DATOS IO

SOLUTION OVERVIEW

The Secret Sauce of RecoverX: CODR

Datos IO provides the industry's first cloud-scale, application-centric, data management platform enabling organizations to protect, mobilize, and monetize all their application data across private cloud, hybrid cloud and public cloud environments.

To learn more, visit www.datos.io

Background

Businesses are in the midst of a digital transformation journey. According to research from IDC, 70 percent of CIOs have a cloud-first strategy with the objective of harnessing the power of the cloud to drive growth by delivering new customer-centric products and services, while also driving greater operational efficiency. As part of this journey, enterprises are now operating IT across a multi-cloud infrastructure, deploying applications on the best suited cloud whether private, public, or managed. More specifically, there are two fundamental shifts occurring:

- Next-generation applications, which are hyper-scale and distributed, are being born 'in-the-cloud', aka they are 'cloud-first' applications. These applications are being deployed on next-generation distributed, non-relational databases such as Apache Cassandra, MongoDB, Apache HBase, and many others. As non-relational databases, they offer high-availability but compromise consistency.
- Traditional applications, originally designed and deployed on traditional data-center infrastructure are migrating 'to-the-cloud'. These applications are still predominantly deployed on traditional, relational databases such as Microsoft SQL Server, but many of the workloads are moving to the cloud.

Like other platforms before it, operating in the cloud needs a data management strategy, but traditional strategies have not kept pace with this rapid change. To address the requirements brought on by multi-cloud, a new cloud-first approach to data management is required. Driven by the need to management both traditional and next-generation mission-critical applications in a multi-cloud environment, and the shift to a multi-cloud infrastructure, three new key data management requirements have emerged:

- A software-only deployment model is required to create flexibility of deployment as well as customer freedom to deploy the storage of their choice.
- Each cloud use different infrastructure, the only common denominator is the "data" itself. There are no ESX virtual machines or SCSI LUNs in the public cloud, and the only common thing that binds all the clouds together is the data itself which does not change representation across VMWare, AWS and Azure.
- The elastic nature of multi-cloud databases necessitates data protection and management to be highly available,scalable and failure resilient.
- The eventually consistent nature of next-generation databases requires novel point-in-time techniques for consistent state across a cluster.

The Genesis of Datos IO

The data management industry thus far has had a scale-up mindset with relational databases and virtual machines being backed up to secondary storage though monolithic media servers. Worse, the media servers were not just a pass-through entity, but would convert the data into proprietary formats resulting in vendor lock-in – creating massive amount of "dark" data for enterprises. At first glance, it was clear that such an architecture was not suitable to support the requirements of traditional scale-up databases as well as the requirements of next generation high volume and velocity scale-out databases where failures are a norm, consistency is not guaranteed and clusters grow in size — and, most importantly where customers want to extract value from their secondary data sets.

With that challenge in mind, Datos IO challenged the status quo, started from scratch, and set out to create a new cloud data management platform that could address the requirements of both traditional and next-generation applications thereby enabling organizations to successfully manage all of their data in across a multi-cloud environment.

Introducing Consistent Orchestrated Distributed Recovery (CODR™)

Consistent Orchestrated Distributed Recovery, or CODR™, is the architecture upon which Datos IO solutions are built. The core principle of CODR™ is a scalable application centric view of data management and data protection that distinguishes it from conventional (media-server or LUN-based) approaches or virtualized approaches (VM-based). The benefit of our application-centric approach is fine-grained and highly space-efficient data protection and mobility that can span clouds over network links. Furthermore, such an application centric view allows the CODR™ engine to enable rich data management services (e.g. data governance, security, masking, etc.), unlike VM-based or LUN-based approaches that treat data as an opaque object and have no application context.

CODR™ is built upon several additional distinguishing principles, including:

- Elastic Compute Based - The CODR architecture a compute only data protection and mobility service that can be auto-scaled (elastic) up and down (scale out) depending on the application change rate - much like cloud compute and cloud storage services that are all elastic in nature. The CODR architecture does not create its own storage or file system, rather it consumes storage as a service that is used for storing versions (backups) of databases. All of this results in highly reduced infrastructure spending for enterprise customers. By contrast, both conventional approaches (backup appliances) and newer approaches (converging media server with backup appliances) incur high fixed storage and compute costs that accrue even if the data protection service is not being used. As a result, these approaches are a complete misfit with the IT movement of moving applications to the cloud, and the PBBA market is declining much faster than before.
- Semantic Deduplication - The CODR architecture introduces the industry-first Semantic Deduplication. There are three important reasons for the demise of traditional de-duplication techniques:
    - Data formats are increasingly compressed leading to the reduced effectiveness of de-duplication that relies on raw duplicated data inside a content stream.
    - In distributed storage systems, data is replicated for availability leading to multiple copies. These copies are not always exactly identical as ordering and flushing of updates differs across nodes. The use of compressed formats complicates the challenge.
    - Even if data were not compressed or replicated, the opportunity for de-duplication is also challenging because of the average de-duplication fragment size is small, exponentially increasing the amount metadata to keep track of de-duplication.

The goal of Semantic Deduplication, therefore, is based on the following insight: while the representations on disk are not physically identical due to compression or replication, they are semantically equivalent. Semantic deduplication identifies all semantically identical data fragments (such as a database column) so that we store only one copy of the data fragment in the secondary storage.

- Parallel Streaming - Another important element of the CODR architecture is the direct parallel and streaming transfer of data from the application to the secondary storage (network storage and cloud storage). This results in customers having secondary storage that can provide low RTO and RPOs for large data sets and high change rates as the data moves directly to secondary storage in parallel. By contrast, backup appliances and converged media server solutions can be deployed only in a single cloud, and will be a choke point if data to be protected is scaled out and globally distributed for performance and availability.
- Globally Distributed Metadata Catalog - The CODR architecture not only enables rich data services but also uses a globally distributed catalog to make the services available across multiple locations. This means that not only can backups happen in any cloud, but the protected data is immediately visible to all clouds in the enterprise. As a result, not only can data be restored to any cloud but rich data services can now be exposed to any location in the enterprise. This also completely eliminates any restriction on how data can move between clouds providing complete cloud flexibility for mobility use-cases. Contrast this to traditional approaches where both protection and recovery is uni-directional. For example, data can be backed up from VMWare to Amazon but not the other way around. Similarly, in the same scenario, data can be restored from Amazon to VMware but again, the reverse direction is not possible.
- Application-Consistent Versioning - CODR introduces the concept of versioning, where a version is defined to be a consistent view of a database. A version of a database is independent of the mechanism by which it is captured – it could be using snapshots, streaming logs, or asynchronous replication. As enterprises increase their pace to the Cloud and everything becomes *-as-a-service, versioning is likely to get abstracted away – consumers of a data protection service are going to be primarily focused on the promise of RTO and RPO rather than “how” of the promise.

A Modular Approach Designed for Flexibility

One of the key elements in the CODR architecture is the use of abstractions to isolate the architecture from the technical variations found in different data sources. One example is the use of the standard Open Database Connectivity (ODBC) API for accessing database management systems (DBMS). An application written using ODBC can be ported to other platforms, both on the client and server side, with few changes to the data access code. Similar other abstractions have allowed us to migrate between applications as well as versions of applications in a matter of weeks rather than months or years.

Additionally, the CODR architecture consists of two complementary software components. The first component is database specific and is referred to as Application Listeners. These listeners are lightweight, stateless and integrate with the databases via standard APIs and stream data in parallel to the secondary storage with no choke point.

The second database-agnostic component is the Scale-out Software Platform that orchestrates data movement, computes consistent space-efficient backups, and orchestrates recovery operations while maintaining native formats. The Scale-out Software Platform reads the data that the Application Listeners have transferred to the file or object in secondary storage and processes the data to make it ready for recovery.

We have isolated functionality of database specific logic in the Application Listeners and used abstractions in a manner such that the core CODR engine is independent of the application. This leads to a pluggable architecture where the support for a new data store can be encapsulated in a plugin.

The Foundation of Datos IO Differentiation

The CODR architecture is the technology foundation that enables Datos IO’s current and future products, to deliver unique benefits not found in any other solution, including:

- Simple Deployment - Datos IO products can be deployed in private, public, and hybrid environments with equal ease.
- Radically Lower TCO - CODR enables

tremendous operation efficiency and cost-savings on several fronts:

- Elasticity - Datos IO software can be scaled-up and scaled-down as needed depending upon workload requirements
- No Media Servers - The elimination of media servers dramatically reduces reliance on expensive, proprietary hardware
- No proprietary hardware or appliances - Datos IO requires no proprietary hardware or appliances. Instead you simply leverage the massive economies of cloud-scale infrastructure
- Unparalleled storage efficiency - Semantic deduplication delivers 10x storage efficiency vs. traditional deduplication, across whatever cloud storage you choose to leverage resulting in orders of magnitude cost savings.
- No Cloud Lock-In - Datos IO functional completely in the control plane and all data protected or managed remains in native format. This eliminates any reliance upon underlying infrastructure and provides complete cloud flexibility.
- Rich, Granular Data Services - Another benefit of application centricity and preserving native formats is that data services can be delivered in-place at a granular, application-specific level providing both flexibility as well as rich insight.

Conclusion

For the burgeoning world of multi-cloud data sources to become a reliable component of modern businesses and institutions, we have built the next-generation CODR architecture that solves important technical challenges to provide cross-cloud data protection and mobility for both traditional and next-generation applications.

About Datos IO

Datos IO is the application-centric data management company for the multi-cloud world. Our flagship Datos IO RecoverX delivers a radically novel approach to data management helping organizations embrace the cloud with confidence by delivering solutions that protect, mobilize, and monetize their data — at scale. Datos IO was recently awarded Product of the Year by Storage Magazine, and was recognized by Gartner in the 2016 Hype Cycle for Storage Technologies. Backed by Lightspeed Venture Partners and True Ventures, Datos IO is headquartered in San Jose, California.