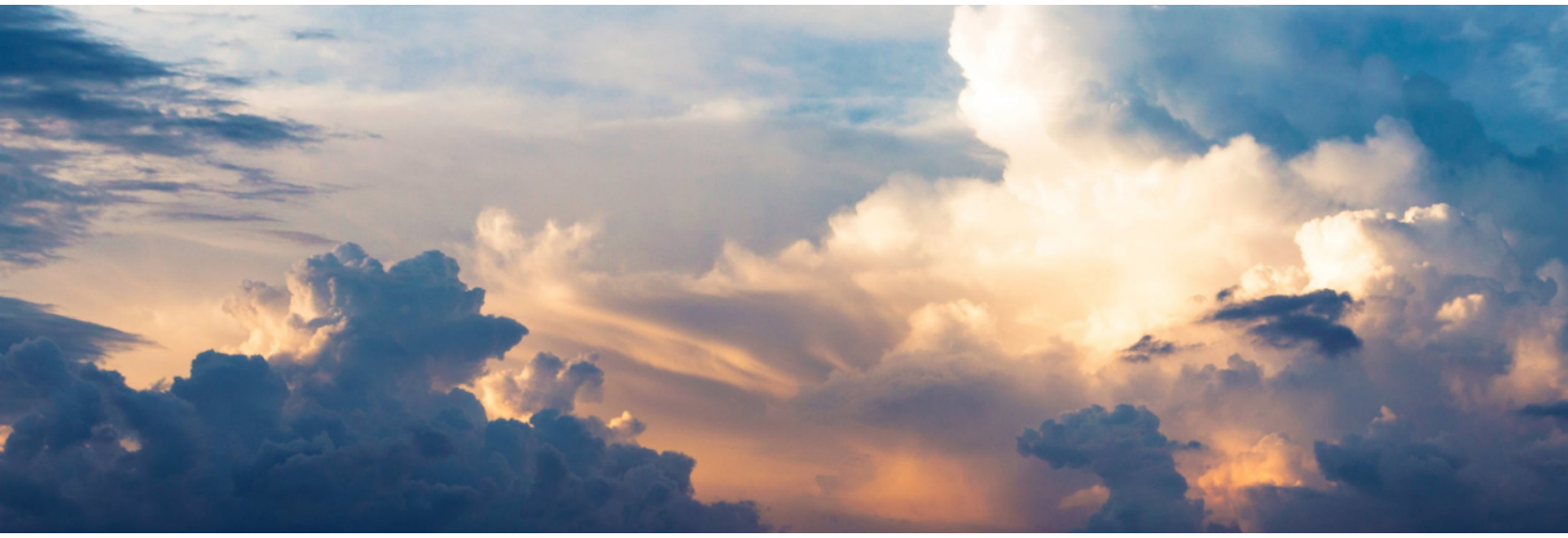
WHITE PAPER

StreamSets



Rearchitecting Data Operations for the Cloud

Introduction

It is rare to find an organization these days that is not looking to migrate at least some of their workload to the cloud. Buying and maintaining hardware remains one of the biggest line items in any IT budget and migrating to the cloud has obvious advantages when it comes to freeing up capital budgets and creating a more flexible business. But it is important to remember that while the hardware may be the most *expensive* part of your system, it is not actually the most *valuable* part. That would be the data! The reason that businesses invest in IT systems is for the data that they consume, manipulate and produce.

Organizations that wish to thrive in a cloud-based world would be well advised to take a step back and consider how they will guarantee the safe and reliable delivery of data across an increasingly fragmented collection of data centers over which they have limited control. At the same time, they need to consider how they will do so in a way that meets ongoing business needs and maintains compliance with an increasingly complex regulatory environment. In short, they need to take a DataOps approach to their cloud strategy.

This paper will introduce the idea of DataOps and specifically explore the role it plays in helping to deliver a successful cloud initiative. It will discuss the steps required to migrate to the cloud while staying ahead of compliance requests and how StreamSets can provide a helping hand.

The Shift to Microservices

Like actual clouds in the sky, an IT cloud deployment may look like a single highly-structured entity from a distance. But up close, things turn out to be much more fragmented and chaotic.

A typical cloud migration plan often starts with a single cloud provider, and it might be easy to imagine that one provider will simply replace your existing data center. But cloud migrations take time and cloud-based offerings evolve rapidly. Unless your total IT environment is small and static, it is unlikely that you will ever achieve 100% migration to just a *single* cloud provider. And it is even less likely that you will remain that way.

Core systems may be predominantly deployed on a primary cloud provider, but it is inevitable that evolving business requirements will introduce secondary cloud providers into the mix.

These services could be anything from a black-box SaaS offering to a secondary fully functional IaaS provider such as Amazon Web Services (AWS), MIcrosoft Azure, or Google Cloud Platform (GCP). And even within a single provider, virtual partitions can fracture your deployment into distinct and relatively isolated pieces.

The result is that instead of a single cohesive set of compute resources—analogous to your old data center—a cloud deployment turns out to be a constantly shifting and evolving collection of smaller environments. These independent services are typically separated by slower, sometimes less-reliable network connections and additional layers of security that must be negotiated.

At the same time, the microservices that live in the cloud are increasingly hungry for data. The increasing use of big data analytics and machine learning to unearth insights that help you win in the marketplace means that more data will need to be made available across service boundaries.

New requirements for the cloud means that having all runtime components interact via real-time connectivity is not realistic. Instead, architects and operators will need to look for ways to decouple runtime components via *near real-time* connectivity based on data streaming and localized data caches. Unlike an enterprise *data lake*, these local *data ponds* will need to be limited to the immediate needs of the microservice in order to optimize performance and costs. The problem is that the data needs of individual microservices will not remain static over time. Application elements and research projects will be created, moved and deprecated and each data pond will need to be adjusted, consuming precious resources and time.

Overcoming the fact that change is now a constant calls for new tools. Tools to create and manage the pipelines that feed batch and streaming data to these data ponds as well as enable accelerated development and testing. Organizations who are on top of this will automate self-service infrastructure that allows data scientists and project owners to request and pull the data they need, without the direct involvement of pipeline developers and other constrained specialists. Enter DataOps.

Introducing DataOps

DataOps is the application of DevOps practices to data management and data integration in order to reduce the cycle time of data analytics, performed with a focus on collaboration, automation and monitoring.

In the same way DevOps has helped companies deliver software quickly, with higher quality and with better responsiveness to market needs, DataOps provides these benefits to data.

In short, frequent ungoverned changes combined with a need for real-time data analysis requires that you integrate your dataflow design and operations into a continuous and agile process.

This dynamic that demands DataOps is ever presence in the cloud. As was mentioned above, the constant change of cloud service providers, the data services across them, and the microservices created in support of cloud operations can greatly hinder the success of any cloud migration initiative. DataOps ensures the continuous integration and continuous delivery (CI/CD) of data services, even in the face of constant change, and provides the operational visibility needed for a complex and dynamic multi-cloud architecture.

With Great Power, Comes Great Regulation

When it comes to cloud, in addition to controlling *who* can access sensitive data, there is also an increasing number of regulations that limit *where* you can transmit such data. Rising public anxiety about how data is being used and protected along with regulations to prevent abuse of personal data are amplified in the cloud.

This can have a serious impact on how you develop and deploy your cloud architecture. Application developers (who typically have not graduated from law school) are not well qualified to interpret and apply the patchwork of government regulations that might apply to their system in any given moment.

Even where developers do succeed in their immediate task, there are still long-term risks to the company. Operators, who do not understand application internals, want to move applications and data quickly and without requiring development resources. In doing so they can easily create hard to detect, but potentially expensive, compliance violations. Furthermore, the evolving nature of these regulations means that small changes within any of the relevant regulatory jurisdictions can impact hundreds, or even thousands, of data pipelines, leading to a large scale rework and retesting effort, potentially across multiple, disparate development teams.

To address this requirement you should abstract data compliance policies so they can be designed separately and decoupled from the data pipelines and then applied to pipelines at runtime. This allows regulatory compliance activities to be centralized in a team of well-trained specialists who are responsible for developing and maintaining a consolidated set of data governance policies. When regulatory changes occur, this compliance team can create a new version of the relevant policies and release them to the application teams, for whom they become part of the development environment. This greatly reduces the overall effort and amount of cross-team coordination required to keep your enterprise in compliance.

Taking Control of your DataOps

Continuously running streaming processes can be a lot more challenging to monitor and manage than batch jobs, and the sheer number of batch and streaming jobs combined can quickly outstrip the practical limits of traditional operational tools.

Aftermarket tools from your cloud provider and 3rd parties provide siloed and incomplete solutions for cloud adoption. Using these types of tools to resolve data communication problems is analogous to trying to understand a two-person conversation by listening to just one party, then going back and listening to the other. Synchronizing the two halves of the conversation in your head takes time. And what do you do when they don't agree with one another?

Additionally these aftermarket tools rely heavily on external agents and gateway processes communicating under the covers across a menagerie of protocols and ports. In the setting of a on-prem data center this is complexity that can be managed, but things become a lot shakier when you start dealing with cloudlets. Disparate security zones, obfuscated network layers, limited access to storage and all the other limitations one can find in a public multi-tenant cloud service reduce the reliability of these aftermarket solutions, making your overall operations brittle. It's not hard to imagine a scenario where an unexpected network change cuts operators off from both the information and the controls they need, all at the moment they are most needed.

A preferable approach to DataOps in the cloud environment is to leverage pipeline development tools that include native operations features to deploy, monitor and repair your pipelines, all managed via a single pane of glass. The obvious advantages of this approach are improved operational efficiency and greater operational reliability by reducing the number of moving parts.

As an added benefit, native operations features typically enable much more fine-grained monitoring as compared to aftermarket tools. Such information might not be an absolute requirement of a cloud migration project, but it can help offset some of the inherent challenges of debugging application problems on a public cloud.

StreamSets Supports DataOps for the Cloud

StreamSets built the industry's first multi-cloud DataOps platform for modern data integration, helping enterprises to continuously flow big, streaming and traditional data to their data scientists and data-intensive applications. It uniquely handles data drift, those frequent and unexpected changes to data that break pipelines and damage data integrity. The platform combines the open source StreamSets Data Collector™ for execution of any-to-any pipelines (the data plane) with the cloud-native StreamSets Control Hub™ for the design, monitoring and performance management of multi-pipeline topologies (the control plane). To learn more, visit www.streamsets.com.

ABOUT STREAMSETS

StreamSets transforms how enterprises flow big and fast data from myriad sources into data centers and cloud analytics platforms. Its DataOps platform helps companies build and operate continuous dataflow topologies, combining award-winning open source data movement software with a cloud-native Control Hub. Enterprises use StreamSets to enable cloud analytics, data lakes, Apache Kafka, IoT and cybersecurity.

Founded by Girish Pancha, former chief product officer of Informatica, and Arvind Prabhakar, a former engineering leader at Cloudera. StreamSets is backed by top-tier Silicon Valley venture capital firms, including Battery Ventures, New Enterprise Associates (NEA), and Accel Partners.

For more information, visit www.streamsets.com

LEARN MORE

To view more StreamSets solutions, go to www.streamsets.com/solutions

For additional information about StreamSets or StreamSets Services, please contact your StreamSets Account Representative or visit us online at www.streamsets.com