Package ‘gbifdb’

May 21, 2022

Version 0.1.2
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 (>= 6.0.1), duckdb (>= 0.2.9), DBI, dplyr
Suggests spelling, dbplyr, testthat (>= 3.0.0), covr, knitr, rmarkdown, aws.s3
URL https://docs.ropensci.org/gbifdb/, https://github.com/ropensci/gbifdb
BugReports https://github.com/ropensci/gbifdb
Language en-US
RoxygenNote 7.1.2
Config/testthat/edition 3
gbif_conn

A [DBI]-style database connection to GBIF data

Description

Returns a database connection to the local GBIF parquet file.

Usage

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

Arguments

- `dir`: the directory containing all parquet files to be read
- `tblname`: name of the table to be created in the duckdb VIEW
- `backend`: Use arrow or duckdb as backend connection?

Value

a DBIconnection object
Examples

```r
gbif.parquet <- gbif_example_data()
con <- gbif_conn(gbif.parquet)
```

---

**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**

```r
gbif_dir()
```

**Value**

path to the gbif home directory directory

**Examples**

```r
gbif_dir()
```
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 \texttt{gbif_dir(\texttt{\})}.

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

Sync parquet files from GBIF public data catalog, \url{https://registry.opendata.aws/gbif/}.
Note that data can also be found on the Microsoft Cloud, \url{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 \url{https://github.com/gbif/occurrence/blob/master/aws-public-data.md}.

Value

\begin{verbatim}
logical indicating success or failure.
\end{verbatim}

Examples

\begin{verbatim}
\texttt{gbif_download(\texttt{\})}
\end{verbatim}

\begin{verbatim}
\texttt{gbif_example_data(\texttt{\})}
\end{verbatim}

Return a path to the directory containing GBIF example parquet data

Description

Return a path to the directory containing GBIF example parquet data

Usage

\begin{verbatim}
\texttt{gbif_example_data(\texttt{\})}
\end{verbatim}

Details

example 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.
gbif_local

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 = "duckdb",
  safe = TRUE
)
```

Arguments

- `dir` the directory containing all parquet files to be read
- `tblname` name of the table to be created in the duckdb VIEW
- `backend` Use arrow or duckdb as backend connection?
- `safe` logical, default TRUE. Should we exclude columns `mediatype` and `issue`? varchar datatype on these columns substantially slows down queries.

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(),
  to_duckdb = FALSE,
  safe = TRUE,
  unset_aws = getOption("gbif_unset_aws", TRUE),
  endpoint_override = Sys.getenv("AWS_S3_ENDPOINT", "s3.amazonaws.com"),
  ...
)
```

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`.
- `to_duckdb` Return a remote duckdb connection or arrow connection?
- `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)
- `endpoint_override` optional parameter to `arrow::s3_bucket()`
- `...` additional parameters passed to the `arrow::s3_bucket()`

Details

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

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
gbif_version

Value

a remote tibble tbl_sql class object (by default), or a arrow Dataset query if to_duckdb is FALSE. In either case, users should call [dplyr::collect] on the final result to force evaluation and bring the resulting data into memory in R.

Examples

```r
gbif <- gbif_remote()
gbif()
```

---

### gbif_version

*Get the latest gbif version string*

Description

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

Usage

```r
gbif_version(
  local = FALSE,
  dir = gbif_dir(),
  bucket = gbif_default_bucket(),
  all = FALSE,
  ...
)
```

Arguments

- **local**: Search only local versions? logical, default FALSE.
- **dir**: local directory (`gbif_dir()`)
- **bucket**: Which remote bucket (region) should be checked
- **all**: show all versions? (logical, default FALSE)
- **...**: additional arguments to `arrow::s3_bucket`

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, 7
arrow::s3_bucket(), 6

gbif_conn, 2
gbif_dir, 3
gbif_dir(), 4, 7
gbif_download, 3
gbif_example_data, 4
gbif_local, 5
gbif_remote, 6
gbif_version, 7

options, 6

tools::R_user_dir(), 3