Arvados - Bug #8189

[FUSE] Listing a project directory is slow when there are many subprojects

01/11/2016 10:12 PM - Jiayong Li

Status:	Resolved	**Start date:**	01/11/2016
Priority:	Normal	**Due date:**	
Assigned To:	Ward Vandewege	**% Done:**	100%
Category:	FUSE	**Estimated time:**	0.00 hour
Target version:	2019-05-22 Sprint		

Description

I created a 2947 sub-projects under the project '1000_genome_exome_raw_reads' with uuid su92l-j7d0g-3c6tenm6q4xn7qm on su92l. As a result, directory operations under that project are slow. For example, 'ls' takes nearly two minutes.

Subtasks:

Task # 8210: review 8189-handle-large-collections-better	**Resolved**

Related issues:

Related to Arvados - Bug #8183: [Workbench] should not look up every group/pr...	**Resolved**	**02/12/2016**

Associated revisions

Revision 37a1505b - 01/14/2016 04:01 PM - Ward Vandewege

Make the Python SDK and workbench effectively default to the API
server's MAX_LIMIT when requesting a list of objects, in those cases
where no explicit limit is set in the client code.

closes #8189

Revision 3fb0f464 - 01/14/2016 08:19 PM - Tom Clegg

Merge branch '8189-handle-large-collections-better' refs #8189

History

#1 - 01/12/2016 06:55 PM - Brett Smith

- Subject changed from [Keep] Directory operations are slow after the creation of a large number of projects. to [FUSE] Listing a project directory is slow when there are many subprojects

#2 - 01/12/2016 10:08 PM - Tom Clegg

arvados_fuse's ProjectDirectory class uses arvados.util.list_all:

```
                contents = arvados.util.list_all(self.api.groups().contents,
                                                 self.num_retries, uuid=self.project_uuid)
```

arvados.util.list_all doesn't set a limit either, so we get the API's default limit of 100 items per page.

Suggest modifying arvados.util.list_all (in source:sdk/python/arvados/util.py#L365) to do something like

```
kwargs.setdefault('limit', sys.maxint)
```

That way, the API server's MAX_LIMIT (currently 1000) will determine the page size.

The rationale is that, once the client is in an API request loop that it won't exit until it gets all of the items, it's never a good idea for it to get *fewer* items per API request. Getting fewer items per page only makes sense if the client has some chance of doing something else (exiting the loop or processing a subset of results) before receiving MAX_LIMIT results.

(ArvadosResourceList#each_page in source:apps/workbench/app/models/arvados_resource_list.rb#177 needs this fix, too.)

#3 - 01/14/2016 04:02 PM - Ward Vandewege

- Assigned To set to Ward Vandewege

- Target version set to 2016-01-20 Sprint

#4 - 01/14/2016 04:05 PM - Ward Vandewege

- Status changed from New to Resolved

- % Done changed from 0 to 100

Applied in changeset arvados|commit:37a1505b607bbf533512f48b47f208c5cde4c435.

#5 - 01/14/2016 04:16 PM - Tom Clegg

- Status changed from Resolved to In Progress

#6 - 01/15/2016 07:51 PM - Jiayong Li

Right now su92l workbench is still considerably slower than qr1hi workbench.

On su92l, I also noticed huge performance difference between read-only mount and writable mount (both freshly mounted to reflect recent changes).

read-only mount:

```
$ time ls keep/home/arvados_genomics_benchmark/1000_genome_exome_raw_reads
real    1m0.422s
user    0m0.020s
sys    0m0.060s
```

writable mount:

```
$ time ls mnt/home/arvados_genomics_benchmark/1000_genome_exome_raw_reads
real    94m40.150s
user    0m0.028s
sys    0m0.080s
```

#7 - 01/20/2016 04:26 PM - Brett Smith

- Target version deleted (2016-01-20 Sprint)

#8 - 01/20/2016 06:44 PM - Brett Smith

- Target version set to Arvados Future Sprints

#9 - 02/03/2016 08:50 PM - Jiayong Li

I tried running a pipeline on su92l, but the "Run a pipeline" button on the workbench homepage is not clickable now.

#10 - 05/06/2019 05:00 PM - Ward Vandewege

- Status changed from In Progress to Resolved

- Target version changed from Arvados Future Sprints to 2019-05-22 Sprint

This was resolved long ago, here's the performance today:

```
wardv@shell:~$ time ls keep/by_id/su92l-j7d0g-3c6tenm6q4xn7qm
```

```
...
```

```
real    0m8.187s
user    0m0.021s
sys    0m0.086s
```

#11 - 05/21/2019 10:27 PM - Tom Morris

- Release set to 15