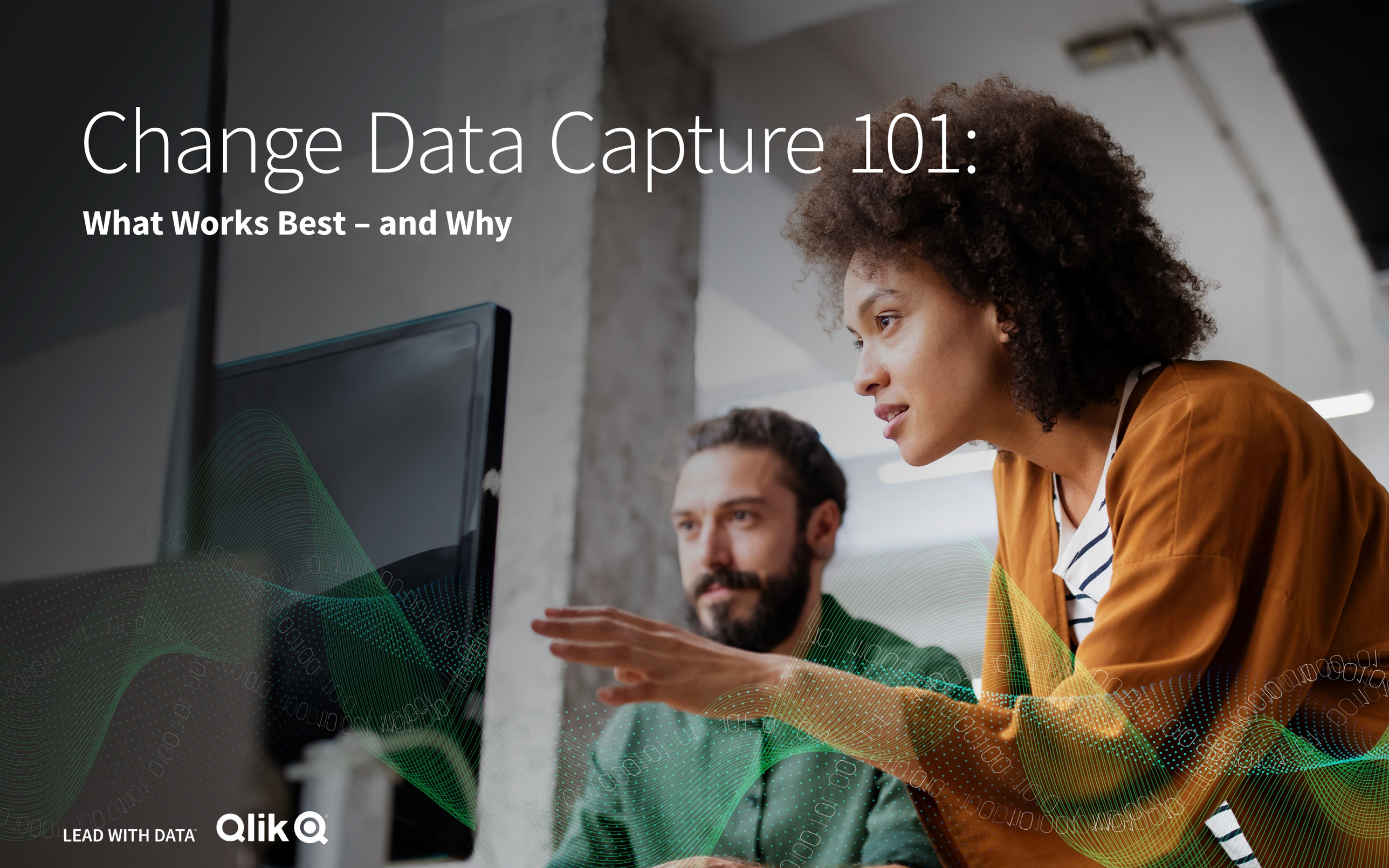
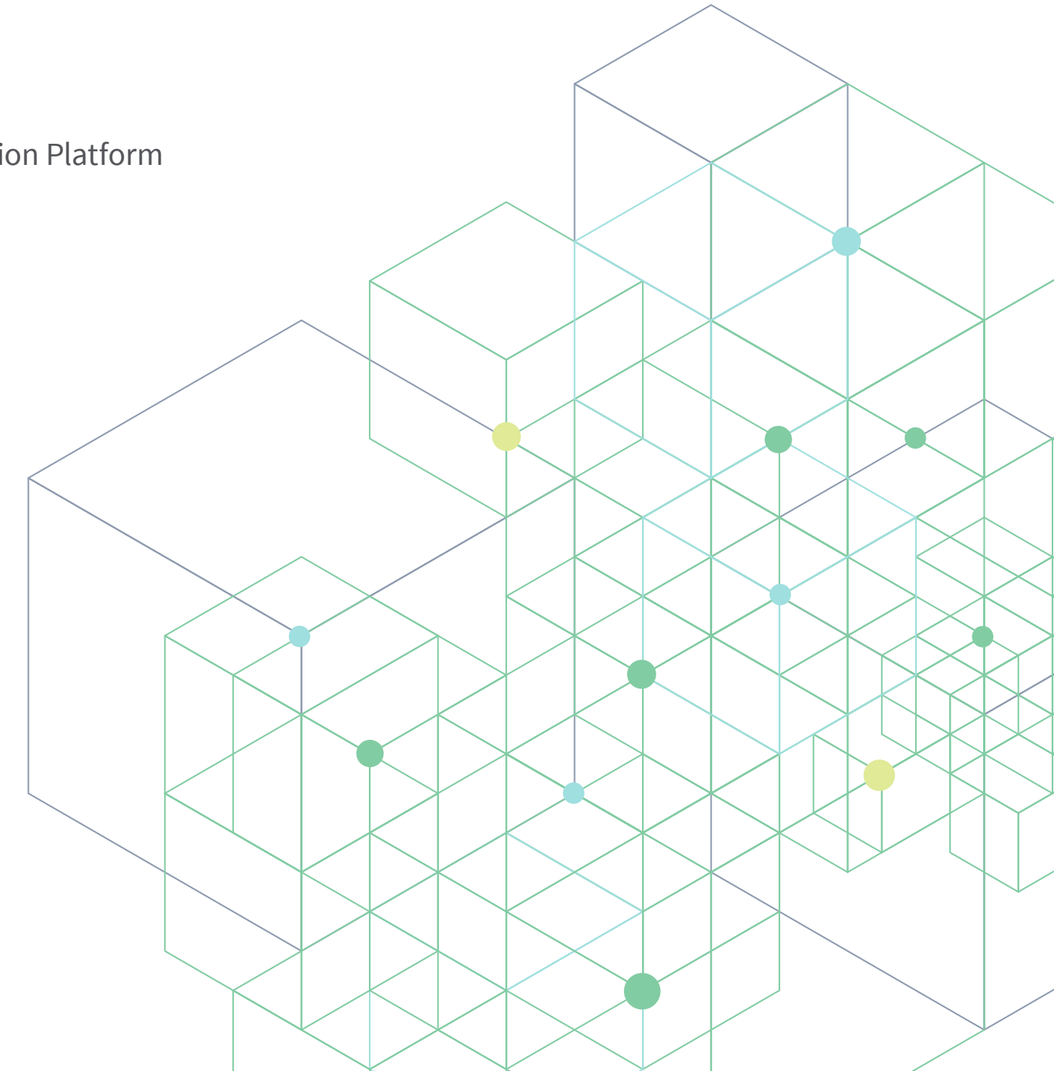
Change Data Capture 101:

What Works Best – and Why



CONTENTS



The new promise of data analytics – and the new demands on data architecture.

Leaner business processes. Personalized user experiences. Smarter approaches to risk. And novel streams of revenue. Modern analytics initiatives have the potential to reinvent business from the inside out. But to take advantage, IT organizations first have to reinvent how they move, store, process and analyze data. And the challenges are real.

Cloud data warehouses, data lakes and real-time data streaming all play a role in modern architectures. These technologies are complementing and sometimes even replacing the enterprise data warehouse, the traditional structured system of record for analytics. But as every data engineer knows, creating an efficient, connected and fast-moving pipeline from raw data to analytics-ready data is tricky stuff.

One primary challenge is integration. Data must be replicated to analytics platforms, often continuously, without disrupting production applications. And because data is being generated at dizzying rates, the processes used to replicate that data must be scalable, efficient and able to absorb high data volumes from many sources. In order to make business sense, they have to do all that without a prohibitive increase in labor or complexity.

The problem with traditional data integration.

Unfortunately, today's data integration demands can't be met by yesterday's data integration processes. Batch replication jobs and manual extract, transform and load (ETL) scripting procedures are slow and inefficient. They disrupt production, tie up talented programmers and create network and processing bottlenecks. And they can't scale sufficiently to support strategic initiatives. As a result, businesses are missing opportunities, losing competitive ground and breaking operational budgets.

Real-life examples in enterprise.

- A Fortune 25 telecom firm was unable to extract data from SAP ERP and PeopleSoft fast enough to its data lake. Laborious, multi-tier loading processes created day-long delays that interfered with financial reporting.
- A Fortune 100 food company ran nightly batch jobs that failed to reconcile orders and production line-items on time, slowing plant schedules and preventing accurate sales reports.
- One of the world's largest payment processors was losing margin on every transaction because it was unable to assess customer-creditworthiness in-house quickly enough. Instead, it had to pay an outside agency.
- A major European insurance company was losing customers because of delays in its retrieval of account information.

The advantages of change data capture.

Traditional, manual data integration processes can't meet today's data demands – but modern, automated technology can. One foundational technology for modernizing the data environment is change data capture (CDC), which continuously identifies and captures incremental changes to data and data structures from a source (or multiple sources) and replicates them to a target (or multiple targets), where the data can then be transformed and delivered to analytics applications.

When designed and implemented well, CDC enables efficient, low-latency data transfer to operational and analytics users, meeting all of today's requirements for scalability, real-time delivery and zero impact.

Why choose CDC over batch replication? It empowers you to:

Enable faster and more accurate decisions by allowing users to capitalize on the most current data available.

Minimize disruptions to production workloads by sending incremental source updates to analytics targets.

Save time and reduce costs by eliminating the need to transfer increasingly massive data stores from on-premises to the cloud.

Free up skilled resources for higher-value business projects by removing the need for manual scripting.

Changes, sources and targets.

Change data capture identifies and captures only the most recent production data and metadata changes that the source has registered during a given time period – typically, a span of seconds or minutes. Then it enables replication software to copy those changes to a separate data repository.

Types of data changes captured:

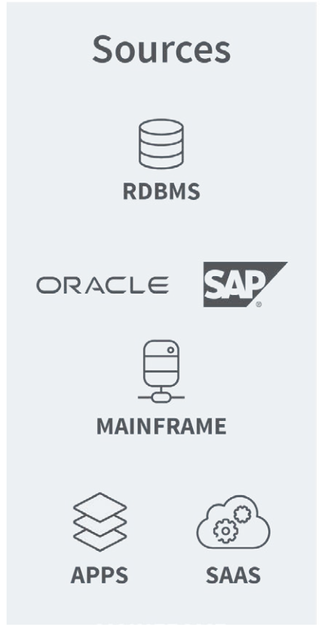
Row changes

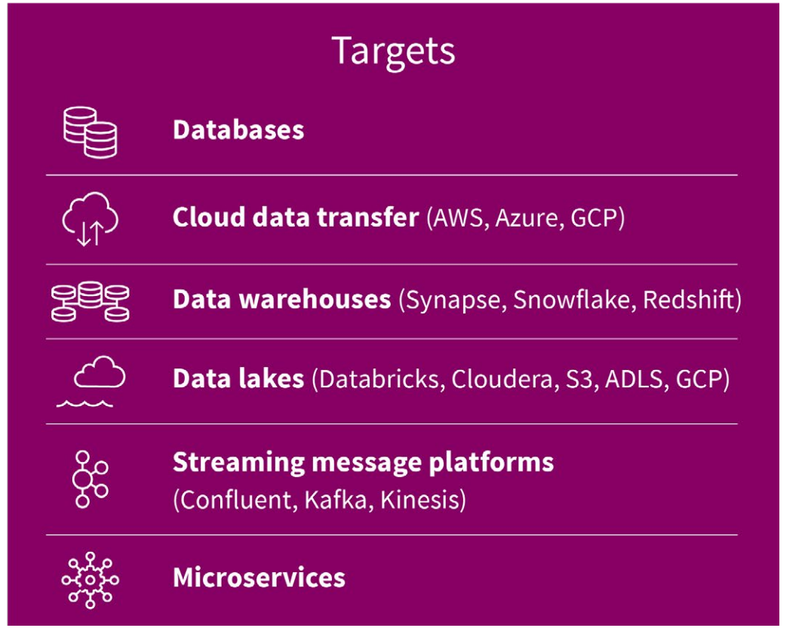
- **INSERTS:** Rows added to a database
- **UPDATES:** Fields changed in existing rows
- **DELETES:** Rows removed from a database

Metadata changes

- **DDL:** The creation, modification and/or removal of database objects like columns, tables and data types using the data definition language.

Sources and targets involved:





Capture and delivery methods.

There are three technology approaches to capturing changes in CDC – and some are more effective than others:

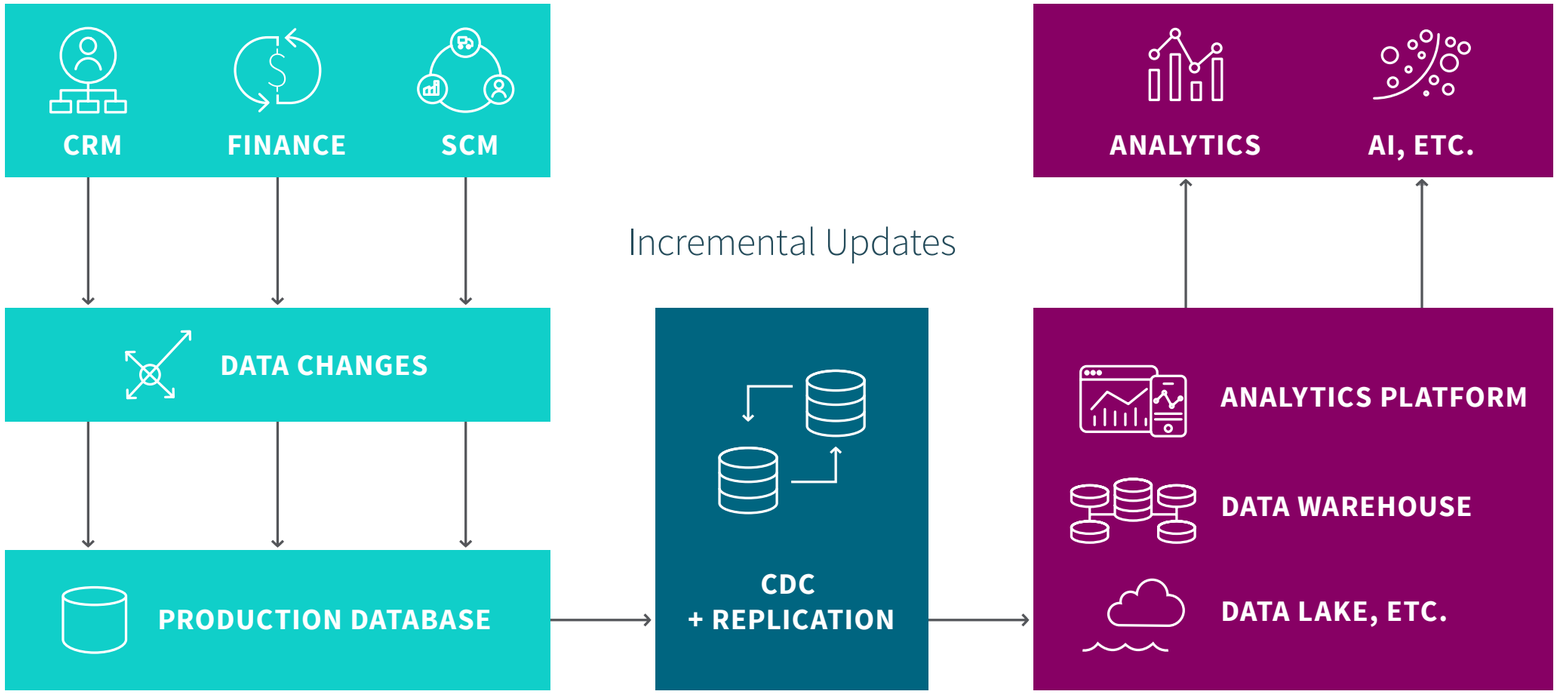
CAPTURE METHOD	PRODUCTION IMPACT
1. Trigger Source transactions "trigger" copies to change-capture table. A preferred method when there's no access to transaction logs.	Medium
2. Query The software flags new transactions in the production table colum with timestamps, version numbers, etc. and the CDC engine periodically asks the production database for updates.	Low
3. Log reader Changes are identified by scanning backup/ recovery transaction logs. The preferred method when log access is available.	Minimal

And there are three options for delivering the replicated data to the target:

DELIVERY METHOD	USE CASE
1. Transactional CDC copies updates – a.k.a. transactions – in the sequence in which they were applied to the source. Appropriate when sequential integrity is more important than ultra-high performance.	Daily financial reports, which need to reflect all completed transactions as of a single point in time.
2. Aggregated (Batch-Optimized) CDC bundles multiple source updates and sends them together to the target. Appropriate when performance is more important than sequential integrity on the target.	Aggregating trend analysis based on the most data points possible.
3. Stream-Optimized CDC replicates source updates into a message stream managed by platforms such as Kafka, Azure Event Hub and Amazon Kinesis. Unlike in the other methods, targets manage data in motion rather than data at rest.	A variety of new use cases, including real-time, location-based customer offers and analysis of continuous stock-trading data.

How change data capture works with analytics.

CDC has evolved to become a critical building block of modern data architectures.



This diagram presents a simplified view of CDC's role.

Agent-Based **vs.** Agentless

There are two primary architectural options for CDC software:

Agent-based. CDC resides on the source server and interacts directly with the production database. CDC agents are not ideal because they direct CPU, memory and storage away from source production workloads, degrading performance.

Agentless architecture is more modern and has zero footprint on the source or target. Instead, the software interacts with the source and target from a separate, intermediate server, minimizing impact and improving ease of use.

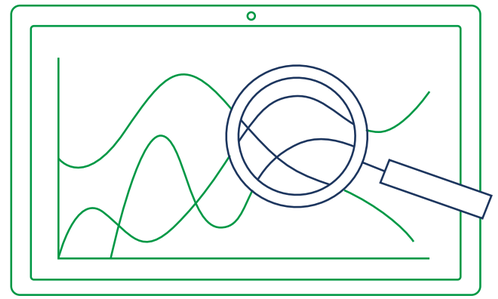
How change data capture fits into modern architectures.

The methods for data transfer vary by target.

TARGET	WHAT CDC DOES – AND HOW YOU BENEFIT
Replication to Databases	Copies the necessary records to reporting databases, enabling you to offload the queries and analytics workload from production – and take advantage of real-time operational dashboards.
Publication to Streaming Platforms	Converts source updates into a stream of messages that can be divided into topics. As a result, you can deliver near-zero latency in real-time analytics.
Microservices	Delivers and synchronizes data across the specialized microservice data repositories, which allows you to provide granular services to a wide range of customers.
Cloud Data Transfer	Continuously synchronizes on-premises and cloud data repositories, enabling you to eliminate repeated, disruptive batch jobs and deliver zero downtime.
ETL and the Data Warehouse	Spreads data extraction and loading over time by continuously processing incremental updates – so you can support real-time transformation while reducing performance impact.
Data Lake Ingestion	Ingests big and wide data into the data lake in near real time as changes occur – empowering your business teams to perform the rapid and real-time analytics that require the very latest data.

Introducing Qlik for CDC.

Qlik Data Integration for CDC Streaming provides a simple, universal solution for converting production databases to live data streams to support modern analytics and microservices. With Qlik, data engineers can entirely automate the end-to-end movement of data – in real time – from multiple sources to multiple targets. Streamlined and agentless configuration, together with a simple graphical interface, make it easy to set up, control and monitor data pipelines.



Qlik enables data management teams to:

Flexibly support one-to-many publication, automated data type mapping and comprehensive metadata integration, all with no hand-coding

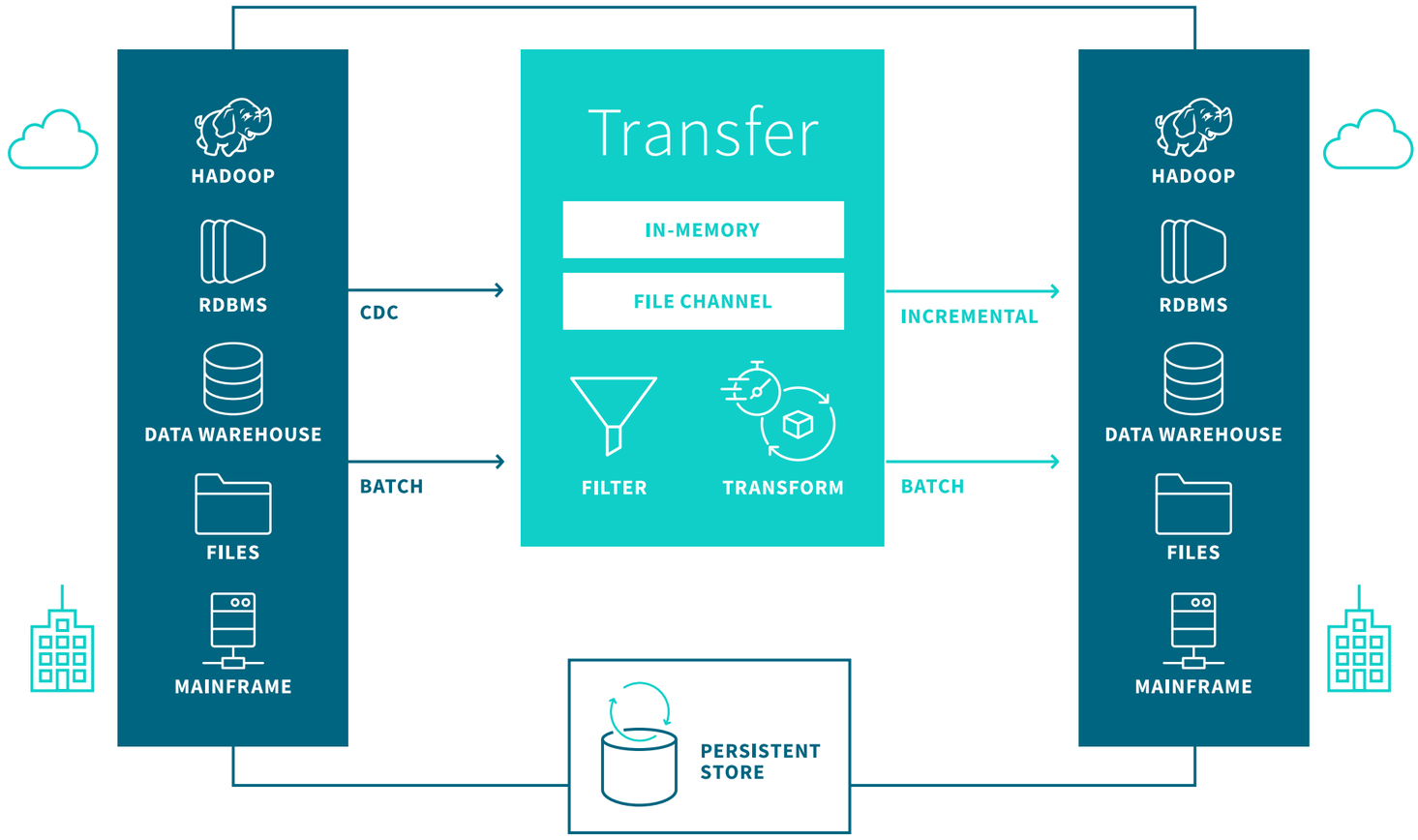
Minimize impact to sources with zero-footprint, log-based CDC

Free up time for big-value projects by greatly reducing your workload, with a 100% automated process and intuitive GUI

Take advantage of cloud optimization and broad platform support, including all major structured data sources (Confluent/Apache Kafka, Amazon Kinesis and Azure Event Hub)

CDC for real-time data transfer

Qlik for CDC Streaming resides on an intermediate server that sits between one or more sources and one or more targets. With the exception of SAP sources, which have special native requirements, no agent software is required on either source or target. Qlik's CDC mechanism captures data and metadata changes through the least-disruptive method possible for each specific source – usually its log reader.



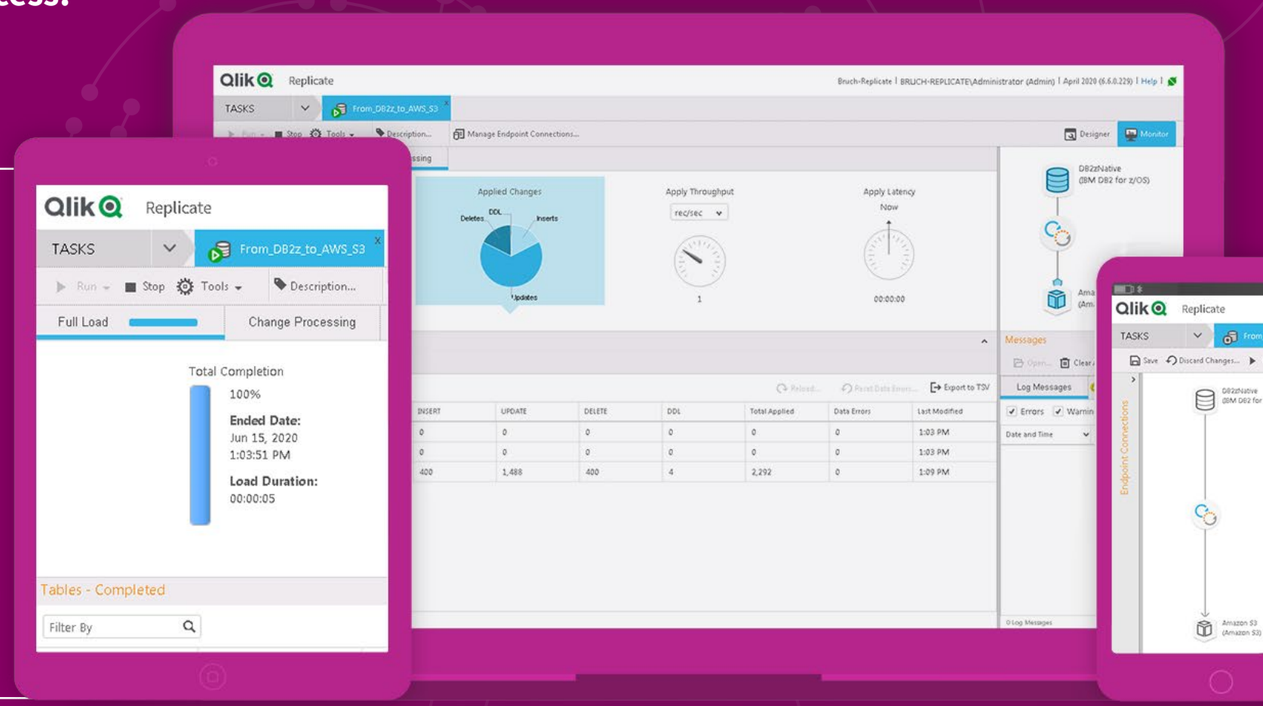
Revolutionizing data delivery.

Organizations around the world are using Qlik for CDC Streaming to reinvent the way they deliver data. Rather than settling for labor-intensive manual coding and the delays inherent in batch processing, they're establishing continuous, real-time streams of reliable data and making it available to business teams across the enterprise for use in analytics and microservices. And they're seeing remarkable results.

In the following pages, you'll find four stories of change data capture success.

The complete platform

The ability to continuously replicate your data is one component of the Qlik Data Integration Platform. You can further accelerate your data pipelines – dramatically – with Qlik's solutions for data warehouse automation and data lake creation.



Powering growth with real-time insights at Aggreko.

A leading global provider of power, heating and cooling solutions, Aggreko was on a mission to become increasingly insight-driven. To do that, they first had to centralize their global operations data.

CHALLENGE

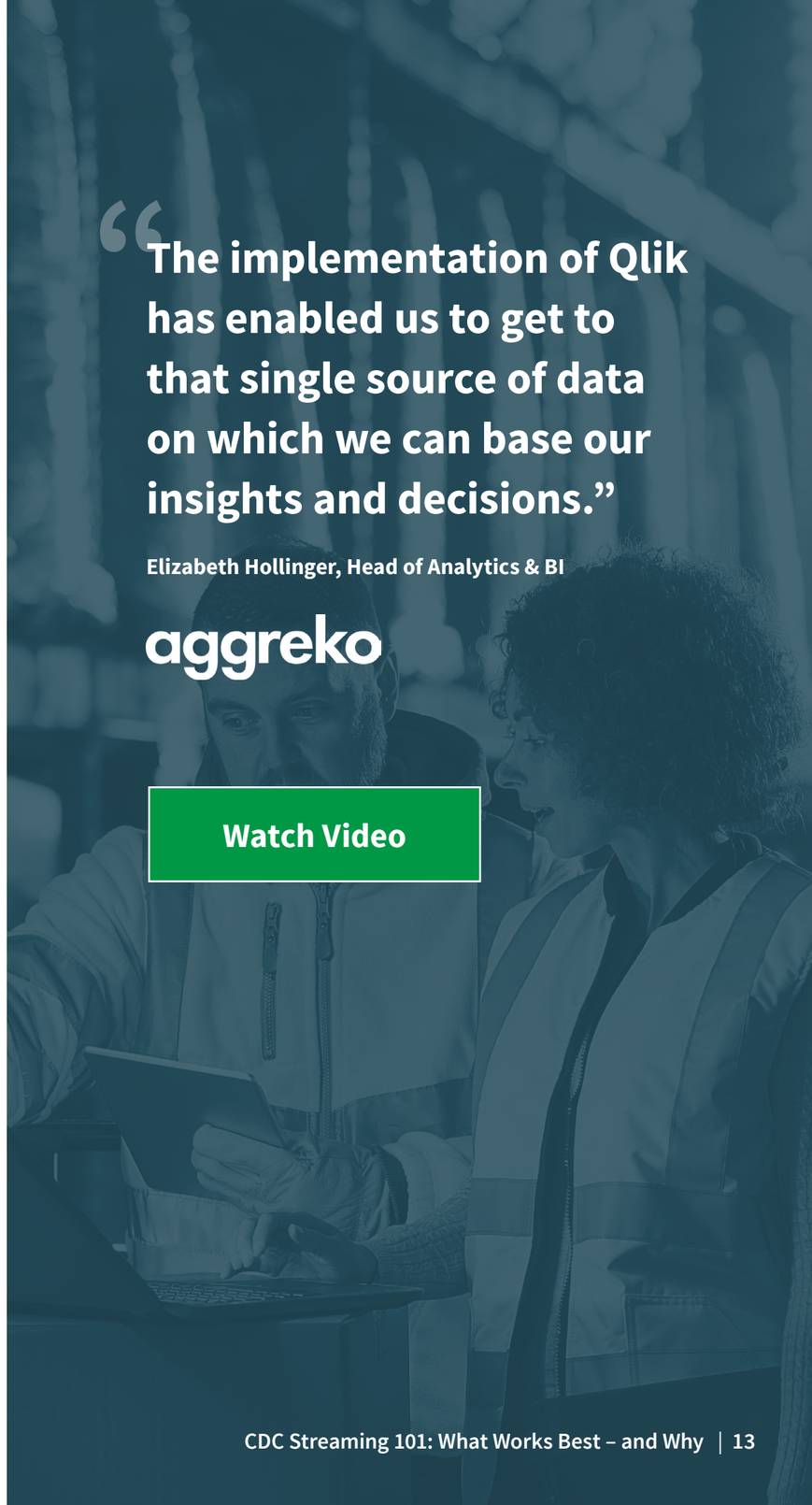
In the quest to establish a clean and consistent process for ingesting data from their source systems, Aggreko's team struggled with the varied requirements of each pipeline. It looked as though they would have to develop a custom solution for every data source, which would quickly become unmanageable.

SOLUTION

The search was on to find a single product that could manage all ingestion pipelines. The team chose Qlik for two reasons: because it's extremely lightweight, with a minimal impact on production systems, and because its ease of implementation would help Aggreko's small team of data engineers do a big job, quickly.

RESULTS

Qlik CDC Streaming was implemented for Aggreko's ERP system within a single week, and no production systems were impacted. Aggreko can now deliver insights in real time and plans to extend the solution to other data sources.



Reducing source-to-target time to under 10 seconds at Generali.

The Swiss arm of a global insurance leader, Generali Switzerland employs 1,800 people across 56 locations and serves 1+ million customers.

CHALLENGE

When core legacy applications were increasingly placing a drag on customer-facing apps and channel applications, the company set out to modernize their data architecture. Their goals: 1) provide up-to-date, high-quality data via multiple channels, and 2) improve IT service delivery.

SOLUTION

From the outset, there was a great deal of data to replicate, and this meant significant integration. The team developed a hybrid connection platform powered by Qlik and Confluent (Apache Kafka), in part because the two solutions work harmoniously together. Qlik sits between the data sources and target destinations, reading and understanding the information and sending it to Generali's target of choice.

RESULTS

Source-to-target time has been reduced from days to under 10 seconds. Data is extracted with no disruptions to the business or applications. A single version of data truth is now available via many applications and channels. With access to accurate, real-time data, customer service and engagement have improved. And finally, Generali can support agile solutions development.

Speeding data integration and boosting analytics performance at Veritix (AXS).

Veritix (now part of AXS) creates digital ticketing, event marketing and CRM applications for professional sports teams and entertainment venues around the world. With their analytics reporting, their customers can target demographic groups more effectively, improve season ticket renewal rates and bring new fans into arenas.

CHALLENGE

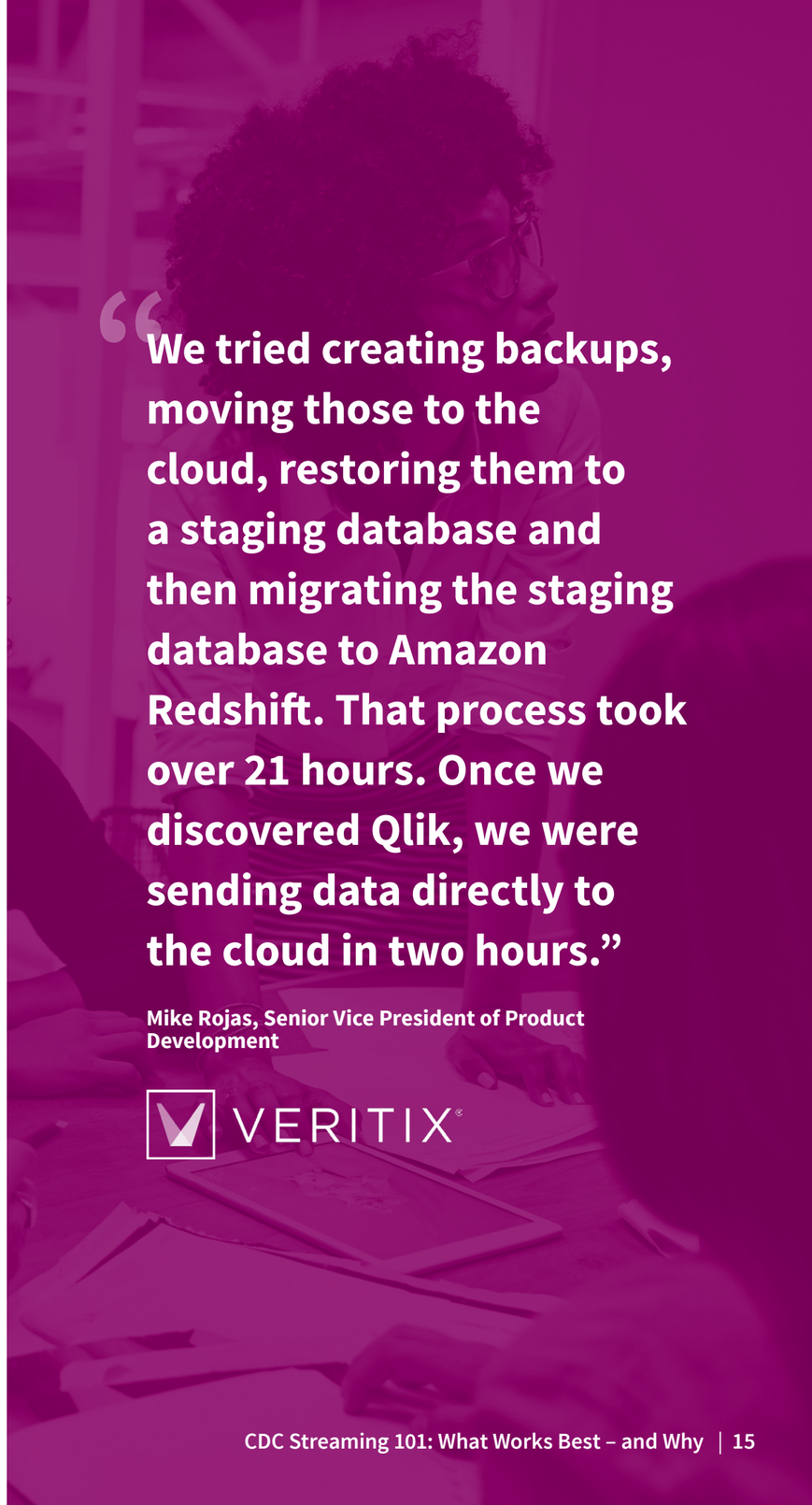
The transactional database at Veritix was used for both production purposes and reporting. As a result, it wasn't optimized for analytics, and performance was a concern. So the team decided to create a separate data warehouse and selected Amazon Redshift for the job.

SOLUTION

Initially, data engineers at Veritix assumed that they would need to build custom tools or rely on an off-the-shelf solution to replicate data to the cloud. They made a first attempt, and the process took 21 hours. Then they discovered Qlik – which took just two hours to install, configure and move hundreds of millions of records to Redshift.

RESULTS

The data warehouse now contains around three terabytes of information, with billions of records available for customer analytics. And not only has Qlik met Veritix's performance requirements; it has also exceeded the team's expectations in ease of use.



Introducing the Qlik Data Integration Platform.

As you modernize your data environment, CDC replication is one piece of the puzzle – or more specifically, one piece of the pipeline. At Qlik, we've built an end-to-end Data Integration Platform that accelerates the discovery and availability of analytics-ready data by automating not only real-time data streaming but also data refinement, cataloging and publishing, too.

Real-Time Data Streaming
Extend enterprise data into live streams to enable modern analytics and microservices.

Managed Data Lake Creation
Automate complex ingestion and transformation processes to provide analytics-ready data lakes.

Agile Data Warehouse Automation
Quickly design, build, deploy and manage purpose-built data warehouses without manual coding.

Enterprise Data Catalog
Enable users across your business to easily find, prepare and share analytics-ready data.

All elements of the Qlik Data Integration Platform work together, enabling you to establish real-time, automated data pipelines that stream from transactional systems, warehouses or data lakes to create trusted, actionable data on demand throughout your organization.

Take advantage of the latest analytics innovations – with the latest data integration technology.

The foundation for modern, efficient data pipelines, Qlik's CDC solution provides automated, real-time and universal data integration across all major data architecture, on-premises and in the cloud. It moves data at high speed. It's simply and easily managed. And it gives you global visibility into and centralized control of data replication across your distributed, hybrid environment.

WHERE TO GO NEXT

Unlock the value of the data in your legacy sources. Enable major data integration projects. Minimize the impact to your production systems. Eliminate direct queries. Save time. Reduce the back-office IT workload. And support your line-of-business teams with the flexibility they need, all while maintaining data security.

Curious to learn more?

Visit Website

Or dive right in, and try Qlik CDC Streaming for yourself.

Free Trial

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. Qlik provides an end-to-end, real-time data integration and analytics cloud platform 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 does business in more than 100 countries and serves over 50,000 customers around the world.

Qlik LEAD WITH DATA™

qlik.com