Exercio HIVE

Exercicio HIVE [obrigatorio]

Neste exercicio realizarás unha serie de consultas sobre un dataset con información do servizo de bikesharing da cidade de Helsinki.

Podes descargar o dataset desde a páxina de Kaggle:

https://www.kaggle.com/datasets/geometrein/helsinki-city-bikes

Para evitar saturación da rede na clase, podes descargar o dataset de lugares alternativos que che proporcione o profesor.

Obxectivos

- Realizar consultas sobre datasets con linguaxes de alto nivel (HiveQL)
- Utilizar un clúster Hadoop para procesar grandes cantidades de datos

Indicacións

NON
FACER PARTE DE PIG

1. Mostra as primeiras filas do dataset para comprender os datos
2. Crea un ficheiro de mostra a partir do ficheiro orixinal para examinar os datos
3. Elimina a primeira liña (Pig non traballa con headers)
4. Usa pig en modo local para experimentar coas consultas con poucos datos
5. Elimina a primeira liña do dataset orixinal
6. Sube o dataset a HDFS
7. Usa pig sobre HDFS para executar as consultas sobre o dataset completo

Preparar os datos

O primeiro é crear a estrutura para dar soporte aos datos (definir a táboa):

```
root@hadoop-master:/vagrant# head database.csv
departure,return,departure_id,departure_name,return_id,retu
rn_name,distance (m),duration (sec.),avg_speed (km/h),depar
ture_latitude,departure_longitude,return_latitude,return_lo
ngitude,Air temperature (degC)

2020-03-23 06:09:44,2020-03-23 06:16:26,86,Kuusitie,111.0,E
sterinportti,1747.0,401.0,0.2613965087281795,60.1952452,24.
9018997,60.1975724,24.9267808,0.9

2020-03-23 06:11:58,2020-03-23 06:26:31,26,Kamppi (M),10.0,
Kasarmitori,1447.0,869.0,0.0999079401611047,60.1686095,24.9
305373,60.1650171805,24.94947287873,0.9
```

Observamos
que os valores están separados por comas e que hai 14 campos.

Crear
táboa:

```
CREATE EXTERNAL TABLE job(

    departure STRING,

    return STRING,

    departure_id STRING,

    departure_name STRING,

    return_id STRING,

    return_name STRING,

    distance STRING,

    duration STRING,
```

```
    avg_speed STRING,

    departure_latitude STRING,

    departure_longitude STRING ,

    return_latitude STRING,

    return_longitude STRING,

    air_temperature STRING) ROW FORMAT DELIMITED

            FIELDS TERMINATED BY ',' LOCATION '/user/xuwira
02/data/jobs' TBLPROPERTIES('skip.header.line.count'='1');
```

Cargar
os datos:

```
hdfs dfs -put database.csv /user/xuwira02/data/jobs
```

Consultas

Utiliza DUMP para mostrar os resultados. Non é necesario que os escribas a disco (STORE).

1. Cantos rexistros ten o dataset
2. Cales son as 5 estacións con maior número de saídas
3. Cales son as 5 estacións con maior número de chegadas
4. Cal é a viaxe coa maior distancia percorrida
5. Canta distancia se ten percorrido en total tendo en conta todas as viaxes
6. Cal é a distancia media percorrida por viaxe
7. Cal é a viaxe de maior duración
8. Cal é a duración media das viaxes
9. Cal é a velocidade media das viaxes

10. Cantas viaxes se fixeron con temperatura menor que 0ºC
11. Cantas viaxes se fixeron con temperatura entre os 15 e 25º
12. Pensa 3 novas consultas que poidan ser interesantes

Pista: podes revisar os notebooks de Kaggle para ver que tipo de preguntas teñen feito ao dataset

Entrega

Pega o código das consultas na páxina de entrega de código no espazo en gris

Pega o resultado das consultas na páxina de entrega de resultados no espazo en gris.

Entrega de código

1. Cantos rexistros ten o dataset

```
select count(*) from job;

12157458
```

2. Cales son as 5 estacións con maior número de saídas

```
select departure_name, count(departure_name) as numero

from job

group by (departure_name)

order by numero desc

limit 5

;
```

3. Cales son as 5 estacións con maior número de chegadas

```
select departure_name, count(departure_name) as numero

from job

group by (departure_name)

order by numero desc

limit 5

;
```

4. Cal é a viaxe coa maior distancia percorrida→ Existen varias

```
SELECT *  FROM jobWHERE distance in (SELECT MAX(CAST(distan
ce as float))FROM job);
```

5. Canta distancia se ten percorrido en total tendo en conta todas as viaxes

```
SELECT SUM(CAST(distance as float)) as distancia_sum FROM j
ob;
```

6. Cal é a distancia media percorrida por viaxe

```
SELECT AVG(CAST(distance as float)) as distancia_media FROM
job;
```

7. Cal é a viaxe de maior duración

```
SELECT *  FROM job WHERE duration in (SELECT MAX(CAST(durat
ion as float)) FROM job);
```

8. Cal é a duración media das viaxes

```
SELECT AVG(CAST(duration as float)) as duracion_media FROM
```

```
job;
```

9. Cal é a velocidade media das viaxes

```
SELECT AVG(CAST(avg_speed as float)) as duracion_media FROM
job;
```

1

0.
Cantas viaxes se fixeron con temperatura menor que 0ºC

```
select count(*) from job where cast(air_temperature as floa
t)<0;
```

11. Cantas viaxes se fixeron con temperatura entre os 15 e 25º

```
select count(*) from job where air_temperature between 15 a
nd 25;
```

12. Pensa 3 novas consultas que poidan ser interesantes

12.1 Años de guardados de datos

```
select year(to_date(departure)) as ano from job group by ye
ar(to_date(departure)) ;
```

12.2
Cuales no cuentan con distancia

```
select count(distance) from job where distance=0;
```

12.3 Registro más reciente de datos

```
select max(to_date(departure))) as ano from job group by ye
ar(to_date(departure)) ;
```