

CIS 602: Big Data – Homework 5

Pradyoth Singenahalli Prabhu - 02071847

Task 1: Launching an Amazon EMR cluster

Creating EMR cluster:

The screenshot shows the 'Create cluster' wizard on the 'Name and applications' step. The 'Name' field is set to 'Hive EMR Cluster'. The 'Amazon EMR release' dropdown is set to 'emr-5.29.0'. Under 'Application bundle', the 'Custom' option is selected. The 'AWS Glue Data Catalog settings' section is collapsed. The 'Custom Amazon Machine Image (AMI)' dropdown is empty. A checkbox for 'Update all installed packages on reboot' is checked. The 'Cluster configuration' section is expanded, showing 'Primary (m4.large), Core (m4.large)' instance groups. The 'Cluster scaling and provisioning' section shows 'Core size: 1 instance'. The 'Networking' section is collapsed. A 'Configure IAM roles' box is present, stating 'You must choose a service role and instance profile before you create this cluster.' A 'Choose IAM roles' button is available. The 'Create cluster' button is highlighted in orange.

The screenshot shows the 'Create cluster' wizard on the 'Cluster configuration' step. The 'Primary' section shows 'm4.large' instances with 2 vCores and 8 GiB memory. The 'Core' section shows 'm4.large' instances with 2 vCores and 8 GiB memory. A 'Node configuration - optional' section is present. The 'EBS root volume - optional' section allows adding up to 48 more task instance groups. The 'Cluster scaling and provisioning' section includes a note about using EMR-managed scaling for releases higher than 5.30.0. The 'Provisioning configuration' section notes that Amazon EMR attempts to provision capacity when launching the cluster. The 'Summary' tab is active, showing the cluster configuration details. A 'Configure IAM roles' box is present, stating 'You must choose a service role and instance profile before you create this cluster.' A 'Choose IAM roles' button is available. The 'Create cluster' button is highlighted in orange.

Creating EMR Key:

The screenshot shows the 'Create key pair' wizard in the AWS EC2 console. The 'Key pair' section is active, displaying the following details:

- Name:** EMR-Key
- Key pair type:** RSA (selected)
- Private key file format:** OpenSSH (selected)
- Tags - optional:** No tags associated with the resource.

At the bottom right, there is a large orange 'Create key pair' button.

Successfully Created the EMR-Key:

The screenshot shows the 'Key pairs' list in the AWS EC2 console. The table displays the following data:

Name	Type	Created	Fingerprint	ID
vockey	rsa	2023/10/28 18:42 GMT-4	94:4da2:11:77:27:93:24:14:df:e9:be:d8...	key-037dbc1801aec55
EMR-Key	rsa	2023/10/28 19:22 GMT-4	8e:11:0e:9c:17:ba:d1:0d:09:d0:19:b7:f4...	key-0a945b5c5d8b11402d

AWS CloudShell

Services Search [Option+S]

Summary Info

Name and applications

Name: Hive EMR Cluster
Amazon EMR release: emr-5.29.0
Application bundle: Custom (Hadoop 2.8.5, Hive 2.3.6)

Cluster configuration

Instance groups: Primary (m4.large), Core (m4.large)
Cluster scaling and provisioning: Provisioning configuration, Core size: 1 instance
Networking

Create cluster

Security configuration and EC2 key pair - optional Info

Security configuration: Select your cluster encryption, authentication, authorization, and instance metadata service settings.
Choose a security configuration, Browse, Create security configuration

Amazon EC2 key pair for SSH to the cluster: Info
EMR-Key, Choose, Browse, Create key pair

Identity and Access Management (IAM) roles Info

The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.

Choose an existing service role: Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.
 Create a service role: Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.

Service role: LabRole

EC2 instance profile for Amazon EMR
The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.

Choose an existing instance profile: Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.
 Create an instance profile: Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

Instance profile: EMR_EC2_DefaultRole

Custom automatic scaling role - optional
When a custom automatic scaling rule triggers, Amazon EMR assumes this role to add and terminate EC2 instances. Learn more

Custom automatic scaling role: EMR_AutoScaling_DefaultRole

CloudShell Feedback © 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Created the cluster successfully:

Your cluster "Hive EMR Cluster" has been successfully created.

Hive EMR Cluster

Summary

Cluster info	Applications	Cluster management	Status and time
Cluster ID: j-2HQYIL2BWCSJ Cluster configuration: Instance groups Capacity: 1 Primary 1 Core 0 Task	Amazon EMR version: emr-5.29.0 Installed applications: Hadoop 2.8.5, Hive 2.5.6	Log destination in Amazon S3: aws-logs-428441136833-us-east-1/elastictmapreduce Primary node public DNS: -	Status: Starting Creation time: October 28, 2023, 20:44 (UTC-04:00) Elapsed time: 0 seconds

Properties Bootstrap actions | Instances (Hardware) | Steps | Applications | Configurations | Monitoring | Events | Tags (0)

Cluster logs Info

Archive log files to Amazon S3 Turned on Amazon S3 location: s3://aws-logs-428441136833-us-east-1/elastictmapreduce/	Encryption for logs Turned off
--	-----------------------------------

Cluster termination Info

Termination option: Manually terminate cluster Idle time: -	Termination protection: Turned off
--	------------------------------------

Network and security Info

Network: Virtual Private Cloud (VPC) Subnet(s) and Availability Zone(s): AZ subnet-00bf5f5af223b9ae (us-east-1a)	Security configuration: None EC2 key pair: EMR-Key	Permissions: Service role for Amazon EMR LabRole EC2 instance profile: EMR_EC2_DefaultRole Custom automatic scaling role: EMR_AutoScaling_DefaultRole
---	---	--

CloudShell Feedback © 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Configure the security group for the main node to allow SSH connections:

Screenshot of the AWS EC2 Security Groups page showing the details of a security group named "sg-0c623df92bfe8905d - ElasticMapReduce-master".

Details

Security group name ElasticMapReduce-master	Security group ID sg-0c623df92bfe8905d	Description Master group for Elastic MapReduce created on 2023-10-23T20:16:20.703Z	VPC ID vpc-0485d570e597607e2
Owner 42844136833	Inbound rules count 7 Permission entries	Outbound rules count 1 Permission entry	

Inbound rules (7)

Name	Security group rule ID	IP version	Type	Protocol	Port range	Source	Description
-	sgr-0da11ac5f815ba65a	-	Custom TCP	TCP	8443	sg-03e8c9f8414a795f...	-
-	sgr-0838838d95d94c...	-	All ICMP - IPv4	ICMP	All	sg-03e8c9f8414a795f...	-
-	sgr-0e50d70e4085ba...	-	All UDP	UDP	0 - 65535	sg-0c623df92bfe8905...	-
-	sgr-0bffa60ce71b96107	-	All UDP	UDP	0 - 65535	sg-03e8c9f8414a795f...	-
-	sgr-0e650421ceea97b7	-	All ICMP - IPv4	ICMP	All	sg-0c623df92bfe8905...	-
-	sgr-01a4379c7b3681be0	-	All TCP	TCP	0 - 65535	sg-03e8c9f8414a795f...	-
-	sgr-0bb0d61bfa7aec72b	-	All TCP	TCP	0 - 65535	sg-0c623df92bfe8905...	-

Screenshot of the "Edit inbound rules" page for the same security group.

Edit inbound rules

Inbound rules control the incoming traffic that's allowed to reach the instance.

Inbound rules info

Security group rule ID	Type	Protocol	Port range	Source	Description
sgr-0da11ac5f815ba65a	Custom TCP	TCP	8443	Custom	pl-f8bd5e91
sgr-0838838d95d94c...	All ICMP - IPv4	ICMP	All	Custom	sg-03e8c9f8414a795f...
sgr-0e50d70e4085ba849	All UDP	UDP	0 - 65535	Custom	sg-0c623df92bfe8905d
sgr-0bffa60ce71b96107	All UDP	UDP	0 - 65535	Custom	sg-03e8c9f8414a795f...
sgr-0e650421ceea97b7	All ICMP - IPv4	ICMP	All	Custom	sg-0c623df92bfe8905d
sgr-01a4379c7b3681be0	All TCP	TCP	0 - 65535	Custom	sg-03e8c9f8414a795f...
sgr-0bb0d61bfa7aec72b	All TCP	TCP	0 - 65535	Custom	sg-0c623df92bfe8905d
-	SSH	TCP	22	Anywhere-IPv4	0.0.0.0/0

Rules warning: Rules with source of 0.0.0.0/0 or ::/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only.

Buttons: Cancel, Preview changes, Save rules.

Successfully Created:

Inbound security group rules successfully modified on security group (sg-0c623df92bfe8905d) | ElasticMapReduce-master

sg-0c623df92bfe8905d - ElasticMapReduce-master

Details

Security group name ElasticMapReduce-master	Security group ID sg-0c623df92bfe8905d	Description Master group for Elastic MapReduce created on 2023-10-23T20:16:20.70Z	VPC ID vpc-0485d570e597607e2
Owner 428444136833	Inbound rules count 8 Permission entries	Outbound rules count 1 Permission entry	

Inbound rules [8]

Name	Security group rule...	IP version	Type	Protocol	Port range	Source	Description
-	sgr-0d811ac5f8158a65a	-	Custom TCP	TCP	8443	pl-f8bd5e91	-
-	sgr-08388369594c...	-	All ICMP - IPv4	ICMP	All	sg-03ebc9f8414a795f...	-
-	sgr-0e30d070e40858a...	-	All UDP	UDP	0 - 65535	sg-0x623df92bfe8905...	-
-	sgr-0ffafaa0ec1b96107	-	All UDP	UDP	0 - 65535	sg-03ebc9f8414a795f...	-
-	sgr-0e650421cecc97b7	-	All ICMP - IPv4	ICMP	All	sg-0x623df92bfe8905...	-
-	sgr-01a4379c7b55681b...	-	All TCP	TCP	0 - 65535	sg-03ebc9f8414a795f...	-
-	sgr-08b0c61bf97aeec72b	-	All TCP	TCP	0 - 65535	sg-0x623df92bfe8905...	-
-	sgr-095dadaa705e7f77d	IPv4	SSH	TCP	22	0.0.0.0/0	-

Successfully Created EMR Cluster (Hive EMR Cluster):

Status changed to Waiting.

Your cluster "Hive EMR Cluster" has been successfully created.

Hive EMR Cluster

Summary

Cluster info	Applications	Cluster management	Status and time
Cluster ID j-2HQYIIL2BWCSJ	Amazon EMR version emr-5.29.0	Log destination in Amazon S3 aws-logs-428444136833-us-east-1/elasticmapreduce	Status Waiting
Cluster configuration	Installed applications	Persistent application UIs	Creation time
Instance groups	Hadoop 2.8.5, Hive 2.5.6	Primary node public DNS ec2-54-234-110-5.compute-1.amazonaws.com	October 28, 2023, 20:44 (UTC-04:00)
Capacity		Connect to the Primary node using SSH	Elapsed time
1 Primary 1 Core 0 Task		Connect to the Primary node using SSM	6 minutes, 54 seconds

Properties **Bootstrap actions** **Instances (Hardware)** **Steps** **Applications** **Configurations** **Monitoring** **Events** **Tags (0)**

Cluster logs Info

Archive log files to Amazon S3 Turned on	Encryption for logs Turned off
Amazon S3 location s3://aws-logs-428444136833-us-east-1/elasticmapreduce/	

Cluster termination Info

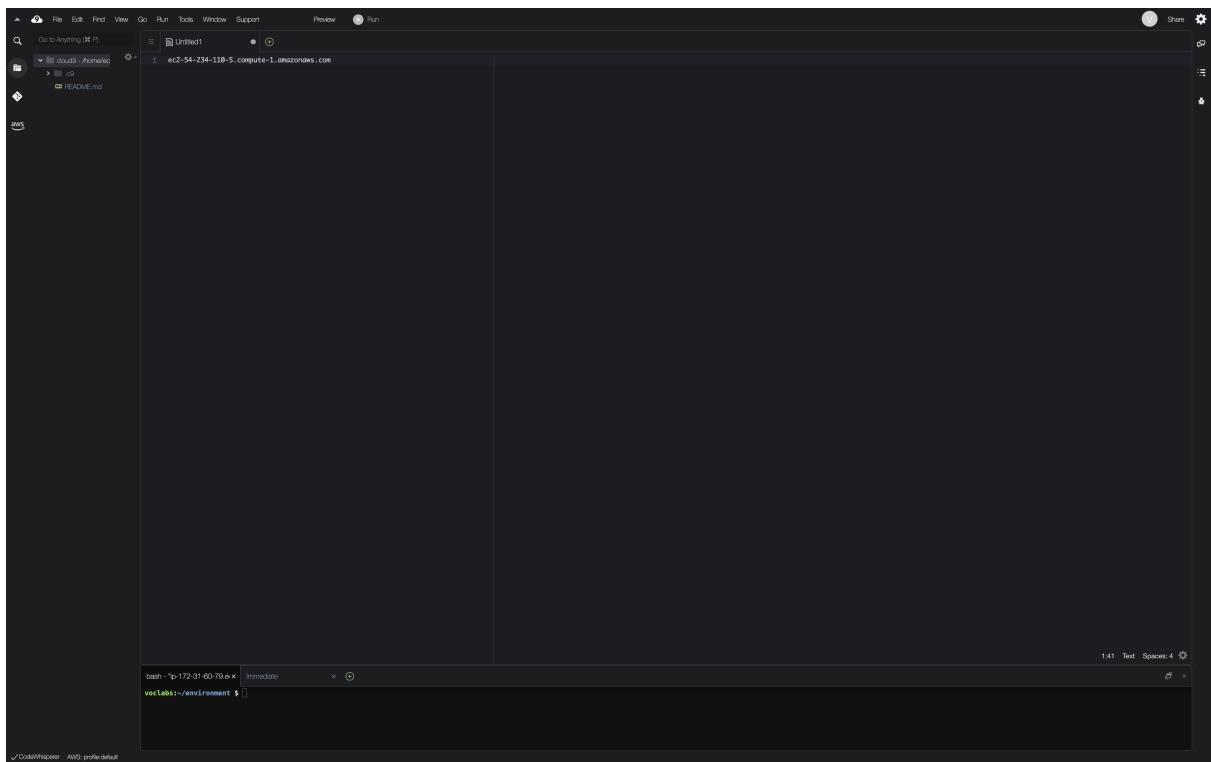
Termination option Manually terminate cluster	Termination protection Turned off
Idle time -	

Network and security Info

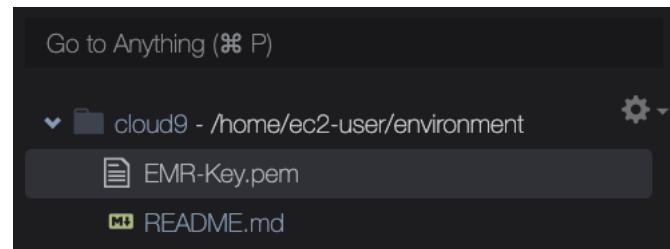
Network	Security configuration	Permissions
Virtual Private Cloud (VPC) vpc-0485d570e597607e2	Security configuration None	Service role for Amazon EMR LambdaRole
Subnets (s) and Availability Zones (AZ) subnet-000f5f5fa0f2339bar us-east-1	EC2 key pair EMR_EKey	EC2 instance profile EMR_EC2_DefaultRole
▼ EC2 security groups (firewall)		Custom automatic scaling role EMR_AutoScaling_DefaultRole
Primary node		
EMR managed security group sg-0x623df92bfe8905d		
Additional security groups		
Core and task nodes		
EMR managed security group sg-03ebc9f8414a795f		
Additional security groups		

Task 2: Connecting to the Hadoop main node by using SSH

Copied Primary node public DNS: ec2-54-234-110-5.compute-1.amazonaws.com



Uploading the EMR-Key:



Modify the permissions on the key pair file so that the SSH software will use it:

Command:

```
chmod 400 EMR-Key.pem
```

Establish an SSH connection to the cluster's main node:

Bash Command:

```
ssh -i EMR-Key.pem hadoop@ec2-54-234-110-5.compute-1.amazonaws.com
```

Successfully established the connection:

```

ssh-keygen -t rsa -b 4096
cat id_rsa.pub | ssh-copy-id ec2-54-234-110-5.compute-1.amazonaws.com
echo "ec2-54-234-110-5.compute-1.amazonaws.com" > /etc/hosts

```

Task 3: Running Hive interactively

To create a logging directory for Hive, running the following commands:

Command:

```

sudo chown hadoop -R /var/log/hive
mkdir /var/log/hive/user/hadoop

```

Successfully Created the folder:

```

[hadoop@ip-172-31-19-97 ~]$ sudo chown hadoop -R /var/log/hive
[hadoop@ip-172-31-19-97 ~]$ mkdir /var/log/hive/user/hadoop
[hadoop@ip-172-31-19-97 ~]$ ls
[hadoop@ip-172-31-19-97 ~]$ ls /var/log/hive/user/
hadoop  hive  root
[hadoop@ip-172-31-19-97 ~]$ 

```

Configure Hive:

To retrieve the name of the S3 bucket, we are running the following command:

Command:

```

aws s3 ls /

```

```
[hadoop@ip-172-31-19-97 ~]$ aws s3 ls /
2023-10-28 23:31:30 aws-logs-428444136833-us-east-1
[hadoop@ip-172-31-19-97 ~]$ █
```

So S3 bucket name is: `aws-logs-428444136833-us-east-1`

Running the hive

```
hive -d SAMPLE=s3://aws-tc-largeobjects/CUR-TF-200-ACDENG-1/emr-lab -d DAY=2009-04-13 -d HOUR=08 -d NEXT_DAY=2009-04-13 -d NEXT_HOUR=08
```

Output:

```
[hadoop@ip-172-31-19-97 ~]$ hive -d SAMPLE=s3://aws-tc-largeobjects/CUR-TF-200-ACDENG-1/emr-lab -d DAY=2009-04-13 -d HOUR=08 -d NEXT_DAY=2009-04-13 -d NEXT_HOUR=09 -d OUTPUT=s3://aws-logs-428444136833-us-east-1
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> █
```

Task 4: Creating tables out of source data by using Hive

To create an external table named `impressions` :

Command:

```
CREATE EXTERNAL TABLE impressions (
    requestBeginTime string,
    adId string,
    impressionId string,
    referrer string,
    userAgent string,
    userCookie string,
    ip string)
PARTITIONED BY (dt string)
ROW FORMAT serde 'org.apache.hive.hcatalog.data.JsonSerDe' with serdeproperties
('paths'='requestBeginTime, adId, impressionId, referrer, userAgent, userCookie, ip')
LOCATION '${SAMPLE}/tables/impressions';
```

```
hive> CREATE EXTERNAL TABLE impressions (
    > requestBeginTime string,
    > adId string,
    > impressionId string,
    > referrer string,
    > userAgent string,
    > userCookie string,
    > ip string)
    > PARTITIONED BY (dt string)
    > ROW FORMAT serde 'org.apache.hive.hcatalog.data.JsonSerDe' with serdeproperties
    > ('paths'='requestBeginTime, adId, impressionId, referrer, userAgent, userCookie, ip')
    > LOCATION '${SAMPLE}/tables/impressions';
OK
Time taken: 18.788 seconds
hive> > █
```

Output:

```
OK
Time taken: 18.788 seconds
```

Updating the Hive metadata for the `impressions` table to include all partitions:

To see how many partitions the `impressions` table currently has:

Command:

```
describe formatted impressions;
```

Output:

```
hive> describe formatted impressions;
OK
# col_name          data_type      comment
requestbegintime    string         from deserializer
adid                string         from deserializer
impressionid        string         from deserializer
referrer             string         from deserializer
useragent            string         from deserializer
usercookie           string         from deserializer
ip                  string         from deserializer

# Partition Information
# col_name          data_type      comment
dt                  string

# Detailed Table Information
Database:           default
Owner:              hadoop
CreateTime:         Sun Oct 29 01:20:16 UTC 2023
LastAccessTime:     UNKNOWN
Retention:          0
Location:           s3://aws-tc-largeobjects/CUR-TF-200-ACDENG-1/emr-lab/tables/impressions
Table Type:         EXTERNAL_TABLE
Table Parameters:
  COLUMN_STATS_ACCURATE  {"BASIC_STATS":"true"}
  EXTERNAL               TRUE
  numFiles               0
  numPartitions          0
  numRows                0
  rawDataSize            0
  totalSize               0
  transient_lastDdlTime  1698542416

# Storage Information
SerDe Library:      org.apache.hive.hcatalog.data.JsonSerDe
InputFormat:         org.apache.hadoop.mapred.TextInputFormat
OutputFormat:        org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:          No
Num Buckets:        -1
Bucket Columns:     []
Sort Columns:        []
Storage Desc Params:
  paths                 requestBeginTime, adId, impressionId, referrer, userAgent, userCook
  serialization.format   1
Time taken: 1.388 seconds, Fetched: 44 row(s)
```

Number of partitions is 0:


```

Repar: Added partition to metastore impressions:dt=2009-04-14-10-15
Repar: Added partition to metastore impressions:dt=2009-04-14-10-28
Repar: Added partition to metastore impressions:dt=2009-04-14-11-09
Repar: Added partition to metastore impressions:dt=2009-04-14-11-15
Repar: Added partition to metastore impressions:dt=2009-04-14-11-18
Repar: Added partition to metastore impressions:dt=2009-04-14-11-25
Repar: Added partition to metastore impressions:dt=2009-04-14-11-28
Repar: Added partition to metastore impressions:dt=2009-04-14-12-08
Repar: Added partition to metastore impressions:dt=2009-04-14-12-13
Repar: Added partition to metastore impressions:dt=2009-04-14-12-18
Repar: Added partition to metastore impressions:dt=2009-04-14-12-25
Repar: Added partition to metastore impressions:dt=2009-04-14-12-28
Repar: Added partition to metastore impressions:dt=2009-04-14-13-09
Time taken: 21.861 seconds, Fetched: 242 row(s)

hive> describe formatted impressions;
# col_name          data_type      comment
# requestbegintime    string        from deserializer
# adid                string        from deserializer
# impressionid       string        from deserializer
# referrer           string        from deserializer
# useragent          string        from deserializer
# usercookie         string        from deserializer
# ip                  string        from deserializer

# Partition Information
# col_name          data_type      comment
# dt                 string

# Detailed Table Information
# Database:          default
# Owner:             hive
# CreateTime:        Sun Oct 29 01:20:16 UTC 2023
# LastAmendmentTime: 2023-10-29 01:20:16
# Retention:         0
# Location:          s3://aws-tc-largeobjects/CUR-TF-200-ACDONG-1/emr-lab/tables/impressions
Table: impressions
Table Parameters:
  Exports:          TRUE
  numFiles:         1466
  numRows:          241
  numPartitions:   0
  radiusPartition:  0
  totalSize:        659817289
  transient_LastDeleteTime: 169842616
  transient_LastUpdate: 169842616

# Storage Information
Serde Library:          org.apache.hive.hcatalog.data.JsonSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:            No
Num Buckets:           -1
Bucket Column Names: []
Sort Columns:          []
Storage Desc Params:
  requestbeginTime:    requestbeginTime, adid, impressionid, referrer, userAgent, userCookie, ip
  serialization.format: 1
Time taken: 0.267 seconds, Fetched: 43 row(s)
hive> ■

```

We can see that now partitions are created.

Creating another external table and again discover its partitions. This table will be named `clicks`, and it references click-stream logs

Command:

```

CREATE EXTERNAL TABLE clicks (impressionId string, adId string)
PARTITIONED BY (dt string)
ROW FORMAT serde 'org.apache.hive.hcatalog.data.JsonSerDe'
WITH SERDEPROPERTIES ('paths'='impressionId')
LOCATION '${SAMPLE}/tables/clicks';

```

Successfully Created the table:

```

hive> CREATE EXTERNAL TABLE clicks (impressionId string, adId string)
  > PARTITIONED BY (dt string)
  > ROW FORMAT serde 'org.apache.hive.hcatalog.data.JsonSerDe'
  > WITH SERDEPROPERTIES ('paths'='impressionId')
  > LOCATION '${SAMPLE}/tables/clicks';
OK
Time taken: 0.651 seconds
hive> ■

```

To return all partitions from the click-stream log data, running the following command:

```
MSCK REPAIR TABLE clicks;
```

Output:


```

# Detailed Table Information
Database: default
Owner: hadoop
CreateTime: Sun Oct 29 01:32:08 UTC 2023
LastAccessTime: UNKNOWN
Retention: 0
Location: s3://aws-tc-largeobjects/CUR-TF-200-ACDONG-1/emr-lab/tables/clicks
Table Type: EXTERNAL_TABLE
Table Parameters:
EXTERNAL             TRUE
numFiles            115
numPartitions       186
numRows             0
rowSize              0
totalSize           23864599
tstable_lastUpdateTime 1698545188
# Storage Information
Serde Library: org.apache.hive.hcatalog.data.JsonSerDe
InputFormat: org.apache.hadoop.mapred.TextInputFormat
OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed: No
Num Buckets: -1
BucketCols: []
Sort Columns: []
Storage Desc Params:
  paths          impressionId
  serialization.format 1
Time taken: 0.000 seconds, Fetched: 38 row(s)
hive> describe formatted clicks;
OK
# col_name      data_type      comment
impressionId    string        from deserializer
adId           string        from deserializer
# Partition Information
# col_name      data_type      comment
dt              string
# Detailed Table Information
Database: default
Owner: hadoop
CreateTime: Sun Oct 29 01:33:08 UTC 2023
LastAccessTime: UNKNOWN
Retention: 0
Location: s3://aws-tc-largeobjects/CUR-TF-200-ACDONG-1/emr-lab/tables/clicks
Table Type: EXTERNAL_TABLE
Table Parameters:
EXTERNAL             TRUE
numFiles            115
numPartitions       186
numRows             0
rowSize              0
totalSize           23864599
tstable_lastUpdateTime 1698545188
# Storage Information
Serde Library: org.apache.hive.hcatalog.data.JsonSerDe
InputFormat: org.apache.hadoop.mapred.TextInputFormat
OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed: No
Num Buckets: -1
BucketCols: []
Sort Columns: []
Storage Desc Params:
  paths          impressionId
  serialization.format 1
Time taken: 0.006 seconds, Fetched: 38 row(s)
hive> 

```

Task 5: Joining tables by using Hive

To create a new external table named `joined_impressions`, running the following command:

Command:

```

CREATE EXTERNAL TABLE joined_impressions (
requestBeginTime string,
adId string,
impressionId string,
referrer string,
userAgent string,
userCookie string,
ip string,
clicked Boolean)
PARTITIONED BY (day string, hour string)
STORED AS SEQUENCEFILE
LOCATION '${OUTPUT}/tables/joined_impressions';

```

Output:

```

Time taken: 0.000 seconds, Fetched: 38 row(s)
hive> CREATE EXTERNAL TABLE joined_impressions (
  > requestBeginTime string,
  > adId string,
  > impressionId string,
  > referrer string,
  > userAgent string,
  > userCookie string,
  > ip string,
  > clicked Boolean)
  > PARTITIONED BY (day string, hour string)
  > STORED AS SEQUENCEFILE
  > LOCATION '${OUTPUT}/tables/joined_impressions';
OK
Time taken: 2.005 seconds

```

To create a new temporary table in HDFS named `tmp_impressions` to store intermediate data, running the following command:

Command:

```
CREATE TABLE tmp_impressions (
    requestBeginTime string,
    adId string,
    impressionId string,
    referrer string,
    userAgent string,
    userCookie string,
    ip string)
STORED AS SEQUENCEFILE;
```

Output:

```
hive> CREATE TABLE tmp_impressions (
    > requestBeginTime string,
    > adId string,
    > impressionId string,
    > referrer string,
    > userAgent string,
    > userCookie string,
    > ip string)
    > STORED AS SEQUENCEFILE;
OK
Time taken: 0.343 seconds
hive> █
```

To insert impression log data for the time duration referenced, running the following command:

Command:

```
INSERT OVERWRITE TABLE tmp_impressions
SELECT
from_unixtime(cast((cast(i.requestBeginTime as bigint) / 1000) as int)) requestBeginTime,
i.adId,
i.impressionId,
i.referrer,
i.userAgent,
i.userCookie,
i.ip
FROM impressions i
WHERE i.dt >= '${DAY}-${HOUR}-00'
AND i.dt < '${NEXT_DAY}-${NEXT_HOUR}-00';
```

Output:

```

hive> INSERT OVERWRITE TABLE tmp_impressions
> SELECT
> from_unixtime(cast((cast(i.requestBeginTime as bigint) / 1000) as int)) requestBeginTime,
> i.adId,
> i.impressionId,
> i.referrer,
> i.userAgent,
> i.userCookie,
> i.ip
>     FROM impressions i
> WHERE i.dt >= '${DAY}-${HOUR}-00'
> AND i.dt < '${NEXT_DAY}-${NEXT_HOUR}-00';
Query ID = hadoop_20231029014340_33c7e875-1b0f-4f07-bac7-c721c08276c3
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1698540576797_0002)

-----  

VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

Map 1 ..... container SUCCEEDED    1       1       0       0       0       0       0  

-----  

VERTICES: 01/01  [=====>>>] 100% ELAPSED TIME: 37.95 s  

-----  

Loading data to table default.tmp_impressions
OK
Time taken: 52.369 seconds
hive> █

```

Create a temporary table named `tmp_clicks` and insert data into the table.

To create the table, running the following command:

Command:

```

CREATE TABLE tmp_clicks ( impressionId string, adId string )
STORED AS SEQUENCEFILE;

```

Output:

```

TIME TAKEN: 52.369 seconds
hive> CREATE TABLE tmp_clicks ( impressionId string, adId string )
      > STORED AS SEQUENCEFILE;
OK
Time taken: 0.062 seconds
hive> █

```

To insert data into the `tmp_clicks` table, running the following command:

Command:

```

INSERT OVERWRITE TABLE tmp_clicks
SELECT impressionId, adId
FROM clicks c
WHERE c.dt >= '${DAY}-${HOUR}-00'
AND c.dt < '${NEXT_DAY}-${NEXT_HOUR}-20';

```

```

hive> INSERT OVERWRITE TABLE tmp_clicks
> SELECT impressionId, adId
> FROM clicks c
> WHERE c.dt >= '${DAY}-${HOUR}-00'
> AND c.dt < '${NEXT_DAY}-${NEXT_HOUR}-20';
Query ID = hadoop_20231029014655_dc53b181-f4b9-4647-ad8c-4c9131c23adf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1698540576797_0002)

-----  

      VERTICES    MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

Map 1 ..... container SUCCEEDED   1       1       0       0       0       0       0
-----  

VERTICES: 01/01  [=====>>] 100% ELAPSED TIME: 36.88 s  

-----  

Loading data to table default.tmp_clicks
OK
Time taken: 38.345 seconds
hive>

```

To create a left outer join of `tmp_clicks` and `tmp_impressions` that writes the resulting data set to the `joined_impressions` table in your S3 output bucket, running the following command:

Command:

```

INSERT OVERWRITE TABLE joined_impressions
PARTITION (day='${DAY}', hour='${HOUR}')
SELECT
i.requestBeginTime,
i.adId,
i.impressionId,
i.referrer,
i.userAgent,
i.userCookie,
i.ip,
(c.impressionId is not null) clicked
FROM tmp_impressions i
LEFT OUTER JOIN tmp_clicks c ON i.impressionId=c.impressionId;

```

Output:

```

hive> INSERT OVERWRITE TABLE joined_impressions
> PARTITION (day='${DAY}', hour='${HOUR}')
> SELECT
> i.requestBeginTime,
> i.adId,
> i.impressionId,
> i.referrer,
> i.userAgent,
> i.userCookie,
> i.ip,
> (c.impressionId is not null) clicked
> FROM tmp_impressions i
> LEFT OUTER JOIN tmp_clicks c ON i.impressionId=c.impressionId;
Query ID = hadoop_20231029014934_adb2c69a-65eb-4eec-84aa-c223c157fa07
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1698540576797_0002)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED   1       1         0         0         0         0         0
Map 2 ..... container  SUCCEEDED   1       1         0         0         0         0         0
-----
VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 17.82 s
-----
Loading data to table default.joined_impressions partition (day=2009-04-13, hour=08)
OK
Time taken: 22.701 seconds
hive> 

```

Task 6: Querying the resulting dataset

Using SQL-like queries to query the resulting dataset.

To instruct the HiveSQL client to print the names of the columns as headers in result sets, running the following command:

Command:

```
set hive.cli.print.header=true;
```

To return the first 10 rows of data in the `joined_impressions` table, running the following SELECT command:

Command:

```
SELECT * FROM joined_impressions LIMIT 10;
```

Output:

```

hive> SELECT * FROM joined_impressions LIMIT 10;
OK
joined_impressions.requestbeginTime    joined_impressions.adid joined_impressions.impressionid joined_impressions.referrer    joined_impressions.userAgent    joined_impressions.usercookie    joined_impressions.ip    joined_impressions.clicked    joined_impressions.day    joined_impressions.hour
2009-04-13 08:01:27  MOHMJITLORSLNcIgHLMp0791828  FlMbd7qyLxtxDPuDfCh4f5C0B  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; FunWebProducts; GTB6; SLCC1; .NET CLR 2.0.58727; Media Center PC 5.0; .NET wcvMMfascoPbGtbdpbauTPPhgPbs 69.191.224
,234 false 2009-04-13 08
2009-04-13 08:01:27  MOHMJITLORSLNcIgHLMp0791828  FlMbd7qyLxtxDPuDfCh4f5C0B  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; FunWebProducts; GTB6; SLCC1; .NET CLR 2.0.58727; Media Center PC 5.0; .NET wcvMMfascoPbGtbdpbauTPPhgPbs 69.191.224
,234 false 2009-04-13 08
2009-04-13 08:01:28  MOHMJITLORSLNcIgHLMp0791828  FlMbd7qyLxtxDPuDfCh4f5C0B  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; FunWebProducts; GTB6; SLCC1; .NET CLR 2.0.58727; Media Center PC 5.0; .NET wcvMMfascoPbGtbdpbauTPPhgPbs 69.191.224
,234 false 2009-04-13 08
2009-04-13 08:02:03  pdcn9VAATfU4zJfrtuVj3bWmc  lgwgfxtxlaedCMhVgBpdg3fFdp  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.2; SV1; .NET CLR 1.1.4322; InfoPath.1) 05b3wkLRA31xIuq0g6NfQ31rCvWd 42.174.193.253 false 2009-04-13 08
2009-04-13 08:02:03  pdcn9VAATfU4zJfrtuVj3bWmc  lgwgfxtxlaedCMhVgBpdg3fFdp  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.2; SV1; .NET CLR 1.1.4322; InfoPath.1) 05b3wkLRA31xIuq0g6NfQ31rCvWd 42.174.193.253 false 2009-04-13 08
2009-04-13 08:02:26  Agu013k0pZx2uTc8B84pGpVvLQ7A  hccNwew11nb08R0kL48p9113ETB83u  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.2; SV1; .NET CLR 1.1.4322; InfoPath.1) 05b3wkLRA31xIuq0g6NfQ31rCvWd 42.174.193.253 false 2009-04-13 08
2009-04-13 08:02:26  Agu013k0pZx2uTc8B84pGpVvLQ7A  hccNwew11nb08R0kL48p9113ETB83u  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.2; SV1; .NET CLR 1.1.4322; InfoPath.1) 05b3wkLRA31xIuq0g6NfQ31rCvWd 42.174.193.253 false 2009-04-13 08
2009-04-13 08:03:25  Wv7cfC63mAt0qfjg8j173XK8bKGf5  jXTxF6kIxcKMnacu3X9gkMjW3IC  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.2; SV1; .NET CLR 1.1.4322; InfoPath.1) 05b3wkLRA31xIuq0g6NfQ31rCvWd 42.174.193.253 false 2009-04-13 08
2009-04-13 08:03:35  V13hDfOOGJXfC9Q0q2e0o37X4Q3y7  05b3wkLRA31xIuq0g6NfQ31rCvWd 42.174.193.253 false 2009-04-13 08
2009-04-13 08:03:26  4fCk4AdgG01tK0k85gBmcq94z5vlg  2x50df6pxuT0hqnquwWhotVz1c  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB6; .NET CLR 1.1.4322)  k0en5phdwWkTUJU8TPNfMFAb0g951.131.29.07  false 2009-04-13 08
2009-04-13 08:03:26  4fCk4AdgG01tK0k85gBmcq94z5vlg  2x50df6pxuT0hqnquwWhotVz1c  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB6; .NET CLR 1.1.4322)  k0en5phdwWkTUJU8TPNfMFAb0g951.131.29.07  false 2009-04-13 08
2009-04-13 08:07:48  wrqahTXLxwvqkA3Gcbqz4349f0  gurfl13BUEtWtvwokXms1HPAk0n10  cartoonnetwork.com  Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB6; .NET CLR 1.1.4322)  k0en5phdwWkTUJU8TPNfMFAb0g951.131.29.07  false 2009-04-13 08
2009-04-13 08:07:48  wrqahTXLxwvqkA3Gcbqz4349f0  gurfl13BUEtWtvwokXms1HPAk0n10  cartoonnetwork.com  Mozilla/5.0 (Macintosh; U; Mac OS X 10.5; en-US; rv:1.9.1.1) Gecko/20090715 Firefox/3.5.1 eM0EVOpjhIn2d73j35iMdgweWkR2.91.173.232 false 2009-04-13 08
Time taken: 0.243 seconds. Fetched: 10 row(s)
hive> 

```

Query:

```
SELECT adid, count(*) AS hits FROM joined_impressions WHERE clicked = true GROUP BY adid ORDER BY hits DESC LIMIT 10;
```

Output:

```

hive> SELECT adid, count(*) AS hits FROM joined_impressions WHERE clicked = true GROUP BY adid ORDER BY hits DESC LIMIT 10;
Query ID = hadoop_20231029015516_71ce43e4-023c-4e70-a099-fba2153013e6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1698540576797_0002)

-----  

    VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

-----  

Map 1 ..... container  SUCCEEDED   1       1       0       0       0       0  

Reducer 2 ..... container  SUCCEEDED   1       1       0       0       0       0  

Reducer 3 ..... container  SUCCEEDED   1       1       0       0       0       0  

-----  

VERTICES: 03/03  [=====>>>] 100%  ELAPSED TIME: 11.94 s  

-----  

OK  

adid      hits  

70n4fCvgAV0wfajgw5xw3QAEiauLMg  5  

FxSntm0CqAx6LejjR2qVelsXofoXL7  4  

2iKJc5AbF16IA3LGrhATQGpd07KhAk  4  

7p32sreTnFKiKsgllSkCmexjX1L8P  4  

HdQw0xhN7o3hQUJIVnoSS4JumGN79w  4  

tbXLLXhArCiqGMj3f1IOTrpGuidIKj  3  

98kBcfJejuIjxpHDfBMxQlk8S4lq16  3  

8NjKIm78sc6i9ccNxHT3v6Ua4N2  3  

vgAQhCFRswb4cFrwGJoaTUL1elqFQc  3  

5CiFcHjh8mk2js8SERhCLofcUrIEdn  3  

Time taken: 12.805 seconds, Fetched: 10 row(s)
hive> █

```

Exiting the hive:

```
exit;
```

Query:

```
hive -e "SELECT referrer, count(*) as hits FROM joined_impressions WHERE clicked = true GROUP BY referrer ORDER BY hits DESC LIMIT 10;"
```

Output:

```

hive> EXIT
[hadoop@ip-172-31-19-97 ~]$ hive -e "SELECT referrer, count(*) as hits FROM joined_impressions WHERE clicked = true GROUP BY referrer ORDER BY hits DESC LIMIT 10;" > /home/hadoop/result.txt
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Query ID = hadoop_20231029015756_252a089d-ffe2-47d8-9d52-6bbb962e8148
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1698540576797_0004)

Map 1: -/          Reducer 2: 0/1  Reducer 3: 0/1
Map 1: 0/1        Reducer 2: 0/1  Reducer 3: 0/1
Map 1: 0/1        Reducer 2: 0/1  Reducer 3: 0/1
Map 1: 0(+1)/1    Reducer 2: 0/1  Reducer 3: 0/1
Map 1: 0(+1)/1    Reducer 2: 0/1  Reducer 3: 0/1
Map 1: 0(+1)/1    Reducer 2: 0/1  Reducer 3: 0/1
Map 1: 1/1        Reducer 2: 0/1  Reducer 3: 0/1
Map 1: 1/1        Reducer 2: 0(+1)/1  Reducer 3: 0/1
Map 1: 1/1        Reducer 2: 1/1  Reducer 3: 0/1
Map 1: 1/1        Reducer 2: 1/1  Reducer 3: 0(+1)/1
Map 1: 1/1        Reducer 2: 1/1  Reducer 3: 1/1
OK
Time taken: 34.456 seconds, Fetched: 10 row(s)
[hadoop@ip-172-31-19-97 ~]$ █

```

Displaying the content of the file `result.txt`:

```
cat /home/hadoop/result.txt
```

Output:

```
[hadoop@ip-172-31-19-97 ~]$ cat /home/hadoop/result.txt
ibibo.com      49
lowes.com      33
odnoklassniki.ru    33
files.wordpress.com 30
media-servers.net   29
tmz.com        29
tattoodle.com    29
tiscali.it      28
mpnrs.com       28
barnesandnoble.com 28
[hadoop@ip-172-31-19-97 ~]$
```

To exit the SSH session, running the following command:

```
exit
```

Analyzing one of the log files that makes up the output from Amazon S3:

To rediscover the name of the bucket that contains the *joined_impressions* table data, running the following command:

Command:

```
aws s3 ls /
```

Output:

```
connection to ec2-54-234-110-3.compute-1.amazonaws.com:443
voclabs:~/environment $ aws s3 ls /
2023-10-28 23:31:30 aws-logs-428444136833-us-east-1
voclabs:~/environment $
```

To download the file that contains the data that you have been querying, running the following command:

Command:

```
aws s3 cp s3://aws-logs-428444136833-us-east-1/output/tables/joined_impressions/day=2009-04-13/hour=08/000000_0 example-log.txt
```

Output:

```
voclabs:~/environment $ aws s3 cp s3://aws-logs-105193538101-us-east-1/output/tables/joined_impressions/day=2009-04-13/hour=08/000000_0 example-log.txt
download: s3://aws-logs-105193538101-us-east-1/output/tables/joined_impressions/day=2009-04-13/hour=08/000000_0 to ./example-log.txt
voclabs:~/environment $
```

Conclusion Remarks:

Launching an Amazon EMR cluster with Hive installed is a fundamental step in setting up a data processing environment. By leveraging the capabilities of EMR, you can efficiently handle large-scale data processing tasks and leverage Hive for data warehousing and analysis.

Establishing a secure connection to the Amazon EMR main node using an SSH client and an Amazon EC2 key pair private key ensures secure access to the cluster. This step is crucial for administering and monitoring the EMR cluster and carrying out various tasks efficiently.

Configuring a logging directory for Hive and initiating an interactive Hive session on the EMR cluster's main node enables seamless data querying and manipulation. This interactive session facilitates data exploration and analysis, providing insights critical for decision-making and further data processing steps.

Creating Hive tables that reference data stored in Amazon S3 forms the backbone of a well-structured data warehouse. Using HiveQL to load data and run queries efficiently organizes the data, laying the groundwork for subsequent analysis and extraction of valuable insights.

The process of joining the impressions and clicks tables provides comprehensive insights into user behavior and aids in optimizing advertising services. Leveraging partitioned tables enhances the efficiency of data retrieval and analysis, enabling more accurate and targeted decision-making for advertising strategies and campaign optimizations.

Completing these tasks demonstrates your proficiency in handling complex data processing and analysis workflows, crucial skills in the dynamic and evolving field of data science.