

E-Commerce Assignment

Problem Statement:

With online sales gaining popularity, tech companies are exploring ways to improve their sales by analyzing customer behavior and deriving insights about product trends. Furthermore, the e-commerce websites make it easier for customers to find the products that they require without much scavenging. The role of big data engineers is among the most sought-after job profiles of this decade. Therefore, as part of this assignment, you, as a big data engineer, will be required to extract data and gather insights from a real-life data set of an e-commerce company

Stepwise Solution to the problem statement:

Step 1: An EMR cluster was created with the desired specifications:

Clone

Terminate

AWS CLI export

⚠️ Auto-termination is not available for this account when using this release of EMR.

Cluster: Ecomm_Assignment Starting Configuring cluster software

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Summary

ID: j-OTLLD04TOMRM

Creation date: 2022-01-23 13:19 (UTC+5:30)

Elapsed time: 6 minutes

After last step completes: Cluster waits

Termination protection: Off [Change](#)

Tags: -- [View All](#) / [Edit](#)

Master public DNS: ec2-52-91-152-244.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

Configuration details

Release label: emr-5.24.0

Hadoop distribution: Amazon 2.8.5

Applications: Ganglia 3.7.2, Hive 2.3.4, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.1

Log URI: s3://aws-logs-943198395872-us-east-1/elasticmapreduce/ [📄](#)

EMRFS consistent view: Disabled

Custom AMI ID: --

Application user interfaces

On-cluster user interfaces [🔗](#): Not Enabled [Enable an SSH Connection](#)

Security and access

Key name: wh_assignment

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

Security groups for Master: [sg-09b60ede4624952a2](#) [🔗](#) (ElasticMapReduce-master)

Security groups for Core & Task: [sg-0d819428d2945fbc4](#) [🔗](#) (ElasticMapReduce-slave)

Network and hardware

Availability zone: us-east-1e

Subnet ID: [subnet-ed0e73dc](#) [🔗](#)

Master: Bootstrapping 1 m4.large

Core: Provisioning 1 m4.large

Task: --

Cluster scaling: Not enabled

[illegible]

```
[hadoop@ip-172-31-51-14 ~]$ wget https://e-commerce-events-ml.s3.amazonaws.com/2019-Oct.csv
--2022-01-23 08:19:43-- https://e-commerce-events-ml.s3.amazonaws.com/2019-Oct.csv
Resolving e-commerce-events-ml.s3.amazonaws.com (e-commerce-events-ml.s3.amazonaws.com)... 52.217.106.140
Connecting to e-commerce-events-ml.s3.amazonaws.com (e-commerce-events-ml.s3.amazonaws.com)[52.217.106.140]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 482542278 (460M) [text/csv]
Saving to: '2019-Oct.csv'

2019-Oct.csv          100%[=====>] 460.19M  48.1MB/s  in 9.5s

2022-01-23 08:19:53 (48.3 MB/s) - '2019-Oct.csv' saved [482542278/482542278]

[hadoop@ip-172-31-51-14 ~]$ wget https://e-commerce-events-ml.s3.amazonaws.com/2019-Nov.csv
--2022-01-23 08:20:13-- https://e-commerce-events-ml.s3.amazonaws.com/2019-Nov.csv
Resolving e-commerce-events-ml.s3.amazonaws.com (e-commerce-events-ml.s3.amazonaws.com)... 52.217.8.68
Connecting to e-commerce-events-ml.s3.amazonaws.com (e-commerce-events-ml.s3.amazonaws.com)[52.217.8.68]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 545839412 (521M) [text/csv]
Saving to: '2019-Nov.csv'

2019-Nov.csv          100%[=====>] 520.55M  52.9MB/s  in 11s

2022-01-23 08:20:23 (48.6 MB/s) - '2019-Nov.csv' saved [545839412/545839412]
```

Step 5: Loading data into the required tables