IBM Cloud Object Storage for Hadoop

Back up and archive data more easily with IBM Big Replicate technology

Benefits

- Streamlines management of your Hadoop inactive data with simplified backup, recovery, migration and expansion
- Helps reduce total cost of ownership by using IBM Cloud Object Storage for backing up and archiving inactive data
- Provides continuous data consistency and virtually zero downtime and data loss
- Increases protection for both data at rest and data in motion
- Avoids the need for standby hardware so that you can use existing servers to dramatically scale up your Hadoop deployments

Analytics and cognitive computing provide some of the fastest-growing use cases across industries for large amounts of unstructured data. The data explosion is accelerating, and enterprises are challenged to analyze and store a growing volume of unstructured data. According to leading analyst firms, digital data is doubling every two years and 80 percent of that data is unstructured[1]. However, enterprises face challenges in deriving value from these vast data repositories. To solve this problem, IT professionals and business leaders are increasingly using analytic tools and cognitive analysis to derive insights from this data.

Approximately 50 percent of firms using big data technologies have built Hadoop data lakes for processing those data sets.[2] Hadoop Distributed File System (HDFS), the default storage platform for Hadoop data sets, can be inherently expensive because of the need to store three copies of each file and scale up compute nodes as storage needs increase. With this in mind, Hadoop software users search for more cost-effective archive and backup alternatives that can still deliver on the most demanding enterprise requirements.

Transforming analytics data stores for greater scale and cost-efficiency without disruption

Enterprises that have deployed Hadoop technology can now experience the benefits of IBM® Big Replicate technology as it delivers higher data consistency and availability with IBM Cloud Object Storage. The new Cloud Object Storage replication option works in combinations of dedicated and public, on-premises or cloud environments, regardless of distance or data source. Hadoop software users can adopt this combined solution to more efficiently replicate data and metadata between Hadoop

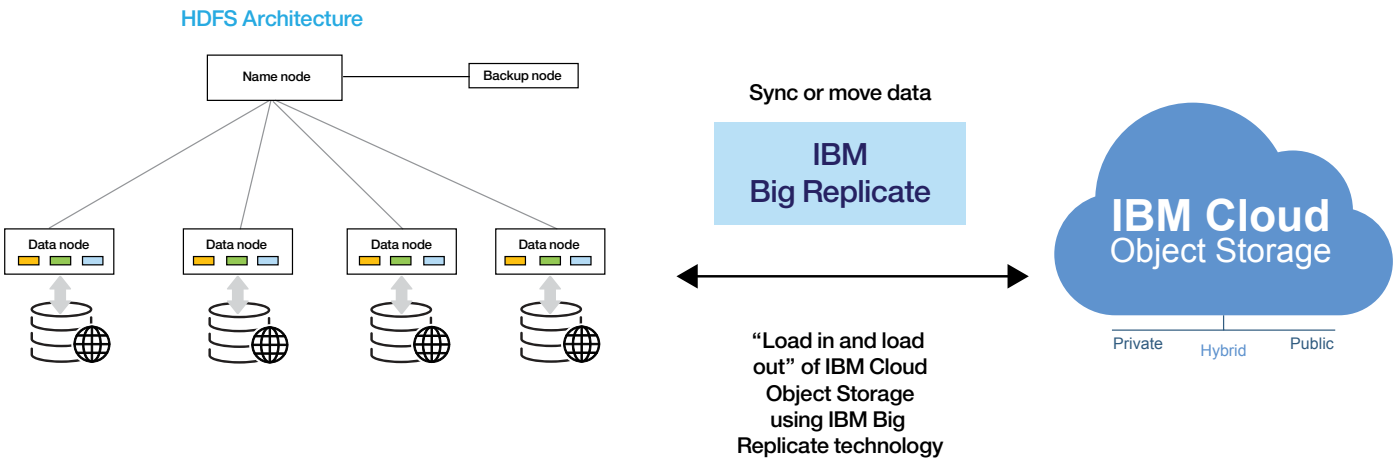

Figure 1: IBM Big Replicate technology loads inactive data in and out of IBM Cloud Object Storage for archive, backup and disaster recovery.

clusters or between Hadoop clusters and IBM Cloud Object Storage. Given its highly efficient and scalable design, IBM Cloud Object Storage and the IBM Big Replicate application can provide simplified backup, recovery, migration and data expansion solutions with significant reductions in total cost of ownership (TCO).

Typical use cases capitalizing on IBM Big Replicate technology include:

- Archiving—Migrate existing inactive data from HDFS by archiving to IBM Cloud Object Storage
- Backup and disaster recovery (DR)—Peer data from HDFS to IBM Cloud Object Storage for greater data protection and durability

With enterprises using analytics more pervasively on production-level data, the consistency and continuous availability of your data has become mission critical—even when it's archived. This level of performance is important to a diverse range of industries such as financial services companies rolling out new fraud detection applications, healthcare research institutes identifying new patient therapies in realtime or gaming companies delivering an immersive experience to their customers while simultaneously working to understand user behavior.

In essence, the IBM technology replicates big data from lab to production or from production or data centers to IBM Cloud Object Storage for archiving, backup and DR. Your data is replicated as it streams in, avoiding the need for files to be fully written and closed before data transfer. This can save time, facilitate smooth operations and accelerate the pace of data analysis. Big Replicate technology offers flexible replication across supported Hadoop distributions that are compliant with HDFS application programming interfaces (APIs) and Hadoop versions running on IBM Cloud Object Storage, which uses an S3 API interface. In effect, you can use this solution to scale storage independent of compute capacity, which helps you avoid purchasing more server capacity than you need and save your budget for other investments while simplifying infrastructure management.

Gaining better data continuity, availability and flexibility

On its own, Big Replicate technology is designed to provide continuous data availability and data consistency. That is, global users experience local network-speed read-and-write access to the same data and the same view of the data in multiple locations—as if from one Hadoop cluster—from virtually any distance. The Big Replicate technology also provides continuous data access in the face of network outages, hardware failures and entire data centers going up and down, supporting virtually complete data resilience. On the other hand, with batch-based, distributed copy (DistCp) solutions, data added since the most recent cluster backup can potentially get lost. Because DistCp replicators tend to consume a high level of resources, larger jobs are often done during off-peak hours, increasing the possibility of data loss throughout the day.

With the IBM Big Replicate active-active replication design, transactions are replicated as they are created and stream in, and virtually every change is replicated to the other participating clusters. You don't need to wait for all data to be fully written and closed before transferring files. This process helps continuously protect your analytics work and data sources.

Data replication with IBM Big Replicate can also be up to 90 percent faster than replication with DistCp solutions such as Cloudera, Inc. Backup and Disaster Recovery (BDR) software without impacting the performance of the other applications running on the Hadoop clusters.[3] In addition, users have the flexibility to selectively replicate subsets of data within Hadoop clusters, giving them greater control and agility for their analytics projects.

When archiving and backing up to IBM Cloud Object Storage, copies of files are automatically geodispersed in multiple locations using patented technology to facilitate high data reliability and availability. You can access these files in the event of a disaster through a built-in, fault-tolerant IBM design to deliver 100 percent uptime at virtually limitless scale. Once data is stored in the IBM Cloud Object Storage public cloud, users can more easily take advantage of IBM Cloud Services and various IBM Watson® analytics capabilities such as fraud detection, facial recognition, visual recognition and weather trends to enhance their analysis.

Over time, users can discover additional benefits when working with IBM Big Replicate and Cloud Object Storage to manage their projects and workflow. They can:

- Migrate production data to the cloud without data loss
- Dynamically add new clusters and data centers to live Hadoop or hybrid cloud deployments without downtime
- Perform staggered upgrades to and migration between Hadoop clusters and IBM Cloud Object Storage without disruption
- Maintain flexibility in moving data and conducting analysis because data remains in its native format
- Scale storage effortlessly across hundreds of thousands of servers and tens of thousands of users
- Manage both on-premises and dedicated cloud deployments using a central console
- Move data in and out of IBM Cloud Object Storage as needed for on-demand burst-out processing and offsite data replication
- Tier data across IBM Cloud Object Storage offerings (Standard, Vault, Cold Vault, Flex) using IBM Big Replicate for greater cost savings as data ages or changes in value

Using cloud storage that is scalable, flexible and simple

Storage capacity can scale more quickly and easily with IBM Cloud Object Storage to meet your growing file sizes; provisioning new capacity is designed to be simple and immediate. There is virtually no limit to the number and size of your files or your file storage capacity. When paired with IBM Cloud Object Storage and its wide array of deployment options (on-premises, dedicated and public cloud), IBM Big Replicate technology gives you access to a more flexible, scalable and simple backup or archiving solution for Hadoop-based data. In addition, because only one stored copy of the data is needed, your team can benefit from more simplicity and cost savings.

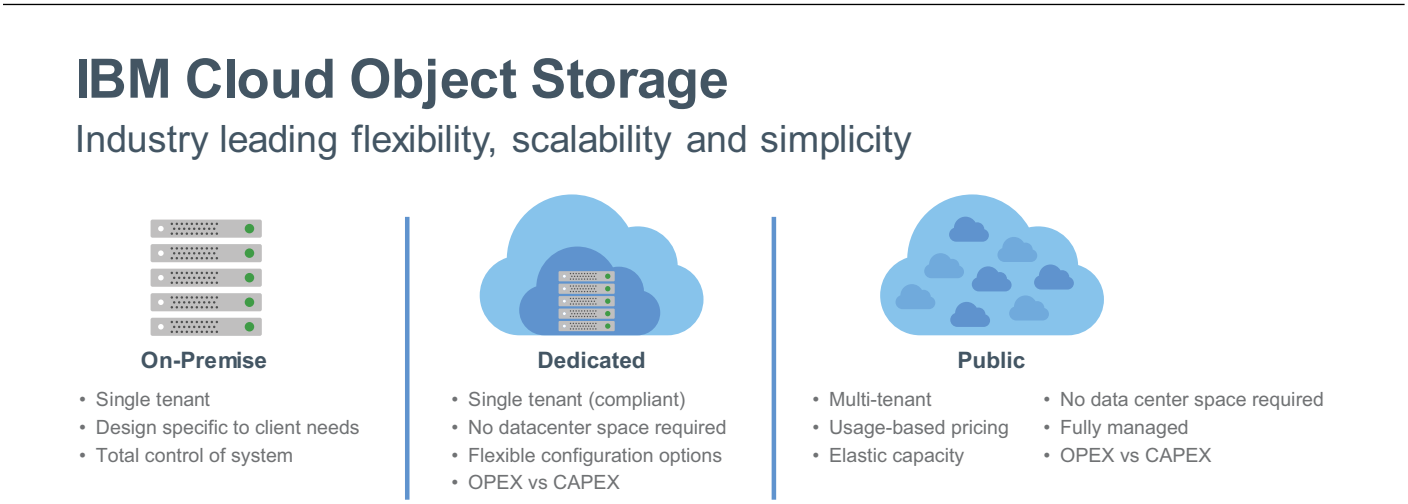

Figure 2: IBM Cloud Object Storage offers flexible deployment options for both on-premises and the cloud.

Enhanced security and data resilience

IBM Cloud Object Storage can inherently provide more data protection and durability than standard Hadoop cluster deployments. Once sent to IBM Cloud Object Storage, your data remains highly available and resilient by design because of the patented geodispersal technology developed by IBM. Furthermore, IBM SecureSlice technology divides one copy of the data into segments and distributes those data segments across multiple data centers for greater availability, cost savings and data protection.

In summary, IBM Big Replicate is a wide-area network (WAN) active replication technology that can deliver continuous availability, streaming backup and uninterrupted migration for hybrid cloud and burst-to-cloud deployments. Combined with IBM Cloud Object Storage, it can meet your most demanding recovery point objectives (RPOs) and recovery time objectives (RTOs) for analytics data storage and access across virtually any combination of Hadoop distributions including on-premises and in the cloud. You can also gain a significantly reduced TCO while receiving exceptional data continuity and high levels of data availability, durability and scalability so that users can focus on their analytics work and leverage enhanced IBM cloud services.

About IBM Big Replicate

Powered by patented technology, IBM Big Replicate enables active-active replication with guaranteed data consistency, virtually zero RTO/RPO and LAN experience at WAN distance. The solution provides continuous availability, streaming backup and uninterrupted migration to the cloud exceeding the most demanding enterprise service level agreements (SLAs). To learn more about IBM Big Replicate, please visit: ibm.com/us-en/marketplace/big-replicate

About IBM Cloud Object Storage

IBM Cloud Object Storage provides the flexibility, scalability and simplicity needed to store, manage and access today's rapidly growing volumes of unstructured data in a private, public or hybrid cloud environment. IBM's solutions can transform storage challenges into business advantages by reducing storage costs while reliably supporting both traditional and emerging cloud-born workloads for enterprise mobile, social, analytics and cognitive computing.

IBM Cloud Object Storage is built on technology from object storage leader IBM Cleversafe®, which was acquired by IBM in 2015. Some of the world's largest repositories rely on IBM Cloud Object Storage. To learn more about IBM Cloud Object Storage, please visit: ibm.com/cloud-computing/products/storage/object-storage

References

1 *Adoption of Object-Based Storage for Hyperscale-Scale Deployments Continues*, IDC Market Spotlight, November 2016. https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WUL12383USEN

2 The Cloudy Future of Hadoop: Hadoop is Here To Stay But The Cloud Will Change What It Is, Forrester Research, Inc., March 1, 2017

3 *Comparing WANdisco Fusion to Cloudera BDR, A Methodology and Results*, Appvance, 2016. https://www.wandisco.com/assets/blt4eb30f518208ab88/Appvance_ComparingWANdiscoFusiontoBDR_Final.pdf

IBM Corporation
IBM Cloud
Route 100
Somers, NY 10589

Produced in the United States of America
June 2017

Please Recycle

WUS12374-USEN-00