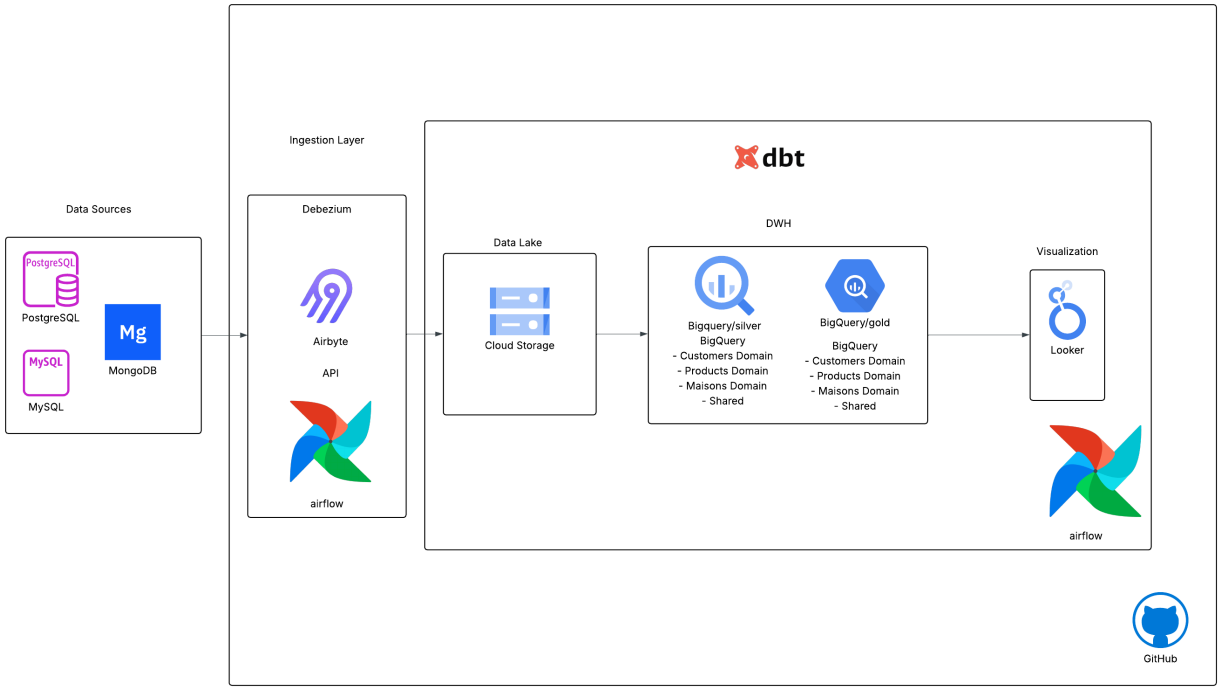
Arquitectura Analítica – Data Mesh en GCP



1. Objetivo

Diseñar una plataforma de datos en Google Cloud que integre múltiples fuentes, sirva analítica (BI y ML) y escale por dominios siguiendo un enfoque Data Mesh.

2. Principios

- Arquitectura Data Mesh con dominios autónomos.
- Servicios cloud-native sobre GCP.
- Uso de tecnologías open-source.
- Automatización con CI/CD.
- Gobernanza federada.

3. Arquitectura General

Flujo principal:

```
Fuentes → Ingesta →Data Lake →Data Warehouse → BI / ML
```

Implementación:

```
Postgres / MySQL / MongoDB / SAP / APIs
            ↓
Airflow + Debezium + APIs
            ↓
CloudStorage(Bronze)
            ↓
BigQuery (Silver / Gold)
            ↓
Looker / Vertex AI
```

---

4. Ingesta

Herramientas:

- **Airflow** como orquestador general.
- **Debezium** para CDC desde bases de datos transaccionales.
- **Pub/Sub** como bus de eventos.
- **Cloud Functions** para consumo de APIs REST.

Resultado:

Todos los datos aterrizan en **Cloud Storage** en formato crudo.

---

5. Data Lake (Bronze)

Tecnología: Google Cloud Storage.

Rol: capa de persistencia inicial de todos los datos.

Estructura lógica:

```
gs://data-lake/bronze/customers/
gs://data-lake/bronze/products/
gs://data-lake/bronze/maisons/
```

Características:

- Datos sin transformar.
- Particionados por fecha.
- Formatos: Parquet / JSON.

---

6. Data Warehouse

Tecnología: BigQuery.

Modelo: un dataset por dominio.

```
customers_domain
products_domain
maisons_domain
shared
```

Capas:

- **Silver:** limpieza, tipado y normalización.
- **Gold:** métricas y modelos de negocio.

---

7. Transformaciones

- **dbt** para transformaciones SQL y validaciones.
- **Spark (Dataproc)** para transformaciones complejas.

Orquestación completa con **Airflow**.

---

8. Data Mesh

Cada dominio es responsable de su ciclo completo de datos:

Dominio	Producto
Customers	customer_360
Products	product_metrics
Maisons	brand_nps

Los datos comunes se publican en el dataset `shared`.

---

9. BI

Herramientas:

- Looker Studio / Looker.
- Alternativa open-source: Lightdash.

Dashboards por dominio y dashboards ejecutivos cross-domain.

---

10. Machine Learning

Stack:

- Vertex AI.
- MLflow.
- Feature Store sobre BigQuery.

Pipeline:

```
BigQuery → Feature Store → Vertex AI → Predicciones → BigQuer
y
```

---

11. Gobernanza

- Dataplex para catálogo y lineage.
- IAM por dominio.
- Column-level security en BigQuery.
- Auditoría con Cloud Logging.

---

12. GitOps / DataOps

Repositorios:

```
data-infra      (Terraform)
data-pipelines  (dbt + Airflow)
ml-models       (ML)
```

Pipeline estándar:

```
PR →Tests →Review →Deploy
```