Tutorial uso de MapReduce con MRJob

Pablo Moreno Quintero

Escuela de Ciencias e Ingeniería, Universidad EAFIT

Pregrado en Ingeniería de Sistemas

Edwin Nelson Montoya Munera

23 de noviembre de 2024

1. Configuración previa

- Tener configurado el AWS CLI y las credenciales de acceso.
- Contar con permisos para crear clústeres EMR.
- Instalar MRJob y sus dependencias en tu máquina local:

```
pip install mrjob boto3
```

2. Escribir el script MRJob

Un ejemplo sencillo para contar palabras de un archivo de texto en Python :

```
from mrjob.job import MRJob


class WordCount(MRJob):
  def mapper(self, _, line):
    # Emitir cada palabra en la línea
    for word in line.split():
      yield word.lower(), 1


  def reducer(self, word, counts):
    # Sumar todas las ocurrencias de una palabra
    yield word, sum(counts)


if __name__ == '__main__':
  WordCount.run()
```

3. Configurar MRJob para AWS EMR

Crea un archivo de configuración llamado **mrjob.conf** para especificar que se utilizará EMR y los detalles del clúster.

```
runners:
 emr:
  aws_access_key_id: YOUR_AWS_ACCESS_KEY
```

```
aws_secret_access_key: YOUR_AWS_SECRET_KEY
region: us-east-1
ec2_instance_type: m5.xlarge
ec2_instance_count: 3
num_core_instances: 2
output_dir: s3://your-output-bucket/mrjob-output/
emr_log_uri: s3://your-log-bucket/emr-logs/
bootstrap_actions:
 - Path: s3://elasticmapreduce/bootstrap-actions/install-py3
release_label: emr-6.10.0
applications:
 - Name: Hadoop
 - Name: Spark
```

4. Subir datos a Amazon S3

Los datos de entrada deben estar disponibles en un bucket de S3.

aws s3 cp datos.txt s3://labsbucket-pmorenoq/Gutenberg-small/*.txt

5. Ejecutar el job de MRJob

Ejecuta el script con el runner de EMR:

python wordcount.py -r emr s3://labsbucket-pmorenoq/Gutenberg-small/*.txt

6. Recuperar los resultados

Los resultados se guardarán en el bucket de salida especificado (s3:// labsbucket-pmorenoq /mrjob-output/).

Descargar los resultados:

aws s3 cp s3://your-output-bucket/mrjob-output/ output/ --recursive

Visualizar los resultados:

cat output/part-00000