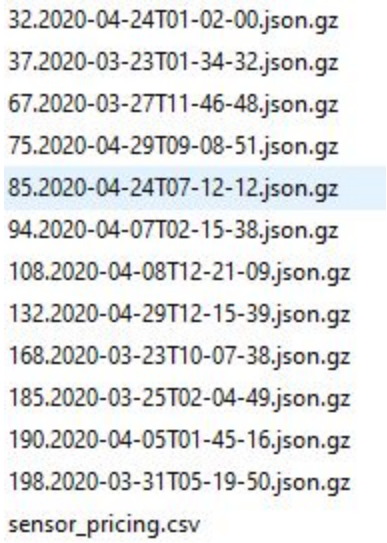
IDC - Cloud Computing Course
Exercise 2
Due Date July 24rd 2020

Submission -

1. This is a very short exercise and you have one month to submit. There will be no extensions.
2. Make sure that you do not publish your AWS access keys. Submission with AWS access key will be graded zero.
3. You will place a zip file in S3 with public access. And fill the following form with the URL and student details. https://forms.gle/kh92DbHy6cr9yLfu7 - The zip file will include all the deliverables required below**.**

The data for this task is here:
https://drive.google.com/file/d/1Wp0JbWx5kVPVboa-xLuCXLymGoYVfz-a/view?usp=sharing

The zip file contains the following files:



These files represent data taken from various cameras on toll roads. Each camera record the following data points on each car:

- Sensor - what sensor took this reading
- LicensePlate - the car's license plate
- Time - the time of the event

The sensor_pricing.csv contains pricing data for each sensor.

Using AWS Athena, you need to implement charges for each one of the cars that is captured in this dataset on a **monthly basis**.

The logic for the computation goes like this:

```
for file in files:
    for reading in readings:
        price = sensors[reading.Sensor].Pricing
        licenses_plates[reading.LicensePlate] += price
```

In other words, for each reading for a license plate, you need to accumulate the price that is configured for the relevant sensor.

For example, the license plate YUYRB78292 shows up 4 times in 32.2020-04-24T01-02-00.json.gz, on sensor 32. The charge for the sensor is 3.24, so the total cost for YUYRB78292 will be 12.96 based on this file alone.

Note that for the total data set, YUYRB78292 should be charged 29.09 for March and 13.76 for April.

The results of this task should be a report with the following fields:

- License Plate
- Total Cost
- Month
- Year
- Number of tolls

Deliverables:

- You need to show your Athena creation scripts, all queries used and the final report.
- Explain your reasoning for the schema and queries you run.
- Provide an estimate for the costs of such a system assuming we have 15,000 sensors and are reading an average of 50,000 cars a day.
- What would be the storage costs? And what would be the compute costs?
- Suggest alternatives to this approach to gather the report and explain advantages and disadvantages over the Athena approach.