Google Cloud

White paper
September 2021

Building a data lakehouse on Google Cloud Platform

Rachel Levy, Steve Thill and Firat Tekiner

Executive summary

For over a decade, the technology industry has searched for optimal ways to store and analyze vast amounts of data. These storage solutions need to handle the volume, latency, resilience, and varying data access requirements demanded by organizations. To tackle these issues, companies have been making the best out of existing technology stacks. This typically involves trying to either make a data lake behave like an interactive data warehouse or trying to make a data warehouse act like a data lake, processing and storing vast amounts of semi-structured data. Both approaches have resulted in unhappy users, high costs, and data duplication across the enterprise. The Google Cloud data lakehouse pattern solves these shortcomings.

Historically, organizations have implemented siloed and separate architectures. Data warehouses store structured, aggregate data (primarily used for BI and reporting), whereas data lakes store large volumes of unstructured and semi-structured data (primarily used for ML workloads). This approach often results in complex ETL pipelines because of extensive data movement, processing, and duplication. Operationalizing and governing this architecture is challenging, costly, and reduces agility. As organizations are moving to the cloud, they want to break these silos.

To address these issues, a new architecture choice has emerged: the data lakehouse. The data lakehouse combines the key benefits of data lakes and data warehouses. This architecture offers a low-cost storage format that is accessible by various processing engines like Spark while also providing powerful management and optimization features.

Operationalizing and governing this architecture is challenging, costly, and reduces agility. As organizations are moving to the cloud, they want to break these silos.

For example, our data center enables a Dataproc environment to connect to either Google Cloud Storage or the BigQuery storage subsystem and read/write data at storage speeds, thanks to our network that achieves petabit bisectional bandwidth. This allows Spark developers to leverage data inside BigQuery without the need for data duplication and cumbersome ETL operations. The speed of the internal Google network enables your organization to bring the processing to the data and avoid data duplication, reducing data latency, processing time, data discrepancies, and cost. In addition, Dataplex, our intelligent data fabric, enables you to manage your distributed data assets while making data securely accessible to all your analytics tools.

Dataplex provides metadata-led data management with built-in data quality and governance capabilities. With trust in the data, you have, you spend more time deriving value out of data and less time wrestling with infrastructure boundaries and inefficiencies.

Additionally, with the integrated analytics experience provided by Dataplex, you are enabled to rapidly create, secure, integrate, and analyze your data at scale. Finally, you can build an analytics strategy that augments existing architecture and meets your financial governance goals.

The landscape of data continues to evolve and grow at an exponential rate. It is important to have flexible patterns and limitless scale to ensure data is used as an investment, rather than a sunk cost.

With trust in the data you have, you spend more time deriving value out of the data and less time wrestling with infrastructure boundaries and inefficiencies.

Introduction

In the ever-evolving world of data architectures and ecosystems, there is a growing suite of tools being offered to enable data management, governance, scalability, and even machine learning. With promises of digital transformation and evolution, organizations often find themselves with sophisticated solutions that have a considerable amount of bolt-on features. However, the ultimate goal should be to simplify the underlying infrastructure and thus enable teams to focus on bringing value to the business. Data engineers should be able to focus on making the raw data more useful to an organization. Data scientists should be able to focus on looking at the data, using tools to exploit hidden information, and producing predictive data models.

Figure 1: Data ecosystem

Google Cloud has taken this approach by using our planet-scale analytics platform to bring together two foundational solutions. The core capabilities for enterprise data operations, data lakes, and data warehouses have been unified — simplifying the management tasks while increasing the value. The centerpiece of this architectural revolution is BigQuery. As shown in Figure 1, BigQuery is at the center of our customers' Data Ecosystem because it is both tightly integrated with Google Cloud and open to partners' technologies. BigQuery provides the lakehouse architecture, which brings the best of the lake and the warehouse without the overhead of both. This unlocks the value of data and ensures a unified governance approach with tools such as Dataplex and Analytics Hub.

Data warehouses are systems that came about when business leaders were looking to gain analytical insight from operational data stores. The legacy systems that may have worked for the past 40 years have proven to be expensive and cannot often address the challenges around data freshness, scaling, and high costs. Also, data warehouses only fit the needs of tabular data, limiting their usability for a rapidly growing variety of data types and structures. Schema was often applied on-write and was driven by a specific analytic use case. This also limited the flexibility of future use of the data for things such as machine learning and advanced analytics.

To solve some data warehouse limitations, new technologies (such as Hadoop-powered Data Lakes) were developed and ushered in the big data era. For example, data lakes were developed as low-cost storage solutions that essentially amounted to distributed storage of files. They looked great on paper by promising low cost and the ability to scale.

In reality, these promises were not realized for many organizations. This was mainly because they were not easily operationalized, productionized, or utilized. To interact with data in the lake, an end user had to be fairly proficient in particular coding paradigms, which limited the set of people who could use the data. All of these in return increased the total cost of ownership. There were also significant data governance challenges created by the data lakes . They did not work well with the existing identity and access management (IAM) and security models. Furthermore, they ended up creating data silos because data was not easily shared across the Hadoop environment.

During the big data era, these two systems co-existed and complemented each other as the two main database management systems of enterprises, residing side by side. Traditionally, structured and processed data was stored in the data warehouse. On the other hand, data lakes provided the ability to land raw data without having to create a schema. This model created silos between teams. Essentially, data warehouse users were closer to the business and had ideas about how to improve analysis, often without the ability to explore the data to drive a deeper understanding. Data lake users, conversely, were closer to the raw data and had the tools and capabilities to explore the data. However, they spend so much time doing this, they were consequently more focused on the data itself than on the business.

The architecture of a data lakehouse reduces operational costs, simplifies transformation processes, and enhances governance. This model is built on convergence of data lakes and warehouses, as well as data teams across organizations. In essence, it implements warehouse-like data structures and data management functions on low-cost storage that is typical of data lakes.

BigQuery as a data lakehouse

In many ways, a data lakehouse becomes a way of centralizing and unifying disparate data sources and engineering efforts across an organization. This means, like any type of data infrastructure, a data lakehouse must include components to ingest, store, and analyze data. A data lakehouse, however, compared to either a data lake or data warehouse, attempts to centralize these components to provide a more unified dataset for all users, regardless of skill set. The ethos of the data lakehouse is about believing that all users can and should be data users, regardless of technical capabilities. By providing an underlying and centralized effort to make the data accessible, different tools can sit on top of the lakehouse that meets each user's capabilities.

To make it possible for all users to have access to the same underlying data, the data lakehouse takes advantage of BigQuery's storage and the compute power to use views rather than materialized tables. This is important because a data lakehouse has the same storage subsystem, enabling shared storage behind the views to minimize unnecessary data replication. This is all done in BigQuery, without the standard storage premium often associated with traditional data warehouses. The permanent location for raw, enriched, and business data then becomes the lakehouse.

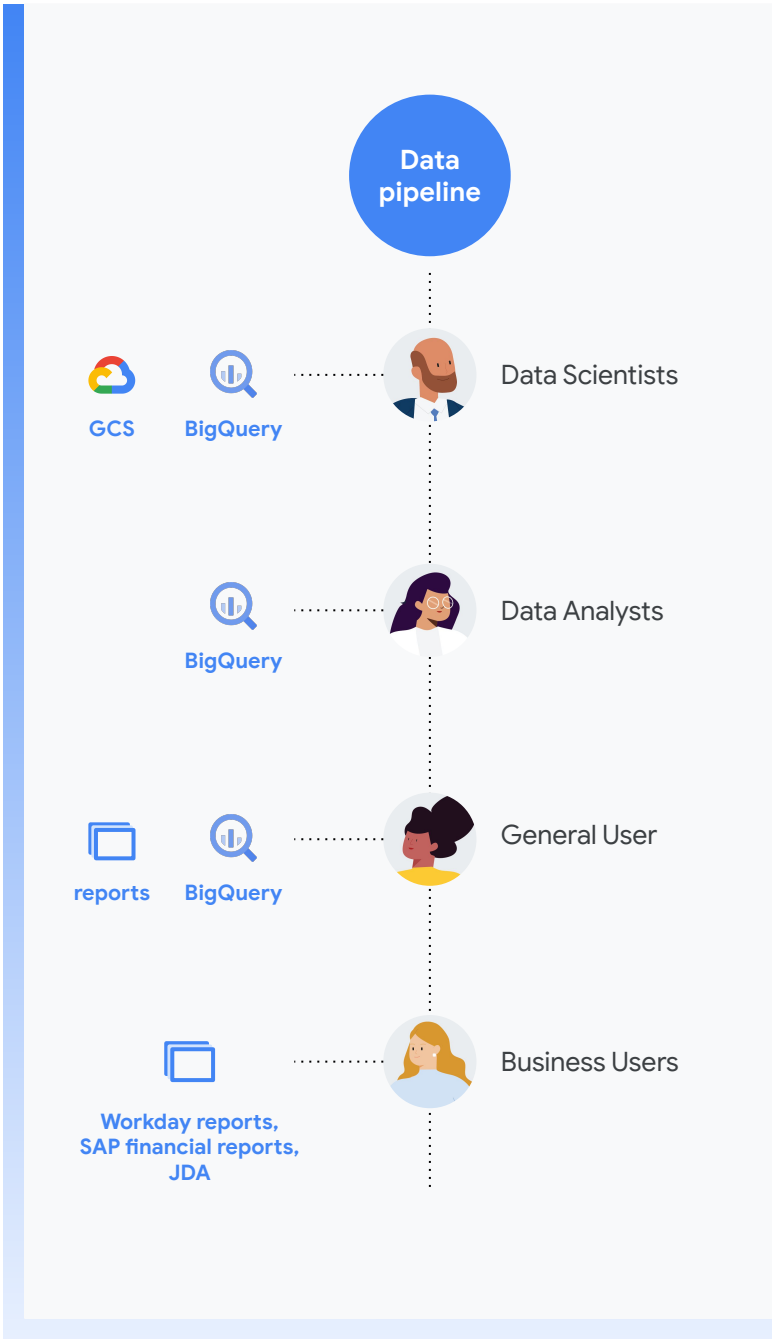

Figure 2: Accessing data for end users

A modern data warehouse like BigQuery can handle massive data volumes and has cost parity with other data storage mechanisms such as Cloud Storage. This reduces operational costs, simplifies transformation processes, and enhances governance. Furthermore, a data warehouse is then used as the data fabric for all the datasets (that are kept and governed in it). In many organizations adopting data lakehouses, the centralized analytics or IT team ingests the data from source systems and provides a standardized set of views that various teams can then leverage for their own use cases. An example of these views is in Figure 3.

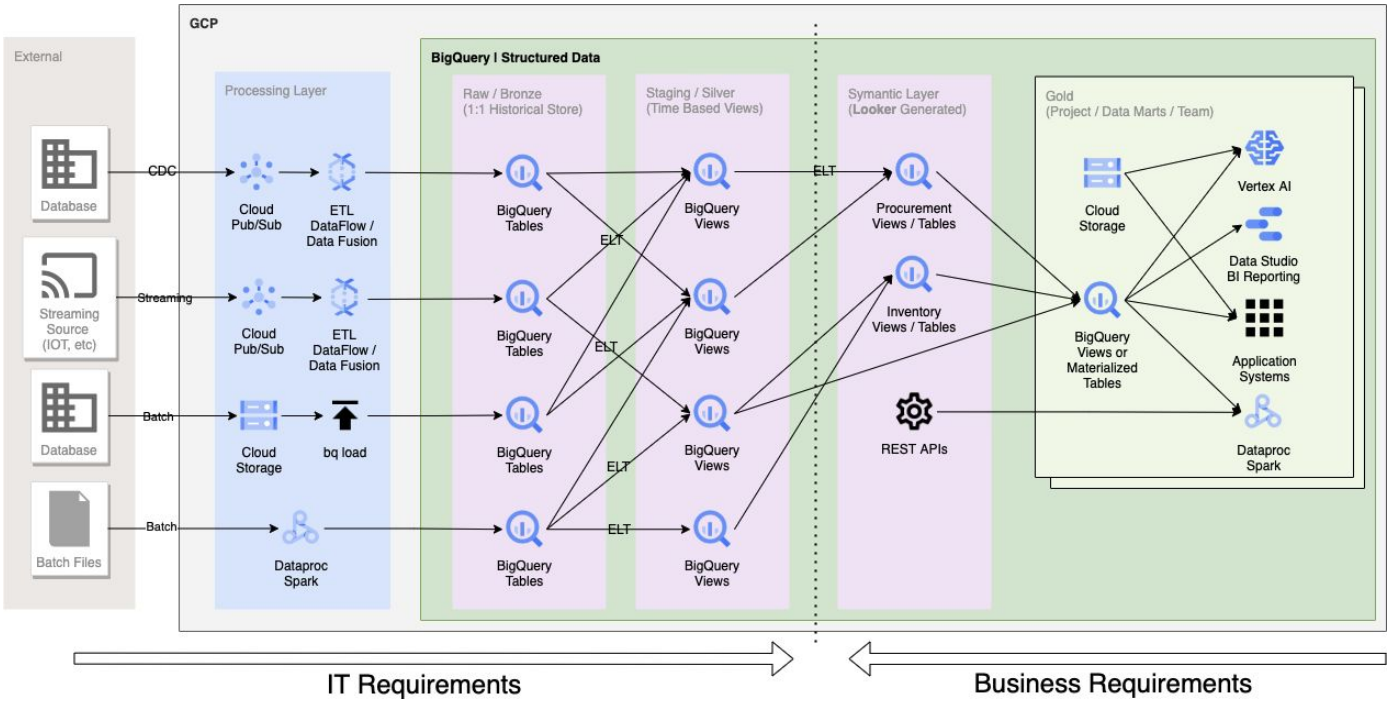

Figure 3: Data architecture

In this example, IT teams can ingest data into the bronze layer and utilize views to cleanse the data through the staging/silver layer (see Figure 3), while the specific use of the data can then be made into curated views in their own projects. A deep dive into each of the architectural components of a data lakehouse on Google Cloud is found in the next sections.

Traditionally, with data warehousing, compute and storage were scarce resources that teams competed for. When those resources were no longer available, it often led to fragmentation of resources (data marts). These arbitrary resource constraints tied the ability to unlock the data's value to the capacity of the hardware, rather than the capacity of the imagination. Data lakehouses, specifically on Google Cloud, remove the artificial constraints to unlock data's value by providing a nearly limitless and instantaneous scalability due, in large part, to the separation of compute and storage.

BigQuery's separation of storage and compute allows for BigQuery compute to be brought to other storage mechanisms through federated queries and have other compute paradigms by using data stored in native BigQuery format through the Storage API. BigQuery has a Storage API that allows the storage (which is separated from the compute clusters) to be treated like structured data in a lake. Rather than reading in parquet or Avro files, Dataproc, Google Cloud's managed Hadoop, can read the data directly from BigQuery storage, run its computations, and write it back to BigQuery. An example of this is seen in Figure 4.

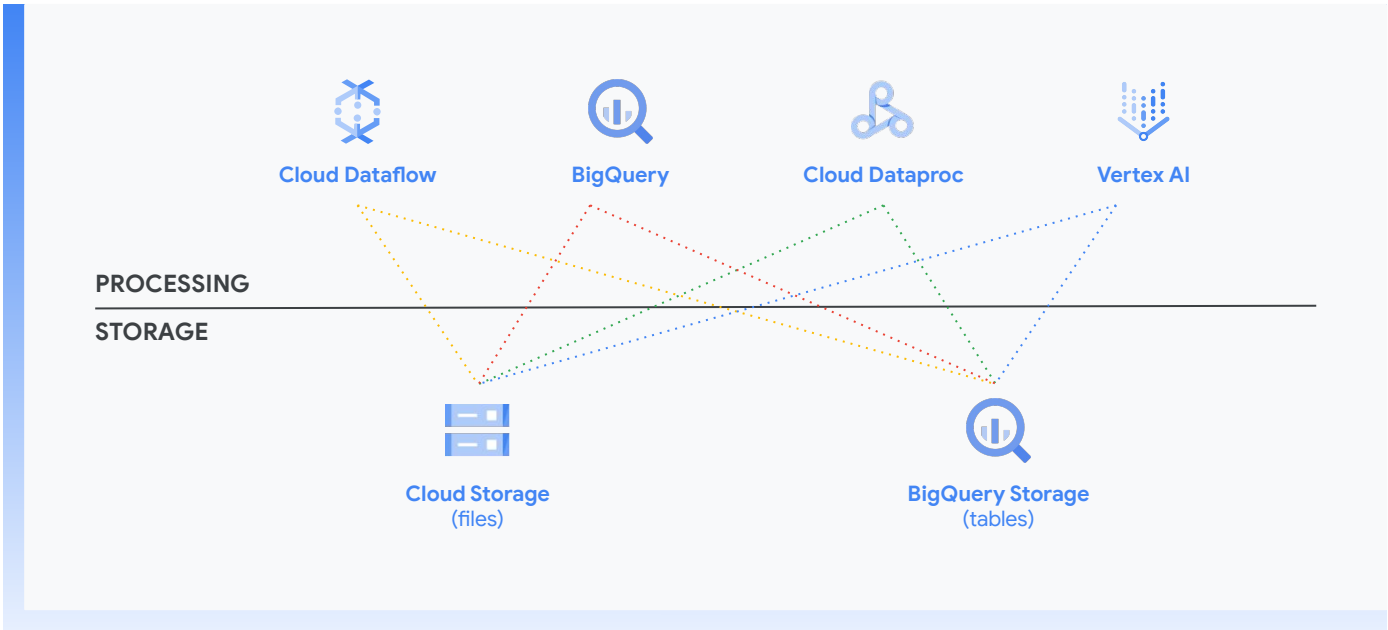

Figure 4: Data processing and storage

Separation of compute and storage is key to managing resources in the cloud. It enables resource sharing across applications with reduced overhead. Furthermore, it enables setting budgets at various levels and stages. It is possible to define caps for different workloads meeting SLAs. For example, flexible slots in BigQuery simultaneously provide SLA guarantees and elasticity. Resource guarantees can be done at minute intervals, monthly intervals, or yearly intervals at BigQuery level. On the other hand, Ephemeral Dataproc clusters let you bring up complex Hadoop (including Spark) clusters within a matter of seconds, running the required workload and shutting it down. A combination of these means you can manage overruns and handle unexpected hikes without requiring considerable capital investment.

Another important aspect of reducing operational overhead is using the capabilities of underlying infrastructure. Consider Dataflow and BigQuery; while they have different underlying infrastructure, Google manages the uptime and mechanics behind the scenes of both services. For example, backups, snapshots, and resiliency are available all out of the box. In turn, this reduces resource and operational overheads. You can also experience better observability by exploiting monitoring dashboards with Cloud Monitoring to lead for operational excellence.

This is how the lines that have traditionally been drawn between data lakes and data warehouses can start to blur. External storage of files in Google Cloud Storage (GCS) can be accessed through BigQuery federated queries and Bigtable's data. The data lakehouse flips the paradigm of infrastructure procurement around and makes data access a permissioning problem and data processing an OpEx/budgeting problem. In traditional IT organizations, a data warehouse and surrounding data marts would be additional hardware purchases for different lines of business or use cases, as depicted in Figure 5.

Instead, creating a single-source-of-truth storage layer accessible by a highly scalable serverless compute cluster enables data access to be controlled by access control listed, as granular as the organization requires. This is one of the key principles in making a data lakehouse work for all users in an enterprise (Figure 6).

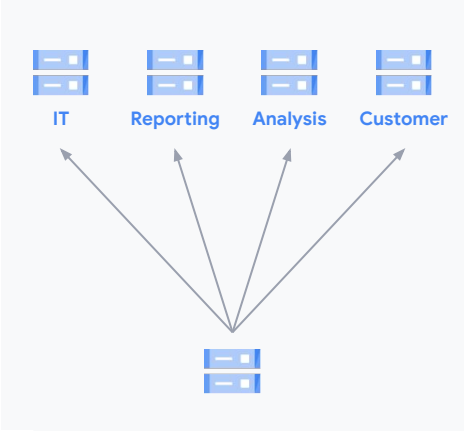

Figure 5: Data warehouse and data marts

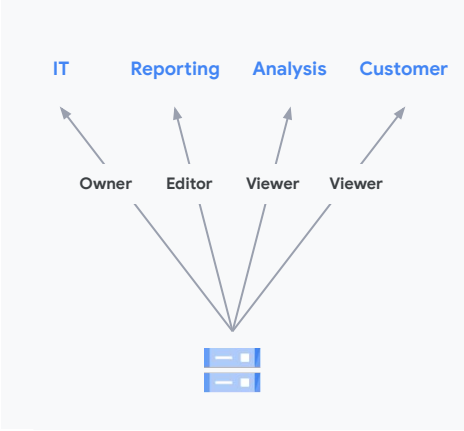

Figure 6: Data access control

Ingesting data into the data lakehouse

Ingesting data into the lakehouse is the first part of the data pipeline. There are several ways to ingest data into BigQuery, but they generally fall into two categories: batch data or streaming data. Depending on the source system and data requirements, data engineering teams may decide how to get the data into the lakehouse.

With a traditional data lake and data warehouse dichotomy, the data files may be dropped into the lake in a semi-structured state. Then a subset of that data would be transformed and structured to be stored in the warehouse where analysts could use the data. With the data lakehouse, however, there is no need to store the data in those separate systems.

Traditionally, applications are loaded in batches, whereby data can be batch loaded directly into a data lake or data warehouse. Batch ingestion is powerful in a system like BigQuery because it uses massive compute clusters that parallelize the ingestion.

This means that even terabytes of data can be ingested into BigQuery in less than a few minutes.

On the other hand, many newer use cases enabled by new technologies allow for data to be streamed directly into the data lakehouse and analyzed in real time. Real-time data may originate from IoT devices or transactional systems that stream data in real time. For example, when streaming data into a lakehouse such as BigQuery, it is best to use an append-only model. This means that historical data will always be in the table, but the query can include a where clause that ensures only the latest version of each record is included in each view. Remember, storage is cheap and organizations can afford to keep historical data that can then be used for additional use cases.

There are several managed services on Google Cloud that help ingest data into the lakehouse. Pub/Sub is a messaging service that runs on our global network and enables asynchronous communication on the order of 100s of milliseconds. The messages are streamed from Pub/Sub to Dataflow for real-time analytics, while Dataproc is used for batched file ingestion.

Once the data lands in BigQuery, it is queryable. Following the layered approach that leverages views over materialized tables, the most up-to-date data end up in every "gold layer" project (see Figure 3) utilizing the data lakehouse architecture. Instead of streaming all of this data into a lake and then structuring it for the warehouse, you have the best of both in BigQuery.

Views can only be used to present the most recent record of the data, but there are also performance benefits to this architecture. It enables the capability to restore data from a point in time if there is a corruption event. In addition, having date-based partitioning provides a performance enhancement and a cost-saving as well. BigQuery charges for the on-demand model based on how much data is read, not the time it takes the compute clusters to process data. This means that you can use partitioning to only query a certain timeframe, minimizing the data scanned and thus the cost incurred.

Storing data in the data lakehouse

One of the key benefits of the data lakehouse is minimizing the copies of data made and stored across an enterprise. Many enterprises suffer from different personas copying a subset of the data, making transformations, and then using that data to make decisions. Without sharing this data across the enterprise, data will lack consistency and become less trustworthy. Furthermore, it becomes a data governance nightmare; imagine a "right to be forgotten" GDPR request coming in and tracking all copies of the data to remove that individual.

To mitigate this, the data lakehouse stores the data in a single-source-of-truth layer, making minimal copies of the data. Data is ingested into BigQuery and then stored in the "raw" layer. Leveraging ELT over ETL, BigQuery enables SQL-based transformations to be stored as logical views. While dumping raw data into data warehouse storage may have been expensive in traditional data warehouses, there is no premium charge on BigQuery storage. Its cost is comparable to blob storage in Cloud Storage. Without a cost premium, there is no longer a required cost-based justification for storage.

When performing ETL, the transformations take 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, so changing them for the sake of new technology might not make sense. If, however, any transformations can be written in SQL, BigQuery is likely a great place to send them. Utilizing some BigQuery tuning best practices like partitioning and clustering is one way to keep costs and performance optimal. It could also be a good opportunity to materialize a table instead of storing the logic as a view.

Furthermore, not all data is structured or tabular enough to be stored in a database. This is increasingly common when the proliferation of unstructured data sources (such as image files or text) is going to be analyzed by an enterprise. Non-tabular or unstructured data does not reside in BigQuery. To use those types of unstructured data, you might need a machine learning platform. It is important, though, to ensure that this is part of the unified data lakehouse and that it can work with the following architecture. This enables the metadata to be stored with the rest of the metadata so that it can still be searchable as part of the underlying data platform.

This type of architecture is the complete data lakehouse with unstructured data. If there is no unstructured data, the top half of Figure 7 still works as previously discussed and uses the same principles to be an enterprise data lakehouse.

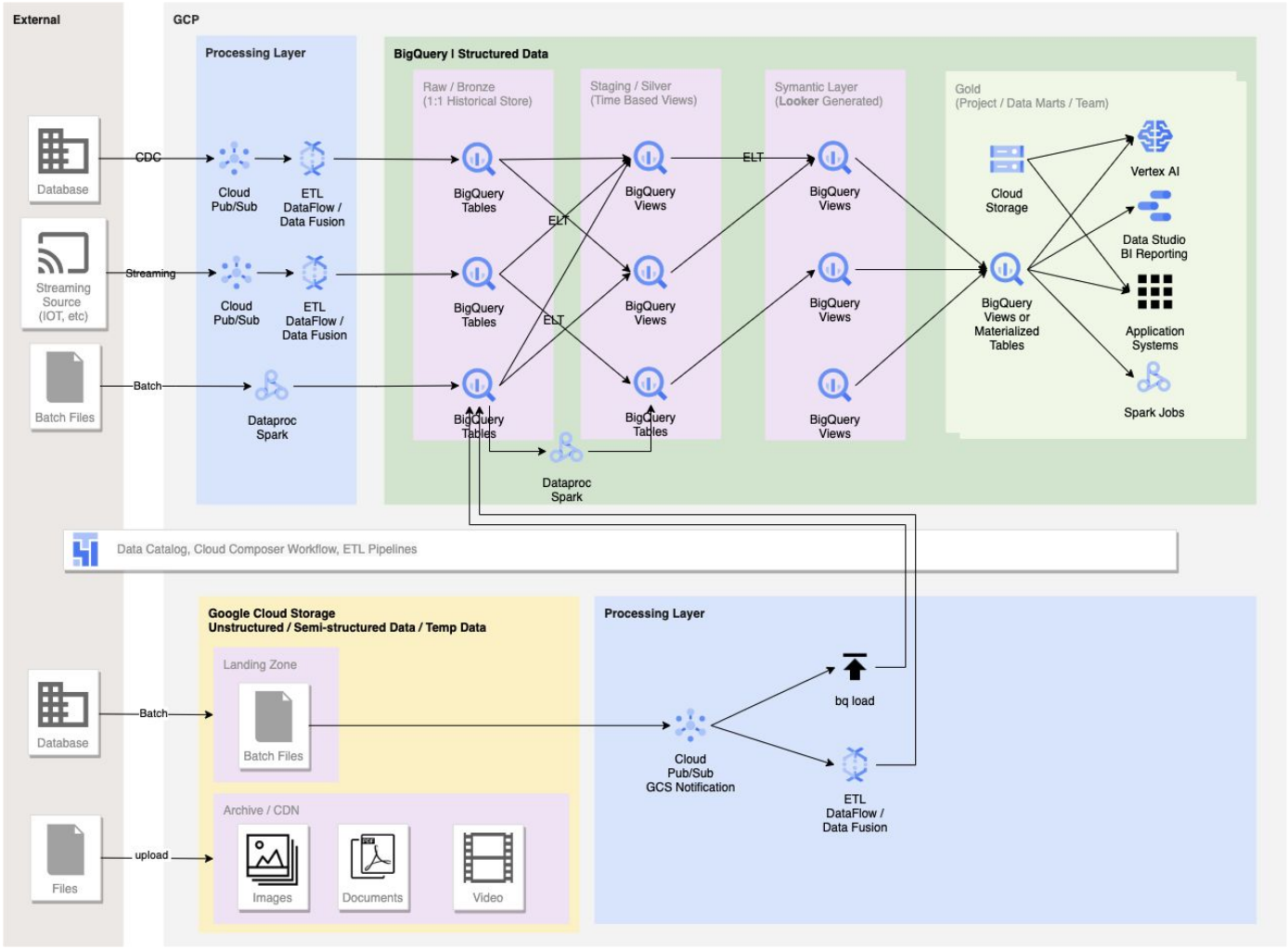

Figure 7: Data lakehouse design pattern

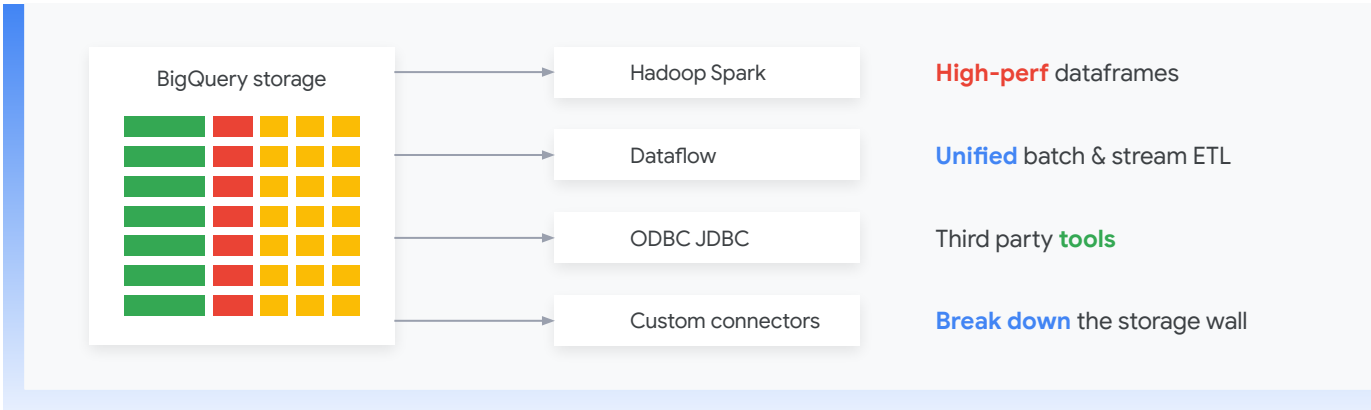

Figure 8: BigQuery Storage API

Analyzing data in the data lakehouse

Once the data is ingested and stored in the data lakehouse, it must be analyzed and activated to drive business value. If the data is not accessible to the right resources, it is not even paying for the storage costs it incurs. To activate the data, an analyst or data scientist must find insight that drives action. Traditional reporting with data in a warehouse is to look back at historical data over the past week, month, quarter, etc. While there is value in understanding these trends in the business, it is also important to use analytics to look forward so that real-time actions can be taken to correct issues before or once they arrive.

There are a few ways to use the data that is stored in BigQuery, and the access method should be based on an end user's skill set. Meeting users at their level of data access including SQL, Python, or more GUI-based methods mean that technological skills do not limit their ability to use data for any job. Data scientists may be working outside traditional SQL-based or BI types of tools. Because BigQuery has the storage API, tools such as Spark running on Dataproc or AI notebooks can easily be integrated into the workflow. The paradigm shift here is that the data lakehouse architecture supports bringing the compute to the data rather than moving the data around. In addition to the BigQuery SQL engine, the following diagram demonstrates other computation frameworks.

The data lakehouse architecture makes it easy to share data with granular access controls across enterprises and with other/partner companies. For example, role-based access methods across a suite of products make it possible to apply the same rules to data in its transformation journey, ensuring data governance and reduced operational cost. Therefore, Spark code using the BigQuery Storage API as well as users using spreadsheets rather than writing SQL would still be leveraging the data lakehouse as their data source. This would allow increased collaboration across the organization and enable the democratization of data.

When data is organized and democratized with a business-driven approach, data can be leveraged as a shareable and monetizable asset within an organization or with partner organizations. To formalize this capability, Google offers a layer on top of BigQuery called Analytics Hub, that can create private data exchanges. Exchange administrators (a.k.a. data curators) give permission to publish and subscribe data in the exchange to specific individuals or groups both internally and externally to business partners or buyers, as depicted in Figure 9.

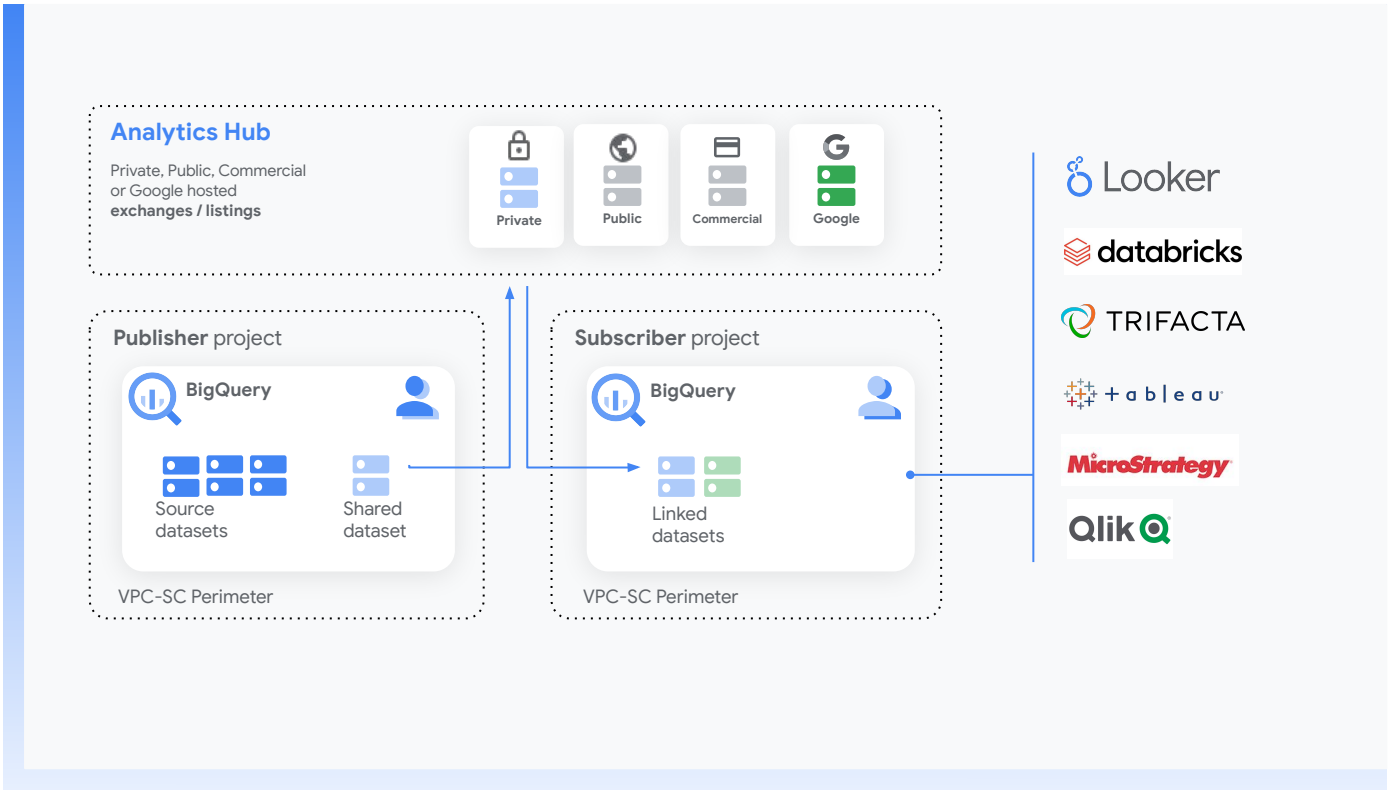

Figure 9: Analytics Hub

You can publish, discover and subscribe to shared assets that are powered by the scalability of BigQuery, including open source formats. 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 similar to how BigQuery public datasets are managed through a Google-managed exchange. You can drive innovation with access to unique Google datasets, commercial/industry datasets, public datasets, or curated data exchanges from your organization or partner ecosystem. These capabilities can be driven when data operations are optimized to provide more valuable opportunities to the organization, rather than spending time feeding and caring for individual, and potentially redundant, systems.

Figure 10: Analytics Hub for secure & scalable sharing

Making it seamless

Dataplex provides a managed lakehouse service that enables enterprises to rapidly build federated lakehouses. Curate, catalog, secure, integrate, and analyze any type of data at any scale with an integrated experience. Dataplex enables quickly built lakes without needing to acquire and worry about different resources. It expands automatic data discovery and schema inference across different systems. And it complements this by automatically registering metadata as tables and filesets as metastores and Data Catalog. Furthermore, with our integrated Cloud Data Loss Prevention API (DLP) and built-in data quality checks, the tagging of sensitive data is tightly integrated.

Dataplex also enables easier management of the lakehouse with simpler security controls. Consistent security policy and enforcement across Cloud Storage and BigQuery is enabled out of the box. Also enabled are managed Data Lake Storage with fine-grained access control, ACID transactions on files, and a BigQuery single pane of glass.

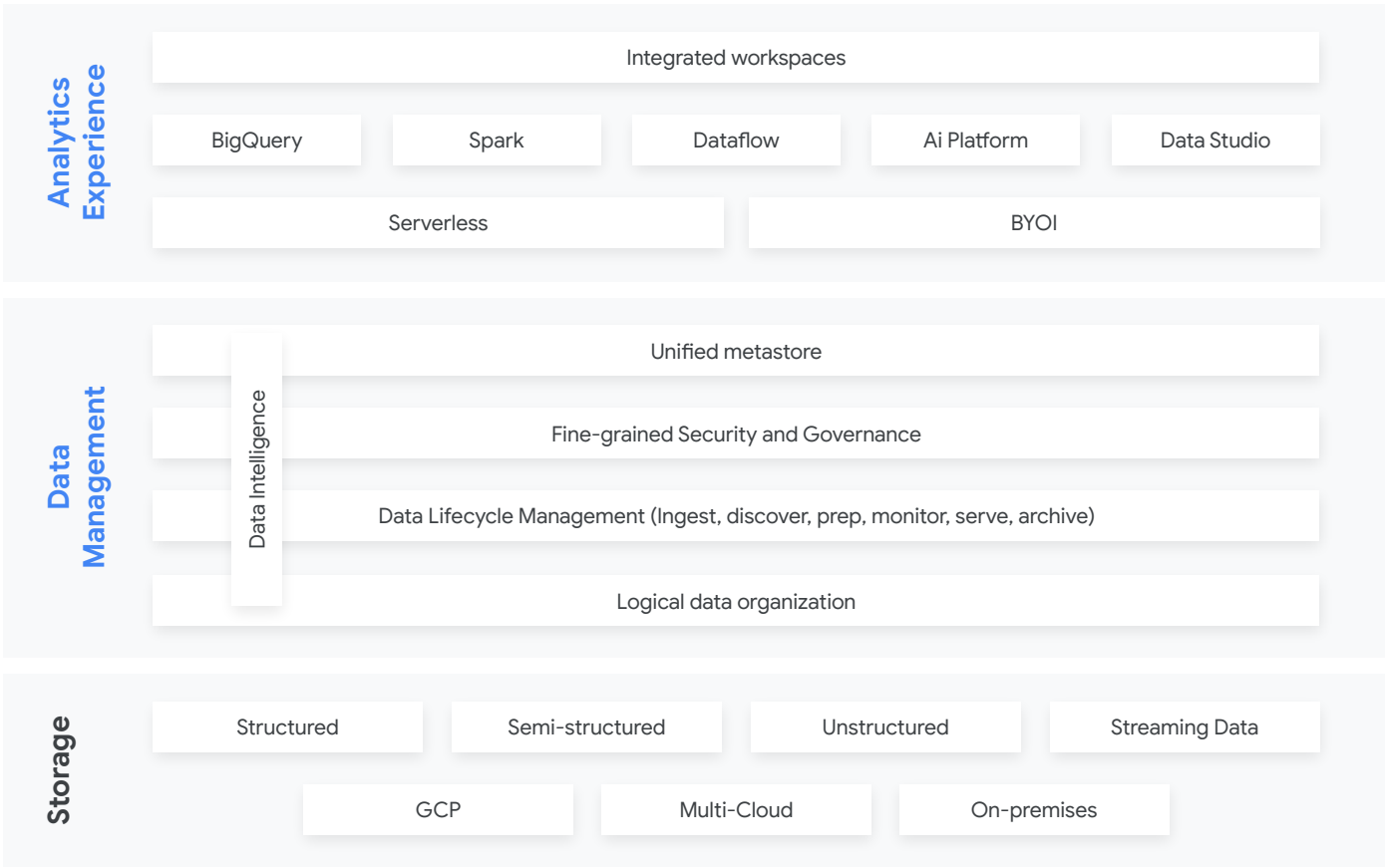

Figure 11: Dataplex

With Dataplex, an integrated analytics workspace is becoming a reality. There is no infrastructure to manage and it provides one-click access to insights for different personas. This means that data administrators are able to set up and manage workspaces together with appropriate environment profiles, including compute parameters, libraries, etc. At the same time, they are able to control user access and manage costs through one seamless interface. Data scientists have one-click access to notebooks. Further, they can discover notebooks by using a notebook repository with links to associated data while being able to save and share notebooks as if they were sharing another asset within the organization. Data analysts are able to use SQL Workspace for ad-hoc analysis without being dependent on any data processing environment. Effectively, through a single pane of glass they will be able to use Presto, Hive, or BigQuery without needing to access various environments.

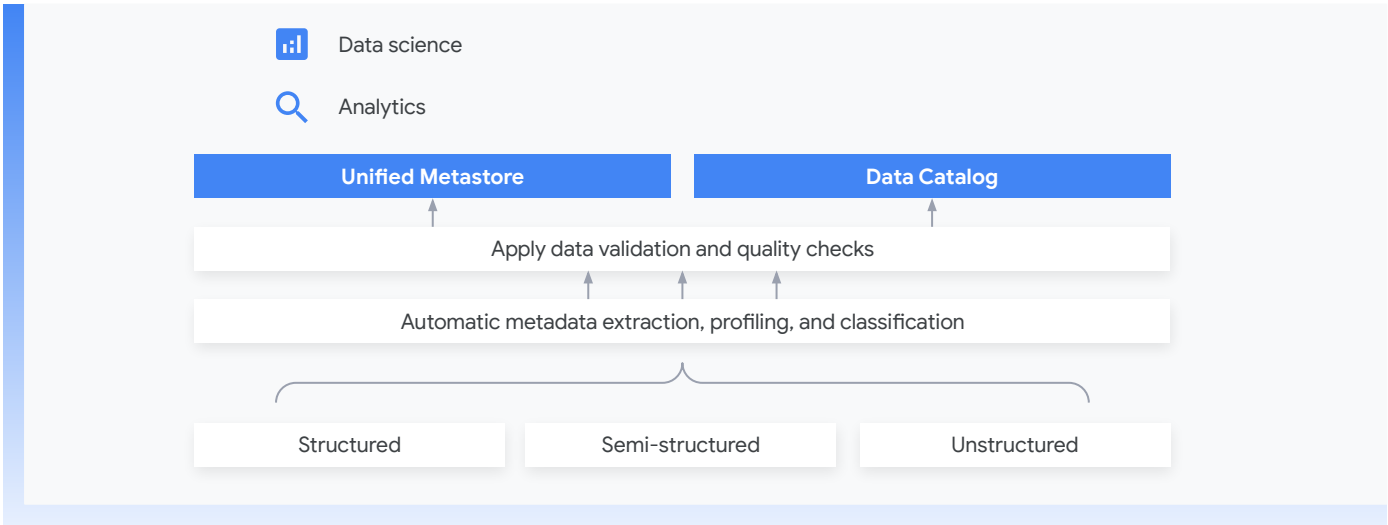

Figure 12: Datalake with Dataplex

Last, but not least, data access is made simple and straightforward. An integrated experience across all Google Cloud Data Analytics services provides virtual lakehouse experiences. This is complemented with an integrated serverless notebook experience with serverless Spark for data science. All Cloud Storage data is automatically made queryable through OSS tools and BigQuery, while enabling search and discovery across the board by using Data Catalog.

Conclusion

We are in a transformative era for data analytics in the Cloud. As data volumes increase and companies become more data-driven, they need to break down data silos and make data more accessible by numerous users across the business. We have seen an increase in the number of unified data platform architecture options that meet the needs of different organization types. Google Cloud's suite of data analytics products is well suited for any modern analytics data platform pattern, including a data lakehouse.

Google is a data company, with a world-class suite of analytics products. But our secret sauce lies in our planet-scale, intelligent infrastructure upon which our products are built. Not only can you develop a lakehouse that meets your data users' needs, but we have you covered in unique hardware and networking, integration that enables streaming at unlimited scale, a serverless data warehouse with an unmatched 99.99% uptime SLA, and flexible and intelligent compute that takes the guesswork out of provisioning servers.

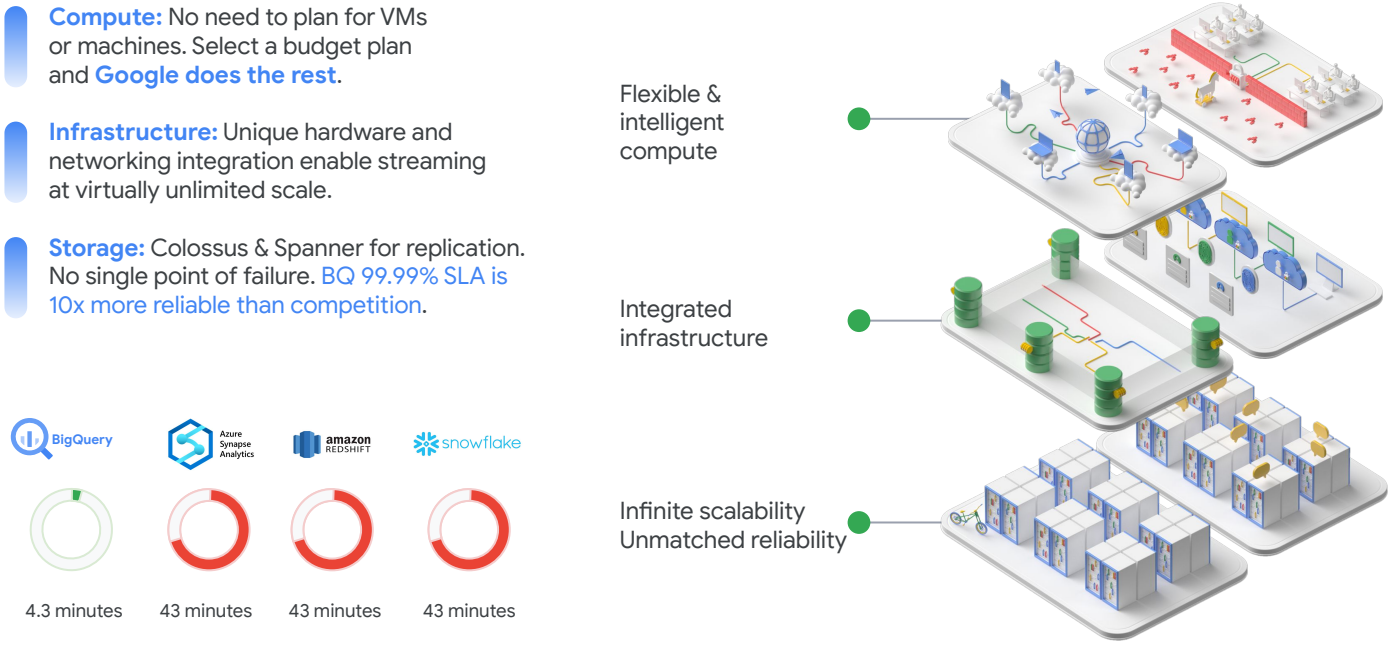

Figure 13: Our secret sauce

Building a data lakehouse on Google Cloud Platform

September 2021

Interested in getting started? Contact us to learn more.

Google Cloud