Objective

To build an ETL Pipeline to process the metrics in the “metrics.csv” and generate KPI’s for reporting purposes.

Pre-processing

We query the advertiser API everyday at 6:00AM and download the campaign metrics for the previous day. The campaign metrics are downloaded as a wide record format as shown below

client_id	campaign_id	metric_date	metric_1	metric_2	metric_3	metric_4	...

The ETL process should read the file, and convert it into a columnar format, wherein each metric is transformed from a column in wide format to a row in columnar format. You should also generate additional columns as described below. Save the columnar records in a “metrics_columnar.csv” file

client_id	campaign_id	metric_date	metric_name	metric_value	customer_id	client_name	campaign_name

customer_id: The customers.csv provides a mapping between client_id and customer_id
client_name: The clients.csv provides a mapping between client_id and client_name
campaign_name: The campaigns.csv provides a mapping between campaign_id and campaign_name

KPI

Generate the below defined Daily KPIs from the “metrics_columnar.csv” for each client

[redacted]

KPI	Formula
Daily Net Spend	[cost_micros] * 10.0E-7
Daily Gross Spend	[Daily Net Spend] / (1.0 - 0.26)
imp000	impressions/1000.0
CTR	clicks/impressions
eCPM	[Daily Net Spend]/[imp000]

Save the KPI's in the below defined columnar format as a Parquet file partitioned on customer_id

customer_id	client_id	date	kpi	value
Partition column				

Logging, Error Handling and Validation

Please provide your thoughts on logging, error handling and validations that you will need to incorporate

Performance

Please provide your thoughts on different ways of improving the speed and performance of the ETL process

[redacted]