Objective : Build a Real Time Analytics Platform for E-commerce data

We have three data sources ,

- CRM Database
- Transactional Database
- Click Stream data

Assumptions

Transactional Data and Click stream data are both high traffic and volume.
The CRM database does not change in real time and is slow changing data.

Data Ingestion Strategy

One approach for a real time / near real time data ingestion would be to ingest data using **pub sub** model. This model provides the data at the destination with the least data latency. Create **Data flow** pipelines to subscribe to these topics and ingest data into BQ.
You can see that we have exposed two services to ingest data into the publisher, the third party clients may write their data to these services and abstract the changes as event based solution.

Data flow **automatically scales up or down** resources and can process data in parallel based on workload and does not involve manual intervention. Utilize dataflow to perform real time data cleansing, transformation and enrichment and ingestion to BQ. Dataflow offers per second billing for resources that the job uses.

BigQuery : BQ makes a good choice for this design as the use case is to utilise this for analytics. It can handle both batch and streaming effectively.We are not choosing BIg Table as it does not support sql queries or joins. The application must be able to ingest data in real time and respond with relevant analytical information straight away.

Cloud Storage

Cloud storage may be used to store raw data where required and is very cost effective solution for storing large data for short duration.

Data Access :

Big Query Studio may be used by users who wish to query the data.
IAM profiles with the right roles may be used to class various access for different persona and protect sensitive data

Looker Studio may be used to create reporting dashboards and visualisation of data