Extraction Service - Implementation

Creator **Michael Clark** Created **Aug 24, 2025, 23:01** Last updated **Aug 24, 2025, 23:09**

Implementation Plan

a. Milestones (MVP → Beta → GA)

MVP (Sept–Oct 2025)

- Local filesystem connector.
- Job API (submit config, track status, retrieve results).
- Configurable extraction schema + optional hints.
- Batch processing (≤100k docs) with retries + idempotency.
- JSONL output (partitioned shards) + Results API.
- Core observability (metrics/logs/traces).
- Security baseline: TLS, API keys, tenant isolation.
- **Cutline:** Demo run on 10k docs processed end-to-end under SLA.

Beta (Nov–Dec 2025)

- S3 connector (LocalStack first, AWS S3 in prod).
- Cost pre-checks, quotas, and fail-fast enforcement.
- Cost metrics exposed in API/dashboards.
- AuditLog & JobMetrics fully populated.
- Step Functions + SQS scaling validated at 50–100k docs.
- **Cutline:** 100k docs job finishes <24h with p95 <10s/doc.

GA (Q1 2026)

- SharePoint connector.
- Multi-tenant RBAC + Tenant API.
- Immutable audit logs (WORM) + PII handling (masking, retention, right-to-delete).
- Production dashboards/alerts for SLO monitoring.
- Compliance reviewed.
- **Cutline:** External tenants onboarded, 100k docs job runs stable under SLA + compliance pass.

Risk Register (Top 5 Risks & Mitigations)

1. **LLM cost overruns**
   - Mitigation: Pre-flight estimator, mid-run budget caps, per-tenant quotas.
2. **SQS backlog / throughput bottleneck**
   - Mitigation: Concurrency auto-scaling, backpressure controls, partition by tenant.
3. **Extraction accuracy <95%**
   - Mitigation: Rule-based fallbacks, test sets, config-based hints, DLQ replay.
4. **Multi-tenant data leakage**
   - Mitigation: Strong IAM policies, tenantId tagging, isolation tests, CloudTrail audits.
5. **Connector fragility (SharePoint/Graph API limits)**
   - Mitigation: Token refresh handling, retry with backoff, circuit breakers, contract testing.

Effort Estimate & Roles

- **API & Orchestration** – build Job API, Results API, Step Functions orchestration (@Backend Engineer).
- **Connectors** – filesystem, S3, SharePoint adapters (@Integration Engineer).
- **Extraction Engine** – rule-based parsing, LLM integration, schema/hints (@ML Engineer).
- **Data Model & Storage** – DynamoDB schemas, JobResults/JobMetrics, audit logging (@Data Engineer).
- **Observability** – metrics/logs/traces wiring, dashboards, alerts (@SRE/DevOps).
- **Security & Compliance** – IAM roles, TLS certs, PII handling, retention, audit readiness (@Security Engineer).
- **PM/Tech Lead** – milestone tracking, risk management, stakeholder comms (@Michael Clark).

Estimate:

- MVP: ~8 weeks (2 engineers + 1 SRE/DevOps + part-time PM).
- Beta: ~6 weeks.
- GA: ~8 weeks.

Build vs Buy Choices

- **LLM**
  - **Buy:** Amazon Bedrock, OpenAI/Anthropic (via Secrets Manager).
  - **Build:** not feasible; provider APIs are commodity.
  - **Decision:** Buy — focus on orchestration, not model training.
- **OCR**
  - **Buy:** AWS Textract or Google Vision.
  - **Build:** not in scope (scanned docs excluded).
  - **Decision:** Exclude OCR from v1; revisit if demand arises.
- **Vector Store / Semantic Search**
  - **Buy:** OpenSearch, Pinecone, Weaviate.
  - **Build:** unnecessary overhead.
  - **Decision:** Not in scope until GA+; if needed, buy/integrate managed.
- **Connectors**
  - **Filesystem, S3:** Build (straightforward).
  - **SharePoint:** Build adapter via Microsoft Graph API; use Secrets Manager for creds.
  - **Decision:** Build connectors in-house for control + auditability.