DATOS IO

WHITEPAPER

Datos IO RecoverX

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

RecoverX

Executive Summary

Enterprises are increasingly adopting next-generation applications and migrating traditional applications to multi-cloud environments. As a result, enterprise IT infrastructure now consists of multiple clouds (public, private and hybrid) in distributed geographies, all connected over multiple networking links. However, every cloud uses unique technology for infrastructure services. For example, private clouds are predominantly powered by the VMware ESX virtualization suite, while AWS and Azure public clouds use Xen and Hyper-V for their virtualization technology. Accordingly, there are no ESX virtual machines and SCSI LUNs in the public cloud, and the only common denominator binding all the clouds together is the data itself, which does not change across VMWare, AWS and Azure.

Because of these shifts, customers now need a comprehensive Cloud Data Management solution. That is why we have evolved RecoverX 2.0 around three key dimensions: Cloud Mobility, Data Protection, and Platform Enhancements.

Multi-Cloud, The New Normal: What This Means For CIOs

- Next-generation applications, which are hyper-scale and distributed, are being born 'in-the-cloud', aka they are 'cloud-first' applications. These applications, deployed on next-generation distributed, non-relational databases such as Apache Cassandra, MongoDB, Redis, Apache HBase, Amazon DynamoDB among others must be protected.
- Traditional applications (often deployed on relational databases such as MySQL, Microsoft SQL Server, and others), originally designed and deployed on traditional data-center infrastructure are migrating 'to-the-cloud' (beginning with non-recovery workloads such as test/dev in the cloud, CI/CD in the cloud, DR in the cloud, et al). These applications need to be mobilized to enable organizations to move them to and from the cloud in an efficient and completely non-disruptive manner. Application owners and line of businesses (LOBs) are driving this momentum for enterprises as look to launch new customer-centric applications and services.

Any enterprise that has a multitude of applications and databases is living in a multi-cloud world and the implications are profound. From a CIO's perspective, there are several strategic takeaways. First, applications dictate the choice of cloud. For example, if you have applications that leverage Oracle's Exadata platform, you are going to move the Oracle Exadata platform to Oracle Cloud. Similarly, for Microsoft SQL Server-specific applications, you will likely move these applications either to Microsoft Azure public cloud or Amazon AWS. Not surprisingly, new and modern applications that are deployed on non-relational and modern databases will be deployed from the get-go on cloud-first infrastructure.

Second, use-cases cross cloud boundaries. In addition to cloud-native protection of enterprise applications that have migrated to the cloud, organizations need to move data sets to the cloud for all non-recovery workloads such as testing, development or analytics, migrating inactive data to the cloud for cost efficiency, and bringing data back on-premises for compliance and governance.

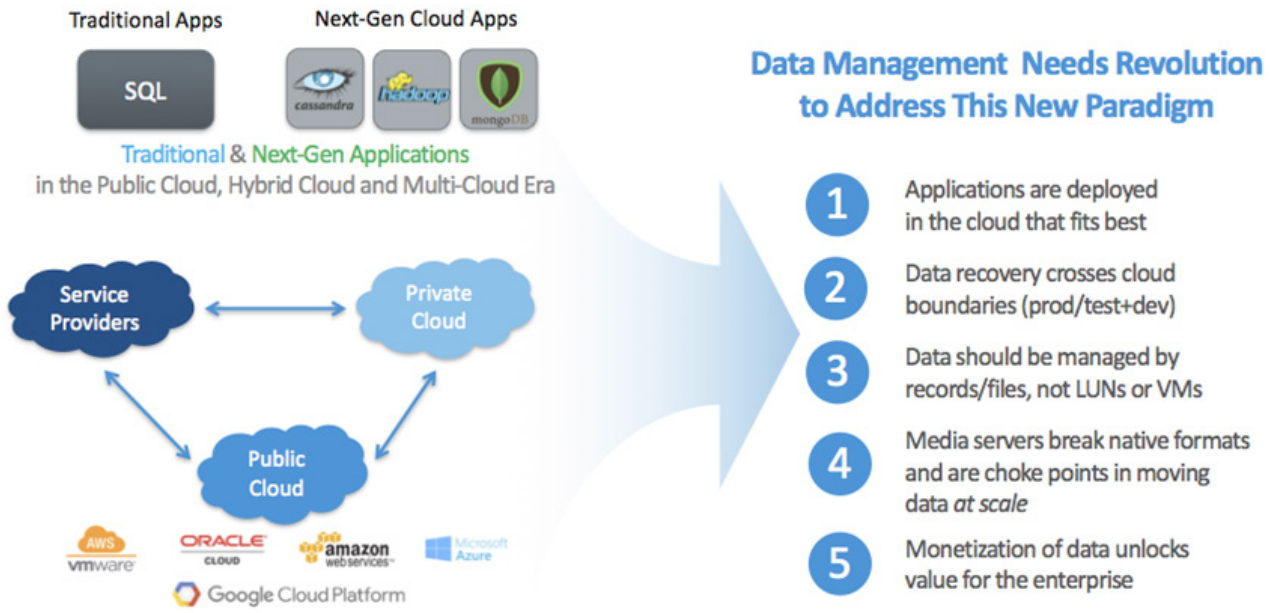
The bottom line is CIOs need a new data management strategy to thrive in the multi-cloud world, a strategy that not only provides data protection for hyper-scale, distributed applications born in-the-cloud, but also provides the freedom to best leverage all their cloud resources as dictated by application requirements. All of this while ensuring that enterprises have a scalable and reliable approach to manage, protect, recover, and monetize data in multi-cloud environments.

Multi-Cloud Requires a New Data Management Approach

The requirements for data protection and data management in a multi-cloud world require a fundamentally different approach. There are a number of critical capabilities to look for when opting for a data management strategy that can keep pace with your overall infrastructure transformation:

- Elastic Scale-out Software – to fully harness the power of the cloud, data management needs to be elastic and its performance needs to scale with the underlying compute infrastructure seamlessly. The common theme of the multi-cloud world, of next-generation applications born in the cloud, and of traditional applications migrating to the cloud is that of hyper-scale. Multi-cloud applications are, by definition, hyper-scale and distributed, therefore any data management strategy must be grounded in addressing protection and mobility at hyper-scale.

The Mult-Cloud Data Management Challenge



- Application-Centric – as applications migrate to the cloud there is no concept of a LUN or an ESX VM in the public cloud. All the underlying infrastructure is exposed as cloud-native services such as elastic block storage (EBS) or elastic compute cloud (EC2). In the cloud, the value is moving up the stack towards applications. Therefore, any data management strategy should be application-centric (table-level, column-level, et al) instead of infrastructure (e.g. LUN, VM) centric, eliminating any dependencies on underlying infrastructure.
- Performance at Scale – multi-cloud data management must eliminate the inherent shortcomings of legacy media-server based architectures. Instead, data must move directly and in parallel from the source to the destination, without any media servers.
- Storage Efficiency at Scale – deduplication technologies found in traditional data protection solutions don't work for the third platform applications and even for cloud-native applications. Rather, next-generation semantic deduplication technique is required that is application-centric and can provide the highest efficiency for protection and mobility operations.
- Global Data Visibility – due to the distributed nature of enterprise applications in the multi-cloud environment, data management needs to provide global data visibility enabling backup anywhere, recover anywhere, and migrate anywhere capabilities.
- Universal Data Portability – to maintain complete independence from the underlying multi-cloud infrastructure, data management should provide backups in native formats, always consistent data versioning enabling complete data recoverability, portability, and mobility.

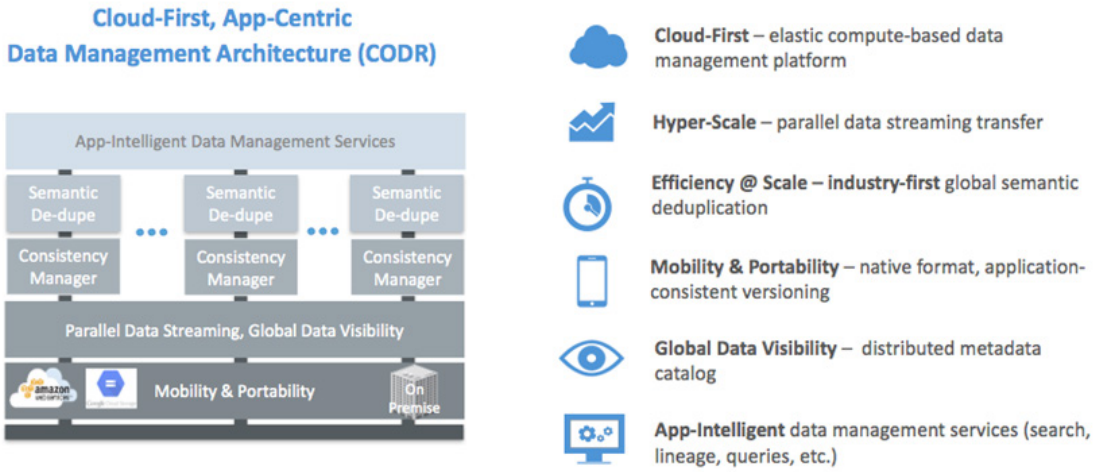
Datos IO RecoverX

Datos IO has leapfrogged the data management industry, creating the world's first application-centric data management platform for the multi-cloud world. Datos IO RecoverX is elastic, scale-out software, that delivers most efficient data protection and data management services for traditional and next-generation cloud applications.

RecoverX CODR™ Architecture

RecoverX is built on top of our seminal data management architecture called Consistent Orchestrated Distributed Recovery (CODRtm) engine, which is not dependent on media servers and transfers data in parallel to and from file-based and object-based secondary storage. The architecture is fully distributed in nature, which provides high availability in failure scenarios and uses elastic compute resources for scalable performance. CODR delivers application-

Reinvention Starts with Seminal Architecture: CODR



Datos IO is "application" centric. Not LUN or VM or Infrastructure Centric.

consistent data protection and management that allows massive storage efficiency, native formats, and sub-table level granular recovery/mobility solutions at scale for traditional and next-generation applications. CODR has two complementary software components:

- Light-weight application listeners that integrate with the data sources via standard APIs and stream data in parallel to the storage target with no choke point (due to no media servers). These Application Listeners compress the changes (if required) and transfer the data in parallel to the backend file or object storage of user's choice. Note that the data transfer is performed (for first full and incremental forever) directly to the backend storage without any involvement of any Datos IO software component. To avoid performance side-effects on the production data sources, Application Listeners are very lightweight and stateless. Application Listeners provide application-centricity and remove any dependency on the underlying infrastructure stack (e.g. VM, physical, public cloud) where the application is deployed. This allows RecoverX to provide universal data portability.
- Scale-out Software Platform that manages data movement, creates consistent space-efficient backups, and orchestrates recovery operations while maintaining native formats. The distributed Datos IO Software Platform is the brain behind the data management operations. It manages the deduplication process either at source or after the data is transferred by Application Listeners to the backend storage depending on the nature of the application that is protected. The Datos IO Software Platform also organizes the metadata to allow low recovery time and recover anywhere, and migrate anywhere capabilities. The storage gateway layer interfaces with various object and file based storage technologies and handles the nuances in different storage APIs. Storage gateway allows for a seamless experience for a user that may want to store their data on-premise, in the cloud or a combination of both.

RecoverX Features and Overview

Datos IO RecoverX industry-first cloud-scale, data-management software enables enterprises to protect, mobilize and manage their traditional applications and next-generation applications in a multi-cloud environment. Datos IO RecoverX provides scalable data protection, single-click recovery, industry-first semantic de-duplication and cross-cloud data mobility for next-generation databases (e.g. Cassandra, MongoDB), big data filesystems (e.g. HDFS) and relational databases (e.g. MS SQL Server). RecoverX allows organizations to protect and mobilize their data at a table-level

Industry-First Product Features

- Scale-out Software
- Scalable Versioning
- Reliable Recovery
- Semantic Deduplication
- Cloud Mobility

Customer Benefits

- Failure resiliency / performance
- Cluster-Consistent Backups
- Recover in minutes, not hours
- **Industry First!** Up to 80% reduction on secondary storage costs
- Operational efficiency in multi-cloud environment

DatosIO RecoverX
Industry-First Application-Centric Cloud Data Management Platform

granularity, reduce the application downtime with fully orchestrated, sub-table level recovery, save more than 80% on secondary storage costs and increase productivity of DevOps teams by mobilizing their data across cloud boundaries.

There are five key features of Datos IO RecoverX that are built to address the data protection and mobility needs of cloud-native applications.

Scale-out Software

Datos IO RecoverX is built to scale-out horizontally to ensure high availability of data protection and mobility infrastructure as well as increase in performance (RPO and RTO) to meet the growing application needs. RecoverX can be deployed in a single node or clustered configuration e.g. 3-node or 5-node. This allows customer to reliably protect their large environments with consistent backup and recovery performance.

- High availability: Like any other enterprise software, there can be internal system process failures or external infrastructure failures especially when commodity hardware is used. A single node deployment creates a single point of failure. Deploying a 3-node or 5-node RecoverX cluster ensures that all tasks handled by the failed RecoverX node are redistributed to the remainder nodes in the RecoverX cluster automatically without any disruptions.
- Higher throughput performance: The hyper-scale and scale-out nature of next generation application allows customers to easily scale their data size depending on application growth. Scale-out architecture of RecoverX brings a high degree of parallelism to achieve higher throughput for lower backup and recovery RPO and super-efficient data movement.

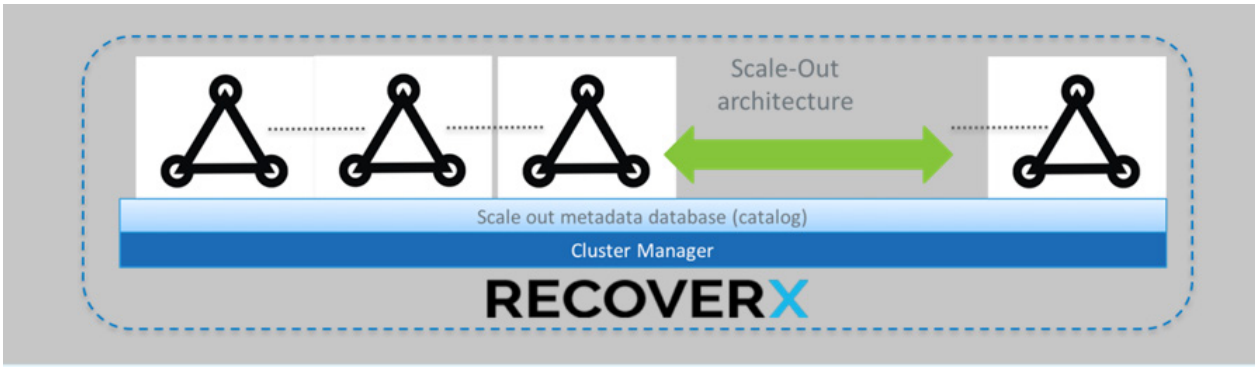
Scalable Versioning

Datos IO RecoverX allows enterprises to protect their enterprise applications for operational recovery in the event of data loss and mobilize the data within a cloud or across the clouds for non-recovery use cases (such as test/dev, DR, cloud on-ramping, performance testing, pre-staging, etc.).

A few noteworthy elements of versioning are:

- Application-consistent backups: Whether it is a relational database, non-relational database or big data filesystem, RecoverX creates a true point-in-time application-consistent backup copy.

Scale & Elasticity (3 to 5 software nodes)



Customer Asks / Use Cases	Datos IO Scale / Elasticity Description
1. Support for large scale clusters (>100TB database sizes, 100+ nodes, large number of tables 500+)	1. RecoverX can be deployed in a 5-node cluster for high performance, availability and scale
2. High change rates (+10%), Low Backup Intervals (1hr)	2. Datos IO RecoverX now available as 1-node, 3-node and 5-node configurations
3. Elastically scale up or scale down compute resources	3. User can configure compute resources to achieve required performance → elastic compute
4. TCO is top of mind: least RecoverX nodes as possible	

- Flexible RPO: RecoverX allows administrators to generate versions of their databases at any user-specified time interval. It provides much flexibility in setting backup intervals based on the RPO requirements of different data sources. Administrators may create backups at an interval as small as few hours to as large as several days.
- Granular Versioning: RecoverX empowers administrators to create versions of their databases at a granular level e.g. column family level for Apache Cassandra, collection level for MongoDB or table level for Microsoft SQL Server or directory level backups for Apache HDFS (including commercial versions of Cloudera and others). For backup and mobility operations, administrators may define policies that include a single table, multiple tables within a database or multiple tables across databases or multiple directories of a filesystem.
- Failure resiliency: Given that infrastructure, failures (network, storage, node, database) are a norm, RecoverX ensures that backup operations are resilient to such failures.
- Parallel Data Streaming: Given the hyper-scale nature of the next-generation applications where a single database or big data filesystem could be 100s of terabytes or more, it is critical that there is no bottleneck to data movement into and out of the cluster. Versioning is highly parallel and streaming in nature, whereby, RecoverX only acts as control plane that orchestrates data movement from data source cluster to version (backup) storage.
- Automated management of Application Listeners: RecoverX inserts and manages Application Listeners automatically without any user intervention. This is especially useful for large data sources such as Hadoop clusters that may have 100s of Datanodes.

Key benefits of versioning include: minimal impact on source applications, resiliency to data source failures, and no repairs when a version is restored that results in reduced application downtime. The backups are stored in data source's native format. Therefore, data recovery and movement processes are easier and vendor lock-in is avoided. Overall, versioning results in reduced data loss, consistent and efficient data movement, and minimal capital and operating expenditure for an enterprise.

Reliable and Granular Recovery

Datos IO RecoverX provides multiple ways to recover data for protection or mobility as described below:

Orchestrated Recovery to Data Source: This feature allows administrators to restore a backup copy from any point-in-time directly back to the source database or filesystem. This means that the end-user (database administrators or DevOps) can run their application directly on the recovered version of the data. The users can recover at a

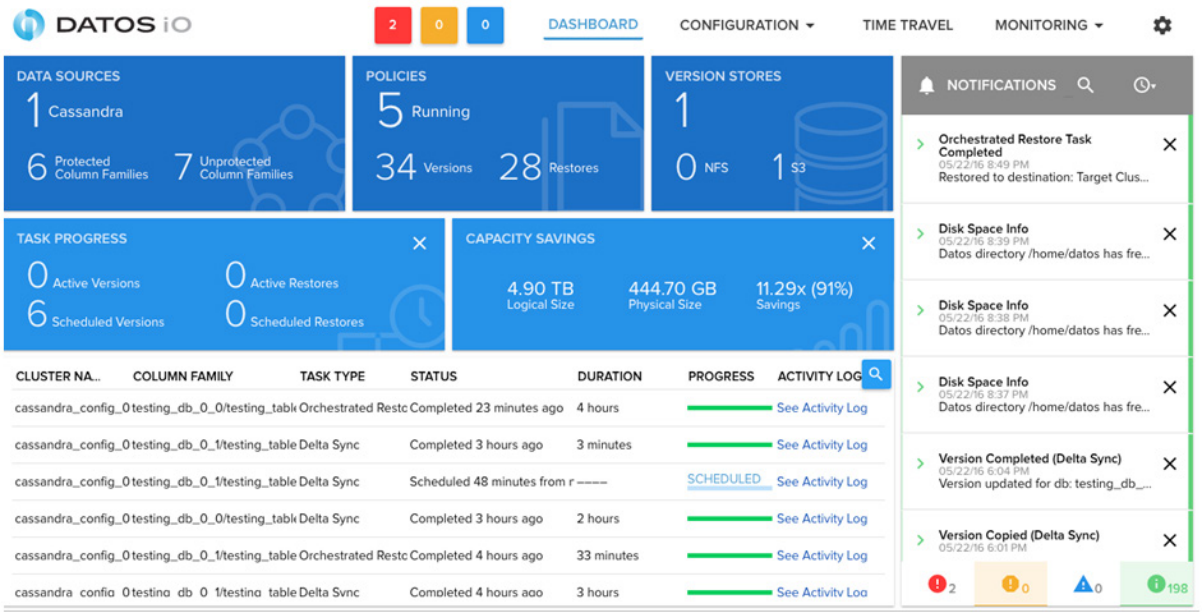


table-level granularity or file-level granularity. For example, a user may choose to recover certain files rather than the entire directory for their Hadoop filesystem cluster. Alternatively, a user may choose to recover a MongoDB collection rather than the entire database.

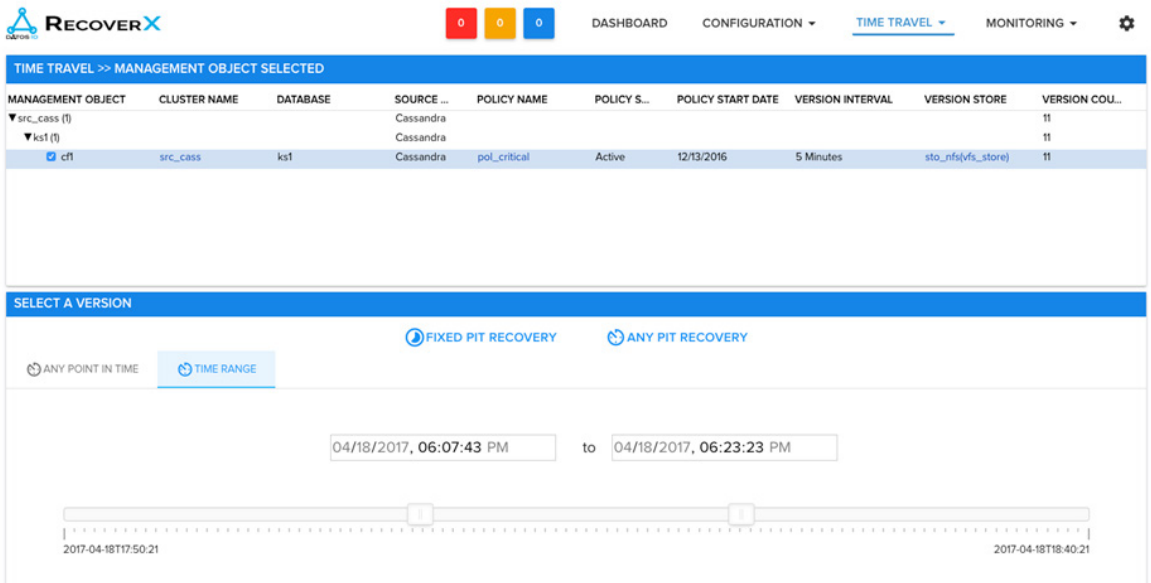
Orchestrated Recovery to a Different Cluster: This feature allows administrators to restore a backup copy from any point-in-time to a staging or test/development cluster that has a different topology (different number of nodes or different version of application) from the source cluster and may be in the same cloud or a different cloud environment (mobility).

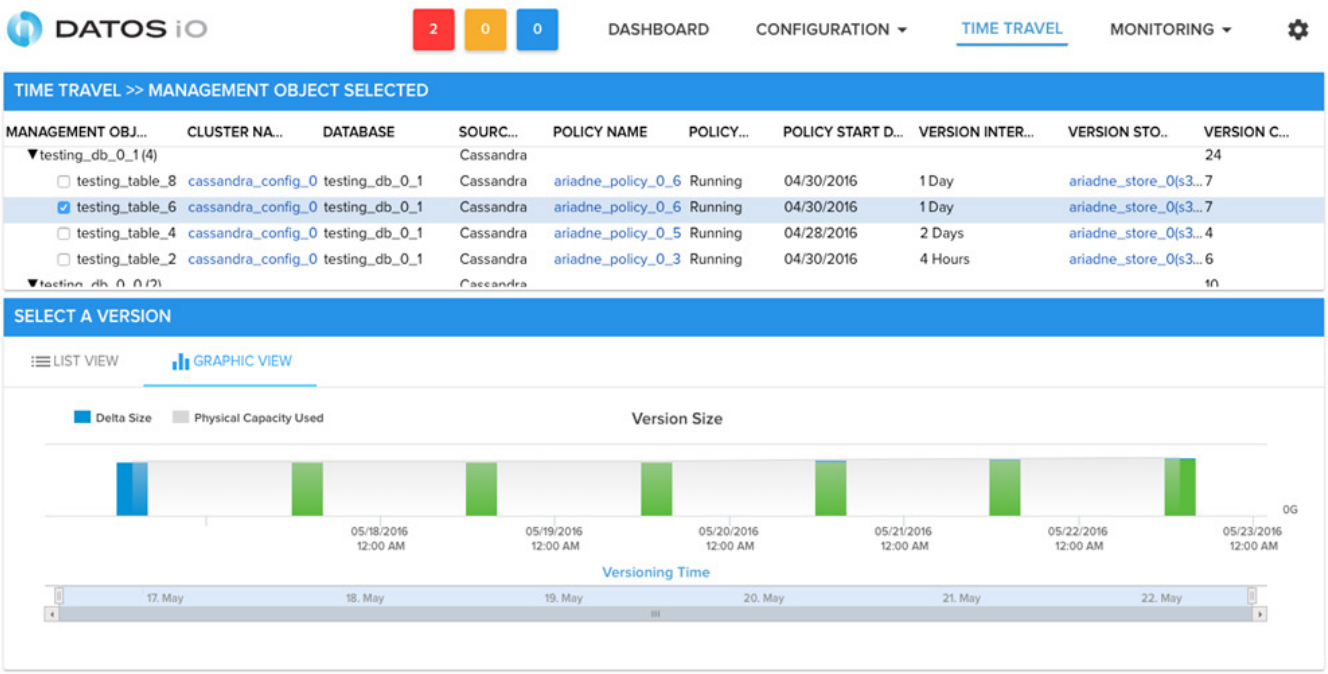
Recovery to clusters with unlike topology is especially useful if a user wants to populate their test or development environment using a version of their production data. For example, a user may have a 12-node production Cassandra database cluster and wishes to populate a 3-node test cluster with the production data. The database administrator configures RecoverX to protect their 12-node production database cluster. The database administrator can utilize any version of the protected production database to restore to the 3-node test cluster with ease. RecoverX dynamically manages the database replication factor and other elements such as TTL data that are needed for restore.

Incremental Recovery: This time-range based incremental restore feature allows administrators to restore data that has changed between two time ranges without recovering the entire table. For example, a customer may restore all data from timestamp: 4/15/2017 09:50:50 to timestamp: 4/18/2017 09:00:00 for a specific table. One of the use cases for this feature is data analytics where a customer may want to dig deeper into specific data sets for compliance, governance, or failure issue analysis. Incremental restores empowers DBAs and application owners to not be bogged down with restoring all the backup data, and rather only a subset of the data

Semantic Deduplication

Semantic de-duplication is an industry-first capability that Datos IO has developed specifically to reduce the cost of storing backup copies of application data over its retention period. Today, most scale-out and cloud databases keep multiple copies of the primary data – also called replicas. However, it is quite inefficient to store multiple copies of the same data on the secondary storage. As part of versioning, Datos IO RecoverX makes sure the final backup has no replicas of primary data set, thus providing de-duplication of source data -- all without losing native formats. For example, if the database uses a replication factor (RF) of 3, using Datos IO RecoverX will save up to 70% secondary storage costs. For big data filesystems and relational databases, the deduplication is done at source to minimize the amount of data that is transferred over the network.





This example shows a sample of storage savings that are achieved for operational recovery use case of Apache Cassandra database:

- Database Size (logical): 10TB (10% daily insert rate)
- Replication Factor: 3x
- Versioning Frequency: Daily (7-day retention)

	Secondary Storage Required (Without Datos IO)	Secondary Storage Required (With Datos IO)
Day 1 (Initial Sync)	30TB	10TB
Day 2 (Incremental)	3TB	1.0TB
Day 3 (Incremental)	3.3TB	1.1TB
Day 4 (Incremental)	3.6TB	1.21TB
Day 5 (Incremental)	4.0TB	1.33TB
Day 6 (Incremental)	4.4TB	1.46TB
Day 7 (Incremental)	4.8TB	1.61TB
Total Storage	**53.1TB**	**17.TB (3x savings)**

Cloud Mobility

Leveraging the storage efficiencies of the CODRtm architecture Datos IO RecoverX allows enterprises to mobilize their data across cloud boundaries for non-backup use-cases such as compliance/governance or continuous development by enabling test/dev refresh across clouds.

Data Governance and Compliance: Enterprises that have production clusters deployed in public cloud environment often need to keep a copy of that data on-premise as well for governance or compliance requirements. These enterprises can leverage RecoverX in public cloud to protect production applications running in public cloud and at the same time move a copy of that data back to an on-premise repository.

Test/Dev Refresh: Oftentimes, different teams within an enterprise leverage different cloud environments. Test or development teams may use a public cloud (e.g. Amazon AWS or Google Cloud) but the core IT teams may use private cloud. Moving data across these cloud boundaries safely is not only an operational nightmare but also extremely error prone. RecoverX natively

allows administrators to move data across such cloud boundaries. In addition, the semantic deduplication ensures that the data is moved extremely efficiently and at a table level granularity.

Beyond these key features, RecoverX provides several other features that makes it an enterprise grade data management platform.

Failure Resiliency

Failures are a norm for next-generation applications that are deployed on commodity hardware. To provide robust data protection in such environment, Datos IO RecoverX has built-in features that enable continuous versioning and recovery even though there are multiple failures.

Operational Visibility

RecoverX provides deep operational visibility for administrators to monitor their environments and better understand their data change patterns. These metrics are available through the GUI and may also be sourced through the RESTful API interface of RecoverX.

- Pre-classified system notifications as critical, warnings, errors, informational alerts
- Comprehensive view of data protection status as number of protected tables
- Incremental data and number of file changes per interval
- Granular storage savings trends
- Backend storage consumption trends
- Table level storage consumption on backend storage
- Status of all active tasks

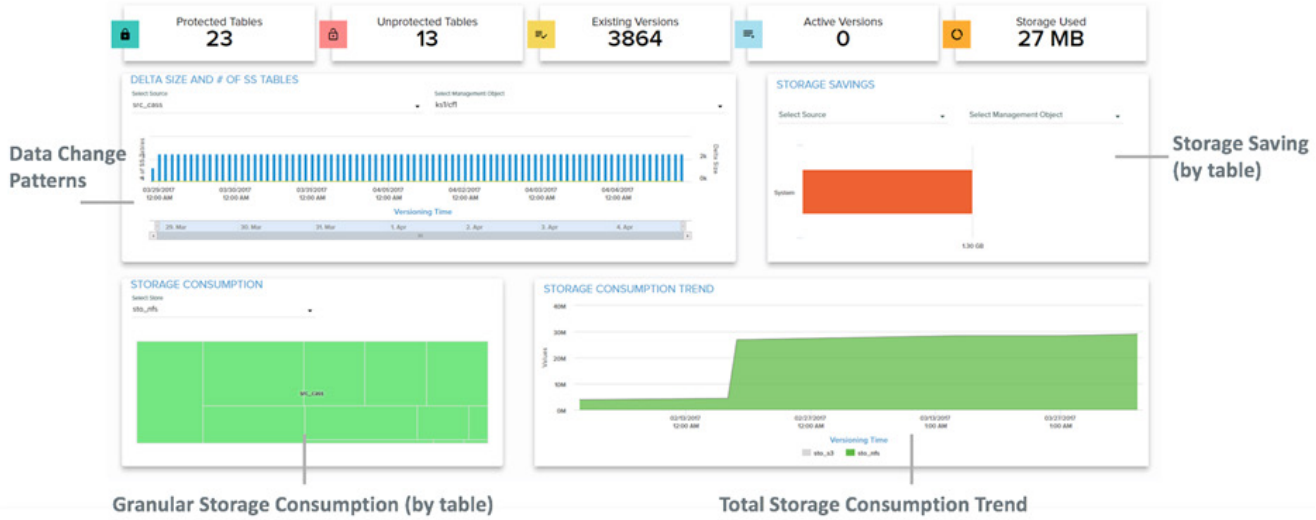
Benefiting Traditional and Next-Generation Personas

As enterprises are shifting into a multi-cloud world, individual's roles evolve and new and different personas are increasingly becoming key players in architecting an overall cloud data management strategy. Data protection and database administrator's' responsibilities are extended from management of traditional to next-generation applications. Meanwhile, DevOps teams responsible for building new applications are also tasked with ensuring those application are fully protected and can be mobilized for developer agility and continuous development. Datos IO RecoverX is built to address the needs of multiple constituents within an organization:

- The first are application architects, application owners, and admins who build the cloud-first applications and owners of traditional applications and are key stakeholders in the proper functioning of the application. Datos IO offers fast, granular protection

	Impact On Versioning	Impact on Recovery
Single Source Node and Source Database failures	None; In the event of failures, RecoverX transfers all accessible data and creates a consistent copy based on this dataset	N/A
Single Destination (target) Node and database failures	N/A	None; in the event of failures after data is copied to the destination cluster, RecoverX completes the recovery operations
Internal RecoverX node failures	None; versioning continues as a normal operation as Datos IO software handles the failures in the distributed architecture	None; recovery continues as a normal operation as Datos IO software handles the failures in the distributed architecture

Rich Analytical Information on Data Patterns and Data Protection



and recovery in case of data or schema corruptions that lead to application downtime, and improves productivity by enabling agile test and development.

- The second groups are data protection administrators, database architects and database administrators (DBAs), who are responsible for ensuring that databases offer the same kind of enterprise-grade data management and recovery capabilities found in traditional relational databases. Architects and DBAs need to ensure that database infrastructure is protected against corruption caused by application developer mistakes; they can recover to a previous, uncorrupted state by using point-in-time versions.
- The final group are DevOps and IT operations teams, who are the end-users, and have emerged in recent years from the growing community of applications and orchestration deployment frameworks such as Chef and Puppet. Datos IO provides this group with simplified cloning for test and development in the same cloud or across clouds, and the API-driven orchestration that reduces operational management complexity through better usability and deployment.

Conclusion

Multi-cloud is the new normal, and to keep pace with this cloud transformation, enterprises must adopt a cloud-first data management strategy. Datos IO's mission is to help organizations accelerate their adoption of multi-cloud by enabling them to protect, mobilize, and monetize their 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.