Multi-cloud for data analytics.

Seizing the opportunity.
Introduction

The days of easily finding your cloud partner are over. Previously, early pioneers signed up to their chosen cloud provider – almost always Amazon – and within minutes the relationship was sealed, and they were ready to go. In the cloud’s initial heyday, very few companies had an enterprise-wide cloud strategy with many of the first moves being department-led.

Today this has changed significantly. Virtually every organization now has at least part of its computing system in some sort of cloud. CIOs are now fully aware of the potential of the technology and are demanding more and more innovation as a result.

At the same time, we’re also witnessing a more competitive scene, where providers are making a big point to differentiate their offerings, with Microsoft and Google slowly eating away at Amazon’s dominance in the space. This has prompted a change of thinking in many companies. Businesses are no longer signing up for one cloud provider alone but turning towards alternative providers to leverage a ‘best-in-breed’ solution that will better meet their IT strategy and requirements.

The shift to the cloud offers not just transformational flexibility in resources; it also gives organizations the opportunity to take control of their enterprise architecture and data, helping define a strategy according to business needs. Organizations once faced limited choice and control. There was a handful of enterprise application vendors, and customers largely did things their way. Similarly, there was a select group of vendors capable of handling enterprise data warehousing and company-wide business intelligence and analytics. Moving from the main enterprise application or data warehousing vendors was costly, risky and slow.

In migrating to the cloud, organizations can rethink these assumptions. The variety of vendors offering applications, data warehousing and analytics is booming. This is because they base their offers on open APIs and standards, allowing the user to define their strategy and bring in vendors to support it – without being locked into a single one.

As such, today the question for many organizations is no longer “shall we move to the cloud?” but “shall we go multi-cloud?” In this white paper, we will explore multi-cloud approaches and how Snowplow Insights can support you in gaining the strongest business advantage.
The dominance of multi-cloud

According to Flexera’s 2019 State of the Cloud survey, 84% of enterprises now have a multi-cloud strategy – a sign of how many companies have changed their approach. This is a view that has been backed up by Gartner. The research company found a similar number of organizations – 81% – are choosing to work with two or more public cloud providers.

The dominance of megavendors in the public cloud services market is driving enterprise buyers to choose multiple cloud providers, says Michael Warrilow, VP analyst at Gartner. “Most organizations adopt a multi-cloud strategy out of a desire to avoid vendor lock-in or to take advantage of best-in-breed solution,” he says. “We expect that most large organizations will continue to willfully pursue this approach.”

Multi-cloud – four reasons why

There are several reasons for the rapid adoption of multi-cloud:

Avoiding vendor lock-in

One of the most widely cited drivers of multi-cloud is to avoid becoming locked in to a particular cloud provider’s infrastructure, add-on services and pricing model. Firstly, a multi-cloud strategy ensures an easier migration from one provider to another as you can design your architecture with portability in mind. Secondly, multi-cloud gives a company the opportunity to shift spend between clouds – to optimize costs, to take advantage of differentiated services or in response to differing service levels that they find the clouds provide when they actually use them.

Geography

A multinational with offices in different parts of the world may well find one provider is stronger in a particular region, either in terms of support or latency.

Meanwhile, larger organizations are building their own content delivery networks (CDNs) to secure availability and performance by locating services closer to end-users. The approach helps avoid latency created by relying on points of presence that are selected by the cloud provider. The CDN approach helps mitigate any geographic downside to a multi-cloud strategy.

Governance

As we have seen with GDPR, the imposition of a new regulatory regime can have a profound impact on the way that companies operate, especially how they store and manage data. It’s also the case that many public sector organizations want their data stored locally – in some countries this is a legal requirement and in others, it’s a preferred option.

Resilience

Multi-cloud deployment is disaster recovery; having multiple providers can help organizations to protect their data better. One of the dangers of putting all your company’s data into one cloud provider is what happens when that provider goes bust? This happened a few years ago when a cloud storage company called Nirvanix went into administration and its customers found that they had just a short time to retrieve all their data and find another place to store it.

---


Multi-cloud offers the best of all worlds in data tools

The public cloud fundamentally offers two kinds of services: commoditized services such as underlying compute and storage and differentiated services, which are higher up in the application stack. These look different across the different clouds, are higher margin and are often used by cloud providers as a “gateway” to get companies to start using their products.

Notably, data tools fall into this second category. The three giant cloud providers offer a raft of tooling to help organizations move, manage and analyze their data according to the application needs. Since cloud solutions commonly have a strong data element, a multi-cloud strategy offers users the best tools from the main cloud providers, as well as the best of the independent and open-source tools on the market. The differentiators are in the details – and a multi-cloud strategy facilitates the use of the data processing tools the user wants, across all cloud providers if required or desired.

For example, if you look at AWS and its data warehouse offering, Redshift, it is fundamentally different from Google’s BigQuery in several ways that a user may want to consider:

- Elastic compute and storage in BigQuery (versus Redshift, where you pay for a fixed number of nodes, and compute and storage are tightly coupled). Redshift costs are more predictable, but scaling Redshift to handle enormous data volumes is more expensive – especially if query patterns are not constant during the day.
- Support for complex data structures including nesting. BigQuery has much better support for complicated data structures than Redshift.
- BigQuery offers users the ability to run machine learning on data via a SQL interface using “BigQuery ML”. Redshift offers no equivalent feature.

Similarly, the toolset for processing data in-stream is very different across the different clouds. AWS’s Lambda offering is broadly comparable to GCP’s Cloud Functions, but there is no real AWS equivalent of Cloud Dataflow, with users on AWS who want to write more involved streaming jobs are forced to use Spark or Flink or AWS EMR.

The point is, a multi-cloud strategy can allow organizations to run applications in one cloud environment while doing data warehousing in another, and big data discovery and machine learning in a third. Additionally, and crucially, a multi-cloud strategy gives a company the ability to experiment, for example with machine learning across two different cloud environments, and to easily switch from one to another if the offering on one cloud moves in a direction that is a better fit. The freedom to shift workloads and experiment is valuable on an ongoing basis.
Multi-cloud mixing is complex

It’s important not to get carried away. Multi-cloud is not a ready-made solution to every problem. There are definite challenges that need to be addressed. For example, a company may have already invested heavily in AWS and may not want to complicate matters by adopting another provider. In particular, the workload involved in moving from a long-established partner can be financially costly and resource-heavy. According to the Flexera survey mentioned previously, optimizing cloud costs is seen as the biggest challenge facing cloud-using CIOs, who are also grappling with concerns about managing software licenses in the cloud.

The other barrier to multi-cloud is a possible skill shortage. The move to cloud can be difficult, and the availability of talent is really critical: both for recruiting and retaining your internal team, but also having external consultants and customer engineers at the cloud provider available to support leveraging the cloud. The genuine insufficiency of skilled and experienced cloud engineers and architects, only heightened with multi-cloud deployment where there can be an extreme lack of cross-cloud training, makes this an acute concern. Put simply, cloud experts are hard enough to find, and people with expertise across more than one cloud are a rare breed. This means that it’s more difficult for a company to take a fully supported multi-cloud approach without significant investment upfront.

The skill shortage is definitely going to have some detrimental effects on multi-cloud deployment. Companies need to ensure that they have planned carefully for the disruption that this can cause.
Approaches to multi-cloud

Given the value in being multi-cloud, and considering the cost, a number of evolving approaches make it easier for companies to manage multi-cloud deployments. The focus of these efforts is to provide a standardized set of technologies (e.g., APIs) that work consistently across different cloud environments. By working with these tools, companies can more effectively work across more than one cloud.

One approach is using software containers as a framework for managing the deployment, running and scaling of applications. Many organizations are deploying containers to help run their applications as they offer a degree of platform independence, speedier deployment, lower overhead on the network and better scalability.

The other problem with this approach, while it will provide the ability to run on multiple clouds independently, is that it will not take advantage of the native services that each cloud provider offers and where the most value can be created. For example, moving to Google Cloud Platform can make it easier to take advantage of GCP’s TensorFlow support and dedicated offerings – these services can create more value than can be saved running separate services that are fully containerized. Technologies like Kubernetes are great for managing commoditized services across clouds, but not so great at the differentiated ones, so Kubernetes may not be the best option for multi-cloud on its own.

Another approach is a solution like Hashicorp, which offers a range of technologies to empower devops to deploy and manage infrastructure that works across multiple cloud providers.

And yet another approach is Snowplow Insights, which offers companies the ability to run customer data pipelines in AWS, GCP or multi-cloud (AWS, GCP and Azure).

Illustration: An example analytics pipeline on GCP

The other problem with this approach, while it will provide the ability to run on multiple clouds independently, is that it will not take advantage of the native services that each cloud provider offers and where the most value can be created. For example, moving to Google Cloud Platform can make it easier to take advantage of GCP’s TensorFlow support and dedicated offerings – these services can create more value than can be saved running separate services that are fully containerized. Technologies like Kubernetes are great for managing commoditized services across clouds, but not so great at the differentiated ones, so Kubernetes may not be the best option for multi-cloud on its own.

Another approach is a solution like Hashicorp, which offers a range of technologies to empower devops to deploy and manage infrastructure that works across multiple cloud providers.

And yet another approach is Snowplow Insights, which offers companies the ability to run customer data pipelines in AWS, GCP or multi-cloud (AWS, GCP and Azure).

Illustration: An example analytics pipeline on GCP

The other problem with this approach, while it will provide the ability to run on multiple clouds independently, is that it will not take advantage of the native services that each cloud provider offers and where the most value can be created. For example, moving to Google Cloud Platform can make it easier to take advantage of GCP’s TensorFlow support and dedicated offerings – these services can create more value than can be saved running separate services that are fully containerized. Technologies like Kubernetes are great for managing commoditized services across clouds, but not so great at the differentiated ones, so Kubernetes may not be the best option for multi-cloud on its own.

Another approach is a solution like Hashicorp, which offers a range of technologies to empower devops to deploy and manage infrastructure that works across multiple cloud providers.

And yet another approach is Snowplow Insights, which offers companies the ability to run customer data pipelines in AWS, GCP or multi-cloud (AWS, GCP and Azure).
The Snowplow approach

At Snowplow, our approach is to give customers the flexibility to work with their data across multiple clouds so they can take advantage of the differentiated offerings from each provider. Whether you have separate teams that want to access Snowplow data in different services across different clouds, have data or applications running in a different cloud from where you have your Snowplow pipeline set up, or want to migrate to another cloud and need to feed the Snowplow data to your other data sets, Snowplow supports your ability to pipe your data seamlessly across clouds in real time and supports:

• Real-time processing in each environment, streaming the data onto Kinesis in AWS, Pub/Sub in GCP and EventHub in Azure, to enable real-time and batch-based computation on the Snowplow data.
• Streaming the data into multiple, cloud-specific storage targets on each cloud, enabling our customers to take advantage of Redshift, Snowflake DB, Elastic and BigQuery data warehousing technologies.
• Streaming the data into AWS S3 and GCP Cloud Storage as data lake solutions.

Data teams – and the wider business – can benefit

It is only natural that businesses want to handle one of their most valuable assets – data – as effectively as possible.

“By using the right services, companies can make use of the many powerful programs available to them and draw on the potential of the cloud.”

We’ve seen how the use of a multi-cloud approach can benefit data analytics teams, freeing them to draw on the best technology for complex queries, but it’s not just about data teams. Other business units can benefit from a multi-cloud approach too.

All data teams will want to benefit from the use of cloud computing. By using the right services, companies can make use of the many powerful programs available to them and draw on the potential of the cloud: the scalability, the instant provisioning and the financial advantages – the capex accounting – will always be attractive.

Given the power multi-cloud puts at the fingertips of data teams it’s important that a cloud strategy is in place. It’s not then a matter for CIOs or CFOs but for the data professionals to pick the right option to find the information that they’re looking for.
About Snowplow

The Snowplow product is a real-time event data pipeline, built for data-informed businesses with data teams. Snowplow lets you track, contextualize, validate and model your customers’ behaviour across your entire digital estate. Your data is available in real-time and is delivered to your data warehouse of choice, where it can be used to power analytics and reporting, business-critical applications and machine-learning models. The Snowplow product is distributed as ‘private SaaS’, running in your own cloud environment and giving you complete ownership and control of your data.