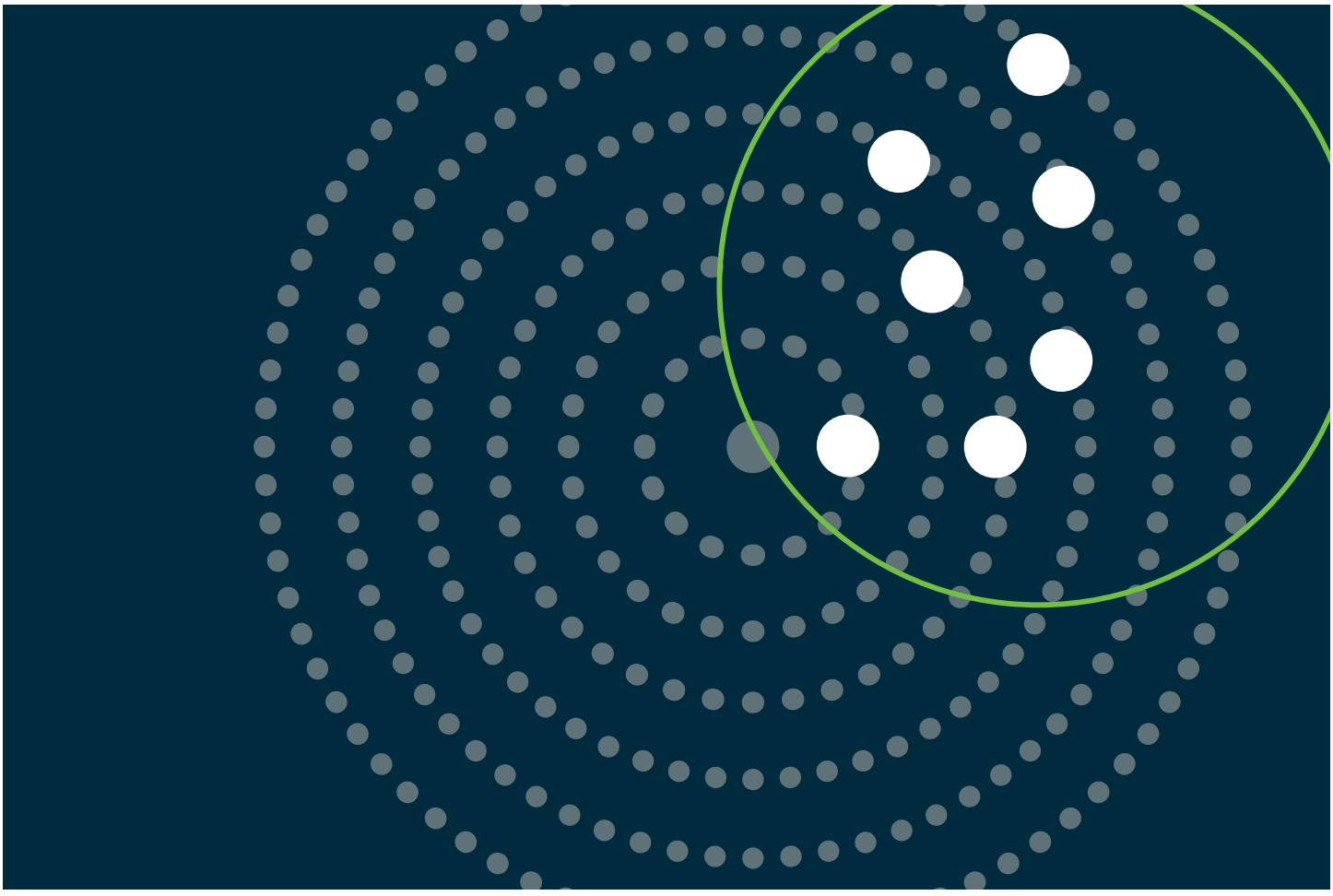
Collibra
United by Data™

Whitepaper

Enterprise-scale Data and Analytics in the Cloud

Three key steps for successfully moving to the cloud



Executive summary

Cloud platforms are increasingly becoming the logical choice for enterprises. Cloud service providers help abstract away the complexity and upfront technological investment to deliver ubiquitous data access with sophisticated analytical tools.

Enterprise-scale data and analytics in the cloud offer unlimited storage and computing power to leverage machine learning. For data producers and consumers, the cloud presents tremendous opportunities for business transformation.

The real-world migration to the cloud is significantly complex, considering the scale and diversity of data sources. A traditional lift-and-shift approach can only move the data, but fails to unlock its potential value. A transformational approach with the right technology can maximize the value of data and analytics in the cloud.

This paper describes:

- Four challenges organizations face in migrating data to the cloud
- Three key steps for successfully migrating data to the cloud
- Adoption of enterprise-scale data and analytics in the cloud

"Business leaders are beginning to understand the importance of using data and analytics to accelerate digital business initiatives."

- Gartner Top 10 Data and Analytics Trends for 2021

Introduction

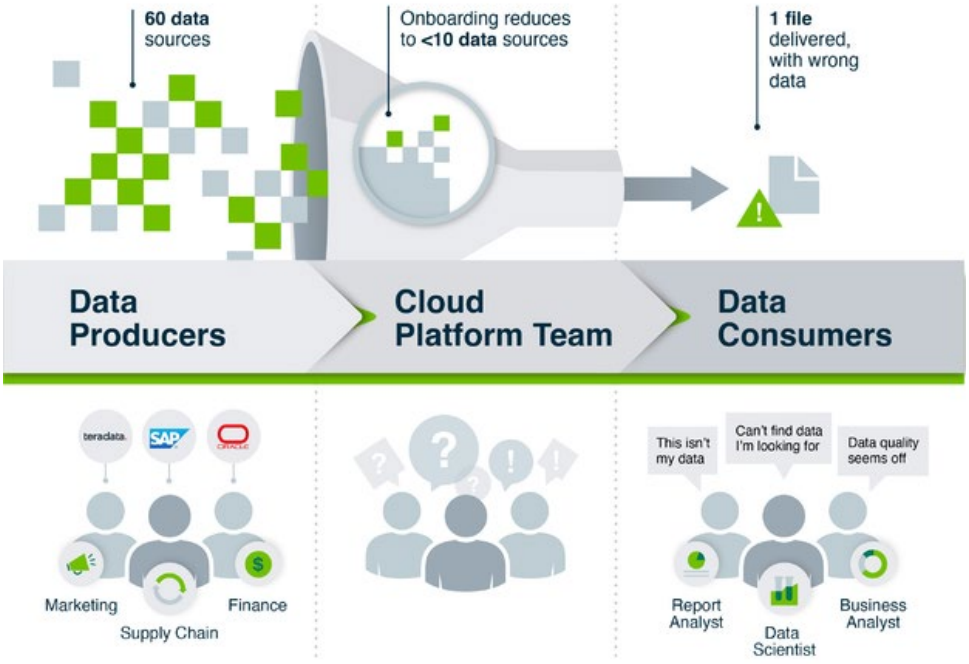
Moving to the cloud is a key element of digital transformation to manage the expanding data volumes and growing analytical complexity. Cloud data platforms empower data consumers - business analysts and data scientists, to shop for business-ready data from a single location. With all their needs for data and analysis fulfilled in one place, they utilize their time more efficiently and focus on critical business challenges. Rapid scaling in the cloud delivers the agility to respond to the ever-changing business landscape.

Automatic data quality capabilites integrated with machine learning technologies and analytics in the cloud produce outstanding benefits. Data quality tools in the cloud help manage streaming data in real time, scale with large volumes of data arriving from multiple channels, and create high-quality data pipelines for trusted cloud analytics.

Organizations are steadily migrating to the cloud and adopting a cloud-first approach for data and analytics. An increasing number of enterprises are turning to Google Cloud's Managed Open Analytics (Dataproc), Managed Lake Solutions (Dataplex), and serverless Data Warehouse (BigQuery) to take advantage of the cost-efficiency and scale offered by Google Cloud. Cloud's agility enables organizations to rapidly achieve scalability and business agility, and turn enterprise data into valuable business insights.

Four challenges of cloud migration

Google Cloud offers sophisticated tools for quickly onboarding data streams, pre-processing streaming data and data visualization. While these tools simplify the migration process, organizations still need to tackle four fundamental challenges.



Victims of their own success, cloud platform teams get overloaded due to lack of governance and a lift and shift to the cloud approach.

1. Data discovery and understanding

Cloud platforms offer data aggregation across siloed systems and operations. Yet, finding the right data is often hard considering the gigantic scale of data in the cloud. Data discovery is surprisingly still a challenge when organizations do not use dedicated tools for it. Data consumers struggle to discover the right data for their project, understand it in the business context and request access if they are provisioned to it. Unlocking the value of enterprise data begins with understating it and then cataloging it for easy data discovery.

2. Data quality and interpretation

The quality of data is a broad term, and its measurement depends on the dimensions that matter to the organization. The typical data quality dimensions include completeness, accuracy, consistency, validity, uniqueness and integrity. For data consumers, the standard quality dimensions that measure if data is “right” is not enough. Data accessibility, timeliness and relevance play a substantial role in shopping for the “right” data.

When enterprise data gets aggregated in the cloud, assuring data quality becomes critical. Uniform interpretation of data becomes challenging without a shared understanding and a common business glossary because definitions and metrics inevitably have nuanced differences across silos. This can introduce inconsistencies in data.

3. Data ownership

A fundamental challenge during migration is about who owns what data, and who is responsible for data quality. Without properly assigned roles and responsibilities, data ownership can be distributed over several people without any accountability or governance. Only the right data owners can appreciate the business impact of data quality or availability. During migration, data owners are responsible for raising concerns over data quality, initiating quality improvement and communicating the progress. The absence of ownership can lead to poor control over source data, leading to overlaps, redundancy or missing data.

4. Data access

The rapidly evolving regulatory requirements such as GDPR and CCPA provide data subjects greater control over their sensitive information. Concerns about regulatory compliance combined with internal policy enforcement, result in organizations imposing unnecessary access restrictions. Unless everyone understands the possible threats, organizations are unable to offer appropriate access to sensitive data.

Any uncertainty over the access to sensitive data results in too much control or a long-drawn, inefficient, compliance process. Enterprises move data to Google Cloud for ease of access and scaling. Though without the proper access mechanism, data assets may get restricted to a select group, limiting the opportunity to unlock their value.

Three key steps to a successful cloud migration

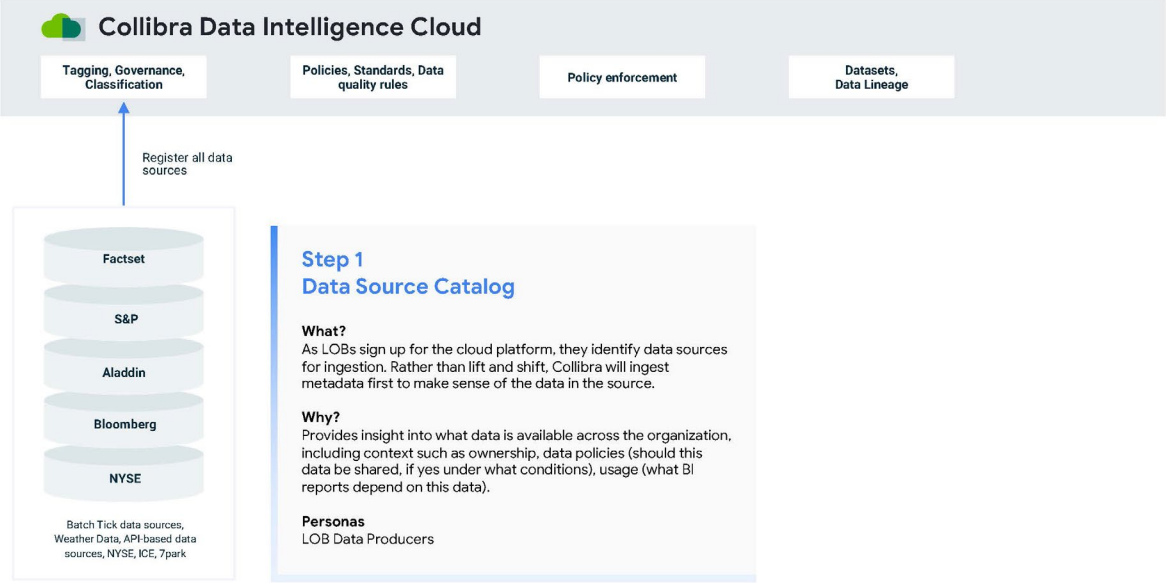
The traditional lift-and-shift approach to migration can deliver only limited benefits, without unlocking the real value of data. Only a business transformational strategy with the right foundational technology can maximize the rewards of cloud migration.

Collibra recommends a comprehensive approach for migrating data and analytics to the cloud, supported by governance and data quality.

1. Data source registration

The first step of the migration process is to inventory enterprise data assets. LOB data producers identify data sources for ingestion, then tag and classify them. A wide range of data sources can be registered including reports, analytical models, data from sensors or legacy data.

Data source registration provides insights into what data assets are available in the organization, with the context of ownership and data policies. The context also includes usage information, such as which analytical reports depend on which data. Registering data sources provide data citizens with a complete understanding of the data at their disposal.



Google Cloud

Collibra provides the ability to rapidly catalog data stored in popular enterprise SaaS applications, which can be in different file formats. With native integrations with leading ETL tools (PowerCenter and BI tools) such as Tableau and Looker, Collibra helps orchestrate data pipelines, register analytical assets and trace data lineage from report to source.

Collibra offers native connectors to register data sources in Collibra Data Catalog and automatically ingest technical metadata, making the inventory process fast and scalable. Profile and classify on-prem resources such as Teradata, and classify other enterprise silos already in Google Cloud Storage or BigQuery.

At the end of this step, enterprise data assets are fully cataloged and ready for moving to the cloud.

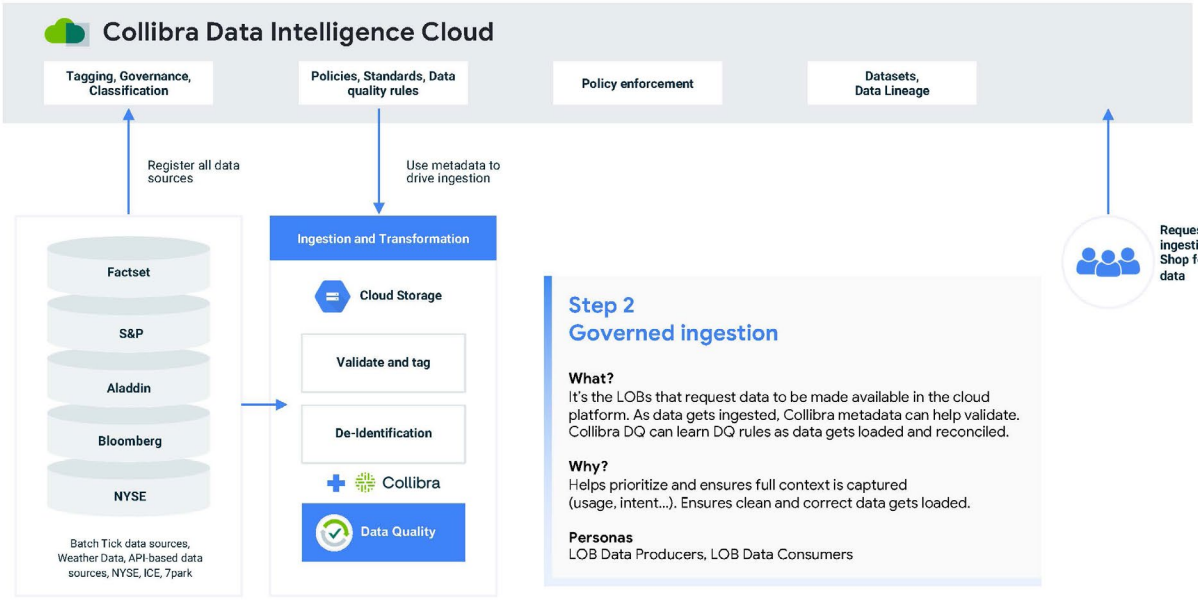
2. Governed ingestion and transformation

At the next step, LOB data producers and data consumers work together on metadata-driven ingestion. An important point to note is that Collibra first ingests metadata to give a context to the source data that gets ingested later. Metadata validates data as it gets ingested. Governed ingestion and transformation help prioritize issues to ensure that clean and correct data is loaded.

Cataloging, profiling and classifying metadata provides information on the type of data. Linking technical metadata to the logical and conceptual model builds up the full context for data. At the same time, data is mapped to the business glossary so that it is discoverable by business terms.

With Collibra, you can create policies about using data and retaining it for a defined time. You can map the policies to data for assured compliance, and data always gets used with the mapped policies. If the source data contains any personally identifiable information (PII), it can be recognized and de-identified, ensuring that the data loaded is secure and privacy-compliant.

Collibra systematizes enterprise data assets for compliant use with a fully built data context.



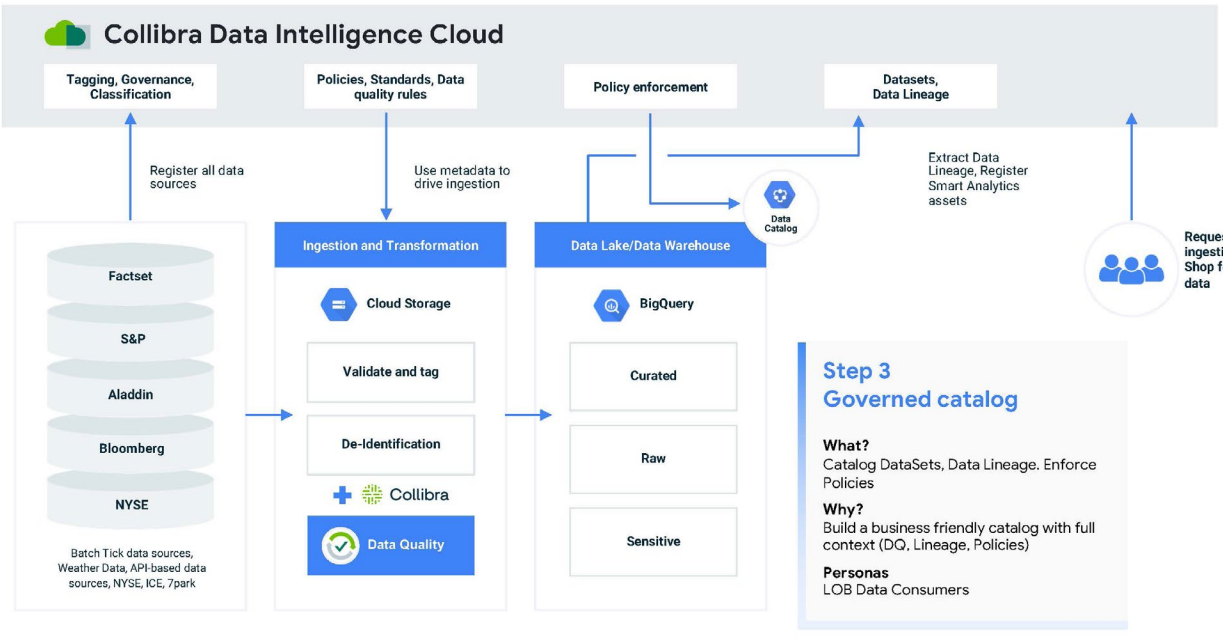
While data gets ingested, Collibra Predictive Data Quality leverages ML to auto-generate SQL-based, explainable and adaptive rules. The rules keep learning as data gets loaded and reconciled. The developed rules are uniquely adapted to the migrated data and later work with the streaming data. The Spark-based architecture of Collibra Predictive Data Quality supports alerting issues at the source. Quick alerts on dashboards and the ability to self-service fixes ensure issues get addressed at source. With ML-driven issue identification, classification and impact-based prioritization, along with automated anomaly detection, Collibra Predictive Data Quality delivers scalable continuous quality in the cloud.

At the end of this step, high-quality data is loaded and made available in the cloud. The predictive data quality rules are ready for continuous monitoring and delivering high-quality data pipelines.

3. Governed catalog for data lake

After having data migrated to the cloud platforms such as Google Cloud, organizations want to maximize its value by making it widely accessible. The typical challenge at this stage is about efficient access control that does not limit the data use while ensuring privacy compliance. Building a business-friendly catalog with the full context, lineage, data quality and policies enables LOB data consumers to access and shop for relevant data fueling their projects.

With Collibra, organizations can centrally manage data policies and accurately identify data for compliance, including data classes such as PII or PHI. They can establish usage restrictions for sensitive data by defining access controls that comply with relevant policies.



At this stage, a successful cloud data migration provides a governed catalog of curated data with policy-driven sensitive data access. Data analysts and data scientists can easily choose the relevant data sets from the governed catalog in agile and scalable cloud data platforms such as Google BigQuery. They can drive their analysis in the cloud-native analytics platforms, Tableau or Looker, with complete intelligence on data quality and privacy.

Collibra integrates directly with services such as BigQuery, Dataproc, Dataplex, Spanner, and CloudSQL. It can also utilize metadata coming out of Google Data Catalog.

Users working within the Google platform can leverage the GCP Data catalog for rapid search and discovery. They can further enrich the data with Collibra Data Catalog. On the enterprise level, users can work directly in Collibra and find rich context coming from the GCP Data Catalog.

Diverse industries have leveraged Collibra and Google Cloud for successful cloud adoption, enabling self-service access to governed data and further innovation.

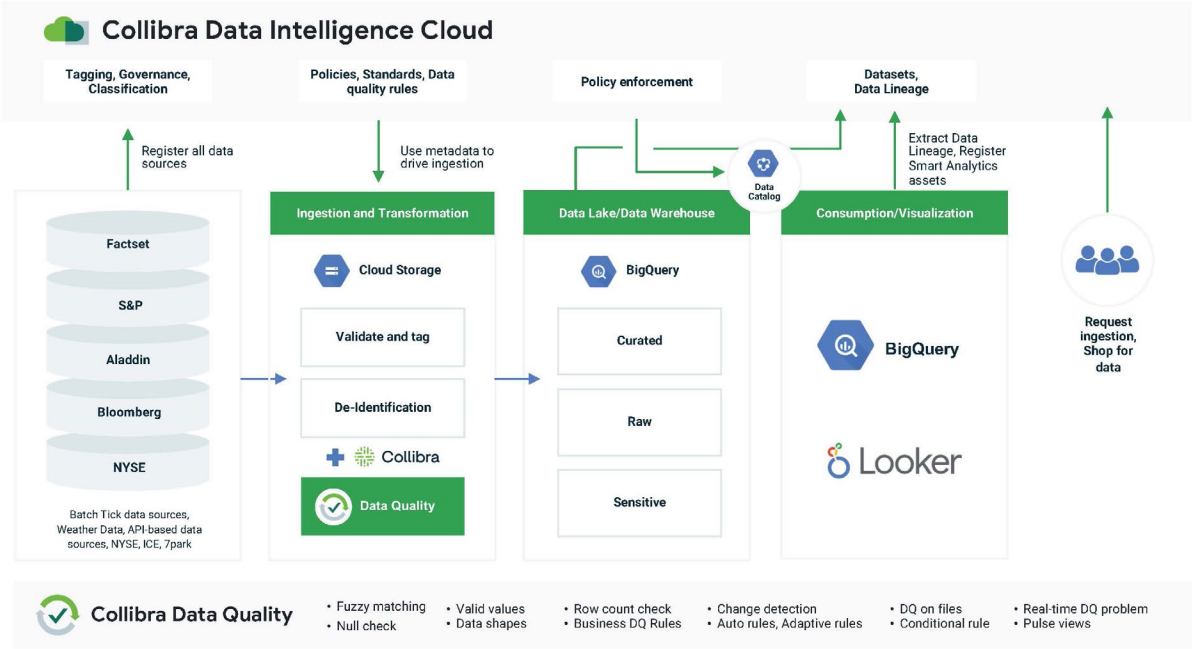
Enabling enterprise-scale data and analytics in the cloud

Gartner recommends cloud-first to leverage the benefits of costs, agility, scalability, security and accessibility.

Analytics in the cloud requires engineering high-quality data pipelines, either scheduled or ad hoc. A typical example of a scheduled pipeline is extracting sales data from diverse sources and uploading it to a cloud data warehouse. It supports the regular business reporting function. Automating the scheduled pipelines is relatively straightforward, considering the task is well-defined and repetitive.

Automating ad-hoc pipelines supporting new use cases can be significantly more challenging. Cataloging with metadata speeds up provisioning these types of pipelines. Along with data, it is essential to catalog analytical assets, including reports, models, metrics and other analytics used to derive insights from that data, for providing the full context. For example, it is possible to register in Collibra Data Catalog the reports created in Looker, along with the individual objects that make up those reports, such as dashboards, looks, tiles, and queries.

A well-cataloged cloud data lake helps data consumers to choose the "right" data for their analysis. Collibra scalable, predictive Data Quality complements this approach by ensuring the data is "right". The ML-driven adaptive data quality rules continuously monitor incoming data for quality. They proactively identify a wide range of issues, including inconsistencies and missing data. The predictive data quality enables high-quality scheduled, as well as ad-hoc data pipelines in real-time.



Statistics and quotes in callouts

"Google cloud and Collibra bring two best in class capabilities to help enterprises unlock the business value of data. We empower them to understand and use data effectively. Our partnership provides a strong governance foundation, enabling them to scale access and accelerate analytics while maintaining privacy."

- Evren Eryurek, Director, Product Management, Data Management & Streaming Analytics, Google Cloud

Conclusion

Cloud platforms offer agility and scalability to overcome data silos and enhance digital transformation efforts with powerful analytics. Successfully migrating data to the cloud with three key steps requires a transformational approach focusing on self-service access to governed, trusted data.

The three pillars of data governance foundation, predictive data quality, and policy-driven access control maximize the value of enterprise-scale data and analytics in the cloud.

To learn more about the integration of Collibra and Google Cloud, visit collibra.com/google