Tapcart Data Engineer Challenge

Tapcart analytics systems are primarily event based. Often times these events need to be enriched and aggregated to provider useful insights to our merchants. Our current datastack is GCP and built in Java. In order to provide these insights we've developed a complex ETL process using a unified data processing model (Apache Beam)

In this evaluation, we are testing the following:

- Ability to setup a basic java project using maven as a dependency manager
- Knowledge of the Beam programming model
- Familiarity with data ingestion from different sources
- Data transformations/enrichment involved multiple sources
- Ability to perform simple aggregations
- Ability to perform multistep compound aggregations
- Knowledge of IO steps of data pipelines

Deliverables:

Create a data pipeline using Apache Beam and Java. The inputs are data files provided to use as CSVs. One CSV will contain a sample dataset of events. The second CSV will be a sample dataset of devices.

Data Models: ` ` ` Event { sessionId: string deviceId: string timestamp: timestamp type: emun(ADDED_TO_CART | APP_OPENED) total_price: 50.00 }

Device { id: string user: string } ` ` `

Enrichment Tasks

- Enrich the event objects with the user property. The user property can be found in the device objects

```
EnrichedEvent {
    sessionId: string
    deviceId: string
    timestamp: timestamp
    type: emun(ADDED_TO_CART | APP_OPENED)
  total_price: 50.00
  user: string
}
```

Calculation Tasks

- Count of APP_OPENED
- Sum of ADDED_TO_CART
- Average Session Length (Every event has a unique sessionId and every sessionId has multiple events associated with it)

IO Tasks

- Write enriched events to an output file (should contain equal amount of lines as the input file)
- Write the calculations in an output file (should contain 3 values)