Errors, Failure

Overview & Scope

System: Boston311 ingestion

Pipelines:

- boston311_daily (incremental, MERGE into service_requests_2025)
- boston311_weekly (full snapshot, guarded overwrite of service_requests_2025)
- boston311_build_filtered_tables (rebuilds chatbot and dashboard gold tables from service_requests_2025)

Storage & Compute: CKAN (Boston open data) is downloaded as JSON and stored in GCS as buckets temporarily and finally stored in BigQuery (Boston 311 dataset, US location)

Downstream:

- Chatbot: boston311.chatbot (query-focused subset)
- Dashboard: boston311.dashboard (aggregated metrics by date, dept, type, neighborhood)

Error Audit

1. CKAN API throttling / network timeouts (fetch_data._do_request)

Error Type	
What it's about	CKAN endpoint intermittently returns 429/timeouts paging via _id stalls and no file is produced for the run.
Error type	Background, System
User stakes	High
Error Signals	
Signals	Multiple retry log lines reduced pages fetched zero-row JSONL missing GCS object for the date

	daily DAG finishes early with ShortCircuit skip.
Error Resolution	
Error rationale	Dashboard/chatbot show yesterday's data operators see "stuck" runs.
Solution type	Retry/backoff tuning and graceful degradation in product
Resolution	Add jitter to backoff, optionally reduce PAGE_SIZE under failure on zero-row ingestion, show a "freshness lag" banner and have the chatbot cite last successful _ingested_at alert via email with a successful baseline comparison.

2. Duplicate/late updates per case_enquiry_id

Error Type	
What it's about	Same case_enquiry_id arrives multiple times across pages/runs MERGE picks the older row if _ingested_at parity.
Error type	Background
User stakes	High
Error Signals	
Signals	Inconsistencies vs. CKAN portal MERGE updates that don't change values trend line "bounce."
Error Resolution	
Error rationale	Users notice mismatches between products/sources
Solution type	Freshness precedence and change detection
Resolution	Always stamp _ingested_at require S._ingested_at > T._ingested_at for updates add change-hash to skip no-op updates expose "case last updated at" and dataset "as of" times in UI.

3. BigQuery job failures: quota/slots or location mismatch

Error Type	
What it's about	Jobs fail due to slot pressure or non-US location settings
Error type	System

User stakes	High
Error Signals	
Signals	Quota exceeded," "resources exceeded," or location error messages prolonged job queues.
Error Resolution	
Error rationale	Data fails to update on expected cadence
Solution type	Capacity planning and serialization
Resolution	Centralize BQ_LOCATION="US" serialize heavy jobs auto-reschedule weekly batch consider reservations/flex slots email on backpressure with ETA for retry window.

4. Weekly overwrite could wipe target if _dedup is empty

Error Type	
What it's about	TRUNCATE then INSERT with an empty _dedup temp table leads to empty gold.
Error type	Background
User stakes	High
Error Signals	
Signals	Row count drops to 0 post-weekly anomaly alerts trigger.
Error Resolution	
Error rationale	All data disappeared.
Solution type	Safety guard and precondition check
Resolution	Add guard to only TRUNCATE+INSERT if _dedup row_count > 0 else skip and alert canary query with expected thresholds before the destructive step.

5. Daily MERGE vs weekly OVERWRITE contention

Error Type

What it's about	Overlapping runs cause transient metric flicker and inconsistent counts.
Error type	Background, System
User stakes	Medium
Error Signals	
Signals	Two success emails close together dashboard counts oscillate between runs.
Error Resolution	
Error rationale	Users see numbers change within minutes without explanation.
Solution type	Orchestration and user messaging
Resolution	Serialize runs (ExternalTaskSensor/trigger rules) pause daily while weekly executes show brief "data updating" banner consider consolidating into a parent orchestration DAG.

Quality assurance

1. Freshness of data

Goal	Keep data < 24h old for daily runs and < 7d for weekly snapshots.
Method	BigQuery scheduled query reads max _ingested_at from service_requests_2025 Airflow pushes metric The dashboard badge shows "Data as of <timestamp>." Email if lag > threshold.
Review Frequency	Daily

2. Ingest volume delta

Goal	Detect unexpected spikes/dips in new rows.
Method	BigQuery scheduled query computes rows/day and 7-day moving average with z-score Looker Studio chart + email if $\|z\| \geq 3$.

Review Frequency	Daily

3. Job health

Goal	Ensure DAGs complete reliably within SLO
Method	Airflow metrics (task duration, retries, status)—weekly report Alert on repeated retries or >P95 runtime.
Review Frequency	Daily glance Weekly deep dive.