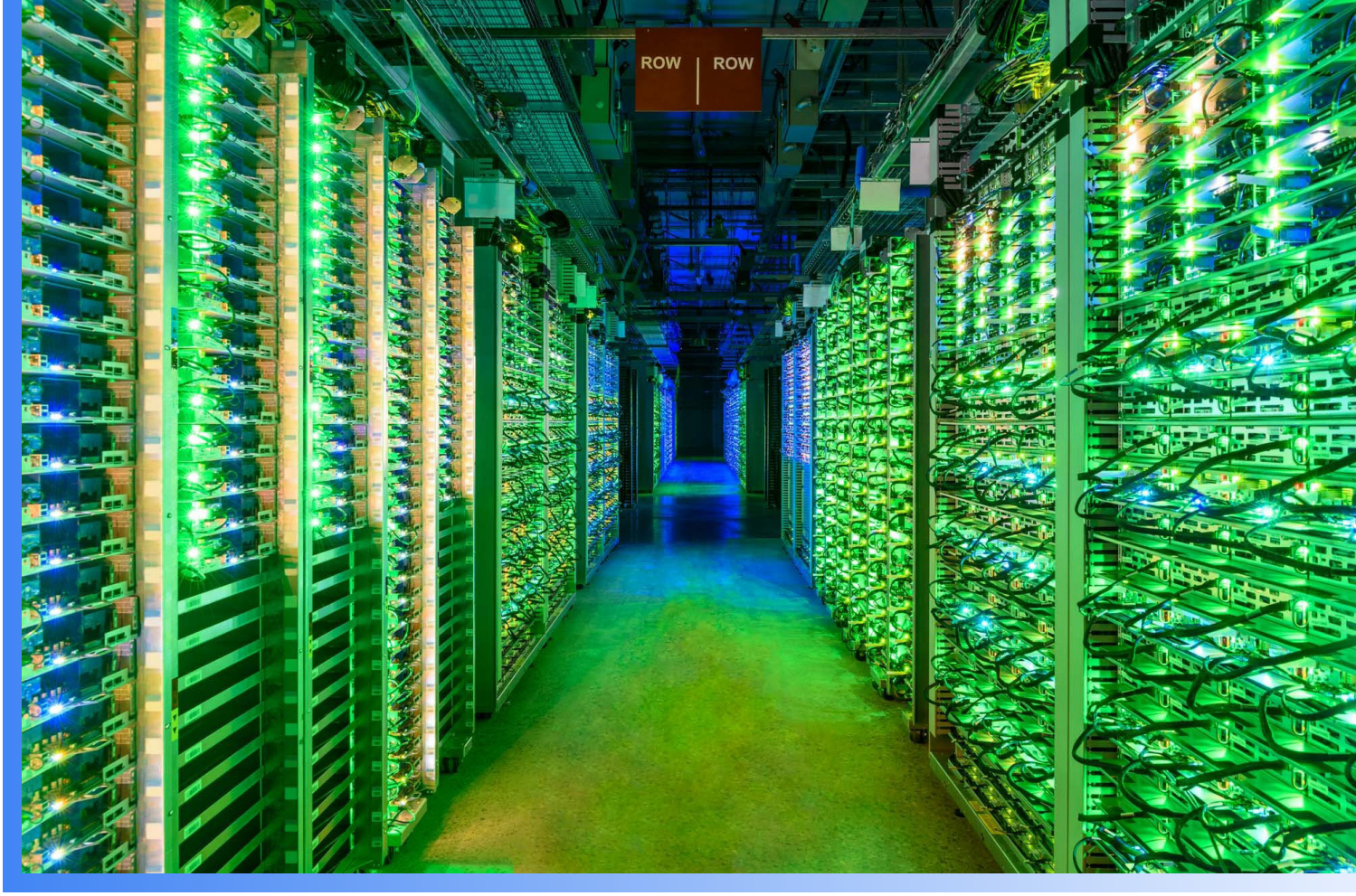


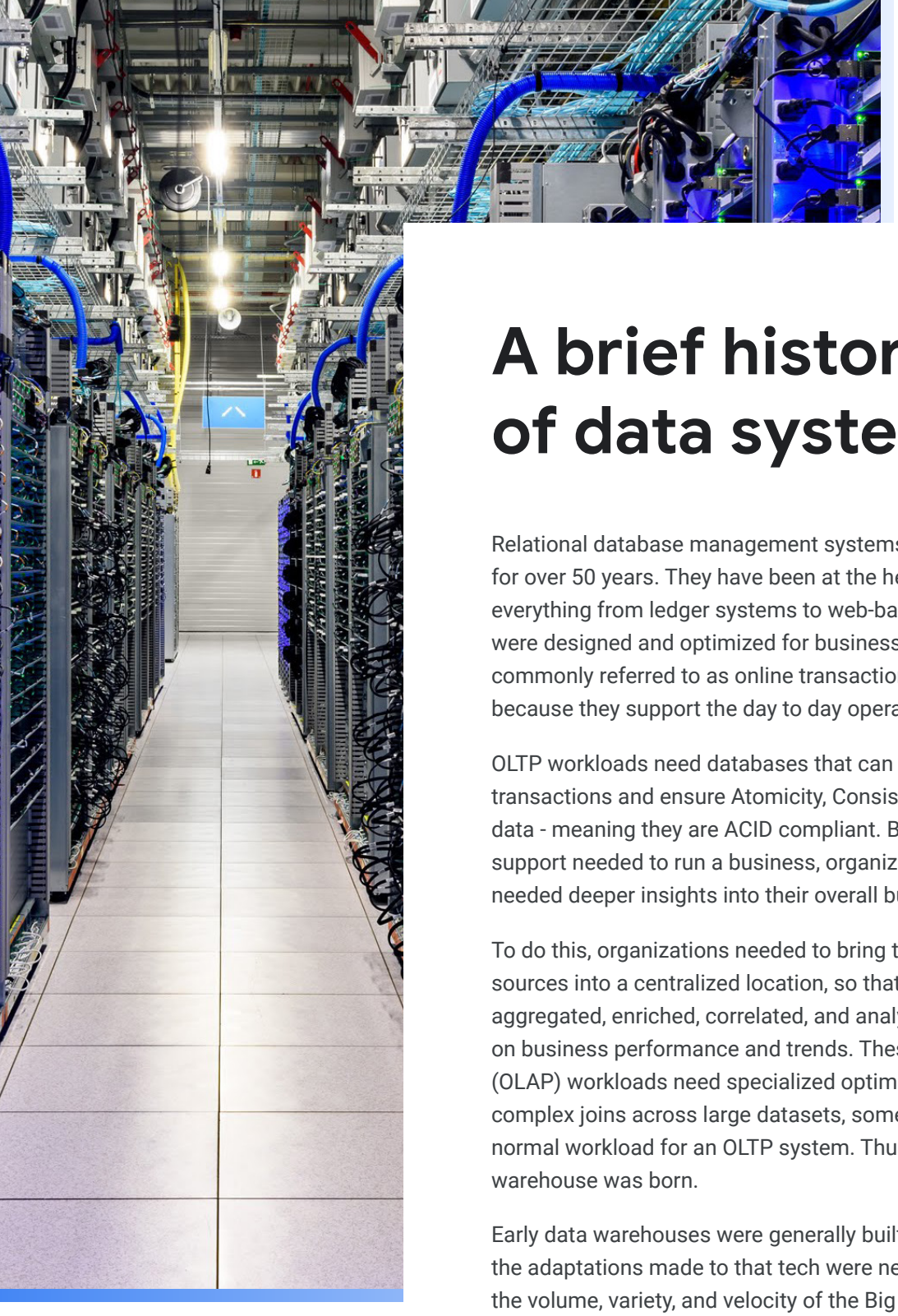
Converging Architectures: Bringing Data Lakes and Data Warehouses Together

June 2021

Firat Tekiner, Rachel Levy and Susan Pierce

Google Cloud

Companies have always been data driven, but over the last 50 years we have seen a striking change in the way companies generate, collect, and use data. By looking at the history of data processing technology and the progression of requirements over time, we can evaluate future looking data platforms. In this paper we will cover advancements in data processing and provide an opinionated perspective for CIOs and CTOs who are building data driven organizations.



A brief history of data systems

Relational database management systems (RDBMS) have been around for over 50 years. They have been at the heart of enterprises, supporting everything from ledger systems to web-based applications. These systems were designed and optimized for business data processing and are commonly referred to as online transactional processing (OLTP) systems because they support the day to day operations of a company.

OLTP workloads need databases that can handle a high volume of transactions and ensure Atomicity, Consistency, Isolation, and Durability of data - meaning they are ACID compliant. But in addition to the transactional support needed to run a business, organizations over time realized that they needed deeper insights into their overall business performance.

To do this, organizations needed to bring together data from multiple sources into a centralized location, so that operational data could be aggregated, enriched, correlated, and analyzed to produce deep insights on business performance and trends. These online analytical processing (OLAP) workloads need specialized optimization of data stores to handle complex joins across large datasets, something that is outside the normal workload for an OLTP system. Thus the idea of a centralized data warehouse was born.

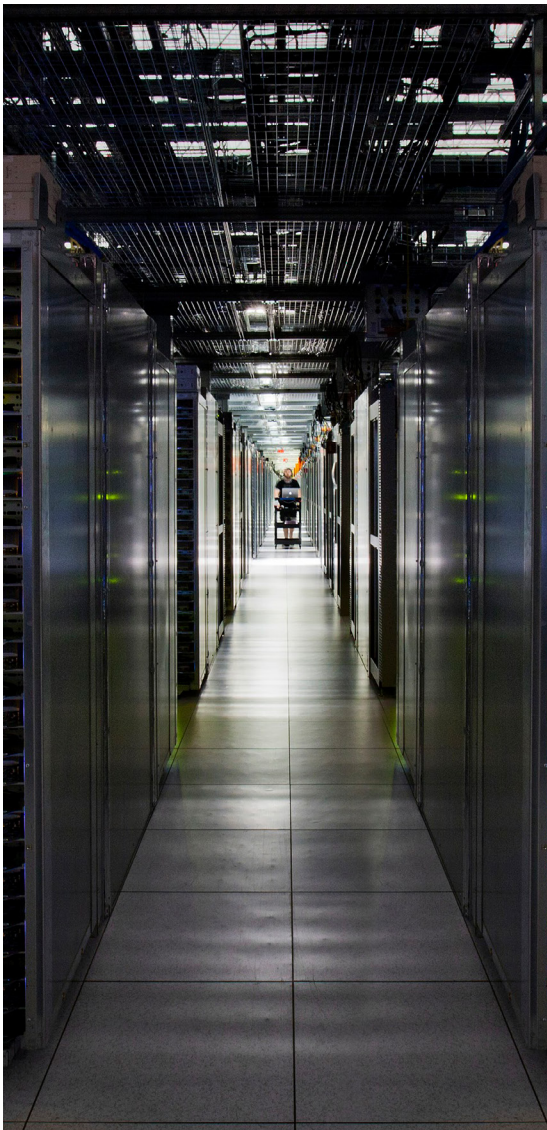
Early data warehouses were generally built on existing RDBMS stacks, and the adaptations made to that tech were never really sufficient to support the volume, variety, and velocity of the Big Data era. As more companies embraced the Internet and digital transformation, data volumes and types also increased dramatically. Up until the mid- to late 1990s, most of the data being generated by and for companies was structured or semi-structured in nature. With the rise of social media, sharing platforms, and IoT devices, the types of data available became more varied. Data warehouses could only handle structured and semi-structured data, and were not the answer for the growing unstructured data ingested from the new sources. A new method of collecting, storing, and exploring these combined data types was needed.

The Big Data explosion of the 2000s changed the rules of the game with a host of new distributed database systems and data engines, mainly from the NoSQL and columnar families. They marked the end of the “one size fits all” paradigm that fueled data warehouses and business intelligence until then.

This gave rise to a new concept called a data lake, which soon became a core pillar of data management alongside the data warehouse. A data lake is a place for enterprises to ingest, store, explore, process, and analyze any type or volume of raw data coming from disparate sources like operational systems, web sources, social media and Internet of Things (IoT). To make the best use of a data lake, data is stored in its original format without the added structure or much pre-processing. In order to facilitate unflawed exploration by analysts and data scientists for some use cases it is necessary to insert data into a data warehouse in its raw format. This would prevent unintentionally contaminating the data by business logic and adding bias into it whilst enriching it.

Hadoop introduced distributed data processing (MapReduce) and the Hadoop Distributed File Service (HDFS) (inspired by Google’s GFS) on cheap commodity servers. Every application running on top of Hadoop was designed to tolerate node failures, thus making it a cost effective alternative for some of the traditional data warehouse workloads.

Hadoop-based data lakes seemed to hit all of the 3 Vs (volume, variety, and velocity) of Big Data and compensated for many of the shortcomings of data warehouses as discussed above. Data could be ingested and stored at high volumes (up to petabyte or even exabyte scale) at a relatively cheap price. Unlike data warehouses that only supported structured and semi-structured data, the ingested data could also be unstructured (images, videos, documents). These characteristics also influenced the velocity of decision-making and led to democratization of data with more personas in the organization, giving data scientists and data engineers more immediate access to data. Organizations could also make further enhancements by layering security, metadata, and governance across the platform in a centralized way.



Whether used as the source system for data warehouses, as data processing and transformation layer, as a platform for experimentation for data scientists and analysts, or as a direct source for self-service BI — it is clear that data warehouses and data lakes complement each other and the main transactional data stores in an organization. Over time, organizations were led to believe that Hadoop in combination with their data warehouse would solve all their analytics needs, but that was not necessarily the case.

1. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation#
2. https://research.google/pubs/pub62/
3. https://research.google/pubs/pub51/

Data lakes began expanding their capabilities beyond storage of raw data to include advanced analytics and data science on large volumes of data. This enabled self-service analytics across the organization, but it required an extensive working knowledge of advanced Hadoop and engineering processes in order to access the data. The Hadoop OSS ecosystem grew in terms of data systems and processing frameworks (Hbase, Hive, Spark, Pig, Presto, TensorFlow, and more) in parallel to the exponential growth in organizations' data, but this led to additional complexity and cost of maintenance.

Meanwhile, data volumes and types are continuing to grow, and conventional data analytics platforms are failing to keep up. According to the 2017 HBR research report[4], companies use <1% of unstructured and less than 50% of structured data for any decision-making, missing out on critical business insights. The cost and complexity of provisioning, maintaining, and scaling data lake clusters has kept organizations from using them to their full potential.

Now, businesses are looking to modernize the data lake and data warehouses by moving them to the cloud because of cost savings and the need to realize value from data by making it available for real-time business insights and artificial intelligence (AI). As more companies optimize to become fully data-driven, AI and real time analytics are in higher demand.

Cloud, is an opportunity to explore the way data warehouses and data lakes have changed and why these two platforms are converging with each other and with the other pillars of smart analytics.

Cloud to the rescue

Facing the shortcomings of traditional data warehouses and data lakes on-premises, data stakeholders struggle with the challenge of scaling infrastructure, finding critical talent, improving costs, and ultimately managing the growing expectation to deliver valuable insights. Furthermore, as enterprises are increasingly becoming data-driven, data warehouses and data lakes play a critical role in an organization's digital transformation journey. Today, companies need a 360 degree real time view of their business to gain a competitive edge.

In order to stay competitive, companies need a data platform that enables data-driven decision making across the enterprise. But this requires more than technical changes; organizations need to embrace a culture of data sharing. Siloed data is silenced data. To broaden and unify enterprise intelligence, securely sharing data across lines of business leads is critical.

When users are no longer constrained by the capacity of their infrastructure, data nirvana is reached when the value-driven data products are only limited by an enterprise's imagination. Utilizing the cloud supports organizations in their modernization efforts because it minimizes the toil and friction by offloading the administrative, low-value tasks.

4. https://hbr.org/2017/05/whats-your-data-strategy

By migrating to the cloud and modernizing these traditional management systems, organizations can:

- Reduce their storage and data processing costs[5].
- Scale to ingest, store, and process and analyze all the relevant data both from internal sources and from external and public ones.
- Increase time to value by enabling real time and predictive analytics.
- Embrace a data culture across the organization and enjoy the best of breed analytics and machine learning (ML).
- Leverage simple and powerful data security and governance across layers.
- Democratize data, which needs to be easily discovered and accessible to the right stakeholder inside and outside of the enterprise in a secure manner. Cloud enables accessibility and offers tools so that skill sets do not define the limitation of a business user embedding data into their daily work. This may look like simplified reporting tools, cloud-back spreadsheet interfaces, and drag-and-drop analytic tools.

5. https://www.esg-global.com/hubfs/pdf/Google-Big-Query-ESG-Economic-Value-Audit-Whitepaper-May-2017.pdf

Data warehouse and data lake convergence

As mentioned previously, some of the key differences between a data lake and a data warehouse relate to the type of data that can be ingested and the ability to land unprocessed (raw) data into a common location. This can happen without the governance, metadata, and data quality that would have been applied in traditional data warehouses.

These core differences explain the changes around the personas using the two platforms:

- Traditional data warehouse users are BI analysts who are closer to the business, focusing on driving insights from data. Data is traditionally prepared by the ETL tools based on the requirements of the data analysts. These users are traditionally using the data to answer questions.
- Data lake users (in addition to analysts), include data engineers and data scientists. They are closer to the raw data with the tools and capabilities to explore and mine the data. They not only transform the data to business data that can be transferred to the data warehouses but also experiment with it and use it to train their ML models and for AI processing. These users not only find answers in the data, but they also find questions.

As a result, we often see these two systems are traditionally managed by different IT departments with different teams. They are split between their use of the data warehouse and the data lake. However, this approach has a number of tradeoffs for customers and traditional workloads. This disconnect has an opportunity cost; organizations spend their resources on operational aspects rather than focusing on business insights. As such, they cannot allocate resources to focus on the key business drivers or on challenges that would allow them to gain a competitive edge.

Additionally, maintaining two separate systems with the same end goal of providing actionable insights from data can cause data quality and consistency problems. By not aligning on the storage and transformations of the data, there may end up being two different values for what is ostensibly one record. With the extra effort required to transform data that have standardized values, such as timestamp, many data users are less compelled to return to the data lake every time they need to use data. These can lead to data puddles across the enterprise, which are datasets stored on individual's machines, causing both a security risk and inefficient use of data.

For example, if an online retailer spends all their resources on managing a traditional data warehouse to provide daily reporting which is key to the business, then they fall behind on creating business value from the data such as leveraging AI for predictive intelligence and automated actions. Hence, they lose competitive advantage as they have increased costs, lower revenues and higher risk. Effectively, it is a barrier to gain a competitive edge. The alternative is to use fully managed cloud environments whereby most of the operational challenges are resolved by the services provided.

Cloud computing introduced new deployment methods of large scale DBMS where the storage is not co-located with the compute servers. Managing both the storage and compute in distributed and elastic clusters with resiliency and security in place still requires administrative overhead to ensure that capacity is in place for the converged system. Cloud's global scale infrastructure and native managed services provide an environment that optimizes the convergence of the data lake and data warehouse, providing the benefits of having both the lake and the warehouse without the overhead of both.

Cloud's nearly limitless scalability is what enables the convergence of data warehouses and lakes. Having data centers full of servers that allocate storage and compute differently enables distributed applications for interactive queries. The amount of data and the compute required to analyze it can scale dynamically from warm pools of compute clusters. When storage is decoupled from the compute, it can be used for many different use cases. The same storage that was once file-based data lakes for structured data, can now be the same storage and data for the data warehouse. This key convergence enables data to be stored once, utilizing views to prepare the data for each specific use case.

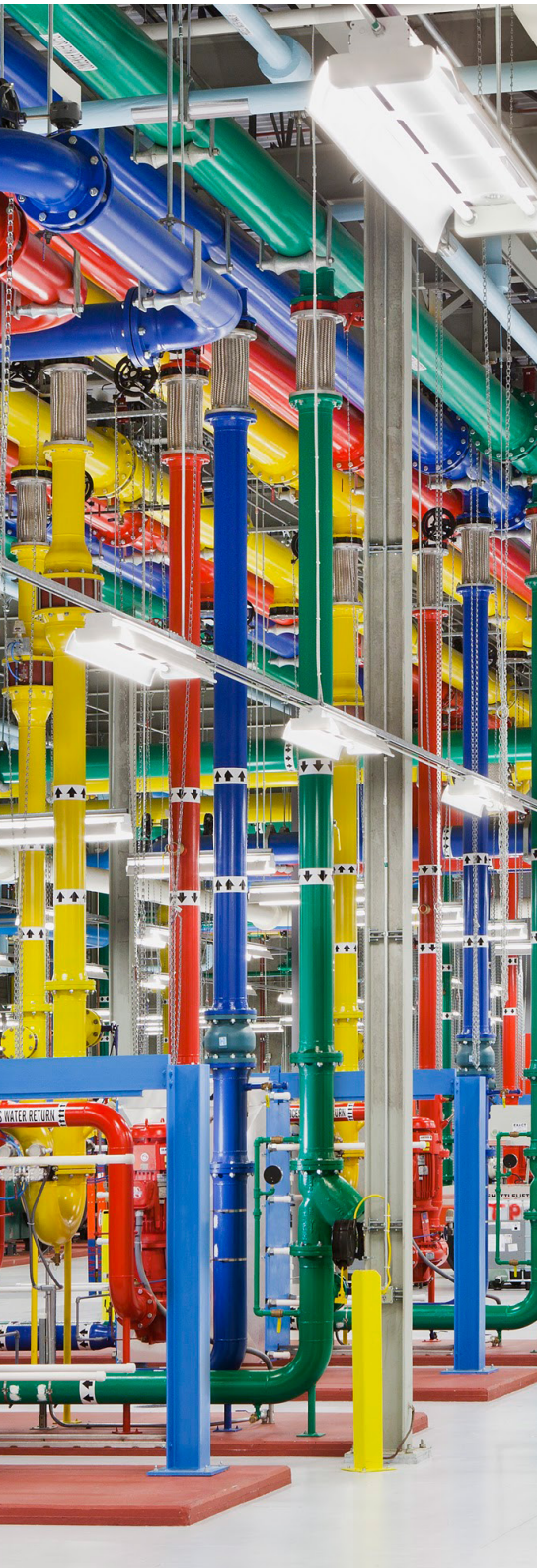
An example of this would be utilizing the same underlying storage for a data warehouse that serves BI reporting for the storage that a Spark cluster uses. This enables Spark code that data lake teams spent years perfecting to take advantage of the more performant storage that is often used as part of a distributed computing system. It allows the compute to move to the data, rather than the data to have to shuffle. This unlocks better speed and performance without requiring high-end infrastructure. Many clouds offer this as a managed service, further abstracting the required management of the infrastructure, much like converging the storage of these two systems.

Our customers face common challenges and trade offs when they try to build a single monolithic platform:

- **IT Challenge:** Data sits across multiple storage types and systems -- data warehouses, data lakes, data marts that may be located on-premise, in a single cloud, or across multiple cloud providers. Customers are forced to either distribute their data governance and replicate the overhead of security, metadata management, data lineage, etc across systems, or copy large amounts of “important” or “sensitive” data into one large system that is more tightly controlled than the rest.
- **Analytics Challenge:** Analytics tools cannot always access the right data and related artifacts. Organizations usually find themselves having to choose between giving their analytics team free reign or limiting data access, which can in turn hamper analytic agility.
- **Business Challenge**: Data trustworthiness is a big issue. Business users want to have more data ownership, which would give them more trust in the data, but freer access to data can potentially lower its quality. Organizations need to decide whether to optimize for more access with a potential lower data quality, or to tightly control access in an attempt to maintain high data quality.

These challenges create unintended tension among teams. Every organization wants a platform that provides secure, high quality data, that is accessible to the right data users. What if they don’t have to compromise?

Dataplex can help companies build in the right balance of governance and access to their data platform. Dataplex is an intelligent data fabric that unifies distributed data to help automate data management and power analytics at scale. It brings together data warehouses, data lakes and data marts through a single pane of glass. By understanding that all end users in an enterprise can and should be a “data person”, user experience tools can help minimize the skill gap which have been a barrier to people getting access to real-time and central data in an enterprise. Utilizing these managed services is considered cloud native. It takes advantage of the cloud’s investment in infrastructure and site reliability engineers to maintain service level agreements (SLAs). Providing a cloud agnostic service, though, is difficult because of the nuanced differences behind the software that interacts with the hardware. Thus, many CxOs will often face the choice of being more cloud native (with less environmental complexities) or cloud agnostic (with less vendor lock-in).



One prominent example of a serverless and fully managed native cloud product that can serve as a complete modern replacement for an Enterprise data warehouse is Google's BigQuery. This powerful columnar distributed system enables scalable analysis over petabytes of data in ANSI SQL. In fact BigQuery is the externalized product of one of Google's powerful analytical engines: Dremel[6], a cloud-powered massively parallel query service used internally by many of its products.

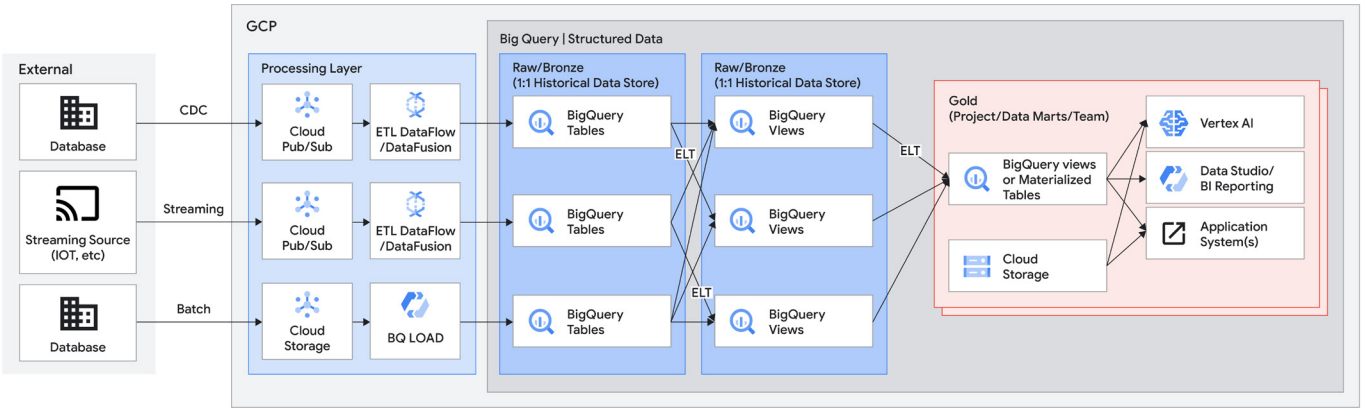
While these systems and teams are working towards converging at an organization, it also makes sense to leverage cloud-native technologies oriented towards file-based or data lake style workloads. Google's Dataproc is a managed offering that enables persistent Hadoop or Spark clusters, among others, to be thought of as serverless job-based tasks. Because it is serverless and only persists for the length of the job, rather than a 24x7 cluster, there can be a paradigm shift towards the way data teams interact with the data lake. It can be modernized as it is converging with the data warehouse. Additionally, Google Cloud offers other tools that can help shift the traditional data lake paradigm to a more modern approach including Dataproc notebooks — a managed Jupyter notebook environment — for easier collaboration on Hadoop-based workloads.

Convergence of the data lake and data warehouse is about simplifying and unifying the ingestion and storage of the data, and leveraging the correct computing framework for a given problem. For structured and semi-structured data, writing all of the data as it is streamed into tables using a change data capture (CDC) tool enables users to use simple SQL to transform the data and build logical views to query the data in a way that aligns with the business use cases. Because views are heavily leveraged in this pattern, there can be column elimination, partitioning, and logic to optimize the speed of the queries while maintaining a historical ledger of the data streamed into the tables.

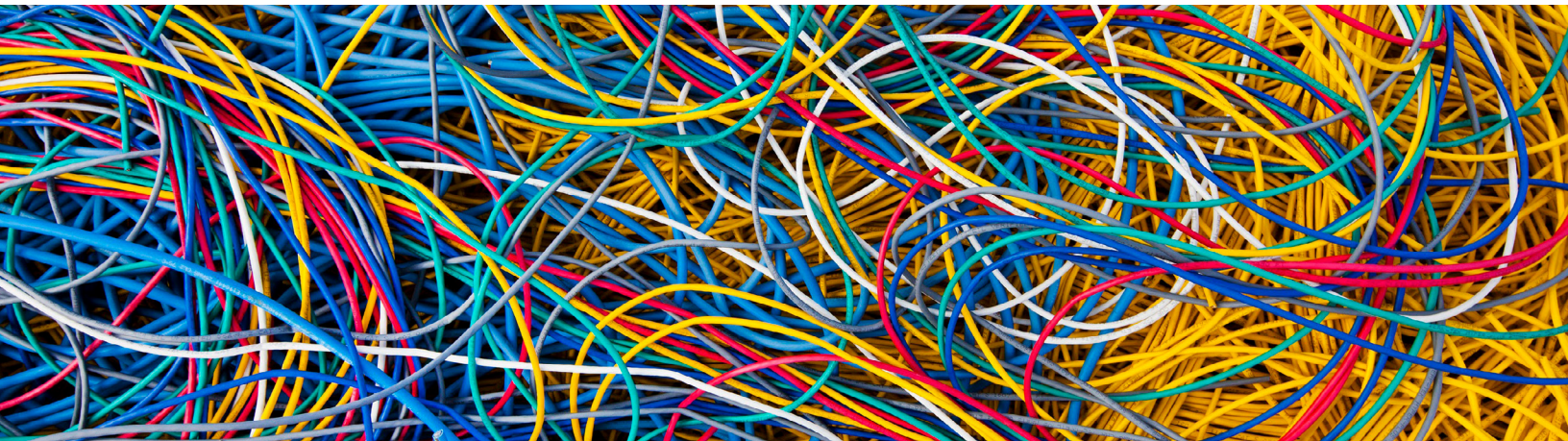
6. https://research.google/pubs/pub36632/

Conversely, data that is ingested via a batch pipeline can use a similar approach by which all of the data is written to a table, and SQL is used to create views with the most recent view of each record. Like the streaming data, a historical ledger is maintained in the raw tables, allowing data scientists to use all of the data for building and testing ML models. In this architecture, users can leverage scheduled queries or an event-based lambda architecture for data ingestion.

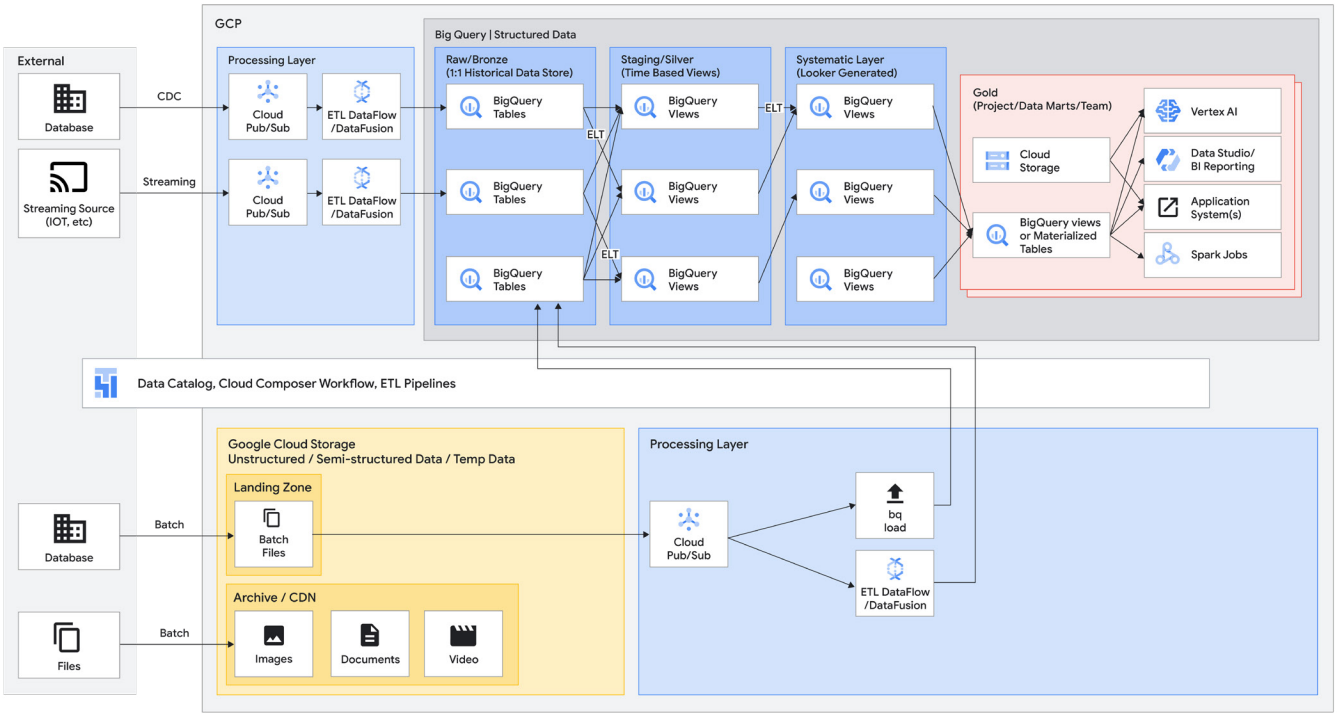
The data pipeline for ingestion could start to look like the following diagram, where the data pipeline will support ingestion by providing a different layer depending on how the data is transformed:



The “Gold” layer of the project can be the business-driven views or materialized tables that are governed and purpose-built. The only place the data is stored is at the “Raw” layer, unless there is a reason to materialize the data in the final layer for performance purposes. The underlying logic and storage provides access to end users and applications alike, allowing for the converged data platform to be used for Hadoop, Spark, analytics, and ML.



The following diagram adds the additional pieces to the previous diagram to demonstrate the convergence of data lakes and warehouses:



It no longer matters if the data is stored within the data warehouse or within the freely floating cloud bucket. This is because behind the scenes it is the similar distributed storage architecture but data is structured differently. For example, data lakes would move the data from HDFS to the same object storage outside the cluster. This is the same as what cloud EDW would use as the backbone of its storage system. As a result, data is easily accessible and managed by both data lake and data warehouse architectures in one place. Therefore, organizations can now apply governance rules around data residing in the lake and the same data accessed by the data warehouse. As a result, we can break down silos, not just by putting data into a central repository, but by enabling processing and query engines to move to wherever that data is. This leads to data warehouses and data lake convergence whereby applications such as Spark and R do ML in a single environment. This allows them to use the same metadata store and governance, enabling data engineers, data scientists and data analysts to work together in collaboration, rather than in siloed systems. Afterall, siloed data is silenced data.

Benefits of convergence

The benefits of the converged data lake and data warehouse environment presents itself in a number of ways. Most of these are driven by the ability to provide managed, scalable and serverless technologies. As a result, the notion of storage and computation is blurred. Now it is no longer important to explicitly manage where data is stored or what format it is stored. Users should be able to access the data regardless of the infrastructure limitations.

This is achieved through convergence of the data lakes and warehouses covered with the below points:

Time to market

Data can be ingested and acted upon straight away whether it is batch or real-time data sources. Rather than employing complex ETL pipelines to process data, data is “staged” in either a messaging bus or through object storage. Then it is transformed within the converged data warehouse / data lakes that enables users to act as the data is received.

Reduced risk

Existing tools and applications continue to work without needing to be rewritten. This reduces the risk and costs associated with change.

Predictive analytics

Moving away from the traditional view of data marts and data mining to real-time decision-making using fresh data increases business value. This is only possible as barriers to entry have been reduced as governance and strictness around DWs have come down.

Data sharing

Converged environment is now the one-stop shop for all, regardless of the type of user (i.e. data analyst / data engineer / data scientist) can all access the same managed environment but different stages of data when it is required. For example, data is ingested and stored in its raw form in the data warehouse, transformed and made available in the presentation layer. Storage is cheap in data warehouses such as BigQuery, as a result all stages of data can be stored within the data warehouse. At the same time different roles can have access to the same data through different layers and this is governed by platform wide access rights. This does not only increase the data governance but also allows simpler access management and auditing throughout the data ecosystem.

ACID transactions

In a typical data warehouse the data integrity is maintained, and multiple users reading and writing the data see a consistent copy of the data. Although ACID is a key feature in the majority of the databases, traditionally this has been rather challenging to provide the same guarantees when it comes to traditional HDFS-based data lakes. There are schemes such as Delta Lake and Apache Iceberg which try to maintain ACID semantics; they store a transaction log with the aim of keeping track of all the commits made to a data source. However, this introduces yet another layer of complexity and is best maintained by modern data warehouses. For example, BigQuery and Snowflake provide such capabilities.

Multi-modal data support

Semi-structured and structured data are key differentiators with the data warehouses and data lakes. Semi-structured data has some organizational properties such as semantic tags or metadata to make it easier to organize, but data still does not conform to a strict schema. In the converged world this is accommodated with extended semi-structured data support. On the other hand, for unstructured use cases, data lakes are still required apart from edge cases. For example, it is possible to store unstructured data such as images in data warehouses environments such as BigQuery and then apply ML models to it.

Breakdown silos and ETL pipelines

Traditionally data capture, ingest, storage, processing and serving are managed by different tools and environments, usually orchestrated by ETL tools. In addition, processing frameworks such as Spark, Storm Beam, etc provide built-in ETL templates to allow organizations to build ETL pipelines. However, with capable Cloud EDWs and integrated Cloud tools this pipeline is now all handled by a single environment. Most of the traditional ETL tasks such as cleanse, de-dupe, join and enrich are done by ELT. This is made possible at different stages of the Data Lake implementation within the DW. Furthermore, with the support of core data warehouses concepts such as stored procedures, scripting, and materialized views are made available through a united environment.

Schema and governance

In reality, business requirements and challenges evolve in time. As a result, associated data changes and accumulates, either by adapting to new data or by introducing new dimensions. As the data changes, applying data quality rules becomes more challenging and requires schema enforcement and evolution. Furthermore, PII data governance becomes more important as new data sources are added as well. There needs to be a data governance solution allowing organizations to have a holistic view of their data environment. In addition, it is paramount to have the ability to identify and mask PII data for different purposes and personas.

Streaming analytics

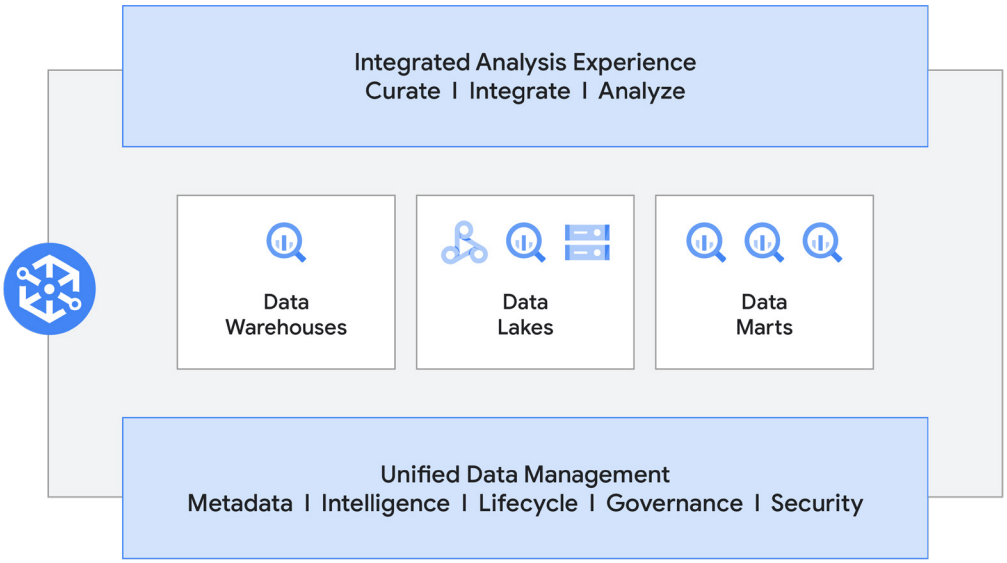
Real-time analytics enables immediate responses and there would be specific use cases where extremely low latency anomaly detection application is required to run. In other words, business requirements would be such that it has to be acted upon as the data arrives on the fly. Processing this type of data or application requires transformation done outside of the warehouse.

Future of smart analytics

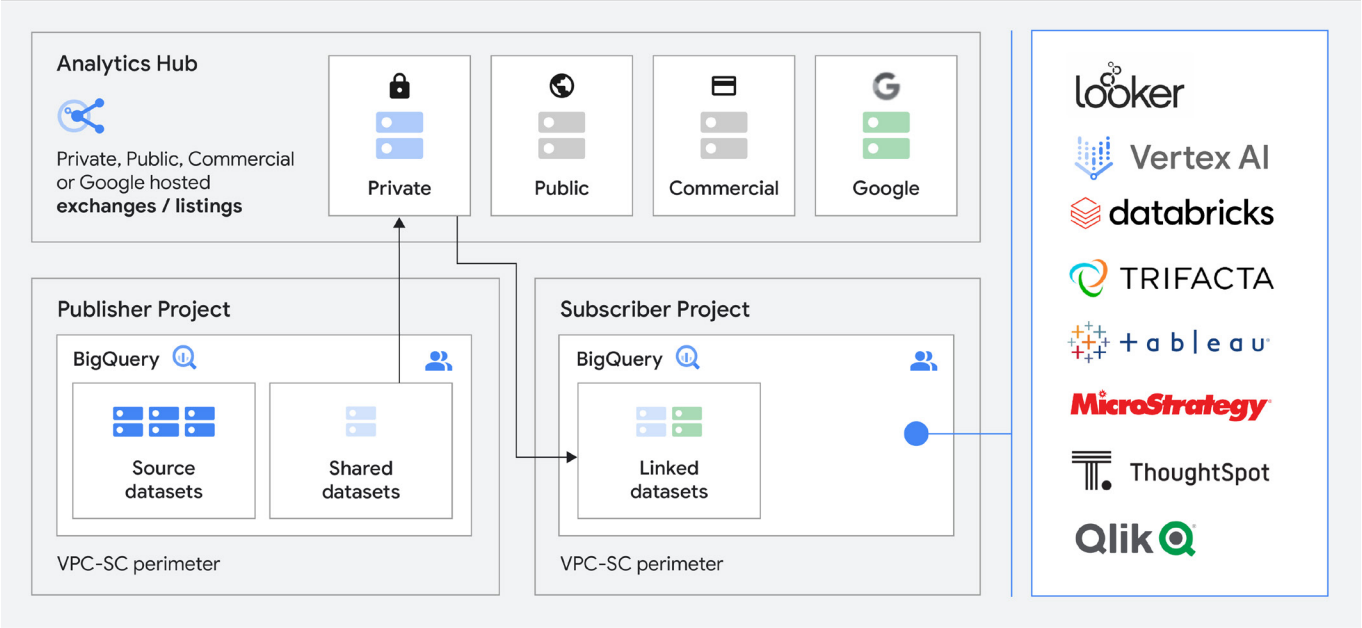
As discussed previously, cloud computing provides a serverless and comprehensive, end-to-end data ecosystem. Further, it allows tools to embed and support ML natively. Each of those are valuable in their own right; it is better to offload machine management. All organizations are looking to do ML to stay ahead of the competition. However, reality is more than that; when services are serverless, enterprises have to calculate the resources needed to run and administer jobs, which increases the overhead. The technical burden is reduced when resources are calculated automatically, thus allowing organizations to focus on using new tools and enabling new personas to analyze existing data. Furthermore, when we have a connected and embedded platform that's built on a serverless infrastructure, we can enable our tools to easily call upon and control other parts of the platform, like we see with Dataflow SQL — creating pipelines from within a data warehouse. In addition, the serverless aspect means there is no provisioning required to make that pipe happen, just the knowledge that it would help the use case. Finally, when we have connected products, and products that are autoscaling with serverless tech, we can deliver machine learning capabilities to multiple aspects of the pipeline. This results in creating a scenario where you could have a net-new, end-to-end ML pipeline that was constructed via SQL without leaving the data warehouse.

We can break down silos, not just by putting data into a central repository, but by enabling processing and query engines to move to wherever that data is. This effectively combines data warehouses and data lakes, so that traditional EDW and Hadoop/Spark workloads are served in one environment. This in return brings up traditional lambda architecture, stream and batch processing under one umbrella. As a result, data engineers can cope easily with real-time deeds without having to deal with it explicitly using another framework. On the other hand, the data silos that exist between on-premises, cloud, and multi-cloud environments are also converged by using technologies that allow data to be seamlessly ingested and/or staged before it is being processed. Last but not least, the areas of ML and DA are converged, meaning that data insights is democratized. Organizations, in return, focus on business insights and predictive analytics (AutoML, ML APIs) rather than investing into net-new organizational capabilities without spending all their effort turning themselves into engineering organizations.

Additional elements that are not included in this paper, but represent important pieces of any data ecosystem are orchestration, data cataloging, metadata management, and governance. For governance, Google Cloud provides Dataplex. This 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 your data at scale. Finally, you can build an analytics strategy that augments existing architecture and meets your financial governance goals. Essentially, it brings all of the pieces together:

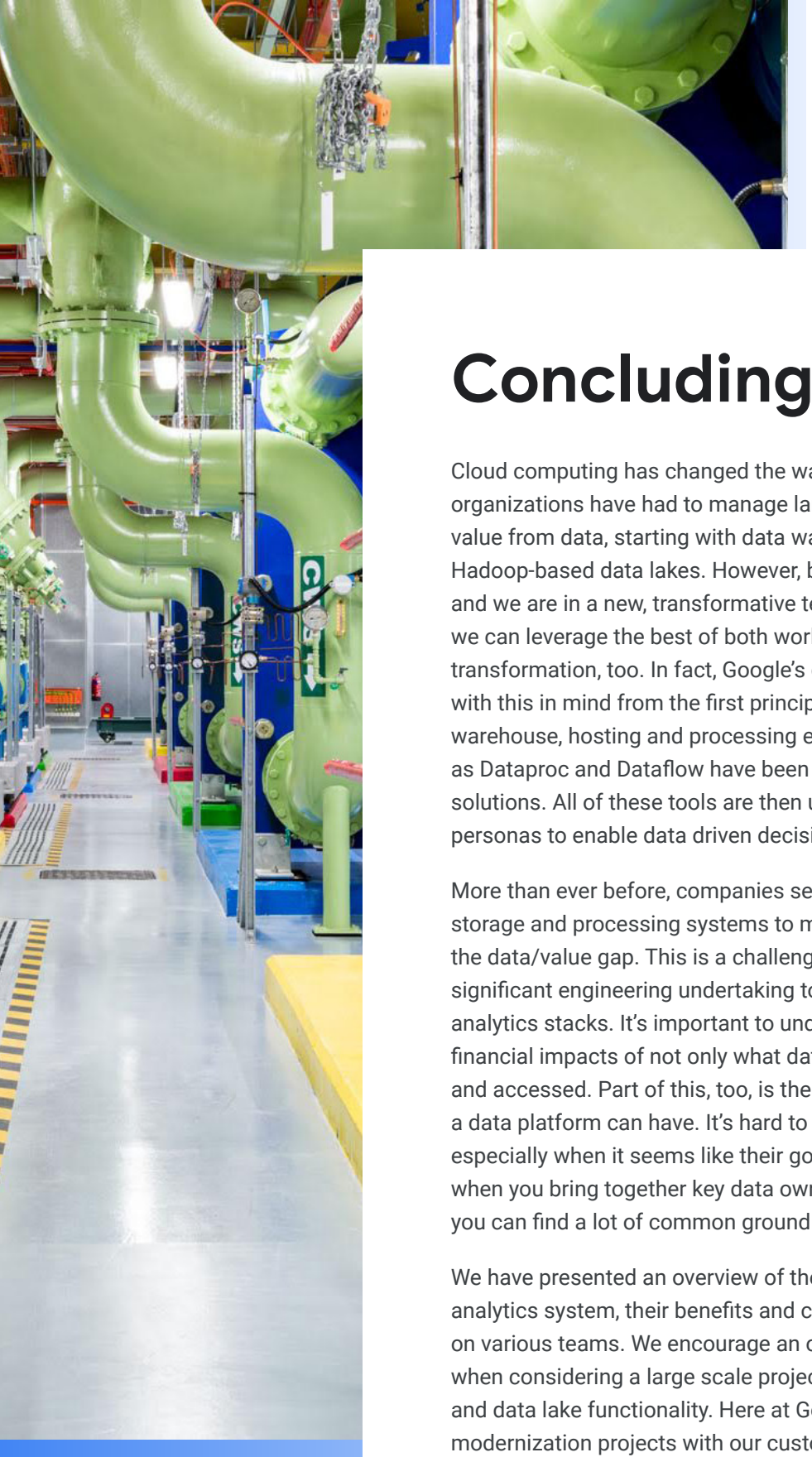


When governance, cataloging, and metadata management are pieced together in a business-driven approach, data can be leveraged as a shareable and monetizable asset within an organization or with partner organizations. To formalize this capability, Google offers a layer on top of BigQuery called Analytics Hub. 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 (within or outside of their organization).



With Analytics Hub, you can 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. Organizations 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 are what 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.

This convergence manifests itself in giving existing personas new skills with minimal training. Therefore the roles separating data analysts, data engineers, data scientists, and ETL developers starts to blur. There are now converged data teams, whereby on both ends of the spectrum data analysts and data scientists are doing simple data engineering tasks. On the one end, data engineers focus on creating reusable data processing pipelines to enable scaling of a number of applications. On the other end, traditional DBAs focus their tasks on governance of the environment rather than maintaining different hardware and low level operational aspects of the database. This in return removes bottlenecks, increases agility, increases employee retention, and delivers a more diverse set of insights that originate closer to the business unit than ever before.



Concluding remarks

Cloud computing has changed the way that we approach data. Traditionally, organizations have had to manage large amounts of infrastructure to extract value from data, starting with data warehouses and leading to the rise of Hadoop-based data lakes. However, both approaches have their challenges, and we are in a new, transformative technical era in cloud computing where we can leverage the best of both worlds. Google has gone through this transformation, too. In fact, Google's data processing environment is built with this in mind from the first principles. BigQuery acts as a massive data warehouse, hosting and processing exabytes of data. Processing engines such as Dataproc and Dataflow have been closely coupled with BigQuery and other solutions. All of these tools are then used seamlessly by different teams and personas to enable data driven decision making and applications.

More than ever before, companies see the need to modernize their data storage and processing systems to manage massive data volumes and close the data/value gap. This is a challenging problem to solve, and it can be a significant engineering undertaking to overhaul and consolidate legacy data analytics stacks. It's important to understand the technical, business, and financial impacts of not only what data is being collected but how it is stored and accessed. Part of this, too, is the organizational impact that changes to a data platform can have. It's hard to bring together multiple stakeholders, especially when it seems like their goals aren't aligned. The good news is that when you bring together key data owners, users, and stewards of data systems, you can find a lot of common ground and agree on areas of compromise.

We have presented an overview of the key technical components in a data analytics system, their benefits and challenges, and the impact they have on various teams. We encourage an open dialog among stakeholders when considering a large scale project to bring together data warehouse and data lake functionality. Here at Google we have deep expertise in data modernization projects with our customers and can help you make the right architecture and implementation decisions to meet your business needs. Contact us for more info.