EDS AI Data Platform: Technical Architecture

Implementation Details and Operational Design

Complete Platform Architecture

Our comprehensive architecture spans AWS and GCP, enabling seamless cross-cloud data processing and AI consumption. The platform orchestrates structured and unstructured data through sophisticated event-driven workflows.

Source Systems (AWS)

Snowflake: Structured data including claims, policies, and customer records

S3: Document repository containing 20M+ files requiring AI processing

EDS Platform Layer

Event Detection: S3 events routed cross-cloud to GCP via EventBridge

Document Processor: OCR, quality validation, and intelligent evaluation

Lifecycle Manager: Automated tier transitions and retention policies

Access Controller: Template-based IAM automation and security

GCP Data Layer

Snowflake Replica: Near real-time structured data replication

Vertex AI Search: Multiple tiered data stores for document intelligence

BigQuery: Metadata, reference data, and comprehensive audit logs

Cloud Storage: Processing workspace and long-term archives

Snowflake Replication Strategy

Following Google Professional Services recommendations, our Phase 1 approach leverages Snowflake's native replication capabilities to establish reliable, low-latency data availability in GCP for immediate data enablement. This proven solution enables rapid implementation while maintaining enterprise-grade reliability, allowing for a gradual transition to a dual-platform strategy.

01

Native Replication Setup

As recommended by Google Professional Services, Snowflake's built-in database replication feature provides continuous AWS → GCP synchronization with sub-5-minute lag. This is fully managed by Snowflake infrastructure, not EDS resources, aligning with the strategy for immediate data enablement.

02

EDS Monitoring Role

EDS fulfills a monitoring role, overseeing replication lag, controlling access permissions, and creating curated views as needed, in line with Google PS guidance. Our focus remains on document intelligence rather than the complexity of data synchronization.

03

DS Team Access Options

As part of the dual-platform strategy endorsed by Google Professional Services, data science teams can query the Snowflake replica directly in GCP or access EDS-curated BigQuery views. This flexible consumption model supports varied analytical workflows.

04

Phase 2 Strategy Evolution

Google Professional Services recommended a Phase 2 strategy evolution: After 6 months of operational data, we will evaluate the BigQuery transition based on cost analysis, performance metrics, and data science team preferences. The recommended options emphasize a hybrid approach, maintaining both Snowflake and BigQuery in parallel, rather than a full migration, to ensure continued data enablement.

End-to-End Document Processing Pipeline

Our document pipeline transforms raw S3 uploads into AI-ready, searchable content within 5-8 minutes. Each stage employs specific GCP services optimized for performance and reliability at enterprise scale.

1

Event Detection (15s)

S3 event triggers SNS → EventBridge → Pub/Sub → Cloud Function chain. Immediate detection when documents upload to S3 buckets.

2

Evaluation (30s)

Cloud Function evaluates document necessity, quality acceptance, and appropriate tier assignment. Routes to HOT/WARM/COLD or rejects based on business rules.

3

Transfer (1-2min)

Storage Transfer Service efficiently copies accepted documents from S3 to GCS. Optimized cross-cloud bandwidth utilization.

4

Processing (1-3min)

Cloud Run services handle format conversion, Document AI OCR, quality validation, and entity extraction. Output: searchable, AI-ready documents.

5

Ingestion (1-2min)

Documents ingested to appropriate Vertex AI Search data stores with semantic indexing. Enables sophisticated AI agent queries.

6

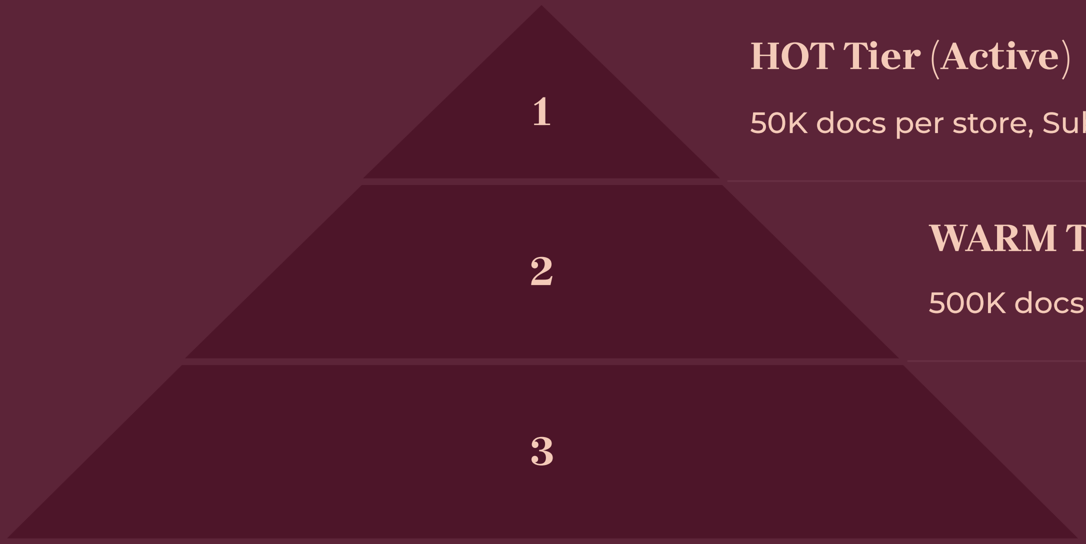
Metadata (seconds)

BigQuery updated with document location, processing metadata, and routing information. Enables intelligent query direction for AI agents.

Total Processing Time: 5-8 minutes from S3 upload to fully searchable document ready for AI consumption.

HOT/WARM/COLD Tiered Storage Implementation

Our intelligent tiering strategy maintains optimal performance while controlling costs. Active claims remain in high-performance HOT tier, while older documents move through WARM and COLD tiers based on access patterns and business rules.



HOT Tier (Active)

50K docs per store, Sub-200ms latency

WARM Tier (Recent)

500K docs per store, Sub-500ms latency

COLD Tier (Archive)

2M docs per store, Sub-1000ms latency

Data Store Organization

HOT Tier: Active claims and recently uploaded documents (last 30 days). Stores like liability-docs-active, fraud-docs-active, property-docs-active deliver sub-200ms query performance for real-time AI decisions.

WARM Tier: Recently closed claims (last 2 years) and training data. Recent stores provide sub-500ms performance for model training and reference lookups with 500K documents per specialized store.

COLD Tier: Historical claims (2-7 years) for compliance retention. Archive stores handle sub-1000ms queries acceptable for audit and compliance scenarios, supporting up to 2M documents each.

Automated Transitions

Cloud Composer DAGs execute daily tier management:

- Claim closure + 7 days → HOT to WARM
- Document age > 2 years → WARM to COLD
- Document age > 7 years → DELETE (retention met)
- Litigation holds prevent automated deletion

This approach keeps HOT tier small and fast while optimizing costs through appropriate storage classes.

Vertex AI Search Scaling Strategy

Managing 20M+ documents requires sophisticated data store partitioning to maintain optimal performance. Our metadata-first routing approach ensures AI agents query only relevant stores while automated capacity management prevents performance degradation.

Multiple Data Store Organization

Partition by use case (liability, fraud, property) and tier (active, recent, archive). Example: liability-docs-active contains 50K documents, liability-docs-recent holds 500K documents.

Metadata-First Routing

AI agents query BigQuery metadata first to identify relevant data stores for each claim. This targeted approach eliminates unnecessary searches across irrelevant stores.

Automated Capacity Management

Daily monitoring tracks document counts per store. At 80% capacity (800K documents), automated Terraform provisioning creates new stores and updates routing logic.

Automated Provisioning Workflow

1

Capacity Alert

EDS detects store at 85% capacity (850K documents)

2

Terraform Generation

Automation generates config for new data store

3

Git PR Creation

Pull request created for platform review

4

Automatic Provisioning

New Vertex data store deployed and routing updated

DS Team Experience: Completely transparent scaling with no code changes required. Query performance maintained automatically regardless of total data volume growth.

BigQuery Metadata Layer Architecture

Our metadata architecture provides intelligent routing for AI agents while maintaining comprehensive audit trails and compliance capabilities. Strategic partitioning and clustering optimize query performance across millions of document records.

1

document_index (Primary Routing)

Core routing table with document_id, claim_id, vertex_data_store_name, data_store_tier, quality_score. Partitioned by creation date, clustered by claim_id and document_type for sub-50ms routing queries.

2

data_store_registry (Capacity Tracking)

Operational table tracking store_name, tier, use_case, document_count, capacity_threshold. Updated daily by monitoring jobs to enable automated capacity management and scaling decisions.

3

access_audit_log (Compliance Trail)

Complete audit trail with timestamp, service_account, resource_name, action, query_hash. 10-year retention for regulatory compliance with automatic logging of every platform access.

4

state_traffic_laws (Reference Data)

Versioned reference data supporting time-travel queries. Fields include law_id, effective_date, superseded_date, enabling AI agents to apply historically accurate law versions.

AI Agent Query Pattern

Optimized two-step process: **Step 1:** Query metadata for routing information (under 50ms). **Step 2:** Query specific Vertex stores (200-500ms). **Total query time: Under 1 second** for complex document searches across multiple data stores.

Layered IAM and Template-Based Access Control

Our three-layer access model scales efficiently while maintaining security. Template-based provisioning enables consistent, automated access grants without manual IAM configuration for each data science team.

Layer 1: Platform Team (Infrastructure)

Creates service accounts per DS team (e.g., ds-auto-liability@project.iam). Grants minimal project permissions for infrastructure access. One-time setup per team.

Layer 2: EDS (Data Products)

Grants access to specific data products using template-based provisioning. Controls which teams access which datasets and Vertex stores. Ongoing access management.

Layer 3: DS Teams (Consumption)

Query data via granted permissions through service accounts. Cannot modify their own access. All usage automatically logged for audit compliance.

Template-Based Provisioning

Template Components (Liability Assessment v2.0):

- BigQuery resources: Specific tables and views
- Vertex AI resources: Designated data stores by tier
- IAM roles: dataViewer, discoveryengine.viewer
- Access conditions: Data classification requirements

Automated Application Process

5-15 Minute Provisioning:

1. DS team requests data product via portal
2. EDS reads appropriate template configuration
3. Template applied to team's service account
4. All permissions granted atomically
5. Access logged to comprehensive audit trail

Scalability Advantage: Adding the 10th team requires identical effort as the 1st team. Templates enable consistent access patterns across unlimited teams without manual configuration.

Time-Travel Queries for Compliance

Regulatory compliance demands that AI decisions use law versions active on incident dates, not current versions. Our versioned reference data architecture enables precise historical accuracy with complete audit trails.

01

Version Storage Architecture

BigQuery stores structured law data with version tracking fields: law_id, effective_date, superseded_date, is_current flag. Vertex AI Search maintains full PDF documents for semantic search capabilities.

02

Time-Travel Query Logic

Filter criteria: effective_date ≤ incident_date AND (superseded_date IS NULL OR superseded_date > incident_date). Returns law version that was legally active on the specific incident date.

03

Automated Version Activation

Daily jobs check for laws with effective_date = today. Updates old version: is_current = FALSE, sets superseded_date. Activates new version: is_current = TRUE while preserving historical versions.

04

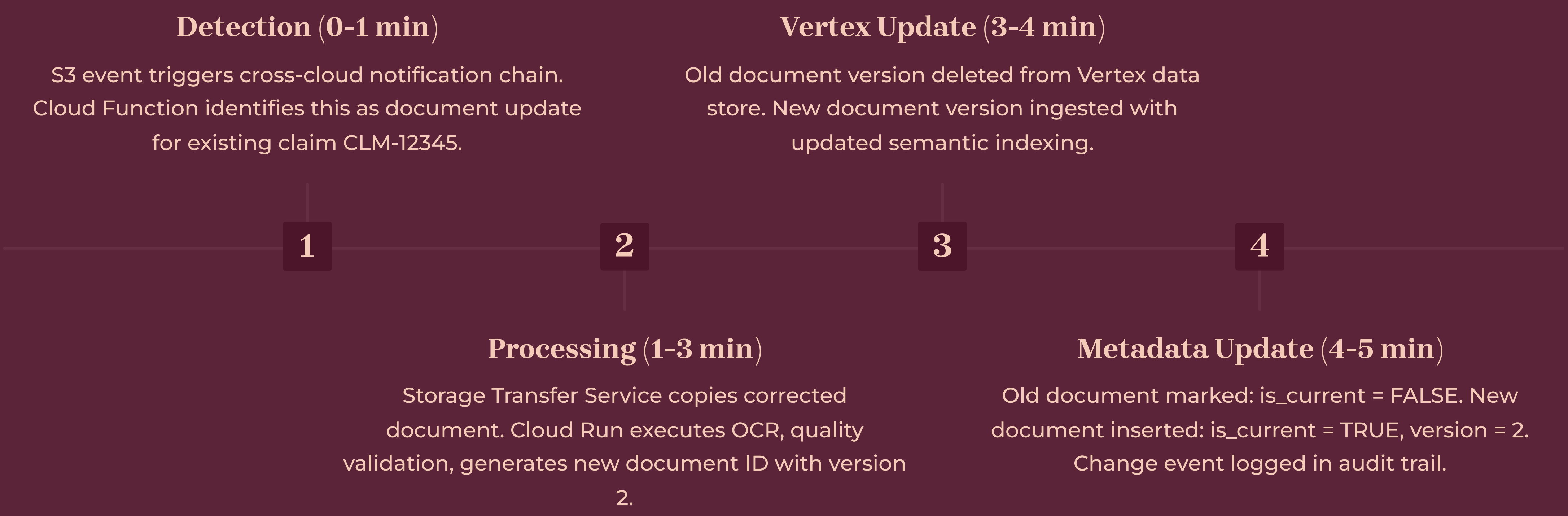
Complete Audit Trail

Every AI decision logs which law version was applied. Enables exact reproduction of historical decisions with legally defensible chain of custody documentation.

> **Compliance Example:** Incident in March 2024, law changed July 2024 - AI must apply March version for that incident, ensuring legally accurate and defensible decisions.

Document Replacement Scenario

When CIOs identify incorrect documents, our automated replacement system ensures AI agents immediately access corrected versions while preserving complete version history and audit trails.



Detection (0-1 min)

S3 event triggers cross-cloud notification chain. Cloud Function identifies this as document update for existing claim CLM-12345.

Processing (1-3 min)

Storage Transfer Service copies corrected document. Cloud Run executes OCR, quality validation, generates new document ID with version 2.

Vertex Update (3-4 min)

Old document version deleted from Vertex data store. New document version ingested with updated semantic indexing.

Metadata Update (4-5 min)

Old document marked: is_current = FALSE. New document inserted: is_current = TRUE, version = 2. Change event logged in audit trail.

Automated Response Workflow

Initial State: Incorrect police report uploaded to S3 for claim CLM-12345, processed and available to AI agents in liability-docs-active store.

CIO Action: Discovers error and uploads corrected version to S3 with same claim association but new filename.

EDS Response: Complete 5-minute automated workflow replaces document without manual intervention. AI agents automatically receive new version for subsequent queries.

Safety Features

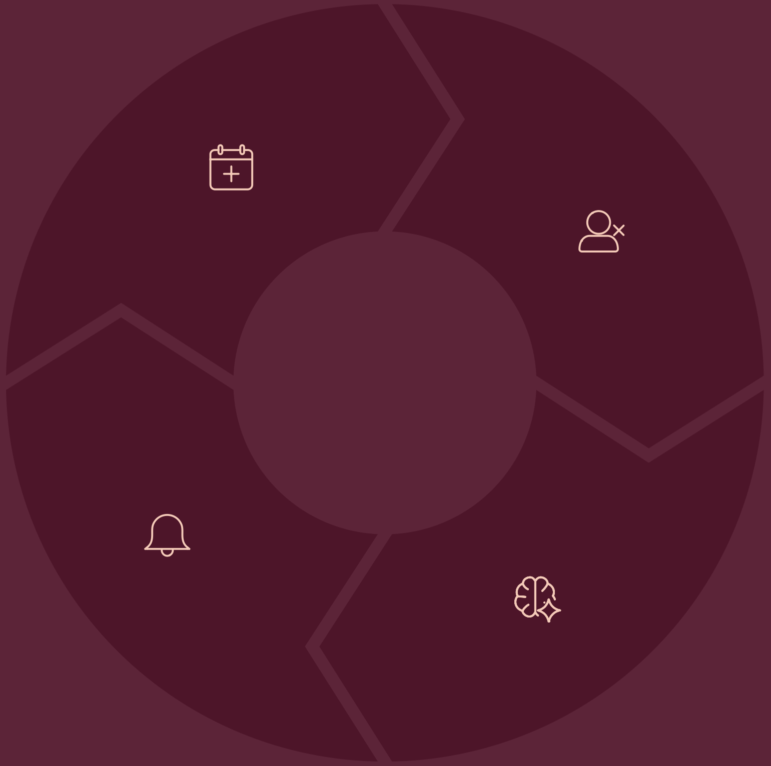
In-Flight Query Protection: If AI agent queries during replacement window, decision flagged for review with automatic re-run capability using correct document.

Version Preservation: Complete history maintained with clear superseded_by relationships and comprehensive audit trail of all changes.

Result: Zero manual intervention required. AI agents automatically use corrected version while maintaining complete audit trail and version history for compliance.

Automated Retention and Deletion

Our comprehensive retention management system ensures regulatory compliance while protecting against accidental deletion. Automated workflows handle 7-year retention policies and GDPR "right to be forgotten" requests with complete audit trails.



Daily Retention Check

Query identifies documents where claim closed > 7 years ago, excluding litigation hold flags.

Cascade Deletion Process

Delete from Vertex AI Search across all tiers, remove from Cloud Storage, soft delete in BigQuery metadata.

Audit Trail Preservation

Maintain metadata records marked as deleted with 10-year audit log retention for compliance.

CIO Notification

Alert CIO that documents eligible for S3 deletion after EDS processing complete.

GDPR "Right to be Forgotten" Process

Deletion Workflow

1. Legal team submits request with customer_id
2. EDS identifies all customer documents across claims
3. Cascade deletion executed per document
4. Completion within 7 days (regulatory requirement)
5. Deletion confirmation report for legal team

Critical Safeguards

- Litigation holds prevent automated deletion
- Manual override requires compliance officer approval
- All deletions logged and auditable
- Multiple verification checks prevent accidents
- Complete chain of custody maintained

Technical Readiness Assessment

Our architecture leverages proven, enterprise-grade technologies with comprehensive automation designed for immediate implementation. All critical components are validated and ready for 8-week deployment timeline.

100%

Architecture Readiness

Phase 1 uses proven Snowflake replication and GCP services. Phase 2 learning approach gathers evidence before major decisions.

10x

Scalability Factor

Current design handles 10x data volume growth without architectural changes through automated provisioning.

8

Weeks to Production

Complete platform deployment with first data product and pilot DS team integration.

Technology Stack Validation

GCP Services: All generally available with enterprise support. Vertex AI Search, BigQuery, Cloud Run, Storage Transfer Service proven at scale.

Snowflake Replication: Native feature available and tested for AWS → GCP scenarios with sub-5-minute lag guarantees.

Cross-Cloud Integration: EventBridge → Pub/Sub routing follows established patterns with reliability testing complete.

Automation Design Status

Data Store Provisioning: Terraform templates defined for automated Vertex AI Search scaling based on capacity thresholds.

Tier Management: Daily job logic specified with Cloud Composer DAG workflows for document lifecycle management.

Access Control: Template-based system architected with complete IAM delegation patterns and audit logging.

Comprehensive Monitoring

Metrics defined for freshness, performance, capacity, quality, and cost. Alert thresholds established for critical issues with operational dashboards designed.

Enterprise Security

Layered IAM delegation model, encryption at rest and in transit, complete audit logging, and compliance support for retention and GDPR requirements.

8-Week Implementation Timeline

Our phased approach delivers a working platform with pilot integration in 8 weeks. Each phase builds systematically on proven foundations while maintaining parallel workstreams for maximum efficiency.

1

Weeks 1-2: Foundation Infrastructure

Platform team provisions GCP project with required services. Setup Snowflake replication for structured data. Deploy cross-cloud event routing via EventBridge → Pub/Sub integration. Establish basic monitoring and security frameworks.

2

Weeks 3-4: Document Pipeline Core

Deploy Cloud Run processing services for OCR and quality validation. Create initial Vertex AI Search data stores for HOT/WARM/COLD tiers. Implement automated tier management with Cloud Composer DAGs. Test end-to-end document processing workflow.

3

Weeks 5-6: Access Control & Metadata

Deploy BigQuery schemas for metadata and audit logging. Implement template-based access automation with IAM delegation. Build operational monitoring dashboards and alerting. Configure automated capacity management and scaling workflows.

4

Weeks 7-8: Pilot Integration

Onboard first DS team (auto liability use case) with complete access provisioning. Integrate pilot AI agent with document search capabilities. Validate performance metrics and measure against SLA targets. Conduct full operational readiness review.

Week 1-2 Deliverables

- GCP project provisioned and configured
- Snowflake replication active and monitored
- Cross-cloud event routing operational
- Security and compliance frameworks deployed

Week 3-4 Deliverables

- Document processing pipeline operational
- Vertex AI Search stores configured
- Tier management automation active
- Processing workflow validated

Week 5-6 Deliverables

- Metadata layer fully operational
- Access automation implemented
- Monitoring dashboards deployed
- Scaling automation configured

Week 7-8 Deliverables

- First DS team successfully onboarded
- AI agent integration validated
- Performance SLAs verified
- Production readiness confirmed

Risk Mitigation and Success Factors

Our implementation approach minimizes technical and operational risks through proven technologies, comprehensive monitoring, and phased rollout strategies. Critical success factors ensure sustainable platform operations and stakeholder alignment.

Technical Risk Mitigation

Proven Technology Stack: Leverage GA GCP services and Snowflake native features rather than custom solutions. **Redundancy:** Multiple data stores and automated failover capabilities. **Monitoring:** Comprehensive observability with proactive alerting for all critical paths.

Operational Risk Management

Automated Recovery: Self-healing systems for common failure modes. **Rollback Capabilities:** Version control and automated rollback for all infrastructure changes. **Security Safeguards:** Multiple approval layers for sensitive operations like deletions.

Scalability Planning

Capacity Monitoring: Automated scaling before performance impact. **Load Testing:** Validate performance under 10x expected load. **Cost Controls:** Automated budget alerts and resource optimization recommendations.

Critical Success Factors

Stakeholder Alignment

- **Platform Team:** Resource commitment for infrastructure provisioning and ongoing support
- **CIO Partnership:** S3 event routing configuration and document governance alignment
- **DS Team Engagement:** Early adopter commitment for pilot validation and feedback
- **Security Approval:** IAM delegation model and cross-cloud data flow approval

Operational Readiness

- **Monitoring Systems:** Complete observability across all platform components
- **Incident Response:** Clear escalation paths and automated recovery procedures
- **Documentation:** Comprehensive operational runbooks and troubleshooting guides
- **Training:** Platform team knowledge transfer for ongoing maintenance

Risk Assessment: Low technical risk due to proven components. Primary risks center on stakeholder coordination and resource allocation rather than technical implementation challenges.

Next Steps and Decision Points

Immediate action required to maintain 8-week delivery timeline. Critical decisions and resource commitments must be finalized to begin infrastructure provisioning and stakeholder coordination activities.

Team & Resource Commitment

Finalize dedicated team assignments and budget approvals for 8-week implementation timeline

Infrastructure Provisioning

Begin Week 1 GCP project setup and Snowflake replication configuration immediately

Stakeholder Partnerships

Establish formal partnerships with CIO and Platform teams for ongoing collaboration

Immediate Action Items (Next 48 Hours)

1 Executive Approval

Secure final budget approval and resource allocation for 8-week implementation. Confirm dedicated team members and their availability for full-time assignment to platform development.

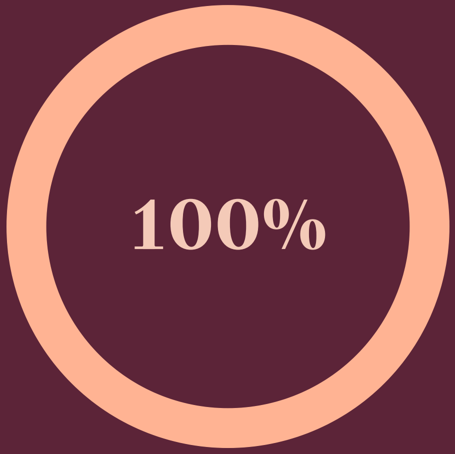
2 Platform Team Coordination

Schedule kick-off meeting with Platform team for GCP project provisioning. Review infrastructure requirements and confirm service account creation processes.

3 CIO Partnership Agreement

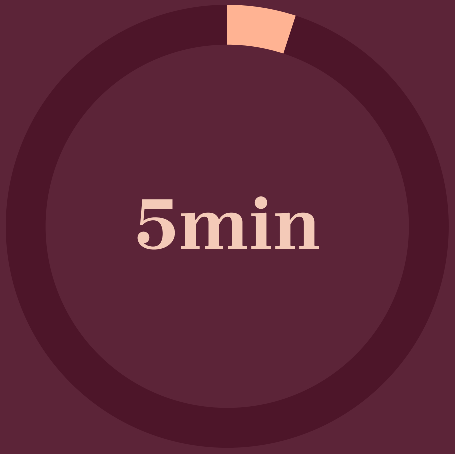
Finalize S3 event routing configuration requirements and document governance policies. Establish ongoing collaboration framework for document lifecycle management.

Week 1 Success Criteria



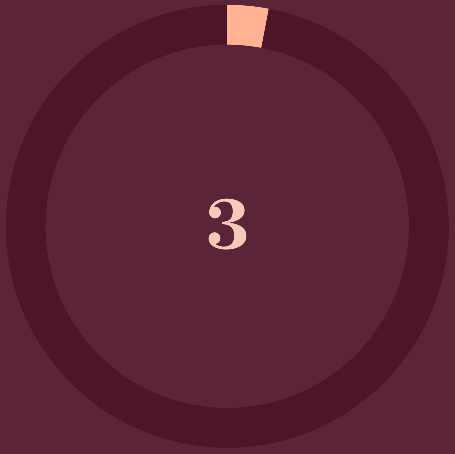
Infrastructure Readiness

GCP project provisioned with all required services enabled and configured



Replication Lag

Snowflake replication active with sub-5-minute lag verification



Team Integration

Platform, EDS, and CIO teams aligned with clear communication channels

> **Commitment Required:** Success depends on immediate stakeholder alignment and resource commitment. Delays in decision-making directly impact 8-week delivery timeline and pilot DS team integration schedule.

Key Takeaways

Our proposed EDS AI Data Platform architecture delivers a robust, scalable, and secure foundation for advanced data analytics and AI/ML initiatives. Success hinges on a clear path forward and immediate collaboration.

Robust Architecture

A comprehensive platform leveraging proven GCP and Snowflake technologies to ensure scalability, reliability, and security for all data needs.

Accelerated Delivery

An 8-week phased implementation timeline ensures rapid deployment, continuous validation, and integration of critical components with minimal risk.

Empowering AI/ML

The platform unlocks powerful capabilities for data science teams, enabling faster experimentation, model training, and deployment of AI-driven solutions.

Collaborative Success

Achieving our goals requires immediate resource commitment and strong partnership across EDS, CIO, and Platform teams for seamless execution.