QLIK DATA INTEGRATION:

Deliver Analytics-Ready Data to Databricks Lakehouse in Real Time

Qlik databricks

Databricks: The lakehouse of choice for many organizations.

In under a decade, Databricks has emerged as a reliable, high-performing, and flexible solution for data storage. And for good reason. As the variety, volume, and velocity of data increase – and as demands on data teams keep intensifying – Databricks makes it possible to get more data to more users more quickly, leading to faster ROI and sharper competitive edge.

Among other features, Databricks enables you to:

- Quickly and reliably store tremendous volumes of data
- Store both structured and unstructured data, in a world where you increasingly need both
- Create a central repository – a single source of truth – for your enterprise data
- Scale up and down quickly as your needs evolve
- Make it faster and easier for data engineers to create and train AI and machine learning (ML) models
- Reduce the costs associated with data warehousing

“We took three key bets: 100% cloud, open source, and machine learning. I think we are today the largest commercial open-source-only vendor that’s independent. That has a lot to do with [the fact that] we were looking far into the future.”[1]

Ali Ghodsi
CEO, Databricks

Accelerating ROI in Databricks.

A data lake is foundational architecture for your data and analytics strategy, especially if it includes AI, ML, and data science initiatives. And to meet today's ever-increasing data demands, you can amp up the power of Databricks – and get to ROI faster – by automating various aspects of data management:

INGESTION

Naturally, you need to bring data into your data lake – and not just a few types of data but every type from every source you have, no matter how it's structured (or isn't). And because so many aspects of business are now real-time, it's imperative to ingest data continuously, as it changes. Traditional batch loading is too slow, and it can't handle today's massive volumes.

TRANSFORMATION

Once your data is in Databricks, it has to be made ready for use in AI, ML, data science, and analytics – immediately, if you plan to stay competitive. But data transformation is a complex undertaking. And even if you hire an army of engineers to perform those transformations manually (and who can afford that?), you'll be introducing countless opportunities for error.

CONSUMPTION

It's also critical to get data to users without involving IT. Users should have a simple way to find the data sets they need, combine them with other sets, and understand where the data came from. And everyone should be confident that the data is reliable – which means you need to keep it governed, permissioned, and secure all the way through the pipeline.

How many data engineers do you need?

"A common starting point is 2-3 data engineers for every data scientist. For some organizations with more complex data engineering requirements, this can be 4-5 data engineers per data scientist."[2]

Jesse Anderson, *O'Reilly*

What to look for in a data integration solution.

When you're ready to automate data lake creation, what should you look for in a platform? The ideal solution will have the following:

1. Universal, real-time data ingestion.

Look for a solution with change data capture (CDC) capabilities. You want the ability to ingest and continually update massive volumes of transactional data from virtually any source – databases, data warehouses, legacy mainframe systems, enterprise apps like SAP, and more – directly into the Databricks Unified Analytics Platform. So the data you use is always up-to-date, whether it's powering your AI, ML, data science, or analytics initiatives.

Features to look for:

Real-time change data capture – identifying and moving only the changes to data sets and metadata as they occur, and capturing source schema changes with ease and without rescripting

Enterprise-wide data ingest – with out-of-the-box connectivity to the widest variety of data sources, and support for loading into Delta Lake through cloud object storage on leading platforms such as Microsoft Azure, Amazon Web Services, and Google Cloud Platform

Centralized monitoring and control – with a single, fully automated console to design, execute, and monitor thousands of data replication tasks with no manual coding

Speed – including faster data replication with Databricks Lakehouse (Delta) endpoint connectors, which allow you to skip saving in intermediate formats

Cost efficiency – using lower-compute Databricks clusters (rather than regular interactive Databricks clusters) – to ingest a higher volume of data at the same cost

2. Analytics-ready data delivery.

You'll want to be able to automate the entire data pipeline – from ingestion to transformation to the creation of analytics-ready data sets – without any coding. By eliminating error-prone, labor-intensive, and time-consuming manual schema creation and ETL scripting processes, you'll reduce risk and accelerate your speed-to-insight.

Features to look for:

Data pipeline automation – from the generation of source system data streams all the way through to the creation of analytics-ready data sets

Automatic generation of transformations – using Apache Spark SQL or another high-performance data processing engine to deliver analytics-ready data sets into the Databricks Platform

Preparation and provisioning at scale – to IT users, who can quickly build scalable data pipelines, and to business users, if they have ad-hoc requirements for blending unmanaged data

3. Trusted, enterprise-grade data.

Search for a platform that builds a secure, fully-governed catalog of all your data – not just in Databricks but across all your enterprise sources – providing trustworthy, transparent data sets for all your initiatives. Data consumers should be able to generate insights in the BI tool of their choice within the Databricks Unified Analytics Platform.

Features to look for:

Smart, integrated data catalog – which uses technical, operational, and business metadata to organize, document, and describe all data in the collection. A modern online shopping experience should enable users to easily self-provision data with powerful search, preview, and selection capabilities

Data integrity and trust – with persisted change history for the entire “raw-to-ready” preparation process for end-to-end data lineage, while also supporting Databricks ACID capabilities

Security and governance – with enterprise-scale data access controls and data obfuscation capabilities to ensure that data is protected and secure, and without any intermediate data stores serving as staging areas

Introducing Qlik® for Databricks.

Qlik provides the only real-time, end-to-end, enterprise-class data ingestion and pipeline automation solution for the Databricks Lakehouse Platform – so you can accelerate your data-driven initiatives and get more ROI, faster.

The Qlik Data Integration platform delivers:

1. **Universal, real-time data ingestion direct to Databricks Delta.** Qlik automates the entire data pipeline by streaming real-time data, at scale, from virtually any source – databases, data warehouses, enterprise systems, and more – directly into the Databricks Unified Analytics Platform, without intervening layers.

2. **Analytics-ready data delivery.** Qlik automatically generates Spark SQL transformations, without manual coding, to deliver analytics-ready data in Delta Lake.

3. **Trusted, enterprise-grade data.** Qlik builds a secure, fully governed, self-service catalog for all data, not just in Databricks but across the enterprise – giving consumers an easily accessible data marketplace to find, understand, and use data.

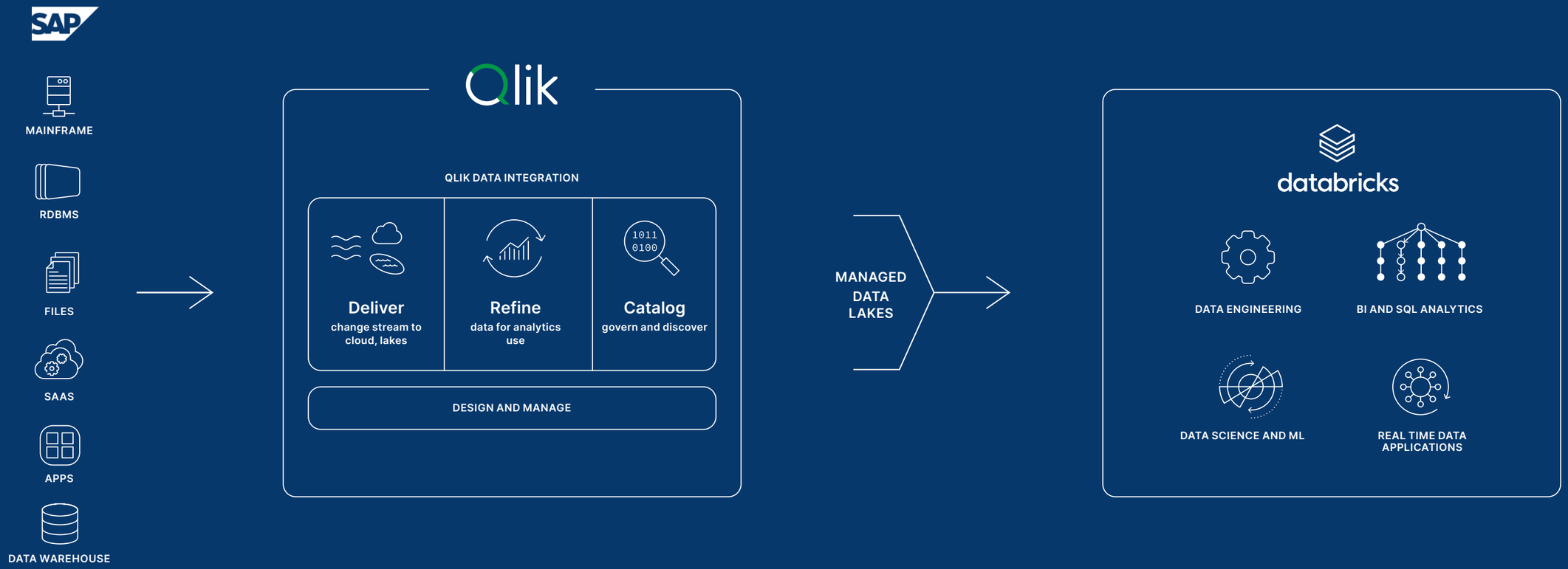
By abstracting the manual, error-prone processes of data modeling, ETL coding, and scripting, Qlik drives agility, reduces risk, and allows you to maximize your Databricks investment. It also provides point-to-point data replication – within an organization's managed infrastructure, without having to pass data through a third party – for improved security and control.

> **"Customers are constantly looking to expand the volume and type of data they can easily move from various sources, especially on the leading cloud platforms, to their data lakehouse. We're excited about the potential of Qlik's Databricks Delta Endpoint to seamlessly and efficiently deliver the data that customers need to drive more value from their investment in Databricks Lakehouse."**

Roger Murff
VP of ISV Partners, Databricks

Automate, orchestrate, and monitor end-to-end data pipelines.

No matter which data sources you need to ingest or which type of data initiatives you want to advance, Qlik makes it possible to for you to transform that data into action much faster and more reliably.



CUSTOMER SUCCESS

J. B. Hunt drives transportation forward.

With Qlik, the transportation logistics company provides analytics-ready data in near real-time to multiple user groups – without deploying an army of data engineers.

CHALLENGE

The implementation of Databricks data lake led to increasing pressure on the operational data stores that serve as the backbone for J.B. Hunt 360, the company's technology solution for shippers and carriers.

SOLUTION

Change data capture delivers near-real-time data from a variety of sources, including legacy mainframe systems and SQL server, directly into Delta Lake. The data is then automatically modeled and transformed for a cloud data warehouse.

OUTCOME

With more real-time data available and latency reduced to just minutes, the user experience has been greatly improved. The company's expertise in supply chain technology has grown, and they've been able to develop automated processes to improve efficiencies.

Get the whole story →

> **"We're seeing more real-time data in J.B. Hunt 360, which gives shippers and carriers up-to-the-minute information on how they are performing."**

Joe Spinelle
Director of Engineering and Technology

Get the most possible value from your Databricks data lake.

To build a high-performing data lake, you need a solution for handling the labor-intensive manual engineering tasks that have traditionally slowed data delivery to a crawl. In other words, you need automated data integration, transformation, and cataloging. And that's exactly what Qlik provides. When you use Qlik Data Integration with Databricks, you'll see:

Faster time-to-value – with real-time data ingestion directly into Delta Lake. Boost your ROI by pumping analytics-ready data into your AI, ML, data science, and analytics platforms.

Higher efficiency and productivity – by serving reliable, governed, analytics-ready data faster and at scale. Automate labor-intensive tasks to free your data engineers while making data readily available for your data consumers.

Lower risk – Deliver trusted data with end-to-end lineage, ACID compliance, and code-free refinement. Ensure enterprise-grade security and governance for all data.

Increased flexibility and agility – Deploy in any cloud configuration. Integrate with ever-growing types of data sources, targets, and platforms. And consume data in the analytics tools of your choice.

Ready to see for yourself?

Test-Drive Qlik for Databricks | **Request a Demo** | **Learn More**

About Qlik

Qlik transforms complex data landscapes into actionable insights, driving strategic business outcomes. Serving over 40,000 global customers, our portfolio leverages advanced, enterprise-grade AI/ML and pervasive data quality. We excel in data integration and governance, offering comprehensive solutions that work with diverse data sources. Intuitive and real-time analytics from Qlik uncover hidden patterns, empowering teams to address complex challenges and seize new opportunities. Our AI/ML tools, both practical and scalable, lead to better decisions, faster. As strategic partners, our platform-agnostic technology and expertise make our customers more competitive.

qlik.com

About Databricks

Databricks is the data and AI company. More than 7,000 organizations worldwide – including Comcast, Condé Nast, H&M, and over 40% of the Fortune 500 – rely on the Databricks Lakehouse Platform to unify their data, analytics, and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Delta Lake, Apache Spark™, and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems.

databricks.com

[1] Sharma, Shubham, "Databricks vs. Snowflake: The race to build [a] one-stop shop for your data," VentureBeat, 5/29/22, https://venturebeat.com/2022/03/29/databricks-vs-snowflake-the-race-to-build-one-stop-shop-for-your-data/.

[2] Anderson, Jesse, "Data Engineers versus Data Scientists," O'Reilly, April 2018, https://www.oreilly.com/radar/data-engineers-vs-data-scientists/.

2-WEB-21