Use of HDFS Snapshots for External Table Replication

Problem Statement

Use of DistCp diff with snapshots for external table replication. The distCp diff with snapshots allows it to be restricted only to the modified entries and removes the burden of building the entire listing of the directory including the files/directories which aren't modified. Thus saving a considerable amount of effort during copy.

Use Case

The distCp diff with snapshots is fundamentally based on the fact that all of the files/directories won't be modified in between two replication events.

In case the user has a case where the modification is very high between two copy events for the same path, then most likely the snapshot diff won't be of much use.

Further Reading on DistCp with Snapshots:
https://blog.cloudera.com/distcp-performance-improvements-in-apache-hadoop/

Limitations

FileSystem Support:

The Snapshot diff copy requires FileSystems which supports snapshots, So, In case either the source or target FileSystem doesn't support Snapshots, The snapshot based replication can not be enabled.

Supported Configurations: OnPrem(HDFS) to OnPrem(HDFS)

SuperUser Privilege

Hive needs to be a super user to allow snapshots. Either the paths should already be snapshottable or hive should have privilege to allow snapshots for those paths. Unable to create a snapshot will fail the replication if the enable snapshot diff config is enabled.

Configurations

If the snapshot diff copy config is enabled, it will work implicitly for the external warehouse location. For the tables outside the warehouse location use the other configuration to pass the exact paths where snapshot diff copy needs to be used. There won't be any failure in case the path doesn't belong to any table.

Name of Config	Value to be Set	Description
hive.repl.externaltable.snapshotdiff.copy	true	Enables the snapshot diff copy
hive.repl.externatable.snapshot.paths	Comma separated paths	Comma separated list of exact paths of table, for which snapshots need to be used apart from the ones in the warehouse location. Paths can be added/removed in the later stages as well.

Design Overview

DistCp Copy Modes

InitialCopy

The mode used when the directory is being copied for the first time using snapshots. This is a normal distCp from the source snapshot directory to target. During the first time only one snapshot is available, so diff can not be computed.

DiffCopy

This is the intended mode of copy, This copy mode is used when we have two snapshots available at source, to compute the diff, and in this mode distcp is done using the snapshot diff option.

FallbackCopy

This mode symbolises regular/normal distcp, This copy mode is used when snapshot is not enabled or snapshot diff copy fails during copy.

Failure Cases

NonSnapshottable Directories

In case the source or target directory aren't snapshottable and we aren't able to allowSnapshot, due to any reason, say inadequate permissions, or any other directory already snapshottable directory in the hierarchy, the replication shall fail and not fallback, and require manual interventions.

Failure to Create Snapshot

If creation of snapshots fail after a considerable number of retries, at source or target, the replication fails, to alarm out the admin to look after the reasons.

> During the dump itself the copy mode is decided and propagated while executing the copy.

Snapshots Creations:

At Source

Two Snapshots are maintained (new & old) each prefixed with the database name which are used to compute the diff. One iteration creates on one snapshot, When there is only one snapshot, InitialCopy mode will be used, from the second iteration we would be having a pair of snapshots for diff, and in further iterations, we will keep on deleting the older snapshots, so as to maintain only a pair of snapshots.

At Target

One Snapshot is maintained which is the same as the older snapshot at the source.(old). The snapshot is created immediately after the copy is done. This snapshot is used for the following

- validations such as no modification at target
- optimizations in case of renames, etc.

Fallback

- If the target is modified, rDiff is called to restore back the state and the copy is re-attempted.
- No Fallback if the snapshots aren't enabled or creation of snapshot fails after a couple of retries as well. We fail with a non-recoverable error.

Metrics

The replication metrics will store the paths for which the snapshot diff copy copy failed and we used the fallback mode of copy along with the reason.

CleanUp

The Snapshots created as part of the replication process are cleared in the following states:

- If the Database location is altered, In that case the snapshots available in the old location are cleared.
- If the table location explicitly configured using the config, is dropped or altered to a new location then the snapshots are cleared.
- If the paths in the configurations are changed, then the path removed will have their snapshots cleared.

Future Work

Add a shutdown hook which can be used when a scheduled query is dropped and which can be used for cleaning up of snapshots when the scheduled query is dropped.

Control Failover:

In case of control failover the source on which the diff was computed will become the target, and the target cluster will become the source of copy.
Source: cluster1 & target Cluster2
Example Directory hdfs://cluster1/table1 to hdfs://cluster2/table1
During this phase:
Cluster1 will be having two snapshots the second & the first snapshot, The first being the latest and the second being the older one.
Cluster2 will contain only one snapshot, the second snapshot, which will be the same as the second snapshot of cluster1 prior to copy and same as the first snapshot post copy is successful.
**The control failover will happen when both the clusters are in sync, so it will lead to a state,

When cluster1 dir and cluster2 dir are same. In that case. First snapshot at cluster1 will be the same as the second snapshot at cluster2.
Now Post Failover:
Source: Cluster2 & target Cluster1
During Dump the Cluster2 will find a second snapshot(which is the same as the first snapshot of cluster1) it will create another snapshot first snapshot and proceed in a regular way as it does.
During Copy, the second snapshot should be the same, but in this very case the first snapshot at the target cluster1 is the same as source cluster2, so we rename the first snapshot as the second snapshot and delete the second snapshot. The copy shall continue.

Problem:
The base of the snapshot replication is the copy of the external warehouse location and using the warehouse path for copy. But in case of replication the External warehouse path isn't preserved, so when moving from cluster2 to cluster1 the database location will be different, so the same DirCopyTask will not be created, hence the snapshots won't be used in this path.

Solution:
1. Preserve the Database External Warehouse Location
2. Add an ability to provide a path for which a single copy task can be created similar it is done for Database External Warehouse Location, and provide that path as part of the with clause.

Code Cases

Scenario	Auto/Manual	Handling
Target Modified	Auto	Calls distcp with rdiff to restor to previous state, and retries. If overwrite is disabled, the replication fails,
Tables added/removed as part of the config	Auto	Cleans up the snapshot by comparing the list of created snapshots during the last dump
Altter of database location	Auto	Cleans up the snapshots by comparing the last location.
Dump fails before starts processing data copy tasks	Auto	Least bothered, we resume as is.
Dump fails in between the dump	Auto	Keeps the track of snapshots created during last run and just updates the latest

		snapshot to increase the diff duration for the snapshots created earlier, for the new ones it continues as is
Failure to create snapshot at source	Manual	We fail with non-recoverable exception. The Admin needs to make sure the path is snapshottable or remove the path from the scope
Failure to create snapshot at target	Manual	Load fails, Make sure it is able to create and retry, The dump will be able to handle mid-level failures.
Enable Snapshots After Couple of Dump & Load Cycle	Auto	The snapshot creation can start after bootstrap/incremental cycles. It will start creating snapshots once enabled.
Re-Bootstrap	Auto	If everything is cleared at target the Database & tables are purged, And then if Bootstrap is triggered, it will regenerate the snapshots at source and start a fresh dropping the old snapshots from INITIAL_COPY stage
Purge Tables(HDFS doesn't allow to delete directories with snapshots.)	Auto	The table location if contains snapshots, in case delete is triggered, it cleans up the replication level snapshots and doesn't fail to delete the directory