

1. Create an EMR cluster with Hadoop, HBase and Sqoop

Amazon EMR > EMR on EC2: Clusters > My cluster

My cluster

Updated 8 minutes ago

Actions

▼ Summary

Cluster info

Cluster ID
j-2XQXCUIIN89W90

Cluster configuration
Instance groups

Capacity
1 Primary 0 Core 0 Task

Applications

Amazon EMR version
emr-6.13.0

Installed applications
HBase 2.4.17, Hadoop 3.3.3, Sqoop 1.4.7

Cluster management

Log destination in Amazon S3
[aws-logs-940702677798-us-east-1/elasticmapreduce](#)

Persistent application UIs
[YARN timeline server](#)

Primary node public DNS
[ec2-3-80-89-67.compute-1.amazonaws.com](#)
[Connect to the Primary Node using SSH](#)

Status and time

Status
Waiting

Creation time
29 September 2023 14:28 (UTC+05:30)

Elapsed time
15 minutes, 26 seconds

Properties

Bootstrap actions

Instances (hardware)

Steps

Applications

Configurations

Monitoring

Events

Tags (0)

Operating system

Amazon Linux release 2.0.20230808.0

Cluster logs

Archive log files to Amazon S3
Turned on

Amazon S3 location
[s3://aws-logs-940702677798-us-east-1/elasticmapreduce/](#)
[Turn on encryption for logs](#)

Cluster termination

Edit cluster termination

Termination option
Manually terminate cluster

Idle time
-

2. Create a free tier RDS Instance

RDS > Databases > database-1

database-1

ModifyActions ▼

Summary

DB identifier database-1	CPU <div>2.87%</div>	Status Available	Class db.t3.micro
Role Instance	Current activity <div>0 Connections</div>	Engine MySQL Community	Region & AZ us-east-1a

Connectivity & security

Monitoring

Logs & events

Configuration

Maintenance & backups

Tags

Connectivity & security

Endpoint & port Endpoint database-1.c10i7rbsadgr.us-east-1.rds.amazonaws.com Port 3306	Networking Availability Zone us-east-1a VPC vpc-0dd43d567aa42e244 Subnet group rds-ec2-db-subnet-group-1	Security VPC security groups rds-ec2-6 (sg-0f22edec0ac55cc55) Active Publicly accessible No Certificate authority info rds-ca-2019
--	--	---

3. Connect to the primary node of the EMR Cluster using SSH

```
(base) hemanthreddyk@Hemanths-MacBook-Air Personal % ssh -i upgrad-lab-key.pem hadoop@ec2-35-172-214-96.compute-1.amazonaws.com
The authenticity of host 'ec2-35-172-214-96.compute-1.amazonaws.com (35.172.214.96)' can't be established.
ED25519 key fingerprint is SHA256:3G8Pj+HtJSsr6ZyznzASQ0p5pVsb1Mu0ZjsA99T/p6g.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-35-172-214-96.compute-1.amazonaws.com' (ED25519) to the list of known hosts.
Last login: Wed Sep 27 19:37:15 2023

  __|  __|_  )
  _| (  /   Amazon Linux 2 AMI
 ---|\\___|___|

https://aws.amazon.com/amazon-linux-2/
94 package(s) needed for security, out of 142 available
Run "sudo yum update" to apply all updates.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory

EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRRRRRRRRR
E:EEEEEEEEEEEEEEEE M:EEEE M:EEEE R:EEEEEEEEEEEE
EE:EEEEEEEEEEEEEEEE M:EEEE M:EEEE R:RRRRRRRRRRRR
E:EE EEEEE M:EEEE M:EEEE RR:R R:R
E:EE M:EEEE M:EEEE M:EEEE R:R R:R
E:EEEEEEEEEEEE M:EE M:EE M:EE M:EE R:RRRRRRRRRR
E:EEEEEEEEEEEE M:EE M:EE M:EE M:EE R:RRRRRRRRRR
E:EEEEEEEEEEEE M:EE M:EE M:EE M:EE R:RRRRRRRRRR
E:EE M:EEEE M:EE M:EE M:EE R:R R:R
E:EE EEEEE M:EE M:EE M:EE R:R R:R
EE:EEEEEEEEEEEE M:EE M:EE M:EE M:EE R:R R:R
E:EEEEEEEEEEEEEEEE M:EE M:EE M:EE M:EE R:R R:R
EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRR RRRRRR

[hadoop@ip-172-31-16-22 ~]$
```

4. Download the required csv files into the EMR cluster

In this activity we need “yellow_tripdata_2017-01.csv” and “yellow_tripdata_2017-02.csv” files, let’s put them at following path ‘/home/hadoop/YellowCabData’

```
mkdir YellowCabsData
```

```
wget -P /home/hadoop/YellowCabsData
```

```
https://nyc-tlc-upgrad.s3.amazonaws.com/yellow\_tripdata\_2017-01.csv
```

```
wget -P /home/hadoop/YellowCabsData
```

```
https://nyc-tlc-upgrad.s3.amazonaws.com/yellow\_tripdata\_2017-02.csv
```

```
ls /home/hadoop/YellowCabsData/
```

```
[hadoop@ip-172-31-16-22 ~]$ mkdir YellowCabData
[hadoop@ip-172-31-16-22 ~]$ cd YellowCabData/
[hadoop@ip-172-31-16-22 YellowCabData]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
--2023-09-27 19:43:13-- https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 54.231.132.1, 52.217.236.57, 3.5.29.194, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|54.231.132.1|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 914029540 (872M) [text/csv]
Saving to: 'yellow_tripdata_2017-01.csv'

100%[=====>] 914,029,540 27.8MB/s in 32s

2023-09-27 19:43:45 (27.5 MB/s) - 'yellow_tripdata_2017-01.csv' saved [914029540/914029540]

[hadoop@ip-172-31-16-22 YellowCabData]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
--2023-09-27 19:43:51-- https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 52.217.163.249, 54.231.235.241, 52.216.60.57, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|52.217.163.249|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 863487050 (823M) [text/csv]
Saving to: 'yellow_tripdata_2017-02.csv'

100%[=====>] 863,487,050 27.1MB/s in 31s

2023-09-27 19:44:22 (27.0 MB/s) - 'yellow_tripdata_2017-02.csv' saved [863487050/863487050]

[hadoop@ip-172-31-16-22 YellowCabData]$ ls
yellow_tripdata_2017-01.csv yellow_tripdata_2017-02.csv
[hadoop@ip-172-31-16-22 YellowCabData]$
```

5. Connect to mysql running in RDS from the EMR cluster using the following command and Create a database named ‘YellowTaxiTripsDB’

mysql -h database-1.c10i7rbsadgr.us-east-1.rds.amazonaws.com -P 3306 -u admin -p

CREATE DATABASE YellowTaxiTripsDB;

USE YellowTaxiTripsDB;

```
[hadoop@ip-172-31-16-22 YellowCabData]$ mysql -h database-1.c10i7rbsadgr.us-east-1.rds.amazonaws.com -P 3306 -u admin -p
Enter password:
Welcome to the MariaDB monitor. Commands end with ; or \g.
Your MySQL connection id is 18
Server version: 8.0.33 Source distribution

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]> CREATE DATABASE YellowTaxiTripsDB;
Query OK, 1 row affected (0.01 sec)

MySQL [(none)]> use YellowTaxiTripsDB;
Database changed
MySQL [YellowTaxiTripsDB]>
```

6. Now create a table 'tripdata' that can hold all the columns present in the csv files with appropriate data types

```
CREATE TABLE tripdata (  
  VendorID INT,  
  tpep_pickup_datetime VARCHAR(20),  
  tpep_dropoff_datetime VARCHAR(20),  
  passenger_count INT,  
  trip_distance FLOAT,  
  RatecodeID INT,  
  store_and_fwd_flag CHAR(1),  
  PULocationID INT,  
  DOLocationID INT,  
  payment_type INT,  
  fare_amount FLOAT,  
  extra FLOAT,  
  mta_tax FLOAT,  
  tip_amount FLOAT,  
  tolls_amount FLOAT,  
  improvement_surcharge FLOAT,  
  total_amount FLOAT,  
  congestion_surcharge FLOAT DEFAULT 0.0,  
  airport_fee FLOAT DEFAULT 0.0  
);
```

```

MySQL [YellowTaxiTripsDB]> CREATE TABLE tripdata (
->   VendorID INT,
->   tpep_pickup_datetime VARCHAR(20),
->   tpep_dropoff_datetime VARCHAR(20),
->   passenger_count INT,
->   trip_distance FLOAT,
->   RatecodeID INT,
->   store_and_fwd_flag CHAR(1),
->   PULocationID INT,
->   DOLocationID INT,
->   payment_type INT,
->   fare_amount FLOAT,
->   extra FLOAT,
->   mta_tax FLOAT,
->   tip_amount FLOAT,
->   tolls_amount FLOAT,
->   improvement_surcharge FLOAT,
->   total_amount FLOAT,
->   congestion_surcharge FLOAT DEFAULT 0.0,
->   airport_fee FLOAT DEFAULT 0.0
-> );

```

Query OK, 0 rows affected (0.04 sec)

```

MySQL [YellowTaxiTripsDB]> describe tripdata;

```

Field	Type	Null	Key	Default	Extra
VendorID	int	YES		NULL	
tpep_pickup_datetime	varchar(20)	YES		NULL	
tpep_dropoff_datetime	varchar(20)	YES		NULL	
passenger_count	int	YES		NULL	
trip_distance	float	YES		NULL	
RatecodeID	int	YES		NULL	
store_and_fwd_flag	char(1)	YES		NULL	
PULocationID	int	YES		NULL	
DOLocationID	int	YES		NULL	
payment_type	int	YES		NULL	
fare_amount	float	YES		NULL	
extra	float	YES		NULL	
mta_tax	float	YES		NULL	
tip_amount	float	YES		NULL	
tolls_amount	float	YES		NULL	
improvement_surcharge	float	YES		NULL	
total_amount	float	YES		NULL	
congestion_surcharge	float	YES		0	
airport_fee	float	YES		0	

19 rows in set (0.00 sec)

```

MySQL [YellowTaxiTripsDB]>

```

7. Load the data from csv files into tripdata table

LOAD DATA LOCAL INFILE

*'/home/hadoop/YellowCabData/yellow_tripdata_2017-01.csv' INTO TABLE tripdata
FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;*

LOAD DATA LOCAL INFILE

*'/home/hadoop/YellowCabData/yellow_tripdata_2017-02.csv' INTO TABLE tripdata
FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;*

select count() from tripdata;*

```
MySQL [YellowTaxiTripsDB]> LOAD DATA LOCAL INFILE '/home/hadoop/YellowCabData/yellow_tripdata_2017-01.csv' INTO TABLE tripdata FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (2 min 35.27 sec)
Records: 9710820 Deleted: 0 Skipped: 0 Warnings: 19421640

MySQL [YellowTaxiTripsDB]> LOAD DATA LOCAL INFILE '/home/hadoop/YellowCabData/yellow_tripdata_2017-02.csv' INTO TABLE tripdata FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (2 min 46.95 sec)
Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 18339550

MySQL [YellowTaxiTripsDB]> select count(*) from tripdata;
+-----+
| count(*) |
+-----+
| 18880595 |
+-----+
1 row in set (52.07 sec)

MySQL [YellowTaxiTripsDB]>
```

Therefore we loaded **18880595** records into **tripdata** table