Extraction Service – Observability & Ops Runbook

Creator **Michael Clark** Created **Aug 24, 2025, 22:58** Last updated **Aug 24, 2025, 23:09**

Observability

Metrics（CloudWatch）

- **Throughput**
  - `docsProcessed`, `recordsEmitted`, `recordsDLQ`
  - `throughputDocsPerSec`, `throughputRecordsPerSec`
- **Latency**
  - `latencyP50`, `latencyP95`, `latencyP99` per document
- **Reliability**
  - `errors`, `retries`, `jobSuccessRate`
  - `sqsQueueDepth`, `sqsMessageAge`
- **Cost**
  - `costEstimatePerJob`, `$PerPage`, `$PerTenant`
- **Quotas**
  - `jobsRunningPerTenant`, `docsPerJob`, `tenantQuotaUtilization`

Logs（CloudWatch Logs）

- **Structured JSON logs** with correlation IDs: `{tenantId, jobId, partId, eventType, timestamp}`
- Key events: `JobCreated`, `JobStarted`, `RecordProcessed`, `RecordFailed`, `JobCompleted`, `BudgetExceeded`
- **Sensitive data sanitized**; PII fields flagged or redacted.

Traces（X-Ray）

- **End-to-end traces** spanning:
  - API Gateway → Lambda → Step Functions → SQS → Worker → LLM → S3
- Each trace tagged with `tenantId`, `jobId`, and `partId`
- Trace maps used to identify bottlenecks (LLM latency vs. S3 I/O vs. parsing).

Dashboards（CloudWatch/Grafana）

- **SLA heatmaps**: doc latency p95 per tenant.
- **Throughput graphs**: docs processed/hour, shard completion.
- **Cost dashboards**: spend per tenant, spend per connector.
- **Error views**: DLQ rate, retry rate, job failures.
- **Queue health**: SQS backlog, message age, worker concurrency.

Ops Runbook

Service Level Objectives (SLOs)

- **Throughput:** 100k docs < 24h.
- **Latency:** p50 < 2s, p95 < 10s per doc.
- **Reliability:** ≥ 99.5% jobs succeed (after retries).
- **Cost:** <$0.01 per page (LLM phase).

Alerts (CloudWatch Alarms → SNS/EventBridge → Pager/Slack)

- **p95 latency > 10s** (sustained 5 min) → alert.
- **Job failure rate > 1%** per tenant over last hour.
- **SQS backlog > 50k messages** or **queue age > 10 min**.
- **DLQ rate > 0.5%** of records.
- **Cost anomaly > +30% day-over-day** or job exceeding budgetCap.
- **Error spikes** (5x above baseline).

Common Incident Scenarios & Playbooks

1. **High latency (p95 breach)**
    - Check X-Ray for LLM vs. S3 bottlenecks.
    - Scale concurrency up/down in Step Functions.
    - If LLM provider throttling → switch fallback provider.
2. **SQS backlog grows**
    - Inspect worker concurrency; scale up Lambda reserved concurrency.
    - Check for poison messages; DLQ drain if necessary.
3. **High error or DLQ rate**
    - Sample DLQ records to find schema/config issues.
    - Verify extraction hints vs. schema.
    - If parsing bug → hotfix worker logic.
4. **BudgetExceeded job cancellation**
    - Confirm cost metrics vs. tenant config.
    - Notify tenant; partial results remain retrievable.
5. **Auth/tenant isolation failure**
    - CloudTrail + AuditLog inspection.
    - Rotate API keys, enforce IAM scoping.