

Increase Your Data Lake ROI

How streaming data pipelines prevent your data lake
from becoming a data swamp

ABSTRACT

Data lakes can store masses of structured or unstructured data in raw format until your enterprise needs that data for analytics. That's why data lakes are now seen as an attractive alternative to traditional data warehouses. However, enterprises like yours struggle to realize the expected return on data lake investments because of unexpected data quality, data governance, and data immediacy challenges. This paper describes how to address these issues and prevent your pristine data lake from devolving into a useless data swamp.

Data Lake Origins

Organizations traditionally built an enterprise data warehouse (EDW) to analyze large volumes of business information. They'd connect it to a query engine and visualize the result with a reporting tool. Yet the approach wasn't ideal. These data warehouses often required expensive and dedicated hardware. Populating them was slow because teams relied on hand-coded ETL scripts. Querying them frequently required specialist data schemas and SQL code knowledge. And report distribution typically lagged the useful half-life of the data.

What a paradox. The warehouse data's true value was only appreciated after teams built and populated the data warehouse, but by then it was too late. The data was out of date. That's why industries started desperately searching for a more flexible, timely, and cost-effective alternative – one many thought they'd found in the data lake.

Genesis of the "Data Lake" Phrase

In 2008, James Dixon, then chief technology officer at Pentaho, allegedly coined the term "data lake." He sought to explain unstructured data in contrast to the established phrase "data mart," a smaller repository of interesting attributes derived from raw data. Dixon thought about water metaphors: thirsty people get bottles from a mart, the mart gets cases from a warehouse, and the warehouse stores bottles filled with water from the source — the lake.

A data lake is a single enterprise data store typically, but not always, built using Hadoop. It stores virtually unlimited volumes of

- Structured data from relational databases (rows and columns)
- Semi-structured data (CSV, logs, XML, JSON)
- Unstructured data (emails, documents, PDFs)
- Binary data (images, audio, video)

The lake can also store raw copies of source system data and transformed data used for tasks such as reporting, visualization, analytics, and machine learning.

Importantly, the underlying data lake hardware didn't require specialized components; it mostly relied on cost-effective, commodity servers, and storage. And teams could use a variety of open source scripting and processing tools to extract value from the data and inform key organizational decisions. The hype surrounding the growing variety and volume of data propelled data lakes into a powerful architectural approach, especially as enterprises adopted cloud, mobile, and Internet of Things (IoT) applications where real-time data delivery was key to success. Unsurprisingly, teams believed they'd found a viable alternative to the EDW.

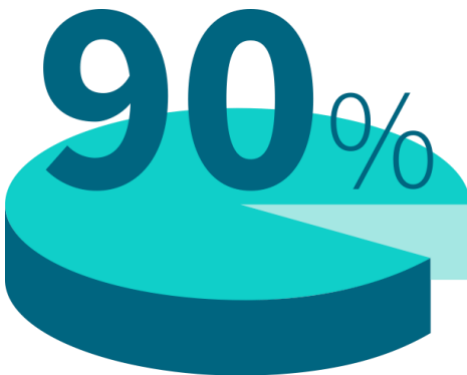
Data Lake Challenges

Many technologists were ready to declare the data warehouse dead – no longer relevant in the high-volume, high-velocity, big data age. But that changed when enterprises began rolling the new concept from development into production.

Complex Technology Stack, Lack of Skills and Best Practices

Among the first signs of trouble: the growing list of technologies observers suggested were necessary to build a data lake. A seemingly infinite array of confusing and contradictory components leading to a never-ending list of deployment opinions and advice. In fact, teams were choosing different technologies for data storage, ingestion, processing, operations, governance, security, and data analytics. Each demanded specialized expertise. Consequently, implementing a data lake turned into a complex and confusing effort with competing architectures, technologies, and methodologies. And teams were left wondering whether to deploy Apache Hadoop HDFS or Apache HBase for the underlying datastore. Apache Kafka or Apache Flink for streaming. Apache Hive or Apache Spark for SQL-like processing. Apache Sqoop or Apache Flume for ETL, and more.

Not to mention organizations being forced to revisit the same questions as new components gained momentum, as well as address new issues regarding implementation resources, readily available skillsets, and best practices. The most important question, however, was always this one: How to keep the data in the lake fresh, timely, and accurate in such a dynamic environment. If the data in the lake wasn't a relevant source of truth, then it would be useless for critical business analytics.



of deployed data lakes will be useless as they are overwhelmed with information assets captured for uncertain use cases.

Gartner, Inc.¹

¹ Reference to Gartner Paper

Gartner : G00332852 : Derive Value From Data Lakes Using Analytics Design Patterns

<https://www.gartner.com/en/documents/3803499/derive-value-from-data-lakes-using-analytics-design-patt>

Low Data Confidence

It turned out the most attractive trait of a data lake was also its Achilles' heel: The practice of storing virtually infinite amounts of data without ever questioning whether it's really needed. Since data decisions were made at retrieval time, practitioners discovered they couldn't be confident in it because they had no idea where the data set originated. Data lake query speed and cluster reliability just didn't matter if someone couldn't determine the data provenance.

This problem surfaced, in part, because of two fundamental data lake assumptions:

- Data lakes used a singular, albeit highly distributed data store
- The data lake ingest process rarely contained the necessary data lineage information or metadata for effective queries

The practice of storing everything without regard to metadata failed to guide governance, inform data provenance, and ensure compliance. As a result, confidence about the data in the lake began to drop as data delivery times rose.

Shifting Market and Technology Conditions

At first, enterprises deployed all Hadoop-based data lake implementations inside corporate data centers because their technologists could tightly control everything about the storage, server clusters, network, and software. Then public cloud vendors began offering pre-packaged, fully managed versions of Hadoop in an on-demand, pay-as-you-go model. Not surprisingly, public cloud solutions like Amazon EMR, Microsoft Azure HD Insight and Google Dataproc became first-choice Hadoop implementations. Teams saw cloud as an advantage for their data lakes for several reasons: better security, faster time to deployment, higher availability, more frequent feature/functionality updates, elasticity, greater geographic coverage, and costs linked to actual utilization.

Ironically, the acceptance of cloud reduced implementation complexity, while raising additional data lake strategy questions. All because companies were now operating both on-premises data lakes as well as deploying in the cloud – with no idea about the best way to deliver the right data to their analytics users.

Which Data Lake Storage Technology?

Traditionally, Hadoop Distributed File System (HDFS) and Hadoop YARN (a.k.a. Yet Another Resource Negotiator) formed the data management layer of Apache Hadoop-based data lakes. YARN is the resource management framework while HDFS provides scalable, fault-tolerant, and reliable storage. But cloud adoption is moving enterprises from on-premises data lakes to new cloud-based data lake storage solutions such as these:



Amazon S3

Amazon S3 is a secure, highly scalable, durable object storage solution with millisecond latency for data access. S3 is built to store any type of data from anywhere – web sites and mobile apps, corporate applications, and data from IoT sensors or devices. It is built from the ground up to store and retrieve any amount of data and delivers 99.999999999% (11 nines) of durability.



Azure Data Lake Storage Gen2

Azure Data Lake Storage Gen2 is a solution dedicated to big data analytics, built on top of Azure Blob storage. It allows teams to interface with their data using both file system and object storage paradigms. This makes Data Lake Storage Gen2 the only cloud-based, multi-modal storage service, allowing organizations to extract analytics value from all of their data.



Google Cloud Storage

Google Cloud Storage enables worldwide storage and retrieval of any amount of data at any time. Teams can use Cloud Storage for a range of scenarios including serving website content, storing data for archival and disaster recovery, or distributing large data objects to users via direct download. Google Cloud Storage is also well suited to serve as the central storage repository for data lakes.

Finding Demonstrable Data Lake ROI

The perception of data lakes to this point was that they created complexity, had unexpected costs, and caused confusion. And the “store now, find value later” ethos never allowed enterprises to move onto the data value phase. That meant data lake initiatives failed to produce a return on investment (ROI). They were too difficult to implement, needed specialized domain expertise, and took months or even years to roll out. It seemed the data lake wasn't the “new data warehouse” after all.

Rethinking the Data Lake

However, not all was lost. Ironically data lake shortcomings could be addressed by applying the lessons learned from 30+ years of data warehousing:

1. Align with the Business

Even if you believe you're creating a great data lake, your business teams may disagree. Or if they don't have visibility into what you're delivering, they may not have adoption plans. Misalignment between data lake builders and analyst teams is the single biggest barrier to getting value out of your data lake.

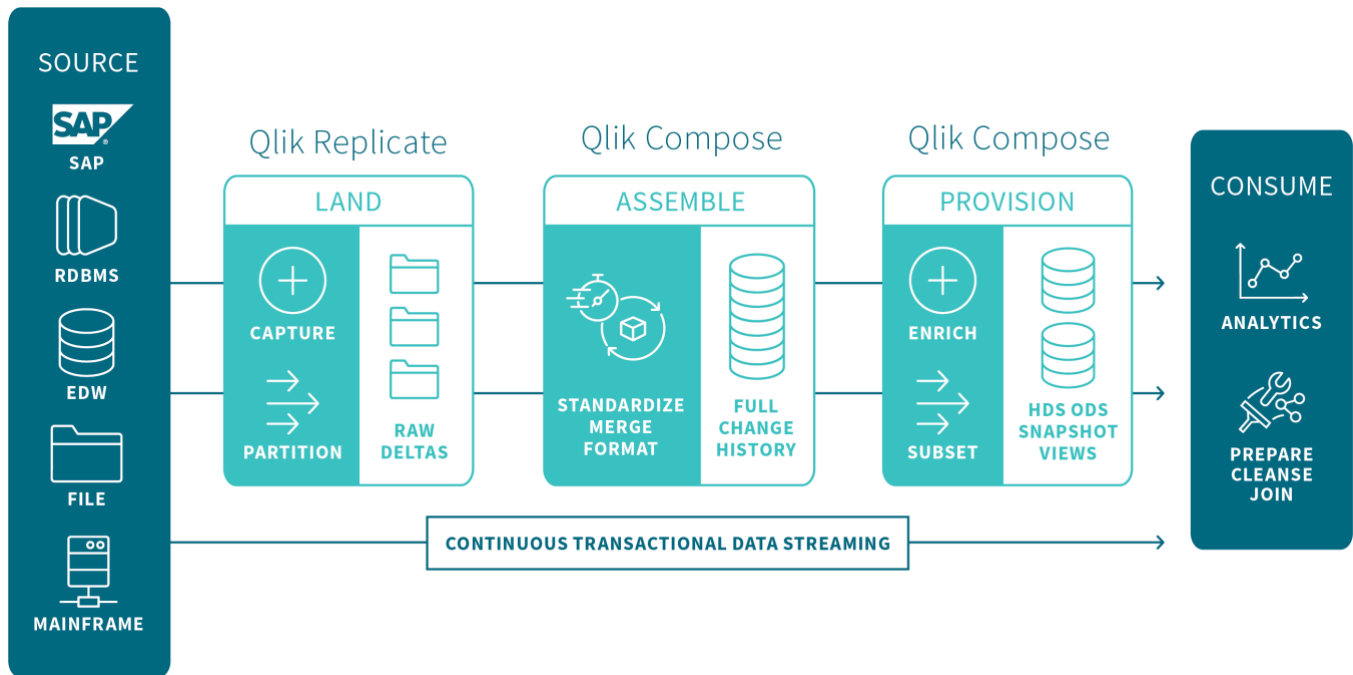
Use case discovery workshops and value-definition frameworks can help bring your teams together to agree on priorities, requirements, and business use to drive adoption. In addition, staying focused on specific use cases and departments early in your data lake journey can help you quickly demonstrate value.

2. Address the Skills Gap

You will still need someone with Hadoop administration skills to manage your data lake. Close any skills gaps you see by augmenting your standard tool set with a streaming data pipeline automation solution. It improves administration productivity. Layering a visual automation solution with familiar platform tools will help you automatically acquire new data sources and keep your provisioned data sets up to date. That way you can spend more time meeting your analysts' needs and less time on repetitive and manual data ingest, update, and data-set provisioning tasks

3. Operationalize Your Processes

Before your analyst teams can get to the right data and perform discovery and analytics, you should operationalize the functions that feed your data to the business. This means establishing a multi-zone data pipeline methodology on top of your singular data lake store. These zones promote the proper use of data lake roles, enforce data security, and assist with compliance.



Multi-zones are used this way:

Landing Zone – Raw data is continually ingested into the data lake from a variety of data sources.

Assemble Zone – Data is standardized, repartitioned, and merged into a transactionally consistent, transformation-ready historical data store.

Provision Zone – Enriched data subsets, created by data engineers, are available for consumption by data analysts or scientists. Note: data scientist no longer work with data from raw sources. Instead, they now have access to the curated data sets and views from the provisioning zone. Data sets such as an historical data store (HDS), an operational data store (ODS) or snapshots of both.

Additional operational capabilities – from scalable execution and detailed data management policies to flexible job scheduling and more – will help you deploy analytic data pipelines to your business.

4. Leverage Metadata and Catalog

Simply finding the right datasets in data lakes is a tremendous challenge for analysts. There's a large volume of data and it comes in different formats, some very complex and cryptic. That's why it's critically important to ensure that your data lake is both integrated with a metadata store that records metadata at every stage of your pipeline, and also leverages a data catalog that highlights how your data is consumed. Automating the metadata management process helps you better understand, utilize, and trust your data as it flows into and along the pipeline to eventual consumption.

5. Data Validation and Enrichment

Data engineers will often bring up data security, especially when tremendous volumes of raw data are streaming into your lake. Strong, granular, and flexible governance in your data pipeline can help you set the right policies at various stages to ensure data is validated, enriched, and protected. For example, validation ensures no data is "lost" during transport between the ingest and pipeline phases. Enrichment improves data quality, so data consumers spend less time prepping data for analysis and the eventual results are more accurate and timelier. Customer analytics is the number-one use case for data lakes, so data engineers need to secure private data. You should use the enrichment stage to apply data masking rules to protect personally identifiable information (PII) when you provision analytics-ready data subsets.

Data Lake vs. Data Warehouse: Which Right Choice?

Data lakes and data warehouses are both widely used to store and analyze masses of data, but they are not interchangeable terms. A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse meanwhile stores structured data that has generally been processed for a specific purpose.

	Data Lake	Data Warehouse
Structure	Raw	Processed – high quality
Schema	Schema-on-read	Schema-on-write
Purpose	Unknown at Time of Ingest	Current
Storage	Distributed File Store	Relational Database
Use case	Predictive Analytics	Operational and Transactional

What is the right choice? The short answer is that organizations often need both. Data lakes were born out of the need to harness the raw, granular data for new analytics uses such as machine learning, predictive analytics and data profiling. Conversely data warehouses are generally used by business users for operational reporting and business intelligence.

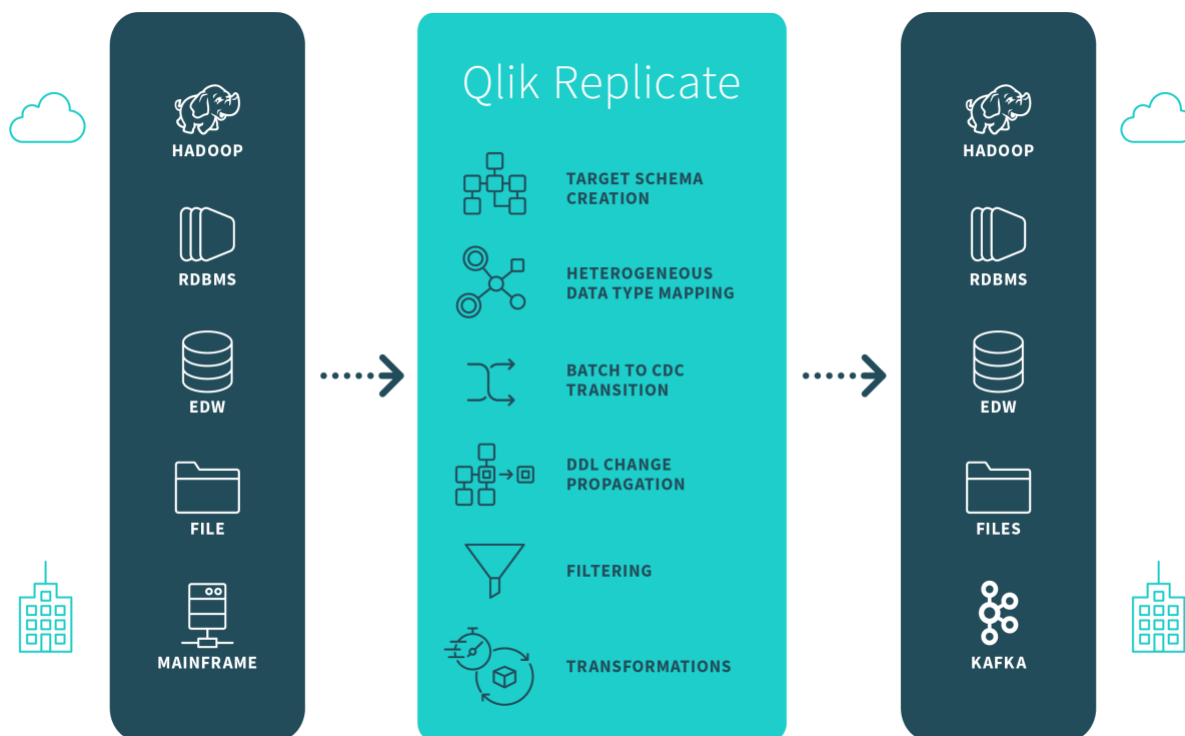
The Qlik Data Lake Solution

Qlik Replicate™

Qlik Replicate (formerly Attunity Replicate) is a simple, universal, real-time data ingestion solution, delivering data efficiently to any major data lake solution. Enterprise architects and database administrators using it eliminate manual coding with a 100% visual interface that quickly and easily configures, controls, and monitors bulk data loading as well as continual real-time updates.

Foundational change data capture (CDC) technology in our Qlik Replicate solution delivers only committed changes made to your enterprise data sources to your data lake without imposing additional overhead on the source system or data lake infrastructure. Here's what else is included:

- **Universal Connectivity** – Supports all major data sources including relational databases, mainframes, enterprise applications such as SAP, streaming solutions such as Apache Kafka, enterprise data warehouses, Big Data technologies, and cloud infrastructure such as Amazon Web Services, Microsoft Azure and Google Cloud Platform.
- **No Coding, Simple GUI** – Has an intuitive interface that lets you quickly and easily configure data feeds.
- **High Performance and Scalable** – Ingests data at high speeds with near linear scalability from hundreds to thousands of data sources.
- **Agentless Architecture** – Built-in log-based, agentless CDC reduces the administration burden and eliminates the source system processing penalty.
- **Real-Time Data Updates** – Continually ingests data with enterprise-class CDC technology that immediately delivers, with virtually no latency.

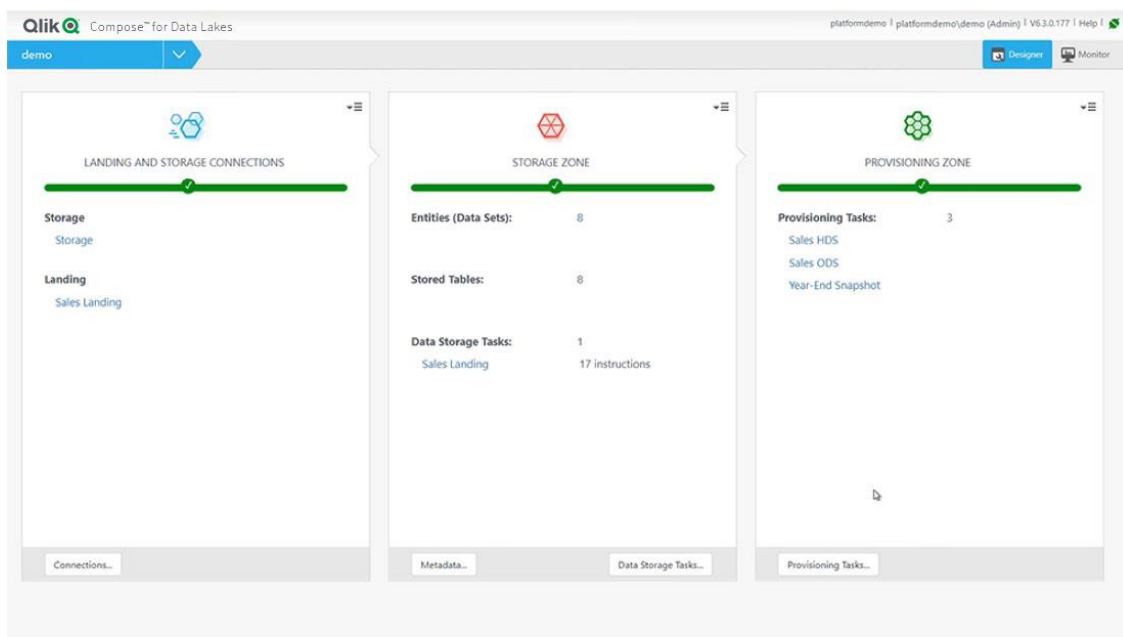


Qlik Compose for Data Lakes

Our Qlik Compose for Data Lakes solution (formerly Attunity Compose) simplifies and automates these manual, time-consuming and repetitive tasks: data pipeline creation, loading, and updates. It improves your data lake ROI and is your fastest route to analytics-ready data.

Realize greater value by automating your data ingest and target schema creation while ensuring continuous data updates to zones. These are features our Qlik Compose for Data Lakes includes:

- **Data Pipeline Designer** – Use our point-and-click designer which automatically generates transformation logic and pushes it to task engine for execution.
- **Hive or Spark Task Engines** – Run transformation tasks as a single, end-to-end process on either Hive or Spark engines.
- **Full Change Data History** – Standardize and combine multiple change streams into a single historical data store ready for downstream processing.
- **Data Set Provisioning** – Easily create analytics-ready data subsets for analysts or further downstream processing.
- **Multiple Export Formats** – Export data sets in several formats including ORC, AVRO, and Parquet.



Apache Hive and Apache Spark

Apache Hive, Apache Spark, and Spark SQL all belong in the SQL-on-Hadoop category.



Apache Hive is an open-source data warehouse and analytic package that runs on top of an Apache Hadoop cluster. Hive scripts use an SQL-like language called Hive QL (query language) that abstracts programming models and supports typical data warehouse interactions. Hive enables you to avoid the complexities of writing Tez jobs based on directed acyclic graphs (DAGs) or MapReduce programs in a lower-level computer language, such as Java.

Hive extends the SQL paradigm by including serialization formats. You can also customize query processing by creating table schema that match your data, without touching the data itself. In contrast to SQL (which only supports primitive value types such as dates, numbers, and strings), values in Hive tables are structured elements, such as JSON objects, any user-defined data type, or any function written in Java.

Apache Spark and Spark SQL



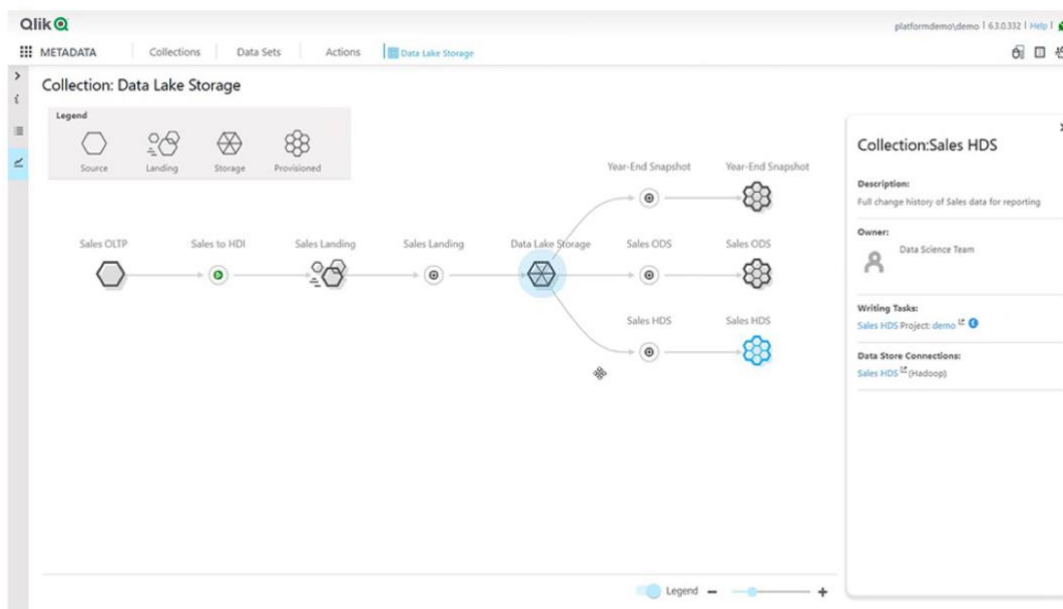
Apache Spark is an open-source, distributed, general-purpose, cluster-computing framework. Originally developed at the University of California, Berkeley's AMPLab, the Spark codebase was later donated to the Apache Software Foundation, which has maintained it since.

Spark SQL is Apache Spark's module for working with structured data – that's any data that has a schema such as JSON, Hive Tables, and Parquet. Spark SQL provides a domain-specific language (DSL) to manipulate data in Scala, Java, or Python and provides SQL language support, with command-line interfaces and ODBC/JDBC server.

Qlik Enterprise Manager

Our Qlik Enterprise Manager (formerly Attunity Enterprise Manager) is the unified command center helping you configure, execute, and monitor your data pipelines across your enterprise. It features an intuitive graphical interface that boosts administration productivity, optimizes performance, and ensures security. Here's what else it includes:

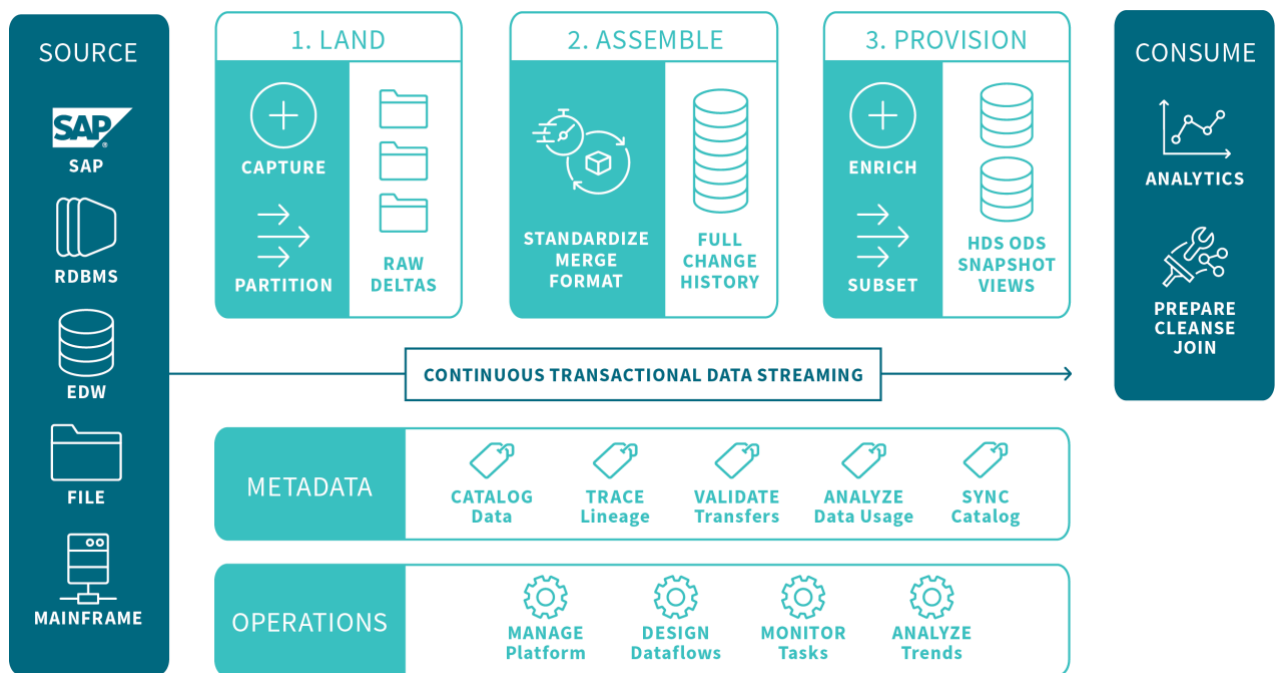
- **Unified Console with Multiple Perspectives** – The single pane of glass provides both operational and metadata views for easier pipeline and data management.
- **Monitor, Analyze, and Control** – You can track and analyze hundreds of operational tasks in real time across your environment to pinpoint, diagnose, and resolve data pipeline issues.
- **Metadata Catalog** – The central metadata repository is automatically populated with information from source and target systems to help your data engineers understand, use, and trust data pipeline flows.
- **Data Profiling and Lineage** – You get a detailed summary of all data attributes in your data lake and pipeline that highlights data provenance and the downstream impact of any data changes.
- **Metadata Directory Interoperability** – Our metadata is open and accessible to leading metadata repositories such as Apache Atlas to help you better understand how your pipeline data fits into your broader data landscape.



How Qlik for Data Lakes Automates Your Data Pipeline

By applying new data lake thinking, your enterprise IT organization can more readily build an on-premises or cloud architecture to meet historical and real-time analytics requirements. Our solution, which includes Qlik Replicate and Qlik Compose for Data Lakes, leverages new concepts and enables your organization to achieve your data delivery objectives.

Here's a sample architecture and description of how our solution can manage data flows at each stage of your data lake pipeline.



Let's start with the Landing Zone. Our Qlik Replicate copies data in raw form (often from traditional sources such as Oracle, SAP, and mainframe) into the data lake Landing Zone. This process showcases Qlik Replicate capabilities, including full load/CDC, time-based partitioning for transactional consistency and auto-propagation of source schema and data definition language (DDL) changes. Data is now ingested and available, but not yet in a form ready for analyst consumption.

In the Assemble Zone, our Qlik Compose software standardizes and combines multiple change streams into a single transformation-ready change history. It automatically merges the multi-table and/or multi-sourced data into a flexible format and structure, retaining full data source transaction history. This is extremely useful if you need to rewind, identify, or remediate data changes. The resulting persisted history provides your data consumers with rapid access to trusted data – without having to understand what underlying transaction processes have taken place to date. Your data architects and engineers, meanwhile, maintain central control of the entire process.

In the Provision Zone, your data architects and engineers provision an enriched data subset to a target, potentially a structured data warehouse, for further consumption by your data scientists and analysts. It's important to note provisioned data sets are automatically and continuously updated. There's no need for your analyst to access original data sources to acquire fresh data sets for their analytics workflows.

Our Qlik® solution for data lakes also provides automated metadata management capabilities to help your data consumers better understand, use, and trust their data as it flows into, and is transformed within, your data lake pipeline. Using both our Qlik Replicate and Qlik Compose, your team can also add, view, and edit data entities (e.g., tables) and attributes (i.e., columns). Our Qlik Enterprise Manager centralizes all the technical metadata so you can track data lineage of any piece of data from source to target, and assess the potential impact of table/column changes across data zones.

In addition, our metadata is open and accessible to leading metadata repositories such as Apache Atlas or data catalogs such as Qlik Catalog™. That helps you better understand how your pipeline data fits into your broader data landscape. By continuing to enrich our metadata management capabilities and contributing to industry initiatives such as ODPi, our solutions help simplify and standardize data lake ecosystems with reference architecture specifications.

The Value of a Full Change History in Your Data Lake

The full change history serves as an interim repository for information before data is provisioned to one or more data marts and sent to other data lakes or warehouses for further processing.

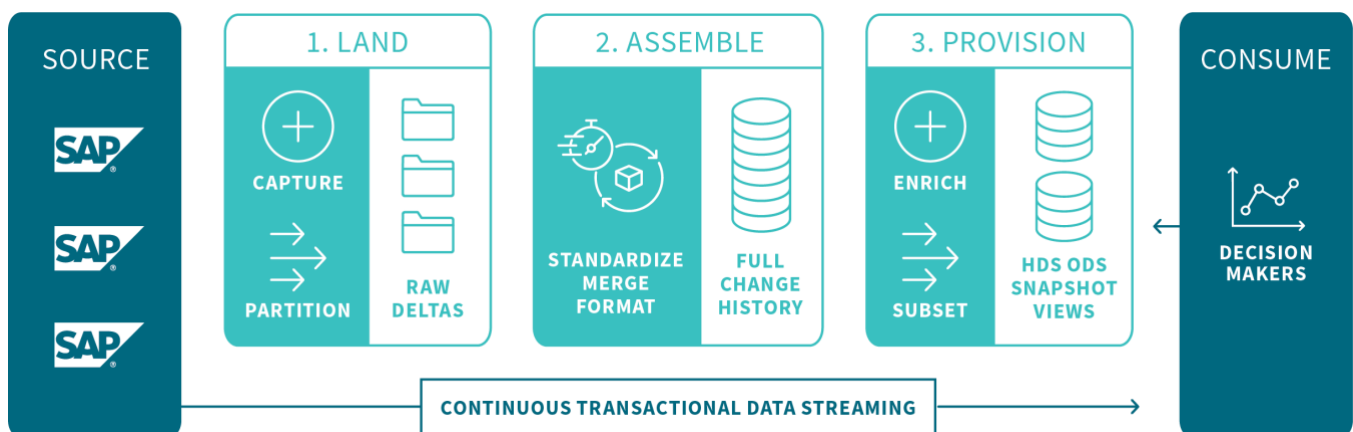
While change histories aren't an entirely new concept – they're from the 1990s – data lake teams are taking a fresh look at adding historical transactional data stores to their data lake strategies. An historical store is often a welcome addition because it not only aggregates multiple data sources inside the lake to speed data mart provisioning, it also insulates transactional systems from direct interaction by data scientists and analytics reporting.

A Case Study

Decision makers at an international food distributor discovered they were unable to match orders and production line item data fast enough to sell tens of millions of food items each week. They needed a current view of production capacity data, customer orders, and purchase orders. Yet they struggled to integrate large datasets – distributed across several business units and systems including SAP Enterprise Resource Planning (ERP) applications. That's when they decided to roll out a new Hadoop-based data lake to streamline analytical processes.

The organization's Data Operations Engineers used our Qlik for Data Lakes solution to completely automate their data pipeline and continuously deliver analytics-ready data sets to the food distributors' decision makers. Now our solution efficiently captures SAP record changes every five seconds, decodes that data from complex source SAP pool and cluster tables, then copies it into the data lake Landing Zone. From there, our solution automatically merges the data into the historical data store in the Assembly Zone where in-memory Spark jobs match orders to production data on a real-time basis. Next, data marts are automatically created and distributed via the Provisioning Zone to various analytics stakeholders – confident in the data provenance.

Today, the global food distributor operates more efficiently and profitably because it was able to automatically unlock data from complex SAP source structures. The team also accelerated sales and product delivery with accurate, real-time operational reporting from the greatly improved data lake.



Increase Your Data Lake ROI

There's no doubt data lakes have become an established, attractive complement to traditional data warehouses for many real-time analytics use cases. Yet many enterprises still struggle to realize the expected ROI on their data lakes due to unanticipated data quality, data governance, and data immediacy challenges.

Our Qlik for Data Lakes solution reduces your time to analytics readiness and improves your data lake ROI by automating your data pipelines and optimizing your data delivery. Get your data lake analytics project back on track with our solution, and realize more value from data at the speed of change.

Next Steps

Visit qlik.com/us/products/qlik-compose-data-lakes to find out more about our Qlik for Data Lakes solution or to arrange a demonstration.



About Qlik

Qlik's vision is a data-literate world, one where everyone can use data to improve decision-making and solve their most challenging problems. Only Qlik offers end-to-end, real-time data integration and analytics solutions that help organizations access and transform all their data into value. Qlik helps companies lead with data to see more deeply into customer behavior, reinvent business processes, discover new revenue streams, and balance risk and reward. Qlik does business in more than 100 countries and serves over 50,000 customers around the world.

qlik.com

© 2020 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.