Arvados - Feature #16573

[keep] deduplication reporting tool

06/29/2020 05:01 PM - Ward Vandewege

Status:	Resolved	**Start date:**	07/10/2020
Priority:	Normal	**Due date:**	
Assigned To:	Ward Vandewege	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	2020-07-15		

Description

Create a tool that, when given a list of 2 or more collection UUIDs, prints out a report that shows the aggregate nominal and the actual size of the collections, and indicates how much space is saved by Keep's deduplication.

Add this tool as a command to *arvados-client*.

Subtasks:

Task # 16579: Review branch 16573-keep-deduplication-reporting-tool	**Resolved**

Related issues:

Related to Arvados Epics - Story #16514: Actionable insight into keep usage	**New**	**06/01/2022**	**09/30/2022**

Associated revisions

Revision 28607876 - 07/13/2020 01:27 PM - Ward Vandewege

Merge branch '16573-keep-deduplication-reporting-tool'

closes #16573

Arvados-DCO-1.1-Signed-off-by: Ward Vandewege <ward@curii.com>

History

#1 - 06/29/2020 05:04 PM - Ward Vandewege

- Related to Story #16514: Actionable insight into keep usage added

#2 - 06/29/2020 05:06 PM - Ward Vandewege

- Subject changed from compare-collections tool to [keep] collection deduplication reporting tool

#3 - 06/29/2020 05:22 PM - Ward Vandewege

- Subject changed from [keep] collection deduplication reporting tool to [keep] deduplication reporting tool

#4 - 06/30/2020 08:22 PM - Ward Vandewege

Ready for review at 2e0648fb2b8a006664e6225826d78916f682eff5 on branch 16573-keep-deduplication-reporting-tool. Tests are at https://ci.arvados.org/view/Developer/job/developer-run-tests/1952/

#5 - 06/30/2020 08:23 PM - Ward Vandewege

- Target version changed from 2020-07-01 Sprint to 2020-07-15

#6 - 07/10/2020 07:50 PM - Tom Clegg

Usage note: "nominal space used by the list of collection" → "list of collections" (or just "collections")

Usage example: instead of assuming server default page size is 100, how about adding --limit 100 and changing "tail -n100" to "tail -n+2"

deDuplicate: map lookups return the zero value when the key is not found, so with a bool map where you only store true values, you can replace if _, ok := seen[uuid]; !ok with just if !seen[uuid] (similar thing in the for _, v := range blocks loop in report())

if len(inputs) < 2 -- could this be < 1? If there's only one UUID after deduplication, I think it would be reasonable to output the simple answer rather than throwing an error (I'm guessing it would just work that way if you remove the error check)

If not... error message "different collections UUIDs" → "different collection UUIDs"

fmt.Println() blank line before the loop in report() seems unnecessary ... but if we keep it, it should be fmt.Fprintln(stdout)

Delete commented-out imports in report_test.go

The OverlappingCollections test cases appear nearly identical - maybe they could be combined into a parameterized test along the lines of TestPermission in source:services/arv-git-httpd/auth_handler_test.go

The manifests in the test cases appear to have hard-coded permission signatures -- signature ending in @5f0de808 will expire on Tue Jul 14 13:14:48 EDT 2020 which might cause tests to fail. Since NewClientFromEnv gives you an admin token for tests, can you leave off the +A signatures entirely?

Feature suggestion / scope creep: instead of relying on the user to provide PDHs for efficiency, the tool could start by doing a collection list with select=pdh, filters=[uuid,in,[...]] (which should always be fast), then deduplicate by pdh, then fetch the manifests.

#7 - 07/13/2020 01:36 AM - Ward Vandewege

Tom Clegg wrote:

> Usage note: "nominal space used by the list of collection" → "list of collections" (or just "collections")

Fixed.

> Usage example: instead of assuming server default page size is 100, how about adding --limit 100 and changing "tail -n100" to "tail -n+2"

Nice, fixed.

> deDuplicate: map lookups return the zero value when the key is not found, so with a bool map where you only store true values, you can replace if _, ok := seen[uuid]; !ok with just if !seen[uuid] (similar thing in the for _, v := range blocks loop in report())

Oh, yes, fixed.

> if len(inputs) < 2 -- could this be < 1? If there's only one UUID after deduplication, I think it would be reasonable to output the simple answer rather than throwing an error (I'm guessing it would just work that way if you remove the error check)

Yes, it does just work; I changed it to < 1 and adjusted the error message accordingly.

> fmt.Println() blank line before the loop in report() seems unnecessary ... but if we keep it, it should be fmt.Fprintln(stdout)

Right! I dropped it.

> Delete commented-out imports in report_test.go

Done.

> The OverlappingCollections test cases appear nearly identical - maybe they could be combined into a parameterized test along the lines of TestPermission in source:services/arv-git-httpd/auth_handler_test.go

Done.

> The manifests in the test cases appear to have hard-coded permission signatures -- signature ending in @5f0de808 will expire on Tue Jul 14 13:14:48 EDT 2020 which might cause tests to fail. Since NewClientFromEnv gives you an admin token for tests, can you leave off the +A signatures entirely?

Nice, done.

> Feature suggestion / scope creep: instead of relying on the user to provide PDHs for efficiency, the tool could start by doing a collection list with select=pdh, filters=[uuid,in,[...]] (which should always be fast), then deduplicate by pdh, then fetch the manifests.

Yeah; if it's OK I'll leave that as a future improvement.

All fixes at 544d7aae0d58a25e8c761c638167c3564de06af5 on branch 16573-keep-deduplication-reporting-tool.

#8 - 07/13/2020 01:21 PM - Tom Clegg

Just a couple of nits

logger.Errorf("...\n") doesn't need "\n" (unlike fmt.Fprintf())

might as well delete this:

```
+       //c.Check(stdout.String(), check.Equals, "")
```

Rest LGTM, thanks!

#9 - 07/13/2020 01:28 PM - Ward Vandewege

- Status changed from In Progress to Resolved

I've made those changes and merged, thanks.

#10 - 08/04/2020 08:48 PM - Ward Vandewege

- Release set to 25