wanDISCO

The World Leader in
Active Data Replication™

WANdisco Fusion

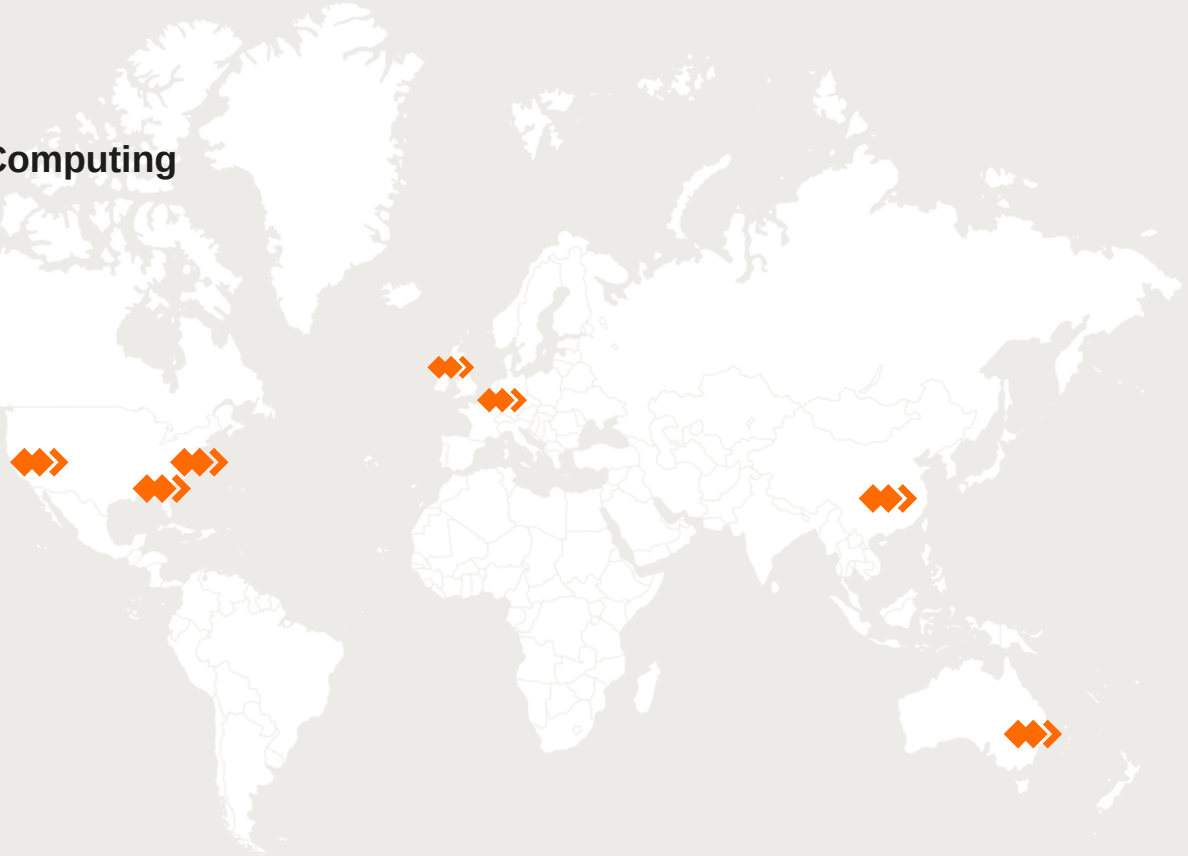
Use Cases and Customer Case Studies

Steve Kilgore
Director, Global Partner Solutions Architecture

WANDISCO

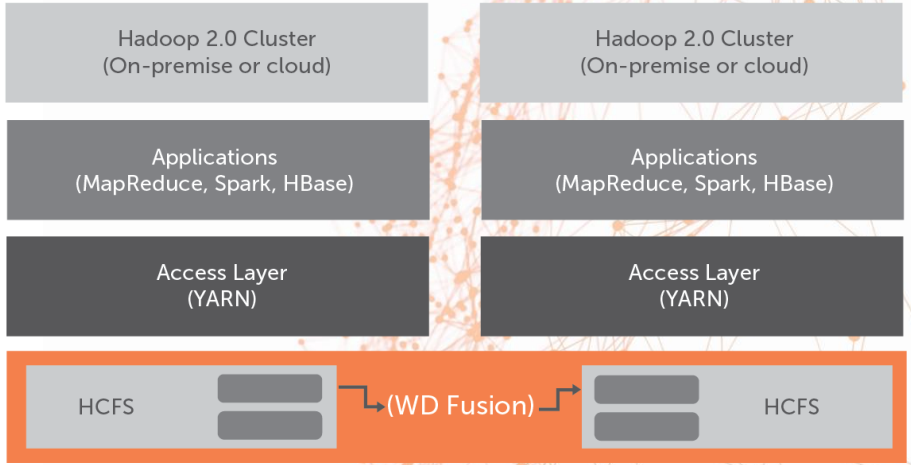
Wide Area Network Distributed Computing

- WAN optimized active-active replication technology
- Founded in 2005. IPO in 2012 (LSE:WAND)
- Apache Software Foundation sponsor, member of the Open Data Platform
- Global presence with 24/7 support



WHAT IS WANDISCO FUSION?

- Active-active replication of Hadoop data across multiple appliances, clusters, distributions
- WAN-capable for geo-replication
  - Selective, opt-in, cross-distribution
- Enabler for extending Big Data solutions – on premise and cloud
  - Can replicate between on-premises and cloud environments
  - Enables bursting from on-premises hardware into Cloud Services
  - Drives adoption of Cloud services
- Supports all major Hadoop distributions and versions, operating as an application without change to the underlying cluster



Fusion Use Cases

Use Cases

Disaster Recovery

Migration

Multiple Ingest Points

Multiple Zones

- QA/Test/Dev
- Resource Isolation
- Security Isolation

Cloud/Hybrid

NFS, Object Store and HDFS

Disaster Recovery

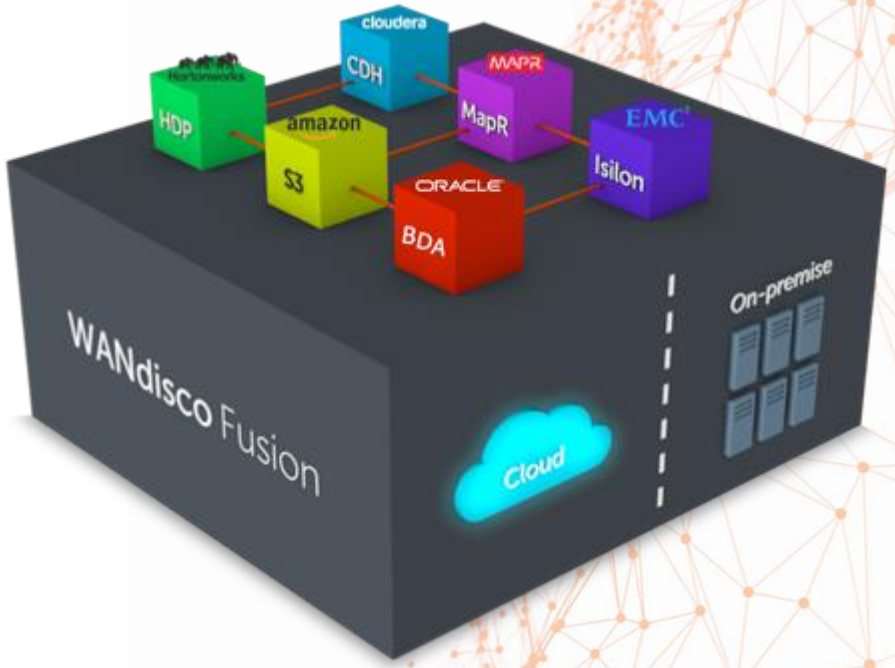
- Data is as current as possible (no periodic synchs)
  - Low RPO
- Virtually zero downtime to recover from regional data center failure
  - Active/Active means you're already running in both locations
- Meets or exceeds strict regulatory compliance around disaster recovery
  - Auditability

Fusion Active Migration

- Migrate between Hadoop distributions
- Staggered upgrades between versions of distributions across multiple clusters and data centers
- WANdisco Fusion ensures continuous data integrity and security with fast, seamless migration from distribution to distribution.
- There is no downtime or disruption during migration
  - Migrate users, applications and data in phases
  - Test things in parallel

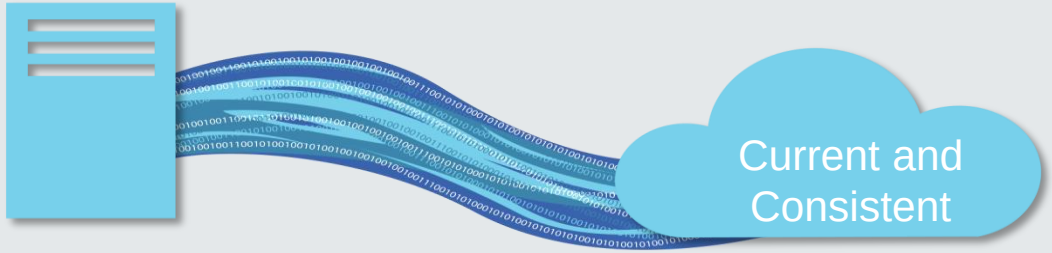
Fusion Active Migration

- Migrate without disruption
- Between distributions
- Between versions
- HCFS compatible file systems
- Cloud Object Storage
- In/Out of the Cloud
- Between Clouds
- No downtime
- No data loss

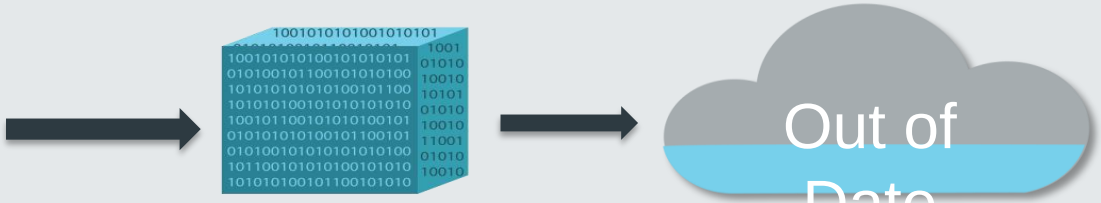


Migrations – Cloud

WANdisco uses
active transactional data
replication to
move data as it changes



Every other solution is
time-based
and batch risking data loss,
downtime and business
disruption



Fusion Cloud Online Migration: Step by Step

- Induct new cluster into WANdisco Fusion as a new zone.
- Establish memberships including new zone and define replication rules for desired directories.
- Use Repair functionality to synchronize updates from source to new cluster on a Replication rule by Replication rule basis.

Fusion Cloud Offline Migration: Step by Step

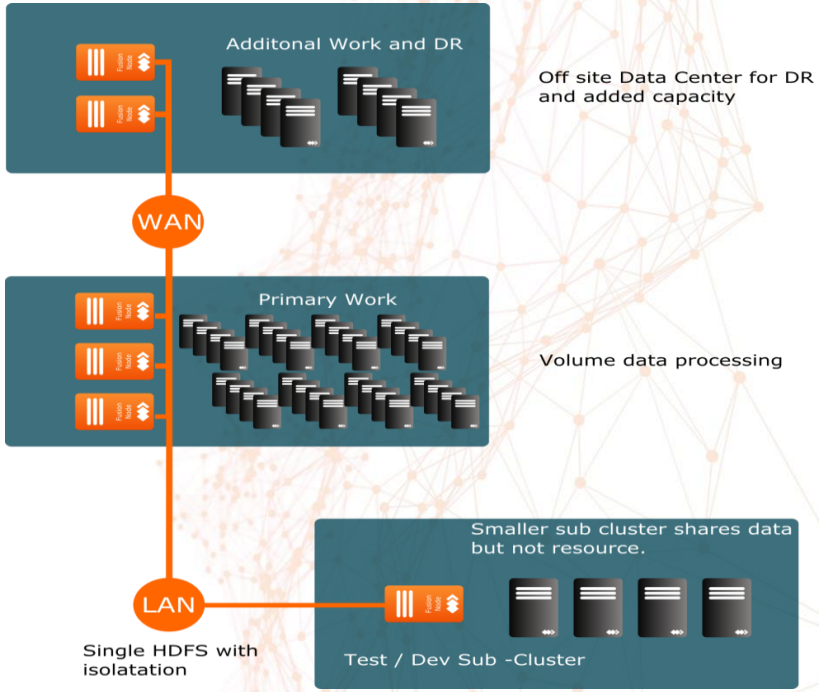
- Copy data from source Hadoop cluster to portable storage.
- Ship storage to destination and upload into new cluster.
- Induct new cluster into WANdisco Fusion as a new zone.
- Establish memberships including new zone and define replication rules for desired directories.
- Use Repair functionality to synchronize updates from source to new cluster on a Replication rule by Replication rule basis.

Multiple Ingest Points

- Ingest and analyze anywhere
- Analyze Everywhere
  - Fraud Detection
  - Equity Trading Information
  - New Business
  - Etc…
- Backup Datacenter(s) can be used for work
  - No idle resource

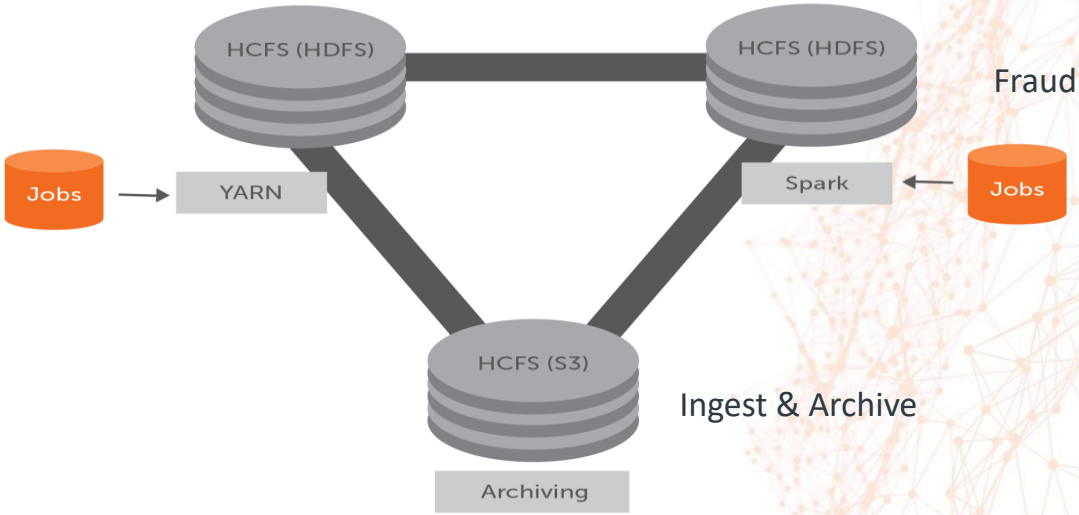
Multiple Zones – QA/Dev/Test

- Share data, not processing
  - Isolate lower priority (dev/test) work
  - Share data not resource
- Maximize Resource Utilization
  - No idle standby
- Mixed Hardware Profiles
  - Memory, Disk, CPU
  - Isolate memory-hungry processing (Storm/Spark) from regular jobs
- Mixed Vendor
  - **CDH, PHD, HDP, MAPR, AliCloud, AWS, AZURE, Google, etc.**



Multiple Zones – Resource Isolation

Customer Analytics

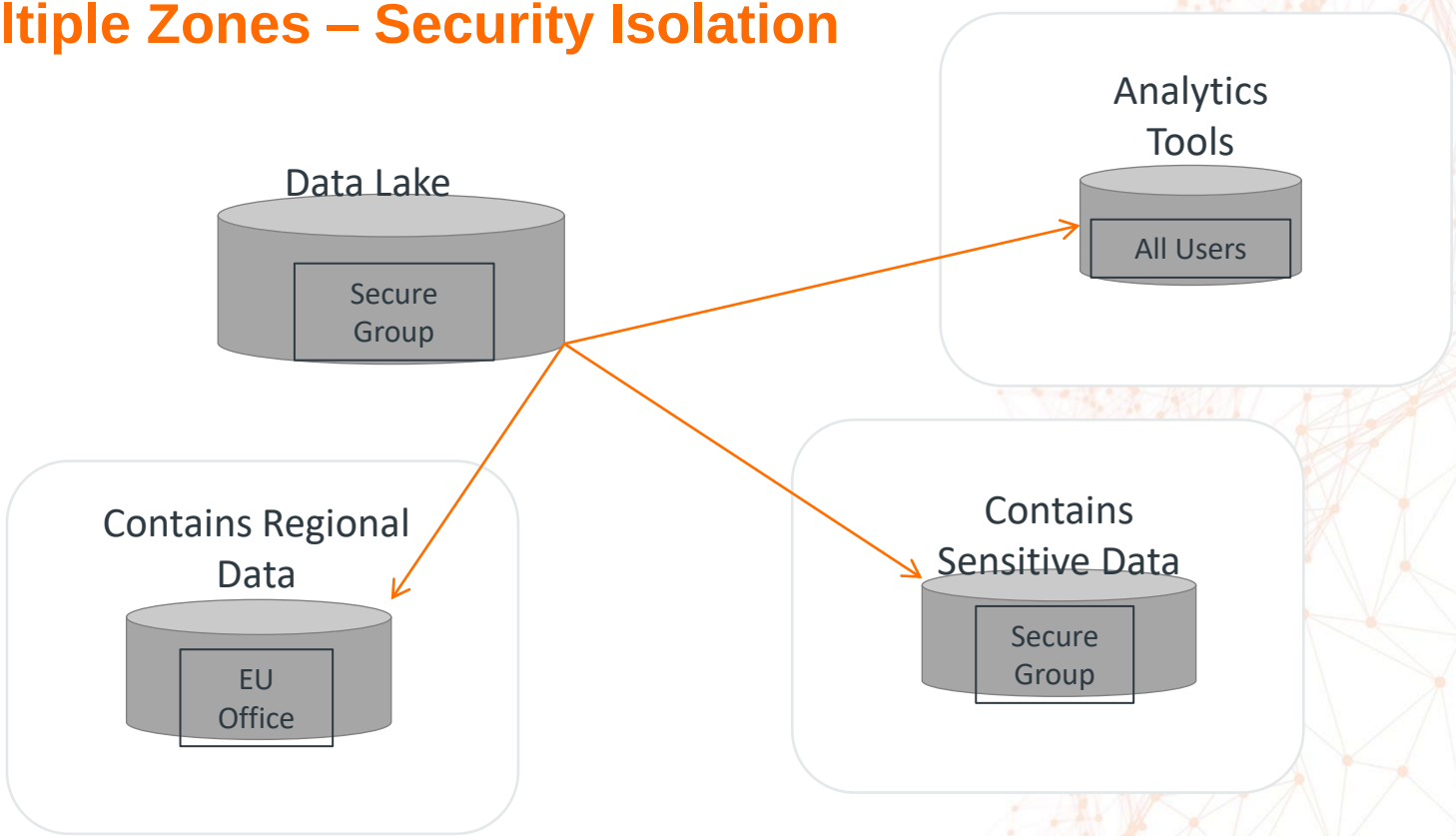


Fraud Analytics

Ingest & Archive

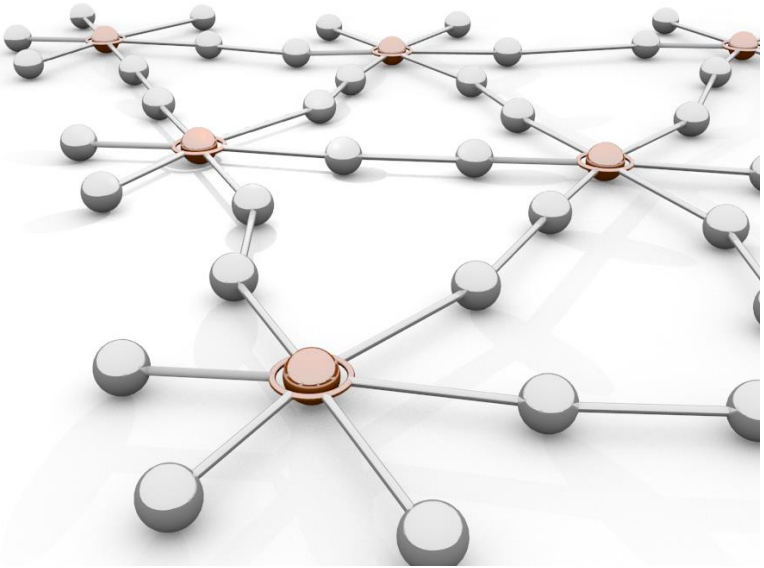
- Provides protection to Hadoop jobs
- Targeted workloads on specific hardware (e.g. Spark jobs on high memory systems)

Multiple Zones – Security Isolation



Security Between Data Centers

- Hadoop clusters do not require direct communication with each other.
  - No n x m communication among datanodes across datacenters
  - Reduced firewall / socks complexities
- Reduced Attack Surface
- Fast network protocols can keep up with demanding network replication



Data Lakes

Use Cases

- 360-degree view of global supply chain
- Trend analysis and predictive analytics
- Compliance monitoring

Requirements

- Data integration
  - Share data globally across a variety of storage platforms at multiple locations
    - Hadoop Clusters running on a mix of distributions, versions and storage
    - Multiple Cloud vendors
- Data ingest across disparate sources and locations
- Data consistency across sources and locations
- Respect data protection and privacy regulations
- Deliver actionable data in minutes
- 24x7 operation with no downtime or data loss

Real-time Analytics

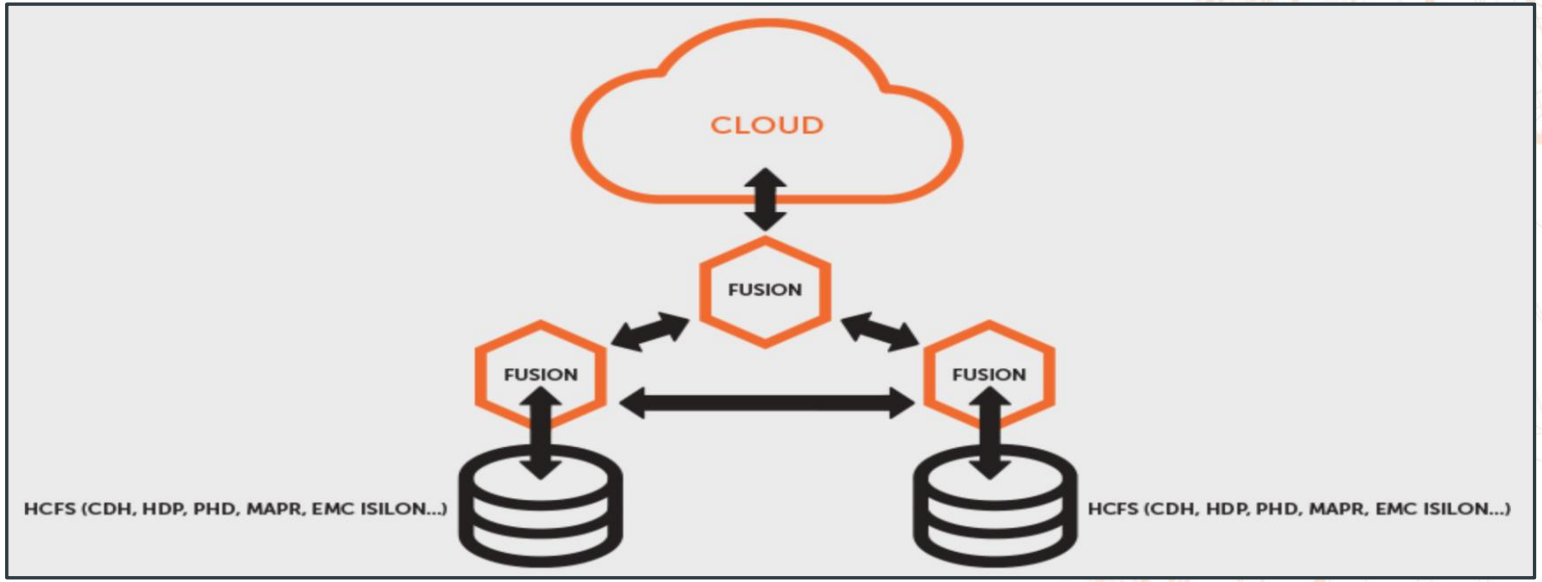
Use Cases

- Fast Data Applications based on volatile event streams
- High frequency securities trading
- Credit card fraud detection
- Industrial sensor data analysis

Requirements

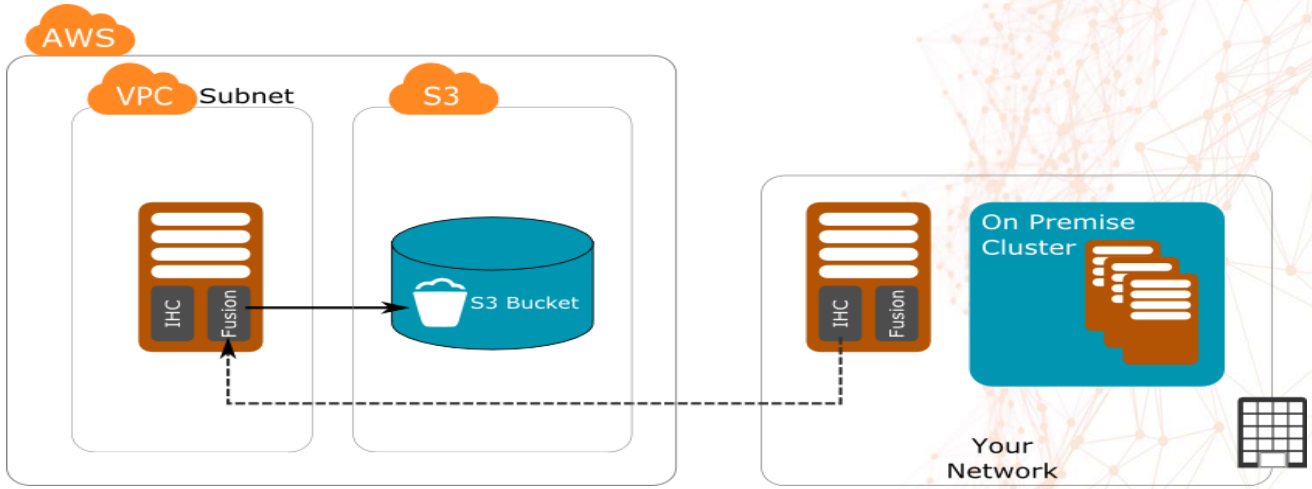
- Ingest Fast (streaming) data from multiple sources and locations simultaneously
- Real-time access to data lake to provide historical context
- Support complex analysis pipelines with data protection requirements
- 24x7 operation with no downtime or data loss

Cloud and Hybrid Use Cases



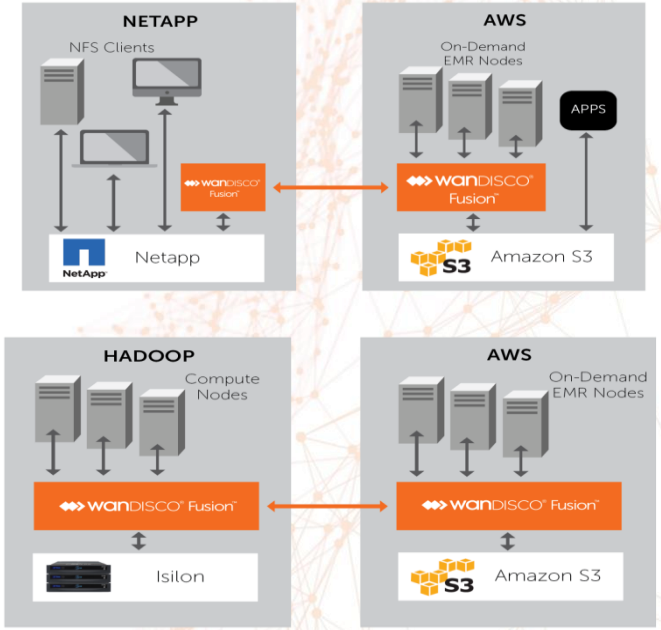
Cloud and Hybrid Use Cases – AWS Examples

- Continuous replication to S3 keeps data relevant
- Continuous replication allows high volume data transfer to keep up with ingest rate
- Continuous replication minimizes WAN spikes common with periodic replication methods



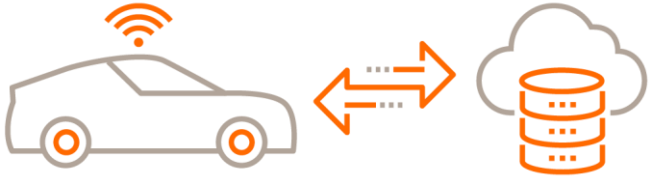
NFS, Object Store and HDFS

- Moves data from any local Posix or NFS file system mounted on Fusion server
- When you define replicated folders you'll view the file system
- AWS has the same configuration as other cloud environments
- Any Combination to Any Combination
  - HDFS→NFS for backup
  - NFS→HDFS for import
  - HDFS/NFS→Object Storage for Cloud Migration
  - Object to Object to migrate from one cloud to another



Customer Case Studies

Big Three US Automotive Manufacturer



- CHALLENGE
  - Autonomous vehicle project generating over 200TB of fast streaming data per day
    - Ingested into clusters at 2 locations to balance analytics workload
    - Needed active-active multi-data center ingest to keep clusters continuously in sync and available
    - Wanted each cluster to provide backup and recovery for the other
  - Tried dual ingest to simulate active-active
    - Too unreliable, admin heavy and unable to handle data created independently on each cluster
- SOLUTION
  - Selected Big Replicate (IBM OEM version of WANdisco Fusion) for:
    - Continuous synchronization with guaranteed consistency across both clusters running under load
    - Full active-active read/write access to both clusters with auto-recovery between them for continuous availability
    - Scalability as the clusters grow in size to over 400PB each

US Bank #1, HQ in Boston, MA (Fortune 300)

- CHALLENGE
  - Tested Cloudera BDR (included with Oracle BDA) and found it unable to meet business and regulatory SLAs.
    - Backups could only run every 24 hours due to resource contention, risking significant data loss after an outage.
- SOLUTION
  - Fusion exceeds SLAs with virtually zero downtime and zero data loss
    - Began production rollout within 30 days after delaying for nearly two years due to lack of an acceptable solution
    - Using Fusion to replicate 20TB per day across two data centers, with plans to expand to four sites
    - Licensed for six racks of BDA.

US Bank #2, HQ in Atlanta, GA (Fortune 350)

- CHALLENGE
  - Required active-active Hive and HDFS replication to keep 800TB Oracle BDA (CDH) clusters ingesting 2TB/day in sync across two data centers
  - Found Cloudera BDR too admin heavy and unable to meet RTO/RPO SLAs.
- SOLUTION
  - After extensive comparison testing of BDR, Fusion, and dual ingest, Fusion was selected
  - Fusion delivers virtually zero downtime and zero data loss vastly exceeding their sub-15 minute RTO/RPO SLAs.

Large US Mutual Insurance Company, HQ in NYC (Fortune 250)

- CHALLENGE
  - Required DR solution for Pivotal, but Falcon required high admin overhead and couldn't meet SLAs for RTO/RPO
  - DR cluster had to run in standby read-only mode to avoid divergence, so they could not get full value from their DR hardware
- SOLUTION
  - Selected Fusion after testing other software and far more costly hardware options from EMC
  - Fusion exceeded their SLAs and enabled them to fully operationalize Hadoop so limited admin staff can easily support it
  - After starting with DR, they now make full active-active use of their three 100TB clusters for analytics

Global Technology Memory and Storage Vendor

- CHALLENGE
  - Keep three 1.2 PB HDP clusters with up to 30 million transactions per day continuously in sync with no admin overhead and no downtime, before and after one cluster moves to another data center
    - Falcon (included with HDP) could not meet their requirements
- SOLUTION
  - Fusion enabled their data center move without downtime
  - They continue to use Fusion for cluster synchronization and DR
  - Fusion enables them to continuously monitor their clusters and the status of data replication across them

Large US Biotechnology Company (Fortune 250)

- CHALLENGE
  - Needed to continuously synchronize two 100TB HDP clusters, growing to 500TB in size, and addition of a third cluster with no downtime
  - Falcon (included with HDP) could not meet their SLAs, and dual ingest was too unreliable
- SOLUTION
  - Deployed Fusion and now use both clusters in full read/write mode
    - Each cluster can failover to and recover automatically from the other without downtime or data loss
  - Testing proved clusters and sites can be added dynamically without disrupting their live implementation.

Global Financial Data Aggregation and Analytics Company

- CHALLENGE
  - Needed to upgrade their production CDH 4.4 clusters to CDH 5.7 without downtime, which was impossible with Cloudera's tools due to significant differences between the releases
- SOLUTION
  - Fusion allowed their existing CDH 4.4 environment to stay fully operational during the upgrade to CDH 5.7 enabling them to test applications in parallel in both environments
  - With Fusion, they now use their former DR cluster in fully active read/write mode, scaling their deployment 2X with existing hardware for major cost savings

Largest Canadian Bank

- CHALLENGE
  - Need guaranteed data consistency and continuous availability across multiple 400 node 5PB real-time clusters to balance workload
  - HDP Falcon, couldn't meet business or regulatory SLAs for RTO/RPO, or guaranteed consistency
  - Standby clusters used for backup and recovery had to be read-only, preventing them from getting full utility from their hardware
- SOLUTION
  - Fusion ensures data consistency across all clusters and enables continuous availability, exceeding SLA's.
  - Fusion enabled full read/write use of their former standby clusters, so they could scale up their deployment without additional hardware expense

Largest UK Grocery Retailer

- CHALLENGE
  - UK's largest grocery retailer running big data analytics on HDP to track customer behavior and anticipate demand at store level and support eStore customers
  - Required continuous availability and performance to guarantee positive experience for millions of online shoppers
  - HDP Falcon required hours of downtime resulting in significant lost revenue
- SOLUTION
  - Fusion enabled continuous availability and data consistency across their two locations
  - Former standby cluster is now fully active enabling them to process online orders at both data centers, improving customer experience.

Healthcare Analytics Provider

- CHALLENGE
  - Customer offers HIPAA compliant cloud-based predictive health care analytics
    - Collects over 200 TB of medical history and real-time patient sensor data each day
  - HIPAA regulations require continuous availability and current backup
    - Cloud vendor offered periodic one-way synching between active and backup data centers
    - Read-only backup data center hours behind meant risking data loss and paying for storage they couldn’ t fully use
- SOLUTION
  - Deployed Fusion in cloud vendor’ s data centers
  - Both data centers are fully active and always in sync
    - Virtually zero RTO/RPO
    - Analytics apps run in both cloud data centers

UK University

- CHALLENGE
  - Move continually changing medical data for 6000 dementia patients between 8 cloud vendors and 6 HPC providers for a variety of analysis
  - Required guaranteed consistency and auto-recovery between on-premises and external sites
  - One-way high speed data transfer solutions
    couldn’ t give them the resilience or data consistency they required.
- SOLUTION
  - Deployed Fusion on-premises and at each external site
  - Fusion’ s active data replication moved raw as well as result data between any combination of sites as it changes
  - Fusion delivers guaranteed consistency with auto-recovery after network or hardware

SUMMARY: FUSION USE CASES - 1

Disaster Recovery

- Reduce RPO and RTO from hours to minutes
- WAN-capable HA/DR
- Central control of data location

100% Use of Cluster Resources

- Run applications on every cluster and appliance
- Maintain cost advantage of Hadoop (convert overhead to production assets)

Truly Hybrid Cloud

- Easy data flow for burst-out processing
- Ingest at multiple data centers, analyze everywhere
- Run applications on the right cluster for each purpose (e.g. Isilon, batch, stream, ingest, sandbox) and maintain performance
- Reduce network security risk and management
- On-premise, Cloud & Hybrid support
- Easier upgrades and migrations - zero down time

SUMMARY: FUSION USE CASES - 2

Cluster Migration

- Both data and underlying metadata are replicated selectively while operation continues in original cluster
- Validate application behavior in target environment with zero risk to continued operations. Rollback at any point, immediately if necessary.

Continuous Availability and Performance

- Delivers continuously synchronized data with automated failover and disaster recovery over LAN and WAN
- LAN-speed read/write access to the same data at every location
- Support for different versions of Hadoop enables continuous operation with staggered upgrades across clusters and data centers

MORE INFORMATION

WANdisco website: http://www.wandisco.com

Software Download https://www.wandisco.com/free-trial

WANdisco contacts:

- Felix Fong, Senior Solution Architect: felix.fong@wandisco.com, +61 411 267 609
- Peter Scott, SVP Business Development: peter.scott@wandisco.com, 925 219 2771
- Steve Kilgore, Director Global Partner Solutions Architecture: steve.kilgore@wandisco.com, 925 365 0641