fs - Feature #24880

pybind/mgr/volumes: restore from snapshot

07/12/2018 03:17 AM - shi liu

Status:	Resolved	**% Done:**	0%
Priority:	Urgent		
Assignee:	Venky Shankar		
Category:			
Target version:	v15.0.0		
Source:	Development	**Affected Versions:**	
Tags:		**Component(FS):**	mgr/volumes
Backport:	nautilus	**Labels (FS):**	
Reviewed:		**Pull request ID:**	32030

Description

Old description:

Manila's cephfs driver does not support recovering data from snapshots,The driver uses the ceph_volume_client library.

If implement cephfs_volume_client's `_cp_r` method[1], it useful for data recovery of manila cephfs drivers.

[1]: https://github.com/ceph/ceph/blob/master/src/pybind/ceph_volume_client.py#L1390

Related issues:

Related to fs - Feature #43349: mgr/volumes: provision subvolumes with config...	**Resolved**
Copied to fs - Backport #44020: pybind/mgr/volumes: restore from snapshot	**Resolved**

History

#1 - 07/12/2018 09:54 AM - John Spray

I just checked who wrote that "TODO" comment, and it turns out it was me, even though I have no memory of it :-)

IIRC, the hope was that there would be a "read only clone" mechanism (i.e. a clone but the new share would have a readonly flag set), that would in reality just map the new volume to the proper .snap subdirectory, and we'd only do the full copy on a writable clone. I'm not sure whether the Manila clone API ended up in a form that provides that distinction, so that would be something to check.

#2 - 07/12/2018 09:54 AM - John Spray

- Project changed from Ceph to fs

#3 - 07/16/2018 01:46 PM - Patrick Donnelly

- Assignee set to Ramana Raja

- Component(FS) VolumeClient added

#4 - 03/07/2019 11:21 PM - Patrick Donnelly

- Target version changed from v14.0.0 to v15.0.0

#5 - 05/22/2019 11:41 AM - Patrick Donnelly

- Subject changed from ceph_volume_client: Implementation of the cp method to pybind/mgr/volumes: restore from snapshot

- Description updated

- Assignee changed from Ramana Raja to Rishabh Dave

- Start date deleted (07/12/2018)

- Backport changed from mimic,luminous to nautilus

- Component(FS) mgr/volumes added

- Component(FS) deleted (VolumeClient)

#6 - 06/12/2019 01:47 PM - Patrick Donnelly

- Assignee changed from Rishabh Dave to Ramana Raja

- Priority changed from Normal to Urgent

ceph-csi ticket: https://github.com/ceph/ceph-csi/issues/411

Ramana, I'm reassigning this to you. We need this done quickly.

#7 - 10/28/2019 03:45 AM - Patrick Donnelly

- Assignee changed from Ramana Raja to Venky Shankar

#8 - 10/29/2019 12:57 PM - Ramana Raja

This feature will be used by Ceph CSI to create a PVC from a snapshot [1], and by OpenStack Manila to create a share from a snapshot [2]. Since snapshot from clone might take a while, we'd want a mgr/volumes call to initiate the asynchronous subvolume creation from a subvolume snapshot, and an another call to check the status of the subvolume. The `ceph fs subvolume create` CLI can be extended to trigger the subvolume creation from snapshot and return immediately,

```
ceph fs subvolume create <volname> <subvolname> [--group_name <group_name>] [--size <size>] [--snapshot_source <src snapshot name>] [--subvol_source <src subvolname>]
```

and have an another CLI say

```
ceph fs subvolume show <volname> <subvolname> [--group_name <group_id>]
```

to return the status of the subvolume creation.

See discussion here for more details,
https://github.com/ceph/ceph-csi/issues/411#issuecomment-503669965

It should also be possible to garbage collect the subvolumes that were partially restored from a snapshot.

[1] https://kubernetes.io/docs/concepts/storage/persistent-volumes/#create-persistent-volume-claim-from-volume-snapshot
[2] https://docs.openstack.org/manila/latest/cli/manila.html#manila-create

#9 - 11/12/2019 01:43 PM - Venky Shankar

clone operation design & interface:

. Interface

Introduce `clone` sub-command in `subvolume snapshot` command

```
$ ceph fs subvolume snapshot clone <source> <target> [<pool-layout>]
```

where,
source: (filesystem, group, subvolume, snapshot) tuple
target: (filesystem, group, subvolume) tuple
pool-layout: optional, default to pool-layout of target group

Clone operation is asynchronous. Also, allow clone to a different subvolume group. Cloning a subvolumegroup is not a requirement.

b. Since clone is asynchronous, what if the snap is removed when clone is in progress? The clone operation should be canceled and marked as failed (user interrupted).

c. Clone operation status
Introduce `clone status` subcommand.

#2 subcommand takes "<source> <target>" and displays its status.

```
$ ceph fs subvolume snapshot clone status <source> <target>
```

A clone operation can be in the following states:

- not started
- in progress
- failed
- complete

`not started and `in progress` are fairly self-explanatory.

Maintain (or figure out) list of failed clones (and this list needs to be available across mgr restarts and failover so as to provide consistent state to CSI). mgr/volumes should skip over failed clones on restart/failover. Provide a CLI command to clear failed clones.

d. Control operations: Support canceling an on-going clone operation (+ CLI command)

Changes to subvolume provisioning by mgr/volumes:

Maintain subvolume metadata in cephfs -- this would allow mgr/volumes to persist subvolume metadata to satisfy consistent `clone status` reporting across manager restarts and failover. Metadata would also carry "version" of a subvolume that would be bumped up as features get added (cloning a subvolume). Backward compatibility would be maintained so no migration of subvolume to the new format is needed.

#10 - 12/17/2019 09:26 AM - Venky Shankar

- Related to Feature #43349: mgr/volumes: provision subvolumes with config metadata storage in cephfs added

#11 - 01/17/2020 07:50 AM - Venky Shankar

- Status changed from New to Fix Under Review

- Pull request ID set to 32030

clone from a snap: https://github.com/ceph/ceph/pull/32030

Most of this work will be required for restoring a subvolume from a snap.

#12 - 02/06/2020 04:46 PM - Ramana Raja

- Status changed from Fix Under Review to Pending Backport

#13 - 02/06/2020 04:47 PM - Ramana Raja

- Copied to Backport #44020: pybind/mgr/volumes: restore from snapshot added

#14 - 02/12/2020 04:30 PM - Ramana Raja

- Status changed from Pending Backport to Resolved