

Nombre: Andrés Felipe Téllez Rodríguez - email: aftellezr@eafit.edu.co

Laboratorios de clase:

Laboratorio 3.0 (EMR):

▼ Nombre y aplicaciones - *obligatorio* [Información](#)

Asigne un nombre a su clúster y elija las aplicaciones que desea instalar en él.

Nombre

Laboratorio


Versión de Amazon EMR [Información](#)

Una versión contiene un conjunto de aplicaciones que se puede instalar en el clúster.


emr-7.3.0 ▼

Paquete de aplicaciones


Spark
Interactive




Core
Hadoop




Flink




HBase




Presto



Trino



Custom



☐ AmazonCloudWatchAgent
1.300032.2

☒ HCatalog 3.1.3

☒ Hue 4.11.0

☒ Livy 0.8.0

☐ Pig 0.17.0

☒ Sqoop 1.4.7

☐ Trino 442

☐ Flink 1.18.1

☒ Hadoop 3.3.6

☒ JupyterEnterpriseGateway 2.6.0

☐ Oozie 5.2.1

☐ Presto 0.285

☐ TensorFlow 2.16.1

☒ Zeppelin 0.11.1

☐ HBase 2.4.17

☒ Hive 3.1.3

☒ JupyterHub 1.5.0

☐ Phoenix 5.1.3

☒ Spark 3.5.1

☐ Tez 0.10.2

☒ ZooKeeper 3.9.1

Configuración del Catálogo de datos de AWS Glue

Utilice el Catálogo de datos de AWS Glue para proporcionar un meta-almacén externo a la aplicación.

☒ Usar para metadatos de la tabla de Hive

☒ Usar para metadatos de la tabla de Spark

Opciones del sistema operativo [Información](#)

☒ Versión de Amazon Linux

☐ Imagen de máquina de Amazon (AMI) personalizada

☒ Aplicar automáticamente las actualizaciones más recientes de Amazon Linux

▼ Configuración del clúster - obligatorio [Información](#)

Elija un método de configuración para los grupos principales, centrales y de nodos tareas para su clúster.

☒ Grupos de instancias uniformes

Elija el mismo tipo de instancia de EC2 y la misma opción de compra (bajo demanda o de spot) para todos los nodos de su grupo de nodos. [Más información](#)

☐ Flotas de instancias flexibles

Elija entre la más amplia variedad de opciones de aprovisionamiento para las instancias de EC2 de su clúster. Diversifique los tipos de instancias y las opciones de compra, y utilice una estrategia de asignación. [Más información](#)

Grupos de instancias uniformes

Principal

Elegir tipo de instancia de EC2

m5.xlarge

4 vCore 16 GiB memoria
Únicamente EBS almacenamiento
Precio bajo demanda: -
Precio de spot más bajo: -

Acciones ▼

☐ Utilice la alta disponibilidad

Lance un clúster más resiliente y de alta disponibilidad con tres nodos principales en instancias bajo demanda. Esta configuración se aplica durante toda la vida útil del clúster. [Más información](#)

► Configuración de nodo - opcional

Central

Elegir tipo de instancia de EC2

m5.xlarge

4 vCore 16 GiB memoria
Únicamente EBS almacenamiento
Precio bajo demanda: -
Precio de spot más bajo: -

Acciones ▼

► Configuración de nodo - opcional

Tarea 1 de 1

Eliminar grupo de instancias

Nombre

Tarea - 1

Elegir tipo de instancia de EC2

m5.xlarge

4 vCore 16 GiB memoria

Únicamente EBS almacenamiento

Precio bajo demanda: -

Precio de spot más bajo: -

Acciones ▼

► Configuración de nodo - *opcional*

Agregar grupo de instancias de tareas

Puede agregar hasta 47 grupos más de instancias de tareas.

Volumen raíz de EBS

El volumen raíz de EBS se aplica a los sistemas operativos y las aplicaciones que instale en el clúster. [Restricciones de relación de volumen raíz de EBS](#)

Tamaño (GiB)

15

15- 100 GiB por volumen SSD de uso general (gp3)

IOPS

3000

3000-16000 IOPS por volumen. Elija una relación máxima de 500:1 entre IOPS y el tamaño del volumen.

Rendimiento (MiB/s)

125

125-1000 MiB/s por volumen. Elija una relación máxima de 0.25:1 entre el rendimiento y las IOPS.

▼ **Aprovisionamiento y escalado de clústeres - obligatorio** [Información](#)

Elija cómo Amazon EMR debe dimensionar su clúster.

Elija una opción

☒ **Establecer el tamaño del clúster manualmente**

Utilice esta opción si conoce los patrones de la carga de trabajo de antemano.

☐ **Utilizar escalado administrado por EMR**

Supervise las métricas clave de la carga de trabajo de modo que EMR pueda optimizar el tamaño del clúster y la utilización de los recursos.

☐ **Utilizar el escalamiento automático personalizado**

Para escalar mediante programación los nodos principales y los nodos de tarea, cree políticas de escalamiento automático personalizadas.

Configuración de aprovisionamiento

Establezca el tamaño del principal y tarea grupos de instancias. Amazon EMR intenta aprovisionar esta capacidad al lanzar el clúster.

Nombre	Tipo de instancia	Tamaño de instancia(s)	Utilizar la opción de compra de spot
Central	m5.xlarge	<input type="text" value="1"/>	<input type="checkbox"/>
Tarea - 1	m5.xlarge	<input type="text" value="1"/>	<input type="checkbox"/>

▼ **Redes - obligatorio** [Información](#)

Elija la configuración de red que determina la forma en que usted y otras entidades se comunican con su clúster.

Virtual Private Cloud (VPC) [Información](#)

[Examinar](#)[Crear VPC](#) 

Subred [Información](#)

[Examinar](#)[Crear subred](#) 

► **Grupos de seguridad de EC2 (firewall)**

▼ Grupos de seguridad de EC2 (firewall)



Aviso de cambio

Hemos actualizado los nombres de algunos grupos de seguridad para utilizar un lenguaje más inclusivo. Por ejemplo, los grupos que incluían términos como “maestro” y “esclavo” ahora utilizan en su lugar los términos “principal” y “central”.

Nodo principal

Grupos de seguridad administrados de EMR

EMR actualizará automáticamente el grupo seleccionado.

ElasticMapReduce-Primary
sg-045a5c5e53598f955



Grupos de seguridad adicionales - *opcional*

Seleccione hasta 4 grupos de seguridad adicionales.

Elegir grupos de seguridad adicionales



Nodos principales y de tareas

Grupos de seguridad administrados de EMR

EMR actualizará automáticamente el grupo seleccionado.

ElasticMapReduce-Core
sg-069ceb68673aa11b9



Grupos de seguridad adicionales - *opcional*

Seleccione hasta 4 grupos de seguridad adicionales.

Elegir grupos de seguridad adicionales



▼ Configuración de software [Información](#)

Anule las configuraciones predeterminadas para aplicaciones específicas de su clúster.

☒ Ingresar la configuración

☐ Cargar JSON desde Simple Storage Service
(Amazon S3)

```
1 [
2   {
3     "Classification": "jupyter-s3-conf",
4     "Properties": {
5       "s3.persistence.bucket": "aftellezrnotebooks",
6       "s3.persistence.enabled": "true"
7     }
8   }
9 ]
```

JSON Ln 1, Col 1 ✖ : 0 ⚠ : 0



▼ Configuración de seguridad y par de claves de EC2 [Información](#)

Elija una configuración de seguridad o cree una nueva que pueda reutilizar con otros clústeres.

Configuración de seguridad

Seleccione la configuración del servicio de cifrado, autenticación, autorización y metadatos de instancia del clúster.



Examinar [↗](#)

[Crear configuración de seguridad](#) [↗](#)

Par de claves de Amazon EC2 para el protocolo SSH al clúster [Información](#)



Examinar

[Crear par de claves](#) [↗](#)

▼ Roles de Identity and Access Management (IAM) - *obligatorio* [Información](#)

Elija o cree un rol de servicio y un perfil de instancia para las instancias de EC2 del clúster.

Rol de servicio de Amazon EMR [Información](#)

El rol de servicio es un rol de IAM que Amazon EMR asume para aprovisionar recursos y realizar acciones de nivel de servicio con otros servicios de AWS.

☒ Elegir un rol de servicio existente

Seleccione un rol de servicio predeterminado o un rol personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con otros servicios de AWS.

☐ Crear un rol de servicio

Deje que Amazon EMR cree un nuevo rol de servicio para que pueda conceder y restringir el acceso a los recursos de otros servicios de AWS.

Rol de servicio



Perfil de instancia de EC2 para Amazon EMR

El perfil de instancia asigna un rol a cada instancia de EC2 de un clúster. El perfil de instancia debe especificar un rol que pueda acceder a los recursos de los pasos y las acciones de arranque.

☒ Elegir un perfil de instancia existente

Seleccione un rol predeterminado o un perfil de instancia personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con sus recursos de Amazon S3.

☐ Crear un perfil de instancia

Deje que Amazon EMR cree un nuevo perfil de instancia para que pueda especificar un conjunto personalizado de recursos a los que tendrá acceso en Amazon S3.

Perfil de instancia



Rol de escalamiento automático personalizado - *opcional*

Cuando se activa una regla de escalamiento automático personalizada, Amazon EMR asume esta función para agregar y finalizar instancias de EC2. [Más información](#) [↗](#)

Rol de escalamiento automático personalizado



[Crear rol de IAM](#) [↗](#)


```
[hadoop@ip-172-31-77-5 ~]$ sudo yum install git
Last metadata expiration check: 2:09:51 ago on Tue Nov 5 22:36:46 2024.
Dependencies resolved.

=====
Package                        Architecture      Version           Repository        Size
=====
Installing:
git                           x86_64            2.40.1-1.amzn2023.0.3  amazonlinux      54 k
Installing dependencies:
git-core                      x86_64            2.40.1-1.amzn2023.0.3  amazonlinux      4.3 M
git-core-doc                  noarch            2.40.1-1.amzn2023.0.3  amazonlinux      2.6 M
perl-Error                    noarch            1:0.17029-5.amzn2023.0.2  amazonlinux      41 k
perl-Git                      noarch            2.40.1-1.amzn2023.0.3  amazonlinux      42 k
perl-TermReadKey              x86_64            2.38-9.amzn2023.0.2    amazonlinux      36 k
perl-lib                      x86_64            0.65-477.amzn2023.0.6    amazonlinux      15 k

Transaction Summary
=====
Install 7 Packages

Total download size: 7.1 M
Installed size: 34 M
Is this ok [y/N]: y
Downloading Packages:
(1/7): git-2.40.1-1.amzn2023.0.3.x86_64.rpm          974 kB/s | 54 kB      00:00
(2/7): git-core-doc-2.40.1-1.amzn2023.0.3.noarch.rpm 30 MB/s | 2.6 MB     00:00
(3/7): perl-Error-0.17029-5.amzn2023.0.2.noarch.rpm  1.1 MB/s | 41 kB     00:00
(4/7): git-core-2.40.1-1.amzn2023.0.3.x86_64.rpm    34 MB/s | 4.3 MB     00:00
(5/7): perl-Git-2.40.1-1.amzn2023.0.3.noarch.rpm     1.0 MB/s | 42 kB     00:00
(6/7): perl-TermReadKey-2.38-9.amzn2023.0.2.x86_64.rpm 944 kB/s | 36 kB     00:00
(7/7): perl-lib-0.65-477.amzn2023.0.6.x86_64.rpm    894 kB/s | 15 kB     00:00
=====
Total                                           36 MB/s | 7.1 MB     00:00
Running transaction check
Transaction check succeeded.
Running transaction test
Transaction test succeeded.
Running transaction
  Preparing                :                               1/1
  Installing               : git-core-2.40.1-1.amzn2023.0.3.x86_64 1/7
  Installing               : git-core-doc-2.40.1-1.amzn2023.0.3.noarch 2/7
  Installing               : perl-lib-0.65-477.amzn2023.0.6.x86_64 3/7
  Installing               : perl-TermReadKey-2.38-9.amzn2023.0.2.x86_64 4/7
  Installing               : perl-Error-1:0.17029-5.amzn2023.0.2.noarch 5/7
  Installing               : perl-Git-2.40.1-1.amzn2023.0.3.noarch 6/7
```

```
A newer release of "Amazon Linux" is available.

Available Versions:

Version 2023.6.20241028:
Run the following command to upgrade to 2023.6.20241028:

dnf upgrade --releasever=2023.6.20241028

Release notes:
https://docs.aws.amazon.com/linux/al2023/release-notes/relnotes-2023.6.20241028.html

Version 2023.6.20241031:
Run the following command to upgrade to 2023.6.20241031:

dnf upgrade --releasever=2023.6.20241031



Release notes:
https://docs.aws.amazon.com/linux/al2023/release-notes/relnotes-2023.6.20241031.html

=====

Installed:
git-2.40.1-1.amzn2023.0.3.x86_64 git-core-2.40.1-1.amzn2023.0.3.x86_64 git-core-doc-2.40.1-1.amzn2023.0.3.n
oarch perl-Error-1:0.17029-5.amzn2023.0.2.noarch perl-Git-2.40.1-1.amzn2023.0.3.noarch
perl-TermReadKey-2.38-9.amzn2023.0.2.x86_64 perl-lib-0.65-477.amzn2023.0.6.x86_64

Complete!
```


1. Evidencias de HDFS y S3 desde EMR

<input type="checkbox"/>	Name	Size	Usuario	Group	Permisos	Date
<input checked="" type="checkbox"/>	 j		hdfs	hdfsadmin	drwxr-xr-x	October 23, 2024 06:58 PM
<input type="checkbox"/>	 .		hadoop	hdfsadmin	drwxrwxrwx	October 23, 2024 07:06 PM
<input type="checkbox"/>	 datasets		hadoop	hdfsadmin	drwxr-xr-x	October 23, 2024 07:34 PM



















```
[hadoop@ip-172-31-73-126 ~]$ sudo yum install python3-pip
Last metadata expiration check: 0:19:24 ago on Wed Nov 6 22:06:52 2024.
Package python3-pip-21.3.1-2.amzn2023.0.8.noarch is already installed.
Dependencies resolved.
Nothing to do.
Complete!
```

```
[hadoop@ip-172-31-73-126 ~]$ sudo pip3 install mrjob
Collecting mrjob
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
    |████████████████████| 439 kB 6.8 MB/s
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib64/python3.9/site-packages (from mrjob) (5.4.1)
Installing collected packages: mrjob
Successfully installed mrjob-0.7.4
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
```

```
[hadoop@ip-172-31-73-126 bigdata]$ cd 02-mapreduce/
[hadoop@ip-172-31-73-126 02-mapreduce]$ ls
README.md  data1.txt  data2.txt  wordcount-local.py  wordcount-mr.py
```

 Inicio

/user/hadoop/datasets/gutenberg-small

<input type="checkbox"/>	Name	Size	Usuario	Group	Permisos	Date
<input type="checkbox"/>	 .		hadoop	hdfsadmin	drwxr-xr-x	November 06, 2024 02:33 PM
<input type="checkbox"/>	 AbrahamLincoln__LincolnLetters.txt	5,7 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__LincolnsFirstInauguralAddress.txt	21,1 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__LincolnsGettysburgAddressGivenNovember-19-1863.txt	1,6 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__LincolnsInauguralsAddressesandLettersSelections.txt	255,9 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__LincolnsSecondInauguralAddress.txt	4,0 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__SpeechesandLettersofAbrahamLincoln1832-1865.txt	504,2 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__StateoftheUnionAddresses.txt	164,0 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__TheEmancipationProclamation.txt	3,8 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__TheLifeandPublicServiceofGeneralZacharyTaylorAnAddress.txt	44,6 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__TheWritingsofAbrahamLincolnVolume1.txt	448,2 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__TheWritingsofAbrahamLincolnVolume2.txt	493,3 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__TheWritingsofAbrahamLincolnVolume3.txt	249,0 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__TheWritingsofAbrahamLincolnVolume4.txt	204,7 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__TheWritingsofAbrahamLincolnVolume5.txt	675,8 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__TheWritingsofAbrahamLincolnVolume6.txt	587,0 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM
<input type="checkbox"/>	 AbrahamLincoln__TheWritingsofAbrahamLincolnVolume7.txt	467,5 KB	hadoop	hdfsadmin	-rw-r--r--	November 06, 2024 02:39 PM

```

[hadoop@ip-172-31-73-126 02-mapreduce]$ python wordcount-mr.py data1.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/wordcount-mr.hadoop.20241106.232735.654859
Running step 1 of 1...
job output is in /tmp/wordcount-mr.hadoop.20241106.232735.654859/output
Streaming final output from /tmp/wordcount-mr.hadoop.20241106.232735.654859/output...
"arriba"      1
"como"      1
"cruel"      1
"el"        2
"esta"      1
"hola"      1
"mundo"     2
"nacional"   1
Removing temp directory /tmp/wordcount-mr.hadoop.20241106.232735.654859...
[hadoop@ip-172-31-73-126 02-mapreduce]$ python wordcount-mr.py data2.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/wordcount-mr.hadoop.20241106.232744.294202
Running step 1 of 1...
job output is in /tmp/wordcount-mr.hadoop.20241106.232744.294202/output
Streaming final output from /tmp/wordcount-mr.hadoop.20241106.232744.294202/output...
"arriba"      1
"bacano"      1
"el"        2
"es"         1
"esta"      1
"mundo"     2
"muy"        1
"patas"      1
Removing temp directory /tmp/wordcount-mr.hadoop.20241106.232744.294202...
[hadoop@ip-172-31-73-126 02-mapreduce]$ |

```

Laboratorio 3.3 (Hive-Sparksql):



Query. Explore. Repeat.

Since this is your first time logging in, pick any username and password. Be sure to remember these, as **they will become your Hue superuser credentials.**

The password must be at least 8 characters long, and must contain both uppercase and lowercase letters, at least one number, and at least one special character.

Create Account

File Browser

Search for file name

Actions

Copy Path

Open in Importer

Upload

New

Inicio

/user/hive/warehouse/hdi

	Name	Size	Usuario	Group	Permisos	Date
			hdfs	hdfsadmingroup	drwxrwxrwt	November 04, 2024 10:14 AM
	.		hadoop	hdfsadmingroup	drwxr-xr-x	November 04, 2024 10:22 AM
	export-data.csv	4,3 KB	hadoop	hdfsadmingroup	-rw-r--r--	November 04, 2024 10:22 AM
	hdi-data.csv	9,0 KB	hadoop	hdfsadmingroup	-rw-r--r--	November 04, 2024 10:22 AM

Show

45

of 2 items

Page 1 of 1

0.23s default

```
1 use usernameDB;
2 show tables;
3 describe hdi;
4
5 select * from hdi;
6
7 select country, gni from hdi where gni > 2000;
```

```
select country, gni from hdi where gni > 2000
INFO : Completed executing command(queryId=hive_20241104203120_8b16c28e-42aa-42c9-8615-0377c7ddb9b8); Time taken: 0.0 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
```

Query HistorySaved QueriesResults (100+)

	country	gni
1	Norway	47557
2	Australia	34431
3	Netherlands	36402
4	United States	43017
5	New Zealand	23737
6	Canada	35166
7	Ireland	29322
8	Liechtenstein	83717
9	Germany	34854
10	Sweden	35837
11	Switzerland	39924
12	Japan	32295

1.85s default

```
1 use usernameDB;
2 CREATE EXTERNAL TABLE EXPO (country STRING, expct FLOAT)
3 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
4 STORED AS TEXTFILE
5 LOCATION 's3://emontoyadatasets/ONU/export/';
```

Error while processing statement: FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. MetaException(message:Got exception: java.io.IOException com.amazon.ws.emr.hadoop.fs.shaded.com.amazonaws.services.s3.model.AmazonS3Exception: The specified bucket does not exist (Service: Amazon S3; Status Code: 404; Error Code: NoSuchBucket; Request ID: DE0XAQNKY2J81HK0; S3 Extended Request ID: uMgPqFMsBo1owbqU00/ScRy1RFUx7LfH32YI4c7uLSiHeJISCoRa4YjUNs1GfIPjniWez5RjOUF11nhw8A3kU9RKu2c+m5; Proxy: null), S3 Extended Request ID: uMgPqFMsBo1owbqU00/ScRy1RFUx7LfH32YI4c7uLSiHeJISCoRa4YjUNs1GfIPjniWez5RjOUF11nhw8A3kU9RKu2c+m5)

Request ID: DE0XAQNKY2J81HK0; S3 Extended Request ID: uMgPqFMsBo1owbqU00/ScRy1RFUx7LfH32YI4c7uLSiHeJISCoRa4YjUNs1GfIPjniWez5RjOUF11nhw8A3kU9RKu2c+m5; Proxy: null), S3 Extended Request ID: uMgPqFMsBo1owbqU00/ScRy1RFUx7LfH32YI4c7uLSiHeJISCoRa4YjUNs1GfIPjniWez5RjOUF11nhw8A3kU9RKu2c+m5)
INFO : Completed executing command(queryId=hive_20241104203147_c2480fa9-53c5-4db4-bd12-76effd32bfb2); Time taken: 0.822 seconds
INFO : Concurrency mode is disabled, not creating a lock manager

0.38s default ?

```
1 CREATE EXTERNAL TABLE docs (line STRING)
2 STORED AS TEXTFILE
3 LOCATION 's3://emontoyadatasets/gutenberg-small/';
```

Error while processing statement: FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. MetaException(message:Got exception: java.io.IOException com.amazon.ws.emr.hadoop.fs.shaded.com.amazonaws.services.s3.model.AmazonS3Exception: The specified bucket does not exist (Service: Amazon S3; Status Code: 404; Error Code: NoSuchBucket; Request ID: XJBW97WEKVDJ0361; S3 Extended Request ID: wsaF0WTqpjF1BIlnadNDgCsG61gHmdaXBYU0ytwYj+QfScN82YPrj2VO5gVpFz+5wep3DcXa0wLCZ2TsIpm5IIqN9/GfnW/w22MbK+qE+QU=; Proxy: null), S3 Extended Request ID: wsaF0WTqpjF1BIlnadNDgCsG61gHmdaXBYU0ytwYj+QfScN82YPrj2VO5gVpFz+5wep3DcXa0wLCZ2TsIpm5IIqN9/GfnW/w22MbK+qE+QU=)

Group: 13, Shaded: com.amazonaws.services.s3.model.AmazonS3Exception: The specified bucket does not exist (Service: Amazon S3; Status Code: 404; Error Code: NoSuchBucket; Request ID: XJBW97WEKVDJ0361; S3 Extended Request ID: wsaF0WTqpjF1BIlnadNDgCsG61gHmdaXBYU0ytwYj+QfScN82YPrj2VO5gVpFz+5wep3DcXa0wLCZ2TsIpm5IIqN9/GfnW/w22MbK+qE+QU=; Proxy: null), S3 Extended Request ID: wsaF0WTqpjF1BIlnadNDgCsG61gHmdaXBYU0ytwYj+QfScN82YPrj2VO5gVpFz+5wep3DcXa0wLCZ2TsIpm5IIqN9/GfnW/w22MbK+qE+QU=)
INFO : Completed executing command(queryId=hive_20241104204239_d926fce1-ff78-48ad-a1d4-3b9e2098a404); Time taken: 0.358 seconds
INFO : Concurrency mode is disabled, not creating a lock manager

Para poder llenar los resultados de la tabla con los resultados del query que se nos da el ejemplo, es decir del wordcount. Se deben de seguir algunos pasos.

El primer paso es crear la tabla de destino. Un ejemplo de esto es:

```
CREATE TABLE IF NOT EXISTS frecuencias_palabras (
    palabra STRING,
    frecuencia INT
);
```

Después se ejecuta el query que se propone. Y de ahí se guarda los resultados en la tabla de destino

```
INSERT INTO TABLE frecuencias_palabras
SELECT palabra, COUNT(*) AS frecuencia
FROM (
    SELECT explode(split(contenido, ' ')) AS palabra
    FROM texto
) AS palabras
GROUP BY palabra;
```

Y por último después de ejecutar el query se revisa que los resultados que se esperan si estén dentro de la tabla de destino:

```
SELECT * FROM frecuencias_palabras;
```

También si es necesario se envía el resultado de la consulta a una carpeta en hue o S3 y subirlo a la base de datos de destino

Laboratorio 3.4 (Spark):



Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analytics.

You can make beautiful data-driven, interactive, collaborative document with SQL, code and even more!

Notebook

Import note

Create new note

Help

Get started with [Zeppelin documentation](#)

Community

Please feel free to help us to improve Zeppelin,
Any contribution are welcome!

Mailing list

Issues tracking

Github

Create New Note



Note Name

Default Interpreter



Use '/' to create folders. Example: /NoteDirA/Note1

Create

```
%spark2| pyspark
# WORDCOUNT COMPACTO
#files_rdd = sc.textFile("s3://emontoyadatasets/gutenberg-small/*.txt")
files_rdd = sc.textFile("hdfs:///datasets/gutenberg-small/*.txt")
wc_unsort = files_rdd.flatMap(lambda line: line.split()).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
wc = wc_unsort.sortBy(lambda a: -a[1])
for tupla in wc.take(10):
    print(tupla)
wc.coalesce(1).saveAsTextFile("hdfs:///tmp/wcout1")
```