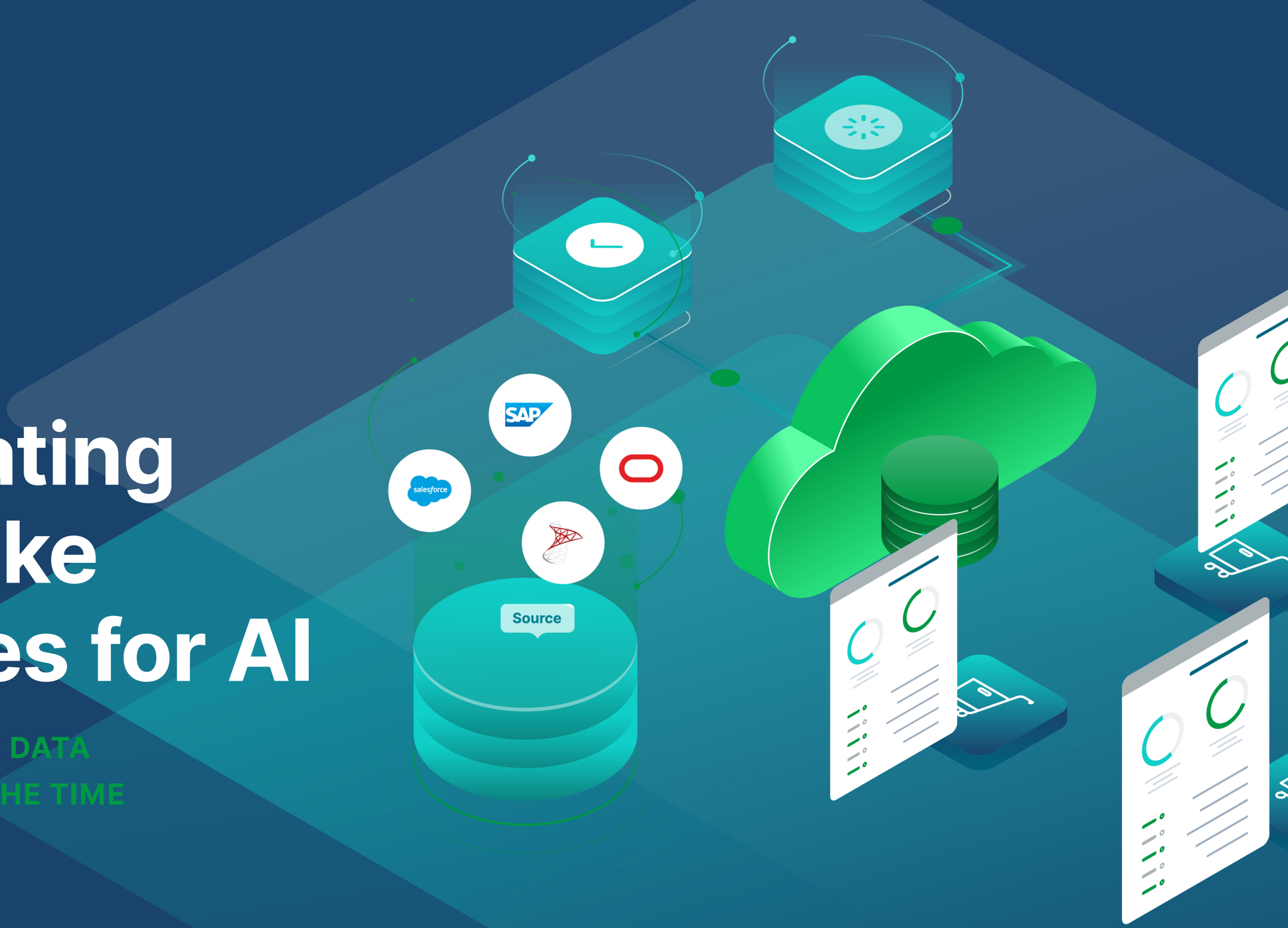
Qlik

Automating Data Lake Pipelines for AI

DELIVER ACCURATE DATA
IN A FRACTION OF THE TIME



The data delivery challenge for AI/ML

Every day, AI becomes more critical to business operations, innovation, and competitive advantage. And every day, data engineers face increasing demands to deliver accurate, model-ready data to AI initiatives as quickly as possible. Rising compliance fines and poor data quality create a double burden — outdated, inaccurate data can cost firms millions annually while data scientists devote vast majority of their time to manual cleaning, not insights.

It's a tall order. To enable AI, you need to prepare, transform, and integrate a wide variety of data types into a high-performing data lake. To do that, traditional data integration tools require time-consuming, labor-intensive coding by highly skilled staff. These tools also often lack the capabilities to fully prepare, cleanse, tag, and catalog the data. What's the answer? Automated data pipelines driven by AI and ML.

In this e-book, you'll learn how to build AI- and ML-enabled data pipelines for delivering AI-ready data.

The two types of data that feed AI

Just as there are many different types of AI — from speech and image recognition to machine learning, deep learning, and generative AI — there are many types of data formats required to power it. Broadly speaking, these data types fall into two categories:

STRUCTURED DATA

Clean, structured data enables models like neural networks to learn robust representations. Thanks to consistent tables with rows and columns, models can understand relationships and features in the data. To facilitate learning, preprocessing (through cleaning, normalization, and transformation) prepares structured data and makes features consistent, complete, and accurate.

UNSTRUCTURED DATA

Text, images, and video contain valuable real-world signals, but AI can't make use of these types of data in their raw form. They require preprocessing — for example, adding metadata tags and labels to extract useful features.

- ☑ Text has to be cleaned, tokenized, and encoded
- ☑ Images require labeling, normalization, and tensor transformation
- ☑ Videos need frame and audio extraction

Though it's a complex undertaking, when unstructured data is preprocessed properly, it augments AI models with nuanced, real-world training data — which enables more sophisticated AI use cases.

Key considerations when you're building data lakes for AI

Regardless of which kind of AI or use case you're enabling, data lakes have been embraced as the most effective foundation for AI data storage. Here's what to consider when you're building a data lake for AI:

STORAGE

Data lakes require highly scalable storage for diverse data. Optimal storage types for providing unlimited scalability, flexibility, and performance include S3, HDFS, and NoSQL databases. An additional layer such as a Delta Lake can enable ACID transactions, scalable metadata handling, and data versioning.

COMPUTE

Analyzing massive datasets in data lakes requires distributed computing engines like Spark, Hadoop, and SQL. These engines provide scalable in-memory processing across the enterprise.

QUALITY

Data lakes should integrate robust data profiling, validation, and observability to measure quality. They should enable collaborative curation workflows for trustworthy data prep and provide end-to-end lineage tracking.

How automation can help

When you're building data pipelines to feed AI, you can deliver data far more quickly, accurately, and efficiently with automation. Current capabilities include:

MACHINE LEARNING

- ✓ **Metadata and schema inference:** ML techniques like clustering and rule mining can automatically analyze datasets and derive metadata (e.g., data types and relationships).

  Role of the data engineer: Review the derived schema

- ✓ **Automated data validation:** Machine learning models can be trained to detect anomalies and errors in datasets (for example, records with outliers or implausible values).

  Role of the data engineer: Review/correct flagged rows rather than checking everything

---

- ✓ **Automated data labeling:** You can use ML to automatically generate labels for image, text, and other datasets.

  Role of the data engineer: Provide the model with a description of the classes

GENERATIVE AI

- ✓ **Writing analytical SQL queries:** ML models can enable more self-service analytics for the end user. Currently, these models provide boilerplate code but struggle with optimization and accuracy.

  Role of the data engineer: Promote less user hand-holding with reduced coding

---

- ✓ **Feature engineering:** ML models can suggest creative ways to transform or combine variables into new features.

  Role of the data engineer: Evaluate the feasibility of generated suggestions

---

- ✓ **Logic-response sequences:** You can chain multiple generative AI models to follow predefined prompts to extract data from sources, transform data, load it, and summarize in destinations for analytics consumption.

  Role of the data engineer: Chain the models to orchestrate more complex operations

How does automation work across the data pipeline?

How, specifically, do ML and generative AI enable you to automate and accelerate data delivery? Here are the details across some key categories:

INGESTION

During ingestion, ML detects anomalies to identify errors and inconsistencies. Over time, reinforcement learning optimizes and improves ingestion pipelines, shifting rigid ETL to flexible, self-optimizing models that understand data the way humans do. ML techniques like natural language processing (NLP) and computer vision interpret unstructured data, automatically identifying structures, relationships, and schemas as new data is ingested.

QUALITY

ML models — e.g., classification algorithms and NLP — identify data quality issues like anomalies, errors, inconsistencies, bias, and completeness. Reinforcement learning optimizes data validation, correction, and cleansing procedures over time for increased automation. ML tools analyze datasets to detect constraints, patterns, and relationships, enabling more robust profiling. With continuous feedback, the tools perform better than static rules, recommending mitigation strategies for quality issues.

PROCESSING

ML algorithms can detect data relationships and patterns to automate optimal mapping and transformation logic for ETL processes. Neural networks can learn complex transformations. Natural language interfaces and pre-trained models like transformers encapsulate common data prep tasks, increasing reuse and automation. With continuous training on new data, ML models adapt and improve their data manipulation abilities over time versus relying on rigid rules.

STORING

ML models can analyze data characteristics, usage patterns, and constraints to automatically determine optimal storage formats, compression, partitioning, and layouts. ML enables the auto-optimization of data storage by continuously monitoring workloads, adaptively modifying data storage, and indexing as usage changes over time.

DISCOVERY

ML models can crawl repositories and catalogs to identify new datasets as they're transformed, automatically classifying and cataloging them based on an analysis of metadata, schemas, and samples. This enables the auto-discovery of new data, relieving analysts of manual documentation and search. ML can also recommend related datasets to users by identifying similar data profiles across sources, ensuring that transformed data is easily discoverable without manual effort.

Observing and improving the pipeline

Once you've built automation into your data pipeline, you can optimize for accuracy, speed, and intelligence by observing and tuning over time.

LOG, MONITOR, AND ALERT

Oversee the pipelines and models in production with comprehensive logging, monitoring, and alerting systems.

TRACE DATA LINEAGE FROM SOURCE TO ML

To support auditing and explainability, enable tracing with the robust logging of key events throughout the data flow.

QUANTIFY DATA QUALITY, MODEL ACCURACY, AND SYSTEM PERFORMANCE

Rigorously monitor data and model quality metrics. When you track performance indicators for all data infrastructure and measure model accuracy on live input data, you can verify that everything in your system is operating as intended.

ANALYZE TELEMETRY FOR CONTINUOUS IMPROVEMENT

Combining telemetry data enables you to detect anomalies and be alerted if critical issues arise. Over time, instrumentation and measurement data are invaluable for continuously analyzing and improving the ML system.

Checklist for success

When you're ready to begin automating your AI data pipeline, follow these guidelines:

1

SELECT AND PREPARE THE DATA

- ✓ Leverage both structured and unstructured data to train models efficiently.
- ✓ Preprocess to extract nuanced, real-world signals and augment model capabilities.
- ✓ Carefully transform all data into clean, consistent, model-ready formats.

2

BUILD AN OPTIMIZED DATA LAKE

- ✓ Build scalable data lakes on cloud object stores using distributed engines like Spark.
- ✓ Choose compute to match workloads, integrating with storage and using in-memory processing.
- ✓ Integrate robust, automated data quality and governance that enables trust.

3

USE THE RIGHT TYPE OF AI

- ✓ Use ML for metadata inference, data validation, and labeling to accelerate dataset documentation and quality checks — but make sure to manually review outputs.
- ✓ Chain generative models to produce and automate ETL sequences — and manually review key steps.

4

CHOOSE THE KEY CAPABILITIES TO AUTOMATE

- ✓ Apply ML to bring automation and intelligence to key areas such as adaptive ingestion, quality checks, processing, and storage optimization while discovering new data sources.
- ✓ Closely monitor ML systems for correct functioning, manually evaluating output quality before full deployment.

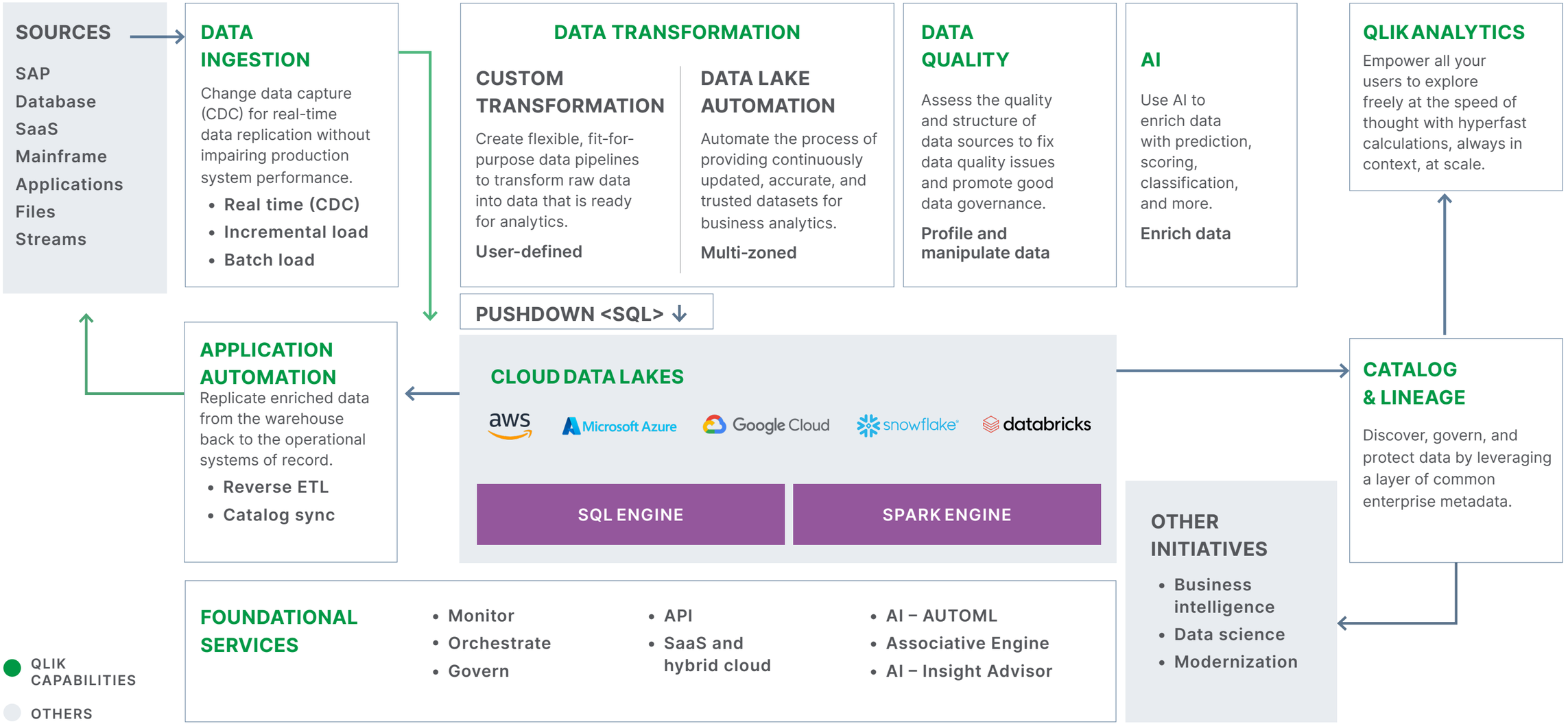
5

CONTINUOUSLY TEST YOUR PIPELINE

- ✓ Implement comprehensive logging, monitoring, and alerting to enable full traceability.
- ✓ Quantify data and model quality, assess system performance, and enable continuous analysis to improve reliability and build trust in end-to-end ML pipelines and models.

End-to-end data pipeline automation with Qlik®

Qlik offers a series of integrated solutions for automating your entire data lake pipeline — from real-time ingestion to processing and refinement — without requiring you to write code.



Accelerate data delivery to AI/ML

Qlik Cloud® Data Integration automates your entire AI data pipeline, from real-time ingestion to the creation and provisioning of AI-ready data. The solution:

- ☑ Supports virtually all industry standard data sources and targets
- ☑ Profiles, cleans, and catalogs all the data in your data lake
- ☑ Maintains end-to-end lineage to ensure data confidence

As a result, your engineers can finally meet the growing demands for AI-ready datasets in real time with confidence.

Ready to see how Qlik can automate your data lake pipelines for AI?

Take the Self-Guided Tour

Qlik

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

24-DI-0001-4-WHM