Package ‘gbifdb’

October 19, 2023

Version 1.0.0
Title High Performance Interface to ‘GBIF’
Description A high performance interface to the Global Biodiversity Information Facility, ‘GBIF’. In contrast to 'rgbif', which can access small subsets of 'GBIF' data through web-based queries to a central server, 'gbifdb' provides enhanced performance for R users performing large-scale analyses on servers and cloud computing providers, providing full support for arbitrary ‘SQL’ or ‘dplyr’ operations on the complete ‘GBIF’ data tables (now over 1 billion records, and over a terabyte in size). 'gbifdb' accesses a copy of the ‘GBIF’ data in 'parquet' format, which is already readily available in commercial computing clouds such as the Amazon Open Data portal and the Microsoft Planetary Computer, or can be accessed directly without downloading, or downloaded to any server with suitable bandwidth and storage space. The high-performance techniques for local and remote access are described in <https://duckdb.org/why_duckdb> and <https://arrow.apache.org/docs/r/articles/fs.html> respectively.

License Apache License (>= 2)
Encoding UTF-8
ByteCompile true
Depends R (>= 4.0)
Imports arrow (>= 8.0.0), dplyr, duckdbfs
Suggests spelling, dbplyr, testthat (>= 3.0.0), covr, knitr, rmarkdown, minioclient
URL https://docs.ropensci.org/gbifdb/, https://github.com/ropensci/gbifdb
BugReports https://github.com/ropensci/gbifdb
Language en-US
RoxygenNote 7.2.3
Config/testthat/edition 3
gbif_dir

VignetteBuilder knitr

NeedsCompilation no

Author Carl Boettiger [aut, cre] (<https://orcid.org/0000-0002-1642-628X>)

Maintainer Carl Boettiger <cboettig@gmail.com>

Repository CRAN

Date/Publication 2023-10-19 20:30:03 UTC

R topics documented:

- gbif_dir ........................................... 2
- gbif_download .................................. 3
- gbif_example_data .............................. 4
- gbif_local ....................................... 4
- gbif_remote ...................................... 5
- gbif_version ..................................... 6

Index 8

---

**gbif_dir**

*Default storage location*

**Description**

Default location can be set with the env var GBIF_HOME, otherwise will use the default provided by `tools::R_user_dir()`

**Usage**

`gbif_dir()`

**Value**

path to the gbif home directory directory

**Examples**

`gbif_dir()`
gbif_download  Download GBIF data using minioclient

Description
Sync a local directory with selected release of the AWS copy of GBIF

Usage
```r
gbif_download(
    version = gbif_version(),
    dir = gbif_dir(),
    bucket = gbif_default_bucket(),
    region = ""
)
```

Arguments
- **version**: Release date (YYYY-MM-DD) which should be synced. Will detect latest version by default.
- **dir**: path to local directory where parquet files should be stored. Fine to leave at default, see `gbif_dir()`.
- **bucket**: Name of the regional S3 bucket desired. Default is "gbif-open-data-us-east-1". Select a bucket closer to your compute location for improved performance, e.g. European researchers may prefer "gbif-open-data-eu-central-1" etc.
- **region**: bucket region (usually ignored? Just set the bucket appropriately)

Details
Note that data can also be found on the Microsoft Cloud, https://planetarycomputer.microsoft.com/dataset/gbif
Also, some users may prefer to download this data using an alternative interface or work on a cloud-host machine where data is already available. Note, these data include all CC0 and CC-BY licensed data in GBIF that have coordinates which passed automated quality checks, see https://github.com/gbif/occurrence/blob/master/aws-public-data.md.

Value
logical indicating success or failure.

Examples
```r
gbif_download()
```
### gbif_example_data

Return a path to the directory containing GBIF example parquet data

#### Description

Return a path to the directory containing GBIF example parquet data

#### Usage

```r
gbif_example_data()
```

#### Details

element data is taken from the first 1000 rows of the 2011-11-01 release of the parquet data.

#### Value

path to the example occurrence data installed with the package.

#### Examples

```r
gbif_example_data()
```

### gbif_local

Local connection to a downloaded GBIF Parquet database

#### Description

Local connection to a downloaded GBIF Parquet database

#### Usage

```r
gbif_local(
dir = gbif_parquet_dir(version = gbif_version(local = TRUE)),
tblname = "gbif",
backend = c("arrow", "duckdb"),
safe = TRUE
)
```

#### Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dir</td>
<td>location of downloaded GBIF parquet files</td>
</tr>
<tr>
<td>tblname</td>
<td>name for the database table</td>
</tr>
<tr>
<td>backend</td>
<td>choose duckdb or arrow.</td>
</tr>
<tr>
<td>safe</td>
<td>logical. Should we exclude columns mediatype and issue? (default TRUE). varchar datatype on these columns substantially slows down queries.</td>
</tr>
</tbody>
</table>
gbif_remote

Details

A summary of this GBIF data, along with column meanings can be found at https://github.com/gbif/occurrence/blob/master/aws-public-data.md

Value

a remote tibble tbl_sql class object

Examples

```r
gbif <- gbif_local(gbif_example_data())
```

Description

Connect to GBIF remote directly. Can be much faster than downloading for one-off use or when using the package from a server in the same region as the data. See Details.

Usage

```r
gbif_remote(
  version = gbif_version(),
  bucket = gbif_default_bucket(),
  safe = TRUE,
  unset_aws = getOption("gbif_unset_aws", TRUE),
  endpoint_override = Sys.getenv("AWS_S3_ENDPOINT", "s3.amazonaws.com"),
  backend = c("arrow", "duckdb"),
  ...
)
```

Arguments

- **version**: GBIF snapshot date
- **bucket**: GBIF bucket name (including region). A default can also be set using the option `gbif_default_bucket`, see options.
- **safe**: logical, default TRUE. Should we exclude columns `mediatype` and `issue`? varchar datatype on these columns substantially slows downs queries.
- **unset_aws**: Unset AWS credentials? GBIF is provided in a public bucket, so credentials are not needed, but having a AWS_ACCESS_KEY_ID or other AWS environmental variables set can cause the connection to fail. By default, this will unset any set environmental variables for the duration of the R session. This behavior can also be turned off globally by setting the option `gbif_unset_aws` to FALSE (e.g. to use an alternative network endpoint)
gbif_version

endpoint_override
  optional parameter to \texttt{arrow::s3\_bucket()}
backend
  duckdb or arrow
  \ldots
  additional parameters passed to the \texttt{arrow::s3\_bucket()}

Details

Query performance is dramatically improved in queries that return only a subset of columns. Consider using explicit \texttt{select()} commands to return only the columns you need.

A summary of this GBIF data, along with column meanings can be found at \url{https://github.com/gbif/occurrence/blob/master/aws-public-data.md}

Value

a remote tibble \texttt{tbl\_sql} class object.

Examples

\begin{verbatim}
gbif <- gbif\_remote()
gbif()
\end{verbatim}

\begin{verbatim}
\begin{tabular}{ll}
gbif\_version & \textit{Get the latest gbif version string} \\
\end{tabular}
\end{verbatim}

Description

Can also return latest locally downloaded version, or list all versions

Usage

\begin{verbatim}
gbif\_version(
  local = FALSE,
  dir = gbif\_dir(),
  bucket = gbif\_default\_bucket(),
  all = FALSE,
  \ldots
)
\end{verbatim}

Arguments

\begin{verbatim}
local Search only local versions? logical, default FALSE.
dir local directory (\texttt{gbif\_dir()})
bucket Which remote bucket (region) should be checked
all show all versions? (logical, default FALSE)
\ldots additional arguments to \texttt{arrow::s3\_bucket}
\end{verbatim}
**gbif_version**

**Details**

A default version can be set using option `gbif_default_version`

**Value**

latest available gbif version, string

**Examples**

```r
## Latest local version available:
gbif_version(local=TRUE)
## default version
options(gbif_default_version="2021-01-01")
gbif_version()

## Latest online version available:
gbif_version()
## All online versions:
gbif_version(all=TRUE)
```
Index

arrow::s3_bucket, 6
arrow::s3_bucket(), 6

gbif_dir, 2
gbif_dir(), 3, 6
gbif_download, 3
gbif_example_data, 4
gbif_local, 4
gbif_remote, 5
gbif_version, 6

options, 5

tools::R_user_dir(), 2