DataReplicator

ENTERPRISE DATA OPERATIONS AND ORCHESTRATION

Infoworks

Challenges

There are multiple issues organizations face that force the need to replicate data across big data clusters:

DISASTER RECOVERY: Company critical data often needs to be replicated to support scenarios that require disaster recovery. Even cloud data needs to be replicated in case of a disaster.

ON-PREMISE TO CLOUD: Organizations are moving their on-premise data to the cloud and now have multi-cloud and hybrid on-premise/cloud deployments that require replication of data across clusters to keep data properly synchronized.

CLOUD MIGRATIONS: Data migration from one environment to another when dealing with large data volumes is a complex process that takes time to complete. When organizations decide to swap out their big data fabric or run multiple fabrics simultaneously, they have to be able to move their data and keep clusters synchronized on an ongoing basis.

The Solution

Infoworks DataReplicator provides high-performance replication of data and metadata for large data volumes and a variety of scenarios including on-premise cluster to cluster, on-premise to cloud, and cloud to cloud. The Infoworks DataReplicator significantly simplifies data and metadata replications with a software-only solution that:

- Provides an automated, code-free, production-ready software platform for bi-directional, fault-tolerant data replication and removes the need for custom scripting
- Eliminates system downtime by incrementally synchronizing data without needing to stop source applications
- Scales to support petabytes of data
- Significantly simplifies ongoing operational management with a single orchestration platform
- Dynamically throttles bandwidth to optimally utilize network capacity
- Incrementally replicates data so clusters remain fully operational while data is being synchronized

CASE STUDY: CREDIT CARD FINANCIAL SERVICES FIRM

Challenges

- Replicate data from legacy Cloudera cluster to new production cluster, with zero downtime
- Over 3 petabytes of data to be migrated while existing applications continue to use and change data in the source cluster
- Make optimal use of network capacity between clusters, while not consuming the entire bandwidth

Alternatives

- The company evaluated open source tooling that required many months of services, was limited in functionality and could not do bi-directional synch

Results

- The project required only 6 weeks to replicate 3 petabytes of data including installation, replication, and validation
- Replication was automatically throttled to optimize bandwidth and eliminate interference with normal business operations

info@infoworks.io

(650) 391-9306

Infoworks

Capabilities Summary

CAPABILITY	DESCRIPTION
No code, automated data replication	Easy to use. Data engineering expertise is not required Data and metadata are both automatically and incrementally synchronized
Incremental, bi-directional replication	As data changes in the file system, the system detects incremental changes and replicates the changes Supports replication of data in two directions including reconciliation of data as data sources are updated
Enterprise scalability with dynamic network throlling	Works in the most data intensive environments. In real world use cases, 1 PB of data was replicated in 24hours with hundreds of data replication jobs Replication and delta computation can be automatically run in parallel Provides fully configurable bandwidth throttling to minimize disruption of normal business operations
Cross-cloud and on-premise orchestration & management	Built in orchestration with a single interface for management of replication across multiple clouds
Start, stop and restart	Ability to start, stop and restart from the last known point

Cloud Migrations

- Migrate on-premise clusters to cloud, or cloud to cloud migrations
- Operate cloud and on-premise in parallel during migration without downtime

Hybrid Cloud Replication

- Synchronize on-premise data to the cloud for transient or ongoing big data processing
- Utilize cloud for elastic burst into the cloud, and process in either location as needed

Disaster Recovery

- Continuously replicate cluster data to an active, highly–available backup cluster in the cloud or on-premise
- Support on-premise or cloud-based disaster recovery with bidirectional switchover from primary to backup

About Infoworks

Infoworks provides the first Enterprise Data Operations and Orchestration (EDO2) software system to automate the development and operationalization of data pipelines from source to consumption in support of business intelligence (BI), machine learning (ML) and artificial intelligence (AI) applications. Infoworks' code-free development environment allows organizations to develop and manage end-to-end data workflows, or migrate existing data and workflows, without requiring an army of big data experts. Infoworks delivers capabilities to automate and simplify development of data ingestion, data preparation, query acceleration and ongoing operationalization of production data pipelines at scale. Infoworks supports cloud, multi-cloud, and on premise deployments, enabling customers to deploy projects to production within days, dramatically increasing business agility and accelerating time to value.

info@infoworks.io

(650) 391-9306
DR201907