Experimentos ETL con Apache Impala

Big Data I

Francisco Luque Sánchez

01/03/2020

1 Introducción

En esta práctica se mostrará el diseño de un experimento ETL utilizando Apache Impala. En primer lugar, describiremos la base de datos que hemos seleccionado, a continuación cómo podemos cargar la misma en el motor de bases de datos de Impala, y finalmente daremos un par de ejemplos de consulta sobre el conjunto de datos.

Las prácticas se han realizado sobre una máquina virtual aportada por Cloudera, la cual trae el sistema de Apache Impala preinstalado y configurado.

2 Base de datos escogida

La base de datos que hemos escogido para realizar las prácticas se encuentra disponible en https://www.kaggle.com/jessemostipak/hotel-booking-demand. En dicha base de datos se recoge información sobre los clientes que han efectuado reservas en dos establecimientos distintos entre los años 2015 y 2017. Se guarda información sobre la fecha de la estancia, la duración, el número de personas, el tipo de habitación reservada… Toda la información personal de los huéspedes ha sido eliminada por cuestiones de privacidad, por lo que los datos están completamente anonimizados de partida.

En total, el conjunto cuenta con 119390 filas y 32 columnas, y se nos aporta como un archivo en formato CSV.

3 Ingesta de los datos en Impala

Lo primero que estudiaremos es la carga de los datos en el sistema de Impala. Lo primero que tenemos que hacer es crear y seleccionar la base de datos con la que trabajamos:

```
-- Database creation
CREATE DATABASE IF NOT EXISTS hotels LOCATION
↪  '/user/impala/impalahotels.db'
-- Database selection
USE hotels;
```

Una vez creada la base de datos, tenemos que crear la tabla con la que vamos a trabajar. Los nombres de las columnas y los tipos de datos de las mismas han sido extraídos del enlace proporcionado anteriormente, utilizando la pestaña de visualización por columnas de Kaggle, y adaptando el tipo de

dato de alguna de las columnas numéricas, que debido a que hay datos perdidos marcados con la cadena NULL, estaban mal marcadas:

```
-- Table creation (last line ignores header)
CREATE TABLE IF NOT EXISTS HotelBooking (
hotel STRING, canceled INT, lead_time INT, arrival_year INT,
arrival_month STRING, arrival_week INT, arrival_day INT,
days_in_weekend INT, days_in_week INT, adults INT,children INT,
babies INT, meal STRING, country STRING, market_segment STRING,
distribution_channel STRING, repeated_guest INT,
previous_cancellations INT, previous_bookings_not_cancelled INT,
reserved_room_type STRING, assigned_room_type STRING,
booking_changes INT, deposit_type STRING, agent INT, company INT,
days_in_waiting INT, customer_type STRING, adr FLOAT,
parking_places INT, special_requests INT, reservation_status STRING,
reservation_status_date TIMESTAMP
) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\,' LINES TERMINATED BY '\n'
tblproperties("skip.header.line.count"="1");
```

Una vez creada la tabla, ingestamos los datos (no mostramos cómo hemos subido a hdfs):

```
-- Data ingestion
LOAD DATA INPATH '/user/impala/input/hotel_bookings.csv'
OVERWRITE INTO TABLE HotelBooking;
```

Una vez cargados los datos en la base de datos, podemos proceder a hacer consultas.

4 Consultas en formato SQL

Se han realizado dos consultas distintas. En la primera de ellas, se trata de conocer cómo ha evolucionado el alquiler de habitaciones por tipo en los años en los que se tiene registro. Se agrupan las reservas por tipo de habitación y año y se recuenta el número de filas. Se ha ordenado el resultado para hacer más cómoda su visualización:

```
-- First query: Statistics about amount and types of reserved rooms by
-- year
SELECT arrival_year, reserved_room_type, COUNT(*)
FROM HotelBooking
GROUP BY reserved_room_type, arrival_year
ORDER BY reserved_room_type, arrival_year;
```

El resultado es el siguiente:

arrival_year	room_type	count(*)	arrival_year	room_type	count(*)
2015	A	17720	2015	F	374
2016	A	40718	2016	F	1424
2017	A	27556	2017	F	1099
2015	B	244	2015	G	260
2016	B	672	2016	G	997
2017	B	202	2017	G	837
2015	C	171	2015	H	82
2016	C	282	2016	H	306
2017	C	479	2017	H	213
2015	D	2186	2015	L	6
2016	D	9421	2016	P	6
2017	D	7594	2017	P	6
2015	E	953			
2016	E	2881			
2017	E	2701			

Esto nos permite sacar información interesante sobre las reservas. En primer lugar, tenemos que el tipo de habitación más reservado es el tipo A, con una diferencia sustancial. Por otra parte, el año 2016 fue el más productivo, ya que para todos los tipos de habitaciones es el año en el que más reservas se hicieron, a excepción del tipo C, que experimenta un repunte en el 2017.

La siguiente consulta que se ha realizado trata de establecer el perfil de los viajeros españoles para el mes de agosto. Estaremos interesados en la composición de las familias, así como el número de familias de cada tipo. En la consulta, por tanto, se seleccionan las filas tales que `arrival_month = "August"` y `country = "ESP"`, se agrupa por número de adultos, niños y bebes, se seleccionan esas columnas y el número de repeticiones de cada combinación. Además, se ordena por conteo descendente. La consulta es la siguiente:

```
-- Second query: We are interested in demographics about Spanish
-- travellers in august, concretely about family compositions
```

```
SELECT adults, children, babies, count(*)
FROM HotelBooking
WHERE country = "ESP" and arrival_month = "August"
GROUP BY adults, children, babies
ORDER BY count(*) DESC;
```

Y el resultado de la misma:

adults	children	babies	count(*)
2	0	0	1179
3	0	0	160
2	2	0	124
2	1	0	121
1	0	0	58
3	1	0	25
2	0	1	21
2	1	1	5
1	1	0	4
0	0	0	2
4	0	0	1
2	3	0	1
0	2	0	1
2	2	1	1
3	0	1	1
3	2	0	1

Donde podemos observar que la composición de viaje más repetida es, con diferencia la pareja de adultos. Después, tenemos tres adultos, que posiblemente hagan referencia a dos padres y un hijo mayor de edad, dos padres con dos hijos, y dos padres con un hijo. También podemos comprobar cómo es poco común viajar con bebés habiendo un toral de 28 viajes en los que hay un bebé involucrado.