

A Guide to Data Entitlements & Access At Scale

A Framework for Secure & Auditable Data Platforms

Author: Phil Goerdt | Director, Enterprise Architecture | Onix

AI Addendum: Steve Berrey | Head of AI/ML | Onix

Table of Contents

Abstract 03

Pattern Recognition and Definition 03

- Physical Data Storage and Security
- Data Classification and Cataloging
- User and Application Personas and Access
- Regulated Data Usage and Compliance
- Auditability
- Testing

Reference Google Cloud Implementation 07

Machine Learning and Artificial Intelligence Addendum 08

- Special considerations for training data and for data scientist
- Enterprise Readiness for Generative AI
  - Data Governance and Privacy
  - Security and Compliance
  - Reliability and Sustainability
  - Safety and Responsibility

What's Next 10

Abstract

The achilles heel of many data projects is data governance. Whether it is too much, too little, too explicit or too vague, data governance and controls often complicate and delay initiatives. While this common pain point exists, and some tooling has been created to assist with point issues, no common framework for entitlements has been created. This document aims to serve as the first framework, along with a reference architecture for Google Cloud Platform.

Pattern Recognition and Definition

One of the challenges in this area is that many of the frameworks are specific to a particular data domain or use case. This makes it particularly difficult to come up with a common pattern that can be used from one end of the spectrum to the other. When thinking about permissions for data, we need to be concerned with several areas:

- Physical data storage and security
- Data classification and cataloging
- User and application personas and access
- Regulated data usage and compliance
- Auditability
- Testing

These different vantage points can span beyond a particular domain or use case, whether that be ingestion, computation and transformation, or analytics and machine learning.

Physical Data Storage and Security

Physical Data Storage and Security will follow standard cloud best practices. This will include things such as:

- Key rotation: This is the process of regularly changing encryption keys to prevent unauthorized access to data.
- Encryption at rest and in transit: This is the process of encrypting data both when it is stored on physical media and when it is being transmitted over a network.
- Multifactor authentication: This is the process of requiring multiple forms of identification, such as a password and a one-time code, to access data.
- Strong password management: This is the process of creating and using strong passwords that are difficult to guess.
- End user and service account hygiene: This is the process of keeping end user and service accounts secure by regularly changing passwords and disabling unused accounts.

These best practices will help to protect data from unauthorized access, use, disclosure, disruption, modification, or destruction.

Data Classification and Cataloging

Data classification is the process of assigning labels to data based on its sensitivity and value. This helps organizations to protect sensitive data and ensure that it is only accessed by authorized users. Data classification also helps organizations to comply with regulations, such as the General Data Protection Regulation (GDPR).

There are many different ways to classify data. Some common methods include:

- By sensitivity: Data can be classified as public, private, or confidential.
- By value: Data can be classified as high value, medium value, or low value.
- By type: Data can be classified as financial, personal, or operational.
- By regulatory requirements: Data can be classified based on the specific regulations that apply to it.

Once data has been classified, it is important to catalog it. This means creating a database of all the data in the organization, along with its classification. This catalog can then be used to track data usage and ensure that data is only accessed by authorized users.

Data classification and cataloging are essential for data security and compliance. By taking these steps, organizations can protect their data and avoid costly fines and penalties.

In addition to the above, data classification can also help organizations to:

- Improve data quality: By understanding the different types of data that are being collected, organizations can identify and correct any errors or inconsistencies.
- Increase data efficiency: By organizing data into categories, organizations can make it easier to find and use the data they need.
- Reduce data risk: By identifying and protecting sensitive data, organizations can reduce the risk of data breaches and other security incidents.
- Improve data governance: By establishing clear policies and procedures for data management, organizations can ensure that data is used in a responsible and compliant manner.

User and Application Personas and Access

User and application personas are essential for guiding data usage and consumption. They are fictional representations of ideal customers that help businesses understand their needs and goals. By understanding users, businesses can ensure that they are collecting the right data and using it in a way that is beneficial to them. Businesses can also use personas to identify potential problems with their data usage and consumption, and to develop solutions to address those problems.

Here are some of the benefits to businesses using user and application personas for data usage and consumption:

- Better understand their users' needs and goals
- Identify the right data to collect
- Use data in a way that is beneficial to their users
- Identify potential problems with their data usage and consumption
- Develop solutions to address those problems

Here are some additional tips for using personas effectively:

- Ensure that personas are based on real data and research.
- Keep personas up-to-date as users' needs and goals change.
- Use personas to inform all aspects of data collection and usage.
- Test personas with real users to make sure they are accurate and helpful.

Access to data should be tailored to the specific needs of each user and application. The personas, the use case, and the catalog together determine which data is available, how it is accessed, for how long, and for what purpose.

Some examples of user personas include:

- A data scientist who needs to write to a particular data store to run analyses.
- A business analyst who needs to query certain data sets to generate reports.
- A customer service representative who needs to view customer data to resolve issues.

The specific access permissions that are granted to each user persona will depend on the specific needs of that persona. These personas and their permissions should also be cataloged and routinely analyzed for compliance, overpermissioning and usage.

Regulated Data Usage and Compliance

Regulation can have a large impact on how data is used. For example, the General Data Protection Regulation (GDPR) in the European Union requires companies to obtain consent from users before collecting or using their personal data. The California Consumer Privacy Act (CCPA) in California has similar requirements. Companies that fail to comply with these regulations can face fines or other penalties.

Understanding the regulations, which data is impacted, and usage of data within the bounds of the law is essential for data security, access, and management. This is because data is a valuable asset that needs to be protected from unauthorized access, use, disclosure, disruption, modification, or destruction. By understanding the regulations, organizations can ensure that they are in compliance with the law and that their data is being protected appropriately. Additionally, by understanding which data is impacted, organizations can prioritize their data security efforts and focus on protecting the most sensitive data. Finally, by understanding the usage of data within the bounds of the law, organizations can avoid potential legal issues and ensure that they are using data in a responsible and ethical manner.

Here are some of the regulations that organizations need to be aware of:

- The General Data Protection Regulation (GDPR)
- The California Consumer Privacy Act (CCPA)
- The Brazil General Data Protection Law (LGPD)
- The Personal Information Protection and Electronic Documents Act (PIPEDA)
- The Cybersecurity and Infrastructure Security Agency (CISA) Cybersecurity Framework
- The National Institute of Standards and Technology (NIST) Cybersecurity Framework

Regulations that have an impact on an organization should be captured in both the data catalog, and within the broader decision making process.

Auditability

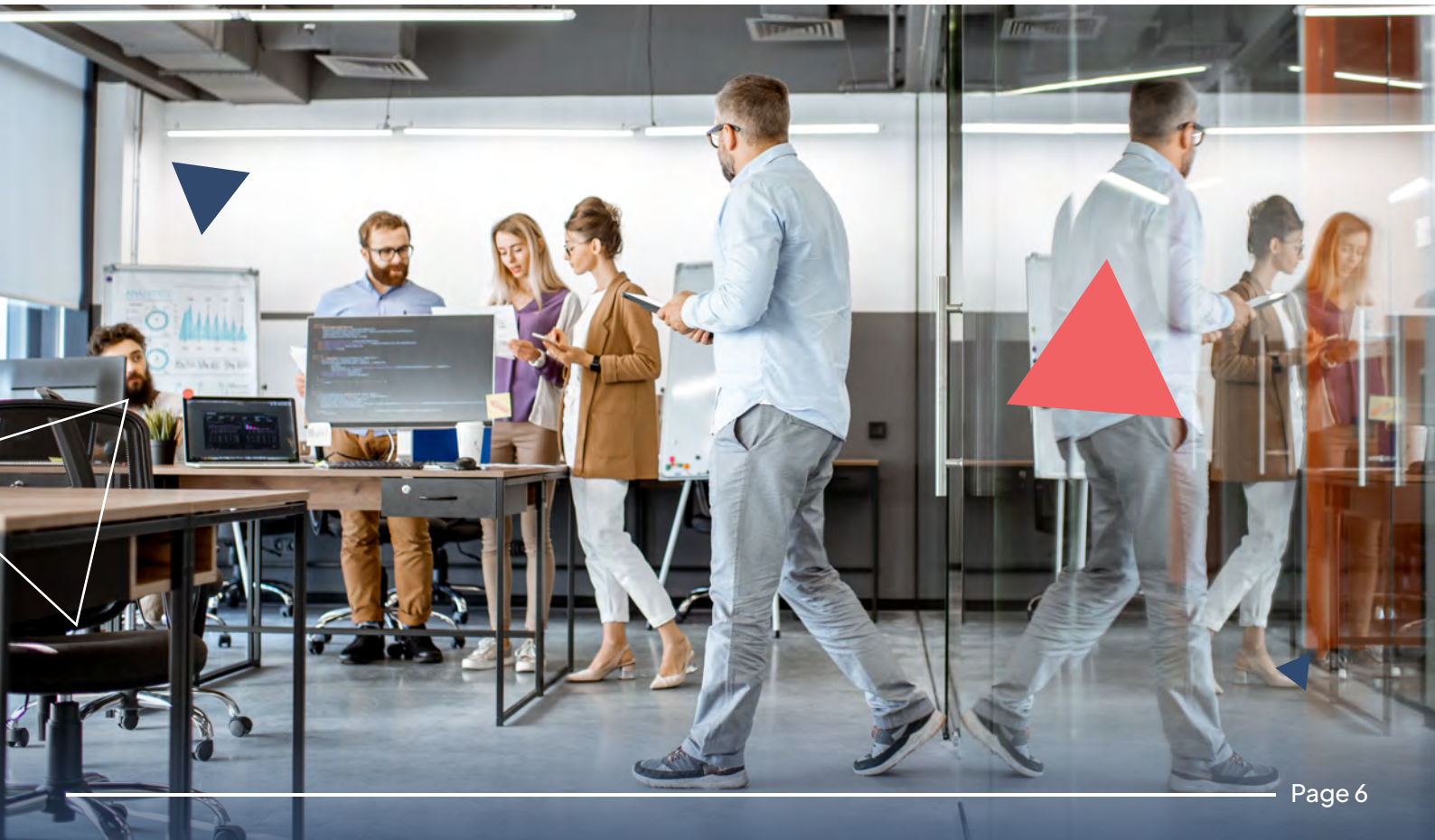
There are several critical needs for auditability for data:

- Data usage (by applications and services): This is important for ensuring that data is being used in a way that is consistent with its intended purpose and that it is not being misused.
- Data movement patterns: This is important for tracking how data is being moved around and used within an organization, and for identifying any potential security risks.
- Data access (by individuals): This is important for ensuring that only authorized individuals have access to data, and that unauthorized access is prevented.
- Data storage: This is important for ensuring that data is stored securely and that it is not lost or corrupted.

All of these aspects of data auditability are critical for maintaining security and compliance with any applicable regulations.

Testing

It is also important to ensure that the best practices outlined above are being correctly implemented. Testing the implementation with automated and ad hoc, person driven testing is a good way to take proactive measures for data security and correct data entitlements. Automated testing can help to identify potential problems with the implementation, while person driven testing can provide a more comprehensive look at how the system works in practice. By testing the implementation in this way, organizations can help to ensure that their data is secure and that users have the correct access to it.

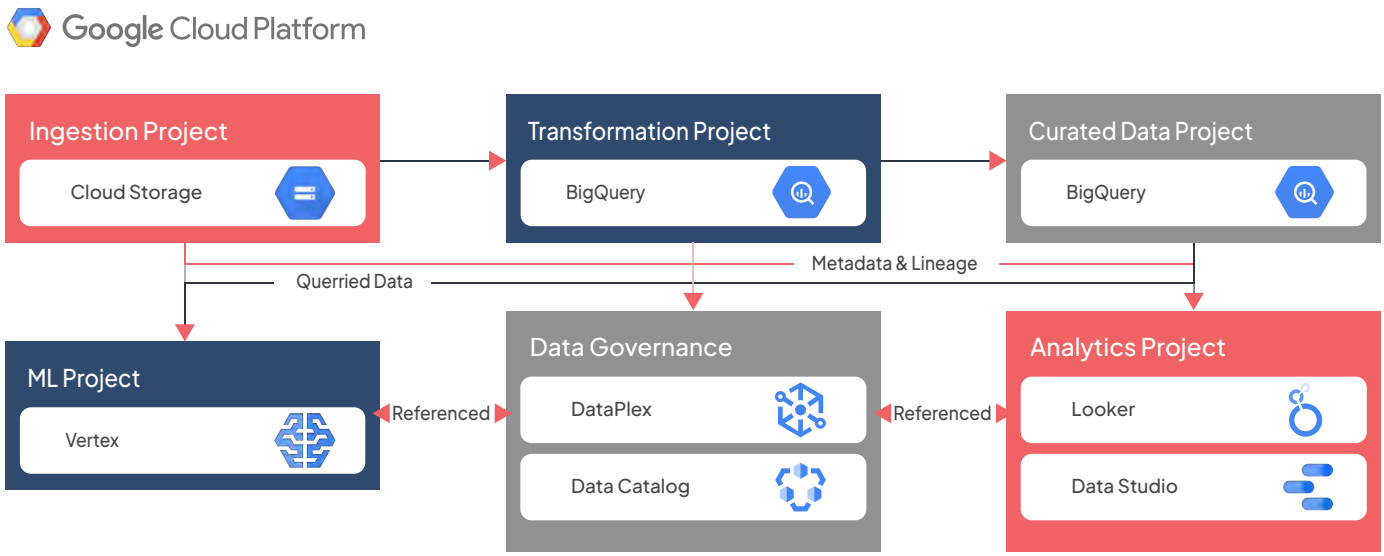


Reference Google Cloud Implementation

This is a reference implementation of Data Entitlements and security at scale for Google Cloud. It includes the following steps:

1. Securing data as it lands in the data lake in Google Cloud Storage. This can be done by using encryption and access control policies.
2. Moving and transforming the data from GCS to BigQuery. This can be done using Dataproc or Dataflow.
3. Cataloging data in data catalog in both the source (GCS) and target (BigQuery) using Data Catalog. This will help to organize and manage the data.
4. Managing permissions in IAM and in BigQuery using groups that are defined by user personas and the use case. This will help to ensure that only authorized users have access to the data.
5. Making data available for consumption via Dataplex. This will provide a centralized platform for users to access and analyze the data.
6. This data governance project (that contains Data Catalog and Dataplex) would be a shared resource, and located in the shared resources folder of the org structure (according to Google & Onix best practices). This will ensure that all users have access to the data governance project.
7. Testing and auditing would also be housed in the data governance project, with strict read permissions on audit data sets. This will help to ensure that the data is accurate and reliable.

This reference implementation can be used as a starting point for building a data governance solution for Google Cloud. This could be extended into MLOps workflows as well, to enable ML personas to access specific data for training or inference purposes.



Machine Learning and Artificial Intelligence Addendum

Special considerations for training data and for data scientist

For machine learning, the inclination is to allow data scientists access to as much data as possible, but you also must control access to the models (for ML and especially for LLMS, see below). It is tempting to allow data scientists access to as much data as possible for machine learning, but you must also control access to the models. A clever user can back into the data from a trained ML model, which is especially true for large language models (LLMs).

LLMs are trained on massive amounts of data, and they can be used to generate text that is indistinguishable from human-written text. The implication here is that if a user has access to an LLM, the user can potentially use the LLM to generate text that contains sensitive information, such as personal data or trade secrets. To protect against this, it is important to control access to LLMs and to monitor the output of these models. The same access model for data should be applied to model access - if the user does not need access to data, they should also not have access to the model trained on that data.

Here are some specific items to implement when controlling access today for data scientist and to control access to machine learning models:

- Be very selective on which data is available to machine learning and LLM models. Do not expose raw data to the data scientist for training, even internally. Instead, use the data warehouse approach, and curate cleaned and sanitized data replicas for training.
- The sanitized data should undergo de-identification of personal data, and re-identification using tooling such as that available in Google's DLP tools.
- Setup and institutionalize model monitoring, using an MLOps methodology. This will ensure the models and their data do not drift away from the original sources and intent of their purpose, and help to identify changes in the performance or changes in the underlying data, which could degrade their impact, or run a skew of current and future regulatory concerns. They also may violate your own internal governance of AI and Machine Learning.
- The purpose for which the model is being used. If the model is being used for a sensitive purpose, such as making decisions about people's lives, it is important to take extra precautions to protect the data, consider bias in the model, and have a responsible AI framework to analyze the efforts before and after the project.
- The potential risks of unauthorized access to the model. If unauthorized users gain access to the model, what could they do with it?

By taking these factors into account, you can help to ensure that machine learning models are used safely and responsibly.

Enterprise Readiness for Generative AI

Publically available Generative AI LLMs all come with varying degrees of privacy and security, some having neither. Using these consumer models for enterprise applications has inherent risk to your data, for exfiltration and your data and prompts being assimilated into the training of future releases for the models. With this in mind, Enterprises should consider securely contained environments for their usage of Generative AI, such as that provided with Google's Vertex AI toolset, as highlighted below.

Data Governance and Privacy

With Vertex AI, you control your data, not Google. The data is not used, by default, to train their models, and is only processed according to your instructions. There are built in compliances for regulation such as GDPR and other privacy best practices. Further reading is available in their published AI/ML Privacy Commitment.

Further, prompts are sent to foundational models encrypted, and not further used to train models. When fine tuning models, the specific model weights that are adjusted to enable the specific tasks are saved only in the customers environment, and they are not saved or stored by Google for modifying the foundational model's settings or training. The customer has sole access to the tuned models and data used for fine tuning.

Security and Compliance

Google Cloud's infrastructure is designed to be secure, with defense-in-depth at scale. Google is responsible for securing the underlying infrastructure, while customers have access to GCP native and partner tools and solutions to manage their own security. Google uses a zero-trust model and operational controls to secure its infrastructure, and customers can choose the level of security responsibility that is right for them.

Additional tooling, some covered above in this paper, is highlighted below:

- Google has a comprehensive security program that includes monitoring, vulnerability detection and remediation, malware prevention, and incident management. The program is designed to protect Google's users and data from a variety of threats.
- Specifically, Google uses a global team of security engineers to monitor for suspicious activity, collect information from a variety of sources, and respond to incidents. Google also uses a variety of tools and techniques to detect and remediate vulnerabilities, prevent malware, and manage incidents.
- Google's security program is constantly being updated and improved to stay ahead of the latest threats. Google is committed to protecting its users and data, and the security program is a key part of that commitment.
- Specific recommended controls
    - VPC Service Controls
    - Customer Managed Encryption Keys
    - Access Transparency

Reliability and Sustainability

Google Cloud's AI infrastructure is optimized for scale, speed, price-performance, and sustainability. It offers a variety of hardware options to support different use cases, and its TPU infrastructure runs on 90% clear energy.

Safety and Responsibility

Responsible AI starts with policies and processes. Defining reviews before undertaking a new project, assessing risk levels for the project, and developing mitigation plans for the risks are example early steps in a company's responsible AI considerations. The final step is to build responsible AI tools and practices across the organization, and actively review all uses of AI.

What's Next

There's a difference between knowing the path, and walking the path. If you're interested in walking the data entitlements at scale with a guide, reach out to Onix.

Get in touch

onixnet.com

connect@onixnet.com