CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computing@CIT

A02 – REPORT ON ADDITIONAL TRANSPORTATION DATASET.

Both Assignment 1 and Assignment 2 have had as reference the public transportation dataset **Dublin Bus GPS sample data from Dublin City Council (Insight Project)**: https://data.gov.ie/dataset/dublin-bus-gps-sample-data-from-dublin-city-council-insight-project

However, there are many other publicly available transportation-related datasets out there (they can be focused on cars, buses, trains, taxis, bikes, scooters, etc).

Write a report of up to 1,000 words where you present and discuss:

- A publicly available transportation-related dataset (different from the Dublin Bus dataset).

Compare and contrast the new dataset w.r.t. the Dublin Bus dataset, including:

- A brief description of the dataset.
- Its main characteristics: URL for accessing to it, size, format of the data, etc.
- Its relevance - Include a potential use-case derived from the dataset you are proposing that cannot be achieved with the Dublin Bus Dataset.
- Its viability:
    - Do not implement the use-case, but briefly discuss in natural language (English and/or psudocode) the main steps that would be needed so as to implement it.
    - Include in the discussion the Spark libraries you will use in case you have had to implement it. Justify your selection.
    - Position the new use-case in terms of difficulty with respect to the other exercises proposed in Assignment 1 and Assignment 2.

Title : **Dublinbikes**

Link : https://data.smartdublin.ie/dataset/dublinbikes-api

Dublin Bikes is a bike sharing scheme in operation from bicycle docks and stations in Dublin City.

Following are the filed :

Column	Type	Description	Example
STATION ID (00)	numeric	Globally unique identifier of station.	1
TIME (01)	timestamp	Time of fetching the data.	01-08-2018 12:30
LAST UPDATED (02)	timestamp	Time of last updated information.	01-08-2018 12:26
NAME (03)	text	Station name.	CLARENDON ROW
BIKE STANDS (04)	numeric	Station total number of bike stands.	31
AVAILABLE BIKE STANDS (05)	numeric	Station available bike stands.	1
AVAILABLE BIKES (06)	numeric	Station available bikes	30
STATUS (07)	text	Station status (Open/Close).	Open
ADDRESS (08)	text	Station address.	Clarendon Row
LATITUDE (09)	numeric	Station latitude.	53.340927
LONGITUDE (10)	Numeric	Station longitude.	-6.262501

Each real-time dataset you will be dealing with contains a number of files file1.csv, file2.csv, file3.csv, file4.csv and represents n batches (of 1 file each) arriving over time for their real-time data analysis.

EXERCISE 1.

Let's assume you live in Dublin, and you have college at 10 am in morning. You live nearby the station CLARENDON ROW and you need to find out the available bike stands from the station 1 to station 69(college location).But you don not know the what are the available bike stands and it's a real time dataset

Evaluate the count of available bike stands from the location station 1 to the 69 within the 07 , 08 and 09 hour

< file1.csv, file2.csv, file3.csv, file4.csv >

Given the aforementioned dataset and program parameters

Station ID = 1

hours = ["07","08","09"]

Once again, please note the batch accumulated average delay:

- Results for Batch 1 contain the average delay for all measurements of < file1.csv >.
- Results for Batch 2 contain the average delay for all measurements of < file1.csv, file2.csv >.
- Results for Batch 3 contain the average delay for all measurements of < file2.csv, file3.csv, file3.csv>.

- Results for Batch 4 contain the average delay for all measurements of < file3.csv, file4.csv>

Implementation :

When we compared to streaming with Structure streaming : The Structure streaming is giving output fast and easy to implement because its same as like SQL queries.

- Filter the data by station ID 1 and the hours 07 ,08 and 09

- Use watermark function to group by the data

- Take the count of data of available bike stands

Output :

```
-------------------------------------------
Batch: 0
-------------------
+---+----+----------+
|day|hour|percentage|
+---+----+----------+
+---+----+----------+

-------------------------------------------
Batch: 1
-------------------
+---+----+----------+
|day|hour|Count|
+---+----+----------+
+---+----+----------+

-------------------------------------------
Batch: 2
-------------------
+---+----+----------+
|day|hour|count|
+---+----+----------+
+---+----+----------+

-------------------------------------------
Batch: 3
-------------------------------------------
+---+----+----------+
|day|hour|count|
+---+----+----------+
| 01|  07|     4|
| 01|  08|     3|
| 01|  08|     7|
| 01|  09|     2|
| 01|  09|     3|
| 01|  09|     1|
+---+----+----------+

-------------------------------------------
Batch: 4
-------------------------------------------
+---+----+----------+
|day|hour|count|
+---+----+----------+
| 02|  08|     2|
| 02|  08|     3|
| 02|  09|     3|
+---+----+----------+

-------------------------------------------
Batch: 5
-------------------------------------------
+---+----+----------+
|day|hour|count|
+---+----+----------+
| 03|  07|     3|
| 02|  07|     2|
| 03|  08|     2|
| 02|  08|     2|
| 03|  08|     1|
| 02|  09|     1|
+---+----+----------+

-------------------------------------------
Batch: 6
----------
+---+----+----------+
|day|hour|count|
+---+----+----------+
| 04|  07|     2|
| 03|  07|     4|
| 04|  08|     3|
| 04|  08|     2|
| 03|  09|     2|
| 03|  09|     1|
+---+----+----------+
```