How Moving Your Couch and Moving Big Data to the Cloud Are Similar

By Daud Khan
Wednesday, January 15, 2020
Share

While many enterprises are moving big data to the cloud, the process is often harder than expected or simply fails. Read this article by Daud Khan, VP Corporate Development, WANdisco on how moving data to the cloud is like moving a couch to a new home - with some critical differences.

Enterprises are moving their big data to the cloud for many reasons - elasticity, speed, flexibility, and more. They're moving *en masse*: Flexera claims that 94% of organizations already use public or private cloud services. And they're moving for both processing and storage purposes - according to a recent survey, 56% of organizations using the cloud have a cloud-based data warehouse, while 48% use a cloud-based data lake.

Yet enterprises seeking the benefits of the cloud often face serious hurdles. Migrating big data to the cloud using existing paradigms and tools demands technical prowess and deep pockets – and it often doesn't work. In fact, more than half of big data migration efforts are either harder to complete than expected or fail outright.

The root of the problem? IT stakeholders continue to apply familiar on-prem assumptions to an inherently different cloud-based universe. They try, for example, to move big data in bulk to the cloud via the digital equivalent of a shipping container.

But the thing is, ***moving data is a lot like moving to a new home***. When you move homes, your furniture stays the same, but the context in which it is used changes. Traditional big data migration models assume that your sofa will always face your fireplace. But what if there is no fireplace in your new digs?

The Reason? Data Flows.

Unlike your couch, data – and the context in which it exists - is not static. It's constantly changing. And business doesn't stop during migration, no matter how that migration occurs.

Often cloud migration is performed in two steps:

The first step is to migrate the data, and the second step is to spend weeks or months porting on-premise applications to cloud stacks. Yet what happens to data that continues to change while this two-step migration is in progress?

Learn More: Software Bugs That "Bug" You And How to "Bug Them Off"!

Data scientists and AI algorithms alike have little use for stale or inconsistent data. This makes the handling of active data in motion - i.e. data that continues to change after the migration to cloud is initiated - one of the most pressing challenges facing enterprise IT today.

This is especially true in the case of data that is physically migrated to the cloud. Both Amazon and Microsoft have launched massive-scale physical data transfer services - Snowmobile and Azure Data Box Heavy. While it's clear that data migrated via truck will safely get to the cloud, it's less clear what

happens in the delta between migration and next use of physically-moved big data, given the inevitable changes to the business while the data was in transit? By the time a dataset is usable, the business ecosystem from which it originated could easily have changed so radically that the data is barely relevant and achieving consistent data is unattainable.

To handle the challenges of active data during migration, several free tools are offered by on-prem Hadoop vendors. These utilities are generally easy to use, but most don't granularly address the core issues facing data migration. Tools that ingest datasets on an end-to-end basis – rather than as files – are taking an overly simplistic and inflexible approach that simply doesn't work in the real-world of data use.

For example, DistCP - a tool that uses MapReduce for reporting, recovery, distribution and error handling – was actually born as a tool for inter-cluster copying, and is complex to use for the nonexpert user and especially deficient when it comes to large-scale migration of active data. The reason? DistCP has no ability to deal with changing data. In order to overcome this, Hadoop vendors have resorted to pushing customers towards a rigid structure where the customer defines the entire data pipeline - where and when it is ingested, how and when it is transformed and when querying is performed. The same problem is found in Hortonworks, Dataplane and Cloudera SDX. The inflexibility of these solutions, and the resulting rigidity of the platforms created when they're used, has led customers to largely eschew this model.

Learn More: 5 common IT stumbling blocks and how to overcome them

The Bottom Line

The reason that data architects are rethinking their plodding, massive data lakes – and that investors are fleeing from Hadoop-centric business models – is that data is not a static, closed body. Data flows and is the lifeblood for most organizations. It's a river that can't be dammed – not a lake. And this river keeps flowing for one simple reason: *businesses don't press pause for migrations, upgrades or downtime.*

The context of data evolves minute-by-minute. And just like you want your familiar comfy couch to fit perfectly in the context of your new living room, big data stakeholders need to ensure ironclad data consistency and usability ***both during and after*** their cloud migration.

About the Author
Daud Khan
Follow
Daud has spent most of his career following and commentating on infrastructure and application software companies and IT service companies. He was a director in equity research at Canaccord Genuity covering UK technology companies. Daud previously served at Berenberg, where he established its global technology research franchise, and has also held senior roles at JP Morgan Cazenove and Merrill Lynch. Daud qualified as an accountant (ACA) from PwC in 1999 and has an MA in Computer Science/Management Studies from the University of Cambridge.