Google Cloud

Whitepaper
August 2021

Build a modern, unified analytics data platform

with Google Cloud

Firat Tekiner & Susan Pierce

Build a modern, unified analytics data platform

with Google Cloud

There is no shortage of data being created. IDC research indicates that worldwide data will grow to 175 zettabytes by 2025[1]. The volume of data being generated every day is staggering, and it is increasingly difficult for companies to collect, store, and organize it in a way that is accessible and usable. In fact, 90% of data professionals say their work has been slowed by unreliable data sources. Around 86% of data analysts struggle with data that is out of date, and more than 60% of data workers are impacted by having to wait on engineering resources each month while their data is cleaned and prepared[2].

Inefficient organizational structures and architectural decisions contribute to the gap that companies have between aggregating data and making it work for them. Companies want to move to the Cloud to modernize their data analytics systems, but that alone doesn't solve the underlying issues around siloed data sources and brittle processing pipelines. Strategic decisions around data ownership and technical decisions about storage mechanisms must be made in a holistic way to make a data platform more successful for your organization.

In this paper, we will discuss the decision points necessary in creating a modern, unified analytics data platform built on Google Cloud Platform.

Big Data has created amazing opportunities for businesses over the last two decades, however, it is complicated for organizations to present their business users with relevant, actionable and timely data. Research shows that 86% of analysts still struggle with outdated data[3] and only 32% of companies feel they are in realizing tangible value from their data[4]. The first issue is data freshness. The second issue stems from the difficulty in integrating disparate and legacy systems across silos. Organizations are migrating to the Cloud, but that does not solve the real problem of older legacy systems that might have been vertically structured to meet the needs of a single business unit.

Maturity	• Reporting & Analytics • Data Discovery & Preparation • Artificial Intelligence	"**80%** of analytics work is still **descriptive**" MIT, 2020
Silos	• Multiple Clouds • Vertical & Disparate Stacks • Data Ownership & Priorities	"**90% of employees say** that their work is slowed by **unreliable data sources**" Dimensional Data, 2020
Complexity	• Volume, Velocity & Variety • Data Recency & Quality Issues • Security & Governance	"**86% of analysts struggle** with data that's **out of date**." Dimensional Data, 2020

1 https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

2 https://www.zdnet.com/article/data-analysts-stretched-lack-engineering-resource-current-data-says-survey/

3 ibid

4 https://www.accenture.com/gr-en/insights/technology/closing-data-value-gap

In planning out organizational data needs, it's easy to over-generalize and consider a single, simplified structure where there is one set of consistent data sources, one enterprise data warehouse, one set of semantics, and one tool for business intelligence. That might work for a very small, highly centralized organization, and could even work for a single business unit with its own integrated IT and data engineering team. In practice, though, no organization is that simple and there are always surprise complexities around data ingestion, processing, and/or usage that complicate matters further.

What we have seen in talking to hundreds of customers is a need for a more holistic approach to data and analytics, a platform that can meet the needs of multiple business units and user personas, with as few redundant steps to process the data as possible. This becomes more than a new architecture or set of software components to purchase; it requires companies to take stock of their overall data maturity and make systemic, organizational changes in addition to technical upgrades.

By the end of 2024, 75% of enterprises will shift from piloting to operationalizing AI, driving a 5X increase in streaming data and analytics infrastructures[5]. It's easy enough to pilot AI with an arms-length data science team, working in a siloed environment. But the fundamental challenge that prevents those insights from getting released into production systems is the organizational and architectural friction that keeps data ownership segmented. As a result, most of the insights that are incorporated into an organization's business operations are descriptive in nature, and predictive analytics are relegated to the realm of a research team.

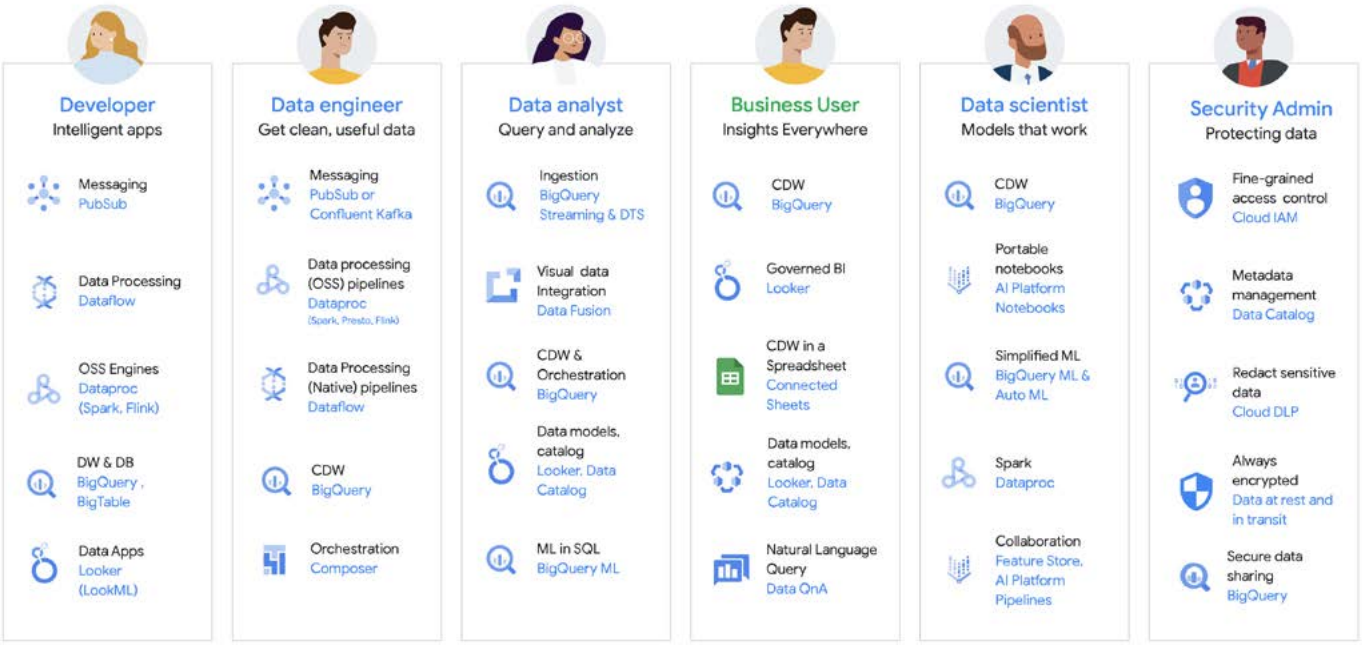
Google Cloud is changing the way businesses think about data, by focusing not just on the tech, but on the users as well.

5 https://emtemp.gcom.cloud/ngw/globalassets/en/doc/documents/721868-100-data-and-analytics-predictions-through-2024.pdf

A Platform for all users throughout the data lifecycle

Data work is rarely done by a single individual; there are many data-related users in an organization who play important roles in the data lifecycle. Each has a different perspective on data governance, freshness, discoverability, metadata, processing timelines, queryability, etc. In most cases, they are all using different systems and software to operate on the same data, at different stages of processing.

Let's look at a machine learning lifecycle, for example. A data engineer may be responsible for ensuring fresh data is available for the data science team, with appropriate security and privacy constraints in place. A data scientist may create training and test datasets based on a golden set of pre-aggregated data sources from the data engineer, build and test models, and make insights available for another team. An ML engineer may be responsible for packaging up the model for deployment into production systems, in a way that is non-disruptive to other data processing pipelines. A product manager or business analyst may be checking in on derived insights, using Data QnA (a natural language interface for analytics on BigQuery data), visualization software or might be querying the result set directly through an IDE or a command-line interface. There are countless users with different needs and we have built a compressive platform to serve them all. Google Cloud meets customers where they are with tools to meet the needs of the business.



The big data decision:

Data warehouse or data lake?

When we talk to customers about their data analytics needs, frequently we'll hear the question, "Which do I need: a data lake or a data warehouse?" Given the variety of data users and needs within an organization, this can be a tricky question to answer that depends on intended usage, types of data, and personnel.

- If you know what datasets you need to analyze, have a clear understanding of its structure, and have a known set of questions you need answered, then you are likely looking at a **data warehouse**.

- On the other hand, if you need discoverability across multiple data types, are unsure of the types of analyses you'll need to run, and are looking for opportunities to explore rather than present preset insights, and you have the resources to effectively manage and explore this environment, a **data lake** is likely going to be more suitable for your needs.

But there's more to the decision, so let's talk through some of the organizational challenges of each.

Data warehouses are often difficult to manage. The legacy systems that have worked well in the past 40 years have proven to be very expensive and pose a lot of challenges around data freshness, scaling, and high costs. Furthermore, they cannot easily provide AI or real-time capabilities without bolting that functionality on after the fact. These issues are not just present in on-premise legacy data warehouses; we even see this with the newly created cloud-based data warehouses as well. Many do not offer integrated AI capabilities, despite their claims. These new data warehouses are essentially the same legacy environments but ported over to the Cloud.

Data warehouse users tend to be analysts, often embedded within a specific business unit. They may have ideas about additional datasets that would be useful to enrich their understanding of the business. They may have ideas for improvements in the analysis, data processing, and requirements for business intelligence functionality.

However, in a traditional organization, they often don't have direct access to the data owners, nor can they easily influence the technical decision makers who decide datasets and tooling. In addition, because they are kept separate from raw data, they are unable to test hypotheses or drive a deeper understanding of the underlying data.

Data lakes have their own challenges. In theory, they are low cost and easy to scale, but many of our customers have seen a different reality in their on-premise data lakes. Planning for and provisioning sufficient storage can be expensive and difficult, especially for organizations that produce highly variable amounts of data. On-premise data lakes can be brittle and maintenance of existing systems takes time. In many cases, the engineers who would otherwise be developing new features are relegated to the care and feeding of data clusters. Said more bluntly, they are maintaining value as opposed to creating new value. Overall, the total cost of ownership is higher than expected for many companies. Not only that, governance is not easily solved across systems, especially when different parts of the organization use different security models. As a result, the data lakes become siloed and segmented, making it difficult to share data and models across teams.

Data lake users typically are closer to the raw data sources and are equipped with tools and capabilities to explore the data. In traditional organizations, these users tend to focus on the data itself and are frequently held at arm's length from the rest of the business. This disconnect means that business units miss out on the opportunity to find insights that would drive their business objectives forward to higher revenues, lower costs, lower risk, and new opportunities.

Given these tradeoffs, many companies end up with a hybrid approach, where a data lake is set up to graduate some data into a data warehouse or a data warehouse has a side data lake for additional testing and analysis. But with multiple teams fabricating their own data architectures to suit their individual needs, data sharing and fidelity gets even more complicated for a central IT team.

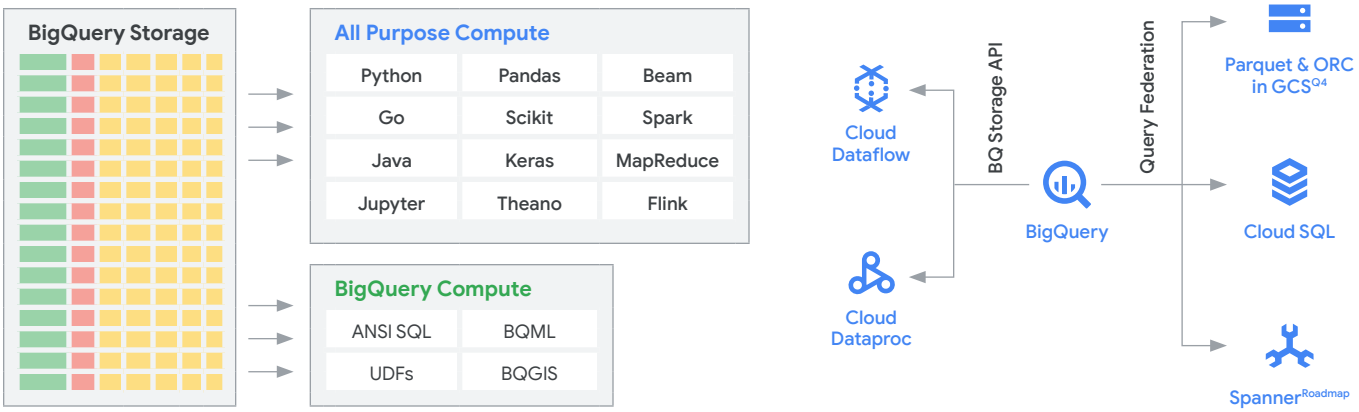
Instead of having separate teams with separate goals — where one explores the business, and another understands the business — you can unite these functions and their data systems to create a virtuous cycle where a deeper understanding of the business drives directed exploration, and that exploration drives a better understanding of the business.

	Data Warehouse (TB scale)	Data Lake (PB Scale)
Use case characteristics	Answer "known" questions Access "known" data	Answer "unknown" questions Access "unknown" data
Data type and access	Structured data SQL access manipulation	Unstructured (raw) and structured data Code-involved access and exploration
	Understanding your business	**Exploring your business**

This requires convergence in both the technology and the approach to understanding and discovering the value in your data.

Treat data warehouse storage like a data lake

You can build a data warehouse or a data lake separately on Google Cloud Platform (GCP), but you don't have to pick one or the other. In many cases, the underlying products that our customers use are the same for both, and the only difference between their data lake and data warehouse implementation is the data access policy that is employed. In fact, the two terms are starting to converge into a more unified set of functionality, a modern analytics data platform. Let's look at how this works in GCP.



BigQuery Storage API provides the capability to use BigQuery Storage like Google Cloud Storage (GCS) for a number of other systems such as Dataflow and Dataproc. This allows breaking down the data warehouse storage wall and enables running high-performance dataframes on BigQuery. In other words, the BigQuery Storage API allows your BigQuery data warehouse to act like a data lake. So what are some of the practical uses for it? For one, we built a series of connectors - MapReduce, Hive, Spark, for example - so that you can run your Hadoop and Spark workloads directly on your data in BigQuery. You no longer need a data lake in addition to your data warehouse! Dataflow is incredibly powerful for batch and stream processing. Today, you can run Dataflow jobs on top of BigQuery data, enriching it with data from PubSub, Spanner or any other data source

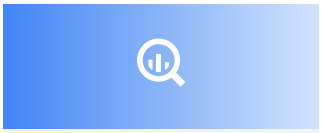
BigQuery can independently scale both storage and compute, and each is serverless, allowing for limitless scaling to meet demand no matter the usage by different teams, tools and access patterns. All of the above applications can run without impacting the performance of any other jobs accessing BigQuery at the same time. In addition, the BigQuery Storage API provides a petabit level network, moving data between nodes to fulfill a query request effectively leading to a similar performance to an in-memory operation. It also allows federating with the popular Hadoop data formats such as Parquet & ORC directly as well as NoSQL and OLTP databases. You can go a step further with the capabilities that are provided by Dataflow SQL, which is embedded in BigQuery. This allows you to join the streams with BigQuery tables or data residing in files, effectively creating a lambda architecture, allowing you to ingest large amounts of batch and streaming data, while also providing a serving layer to respond to queries. BigQuery BI Engine and Materialized Views make it even easier to increase efficiency and performance in this multi-use architecture.

Google's smart analytics platform powered by BigQuery

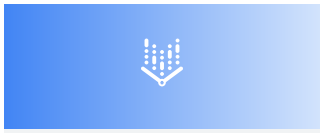
Serverless data solutions are absolutely necessary to allow your organization to move beyond data silos and into the realm of insights and action. All of our core data analytics services are serverless and tightly integrated.

Cloud Pub/Sub is a global message queue that does not require you to manage any infrastructure

Cloud Dataflow is an autoscaling ETL tool that unifies batch and streaming analytics



Google BigQuery provides iinteractive querying across organizational boundaries and can do federated queries on common Big Data formats in Google Cloud Storage



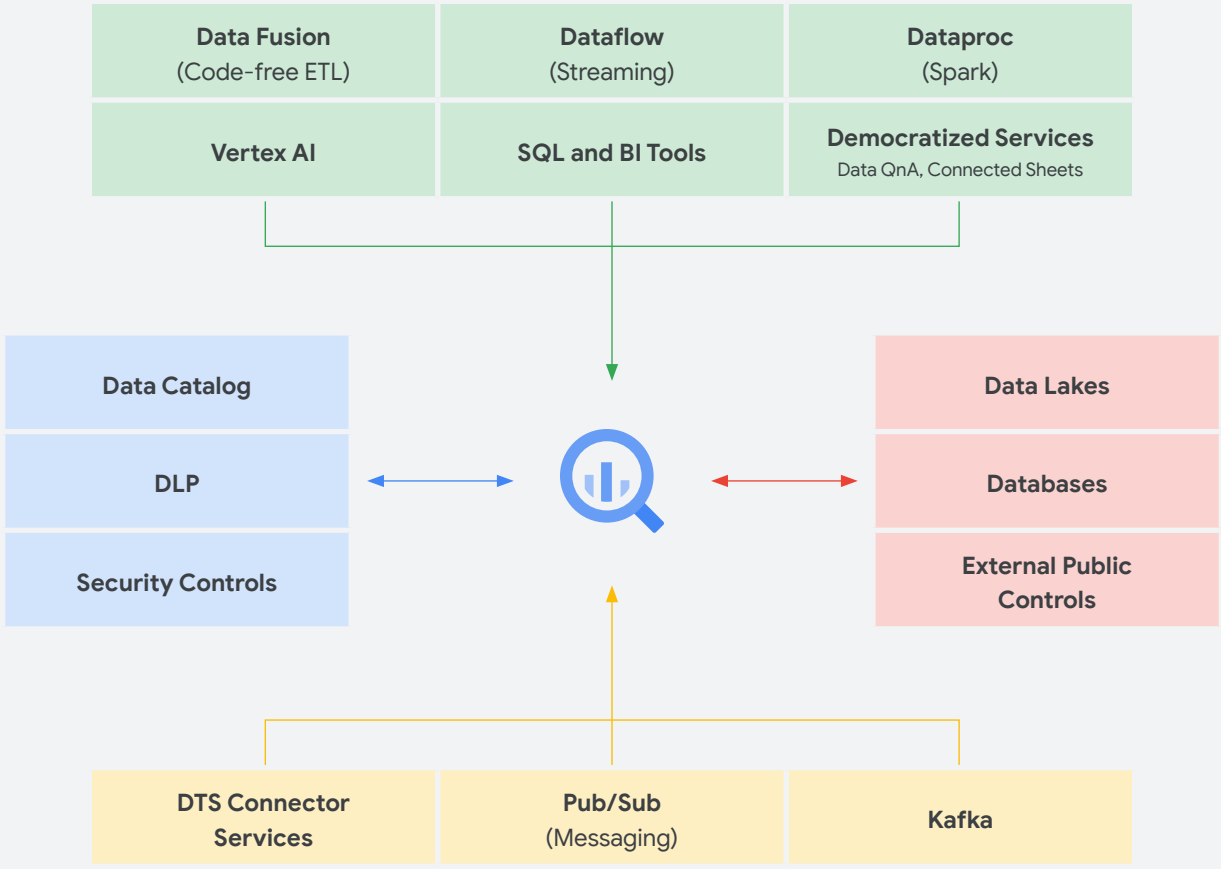
Vertex AI provides state-of-the-art, autoscaling training, deployment, and workflow management

All these services connect transparently to each other due to clear design and clean implementation.

Change management is often one of the hardest aspects of incorporating any new technology into an organization. Google Cloud seeks to meet our customers where they are by providing familiar tools, platforms and integrations for developers and business users alike. Our mission is to accelerate your organization's ability to digitally transform and reimagine your business through data-powered innovation, together. Instead of creating vendor lock-in, Google Cloud provides companies with options for simple, streamlined integrations with on-premise environments, other Cloud offerings and even the Edge to form a truly hybrid Cloud:

- BigQuery Omni removes the need for data to be ported from one environment to another and instead takes the analytics to the data regardless of the environment.
- Apache Beam, the SDK leveraged on Cloud Dataflow, provides transferability and portability to runners like Apache Spark and Apache Flink.
- For organizations looking to run Apache Spark or Apache Hadoop, Google Cloud provides Dataproc.

Most data users care about what data they have, not which system it resides in. Having access to the data they need when they need it is the most important thing. So for the most part, the type of platform does not matter for users, so long as they are able to access fresh, usable data with familiar tools - whether they are exploring datasets, managing sources across data stores, running ad hoc queries or developing internal business intelligence tools for executive stakeholders.



Emerging Trends

Continuing on this idea of the convergence of a data lake and a data warehouse into a unified analytics data platform, there are some additional data solutions that are gaining traction. We have been seeing a lot of concepts emerging around Lake-house and Data Mesh, for example. You may have heard some of these terms before. Some are not new and have been around in different shapes and formats for years. However, they work very nicely within the Google Cloud environment. Let's take a closer look into what a Data Mesh and a Lakehouse would look like in Google Cloud and what they mean for data sharing within an organization.

Lakehouse and Data Mesh are not mutually exclusive, but they help solve different challenges within an organization. But one favors enabling data, while the other enables teams. Data Mesh empowers people to avoid being bottlenecked by one team, and therefore enables the entire data stack. It breaks silos into smaller organizational units in an architecture that provides access to data in a federated manner. Lakehouse brings the data warehouse and data lake together, allowing different types and higher volumes of data. This effectively leads to schema-on-read instead of schema-on-write, a feature of data lakes that was thought to close some of the performance gaps in enterprise data warehouses. As an added benefit, this architecture also borrows more rigorous data governance, something that data lakes typically lack.

Data Lakehouse

- **Removes the overhead of Data Lakes and Data Warehouses**
- **Data warehouse gets the capabilities of a data lake**
- **Data Lake gets the capabilities of the Data Warehouses**
- **Benefits:**
  - Multimodal data access with higher volumes of data
  - Schema on read
  - The governance that Data Lakes lack but DWHs provide
  - Enables unified access to batch and real-time data

Empowering Technology

Data Mesh

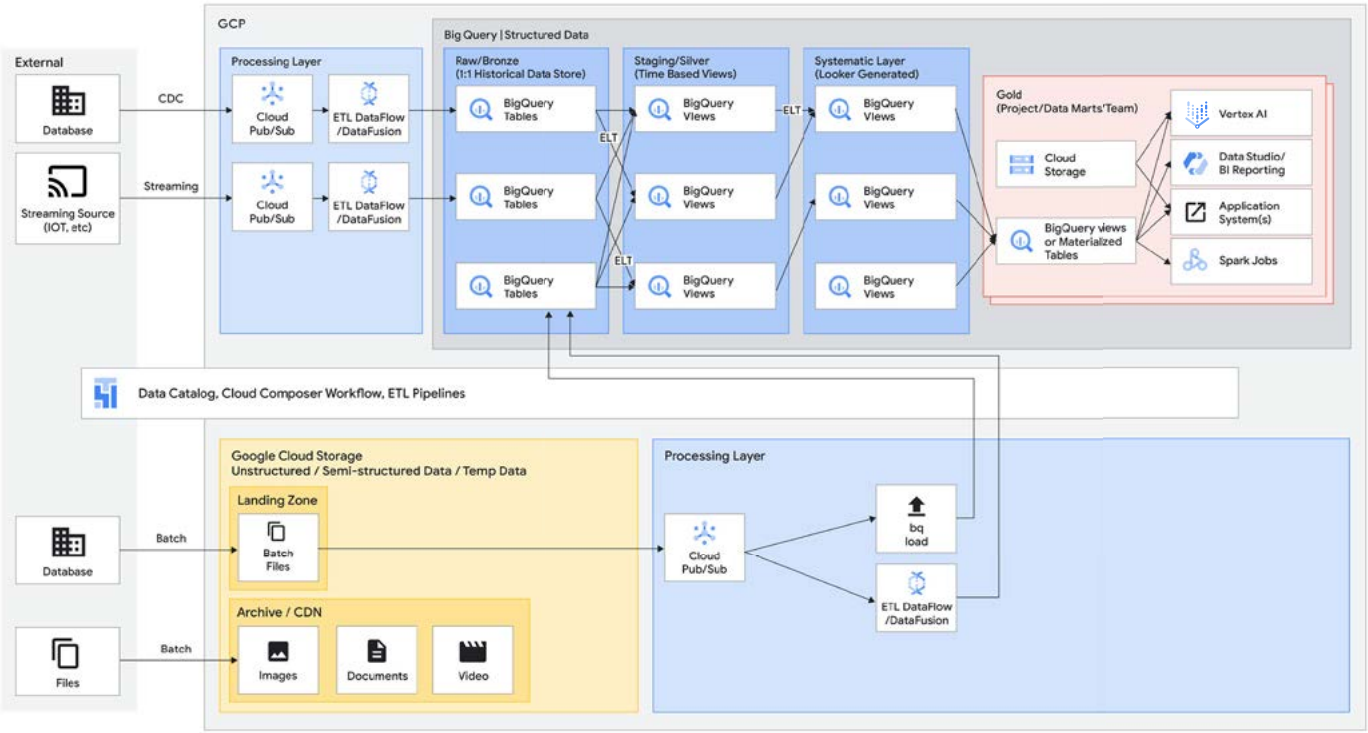
- **Removes the organizational barriers becoming the bottleneck**
  - Federates data ownership
  - Focuses on data as product
  - Allows for the creation of agile teams and shorter time to insights
- **Teams own their data & technology**
  - Provides API / access to other teams
  - Decentralized raw and processed data
- **Benefits:**
  - Well defined, governed and secure
  - Ability to leverage several domains with no data movement
  - Leverages DataOps methodologies (builds on lessons learned in DevOps)

Empowering People

Lakehouse

As mentioned above, BigQuery's Storage API lets you treat your data warehouse like a data lake. Spark jobs running on Dataproc or similar Hadoop environments can use the data stored on BigQuery rather than requiring a separate storage medium by taking storage out of the data warehouse.

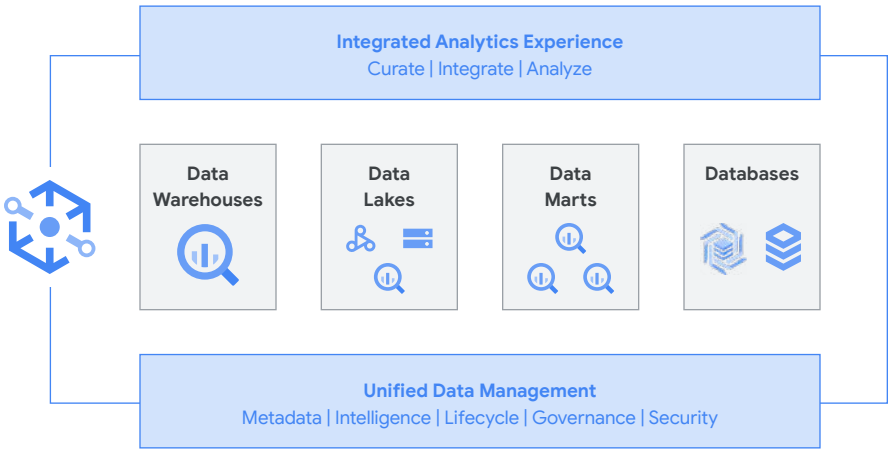
The sheer compute power that is decoupled from storage within BigQuery enables SQL-based transformation and utilizes views across different layers of these transformations. This then leads to an ELT-type approach and enables a more agile data processing platform. Leveraging ELT over ETL, BigQuery enables SQL-based transformations to be stored as logical views. While dumping all of the raw data into data warehouse storage may be expensive with a traditional data warehouse, there is no premium charge for BigQuery storage. Its cost is fairly comparable to blob storage in GCS.



When performing ETL, the transformations are taking place outside of BigQuery, potentially in a tool that does not scale as well. It might end up transforming the data line-by-line rather than parallelizing the queries. There may be instances where Spark or other ETL processes are already codified and changing them for the sake of new technology might not make sense. If, however, there are transformations that can be written in SQL, BigQuery is likely a great place to do them.

In addition, this architecture is supported by all the GCP components like Composer, Data Catalog or Data Fusion. It provides an end-to-end layer for different user personas. Another important aspect of reducing operational overhead can be done by leveraging the capabilities of the underlying infrastructure. Consider Dataflow and BigQuery, all run on containers and let us manage the uptime and the mechanics behind the scenes. Once this is extended to third-party and partner tools, and when they start exploiting similar capabilities such as Kubernetes, then it becomes much simpler to manage and portable. In turn, this reduces resource and operational overheads. Furthermore, this can be complemented by better observability by exploiting monitoring dashboards with Cloud Composer to lead for operational excellence.

Not only can you build a data lake by bringing together data stored in GCS and BigQuery, without any data movement or duplication, but we are offering additional administrative functionality to manage your data sources. Dataplex enables a Lakehouse by offering a centralized management layer to coordinate data in GCS and BigQuery. Doing this enables you to organize your data based on your business needs, so you are no longer restricted by how or where that data is stored.



Dataplex is an intelligent data fabric that enables you to keep your data distributed for the right price/performance while making this data securely accessible to all your analytics tools. It provides metadata-led data management with built-in data quality and governance so you spend less time wrestling with infrastructure boundaries and inefficiencies, trust the data you have and spend more time deriving value out of this data. Additionally, it provides an integrated analytics experience, bringing the best of GCP and open-source together, to enable you to rapidly curate, secure, integrate and analyze their data at scale. Finally, you can build an analytics strategy that augments existing architecture and meets your financial governance goals.

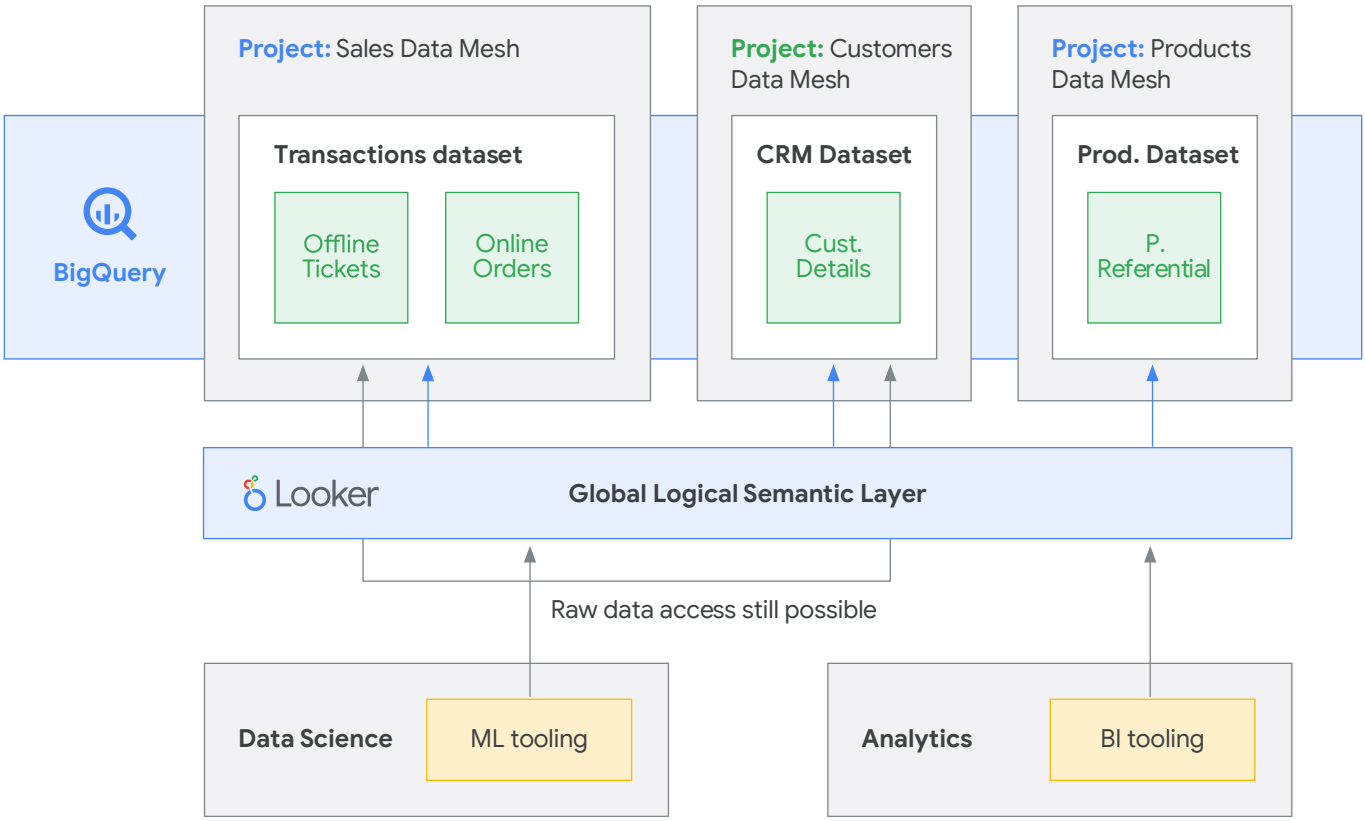
Data Mesh

Data Mesh is built on a long history of innovation from across data warehouses and data lakes, combined with the unparalleled scalability performance pay models, APIs, DevOps and close integration of Google Cloud products.

With this approach, you can effectively create an on-demand data solution. A Data Mesh decentralizes data ownership among domain data owners, each of whom are held accountable for providing their data as a product in a standard way. A Data Mesh also facilitates communication between different parts of the organization to distribute datasets across different locations.

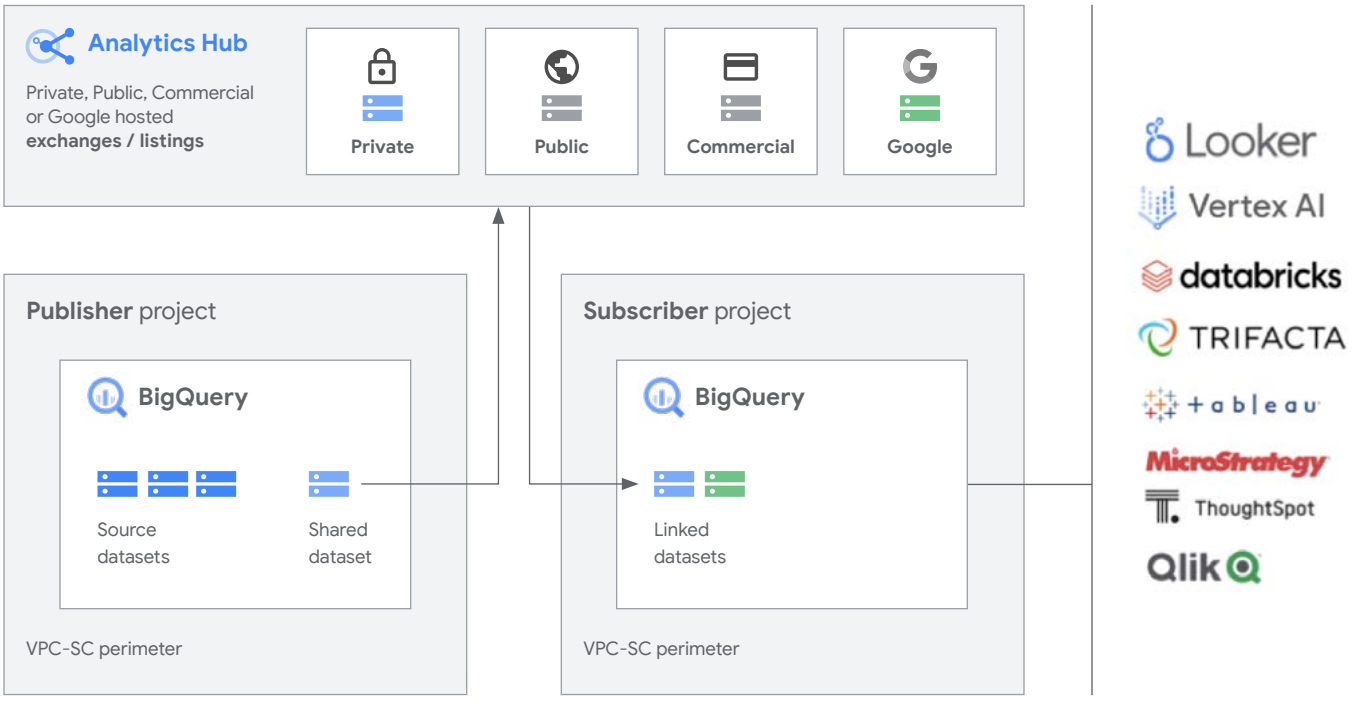
In a Data Mesh, the responsibility for generating value from data is federated to the people who understand it best; in other words, the people who created the data or brought it into the organization must also be responsible for creating consumable data assets as products from the data they create.

In many organizations, establishing a "single source of truth" or "authoritative data source" is challenging due to the repeated extraction and transformation of data across the organization without clear ownership responsibilities over the newly-created data. In the Data Mesh, the authoritative data source is the Data Product published by the source domain, with a clearly assigned Data Owner and Steward who is responsible for that data.



In summary, the Data Mesh promises a domain-oriented, decentralized data ownership and architecture. This is enabled by having federated computation and access layers just like we provide in GCP. Furthermore, if your organization is looking to get more functionality, you can use something like Looker, which can provide a unified layer to model and access the data. Looker's platform offers a single pane UI to access the truest, most up-to-date version of your company's data and business definitions. With this unified view into the business, you can choose or design data experiences that assure people and systems get data delivered to them in a way that makes the most sense for their needs. It fits in perfectly as it allows data scientists, analysts and even business users to access their data with a single semantic model. Data scientists are still accessing the raw data, but without the data movement and duplication.

We're building additional functionality on top of our workhorse products like BigQuery, to make the creation and management of datasets easier. Analytics Hub provides the ability to create private data exchanges, in which exchange administrators (a.k.a. data curators) give permissions to publish and subscribe to data in the exchange to specific individuals or groups both inside the company and externally to business partners or buyers.



Publish, discover and subscribe to shared assets, including open source formats, powered by the scalability of BigQuery. Publishers can view aggregated usage metrics. Data providers can reach enterprise BigQuery customers with data, insights, ML models or visualizations and leverage Cloud marketplace to monetize their apps, insights or models. This is also similar to how BigQuery public datasets are managed through a Google-managed exchange. Drive innovation with access to unique Google datasets, commercial/industry datasets, public datasets or curated data exchanges from your organization or partner ecosystem.

Dealing with the Legacy

While it sounds great to build a brand new data platform from the ground up, we understand that not every company is going to be in a position to do that. Most are dealing with existing legacy systems that they need to migrate, port, or patch until they can be replaced. We have worked with customers at every stage of their data platform journey and we have solutions to meet your situation.

There are typically three categories of migration that we see among customers: Lift and Replatform, Lift and Rehome and full Modernization. For most businesses, we suggest starting with the Lift and Replatform, as it offers a high-impact migration with as little disruption and risk as possible. With this strategy, you migrate your data into BigQuery or Dataproc from your legacy data warehouses and Hadoop clusters. Once data is moved, you can then optimize your data pipelines and queries for performance. With a Lift and Re-platform migration strategy, you can do this in phases, based on the complexity of your workloads. We recommend this approach for large enterprise customers with centralized IT and multiple business units, given their complexity.

The second migration strategy we see most often is a full modernization as the first step. This provides a clean break from the past because you are going full in with a Cloud-native approach. It is built native on GCP, but because you are changing everything in one go, migration can be slower if you have multiple, large legacy environments.

Lift & Rehome

- **Conservative** approach
- Fast migration from existing services such as TD Vantage, Databricks on to GCP
- No modernization or improving existing solutions apart from running them over GCP as a tactical intermediate decision

Lift & Replatform

- Optimal phased approach, low disruption, low risk and high impact
- Migrate data into BQ from legacy EDW
- Migrate data into Dataproc from on-premise Hadoop cluster
- Optiize queries and data pipelines for performance
- Up to **57% lower TCO than on-prem**

Modernize

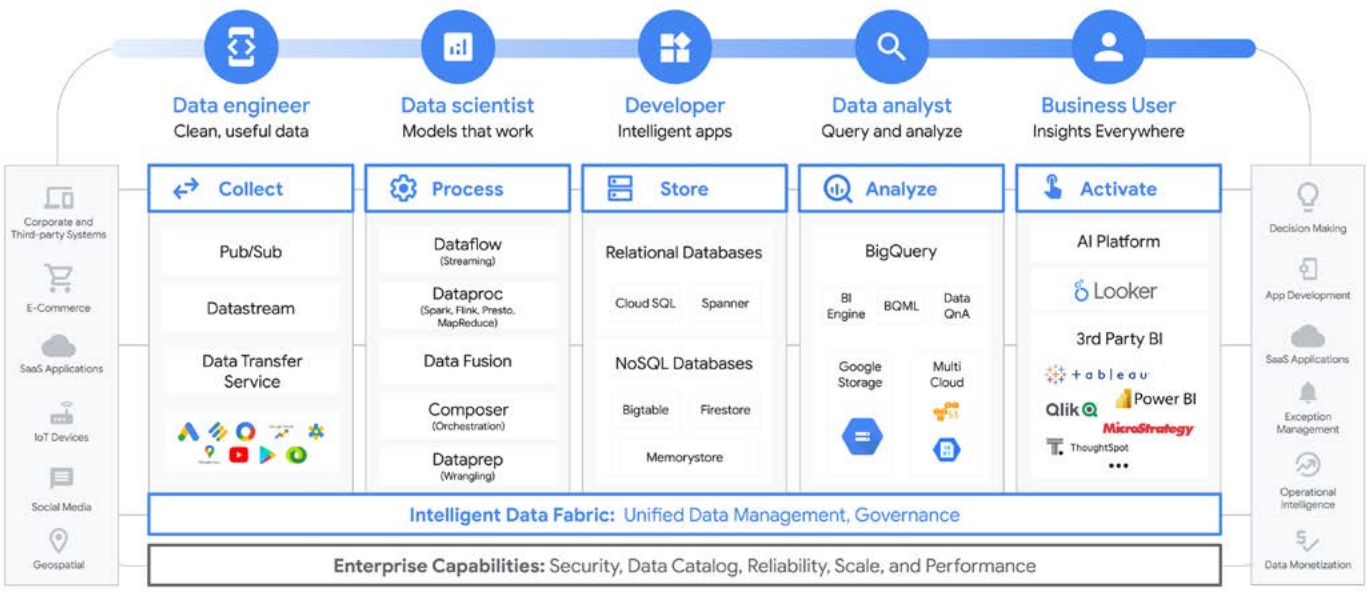
- All in on cloud-native, clean break from the past
- Built natively on GCP
- Can be slower as it requires rewriting jobs
- Greatest development velocity and agility
- 60-88% lower TCO than on-prem, plus value from Google **AI on unstructured data**

A clean legacy break requires rewriting jobs and changing different applications. However, it provides higher velocity and agility as well and the lowest total cost of ownership in the long run compared to the other approaches. This is because of two main reasons: your applications are already optimized and don't need to be retrofitted, and once you migrate your data sources, you don't have to manage two environments at the same time. This approach is best suited for digital natives or engineering-driven organizations with few legacy environments.

Lastly, the most conservative approach is a Lift and Rehome, which we recommend as a short-term tactical solution to move your data estate onto Cloud. You can lift and rehome your existing platforms and carry on using them as before but in the GCP environment. This is applicable for environments such as Teradata and Databricks for example, to reduce the initial risk and allow applications to run. However, this brings the existing siloed environment to the Cloud rather than transforming it, so you won't benefit from the performance of a platform built natively on GCP. However, we can help you with a full migration into Google Cloud native products, so you can take advantage of interoperability and create a fully modern analytics data platform on Google Cloud.

Tactical or strategic?

We think the key differentiators of an analytics data platform built on GCP are that it is open, intelligent, flexible and tightly integrated. There are a lot of solutions in the market that provide tactical solutions that may feel comfortable and familiar. However, these generally provide a short-term solution and just compound organization and technical issues over time.



Google Cloud significantly simplifies data analytics. You can unlock the potential hidden in your data with a cloud-native, serverless approach that decouples storage from compute and lets you analyze gigabytes to petabytes of data in minutes. This allows you to remove the traditional constraints of scale, performance and cost to ask any question of data and solve business problems. As a result, it becomes easier to operationalize insights across the enterprise with a single, trusted data fabric.

What are the benefits?

- Keeps your focus purely on analytics instead of infrastructure
- Solves for every stage of the data analytics lifecycle, from ingestion to transformation and analysis, to business intelligence and more
- Creates a solid data foundation on which to operationalize machine learning
- Enables ability to leverage the best open source technologies for your organization
- Scales to meet the needs of your enterprise, particularly as you increase your use of data in driving your business and through your digital transformation

A modern, unified analytics data platform built on GCP gives you the best capabilities of a data lake and a data warehouse, but with closer integration into the AI platform. You can automatically process real-time data from billions of streaming events and serve insights in up to milliseconds to respond to changing customer needs. Our industry-leading AI services can optimize your organizational decision making and customer experiences, helping you to close the gap between descriptive and prescriptive analytics without having to staff up a new team. You can augment your existing skills to scale the impact of AI with automated, built-in intelligence.

Build a Unified Data Platform with Google

August 2021

Interested in learning more about how the Google data platform can transform the way your business deals with data? Contact us to get started.

Google Cloud