



| WHITE PAPER

Speed Your Data Lake ROI

Five Principles for Effectively Managing
Your Data Lake Pipeline

Introduction

Being able to analyze high-volume, varied datasets is essential in nearly all industries. From fraud detection and real-time customer offers to market trend and pricing analysis, analytics use cases are boosting competitive advantage. In addition, the advent of the Internet of Things (IoT) and Artificial Intelligence (AI) are also driving up the volume and variety of data that organizations like yours want and need to analyze. The challenge: as the speed of business accelerates, data has increasingly perishable value. The solution: real-time data analysis.

Data lakes have emerged as an efficient and scalable platform for IT organizations to harness all types of data and enable analytics for data scientists, analysts, and decision makers. But challenges remain. It's been too hard to realize the expected returns on data lake investments, due to several key challenges in the data integration process ranging from traditional processes that are unable to adapt to changing platforms and data transfer bottlenecks to cumbersome manual scripting, lack of scalability, and the inability to quickly and easily extract source data.

Qlik®, which includes the former Attunity data integration portfolio, helps your enterprise overcome these obstacles with fully automated, high-performance, scalable, and universal data integration software.

Evolution of the Data Lake

Combining efficient distributed processing with cost-effective storage for mixed data sets analysis forever redefined the economics and possibilities of analytics. Data lakes were initially built on three pillars: the Hadoop foundation of MapReduce batch processing, the Hadoop Distributed File System (HDFS), and a “schema on read” approach that does not structure data until it’s analyzed. These pillars are evolving:

- The Apache ecosystem now includes new real-time processing engines such as Spark to complement MapReduce.
- The cloud is fast becoming the preferred platform for data lakes. For example, the Amazon S3 distributed object-based file store is being widely adopted as a more elastic, manageable, and cost-effective alternative to HDFS. It integrates with most other components of the Apache Hadoop stack, including MapReduce and Spark. The Azure Data Lake Store (ADLS) is also gaining traction as a cloud- based data lake option based on HDFS.
- Enterprises are adopting SQL-like technologies on top of data stores to support historical or near- real time analytics. This replaces the initial “schema on read” approach of Hadoop with the “schema on write” approach typically applied to traditional data warehouses.

While the pillars are evolving, the fundamental premise of the data lake remains the same: organizations can benefit from collecting, managing, and analyzing multi-sourced data on distributed commodity storage and processing resources.

Requirements and Challenges

As deployments proceed at enterprises across the globe, IT departments face consistent challenges when it comes to data integration. According to the TDWI survey (Data Lakes: Purposes, Practices, Patterns and Platforms), close to one third (32%) of respondents were concerned about their lack of data integration tools and related Hadoop programming skills.

Traditional data integration software tools are challenging, too, because they were designed last century for databases and data warehouses. They weren't architected to meet the high-volume, real-time ingestion requirements of data lake, streaming, and cloud platforms. Many of these tools also use intrusive replication methods to capture transactional data, impacting production source workloads.

Often, these limitations lead to rollouts being delayed and analysts forced to work with stale and/or insufficient datasets. Organizations struggle to realize a return on their data lake investment. Join the most successful IT organizations in addressing these common data lake challenges by adopting the following five core architectural principles.

Five Principles of Data Lake Pipeline Management

1. Plan on Changing Plans

Your architecture, which likely will include more than one data lake, must adapt to changing requirements. For example, a data lake might start out on premises and then be moved to the cloud or a hybrid environment. Alternatively, the data lake might need to run on Amazon Web Services, Microsoft Azure, or Google platforms to complement on-premises components.

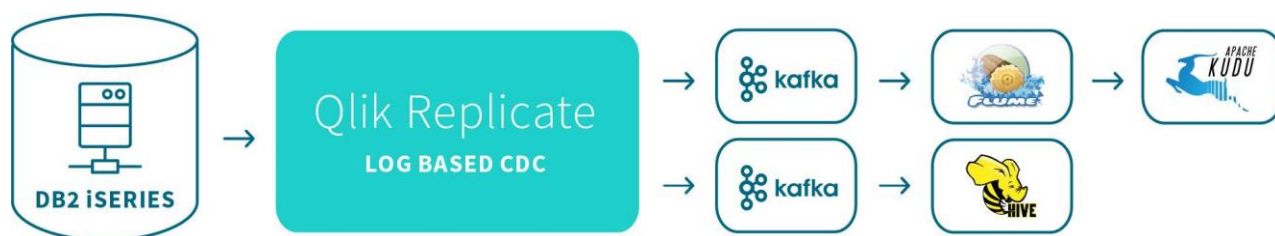
To best handle constantly changing architectural options, you and your IT staff need platform flexibility. You need to be able to change sources and targets without a major retrofit of replication processes.

Qlik Replicate™ (formerly Attunity Replicate) meets these requirements with a 100% automated process for ingesting data from any major source (e.g., database, data warehouse, legacy/mainframe, etc.) into any major data lake based on HDFS or S3. Your DBAs and data

architects can easily configure, manage, and monitor bulk or real-time data flows across all these environments.

You and your team also can publish live database transactions to messaging platforms such as Kafka, which often serves as a channel to data lakes and other Big Data targets. Whatever your source or target, our Qlik Replicate solution provides the same drag-and-drop configuration process for data movement, with no need for ETL programming expertise.

Two Potential Data Pipelines — One CDC Solution



2. Architect for Data in Motion

For data lakes to support real-time analytics, your data ingestion capability must be designed to recognize different data types and multiple service-level agreements (SLAs). Some data might only require batch or microbatch processing, while other data requires stream processing tools or frameworks (i.e., to analyze data in motion). To support the complete range, your system must be designed to support technologies such as Apache Kafka, Amazon Kinesis, Azure Event Hubs, and Google Cloud Pub/Sub as needed.

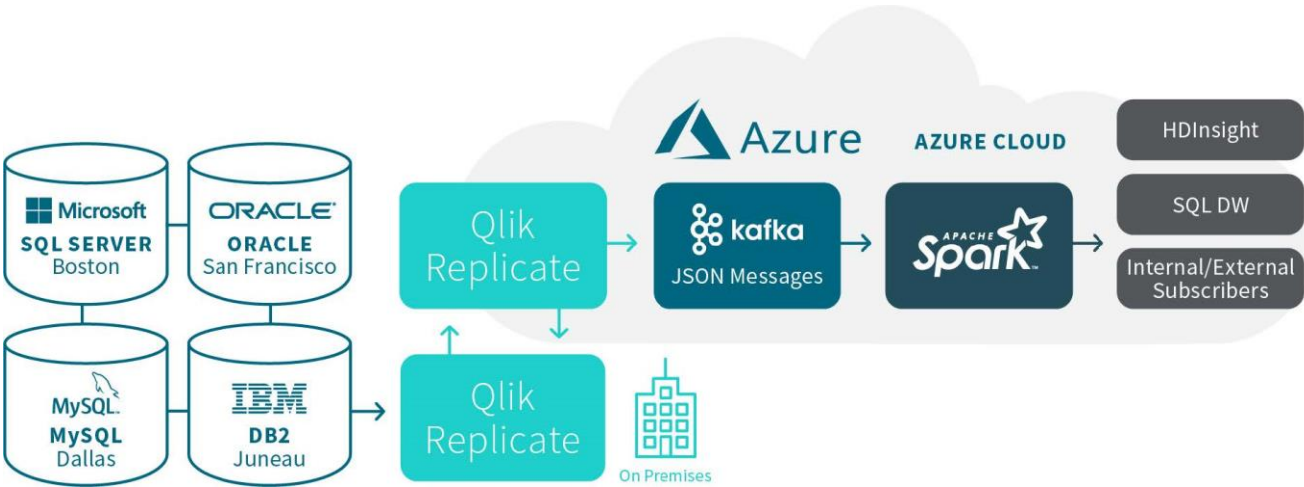
Additionally, you'll need a system that ensures all replicated data can be moved securely, especially when sensitive data is being moved to a cloud-based data lake. Robust encryption and security controls are critical to meet regulatory compliance, company policy, and end-user security requirements.

Qlik Replicate CDC technology non-disruptively copies source transactions and sends them at near-zero latency to any of the real-time/messaging platforms listed above. Using log reader technology, it copies source updates from database transaction logs – minimizing impact on production workloads – and publishes them as a continuous message stream. Source DDL/schema changes are injected into this stream to ensure analytics workloads are fully aligned

with source structures. Authorized people also can transfer data securely and at high speed across the wide-area network (WAN) to cloud-based data lakes, leveraging AES-256 encryption and dynamic multipathing.

As an example, a US private equity and venture capital firm built a data lake to consolidate and analyze operational metrics from its portfolio companies. This firm opted to host its data lake in the Microsoft Azure cloud rather than taking on the administrative burden of an on-premises infrastructure. Qlik Replicate CDC captures updates and DDL changes from source databases (Oracle, SQL Server, MySQL, and DB2) at four locations in the US. Qlik Replicate then sends that data through an encrypted File Channel connection over a WAN to a virtual machine-based instance of Qlik Replicate in Azure cloud.

This Qlik Replicate instance publishes the data updates to a Kafka message broker that relays those messages in the JSON format to Spark. The Spark platform prepares the data in micro-batches to be consumed by the HDInsight data lake, SQL data warehouse, and various other internal and external subscribers. These targets subscribe to topics that are categorized by source tables. With the CDC-based architecture, this firm is now efficiently supporting real-time analysis without affecting production operations.



3. Architect for Data in Motion

Your data lake runs the risk of becoming a muddy swamp if there is no easy way for your users to access and analyze its contents. Applying technologies like Hive on top of Hadoop helps to provide an SQL-like query language supported by virtually all analytics tools. Organizations like yours often need both an operational data store (ODS) for up-to-date business intelligence (BI) and reporting as well as a comprehensive historical data store (HDS) for advanced analytics. This requires thinking about the best approach to building and managing these stores to deliver the agility the business needs.

This is more easily said than done. Once data is ingested and landed in Hadoop, often IT still struggles to create usable analytics data stores. Traditional methods require Hadoop-savvy ETL programmers to manually code the various steps – including data transformation, the creation of Hive SQL structures, and reconciliation of data insertions, updates, and deletions to avoid locking and disrupting users. The administrative burden of ensuring data is accurate and consistent can delay and even kill analytics projects.

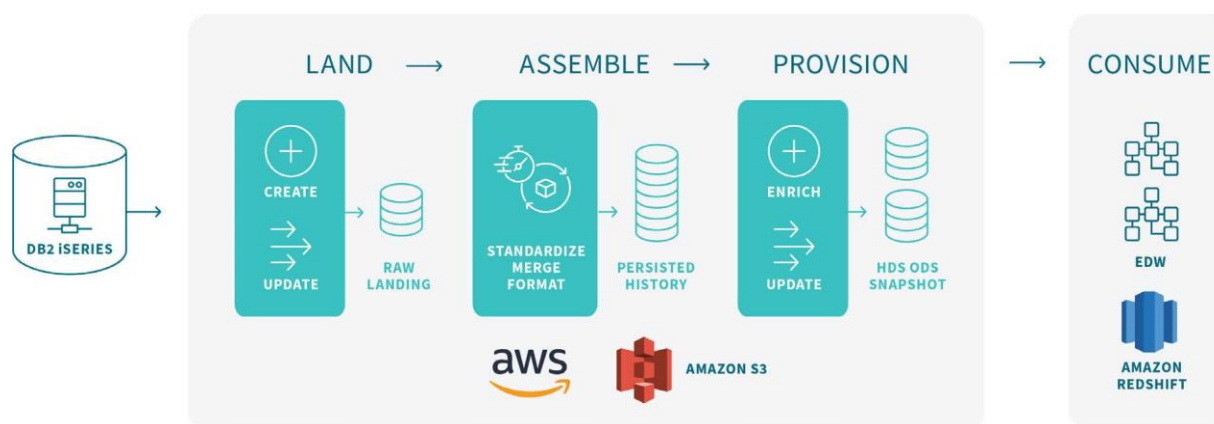
Qlik Compose™ for Data Lakes (formerly Attunity Compose for Data Lakes) solves these problems by automating the creation and loading of Hadoop data structures, as well as updating and transforming enterprise data within the data store. You, your architects, or DBAs can automate the pipeline of BI ready data into Hadoop, creating both an ODS and HDS. Because our solution leverages the latest innovations in Hadoop such as the new ACID Merge SQL capabilities, available today in Apache Hive you can automatically and efficiently process data insertions, updates, and deletions. Qlik Replicate integrates with Qlik Compose for Data Lakes to simplify and accelerate your data ingestion, data landing, SQL schema creation, data transformation, and ODS and HDS creation/updates.

As an example of effective data structuring, Qlik works with a major provider of services to the automotive industry to more efficiently feed and transform data in a multi-zone data lake pipeline. The firm's data is extracted from DB2 iSeries and then landed as raw deltas in an Amazon S3-based data lake. In the next S3 zone, tables are assembled (i.e., cleansed and merged) with a full persisted history available to identify potential errors and/or rewind, if necessary. Next these tables are provisioned/presented via point-in-time snapshots, ODS, and comprehensive change histories. Finally, analysts consume the data through an Amazon Redshift data warehouse. In this

case, the data lake pipeline transformed the data while structured data warehouses perform the actual analysis. The firm is automating each step in the process.

A key takeaway here is that the most successful enterprises automate the deployment and continuous updates of multiple data zones to reduce time, labor, and costs. Consider the skill sets of your IT team, estimate the resources required, and develop a plan to either fully staff your project or use a technology that can reduce anticipated skill and resource requirements without compromising your ability to deliver.

Automating the Data Lake Pipeline



4. Architect for Data in Motion

Your data management processes should minimize production impact and increase efficiency as your data volumes and supporting infrastructure grow. Quantities of hundreds or thousands of data sources affect implementation time, development resources, ingestion patterns (e.g., full data sets versus incremental updates), the IT environment, maintainability, operations, management, governance, and control.

Here again organizations find automation reduces time and staff requirements, enabling staff to efficiently manage ever- growing environments. Best practices include implementing an efficient ingestion process, eliminating the need for software agents on each source system, and centralizing management of sources, targets, and replication tasks across the enterprise.

With Qlik Replicate, your organization can scale to efficiently manage data flows across the world's largest enterprise environments. Our zero-footprint architecture eliminates the need to install, manage, and update disruptive agents on sources or targets. In addition, Qlik Enterprise Manager™ (formerly Attunity Enterprise Manager) is an intuitive and fully automated, single console to configure, execute, monitor, and optimize thousands of replication tasks across hundreds of end points. You can track key performance indicators (KPIs) in real time and over time to troubleshoot issues, smooth performance, and plan the capacity of Qlik Replicate servers. The result: the highest levels of efficiency and scale.

5. Depth matters

Whenever possible, your organization should consider adopting specialized technologies to integrate data from mainframe, SAP, cloud, and other complex environments. Here's why:

Enabling analytics with SAP-sourced data on external platforms requires decoding data from SAP pooled and clustered tables and enabling business use on a common data model. Cloud migrations require advanced performance and data encryption over WANs.

And deep integration with mainframe sources is needed to offload data and queries with sufficient performance. Data architects have to take these and other platform complexities into account when planning data lake integration projects.

Qlik Replicate provides comprehensive and deep integration with all traditional and legacy production systems, including Oracle, SAP, DB2 z/ OS, DB2 iSeries, IMS, and VSAM. Our company has invested decades of engineering to be able to easily and non-disruptively extract and decode transactional data, either in bulk or real time, for analytics on any major external platform.

When decision makers at an international food industry leader needed a current view and continuous integration of production-capacity data, customer orders, and purchase orders to efficiently process, distribute, and sell tens of millions of chickens each week, they turned to Qlik. The company had struggled to bring together its large datasets, which were distributed across several acquisition-related silos within SAP Enterprise Resource Planning (ERP) applications. The company relied on slow data extraction and decoding processes that were unable to match orders and production line-item data fast enough, snarling plant operational scheduling and preventing sales teams from filing accurate daily reports.

The global food company converted to a new Hadoop Data Lake based on the Hortonworks Data Platform and Qlik Replicate. It now uses our SAP-certified software to efficiently copy SAP record changes every five seconds, decoding that data from complex source SAP pool and cluster tables. Qlik Replicate injects this data stream – along with any changes to the source metadata and DDL changes – to a Kafka message queue that feeds HDFS and HBase consumers subscribing to the relevant message topics (one topic per source table).

Once the data arrives in HDFS and HBase, Spark in-memory processing helps match orders to production on a real-time basis and maintain referential integrity for purchase order tables within HBase and Hive. The company has accelerated sales and product delivery with accurate real-time operational reporting. Now, it operates more efficiently and more profitably because it unlocked data from complex SAP source structures.

The global food company converted to a new Hadoop Data Lake based on the Hortonworks Data Platform and Qlik Replicate. It now uses our SAP-certified software to efficiently copy SAP record changes every five seconds, decoding that data from complex source SAP pool and cluster tables. Qlik Replicate injects this data stream – along with any changes to the source metadata and DDL changes – to a Kafka message queue that feeds HDFS and HBase consumers subscribing to the relevant message topics (one topic per source table).

Once the data arrives in HDFS and HBase, Spark in-memory processing helps match orders to production on a real-time basis and maintain referential integrity for purchase order tables within HBase and Hive. The company has accelerated sales and product delivery with accurate real-time operational reporting. Now, it operates more efficiently and more profitably because it unlocked data from complex SAP source structures.

Streaming Data to a Cloud-based Data Lake and Data Warehouse

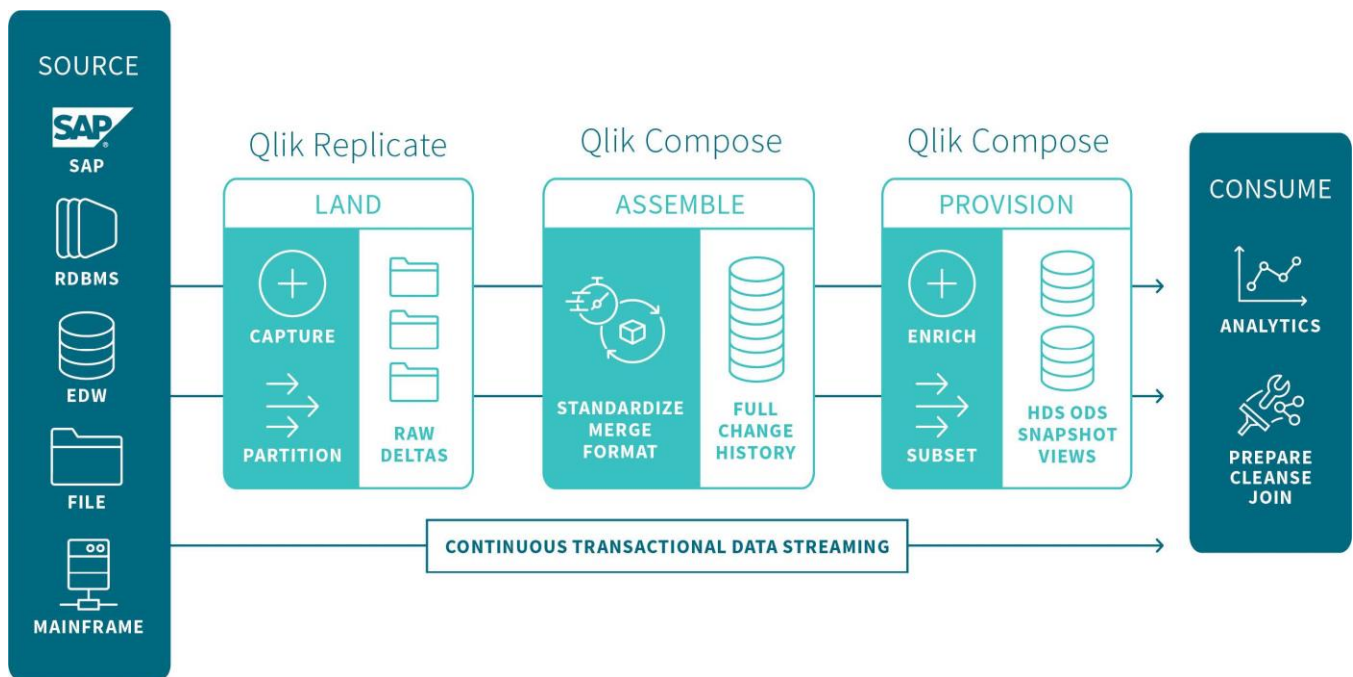


How Qlik Automates the Data Lake Pipeline

By adhering to these five principles, your enterprise IT organization can strategically build an architecture on premises or in the cloud to meet historical and real-time analytics requirements. Our solution, which includes Qlik Replicate and Qlik Compose for Data Lakes, addresses key challenges and moves you closer to achieving your business objectives.

The featured case studies and this sample architecture and description show how Qlik manages data flows at each stage of a data lake pipeline.

Your Data Lake Pipeline



Take a closer look, starting with the Landing Zone. First, Qlik Replicate copies data – often from traditional sources such as Oracle, SAP, and mainframe – then lands it in raw form in the Hadoop File System. This process illuminates all the advantages of Qlik Replicate, including full load/CDC capabilities, time-based partitioning for transactional consistency and auto-propagation of source DDL changes. Now, data is ingested and available as full snapshots or change tables, but not yet ready for analytics.

In the Assemble Zone, Qlik Compose for Data Lakes standardizes and combines change streams into a single transformation-ready data store. It automatically merges the multi-table and/or multi-sourced data into a flexible format and structure, retaining full history to rewind and identify/remediate bugs, if needed. The resulting persisted history provides consumers with rapid access to trusted data, without having to understand or execute the structuring that has taken place. Meanwhile, you, your data managers, and architects maintain central control of the entire process.

In the Provision Zone, your data managers and architects provision an enriched data subset to a target, potentially a structured data warehouse, for consumption (curation, preparation, visualization, modeling, and analytics) by your data scientists and analysts. Data can be continuously updated to these targets to maintain fresh data.

Our Qlik software also provides automated metadata management capabilities to help your enterprise users better understand, utilize, and trust their data as it flows into and is transformed within their data lake pipeline. With our Qlik Replicate and Qlik Compose solutions you can add, view, and edit entities (e.g., tables) and attributes (i.e., columns). Qlik Enterprise Manager centralizes all this technical metadata so anyone can track the lineage of any piece of data from source to target, and assess the potential impact of table/column changes across data zones. In addition, Qlik Enterprise Manager collects and shares operational metadata from Qlik Replicate with third-party reporting tools for enterprise-wide discovery and reporting. And our company continues to enrich our metadata management capabilities and contribute to open-source industry initiatives such as ODPI to help simplify and standardize Big Data ecosystems with common reference specifications.

Conclusion

You improve the odds of data lake success by planning and designing for platform flexibility, data in motion, automation, scalability, and deep source integration. Most important, each of these principles hinge on effective data integration capabilities.

Our Qlik technology portfolio accelerates and automates data flows across your data lake pipeline, reducing your time to analytics readiness. It provides efficient and automated management of data flows and metadata. Using our software, you and your organization can improve SLAs, eliminate data and resource bottlenecks, and more efficiently manage higher-scale data lake initiatives. Get your analytics project back on track and help your business realize more value faster from your data with Qlik.

Next Steps

Call or visit <https://www.qlik.com/us/products/data-integration-products> today to learn more about our software and get a free trial version.








Appendix: Platform Support

Qlik Replicate supports the following end points.

SOURCES

RDBMS  <ul style="list-style-type: none"> Oracle SQL Server DB2 iSeries DB2 z/OS DB2 LUW MySQL PostgreSQL Sybase ASE Informix 	DW  <ul style="list-style-type: none"> Exadata Teradata Netezza Vertica Pivotal 	SAP  <ul style="list-style-type: none"> ECC on Oracle ECC on SQL ECC on DB2 ECC on HANA S4 HANA 	MAINFRAME  <ul style="list-style-type: none"> DB2 for z/OS IMS/DB VSAM
CLOUD  <ul style="list-style-type: none"> AWS RDS Amazon Aurora* Salesforce 	OTHER LAGACY  <ul style="list-style-type: none"> SQL/MP Enscribe RMS 	FLAT FILES  <ul style="list-style-type: none"> Delimited (e.g., CSV, TSV) 	

TARGETS

DATABASE  <ul style="list-style-type: none"> Oracle SQL Server DB2 LUW MySQL PostgreSQL Sybase ASE Informix MemSQL* 	EDW  <ul style="list-style-type: none"> Microsoft PDW Exadata Teradata Netezza Vertica Sybase IQ Amazon Redshift Actian Vector SAP HANA 	CLOUD  <ul style="list-style-type: none"> Amazon RDS Amazon Redshift Amazon EMR Amazon S3 Amazon Aurora Google Cloud SQL Google Dataproc Azure SQL DW Azure SQL DB Snowflake Azure ADLS* 	DATA LAKE  <ul style="list-style-type: none"> Hortonworks Cloudera MapR Amazon EMR HDInsight
STREAMING  <ul style="list-style-type: none"> Azure Event Hubs MapR Kafka 	SAP  <ul style="list-style-type: none"> HANA 	FLAT FILES  <ul style="list-style-type: none"> Delimited (e.g., CSV, TSV) 	



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](https://www.qlik.com)

© 2024 QlikTech International AB. All rights reserved. All company and/or product names may be trade names, trademarks and/or registered trademarks of the respective owners with which they are associated. For the full list of Qlik trademarks please visit: <https://www.qlik.com/us/legal/trademarks>