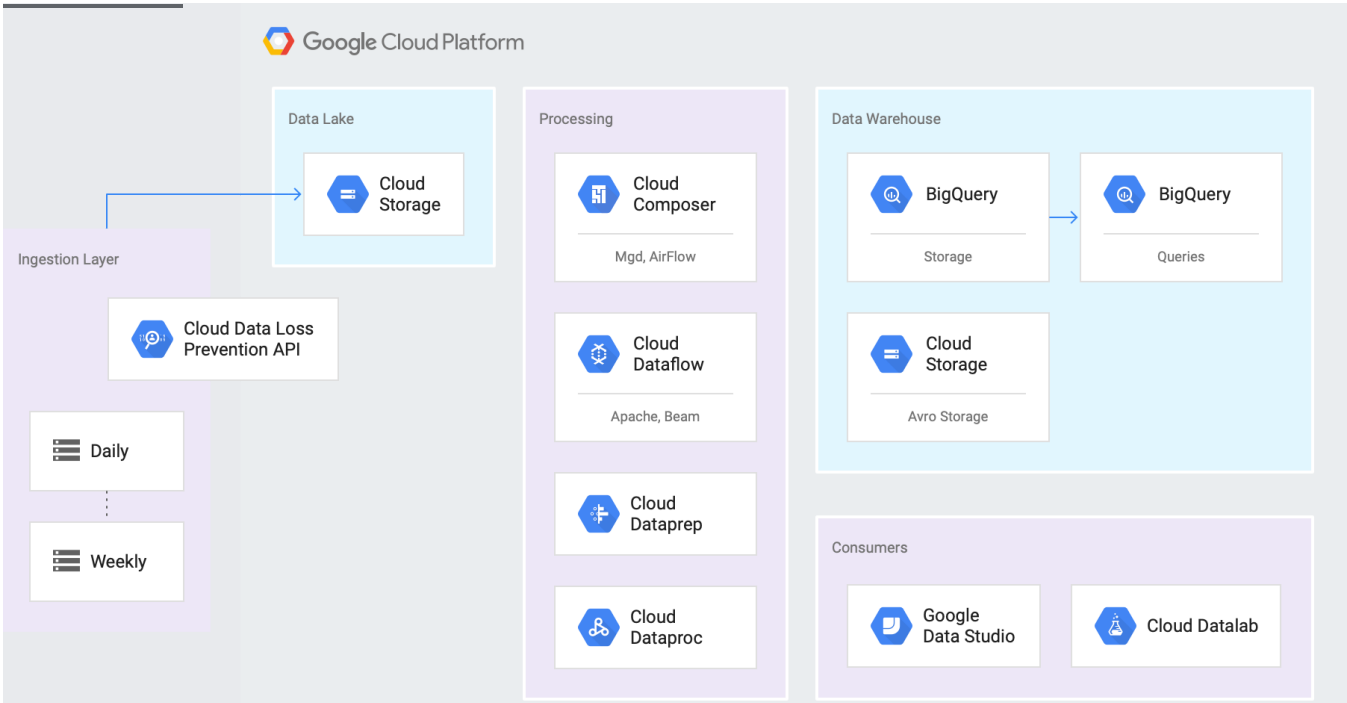
This article discusses the security controls designed to help manage data access to and prevent data exfiltration of the pipeline from your data lake to your data warehouse.

The article uses an example pipeline to show the following:

- Configuring Cloud IAM permissions to grant access to a set of personas who need to access data stored in a data lake-to-data warehouse pipeline.
- Configuring network controls to manage access paths to your data and help prevent data exfiltration.
- Implementing policies with the Organization Policy Service and Cloud IAM to help enforce your controls.
- Using Cloud KMS as part of your encryption strategy.
- Using the Cloud Data Loss Prevention API as part of the pipeline to classify and redact (or tokenize) sensitive data.
- Using auditing tools to see who has accessed your data.

The following diagram shows the example pipeline architecture.



You can use this architecture as the basis for various data lake use cases. To help identify an architecture that best suits your use case, see Build a data lake (/solutions/build-a-data-lake-on-gcp).

The example in this article resembles the Build a data lake (/solutions/build-a-data-lake-on-gcp) architecture, with a few differences. This batch analytics architecture also does the following:

- Uses Cloud Data Loss Prevention (/dlp/) (Cloud DLP). All data is first scanned and processed using Cloud DLP to identify and tokenize data that is classified as sensitive before uploading it into the data lake. Data goes through a workflow sorting and mining to cleanse, refine, and making data available for consumption.
- Uses both BigQuery and Cloud Storage as the final destination for the processed data (that is, the data warehouse).
- Supports using Cloud Data Studio and Datalab to query data stored in BigQuery and Cloud Storage.

Data lake (https://wikipedia.org/wiki/Data_lake). A repository that stores data in its native format. This example architecture uses Cloud Storage (/storage/), which is explained in the Cloud Storage as the data lake (/solutions/build-a-data-lake-on-gcp#cloud-storage-as-data-lake) section of Build a data lake (/solutions/build-a-data-lake-on-gcp).

Data warehouse (https://wikipedia.org/wiki/Data_warehouse). A central repository of integrated data from one or more disparate sources. A data warehouse stores current and historical data in one place, where it can be used for analytics. In this example architecture, the data warehouse includes data stored both in Cloud Storage and in BigQuery. These services replace the typical setup for a traditional data warehouse. That is, they serve as a collective home for all the analytical data in an organization.

The following table lists people and services associated with a data lake-to-data warehouse pipeline.

Persona	Activities
Data uploader	A service account or a person who writes data to the Cloud Storage data lake. A service account running automated uploads. People might also perform ad hoc uploads.
Data viewer	Person who consumes data from BigQuery reporting tables through Cloud Data Studio and other reporting tools, such as SAP Business Objects.
Data analyst (no SQL knowledge)	Person who prepares data in Dataprep (for example, joins BigQuery denormalized tables), and develops Google Data Studio reports and reports that use other tools.
Data analyst (SQL knowledge)	Person who performs ad hoc analysis in BigQuery (using denormalized tables), prepares reporting tables in BigQuery, and develops Data Studio reports and reports that use other tools.
Data scientist	Person who performs data science tasks, including statistical data analysis and machine learning model development, using various solutions. Solutions might include AI Platform, Datalab, R Studio, and SAS. Data scientists might perform ad hoc activities and develop models.
Data engineer	Person who develops pipelines for moving data to Cloud Storage and BigQuery. Creates tables in BigQuery, implements ML models and pipelines developed by data scientists. Uses solutions such as Dataflow, Dataproc, Cloud Composer, and Dataprep by Trifacta in addition to other data science solutions.

Persona	Activities
Operations	Person who implements development work performed by data engineers, using orchestration, CI/CD, and other tooling into production. They provide environments for use by both data engineers and data scientists, build VM images for use of third-party data science solutions, such as R Studio and Anaconda. They set up Cloud Storage bucket structures for the data lake and data warehouse, and create the BigQuery datasets.
Operational identities	Service accounts used to run pipelines.

The preceding table treats personas as individuals for clarity, but best practice is to use groups to manage access to e Cloud resources (/iam/docs/using-iam-securely#policy_management).

The following table lists the customer job roles and activities, and how they map to preceding personas for this article's example architecture.

Customer job role	Activities	Personas
Ad hoc data uploader	Uploading data directly to Cloud Storage.	Data uploader
Data engineer	Developing pipelines for moving data to Cloud Storage and BigQuery, creating tables in BigQuery from Cloud Storage, and making pipelines operational.	Data engineer, operations
Data warehouse business analyst (DW-BA)	Views their own data after it is loaded into the data warehouse.	Data analyst, Data viewer
Cross-sectional business analyst (CS-BA) Marketing analyst	Views a predefined set of data after it is loaded into the data warehouse. (Equivalent to the access of multiple DW-BAs.)	Data analyst, Data viewer
Super business analyst	Views all data in the data lake and data warehouse, uses tools like Dataprep in the data lake.	Data analyst, Data viewer

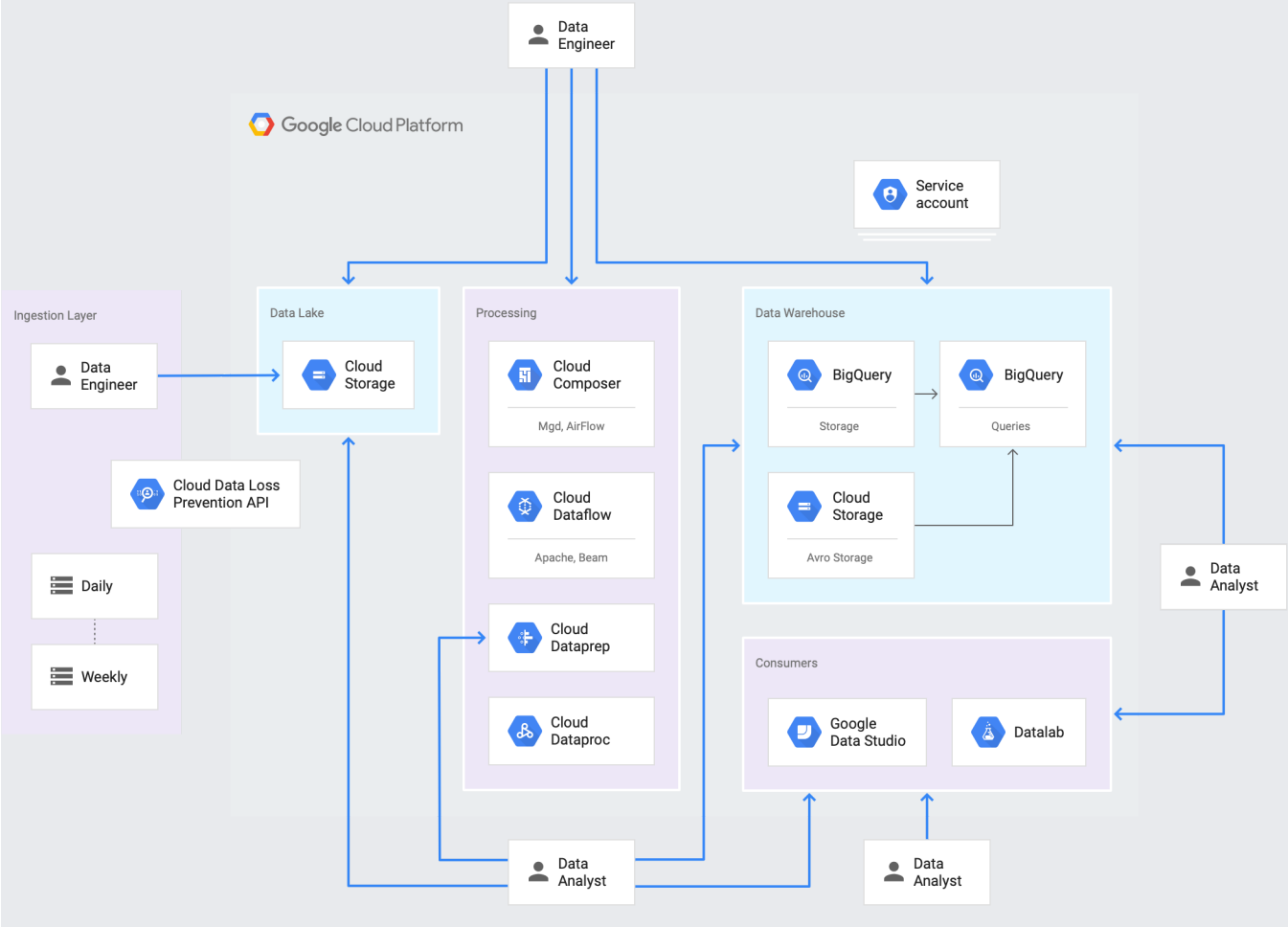
Often customer job roles don't map directly to the preceding personas. Job roles might exist that combine individual personas. The key is to assign someone with the skills to perform the activities.

It's a good idea to follow the principle of least privilege (https://wikipedia.org/wiki/Principle_of_least_privilege). This section discusses your access control options in Cloud IAM, BigQuery, and Cloud Storage.

You use Cloud IAM (/iam/docs/overview) to grant permissions to the Google Cloud resources that make up the architecture. Before you grant permissions, you must understand exactly which activities people perform along the pipeline. This helps you determine the access levels required by each job role.

Follow Cloud IAM best practices guidance (/iam/docs/using-iam-securely) to define your Cloud IAM access control policies.

The following diagram shows how the job roles in the example architecture interact with the data and services, and shows where those interactions occur.



Data is uploaded to the data lake through an automated process, and occasionally through an ad hoc process. The automated process can only add data to the data lake (that is, it can't read, delete, or modify data in the data lake).

The service account that runs the automated upload process, and any ad hoc data loaders, must have the following Cloud IAM role to work with the resources they need.

Role	Resources	Members	Permissions
	Cloud Storage buckets	Service account to run automated uploads Ad hoc data uploader (Super business analyst)	Allows the uploader application to create objects (but not view, delete, or overwrite objects).

In the example architecture, data is uploaded to specified buckets. The bucket location is defined when you invoke the uploader process. You can create separate uploader groups (and, if necessary, service accounts) to further segregate who can access which buckets.

Before you upload the data into the data lake, you might be required to tokenize or redact sensitive data. You can do so using the Cloud DLP API, which is explained later in this article.

Because the data lake stores data in its raw format, often the data must be processed before you can load it into the data warehouse. For example, processing might include data cleaning, deduplication, and transformation. Google Cloud tools for data processing include Cloud Composer, Dataproc, Dataflow, BigQuery, and Dataprep.

Data engineers, and the service account that runs Cloud Composer (/composer/docs/concepts/overview) jobs, must have the following Cloud IAM roles to work with the resources they need.

Role	Resources	Members	Permissions
	project	Data engineer	Grants a data engineer full control of Cloud Composer resources. (This role does not give direct access to the data in the buckets.)

Role	Resources	Members	Permissions
	project	Service account for running Cloud Composer	Grants a service account the permissions required to run a Cloud Composer environment on the VM it is associated with.

Data engineers can access the Cloud Composer web interface (/composer/docs/how-to/accessing/airflow-web-interface#accessing_the_web_interface_via_the) through Identity-Aware Proxy (IAP) (#define_policies_that_grant_access_with_cloud_iam). Cloud Composer, a workflow orchestration service built on Apache Airflow, makes calls to more APIs, for example, to start up a Dataproc cluster. The service account associated with your Cloud Composer environment (/composer/docs/how-to/managing/creating#before_you_begin) needs permissions so it can use those resources. At a minimum, the service account must have the `roles/composer.worker` permissions (/composer/docs/how-to/access-control#roles).

The Dataflow service account must have the following Cloud IAM roles to work with the resources it needs.

Role	Resources	Members	Permissions
	organization	Service account for executing Dataflow work units.	Grants a service account permission to execute work units for a Dataflow pipeline.

You can visually explore and transform raw data from disparate and large datasets with Dataprep. You use Dataprep separately from the automated process that Dataflow and Cloud Composer use.

:ant: This is a service provided with Trifacta, a third-party partner of Google Cloud. Sampling data is processed outsi
roject.

The Dataprep service account and the business analyst must have the following Cloud IAM roles to work with the resources they need.

Role	Resources	Members	Permissions
	project	Dataprep service account	Grants the Dataprep service account permission to access and modify datasets and storage, and run and manage Dataprep jobs within a project.
	project	Data analysts	Allows a person to run the Dataprep application.

In the example scenario, different data analysts have different levels of access to the data in the data warehouse.

The data warehouse business analyst (DW-BA) has a view of their data after it has been loaded into the data warehouse.

The DW-BA has the following Cloud IAM roles to work with the resources they need.

Role	Resources	Members	Permissions
	project	DW-BA	Allows the DW-BA to run queries against datasets defined by permissions granted in the **`bigquery.dataviewer`** role.
	BigQuery datasets	DW-BA	Allows DW-BAs with the **`bigquery.user`** role on the project to run queries against their data in the specified dataset.
	bucket	DW-BA	Grants permission to view the federated data source (because data in the data warehouse is stored in Cloud Storage and BigQuery).

The DW-BA can start up Dataproc clusters and run Hive queries across their data with a user-managed service account (/dataproc/docs/concepts/configuring-clusters/service-accounts). The service account must have the following Cloud IAM roles to work with the resources it needs.

Role	Resources	Members	Permissions
	project	DW-BA	Grants permissions to start Dataproc clusters and run jobs, which is necessary to run Hive queries across the data stored in Cloud Storage.

The DW-BA does not require the ability to upload files to Cloud Storage, because the files are already uploaded.

The DW-BA must have the following Cloud IAM roles to view job output on the specified bucket.

Role	Resources	Members	Permissions
	bucket	DW-BA	Grants permission to view the federated data source.

Follow the best practice of using a user-managed service account to start Dataproc clusters. That way, long-running jobs can carry on after the user who originally started the job has access rescinded. You can also create fine-grained access and control for clusters.

The user-managed service account must have the following Cloud IAM role to work with the resources it needs.

Role	Resources	Members	Permissions
	project	Dataproc service account	Allows the service account to start and run Dataproc clusters.

To achieve further granularity with your permissions, see Dataproc granular Cloud IAM (/dataproc/docs/concepts/iam/granular-iam).

In this scenario, the cross-sectional business analyst (CS-BA) or marketing analyst can view a predefined set of data after it is loaded into the data warehouse. The access is the equivalent of multiple DW-BA views.

In our example, the datasets and Cloud Storage buckets that the CS-BA can view are located in the same project.

The CS-BA must have the following Cloud IAM roles to work with the resources they need.

Role	Resources	Members	Permissions
	project	CS-BA	Allows the CS-BA to run queries against datasets for which they have the **`bigquery.dataviewer`** role.
	project	CS-BA	Allows the CS-BA to enumerate all datasets in the project and read dataset metadata, list tables in the dataset, and read data and metadata from the dataset tables.
	project	CS-BA	Grants permission to view the federated data source.

If you're working with a manageable number of datasets and federated buckets, using DW-BA configuration is enough. (You must grant the appropriate permissions on each bucket and each dataset, rather than granting permissions at the project level.)

In this scenario, the super business analyst (S-BA) can view all data in the data lake after it is loaded into the data warehouse.

The S-BA must have the following Cloud IAM roles to work with the resources they need.

Role	Resources	Members	Permissions
	organization	S-BA	Allows the S-BA to run queries against datasets for which they have the `bigquery.dataviewer` role.

The `bigquery.user` role does not give users permission to query data, view table data, or view table schema details for datasets the user did not create.

Role	Resources	Members	Permissions
	organization	S-BA	Allows S-BAs to enumerate all datasets in the project and read datasets metadata, list tables in the dataset, and read data and metadata from the datasets tables.
	organization	S-BA	Grants permission to view the federated data source.

The S-BA can also use Dataprep to help transform the data to be loaded into the data warehouse.

Business Analysts can also use reports that are generated by Data Studio and Datalab. Datalab uses the Datalab VM service account. Before you can run the notebook to generate the report, you must grant the business analyst and service account the following Cloud IAM roles.

Role	Resources	Members	Permissions
	service account	Business analyst	Grants the business analyst access to connect to the Datalab instance. They must have the `serviceAccountUser` role for the service account that started the Datalab instance.

Datalab requires that individual users be granted access to a single instance.

You must grant access to BigQuery, Cloud Storage, and Dataflow to the service account used to start the Datalab instance. The service account must have the following Cloud IAM roles to work with the resources they need.

Role	Resources	Members	Permissions
	organization	service account	Allows the Datalab service account to run queries on datasets on which it has the `bigquery.dataviewer` role.
	organization	service account	Allows the service account to enumerate all datasets in the project, read metadata, list tables in the dataset, and read data and metadata from the tables.
	organization	service account	Grants permission to view the federated data source.

Data Studio must have credentials for the appropriate data source in order to access to BigQuery datasets. For information, see Data source credentials. (https://support.google.com/datastudio/answer/6371135)

You can manage fine-grained access control of the views of the data in BigQuery, for example, when you have several business analysts who need different levels of access. Here's a scenario:

- The data warehouse business analyst (DW-BA) has a view of their data.
- The cross-sectional business analyst (CS-BA) or marketing analyst has a view of a predefined set of data after it is loaded into the data warehouse. The access required is the equivalent of multiple DW-BA views.
- The super business analyst (S-BA) has a view of all data in the data lake, or after it is loaded into the data warehouse.

In addition to Cloud IAM permissions, you must configure authorized views (/bigquery/docs/authorized-views).

Authorized views allow you to share query results with particular users and groups without giving them access to the underlying tables. In the example scenario, you can provide the curated views for the DW-BA and CS-BA. The view is created in a dataset that is separate from the source data that is queried by the view. You grant the business analyst access to the dataset based on the view.

For guidance on implementing restricted access to BigQuery datasets, see Secure data workloads use case: Limit access to data for specific identities

(/solutions/secure-data-workloads-use-cases#limit_access_for_specific_identities).

In this scenario, you don't need to configure row-level permissions (/bigquery/docs/authorized-views#row-level-permissions). You can, however, display different rows to different users if your scenario requires it. You add another field to your tables that contains the user who is allowed to see the row. Then, you create a view that uses the `SESSION_USER()` function. The `SESSION_USER()` function returns the current user (the email address they authenticate against Google Cloud with). If `SESSION_USER()` returns a user that is contained in the field you added, the user can view the data in that row.

Usually, Cloud IAM is the right choice for managing access to buckets and the objects in them, which is the approach shown in the pipeline. Cloud Storage has more access control mechanisms (/storage/docs/access-control/), however, which you might use in a pipeline when you want to grant access to a specific object within a bucket. To do so, you use access control lists (/storage/docs/access-control/lists) or signed URLs (/storage/docs/access-control/signed-urls).

You can use the Organization Policy (/resource-manager/docs/organization-policy/overview) to configure restrictions on supported resources. You configure constraints (/resource-manager/docs/organization-policy/overview#constraints) against the supported resources. The constraints that apply to the sample pipeline are the following:

- **Domain-restricted sharing.** Restrict the set of users who can be added to Cloud IAM policies in the organization where your pipeline is configured. The allowed/denied list must specify one or more G Suite or Cloud Identity customer IDs.

  To use Cloud Composer you must disable the policy constraint before creating an environment (/composer/docs/how-to/managing/creating#before_you_begin) so that Cloud Composer can apply the required ACLs to the Cloud Storage bucket for your environment. You can re-enable the policy constraint after you create the environment.

  For guidance on implementing domain-restricted sharing, see Secure data workloads use cases: prevent access by non-domain identities (/solutions/secure-data-workloads-use-cases#prevent_access_by_non-domain_identities)

- **Disable service account key creation.** Prevents the creation of service account external keys by setting this constraint to `TRUE`.

- **Enforce bucket policy only.** Disables the evaluation of ACLs assigned to Cloud Storage objects in the bucket, so that only Cloud IAM policies grant access to objects in these buckets.

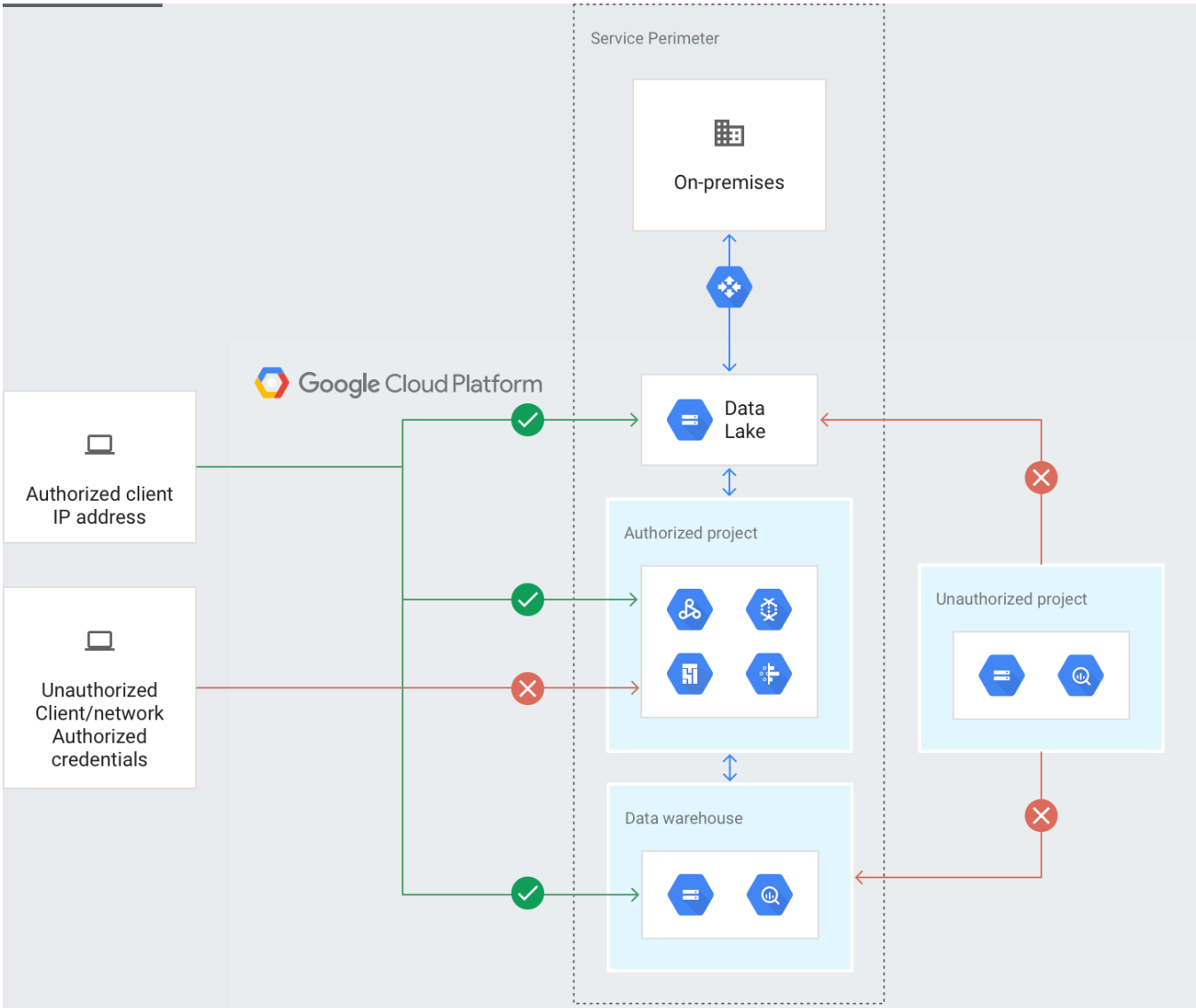
Identity-Aware Proxy (/iap/) (IAP) establishes a central authorization layer for applications accessed by HTTPS that are hosted on Google Cloud. When an application or resource is protected by IAP, it can only be accessed through the proxy by users who have the correct Cloud IAM role. When a user tries to access an IAP-secured resource, IAP performs authentication and authorization checks. In the example pipeline, IAP is used to access the web interface (/composer/docs/how-to/accessing/airflow-web-interface#accessing_the_web_interface_via_the) for Cloud Composer.

By configuring VPC Service Controls, you can define a security perimeter around Google Cloud resources like Cloud Storage buckets and BigQuery datasets. You constrain data within a Virtual Private Cloud (VPC) (/vpc/), which helps to mitigate data exfiltration risks.

Private Google Access for on-premises (/vpc/docs/configure-private-google-access-hybrid)enables on-premises hosts to reach Google APIs and services over a Cloud VPN (/vpn/docs) or Cloud Interconnect (/interconnect/docs) connection from your data center to Google Cloud. On-premises hosts don't need external IP addresses; instead, they use internal RFC 1918 (https://tools.ietf.org/html/rfc1918) IP addresses.

The following sample architecture restricts access to the projects that contain the data lake and data warehouse by complementing the Cloud IAM controls with Private Google Access and VPC Service Controls (/vpc-service-controls/docs/on-premises-access). Emergency access for data engineers and operators is implemented in case private communication between on-premises and Google Cloud is unavailable. Context-Aware Access controls are also configured.

This configuration is illustrated in the following architecture diagram:



For guidance on implementing VPC Service Controls to help mitigate data exfiltration, see mitigate data exfiltration for apps (/solutions/secure-data-workloads-use-cases#mitigate_data_exfiltration_for_apps) and mitigate data exfiltration for people (/solutions/secure-data-workloads-use-cases#mitigate_data_exfiltration_for_people).

For guidance on implementing managed access to Google Cloud APIs, see Managed access to Google Cloud APIs (/solutions/secure-data-workloads-use-cases#managed_access_to_gcp_apis).

Cloud Audit Logs (/logging/docs/audit/) consists of three audit log streams for each project, folder, and organization:

- Admin activity
- System event
- Data access

Google Cloud services write audit log entries to these logs to help you answer the questions "who did what, where, and when?" within your Google Cloud projects.

Admin activity logs (/logging/docs/audit/#admin-activity) contain log entries for API calls or other administrative actions that modify the configuration or metadata of resources. Admin activity logs are always enabled. There's no charge for admin activity audit logs, and they're retained for 13 months (400 days).

Data access logs (/logging/docs/audit/#data-access) record API calls that create, modify, or read user-provided data. Data access audit logs are disabled by default except in BigQuery, because they can grow to be large.

System Event logs (/logging/docs/audit/#system-event) contain log entries for when Compute Engine performs a system event. For example, each live migration (/compute/docs/instances/live-migration) is recorded as a system event. There is no charge for your System Event audit logs.

In the example pipeline, you audit both admin and data access logs. The following services have data access audit logs configured (/logging/docs/audit/configure-data-access) for the example architecture:

- BigQuery (/bigquery/docs/reference/auditlogs/)
- Dataproc (/dataproc/)
- Cloud Storage (/storage/docs/audit-logs)
- Cloud DLP (/dlp/docs/audit-logging) (Cloud DLP)
- Cloud Key Management Service (/kms/docs/logging) (Cloud KMS)

BigQuery data access logs are enabled by default and do not count against your logs allotment.

Audit logging Cloud IAM roles are configured to restrict access to the logs (/iam/docs/roles-audit-logging). Log exports (/logging/docs/export/) (not shown) are also configured to provide a way to collate and retain logs beyond the default retention period. See Design patterns for exporting Stackdriver Logging (/solutions/design-patterns-for-exporting-stackdriver-logging) for examples of scenarios and how to configure an export logging strategy.

Personally identifying information (https://wikipedia.org/wiki/Personal_data), or PII, is any information related to identifying a specific individual.

Google Cloud encrypts customer data stored at rest (/security/encryption-at-rest/)by default, with no additional action required from you.

Data in Google Cloud is broken into subfile chunks for storage, and each chunk is encrypted at the storage level with an individual encryption key. The key used to encrypt the data in a chunk is called a *data encryption key (DEK)*. Because of the high volume of keys at Google, and the need for low latency and high availability, these keys are stored near the data that they encrypt. The DEKs are encrypted with (or "wrapped" by) a key encryption key (KEK). Customers can choose which key management solution they prefer for managing the KEKs that protect the DEKs that protect their data.

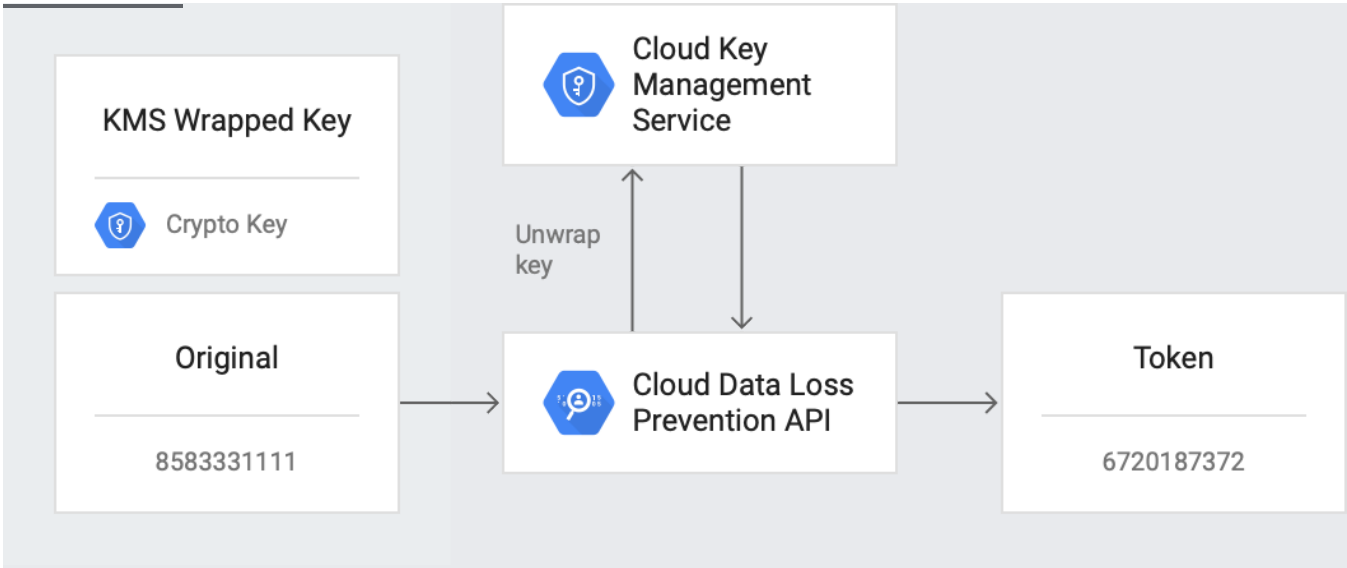
For sensitive operations, you may need to generate and manage your own key encryption keys using customer-supplied encryption keys (CSEK) or you can manage encryption keys using Cloud KMS.

In our example pipeline, we have a requirement to manage keys using Cloud KMS (/kms/)when using Cloud DLP to tokenize sensitive data.

The Cloud DLP (DLP (/dlp/)) API provides programmatic access to a powerful sensitive data inspection, classification, and deidentification platform.

Data is processed by the DLP API. Then the processed data can be written to a sink.

If you are required for policy or compliance reasons to identify sensitive data items and tokenize those items before writing data to the data lake, you can use the DLP API together with Cloud KMS. The DLP API can be used to tokenize sensitive data items as part of the upload process. If you also need to de-tokenize (reveal the original raw data item), you can use the KMS key and cryptographic hash used to initially tokenize the data items.



For details on how to implement the tokenization/de-tokenization process, see deidentifying sensitive data in text (/dlp/docs/deidentify-sensitive-data#cryptoreplaceffxfpeconfig).

The sample pipeline architecture uses Cloud DLP at the ingestion stage to classify the data when the data is uploaded to the data lake. Any sensitive data detected is tokenized using the key managed by Cloud KMS.

- To learn about building a data warehouse using BigQuery, see BigQuery for data warehouse practitioners (/solutions/bigquery-data-warehouse).
- For information on building a data lake using Google Cloud, see Build a data lake (/solutions/build-a-data-lake-on-gcp).
- Learn about the Google Cloud products that help to secure data workloads (/solutions/secure-data-workloads-gcp-products).
- For information about envelope encryption, see envelope encryption (/kms/docs/envelope-encryption).
- Try out other Google Cloud features for yourself. Have a look at our tutorials (/docs/tutorials).