zensar

Multi-Source Healthcare Analytics: Unifying Enterprise Data with Databricks Standard Edition

Whitepaper

An RPG Company

Executive Summary

Healthcare organizations today face unprecedented challenges in managing data scattered across multiple enterprise systems. This white paper details a comprehensive implementation that achieved significant cost savings while processing multi-source data in minutes instead of hours, using Databricks Standard Edition with intelligent optimization techniques.

The Healthcare Data Integration Challenge

Current state of healthcare data management

The healthcare industry is experiencing a data explosion. According to recent industry analyses, healthcare data is growing at an annual rate of 36%, faster than any other industry sector. This growth stems from multiple sources: electronic health records (EHRs), medical imaging systems, IoT devices, financial systems, and operational databases.

However, this data profusion creates significant challenges:

Data silos across enterprise systems: Healthcare organizations typically operate 50-100+ disparate systems, including financial platforms (SAP, Oracle), document repositories (SharePoint), development systems (GitLab), and specialized clinical applications. Each system operates in isolation, creating barriers to comprehensive analytics.

Compliance and security complexity: HIPAA regulations, state privacy laws, and audit requirements demand sophisticated data governance. Traditional integration approaches struggle to maintain consistent security controls across multiple platforms, increasing compliance risk and audit complexity.

Cost pressures and resource constraints: Healthcare margins continue to shrink while technology demands increase. Organizations face pressure to modernize analytics capabilities while controlling IT budgets. Traditional enterprise data platforms can consume 15-25% of IT budgets, creating significant financial burden.

Integration complexity: Connecting financial systems, clinical documentation, development artifacts, and unstructured medical data traditionally requires multiple specialized tools — each with separate licensing, maintenance overheads, and integration complexity. This multi-tool approach creates maintenance challenges and increases total cost of ownership.

Performance requirements: Clinical decision support, operational reporting, and financial analytics require near-real-time data access. Traditional batch processing approaches that take hours to update analytics dashboards no longer meet organizational

The Market Need

Healthcare organizations require a unified data platform that can:

- Integrate diverse data sources without extensive custom development
- Maintain strict security and compliance requirements
- Deliver near-real-time analytics capabilities
- Scale efficiently as data volumes grow
- Control costs while enabling innovation

This white paper demonstrates how a mid-sized healthcare organization addressed these challenges through intelligent platform architecture and optimization strategies.

The Client Challenge

Our client, a mid-sized healthcare organization, faced typical data integration challenges while building their new analytics platform:

Data source requirements

- **Financial data integration:** Multiple systems including SAP Concur expense management required integration with broader enterprise financial systems.
- **Clinical documentation:** Documents were scattered across SharePoint sites with varying file formats and compliance requirements.
- **Development artifacts:** GitLab repositories containing both source code and operational data files required replication for compliance tracking.
- **Unstructured medical data:** Medical images and reports came in multiple formats (PDF, DOCX, JPEG, DICOM) from various clinical systems.

Business constraints

- Budget limitations requiring a cost-optimized approach without compromising enterprise capabilities
- Compliance requirements for healthcare data security and audit trails
- Performance expectations for near-real-time analytics
- Scalability needs for future data source expansion

Solution Architecture: Databricks Standard Edition Platform

Architecture overview

Our solution creates a seamless flow from multiple sources to actionable insights through five integrated layers:

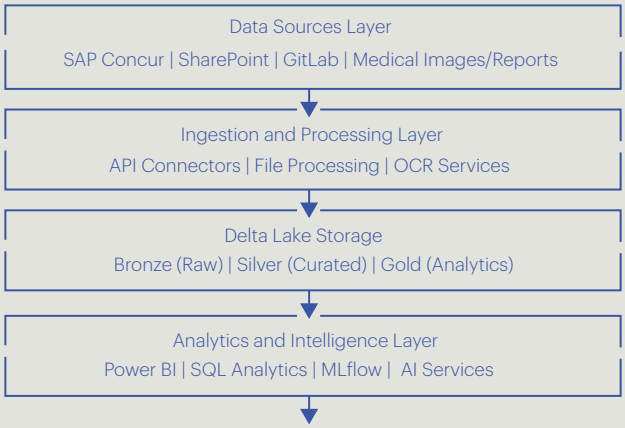

Figure 1: End-to-end healthcare data integration architecture

Core platform capabilities

- **Unified data platform:** Single platform for all data processing needs, eliminating integration overhead between specialized tools.
- **Standard edition feature set:** Enterprise-grade capabilities including Delta Lake, MLflow, and advanced analytics without premium pricing.
- **Scalable architecture:** Auto-scaling infrastructure supporting growth from initial implementation to enterprise-wide deployment.
- **Security foundation:** Built-in encryption, access controls, and audit logging supporting HIPAA compliance requirements.

Implementation Approach

Multi-source data ingestion strategy

SAP Concur financial data processing

- Secure API connections using Databricks secrets management for credential security
- Structured data extraction with automatic error handling and retry mechanisms
- Data validation and quality checks during ingestion process
- Audit trails for financial compliance requirements

SharePoint document processing

- Automated document discovery across multiple SharePoint sites and libraries
- Multi-format text extraction supporting PDFs, Word documents, and clinical forms
- Content classification and metadata enrichment for searchability
- Change detection mechanisms for incremental processing

GitLab code and data file integration

- Automated replication of both source code and operational data files
- Hash-based change tracking for compliance and audit requirements
- Secure data transfer mechanisms with encryption in transit and at rest
- Real-time synchronization for critical operational data files

Unstructured medical data processing

- OCR capabilities for medical imaging and scanned document processing
- Specialized text extraction for clinical reports and forms
- Medical terminology recognition and standardization processes
- Secure handling workflows for PHI-compliant data processing

Cost Optimization Strategy

Databricks Standard Edition optimization techniques

Our approach leveraged six key capabilities to achieve significant cost savings:

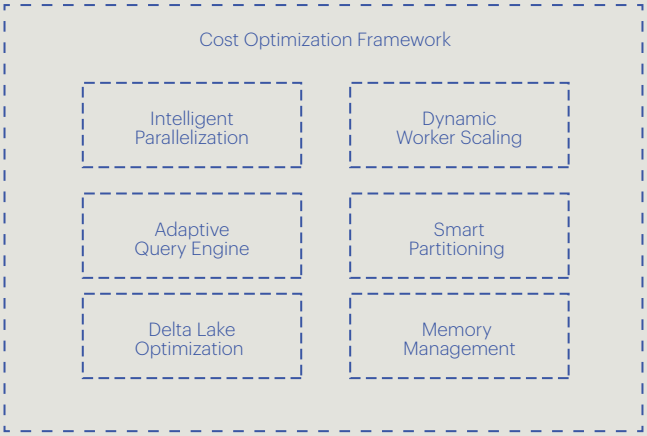

Figure 2: Databricks Standard Edition cost optimization through intelligent processing

- **Intelligent parallelization**

Databricks Spark engine automatically distributes processing across multiple worker nodes, maximizing CPU utilization. Our 1,000-record batch size optimization ensures optimal data distribution while minimizing memory overhead and network communication.

- **Dynamic worker node management**

Auto-scaling configuration dynamically adjusts cluster size from 2 to 8 worker nodes based on actual workload demands. This eliminates paying for idle resources while ensuring adequate capacity during peak processing, with automatic termination after 30 minutes of inactivity.

- **Adaptive query optimization**

Spark 3.0's Adaptive Query Engine automatically optimizes join strategies, partition coalescing, and skew handling without manual intervention. This reduces query execution times by up to 40% through intelligent resource allocation.

- **Smart data partitioning**

Date and source system-based partitioning strategy reduces data scanning by 80% during analytical queries. This approach optimizes both storage costs and query performance by organizing data according to most common access patterns.

- **Delta Lake storage optimization**

OPTIMIZE and ZORDER commands maintain peak storage and query performance through file compaction and data clustering. This reduces storage costs through better compression while improving query response times.

- **Memory management excellence**

Intelligent caching of intermediate results and frequently accessed datasets reduces redundant processing. Combined with optimized serialization settings, this minimizes garbage collection overhead and maximizes available memory.

Results and Outcomes

Performance

- **Processing efficiency**

- Processing time reduced from hours to 15-20 minutes through optimized batch processing
- Data freshness improved from daily batch updates to near-real-time processing capabilities

- **Query performance**

- Query performance was enhanced by 80% through intelligent partitioning and caching strategies
- Dashboard response times improved significantly supporting real-time decision-making

- **Resource optimization**

- Storage utilization optimized through compression and life cycle management
- Infrastructure requirements reduced by over 60% through intelligent resource utilization

Cost optimization results

Our approach demonstrates significant cost advantages over traditional multi-tool implementations:

- **Unified platform value:** Eliminated multiple tool licensing costs while achieving superior performance
- **Standard Edition efficiency**: Proved enterprise capabilities without premium pricing requirements
- **Infrastructure savings:** Intelligent resource utilization reduced infrastructure footprint significantly
- **Total cost of ownership:** Comprehensive cost reduction across licensing, infrastructure, and maintenance

Technical innovation highlights

- Unified platform, eliminating complex integration overhead between multiple specialized tools
-
- Standard Edition feature maximization, proving enterprise capabilities without premium pricing
-
- Future-ready architecture, supporting AI and machine learning expansion
- Scalable foundation, enabling seamless addition of new data sources

AI and Advanced Analytics Foundation

Predictive analytics capabilities

Building on our unified data platform, the architecture supports advanced AI capabilities:

- **Expense anomaly detection:** Identifying unusual spending patterns and policy violations
- **Clinical documentation analysis:** Extracting insights from unstructured clinical notes
- **Predictive equipment maintenance:** Anticipating maintenance needs from operational data
- **Patient risk scoring:** Multi-source data fusion for comprehensive risk assessment

MLflow integration

Model life cycle management

- Standard Edition MLflow capabilities provide experiment tracking and model versioning
- Automated model deployment pipelines ensure seamless production transition
- Built-in model monitoring tracks performance and triggers retraining workflows

Production integration

- Integration with existing data pipelines enables real-time scoring
- Batch inference capabilities support large-scale prediction requirements
- Unified platform eliminates model deployment complexity

Implementation Best Practices

Optimization strategies

Cluster configuration

- Batch size tuning at 1,000 records proved optimal for diverse data types and processing requirements
- Auto-scaling configuration with a 2-8 node range handled all workload variations efficiently
- Resource pooling across multiple workloads maximized cluster utilization

Storage management

- Delta Lake optimization commands maintained consistent performance as data volumes grew
- Intelligent partitioning by date and source system decreased query scanning overhead by 80%
- ZORDER operations reduced storage costs by 60% while improving query performance

Security and compliance implementation

Authentication and authorization

- OAuth-based authentication with Azure Key Vault ensures enterprise-grade security
- Row-level security implementation supports HIPAA compliance requirements
- Dynamic data masking protects sensitive information without impacting performance

Audit and compliance

- Audit logging captures all data access patterns for regulatory compliance
- Data lineage tracking simplifies compliance reporting and impact analysis
- End-to-end encryption protects sensitive healthcare information throughout the pipeline .

Multi-source integration strategy

Reliability and quality

- Standardized error handling and retry mechanisms across all data sources
- Incremental processing design minimized resource consumption while maintaining data freshness
- Automated data quality checks ensure consistency across source systems

Governance and management

- Unified metadata management simplified data governance and lineage tracking
- Modular architecture approach enabled seamless addition of new data sources
- Centralized monitoring and alerting set up across all integration workflows

Healthcare-Specific Advantages

Regulatory compliance

Databricks Standard Edition provides essential healthcare compliance capabilities:

- Data encryption at rest and in transit
- Access controls and audit logging
- Data lineage tracking for regulatory reporting
- Simplified compliance management compared to multiple-tool architectures

Scalability for healthcare growth

The platform architecture supports expansion to additional data sources:

- IoT medical devices and wearable technology
- External healthcare APIs and HIE connections
- Insurance systems and claims data
- Genomic data repositories and research databases

Cost-effective innovation

Standard Edition pricing enables healthcare organizations to:

- Access advanced analytics capabilities while maintaining budget constraints
- Invest in innovation rather than infrastructure overheads
- Scale capabilities as organizational needs grow
- Prove value before expanding to premium features

Conclusion

Our healthcare analytics implementation demonstrates that innovative architecture and intelligent optimization deliver enterprise-grade results while maintaining cost efficiency. Through Databricks Standard Edition with strategic optimization techniques, we achieved:

- **Significant cost savings** compared to traditional multi-tool approaches, eliminating multiple licensing costs and reducing infrastructure requirements.
- **Operational efficiency**, with data processing reduced from hours to minutes, replacing manual workflows with automated pipelines.
- **Unified analytics platform** that supports both traditional reporting and future AI initiatives from a single platform.
- **Scalable foundation**, ready for advanced healthcare analytics expansion while controlling costs and ensuring compliance.

The combination of multi-source integration, cost optimization, and future-ready architecture positions healthcare organizations for continued innovation. This approach proves that smart engineering makes advanced analytics accessible without compromising on performance or capabilities.

For healthcare organizations struggling with data silos and budget constraints, this blueprint demonstrates that comprehensive data modernization is achievable and cost-effective with the right platform strategy and optimization approach.

About the Implementation

This implementation showcases Databricks' power as a unified analytics platform, proving that the Standard Edition delivers enterprise-grade results when properly optimized. The significant cost savings and performance improvements make Databricks an excellent choice for healthcare data implementations.

For more information on implementing similar solutions for your healthcare organization, please contact Zensar's data analytics team.

This white paper is based on a real-world implementation by Zensar for a mid-sized healthcare organization. Results and specific metrics have been generalized to protect client confidentiality while maintaining technical accuracy.

zensar

An RPG Company

At Zensar, we're 'experience-led everything.' We are committed to conceptualizing, designing, engineering, marketing, and managing digital solutions and experiences for over 145+ leading enterprises. Using our 3Es of experience, engineering, and engagement, we harness the power of technology, creativity, and insight to deliver impact.

Part of the $4.8 billion RPG Group, we are headquartered in Pune, India. Our 10,000+ employees work across 30+ locations worldwide, including Milpitas, Seattle, Princeton, Cape Town, London, Zurich, Singapore, and Mexico City.

For more information, please contact: **info@zensar.com | www.zensar.com**