Universidad Católica
San Pablo

PROGRAMA PROFESIONAL

Ciencias de la Computación

CURSO

Big Data

TEMA

Ejercicios Hive - Hadoop

Integrantes:

- Loayza Huarachi Angel Josue

SEMESTRE: VIII

AÑO: 2024

“Los alumnos declaran haber realizado el presente trabajo de acuerdo a las normas de la Universidad Católica San Pablo

1. Introducción

Este informe presenta los resultados de una serie de ejercicios realizados con Hive. A través de un conjunto de consultas SQL, se abordaron diversas tareas analíticas sobre un conjunto de datos simulado. Los ejercicios incluyen el clásico análisis de **Wordcount**, el cálculo del número de entradas en los logs por usuario, la determinación del promedio de visitas por usuario, y la identificación de los usuarios que acceden, en promedio, a las "páginas mejor rankeadas". El código utilizado para estas consultas se encuentra en el archivo `script.sql` del repositorio adjunto:

2. WordCount

- Código Necesario:

```sql
-- Creacion de la Tabla (input)
CREATE EXTERNAL TABLE IF NOT EXISTS wordcount (
    line STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\n'
STORED AS TEXTFILE
LOCATION '/user/hive/warehouse/employees/';

-----------------------------------------
-- Creacion de la tabla resultados
CREATE TABLE IF NOT EXISTS wordcount_results AS
SELECT word, COUNT(*) AS COUNT
FROM (
    SELECT explode(split(line, ' ')) AS word
    FROM wordcount
) tmp
GROUP BY word;
```

- Input del ejercicio

```
hive> !cat wordcount_input.csv;
Hola,mundo
Hola,OpenAI
Mundo,de,inteligencia,artificial
Inteligencia,artificial,y,aprendizaje,automático
Hola,a,todos,en,el,mundo
hive>
```

- Creación de la tabla de resultados

```
hive> CREATE TABLE IF NOT EXISTS wordcount_results AS
    > SELECT word, COUNT(*) as count
    > FROM (
    >     SELECT explode(split(line, ',')) as word
    >     FROM wordcount
    > ) tmp
    > GROUP BY word;
Query ID = hadoop_20241010070054_6e888797-28d1-4f1e-b2d5-5ed8c38adb4e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1728538810507_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.52 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-30-76.ec2.internal:8020/user/hive/warehouse/wordcount_results
OK
Time taken: 4.849 seconds
hive>
```

- Ver contenido de la tabla de resultados

```
hive> SELECT * FROM wordcount_results ORDER BY count DESC;
Query ID = hadoop_20241010070230_e05e8a11-8314-470b-9432-04bf4e119341
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1728538810507_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.71 s
--------------------------------------------------------------------------------
OK
Hola    3
artificial      2
mundo   2
Inteligencia    1
Mundo   1
OpenAI  1
automático      1
el      1
en      1
todos   1
y       1
a       1
aprendizaje     1
de      1
inteligencia    1
Time taken: 4.926 seconds, Fetched: 15 row(s)
hive>
```

3. Calculando el número de entradas en el log por cada usuario

- Código Necesario:

```
-- Creacion de la tabla logs
CREATE EXTERNAL TABLE IF NOT EXISTS logs (
    user_a STRING,
    time STRING,
    QUERY STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
LOCATION '/user/hive/warehouse/logUser/';

----------------------------------------
-- Creacion de la tabla de resultados
CREATE TABLE IF NOT EXISTS result AS
SELECT user_a, COUNT(1) AS log_entries
FROM logs
GROUP BY user_a
ORDER BY user_a;
```

- Input del ejercicio:

```
hive> !cat logUser_input.csv;
user123456      1234567890      yahoo chat
user234567      1234567891      foods
user123456      1234567892      yahoo
user345678      1234567893      spiders
user234567      1234567894      yahoo,chat
user123456      1234567895      foods
user345678      1234567896      yahoo chat
user123456      1234567897      spiders
user234567      1234567898      foods
user345678      1234567899      yahoo
hive>
```

- Creación de la tabla de resultados:

```
hive> CREATE TABLE IF NOT EXISTS logsUser_result AS
    > SELECT user_a, COUNT(1) AS log_entries
    > FROM logsUser
    > GROUP BY user_a
    > ORDER BY user_a;
Query ID = hadoop_20241010072801_0dcf7576-7035-40b7-864d-aa860ff55a48
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1728538810507_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.27 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-30-76.ec2.internal:8020/user/hive/warehouse/logsuser_result
OK
Time taken: 4.812 seconds
hive>
```

- Ver contenido de la tabla de resultados

```
hive> SELECT * FROM logsUser_result ORDER BY log_entries DESC;
Query ID = hadoop_20241010073032_2b3e10fb-21f3-4b88-b0a9-e62b5640169e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1728538810507_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.56 s
--------------------------------------------------------------------------------
OK
user123456      4
user234567      3
user345678      3
Time taken: 7.62 seconds, Fetched: 3 row(s)
hive>
```

4. Calculando el promedio de visitas por cada usuario

- Código necesario:

```
-- Creacion de la tabla visitUser
CREATE EXTERNAL TABLE IF NOT EXISTS visitUser (
    name STRING,
    url STRING,
    time_a STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
LOCATION '/user/hive/warehouse/visitsUser/';

-----------------------------------------
-- Creacion de la tabla de resultados
CREATE TABLE IF NOT EXISTS visitUser_result AS
SELECT AVG(num_pages) AS avg_visits
FROM (
    SELECT name, COUNT(1) AS num_pages
    FROM visitUser
    GROUP BY name
) np;
```

- Input del ejercicio

```
[hadoop@ip-172-31-30-76 ~]$ cat visitUser_input.csv
user123456      http://example.com/page1        12:00
user234567      http://example.com/page2        12:05
user123456      http://example.com/page3        12:10
user345678      http://example.com/page4        12:15
user234567      http://example.com/page5        12:20
user123456      http://example.com/page1        12:25
user345678      http://example.com/page6        12:30
user234567      http://example.com/page7        12:35
user345678      http://example.com/page8        12:40
[hadoop@ip-172-31-30-76 ~]$
```

- Creación de la tabla de resultados

```
hive> CREATE TABLE IF NOT EXISTS visitUser_result AS
    > SELECT AVG(num_pages) AS avg_visits
    > FROM (
    >     SELECT name, COUNT(1) AS num_pages
    >     FROM visitUser
    >     GROUP BY name
    > ) np;
Query ID = hadoop_20241010075721_e3241129-7128-4972-818e-1c6ef9db6f4f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1728538810507_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.59 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-30-76.ec2.internal:8020/user/hive/warehouse/visituser_result
OK
Time taken: 7.264 seconds
hive>
```

- Ver el contenido de la tabla de resultados

```
hive> SELECT * FROM visitUser_result;
OK
3.0
Time taken: 0.057 seconds, Fetched: 1 row(s)
hive>
```

5. Identificar cuáles usuarios visitan “Página mejores rankeadas” en promedio

- Código necesario

```
-- Creacion de las tablas visits y pages:
CREATE EXTERNAL TABLE IF NOT EXISTS rankVisits (
    name STRING,
    url STRING,
    time_a STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
LOCATION '/user/hive/warehouse/userRank1';

CREATE EXTERNAL TABLE IF NOT EXISTS rankPages (
    url STRING,
    pagerank FLOAT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
LOCATION '/user/hive/warehouse/userRank2';
```

```
-- Crear la tabla de resultados:
CREATE TABLE IF NOT EXISTS rank_results AS
SELECT pr.name
FROM (
    SELECT V.name, AVG(P.pagerank) AS prank
    FROM rankVisits V
    JOIN rankPages P ON (V.url = P.url)
    GROUP BY V.name
) pr
WHERE pr.prank > 0.5;
```

- Input del ejercicio

```
[hadoop@ip-172-31-30-76 ~]$ cat visitsRank.log
user123456      http://example.com/page1        12:00
user234567      http://example.com/page2        12:05
user123456      http://example.com/page3        12:10
user345678      http://example.com/page4        12:15
user234567      http://example.com/page5        12:20
user111111      http://example.com/page1        12:25
user345678      http://example.com/page6        12:30
user222222      http://example.com/page7        12:35
user345678      http://example.com/page8        12:40
[hadoop@ip-172-31-30-76 ~]$ cat pagesRank.log
http://example.com/page1        0.6
http://example.com/page2        0.4
http://example.com/page3        0.7
http://example.com/page4        0.3
http://example.com/page5        0.8
http://example.com/page6        0.9
http://example.com/page7        0.5
http://example.com/page8        0.1
[hadoop@ip-172-31-30-76 ~]$
```

- Creación de la tabla de resultados

```
hive> CREATE TABLE IF NOT EXISTS rank_results AS
    > SELECT pr.name
    > FROM (
    >     SELECT V.name, AVG(P.pagerank) AS prank
    >     FROM rankVisits V
    >     JOIN rankPages P ON (V.url = P.url)
    >     GROUP BY V.name
    > ) pr
    > WHERE pr.prank > 0.5;
Query ID = hadoop_20241010082303_f94fd465-2506-4481-a060-26de77467dd7
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1728538810507_0013)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 7.57 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-30-76.ec2.internal:8020/user/hive/warehouse/rank_results
OK
Time taken: 13.573 seconds
```

- Ver contenido de la tabla de resultados

```
hive> SELECT * FROM rank_results;
OK
user111111
user123456
user234567
Time taken: 0.055 seconds, Fetched: 3 row(s)
hive>
```