Arvados - Story #7824

[SDKs] arv-get and arv-ls should use new PySDK Collection APIs

11/19/2015 12:51 AM - Sarah Guthrie

Status:	Resolved	**Start date:**	11/19/2015
Priority:	Normal	**Due date:**	
Assigned To:	Lucas Di Pentima	**% Done:**	100%
Category:	SDKs	**Estimated time:**	0.00 hour
Target version:	2017-04-12 sprint		

Description

Fix

- Rewrite arv-ls and arv-get to use modern CollectionReader API (keys() and open()) instead of legacy methods all_streams() and all_files().
- To be consistent with other tools, move main code of arv-get into arvados.commands module and replace bin/arv-get with a stub that calls it.
- Update both tools to consistently use logging instead of print >>sys.stderr
- Must not change command line behavior of existing arv-get

Original report

In 1.5 hrs, 8MiB of a 55MiB file was downloaded using the command: arv keep get 215dd32873bfa002aa0387c6794e4b2c+54081534/tile.csv .

A top on the computer running the "arv keep get" command results in:

```
top - 19:47:07 up 2 days,  9:09,  8 users,  load average: 1.12, 1.26, 1.32
Tasks: 223 total,   3 running, 217 sleeping,   0 stopped,   3 zombie
%Cpu(s): 43.5 us,  8.7 sy,  0.0 ni, 47.8 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem:  15535256 total, 12281116 used,  3254140 free,  1069760 buffers
KiB Swap: 15929340 total,   221892 used, 15707448 free.  5467732 cached Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND

14366 sguthrie  20   0 2498672 2.173g   7204 R 100.0 14.7  98:02.16 arv-get
```

Downloads from workbench on this collection generate a timeout before allowing the user to choose where to download the file.

Story #7729 requires multiple downloads from this qr1hi collection (qr1hi-4zz18-wuld8y0z7qluw00) and ones with similarly large manifests. To unblock #7729 I would need one of:

- A recipe that allows a user to alter the manifest to be well behaved
- Faster downloads from collections with very large manifests

Update by Ward:

I investigated a bit while this was ongoing. There was no discernable extra load on keepproxy, or on the API server, or on Postgres while Sally's download was ongoing. But when I tried to run the command locally, after a while I saw arv-get suck up 100% cpu (one core) and peak ram at 3GiB (resident!) until I killed it.

Subtasks:

Task # 11272: Review 7824-arvls-arvput-collection-api-usage	**Resolved**

Related issues:

Related to Arvados - Story #10387: Faster downloading using arv-get	**New**	**10/27/2016**
Related to Arvados - Task #5449: Update examples to use new Python Collection...	**Resolved**	**02/13/2015**

Associated revisions

Revision 1affdcd3 - 04/06/2017 07:17 PM - Lucas Di Pentima

Merge branch '7824-arvls-arvput-collection-api-usage'

Closes #7824

History

#1 - 11/19/2015 01:00 AM - Ward Vandewege

- Description updated

#2 - 11/19/2015 02:03 AM - Peter Amstutz

arv-get uses the legacy all_streams() and StreamReader code, which is probably less efficient than the more recent Collection and ArvFile code.

#3 - 11/19/2015 03:10 AM - Peter Amstutz

Here's a collection downloader using the newer Collection API:

```
import arvados
import arvados.collection

c = arvados.collection.CollectionReader("215dd32873bfa002aa0387c6794e4b2c+54081534")

for i in c:
    print i
    with c.open(i) as af:
        with open(i, "w") as f:
            while True:
                d = af.read(1024*1024)
                if not d:
                    break
                f.write(d)
```

This holds steady at about 1 GiB of RAM and for me it is bandwidth, not CPU limited.

#4 - 11/19/2015 03:11 AM - Ward Vandewege

Another update; I was also trying to look via arv mount. The ls listing took a long time (minutes, maybe many). When I finally returned to that terminal, it had printed output, but the arv-mount process had apparently also been killed by the oom-killer:

```
wardv@shell.qr1hi:~$ ls keep/by_id/215dd32873bfa002aa0387c6794e4b2c+54081534/
240_library.csv  250_library.csv  260_library.csv  2c9_library.csv
241_library.csv  251_library.csv  2ba_library.csv  2ca_library.csv
242_library.csv  252_library.csv  2bb_library.csv  2cb_library.csv
243_library.csv  253_library.csv  2bc_library.csv  2cc_library.csv
244_library.csv  254_library.csv  2bd_library.csv  2cd_library.csv
245_library.csv  255_library.csv  2be_library.csv  2ce_library.csv
246_library.csv  256_library.csv  2bf_library.csv  2cf_library.csv
247_library.csv  257_library.csv  2c0_library.csv  2d0_library.csv
248_library.csv  258_library.csv  2c1_library.csv  2d1_library.csv
249_library.csv  259_library.csv  2c2_library.csv  path_lengths.txt
24a_library.csv  25a_library.csv  2c3_library.csv  tile.csv
24b_library.csv  25b_library.csv  2c4_library.csv  tilelocusannotation.csv
24c_library.csv  25c_library.csv  2c5_library.csv  tilevariant.csv
24d_library.csv  25d_library.csv  2c6_library.csv
24e_library.csv  25e_library.csv  2c7_library.csv
24f_library.csv  25f_library.csv  2c8_library.csv
wardv@shell.qr1hi:~$ ls keep/by_id/215dd32873bfa002aa0387c6794e4b2c+54081534/
ls: cannot access keep/by_id/215dd32873bfa002aa0387c6794e4b2c+54081534/: Transport endpoint is not connected

[4001831.812294] atop invoked oom-killer: gfp_mask=0x280da, order=0, oom_score_adj=0
[4001831.812298] atop cpuset=/ mems_allowed=0
[4001831.812303] CPU: 1 PID: 25736 Comm: atop Not tainted 3.19.0-28-generic #30~14.04.1-Ubuntu
[4001831.812305] Hardware name: Microsoft Corporation Virtual Machine/Virtual Machine, BIOS 090006  05/23/2012
[4001831.812306]  0000000000000000 ffff880106353968 ffffffff817aed97 0000000000003e80
[4001831.812308]  ffff8801b80e75c0 ffff880106353a08 ffffffff817a9ccf ffff880106353a28
[4001831.812310]  0000000000000000 0000000000000000 0000000000000002 0000000000000001
[4001831.812312] Call Trace:
[4001831.812320]  [<ffffffff817aed97>] dump_stack+0x45/0x57
[4001831.812322]  [<ffffffff817a9ccf>] dump_header+0x7f/0x1f1
[4001831.812327]  [<ffffffff8117991b>] oom_kill_process+0x22b/0x390
[4001831.812332]  [<ffffffff8107dd2e>] ? has_capability_noaudit+0x1e/0x30
[4001831.812334]  [<ffffffff8117a112>] out_of_memory+0x4d2/0x520
[4001831.812338]  [<ffffffff8117f540>] __alloc_pages_nodemask+0x940/0xa60
[4001831.812341]  [<ffffffff811c66f7>] alloc_pages_vma+0x97/0x150
[4001831.812344]  [<ffffffff811a6014>] handle_mm_fault+0xd94/0x10e0
[4001831.812346]  [<ffffffff811a07a0>] __get_user_pages+0x100/0x600
[4001831.812349]  [<ffffffff811aaf3f>] ? vma_set_page_prot+0x3f/0x60
```

```
[4001831.812351]  [<ffffffff811a83e8>] __mlock_vma_pages_range+0x68/0x70
[4001831.812353]  [<ffffffff811a8aac>] __mm_populate+0x9c/0x130
[4001831.812355]  [<ffffffff81192153>] vm_mmap_pgoff+0xb3/0xc0
[4001831.812358]  [<ffffffff811aad06>] SyS_mmap_pgoff+0x116/0x270
[4001831.812361]  [<ffffffff8101a992>] SyS_mmap+0x22/0x30
[4001831.812364]  [<ffffffff817b674d>] system_call_fastpath+0x16/0x1b
[4001831.812366] Mem-Info:
[4001831.812367] Node 0 DMA per-cpu:
[4001831.812369] CPU    0: hi:    0, btch:   1 usd:   0
[4001831.812370] CPU    1: hi:    0, btch:   1 usd:   0
[4001831.812371] Node 0 DMA32 per-cpu:
[4001831.812372] CPU    0: hi:  186, btch:  31 usd:   0
[4001831.812373] CPU    1: hi:  186, btch:  31 usd:   0
[4001831.812374] Node 0 Normal per-cpu:
[4001831.812375] CPU    0: hi:  186, btch:  31 usd:  25
[4001831.812376] CPU    1: hi:  186, btch:  31 usd:   0
[4001831.812379] active_anon:801444 inactive_anon:107 isolated_anon:0
[4001831.812379]  active_file:23 inactive_file:51 isolated_file:0
[4001831.812379]  unevictable:8409 dirty:0 writeback:0 unstable:0
[4001831.812379]  free:23414 slab_reclaimable:13788 slab_unreclaimable:9469
[4001831.812379]  mapped:883 shmem:3757 pagetables:4485 bounce:0
[4001831.812379]  free_cma:0
[4001831.812382] Node 0 DMA free:13976kB min:304kB low:380kB high:456kB active_anon:1392kB inactive_anon:8kB a
ctive_file:0kB inactive_file:80kB unevictable:40kB isolated(anon):0kB isolated(file):0kB present:15992kB manag
ed:15908kB mlocked:40kB dirty:0kB writeback:0kB mapped:0kB shmem:56kB slab_reclaimable:16kB slab_unreclaimable
:144kB kernel_stack:0kB pagetables:12kB unstable:0kB bounce:0kB free_cma:0kB writeback_tmp:0kB pages_scanned:0
 all_unreclaimable? no
[4001831.812386] lowmem_reserve[]: 0 416 3423 3423
[4001831.812388] Node 0 DMA32 free:20192kB min:8176kB low:10220kB high:12264kB active_anon:383156kB inactive_a
non:40kB active_file:0kB inactive_file:4kB unevictable:6748kB isolated(anon):0kB isolated(file):0kB present:50
7840kB managed:428080kB mlocked:6748kB dirty:0kB writeback:0kB mapped:224kB shmem:1740kB slab_reclaimable:5476
kB slab_unreclaimable:3548kB kernel_stack:336kB pagetables:1948kB unstable:0kB bounce:0kB free_cma:0kB writeba
ck_tmp:0kB pages_scanned:0 all_unreclaimable? no
[4001831.812392] lowmem_reserve[]: 0 0 3007 3007
[4001831.812394] Node 0 Normal free:59488kB min:59100kB low:73872kB high:88648kB active_anon:2821228kB inactiv
e_anon:380kB active_file:92kB inactive_file:120kB unevictable:26848kB isolated(anon):0kB isolated(file):0kB pr
esent:3145728kB managed:3079180kB mlocked:26848kB dirty:0kB writeback:0kB mapped:3308kB shmem:13232kB slab_rec
laimable:49660kB slab_unreclaimable:34184kB kernel_stack:5136kB pagetables:15980kB unstable:0kB bounce:0kB fre
e_cma:0kB writeback_tmp:0kB pages_scanned:8 all_unreclaimable? no
[4001831.812398] lowmem_reserve[]: 0 0 0 0
[4001831.812400] Node 0 DMA: 4*4kB (U) 9*8kB (UEM) 2*16kB (UM) 3*32kB (UEM) 3*64kB (UEM) 2*128kB (EM) 2*256kB
 (EM) 1*512kB (E) 2*1024kB (UE) 3*2048kB (EMR) 1*4096kB (M) = 13976kB
[4001831.812409] Node 0 DMA32: 302*4kB (UEM) 281*8kB (UE) 160*16kB (UEM) 128*32kB (UE) 62*64kB (UEM) 20*128kB
 (UE) 6*256kB (UE) 2*512kB (EM) 1*1024kB (M) 0*2048kB 0*4096kB = 20224kB
[4001831.812418] Node 0 Normal: 963*4kB (UEM) 1482*8kB (UEM) 765*16kB (UE) 693*32kB (UEM) 146*64kB (UEM) 0*128
kB 0*256kB 0*512kB 0*1024kB 0*2048kB 0*4096kB = 59468kB
[4001831.812425] Node 0 hugepages_total=0 hugepages_free=0 hugepages_surp=0 hugepages_size=2048kB
[4001831.812426] 4716 total pagecache pages
[4001831.812427] 0 pages in swap cache
[4001831.812428] Swap cache stats: add 0, delete 0, find 0/0
[4001831.812429] Free swap  = 0kB
[4001831.812430] Total swap = 0kB
[4001831.812431] 917390 pages RAM
[4001831.812432] 0 pages HighMem/MovableOnly
[4001831.812432] 36598 pages reserved
[4001831.812433] 0 pages cma reserved
[4001831.812434] 0 pages hwpoisoned
[4001831.812435] [ pid ]   uid  tgid total_vm      rss nr_ptes swapents oom_score_adj name
[4001831.812439] [  555]     0   555     3816      229      12        0             0 upstart-socket-
[4001831.812441] [  803]     0   803     2557      574       8        0             0 dhclient
[4001831.812443] [ 1016]     0  1016     3820       62      11        0             0 upstart-file-br
[4001831.812445] [ 1087]   102  1087    10225      907      24        0             0 dbus-daemon
[4001831.812446] [ 1133]     0  1133     3636       39      12        0             0 getty
[4001831.812448] [ 1136]     0  1136     3636       41      12        0             0 getty
[4001831.812449] [ 1141]     0  1141     3636       38      12        0             0 getty
[4001831.812451] [ 1142]     0  1142     3636       42      12        0             0 getty
[4001831.812452] [ 1144]     0  1144     3636       41      13        0             0 getty
[4001831.812454] [ 1175]     0  1175     4786       41      13        0             0 atd
[4001831.812455] [ 1176]     0  1176     5915      334      18        0             0 cron
[4001831.812457] [ 1219]     0  1219     1093       37       8        0             0 acpid
[4001831.812458] [ 1403]     0  1403     1528       28       8        0             0 getty
[4001831.812460] [ 1416]     0  1416     1090       28       8        0             0 getty
[4001831.812462] [25188]     0 25188     2704       47      10        0             0 xinetd
[4001831.812464] [31864]   108 31864    10960      304      23        0             0 exim4
[4001831.812465] [42796]     0 42796    13773      160      29        0         -1000 sshd
```

```
[4001831.812467] [29141]   110 29141     5330       55      14        0             0 dnsmasq
[4001831.812469] [32851]     0 32851    88015      873      40        0             0 bareos-fd
[4001831.812470] [37847]     0 37847    24518      173      18        0             0 monit
[4001831.812472] [39539]   109 39539    10212      928      21        0             0 snmpd
[4001831.812473] [ 2482]  4040  2482    96833     4956      87        0             0 arv-mount
[4001831.812475] [16139]  4081 16139   254719    75313     233        0             0 arv-mount
[4001831.812477] [18470]  4118 18470   113150     5403      85        0             0 arv-mount
[4001831.812478] [15692]  4035 15692   117953    12282     103        0             0 arv-mount
[4001831.812480] [ 3788]  4085  3788   134036     8338      99        0             0 arv-mount
[4001831.812481] [57572]  4077 57572   113217     4988      88        0             0 arv-mount
[4001831.812483] [ 7765]   112  7765    22324      435      17        0             0 shellinaboxd
[4001831.812484] [ 7768]   112  7768     5918      359      15        0             0 shellinaboxd
[4001831.812486] [29411]  4050 29411   153167    16228     109        0             0 arv-mount
[4001831.812488] [ 7770]  4115  7770    96311     4346      81        0             0 arv-mount
[4001831.812489] [11526]  4202 11526   138408    13586     106        0             0 arv-mount
[4001831.812491] [  314]  4085   314     6479      322      16        0             0 screen
[4001831.812492] [  316]  4085   316     5400     1030      15        0             0 bash
[4001831.812494] [15760]     0 15760    53630     3956     106        0             0 shellinaboxd
[4001831.812495] [16197]  4202 16197     3821     1493      13        0             0 bash
[4001831.812497] [27064]  4202 27064    55541     5517      71        0             0 arv-run
[4001831.812498] [  483]     0   483    53630     3956     106        0             0 shellinaboxd
[4001831.812501] [  876]  4202   876     3821     1494      13        0             0 bash
[4001831.812503] [26886]  4202 26886    55546     5646      77        0             0 arv-run
[4001831.812504] [40742]     0 40742    53631     3957     106        0             0 shellinaboxd
[4001831.812506] [41165]  4202 41165     3821     1493      13        0             0 bash
[4001831.812508] [45224]  4202 45224    55541     5632      77        0             0 arv-run
[4001831.812509] [33921]     0 33921    53621     3960     108        0             0 shellinaboxd
[4001831.812511] [34520]  4202 34520     5933     1504      17        0             0 bash
[4001831.812512] [36486]  4202 36486    56584     5460      76        0             0 arv-run
[4001831.812514] [27607]     0 27607     4870       46      13        0             0 upstart-udev-br
[4001831.812515] [16508]  4085 16508     6479      567      16        0             0 screen
[4001831.812517] [16509]  4085 16509     5400     1054      15        0             0 bash
[4001831.812518] [49124]  4202 49124    97748     4937      87        0             0 arv-mount
[4001831.812520] [36654]  4018 36654   184267    90763     272        0             0 arv-mount
[4001831.812522] [32526]  4042 32526   561519   350058     788        0             0 arv-mount
[4001831.812523] [ 8540]  4073  8540   259859    48943     186        0             0 arv-mount
[4001831.812525] [57991]  4122 57991   254108    72095     233        0             0 arv-mount
[4001831.812527] [33129]     0 33129    28226    17873      60        0             0 systemd-logind
[4001831.812528] [33194]     0 33194     4799      396      13        0             0 irqbalance
[4001831.812534] [36802]     0 36802   142128     2438      41        0             0 docker.io
[4001831.812536] [46168]   992 46168     6178      563      16        0             0 tmux
[4001831.812537] [46183]   992 46183     5358      988      15        0             0 bash
[4001831.812539] [56483]  4069 56483   113737     4511      84        0             0 arv-mount
[4001831.812541] [39091]  4205 39091   113822     4807      87        0             0 arv-mount
[4001831.812542] [ 5678]  4205  5678     6107      545      17        0             0 tmux
[4001831.812544] [ 5679]  4205  5679     6345     1962      17        0             0 bash
[4001831.812545] [25736]     0 25736    12781     8393      30        0             0 atop
[4001831.812547] [42011]     0 42011    26410      680      53        0             0 sshd
[4001831.812549] [42186]  4042 42186    26410      250      52        0             0 sshd
[4001831.812550] [42187]  4042 42187     6345     1992      17        0             0 bash
[4001831.812552] [60902]     0 60902    25891      689      52        0             0 sshd
[4001831.812553] [60903]     0 60903    14911      488      35        0             0 cron
[4001831.812555] [60905]     0 60905     1112      165       7        0             0 sh
[4001831.812556] [60906]     0 60906     3782     1434      12        0             0 bash
[4001831.812558] [60955]     0 60955     1112      137       7        0             0 sh
[4001831.812560] [60956]     0 60956     1085      167       8        0             0 run-parts
[4001831.812561] [60973]     0 60973     1112      166       7        0             0 50-landscape-sy
[4001831.812563] [60980]     0 60980    14167     2360      31        0             0 landscape-sysin
[4001831.812564] [60986]     0 60986     8314     1758      21        0             0 ruby
[4001831.812566] [60987]     0 60987     2617      396      10        0             0 grep
[4001831.812567] Out of memory: Kill process 32526 (arv-mount) score 398 or sacrifice child
[4001831.835080] Killed process 32526 (arv-mount) total-vm:2246076kB, anon-rss:1400232kB, file-rss:0kB
```

#5 - 11/19/2015 03:12 AM - Ward Vandewege

- Project changed from Arvados to Arvados Private

#6 - 11/19/2015 03:44 AM - Peter Amstutz

- Project changed from Arvados Private to Arvados

- Category set to Keep

Some very rough numbers using the program from note-3:

- Upon requesting the manifest, it takes around 20 seconds before data starts flowing (presumably API server is building a signed manifest in this time)
- It takes me about 40 seconds to receive the manifest (but this is at home over DSL)
- It takes another 40 seconds on my laptop to actually parse out the manifest into the data structures

Also, using arv-mount:

- Using "ls" works, but as expected takes 90+ second time to respond.
- Using "cp" to copy files from arv-mount into a directory, I am also seeing resident memory usage of arv-mount creeping up, but it seems to level off (even go down slightly) around 1.7 GiB.

#7 - 12/03/2015 07:55 PM - Brett Smith

- Subject changed from arv-get on qr1hi collections with large manifests takes too long to [SDKs] arv-get should use new PySDK Collection APIs

- Description updated

- Category changed from Keep to SDKs

I am updating this story to mean "make arv-get use Peter's suggested code," since we know that works.

Other issues can be or have been addressed in other tickets. For instance, #7832 will also help with RAM utilization.

I'm not making it a functional requirement, but if we took this opportunity to make arv-get testable and wrote some tests for it, like we did for arv-ls, I wouldn't complain.

#8 - 12/03/2015 07:56 PM - Brett Smith

- Target version set to Arvados Future Sprints

#9 - 02/28/2017 07:40 PM - Peter Amstutz

- Subject changed from [SDKs] arv-get should use new PySDK Collection APIs to [SDKs] arv-get and arv-ls should use new PySDK Collection APIs

#10 - 02/28/2017 07:47 PM - Lucas Di Pentima

- Assigned To set to Lucas Di Pentima

#11 - 03/07/2017 07:57 PM - Tom Morris

- Target version changed from Arvados Future Sprints to 2017-04-12 sprint

- Story points set to 1.0

#12 - 03/07/2017 08:12 PM - Peter Amstutz

- Description updated

- Target version changed from 2017-04-12 sprint to Arvados Future Sprints

#13 - 03/07/2017 08:13 PM - Peter Amstutz

- Target version changed from Arvados Future Sprints to 2017-04-12 sprint

#14 - 03/15/2017 08:07 PM - Tom Morris

- Tracker changed from Bug to Story

- Target version changed from 2017-04-12 sprint to 2017-03-29 sprint

#15 - 03/22/2017 06:29 PM - Lucas Di Pentima

- Status changed from New to In Progress

#16 - 03/29/2017 07:04 PM - Lucas Di Pentima

- Target version changed from 2017-03-29 sprint to 2017-04-12 sprint

#17 - 03/31/2017 04:49 PM - Lucas Di Pentima

Branch 7824-arvls-arvput-collection-api-usage at c820bfc91be7635739bad0857ba3a385d1334b6a*(I've realized the branch is misnamed after pushing it)*
Test run: https://ci.curoverse.com/job/developer-run-tests/217/

#18 - 03/31/2017 06:47 PM - Peter Amstutz

get.py L170

```
                if 0 != string.find(os.path.join(s.stream_name(), f.name),
                                    '.' + get_prefix):
                    continue
```

I think you want os.path.join(s.stream_name(), f.name).startwith('.' + get_prefix) ?

If I try to arv-get a non-existent file from an existing collection, it silently exits (this seems to be consistent with original arv-put, but it doesn't seem desirable):

```
$ arv-get 34t0i-4zz18-mx2hgyidthggk6n/abcdefg
$ echo $?
0
```

Same with directories:

```
$ arv-get 34t0i-4zz18-mx2hgyidthggk6n/abcdefg/ .

$ echo $?
0
```

This also happens (silently exit 0) if I try to arv-get a subdirectory that does exist, but don't provide the trailing '/'

The error trying to arv-get a non-existing collection is not so good:

```
$ arv-get 34t0i-4zz18-abcdeabcdeabcde
Traceback (most recent call last):
  File "/home/peter/work/scripts/venv/bin/arv-get", line 4, in <module>
    __import__('pkg_resources').run_script('arvados-python-client==0.1.20170331164023', 'arv-get')
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/pkg_resources/__init__.py", line 739,
in run_script
    self.require(requires)[0].run_script(script_name, ns)
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/pkg_resources/__init__.py", line 1494,
 in run_script
    exec(code, namespace, namespace)
  File "/home/peter/work/scripts/venv/lib/python2.7/site-packages/arvados_python_client-0.1.20170331164023-py2
.7.egg/EGG-INFO/scripts/arv-get", line 7, in <module>
    sys.exit(main(sys.argv[1:], sys.stdout, sys.stderr))
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.201703311640
23-py2.7.egg/arvados/commands/get.py", line 138, in main
    reader = arvados.CollectionReader(collection, num_retries=args.retries)
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.201703311640
23-py2.7.egg/arvados/collection.py", line 1680, in __init__
    super(CollectionReader, self).__init__(manifest_locator_or_text, *args, **kwargs)
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.201703311640
23-py2.7.egg/arvados/collection.py", line 1236, in __init__
    self._populate()
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.201703311640
23-py2.7.egg/arvados/collection.py", line 1353, in _populate
    raise error_via_api
arvados.errors.ApiError: <HttpError 404 when requesting https://192.168.5.2:8000/arvados/v1/collections/34t0i-
4zz18-abcdeabcdeabcde?alt=json returned "Path not found">
```

Same for arv-ls

```
$ arv-ls 34t0i-4zz18-mx2hgyidthggk6n/abc
Traceback (most recent call last):
  File "/home/peter/work/scripts/venv/bin/arv-ls", line 4, in <module>
    __import__('pkg_resources').run_script('arvados-python-client==0.1.20170331164023', 'arv-ls')
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/pkg_resources/__init__.py", line 739,
in run_script
    self.require(requires)[0].run_script(script_name, ns)
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/pkg_resources/__init__.py", line 1494,
 in run_script
    exec(code, namespace, namespace)
  File "/home/peter/work/scripts/venv/lib/python2.7/site-packages/arvados_python_client-0.1.20170331164023-py2
.7.egg/EGG-INFO/scripts/arv-ls", line 7, in <module>
    sys.exit(main(sys.argv[1:], sys.stdout, sys.stderr))
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.201703311640
23-py2.7.egg/arvados/commands/ls.py", line 49, in main
    num_retries=args.retries)
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.201703311640
23-py2.7.egg/arvados/collection.py", line 1680, in __init__
    super(CollectionReader, self).__init__(manifest_locator_or_text, *args, **kwargs)
```

```
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.20170331164023-py2.7.egg/arvados/collection.py", line 1236, in __init__
    self._populate()
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.20170331164023-py2.7.egg/arvados/collection.py", line 1353, in _populate
    raise error_via_api
arvados.errors.ApiError: <HttpError 404 when requesting https://192.168.5.2:8000/arvados/v1/collections/34t0i-4zz18-mx2hqyidthggk6n%2Fabc?alt=json returned "Path not found">
```

If args.destination "-" then it should check that len(todo) 1

It is missing an explicit outfile.close() at the end of the main read-write loop.

Other thoughts

A couple things I thought of while reviewing. These are not in the original story, so you probably shouldn't do them.

It might be useful to have an --append-existing flag for resuming downloads. If a local file already exists, it would be opened in append mode and it would use file_reader.seek() to skip ahead and only fetch the remaining data in the file.

arv-ls doesn't do directories, and returns a stack trace if you try:

```
$ arv-ls 34t0i-4zz18-gnl9h5j3p026ny9/.git/
Traceback (most recent call last):
  File "/home/peter/work/scripts/venv/bin/arv-ls", line 4, in <module>
    __import__('pkg_resources').run_script('arvados-python-client==0.1.20170331164023', 'arv-ls')
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/pkg_resources/__init__.py", line 739, in run_script
    self.require(requires)[0].run_script(script_name, ns)
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/pkg_resources/__init__.py", line 1494, in run_script
    exec(code, namespace, namespace)
  File "/home/peter/work/scripts/venv/lib/python2.7/site-packages/arvados_python_client-0.1.20170331164023-py2.7.egg/EGG-INFO/scripts/arv-ls", line 7, in <module>
    sys.exit(main(sys.argv[1:], sys.stdout, sys.stderr))
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.20170331164023-py2.7.egg/arvados/commands/ls.py", line 49, in main
    num_retries=args.retries)
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.20170331164023-py2.7.egg/arvados/collection.py", line 1680, in __init__
    super(CollectionReader, self).__init__(manifest_locator_or_text, *args, **kwargs)
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.20170331164023-py2.7.egg/arvados/collection.py", line 1236, in __init__
    self._populate()
  File "/home/peter/work/scripts/venv/local/lib/python2.7/site-packages/arvados_python_client-0.1.20170331164023-py2.7.egg/arvados/collection.py", line 1353, in _populate
    raise error_via_api
arvados.errors.ApiError: <HttpError 404 when requesting https://192.168.5.2:8000/arvados/v1/collections/34t0i-4zz18-gnl9h5j3p026ny9%2F.git%2F?alt=json returned "Path not found">
```

#19 - 03/31/2017 07:00 PM - Peter Amstutz

One last thought. It seems a bit inefficient to iterate over every single file in the collection to only extract a subset (or a single file). Consider using reader.find() to pull out a single file or subdirectory.

#20 - 03/31/2017 07:14 PM - Peter Amstutz

arv-get should get some test cases, as well. At minimum, something like:

- Get a single file from a collection
- Get all files in a collection
- Get collection manifest text
- Error cases: invalid collection, invalid file, invalid destination, destination already exists

#21 - 04/03/2017 09:25 PM - Lucas Di Pentima

Updates at 313415e33
Test run: https://ci.curoverse.com/job/developer-run-tests/218/

Note: tests are pending.

Updates:

- arv-get & arv-ls printed a stack trace when asked for a non existant collection, now they print the error message to the logging facility.
- arv-get did nothing when asked for a non existant file/subdir inside an existing collection, now it errors out.

- arv-ls has the ability to list a subdirectory.
- arv-get no longer traverses all the collection when asked to download just a subcollection.
- Written outfile is close()d when finished downloading.
- Several error message enhancements.

#22 - 04/04/2017 09:10 PM - Lucas Di Pentima

Added test cases at dfd3260e8
Test run: https://ci.curoverse.com/job/developer-run-tests/219/

#23 - 04/06/2017 06:51 PM - Peter Amstutz

LGTM @ dfd3260

#24 - 04/06/2017 07:20 PM - Lucas Di Pentima

- Status changed from In Progress to Resolved

- % Done changed from 0 to 100

Applied in changeset arvados|commit:1affdcd3cd34424a817ae350fd9ca236927fd538.