Google Cloud

☰ Blog

Google Cloud
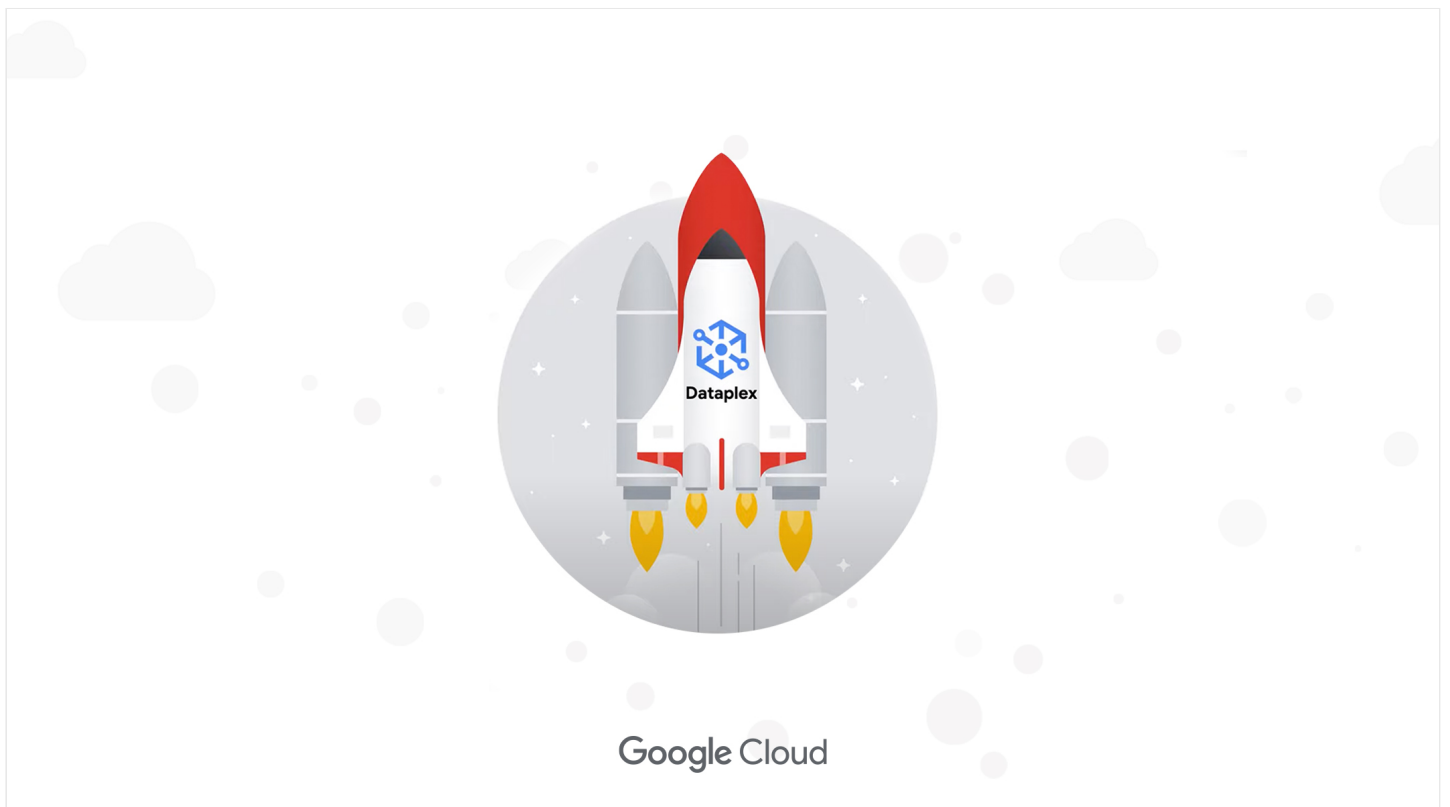
Blog

Google Cloud

Blog

now generally available

February 23, 2022

Prajakta Damle
Sr. Product Manager, Google Cloud

Nikhil Kothari
Engineering Lead, Dataplex and Open Data Analytics

Democratizing data insights and accelerating data-driven decision making is a
top priority for most enterprises seeking to build a data cloud. This often requires
building a self-serve data platform that can span data silos and enable at-scale

Google Cloud

Blog

distributed data landscape.

Today we are excited to announce the general availability of [Dataplex](), an intelligent data fabric that enables you to centrally manage, monitor, and govern data across data lakes, data warehouses, and data marts, and make this data securely accessible to a variety of analytics and data science tools.

With Dataplex, enterprises can easily delegate ownership, usage, and sharing of data, to data owners who have the right business context, while still having a single pane of glass to consistently monitor and govern data across various data domains in their organization. With built-in data intelligence, Dataplex automates the data discovery, data lifecycle management, and data quality, enabling data productivity and accelerating analytics agility.

Here is what some of our **customers** have to say,

"We have PBs of data stored in GCS and BigQuery in GCP, accessed by 1000s of internal users daily" said **Saral Jain, Director of Engineering, Snap Inc**. *"Dataplex enables us to deliver a business domain specific, self-service data*

Google Cloud

Blog

Dataplex as a central component for building a unified data mesh across our analytics data."

"As the central data team at Deutsche Bank, we are building a data mesh to standardize data discovery, access control and data quality across the distributed domains," said **Balaji Maragalla, Director Big Data Platform at Deutsche Bank**. *"To help us on this journey, we are excited to use Dataplex to enable centralized governance for our distributed data. Dataplex formalizes our data mesh vision and gives us the right set of controls for cross-domain data organization, data security, and data quality."*

"As one of the largest entertainment companies in Japan, we generate TBs of data everyday and use it to make business critical decisions", said **Kazumasa Iwao, Director of Data Analytics at DeNA**. *"While we manage each product independently as a separate domain, we want to centralize governance of data across our products. Dataplex enables us to effectively manage and standardize data quality, data security, and data privacy for data across these domains. We are looking forward to building trust in our data with Google Cloud's Dataplex."*

One of the key use cases that Dataplex enables is a data mesh architecture. Let's take a closer look at how you can use Dataplex as the data fabric that enables a data mesh.

What is a Data Mesh?

With enterprise data becoming more diverse and distributed, and the number of tools and users that need access to this data growing, organizations are moving away from monolithic data architectures that are domain agnostic. While monolithic, centrally managed architectures create data bottlenecks and impact analytics agility, a completely decentralized architecture where business domains maintain their own purpose-built data lakes also has its pitfalls and results in data duplication and silos, making governance of this data impossible.

Google Cloud

Blog

describes a modern data stack that moves away from a monolithic data lake or data warehouse architecture to a distributed domain-specific architecture that enables autonomy of data ownership, provides agility with decentralized domain aware data management while providing the ability to centrally govern and monitor data across domains. To learn more, refer to this [Build a Modern Distributed Data Mesh Whitepaper](#).

How to make Data Mesh real with Google Cloud

Dataplex provides a data management platform to easily build independent data domains within a data mesh that spans your organization while still maintaining central controls for governing and monitoring the data across domains.

"Dataplex is embodying the principles of Data Mesh as we have envisioned in Adeo. Having a first party, cloud-native, product to architect a Data Mesh in GCP is crucial for effective data sharing and data quality amongst teams. Dataplex streamlines productivity, allowing teams to build data domains and orchestrate data curation across the enterprise. I only wish we had Dataplex three years ago." —**Alexandre Cote, Product Leader with ADEO**

Imagine you have the following domains in your organization,

Google Cloud

Blog
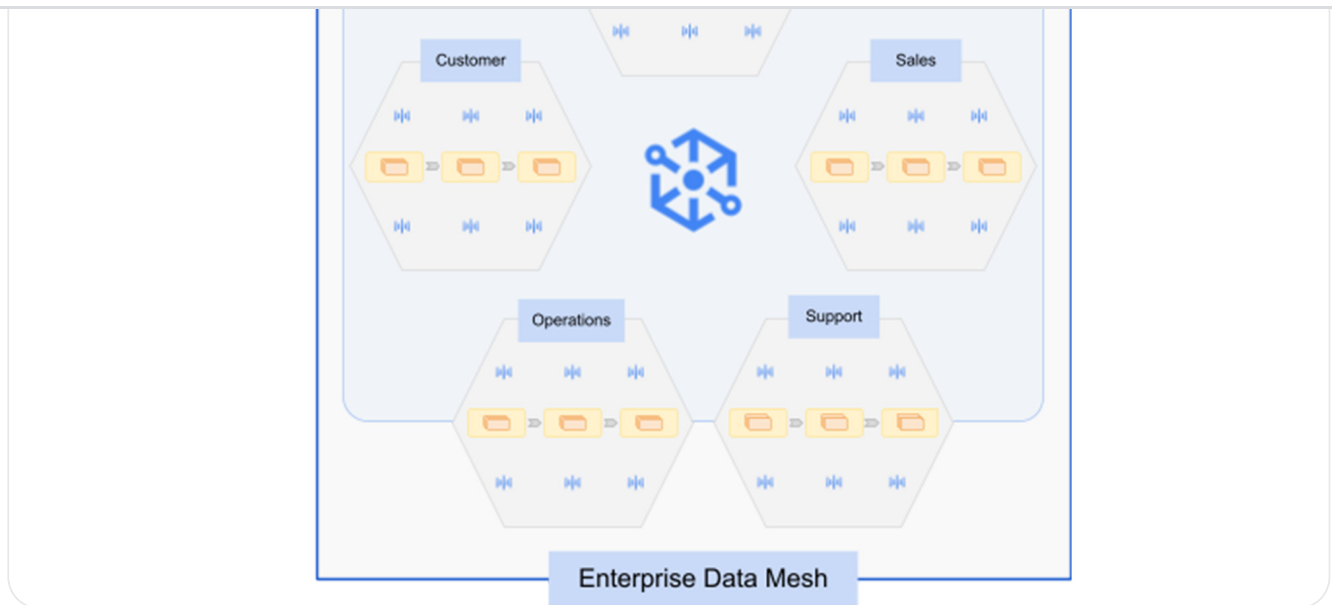
With Dataplex you can logically organize your data and related artifacts such as code, notebooks, and logs, into a Dataplex Lake which represents a data domain.

You can model all the data in a particular domain as a set of Dataplex Assets within a lake without physically moving data or storing it into a single storage system. Assets can refer to [Cloud Storage](#) buckets and [BigQuery](#) datasets, stored in multiple Google Cloud projects, and manage both analytics and operational data, structured and unstructured data that logically belongs to a single domain. Dataplex Zones enable you to group assets and add structure that capture key aspects of your data - its readiness, the workloads it is associated with, or the data products it is serving.

The lakes and data zones in Dataplex enable you to unify distributed data and organize it based on the business context. This forms the foundation for managing metadata, setting up governance policies, monitoring data quality and so on, giving you the ability to manage your distributed data at scale.

Now let's take a look at one of the domains in a little more detail.

Google Cloud

Blog
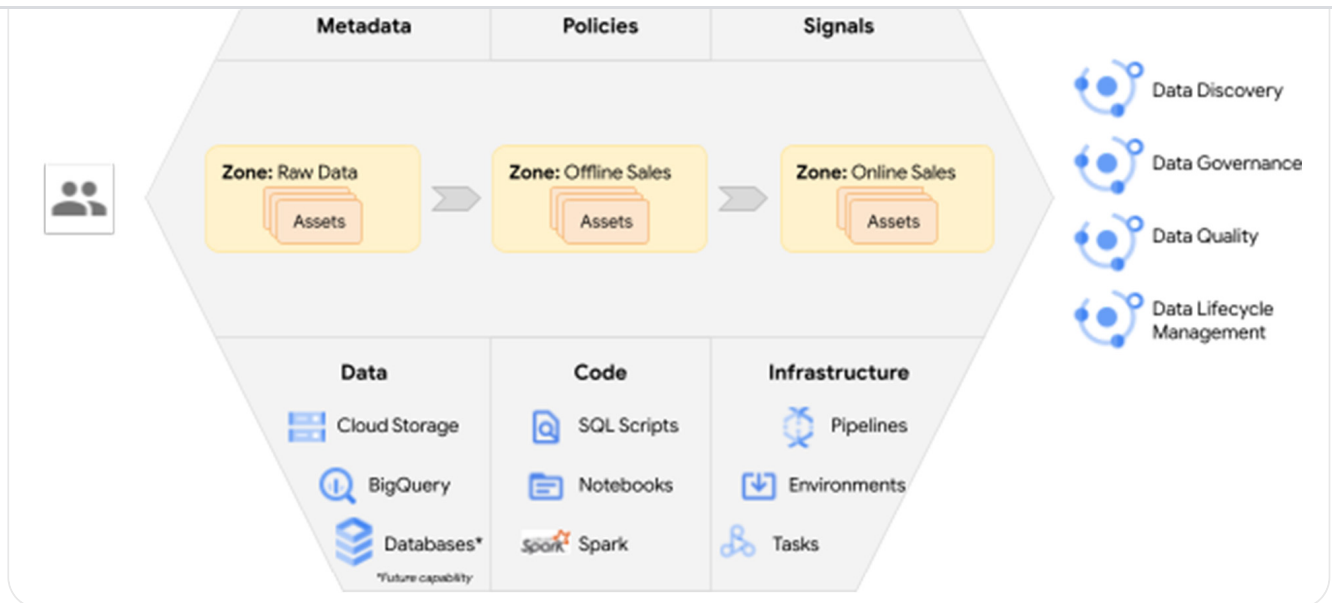
Automatically discover metadata across data sources: Dataplex provides [metadata management](#) and cataloging that enables all members of the domain to easily search, browse and discover the tables and filesets as well as augment them with business and domain-specific semantics. Once data is added as assets, Dataplex automatically extracts associated metadata and keeps it up-to-date as data evolves. This metadata is made available for search, discovery, and enrichment via integration with [Data Catalog](#).

Enable interoperability of tools: The [metadata](#) curated by Dataplex is automatically made available as runtime metadata to power federated open source analytics via Apache SparkSQL, HiveQL, Presto, and so on. Compatible metadata is also automatically published as external tables in BigQuery to enable federated analytics via BigQuery.

Govern data at scale: Dataplex enables data administrators and stewards to consistently and scalably manage their IAM data policies to [control data access](#) across distributed data. It provides the ability to centrally govern data across domains while enabling autonomous and delegated ownership of data. It provides the ability to manage reader/writer permissions on the domains and the

Google Cloud

Blog

as data quality tasks across your data in BigQuery and GCS.

One-click data exploration: Dataplex enables data engineers, data scientists and data analysts with a built-in, self-serve, serverless [data exploration experience](#) to interactively explore data and metadata, iteratively develop scripts, and deploy and monitor data management workloads. It provides content management across SQL scripts and Jupyter notebooks that makes it easy to create domain-specific code artifacts and share or schedule them from that same interface.

Data management: You can also leverage the built-in data management tasks that address common tasks such as tiering, archiving or refining data. It integrates with Google Cloud's native data tools such as [Dataproc Serverless](#), [Dataflow](#), [Data Fusion](#), and BigQuery to provide an integrated data management platform.

With the collective of data, metadata, policies, code, interactive and production analytics infrastructure, and data monitoring, Dataplex delivers on the core value proposition of a data mesh: data as the product.

"Consistent data management and governance of distributed data remains a top priority for most of our clients today. Dataplex enables a business-centric data mesh architecture and significantly lowers the administrative overhead associated with managing, monitoring, and governing distributed data. We are excited to collaborate with the Dataplex team to enable enterprise clients to be more data-driven and accelerate their digital transformation journeys."—**Navin Warerkar, Managing Director, Deloitte Consulting LLP, and US Google Cloud Data & Analytics GTM Leader**

Next steps

Google Cloud

Blog
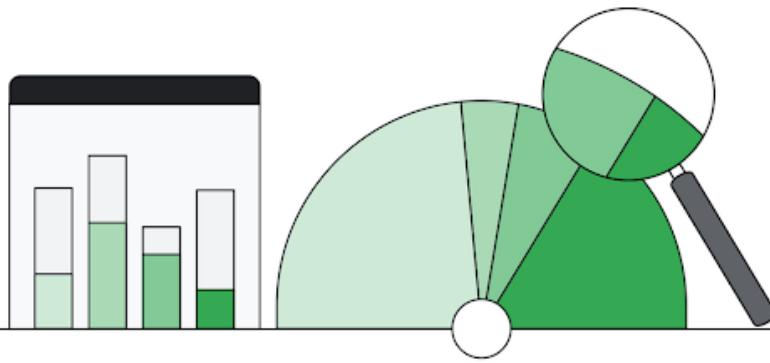
Data Analytics

Introducing Dataplex—an intelligent data fabric for analytics at scale

Dataplex unifies distributed data to help automate data management and power analytics at scale.

By Prajakta Damle • 6-minute read

Posted in Data Analytics—Google Cloud

Related articles

AI & Machine Learning

Perform hyperparameter tuning using R and caret on Vertex AI

By Fabian Hirschmann • 5-minute read

Google Cloud

Blog

Data Analytics

Automate data governance, extend your data fabric with Dataplex-BigLake integration

By Uri Gilad • 2-minute read

Data Analytics

Understand and optimize your BigQuery analytics queries using the query execution graph

By Vinay Yerramilli • 2-minute read

Follow us

Twitter   YouTube   LinkedIn   Instagram   Facebook

Google Cloud   Google Cloud Products   Privacy   Terms

? Help   English