

TECH BLUEPRINT

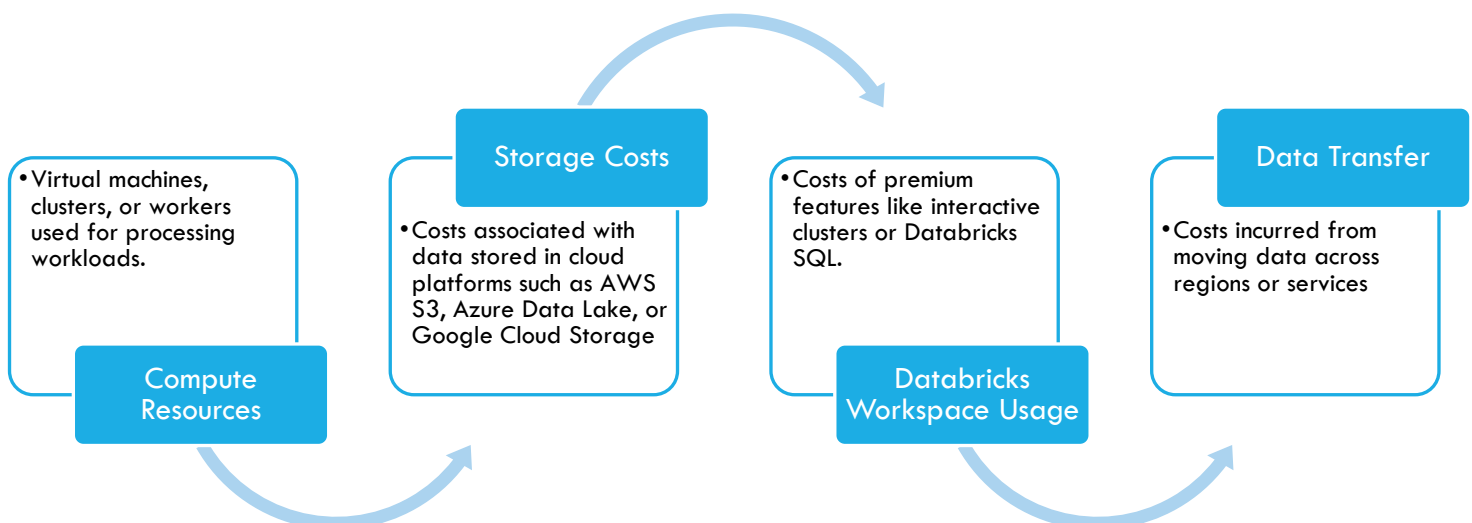
Databricks Cost Optimization Framework: Tailored for the Mortgage Industry

Introduction

Cost optimization is a vital aspect of managing workloads in a Databricks environment, especially in industries like mortgage, where large-scale, complex datasets such as loan origination, payment histories, credit evaluations, and property appraisals are managed. This white paper presents a comprehensive framework for cost optimization tailored specifically to the unique needs of the mortgage industry.

Understanding Cost Drivers in Databricks

Before diving into optimization, it is essential to understand the primary cost components:



Framework for Cost Optimization: Mortgage Industry Focus

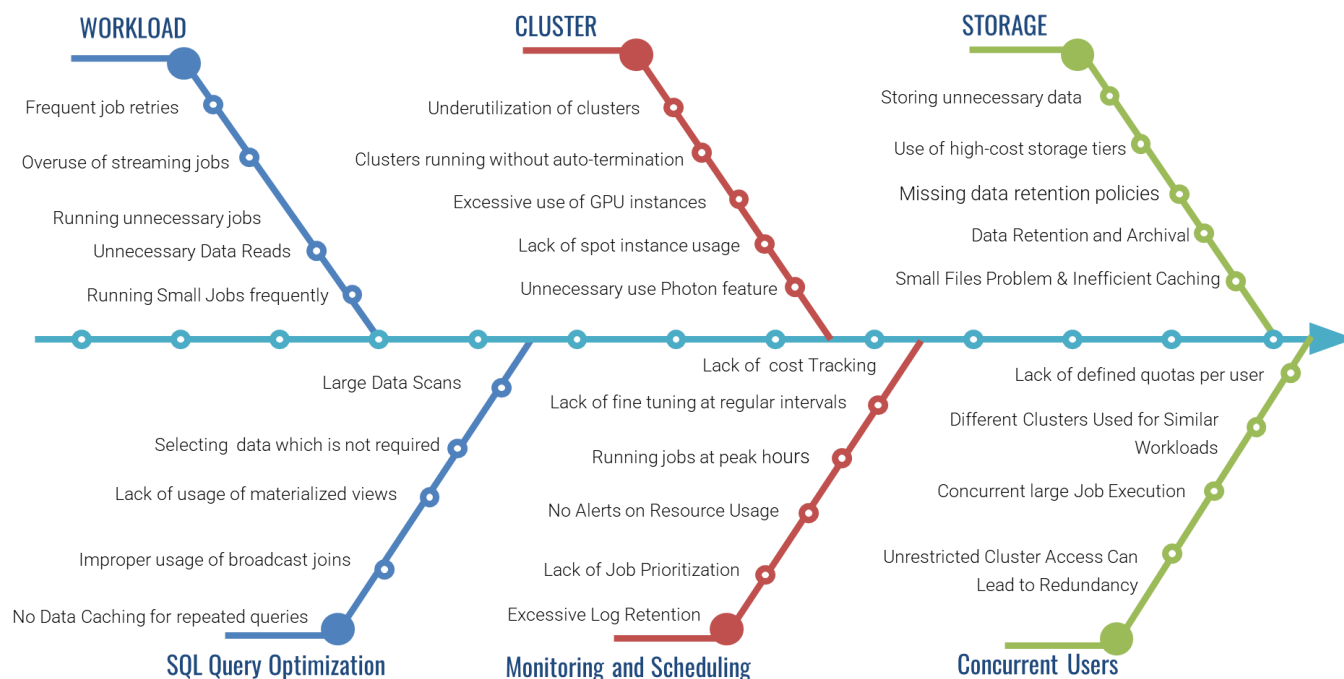
This framework is designed to address the unique challenges faced by mortgage companies, which deal with extensive datasets like loan origination, payment histories, and property appraisals. By focusing on these steps, organizations can achieve cost efficiency while ensuring high performance in their Databricks environment.

The framework is structured into six key steps, each tailored with examples from the mortgage industry:

Steps	Example Scenarios	Optimization(s)
Evaluate Cluster Utilization	Loan origination data (structured data) is ingested daily into a Delta Lake. Queries calculate aggregate metrics such as loan-to-value (LTV) ratios and risk scores.	• Configure auto-scaling clusters to handle ingestion and transformation workloads • Enable Photon runtime for faster SQL query execution and reduced costs
Optimize Workloads	Monthly batch jobs analyze historical delinquency trends in loan performance	• Schedule jobs during off-peak hours to leverage spot instance pricing • Refactor queries to minimize resource consumption • Cache intermediate results using Delta Cache to accelerate repetitive computations
Storage Optimization	Mortgage datasets contain millions of records grouped by state and year of origination	• Partition data by state and year to improve query performance • Implement data lifecycle policies to archive older data (e.g., property appraisal data older than 3 years) to cold storage
Delta Lake Compaction	Payment status updates and loan term modifications generate small files in Delta Lake	• Schedule nightly data compaction jobs to merge small files, improving read performance and reducing storage costs • Use Z-order clustering for frequent queries

		such as “Retrieve all loans with outstanding balances > \$1M in California
Monitor and Govern Usage	A mortgage company wants to track costs by department for different loan types (e.g., FHA, VA, Jumbo Loans)	• Implement resource tagging for team or project-level cost tracking, ensuring each department’s resource consumption is visible • Monitor usage metrics with Databricks’ native tools and set up alerts for anomalies using tools like Azure Monitor or AWS CloudWatch • Establish governance policies such as: ○ Enforcing maximum cluster lifespan to prevent long-running unused clusters. ○ Creating predefined cluster configurations to ensure resource consistency across teams ○ Mandating approval workflows for high-cost operations or clusters • Use third-party anomaly detection tools, such as CloudHealth, to automate alerts for unusual spikes in costs or usage patterns, helping identify potential misconfigurations or waste quickly
	A mortgage company wants to track costs by department for different loan types (e.g., FHA, VA, Jumbo Loans)	• Implement resource tagging for team or project-level cost tracking • Monitor usage metrics with Databricks’ native tools and set up alerts for anomalies • Establish governance policies to enforce optimal configurations and prevent misuse
Collaboration and Training	Analysts frequently re-run similar queries on loan origination and default prediction datasets	• Train teams to reuse existing notebooks and leverage collaborative features in Databricks • Use Databricks dashboards to share query results, avoiding redundant computations

General Checklist Items – Fishbone Analysis



Continuous Optimization Cycle

Optimization in a Databricks environment is not a one-time activity. Mortgage companies must adopt a continuous optimization cycle, assigning specific roles to ensure accountability.

Monthly Reviews

- Data engineers and IT administrators analyze billing data and resource utilization to identify inefficiencies.

Periodic Audits

- Governance teams and cost analysts perform audits to uncover cost-saving opportunities in workloads and storage.

Forecasting

- Business analysts use historical trends to predict future resource needs, ensuring proactive cost management

By involving these roles, organizations can ensure that the optimization process is actionable and aligned with operational goals.

Cost Optimization Workflow – Mortgage Industry Examples

The following blocks outlines a few use cases and action driven approach for optimizing workloads in a Databricks environment tailored for the mortgage industry, with examples of cost-saving strategies:

Mortgage Dataset Ingestion • Action: Load loan origination and payment data into Delta Lake • Cost-Saving Strategy: Use auto-scaling clusters to handle data ingestion efficiently, avoiding over-provisioning	Delta Table Storage • Action: Store datasets efficiently in Delta format • Cost-Saving Strategy: Partition data by relevant dimensions (e.g., state, year) to improve query performance and reduce compute costs	Cluster Utilization Review • Action: Optimize clusters for auto-scaling and right-sizing • Cost-Saving Strategy: Right-size clusters based on workload needs and enforce termination policies for idle clusters
Delta Compaction Job • Action: Schedule nightly jobs to merge small files • Cost-Saving Strategy: Reduce storage costs and improve query performance by compacting Delta Lake files regularly	Query Optimization • Action: Use Photon runtime, caching, and partitioning for frequent queries • Cost-Saving Strategy: Leverage Photon runtime for faster query execution and minimize redundant computations using Delta Cache	

Tools for Cost Monitoring

Companies can leverage the following tools to monitor and manage costs:

Databricks Cost Insights: Analyze workspace costs and usage trends. Integrate this tool into daily workflows by setting up automated reports and dashboards for regular team reviews.

Cloud-Specific Tools: AWS Cost Explorer, Azure Cost Management, or Google Cloud Billing. Use these tools to compare cost patterns across different cloud services and identify anomalies or optimization opportunities.

Third-Party Tools: Platforms like CloudHealth and Harness provide granular cost insights. Evaluate their features such as multi-cloud support and predictive analytics to determine the best fit for your organization's needs. Implement these tools as part of your governance framework to automate alerts and recommendations for cost-saving measures.

Conclusion

Cost optimization in Databricks for the mortgage industry involves a combination of technical best practices, governance, and continuous monitoring. By adopting this framework, organizations can manage resources effectively, reduce operational expenses, and maintain high performance. This tailored approach ensures that mortgage companies can scale efficiently while focusing on their core mission of delivering superior financial services.

About Kadel Labs

At Kadel Labs we have a dedicated practice for Databricks and Apache Spark Consulting. We specialize in providing cutting-edge data engineering solutions, helping businesses optimize their applications for performance, scalability, and reliability.

With our deep expertise in Data and AI engineering, we focus on improving performance across all levels of your data environment — from software logic to infrastructure. Our mission is to provide tailored solutions that reduce operational costs, enhance user experiences, and ensure your data ecosystem is built to scale with maximum efficiency.

Contact Us

Email: contact@kadellabs.com

Website: www.kadellabs.com