2/5/2020 OneNote

Hive & Sqoop

miércoles, 29 de abril de 2020 15:52

• Conexión con la DB

```
[ec2-user@ip-172-31-93-194 ~]$ mysql -u admin -p -h database-1.czsfk9ugozci.u:
ast-1.rds.amazonaws.com
inter password:
Welcome to the MySQL monitor. Commands end with ; or \g.
Kour MySQL connection id is 2152
Server version: 8.0.17 Source distribution
  opyright (c) 2000, 2018, Oracle and/or its affiliates. All rights reserved.
Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
```

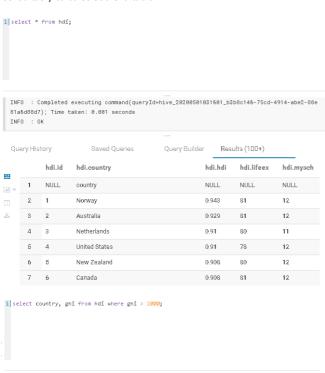
· Crear base de datos



• Crear tabla "hdi" con datos de s3

```
1 SHOW TABLES;
2 DESCRIBE hdi;
 INFO : Compiling command(queryId=hive_20200501031410_4ba81de5-8712-42f4-b15e-0516e72c7be
DESCRIBE hdi
                      Saved Queries Query Builder Results (7)
 Query History
                        data_type
112
     1 id
                                  int
ail w
      2 country
                                   string
     3 hdi
                                   float
      4 lifeex
                                   int
      5 mysch
    6 evsch
                                  int
```

• Consultas y cálculos sobre la tabla HDI:

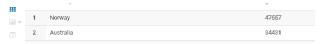


INFO : Compiling command(queryId=hive_20200501031816_132d0c2f-daa7-455e-baa4-d3c99796455

Saved Queries Query Builder Results (100+)

3): select country, gni from hdi where gni > 2000 INFO : Semantic Analysis Completed

Query History



• Ejecutar join con hive

Se crea la tabla expo

```
CREATE EXTERNAL TABLE EXPO (country STRING, expct FLOAT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION
's3://amartinezvdatasets/datasets/onu/datasets/onu/export/'
hace unos segundos 🗸 🗸
```

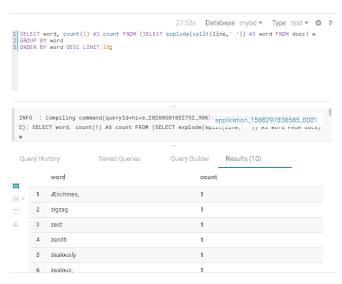
Join

WORDCOUNT EN HIVE:

• Creamos la tabla que tendrá todos los datos de gutenbergsmall

```
CREATE EXTERNAL TABLE docs (line STRING)
STORED AS TEXTFILE
LOCATION
'331/Jamartinezvdatasets/datasets/onu/datasets/gutenberg-small'
```

ordenado por palabra



ordenado por frecuencia de menor a mayor

```
I SELECT word, count(1) AS count FRDM (SELECT explode(split(line,' ')) AS word FRDM docs) w 2 GROUP BY word 3 ORDER BY count DESC LIMIT 10;
  INFO : Compiling command(queryId=hive_28288581833883_c3b5 application_1588297838585_0001
f): SELECT word, count(1) AS count FROM (SELECT explode(spilitizing, )) As word FROM doce)
  Query History
                                   Saved Queries
                                                                  Query Builder
                                                                                        Results (10)
112
         1 the
                                                                       44647
         2 of
                                                                       28020
                                                                       27298
         4 to
                                                                       23208
         5
               and
                                                                       20444
         6
              in
                                                                        13174
```

RETO:

¿cómo llenar una tabla con los resultados de un Query? por ejemplo, como almacenar en una tabla el diccionario de frecuencia de palabras en el wordcount?

Creamos una tabla y guardamos los resultados ahí:

```
CREATE TABLE result as (SELECT word, count(1) AS count FROM (SELECT explode(split(line, ' ')) AS word FROM docs) w GROUP BY word ORDER BY count DESC LIMIT 10)
```

Consultamos la tabla result



/ user / hive / warehouse / mybd.db / result / 000000_0

```
theD44647
of□28020
□27298
to023208
and 0 2 0 4 4 4
inD13174
that 12265
ID10880
a 🗆 10431
isD7776
```

Apache Sqoop

· Nos conectamos al clúster vía ssh y buscamos la lib correspondiente

```
hadoop@ip-172-31-86-245 ~]$ hdfs dfs -ls /user/oozie/share/lib/
Found 1 items
drwxr-xr-x
             - oozie oozie
                                      0 2020-05-01 20:01 /user/oozie/share/lib/lib_20200501200113
hadoop@ip-172-31-86-245 ~]$
```

Configuraciones

```
[hadoop@ip-172-31-86-245 ~]$ hdfs dfs -put /usr/share/java/mysql-connector-java.jar /user/oozie/share/lib/lib_20200501200113/sqoop/mysql-connector-java.jar': file exists [hadoop@ip-172-31-86-245 ~]$ hdfs dfs -chown oozie /user/oozie/share/lib/lib_2020050120013/sqoop/mysql-connector-java.jar [hadoop@ip-172-31-86-245 ~]$ hdfs dfs -chown oozie /user/oozie/share/lib/lib_2020050120013/sqoop/mysql-connector-java.jar [hadoop@ip-172-31-86-245 ~]$ hdfs dfs -cp /user/oozie/share/lib/lib_2020050120013/sqoop/mysql-connector-java.jar [hadoop@ip-172-31-86-245 ~]$ hdfs dfs -cp /user/oozie/share/lib/lib_2020050120013/sqoop/sqoop/spackeds-sesors-smart-1.2.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/apacheds-i18n-2.0.0-H15.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/apacheds-i18n-2.0.0-H15.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/apacheds-harb-1.0.0-H20.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/apa-sn1-api-1.0.0-H20.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/apa-sn1-api-1.0.0-H20.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/commons-codec-1.4.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/commons-codec-1.4.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/commons-collections-3.2.2.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/commons-jex1-2.1.1.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/commons-jex1-2.1.1.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/commons-longing-1.1.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/commons-longing-1.1.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/commons-longing-1.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/commons-longing-1.jar': file exists cp: /user/oozie/share/lib/lib_20200501200113/sqoop/commons-longing-1.jar': file exists cp: /user/oozie/share/lib/lib_202005
        cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/guava-11.0.2.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/hadoop-auth-2.8.5-amzn-4.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/httpclient-4.5.9.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/httpcore-4.4.11.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/jcip-annotations-1.0-1.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/jcip-annotations-1.0-1.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/jcip-annotations-1.0-1.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/jsino-smart-2.3.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/jsr305-3.0.0.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/nombus-jose-jwt-4.41.1.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/slf4j-api-1.6.6.jan': File exists
cp: \/user/oozie/share/lib/lib_20200501200113/sqoop/slf4j-api-1.6.6.jan': File exists
[hadoop@ip-172-31-86-245 ~]$ hdfs dfs -chown oozie /user/oozie/share/lib/lib_20200501200113/sqoop/*
[hadoop@ip-172-31-86-245 ~]$ hdfs dfs -chown oozie /user/oozie/share/lib/lib_20200501200113/sqoop/*
[hadoop@ip-172-31-86-245 ~]$
```

• Creamos la base de datos cursodb y la tabla employee

```
use changed
[cursodb]> CREATE TABLE `cursodb`.`employee` (   `emp_id` INT NOT NULL,
/ARCHAR(45),    `salary` INT, PRIMARY KEY (`emp_id`));
OK, O rows affected (0.04 sec)
       [cursodb]> CREATE USER 'curso'@'%' IDENTIFIED BY 'curso'; OK, O rows affected (0.01 sec)
ySQL [cursodb]> GRANT ALL PRIVILEGES ON cursodb.* TO 'curso'@'%';
uery OK, O rows affected (0.01 sec)
```

Llenamos la tabla

```
MySQL [cursodb]> insert into employee values (101, 'name1', 1800);
```

2/5/2020 OneNote

```
luery OK, 1 row affected (0.01 sec)
MySQL [cursodb]> insert into employee values (102, 'name2', 1500);
Query OK, 1 row affected (0.01 sec)
MySQL [cursodb]> insert into employee values (103, 'name3', 1000);
Query OK, 1 row affected (0.01 sec)
MySQL [cursodb]> insert into employee values (104, 'name4', 2000);
Query OK, 1 row affected (0.01 sec)
MySQL [cursodb]> insert into employee values (105, 'name5', 1600);
Query OK, 1 row affected (0.01 sec)
```

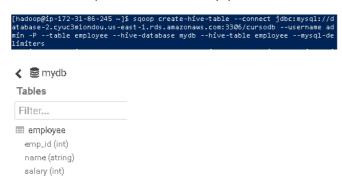
Transferir datos de una base de datos (tipo mysql) hacia HDFS:

t jdbc:mysql://database-2.cyuc3m1ondou.us-east-1.rds.amazonaws.com:3306/curso /user/admin/mysqlOut -m 1 [hadoop@ip-172-31-86-245 ~]\$ sqoop import --connect --username admin -P --table employee --target-dir

· Listamos los archivos:

```
op@ip-172-31-86-245 ~]$ hdfs dfs -ls /user/admin/mysqlOut
2 items
                                       0 2020-05-01 21:48 /user/admin/mysqlOut/_SUCCESS
75 2020-05-01 21:48 /user/admin/mysqlOut/part-m-0
         / user / admin / mysqlOut / part-m-00000
101, name1, 1800
102, name2, 1500
103, name3, 1000
104, name4, 2000
105,name5,1600
```

• Crear tabla HIVE a partir de definición tabla Mysql:



• Transferir datos de una base de datos (tipo mysql) hacia HIVE vía HDFS:

[hadoop@ip-172-31-86-245 ~]\$ sqoop import --connect jdbc:mysql://database-2.cyuc3mlondou.us-east-1.rds.amazo naws.com:3306/cursodb --username admin -P --table employee --hive-import --hive-database mydb --hive-table e mployee --mysql-delimiters

