

## 1. EMR HIVE / SPARK SQL

- Realice las actividades de hive con
  - tablas nativas

### Datos:

Search data and saved documents...

File Browser

Search for file name

Actions Delete forever

Upload New

Home / user / hive / warehouse / hdi

Name	Size	Usuario	Group	Permisos	Date
.		hdfs	hdfsadmingroup	drwxrwxrwt	June 01, 2023 10:46 AM
.		hadoop	hdfsadmingroup	drwxr-xr-x	June 01, 2023 10:49 AM
hdi-data.csv	9.0 KB	hadoop	hdfsadmingroup	-rw-r--r--	June 01, 2023 10:49 AM

### Tablas:

Hive

Add a name... Add a description...

0.44s default

```
select * from hdi
```

INFO : Completed executing command(queryId=hive\_20230601175130\_b36f8478-4b08-47c7-b297-84072e22982b); Time taken: 0.0 seconds

INFO : OK

INFO : Concurrency mode is disabled, not creating a lock manager

Query History Saved Queries Results (100+)

hdi.id	hdi.country	hdi.lifeex	hdi.mysch	hdi.eysch	hdi.eysch
1	NULL	country	NULL	NULL	NULL
2	1	Norway	0.943	81	12
3	2	Australia	0.929	81	12
4	3	Netherlands	0.91	80	11
5	4	United States	0.91	78	12
6	5	New Zealand	0.908	80	12
7	6	Canada	0.908	81	12
8	7	Ireland	0.908	80	11

- tablas externas

### Datos:

Search for file name

Actions Delete forever

Upload New

Home / user / hadoop / onu

Name	Size	Usuario	Group	Permisos	Date
.		hadoop	hdfsadmingroup	drwxrwxrwt	June 01, 2023 10:43 AM
.		hadoop	hdfsadmingroup	drwxr-xr-x	June 01, 2023 10:43 AM
hdi-data.csv	9.0 KB	hadoop	hdfsadmingroup	-rw-r--r--	June 01, 2023 10:43 AM

Show 45 of 1 items

Page 1 of 1

### Tablas:



default

Tables (3) +

Filter...

hdi

hdi\_ext

hdi\_s3

1

select \* from hdi\_s3

▶

0.41s default

?

INFO : Completed executing command(queryId=hive\_20230601180206\_b42fd7c2-c763-4e56-9793-35b29f8b7d65); Time taken: 0.0 seconds

INFO : OK

INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (100+)

	hdi_s3.id	hdi_s3.country	hdi_s3.hdi	hdi_s3.lifeex	hdi_s3.mysch	hdi_s3.eysch	hdi_s3.g
1	NULL	airline	NULL	NULL	NULL	NULL	NULL
2	10001	Delta Air Lines	NULL	NULL	7	NULL	4
3	10002	Delta Air Lines	NULL	NULL	0	NULL	2
4	10003	Delta Air Lines	NULL	NULL	0	NULL	1
5	10004	Delta Air Lines	NULL	NULL	9	NULL	4
6	10005	Delta Air Lines	NULL	NULL	7	NULL	3
7	10006	Delta Air Lines	NULL	NULL	9	NULL	5
8	10007	Delta Air Lines	NULL	NULL	0	NULL	1

## Consultas:

default

Tables (3) +

Filter...

hdi

hdi\_ext

hdi\_s3

1

select country, gni from hdi where gni > 2000;

▶

0.43s default

?

INFO : Completed executing command(queryId=hive\_20230601180917\_e819771b-8f65-4b02-b197-8c13f78f25a); Time taken: 0.001 seconds

INFO : OK

INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (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

default

Tables (4) +

Filter...

docs

hdi

hdi\_ext

hdi\_s3

1

SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w

2

GROUP BY word

3

ORDER BY word DESC LIMIT 10;

▶

35.23s default

?

INFO : Completed executing command(queryId=hive\_20230601181538\_b053290a-dc35-4-application\_1685639854993\_0001

35.229 seconds

INFO : OK

INFO : Concurrency mode is disabled, not creating a lock manager

Query History

Saved Queries

Results (10)

	word	count
1	Æschines,	1
2	zigzag	1
3	zest	1
4	zenith	1
5	zealously	1
6	zealous,	1
7	zealous	5
8	zeal.	3

12.76s default

```

1 SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w
2 GROUP BY word
3 ORDER BY count DESC LIMIT 10;

```

INFO : Completed executing command(queryId=hive\_20230601181948\_de45a83f-c85c-4... application\_1685639854993\_0001  
12.736 seconds  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager

Query History Saved Queries Results (10)

word	count
1 the	44647
2 of	28020
3 to	27298
4 and	23208
5 in	20444
6 that	13174
7 I	12265
8 a	10880

- b. Realice la ejecución de consultas SQL desde Apache SparkSQL hacia tablas Hive y hacia tablas AWS Glue.

```

In [1]: df = spark.sql("show databases")
Starting Spark application

```

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
1	application_1685639854993_0003	pyspark	idle	<a href="#">Link</a>	<a href="#">Link</a>	None	✓

SparkSession available as 'spark'.

```

In [2]: df.show()
+-----+
|namespace|
+-----+
| default|
+-----+

In [6]: df = spark.sql("select * from default.hdi_ext")
In [7]: df.show()
+-----+-----+-----+-----+-----+-----+
|id|country|hdi|lifeex|mysch|eysch|gni|
+-----+-----+-----+-----+-----+-----+
|null|country|null|null|null|null|null|
|1|Norway|0.943|81|12|17|47557|
|2|Australia|0.929|81|12|18|34431|
|3|Netherlands|0.91|80|11|16|36402|
|4|United States|0.91|78|12|16|43017|
|5|New Zealand|0.908|80|12|18|23737|
|6|Canada|0.908|81|12|16|35166|
|7|Ireland|0.908|80|11|18|29322|
|8|Liechtenstein|0.905|79|10|14|83717|
|9|Germany|0.905|80|12|15|34854|
|10|Sweden|0.904|81|11|15|35837|
|11|Switzerland|0.903|82|11|15|39924|
|12|Japan|0.901|83|11|15|32295|
|13|Hong Kong China (...)|0.898|82|10|15|44005|
|14|Iceland|0.898|81|10|18|29354|
|15|Korea (Republic of)|0.897|80|11|16|28238|
|16|Denmark|0.895|78|11|16|34347|
|17|Israel|0.888|81|11|15|25849|
|18|Belgium|0.886|80|10|16|33357|
|19|Austria|0.885|80|10|15|35719|
+-----+-----+-----+-----+-----+-----+

Crawler successfully starting
The following crawler is now starting: "Reto6.1"

AWS Glue > Tables > hdi
hdi
Last updated (UTC) June 1, 2023 at 18:57:26 Version 0 (Current version) Actions
Table overview Data quality New
Table details Advanced properties
Name hdi Description - Database st0263d6 Classification csv
Location s3://st0263jasanchez/datasets/onu/hdi/ Connection - Deprecated Last updated June 1, 2023 at 18:57:26
Input format org.apache.hadoop.mapred.TextInputFormat Output format org.apache.hadoop.hive.q1.io.HiveIgnoreKeyTextOutputFormat Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

```

```
In [9]: df = spark.sql("show databases")

In [10]: df.show()

+-----+
|namespace|
+-----+
| default|
| st0263db|
+-----+

In [11]: df = spark.sql("select * from st0263db.hdi")

In [12]: df.show()

+-----+-----+-----+-----+-----+-----+-----+-----+
| id|country|hdi|lifeex|myschool|eyschool|gn1|gn12|nihdi|
+-----+-----+-----+-----+-----+-----+-----+-----+
| null|country| null| null| null| null| null| null| null|
| 1|Norway|0.943|81.1|12.6|17.3|47557|6|0.975|
| 2|Australia|0.929|81.9|12.0|18.0|34431|16|0.979|
| 3|Netherlands|0.91|80.7|11.6|16.0|36402|9|0.944|
| 4|United States|0.91|78.5|12.4|16.0|43017|6|0.931|
| 5|New Zealand|0.908|80.7|12.5|18.0|23737|30|0.978|
| 6|Canada|0.908|81.0|12.1|16.0|35166|10|0.944|
| 7|Ireland|0.908|80.6|11.6|18.0|29322|19|0.959|
| 8|Liechtenstein|0.905|79.6|10.3|14.7|83717|-6|0.877|
| 9|Germany|0.905|80.4|12.2|15.9|34854|8|0.94|
| 10|Sweden|0.904|81.4|11.7|15.7|35837|4|0.936|
| 11|Switzerland|0.903|82.3|11.0|15.6|39924|0|0.926|
| 12|Japan|0.901|83.4|11.6|15.1|32295|11|0.94|
| 13|Hong Kong China (...)|0.898|82.8|10.0|15.7|44805|-4|0.91|
| 14|Iceland|0.898|81.8|10.4|18.0|29354|11|0.943|
| 15|Korea (Republic of)|0.897|80.6|11.6|16.9|28230|12|0.945|
| 16|Denmark|0.895|78.8|11.4|16.9|34347|3|0.926|
| 17|Israel|0.888|81.0|11.9|15.5|25849|14|0.939|
| 18|Belgium|0.886|80.0|10.9|16.1|33357|2|0.914|
| 19|Austria|0.885|80.9|10.8|15.3|35719|-4|0.908|
+-----+-----+-----+-----+-----+-----+-----+-----+

only showing top 20 rows
```

## 2. IMPLEMENTACIÓN DE UN DATA WAREHOUSE SENCILLO CON AWS ATHENA

### Catalogando datos:

AWS Glue > Databases > onudb

onudb Last updated (UTC) June 1, 2023 at 20:17:57 [Refresh](#) [Edit](#) [Delete](#)

**Database properties**

Name onudb	Description -	Location -	Created on (UTC) June 1, 2023 at 20:16:16
---------------	------------------	---------------	--

**Tables (1)** Last updated (UTC) June 1, 2023 at 20:17:58 [Refresh](#) [Delete](#) [Data quality](#) [New](#) [Add tables using crawler](#) [Add table](#)

View and manage all available tables.

<input type="checkbox"/>	Name	Database	Location	Classification	Deprecated	View data
<input type="checkbox"/>	hdi	onudb	s3://st0263jasanchez/datasets/onu/h	csv	-	<a href="#">Table data</a>

### Consultas:

#### Athena

Data

Data source

AwsDataCatalog

Database

onudb

Tables and views

Create

Filter tables and views

Tables (1)

< 1 >

hdi

Views (0)

< 1 >

Query 6

1 SELECT \* FROM "onudb"."hdi" limit 10;

SQL Ln 1, Col 1

Run again

Explain

Cancel

Clear

Create

Query results

Query stats

Completed

Time in queue: 311 ms

Run time: 832 ms

Data scanned: 1.08 KB

Results (10)

Copy

Download results

Search rows

#	id	country	hdi	lifeex	myschool	eyschool	gni	gni2	nihdi
1		"Norway"							
2		"Australia"							
3		"Netherlands"							

## Hive

Hive

Add a name...

Add a description...

0.42s onudb

?

onudb

Tables (1)

Filter...

hdi

select \* from onudb.hdi limit 10

INFO : Compiling command(queryId=hive\_20230601202613\_d92d4c4a-4c70-458e-85d5-9711ecc28e86): select \* from onudb.hdi limit 10

INFO : Concurrency mode is disabled, not creating a lock manager

INFO : Semantic Analysis Completed (retrial = false)

INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name=hdi.id, type=bigint, comment=null), FieldSchema(name=hdi.country, type=string, comment=null), FieldSchema

Query History

Saved Queries

Results (10)

	hdi.id	hdi.country	hdi.hdi	hdi.lifeex	hdi.myschool	hdi.eyschool	hdi.gni	hdi.gni2	hdi.nihdi
1	NULL	"Norway"	NULL	NULL	NULL	NULL	NULL	NULL	NULL
2	NULL	"Australia"	NULL	NULL	NULL	NULL	NULL	NULL	NULL
3	NULL	"Netherlands"	NULL	NULL	NULL	NULL	NULL	NULL	NULL
4	NULL	"United States"	NULL	NULL	NULL	NULL	NULL	NULL	NULL
5	NULL	"New Zealand"	NULL	NULL	NULL	NULL	NULL	NULL	NULL
6	NULL	"Canada"	NULL	NULL	NULL	NULL	NULL	NULL	NULL
7	NULL	"Ireland"	NULL	NULL	NULL	NULL	NULL	NULL	NULL
8	NULL	"Liechtenstein"	NULL	NULL	NULL	NULL	NULL	NULL	NULL
9	NULL	"Germany"	NULL	NULL	NULL	NULL	NULL	NULL	NULL
10	NULL	"Sweden"	NULL	NULL	NULL	NULL	NULL	NULL	NULL

## Spark

```
In [2]: df = spark.sql("show databases")

In [3]: df.show()

+-----+
|namespace|
+-----+
| default|
| onudb|
+-----+

In [9]: df = spark.sql("select * from onudb.hdi")

In [10]: df.show()

+-----+-----+-----+-----+-----+-----+-----+-----+
| id|country|hdi|lifeex|myschool|eyschool|gni|gni2|ni|hdi|
+-----+-----+-----+-----+-----+-----+-----+-----+
| null|country| null| null| null| null| null| null| null|
| 1|Norway|0.943|81.1|12.6|17.3|47557|6|0.975|
| 2|Australia|0.929|81.9|12.0|18.0|34431|16|0.979|
| 3|Netherlands|0.91|80.7|11.6|16.8|36402|9|0.944|
| 4|United States|0.91|78.5|12.4|16.0|43017|6|0.931|
| 5|New Zealand|0.908|80.7|12.5|18.0|23737|30|0.978|
| 6|Canada|0.908|81.0|12.1|16.0|35166|10|0.944|
| 7|Ireland|0.908|80.6|11.6|18.0|29322|19|0.959|
| 8|Liechtenstein|0.905|79.6|10.3|14.7|83717|-6|0.877|
| 9|Germany|0.905|80.4|12.2|15.9|34854|8|0.94|
| 10|Sweden|0.904|81.4|11.7|15.7|35837|4|0.936|
| 11|Switzerland|0.903|82.3|11.0|15.6|39924|0|0.926|
| 12|Japan|0.901|83.4|11.6|15.1|32295|11|0.94|
| 13|Hong Kong China (...)|0.898|82.8|10.0|15.7|44805|-4|0.91|
| 14|Iceland|0.898|81.8|10.4|18.0|29354|11|0.943|
| 15|Korea (Republic of)|0.897|80.6|11.6|16.9|28230|12|0.945|
| 16|Denmark|0.895|78.8|11.4|16.9|34347|3|0.926|
| 17|Israel|0.888|81.6|11.9|15.5|25849|14|0.939|
| 18|Belgium|0.886|80.0|10.9|16.1|33357|2|0.914|
| 19|Austria|0.885|80.9|10.8|15.3|35719|-4|0.908|
+-----+-----+-----+-----+-----+-----+-----+-----+

only showing top 20 rows
```

3. IMPLEMENATCIÓN DE UN DATA WAREHOUSE CON AWS REDSHIFT

Detalles del rol

Nombre del rol

Ingresar un nombre significativo para identificar a este rol.

myspectrum\_rol

Máximo de 64 caracteres. Utilice caracteres alfanuméricos y "-","\_","."

Descripción

Agregue una breve explicación para este rol.

Allows Redshift clusters to call AWS services on your behalf.

1000 caracteres como máximo. Utilice caracteres alfanuméricos y "-","\_","."

Paso 1: seleccionar entidades de confianza

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

{

"Version": "2012-10-17",

"Statement": {

{

"Effect": "Allow",

"Action": [

"sts:AssumeRole"

],

"Principal": {

"Service": {

"redshift.amazonaws.com"

}

}

}

}

}

}

Paso 2: agregar permisos

Resumen de la política de permisos

Nombre de la política

Tipo

Adjuntado como

AmazonAthenaFullAccess

Administrada por AWS

Política de permisos

AWSGlueConsoleFullAccess


Administrada por AWS

Política de permisos

AmazonS3ReadOnlyAccess

Administrada por AWS

Política de permisos

**No se pudo crear el rol myspectrum\_role.**  
User: arn:aws:sts:725662720187:assumed-role/vodabs/user1891167-jasanchez@eaffl.edu.co is not authorized to perform: iam:CreateRole on resource: arn:aws:iam:725662720187:role/myspectrum\_role because no identity-based policy allows the iam:CreateRole action

[IAM](#) > [Roles](#) > [Crear rol](#)

Paso 1  
[Seleccionar entidad de confianza](#)

Paso 2  
[Agregar permisos](#)

Paso 3  
**Asignar nombre, revisar y crear**

## Asignar nombre, revisar y crear

### Detalles del rol

#### Nombre del rol

Ingrese un nombre significativo para identificar a este rol.

myspectrum\_role

Máximo de 64 caracteres. Utilice caracteres alfanuméricos y '+', '@', '-'.

#### Descripción

Agregue una breve explicación para este rol.

Allows Redshift clusters to call AWS services on your behalf.

ARN: arn:aws:iam::725662720187:role/myspectrum\_role