Data Platform Incremental Loading Strategy

Executive Presentation - Slide Content

Slide 1: Executive Summary & Recommendation

Current Challenge: Unreliable incremental loading based on application-maintained audit columns

Option	Investment	Timeline	Risk Level	Recommendation
Option 1: Current State	$0	Ongoing	🔴 High	Continue with enhanced monitoring
Option 2: GoldenGate CDC	$400K-600K annually	9-12 months	🟡 Medium	Future consideration for critical systems
Option 3: Debezium CDC	$200K-350K annually	6-9 months	🟡 Medium	✅ **Recommended pilot approach**

Strategic Recommendation: Phased hybrid approach - pilot Option 3 while maintaining Option 1 as fallback

---

Slide 2: Current State Assessment (Option 1)

✅ **Strengths:**

- Already in production with known operational model
- Minimal additional infrastructure investment required
- Team expertise and established processes

🔴 **Critical Risks:**

- Unreliable application-maintained audit columns
- Data gaps and inconsistencies impacting downstream analytics
- Manual intervention required for data quality issues

Current Mitigation:

- Weekly full loads to catch missed updates
- Month-end processing full reload
- Manual reconciliation processes

Business Impact: Data accuracy issues affecting critical reporting and analytics

---

Slide 3: Technology Comparison Matrix

Criteria	Current State	GoldenGate CDC	Debezium CDC
Data Reliability	🔴 Poor	🟢 Excellent	🟢 Excellent
Real-time Capability	❌ Batch only	✅ Sub-second	✅ 1-5 seconds
Total Annual Cost	Baseline	+$400K-600K	+$200K-350K
Implementation Complexity	⭐ Simple	⭐⭐⭐⭐⭐ Very High	⭐⭐⭐⭐ High
Vendor Lock-in Risk	🟢 Low	🔴 High (Oracle)	🟡 Medium
Cloud Readiness	✅ Current	⚠️ Limited	✅ Native
Skills Required	Current team	+2-3 specialists	+1-2 specialists

---

Slide 4: Infrastructure & Cost Analysis

GoldenGate CDC (Option 2)

Component	Annual Cost	Notes
Oracle GoldenGate License	$50K-200K	Per processor core model
Confluent Kafka Platform	$30K-100K	Per connector + cluster size
Additional Infrastructure	$180K-360K	Compute + storage + operations
Total Annual Investment	**$400K-600K**	**Enterprise-grade solution**

Debezium CDC (Option 3)

Component	Annual Cost	Notes
Confluent Kafka Platform	$30K-100K	Per connector + cluster size
Additional Infrastructure	$120K-240K	Compute + storage + operations
Support & Operations	$50K-100K	Team augmentation
Total Annual Investment	**$200K-350K**	**Cost-effective CDC solution**

Slide 5: Implementation Timeline & Phases

Option 2: GoldenGate CDC Timeline

Phase	Duration	Key Activities	Investment
Phase 1	Months 1-3	Infrastructure setup, GG installation	$100K-150K
Phase 2	Months 4-6	Oracle/SQL Server configuration	$150K-200K
Phase 3	Months 7-9	MongoDB integration, testing	$100K-150K
Phase 4	Months 10-12	Production rollout, optimization	$50K-100K

Option 3: Debezium CDC Timeline

Phase	Duration	Key Activities	Investment
Phase 1	Months 1-2	Kafka cluster setup, Debezium deployment	$50K-75K
Phase 2	Months 3-4	Connector configuration, testing	$75K-100K
Phase 3	Months 5-6	Production pilot, MongoDB integration	$50K-75K
Phase 4	Months 7-9	Full rollout, optimization	$25K-50K

Slide 6: Risk Assessment & Mitigation

Risk Category	Current State	GoldenGate CDC	Debezium CDC
Data Accuracy	🔴 High - Missing updates	🟢 Low - Complete CDC	🟡 Medium - Config dependent
Operational Complexity	🟢 Low - Known processes	🔴 High - Multi-component	🟡 Medium - Kafka expertise
Vendor Dependency	🟢 Low - Open stack	🔴 High - Oracle ecosystem	🟡 Medium - Confluent
Cost Predictability	🟢 High - Fixed costs	🔴 Low - Complex pricing	🟢 High - Transparent model
Skills Availability	🟢 High - Current team	🔴 Low - Specialized	🟡 Medium - Market available

Key Mitigation Strategy: Implement hybrid approach with gradual CDC adoption while maintaining current system as fallback

---

Slide 7: Business Impact Analysis

Current State Challenges

- **Data Quality Issues:** 15-20% of incremental loads miss updates
- **Manual Effort:** 40+ hours monthly for data reconciliation
- **Reporting Delays:** Week-end processing delays impact Monday reporting
- **Compliance Risk:** Audit trail gaps for regulatory reporting

CDC Benefits (Options 2 & 3)

- **Real-time Data:** Enable near real-time analytics and reporting
- **Data Integrity:** 99.9%+ accuracy with complete change capture
- **Operational Efficiency:** Eliminate manual reconciliation processes
- **Business Agility:** Support real-time decision making and ML pipelines

ROI Projection (3-year)

Benefit Category	Annual Value	3-Year Total
Reduced Manual Effort	$120K	$360K
Improved Decision Speed	$200K	$600K
Compliance Automation	$80K	$240K
Total Business Value	**$400K**	**$1.2M**

Slide 8: Source System Impact Assessment

No Application Code Changes Required

Source System	Current Method	CDC Method	Application Impact
Oracle OLTP	Audit column queries	Transaction log CDC	❌ No changes
SQL Server	Modified date tracking	Transaction log CDC	❌ No changes
MongoDB	ObjectId timestamps	Change streams	❌ No changes

Optional Enhancement: Outbox Pattern (Option 3)

Benefits if Implemented:

- Guaranteed event delivery and ordering
- Application-controlled event schema
- Enhanced reliability for critical transactions

Implementation Effort: 2-3 months additional development per application

---

Slide 9: Disaster Recovery & Scalability

Capability	Current State	GoldenGate CDC	Debezium CDC
RTO (Recovery Time)	4-6 hours	15-30 minutes	15-30 minutes
RPO (Data Loss)	Up to 1 week	< 1 minute	< 1 minute
Replay Capability	Manual reload	Trail file replay	Kafka topic replay
Horizontal Scaling	PySpark scaling	Enterprise-grade	Good scaling
Cloud Portability	✅ High	⚠️ Limited	✅ High
Multi-region Support	Manual setup	Built-in replication	Kafka replication

Key Advantage: CDC approaches provide enterprise-grade DR capabilities with minimal data loss

---

Slide 10: Recommended Implementation Strategy

Phase 1: Foundation (Months 1-3)

- **Continue Option 1** with enhanced monitoring and alerting
- **Pilot Option 3** for 2-3 high-value tables (Customer, Orders)
- Establish Kafka infrastructure on OpenShift
- Team training and skill development

Phase 2: Expansion (Months 4-6)

- Extend Debezium CDC to remaining critical tables
- Implement comprehensive monitoring and alerting
- Establish operational procedures and runbooks
- Performance optimization and tuning

Phase 3: Full Migration (Months 7-12)

- Migrate all suitable tables to CDC approach
- Decommission batch-based incremental processes
- Implement advanced features (schema evolution, data quality)
- Consider GoldenGate for ultra-critical systems if needed

Success Criteria

- 99%+ data accuracy across all ingestion pipelines
- < 5 minute data freshness for critical business processes
- 50%+ reduction in manual data operations effort

Slide 11: Next Steps & Decision Points

Immediate Actions (Next 30 Days)

1. **Stakeholder Alignment:** Secure executive sponsorship and budget approval
2. **Vendor Engagement:** Initial discussions with Confluent for licensing and support
3. **Pilot Planning:** Define pilot scope, success criteria, and timeline
4. **Team Preparation:** Identify training needs and skill gaps

Decision Points

- **Budget Approval:** $200K-350K for Debezium CDC implementation

- **Resource Allocation:** 1-2 additional team members or contractor support
- **Pilot Scope:** Select 2-3 tables for initial CDC implementation
- **Timeline Commitment:** 6-9 month implementation timeline

Risk Mitigation

- Maintain current system as fallback during migration
- Implement comprehensive testing and validation
- Phased rollout with incremental value delivery
- Regular checkpoint reviews and course correction

Recommendation: Proceed with Option 3 (Debezium CDC) pilot while maintaining Option 1 operational