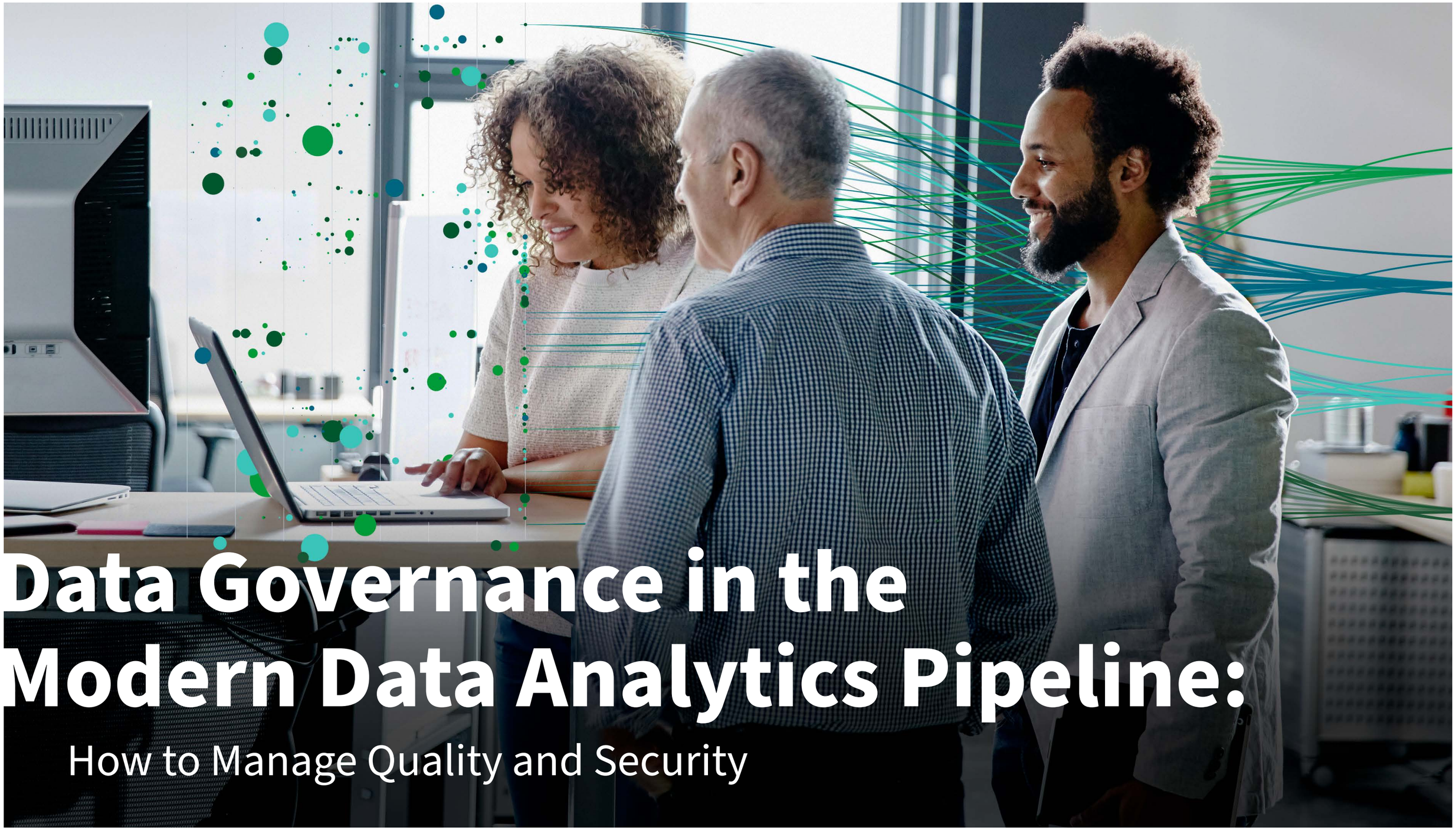


Data Governance in the Modern Data Analytics Pipeline:

How to Manage Quality and Security

The business is relying on data. And users want it fast.

Today's C-suite is increasingly focused on leveraging data for business-critical initiatives. And for good reason. In organizations that have invested strongly in data management and analytics[1]:

76% report that operational efficiency has improved

75% report that revenue has increased

74% report that profit has increased

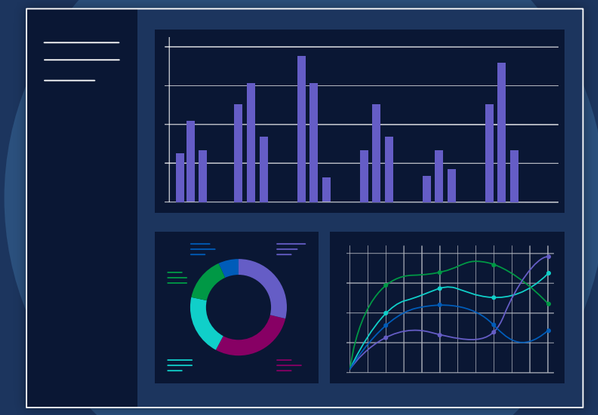
For these ever-increasing data-driven initiatives, speed is a top priority. In fact, when naming their top digital ambitions, 29% of executives cite speed.[2] But in data as in life, speed introduces risk. In the case of data, speed introduces risks in both quality and security. That's why it's not surprising that as they ramp up more ambitious digital initiatives, executives are also investing heavily in security[3]:

55% are increasing security budget

51% are increasing security headcount

50% report that security will be baked into every decision or plan

The answer: modern, secure, efficient data pipelines.



As a result of all the new data initiatives, IT and data teams are seeing a dramatic increase in data demands. And that's on top of the massive explosion in data volume and the vast array of new capabilities.

How can you balance these rising demands for data – at speed – against the ever-present quality risks and the ever-growing security threats? With modern, secure, efficient data-to-analytics pipelines. But what are those pipelines, exactly? They're more than just data migration.

The Modern Data Analytics Pipeline

IDENTIFY DATA

Starts by identifying data, both internally and externally, that could be valuable to the organization.

GATHER DATA

Ingests raw data in a broad range of formats from a vast array of sources.

TRANSFORM DATA

Applies a set of actions to make the data understandable and useful to a business user. Actions can include but are not limited to: transformation, standardization, sorting, deduplication, filtering, validation, and verification.

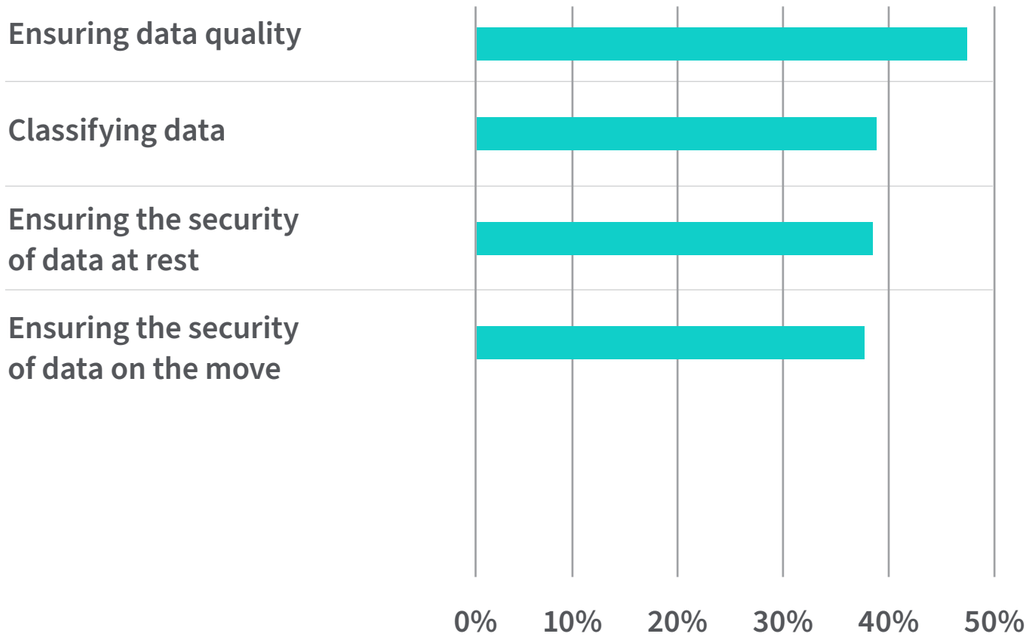
DELIVER DATA

And finally, the pipeline applies the analytics-ready data to a store (e.g., a data warehouse or lake) or directly into an analytics application.

Creating valuable data: The struggle is real.

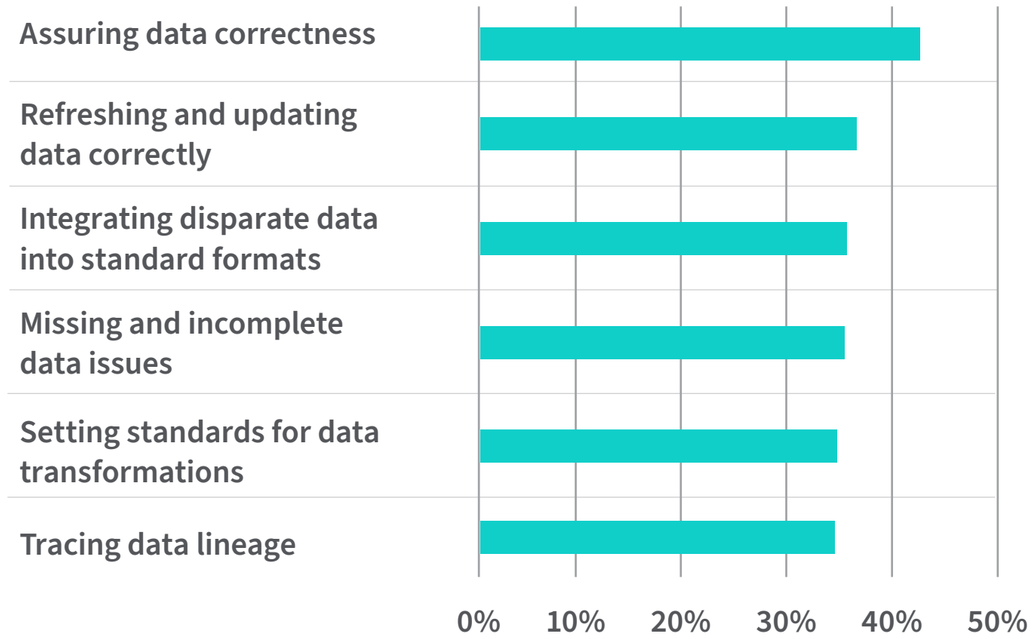
Today, over 60% of organizations experience significant challenges in assessing the value of data and identifying valuable sources. When IDC asked data engineers about their main challenges in capturing new data and moving it before processing, this is what they heard[4]:

Main challenges in capturing new data and moving it before processing



And there are just as many challenges farther along in the pipeline. When IDC asked data engineers about their top challenges in processing or transforming the data into an analytics-ready form, they heard the following[4]:

Main challenges in processing data into an analytics-ready form



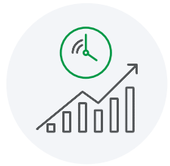
Data quality defined.

The challenges of establishing a data analytics pipeline can be divided into two categories: data quality and data security. Let's look at quality first.

Data quality can be defined as the condition of the data on several axes:

How accurate it is

How complete it is

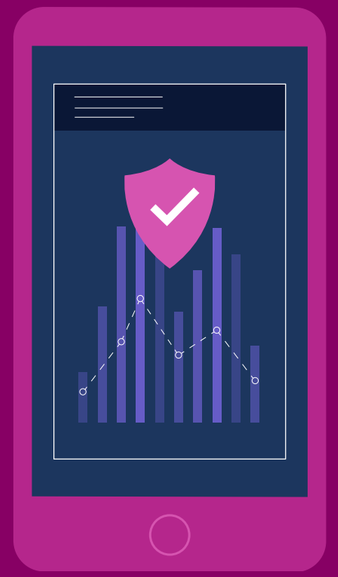


How current it is

Measuring and maintaining high data quality helps organizations identify (and resolve) errors before the data is analyzed, when those errors will impact the business. And importantly, consistently delivering high-quality data also builds trust among data users.

> **"Data in the wild is often not well understood, or even usable. This can be because it's in a structure that business analysts don't understand. In these cases, data has to be profiled to assure its quality. Otherwise, you can fall foul to the most common reason data analytics projects fail to meet their objectives: the quality of data not being good enough."**

Joe DosSantos, CDO, Qlik



Trust matters. A lot.

> Data debt – the cost attached to poor governance of data in a business – is a challenge for 78% of organizations … [And] two in five organizations say individuals within the business do not trust data insights."

Experian[5]

How to deliver high-quality data.

A modern data analytics pipeline delivers high-quality data by offering – and automating – all of the following capabilities:

PROFILING

- Assess the data – how bad is it? – and categorize clean vs. dirty records
- Quarantine data that doesn't meet quality standards
- Identify additional issues the architect may need to address

FILTERING AND CLEANSING

- Refine and merge data into analytics-ready structures
- Automatically clean the dirty records according to organization-specific rules

TRANSFORMATION

- Perform lightweight transformations on the fly, such as filtering (e.g., by current year) and apply global transformations (e.g., consistency with one date format)
- Use data quality rules to trim, merge, and calculate
- Dump problem data into error marts
- Transform data via business rules; once defined, the rules can be broadened to other pipelines

STANDARDIZATION

- Apply business standards across all data
- For example, make sure that the same data filters are used against all sources

Data security within the modern analytics pipeline.

Establishing data security requires protecting data from unauthorized access at every step of the journey from raw to ready. A modern data analytics pipeline protects data security in the following ways:

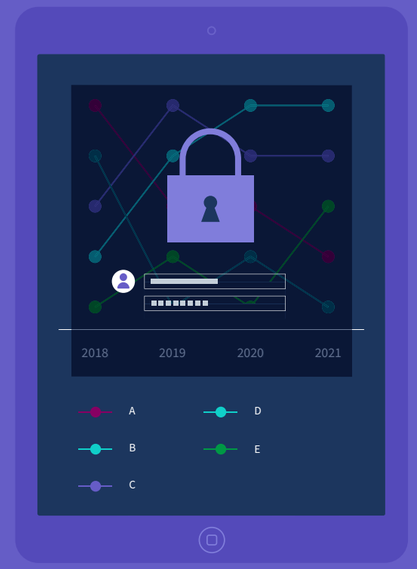
1. **It authenticates users, verifying that they are who they say they are**
2. **It provides detailed access controls that can be enabled by role or individual user**
3. **It offers data-field obfuscation, masking sensitive information**
4. **It encrypts data at motion and at rest (i.e., during the transport and storage stages)**

→ **Important for GDPR compliance**

The cost of a data breach.

In 2020, the average cost of a data breach was **$3.86 million,** and the average time to recover was 280 days.

IBM[6]



Want security? Control access.

47% of global business leaders say that giving the right people access to the right data at the right time is a significant challenge. Whenever data is in play, establishing correct access is just as important for security and compliance as it is for insights.

IDC[7]

For quality and security, get data governance.

How can you ensure both data quality and data security across the modern data analytics pipeline? With data governance – the exercising of authority and control over data assets. That includes tracking, maintaining, and protecting data at every stage of the lifecycle. And this is important: You have to control not just which functionality each person can use but also which data sets they can access.

As it is elsewhere, in a data analytics pipeline, governance is about enabling users to get what they need, not preventing them from doing so. What does that look like?

ACCESS

Can you offer table-based or column-based access control?

ORGANIZATION

Can you arrange data into understandable and accessible datasets?

LINEAGE

Can you leverage metadata to automatically build data lineage views?

AVAILABILITY

Can you create a single, trusted view of data where users can “shop” for what they need?

Two indispensable tools: automation and cataloging.

Two modern technologies help you achieve successful governance across the data pipeline:

1

AUTOMATION

The latest technology for data delivery provides a new set of automation features, and moving away from manual processes gives you more control – in several ways.

Automation allows embedded governance with rules and policies to influence discovery early and often. It can inject data-quality improvements based on the proper context for individuals and workflows. And when you automate data property identification along the pipeline, you prevent users from seeing what they shouldn't.

2

CATALOGING

How can you make different kinds of data available to different users quickly, without adding risk? Establish a use case-based catalog with the ability to onboard, profile, describe, secure, and even prepare and obfuscate data quickly in anticipation of analytics sprints.

Catalogs enable everyone within an organization to see which data is available to them in a single, understandable marketplace where they can "shop" for the sets they need. Catalogs manage, secure, and control access through enterprise-wide data schemas. And they enforce security by identifying and masking the data on user types and access rights.

The data catalog as enabler.

In a data catalog, governance can set limits without stifling availability. In fact, it can provide access to more people, because the data is trusted as employees encounter it.

Rapid, secure, and governed data delivery for next-generation analytics.

With the end-to-end Qlik® Data Integration Platform, you can vastly accelerate the availability of real-time, analytics-ready data by automating data streaming, refinement, cataloging, and publishing. And because it's an open platform, Qlik Data Integration can integrate with your cloud platform or existing third-party quality, security, and governance services.

Qlik's Data Integration platform includes Qlik Catalog, which creates an enterprise-scale repository of all the data your business has available for analytics. It provides data consumers with a single, go-to catalog where they can find, understand, and gain insights from any underlying enterprise data source.

This infrastructure enables the creation of reliable, governed modern data pipelines. And it gives you a head start on the journey to Active Intelligence – a state of continuous intelligence that uses real-time data pipelines to trigger immediate action.

Explore Qlik Data Integration | **Explore Qlik Catalog**

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 offers real-time data integration and analytics solutions, powered by Qlik Cloud, 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 serves more than 38,000 active customers in over 100 countries.

qlik.com

[1] IDC, Infobrief sponsored by Qlik, *"Data as the New Water: The Importance of Investing in Data and Analytics Pipelines,"* June 2020.

[2] PwC, *"2021 Global Digital Trust Insights: Cybersecurity comes of age,"* https://www.pwc.com/us/en/services/consulting/cybersecurity-privacy-forensics/library/global-digital-trust-insights.html.

[3] PwC, *"2021 Global Digital Trust Insights: Cybersecurity comes of age,"* https://www.pwc.com/us/en/services/consulting/cybersecurity-privacy-forensics/library/global-digital-trust-insights.html.

[4] IDC, Infobrief sponsored by Qlik, *"Data as the New Water: The Importance of Investing in Data and Analytics Pipelines,"* June 2020.

[5] Experian, *"The Cost of Data Debt Rises as Businesses Face the Challenge of Data Literacy,"* February 2020, https://www.experianplc.com/media/news/2020/the-cost-of-data-debt-rises-as-businesses-face-the-challenge-of-low-data-literacy/.

[6] IBM, *"Cost of a Data Breach Report 2020,"* https://www.ibm.com/security/data-breach.

[7] IDC, Infobrief sponsored by Qlik, *"Data as the New Water: The Importance of Investing in Data and Analytics Pipelines,"* June 2020.