Informatica | databricks

Solution Brief

# Accelerate Data Engineering Pipelines for AI & Analytics

**Key Benefits**

- Accelerate innovation and increase business agility with successful AI projects
- Gain insights faster with reliable, high-volume ingestion of hybrid data
- Lower the costs of creating and maintaining data engineering pipelines at scale
- Accelerate model training by enabling data scientists to quickly discover the right datasets
- Increase productivity and do more with existing resources

## Build End-to-End Intelligent Data Pipelines With Databricks and Informatica

For businesses looking to gain competitive advantage with advanced analytics, artificial intelligence (AI) and machine learning (ML) projects represent a key priority. But despite the potential business impact of these technologies, industry research shows that very few AI projects are successful.[1] The reason? Writing ML code is just one small part of what goes into successful projects. For AI to successfully deliver predictive analytics, fraud analytics, and other high-value insights, a complex surrounding infrastructure is required.

Without fast access to accurate and prepared datasets, data teams are challenged to build accurate models. In addition, inaccurate or incomplete data can skew results and undermine confidence in AI and ML projects generally. What's needed is an end-to-end solution to ingest high volumes of hybrid data at high speed, create high volume data integration for fast and reliable processing, and provide data discovery and lineage across the enterprise. Informatica® and Databricks have partnered to provide a scalable data and machine learning solution with faster data discovery, ingestion, and preparation that accelerates development for machine learning and advanced analytics.

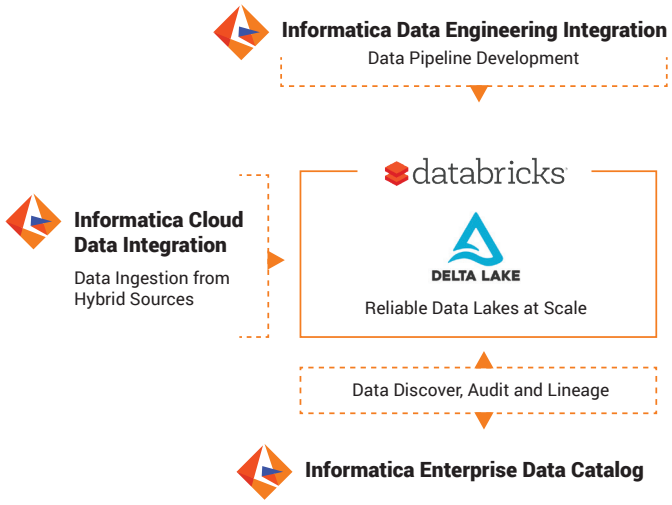


**Figure 1. The Informatica and Databricks joint solution accelerates data engineering pipelines for AI and analytics.**

[1] Databricks, 2018 Trend Report: Enterprise AI Adoption

## Key Capabilities

### Create High-Volume Data Pipelines at Scale

Accelerate data pipeline development with a no-code, visual development environment that increases developer productivity by up to 5x compared to hand-coding. Drag and drop pipelines created with Informatica Data Engineering Integration can be pushed down to Databricks for processing in an optimized Apache Spark implementation. Take advantage of hundreds of prebuilt transforms, connectors, and parsers, along with dynamic mapping and templates. With Data Engineering Integration, you can prepare the data for machine learning, improve productivity and future-proof against any open source technology changes with the ability to push onto any compute engine.

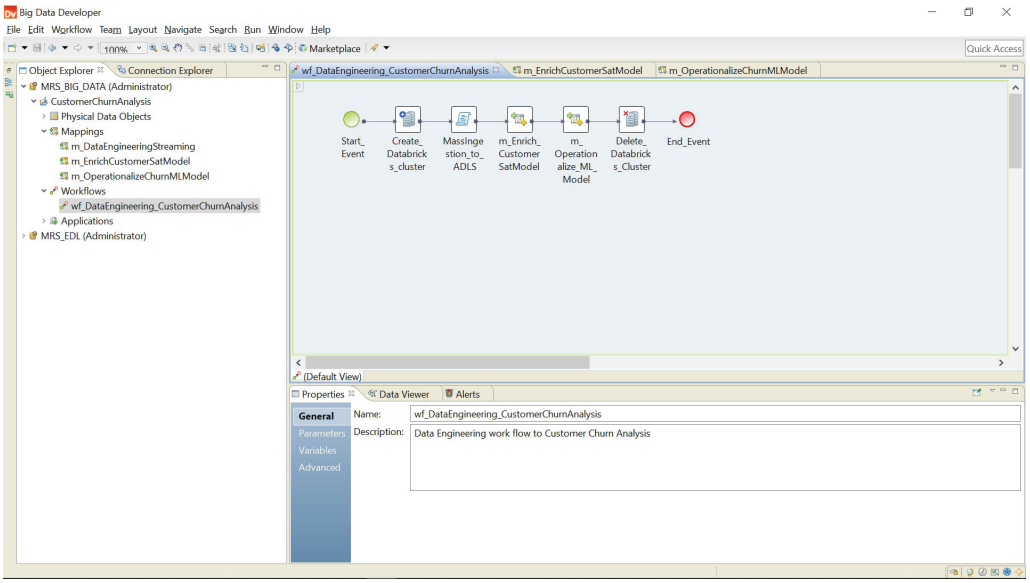


**Figure 2. Operationalize the end-to-end data pipeline.**

### Enable High-Speed Ingestion of Hybrid Data into a Managed Delta Lake

Data ingestion using Informatica Cloud Data Integration with Delta Lake enables ingestion of high volumes of data from multiple hybrid sources into a data lake, with Delta Lake providing high reliability and performance through schema enforcement, compaction, and other file enhancements. Build integrations faster with flow wizards and out-of-the-box templates. Comprehensive transformation capabilities to handle structured and unstructured data. Orchestrate your integration processes with advanced exception handling capabilities. Automate your integration processes and seamlessly integrate with continuous integration and continuous delivery systems.

### Enable Data Scientists to Discover, Classify, and Organize Data Assets Across the Enterprise

Gain complete visibility into how your data is moving through your data stack. With data lineage capabilities understand where your data comes from and how it is being used. Informatica Enterprise Data Catalog enables you to easily find and discover trusted data across your hybrid and multi-cloud environment. Explore holistic data relationships and drill down into end-to-end lineage and impact analysis. An integrated business glossary, crowd-sourced curation of business assets, and semantic search enable users to better understand data and rapidly use it.

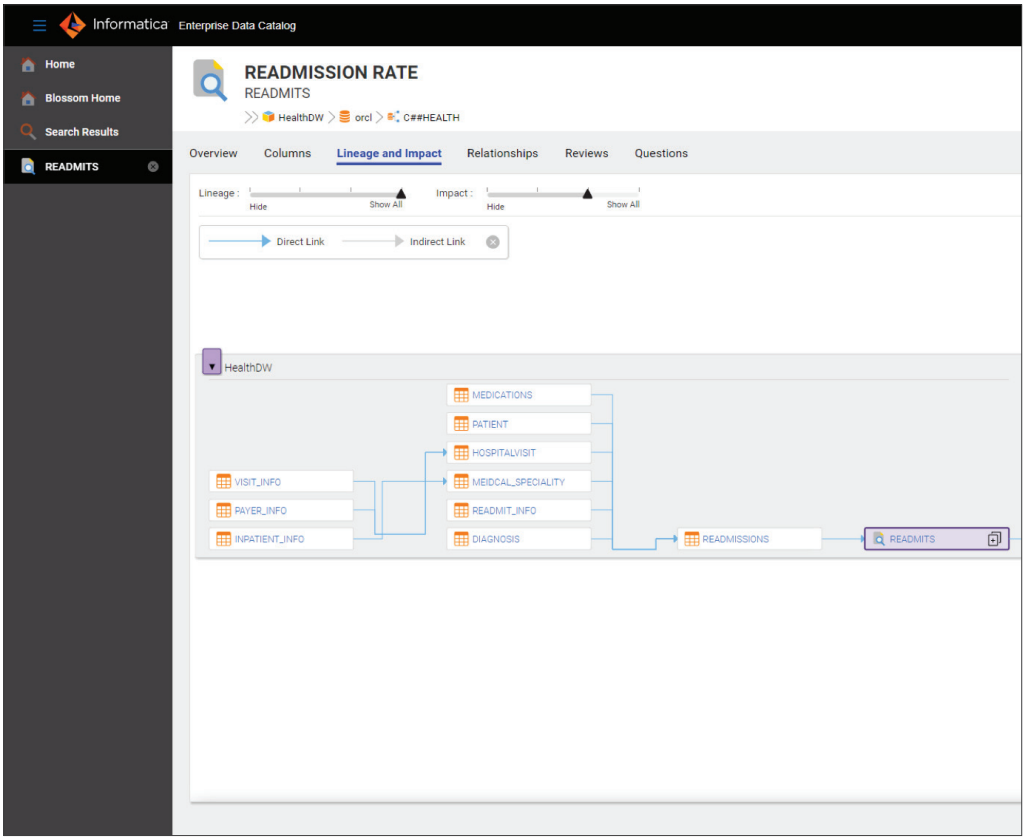


**Figure 3. Easily find and discover trusted data building machine learning models.**

### Optimize Data Lakes to Make Data Available for Machine Learning

The managed Delta Lake on Databricks optimizes data lakes to make massive amounts of data available for machine learning. By providing structure on top of your existing data lake, Delta Lake ensures that data is correct and in the desired format. It provides serializable isolation levels, so you never see inconsistent data. With ACID transactions and schema enforcement, it brings reliability at scale to data lakes and makes high-quality datasets ready for downstream analytics.

### Manage the End-to-End Machine Learning Lifecycle

The Databricks workspace manages the lifecycle for the machine learning process, enabling you to manage your experiments from development to production. With collaborative Databricks Notebooks, you have a workspace to create reusable AI and ML tools with support for multiple languages including R, Python, Scala, and SQL.

## About Informatica

Digital transformation changes expectations: better service, faster delivery, with less cost. Businesses must transform to stay relevant and data holds the answers.

As the world's leader in Enterprise Cloud Data Management, we're prepared to help you intelligently lead—in any sector, category, or niche. Informatica provides you with the foresight to become more agile, realize new growth opportunities, or create new inventions. With 100% focus on everything data, we offer the versatility needed to succeed.

We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption.

## Benefits

**Accelerate Innovation and Increase Business Agility With Successful AI Projects**

Investment in AI is increasing as organizations look to accelerate innovation and go to market faster with initiatives such as fraud detection and prediction, supply chain optimization, personalized digital experiences that increase customer engagement, improved patient outcomes, and more. The joint Informatica and Databricks solution enables organizations to build and iterate machine learning models faster to address rapid go-to-market demands.

**Gain Insights Faster With Reliable, High-Volume Ingestion of Hybrid Data**

Seamless integration between Databricks and Informatica enables data engineers to quickly ingest high volumes of data from multiple hybrid sources into a data lake with high reliability and performance.

**Lower the Costs of Creating and Maintaining Data Engineering Pipelines at Scale**

With an easy to use drag-and-drop user interface that pushes processing down to an optimized Apache Spark implementation in the cloud, customers experience faster and lower cost development of high-volume data pipelines.

**Enable Data Scientists to Discover the Right Datasets for Model Training**

With comprehensive data discovery, data scientists can quickly build more accurate models based on the right dataset. End-to-end lineage addresses compliance with GDPR and other regulations while enabling data scientists to verify lineage of the data used for model creation and analytics.

**Increase Productivity and Do More With Existing Resources**

The joint Informatica and Databricks solution enables you to increase productivity and leverage your existing infrastructure. A no-code visual design interface can increase developer productivity up to 5x as compared to hand-coding. Decrease overall total cost of ownership by optimizing infrastructure resources for AI and ML projects.

## Next Steps

Learn more about the Informatica and Databricks partnership at informatica.com/databricks and databricks.com/informatica.

Informatica

**Worldwide Headquarters** 2100 Seaport Blvd., Redwood City, CA 94063, USA Phone: 650.385.5000, Toll-free in the US: 1.800.653.3871

IN17_1019_03708