Data Platform Incremental Loading Options - Executive Summary

Options Comparison Matrix

Assessment Criteria	Option 1: Current State (Audit Column Based)	Option 2: GoldenGate CDC → Confluent Kafka	Option 3: Debezium CDC → Confluent Kafka
Current Status	✅ **In Production**	❌ **Not Implemented**	❌ **Not Implemented**
Risk Level	🔴 **High** - Unreliable audit columns	🟡 **Medium** - Complex but mature	🟡 **Medium** - Open source dependencies
Mitigation Strategy	Weekly full loads + month-end catch-up	Real-time CDC eliminates data gaps	Real-time CDC eliminates data gaps

Infrastructure & Cost Assessment

Component	Option 1: Current State	Option 2: GoldenGate CDC	Option 3: Debezium CDC
Infrastructure Setup			
Complexity	⭐ **Simple** - Existing PySpark jobs	⭐⭐⭐⭐⭐ **Complex** - Multi-tier setup	⭐⭐⭐⭐ **Complex** - Kafka + connectors
Additional Components	None (current stack)	GoldenGate + Confluent Kafka + Schema Registry	Confluent Kafka + Debezium + Schema Registry
Compute Requirements	Current OpenShift resources	+30-50% compute for GG + Kafka	+25-40% compute for Kafka cluster
Storage Requirements	Existing Iceberg storage	+Kafka retention + GG trail files	+Kafka retention + connector state
Cost Structure			
Licensing Costs	$0 (existing)	**High**: Oracle GG (~~$50K-200K/year) + Confluent~~ ($30K-100K/year)	**Low**: Confluent only (~$30K-100K/year)
Infrastructure Costs	Existing	+$15K-30K/month (GG infra + Kafka)	+$10K-20K/month (Kafka cluster)
Operational Overhead	Current team	+2-3 FTE specialists	+1-2 FTE specialists
Total Annual Cost	Existing baseline	+$400K-600K annually	+$200K-350K annually
Licensing Model			

Component	Option 1: Current State	Option 2: GoldenGate CDC	Option 3: Debezium CDC
Oracle GoldenGate	N/A	Per processor core + support (20-22% annually)	N/A
Confluent Platform	N/A	Per connector + cluster size	Per connector + cluster size
Support Model	Internal team	Oracle + Confluent enterprise support	Confluent + community/commercial Debezium

Implementation & Operations

Aspect	Option 1: Current State	Option 2: GoldenGate CDC	Option 3: Debezium CDC
Phased Timeline			
Phase 1 (Months 1-2)	✅ **Complete**	Infrastructure provisioning + GG setup	Kafka cluster setup + Debezium deployment
Phase 2 (Months 3-4)	Ongoing maintenance	Oracle/SQL Server GG configuration	Oracle/SQL Server connector configuration
Phase 3 (Months 5-6)	Enhancement projects	MongoDB integration + testing	MongoDB integration + testing
Phase 4 (Months 7+)	BAU operations	Production rollout + optimization	Production rollout + optimization
Disaster Recovery			
Strategy	Iceberg snapshot restoration	Active-passive GG + Kafka replication	Kafka cluster replication + connector failover
RTO Target	4-6 hours (full reload)	15-30 minutes	15-30 minutes
RPO Target	Up to 1 week (audit gap risk)	< 1 minute	< 1 minute
DR Testing	Quarterly full load tests	Monthly failover tests	Monthly failover tests
Replay Capabilities			
Data Recovery	Manual full/incremental reload	GG trail file replay	Kafka topic replay
Time Range Recovery	Limited (audit column constraints)	Precise timestamp recovery	Kafka retention window
Granularity	Table-level	Transaction-level	Message-level
Scalability			

Aspect	Option 1: Current State	Option 2: GoldenGate CDC	Option 3: Debezium CDC
Current Capacity	Handles current volume with gaps	Enterprise-grade horizontal scaling	Good horizontal scaling
Scale-out Model	PySpark cluster scaling	GG extract processes + Kafka partitioning	Kafka partitioning + connector scaling
Performance Ceiling	Limited by batch windows	Very high throughput	High throughput
Cloud Readiness			
Cloud Native	✅ OpenShift compatible	⚠️ Traditional enterprise (cloud adaptable)	✅ Kubernetes/OpenShift native
Managed Services	Uses existing managed services	Limited cloud-managed GG options	Confluent Cloud available
Container Support	✅ Current containerized setup	Partial containerization	✅ Full containerization
Multi-cloud	✅ IBM Cloud Object Storage	Limited portability	✅ Cloud agnostic

Source Application Architecture Changes

Requirement	Option 1: Current State	Option 2: GoldenGate CDC	Option 3: Debezium CDC
Application Changes Required			
Oracle Applications	❌ **None**	❌ **None** - Log-based CDC	❌ **None** - Log-based CDC
SQL Server Applications	❌ **None**	❌ **None** - Log-based CDC	❌ **None** - Log-based CDC
MongoDB Applications	❌ **None**	❌ **None** - Oplog/Change streams	❌ **None** - Change streams
Outbox Pattern Implementation			
Required for Current	❌ **No**	❌ **No** - CDC bypasses outbox need	🟡 **Optional** - Enhanced reliability
Implementation Effort	N/A	N/A	Low-Medium if implemented
Application Code Changes	N/A	N/A	Transactional outbox table + event publishing
Benefits if Implemented	N/A	N/A	Guaranteed event delivery + ordering

Outbox Pattern Considerations (Option 3 Enhancement)

Component	Without Outbox Pattern	With Outbox Pattern
Reliability	Depends on Debezium CDC reliability	Guaranteed event delivery
Ordering	Database transaction order	Application-controlled ordering
Application Changes	None required	Moderate - add outbox tables + logic
Event Schema	Database schema driven	Application domain driven
Implementation	Ready to use	2-3 months additional development

Risk Assessment & Mitigation

Risk Category	Option 1: Current	Option 2: GoldenGate	Option 3: Debezium
Technical Risks			
Data Accuracy	🔴 **High** - Missing updates	🟢 **Low** - Complete CDC	🟡 **Medium** - Config dependent
Vendor Lock-in	🟢 **Low** - Open stack	🔴 **High** - Oracle ecosystem	🟡 **Medium** - Confluent dependency
Technology Obsolescence	🟢 **Low** - Standard approach	🟡 **Medium** - Enterprise CDC	🟢 **Low** - Modern stack
Operational Risks			
Skills Gap	🟢 **Low** - Current expertise	🔴 **High** - Specialized skills	🟡 **Medium** - Kafka expertise
Complexity	🟢 **Low** - Simple pipeline	🔴 **High** - Multi-component	🟡 **Medium** - Kafka ecosystem
Business Risks			
Cost Overrun	🟢 **Low** - Known costs	🔴 **High** - Complex pricing	🟡 **Medium** - Predictable costs
Timeline Risk	🟢 **Low** - In production	🔴 **High** - 9-12 month timeline	🟡 **Medium** - 6-9 month timeline

Executive Recommendation Summary

Criteria	Recommendation	Timeline	Investment
Immediate (0-6 months)	**Continue Option 1** with enhanced monitoring	Q1-Q2	Minimal
Short-term (6-12 months)	**Pilot Option 3** for high-value tables	Q3-Q4	$200K-350K
Long-term (12+ months)	**Full Option 3** deployment with selective Option 2 for critical systems	Year 2+	$400K+
Risk Mitigation	Implement **hybrid approach**: maintain Option 1 as fallback during CDC rollout	Ongoing	10-15% overhead