CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Computing@CIT

A01 – REPORT.

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

- A novel exercise to be included in the data analysis of the Dublin Bus dataset.

There is no need to implement the new exercise, you just need to discuss it in terms of:

- Its originality - It has to be different from the 4 exercises proposed.
- Its relevance - Include a potential use-case derived from the exercise you are proposing.
- Its viability
  - Do not implement the exercise, but briefly discuss in natural language the main steps that would be needed so as to implement it.
  - Include in the discussion whether you will choose to implement it in Spark Core or Spark SQL, justying your selection.
  - Position the new exercise in terms of difficulty with respect to the four exercises proposed.

Let's assume you live in Dublin, Its Saturday morning. My friend called me and he invited me to his place for a birthday party. I was very excited to celebrate his birthday but I had to buy some groceries from the city centre .I didn't know where my bus stop was ,but I knew the bus line(40) which would take me to his place. I needed to find the bus stop and timing schedules for that particular day.

Bus_line: 40

Date : 9th January 2013

The bus line number is 33488 and the date is 9th January 2013. We need to find out the bus stop and at what time the bus will arrive.

Given a program passing by parameters:

- The bus vehicle "bus_line" (e.g., 40)

- The current date "current_time" (e.g., "2013-01-09 08:59:59" take the date "09")

Your task is to:

- Find out the **Closer_Stop and the timing the bus will arrive**

EXAMPLE - SMALL DATASET
• 2013-01-09 08:00:36,40,015B1002,0,-6.258078,53.339279,300,33488,279,1
• 2013-01-09 08:25:36,40,015B1002,0,-6.258078,53.339279,-200,33488,282,1
• 2013-01-09 08:36:36,40,015B1002,0,-6.258078,53.339279,100,33488,290,1
• 2013-01-09 08:55:36,40,015B1002,0,-6.258078,53.339279,300,33488,292,1
• 2013-01-09 09:15:36,40,015B1002,0,-6.258078,53.339279,-200,33488,298,1
• 2013-01-09 08:00:36,44,015B1002,0,-6.258078,53.339279,100,33488,279,1
• 2013-01-29 08:00:36,40,015B1002,0,-6.258078,53.339279,300,33488,279,1
• 2013-01-29 08:25:36,40,015B1002,0,-6.258078,53.339279,-200,33488,279,1
• 2013-01-29 08:50:36,40,015B1002,0,-6.258078,53.339279,100,33488,279,1
• 2013-01-09 09:00:36,40,015B1002,0,-6.258078,53.339279,100,33488,279,1
• 2013-01-09 09:25:36,40,015B1002,0,-6.258078,53.339279,-100,33488,279,1
• 2013-01-09 09:50:36,40,015B1002,0,-6.258078,53.339279,150,33488,279,1

SOLUTION EXAMPLE - SMALL DATASET

--- SPARK CORE ---

• solutionRDD:

< (40, [('2013-01-09 08:00:36, 279), ('2013-01-09 08:25:36, 282), ('2013-01-09 08:36:36, 290), ('2013-01-09 08:55:36, 292), ('2013-01-09 09:15:36, 298)]) >

• solutionRDD printed by the screen:

```
(40, [('2013-01-09 08:00:36, 279),
      ('2013-01-09 08:25:36, 282),
      ('2013-01-09 08:36:36, 290),
      ('2013-01-09 08:55:36, 292),
      ('2013-01-09 09:15:36, 298)
      ]
)
```

The above output gives you the bus line 40 will arrive on bus stop (279,282,290,292,298) at the different time.

Steps to compute this problem

- Filter the data on basis of bus line number(40) and date (09)
- groupby the data according to bus line and select the datetime and bus stop number

To implement this exercise, I would prefer to solve this on Spark SQL

- Spark SQL is used for structured data processing. Compare to spark core the Spark SQL provides spark with more information about the structure of the data and computation being performed.

- Spark SQL is used to execute SQL queries and using existing hive installation it can read the data.

- The result of this it will returned as a Dataset or Data Frame.

According to me If we can solve the 4th exercise then this exercise is not much difficult to solve.