Cross-Cloud Data Engineering: Unified Big Data Workflows with AWS and GCP

Praveen Kodakandla

Abstract

Using cloud computing has made it accessible for organizations to benefit from several providers and boost their flexibility, safety, and inventive ideas when handling data. This approach is gaining popularity in big data architecture because it enables businesses to formulate and handle data operations that work on AWS and GCP. With this approach, it's easier to change vendors, balance costs with performance, and improve compliance with the rules for handling data. Also, it leads to difficulties in things like joining various types of data, coordinating different services, ensuring safety, and checking the system's health. This article goes over the principles, tools, and patterns needed to develop effective cross-cloud big data workflows. The analysis covers important AWS and GCP services, explains common ways to tie them together, and discusses performance as well as governance matters. By studying a real case, we explain how companies can create data pipelines that easily expand, are fault-tolerant, and secure using all their cloud ecosystems. The author ends the article by introducing new trends that will play a key role in cross-cloud data engineering, such as cloud-agnostic platforms, AI-based orchestration, and responsible cost management.

1. Introduction

As the amount of data keeps increasing and organizations use more cloud services, companies are changing their data infrastructure designs and how they manage them. Businesses now tend to rely on multiple cloud platforms, since these strategies make it easier to grow and save money. Merging data workflows through multiple cloud platforms, notably AWS and GCP, has turned out to be a main strategy that fully utilizes what these providers have to offer.

Single-cloud setups help manage things easier, but they may tie a business to a particular vendor, open them to temporary regional interruptions of service, and reduce the opportunities to get the most out of their resources and performance. Thanks to the cross-cloud model, enterprises can benefit from the superior services of both suppliers and make use of Amazon S3 and GCP's BigQuery at the same time.

Still, merging big data operations on AWS and GCP brings many technical and operational issues. These consist of allowing devices to exchange information, following the same security rules, dealing with the delays in cross-cloud communications, and obeying regulations in different parts of the world. For a distributed cloud and data environment to be fault-tolerant, scalable, and governed well, high-quality orchestration tools, clear architectural structures, and good teamwork are needed.

The article gives a thorough look at how data engineering works between AWS and GCP. The first step is to check the reasons behind using multiple clouds and compare the main data services they offer. After that, it explains the principles of designing workflows, the different tools used for integration, and important strategies to save on costs and increase security. The article ends by discussing future trends that will play a part in the development of cross-cloud data systems.

2. The Rise of Multi-Cloud Architectures

Modern enterprises now need to use multi-cloud architectures, which wasn't true just a few years ago. A multi-cloud strategy involves placing applications, workloads, and data assets on different cloud solutions, usually by using both AWS and GCP, so businesses can remain agile, enhance how their systems run, and reduce the risks caused by depending on only one provider.

Using the multi-cloud approach, organizations can choose AWS Glue for data ingestion and BigQuery from GCP for advanced analysis in data engineering. With this method, you can distribute computing tasks to platforms that are skilled in certain fields.

Reasons Leading to the Rise of Multi-Cloud

a) By not relying on one cloud vendor, companies lower their risk of changes in cloud pricing, experiencing outages, and fewer innovative options, enabling them to negotiate better and use modern solutions.
b) Having multiple cloud providers means users can access servers in various areas and use unique services, achieving speedy responses and greater data transfer for globally spread customers.
c) According to data residency laws (like GDPR and HIPAA), all sensitive information needs to be processed and stored where they originated. multi cloud architectures lets enterprises select solutions specifically for each region that are governed by local legal requirements.
d) With the right planning, companies in the cloud can keep costs down as they still maintain the quality of their services.
e) Using multiple cloud providers for the same data makes it possible to immediately use a backup version in the event of a disruption in one cloud service.

Figure 1: Primary drivers motivating enterprise adoption of multi-cloud architectures, based on recent industry survey data.

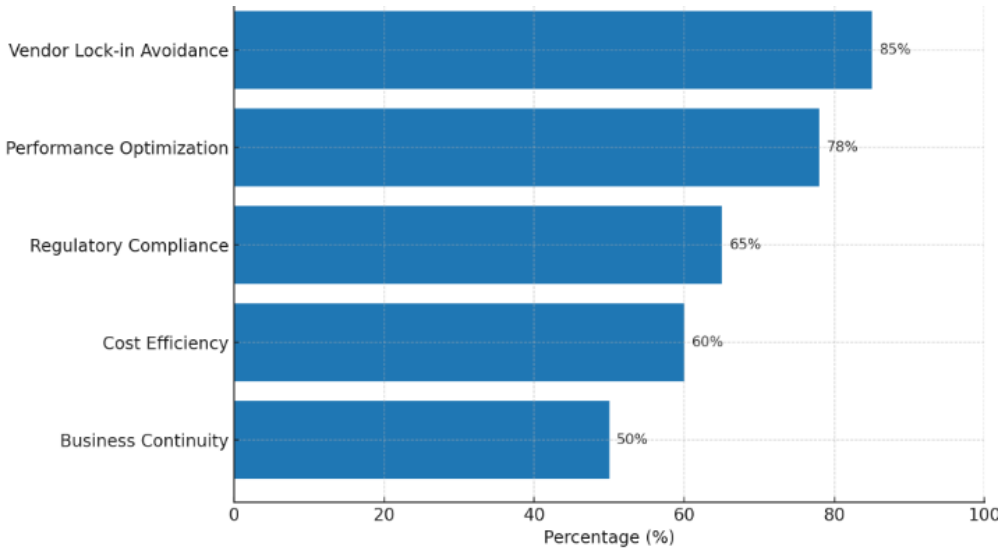


Figure 1 makes it clear that, besides performance benefits, people also choose multi-cloud solutions to not depend on one provider. Ensuring compliance with laws, saving money, and continuing operations are important for the business as well. Although the positive aspects are easy to see, businesses have to deal with difficulty in combining tools, following company rules, and setting up governance. This article will focus on those issues in the following sections.

3. Core Data Engineering Services: AWS vs. GCP

It is important to know the key data engineering services from AWS and GCP to design cross-cloud data workflows well. Every cloud platform comes with a range of tools and services appropriate for different steps of dealing with data, such as bringing data in, storing it, transforming it, analyzing it, and managing it. Comparing these factors from various resources allows companies to make the best choices about their systems, the balance between costs and performance, and how they will work with other technology.

3.1 The comparison of storage between Amazon S3 and Google Cloud Storage

Each of these cloud ecosystems is strongly supported by Amazon S3 and Google Cloud Storage as basics. While both platforms provide the same high durability, scalability, and consistency, how they are set up differently has a major effect on cross-cloud data engineering.

Supporting an array of functions, Amazon S3 has Intelligent-Tiering that helps to reduce costs by moving your data into cheaper storage spots depending on how it is used. Dataflow's policies make it possible to set up automatic data provisioning, so when matched with AWS Glue, Lake Formation, and Redshift Spectrum, it is great for data lakes. S3 stores different versions of each object, encrypts them by default, and provides access control with the help of IAM policies.

In contrast, Google Cloud Storage allows hassle-free permissions by giving access to all resources in a bucket along with strong integration with BigQuery, Dataflow, and Vertex AI services within GCP. Using multi-regional and nearline/coldline storage allows businesses to choose the balance between affordability and performance for storage of rarely needed data. Real-time analytics are the main priority in GCS, thanks to quick read times and effective streaming data handling.

When you are using multiple clouds, being aware of their cost plans, how clouds work together, and how much they can integrate on their own helps you choose the right storage solution for each pipeline bit.

Table 1: Feature Comparison – Amazon S3 vs. Google Cloud Storage

Feature	Amazon S3	Google Cloud Storage
Durability	99.999999999% (11 nines)	99.999999999% (11 nines)
Storage Classes	Standard, Intelligent-Tiering, Glacier	Standard, Nearline, Coldline, Archive
Lifecycle Management	Supported	Supported
Data Encryption	SSE-S3, SSE-KMS, Client-side	CMEK, CSEK, Tink-supported
Object Versioning	Yes	Yes
Access Control	IAM, ACL, Bucket Policies	Uniform bucket-level access, IAM
Integration	Glue, Redshift, Athena	BigQuery, Dataflow, Vertex AI
Event Triggering	S3 Events (Lambda, SNS, SQS)	Cloud Functions, Pub/Sub
Availability Options	Regional, Multi-AZ	Regional, Multi-Regional
Cross-region Replication	Yes	Yes (via Transfer Service or AP

A description of main points in comparison between Amazon S3 and Google Cloud Storage that matter for data engineers working with cloud services.

3.2 The third part of this question is: AWS Glue or Google Dataflow for data processing.

AWS Glue is a serverless and fully managed system meant for preparing data in batch operations. It processes Spark and Python workloads and allows managing both data schemas and metadata in a central place, important for a data lake architecture. Since Glue works with S3, Redshift, and Athena, it is fit for creating data pipelines that apply changes after data is stored.

With Google Dataflow, which is based on Apache Beam, you get one way to program both stream and batch processing. The way it scales automatically, runs pipelines from multiple environments, and re-distributes jobs in motion make it perfect for use in real-time event-driven analysis. By making use of Pub/Sub, BigQuery, and Cloud Storage, Dataflow can manage a great volume of data while providing fast results.

If a company needs to process time-critical data such as for fraud or IoT, Dataflow is frequently preferred due to its aim at streaming. AWS Glue is mostly used for handling data with a clear structure and reliable schema in one central location. Because Apache Beam is portable, Dataflow pipelines built with Apache Beam can run the same code in services from both AWS and GCP.

Table 2: Feature Comparison – AWS Glue vs. Google Dataflow

Feature	AWS Glue	Google Dataflow
Processing Model	Batch-oriented	Unified Batch and Stream
Underlying Engine	Apache Spark	Apache Beam
Orchestration Support	Integrated Job Scheduler	Pipeline DAG via Apache Beam SDK
Scaling	Serverless (auto-scaling Spark)	Dynamic, auto-scaling workers
Latency	High (batch-focused)	Low (supports near real-time)
Use Case Fit	ETL for structured data, data lakes	Real-time analytics, event-driven pipelines
Integration Ecosystem	S3, Redshift, Athena, Glue Data Catalog	Pub/Sub, BigQuery, Cloud Storage
Monitoring & Logging	CloudWatch, AWS Glue Console	Cloud Monitoring, Dataflow Console
Pipeline Portability	AWS-only	Portable across cloud providers (Beam SDK)
Cost Model	Per-DPU-hour	Per-GB processed + compute time

Framing AWS Glue and Google Dataflow, important differences for users include the modes of data processing, how quickly results are available, how each platform is organized, the supported integration tools, and the ability to use analytics in real time.

3.3 Analytics: Amazon Athena vs. Google BigQuery

Analytics is a main feature in data engineering, especially when managing data across different cloud services since fast access to analytics helps a lot. With Amazon Athena and Google BigQuery, users can get serverless analytics without having to take care of infrastructure.

Amazon Athena gives you the ability to query any data saved in Amazon S3 using standard SQL. Thanks to Presto, Athena is the right fit for quick analysis on data stored in Parquet, ORC, JSON, and Avro. Because you only pay for what you use and the data can be split and compressed, it is affordable for anyone trying to analyze data. Athena joins forces with AWS Glue to track metadata and makes it possible to query different data sources like RDS and Redshift together in one task.

Unlike the previous examples, Google BigQuery is a managed data warehouse made for dealing with huge amounts of data. It makes it so that storage and processing can be expanded separately. BigQuery comes with SQL queries, federated data queries, materialized views, and supports both types of data, either synchronized or real time. Those features make it easy for users to do machine learning and create dashboards in Bigquery thanks to BigQuery ml and BI Engine.

Since athena works well for cases that require affordable read analysis based on schemas, BigQuery handles complex joins, aggregation operations, and AI models on big data in S3. In this setup using BigQuery for analytics and AWS S3 to provide data, you can create a very powerful mix of tools in the cloud.

4. Cross-Cloud Data Integration and Interoperability

When organizations move to using various clouds, connecting their data from one cloud to another in real time becomes very necessary. Many enterprises make use of certain AWS and GCP tools, which means it is important for their workflows to be smooth between the two. To accomplish complete interoperability, projects need large-scale designs, common and easy-to-move tools, and standardized information. The section covers three wonderful aspects: real-time replication, federated querying, and the management of schemas.

4.1 CDC (Change Data Capture) technology allows data replication to happen in real time.

Having real-time replication is very important in cross-cloud data strategies because it is needed for consistent operations and hybrid analytics. Most of the time, Change Data Capture (CDC) is used to make replication possible. The process detects and collects any new information from the source and applies it right away to the target system. There are many cases in which AWS DMS, Debezium, and Striim are applied to CDC among cloud environments.

There might be a need for a healthcare firm to copy transactions from an AWS hosted PostgreSQL database, and instantly duplicate them in a BigQuery analytics layer in GCP. It makes it possible to analyze data that is genuinely current and yet, no re-importing is necessary. They are designed to ensure business continuity whether the system is moving to the cloud or going through a failure situation.

Greater learning happens when the following best practices are used:

a) Using WAL or similar mechanisms to keep track of the alterations in the database
b) Ensure that the transformation is consistent by monitoring the changes in the schema and data's quality.
c) Traffic should remain hidden and encrypted through AWS Direct Connect or GCP Partner Interconnect and also using Private Service Connect, a private AWS tool.

By means of CDC, event-driven mechanisms can be built, streaming use cases can be honored, and you can adjust exactly which data is sent to downstream services.

4.2 Federated Queries and Cross-Cloud Virtualization

For cases where putting copies of data in multiple locations is not possible due to data storage rules or performance reasons, federated queries work very well. With these queries, users can combine and work with data across several clouds by staying on their own platform. Thanks to this, data scientists and analysts can query only one view of the entire system instead of each data set separately.

It's possible for analysts to use BigQuery Omni, running standard SQL from GCP's BigQuery even when accessing data kept in Amazon S3 or Azure Blob Storage. Besides, Presto, Trino, and Apache Drill are examples of platforms that can query distributed data using memory-efficient performance.

Using federated approaches makes data egress costs much lower and also stops the risks that data might be duplicated. Besides all this, data locality regulations such as GDPR or HIPAA can be followed, allowing users to still get data from one place.

Technical Considerations:

a) Make sure that users are governed by the same IAM setup for accessing information no matter which cloud service they are in
b) Make performance better by pushing predictions in queries, filtering data, and saving it in formats like Parquet and ORC
c) Look at the query logs and statistics to find issues with connectors and fix them
d) It is effective for any business group anxious to analyze data on the spot with little need for data shifting and extra time.

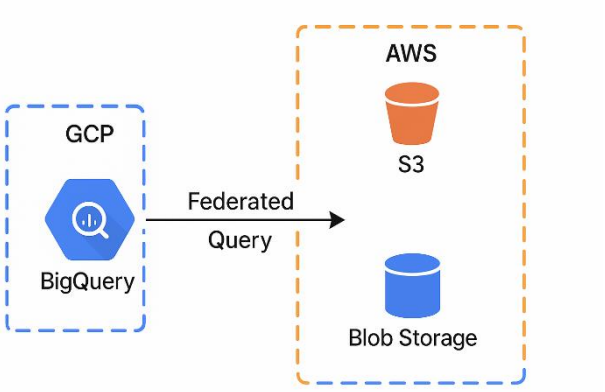


Fig 2: Federated querying across cloud providers: BigQuery on GCP executes queries directly on data stored in AWS S3 and Azure Blob Storage, enabling cross-cloud analytics without data duplication.

4.3 Schema Evolution and Metadata Consistency

The same schema and metadata must be used on AWS and in GCP to guarantee that cross-cloud extract-transform-load (ETL) processes run smoothly. A minor adjustment to table design, the data types, or naming things in a different way could cause ETL jobs to fail or lead analytics to give faulty results.

The problem can be solved by having one central layer for data and ensuring schema governance through using version control, contracts, and registry systems. These tools get the job done by storing metadata and different schema versions in one place.

As an example, a company combining Kafka on AWS with BigQuery on GCP could use Avro or Protobuf schemas kept in Confluent Schema Registry to support compatibility in its data streams and analytical tables.

Some of the Best Practices Are:

a) Carry out schema evolution by following rules, such as not deleting existing elements until some time window is applied
b) Configuring metadata syncs between cloud services to ensure that it stays similar Creating contracts for data that clearly explain what information, as well as its format, is passed between producers and consumers
c) If you make sure that schema and metadata are consistent, it becomes less likely that data integration will fail and observability increases in your data pipelines.

Table 3: Cross-Cloud Integration Tools and Use Cases

Tool/Service	Function	Cloud Compatibility	Primary Use Case
AWS DMS	Change data capture & replication	AWS to GCP (via connectors)	Database replication to BigQuery
Striim	Real-time data streaming and replication	AWS, GCP, Azure	Streaming analytics, multi-cloud sync
BigQuery Omni	Federated querying	GCP interface to AWS/Azure	Unified analytics across data lakes
Apache Beam	Unified batch/stream data processing	AWS, GCP (Dataflow/EMR)	Cloud-agnostic ETL pipelines
Confluent Schema Registry	Schema management	Multi-cloud	Kafka-compatible schema versioning

Important instruments that provide live data transfer, join queries spanning both cloud platforms, and proper coordination of metadata.

5. Security, Compliance, and Identity Federation

Modern IT infrastructures, mainly those using cloud or hybrid environments, strongly rely on sturdy security, sticking to regulation, and identity federation. The purpose of this section is to look closely at these themes to point out their key features and guidelines.

5.1 Security plays a major role in modern IT infrastructure.

Every infrastructure needs to be secure from the start. Because cloud services and AI are now widely used, the area that cybercriminals can target has grown bigger, so better security approaches are required such as:

Zero Trust means that every access request, inside or outside the original network, is double-checked for identity instead of trusting the outer network perimeter.

a) Both when it is not in use and while traveling, using strong encryption keeps your sensitive data safe from others.

b) Breaking the network into individual parts to block the spread of attacks if a breach happens.

c) Using AI for Security Making sure to spot unusual behavior in the network as soon as it happens.

5.2 The second important consideration relates to compliance with laws and regulations.

A lot of organizations that manage healthcare or financial data are required to follow industry regulations and standards such as GDPR, HIPAA, and SOC 2.

a) Cloud providers give tools that continually review your setup to ensure it remains compliant.

b) Maintaining logs is very important for conducting forensic investigations and for auditing purposes.

c) Where Data is Kept: Organizations are required to make sure that data is kept and processed according to laws regarding country or region.

5.3 Identity Federation and Access Management

It is not easy to handle identities and access within hybrid or multi-cloud systems. With identity federation, people can access many systems using only one set of login credentials.

With Single Sign-On (SSO), people don't have to enter their usernames and passwords repeatedly and can access many services easily with one login.

OAuth, SAML, and OpenID Connect services make it possible for people to use their identities anywhere within a federation.

With RBAC and ABAC, permissions are given only when someone needs them for their role or attributes, this helps to enforce least privilege.

5.4 Emerging Trends in Security and Identity Federation

a) More and more often, people are opting to use biometrics or physical devices to authenticate rather than passwords.

b) Giving individuals the ability to manage their digital identities on distributed ledgers is called Decentralized Identity (DID).

c) Using AI, security automation can automatically predict, find, and deal with threats.

6. Use Cases and Industry Applications

6.1 Real-Time Analytics Across AWS and GCP

Cross-cloud analytics allows companies to quickly analyze and respond to constant data by using the top services from both AWS and GCP at the same time. Usually, Google Cloud receives events using Pub/Sub, but AWS handles the processing and transformation with Kinesis and Lambda. With the help of this model, data from different systems can be combined smoothly to activate dashboards, alerts, and AI models almost in real-time.

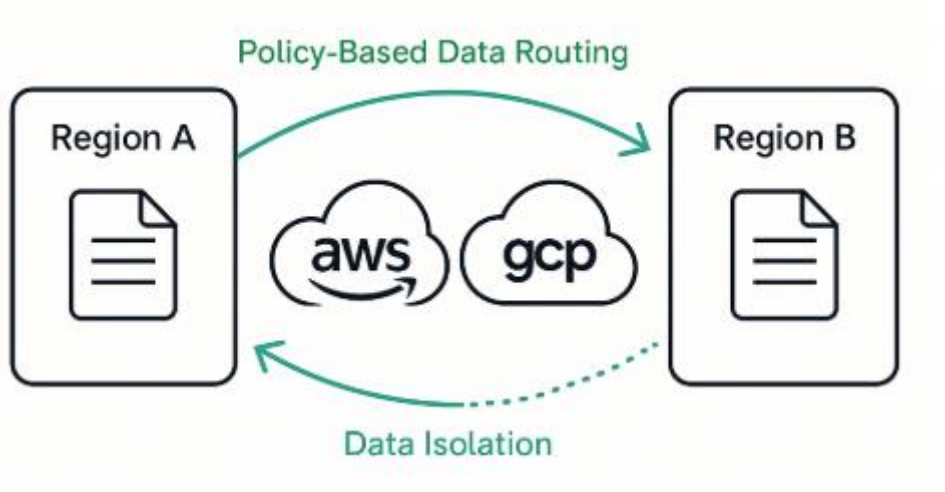
The main advantage here is that GCP's advanced machine learning services in Vertex AI can be used at the same time as AWS's strong data lake, created with S3 and Lake Formation. By connecting Looker (GCP) and Amazon QuickSight (AWS) with federated query layers, you can get useful information from various places in your data.

Retailers and banks worldwide have chosen this approach to support all types of customer engagements and detect fraud. In real time, companies can alter the amount of inventory, respond to users' actions, and

identify any unusual activities in different regions. Data pipelines built with Apache Beam and Apache Airflow are meant to help guarantee both high availability and the ability to scale the system.

All in all, using real-time cross-cloud analytics makes it possible to make better decisions with little compromise. If you apply strong identity management and observability solutions, this approach results in a single, lively, and speedy data infrastructure for digital transformation.

Fig 3: Compliance and Data Residency Management in Cross-Cloud Architectures



Showing the processes for routing policy-based data and setting data isolation regulations in AWS and GCP across Regions A and B to keep data within each region in order to comply with legal regulations.

6.2 Multi-Cloud Disaster Recovery and Data Redundancy

Having multiple disaster recovery plans and extra data copies is crucial for the continuity of businesses whose data is essential. Spreading their systems among AWS and GCP can help businesses avoid dangers linked to disruptions or breaches specific to one provider or data corruption.

Usually, a strong DR setup across clouds uses Velero, gsutil, or Veeam to replicate data between Amazon S3 and Google Cloud Storage. Some organizations may choose to download data from GCP and store them in AWS S3 Glacier or Backup to ensure that they are still backed up in a cold storage. If the primary system goes offline, there is an almost instant backup available from another cloud at little to no loss of function.

Besides just storing data, a compute failover strategy duplicates containerized services or VMs using Kubernetes, having different clusters on GCP and AWS connected together. Through Terraform infrastructure as code, there is a consistency in deployment between the development and test environments. With this plan in place, recovery is quick and human intervention is reduced.

Furthermore, organizations set up policies that direct traffic from users to different cloud areas or providers using AWS Route 53 and Google Cloud DNS. Not only does it ensure services are always available, but is also optimized based on location for apps that need fast response.

Organizations are now required by regulators to show that their business can continue and that their data is protected. A well-prepared plan for cloud DR meets auditing standards and boosts customers' trust. Such architectures are especially important for healthcare and finance since any problems could have serious legal, ethical, or financial outcomes.

In short, using AWS and GCP as part of a multi-cloud DR strategy avoids operational problems, checks regulatory requirements, and helps maintain 24/7 service wherever needed globally.

6.3 Compliance and Data Residency Management

Having proper compliance and choosing where data will reside are crucial issues in data engineering for the cloud, mainly in areas regulated by strict laws such as finance, healthcare, and governance. If a business links AWS and GCP, it should ensure that rules of data security, like GDPR, HIPAA, and laws that direct data should be uniformly applied in both cloud environments.

It can be difficult in using multiple clouds to make sure sensitive data is kept within the right boundaries. To stop data from going across regions without permission, policies are put in place to guide the routing of all traffic. Both AWS and GCP make it possible to set up specific data management rules by offering tools including AWS Lake Formation with region-based access and Google Cloud's Assured Workloads.

It is shown in the figure that data gathered from the European Union enters GCP, where it is both collected and processed on-site. At the same time, shorter data outside GCP's EU storage (technical information) is made available to other AWS systems, located in the United States. The choice of using two cloud services enables regulatory conformity and also social media marketing benefits derived from each service.

Encryption should not be ignored. Both allow for managing customer encryption keys and KMS, which makes it easier to oversee who can view data and the logs of those actions. RBAC and identity federation with services from AWS IAM and Google Cloud IAM make sure that just authorized people and systems can access confidential data.

Unified logging and monitoring should be maintained across all the clouds. When AI or SIEM systems are mixed with AWS CloudTrail and Google Cloud's Operations Suite, users can monitor data activities at all times and ensure they remain compliant.

All in all, working with data in different clouds calls for a well-designed infrastructure, automatic shielding, and teamwork. If implemented suitably, this puts regulation, creativity, and speed in harmony.

Conclusion

As enterprise data architectures switch to using multiple clouds, data engineering across clouds has become very important for organizations aiming to be flexible, work on different platforms, and conduct analytics on a large-scale. This article shows how you can use Amazon Web Services (AWS) and Google Cloud Platform (GCP) to build flow of data that goes past the borders of single clouds.

Thanks to AWS's secure data transfer and data processing technology (Amazon S3, Glue, EMR), as well as GCP's powerful query and analytics systems (BigQuery and Dataflow), engineering teams may create pipelines that are highly reliable, perform well, are cost-effective, and work with any vendor. Thanks to hybrid cloud, firms can make use of leading cloud services between different providers, thus reducing the danger of being dependent on a single service and achieving greater flexibility in their IT set-up.

In order to fully use cross-cloud data engineering across their systems, companies should use the following guidelines:

a) Make sure the company's business and technology goals are clear so that the architecture design supports real-time analytics, machine learning, or compliance with rules.
b) Resort to using Apache Airflow, Apache Beam, or dbt, and other frameworks, since these are cloud-agnostic and help avoid tight links with a specific platform.
c) Maintain proper boundaries in data and compliance management so that all identity access practices, encryption policies, and regulation compliance is equivalent in every environment.

d) Ensure you use common tools that gather data for monitoring, logging, and dealing with issues across the clouds to make systems trustworthy and maintain good data quality.
e) Start with a small, easily made illustration to make sure the system works and fits as planned before carrying out more important functions.
f) Using such a design, companies can access untapped business benefits, plan budgets for data infrastructure wisely, and ready their analytical systems for any changes.

References:

1) Jiang, F., Ferriter, K., & Castillo, C. (2020, April). A cloud-agnostic framework to enable cost-aware scheduling of applications in a multi-cloud environment. In *NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium* (pp. 1-9). IEEE. https://doi.org/10.1109/NOMS47738.2020.9110325

2) Hummer, W., Muthusamy, V., Rausch, T., Dube, P., El Maghraoui, K., Murthi, A., & Oum, P. (2019, June). Modelops: Cloud-based lifecycle management for reliable and trusted ai. In *2019 IEEE International Conference on Cloud Engineering (IC2E)* (pp. 113-120). IEEE. https://doi.org/10.1109/IC2E.2019.00025

3) Bergmayr, A., Breitenbücher, U., Ferry, N., Rossini, A., Solberg, A., Wimmer, M., ... & Leymann, F. (2018). A systematic review of cloud modeling languages. *ACM Computing Surveys (CSUR)*, *51*(1), 1-38.

4) Torkura, K. A., Sukmana, M. I., Cheng, F., & Meinel, C. (2019, September). Slingshot-automated threat detection and incident response in multi cloud storage systems. In *2019 IEEE 18th International Symposium on Network Computing and Applications (NCA)* (pp. 1-5). IEEE. https://doi.org/10.1109/NCA.2019.8935040

5) Dubuc, T., Stahl, F., & Roesch, E. B. (2020). Mapping the big data landscape: technologies, platforms and paradigms for real-time analytics of data streams. *IEEE Access*, *9*, 15351-15374. https://doi.org/10.1109/ACCESS.2020.3046132

6) Ahmed, A. E., Heldenbrand, J., Asmann, Y., Fadlelmola, F. M., Katz, D. S., Kendig, K., ... & Mainzer, L. S. (2019). Managing genomic variant calling workflows with Swift/T. *PloS one*, *14*(7), e0211608.

7) Natu, V., & Ghosh, R. (2019, January). EasyDist: An End-to-End distributed deep learning tool for cloud. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data* (pp. 265-268).

8) Nesztler, T., & Georgescu, M. (2019). Advances And Challenges For Scalable Cloud-based Infrastructure For Building Data Analysis And Simulation. *Building Simulation 2019*, *16*, 2721-2728.

9) Šipek, M., Muharemagić, D., Mihaljević, B., & Radovan, A. (2020, September). Enhancing performance of cloud-based software applications with GraalVM and Quarkus. In *2020 43rd international convention on information, communication and electronic technology (MIPRO)* (pp. 1746-1751). IEEE.

10) Barisits, M., Beermann, T., Berghaus, F., Bockelman, B., Bogado, J., Cameron, D., ... & Wegner, T. (2019). Rucio: Scientific data management. *Computing and Software for Big Science*, *3*, 1-19.

11) Watada, J., Roy, A., Kadikar, R., Pham, H., & Xu, B. (2019). Emerging trends, techniques and open issues of containerization: A review. *IEEE Access*, *7*, 152443-152472. https://doi.org/10.1109/ACCESS.2019.2945930

12) Leitner, P., Wittern, E., Spillner, J., & Hummer, W. (2019). A mixed-method empirical study of Function-as-a-Service software development in industrial practice. *Journal of Systems and Software*, *149*, 340-359.

13) Jiang, F. (2020). *On Improving Efficiency of Data-Intensive Applications in Geo-Distributed Environments* (Doctoral dissertation, The University of North Carolina at Chapel Hill).

14) Fracchia, C. (2020). Secure and Scalable Collection of Biomedical Data for Machine Learning Applications. In *Artificial Neural Networks* (pp. 317-336). New York, NY: Springer US.

15) Torkura, K. A., Sukmana, M. I., Strauss, T., Graupner, H., Cheng, F., & Meinel, C. (2018, November). Csbauditor: Proactive security risk analysis for cloud storage broker systems. In *2018 IEEE 17th International Symposium on Network Computing and Applications (NCA)* (pp. 1-10). IEEE. https://doi.org/10.1109/NCA.2018.8548329