ceph-volume - Bug #24795

Unclear documentation on preparing devices

07/06/2018 01:57 PM - Sébastien Han

Status:	Resolved		**Start date:**	07/06/2018
Priority:	Normal		**Due date:**	
Assignee:			**% Done:**	0%
Category:			**Estimated time:**	0.00 hour
Target version:				
Source:			**Reviewed:**	
Tags:			**Affected Versions:**	v13.2.0
Backport:			**ceph-qa-suite:**	
Regression:	No		**Pull request ID:**	
Severity:	3 - minor			

Description

This is the doc I have for 13.2.0:

```
[root@ceph-osd0 /]# ceph-volume lvm prepare

Prepare an OSD by assigning an ID and FSID, registering them with the
cluster with an ID and FSID, formatting and mounting the volume, and
finally by adding all the metadata to the logical volumes using LVM
tags, so that it can later be discovered.

Once the OSD is ready, an ad-hoc systemd unit will be enabled so that
it can later get activated and the OSD daemon can get started.

Encryption is supported via dmcrypt and the --dmcrypt flag.

Existing logical volume (lv):

    ceph-volume lvm prepare --data {vg/lv}

Existing block device, that will be made a group and logical volume:

    ceph-volume lvm prepare --data /path/to/device

Optionally, can consume db and wal devices or logical volumes:

    ceph-volume lvm prepare --data {vg/lv} --block.wal {device} --block.db {vg/lv}
```

Looking at the section "Existing block device, that will be made a group and logical volume:"
So it seems that in a all in one fashion (block/wal/db on the same device) passing a block device to ceph-volume works.

However looking at "Optionally, can consume db and wal devices or logical volumes:", it says it can consume db and wal "devices"
which makes me think I can use a device just like in the previous command.
However, it seems that we lost the ability to use a block device for --data because the doc says {vg/lv} (so it's expecting an LV now).

1) Is this intentional?

But I didn't stop here and tried with a block device for --data anyway and here is the result:

```
[root@ceph-osd0 /]# ceph-volume lvm prepare --data /dev/sdb --block.wal /dev/sdc
Running command: /bin/ceph-authtool --gen-print-key
Running command: /bin/ceph --cluster ceph --name client.bootstrap-osd --keyring /var/lib/ceph/boot
strap-osd/ceph.keyring -i - osd new 5fafadcf-662a-47cf-9bb4-f1912fa8e33d
```

```
Running command: /usr/sbin/vgcreate --force --yes ceph-fca417c9-7960-4627-82c9-5954d7d8e7b7 /dev/s
db
 stdout: Physical volume "/dev/sdb" successfully created.
 stdout: Volume group "ceph-fca417c9-7960-4627-82c9-5954d7d8e7b7" successfully created
Running command: /usr/sbin/lvcreate --yes -l 100%FREE -n osd-block-5fafadcf-662a-47cf-9bb4-f1912fa
8e33d ceph-fca417c9-7960-4627-82c9-5954d7d8e7b7
 stdout: Logical volume "osd-block-5fafadcf-662a-47cf-9bb4-f1912fa8e33d" created.
--> blkid could not detect a PARTUUID for device: /dev/sdc
--> Was unable to complete a new OSD, will rollback changes
--> OSD will be fully purged from the cluster, because the ID was generated
Running command: /bin/ceph osd purge osd.2 --yes-i-really-mean-it
 stderr: purged osd.2
-->  RuntimeError: unable to use device
```

I would expect based on the doc that the error will be on the --data portion so on /dev/sdb but instead it's on /dev/sdc.
So ceph-volume tells me it cannot find a PARTUUID on /dev/sdc which is normal because there is no partition on /dev/sdc.

So my last question is, should I create a partition? (if so we should clarify the help command) How is this supposed to work?
Thanks.

History

#1 - 07/06/2018 02:30 PM - Alfredo Deza

The --help menu isn't meant to give a granular description of everything the tool can do. If that is the only thing you were following it is easy to get into trouble.

In both the filestore and bluestore sections we emphasize that it **has to be a partition**:
http://docs.ceph.com/docs/master/ceph-volume/lvm/prepare/#bluestore

```
    If a block.db or a block.wal is needed (they are optional for bluestore) they can be specified with --bloc
k.db and --block.wal accordingly. These can be a physical device (they must be a partition) or a logical volum
e.
```

There are so many options and permutations of what is accepted that pushing all that into the --help would make it easy to miss as well.

These requirements for a partition are also in the man page for ceph-volume http://docs.ceph.com/docs/master/man/8/ceph-volume/

```
[-block.wal] Path to a bluestore block.wal logical volume or partition
[-block.db] Path to a bluestore block.db logical volume or partition
```

We emphasize **a lot** the need for a PARTUUID in the docs:

```
When using a partition, it must contain a PARTUUID discoverable by blkid, so that it can later be identified c
orrectly regardless of the device name (or path).
```

#2 - 07/09/2018 11:39 AM - Sébastien Han

I agree we can't emphasize all the permutations. However, the --help is misleading so it should be fixed Anything that is called a 'device' should be called a 'partition' if this is what we expect.
Also, the manpage must be the reference and should be self-sufficient, you shouldn't expect people to look in
http://docs.ceph.com/docs/master/ceph-volume/lvm/prepare/#bluestore. Consultants on site don't always have internet access.

Thanks.

#3 - 07/19/2018 08:32 AM - Michael Jones

Personally I consider it a regression from ceph-disk if I can't simply provide a block device (e.g. /dev/sda) and ceph-volume just figures out what to do.

What's the problem? ceph-disk was able to do it just fine.

#4 - 07/19/2018 08:34 AM - Michael Jones

The same documentation you link also clearly says

```
The bluestore objectstore is the default for new OSDs. It offers a bit more flexibility for devices. Bluestore
 supports the following configurations:

A block device, a block.wal, and a block.db device
A block device and a block.wal device
A block device and a block.db device
A single block device
*It can accept a whole device (or partition)*, or a logical volume for block.
```

So I'm quite confused what my options are with this tool.

#5 - 07/19/2018 11:57 AM - Alfredo Deza

- Status changed from New to Verified

@Michael Jones: this is not a regression. I believe you've opened #24993 which tracks the problem you have which is unrelated to this bug.

@Sebastien, yes I wholeheartedly agree that this has to be clear on the man page and we need to reduce the ambiguity in the help menu. Thanks for point that out

#6 - 10/03/2018 12:12 PM - Alfredo Deza

- Status changed from Verified to In Progress

master PR: https://github.com/ceph/ceph/pull/24394

#7 - 10/29/2018 09:52 AM - Alfredo Deza

- Status changed from In Progress to Resolved

mimic PR: https://github.com/ceph/ceph/pull/24449
luminous PR: https://github.com/ceph/ceph/pull/24451