What is Dataplex?

Dataplex is an intelligent data fabric that helps you unify distributed data and automate data management and governance across that data to power analytics at scale.

Additionally, you can discover and curate metadata across various silos using catalog capabilities in Dataplex. See Data Catalog Overview (/data-catalog/docs/concepts/overview).

For data stored in Cloud Storage and BigQuery, Dataplex enables you to:

- Build a domain-specific data mesh across data stored in multiple Google Cloud projects, without any data movement.
- Consistently govern and monitor data with a single set of permissions.
- Automate metadata discovery and make the data securely accessible and available for querying via BigQuery as external tables, and open source applications such as SparkSQL, Presto, and HiveQL.
- Run data quality and data lifecycle management tasks, including serverless Spark tasks.
- Explore data using fully-managed, serverless Spark environments with simple access to notebooks and SparkSQL queries.

Why use Dataplex?

Enterprises have data distributed across data lakes, data warehouses, and data marts. Dataplex enables you to discover, curate, and unify this data without any data movement, organize it based on your business needs, and centrally manage, monitor, and govern this data. Dataplex enables standardization and unification of metadata, security policies, governance, classification, and data lifecycle management across this distributed data.

How Dataplex works

Dataplex manages data in a way that doesn't require data movement or duplication. As you identify new data sources, Dataplex harvests the metadata for both structured and unstructured data, using built-in data quality checks to enhance integrity.

Dataplex automatically registers all metadata in a unified metastore. You can also access data and metadata through a variety of Google Cloud services, such as BigQuery, Dataproc Metastore, Data Catalog, and open source tools, such as Apache Spark and Presto.

Terminology

Dataplex abstracts away the underlying data storage systems, by using the following constructs:

- **Lake**: A logical construct representing a data domain or business unit. For example, to organize data based on group usage, you can set up a lake per department (for example, Retail, Sales, Finance).
- **Zone**: A sub-domain within a lake, useful to categorize data by stage (for example, `landing`, `raw`, `curated_data_analytics`, `curated_data_science`), usage (for example, data contract), or restrictions (for example, security controls, user access levels). Zones are of two types, raw and curated.
    - **Raw zone**: Data that is in its raw format and not subject to strict type-checking.

    - **Curated zone**: Data that is cleaned, formatted, and ready for analytics. The data is columnar, Hive-partitioned, in Parquet, Avro, Orc files, or BigQuery tables. Data undergoes type-checking, for example, to prohibit the use of CSV files because they do not perform as well for SQL access.
- **Asset**: An asset maps to data stored in either Cloud Storage or BigQuery. You can map data stored in separate Google Cloud projects as assets into a single zone.
- **Entity**: An entity represents metadata for structured and semi-structured data (table) and unstructured data (fileset).

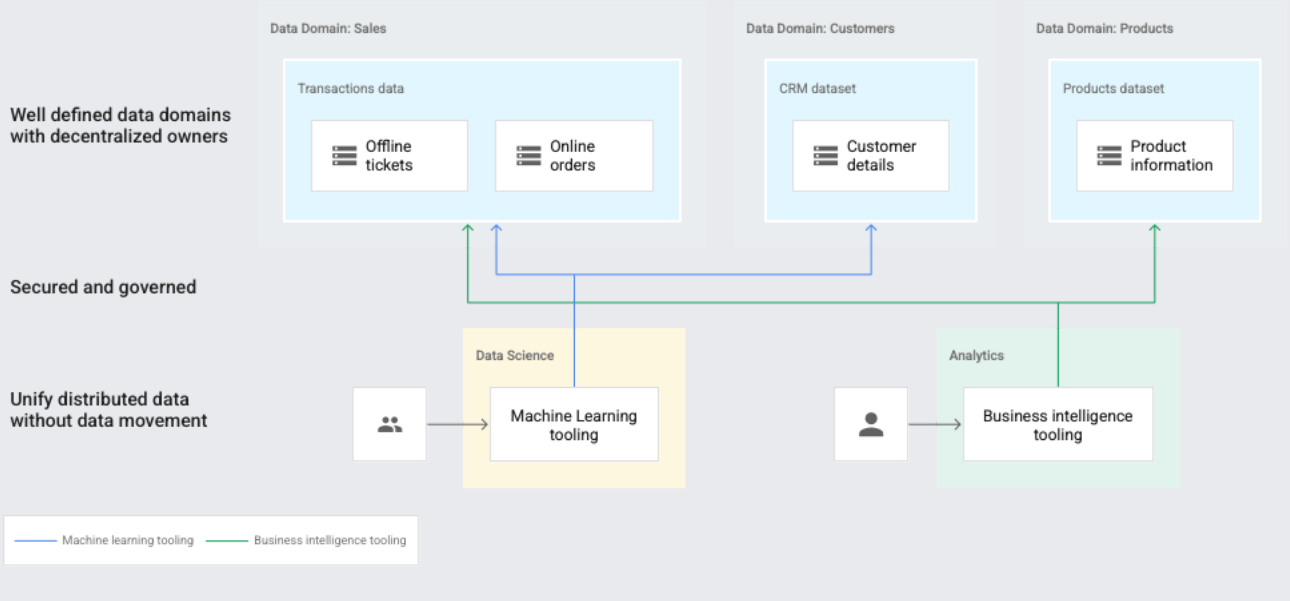
Common use cases

This section outlines the most common use cases for using Dataplex.

A domain-centric data mesh

With this type of data mesh, data is organized into multiple domains within an enterprise, for example, Sales, Customers, and Products. Ownership of the data can be decentralized. Users could subscribe to data from different domains. For example, data scientists and data analysts could pull from different domains to accomplish business objectives like machine learning and business intelligence.

In the following diagram, domains are represented by Dataplex lakes and owned by separate data producers. Data producers own creation, curation, and access control in their domains. Data consumers can then request access to the lakes (domains) or zones (sub-domains) for their analysis.



In this case, data stewards need to retain a holistic view of the entire data landscape.

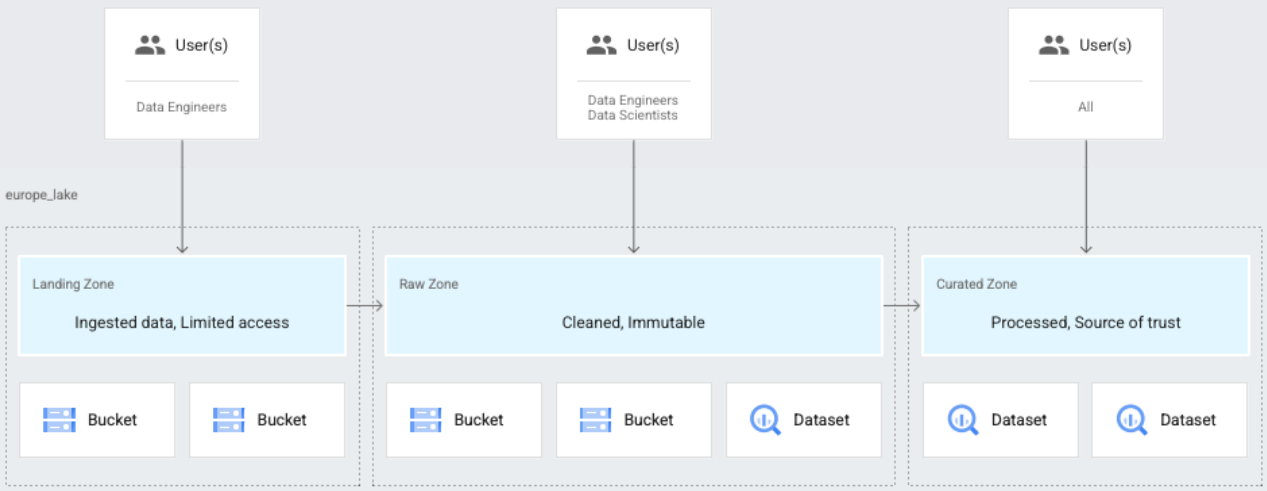
- Dataplex: A mesh of multiple data domains.
- Domain: Lakes for Sales, Customers, and Products data.
- Zone within a domain: For individual teams or to provide managed data contracts.
- Assets: Data stored in either a Cloud Storage bucket or a BigQuery dataset, which can exist in a separate Google Cloud project from your Dataplex mesh.

You can extend this scenario by breaking down data within zones into raw and curated layers. Do this by creating zones for each permutation of a domain and raw or curated data:

- Sales Raw
- Sales Curated
- Customers Raw
- Customers Curated
- Products Raw
- Products Curated

Data tiering based on readiness

Another common use case is when your data is accessible only to data engineers and is later refined and made available to data scientists and analysts. In this case, you can set up a lake to have a raw zone for the data that the engineers have access to, and a curated zone for the data that is available to the data scientists and analysts.



What's next

- Get started (/dataplex/docs/quickstart-guide) with Dataplex
- Build a data mesh (/dataplex/docs/build-a-data-mesh)
- Learn best practices (/dataplex/docs/best-practices)
- Create a lake (/dataplex/docs/create-lake)
- Add zones to your lakes (/dataplex/docs/add-zone)
- Attach assets to your zones (/dataplex/docs/manage-buckets)
- Discover catalog capabilities in Dataplex (/data-catalog/docs/concepts/overview)

Last updated 2022-09-30 UTC.