BigQuery Dataset Sync — corp\_sec\_sem ➜ rto

# 📘 Project Overview

Document Title: BigQuery Dataset Sync — corp\_sec\_sem.table ➜ rto.table  
Prepared By: [Your Name]  
Date: 2025-06-26  
Environment: GCP BigQuery – Internal Data Sync  
Purpose: To maintain a consistent, up-to-date copy of historical or semantic-layer data from corp\_sec\_sem to rto for downstream consumption and reporting.

# 1. 🔍 Objective

To regularly sync data from:  
  
- Source: project\_id.corp\_sec\_sem.source\_table  
- Target: project\_id.rto.target\_table  
  
Both tables share identical schema and are used for analysis, reporting, and backups.

# 2. 🎯 Use Case and Business Need

This sync supports:  
  
- Consistent analytics across environments  
- Isolated access for reporting teams  
- Ability to run transformations without impacting raw data

# 3. ✅ Selected Sync Strategy

Strategy Chosen: MERGE (Upsert)  
Reason: Ensures that updated rows are refreshed and new rows are inserted without duplicating or overwriting the entire table.

# 4. 🧠 SQL Logic

MERGE `project\_id.rto.target\_table` T  
USING `project\_id.corp\_sec\_sem.source\_table` S  
ON T.primary\_key = S.primary\_key  
WHEN MATCHED THEN  
 UPDATE SET T.col1 = S.col1, T.col2 = S.col2, T.last\_updated = S.last\_updated  
WHEN NOT MATCHED THEN  
 INSERT (primary\_key, col1, col2, last\_updated)  
 VALUES (S.primary\_key, S.col1, S.col2, S.last\_updated);

# 5. ⏱ Automation & Scheduling

- Tool: Cloud Composer (Airflow)  
- Schedule: Daily at 1:00 AM IST  
- DAG Name: sync\_corp\_to\_rto\_table  
- Task: BigQueryInsertJobOperator  
- Retries: 2 retries on failure (5 min delay)

# 6. 🔍 Data Validation Logic

Row Count Match:  
SELECT COUNT(\*) FROM `project\_id.corp\_sec\_sem.source\_table`;  
SELECT COUNT(\*) FROM `project\_id.rto.target\_table`;  
  
Mismatch Check:  
SELECT \* FROM `project\_id.corp\_sec\_sem.source\_table`  
EXCEPT DISTINCT  
SELECT \* FROM `project\_id.rto.target\_table`;

# 7. 📈 Monitoring

- Logging via Stackdriver  
- Email alerts to: dataops@example.com  
- Auto-alert on: DAG failure, row count mismatch

# 8. 📂 Version Control

- SQL and DAGs stored in GitHub repo: github.com/org/data-engineering-bq-sync  
- Changes tracked using PRs with mandatory peer review

# 9. 💬 Notes

- Table is partitioned by event\_date for efficiency  
- Only the last 60 days of data are updated during each sync  
- MERGE chosen over INSERT to avoid duplication

# ✅ Summary

|  |  |
| --- | --- |
| Item | Value |
| Sync Type | Incremental (Upsert via MERGE) |
| Tool Used | Cloud Composer |
| Schedule | Daily at 1 AM IST |
| Source Table | corp\_sec\_sem.source\_table |
| Target Table | rto.target\_table |
| Schema Match | ✅ Yes |
| Partitioned? | ✅ Yes (event\_date) |
| Monitoring & Alerts | Stackdriver + Email on Failures |