Project 2 — Financial Transaction Monitoring

Overnight Batch Data Pipeline (Design-Only)

1. Scenario Overview & Scope

This project designs an overnight batch pipeline to detect anomalous payment activity across **Canada, India, and the UK**.
Each night, ~**2 TB** of logs arrive from regional payment processors.
The pipeline aggregates and flags possible **anti-money-laundering (AML)** events while meeting privacy, regulatory, and cost constraints.

- **Batch window:** 22:00 → 06:00 local time (8 h)
- **SLA:** p95 completion ≤ 6 h
- **Cost ceiling:** ≤ $500 / month (edu tier)
- **Scope:** Design-only — no code or running system required.

2. Data Sources & Jurisdictions

Region	Law / Regulator	Approx Volume	Residency Note
Canada	PIPEDA	0.8 TB	ca-central-1
India	DPDP Act 2023	0.7 TB	ap-south-1
UK	DPA 2018 / UK GDPR	0.5 TB	eu-west-2

Cross-border replication blocked; only aggregates shared to HQ.

3. Baseline → Improved Architecture

Baseline: manual ETL on EC2 → timeouts and $600/mo cost.
Improved: AWS Batch + Glue + Athena + CloudWatch.

```
flowchart LR
  A["Regional Processors"] --> B["S3 Regional Buckets"]
  B --> C["Glue Crawler to Data Catalog"]
  C --> D["Athena Transform Queries"]
  D --> E["S3 Aggregate Store"]
  E --> F["Lambda Notifier to Regulator API"]
  E --> G["CloudWatch Metrics and Logs"]

  subgraph Guardrails
    H["Encryption at Rest (KMS)"]
    I["Replication Off"]
    J["TTL <= 90 days (raw), 180 days (agg)"]
```

```
  K["Audit Role Logging"]
end

%% connect to nodes inside the subgraph (not to the subgraph label)
G -. audit .-> H
G -. alerts .-> K
```

Throughput Feasibility Calculation
2 TB ÷ 6 h = ≈ 92 MB/s aggregate throughput.
Athena + Glue combined throughput > 200 MB/s per query in AWS benchmarks,
so 2 TB fits comfortably within the 6 h SLA even with 2× retries.

Trade-offs

Decision Point	Option A	Option B	Chosen	Rationale / Trade-off
Batch Compute	**AWS Glue** (serverless ETL)	**Amazon EMR** (managed Spark cluster)	✅ Glue	Glue auto-scales for short nightly jobs ≤ 2 TB, reducing ops overhead. EMR gives more control but costs ≈ 2× more and requires cluster lifecycle management.
Query Engine / Transform	**Athena** (pay-per-query)	**Redshift** (provisioned warehouse)	✅ Athena	The workload runs once nightly, so Athena's on-demand billing avoids idle compute costs. Redshift is only justified for > 10 TB continuous workloads.
Storage Tier	**S3 Standard + Lifecycle → Glacier**	—	✅ S3 Lifecycle	Raw logs need low-latency access for 6 h; older data automatically moves to Glacier, cutting storage cost ≈ 40 %.
Orchestration	**Step Functions** (serverless state machine)	**Airflow on ECS**	✅ Step Functions	Simpler to maintain; built-in retries + kill hooks. Airflow adds infra cost and complexity for a once-per-day job.
Monitoring & Alerts	**CloudWatch + SNS**	Third-party (Datadog, New Relic)	✅ CloudWatch	Native AWS integration and zero extra cost; sufficient for p95 SLA + cost alarms. External tools add ≈ $50 / mo.

4. Failure & Retry Playbook

- Auto-retry 2× on transient S3/Glue errors (≤ 2 h delay).
- Kill-switch: Lambda stops jobs > 8 h or >$15/day.
- Recovery: next run reprocesses ± 1 day.

5. Metrics / SLA / Cost Plan

Metric	Target	Test Method
p95 completion	≤ 6 h	Synthetic clock probe
Cost / run	≤ $15	Budget Alert
Detection precision gain	+10 %	Offline sample
Data retention	≤ 90 d raw	Lifecycle unit test

Monthly Cost Estimate

Service	Usage	$/unit	$/mo
Glue Job	60 h	0.44	26
Athena Queries	4 TB scan	5	20
S3 Storage (raw + agg)	3 TB-mo	0.023	69
CloudWatch Logs	20 GB	0.60	12
Lambda Notifications	0.5 M calls	0.20	1
Total			**≈ $128 / mo**

Sensitivity Analysis: +25 % storage growth → +$17 / mo (≈ $145 total); still < $500 limit.

6. Privacy & Ethics Excerpt (PIA Link)

- **Purpose:** AML pattern detection without profiling legitimate users.
- **Fields:** txn_amount, merchant_id, region, timestamp, hashed payer_id.
- **Minimization:** No names / account numbers.
- **Retention:** raw ≤ 90 d; agg ≤ 180 d; secure delete via S3 lifecycle.
- **Access:** analyst RO role; audited 1 y.

Telemetry Matrix

Signal	Value	Invasive?	Effort	Keep?
txn_amount	High	Low	Low	✅
merchant_id	Medium	Low	Low	✅
payer_id (hash)	Low	Medium	Medium	✅
IP address	High	High	Low	❌

Recourse & Remedy: Deletion requests via bank portal → auto-ticket → Data Steward manual verify + delete.
Full PIA available at pia_financial.md

7. Empty Chair Perspectives

A. Unbanked Customer (India) — “Why is my data analyzed without consent?” → `Test test_opt_in_enforced()`
B. Small Merchant (UK) — “Logs expose my sales volume.” → `Test mask_logs ledger entry`

8. Clause → Control → Test Bundle Summary

Jurisdiction	Clause	Control	Test (Harm)
CA	PIPEDA s.4.5	S3 TTL ≤ 90 d	Over-retention
IN	DPDP § Consent	Opt-in join filter	Unlawful processing
UK	DPA § Anon	Salted hash rotation	Re-id risk ≥ 1 %

Red-bar pytest files in `/tests/ethics/`; ledger links owners + actions.

9. Risks & Ledger Integration

ID	Risk	Harm	Owner	Status
R1	Over-retention	Privacy breach	Data Platform	🔴
R2	Consent missing	Legal risk	Pipeline Eng	🔴
R3	Merchant logs unmasked	Competitive harm	Obs Team	🟡

10. Ethical Operations Badge (+5%)

This submission includes one red-bar → green-bar transition:
India — DPDP Act §7 (Consent & Lawful Purpose) protecting unbanked / opt-out users.
The full CI output, diff, and updated ledger are available in Ethical_Operations_Bonus.md.