6 best practices for cloud data integration

This guide describes the challenges involved in achieving cloud interoperation and provides best practices and solutions, including log-based change data capture.

Fivetran

Cloud adoption is accelerating. Given its central role in the enterprise, IDC forecasts "whole cloud" spending — which includes total worldwide spending on cloud services, the hardware and software components underpinning the cloud supply chain, and the professional/managed services opportunities around cloud services — will surpass $1.3 trillion by 2025.

One reason for the growth: Organizations often adopt cloud technologies for analytic use cases. The cloud enables access to robust analytics services at scale using a pay-for-use scheme. Importantly, companies can avoid significant upfront investments. Instead of building a configuration in a data center, they can experiment with different options available to them — and quickly scale up or down with the cloud provider's consumption-based model. Cloud will continue to play an even greater role as on-premises systems reach end-of-life and businesses focus on delivering greater efficiency, flexibility and faster innovation.

Most large organizations adopting cloud solutions will, at least initially, run a hybrid cloud. The term hybrid cloud refers to two types of deployments. Hybrid clouds can consist of multiple cloud-based solutions from one or more cloud vendors. Alternatively, they can combine cloud solutions with on-premises systems.

Two factors to consider before adopting the hybrid cloud

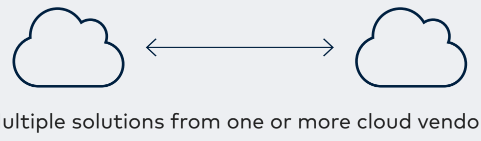
Hybrid cloud is different from multiple geographically separated data centers for two reasons:

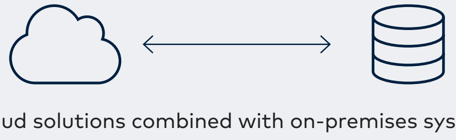
1. You have less control over your deployment. You must trust that the cloud provider's solution is equally (if not more) highly available and secure than your current choice.

2. The cloud offers a variety of deployment models. For example, in the pure on-premises world, you manage your databases. In a hybrid cloud environment, you can manage your databases or sign up to use a DBaaS, or database-as-a-service. This follows a similar path laid out by Salesforce when they showed the world they could access their CRM system as a SaaS platform.

In the cloud, you have more options relative to the on-prem world, each with benefits and disadvantages. With cloud access, you are more likely to consider those alternatives.

Two common types of hybrid cloud environments


Multiple solutions from one or more cloud vendors


Cloud solutions combined with on-premises systems

Why go hybrid?

Organizations often adopt a hybrid cloud strategy when deploying their first analytical solution to the cloud. The cloud combines sheer infinite scalability with consumption-based pricing. These are desirable attributes for analytical environments that require scalability and can be very expensive when scaled for maximum capacity over an extended period.

An organization's primary business process is generally supported by one or more operational systems. Because these environments are crucial, businesses must have optimum access and continuous availability.

Deployments and systems can be quite complex. Additionally, because these operational systems are core to the organization's primary business processes, their data should be included in the analytical environment.

Migrations of operational systems require considerable testing to ensure that the cloud environment has similar or better performance than the existing, often on-prem, systems. These systems must have the same or higher levels of availability and data must be secure. Environments also have to perform well under heavy user load.

The IT team must perform further testing to ensure everything works properly after the migration. Integrations must be rewritten to work with the new system, too. Depending on the criticality of the system, organizations may also need a fallback for some time after the initial migration.

Hybrid cloud data integration best practices

Hybrid cloud integration is not easy. Common challenges and considerations during data integration in hybrid environments include:

1. Impact on the data sources
2. Network efficiency
3. Data security
4. Compatibility across heterogeneous environments
5. Fallback
6. Latency

The following six best practices will ensure a smooth data integration that effectively delivers high performance, security and availability across all heterogeneous data sources.

▷ 1. Learn the impact on operations

Organizations need the data in operational systems that drive the business for consolidatedanalytical environments.

For example, if you are a manufacturer who wants to sell your customers preventive maintenance solutions, you need access to factory data. What items are planned to be produced when and where? How do you get them to your customers? Do they need one of your experts to help with installation?

Operational systems drive the primary business process. As much as we want access to data, we don't want these systems to slow down. Therefore, we must find a solution that captures the changes going into your operational system with minimal overhead.

For many self-hosted applications, consider a database-level change data capture (CDC) solution, of which log-based CDC is widely considered the least intrusive. Because critical systems contain the most important data to help drive decisions, real-time access to this data is required to be more competitive. Log-based CDC handles the highest volumes of change data in real-time — enabling organizations to make informed, data-driven decisions more quickly.

Depending on where you are in your hybrid cloud adoption, you may consider alternatives for your current deployment model. For example, if you currently host your organization's ERP in an on-prem data center, you may consider cloud-hosted or software-as-a-service options. Will your future deployment option provide a similar level of flexibility as your current deployment model? As you optimize for no operational impact on your environment today, will you be able to leverage the same solution in the future? If not, do you have alternative options that address your data integration requirements?

Also, you may need the primary system to synchronize with the new or old configuration during the migration to allow for testing and provide a fallback option.

▷ 2. Increase network efficiency

With widespread access to high-speed connectivity, is network efficiency still relevant? Consider these aspects:

1. Network bandwidth is finite. When bandwidth reaches its peak, you cannot use more of it. When you add more load on a network that has reached capacity, you may increase latency. Higher latency results in lower bandwidth because network transfers require confirmation.

2. Cloud ingress is free for most cloud providers. However, egress typically is not. If you transfer data out of a cloud — and remember hybrid cloud may involve multiple clouds — you pay less if you transfer less data.

To improve data transfer rates beyond maximum bandwidth, you can use compression: Transfer fewer data based on an agreed compression algorithm. If you can achieve 5x compression, then you effectively magnified your bandwidth 5x. And cloud egress costs are lower by 5x.

Likewise, using CDC relative to full extracts will limit the required bandwidth you need. Even better: Filter the data before it is sent across the wire. For example, use a filter condition on data retrieval or an agent to identify the required changes.

Lastly, consider the efficiency of network communication. Sending fewer large data blocks is better than transferring many small data blocks. Across a wide area network (WAN), you want to avoid a chatty communication protocol.

▷ 3. Don't assume all communication is behind a firewall

Traditionally, data transfers between systems have taken place within the confines of data centers. As organizations built out disaster recovery environments, multiple data centers were implemented with direct connectivity among them, still behind a corporate firewall.

As we adopt cloud technologies, we can no longer assume that all communication is behind a firewall managed by our network team. Of course, there are PrivateLink and Direct Connect options with cloud providers. However, many SaaS platforms simply fall outside of these.

Your organization may not be willing or able to invest in Direct Connect to hook up on-premises systems with the cloud environments you have the ability to access. As a result, a security condition may develop as communication is exposed.

The first consideration is to use encryption whenever possible. Given your organization has no control over the end-to-end network connection, you want to look for application-level encryption. Ensure the technology you use encrypts data using TLS 1.2 or higher.

A second consideration is to lock down firewalls. SaaS platforms may reach out to pull data. Lock down the firewall to just the IP addresses the vendor uses. Or, even better, use a solution that reaches out from your (on-prem) environment into the cloud. From there, based on the stateful properties of a firewall, bi-directional communication can then begin.

Finally, validate the vendor's security certifications. Organizations send out long security surveys to determine vendor approval. Many questions receive default answers based on the deployment model, or through industry certifications.

Look for a cloud vendor/provider who has SOC2 type 2 and/or ISO 27001 certifications.

▷ 4. Remember everything changes

While you may appreciate the configuration your organization uses, it will change — systems wear down and must get replaced. Software solutions become incompatible with infrastructure. Your organization may decide to use different solutions. Or, in the context of hybrid cloud, you may decide to change the deployment configuration.
Most hybrid cloud deployments are mixed environments with a range of heterogeneous on-premises systems and diverse cloud services. Obviously, any data integration solution must work in the initial environment.

The environments are also likely to change over time. The cloud provides flexibility because of its technology or services and its "pay for use" pricing. It's relatively easy to switch deployment platforms quickly, and a platform that works well today may not be tomorrow's platform of choice. Organizations continually evaluate solutions, and available technology options often shift. An application running on a relational database today, deployed in the cloud or using a database as a service, may one day be replaced by a SaaS solution.

A best practice is to use a data integration technology that ensures compatibility across the broadest possible array of databases, file systems, applications, platforms and cloud services — including IaaS, PaaS and SaaS. Such a solution delivers a wide range of deployment choices, and offers the flexibility to make changes in a heterogeneous environment. Since the cloud provides many options to store data, the ability to quickly add destinations to an existing data integration flow is an added bonus.

CUSTOMER STORY

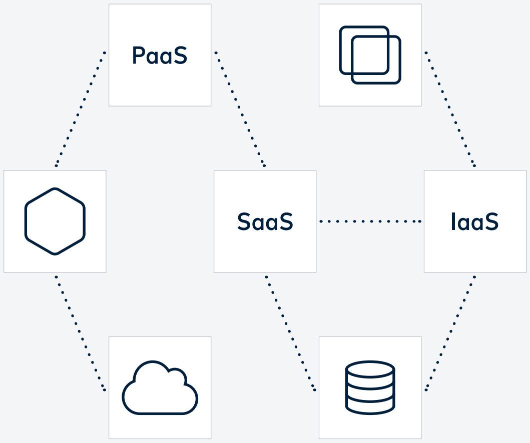
An industry leader in water processing technologies went through such a transition, partly driven by a spinout. Like many companies around the world, SAP ECC was its core ERP system. It used SAP data in Oracle databases for forecasting of purchasing, materials and inventory to assist in improving business decisions.

The enterprise faced a number of challenges:

1. Bulk extracts put a massive load on the source SAP transactional database.
2. Data needed to be fresher; ETL latency was too long and only worked for historical reporting, not real-time analytics.
3. Detecting deletes was labor-intensive and inconsistent resulting in some bad data.
4. Analytics on Oracle was slow because tables contained a very large number of rows and columns.

An on-premises data warehouse solution, limited to a single node, was replaced with a scalable cloud-based relational data warehouse, AWS Redshift. An additional cloud-based database as a service destination was added for operational reporting. The on-premises SAP system running on Oracle was moved into a cloud-hosted environment on AWS. The data warehouse was augmented with a data lake hosted on cloud storage, replacing a lot of the functions the data warehouse previously performed.

One of the few technologies that did not change throughout the transitions was the data replication between sources and destinations. Fivetran's high-volume data replication solution provided the flexibility to support different sources and destinations to meet the customer's needs throughout the transitions



5. Consider bi-directional data movement

You may have started your hybrid cloud journey with one or more analytical use cases. Over time you consider what to do with your operational environments. The question is not whether you will run your operational systems in the cloud. The question is: When will you decide to do so?

Any migration is daunting, especially one that affects your organization's primary business process. What's going to happen when all users switch to the new environment? And, if things don't work out, what's your fallback option?

Consider bi-directional data movement. You probably don't need active/active replication because most applications are not prepared to run in active/active mode. However, running active/passive replication is a powerful way to mitigate data loss if a fallback is required. Instead of asking users to redo their work or re-run routines that were processed already, you replicate the data to the source.

If the migration is not successful, then you switch back. Data processing continues with minimal disruption. How long you keep the old system around is a risk assessment. Some organizations want to see initial successes to feel comfortable about not needing the old system. Others want to see at least a couple of months of successful processing before giving up the fallback option.

6. Low latency is a must-have during migration

Business requirements will determine the maximum allowable latency. Cloud-based environments are built to be available 24x7, and users (as well as customers) have become accustomed to instant access to information. These combined factors drive organizations to look for near real-time or continuous data integration solutions.

Consider the competitive differentiation you can achieve with consolidated data available for analytics closer to real-time. A solution you sell to your customers may become more valuable. Your team may become better equipped to identify fraudulent behavior. You may have opportunities to save costs simply by reacting more quickly.

During the data migration, low latency is a must. If a critical operational system does not meet expectations post-migration, you want to lose no time and resume processing on the old environment. However, it must be up to date with the latest changes.

How to continually integrate data between on-prem and cloud for real-time analysis

As organizations migrate to the cloud, they'll likely need to operate — at least for a time — in a hybrid environment.

Whether data arrives from a SaaS platform or directly from a database, change data capture methodology can enable near real-time updates to the analytical environment. Log-based CDC — reading changes from a database transaction log — is widely considered the least-intrusive method to retrieve database changes.

Fivetran offers CDC as a feature for most of our connectors to applications — and all connectors to databases. After the initial sync of your historical data, Fivetran performs incremental updates of any new or modified data from your source system. We use your database's native transaction log during incremental syncs to request only the data that has changed since our last sync, including deletes. Each database uses a different change capture mechanism.

During incremental syncs, Fivetran maintains an internal set of progress cursors that allow us to track the exact point where our last successful sync left off. If there is an interruption in your service (such as your destination going down), we automatically resume syncing where it was left off — even hours or days later, as long as log data is still present. You can also track deletions to view your archived records.

The hybrid cloud is a reality for many companies undergoing digital transformation.

If you want to learn more about our approach to cloud data integration, sign up for a 14-day free trial and test our system for yourself.

Fivetran is the global leader in modern data integration. Our mission is to make access to data as simple and reliable as electricity. Built for the cloud, Fivetran enables data teams to effortlessly centralize and transform data from hundreds of SaaS and on-prem data sources into high-performance cloud destinations. Fast-moving startups to the world's largest companies use Fivetran to accelerate modern analytics and operational efficiency, fueling data-driven business growth. For more info, visit **Fivetran.com**.