ECV-EMR-221

How to process apache log in EMR cluster?

2017.09.18

Version 1.1

Agenda

About this lab

Scenario

Analysis apache log to get the insight which knew what kind of visitor to your website. Also to get more information for setup the customization content for the visitors.

This lab introduces you to Amazon Elastic Mapreduce (Amazon EMR) using the AWS Management Console. Also let you know how to engage S3 with EMR to process the apache log.

AWS EMR introduction

What is Amazon EMR?

Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data. By using these frameworks and related open-source projects, such as Apache Hive and Apache Pig, you can process data for analytics purposes and business intelligence workloads. Additionally, you can use Amazon EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.

AWS S3 introduction

What is Amazon EMR?

Amazon S3 is storage for the Internet. It's a simple storage service that offers software developers a highly-scalable, reliable, and low-latency data storage infrastructure at very low costs.

Amazon S3 provides a simple web service interface that you can use to store and retrieve any amount of data, at any time, from anywhere on the web. Using this web service, developers can easily build applications that make use of Internet storage. Since Amazon S3 is highly scalable and you only pay for what you use,

developers can start small and grow their application as they wish, with no compromise on performance or reliability.

Amazon S3 is also designed to be highly flexible. Store any type and amount of data that you want; read the same piece of data a million times or only for emergency disaster recovery; build a simple FTP application, or a sophisticated web application such as the Amazon.com retail web site. Amazon S3 frees developers to focus on innovation, not figuring out how to store their data.

Prerequisites for this lab

Download file:

https://s3-us-west-2.amazonaws.com/ecv-training-material/emr-lab/apache-log-analysis/access_log-20140105

https://s3-us-west-2.amazonaws.com/ecv-training-material/emr-lab/apache-log-analysis/do-reports2-edit.pig

The lab region will be in 'Oregon'

Lab Tutorial

Upload files to S3

Click S3

Click 'Create bucket' button

For bucket name: 'ecv-training-your-name'

For Region: choose 'US West (Oregon)'

Click 'Create'

Click the bucket which you created

Click 'Create folder' button

For folder name: 'apache-log-analysis'

Click 'Save'

Click the folder which you created

Upload two files which you download before

Click 'Create folder' button

For folder name: 'output'

Click 'Save'

Click 'Create folder' button

For folder name: 'log'

Click 'Save'

Launch an EMR cluster

Click EMR

Click 'Go to Advanced options'

In Step1: Software and Steps:

For Release, choose EMR 4.7 (It will automatically select Hadoop 2.7.2/Hive 1.0.0/ Hue 3.7.1/ Pig 0.14.0)

In the Add Step dialog:

- For Step type, choose Pig program and click configure
- For Name, accept the default name (Pig program) or type a new name.
- For Script S3 location, type the location of the Pig script. For example:s3://elasticmapreduce/samples/pig-apache/do-reports2.pig.
- For Input S3 location, type the location of the input data. For example:s3://elasticmapreduce/samples/pig-apache/input.
- For Output S3 location, type or browse to the name of your Amazon S3 output bucket.
- For Arguments, leave the field blank.
- For Action on failure, select 'Terminate cluster'.
- Click 'Add'

Select Auto-terminate cluster after the last step is completed

Click 'Next ' button

In Step2: Hardware Configuration:

For Network: leave the default setting

For Subnet: leave the default setting

For Root device EBS volume size: leave the default setting

In Node type dialog:

For Master node, choose m3.xlarge as 1

For Core node, choose m3.xlarge as 2

For Task node, cancel the instance requirement

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m3.xlarge 8 vCPU, 15 GiB memory, 80 SSD GB storage EBS Storage: none	1 Instances	On-demand Spot Maximum bid price: $
Core Core - 2	m3.xlarge 8 vCPU, 15 GiB memory, 80 SSD GB storage EBS Storage: none	2 Instances	On-demand Spot Maximum bid price: $

Click 'Next ' button

In Step3: General Cluster Setting:

In General Configuration dialog

For Cluster name: type 'My first EMR Cluster'

Leave the default setting

For Logging: Please change the path to the log folder which you created before

In Tag dialog:

For Key: type 'Name'

For Value: type 'EMR Cluster'

Click 'Next'

In Step4: Security:

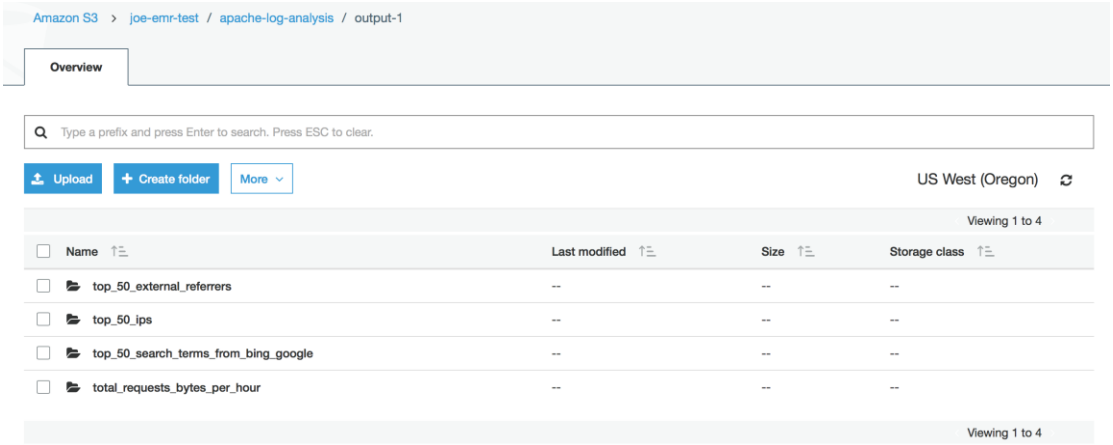
In Security Options dialog

Leave the default setting, DO NOT change the setting

Click 'Create Cluster'

Go back to S3 bucket to check the result

You will see the result which generated by EMR cluster and stored into the bucket.



Conclusion

Congratulations! You now have learned how to:

- Logged into Amazon Management Console
- Create S3 bucket and upload files into it
- Create an EMR cluster to process the data
- Find your resault in the S3 bucket