**Request 1 solution -**

**Key Considerations -**

1. Event data which is being generated with a varying rate requires an ingestion service which can scale accordingly.
2. Medium complexity of transformation required with minimum latency

**Assumptions -**

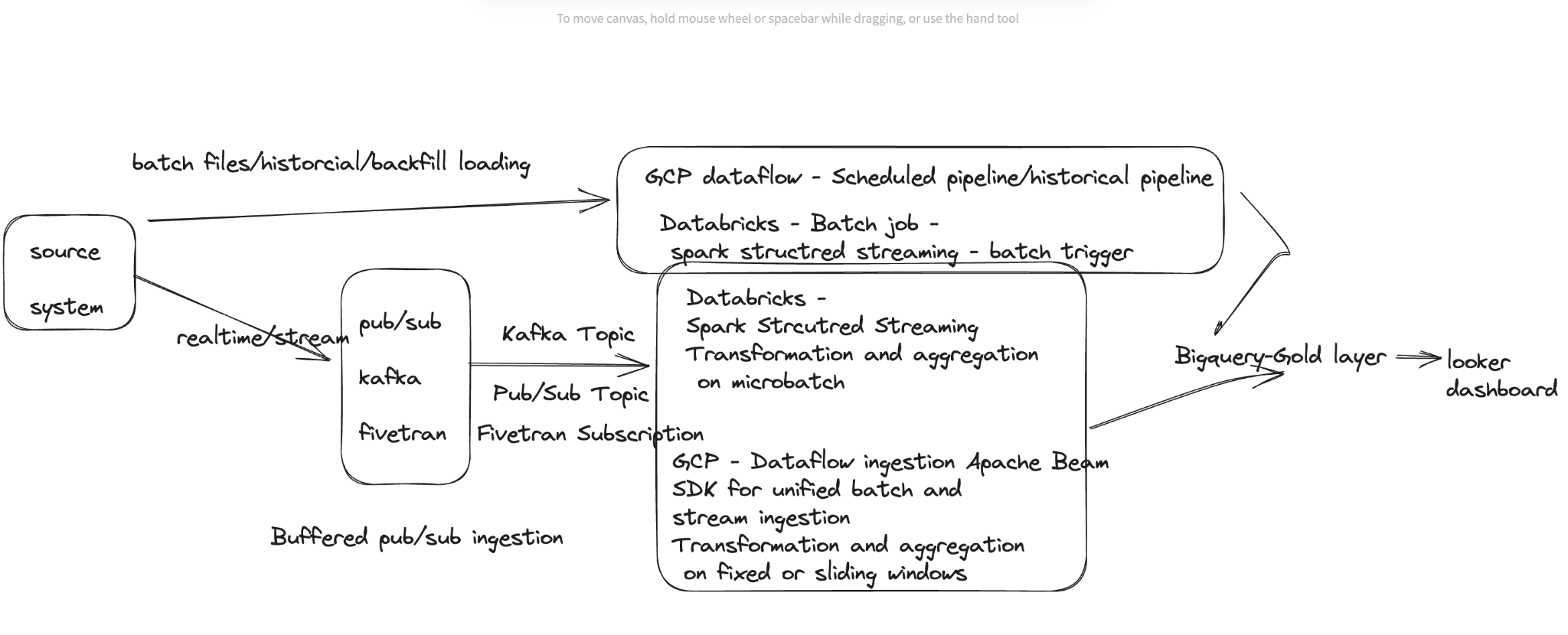
1. Source system is capable of publishing real time Event data to pub/sub based message queues/topics like kafka and GCP pub sub
2. Source system is capable of producing historical or backfill batch files.
3. Microbatch approach with some latency (processing time )is acceptable as a solution
4. Data Governance framework is in place to ensure data cataloging , lineage, quality , auditing and lifecycle management.

**Logical approach -**

**Approach 1 -** Unified model for batch and stream ingestion using Spark Structured Streaming on Databricks - leveraging features like autoloader, intelligent file sensors, delta live table and implementing a deltalake and loading data into Cloud Datawarehouse Bigquery

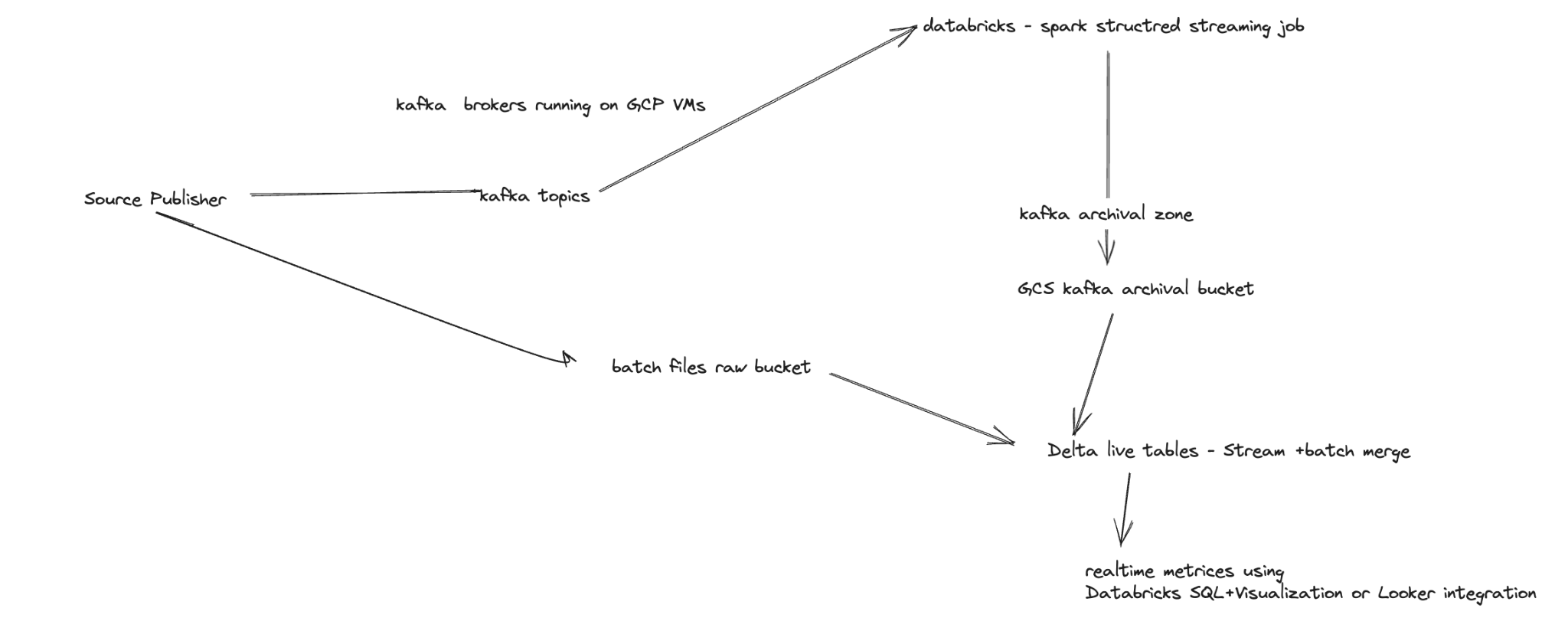
**Approach 2 -** Unified model using Dataflow -Apache Beam SDK - loading into Bigquery.

**Both models are capable of generating microbatch based statistics.**

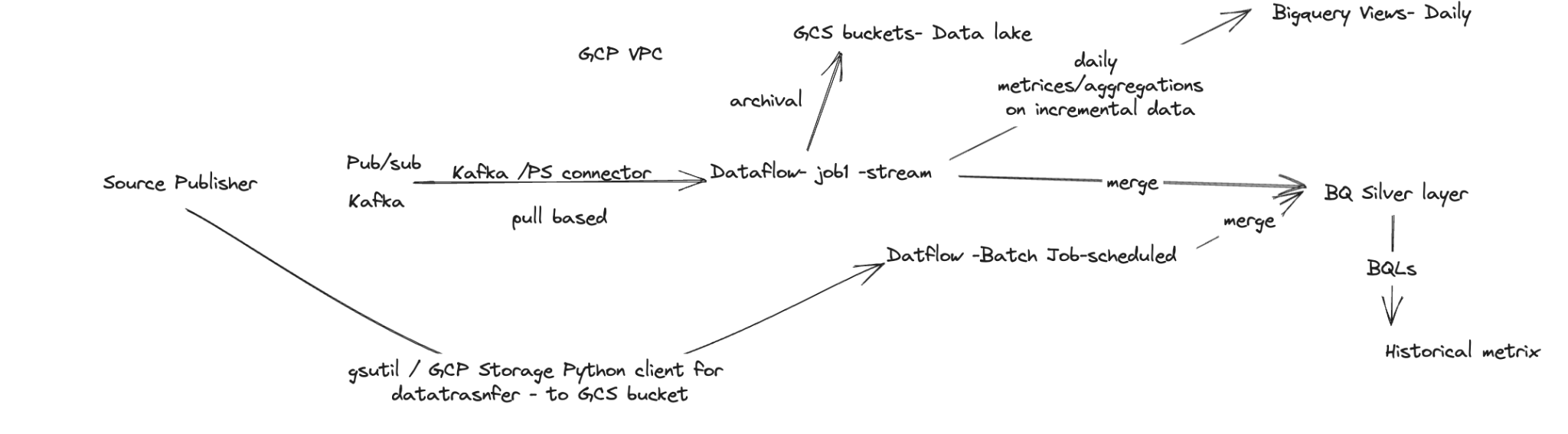
****

**Physical architecture -**

**Approach 1- Deltalake - raw, trusted, refined layers**

****

**Approach 2- Pub/Sub , Dataflow, Bigquery**

****

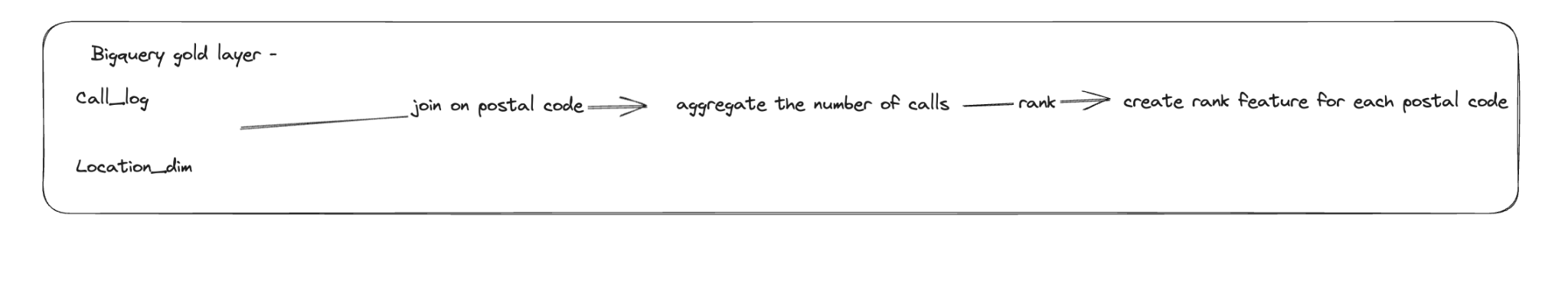
**Tools /Choices-**

| **Ingestion- realtime** | **Kafka, Confluent, Pub/Sub** |
| --- | --- |
| **Ingestion - batch** | **Cloud Storage SDK, gsutil, Fivetran, Storage Transfer Service** |
| **Transformation/ Real time aggregation** | **Dataflow, Databricks - Delta live tables** |
| **Orchestration** | **Airflow, Composer, Databricks- Scheduler, Nifi** |
| **Schema registry** | **Confluent, Liquibase** |
| **Cloud Data Warehouse** | **Bigquery - Supports wide range of analytical functions + Machine learning Capabilities** |
| **Real time visualization** | **Looker, Tableau** |
| **Alerting** | **GCP Alerts** |
| **Cloud Data Lake** | **GCS bucket** |
| **Devops** | **Gitlab, Google Source Repos, Cloud Build , Container Registry, Terraform or Ansible** |

**Real time recommendation engine-**

**Example formula-**

A simple recommendation can be created based on the ranking of the areas based on demand.The higher number of calls the higher demand -



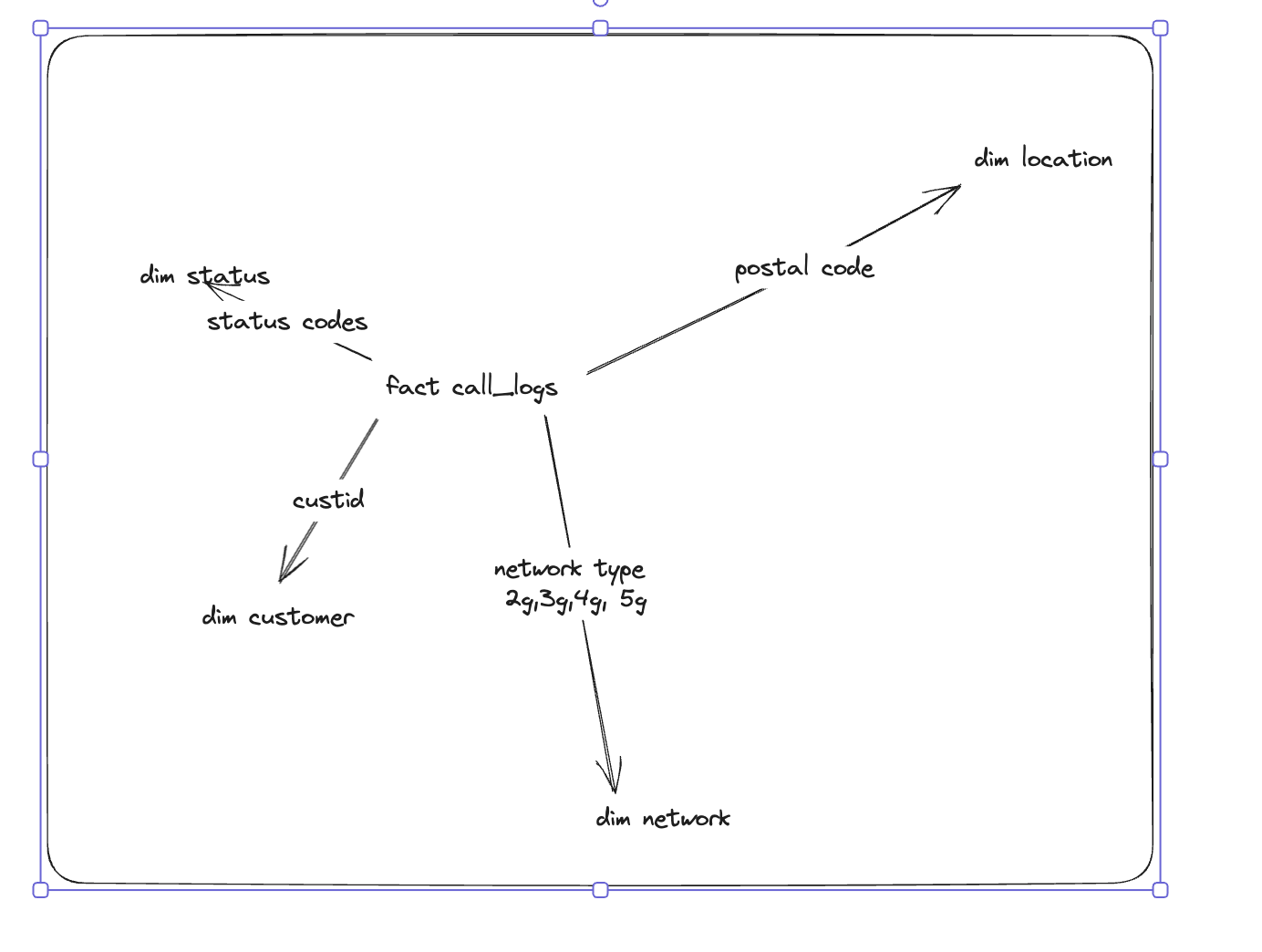
**Real Time alerting for network outages-**

Spark Structured Streaming job

spark streaming job -> time based sliding windows based time feature in call\_log-> moving average of ASU(Signal strength feature)->

trigger every 5 minutes to GCP logging -> setup alert if ASU less than a threshold value -> trigger alert using GCP logging

**Analytical Data Models -**

****