Package: purrrow (via r-universe)

August 11, 2024

Title Out-of-memory data collation into Arrow datasets

Version 0.0.0.9000

Description Iterate over a function and collate its output into an Arrow dataset, without loading the whole result set into memory.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.1

Imports arrow (>= 2.0.0), purrr (>= 0.3.4), lifecycle, magrittr

URL http://petrbouchal.xyz/purrrow, https://github.com/petrbouchal/purrrow

BugReports https://github.com/petrbouchal/purrrow/issues

Suggests dplyr (>= 1.0.3), testthat (>= 3.0.0)

Config/testthat/edition 3

Repository https://petrbouchal.r-universe.dev

RemoteUrl https://github.com/petrbouchal/purrrow

RemoteRef HEAD

RemoteSha 6bf3d0930ae397c0739dbf04c486ab62efcde725

Contents

`marrow_dir` *Iteratively collate output of function into an Arrow dataset out of memory*

Description

Experimental map + arrow: iterate over a function and collate the results into an Arrow dataset. This happens without the whole dataset being in memory, so is suitable for large data objects. The function must return a data.frame or tibble. The returned value is a path to the directory containing the Arrow dataset.

Usage

```
marrow_dir(.x, .f, ..., .path, .partitioning = c(), .format = "parquet")

marrow_ds(.x, .f, ..., .path, .partitioning = c(), .format = "parquet")

marrow_files(.x, .f, ..., .path, .partitioning = c(), .format = "parquet")
```

Arguments

`.x`	vector or list of values for .f to iterate over
`.f`	function; must return a data.frame/tibble
`...`	other arguments to .f
`.path`	path to directory where collated Arrow dataset will be stored. will be created if it does not exist
`.partitioning`	character vector of columns to use for partitioning. Columns must exist in output of .f.
`.format`	"parquet" (the default) or "arrow".

Value

path to new dataset directory; character string of length one.

an Arrow Dataset

character vector containing paths to all files in dataset dir

Functions

- `marrow_dir`: Return path to directory containing dataset
- `marrow_ds`: Return Arrow Dataset
- `marrow_files`: Return paths to all files in dataset dir

Examples

```
months <- unique(airquality$Month)
td <- tempdir()
part_of_aq <- function(month) {
  airquality[airquality$Month==month,]
}

aq_arrow <- purrrow:::marrow_dir(months, part_of_aq,
                                 .path = td)
```

Index