

4. Spark

miércoles, 29 de abril de 2020 15:59

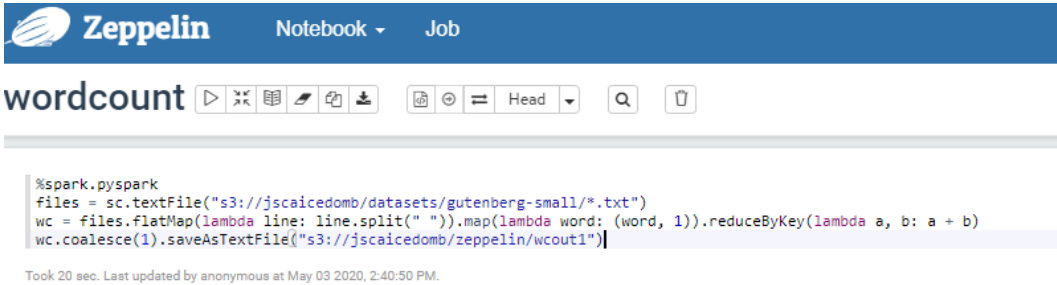
Creo una carpeta notebooks en mi bucket de s3 para los notebooks.

Realizo el tutorial de zeppelin en el clúster

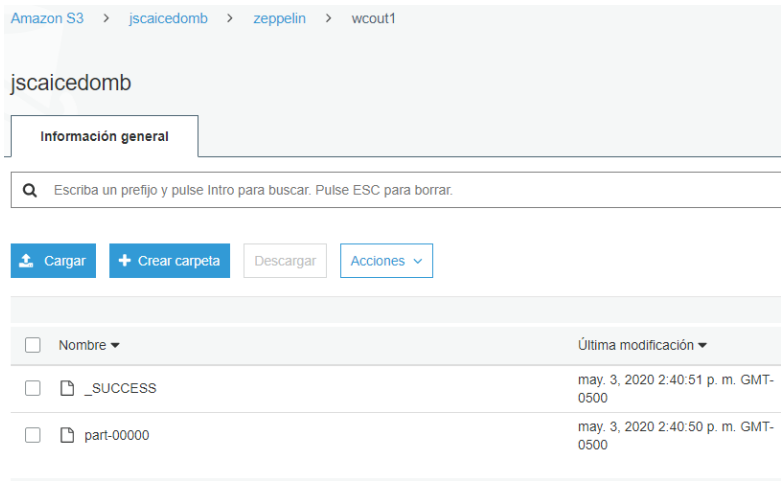


Wordcount en zeppelin

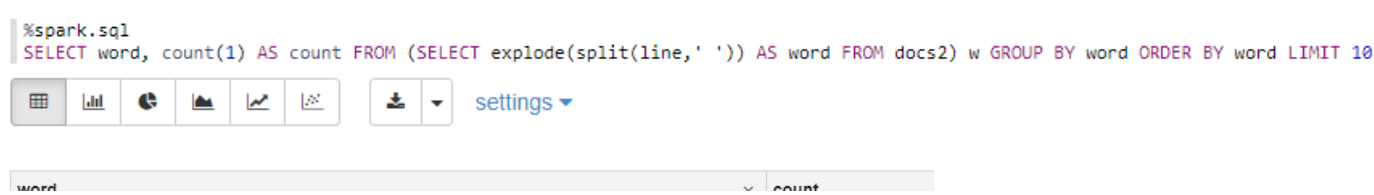
Creamos un notebook y realizamos el wordcount con datos en s3 y lo guardamos en s3



Verificamos en s3



Wordcount con sql en zeppelin



word	count
	27298
"	8
"ARTICLE	1
"But	1
"FROM	1
"KANSAS	1
"Nothing	1
"One	1
<	

EMR NOTEBOOK

Creo un clúster y un notebook asociado a él con ubicación en la carpeta que creé para los notebooks

Bloc de notas: wordcount Listo Notebook is ready to run jobs on cluster j-3POOKA5UN2KCK.

Abrir en JupyterLab

Abrir en Jupyter

Detener

Eliminar

Bloc de notas

ID del bloc de notas: e-2PRSW4B3GBSLGFPIOA79ECZAL

Descripción: --

Última modificación: Hace 4 minutos ⓘ

Modificado por última vez por: ...assumed-role/vocstartsoft/user592210=jscaicedom@eafit.edu.co ⓘ

Fecha de creación: 2020-05-03 14:54 (UTC-5)

Creado por: ...assumed-role/vocstartsoft/user592210=jscaicedom@eafit.edu.co ⓘ

Rol de IAM de EMR_Notebooks_DefaultRole ⓘ

Etiquetas del bloc de notas: creatorUserId = AROAV7WGGMPEKXZ6GEGT2:user592210=jscaicedom@eafit.edu.co Ver todo

Ubicación del bloc de notas: s3://jscaicedomb/notebooks/ ⓘ

Clúster

Clúster: BigDataCluster

ID del clúster: j-3POOKA5UN2KCK

Estado del clúster: Esperando Cluster ready after last step completed.

Etiquetas del clúster: --

Lo abro y subo los ejemplos

Files

Running

Clusters

Select items to perform actions on them.

0 /

☐

Data_processing_using_PySpark.ipynb

☐

wordcount-spark.ipynb

☐

wordcount.ipynb

Wordcount_spark.ipynb

[Archivo completo](#)

Data_processing_using_PySpark.ipynb

[Archivo completo](#)

Primero configuramos python 3

```
In [1]: %%configure -f
{ "conf":{
  "spark.pyspark.python": "python3",
  "spark.pyspark.virtualenv.enabled": "true",
  "spark.pyspark.virtualenv.type": "native",
  "spark.pyspark.virtualenv.bin.path": "/usr/bin/virtualenv"
}}
```

Current session configs: {'conf': {'spark.pyspark.python': 'python3', 'spark.pyspark.virtualenv.enabled': 'true', 'spark.pyspark.virtualenv.type': 'native', 'spark.pyspark.virtualenv.bin.path': '/usr/bin/virtualenv'}, 'kind': 'pyspark'}

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1588531763854_0002	pyspark	idle	Link	Link	

Después de iniciar la sesión de spark instalamos unas librerías que necesitaremos después:

```
sc.install_pypi_package("Pyarrow==0.14.1")
sc.install_pypi_package("Pandas")
```

```
In [2]: #import SparkSession
from pyspark.sql import SparkSession
```

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
3	application_1588531763854_0005	pyspark	idle	Link	Link	✓

SparkSession available as 'spark'.

```
In [3]: #create spar session object
spark=SparkSession.builder.appName('data_processing').getOrCreate()
```

```
In [4]: sc.install_pypi_package("Pyarrow")
```

Collecting Pyarrow
 Using cached pyarrow-0.17.0-cp36-cp36m-manylinux2014_x86_64.whl (63.8 MB)
 Requirement already satisfied: numpy>=1.14 in /usr/local/lib64/python3.6/site-packages (from Pyarrow) (1.14.5)
 Installing collected packages: Pyarrow
 Successfully installed Pyarrow-0.17.0

```
In [5]: sc.install_pypi_package("Pandas")
```

Collecting Pandas
 Using cached pandas-1.0.3-cp36-cp36m-manylinux1_x86_64.whl (10.0 MB)
 Collecting python-dateutil>=2.6.1
 Using cached python_dateutil-2.8.1-py2.py3-none-any.whl (227 kB)

Luego cargamos el archivo csv desde s3 y ejecutamos todo

```
In [6]: # Load csv Dataset
df=spark.read.csv('s3://jscaicedomb/datasets/spark/sample_data.csv',inferSchema=True,header=True)
```

▶ Spark Job Progress

```
In [7]: #columns of dataframe
df.columns
```

['ratings', 'age', 'experience', 'family', 'mobile']

El archivo completo está en github, no hubo problemas corriéndolo.

Guardamos tipo csv y tipo parquet:

```
In [31]: # saving file (csv)
```

▶ Spark Job Progress

```
In [33]: #target directory
write_uri='s3://jscaicedomb/df_csv'
```

```
In [34]: #save the dataframe as single csv
```

```
In [34]: #save the dataframe as single csv
df.coalesce(1).write.format("csv").option("header","true").save(write_uri)

In [35]: # parquet

In [38]: #target location
parquet_uri='s3://jscaicedomb/df_parquet'

In [39]: #save the data into parquet format
df.write.format('parquet').save(parquet_uri)
```

Verificamos que estén en s3 (jscaicedomb)

jscaicedomb

Información general Propiedades Permisos

Q Escriba un prefijo y pulse Intro para buscar. Pulse ESC para borrar.

Cargar + Crear carpeta Descargar Acciones ▾

☐ Nombre ▾

- ☐ datasets
- ☐ df_csv
- ☐ df_parquet
- ☐ notebooks
- ☐ zeppelin

LAB EVALUABLE de SPARK

[ARCHIVO COMPLETO](#)

Bucket público en s3

jscaicedomb

Información general Propiedades Permisos Administración Puntos de acceso

Bloquear acceso público Lista de control de acceso Política de bucket Configuración de CORS

Bloquear acceso público (configuración del bucket)

Se concede acceso público a buckets y objetos a través de listas de control de acceso (ACL), políticas de bucket, políticas de puntos de acceso o todas las anteriores. A fin de garantizar que se bloquee el acceso público a todos sus buckets y objetos de S3, active Bloquear todo acceso público. Esta configuración se aplica en exclusiva a este bucket y a sus puntos de acceso. AWS recomienda activar Bloquear todo acceso público pero, antes de aplicar cualquiera de estos ajustes, asegúrese de que sus aplicaciones funcionarán correctamente sin acceso público. Si necesita cierto nivel de acceso público a sus buckets u objetos, puede personalizar los valores de configuración individuales a continuación para que se ajusten mejor a sus necesidades específicas de almacenamiento. [Más información](#)

Bloquear todo acceso público
Desactivado [Editar](#)

Bloquear el acceso público a buckets y objetos concedido a través de nuevas listas de control de acceso (ACL)
Desactivado

Bloquear el acceso público a buckets y objetos concedido a través de cualquier lista de control de acceso (ACL)

Desactivado

Amazon S3

>

jscairedomb

>

lab_spark

jscairedomb

Información general

Q

Escriba un prefijo y pulse Intro para buscar. Pulse ESC para borrar.

Cargar

Crear carpeta

Descargar

Acciones

<input type="checkbox"/>	Nombre	Última modificación
<input type="checkbox"/>	_SUCCESS	may. 3, 2020 6:55:40 p. m. GMT-0500
<input type="checkbox"/>	part-00000-ed2954dc-e252-4f14-82da-bf2221b17e84-c000.csv	may. 3, 2020 6:55:39 p. m. GMT-0500