Exercicio Hive

Ricardo...

Índice

Dataset	2
Estudando o dataset	
Preparando o dataset	4
Preguntas a resolver	7
Creación de estrutura e carga de datos	8
Consultas Hive	8
Consulta 1: Cantos rexistros correspondentes a viaxes contén o dataset	8
Consulta 2: Cales son as 5 estacións con maior número de saídas	8
Consulta 3: Cales son as 5 estacións con maior número de chegadas	8
Consulta 4: Cal é a viaxe coa maior distancia percorrida	8
Consulta 5: Canta distancia percorreuse en total entre tódalas viaxes	
Consulta 6: Cal é a distancia media percorrida por viaxe	
Consulta 7: Cal é a viaxe de maior duración	
Consulta 8: Cal é a duración media por viaxe	9
Consulta 9: Cal é a velocidade media da viaxes	
Consulta 10: Cantas viaxes se fixeron cando a temperatura era menor de 0°C	9
Consulta 11: Cantas viaxes fixéronse cando a temperatura estaba entre 15 e 25 °C	9
Consulta 12: Datos da primeira viaxe segundo data	9
Resultado da execución das consultas Hive	
Consulta 1: hive_01_viaxes_count	10
Consulta 2: hive_02_estacions_mais_saidas	10
Consulta 3: hive_03_estacions_mais_chegadas	10
Consulta 4: hive_04_viaje_max_distancia	10
Consulta 5: hive_05_total_distancia	10
Consulta 6: hive_06_distancia_media	10
consulta 7: hive_07_viaxe_duracion_maior	10
consulta 8: hive_08_duracion_media_viaxes	10
consulta 09: hive_09_velocidade_media_viaxes	10
consulta 10: hive_10_viaxes_frio	
consulta 11: hive_11_viaxes_temp_ok	11
consulta 12: hive_12_datos_primeira_viaxe	11

Dataset

Estudando o dataset

Nome orixinal do ficheiro:

database. csv

Renomeado como:

bikes database. csv

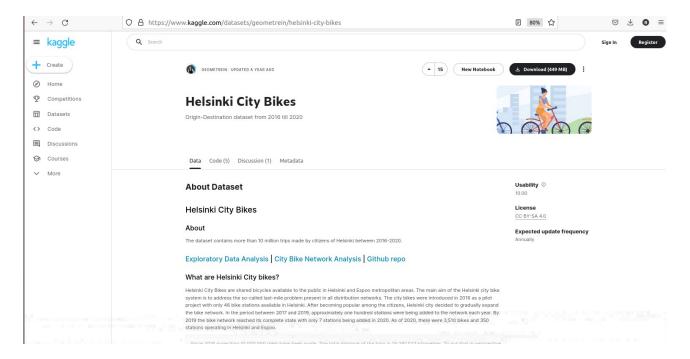
Orixe:

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

Descrición:

Helsinqui City Bikes é o sistema de rede de bicicletas públicas compartidas nas áreas metropolitanas de Helsinqui e Espoo (Finlandia). En 2020 había funcionando 350 estacións e 3.510 bicicletas.

Entre 2016 e 2020 realizáronse máis de 10 millóns de desprazamentos. A distancia total dos devanditos traxectos foi de 25.291.523 Km.



Tamaño do dataset:

2,0 GB (1985327688 bytes)

Tamaño comprimido zip:

439,9 MB (439875920 bytes)

Tamaño comprimido tar. xz:

232,5 MB (232489532 bytes)

Contido:

O dataset contén máis de 10 millóns de rexistros, cada un con datos de desprazamentos en bicicleta

realizados por cidadáns de Helsinqui entre 2016 e 2020, usando a rede de bicicletas compartidas da área Helsinqui e Espoo.

Número total de rexistros:

12.157.458 rexistros

Campos:

14 campos

Nome, tipo e contido dos campos:

float64(8), object(6)

# Colum	Dtype	Contido
0 departure	object	data e hora de partida
1 return	object	data e hora de chegada
2 departure_ide	object	id de saída
3 departure_name	object	nome estación de saída
4 return_ide	object	id de chegada
5 return_name	object	nome estación de chegada
6 distance (m)	float64	distancia (estimada) do traxecto en metros
7 duration (sec.)	float64	duración do traxecto en segundos
8 avg_speed (km/ h)	float64	velocidade media do traxecto (km/h)
9 departure_latitude	float64	latitude estación saída
10 departure_longitude	float64	lonxitude estación saída
11 return_latitude	float64	latitude estación chegada
12 return_longitude	float64	lonxitude estación chegada
13 Air temperature (degC)	float64	temperatura do aire en data do traxecto (°C)

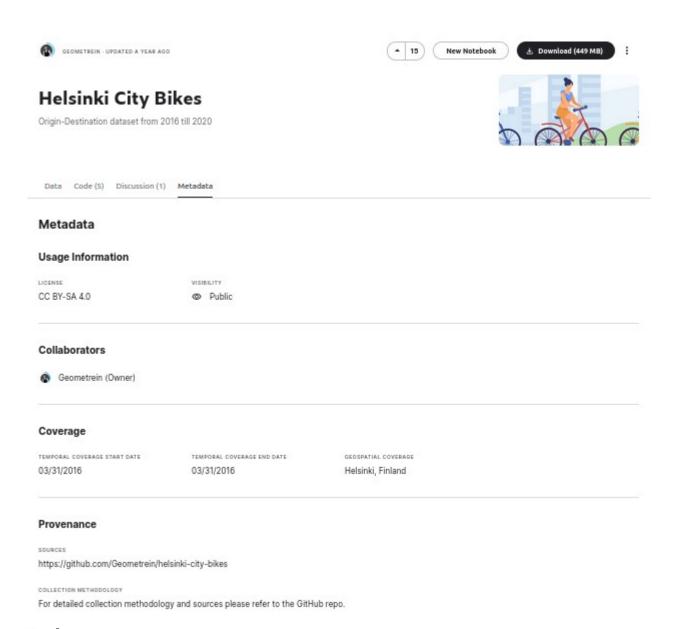
Mostra:

1 df.sam	ple(5)													
	departure	return	departure_id	departure_name	return_id	return_name	distance (m)	duration (sec.)	avg_speed (km/h)	departure_latitude	departure_longitude	return_latitude	return_longitude	Air temperature (degC)
5248712	2019-07-20 23:56:28	2019-07-21 00:05:04	161	Eteläesplanadi	202.0	Merihaka	1731.0	511.0	0.203249	60.167231	24.947466	60.178066	24.958452	16.7
2163223	2020-08-14 15:42:08	2020-08-14 15:58:36	85	Jalavatie	208.0	Valimotie	3618.0	984.0	0.220610	60.193470	24.905889	60.215922	24.876465	18.6
9071247	2018-08-08 21:02:00	2018-08-08 21:15:00	67.0	Perämiehenkatu	41.0	Ympyrätalo	2854.0	804.0	0.212985	60.160088	24.934066	60.180863	24.949400	19.5
214620	2020-04-21 17:28:56	2020-04-21 18:00:10	30	Itämerentori	118.0	Fleminginkatu	6138.0	1873.0	0.196626	60.163531	24.914517	60.189542	24.952160	15.4
3650040	2019-05-08 16:31:40	2019-05-08 16:32:01	18	Porthania	18.0	Porthania	0.0	16.0	0.000000	60.169862	24.948146	60.169862	24.948146	10.4

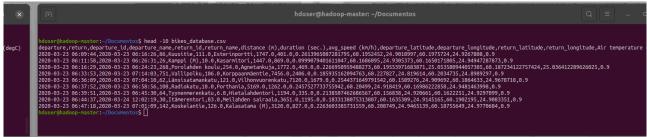
Resumo estatístico:

1 df.	describe()							
	distance (m)	duration (sec.)	avg_speed (km/h)	departure_latitude	departure_longitude	return_latitude	return_longitude	Air temperature (degC)
count	1.215746e+07	1.215746e+07	1.215391e+07	1.215746e+07	1.215746e+07	1.215746e+07	1.215746e+07	1.214156e+07
mean	2.295275e+03	9.597751e+02	3.355556e-01	6.017981e+01	2.492023e+01	6.017971e+01	2.492023e+01	1.565044e+01
std	2.452067e+04	7.346528e+03	3.428006e+01	1.733003e-02	5.764062e-02	1.738792e-02	5.783290e-02	5.497952e+00
min	-4.292467e+06	0.000000e+00	-4.689001e+02	6.014792e+01	2.472137e+01	6.014792e+01	2.472137e+01	-5.200000e+00
25%	1.000000e+03	3.440000e+02	1.467403e-01	6.016723e+01	2.490969e+01	6.016689e+01	2.490969e+01	1.230000e+01
50%	1.739000e+03	5.860000e+02	1.863679e-01	6.017608e+01	2.493407e+01	6.017559e+01	2.493407e+01	1.640000e+01
75%	2.869000e+03	9.710000e+02	2.204348e-01	6.018964e+01	2.495029e+01	6.018964e+01	2.495029e+01	1.930000e+01
max	3.681399e+06	5.401659e+06	1.699104e+04	6.023911e+01	2.510620e+01	6.023911e+01	2.510620e+01	3.290000e+01

Metadatos:



Head:



Obsérvase que o dataset ten ringleira de encabezado e que os campos están separados por coma.

Preparando o dataset

Antes de empeza-las consultas, adecuamo-lo dataset, para o que usaremos *python* e *pandas*, por exemplo.

```
import pandas as pd
df=pd.read_csv('ruta/bikes_database.csv')
```

Hai unha serie de columnas que, alomenos de momento, non se van a usar: departure_id return_id

```
departure_latitude
departure_longitude
return_latitude
return_longitude
```

```
Quitamos con drop as columnas que non usaremos:
df.drop('nome columna', inplace = True, axis = 1)
```

Convertemo-lo tipo de dato object (string) a data/hora nas columnas que conteñen a data e hora de chegada e partida:

```
df['fecha_salida'] = df['fecha_salida'].astype('datetime64')
df['fecha_llegada'] = df['fecha_llegada'].astype('datetime64')
```

Info do dataset resultante:

```
RangeIndex: 12157458 entries, 0 to 12157457

Data columns (total 8 columns):

# Column Dtype
--- 0 fecha_salida datetime64[ns]

1 fecha_llegada datetime64[ns]

2 estacion_salida object

3 estacion_llegada object

4 distancia_m float64

5 duracion_sec float64

6 velocidad_media float64

7 temperatura float64

dtypes: datetime64[ns](2), float64(4), object(2)
```

Mostra do dataset resultante:

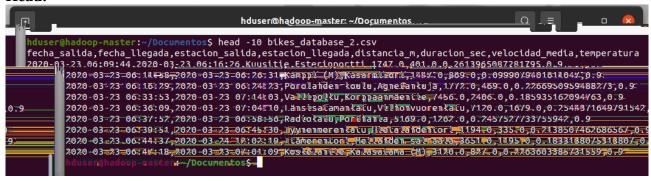
1 df								
	fecha_salida	fecha_llegada	estacion_salida	estacion_llegada	distancia_m	duracion_sec	velocidad_media	temperatura
0	2020-03-23 06:09:44	2020-03-23 06:16:26	Kuusitie	Esterinportti	1747.0	401.0	0.261397	0.9
1	2020-03-23 06:11:58	2020-03-23 06:26:31	Kamppi (M)	Kasarmitori	1447.0	869.0	0.099908	0.9
2	2020-03-23 06:16:29	2020-03-23 06:24:23	Porolahden koulu	Agnetankuja	1772.0	469.0	0.226695	0.9
3	2020-03-23 06:33:53	2020-03-23 07:14:03	Vallipolku	Korppaanmäentie	7456.0	2406.0	0.185935	0.9
4	2020-03-23 06:36:09	2020-03-23 07:04:10	Länsisatamankatu	Vilhonvuorenkatu	7120.0	1679.0	0.254437	0.9

12157453	2017-10-30 23:43:00	2017-10-30 23:55:00	Tyynenmerenkatu	Tyynenmerenkatu	918.0	714.0	0.077143	0.4
12157454	2017-10-30 23:49:00	2017-10-31 04:49:00	Brahen puistikko	Sörnäinen (M)	822.0	252.0	0.195714	0.4
12157455	2017-10-30 23:52:00	2017-10-31 00:02:00	Koskelantie	Intiankatu	1817.0	594.0	0.183535	0.4
12157456	2017-10-30 23:57:00	2017-10-31 00:00:00	Lastenlehto	Kamppi (M)	416.0	152.0	0.164211	0.4
12157457	2017-10-30 23:59:00	2017-10-31 00:12:00	Kaisaniemenpuisto	Ratapihantie	2856.0	771.0	0.222257	0.4
12157458 rd	ws × 8 columns							

Gárdase o dataset modificado con outro nome, en formato csv:

df.to_csv('bikes_database_2.csv' ,index= False)

Head:



Quitamo-los encabezados para traballar con só os datos en Hive: sed -i "1d" bikes_database_2.csv

```
hduser@hadoop-master:~/Documentos$ sed -i "1d" bikes_database_2.csv
hduser@hadoop-master:~/Documentos$ sed -i "1d" bikes_database_2.csv
2020-03-23 06:09:44,2020-03-23 06:16:26,Kuusitie,Esterinportti,1747.0,401.0,0.2613965087281795,0.9
2020-03-23 06:11:58,2020-03-23 06:26:31,Kamppi (M),Kasarmitori,1447.0,869.0,0.0999079401611047,0.9
2020-03-23 06:16:29,2020-03-23 06:24:23,Porolahden koulu,Agnetankuja,1772.0,469.0,0.2266950959488273,0.9
2020-03-23 06:33:53,2020-03-23 07:14:03,Vallipolku,Korppaanmäentie,7456.0,2406.0,0.185935162094763,0.9
2020-03-23 06:36:09,2020-03-23 07:04:10,Länsisatamankatu,Vilhonvuorenkatu,7120.0,1679.0,0.2544371649791542,0.9
2020-03-23 06:37:52,2020-03-23 06:58:56,Radiokatu,Porthania,5169.0,1262.0,0.2457527733755942,0.9
2020-03-23 06:39:51,2020-03-23 06:45:30,Tyynenmerenkatu,Hietalahdentori,1194.0,335.0,0.2138507462686567,0.9
2020-03-23 06:44:37,2020-03-24 12:02:19,Itämerentori,Meilahden sairaala,3651.0,1195.0,0.1833138075313807,0.9
2020-03-23 06:47:18,2020-03-23 07:01:09,Koskelantie,Kalasatama (M),3120.0,827.0,0.2263603385731559,0.9
2020-03-23 06:47:22,2020-03-23 06:56:22,Rautatieläisenkatu,A.I. Virtasen aukio,2070.0,535.0,0.2321495327102803,0.9
hduser@hadoop-master:~/bocumentos$
```

Para evitar problemas, cambiamo-la coma que fai de separadora de columnas polo tabulador:

Unha vez que o dataset xa está preparado e depurado, podemos realiza-las consultas.

Preguntas a resolver

- 1. Cantos rexistros correspondentes a viaxes contén 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 percorreuse en total entre 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 por viaxe
- 9. Cal é a velocidade media das viaxes
- 10. Cantas viaxes fixéronse cando a temperatura era menor de 0° C
- 11. Cantas viaxes fixéronse cando a temperatura estaba entre 15 e 25°C
- 12. Datos da primeira viaxe segundo data

Máis ideas:

- Preguntas de viaxes por datas (por meses, días da semana, por horas, etc)
- Preguntas de viaxes segundo temperaturas (duración de viaxes a temperaturas inferiores a 0°C, etc)

Pendente:

Join con dataset días choiva e preguntas de viaxes segundo días choiva https://en.ilmatieteenlaitos.fi/download-observations

Creación de estrutura e carga de datos

```
CREATE EXTERNAL TABLE IF NOT EXISTS viaxes (
data_ini TIMESTAMP,
data_fin TIMESTAMP,
saida STRING,
chegada STRING,
distancia FLOAT,
duracion FLOAT,
velocidade FLOAT,
temperatura FLOAT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE;

LOAD DATA INPATH '/user/hduser/input/bikes_2.csv' INTO TABLE viaxes;
```

Consultas Hive

Para que as saídas das consultas se vaian gardando no sistema hdfs, iniciamos cada consulta con: INSERT OVERWRITE DIRECTORY '/user/hduser/output/nome ficheiro'

Consulta 1: Cantos rexistros correspondentes a viaxes contén o dataset

INSERT OVERWRITE DIRECTORY '/user/hduser/output/hive_01_viaxes_count'
SELECT COUNT(*) FROM viaxes

Consulta 2: Cales son as 5 estacións con major número de saídas

INSERT OVERWRITE DIRECTORY 'output/hive_02_estacions_mais_saidas'
SELECT saida, COUNT(*) AS num_rex FROM viaxes GROUP BY saida ORDER BY
num_rex DESC LIMIT 5;

Consulta 3: Cales son as 5 estacións con maior número de chegadas

INSERT OVERWRITE DIRECTORY 'output/hive_03_estacions_mais_chegadas'
SELECT chegada, COUNT(*) AS num_rec FROM viaxes GROUP BY chegada ORDER BY
num_rex DESC LIMIT 5;

Consulta 4: Cal é a viaxe coa maior distancia percorrida

INSERT OVERWRITE DIRECTORY 'output/hive_04_max_distancia'
SELECT distancia FROM viaxes ORDER BY distancia DESC LIMIT 1;

Consulta 5: Canta distancia percorreuse en total entre tódalas viaxes

INSERT OVERWRITE DIRECTORY 'output/hive_05_total_distancia'
SELECT SUM(distancia) FROM viaxes;

Consulta 6: Cal é a distancia media percorrida por viaxe

INSERT OVERWRITE DIRECTORY 'output/hive_06_distancia_media'
SELECT AVG(distancia) FROM viaxes;

Consulta 7: Cal é a viaxe de major duración

INSERT OVERWRITE DIRECTORY 'output/hive_07_max_duracion'
SELECT duracion FROM viaxes ORDER BY duracion DESC LIMIT 1;

Consulta 8: Cal é a duración media por viaxe

INSERT OVERWRITE DIRECTORY 'output/hive_08_duracion_media'
SELECT AVG(duracion) FROM viaxes;

Consulta 9: Cal é a velocidade media da viaxes

INSERT OVERWRITE DIRECTORY 'output/hive_09_velocidade_media'
SELECT AVG(velocidad) FROM viaxes;

Consulta 10: Cantas viaxes se fixeron cando a temperatura era menor de 0°C

INSERT OVERWRITE DIRECTORY 'output/hive_10_viaxes_frio'
SELECT COUNT(*) FROM viaxes WHERE temperatura < 0;</pre>

Consulta 11: Cantas viaxes fixéronse cando a temperatura estaba entre 15 e 25 °C

INSERT OVERWRITE DIRECTORY 'output/hive_11_viaxes_temp_ok'
SELECT COUNT(*) FROM viaxes WHERE temperatura>=15 AND temperatura<26;</pre>

Consulta 12: Datos da primeira viaxe segundo data

INSERT OVERWRITE DIRECTORY 'output/hive_12_datos_primeira_viaxe'
SELECT * FROM viaxes ORDER BY data_ini LIMIT 1;

Resultado da execución das consultas Hive

Consulta 1: hive 01 viaxes count

12157458

Consulta 2: hive_02_estacions_mais_saidas

Itämerentori 330397 Töölönlahdenkatu 242555 Kamppi (M) 201560 Rautatientori / länsi 175358 Ympyrätalo 172776

Consulta 3: hive_03_estacions_mais_chegadas

Itämerentori 332453 Töölönlahdenkatu 243592 Kamppi (M) 195787 Rautatientori / länsi 178954 Ympyrätalo 177476

Consulta 4: hive_04_viaje_max_distancia

359383.34 (359,38 km)

Consulta 5: hive_05_total_distancia

2.5475649018170776 E9 (25.475.649. km)

Consulta 6: hive_06_distancia_media

2587.072079543791 (2,587 km)

consulta 7: hive_07_viaxe_duracion_maior

5012669.0 (58 días)

consulta 8: hive 08 duracion media viaxes

1060.442647813547 (17 min. 40 sec.)

consulta 09: hive_09_velocidade_media_viaxes

10.870777369952668 (10,87 km/h)

consulta 10: hive_10_viaxes_frio

3769

consulta 11: hive_11_viaxes_temp_ok

260036

consulta 12: hive 12 datos primeira viaxe