Hortonworks

O'REILLY Strata CONFERENCE + HADOOP WORLD

HDFS Snapshots and Beyond

Tsz-Wo (Nicholas) Sze
Jing Zhao

October 29, 2013

© Hortonworks Inc. 2013

About Us

- **Tsz-Wo Nicholas Sze, Ph.D.**
  - Software Engineer at Hortonworks
  - PMC Member at Apache Hadoop
  - One of the most active contributors/committers of HDFS
    - Started in 2007
  - Used Hadoop to compute Pi at the two-quadrillionth ($2x10^{15}$th) bit
    - It is the current World Record.

- **Jing Zhao, Ph.D.**
  - Software Engineer at Hortonworks
  - Committer at Apache Hadoop
  - Active Hadoop contributor too
    - Contributed ~150 patches in about a year

© Hortonworks Inc. 2013

Before Snapshots…

- **Deleted files cannot be restored**
  - Trash is buggy and not well understood
  - Trash works only for CLI based deletion
- **No point-in-time recovery**
- **No periodic snapshots to restore from**
  - No admin/user managed snapshots

© Hortonworks Inc. 2013

HDFS Snapshot

Point-in-time image of the file system

Read-only

Copy-on-write

© Hortonworks Inc. 2013

Use Cases

Protection against user errors

Backup

Experimental/Test setups

© Hortonworks Inc. 2013

Example: Periodic Snapshots for Backup

- **A typical snapshot policy:**

  Take a snapshot in
  - every 15 mins and keep it for 24 hrs
  - every 1 hr, keep 2 days
  - every 1 day, keep 14 days
  - every 1 week, keep 3 months
  - every 1 month, keep 1 year

© Hortonworks Inc. 2013

Design Goal: Efficiency

- **Storage efficiency**
  - No block data copying
  - No metadata copying for unmodified files
- **Processing efficiency**
  - No additional costs for processing current data
- **Cheap snapshot creation**
  - Must be fast and lightweight
  - Must support for a very large number of snapshots

© Hortonworks Inc. 2013

Design Goal: Features

- **Read-only**
  - Files and directories in a snapshot are immutable
  - Nothing can be added to or removed from directories
- **Hierarchical snapshots**
  - Snapshots of the entire namespace
  - Snapshots of subtrees
- **User operation**
  - Users can take snapshots for their data
  - Admins manage where users can take snapshots

© Hortonworks Inc. 2013

HDFS-2802: Snapshot Development

- **Available in Hadoop 2 GA release (v2.2.0)**
- **Community-driven**
  - Special thanks to who have provided for the valuable discussion and feedback on the feature requirements and the open questions
- **136 subtask JIRAs**
  - Mainly contributed by Hortonworks
- **The merge patch has about 28k lines**
- **~8 months of development**

© Hortonworks Inc. 2013

Hadoop HDFS / HDFS-2802

Support for RW/RO snapshots in HDFS

Edit | Comment | Assign | More | Submit Patch | Resolve Issue

Sub-Tasks

1.	Snapshot of Being Written Files	Resolved	Jing Zhao	
2.	Support snapshot of single files	Resolved	Tsz Wo (Nicholas), SZE	
3.	Support snapshottable INodeDirectory	Resolved	Tsz Wo (Nicholas), SZE	
4.	Handle replication in snapshots	Resolved	Tsz Wo (Nicholas), SZE	
5.	Add SnapshotManager	Resolved	Tsz Wo (Nicholas), SZE	
6.	Add editlog opcodes for snapshot create and delete operations	Resolved	Suresh Srinivas	
7.	Protocol changes for snapshots	Resolved	Suresh Srinivas	
8.	provide CLI support for allow and disallow snapshot on a directory	Resolved	Brandon Li	

© Hortonworks Inc. 2011

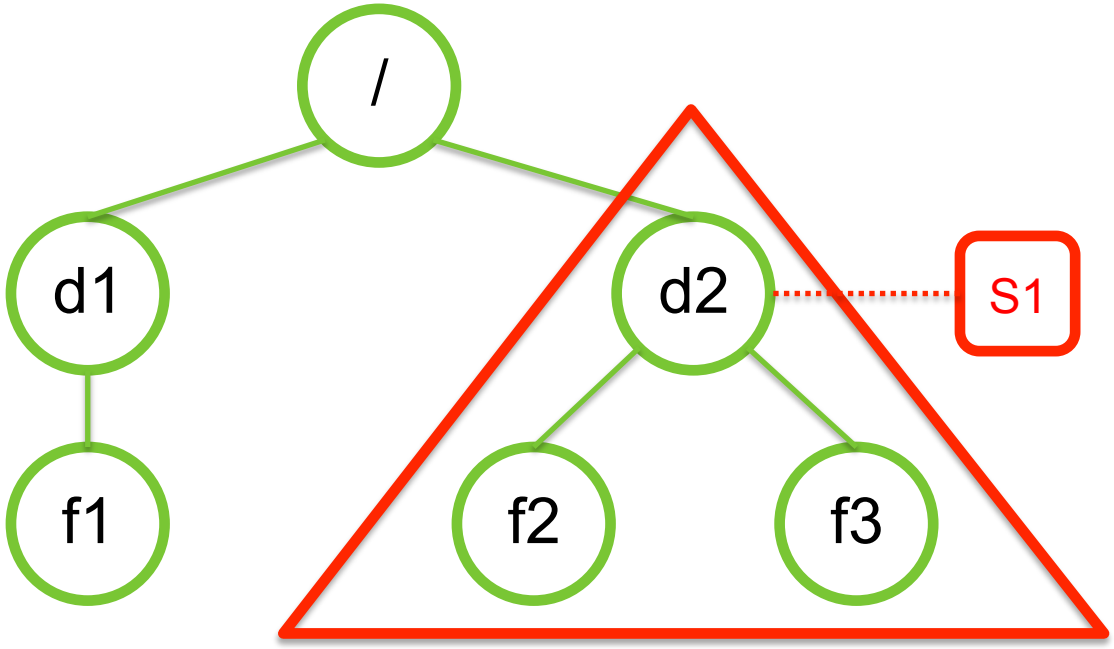
NameNode Only Operation

- **No complicated distributed mechanism**
- **Snapshot metadata stored in NameNode**
- **DataNodes have no knowledge of snapshots**
- **Block management layer also don't know about snapshots**

© Hortonworks Inc. 2013

Fast Snapshot Creation

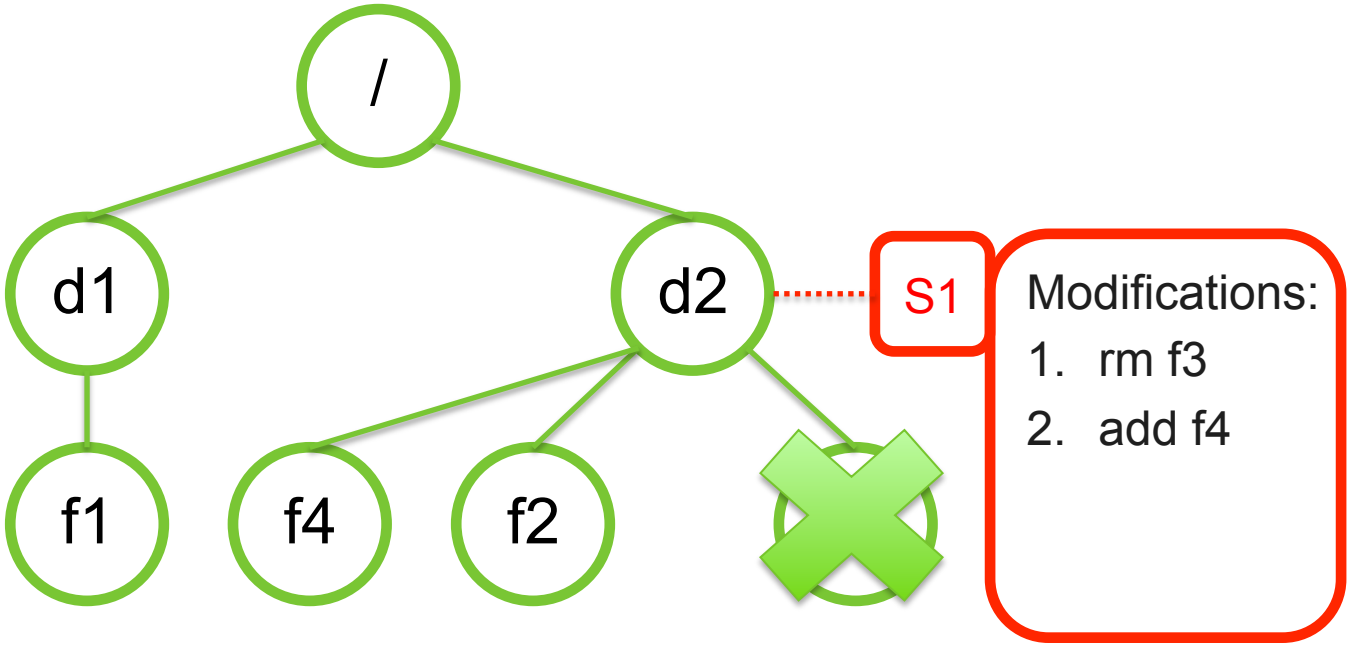
- **Snapshot Creation: *O(1)***
  - It just adds a record to an inode



© Hortonworks Inc. 2013

Low Memory Overhead

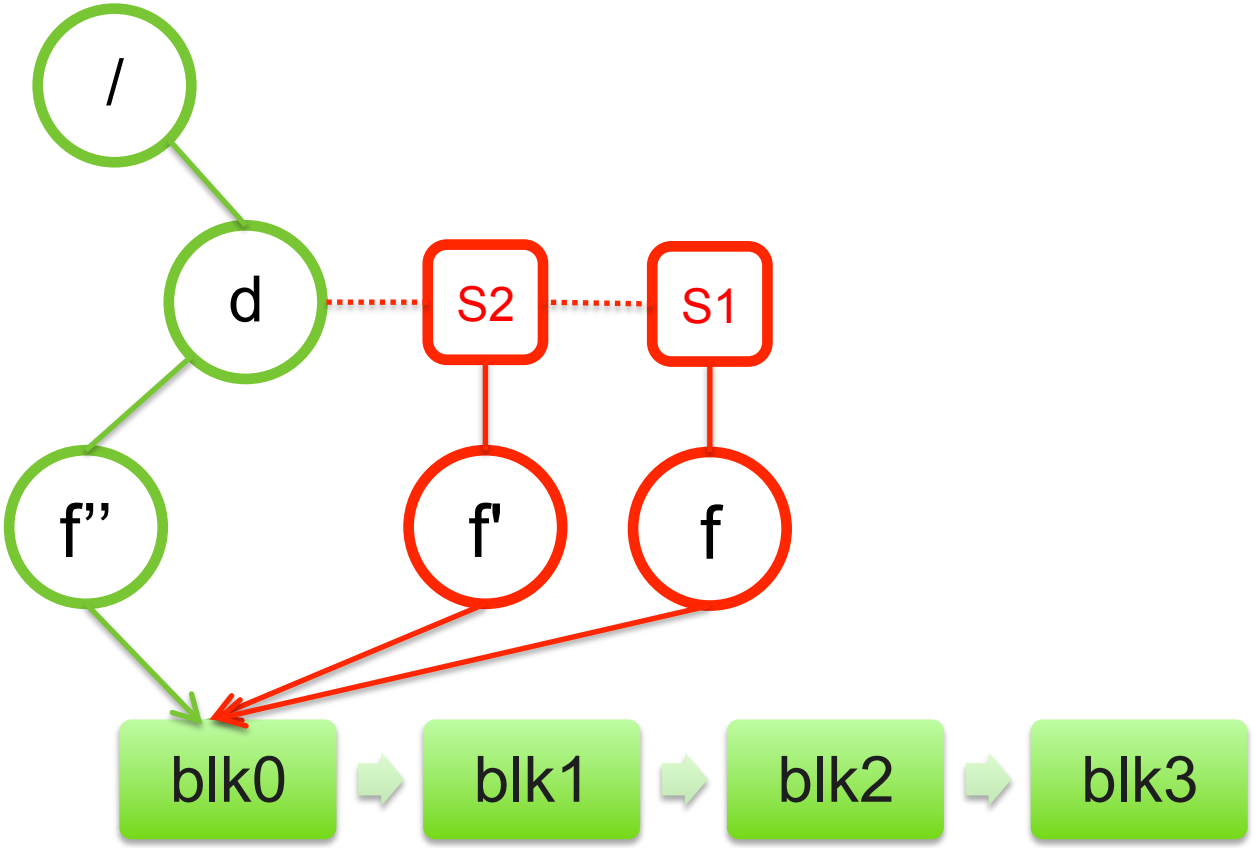
- **NameNode memory usage: *O(M)***
  - *M* is the number of modified files/directories
  - Additional memory is used only when modifications are made relative to a snapshot



© Hortonworks Inc. 2013

File Blocks Sharing

- **Blocks in datanodes are not copied**
  - The snapshot files record the block list and the file size
  - No data copying



© Hortonworks Inc. 2013

Persistent Data Structures

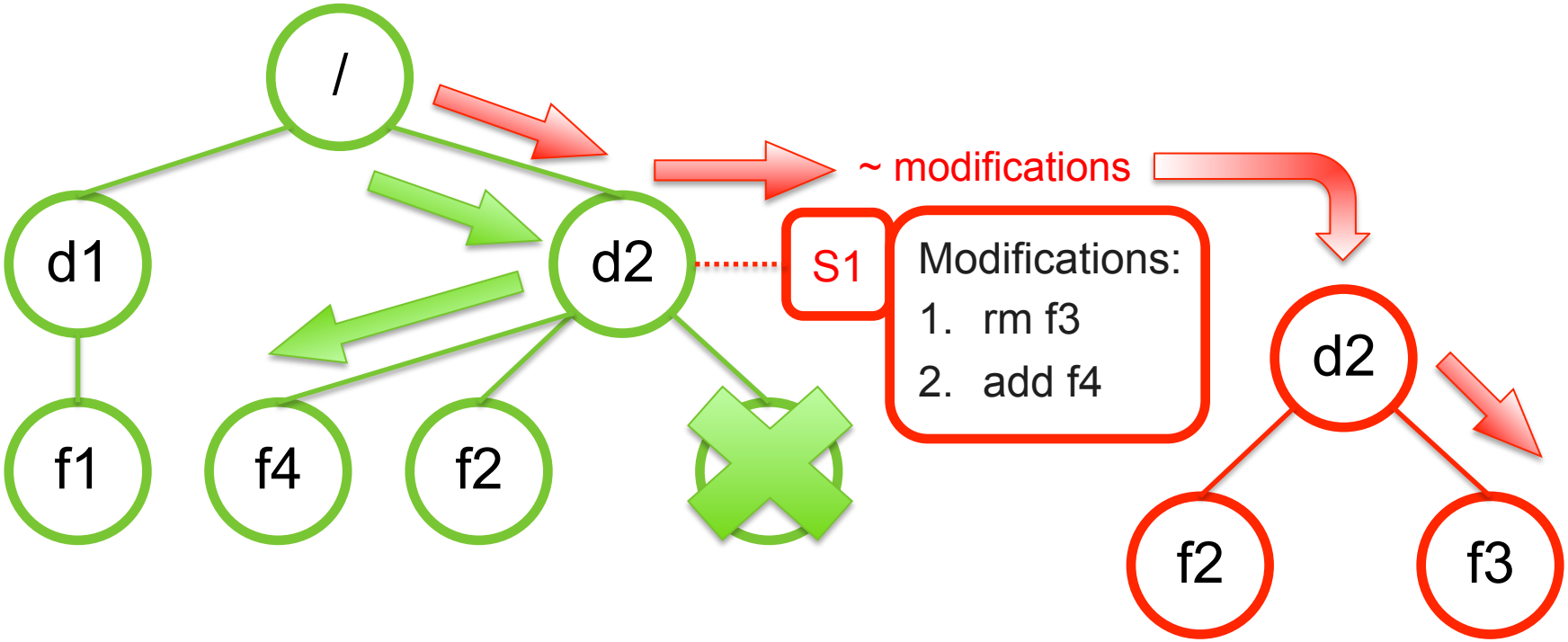
- **A well-known data structure for "time travel"**
  - Support querying previous version of the data
- **Access slow down**
  - The additional time required for the data structure
- **In traditional persistent data structures**
  - There is slow down on accessing current data and snapshot data
- **In our implementation**
  - No slow down on accessing current data
  - Slow down happens only on accessing snapshot data

© Hortonworks Inc. 2013

No Slow Down on Accessing Current Data

- **The current data can be accessed directly**
  - Modifications are recorded in reverse chronological order

Snapshot data = Current data – Modifications



© Hortonworks Inc. 2013

Easy Management

- **Snapshots can be taken on any directory**
  - Set the directory to be ***snapshottable***
- **Support 65,536 simultaneous snapshots**
- **No limit on the number of snapshottable directories**
  - Nested snapshottable directories are currently NOT allowed

© Hortonworks Inc. 2013

Admin Ops

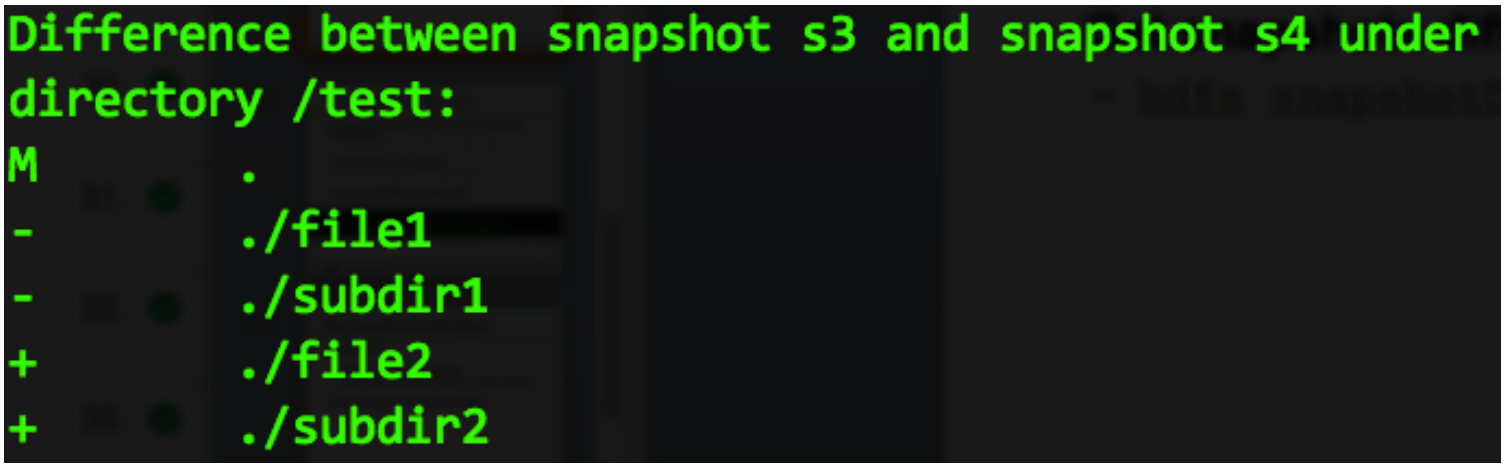
- **Allow snapshots on a directory**
  - `hdfs dfsadmin –allowSnapshot <path>`
- **Reset a snapshottable directory**
  - `hdfs dfsadmin –disallowSnapshot <path>`
- **Example**

```
$ hdfs dfsadmin -allowSnapshot /test
Allowing snapshot on /test succeeded
```


© Hortonworks Inc. 2013

User Ops

- **Create/delete/rename snapshots**
  - `hdfs dfs -createSnapshot <path> [<snapshotName>]`
  - `hdfs dfs -deleteSnapshot <path> <snapshotName>`
  - `hdfs dfs -renameSnapshot <path> <oldName> <newName>`
- **Get snapshottable directory listing**
  - `hdfs lsSnapshottableDir`
- **Get snapshots difference report**
  - `hdfs snapshotDiff <path> <from> <to>`



© Hortonworks Inc. 2013

Use snapshot paths in CLI

- **All regular commands and APIs can be used against snapshot path**
  - `/<snapshottableDir>/.snapshot/<snapshotName>/foo/bar`
- **List all the files in a snapshot**
  - `ls /test/.snapshot/s4`
- **List all the snapshots under that path**
  - `ls <path>/.snapshot`

```
Jing-Zhaos-MacBook-Pro:trunk jing$ hdfs dfs -ls /test/.snapshot
Found 4 items
drwxr-xr-x   - jing supergroup          0 2013-10-29 00:12 /test/.snapshot/s1
drwxr-xr-x   - jing supergroup          0 2013-10-29 00:12 /test/.snapshot/s2
drwxr-xr-x   - jing supergroup          0 2013-10-29 00:14 /test/.snapshot/s3
drwxr-xr-x   - jing supergroup          0 2013-10-29 00:16 /test/.snapshot/s4
```


© Hortonworks Inc. 2013

Test Snapshot Functionalities

- **~100 unit tests**
- **~1.4 million generated system tests**
  - Covering most combination of (snapshot + rename) operations
- **Automated long-running tests for months**

© Hortonworks Inc. 2013

Future Work

- **Restoring/Rolling back to a snapshot**
- **Read-write snapshots**
- **Excluding temp files**
- **Use snapshots beyond HDFS**
  - DistCp
  - Hive exports
  - HBase snapshots

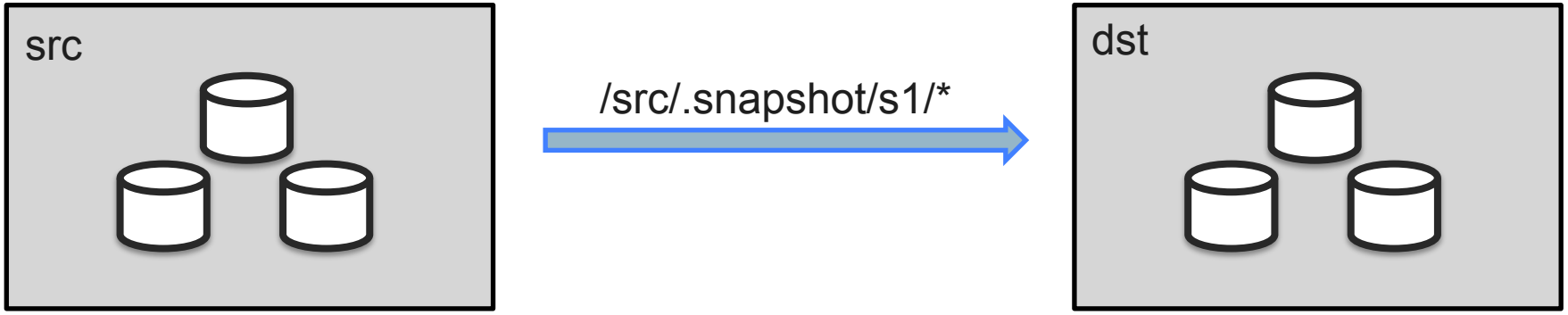
© Hortonworks Inc. 2013

DistCp

- **Copy data to remote cluster**
  - List of files to copy → MapReduce
  - No updates on source files while DistCp
- **Take snapshot before DistCp**
  - Use snapshot paths for source files
- **Use snapshots for incremental backup**
  - Only copy the snapshot diff

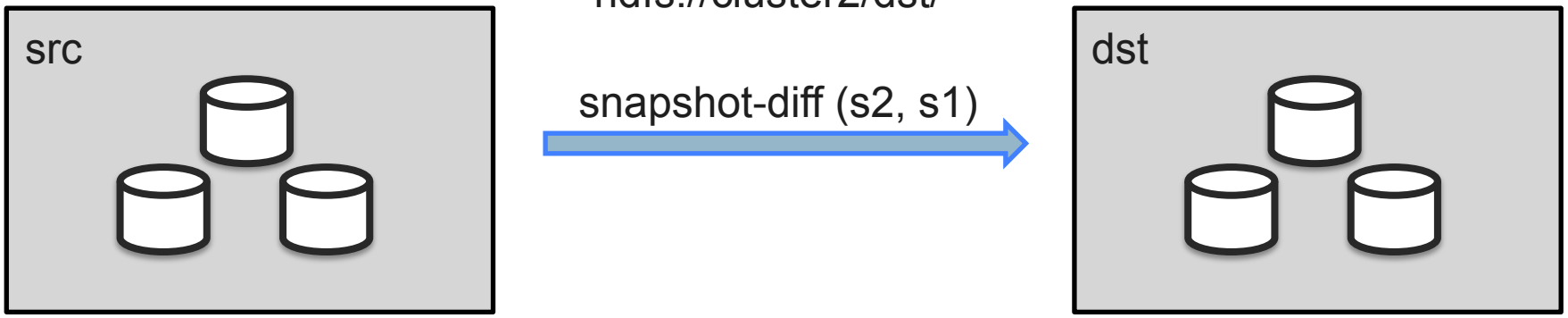
© Hortonworks Inc. 2013

distcp hdfs://cluster1/src/.snapshot/s1/
hdfs://cluster2/dst/



15 min later

distcp hdfs://cluster1/src/.snapshot/s2/
hdfs://cluster2/dst/



© Hortonworks Inc. 2011

Hive Export

- **Current mechanism**
  - Copy data to a new directory in HDFS
  - Export metadata: dump from mySQL and store in HDFS

- **Avoid data copy using snapshots**
  - Snapshot(s) on source directories
  - Create symlinks for snapshot files/directories

© Hortonworks Inc. 2013

HBase Snapshots

- **Current mechanism**
  - Reference all the HFiles
  - Copy the current table info, region info, pending recovered edits
  - NameNode load concern

- **Take HDFS snapshots**
  - Avoid creating references for all HFiles

© Hortonworks Inc. 2013

Q & A

- **Myths and misinformation of HDFS**
  - ~~Not reliable (was never true)~~
  - ~~Namenode dies, all state is lost (was never true)~~
  - ~~Does not support disaster recovery (distcp in Hadoop0.15)~~
  - ~~Hard to operate for new comers~~
  - Performance improvements (always ongoing)
    - Major improvements in 1.2 and 2.x
  - ~~Namenode is a single point of failure~~
  - ~~Needs shared NFS storage for HA~~
  - ~~Does not have point in time recovery~~

Thank You!

© Hortonworks Inc. 2013