Technical PRD: Extraction Service

a. Problem Statement, Goals & Non-Goals

Problem Statement:

Customers in regulated environments (e.g., government or defense sectors) often have large repositories of unstructured research white papers (up to 100k documents) stored in sources like AWS S3 or SharePoint. Manually extracting metadata such as author names, publish dates, abstract summaries, or code snippets is time-consuming, error-prone, and unscalable. There's a need for an automated service that uses LLMs to parse these documents into normalized JSON, while adhering to strict hosting requirements in AWS GovCloud and avoiding proprietary third-party APIs like OpenAI.

Goals:

- Enable users to configure custom metadata extraction and process batches of documents asynchronously.
- Deliver structured JSON outputs per document, with high accuracy for common fields.
- Ensure full compliance with AWS GovCloud with all components GovCloud-compatible managed services or open-source alternatives deployable there.
- Support initial scale to 100k docs with cost-effective processing targeting < $.01 per doc.
- Provide a PoC that demonstrates end-to-end workflow from ingestion to delivery.

Non-Goals:

- Real-time processing.
- Support for non-PDF/document formats (e.g., no images, videos, or proprietary binaries).
- Advanced NLP features beyond basic extraction.
- Integration with non-AWS storage beyond S3/SharePoint connectors.
- Production-grade UI/UX.

b. Target Users & Personas

Target Users:

- Data analysts and researchers in government agencies or contractors who need quick access to structured insights from white papers.
- DevOps engineers deploying the service in secure environments.
- Compliance officers auditing data handling.

Personas:

- **Adrian the Analyst:** Mid-level researcher at a gov agency, non-technical, wants to upload a config file specifying fields like "Author Names" and "Abstract Summary," then

get JSON results via email or S3 without worrying about infrastructure. Pain point: Manual extraction takes days; needs accuracy and ease.
- **Dana the Developer:** SRE in GovCloud, focuses on deployment, scaling, and monitoring. Needs clear docs for idempotent workers, observability, and cost controls. Pain point: Ensuring no data leaks or non-compliant services.
- **Cooper the Compliance Lead:** Oversees PII and audit trails. Requires features like data anonymization and logging for every extraction job.

c. Success Metrics / SLAs

Success Metrics:

- **Throughput:** Process 1k docs/hour per worker instance (scalable via auto-scaling).
- **Latency:** End-to-end job completion <24 hours for 100k docs; per-doc extraction <5 minutes.
- **Accuracy:** >95% for structured fields (e.g., dates, authors) via manual sampling; >85% for summaries/snippets (validated against ground truth datasets).
- **Cost Bounds:** <$1k total for 100k docs (assuming AWS pricing; track via CloudWatch billing metrics).
- **Uptime/Availability:** 99% for the service API over 30 days.

SLAs:

- Response time for job submission: <1 second.
- Error rate: <1% failed extractions (with retries).
- Cost cap: Alert if projected spend exceeds $500/100k docs.

d. Functional Requirements

- **Multi-Tenant:** Support isolated tenants via AWS IAM roles and resource tagging.
- **Storage Connectors:** Integrate with AWS S3 GovCloud and SharePoint via AWS-managed APIs.
- **Extraction Config:** Users provide a JSON config specifying fields (e.g., {"fields": ["author_names", "publish_date", "abstract_summary", "code_snippets"]}). Service uses open-source LLMs to extract/normalize.
- **Result Delivery:** Output normalized JSON per doc delivered to S3 bucket or via API callback. Support batch zipping for large jobs.

e. Non-Functional Requirements

- **Security:** All data encrypted at rest/transit. Role-based access (least privilege IAM). No external internet calls.
- **PII Handling:** Scan and redact PII before processing; flag docs with potential PII.
- **Auditability:** Log all actions via CloudTrail with job IDs for traceability.
- **Observability:** Metrics/logs via CloudWatch; traces with X-Ray. Basic dashboards for job status.

- **Cost Controls:** Auto-scaling groups with budgets; Lambda/SageMaker for on-demand compute to minimize idle costs.

f. Acceptance Criteria & Phased Rollout Plan

Acceptance Criteria:

- Job submission API accepts config and source path, returns job ID.
- Service crawls 10 sample docs, extracts configured fields, delivers JSON with >90% accuracy.
- Full compliance: Deploys in GovCloud without third-party deps; passes basic security scan.
- Handles failures: Retries failed extractions 3x; idempotent.

Phased Rollout Plan:

- **Phase 1 (PoC/MVP - 1 week):** Core extraction for S3 docs, 1 tenant, basic fields. Test with 100 docs.
- **Phase 2 (Beta - 2 weeks):** Add SharePoint connector, multi-tenant, PII handling. Scale to 1k docs.
- **Phase 3 (GA - 4 weeks):** Full observability, cost caps, accuracy benchmarks. Rollout to pilot customers with monitoring.