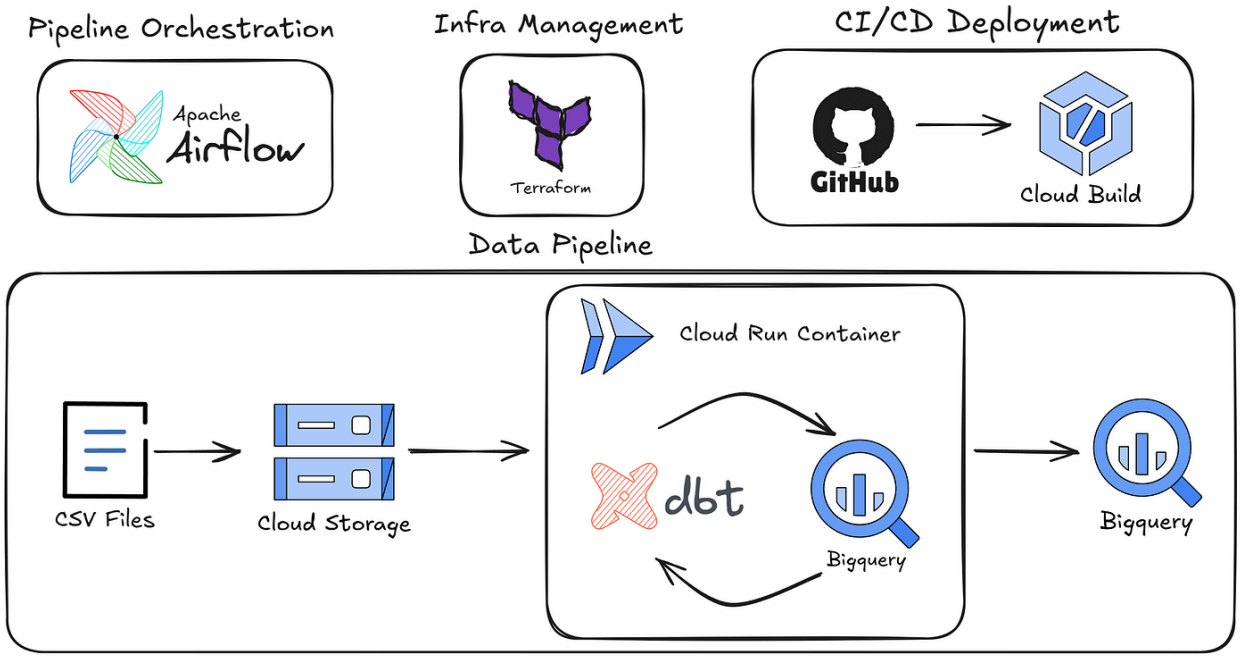
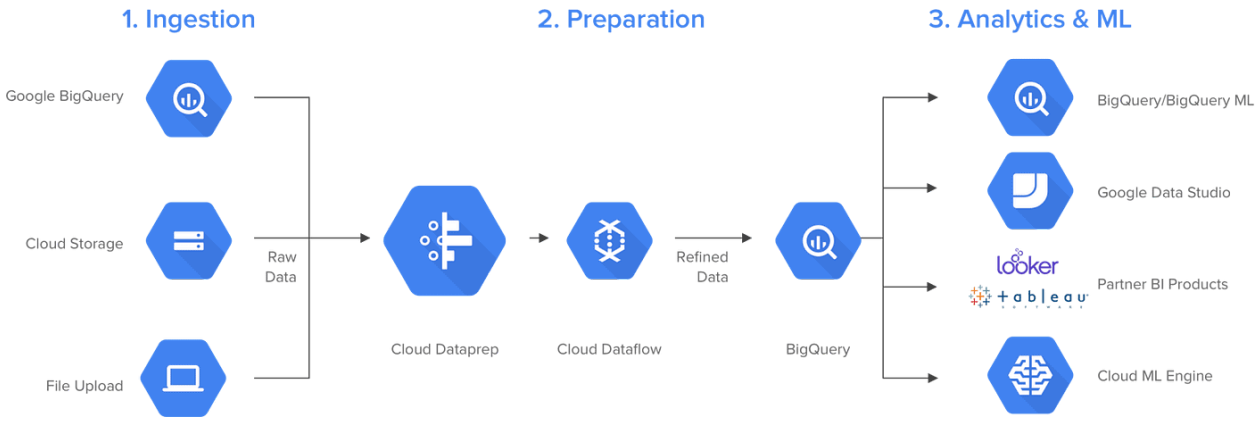
1. Project context
2. High-level flow
3. Detailed step-by-step tasks (end to end)
4. What gets validated at each step

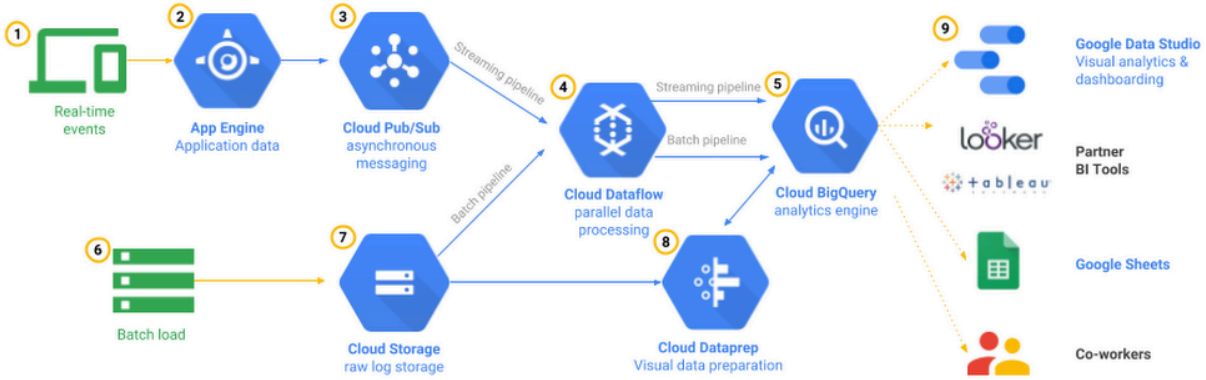
🧩 PROJECT CONTEXT (REFERENCE USE CASE)

Project: Analytics Platform using BigQuery
Goal: Ingest raw data → clean & model → analytics-ready → BI/ML
Core Service: BigQuery
Supporting Services: GCS, Cloud Run/Dataflow, Looker

🧱 HIGH-LEVEL FLOW








```
Source Data
    ↓
Cloud Storage (Raw)
    ↓
BigQuery (Staging)
    ↓
BigQuery (Curated)
    ↓
BigQuery (Analytics)
    ↓
Looker / ML / APIs
```

✅ STEP-BY-STEP BIGQUERY PROJECT TASK BREAKDOWN

🔵 PHASE 1 – PROJECT & ENVIRONMENT SETUP

Step 1: Create GCP Project

- Create a dedicated analytics project
- Enable billing
- Define environment (dev / prod if needed)

✅ Output: GCP project ready

Step 2: Enable Required APIs

Enable:

- BigQuery
- Cloud Storage
- IAM & Logging

✅ Output: Services available

Step 3: IAM & Access Setup

- Create service accounts
- Assign roles:
    - BigQuery Admin / Data Editor
    - Storage Object Viewer
- Grant BI users read-only access

✅ Output: Secure access model

🔵 PHASE 2 – RAW DATA INGESTION (BRONZE)

Step 4: Create Cloud Storage Buckets

```
gs://project-raw-data/
   ├── transactions/
   ├── customers/
   └── reference/
```

- Raw = immutable
- No transformation here

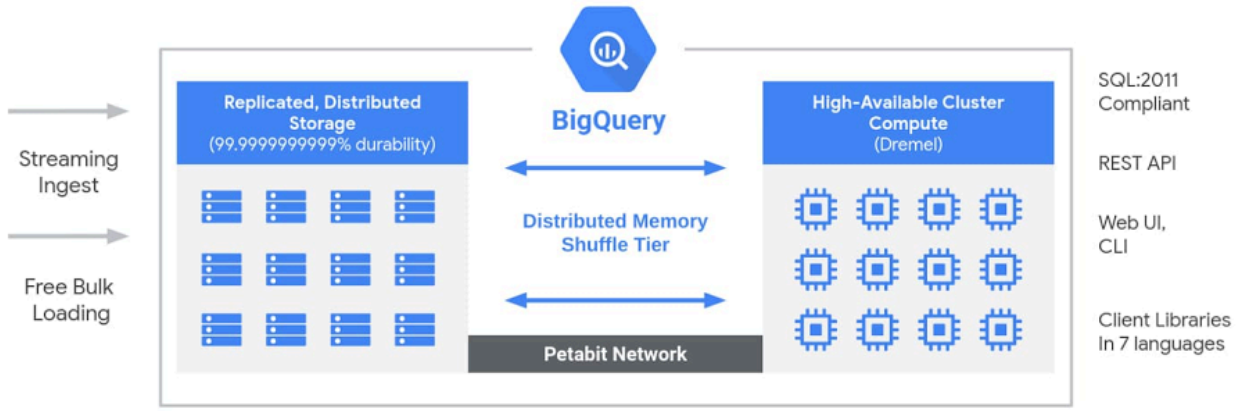
✅ Output: Raw landing zone

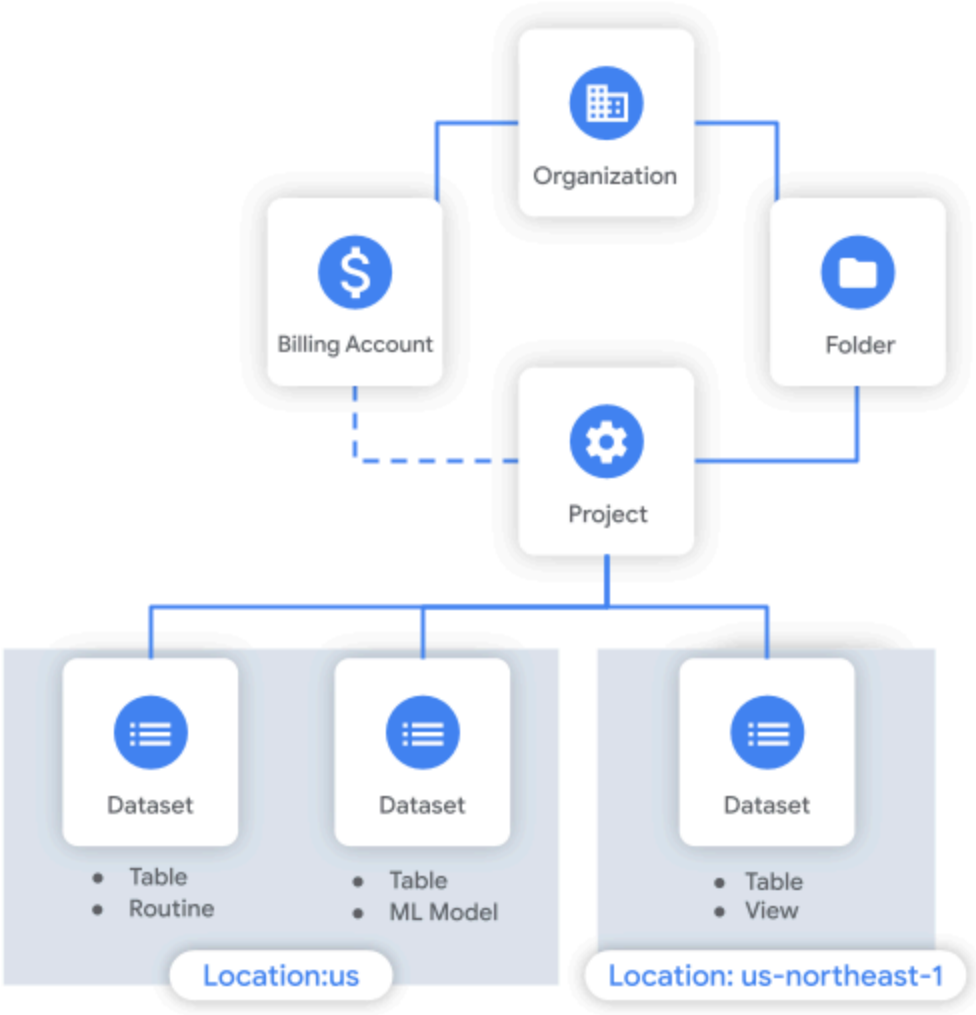
Step 5: Upload Raw Data

- CSV / JSON / Parquet files
- Preserve original schema & timestamps

✅ Output: Source data available

🔵 PHASE 3 – BIGQUERY STAGING LAYER





Step 6: Create BigQuery Datasets

```
bq_raw
bq_staging
bq_curated
bq_analytics
```

✅ Output: Dataset structure created

Step 7: Load Raw Data into BigQuery (Staging)

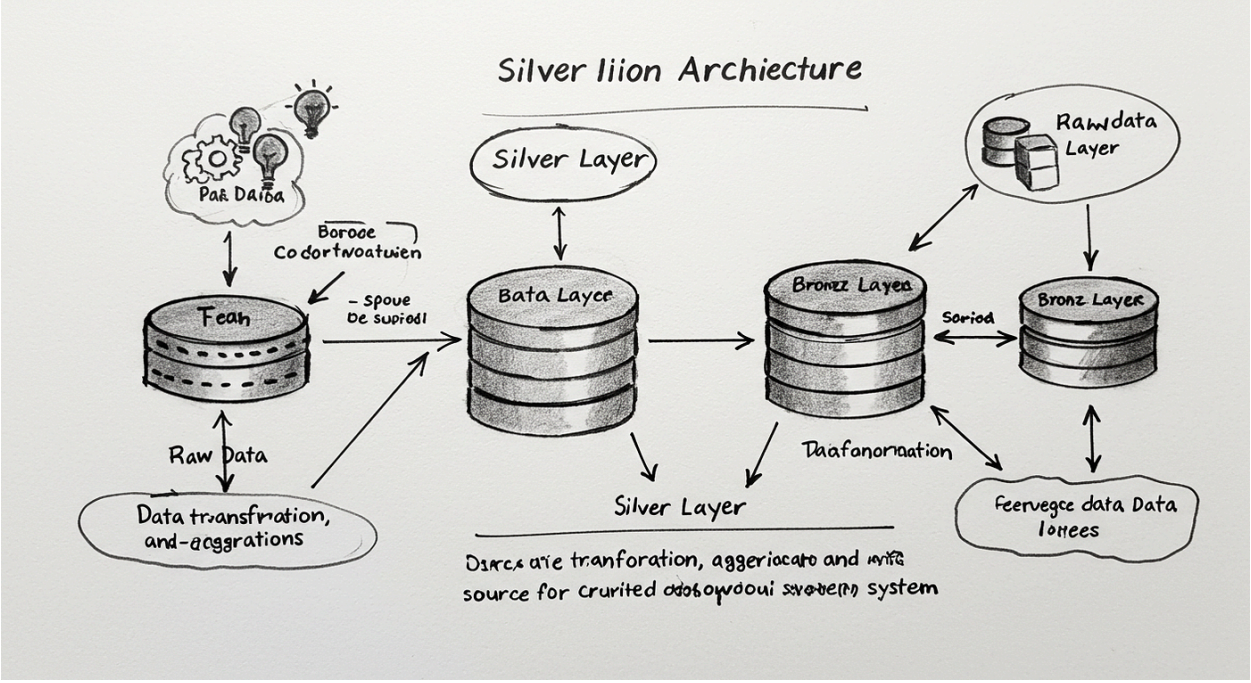
Options:

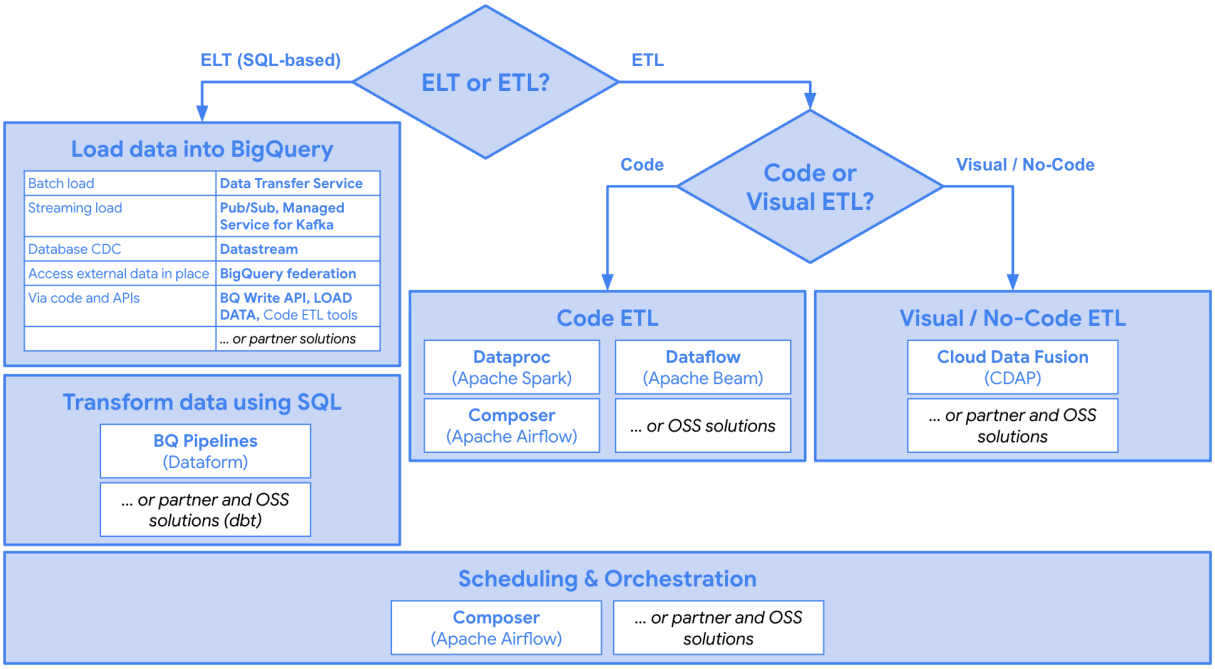
- Batch load from GCS
- External tables (optional)

```
LOAD DATA INTO `bq_raw.transactions`
FROM FILES (
  format = 'CSV',
  uris = ['gs://project-raw-data/transactions/*.csv']
);
```

✅ Output: Raw tables in BigQuery

🔵 PHASE 4 – DATA CLEANING & STANDARDIZATION (SILVER)





Step 8: Data Quality Checks

- Null validation
- Duplicate detection
- Data type validation

```
SELECT COUNT(*)
FROM bq_raw.transactions
WHERE amount IS NULL;
```

✅ Output: Data issues identified

Step 9: Clean & Standardize Data

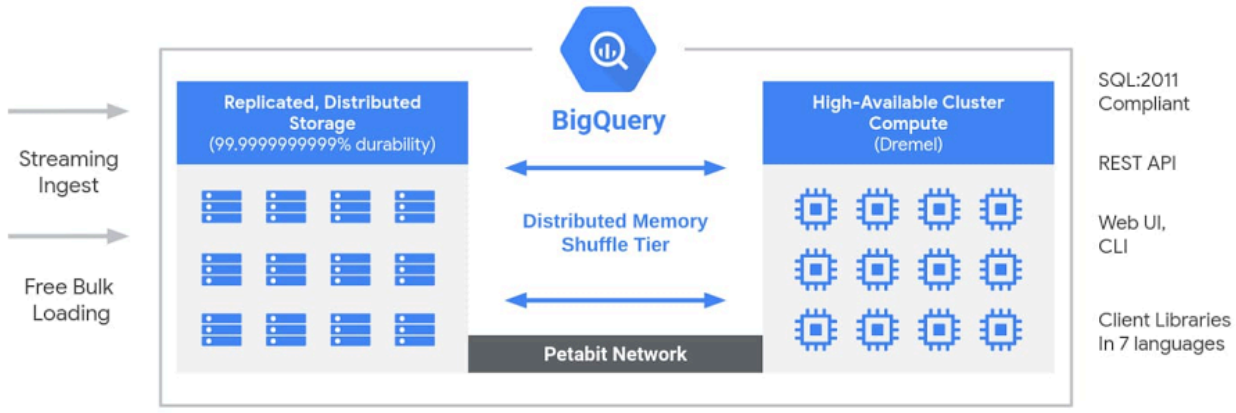
- Fix date formats
- Normalize categories
- Remove duplicates

```
CREATE OR REPLACE TABLE bq_curated.transactions AS
SELECT DISTINCT
  SAFE_CAST(transaction_date AS DATE) AS transaction_date,
```

```
  category,
  amount
FROM bq_raw.transactions
WHERE amount IS NOT NULL;
```

✅ Output: Clean curated tables

🔵 PHASE 5 – DATA MODELING (GOLD)



Step 10: Design Analytics Schema

- Fact tables
- Dimension tables

Example:

- fact_transactions
- dim_date
- dim_category

✅ Output: Logical data model

Step 11: Create Analytics Tables

```
CREATE TABLE bq_analytics.fact_transactions
PARTITION BY transaction_date
```

```
CLUSTER BY category AS
SELECT * FROM bq_curated.transactions;
```

✅ Output: Performance-optimized tables

🔵 PHASE 6 – PERFORMANCE OPTIMIZATION

Step 12: Apply Partitioning

- Date-based partitioning
- Reduces scanned data

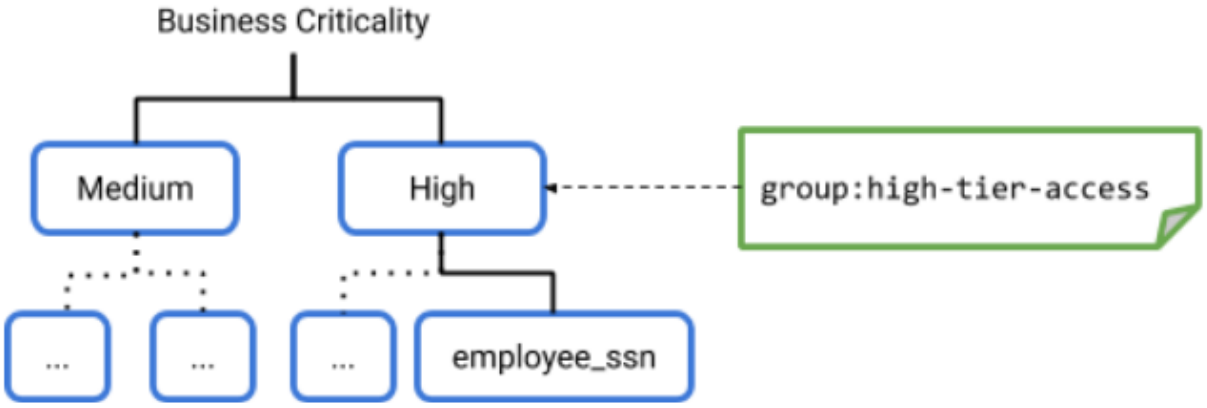
✅ Output: Lower query cost

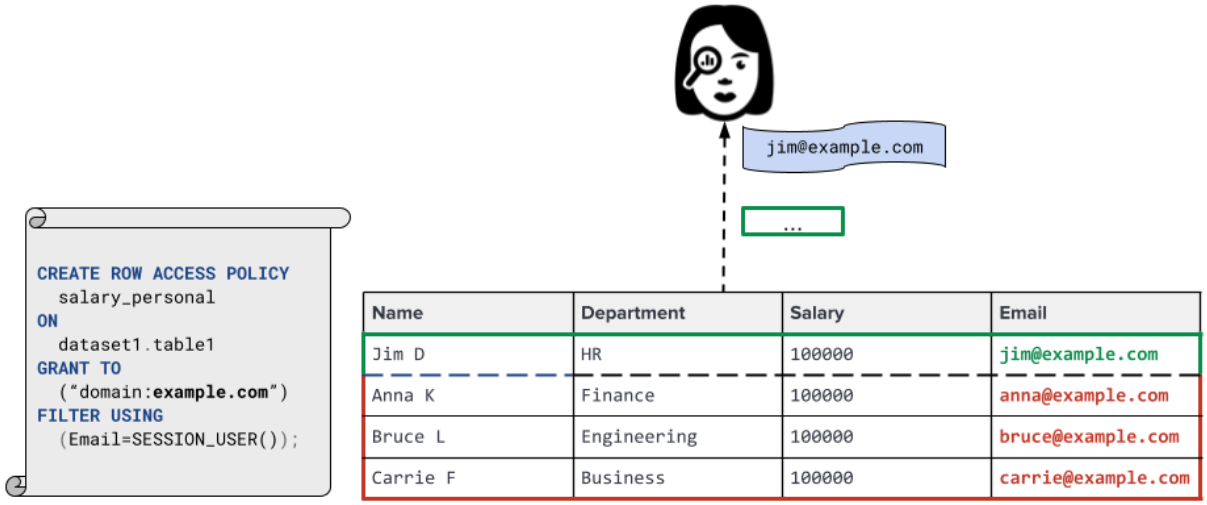
Step 13: Apply Clustering

- High-filter columns
- Improves aggregation speed

✅ Output: Faster analytics

🔵 PHASE 7 – SECURITY & GOVERNANCE





Step 14: Dataset & Table IAM

- Analyst: read-only
- Engineers: edit
- Service accounts: controlled access

Step 15: Column & Row Security

- Policy Tags for PII
- Row Access Policies for data segregation

```
CREATE ROW ACCESS POLICY region_filter
ON bq_analytics.fact_transactions
GRANT TO ('group:us_analysts')
FILTER USING (region = 'US');
```

✅ Output: Compliance-ready setup

🔵 PHASE 8 – AUTOMATION

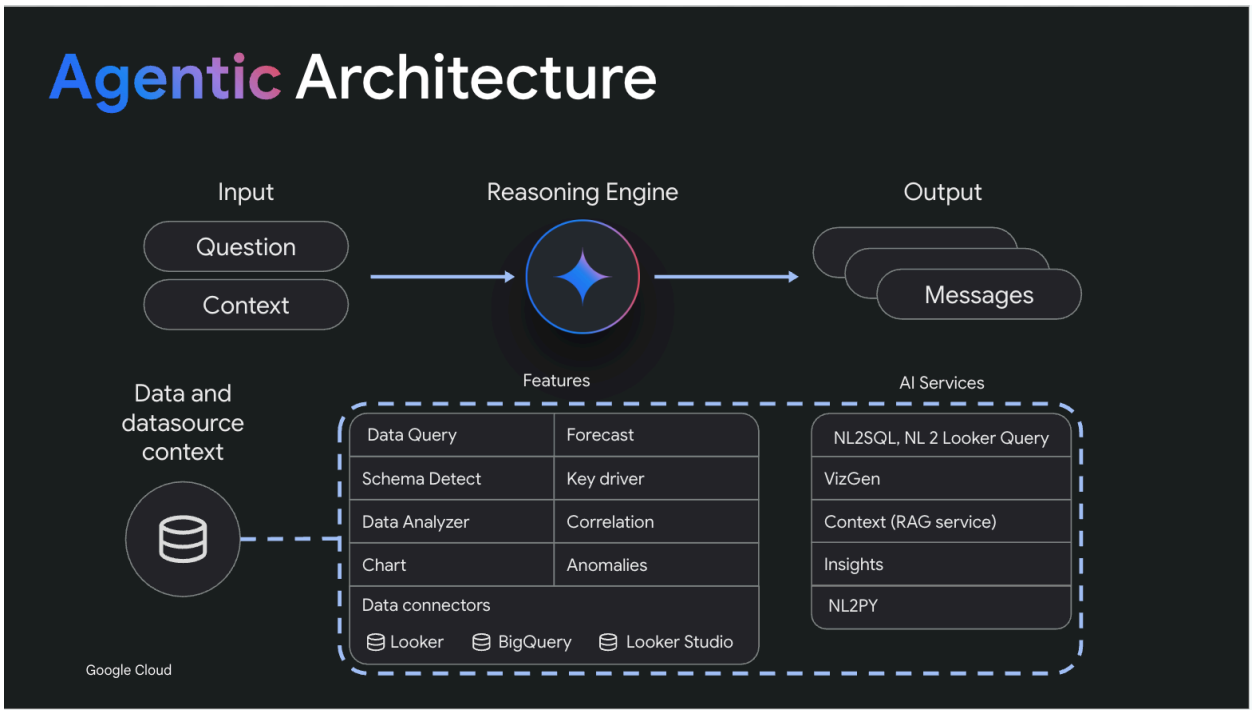
Step 16: Schedule Data Loads

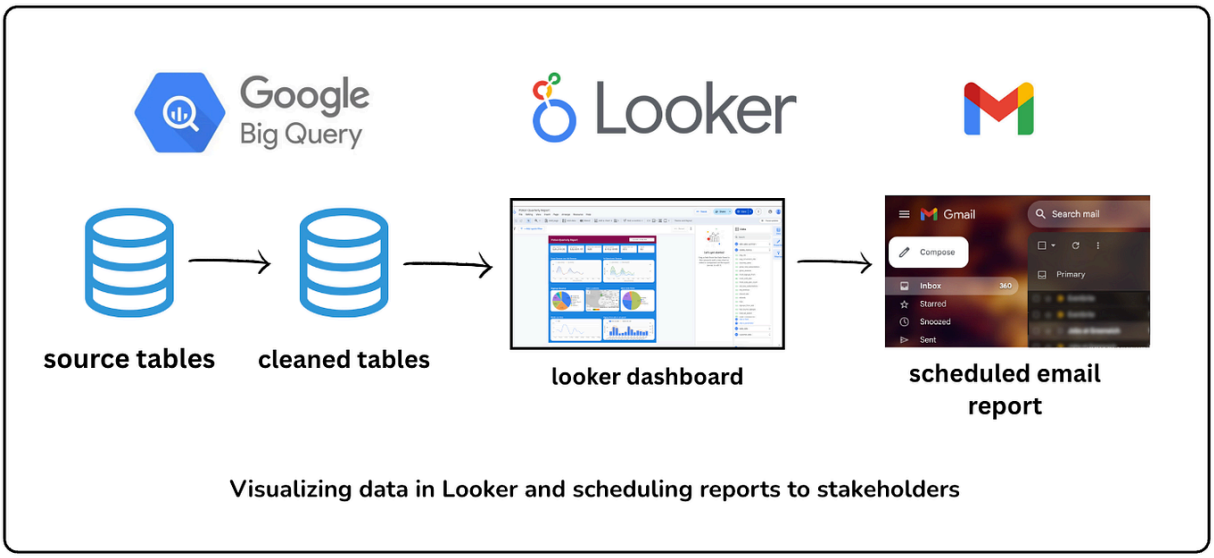
- Cloud Scheduler / Composer

- Daily or hourly refresh

✅ Output: Fully automated pipeline

🔵 PHASE 9 – ANALYTICS & BI





Step 17: Connect BigQuery to BI

- Looker / Looker Studio
- Define semantic metrics

Step 18: Build Dashboards

- Trends
- KPIs
- Drill-downs

✅ Output: Business-ready insights

🔵 PHASE 10 – VALIDATION & HANDOVER

Step 19: Validate Results

- Data reconciliation
- Query performance
- Cost monitoring

Step 20: Documentation & Demo

- Architecture diagram
- SQL samples
- Cost explanation
- Interview walkthrough

🏁 FINAL OUTPUTS

✔ End-to-end BigQuery analytics platform
✔ Optimized, secure, scalable
✔ Interview & production ready
