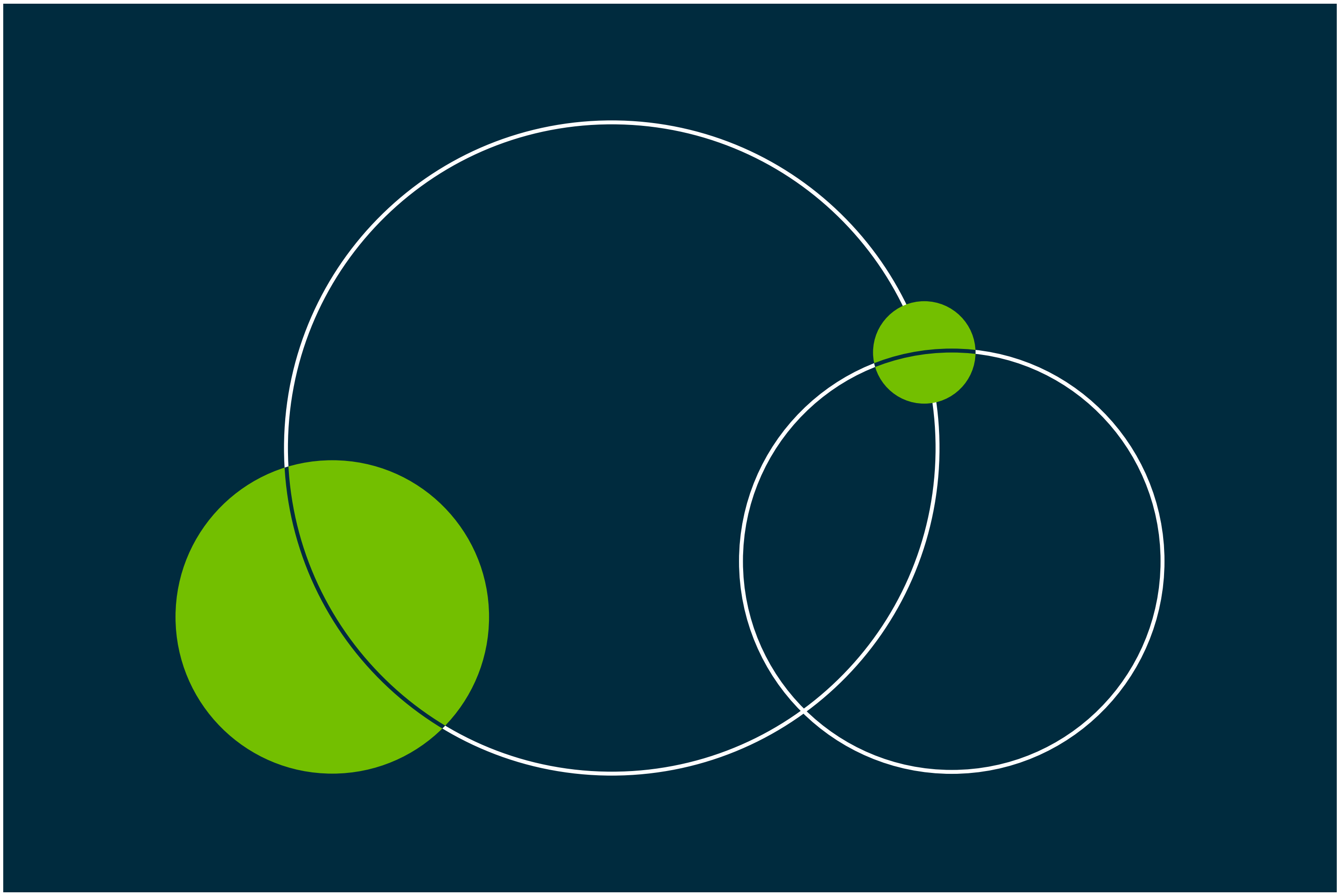
Collibra | Google Cloud

Whitepaper

Building a governed cloud data platform with GCP and Collibra



Executive summary

Cloud data platforms have become a key component of enterprise data architectures, playing a central role in many organizations' digital transformation strategies. That is not simply because they offer multipurpose, scalable and cost-effective storage. It is because they foster more agile data operations, cut through siloed architectures and unlock the potential of artificial intelligence and machine learning to drive new trusted business insights.

Those core benefits help address many of the challenges that companies face in executing their digital transformation programs. Across all industry verticals, companies are collecting much greater volumes of data, which points to the need for scalable, cost-effective solutions. They are also collecting a greater variety of data (including structured, semi-structured and unstructured data) that is difficult to describe via a single schema. This points to the need for multipurpose storage. Most importantly, they need to quickly derive insights from these diverse datasets, which points to the need for agile data operations and sophisticated analytics (particularly AI/ML capabilities).

However, some benefits can also result in unintended consequences. The ease with which organizations can store greater volumes and a broader variety of data, at a lower cost, has the potential to contribute to poor housekeeping. Without proper governance, organizations are likely to run into challenges relating to data quality, data discovery and compliance.

The key to unlocking the benefits of cloud data platforms, while mitigating any unintended consequences, lies in ensuring that your data is well governed.

This paper provides:

- An overview of how we define a cloud data platform
- An explanation for why data needs to be well governed on a cloud data platform
- An architectural overview of the GCP/Collibra governed cloud data platform
- Insights into four strategies to ensure a successful implementation

What is a cloud data platform?

As with many terms in the world of enterprise technology, a "cloud data platform" can be somewhat ambiguous, which is why we will start by defining exactly what we mean by it. From a functional perspective, when we refer to a "cloud data platform" we reference all the tools that an enterprise needs to collect, process, store, analyze and visualize data. While Google Cloud Platform's (GCP) capabilities are always evolving, as new services are launched, below are some insights into tools currently on offer in each of those categories:

Data collection

Cloud data platforms need to aggregate data from a number of different sources, including real-time updates and batch transfers. GCP offers the ability to on-board streaming data via Pub/Sub and IoT Core services, and provides a range of batch upload options via Data Transfer.

Data processing

Data on-boarded from source systems typically needs to be pre-processed before it is stored to support further analysis. GCP offers a range of tools to support these processes, including Dataflow for streaming data, Dataproc for Hadoop/Spark stacks, Data Fusion (for integrating data from multiple sources) and Dataprep (for data wrangling).

Data storage

Most enterprises need a combination of data lake and data warehouse technologies to support their business intelligence and data science teams. Data lakes like Google Cloud Storage are required to accommodate a full variety of data types - particularly unstructured data sources, but also structured data in its raw form (before it has been pre-processed). This versatility lends itself to several use cases. Data lakes serve as a repository for raw source data, a staging area for data as it is prepared for further analysis, a central hub for self-service business intelligence – or a combination of those functions. Data warehouses, on the other hand, serve as a central hub for structured data that has been processed into a common schema and is therefore ready for further analysis. GCP's data warehouse solution BigQuery is particularly relevant for scalable enterprise analytics. Its distributed architecture not only offers highly available and durable storage, but also helps to support query performance and scalability.

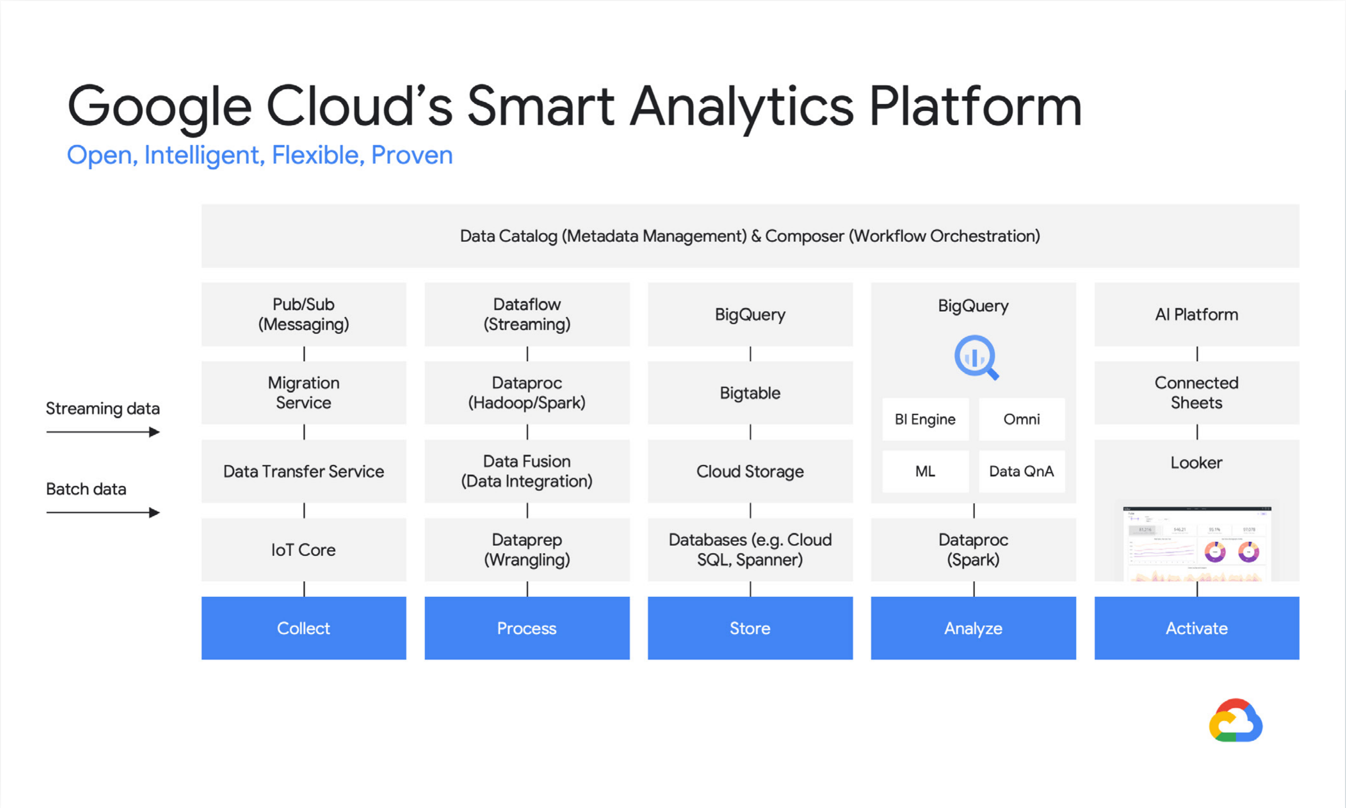
It is important to note that processing unstructured data typically yields outputs that are structured by nature. Take the example of a firm looking to analyze customer conversations collected via multiple channels (call centre audio files, email and instant messaging). The textual content of those conversations (once audio files are transcribed using speech-to-text) is unstructured by nature and would not necessarily be ingested directly into a data warehouse. However, natural language processing engines can be used to score those conversations and determine levels of customer satisfaction (or dissatisfaction). In doing so, raw unstructured data is turned into structured data that can be further analyzed in combination with other sources as part of broader investigations into customer churn.

Data analysis

Once data has been collected, processed and stored in its required structure, it is ready for further analysis. This can take a number of shapes – from a simple query to more complex calculated metrics and analytics, all the way through to machine learning models designed to detect new patterns or drive predictions. BigQuery offers a full range of analytical functions, including specialised capabilities for streaming and geospatial analysis, as well as machine learning.

Data visualization

Once data has been analyzed it needs to be presented in a visually intuitive manner. From simple line and bar charts through to more complex geospatial visualizations, the key to most business intelligence tools is to clearly display patterns and allow data to be more easily understood. GCP recently acquired Looker as its own in-house business intelligence solution, but also partners with a range of third parties, including Tableau and Qlik.



Google Cloud's Smart Analytics Platform

Open, Intelligent, Flexible, Proven

Data Catalog (Metadata Management) & Composer (Workflow Orchestration)

Streaming data →

Batch data →

Collect	Process	Store	Analyze	Activate
Pub/Sub (Messaging)	Dataflow (Streaming)	BigQuery	BigQuery	AI Platform
Migration Service	Dataproc (Hadoop/Spark)	Bigtable	BI Engine Omni	Connected Sheets
Data Transfer Service	Data Fusion (Data Integration)	Cloud Storage	ML Data QnA	Looker
IoT Core	Dataprep (Wrangling)	Databases (e.g. Cloud SQL, Spanner)	Dataproc (Spark)	

Drivers for adoption

Cloud data platforms are a key component to most organizations' digital transformation projects because of three overarching trends:

The ongoing data deluge

As organizations look to be more data-driven in their operations, they invariably collect more data at ever faster velocities. This ongoing data deluge is only gathering pace and supports the need for scalable, cost effective storage.

Variety of requirements

Traditional efforts to consolidate enterprise data focused on building data warehouses and finding an overarching schema that could describe a company's data assets. Yet legacy data warehouse technologies ran into issues relating to scalability and cost. At the same time, the inputs used to derive insights have also evolved. With data being collected from an ever-increasing variety of endpoints - from mobile apps through to IoT devices - requirements for scalable storage of text, audio and video content have also grown.

Analytical complexity

Digital transformation is impacting every industry - and data lies at the heart of that transformation. Companies are looking to use data to build better products, better understand their customers, optimize their operations and mitigate risks. Doing that means analyzing a broader variety and greater volumes of data. The need to cut across operational silos and incorporate data sourced from a greater number of end-points, has also meant data scientists need to be supported by agile data infrastructures, along with sophisticated AI/ML capabilities to derive insights from those diverse datasets.

Cloud data platforms have proven to be very effective at addressing these trends because:

- They benefit from economies of scale and shared services to offer compelling total cost of ownership
- They are highly scalable, not only in terms of storage and processing capacity, but also concurrent use across teams of business analysts and data scientists
- They foster agile operations, helping to speed time to provision resources, not only compute and storage but also automating data pipelines and more complex workflows
- They have been architected with advanced analytics in mind - including specialised geospatial and real-time analytics, as well as cutting edge machine learning and natural language processing models

Why does data need governance on a cloud platform?

Cloud data platforms have proved very effective at fostering agile data operations. That is not only because they negate the need to manage hardware, but also because of the variety of tools and depth of automation available to support core data management and analytical processes.

However, some of the benefits offered by cloud data platforms can also result in unintended consequences. Without proper governance, organizations are likely to run into challenges relating to:

Data quality

A lack of ownership and accountability can lead to poor control over source data, resulting in high levels of overlap or redundancy. A lack of contextual information also makes it hard to ascertain which source is the most complete, accurate or current.

Data discovery

Cloud data platforms hold great promise for business analysts and data scientists. Being able to access enterprise data assets from a single location may sound like a panacea. But without contextual information, they cannot know which sources to choose, how to interpret that data, or whether to trust in its accuracy.

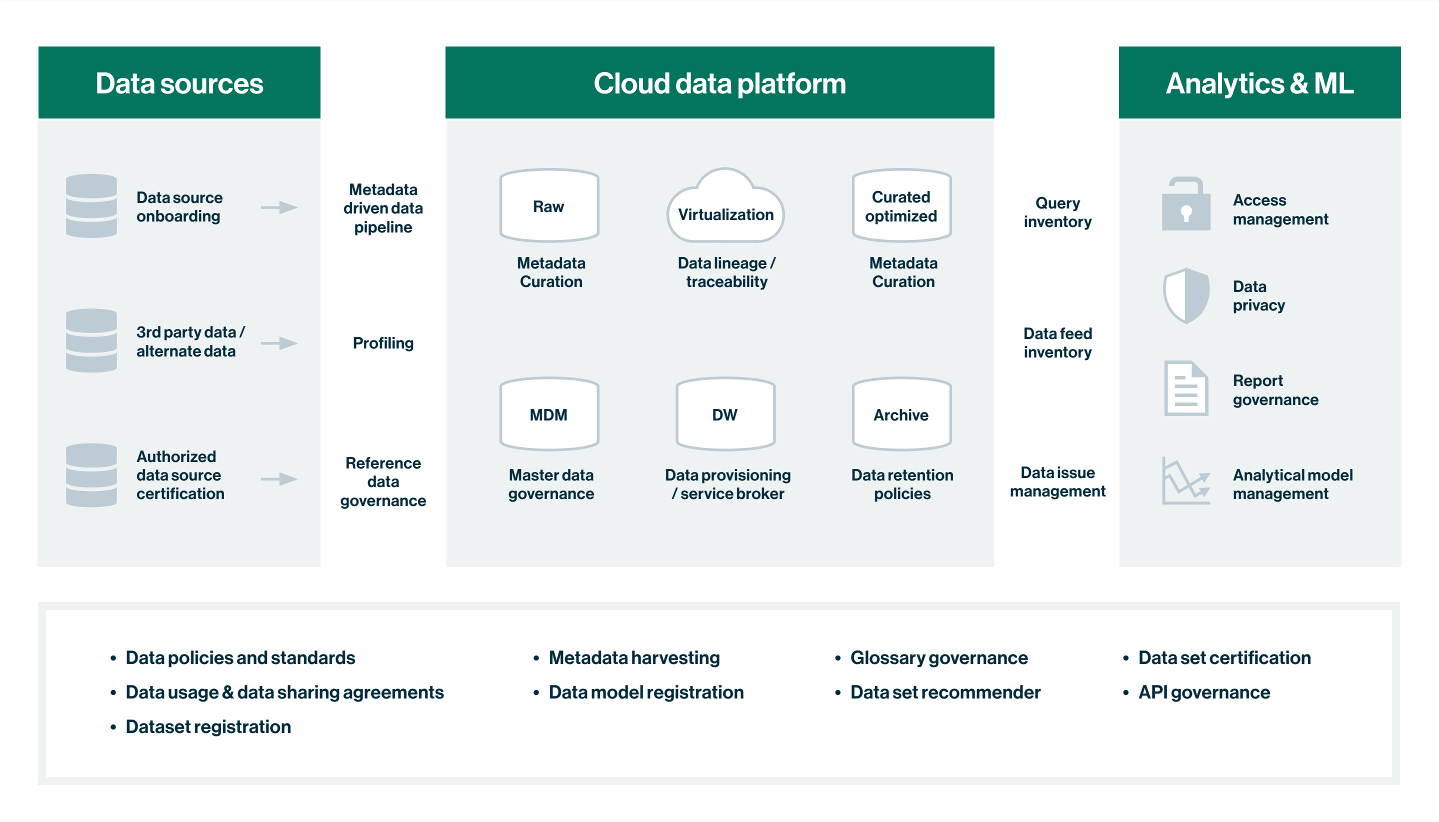
Compliance

Poor data governance can also pose risks. Most organizations face a myriad of rules that impact the way they manage data. Some regulations, like GDPR and CCPA, provide data subjects with greater rights over data (such as the right to have their personal information deleted) and place an onus on organizations to uphold those rights. Other rules set by industry regulators and tax authorities require organizations to retain data for audit purposes. Faced with an ever-more complex set of regulations, organizations need to take a data-centric approach to compliance – knowing where all sensitive information is stored, which policies apply to each dataset, what type of processing is permitted and how access should be controlled.

Addressing these challenges is precisely what has brought the concept of a ‘governed’ cloud data platform to the fore and has been the driving force behind the partnership between Collibra and Google Cloud Platform (GCP).

GCP and Collibra governed cloud data platform architecture

Governing a cloud data platform requires a range of functional components. Below we provide a high level architecture of GCP/Collibra's governed cloud data platform, including an explanation of key components supporting the governance framework.



Data source onboarding

This covers registration of new data sources and the GCP onboarding process. As data is onboarded, insights are captured into what the data looks like, its context, data quality metrics etc.

Third party data / alternative data

It is important to capture metadata not only to describe proprietary datasets, but also third party sources. This enables consumers to evaluate the cost and quality of a particular dataset, and gauge whether that cost is worthwhile relative to alternatives. Analyzing third party data can help identify opportunities to consolidate providers and eliminate duplicative spend. It is also important to be able to detect feed changes so that end users can be notified.

Authorized data source (ADS)

Authorized data sources should be highlighted to help consumers find trusted data more easily. These sources will have been certified and are overseen by a data custodian that accepts accountability for data quality and timeliness.

Metadata curation

This involves gathering all relevant information to describe each dataset. Metadata can include data classes (such as personally identifiable information, sensitive information etc), categories (such as product data, reference data etc), tags and fields (ideally mapped to a standardized set of business terms and definitions). Some metadata can be captured automatically by profiling data sets using Natural Language Processing (NLP) techniques during the on-boarding process. Other information is also gathered through feedback from users, analysts and subject matter experts. As part of that process it is important to support a collaborative framework so that dispersed teams can work together to promote data quality. The end goal is to ensure datasets can be easily tracked, identified by business terms, governed and managed.

Reference data

Reference data defines permissible values to be used by other data fields. For example, entering an address into an online form typically requires choosing from a list of countries rather than entering that information free form. This list is an example of reference data. Just as different systems and applications have different ways of naming and organizing fields, they also use different codes and values to define reference data. Mapping those into a common set of values makes it possible to translate and interpret data consistently across a diverse ecosystem.

Data lineage

Data lineage describes the journey taken by data from source through to destination, including every hop and transformation along the way. It is important that lineage can be mapped not only on a technical level (defining the physical path taken by data), but also at a logical level. It is possible for one type of storage structure and file format to be optimized for a particular workload, but not suitable for another. Situations like these, particularly given the low cost of storage, means that multiple copies of the same data are often created with different underlying storage structures (partitions, folders) and file formats (e.g. ORC vs Parquet). By mapping data lineage both logically and physically, these nuances can be easily understood and captured.

Report watermarking

This process enables organizations to demonstrate accountability throughout the report lifecycle by interacting with all involved parties (internal/external) and supporting technologies that store and process report information.

Analytical model management (AI/ML)

Just as metadata can be curated to gain a better understanding of different datasets, it can also be captured to understand analytical models. This helps to document the purpose, inputs and outputs of the model (including data sources and consumers).

API governance

APIs are a crucial way to access data from many systems. Collibra includes the ability to search API endpoints to find and access the right data.

Data usage and sharing agreements

It is important to be able to track data usage, not only through access by end users but also via batch jobs, queries and APIs. This helps to ensure sharing agreements are adhered to.

Data provisioning

Collibra Data Catalog serves as an entry point for data scientists and business analysts to find raw and enriched datasets that suit their analyses and to be able to access that content (if properly permissioned) in the format they need.

Dataset recommendations

By understanding which datasets have proved useful in the past, Collibra can also provide automated recommendations to help save time in locating datasets.

Dataset certification

Certifying data helps to assure data consumers that they can trust in its accuracy. Users should not only be able to see if a dataset has been certified, but also who certified it, whether they have provided any descriptive notes, which data elements are contained, where they are sourced from, as well as relevant data quality metrics.

Data retention policies

Defining and executing data retention policies helps ensure that data required for audit purposes is stored for as long as needed. Other data can be kept for as long it is actively used to reduce the unnecessary storage costs.

Data model registration

To support automation of data pipelines, it is important to register target data models / data structures. This ensures that source data can be mapped, including transformations is necessary, into the desired format of the target data structure (JSON, XML, Parquet, etc).

Query inventory

Keeping an inventory of queries commonly used by data scientists and analysts enables them to be re-used more easily, saving time and helping to promote data intelligence across the organization.

Four strategies for a successful implementation

Implementing a governed cloud data platform is a complex endeavour. There are many functional components to consider, as evidenced by the previous architectural overview, and not every implementation will require the same mix of capabilities. However, all projects need to remain focused on attaining certain strategic objectives. Here, we highlight four key strategies to ensure governance efforts remain focused on enhancing data quality, data discovery and ensuring compliance.

Strategy 1: Controlled ingestion

GCP's cloud data platform is highly scalable and capable of ingesting massive volumes of data from almost any source, including Internet of Things (IoT) sensors, clickstream activity on websites, online transaction processing (OLTP) data, and on-premises data, to name a few.

This level of scalability makes it all the more important to catalog data before it is ingested. Ultimately, the goal of cataloging data is to ensure data consumers know exactly:

What data is available

This is the easiest step as it can be automated by profiling datasets to capture technical metadata.

What the data means

This is where governance plays its biggest role. It is important that data can be interpreted in the right context, which can only be done by ensuring business definitions are clear and applied consistently.

Where data originated

Lineage and provenance are crucial to add understanding. Anyone who wants to use data will have questions about where data came from and how it was transformed. Technical lineage can be harvested as data is ingested into GCS, allowing end users to identify the source system for any dataset. By registering the data and adding human oversight, more detailed questions can be answered, such as:

- Who produces the data and where?
- Which legal entity owns the data?
- What country did the data come from?

Whether data is complete, accurate and consistent

In order to assure data consumers that data can be trusted it is important to provide them with metrics relating to data quality. To make informed decisions, consumers need to know that data is consistent, complete and accurate. It is important to note that gauging these attributes will typically require some expertise from data stewards and can not be entirely automated.

How to access data

Detailed metadata not only helps data consumers find what they are looking for, but can also help speed the provisioning of data. It is also important to capture any legal, security or strategic considerations that need to be taken into account when authorizing access to data, as this will help support the fourth strategy - access governance.

Populating a data catalog with enough information to answer the above questions is not a simple process. While a significant amount of technical metadata can be automatically captured as part of the ingestion process, this will typically need to be supplemented with insights from analysts and subject matter experts. It is also important to note that metadata continues to be added as data consumers feed their insights back into the data catalog.

The following four tips are useful in ensuring the successful implementation of a governed data catalog:

1. Start with the end in mind and consider what a user sees when they browse the catalog, in terms of topics, categories, data areas etc.
2. Make at least one dataset available per category to ensure that users can be served and do not have to look elsewhere.
3. It is important to track who uses and requests which datasets. These usage insights can help define governance priorities, with the most popular datasets being afforded tighter governance and quality checks.
4. Usage statistics can be equally important to aid data discovery and make recommendations, helping to highlight the most valuable datasets.

Strategy 2: Building pipelines

In order to drive business insights, organizations typically need to combine raw data from multiple source systems and make it available for enterprise analysis. This requires a lot of pre-processing, particularly when new sources need to be on-boarded.

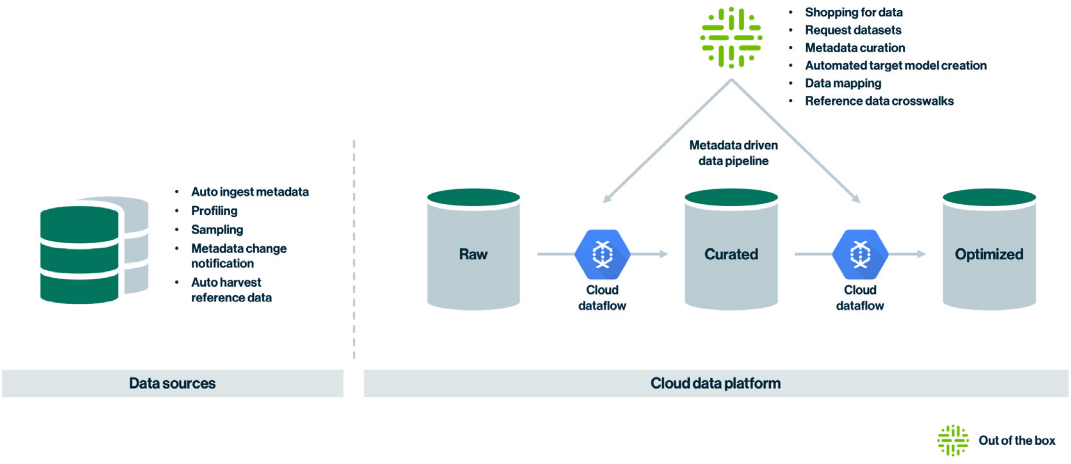
Take the following scenario: a data scientist works for a large multinational organization that recently acquired a new regional subsidiary. He would like to on-board data from that subsidiary into his global analysis.

The traditional approach

In the traditional approach to enterprise data analytics, such a request may end up taking weeks to be fulfilled. It would typically begin with the data scientist contacting the enterprise data warehouse team. In turn, they would need to contact the owners of the source data and make changes to the data warehouse schema to accommodate the new source. They would also need to transform the source data to ensure it conforms to the data warehouse schema; and finally, make the data available to the analyst in his requested structure. Each of these steps would require a significant amount of manual intervention and configuration.

Metadata-driven data pipelines (DataOps)

By contrast, metadata-driven data pipelines can automate the bulk of this process by providing ETL (extract, transform and load) instructions to carry data from its 'raw' source state into the 'optimized' target state that the analyst has defined. This approach is often characterized as 'DataOps' given its similarities with DevOps software development methodologies. DevOps enables smaller code releases on a more frequent basis, while also placing the focus on development teams to automate various aspects of the software development lifecycle, such as integration and testing. Similarly, DataOps enables data to be provisioned on a more agile basis, taking the focus away from centralized teams to incorporate source data into a common schema, and enabling data scientists to define and automate their own pipelines.



In the context of Collibra and GCP, the data scientist uses Collibra Data Catalog to shop for the right sources and request the specific data they are looking for. Knowledge of source metadata, as well as target data models, allows Collibra to pass relevant instructions for GCP Dataflow to automate the ETL processes that transform raw data into its desired target state. This workflow includes the following processes:

Automated target model creation

Automatically generate a target data model / table structure that has been optimized for the relevant datasets being requested.

Data mapping

Mapping data from their source structure to the optimized target model that has been defined.

Reference data crosswalks

Ability to recognize different terms with the same business definition across multiple sources.

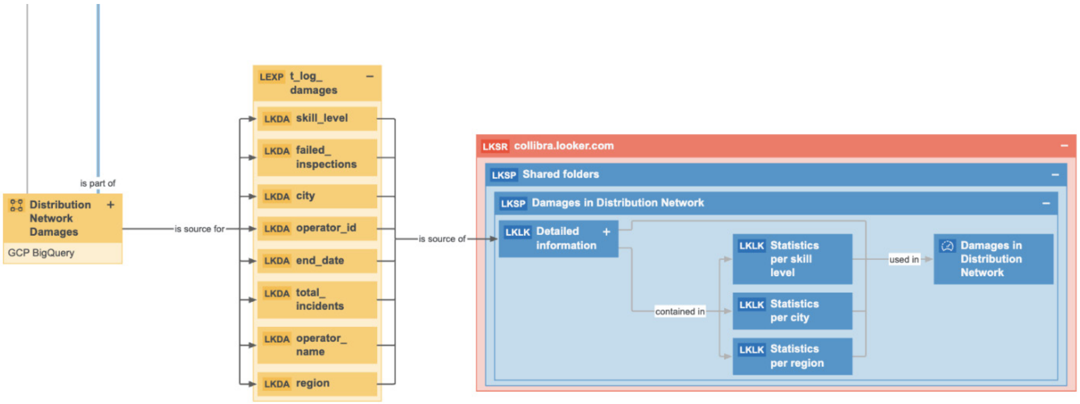
Strategy 3: Certifying information assets

Data governance is a discipline that can be applied to more than just source data. It can be just as important to govern the components that facilitate analysis — everything from specific queries, API calls, analytics and machine learning models, through to reports, worksheets, notebooks, dashboards and cubes. Registering these information assets in a catalog means business analysts can not only share their insights, but also the tools that generated those insights. This helps the entire organization to be more data intelligent — speeding the time taken to turn raw data into meaningful conclusions.

To illustrate that point, take the following example: in any given enterprise, it is not uncommon to have several thousands of reports, dashboards and metrics used to manage finances. Different data marts are often created for each project and reports generated using several different tools. By registering all of those information assets in a catalog, organizations can get better at sharing their insights, as well as the means by which those insights were generated, which helps save time and drive consistency in analytical processes. For example, if a user discovers that a dataset is already used in Looker, rather than starting in

BigQuery and (re)building a new BI object, the user can simply request access to an already certified asset and start making decisions much faster.

Cataloging reports also makes it simpler to track data lineage, which in turn mitigates risks from change management. For example, if a column in BigQuery is deprecated, identifying affected reports (without an automated solution) can potentially be very labor intensive and error prone. By ensuring all reports are registered in a catalog, technical lineage can be captured automatically to ensure that process is seamless.



Strategy 4: Data access governance

Stricter privacy regulations require organizations to rethink how they manage access to customer data. Organizations have to strike a difficult balance between data democratization, which is required to stay competitive, and privacy & risk management, which is necessary to maintain customers' trust. Finding a solution to that challenge means maintaining detailed knowledge of where personal information is stored and administering granular controls over that data to ensure the right policies are applied.

Unfortunately, many organizations manage access controls using an application-centric framework that is owned by IT and is largely manual. This has the potential to stifle innovation by slowing down the provisioning process, while also missing the mark by being too coarse.

By combining GCP's fine-grained access controls with Collibra's data intelligence and policy management, organizations can adopt a data-centric approach to access governance. This means configuring access policies at the level of individual data categories and data elements, and then holistically enforcing those policies across all data stored on GCP (independent of how that data is stored). Such access policies can take into account the full context of the access request, which can include the purpose of the query, the business unit the data consumer belongs to, and the application they are querying the data from.

With the need for finer-grained access controls, the new privacy regulations have also highlighted the need for record level access management. The integration between Collibra and GCP allows organizations to configure holistic record level policies to enforce consent and privacy preferences, retention policies and data residency policies, as well as operationalize individual requests such as the right to access personal information. This leverages the upcoming Collibra Data Matching product, which lets organizations extract metadata at the record level, such as a customer's consent, date of onboarding, or country of residence, as inputs to authorize access to each record.

A holistic data-centric approach to access management greatly reduces the cost of compliance with privacy regulations across the globe, and reduces the risk that comes with access management at the perimeter of the application.

Conclusion

Organizations looking to execute digital transformation strategies will invariably be drawn to cloud data platforms. Governance of data and analytics will play a key role in unlocking the full benefits of those platforms, while mitigating potential risks.

If you are interested in learning more, please visit our website and request a demo at collibra.com/request-a-demo

For additional questions, contact us at:

United States	United Kingdom	All other locations	By email
+1 646 893 3042	+44 203 695 6965	+32 2 894 79 60	info@collibra.com