THE CLOUD IS AN IMPORTANT COMPONENT OF MODERN ENTERPRISE DATA SYSTEMS

Sharing data between on premise and Cloud storage

• Data is generated on premise, and in the cloud
  – Analysis makes most sense when all data is included
• Replication of data across the WAN is challenging
ON-PREMISE BIG DATA

• Hadoop is the most popular system
• Storage is co-located with compute (HDFS)
• Batch processing
  – Ingest data once in a while, then run compute
    (SQL like queries) using Hive (Map Reduce or Spark)
• Real time
  – Data is stored in Key Value stores such as HBase
• (New) Stream processing (Spark Streaming and Storm)
  – Data is continuously streaming in, and is appended to existing data
  – Analysis is performed continuously on data from a slice of time in the past (micro batch)
ON-PREMISE BIG DATA – NOMINAL ARCHITECTURE

• Machines are loaded with disk, memory and CPU
• Compute is co-located with disk, hence network is less significant
• Cluster is Always On, multi tenancy at the Hadoop layer
ON PREMISE BIG DATA – HDFS CANNOT OPERATE OVER THE WAN

HDFS is limited to operating in a single data center
The naive approach: Why not drop some DataNodes in a remote Data Center?
• DataNode heartbeats over the WAN do not work well
• Write pipeline that includes a WAN link is fragile
CLOUD BIG DATA

• Amazon AWS is the most popular cloud
• Two modes of operation in AWS
  1. Shared Storage, accessed through a network connection
     1. EMR (Hive, etc.) with S3 for storage
     2. Local Storage (DAS)
        1. Redshift Enterprise Data Warehouse
        2. Hadoop with HDFS on instance storage
CLOUD BIG DATA
SHARED (NETWORK) STORAGE ARCHITECTURE

- EMR on AWS with S3 Storage
- Machines are loaded with CPU and memory
- Machines do not have much storage
- Storage is accessed through a network connection
  - We’ve come a full circle back to networked storage
  - Storage network needs to be strong
  - Modern concepts such as mesh networking are being investigated
- Compute nodes are transient, i.e. VMs
- Start up a number of VMs, run your job, then terminate VMs
- Multi-tenancy is accomplished using VMs
CLOUD OBJECT STORES ARE ALSO SINGLE DATACENTER SYSTEMS!

- On S3, you choose a Region for your bucket
- Replication can be configured between regions, however it is eventually consistent
THE DATA REPLICATION PROBLEM

• Different Big Data Solutions work well for different problems
• Keeping data replicated and consistent across these different solutions is hard
• Solutions such as Distcp result in inconsistencies
• The Volume of Big Data is such that manual reconciliation of inconsistent data is impossible
THE DATA REPLICATION PROBLEM
HACKS IN USE TODAY: PERIODIC SYNCHRONIZATION

DISTCP
• Runs as Map reduce
• DR Data Center is read only
• Over time, Hadoop clusters become inconsistent
• Manual and labor intensive process to reconcile differences
• Inefficient use of the network
THE DATA REPLICATION PROBLEM
HACKS IN USE TODAY: PARALLEL (DUAL) INGEST

- Hiccups in either of the Hadoop cluster causes the two file systems to diverge
- Potential to run out of buffer when WAN is down
- Requires constant attention and sys-admin hours to keep running
- Data created on the cluster is not replicated
- Use of streaming technologies (like kafka) for data redirection are only for streaming
WANdisco Fusion: Paxos based continuous and strongly consistent replication of data to and from any Hadoop to any Cloud

WANdisco’s products leverage our patented replication engine to achieve unmatched speed and resilience in global Hadoop, Subversion, and Git deployments.
WANdisco’s patented WAN capable paxos implementation

- Mathematically proven
- Provides distributed co-ordination of File system metadata
  - Active/Active (All locations)
  - Create, Modify, Delete
  - Shared nothing (No Leader)

- No restrictions on distance between datacenters

- US Patent granted for time independent implementation of Paxos

Not based on SAN block device synchronization such as EMC SRDF

Paxos is an algorithm for solving consensus in a network of unreliable processors. Consensus is the process of agreeing on one result among a group of participants. This problem becomes difficult when the participants or their communication medium may experience failures.
How DConE Works
WANdisco Active/Active Replication

Majority Quorum
• A fixed number of participants
• The Majority must agree for change

Failure
• Failed nodes are unavailable
• Normal operation continue on nodes with quorum

Recovery / Self Healing
• Nodes that rejoin stay in safe mode until they are caught up

Disaster Recovery
• A complete loss can be brought back from another replica
STRONG CONSISTENT REPLICATION OF DATA BETWEEN HDFS INSTANCES

HDFS Clients (Hive, Spark)

Fusion HDFS Proxy

HDFS

PAXOS

HDFS Clients (Hive, Spark)

Fusion HDFS Proxy

HDFS

PAXOS

HDFS Clients (Hive, Spark)

Fusion HDFS Proxy

HDFS

PAXOS

HDFS Clients (Hive, Spark)

Fusion HDFS Proxy

HDFS
Fusion workflow

1. User makes a request to create/change a file
1. User makes a request to create/change a file
2. Fusion coordinates File Open to other clusters involved (membership)
1. User makes a request to create/change a file
2. Fusion coordinates File Open to other clusters involved (membership)
3. File is added to underlying storage
Fusion workflow

1. User makes a request to create/change a file
2. Fusion coordinates File Open to other clusters involved (membership)
3. File is added to underlying storage
4. Fusion coordinates at configurable write increments and File Close with other clusters (membership)
Fusion workflow

1. User makes a request to create/change a file
2. Fusion coordinates File Open to other clusters involved (membership)
3. File is added to underlying storage
4. Fusion coordinates at configurable write increments and File Close with other clusters (membership)
5. Fusion server at remote cluster pulls data from IHC server on source cluster
Fusion workflow

1. User makes a request to create/change a file
2. Fusion coordinates File Open to other clusters involved (membership)
3. File is added to underlying storage
4. Fusion coordinates at configurable write increments and File Close with other clusters (membership)
5. Fusion server at remote cluster pulls data from IHC server on source cluster
6. Fusion server at remote site writes data to its local cluster
VOLUME, VELOCITY AND VARIETY OF DATA

WANdisco Fusion is designed specifically to tackle the three Vs of Big Data (courtesy Gartner)

Volume: Synchronous Metadata and Asynchronous Data

Velocity: Boxcars (batching), and pipelining

Variety: Fusion Plugins
VOLUME: SYNCHRONOUS METADATA AND ASYNC DATA

- Metadata operations such as File Create, Append etc. are coordinated using DConE (Paxos)
VOLUME: SYNCHRONOUS METADATA AND ASYNC DATA

- Asynchronous data write
What happens if the client renames the file before, during or after the remote data center starts pulling the file asynchronously?!?
The ExecutedAgreementsCache
A cache of agreements executed – helps the remote end find out what happened to the file it was pulling
VELOCITY: BATCHING PROPOSALS A.K.A BOXCARS

Batch a number of proposals into a single Paxos agreement
VELOCITY: BATCHING PROPOSALS A.K.A BOXCARS

Fusion Server 1

m c d

boxcar

Fusion Server 2

a c m

boxcar

Fusion Server 3
Amortize the cost of a Paxos agreement over multiple proposals
VARIETY: FUSION PLUGIN ARCHITECTURE

- Mapper/Reducer
- Fusion HCFS Client
- S3 Client
- S3 Client
- S3 Client
- S3 Network Proxy
- HCFS
- S3
- Hive
- NFS

FUSION SERVER
WEBLOG PROCESSING STANDARD ARCHITECTURE

Enterprise DC is effectively a single point of failure
FUSION EXAMPLE USE CASE: WEBLOG PROCESSING

Cloud Region 1
Web Server
Web Server
Web Server

S3

File folder level replication using Fusion

Cloud Region 2
Web Server
Web Server

Spark/Hive
HDFS

Enterprise Data Center
WANDISCO FUSION: RECAP

- Strongly consistent replication of data from any storage system to any other storage system, across the WAN
- Delivered as software installed on premise, or as a replication service in the cloud
Thanks for attending
jagane.sundar@wandisco.com