Package 'rdtLite'

July 13, 2020

Title Provenance Collector

Version 1.3

Date 2020-07-09

Copyright President and Fellows of Harvard College, Trustees of Mount Holyoke College

Depends R (>= 3.6.0)

Description Defines functions that can be used to collect provenance as an R script executes or during a console session. The output is a text file in PROV-JSON format.

License GPL-3 | file LICENSE

URL https://github.com/End-to-end-provenance/rdtLite

BugReports https://github.com/End-to-end-provenance/rdtLite/issues

Imports curl, digest, grDevices, gtools, jsonlite, knitr, methods, provSummarizeR, provViz (>= 1.0.6), rlang, rmarkdown, sessioninfo, stringi, tools, utils, XML

Suggests ggplot2, roxygen2, testthat

VignetteBuilder knitr

RoxygenNote 7.1.0

NeedsCompilation no

Author Barbara Lerner [aut, cre],
Emery Boose [aut],
Elizabeth Fong [aut],
Luis Perez [aut],
Thomas Pasquier [ctb],
Matthew Lau [ctb],
Yada Pruksachatkun [ctb],
Alex Liu [ctb],
Moe Pwint Phyu [ctb],
Connor Gregorich-Trevor [ctb],
Aaron Ellison [res],
Margo Seltzer [res],

Joe Wonsil [res],
Orenna Brand [res]

Maintainer Barbara Lerner <blerner@mtholyoke.edu>

Repository CRAN

Date/Publication 2020-07-13 19:50:03 UTC

R topics documented:

`prov.init` *Provenance Collection Functions*

Description

prov.init intializes a new provenance graph. This function can be executed in the console or placed inside an R script.

prov.save saves the current provenance graph to a prov-json file. If more R statements are executed, the provenance for these statements is added to the graph. The graph is finalized with prov.quit. This function can be executed in the console or placed inside an R script.

prov.quit saves and closes the current provenance graph. This function can be executed in the console or placed inside an R script.

prov.run initiates execution of a script and collects provenance as the script executes. This function should be used if you want to collect provenance for a script that is in an R file and you do not want to modify the R script directly to include calls to prov.init, prov.save and prov.quit. It essentially wraps the execution of the script with calls to prov.init and prov.quit.

prov.source loads an R script and executes it, collecting provenance as it does so. It assumes that provenance has already been initialized, either via a call to prov.init, or because the R script was executed using prov.run. If you want to collect provenance inside scripts that are loaded with R's source function, you should replace calls to source with calls to prov.source.

Usage

```
prov.init(
  prov.dir = NULL,
  overwrite = TRUE,
  snapshot.size = 0,
  hash.algorithm = "md5",
  save.debug = FALSE
)

prov.save(save.debug = FALSE)
```

```
prov.quit(save.debug = FALSE)

prov.run(
  r.script.path,
  prov.dir = NULL,
  overwrite = TRUE,
  details = TRUE,
  snapshot.size = 0,
  hash.algorithm = "md5",
  save.debug = FALSE,
  exprs,
  ...
)

prov.source(file, exprs, ...)
```

Arguments

prov.dir	the directory where the provenance graph will be saved. If not provided, the directory specified by the prov.dir option is used. Otherwise the R session temporary directory is used.
overwrite	if FALSE, includes a time stamp in the provenance graph directory name.
snapshot.size	the maximum size for snapshot files. If 0, no snapshots are saved. If Inf, the complete state of an object is stored in the snapshot file. For other values, the head of the object, truncated to a size near the specified limit, is saved. The size is in kilobytes.
hash.algorithm	the hash algorithm to use for files. Choices are md5 (default), sha1, crc32, sha256, sha512, xxhash32, xxhash64 and murmur32. This feature uses the digest function from the digest package.
save.debug	If TRUE, debug files are saved to the debug directory. This is intended for developers of the rdt / rdtLite package.
r.script.path	the full path to the R script file that is being executed. A copy of the script will be saved with the provenance graph.
details	if FALSE, provenance is not collected for top-level statements.
exprs	Instead of specifying file, an expression, call, or list of call's, can be passed in to be executed.
...	parameters passed on to the source function. See documentation of source for details.
file	the name of the R script file to source.

Details

rdtLite is an R package that collects provenance as an R script executes. The resulting provenance provides a detailed record of the execution of the script and includes information on the steps that were performed and the intermediate data values that were created. The resulting provenance can

be used for a wide variety of applications that include debugging scripts, cleaning code, and reproducing results.

There are two ways in which a user can collect provenance. To collect provenance from commands stored in a script file, use prov.run. This will execute the commands that are in the script, collecting provenance as it does so.

The user can also collect provenance while executing commands in the console. To do this, first execute prov.init. Then enter console commands as normal. When done with the commands for which you want provenance, use prov.quit. If you want to save the current provenance without turning off provenance collection, call prov.save instead of prov.quit. You can call prov.save multiple times before calling prov.quit. Each call will append to the same provenance file.

The provenance is stored in PROV-JSON format. For immediate use it may be retrieved from memory using the prov.json function. For later use the provenance is also written to the file prov.json. This file and associated files are written by default to the R session temporary directory. The user can change this location by (1) using the optional parameter prov.dir in the prov.run or prov.init functions, or (2) setting the prov.dir option (e.g. by using the R options command or editing the Rprofile.site or .Rprofile file). If prov.dir is set to ".", the current working directory is used.

If prov.source is called when provenance is not initialized, it will just source the file. No provenance will be collected.

Value

prov.init initializes the provenance collector. The prov.init function does not return a value.

prov.save writes the current provenance to a file but does not return a value.

prov.quit writes the current provenance to a file but does not return a value.

prov.run runs a script, collecting provenance as it does so. It does not return a value.

The prov.source function does not return a value.

See Also

prov.json for access to the JSON text of the provenance,

Examples

```
## Not run: prov.run ("script.R")
## Not run: prov.source ("script.R")
prov.init()
a <- 1
b <- 2
prov.save()
ab <- a + b
prov.quit()
```

prov.json *Provenance Access Functions*

Description

prov.json returns the current provenance graph as a prov-json string.

prov.dir returns the current provenance directory.

prov.visualize displays the current provenance as a graph.

prov.summarize outputs a text summary to the R console

Usage

```
prov.json()

prov.dir()

prov.visualize()

prov.summarize(save = FALSE, create.zip = FALSE)
```

Arguments

save	if true saves the summary to the file prov-summary.txt in the provenance directory
create.zip	if true all of the provenance data will be packaged up into a zip file stored in the current working directory.

Details

rdtLite collects provenance as a script executes. Once collected, prov.json can be called to access the provenance as a JSON string. This is useful for applications that operate on the provenance. The JSON is consistent with the PROV-JSON standard.

One such application is a graphic visualizer built into rdt. To view the provenance graphically, call prov.visualize. In the provenance graph, the nodes are data values and operations, with edges connecting them to show data and control flow dependencies. The visualizer also allows the user to view intermediate values of variables, and to graphically view the lineage of how a value was computed, or to look at how a value is used moving forward in the computation. The user can also search for specific data or operation nodes, files, or error messages in the provenance.

Creating a zip file depends on a zip executable being on the search path. By default, it looks for a program named zip. To use a program with a different name, set the value of the R_ZIPCMD environment variable. This code has been tested with Unix zip and with 7-zip on Windows.

Value

prov.json returns the current provenance graph as a prov-json string

prov.dir returns the current provenance directory.

prov.visualize loads and displays the current provenance graph in DDG Explorer. The prov.visualize function does not return a value.

References

PROV-JSON standard: https://www.w3.org/Submission/2013/SUBM-prov-json-20130424/

PROV-JSON output produced by rdtLite: https://github.com/End-to-end-provenance/ExtendedProvJson/blob/master/JSON-format.md

Applications that use the provenance: https://github.com/End-to-end-provenance/End-to-end-provenance.github.io/blob/master/RTools.md

See Also

prov.init and prov.run for functions to collect provenance

Examples

```
prov.init()
a <- 1
b <- 2
ab <- a + b
prov.quit()
str <- prov.json()
pdir <- prov.dir()
## Not run: prov.visualize()
```

Index