DATOS IO

SOLUTION OVERVIEW

Petabyte-Scale Data Protection for Big Data Filesystems (HDFS)

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. Achieve fast, reliable, and automated data protection for Hadoop Distributed File Systems (HDFS) distributions

To learn more, visit www.datos.io

hadoop

Key Benefits

Minimize Application Downtime

- Directory-level backup of large Hadoop filesystems
- Single-click orchestrated recovery
- Granular file-level recovery

Increase Operational Efficiency

- Enterprise policy management
- Recovery to different topology Hadoop clusters (test/dev use case)
- Automated failure handling

Increase Storage Efficiency

- Industry-first semantic deduplication
- Policy-based archival for long-term retention

Simplify Deployment

- On-premises or public cloud deployment
- Compatible with any NFS or Object based backup storage

The Data Protection Challenge

Rapid proliferation in social, mobile, cloud, and Internet-of-Things are driving enterprises to deploy hyper-scale, distributed, data-centric applications such as customer analytics, e-commerce, security, surveillance, and business intelligence. To meet storage and analytics requirements of these high-volume, high-ingestion-rate, and real-time applications, enterprises are rapidly adopting massively scalable data lake platforms such as Hadoop.

Hadoop is primariy used as an analytics platform to capture large-scale data from various sources. As use cases for big data analytics rapidly proliferate the enterprise, big data applications that leverage Hadoop platforms are becoming mission critical and require enterprise-grade data protection. Although the Hadoop (HDFS) filesystem (also termed as "big data filesystem") offer replication and local snapshots, it lacks point-in-time backup and recovery capabilities. Any logical or human error can result in data loss and is detrimental to business applications. Oftentimes, for analytical data stores, data may be reconstructed from the respective data source but it takes long time and is operationally inefficient, leading to data loss. Given the large scale (node count and data set sizes) and the use of direct-attached storage in Hadoop clusters, traditional backup and recovery products will not work, leaving a critical data protection gap.

The Solution: Datos IO RecoverX for Scalable Data Protection of Hadoop

Datos IO has developed an industry-leading software product, RecoverX that is purpose-built, to meet the data protection requirements of petabyte scale HDFS-based filesystem environments. Enterprises can now leverage the performance, scalability, and storage efficiency of this solution to achieve faster recovery and economical storage of data throughout its lifecycle. Key use cases are:

- Point-in-time backup and recovery
- Archival and long-term retention
- Test/Dev refresh

Scale-Out Architecture

Datos IO RecoverX is founded upon Consistent Orchestrated Distributed Recovery (CODR), Datos IO's scale-out software architecture that enables customers to scale data management performance in real time. CODR uses elastic compute services that can be auto-scaled with application load and removes the dependency on media servers. CODR also transfers data in parallel to and from file-based and object-based secondary storage for multiple workloads, including data protection and testing and development.
CODR offers:

- High Availability: Deploying RecoverX in a cluster configuration prevents any software process or external hardware (node) failures from compromising backup and recovery operations.
- Enhanced Backup Performance: With the scale-out architecture of Hadoop, users can easily scale their platform according to application growth. Similarly, the scale-out capability of RecoverX with no reliance on media servers helps customers increase backup and recovery throughput, leading to sustained performance.

Industry-First Semantic Deduplication

Semantic deduplication is Datos IO's an industry-first capability that produces 10x more storage efficiency than traditional deduplication. Most distributed filesystems keep multiple copies of the primary data, called *replicas*. As part of backup operations, RecoverX applies application-specific intelligence via file-level deduplication techniques that scale to petabytes of filesystem data that needs to be protected with consistently scalable backup and recovery performance.

Fully Orchestrated and Granular Recovery

Granular file-level recovery: Datos IO RecoverX provides single-click, granular, point-in-time recovery of HDFS based filesystems. With RecoverX, customers can recover data directly back into the same Hadoop cluster (operational recovery) or to a different cluster (testing-and-development refresh) with a different topology (number of nodes). Granular recovery obviates the need for full directory restore and allows administrators to restore relevant files into smaller test/dev clusters. This option reduces the operational burden of refreshing test/dev clusters for continuous development DevOps environments. During recovery, the data is directly (with no reliance on media servers) transferred from secondary (backup) storage into target HDFS cluster, resulting in a very low RTO.

Scalable Versioning

By using native application intelligence, RecoverX creates a point-in-time backup of Hadoop directories at user-specified backup intervals. It automatically deploys Application Listeners on `DataNodes`. Application Listeners are lightweight scripts that orchestrates data movement of new or changed files from `DataNodes` to backup storage in parallel. This approach allows RecoverX to handle petabyte-scale Hadoop clusters. By allowing administrators to generate backups at any user-specified time interval and at any granularity (directory-level), RecoverX simplifies operational use for IT operations, Infrastructure and DevOps teams.

Reference Deployment

Datos IO RecoverX can be deployed on a physical server, a virtual machine, or any cloud compute instance (for example, Amazon EC2, Amazon EC2, Oracle Compute, Google Compute, Azure Compute, et al)). RecoverX communicates with the Hadoop cluster through Application Listeners that are lightweight scripts automatically deployed by RecoverX on `DataNodes`.

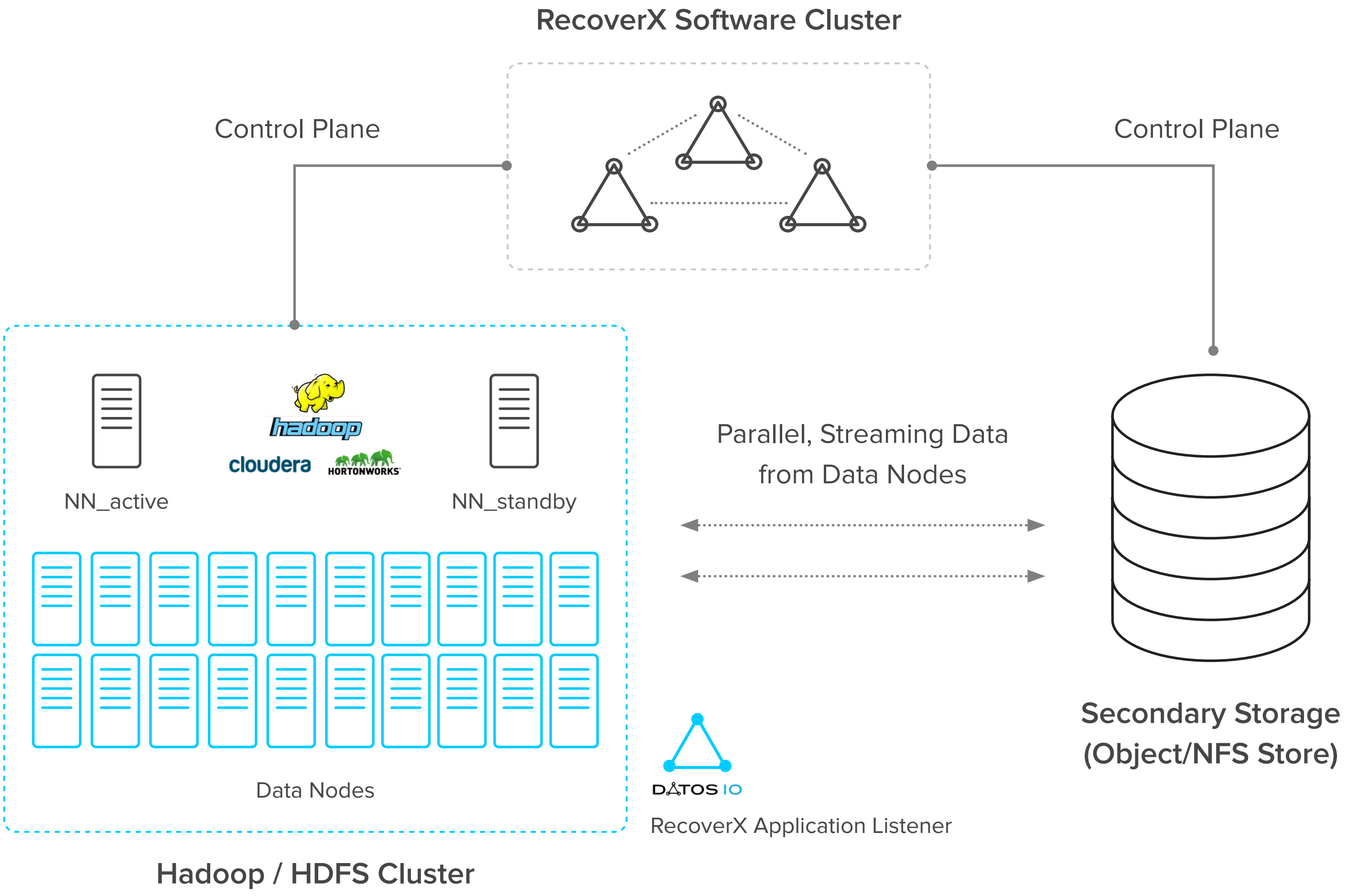

Figure: Reference deployment of Hadoop and Datos IO RecoverX

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.