German International University of Applied Sciences
Informatics and Computer Science
Dr. Nada Sharaf
Eng. Omaima Ahmed
Eng Aya Abdalla

Big Data & NoSQL Databases, Spring 2024
Assignment 1
Due date is March 14th, 2024 at 11:59 PM
Submitted in groups of maximum 2

For this assignment, you will be using a dataset that compiles information about NYC yellow taxi cab trips. You are required to implement the following functionalities and answer some questions using findings of queries you should create on your own. You can perform any necessary alterations to the data that may improve the usability of these queries or even make them possible. You should implement the task twice: once per NoSQL database (**MongoDB** and **Cassandra**).

The dataset is divided into 2 csv files:

- *taxitripdata.csv* contains the relevant trip data and pickup and drop-off "zones"
- *taxizonegeo.csv* contains the longitude & latitude coordinates of the pickup zone areas. These are all list-like locations of the polygon vertices.

You are required to do/answer the following :-

a) Remove the columns "store_and_fwd_flag", "rate_code" and "total_amount" from *taxitripdata*

b) Drop rows with missing essential details that would be required to fulfill the upcoming queries

c) Insert the data in the database as you see fit

d) Calculate the duration for each trip and add it as a new field in your database

e) Use "fare_amount", "extra", "mta_tax", "tip_amount", "tolls_amount" and "imp_surcharge" to calculate the total trip cost and add it as a new field in your database

f) What is the most common payment type used per time of day?
Hint: time of day meaning morning, afternoon or evening

g) What is the average tip amount per passenger count?

h) What are the best 5 locations for drivers to pick up passengers from?

Bonus:

- Is there a correlation between trip distance and the tip amount? (not to be done using the correlation calculation)
- Display the results of f, g and h using visualizations (will be graded based on the creativity and efficiency of the visualizations)

Column	***Type***	***Nullable***	***Description***
vendor_ id	text	No	A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc
pickup_ datetime	datetime	Yes	The date and time when the meter was engaged.
dropoff_ datetime	datetime	Yes	The date and time when the meter was disengaged.
passenger_ count	numeric	Yes	The number of passengers in the vehicle. This is a driver-entered value
trip_ distance	numeric	Yes	The elapsed trip distance in miles reported by the taximeter.
rate_ code	String	Yes	The final rate code in effect at the end of the trip. 1= Standard rate, 2=JFK, 3=Newark, 4=Nassau or Westchester, 5=Negotiated fare, 6=Group ride
store_ and_ fwd_ flag	String	Yes	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward", because the vehicle did not have a connection to the server. Y= store and forward trip N= not a store and forward trip
payment_ type	String	Yes	A numeric code signifying how the passenger paid for the trip. 1= Credit card, 2= Cash, 3= No charge, 4= Dispute, 5= Unknown, 6= Voided trip
fare_ amount	numeric	Yes	The time-and-distance fare calculated by the meter
extra	numeric	Yes	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
mta_ tax	numeric	Yes	$0.50 MTA tax that is automatically triggered based on the metered rate in use
tip_ amount	numeric	Yes	This field is automatically populated for credit card tips. Cash tips are not included
tolls_ amount	numeric	Yes	Total amount of all tolls paid in trip.
imp_ surcharge	numeric	Yes	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
total_ amount	numeric	Yes	The total amount charged to passengers. Does not include cash tips
pickuplocationid	String	Yes	TLC Taxi Zone in which the taximeter was engaged
dropofflocationid	String	Yes	TLC Taxi Zone in which the taximeter was disengaged

Deliverables

- Your code is to be submitted to **bigdata602.s24@gmail.com** as a zip file containing your notebook (include the names and IDs of the team members in the body of the email with **subject:** "Assignment 1 S24").

PLAGIARISM IS NOT TOLERATED AND COPIED WORK WILL BE AWARDED 0 POINTS FOR BOTH TEAMS INVOLVED or IF YOU COPIED IT FROM THE INTERNET OR ELSEWHERE (NO. EXCEPTIONS.)! **There will be an individual evaluation for each team**

Note: you will be asked for the reasoning of any actions, queries or decisions you have taken in the implementation of this task during the evaluations that have led to your answers/results.