DATOS IO

Datos IO And Cassandra: Solution Brief

DATOS IO

+

cassandra

Datos IO RecoverX is purpose-built data protection software for Apache Cassandra

DEPLOYMENT OPTIONS

Public Cloud (IaaS, PaaS)

- Fully compatible with cloud compute and storage services from Amazon AWS and Google GCP

On Premise

- Fully integrates with existing Secondary Storage appliances (NFS)

The Challenge: Reliable Recovery for Cloud Applications

Enterprise growth, innovation and, occasionally, even competition require that today's companies adopt new, high-value, data-centric applications such as content-management systems, real-time intrusion detection, customer analytics, internet-of-things, and digital advertising. To handle the data requirements of this new generation of high-volume, high-ingestion rate and real-time 3rd Platform applications, enterprises are turning to scalable, eventually consistent storage systems (such as Apache Cassandra) rather than traditional scale-up database and storage approaches.

However, this fundamental shift raises critical issues in the lifecycle of data management. Traditional backup and recovery products were originally designed for small-scale databases, tape-based storage media, and legacy architectures that were designed for on-premise deployments. This leaves the next-generation of distributed, reliable recovery solutions underneath modern database architectures with a critical gap.

The Solution: Datos IO RecoverX

Datos IO RecoverX is purpose-built to address the data protection needs of cloud-native applications deployed on Apache Cassandra and DataStax Enterprise database.

Scale-Out Software:

- Highly available data protection infrastructure
- Performance (RPO/RTO) scales per application needs
- High availability: Like any other enterprise software, there can be an internal system process failures or external infrastructure failures especially when commodity hardware is used. A single node deployment creates is a single point of failure. Deploying a 3-node RecoverX cluster ensures that all tasks handled by the failed RecoverX node are redistributed to the remainder nodes in the Datos IO cluster.
- Higher throughput performance: The scale-out architecture of Apache Cassandra, 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 backup and recovery throughput leading to lower RPO.

Scalable Versioning: The versioning operation is also highly parallel in nature, whereby, RecoverX only acts as control plane that orchestrates data movement from data source cluster to version (secondary) storage. This allows Datos IO to handle large clusters and workloads. Databases protected and versioned by RecoverX are stored in the databases' native format. Therefore, data recovery process is made easier and vendor lock-in is avoided.

RecoverX also brings operational ease of use by allowing administrators to generate versions of their databases at any user-specified time interval and at any granularity (table-level or entire database). Finally, given that most distributed databases run on commodity infrastructure, failures (network, storage, node, database) are a norm. RecoverX ensures that backup operations (full backup and incremental backups) are resilient to such failures.

Reliable Recovery: At the time of recovery, the user user can either choose a specific version to restore or restore at any PIT a point-in-time (T) from which data has to be restored. RecoverX creates the schema in the target cluster and then proceeds to stream data into the cluster directly from the backup storage. No processing is performed on the data at the time of recovery. This is one of the key reasons for the performance advantages of RecoverX compared to traditional solutions. One of the key tenets of RecoverX solution is that, at the end of the restore, the target cluster will

have a copy of the data that is cluster consistent as of time T, thus helping administrators achieve low application downtimes. Additionally, Datos IO RecoverX supports advanced recovery features such as node level recovery, any PIT, and masking.

Industry-First Semantic Deduplication: Semantic deduplication is an industry-first capability that Datos IO has developed specifically to reduce the cost of storing backup data over its retention period. Today, most cloud databases keep multiple copies of the primary data – also called replicas. As part of the versioning process, RecoverX removes the redundant data sets to make sure that the backup has no replicas of a primary data set, thus providing de-duplication of source data across all replicas. This ground-breaking semantic de-duplication feature results in up to ~70% reduction in secondary storage.

Reference Deployment

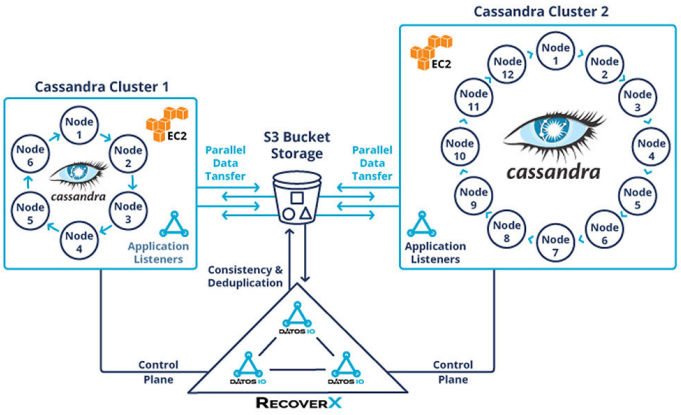
There are three main components in Datos IO RecoverX deployment for Apache Cassandra.

1. Data Source:

- The Apache Cassandra cluster that is protected and managed by RecoverX. Datos IO RecoverX supports Apache Cassandra versions v2.0 and v2.1 and DataStax Enterprise Editions v4.5, v4.6, v4.7 and v4.8.

2. RecoverX, Data Protection Software For Cassandra

- A 3-node (or 1-node) RecoverX software deployment. RecoverX can be deployed on physical servers, virtual machines (VM), EC2 instances on Amazon AWS or Compute Engine on Google Cloud Platform.



3. Backup (secondary) storage

- The secondary storage used by RecoverX to store backups. Supported targets includes NFS based storage, Cloud storage (Amazon AWS S3 and Google Cloud Storage)

Datos IO RecoverX for Cassandra: Compatibility Matrix

Apache Cassandra	Apache Cassandra v2.0, Apache Cassandra v2.1		
DataStax Cassandra	DSE v4.5, DSE v4.6, DSE v4.7, DSE v4.8		
Secondary Storage Type Supported	NFS Storage for On-Premise Deployments	AWS S3 Google Cloud Storage	
Datos IO RecoverX Node	RHEL/Centos 6.x	RHEL/Centos 6.x	RHEL/Centos 6.x
	8-core physical or virtual machine	EC2 M4.2xlarge or above EC2 C3.2xlarge or above	Standard 8 vCPU, 30GB RAM
	16GB Memory (minimum)		
	128GB Local Storage (SSD)	128GB EBS or Local SSD	128GB SSD

Conclusion

Datos IO provides the industry's first cloud-scale data protection software for next-generation applications and cloud databases. Able to deliver reliable recovery at scale in minutes (vs. hours) and reduction of up to 70 percent in secondary storage costs, Datos IO RecoverX increases the productivity of application owners, IT Ops and DevOps teams. Datos IO has been recognized by Gartner in 2016 Hype Cycle for Storage Technologies for new category of Cloud Data Backup. Backed by Lightspeed Venture Partners and True Ventures, Datos IO is headquartered in San Jose, California.