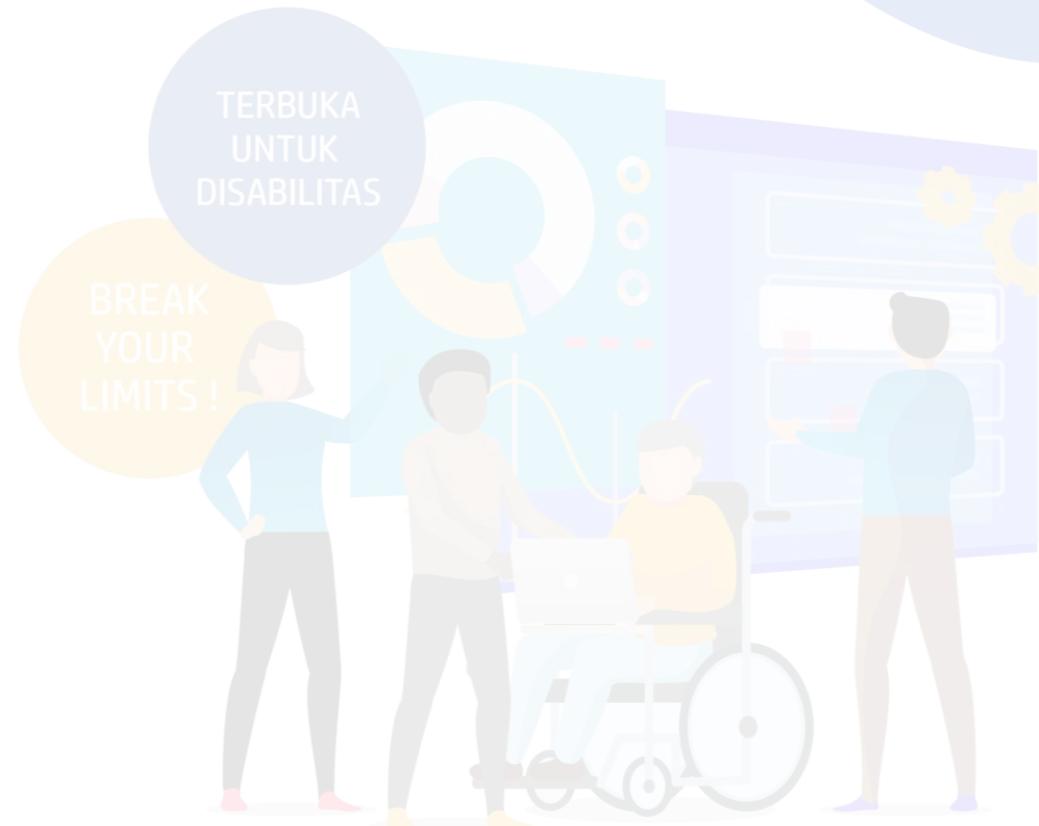
select place from tweet where place like 'a%' limit 10;

The screenshot shows the Hue web interface running on a Windows desktop. The browser title bar reads "EMR - AWS Console" and "Hue - Editor". The URL is "ec2-18-205-56-147.compute-1.amazonaws.com:8888/hue/editor?editor=10#!id=application_1564687472748_0003". The interface includes a sidebar with database management tools and a central pane for querying data. The query results show 10 entries under the column "place", all of which are "amoviepitchbot".

| place |
|-------------------|
| 1 amoviepitchbot |
| 2 amoviepitchbot |
| 3 amoviepitchbot |
| 4 aliasgenerator |
| 5 asklistfeedy |
| 6 ajtheparkour |
| 7 ayatsujibot |
| 8 arenatanbot |
| 9 arnarleir |
| 10 amoviepitchbot |

Penyedia Otentikasi

1. **LDAP**
2. **PAM**
3. **SPNEGO**
4. **OpenID**
5. **Oauth**
6. **SAML2**



Tugas Individu

1. Buatlah rangkuman materi dengan cara berikut:

- Lalukan ulang seperti yang ada di All slide
- Cek plagiasi diturnitin (tiap minggu) dari hasil rangkuman tersebut
 - > Register ke turnitin
 - > Masukkan **id class**: 21563495 & **enrool key**: filkomub9302



DIGITAL
TALENT
SCHOLARSHIP



DIGITAL TALENT SCHOLARSHIP 2019

Big Data Analytics



Terimakasih

Oleh: Imam Cholissodin | imamcs@ub.ac.id, Putra Pandu Adikara, Sufia Adha Putri

Asisten: Guedho, Sukma, Anshori, Aang dan Gusti

Fakultas Ilmu Komputer (Filkom) Universitas Brawijaya (UB)