Extraction Service – Data Model

Creator **Michael Clark** Created **Aug 24, 2025, 22:31** Last updated **Aug 24, 2025, 22:43**

Data Model

Tenant

Represents a customer (org or user group) using the service.

- **tenantId (PK)** – Unique identifier for the tenant.
- **name** – Display name for the tenant.
- **createdAt** – Timestamp when the tenant was onboarded.
- **status** – Active / Suspended / Deleted.
- **quota** – Configured job/document limits for this tenant.
- **billingTag** – Tag/ID for cost attribution and budgets.

TenantConfig

Configuration and security settings for a tenant.

- **tenantId (FK → Tenant)** – Links back to Tenant.
- **apiKeys** – List of active API keys with scopes/roles.
- **rbacRoles** – Role-based access control definitions.
- **connectors** – Configured sources (S3 buckets, SharePoint creds, etc).
- **piiPolicy** – PII handling config (masking, retention, right-to-delete).
- **retentionPolicy** – Default job/results/log retention in days.

Job

Represents a single extraction request submitted by a tenant.

- **jobId (PK)** – Unique identifier for the job.
- **tenantId (FK → Tenant)** – Tenant who owns this job.
- **status** – Pending / Running / Completed / Failed.
- **submittedAt** – Timestamp job was submitted.
- **completedAt** – Timestamp job finished.
- **documentCount** – Number of documents in this job.
- **metricsRef** – Pointer to job metrics (S3 `_metrics.json`).
- **manifestRef** – Pointer to job manifest (S3 `_manifest.json`).

JobConfig

Schema and parameters for an extraction job.

- **jobId (FK → Job)** – Links back to Job.
- **fields** – List of metadata fields to extract (e.g., `authors`, `date`).
- **hints** – Optional extraction hints (regex, date formats, section markers).
- **concurrency** – Max number of concurrent workers.
- **budgetCap** – Max cost allowed for this job.
- **source** – Input connector + location (S3 path, SharePoint folder, FS path).
- **outputLocation** – S3 path or tenant-configured sink for JSONL results.

JobResults

Normalized structured outputs and tracking for results retrieval.

- **jobId (FK → Job)** – Job these results belong to.
- **shards** – List of JSONL shard files (`part-0000.jsonl.gz` etc).
- **totalRecords** – Total JSON objects produced.
- **dlqRef** – Pointer to dead-letter queue output (if any).
- **checksumRef** – Checksums per shard for integrity.
- **retrievedAt** – Last time results were downloaded.

JobMetrics

Operational KPIs collected during processing.

- **jobId (FK → Job)** – Associated job.
- **docsProcessed** – Total docs processed.
- **errors** – Count of extraction errors.
- **retries** – Count of retries performed.
- **latencyP50 / latencyP95 / latencyP99** – Extraction latencies.
- **throughput** – Docs/sec throughput observed.
- **costEstimate** – Estimated cost ($/pages/tokens).

AuditLog

Immutable logs of important actions for compliance.

- **logId (PK)** – Unique event ID.
- **tenantId** – Tenant associated.
- **jobId** – Optional link to a job.
- **actor** – Who performed the action (API key, user).
- **action** – e.g., JobCreated, JobDeleted, TenantCreated.
- **timestamp** – When action occurred.
- **details** – Structured JSON with extra context.

Relationships (UML-style overview)

- **Tenant** **TenantConfig** (1:1)
- **Tenant** **Job** (1:N)
- **Job** **JobConfig** (1:1)
- **Job** **JobResults** (1:1)
- **Job** **JobMetrics** (1:1)
- **All entities** emit **AuditLog** entries for traceability.