WHITE PAPER

More Data, Analytics-Ready

Accelerate Data Delivery to Data Lakes, Data Warehouses, Streaming and Cloud Architectures

DataIQ

Qlik

INTRODUCTION

Companies everywhere now need to cost-effectively analyze a wide variety of data. That's because analytics use cases such as fraud detection, real-time customer offers, market trend/pricing analysis, social media monitoring, and more are becoming the best way to stay competitive – and the only way to truly become a disruptor. Today, as the use of artificial intelligence and machine learning (AI/ML) algorithms, drawing on sources such as Internet of Things (IoT) sensors, further boost the volume, variety, and velocity of data, how can your business keep up? How can you speed your organization's data analysis?

Integrating all required data at scale can overburden your already taxed IT team. There's complex manual coding (often varying by platform type) and procedures that disrupt production sources. And data architects and database administrators (DBAs) are already struggling to efficiently execute and track replication across the enterprise. Without the right tools, it's nearly impossible for them to efficiently manage the hundreds or potentially thousands of integration tasks that your initiatives entail.

Here's the rub: modern data requirements break traditional data integration tools. But modern data requirements don't break our Qlik Replicate™ solution. That's why it's an ideal modern data integration solution for efficiently delivering more data, ready for agile analytics, to a diverse range of data lake, data warehouse, streaming, and cloud architectures.

Modern Data Integration

Qlik Replicate, formerly Attunity Replicate, modernizes your environment by moving data at high speed across all major source and target platforms with a single “Click to Load” interface that completely automates the end-to-end replication process. Our software gives your administrators and data architects a way to easily configure, control, and monitor bulk loads and real-time updates with enterprise-class change data capture (CDC) capabilities. This enables instantaneous database replication of changed data to the target. And our zero-footprint CDC eliminates any risk of production impact.

Our software accelerates both heterogeneous and homogeneous data replication, and controls data flow across hybrid multi-platform environments. It supports all major databases, including Oracle, Microsoft SQL Server, and IBM DB2. Beyond transactional database support, Replicate integrates with major analytics platforms, including Microfocus Vertica, IBM Integrated Analytics PureData (formerly Netezza), Microsoft Synapse Analytics, Oracle Exadata, and Teradata. Also Hadoop distributions from Cloudera and Azure HDInsight, and streaming systems such as Apache Kafka. Replicate leverages native utilities and APIs to guarantee fast, optimized, and secure data capture and loading.

Ease of Use and Automation

A key differentiator is the solution’s “Click-2-Replicate” user interface, supporting drag-and-drop functionality. This intuitive approach makes data replication easy to learn and fast to implement. Because our Qlik Replicate solution automates the steps required to build a replication solution, it shields your users from complexity, eliminating the need for master DBA skills, custom scripting, or consultants. Our “Click-2-Replicate” designer is a web-based interface for your users to access from anywhere when they want to configure database schema mappings between sources and targets, transformations, and filtering—all with a graphical task map. Our designer helps them create table selection patterns, configure transformations, and define filters easily and rapidly.

Industry’s Broadest Platform Support

Our solution showcases deep partnerships and broad product integration with industry leaders. It supports all major source and target systems for data replication, including relational database, data warehouse, data lake, Hadoop, cloud, and mainframe platforms. It also supports MongoDB as a NoSQL target and writes CDC as messages to all major streaming platforms. See Appendix for a detailed list of supported source and target platforms.

Capabilities and Architecture

With its multi-server, multi-task, and multi-threaded architecture, you can scale your on-premises and cloud implementations to thousands of distributed servers and data centers worldwide. The Qlik Replicate architecture is comprised of three domains: sources (databases, etc.), replication server(s), and targets (databases, data warehouses, data lakes, cloud, etc.). Its key architectural principles include:

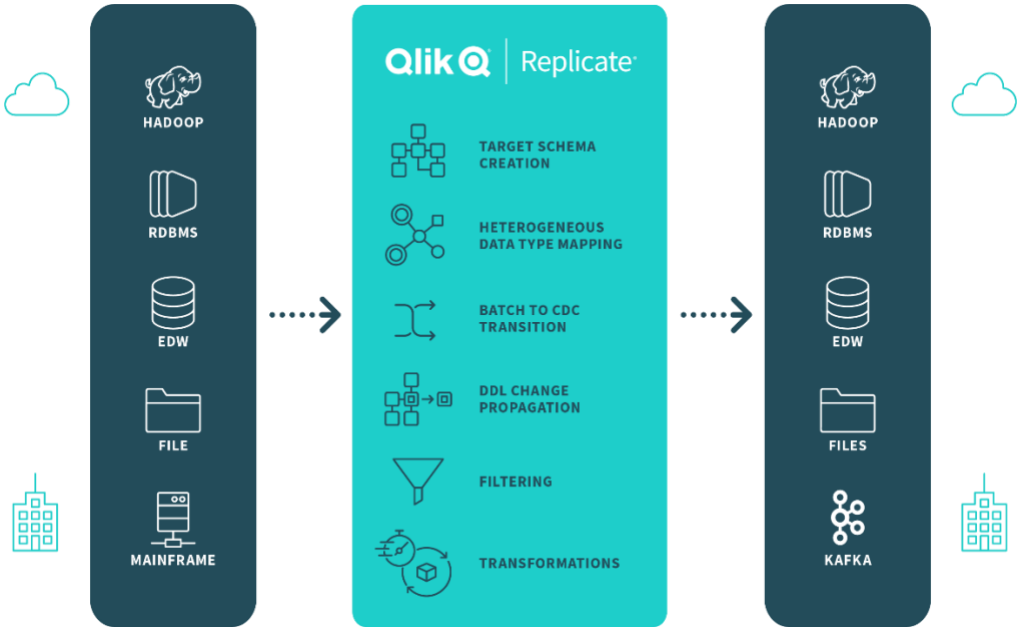
- Full-load and CDC replication
- Agentless, zero-footprint software at source and target
- Scalability and flexibility
- The zero-code user interface
- Centralized visibility and control

Full-Load Replication

With full-load replication, our Qlik Replicate software takes all the tables from the source and creates copies at the target. Then, it automatically defines metadata required by the target and populates the tables with data from the source.

Your data is loaded into one or multiple tables to improve efficiency. Although source tables may be subject to update activity during the full-load process, there's no need to stop applications in the source. Our unique CDC process automatically activates when table loading starts. However, changes are not applied to the target until after loading is complete. And although data on the target may not be consistent while the load is active, at completion, the target has full data consistency and integrity.

If you need to though, you can interrupt the loading process. When it restarts, our software will continue where it stopped. You can add new tables to an existing target without reloading existing tables. Similarly, you can add or drop columns in previously populated target tables without reloading.



Schema/Data Definition Language Replication

Our Qlik Replicate solution automatically generates target databases based on metadata definitions in the source schema. Any data definition language (DDL) changes to that schema, such as the addition of new tables and columns or changes to data types, can be replicated dynamically to the target.

Incremental Replication or CDC

Our CDC process copies updates as they occur in the source data or metadata and applies them to the target endpoint in real time. With Qlik Replicate CDC, you can move large volumes of data changes into target databases or cloud environments with efficiency and ease—at speed.

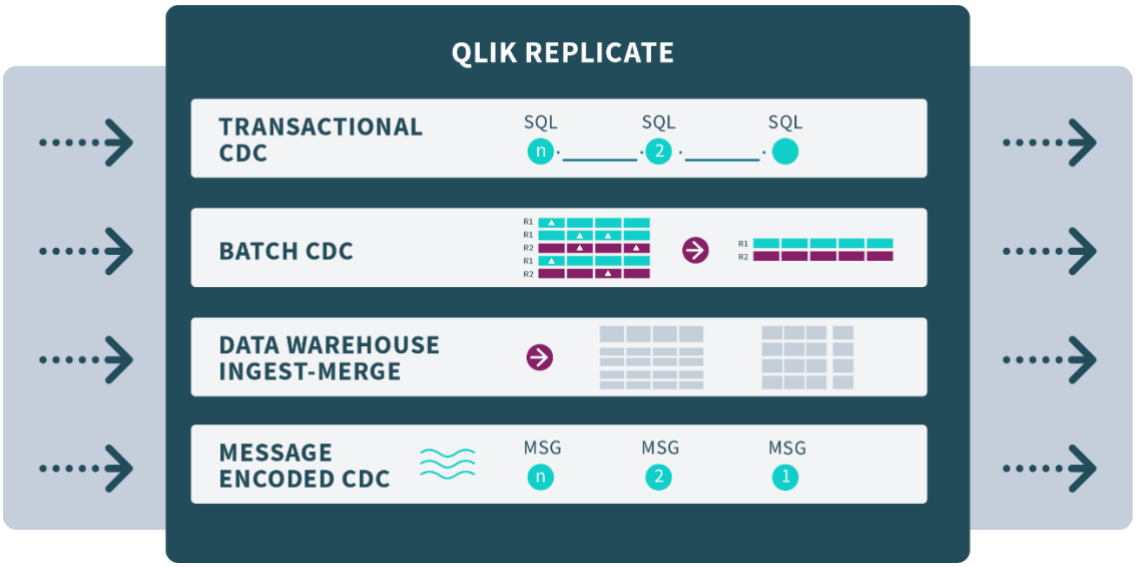
Our Qlik Replicate software gives you the following CDC options:

- **Log-based Capture**

  Our CDC reads the recovery log file of the source endpoint management system and groups together entries for each transaction. Its process techniques ensure efficiency without impacting target data latency. If the CDC process can't apply the changes to the target in a reasonable period (e.g., when the target isn't accessible), then it buffers the changes on the replication server for as long as needed. No rereading of the source database logs that may take hours or days!

- **Query-based Capture**

  When log-based capture isn't available then our software queries the source tables using context columns, such as TIMESTAMP, to identify and capture changes efficiently from source enterprise data warehouse platforms.



Advanced CDC technology has several options when delivering data to targets:

- **Transactional CDC**

  This is for standard database targets where transactional consistency is more important than high performance. Qlik Replicate streams changes on a transaction-by-transaction basis to maintain transactional integrity at any point in time.

- **Batch CDC**

  In cases where high transaction rates and low latency are required, Batch Optimized Apply for CDC, also known as Batch CDC, is the right choice. This process includes a pre-processing action that groups transactions into batches in the most efficient way.

- **Data Warehouse Ingest-Merge**

  This process follows the same step as Batch CDC, then adds this final step: leveraging the target data warehouses' performance-optimized native utilities as it delivers the data.

- **Message Encoded CDC**

  Through integration with Apache Kafka and other streaming systems, our Qlik Replicate software ingests high data volumes from many data sources, enabling your administrators to deliver data to HBase, Cassandra, Spark and other Big Data platforms. Your users can load data either in bulk or through Qlik Replicate CDC, which relies on Kafka message brokers to relay source changes automatically through in-memory streaming. Our software supports multi-topic and multi-partitioned data publication, and like other platforms, provides integrated management and monitoring through an intuitive console. It supports additional message streaming targets such as Azure Event Hubs and Amazon Kinesis.

Zero-Footprint Software

Data replication has to ensure performance and continuous availability – both for production applications and your analytics users. Our Qlik Replicate solution has a unique zero-footprint architecture, designed so that CDC processes can identify and replicate production transactions real time with a remote transaction log reader. No agents required on the source or target databases. This eliminates mission-critical systems' administrative and performance overhead. And it provides a strong advantage over alternative CDC tools that require intrusive triggers and/or shadow tables, slowing production application and replication performance and creating more processing and administrative overhead.

For example, agentless endpoint for IBM DB2 for z/OS and IBM DB2 for iSeries deliver significant optimizations to improve performance and reduce footprints when capturing changes from these platforms. Lab tests demonstrate reductions of 85 percent in source MSU (million service units), 75 percent in replication latency, and 95 percent in loading time.

Our software is designed for maximum flexibility. The transaction log reader can be installed either on the replication server to achieve a zero-footprint impact or on the source database server. As a result, your users can filter source rows on either the source database or replication server.

Time-based Partitioning

Time-based partitioning lets you process transactions (insert/update/ delete) from many tables to Data Lake targets in a consistent way. This feature in our Qlik Replicate solution only makes completed transactions available for a predefined time interval (minutes, hours, or days). Any partial transaction updates will be deferred until the following time. This gives analysts full confidence in both the integrity of the data they query and their query results.

Filtering and Compression

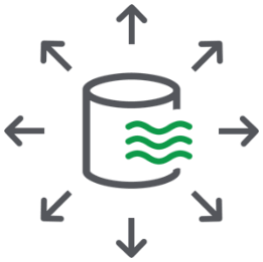
Whenever filtering conditions are defined on the values of one or more source columns, our solution discards irrelevant rows and columns before they're replicated to the target database. This may occur, for example, when a column is not present in the target database schema or when a row doesn't pass the user-defined predicates on the rows within the source tables.

Transformation

There may be circumstances where data to be included in the replicated tables isn't an exact copy of the source data. When this happens, our Qlik Replicate solution allows your users to define and automatically apply those changes to tables and columns. Examples include:

- Renaming the target table
- Renaming any column in the target table
- Deleting a target column
- Changing the data type and/or the length of any target column, and
- Adding target columns

Our software performs data type transformations as required, calculating the values of computed fields and applying the changes as one transaction to the target. When no user defined transformation is set – but replication is done between heterogeneous databases – some transformation between different database data types may be required. In these cases, our solution automatically takes care of the required transformations and computations during the load or CDC execution.



Universal Stream Generation

Our solution separates data and metadata into different topics, allows for smaller data messages and easier integration to major streaming services such as Apache Kafka, Confluent, Microsoft Azure Event Hubs and Amazon Kinesis. Flexible message formats such as JSON and Avro enable easier integration of metadata into various schema registries.

Optimized Cloud Transfer

Having geographically distributed business units can require data storage off-premises or in the cloud. From every location, each group needs timely access to its own subset of data. Our Qlik Replicate software solves this challenge by providing an innovative, highly secure, and resilient wide-area network (WAN) transfer engine that optimizes transfer speeds to target databases based on available bandwidth. Our algorithms compress large tables, which are then

split into multiple, configurable streams which are then combined into small table batches or CDC streams together that improve efficiency and speed.

Since network outages and other unpredictable events can impact data flow to and from the cloud as well as other remote data repositories, our solution offers seamless recovery from interrupted transfers – from the exact point of failure. Our solution first stages all source data in files located in a temporary target directory. It then moves files to the target directory and validates content between the source and target files. After successful validation, it loads the data into the target database.

Security Capabilities

Our solution addresses security issues related to data transfer to the cloud by establishing a three-level, secure data transfer mechanism:

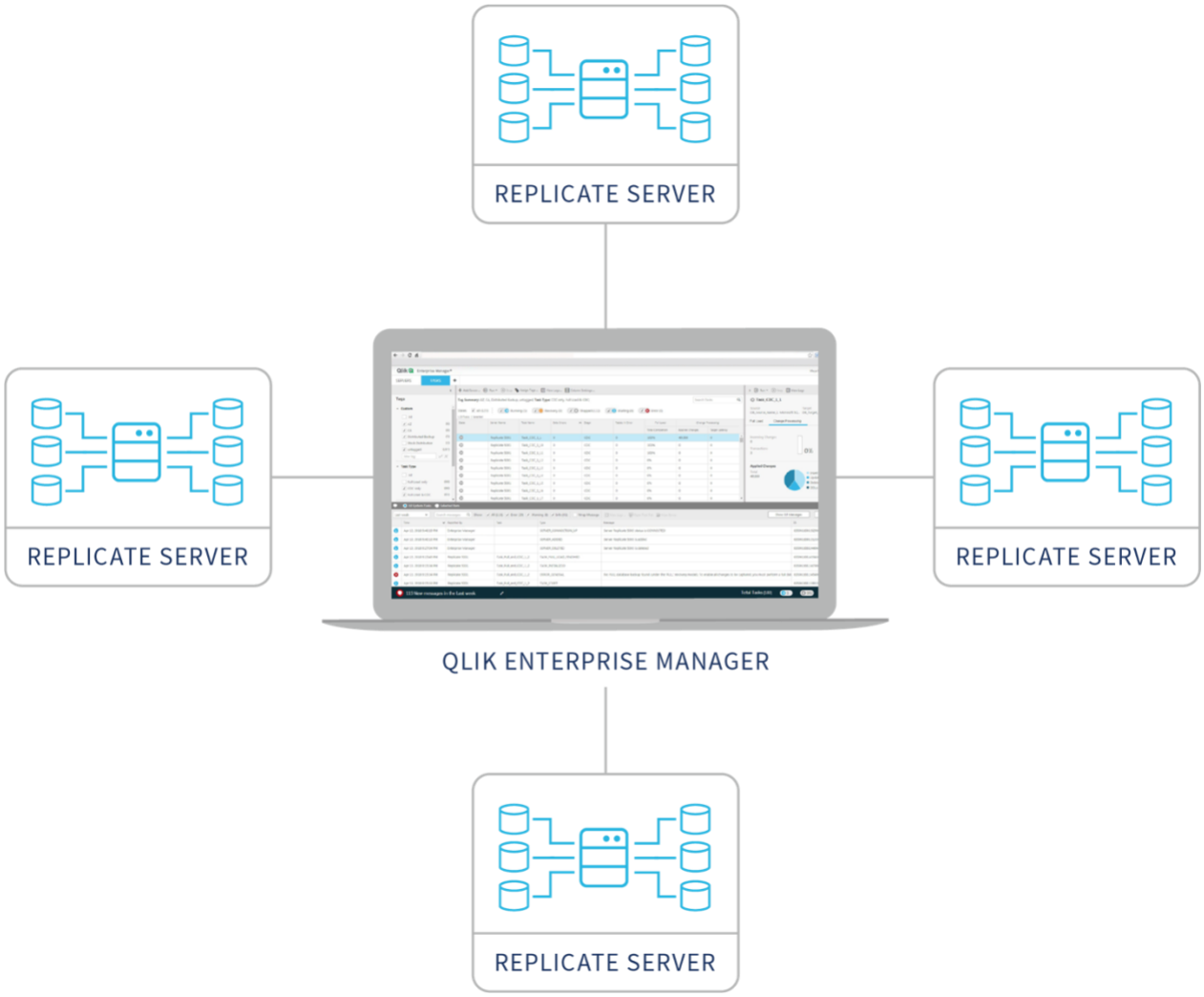
1. Establishing a secure client-server connection through key exchange.
2. An agreed-upon password is then used to scramble the keys, eliminating man-in-the-middle attacks.
3. Files are secured during transfer using advanced, NSA-approved (AES-256) encryption.

Intelligent Management and Control of Enterprise Data Integration

Through our Qlik Enterprise Manager, formerly Attunity Enterprise Manager, solution you gain efficient, high-scale data replication for initiatives such as data lake consolidation. Our software ensures centralized control of Qlik Replicate tasks and data flow across distributed environments, enabling your enterprise to scale easily and monitor thousands of integration tasks in real time through KPIs and alerts. With our Manager, you can monitor distributed Replicate servers across multiple data centers and control data flow across distributed environments, on premises and in the cloud, from a single console.

It’s an intuitive and logical method of improving efficiency, performance, and compliance. Customizable views allow you to define how search results are presented. For example, grouping tasks by server, database source or target,

by specific application, or even by physical location, you can incorporate the enterprise business logic regulations mandate. Granular searching and filtering capabilities offer actionable insight into data loading and task status. By drilling down from the main console, you can view current task status to identify and remediate issues, so you continue to meet performance service-level agreements (SLAs).



Microservices API

Our Qlik Enterprise Manager provides intelligent REST and .NET APIs that are designed for invoking and managing Qlik Replicate services.

Use Cases and Best Practices

Our Qlik Replicate software improves efficiencies and business operations for a variety of enterprise use cases. The most popular are data replication for operational analytics or query offloading, high-scale data lake consolidation, line of business workload offloading, and enablement of cloud analytics.

Enterprise Data Replication

Using Qlik Replicate as a unified replication platform reduces the time you spend and the complexity you experience to maintain data availability across heterogeneous and distributed environments. It replicates, synchronizes, distributes, consolidates, and ingests data across all major databases, data warehouses, and data lakes – whether deployed on-premises or in the cloud. With our solution, your organization can scale its architecture to move data across thousands of databases with centralized monitoring and management. You can more effectively optimize workloads and support business operations, applications, and analytics needs.

Real-Time Data Warehousing

Qlik Replicate delivers database updates to your data warehouse with low latency. A continuous and efficient data acquisition process, it ensures real-time data availability and eliminates the complexity, cost, and delays associated with homegrown and traditional ETL software. Our Qlik Replicate log-based CDC also minimizes impact on your source production operations

Query Offload and Live Reporting

When you use our Qlik Replicate CDC, you can create live replicas from production applications for separate reporting and analytics databases. That lets you scale real-time BI and analytics with zero production impact. By offloading queries and workloads like this, you can continuously analyze fresh data while meeting production application SLAs and minimizing cost.

Easy Ingestion of Structured Data into Data Lakes

Our Qlik Replicate software delivers high-performance data loading and publication to data lake ecosystems through native APIs. And it's certified with the leading Hadoop distributions like Cloudera, Hortonworks, and Azure HD Insight.

Accelerated Cloud Migrations and Analytics

Qlik Replicate accelerates data flow to the cloud. Your organization can easily and securely transfer data over WANs at high performance with encrypted multi-pathing technology to all the major public cloud platforms, including Amazon Web Services (all RDS databases and Amazon Redshift), Microsoft Azure SQL Database and Synapse Analytics, and Google Cloud.

SAP Analytics Enablement

Choose Qlik Replicate to easily and securely transfer your SAP data, documents, and transactions into external data platforms for downstream analytics initiatives. Our solution supports data collection from all core SAP business applications such as ERP, HR, CRM, industry-specific applications, and loading data into SAP HANA.

Streaming Data Delivery

Use Qlik Replicate to deliver high-data volumes at low latency in real-time from many data sources to Apache Kafka, Azure Event Hubs, Google Cloud Pub Sub and other streaming platforms.

CONCLUSION

Changing data integration by enabling your IT staff to deliver more data, ready for analytics, to data lakes, data warehouses, streaming and cloud architectures is our mission. Unlike traditional, batch-oriented, and inflexible ETL approaches of the last decade, our solutions are modern with the real-time architecture enterprises like yours require to harness the agility and efficiencies of new data lakes, data warehouses and cloud offerings. If you're seeking a single solution to improve data delivery for your agile analytics initiatives, consider our Qlik Replicate solution.

Data IQ - Qlik Elite Master Reseller
Av. Libertador 5478, piso 5 - Buenos Aires - Argentina

Contact details

https://dataiq.com.ar/
+54 11 4780-3700
marketing@dataiq.com.ar

About Qlik

Qlik's vision is a data-literate world, where everyone can use data and analytics to improve decision-making and solve their most challenging problems. Our cloud-based Qlik Active Intelligence Platform delivers end-to-end, real-time data integration and analytics cloud solutions to close the gaps between data, insights and action. By transforming data into Active Intelligence, businesses can drive better decisions, improve revenue and profitability, and optimize customer relationships.

qlik.com

© 2020 QlikTech International AB. All rights reserved. Qlik products and QlikTech logos® are trademarks of QlikTech International AB that, where indicated by an "®", have been registered in one or more countries. Other marks and logos mentioned herein are trademarks that are not yet registered. For full trademark list, visit our website.