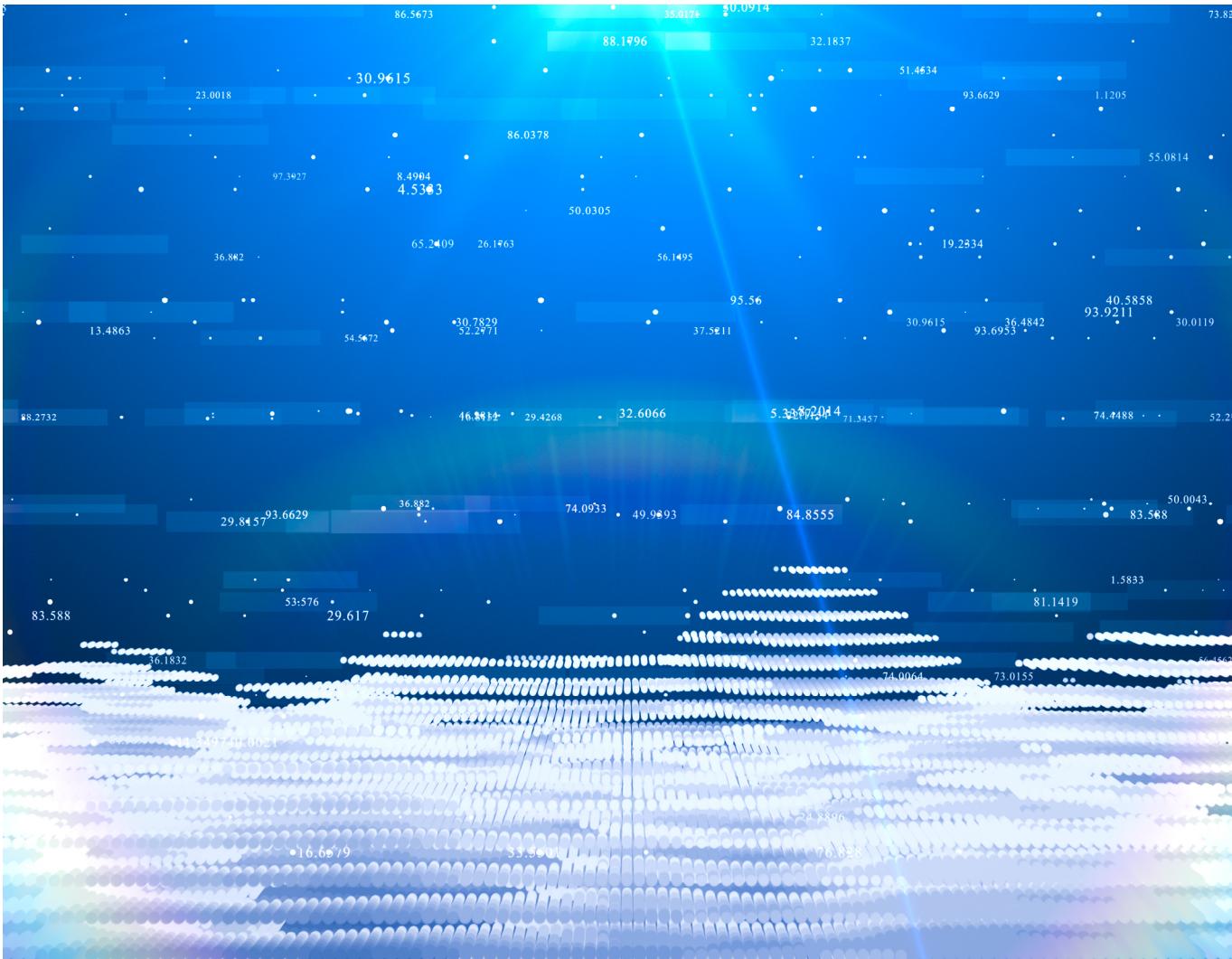
Google Cloud

Google Cloud Whitepaper

Data analytics decisions: Steps to building a modern data warehouse



Google Cloud

Table of Contents

Editor's note

As data grows and customers demand real-time experiences, a strong data foundation is essential. When we hear about typical problems and growing pains that enterprises are encountering, they usually have to do with data. These include legacy data warehouses collapsing under the weight of all the new data that's coming in, unmet analytics needs across the company, and falling behind on modern tools like machine learning and artificial intelligence.

Ever faster and larger data streams, global business needs, and tech-savvy users are all putting the pressure on IT teams to move more quickly and with more agility.

Despite all of these changes, it's often the traditional, legacy data warehouses where many of the data analytics tasks take place, and these systems are underprepared for new demands. When we talk to people working in data today, we hear a lot about the constraints that come with operating legacy technology while trying to build a modern data strategy. Those legacy data warehouses likely aren't cutting it anymore.

To scale a data warehouse, it helps to automate the work of systems engineering away from the work of data analysis, one of the benefits that BigQuery enables. Once those functions are separated, the analytics work can take center stage and users can become less dependent on administrators. BigQuery also helps remove the user access issues that are common with legacy data warehouses. Then users can focus on building reports, exploring datasets, and sharing trusted results easily.

In this whitepaper, we'll explore the key components of a modern data warehouse, what you should consider as you're adopting a cloud data warehouse, and how Google's BigQuery was designed to meet modern needs.

Chapter 1
Making the right infrastructure choices

As business users and teams start to understand the possibilities of data analytics, IT teams are feeling the pressure. Everyone wants access to the latest data, and to be able to work with it and use it right away, but a lot of companies just aren't prepared to support that.

In our engagements with customers, we often observe that a majority of their time is spent on systems engineering, while only about 15% is spent analyzing data[1]. That's a lot of time spent on maintenance work. Because legacy infrastructure is complex, we often hear that businesses continue to invest in hiring people to manage those outdated systems, even though they're not advancing data strategy or agility.

Back in the days of hardware dependency, your data warehouse required tuning and upkeep. (This may still be your model, even with some cloud data warehouses.) The serverless model changes that, and opens up a lot more possibilities once maintenance is a thing of the past.

Everyone wants access to the latest data, but a lot of companies just aren't prepared to support that.

Like any cloud deployment, it's possible to port an inefficient architecture into the public cloud.

What to consider when choosing infrastructure

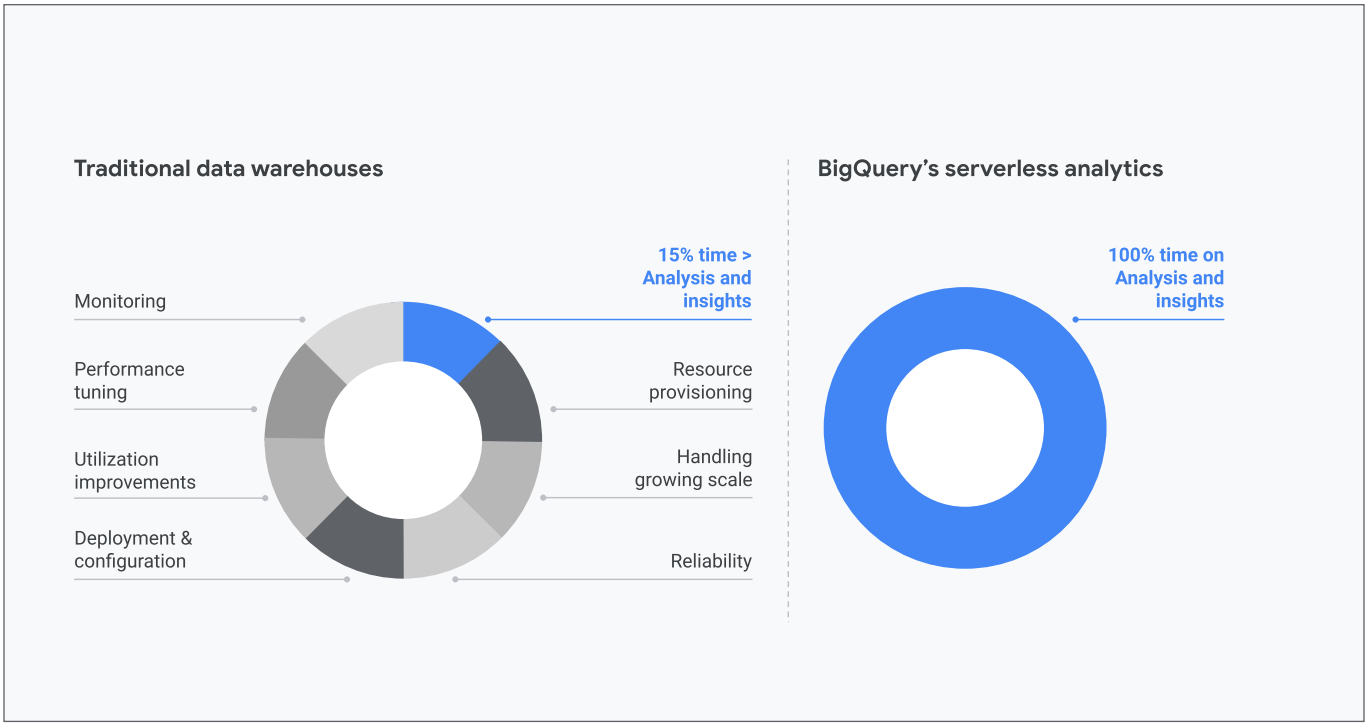
Choosing the right infrastructure to support your modern analytics strategy is essential. Like any cloud deployment, it's possible to simply port an inefficient legacy architecture into the public cloud. To avoid that, it helps to think about the total cost of ownership (TCO) for data warehouses, because it captures the full picture of how legacy technology costs and business agility aren't matching up. For

1 Google internal data, January 2020.

example, moving to BigQuery isn't just moving to the cloud—it's moving to a new cost model. One where you're cutting out that underlying infrastructure and systems engineering—and the costs that they bring. (You can get more details on cloud data warehouse TCO comparisons from ESG.)

The cloud can offer much more cost flexibility, meaning you're not paying for, or managing, the entire underlying infrastructure stack. From a data infrastructure perspective, separating the compute and storage layers is essential to achieving business agility. When storage and compute are decoupled and can scale independently, and on demand, you don't have to keep expensive compute resources up and running all of the time. This is very different from traditional node-based cloud data warehouse solutions or on-premises massively parallel processing (MPP) systems. It also allows for stateless, resilient computing—and a logical database model, rather than a physical one.

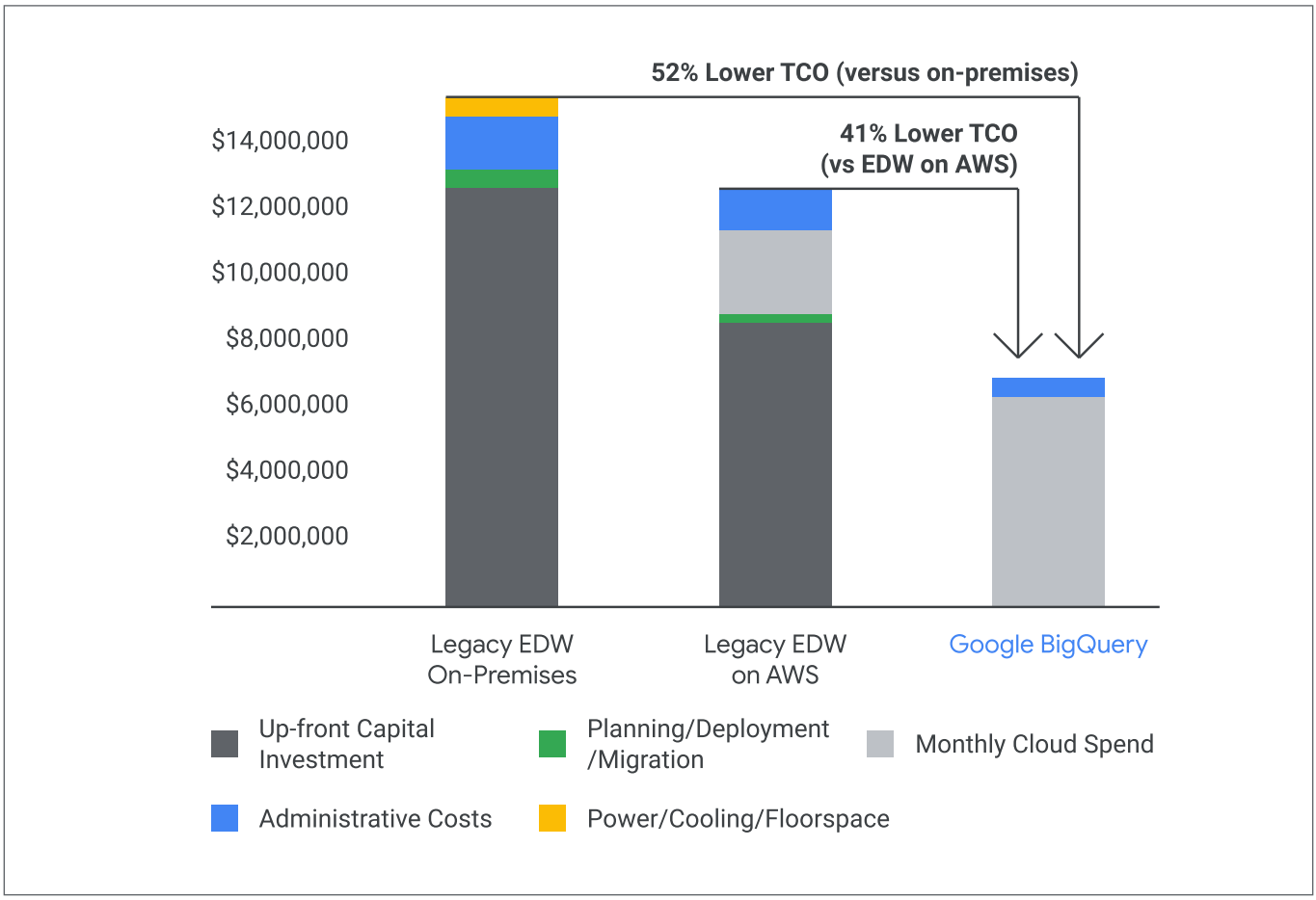
At Google, we designed our data warehouse to meet modern business needs. This serverless concept is simple, but powerful when applied to gigabyte- or petabyte-scale data analytics. BigQuery's on-demand analysis engine is provisioned on the fly, based on the computational requirements of a specific query. There is no need for customers to define nodes or clusters. BigQuery also automatically manages query performance based on the volume of data it needs to process. This is a fundamentally different approach from on-premises or other cloud infrastructure. So, for example, if you haven't queried a table in a couple of weeks, it will perform just as well as the one that you query all day. You can spend less time optimizing queries or creating multiple data copies or warehouses.



How BigQuery's serverless model saves money and time

Our approach at Google Cloud is to decouple system engineering tasks from data analysis—and then automate those tasks—to simplify the need for data operations. A cloud data warehouse like BigQuery doesn't require the constant maintenance of the past, so database administrators (DBAs) can spend more time on planning and innovation, and less time on upkeep. This all makes getting started and scaling up much faster and easier.

When a data warehouse can handle your scalability needs and self-manage performance, that's when you can really start being proactive.



Expected three-year total cost of ownership for BigQuery

ESG found that BigQuery can reduce overall three-year costs by 52% when compared to an on-premises solution and by 41% when compared to AWS[2].

2 ESG Publication, "The Economic Advantages of Migrating Enterprise Data Warehouse Workloads to Google BigQuery," March 2019.

Chapter 2
Data security and management, done right and made simple

Data has been mission-critical in every era of human history. It was true for Neolithic farmers when they planned the next year's harvest and for 17th-century navigators whose survival depended on having decent maps. It's true for today's most sophisticated researchers as they seek answers to problems like disease and climate change.

But the scale of the world's data is larger than ever, with IDC predicting that the global datasphere will increase more than fivefold between 2018 and 2025[3]. To remain competitive in today's world, businesses must activate all that information by transforming it into faster insights, better customer experiences, smarter predictions, and products and services that meet changing expectations. Data has become every organization's most valuable asset. This means that data governance—how businesses protect and manage their most precious resource—can't be an afterthought, especially when moving data to the public cloud.

The stakes are high for businesses that don't prioritize data security. Breaches can gravely damage a company's reputation and profitability: In 2018, the average cost of cybercrime for an organization reached $13 million[4]. A solid data governance framework can help you minimize this risk while addressing other fundamental concerns related to data migration, including questions surrounding access control, privacy, and regulatory compliance.

3 IDC White Paper, sponsored by Seagate, "Data Age 2025: The Digitization of the World From Edge to Core," November 2018.

4 Accenture, "Ninth Annual Cost of Cybercrime Study," March 2019.

Data governance principles and best practices

As you build out your governance capabilities, keep in mind the entire data lifecycle, including intake and ingestion, cataloging, persistence, retention, storage management, sharing, archiving, backup, recovery, disposition, and removal and deletion.

We've put together seven best practices and guidelines for governance in the cloud.

1. Discover and assess

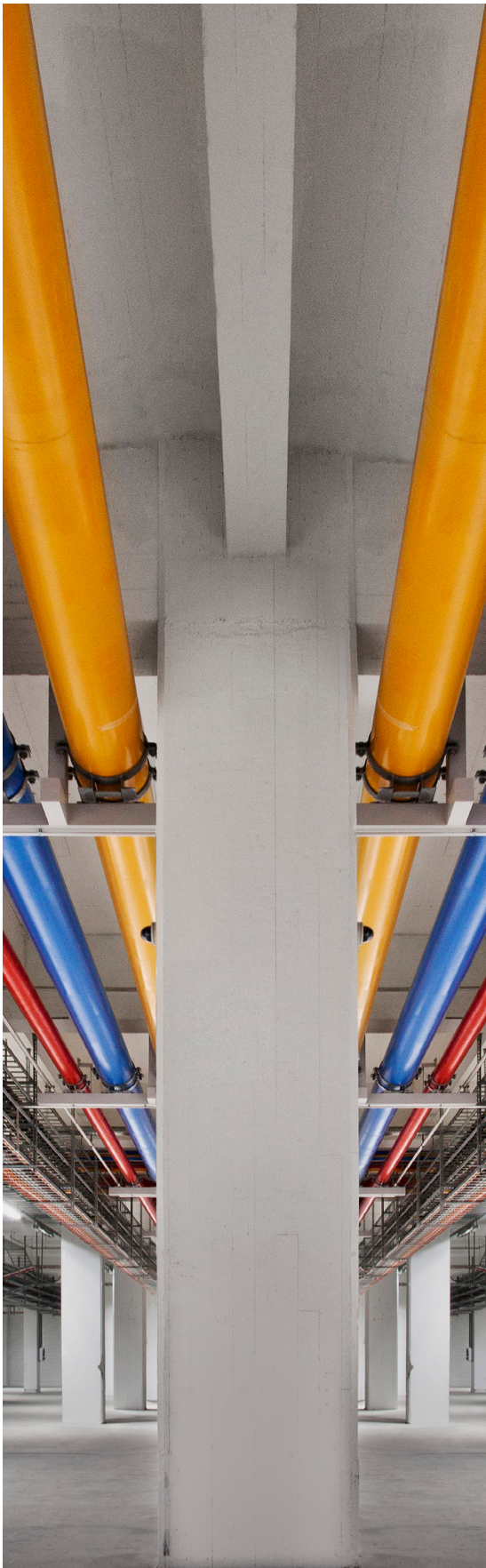
Cloud-based environments often offer an economical option for creating and managing data lakes, but the risk of ungoverned migration of data assets remains. This risk represents a potential loss of knowledge of what data assets are in the data lake, what information is contained within each object, and where those data objects originated from. A best practice for data governance in the cloud is to institute an initial data discovery and assessment phase in order to know what data assets you have.

2. Classify and organize

Properly evaluating a data asset and scanning the content of its different attributes can help categorize it for subsequent organization. This process can also infer whether the object contains sensitive data and, if so, classify it in terms of the different levels of data sensitivity. This helps inform which governance policies and procedures apply to the data.

3. Catalog your data

Once your data assets are assessed and classified, it's crucial that you document your learnings so that your data consumer communities have visibility into your organization's data landscape. You need to maintain a data catalog that contains structural metadata, data object metadata, and sensitivity-level assessments in relation to the governance directives (such as compliance with one or more data privacy regulations).



4. Validate and assess quality

Different data consumers may have different data quality requirements, so it's important to provide a means to document data quality expectations, as well as techniques and tools to support the data validation and monitoring process. Data consumers also need to know data lineage, so they know what they're getting in results. Data quality management processes include creating controls for validation, enabling quality monitoring and reporting, supporting the triage process for assessing the level of incident severity, enabling root cause analysis and remedy recommendations for data issues, and tracking data incidents.

5. Manage data access

There are two aspects of governance for data access. The first aspect is the provisioning of access to available assets. It's important to provide services that allow consumers to access their data, and fortunately, most cloud platform providers offer methods for developing data services. The second aspect is the prevention of unauthorized access. To establish a level of managed access, it's important to define identities, groups, and roles, and to assign access rights.

6. Audit, track, and monitor

Organizations must be able to assess their systems to ensure they're working as designed. Monitoring, auditing, and tracking (for example, who did what and when and with what information) help security teams gather data, identify threats, and act on them before they result in business damage or loss. It's important to perform regular audits, checking the effectiveness of controls, in order to quickly mitigate threats and evaluate overall security health.

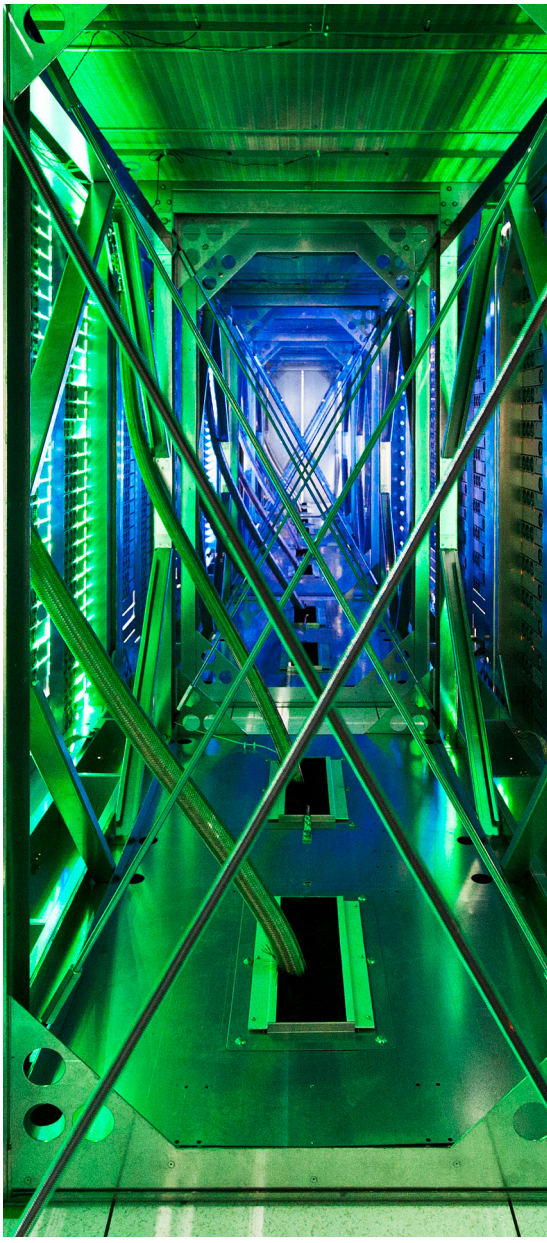
7. Protect your data

Despite the efforts of IT security groups to establish perimeter security as a way to prevent unauthorized individuals from accessing data, perimeter security alone is insufficient for protecting sensitive data. It's difficult to perfectly defend against security breaches, and particularly difficult to prevent insider exfiltration. This makes it

important to institute additional methods of data protection to ensure that exposed data cannot be read, including encryption at rest, encryption in transit, data masking, and permanent deletion.

Using these methods can help you build a data governance strategy and operating model with these three key components:

1. A framework that enables people to define, agree to, and enforce data policies
2. Effective processes for control, oversight, and stewardship over all data assets across on-premises systems, cloud storage, and data warehouse platforms
3. The right tools and technologies for operationalizing data policy compliance



These components help establish control over and maintain visibility into your company's data assets, gaining your company an advantage over the competition. You'll notice significant benefits as you continue to promote a data-driven culture, including improved decision-making, better risk management, and the ability to meet changing regulatory compliance requirements.

The BigQuery approach to governance and security

Google Cloud's BigQuery was designed to make it easy to build an effective data governance strategy and operating model. It offers built-in data protection at scale, tools and technology to govern data, compliance with industry standards, and trust through transparency. Here are more details:

Built-in data protection at scale

Data is automatically encrypted while in transit and at rest, and it can only be accessed by the authorized roles and services that have audited access to the encryption keys. You can manage cryptographic keys for your cloud services the same way you do for your on-premises systems. A multilayered security approach across hardware, services, user identities, storage, internet communication, and operations provides redundancy and reliability. Purpose-built chips, servers, storage, networks, and data centers protect against hardware-level intrusion.

Tools and technology to efficiently govern data

There are some Google Cloud tools that accompany BigQuery to help you track and govern data: Cloud Data Loss Prevention (DLP) automatically helps you discover, classify, and redact sensitive data—such as credit card numbers, names, and social security numbers—in Google Cloud, in other clouds, or in your on-premises environment, according to your data policies. Cloud DLP also includes a feature that shows you whether a de-identified piece of data could make other data identifiable, such as when there's only one person in a certain age bracket in a ZIP code. Data Catalog provides a fully managed metadata management service that lets you quickly discover, manage, and understand your data. Cloud Identity and Access Management (IAM) offers access control and visibility as well as a unified view into security policies.

Compliance with industry standards

Google Cloud products regularly undergo independent third-party verification of security, privacy, and compliance controls. In addition, Google Cloud offers data privacy, data portability, and threat protection products and features that can support your compliance efforts. BigQuery offers high availability and a 99.9% service-level agreement so you can have peace of mind, and data is automatically replicated, restored, and backed up to ensure business continuity. It's been certified in customer environments as HIPAA- and PCI-certified, for example.

145

Average number of security breaches in 2019

+11%

increase in security breaches last year

67%

increase in security breaches in the last five years

$5.2T

Organizations could benefit from US$5.2 trillion of future revenues over the next five years

***Source**: Ninth Annual Cost of Cybercrime Study*

Trust through transparency

Access Transparency gives you near-real-time logs when Google Cloud administrators access your content. It also allows you to perform regular audits of access by administrators as a check on the effectiveness of our controls.

To see these innovations in action, look no further than GO-JEK, Southeast Asia's on-demand multiservice platform and digital payment technology group. The company used Google Cloud to build a centralized search engine for their enormous and growing collection of data, which comes from hundreds of thousands of different sources and serves millions of users. BigQuery serves as their data warehouse, and their search engine is powered by Data Catalog APIs. BigQuery's integration with Cloud IAM provides access control and visibility each time one of GO-JEK's thousands of data and business analysts runs a query.

"Data Catalog gives us the flexibility we need in metadata management. Integration with Cloud Identity and Access Management (IAM) means that data discovery is ACL-ed through the Data Catalog search index, giving us peace of mind."

Ajey Gore, Group Chief Technology Officer, GO-JEK

Chapter 3
Getting users what they need, when they need it

In the past, data analytics tasks likely fell on a single team of analysts or data scientists who gathered metrics and created reports, either regularly or as needed by business teams. But it's not enough anymore to have a handful of dedicated data specialists. These days, every team needs access to advanced analytics. Data is valuable currency, and it lets your users collaborate better and faster. Separating data growth from data operations growth is the start to opening up user access and insights.

It's not enough anymore to have a handful of dedicated data specialists.

Massive data growth, customer demand, and the emergence of real-time streaming data capabilities have also changed the landscape for business data analytics. While overnight data operations and pre-canned reporting used to be the norm, the global opportunities for businesses mean that a data warehouse now has to load streaming and batch data while also supporting simultaneous queries. Hardware and high costs are the main constraints for legacy systems as they struggle to keep up.

Users share some common needs—using data to make better, quicker decisions.

Across industries, users share some common needs—using data to make better, quicker decisions. And users at many businesses are trying to stay ahead of the competition with data-driven decisions. Specific industry use cases for real-time data include:

- **Ecommerce**: clickstream analysis and dynamic user segmentation
- **Retail**: processing of point-of-sale transactions for real-time inventory positions
- **Mobile gaming**: in-game prompts and adjustments based on user behavior
- **Manufacturing**: IoT data analysis for improving operational efficiency and predictive maintenance

Other kinds of important real-time data that business users are contending with include social media, mobile interaction data, network logs, and more. All too often, batch and streaming data analytics are running in parallel within separate systems, so it's hard to unite those patterns to get a holistic view of how all of that data can be used to help the business.

Connecting customers with data to modern technology

Moving your existing architecture into the cloud often means moving your existing issues into the cloud, and we hear from businesses that the shift still doesn't allow for real-time streaming. That's a key component for data analysts and users.

When streaming data is working well, it lets businesses personalize the online customer experience, predict real-time maintenance requirements in complex manufacturing systems, gain a consistent view of inventory positions based on real-time POS transaction data, calculate financial risks, detect fraud, and improve mobile gaming experiences for hundreds of thousands of players.

At Google Cloud, we're helping customers perform streaming data analytics without introducing complexity with our data-processing tool, Dataflow. Unifying batch and stream processing through an open-source SDK, such as Apache Beam, ensures portability and lets developers choose their language. Then, they can reuse code across stream, batch, and open-source processing frameworks. Dataflow even provides data analysts with the capability to deploy a streaming pipeline with SQL semantics, giving them the capabilities of a data engineer with no additional training required.

Along with BigQuery, these solutions are designed to be easy to use across an organization. BigQuery uses a standard SQL interface, allowing anyone to write queries. It's also simple to create reports and build visualizations, such as newly integrated Looker capabilities. Additionally, BigQuery lets you ingest up to 1 million rows per second

and analyze data on the fly. This is essential for modern data warehousing, so users can see and use the latest information when making decisions and serving customers.

Using BigQuery means you're able to move your analytics capabilities into the data warehouse itself, so BigQuery scales as more and more users are accessing analytics. Decoupled and scalable compute, combined with reasonable cost controls, is a pretty good way to help your business become digital. Instead of playing catch-up with user requests, you can focus on developing new features. Cloud brings added security, too, with the ability of cloud data warehouses to do things like automatically replicate, restore and back up data, and offer ways to classify and redact sensitive data.

Google Cloud customer 20th Century Fox was trying to improve their audience yield by more effectively targeting the right customer segments for their upcoming movies. Previously, a lean team of data scientists had done exploratory analytics using past movie box office performance data stored in BigQuery to correlate movie and trailer performance data. Using Google Cloud, with our simple standard SQL interface, let 20th Century Fox give their marketing analysts secure access to all of the relevant data. Analysts can now explore, uncover correlations, and, using BigQuery ML, create audience segmentation and targeting models to optimize media plans. The analytics insights are no longer only available to data scientists. Connected business intelligence lets them share and consume those insights with a few clicks. Their marketing teams are now able to launch targeted campaigns based on this data-driven user segmentation and improve campaign efficacy. These types of successes can lead to better customer experiences and business outcomes.

"We knew data was a big part of making decisions in the future. So we needed a platform that could scale to meet our growing appetite for it. Google Cloud Platform—in particular BigQuery—was ideal for this task."

Lye Kong Wei, Chief of Data Science, Group Head, AirAsia

Chapter 4
Bringing the technology of the future to the present

Using tools like artificial intelligence and machine learning may seem futuristic, but you can get started with them more easily than you might think. Applying ML to data adds efficiency way beyond what businesses have experienced without these tools, which can open doors to innovation and competitive advantage. The growth of data today, and the speed required to stay ahead of customer needs, means that new tools are needed too. Human speed just isn't fast enough, and AI and ML have matured and become easier to deploy to support your analytics needs.

All businesses generate data, but only some adopt ML to truly understand it. And there are many reasons why. Data analysts, proficient in SQL, don't typically have proficiency in programming languages like R or Python, or a deep understanding of feature engineering, model selection, and hypertuning processes. Hiring a team of data scientists to build predictive analytics solutions can be prohibitively expensive. And moving data to and from an enterprise data warehouse can be complex, time-consuming, and costly.

There are other challenges, too. According to a 2019 Gartner study[5], the top challenges to adopting AI for respondents were:

- Skills of staff (56%)
- Understanding AI benefits and uses (42%)
- Data scope or quality (34%)

5 Gartner, "Survey Analysis: AI and ML Development Strategies, Motivators and Adoption Challenges," Jim Hare and Whit Andrews, June 2019

We hear from customers that many of them are tasked with simplifying infrastructure and adding modern capabilities like AI, ML, and self-service analytics for business users. The best stories about digital transformation are those where the technological changes and the business or cultural needs occur at the same time. One customer told us that because BigQuery uses a familiar SQL interface, they were able to shift the work of data science away from a small, overworked group of data scientists and into the hands of many more workers. Doing so also eliminated the need to create siloed data lakes for the data that data scientists had extracted from one project at a time and placed into various repositories in order to train ML models.

These large-scale computational possibilities not only save time and overhead, but also let businesses explore new avenues of growth. AI and ML are already changing the face of industries like retail, where predictive analytics can provide forecasting and other tasks to help a business make better decisions. BigQuery lets you take on sophisticated machine learning tasks without moving data or using a third-party tool.

Using built-in advanced analytics

It's possible to start using this type of analytics now. We design our tools so that AI and advanced analytics are more accessible, bringing them closer to the data and users.

Google Cloud's BigQuery data warehouse lets users analyze large datasets interactively, enabling them to share insights and use customer analytics in order to build better product experiences. And its built-in BigQuery ML capabilities allow data scientists and data analysts to build and deploy machine learning models on massive, structured or semi-structured, datasets directly inside BigQuery using simple SQL statements. It's possible to perform predictive analytics, such as forecasting sales and creating customer segments, right at the data source.

Bringing together predictive analytics and past performance can then be viewed together through integrated tools like Looker and BigQuery for intuitive visualizations and reports.

Ecommerce company Zulily uses machine learning to offer personalized customer experiences for online shoppers. Zulily collects billions of clickstream data points from their website and analyzes them in real time to take action and ultimately increase sales and keep employees focused on innovation. Zulily moved their data pipeline and big data analytics to Dataproc and BigQuery to support real-time decision-making for customers and their more than 500 "merchants"—employees who interact directly with product suppliers—to market their offerings and manage inventory. With the freedom to scale easily and cost-effectively, Zulily increased their daily data collection from 50 million events to 5 billion, giving them richer data to increase sales conversions and improve marketing effectiveness.

"By collecting more data, we're innovating faster and making smarter decisions. Using real-time analytics on Google Cloud Platform, we were able to drive a significant increase in sales conversion in just a few weeks. Without Google Cloud Platform, we would never have been able to scale the clickstream data collection 100-fold in a fraction of the time."

Bindu Thota, Director of Product Management, Zulily

Getting started

Businesses are already seeing success from moving their infrastructure onto the cloud, and with the strength, security, and flexibility of BigQuery, you can design a data warehouse that adapts to your changing needs. As we've seen, a strong cloud-based data warehouse helps your employees make sense of the ever-growing flow of data and do more, giving your business the capacity to disrupt while avoiding being disrupted. And since this technology is already being used across various industries, the time to modernize your data warehouse is now.

Is your data warehouse ready to support your business innovations? Take a free assessment of your data warehouse's maturity and get your results immediately.

Google Cloud