

Project 2 - Analyzing 10Gb of Yelp Reviews Data

Data Set

- Yelp's Reviews and Businesses dataset (about 10gb) from [Kaggle](#)
- Data Frames – Business, Review, User

Introduction

- The dataset is loaded onto a S3 bucket, and this URI is used to load them into spark cluster.
- We provision a Spark cluster on AWS EMR for loading and analyzing the above-mentioned dataset.
- The whole analysis is produced in Jupyter Notebook created on that cluster. (Analysis.ipynb)

Steps to configure Spark Cluster and Notebook

- 1) Creating a cluster on EMR
 - Provision a “cluster” – with a single master node and two worker nodes in EMR. This is the hardware necessary to run our Spark jobs.
 - Open AWS EMR page and click on ‘Create Cluster’.
 - In the advanced configuration options, select emr-5.31.0 as the Release option. Unselect Pig, Select Spark and Livy.
 - Select the instance types as m5.xlarge for both master as well as the core nodes
 - Give the cluster a relevant name and click on create cluster
 - Refer to the cluster configuration for more information
 - Click on the create notebook option under EMR Notebooks option
 - Give a relevant name to the notebook, choose the previously created cluster.
 - Once the notebook is ready, it can be opened in jupyter host for writing your analysis
 - Refer to the notebook configuration for more information
- 2) Running Spark cluster tasks via Jupyter Notebook
 - Change the kernel to PySpark before running any jobs
 - Type in `%%info` to check that everything is working as expected

Cluster configuration

Cluster: spc-project2 Waiting Cluster ready after last step completed.

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Summary

Configuration details

Application user interfaces

Network and hardware

ID: j-3S69C7D5QCFJN

Creation date: 2022-04-30 07:26 (UTC-4)

Elapsed time: 5 hours, 8 minutes

After last step completes: Cluster waits

Termination protection: Off [Change](#)

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

Master public DNS:
ec2-3-16-148-169.us-east-2.compute.amazonaws.com [🔗](#)
[Connect to the Master Node Using SSH](#)

Release label: emr-5.31.0

Hadoop distribution: Amazon 2.10.0

Applications: Hive 2.3.7, Hue 4.7.1, Spark 2.4.6, Livy 0.7.0

Log URI: s3://aws-logs-465003511108-us-east-2/elasticmapreduce/ [📁](#)

EMRFS consistent view: Disabled

Custom AMI ID: --

Persistent user interfaces [🔗](#): [Spark history server](#), [YARN timeline server](#), [Tez UI](#)

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

Availability zone: us-east-2c

Subnet ID: [subnet-00649d4c](#) [🔗](#)

Master: Running 1 m5.xlarge

Core: Running 2 m5.xlarge

Task: --

Cluster scaling: Not enabled

Auto-termination: Not enabled

Notebook configuration

Amazon EMR

EMR Studio

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Notebook: project2-spc Ready Workspace(notebook) is ready to run jobs on cluster j-3S69C7D5QCFJN.

Open in JupyterLab

Open in Jupyter

Stop

Delete

Notebook

Cluster

Notebook ID: e-7NZ9CRQT46DL6KD62X6JYH4X8

Description: --

Last modified: 8 minutes ago [🔗](#)

Last modified by: ...root [🔗](#)

Created on: 2022-04-29 08:50 (UTC-4)

Created by: ...root [🔗](#)

Service IAM role: [EMR_Notebooks_DefaultRole](#) [🔗](#)

Security groups for master instance: [sg-0a0f6de4f978bf3f8](#) [🔗](#)

Security groups for notebook instance: [sg-0ed8899444eb9e8b3](#) [🔗](#)

Notebook tags: creatorUserId = 465003511108 [View All / Edit](#)

Notebook location: s3://aws-emr-resources-465003511108-us-east-2/notebooks/ [📁](#)

Cluster: spc-project2

Cluster Id: [j-3S69C7D5QCFJN](#)

Packages installed

- Matplotlib 3.2.1
- Pandas 1.0.3
- Scipy 1.7.1
- Seaborn 0.11.2

Analysis

Various questions related to the dataset have been answered using pyspark data frame capabilities.

Utilized various libraries to filter, group, transform, and render visualizations in jupyter notebook.