IBM Big Replicate

Enterprise-class replication for Apache Hadoop and object stores

Highlights

- Active-active data replication of Hadoop data and object stores
- Resynchronize your data automatically in the event of hardware and network outages
- Selectively and consistently replicate changes across different environments
- Automatic failover and recovery with no administrator or third-party intervention
- Collaborate with teams around the world as one, without disruption, and continue to use the same familiar tools

Organizations across the globe are intensely investigating Hadoop and object stores to better utilize and leverage data and take advantage of the significant cost savings Hadoop offers over traditional data warehouse solutions. However, some IT executives in the industry have shared concerns that Hadoop is not as mature as established database technologies. They point to specific perceived shortcomings preventing Hadoop from fully addressing both business and regulatory requirements. Foremost among these concerns are:

- **Backup reliability and data consistency across clusters and locations.** Replication, backup and recovery tools provided by the Hadoop distribution vendors are based on distributed copy (DistCp), which is primarily designed for copying files, not consistent replication of an entire cluster for backup and recovery.
- **Performance in distributed environments.** Without consistency across locations, the same applications will return different results depending on *where* they run, introducing business risk whenever decisions are made, based on out-of-date, inaccurate data.
- **Meet application service-level agreements (SLAs).** The deployment of critical, real-time analytics applications with stringent SLAs, on high-performance systems with extra RAM to support in-memory computing can be a daunting IT challenge.
- **Contain on-premises IT costs, and support offsite backup and disaster recovery.** Organizations need to find a solution to reduce their on-premises infrastructure costs rather than purchasing software and servers for use in-house that result in higher costs and excess capacity. The cloud is a way to use additional capacity only when it's needed.

IBM® Big Replicate helps ensure business continuity by providing real-time data replication between two or more Hadoop clusters with virtually zero recovery point objective (RPO) and recovery time objective (RTO). The patented replication engine provides data consistency with automated recovery and data synchronization in an outage scenario. Big Replicate offers flexible replication across supported Hadoop distributions and versions running on supported, compatible Hadoop storage systems and object stores.

IBM Big Replicate technology delivers:

- Virtually zero RPO and RTO
- Backup replication to object stores, such as S3 and IBM Cloud object stores
- Reliable data consistency
- Uninterrupted migration to the cloud and the hybrid cloud
- Performance that meets the most demanding enterprise SLAs

IBM Big Replicate industry-specific use cases

Financial industry

Overcome technical, business and regulatory barriers by using Hadoop for mission-critical applications.

Industry challenges

- Sustaining backup reliability and data consistency across clusters and locations
- Introducing business risk in distributed environments without data consistency across locations, resulting in the same applications returning different outcomes depending on where they run
- Meeting application SLAs
- Containing on-premises IT costs and supporting offsite backup and disaster recovery using the cloud

Solution overview

IBM Big Replicate provides a solution that enables a customer to realize the full potential of Hadoop on premises and in the cloud, while helping to ensure the firm can manage costs, exceed the most stringent SLAs and achieve regulatory compliance.

Solution highlights

- Continuous availability and data consistency make it easier to comply with stringent regulations for data access, bulletproof backup and data consistency.
- Multi-data center ingest has enabled real-time decision-making in remote offices far from the main data centers.
- Full utilization of compute resources eliminates the expense of underutilized servers for backup and disaster recovery, saving up to 50 percent on its hardware costs and scaling production Hadoop deployments without adding additional hardware.
- Selective replication allows applications to use global data for rollup analysis without moving sensitive information across borders, in violation of data security and privacy protection regulations.
- By using the cloud, data transfers as it changes, either on premises or in other cloud environments, with reliable data consistency.
- Flexible and futureproof cloud computing helps eliminate vendor lock-in.

Telecommunications industry

Improve data availability and simplify data flow while using Hadoop for data replication and transfer between clusters.

Industry challenges

- Multiple data centers complicated data flow. More than a terabyte of raw data flowed between the data centers each day.
- The velocity of data prevented effective disaster recovery. The high volume of new data created in several locations required a complex backup system. Backups and data transfers had to be scheduled during off hours.

Solution overview

Telcos use the analytical power of Hadoop to understand and improve the user experience for mobile and location services. They use the massive flow of data from network and consumer devices to improve coverage and build value-added, location-aware services.

Solution highlights

- Improved data availability. Selected vital data is replicated continuously and available on other clusters within seconds of ingest.
- Simplified data flow. Continuous active-active data replication between clusters replaces the transfer processes and makes data available for processing in near real time.

Utilities industry

Improve data backup and disaster recovery times in the Hadoop deployment while simplifying the data transfer process.

Industry challenges

- Lack of enterprise-grade availability
- Performance challenges
- Inability to reliably ingest all the data into a single cluster in one location as more data streams were brought online
- Complicated the data lake by requiring new data pipelines for the batch transfer of data between clusters

Solution overview

Firms in the utilities industry adopt Hadoop to power new initiatives for energy savings, operational efficiency and new revenue opportunities.

Solution highlights

IBM Big Replicate provides an active-active replication platform to address these challenges. Utilities organizations can achieve:

- Improved data availability because the data on the failover cluster is kept current within minutes and the system is restored automatically after a failure
- Improved performance by running replication continuously with no overhead on the cluster
- Improved efficiency by increasing resource utilization and load balancing
- Quicker data analysis because it is ingested and processed at any cluster and immediately available on other clusters

Insurance industry

Deliver the high levels of availability, performance and data consistency required from both a business and regulatory perspective.

Industry challenges

- Significant administrator involvement for setup, maintenance and monitoring
- Sharing resources among production clusters and other applications severely impacts performance
- Manual intervention to handle out-of-sync conditions
- Manual intervention after a system or network outage

Solution overview

Hadoop is a good fit for handling data ingest and analytic workloads in terms of cost effectiveness and performance.

Solution highlights

- This solution helps provide continuous availability and consistency with patented active-active replication for the lowest possible RTO and RPO across any number of clusters, any distance apart, whether on premises or in the cloud.
- Data is available when and where it's needed, and automatically resynchronizes as bandwidth allows.
- This solution helps eliminate read-only backup servers by making every cluster fully writable, as well as readable, and capable of sharing data and running applications regardless of location.
- Selective replication on a per-folder basis allows administrators to define replication policies that control what data is replicated between Hadoop clusters, on-premises file systems and cloud storage.
- Data security risks are minimized by working with all the on-disk and network encryption technologies available for Hadoop. Big Replicate only requires its servers to be exposed through the firewall for replication between on-premises data centers and the cloud. This process dramatically reduces the attack surface available to hackers.

In a challenging IT environment, enterprises must proactively address complex data sovereignty and compliance issues while continuing to extract value from data. IBM Big Replicate enables data availability, business intelligence (BI) and data science to control and automate what data is replicated to where, to help ensure data sovereignty requirements are met.

For more information

To learn more about IBM Big Replicate, please contact your IBM representative or IBM Business Partner, or visit ibm.com/bb-en/marketplace/big-replicate.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
March 2018

IMS14524-USEN-00