Arvados - Story #11876

[R SDK] Create a Bioconductor/R SDK

06/20/2017 07:31 PM - Tom Morris

Status:	Closed	**Start date:**	06/20/2017
Priority:	Normal	**Due date:**	
Assigned To:	Fuad Muhic	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	2018-04-25 Sprint		

Description

Overview:

As an R programmer I'd like to have the ability to query the Arvados APIs directly from R using a package which integrates well with and is published with the rest of the Bioconductor packages. The SDK should ideally allow me to do everything a Python programmer can do using the Python SDK.

As a first step, the R SDK should allow me to allow to find collections and files in Keep using filtering on metadata, load the files into R, process them and then write the results back to a collection.

As an optional second stage, it'd be useful to be able to submit CWL jobs and monitor their progress.

The SDK should work on Windows, OS X, and Linux, which implies that depending on arv-mount for file reading and writing is not an acceptable option. Instead, we will use the webdav support in keep-web. Read-only support is already available (completed in issue #12216). Write support is forthcoming, see issue #12483.

A potential supporting component might be googleAuthR http://code.markedmondson.me/googleAuthR/ which could be used in a similar way to googleComputeEngineR https://cloudyr.github.io/googleComputeEngineR/ and other packages which are layered on it. googleAuthR can be used for API generation and response parsing, but needs to be reworked to not assume Google authentication or endpoints. Instead of the OAuth2 dance, it needs to be able to use an API token.

This code snippet will generate an entire R stub package using googleAuthR:

```
library('googleAuthR')
url="https://qr1hi.arvadosapi.com/discovery/v1/apis/arvados/v1/rest"
req <- httr::RETRY("GET", url)
httr::stop_for_status(req)
content <- httr::content(req,as="text")
api_description <- jsonlite::fromJSON(content)
paste("Loaded API description ", api_description$name, api_description$version)

"Generating API skeleton"
gar_create_api_objects(filename = "arvados_objects.R",api_json = api_description)
gar_create_api_skeleton('arvados_functions.R', api_description, format=TRUE)
"API Generation complete"

# Make sure we can load our newly generated code
source('arvados_functions.R')
source('arvados_objects.R')

# Generate the whole package at once
gar_create_package(api_description, '/tmp/aRv', rstudio = TRUE, check = TRUE, github = FALSE)
```

There gar_create_package call does the whole thing including man pages, README, etc, but the gar_create_api_objects and gar_create_api_skeleton, can be used to just do a part of the process.

The generator assumes the context of a Google API, so has a bunch of built-in assumptions that need to be cleaned up. Below is a non-exaustive list:

- authentication - switch from Google auth to Arvados token based authentication, remove/fix all references to googleAuthR::gar_auth() and Google API scopes
- fixed base URL - in the above example qr1hi.arvadosapi.com is hardwired into the API. This needs to be configurable by the caller.

- man page generation - there's a bunch of warnings due formatting in the docs
- Bioconductor packaging, types, conventions,tests - the core generator targets CRAN tests. This may need to be extended for Bioconductor
- LICENSE & AUTHOR - these are wrong need to figure out where their contents come from

Arvados specific things to pay attention to:

- URL encoding of JSON in query string
- Arvados objects - Collections - manifest parsing, updating, etc.
- WebDAV client to Arvados WebDAV server (depends
- Remove unused / disabled APIs e.g. Crunch1
- Add Jenkins CI job

It is desirable that changes to the code generator be done in such as way that they can be adopted by the upstream project as parameterizable options, but it's not mandatory.

There are also additional things which need to be added:

- tests
- vignettes/examples

Some hints on testing and other advanced API topics are here:
http://code.markedmondson.me/googleAuthR/articles/advanced-building.html

There are two relevant packages, SevenBridges "sevenbridges" and Illumina's "BaseSpaceR", which could be used to compare against or as sources for code (they are both Apache licensed).

http://bioconductor.org/packages/release/bioc/html/sevenbridges.html
https://github.com/sbg/sevenbridges-r
http://bioconductor.org/packages/release/bioc/html/BaseSpaceR.html
https://developer.basespace.illumina.com/docs/content/documentation/sdk-samples/r-sdk-overview

Subtasks:

Task # 12643: Review	**Resolved**
Task # 13033: Integrate into run-test.sh	**Resolved**

Related issues:

Related to Arvados - Feature #12216: [keep-web] machine-readable file listings	**Resolved**	**10/11/2017**
Related to Arvados - Feature #12483: [keep-web] writable webdav	**Resolved**	**10/25/2017**
Related to Arvados - Story #12706: [SDK] R SDK support for Collections	**Resolved**	**01/17/2018**
Related to Arvados - Story #13391: Get R SDK into Bioconductor	**Duplicate**	
Blocks Arvados - Story #13077: R SDK submit to Bioconductor	**New**	

Associated revisions

Revision 7c8c1719 - 12/21/2017 02:22 AM - Ward Vandewege

Fix typo

refs #11876

Arvados-DCO-1.1-Signed-off-by: Ward Vandewege <wvandewege@veritasgenetics.com>

Revision 082df78f - 12/21/2017 02:50 AM - Ward Vandewege

Expand the README a bit, fix some syntax errors, add instructions for
building/installing from source.

refs #11876

Arvados-DCO-1.1-Signed-off-by: Ward Vandewege <wvandewege@veritasgenetics.com>

Revision b37da0c1 - 12/28/2017 09:18 PM - Ward Vandewege

Dependency update.

refs #11876

Arvados-DCO-1.1-Signed-off-by: Ward Vandewege <wvandewege@veritasgenetics.com>

Revision 8fb1de64 - 02/06/2018 06:15 PM - Peter Amstutz

Merge branch '11876-r-sdk' refs #11876

Arvados-DCO-1.1-Signed-off-by: Peter Amstutz <pamstutz@veritasgenetics.com>

Revision 53e11d8a - 02/09/2018 07:20 PM - Peter Amstutz

Merge branch '11876-r-sdk' documentation, refs #11876

Arvados-DCO-1.1-Signed-off-by: Peter Amstutz <pamstutz@veritasgenetics.com>

Revision f344ada0 - 02/15/2018 04:07 PM - Peter Amstutz

R SDK documentation link fix refs #11876

Arvados-DCO-1.1-Signed-off-by: Peter Amstutz <pamstutz@veritasgenetics.com>

Revision f53e3ede - 02/15/2018 07:51 PM - Peter Amstutz

Merge branch '11876-R-deps' refs #11876

Arvados-DCO-1.1-Signed-off-by: Peter Amstutz <pamstutz@veritasgenetics.com>

History

#1 - 06/20/2017 07:45 PM - Tom Morris

- Description updated

- Assigned To set to Radhika Chippada

Radhika - please research and refine

#2 - 06/26/2017 02:39 PM - Radhika Chippada

https://cran.r-project.org/web/packages/googleAuthR/README.html

#3 - 07/20/2017 01:16 AM - Tom Morris

- Description updated

- Assigned To changed from Radhika Chippada to Tom Morris

#4 - 10/18/2017 06:57 PM - Tom Morris

- Description updated

#5 - 10/25/2017 04:59 PM - Ward Vandewege

- Description updated

#6 - 10/25/2017 05:46 PM - Ward Vandewege

- Description updated

#7 - 11/07/2017 04:44 PM - Tom Morris

- Description updated

- File arvados_objects.R added

- File arvados_functions.R added

#8 - 11/07/2017 07:48 PM - Tom Morris

- Description updated

#9 - 11/21/2017 06:02 PM - Peter Amstutz

https://github.com/fmuhic/ArvadosSDK

#10 - 11/22/2017 03:17 PM - Peter Amstutz

Here's an example curl line to get collection contents via WebDAV:

```
curl -v -X PROPFIND -H "Authorization: OAuth2 4invqy35tf70t7hmvdc83ges8ug9cklhgqq1l8gj2cjn18teuq" https://coll
```

```
ections.4xphq.arvadosapi.com/c=4xphq-4zz18-9d5b0qm4fgijeyi/_/
```

#11 - 11/22/2017 08:02 PM - Tom Morris

- Target version changed from Arvados Future Sprints to 2017-12-06 Sprint

#12 - 11/22/2017 08:21 PM - Tom Morris

- Assigned To changed from Tom Morris to Peter Amstutz

#13 - 11/27/2017 09:59 PM - Peter Amstutz

- Assigned To changed from Peter Amstutz to Fuad Muhic

#14 - 11/28/2017 07:16 PM - Peter Amstutz

Hi Fuad:

I've created a branch for the R sdk work. The code should go in arvados/sdk/R

Instructions:

```
git clone git@git.curoverse.com:arvados.git
git checkout --track -b 11876-r-sdk origin/11876-r-sdk
```

Please use "git push" to synchronize your changes daily.

Your commit messages will need a Developer Certificate of Origin (DCO), that means the commit message needs to contain this line:

```
Arvados-DCO-1.1-Signed-off-by: Fuad Muhic <fmuhic@capeannenterprises.com>
```

#15 - 11/29/2017 06:33 PM - Tom Morris

- Status changed from New to In Progress

#16 - 11/29/2017 09:35 PM - Peter Amstutz

- Related to Story #12706: [SDK] R SDK support for Collections added

#17 - 11/30/2017 02:25 PM - Peter Amstutz

Fuad:

Please put the "DCO" line *after* the main commit message, othewise it clutters up the commit log.

#18 - 12/05/2017 06:24 PM - Peter Amstutz

Fuad:

In order to find the keep-web (WebDAV) server for a cluster, you will need to look it up in the discovery document.

This is located on the API server at /discovery/v1/apis/arvados/v1/rest

This is a JSON document which stores configuration parameters for the cluster, which the SDK should fetch during initialization. Specifically the key keepWebServiceUrl is the base URL for the WebDAV service, for example "https://downloads.4xphq.arvadosapi.com/".

However, I just today noticed that this configuration value is missing. I have a branch to fix it, tracked in issue #12752. It will be merged soon but will probably require a few days to get deployed.

#19 - 12/06/2017 08:06 PM - Tom Morris

- Target version changed from 2017-12-06 Sprint to 2017-12-20 Sprint

#20 - 12/07/2017 03:44 PM - Peter Amstutz

Hi Fuad,

Can you add some examples / test cases for doing various operations with the R SDK, and update the README with instructions for running them?

#21 - 12/19/2017 02:16 PM - Peter Amstutz

In addition to collections, we also need support for querying projects.

A project is simply a "group" record with the field "group_class" = "project" http://doc.arvados.org/api/methods/groups.html

The REST API is otherwise the same as collections (same get/put/post/delete semantics and filters), so depending on how you implemented it, the quickest way to implement this may be a cut and paste of the collections code with the endpoint changed.

#22 - 12/21/2017 08:54 PM - Peter Amstutz

Code review comments: 11876-r-sdk @ 99bec637f6d4384a8d6f3c2cb27eb32d13c14f21

- Needs unit testing, and integration into run-test.sh. The run-tests.sh framework will provide a test server for R SDK tests to communicate with. Recommended test framework for R seems to be "testthat"

https://github.com/r-lib/testthat

- This should be callable without any filters:

  ```
  arv$listCollections()
  Error in names(filters) <- c("collection") :
  ```

- I'm getting an error creating a collection:

  ```
  collection <- Collection$new(arv, "c97qk-4zz18-k1kpkv1ign5kcdu")
  Error in curl::curl_fetch_memory(url, h) : <url> malformed
  ```

  On further research, it looks like the upstream cluster has "keepWebServiceUrl" misconfigured, so this isn't really due to a bug in your code (although the error reporting should be better.)

- I think it is later, so you can remove this comment?

  ```
          #Todo(Fudo): Hardcoded credentials to WebDAV server. Remove them later
  ```

- You should add something similar to the "listAll" function in the Python SDK. This uses the "offset" query parameter to get all items when the number of items is > maximum API limit. https://dev.arvados.org/projects/arvados/repository/revisions/master/entry/sdk/python/arvados/util.py#L375

- Instead of having two different modes for Collection$add, I suggest renaming the first one to "create" and having it return an ArvadosFile:

  ```
  collectionFile <- collection$create("main.cpp", "cpp/src/")
  ```

  Then the "add" method only does one thing.

- Instead of "collection$getFileContent()" I suggest calling it "getFileListing()"

#23 - 12/21/2017 09:03 PM - Peter Amstutz

It looks like the R concept of a generalized IO object is called a "connection":

https://www.rdocumentation.org/packages/base/versions/3.4.3/topics/connections

https://stackoverflow.com/questions/30445875/what-exactly-is-a-connection-in-r

Figure out how to expose ArvadosFile as a "connection" so that it can be used directly for loading and saving data.

#24 - 01/02/2018 06:09 PM - Tom Morris

- Target version changed from 2017-12-20 Sprint to 2018-01-17 Sprint

#25 - 01/17/2018 07:41 PM - Peter Amstutz

- Target version changed from 2018-01-17 Sprint to 2018-01-31 Sprint

#26 - 02/01/2018 08:51 PM - Tom Morris

- Target version changed from 2018-01-31 Sprint to 2018-02-14 Sprint

#27 - 02/02/2018 06:21 PM - Peter Amstutz

On packaging: http://r-pkgs.had.co.nz/

#28 - 02/02/2018 07:16 PM - Peter Amstutz

note to self

```
.libPaths( "/var/lib/arvados/Rstuff")
devtools::install_dev_deps()
```

#29 - 02/03/2018 04:16 AM - Peter Amstutz

I pushed a commit to 11876-r-sdk which adds support for running the R SDK unit tests as part of the overall Arvados test suite. I also noticed several tests are failing, can you confirm?

#30 - 02/05/2018 03:35 PM - Peter Amstutz

```
                   ********** Running sdk/R tests **********

> results <- devtools::test()
Loading ArvadosR
Loading required package: testthat
Testing ArvadosR
Arvados API: .........................
ArvadosFile: ........................
Collection: .........................
CollectionTree: ...............
Http Parser: .....
Http Request: ...
REST service: ........1....2.......3.....4.....5.....................................
Subcollection: .................................
Utility function: ..........

Failed -------------------------------------------------------------------------
1. Error: getResource raises exception if response contains errors field (@test-RESTService.R#76) 
is.character(regexp) is not TRUE
1: expect_that(REST$getResource("collections", resourceUUID), throws_error(404)) at /usr/src/arvados/sdk/R/tes
ts/testthat/test-RESTService.R:76
2: condition(object)
3: expect_error(x, regexp, ...)
4: expect_match(error$message, regexp, ..., info = info)
5: stopifnot(is.character(regexp), length(regexp) == 1)
6: stop(sprintf(ngettext(length(r), "%s is not TRUE", "%s are not all TRUE"), ch), call. = FALSE, 
       domain = NA)

2. Error: listResources raises exception if response contains errors field (@test-RESTService.R#117) 
is.character(regexp) is not TRUE
1: expect_that(REST$listResources("collections"), throws_error(404)) at /usr/src/arvados/sdk/R/tests/testthat/
test-RESTService.R:117
2: condition(object)
3: expect_error(x, regexp, ...)
4: expect_match(error$message, regexp, ..., info = info)
5: stopifnot(is.character(regexp), length(regexp) == 1)
6: stop(sprintf(ngettext(length(r), "%s is not TRUE", "%s are not all TRUE"), ch), call. = FALSE, 
       domain = NA)

3. Error: deleteCollection raises exception if response contains errors field (@test-RESTService.R#190) 
is.character(regexp) is not TRUE
1: expect_that(REST$deleteResource("collections", resourceUUID), throws_error(404)) at /usr/src/arvados/sdk/R/
tests/testthat/test-RESTService.R:190
2: condition(object)
3: expect_error(x, regexp, ...)
4: expect_match(error$message, regexp, ..., info = info)
5: stopifnot(is.character(regexp), length(regexp) == 1)
6: stop(sprintf(ngettext(length(r), "%s is not TRUE", "%s are not all TRUE"), ch), call. = FALSE, 
       domain = NA)

4. Error: updateResource raises exception if response contains errors field (@test-RESTService.R#238) 
is.character(regexp) is not TRUE
1: expect_that(REST$updateResource("collections", resourceUUID, newResourceContent), 
       throws_error(404)) at /usr/src/arvados/sdk/R/tests/testthat/test-RESTService.R:238
2: condition(object)
3: expect_error(x, regexp, ...)
4: expect_match(error$message, regexp, ..., info = info)
5: stopifnot(is.character(regexp), length(regexp) == 1)
6: stop(sprintf(ngettext(length(r), "%s is not TRUE", "%s are not all TRUE"), ch), call. = FALSE, 
       domain = NA)

5. Error: createResource raises exception if response contains errors field (@test-RESTService.R#288) 
```

```
is.character(regexp) is not TRUE
1: expect_that(REST$createResource("collections", resourceContent), throws_error(404)) at /usr/src/arvados/sdk
/R/tests/testthat/test-RESTService.R:288
2: condition(object)
3: expect_error(x, regexp, ...)
4: expect_match(error$message, regexp, ..., info = info)
5: stopifnot(is.character(regexp), length(regexp) == 1)
6: stop(sprintf(ngettext(length(r), "%s is not TRUE", "%s are not all TRUE"), ch), call. = FALSE, 
       domain = NA)

DONE ===========================================================================
> any_error <- any(as.data.frame(results)$error)
> if (any_error) {
+   q("no", 1)
+ } else {
+   q("no", 0)
+ }

             ********** !!!!!! sdk/R tests FAILED !!!!!! **********
```

#31 - 02/06/2018 02:48 PM - Peter Amstutz

Tests failing on CI: https://ci.curoverse.com/job/developer-run-tests-remainder/581/consoleFull

#32 - 02/08/2018 09:55 PM - Nico César

for the record, if dependencies are there this will work:

```
> install.packages('http://r.arvados.org/ArvadosR_0.0.3.tar.gz', repos = NULL, type="source")
```

this is a small step towards getting it into Bioconductor Package system

#33 - 02/14/2018 08:18 PM - Tom Morris

- Target version changed from 2018-02-14 Sprint to 2018-02-28 Sprint

#34 - 02/28/2018 08:16 PM - Tom Morris

- Target version changed from 2018-02-28 Sprint to 2018-03-14 Sprint

#35 - 03/14/2018 07:11 PM - Peter Amstutz

- Target version changed from 2018-03-14 Sprint to 2018-03-28 Sprint

#36 - 03/28/2018 03:20 PM - Tom Morris

- Target version changed from 2018-03-28 Sprint to 2018-04-11 Sprint

#37 - 04/11/2018 03:20 PM - Tom Morris

- Target version changed from 2018-04-11 Sprint to 2018-04-25 Sprint

#38 - 04/25/2018 03:26 PM - Tom Morris

- Related to Story #13391: Get R SDK into Bioconductor added

#39 - 04/25/2018 03:29 PM - Tom Morris

- Status changed from In Progress to Closed

#40 - 06/11/2018 08:22 PM - Tom Morris

- Blocks Story #13077: R SDK submit to Bioconductor added

#41 - 07/23/2018 07:00 PM - Tom Morris

- Release set to 13

Files

arvados_objects.R	58.2 KB	11/07/2017	Tom Morris
arvados_functions.R	263 KB	11/07/2017	Tom Morris