HDFS RBF and Storage Tiering

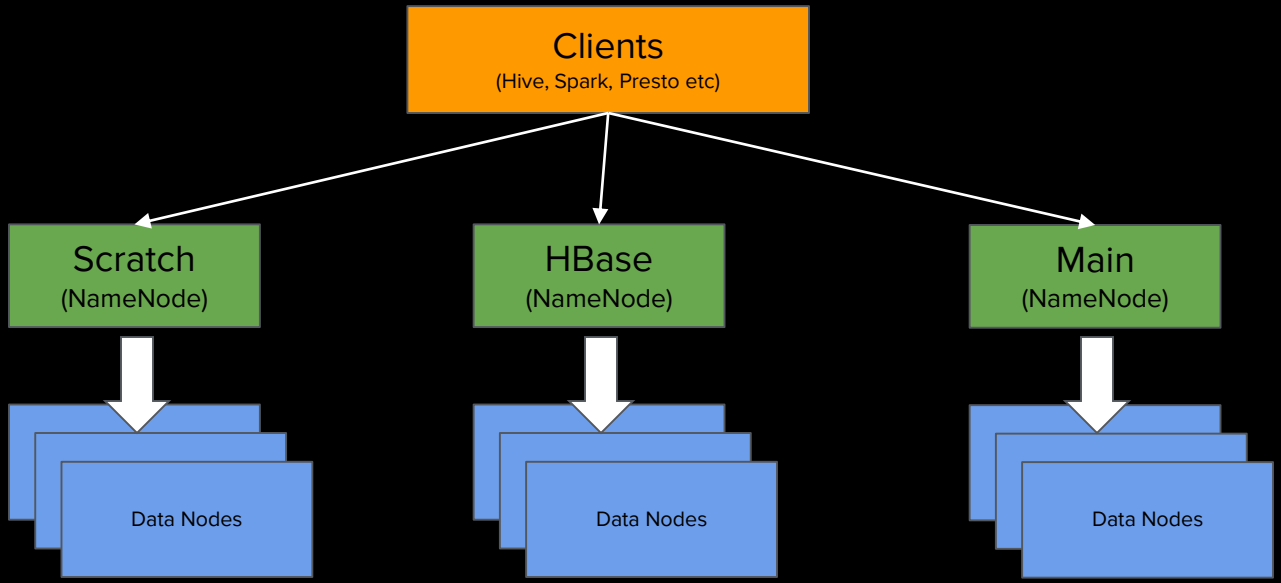
CR Hota
Ekanth Sethuramalingam

January 30th,
2019

Agenda

- Current set-up
- New Use cases
- Router Based Federation
- Security in Router Based Federation
- Warm tier use case Deep-dive

Current Set-up



Tens of Nodes
~10 M files

~100 Nodes
~1 M files

~3500 Nodes
~300 M files
~150 PB

Current Set-up continued ....

- Running multiple clusters using ViewFS
- Viewfs helped hide a small secured and larger non-secured cluster.
- For the most part it got the job done.

... But new cases came up which needs us to re-look at viewfs

New Use cases

- Tiered (Warm) HDFS
- Sharding (Ingestion vs ETL)
- Security - All clusters need to be secured.
- Erasure coding
- Operating multiple versions of hdfs.

Router Based Federation (aka RBF)

- Introduced in Hadoop 2.9.0 and originally developed at Microsoft
- Server side mount table
- Unified view of filesystem

Deployment and stabilization

- POC
  - Routing hive scratch dir to *scratch* cluster
- Backported 2.9 to 2.8.2
- Migrated and now use 3.1 (for EC compatibility)
- Added various fine grained metrics internally for measurement at each step

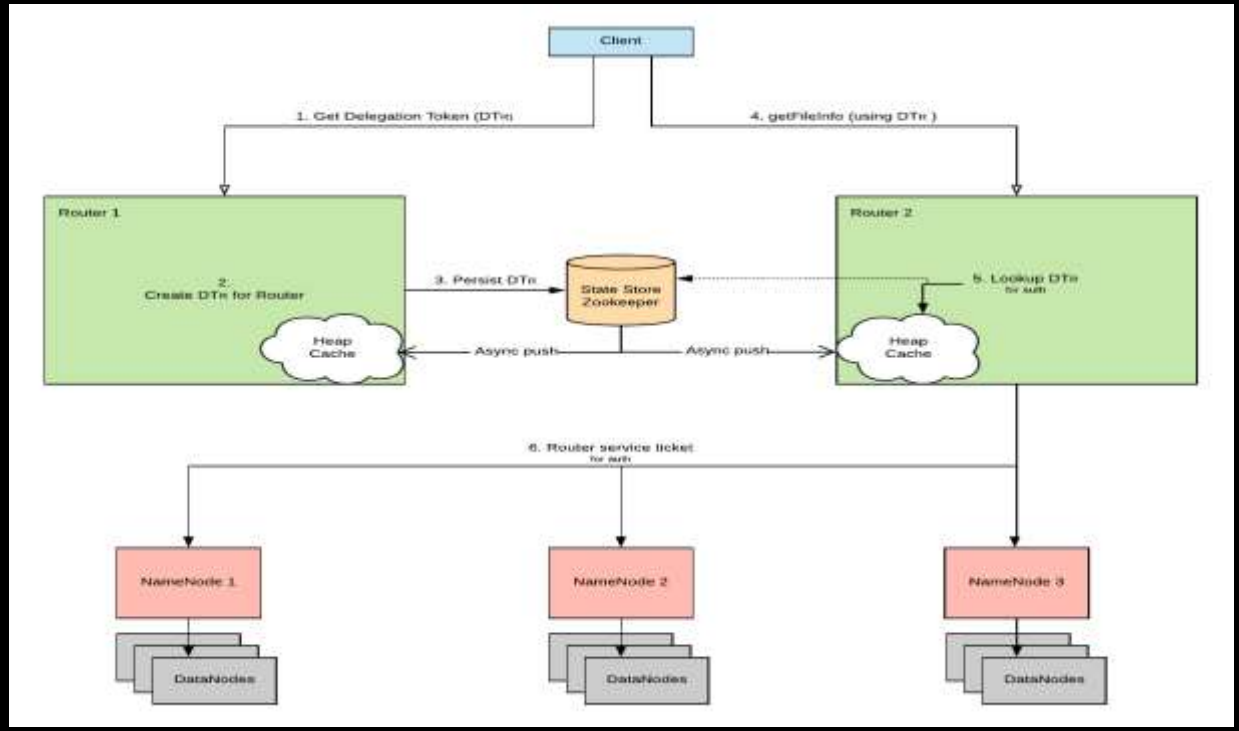
Deployment and stabilization conti ...

- Testing
  - Custom python scripts for web and java for RPC
- Critical bugs
  - HDFS-13834 and HDFS-13637
- Performance bugs
  - HDFS-13230 and HDFS-13232
- Tuning parameters, such as thread pool size, size of connection pools, clean up durations etc.

Security in RBF

- All clusters and services need to be secured at Uber.
- Backported kerberos patch (HDFS-12284)
- HDFS-13532
  Design and development of delegation model.
  - Tested by community
  - Collaboration with Microsoft, LinkedIn, Hortonworks
  - Initial patch available

Security in RBF continued ...



Next steps/Enhancements

- *HDFS-14090*: RPC Queue isolation/throttling across sub clusters.
  - Slow namenode issue
- *HDFS-14118:* Routers behind DNS
- *HDFS-13522:* Support for Observer/Read-Only NN
- *HDFS-13633:* Better connection management - Sync etc
  - Fine grained configs per user
- *HDFS-xxx:* Support for rename handling/custom exception
- *HDFS-xxx:* Database token store implementation

Community Support

Microsoft

LinkedIn

Hortonworks

Cloudera

Huawei

Others ...

Where were we

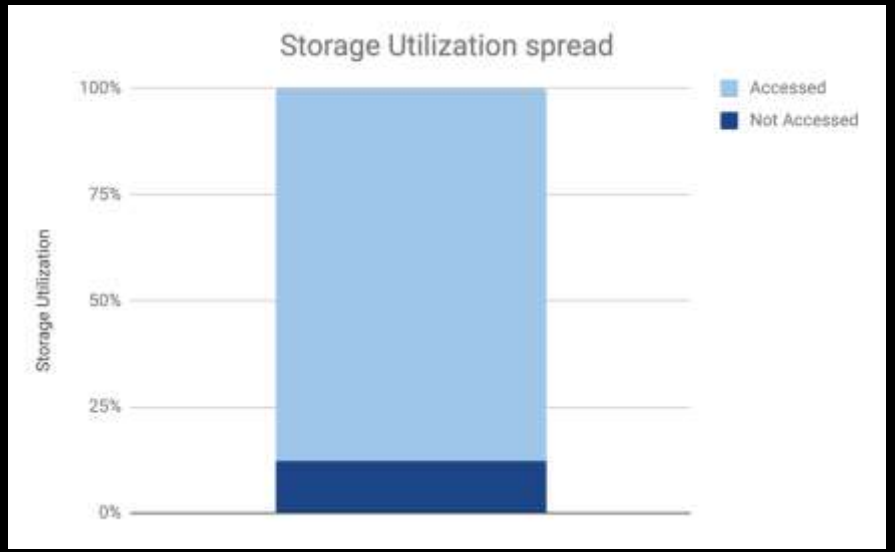
- Disaggregated compute and storage (late 2017 onwards)
- Storage needs continue to grow faster
- Teasing single cluster limits (namenode memory, RPC latency)

Storage Tiering

- Increase cost efficiency
- No changes to SLA guarantees for users
- Support a multi-tiered model for storage
- Bonus: solution augments solving cluster scaling challenges

Why Warm?

- Most of the data is accessed
  - Not really cold



Analysis and Observations

- Average resource utilization low (IO and network util)
- Temporal properties exist
- Large datasets in external tables
- Use high density storage

How?

Choices

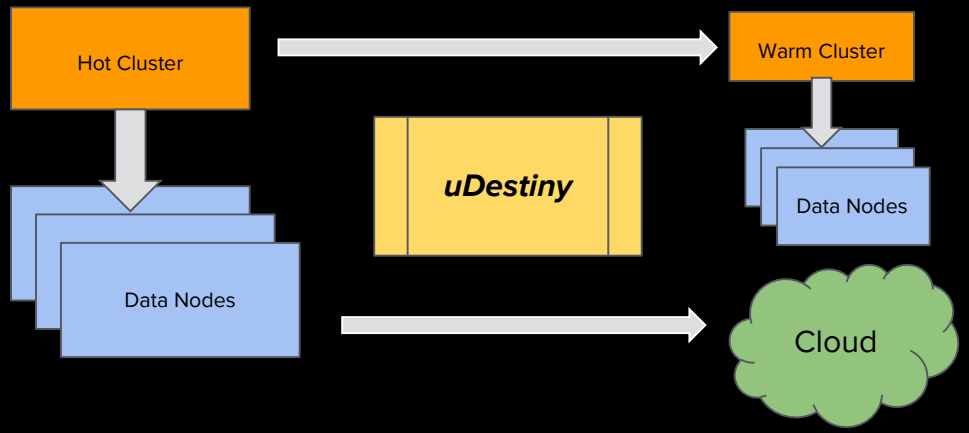
- Intra-server
  - Needs upfront vision for optimal server configuration
- Intra-cluster
  - Single-cluster scaling challenges remain
- Inter-cluster
  - Multi-cluster (alleviates single-cluster scaling challenges)

Building Blocks

- Custom high density storage SKU + Warm HDFS Cluster
  - 3x denser
- Temperature Monitoring Pipelines
- Router Based Federation
  - Add mount-points at run-time with no client side changes
- New service to move data between clusters
  - Implements multi-step workflows to move data between clusters

New Service: uDestiny

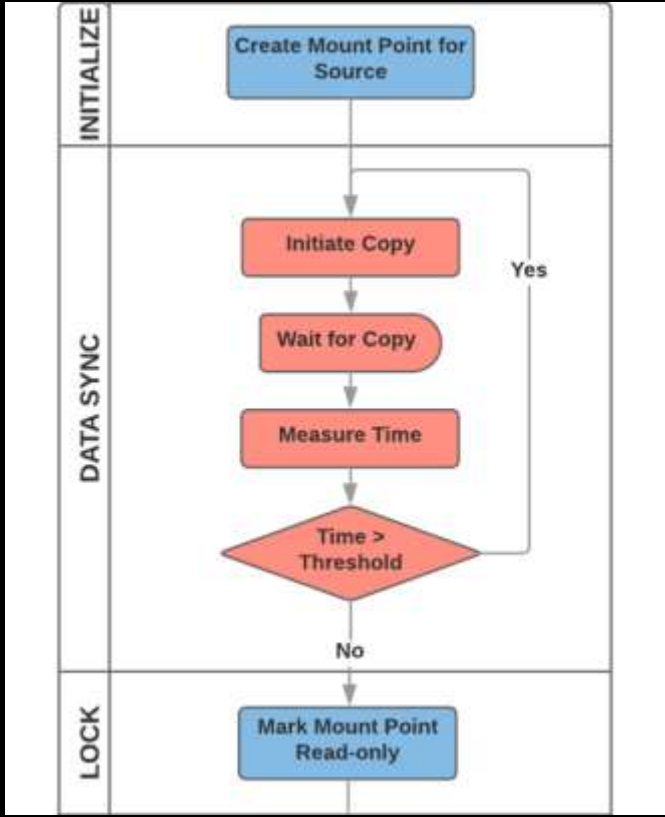
- Data Lifecycle Tasks
  - Supports move and copy workflows for tiering
- APIs for clusters and tasks
- Extensible for other tiers in future (cold storage, cloud)
- Generalized for building other lifecycle tasks (Eg: retention)

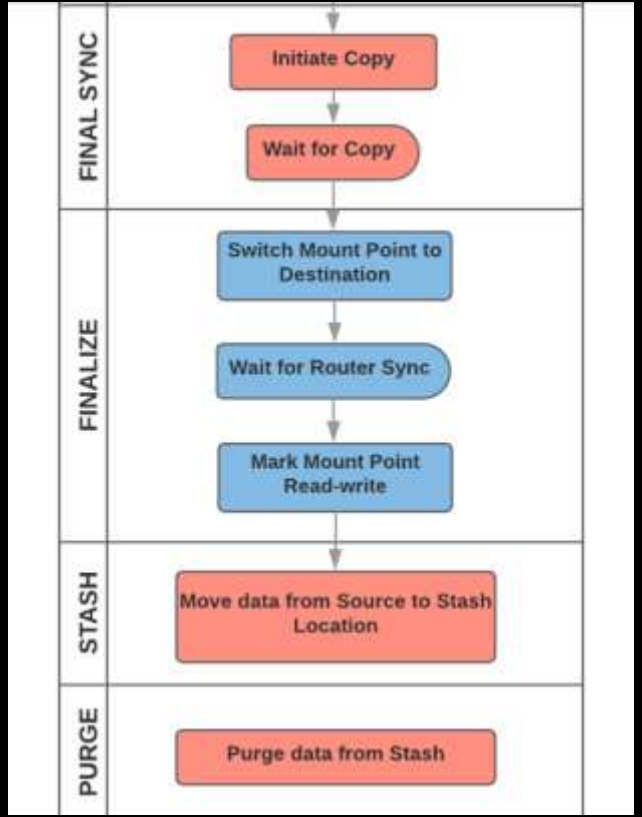


Move Workflow

- Multi-step process
- Safely moves data between clusters
  - Uses RBF for mount-point creation and flipping
- Safe deletion from source

Move Workflow





Operationalizing

- Configuration rollout for RBF
- On-board Hive tables on RBF (update Hive partition location)
- Move partitions to warm cluster using the move workflow
- 10% data served out of warm cluster

What’s next?

- Managed Hive table support
- Copy performance tuning
- Support for secured HDFS clusters
- Erasure Coding

Q&A

Thank You!