

# Package ‘mlr3db’

October 13, 2022

**Title** Data Base Backend for 'mlr3'

**Version** 0.5.0

**Description** Extends the 'mlr3' package with a backend to transparently work with databases such as 'SQLite', 'DuckDB', 'MySQL', 'MariaDB', or 'PostgreSQL'. The package provides two additional backends: 'DataBackendDplyr' relies on the abstraction of package 'dplyr' to interact with most DBMS. 'DataBackendDuckDB' operates on 'DuckDB' data bases and also on Apache Parquet files.

**License** LGPL-3

**URL** <https://mlr3db.mlr-org.com>, <https://github.com/mlr-org/mlr3db>

**BugReports** <https://github.com/mlr-org/mlr3db/issues>

**Depends** mlr3 (>= 0.13.0), R (>= 3.1.0)

**Imports** R6, backports, checkmate, data.table, mlr3misc (>= 0.10.0)

**Suggests** DBI, RSQLite, dbplyr, dplyr, duckdb (>= 0.4.0), future, future.apply, future.callr, lgr, testthat (>= 3.0.0), tibble

**Encoding** UTF-8

**Config/testthat/edition** 3

**RoxygenNote** 7.2.1

**NeedsCompilation** no

**Author** Michel Lang [cre, aut] (<<https://orcid.org/0000-0001-9754-0393>>)

**Maintainer** Michel Lang <michellang@gmail.com>

**Repository** CRAN

**Date/Publication** 2022-08-08 10:10:02 UTC

## R topics documented:

mlr3db-package . . . . .	2
as_duckdb_backend . . . . .	2
as_sqlite_backend . . . . .	3
DataBackendDplyr . . . . .	4
DataBackendDuckDB . . . . .	8

---

mlr3db-package	<i>mlr3db: Data Base Backend for 'mlr3'</i>
----------------	---

---

### Description

Extends the 'mlr3' package with a backend to transparently work with databases such as 'SQLite', 'DuckDB', 'MySQL', 'MariaDB', or 'PostgreSQL'. The package provides two additional backends: 'DataBackendDplyr' relies on the abstraction of package 'dbplyr' to interact with most DBMS. 'DataBackendDuckDB' operates on DuckDB data bases and also on Apache Parquet files.

### Options

- `mlr3db.sqlite_dir`: Default directory to store SQLite databases constructed with `as_sqlite_backend()`.
- `mlr3db.duckdb_dir`: Default directory to store DuckDB databases constructed with `as_duckdb_backend()`.

### Author(s)

**Maintainer:** Michel Lang <michellang@gmail.com> ([ORCID](#))

### See Also

Useful links:

- <https://mlr3db.mlr-org.com>
- <https://github.com/mlr-org/mlr3db>
- Report bugs at <https://github.com/mlr-org/mlr3db/issues>

---

as_duckdb_backend	<i>Convert to DuckDB Backend</i>
-------------------	----------------------------------

---

### Description

Converts to a [DataBackendDuckDB](#) using the **duckdb** database, depending on the input type:

- `data.frame`: Creates a new [DataBackendDataTable](#) first using `as_data_backend()`, then proceeds with the conversion from [DataBackendDataTable](#) to [DataBackendDuckDB](#).
- `mlr3::DataBackend`: Creates a new DuckDB data base in the specified path. The filename is determined by the hash of the [DataBackend](#). If the file already exists, a connection to the existing database is established and the existing files are reused.

The created backend automatically reconnects to the database if the connection was lost, e.g. because the object was serialized to the filesystem and restored in a different R session. The only requirement is that the path does not change and that the path is accessible on all workers.

**Usage**

```
as_duckdb_backend(data, path = getOption("mlr3db.duckdb_dir", ":temp:"), ...)
```

**Arguments**

data	(data.frame()   <a href="#">mlr3::DataBackend</a> ) See description.
path	(character(1)) Path for the DuckDB databases. Either a valid path to a directory which will be created if it not exists, or one of the special strings: <ul style="list-style-type: none"> <li>":temp:" (default): Temporary directory of the R session is used, see <a href="#">tempdir()</a>. Note that this directory will be removed during the shutdown of the R session. Also note that this usually does not work for parallelization on remote workers. Set to a custom path instead or use special string ":user:" instead.</li> <li>":user:": User cache directory as returned by <a href="#">R_user_dir()</a> is used.</li> </ul> The default for this argument can be configured via option "mlr3db.sqlite_dir" or "mlr3db.duckdb_dir", respectively. The database files will use the hash of the <a href="#">DataBackend</a> as filename with file extension ".duckdb" or ".sqlite". If the database already exists on the file system, the converters will just established a new read-only connection.
...	(any) Additional arguments, passed to <a href="#">DataBackendDuckDB</a> .

**Value**

[DataBackendDuckDB](#) or [Task](#).

---

as_sqlite_backend	<i>Convert to SQLite Backend</i>
-------------------	----------------------------------

---

**Description**

Converts to a [DataBackendDplyr](#) using a **RSQLite** database, depending on the input type:

- `data.frame`: Creates a new [DataBackendDataTable](#) first using [as\\_data\\_backend\(\)](#), then proceeds with the conversion from [DataBackendDataTable](#) to [DataBackendDplyr](#).
- `mlr3::DataBackend`: Creates a new SQLite data base in the specified path. The filename is determined by the hash of the [DataBackend](#). If the file already exists, a connection to the existing database is established and the existing files are reused.

The created backend automatically reconnects to the database if the connection was lost, e.g. because the object was serialized to the filesystem and restored in a different R session. The only requirement is that the path does not change and that the path is accessible on all workers.

**Usage**

```
as_sqlite_backend(data, path = getOption("mlr3db.sqlite_dir", ":temp:"), ...)
```

**Arguments**

data	(data.frame()   <a href="#">mlr3::DataBackend</a> ) See description.
path	(character(1)) Path for the DuckDB databases. Either a valid path to a directory which will be created if it not exists, or one of the special strings: <ul style="list-style-type: none"> <li>"temp:" (default): Temporary directory of the R session is used, see <a href="#">tempdir()</a>. Note that this directory will be removed during the shutdown of the R session. Also note that this usually does not work for parallelization on remote workers. Set to a custom path instead or use special string "user:" instead.</li> <li>"user:": User cache directory as returned by <a href="#">R_user_dir()</a> is used.</li> </ul> The default for this argument can be configured via option "mlr3db.sqlite_dir" or "mlr3db.duckdb_dir", respectively. The database files will use the hash of the <a href="#">DataBackend</a> as filename with file extension ".duckdb" or ".sqlite". If the database already exists on the file system, the converters will just established a new read-only connection.
...	(any) Additional arguments, passed to <a href="#">DataBackendDplyr</a> .

**Value**

[DataBackendDplyr](#) or [Task](#).

---

DataBackendDplyr	<i>DataBackend for dplyr/dbplyr</i>
------------------	-------------------------------------

---

**Description**

A [mlr3::DataBackend](#) using [dplyr::tbl\(\)](#) from packages [dplyr/dbplyr](#). This includes [tibbles](#) and abstract database connections interfaced by [dbplyr](#). The latter allows [mlr3::Tasks](#) to interface an out-of-memory database.

**Super class**

[mlr3::DataBackend](#) -> [DataBackendDplyr](#)

**Public fields**

levels (named list())	List (named with column names) of factor levels as <code>character()</code> . Used to auto-convert character columns to factor variables.
connector (function())	Function which is called to re-connect in case the connection became invalid.

**Active bindings**

- `rownames` (`integer()`)  
Returns vector of all distinct row identifiers, i.e. the contents of the primary key column.
- `colnames` (`character()`)  
Returns vector of all column names, including the primary key column.
- `nrow` (`integer(1)`)  
Number of rows (observations).
- `ncol` (`integer(1)`)  
Number of columns (variables), including the primary key column.
- `valid` (`logical(1)`)  
Returns NA if the data does not inherit from "tbl\_sql" (i.e., it is not a real SQL data base).  
Returns the result of `DBI::dbIsValid()` otherwise.

**Methods****Public methods:**

- `DataBackendDplyr$new()`
- `DataBackendDplyr$finalize()`
- `DataBackendDplyr$data()`
- `DataBackendDplyr$head()`
- `DataBackendDplyr$distinct()`
- `DataBackendDplyr$missings()`

**Method** `new()`: Creates a backend for a `dplyr::tbl()` object.

*Usage:*

```
DataBackendDplyr$new(
  data,
  primary_key,
  strings_as_factors = TRUE,
  connector = NULL
)
```

*Arguments:*

`data` (`dplyr::tbl()`)

The data object.

Instead of calling the constructor yourself, you can call `mlr3::as_data_backend()` on a `dplyr::tbl()`. Note that only objects of class "tbl\_lazy" will be converted to a `DataBackendDplyr` (this includes all connectors from **dbplyr**). Local "tbl" objects such as `tibbles` will be converted to a `DataBackendDataTable`.

`primary_key` (`character(1)`)

Name of the primary key column.

`strings_as_factors` (`logical(1) || character()`)

Either a character vector of column names to convert to factors, or a single logical flag: if FALSE, no column will be converted, if TRUE all string columns (except the primary key). For conversion, the backend is queried for distinct values of the respective columns on construction and their levels are stored in `$levels`.

connector (function())\cr If not NULL', a function which re-connects to the database in case the connection has become invalid. Database connections can become invalid due to timeouts or if the backend is serialized to the file system and then de-serialized again. This round trip is often performed for parallelization, e.g. to send the objects to remote workers. `DBI::dbIsValid()` is called to validate the connection. The function must return just the connection, not a `dplyr::tbl()` object! Note that this this function is serialized together with the backend, including possible sensitive information such as login credentials. These can be retrieved from the stored `mlr3::DataBackend/mlr3::Task`. To protect your credentials, it is recommended to use the **secret** package.

**Method finalize():** Finalizer which disconnects from the database. This is called during garbage collection of the instance.

*Usage:*

```
DataBackendDplyr$finalize()
```

*Returns:* `logical(1)`, the return value of `DBI::dbDisconnect()`.

**Method data():** Returns a slice of the data. Calls `dplyr::filter()` and `dplyr::select()` on the table and converts it to a `data.table::data.table()`.

The rows must be addressed as vector of primary key values, columns must be referred to via column names. Queries for rows with no matching row id and queries for columns with no matching column name are silently ignored. Rows are guaranteed to be returned in the same order as rows, columns may be returned in an arbitrary order. Duplicated row ids result in duplicated rows, duplicated column names lead to an exception.

*Usage:*

```
DataBackendDplyr$data(rows, cols, data_format = "data.table")
```

*Arguments:*

`rows` `integer()`

Row indices.

`cols` `character()`

Column names.

`data_format` (`character(1)`)

Desired data format, e.g. "data.table" or "Matrix".

**Method head():** Retrieve the first n rows.

*Usage:*

```
DataBackendDplyr$head(n = 6L)
```

*Arguments:*

`n` (`integer(1)`)

Number of rows.

*Returns:* `data.table::data.table()` of the first n rows.

**Method distinct():** Returns a named list of vectors of distinct values for each column specified. If `na_rm` is TRUE, missing values are removed from the returned vectors of distinct values. Non-existing rows and columns are silently ignored.

*Usage:*

```
DataBackendDplyr$distinct(rows, cols, na_rm = TRUE)
```

*Arguments:*

rows integer()

Row indices.

cols character()

Column names.

na\_rm logical(1)

Whether to remove NAs or not.

*Returns:* Named list() of distinct values.

**Method** `missings()`: Returns the number of missing values per column in the specified slice of data. Non-existing rows and columns are silently ignored.

*Usage:*

```
DataBackendDplyr$missings(rows, cols)
```

*Arguments:*

rows integer()

Row indices.

cols character()

Column names.

*Returns:* Total of missing values per column (named `numeric()`).

## Examples

```
# Backend using a in-memory tibble
data = tibble::as_tibble(iris)
data$Sepal.Length[1:30] = NA
data$row_id = 1:150
b = DataBackendDplyr$new(data, primary_key = "row_id")

# Object supports all accessors of DataBackend
print(b)
b$row
b$col
b$colnames
b$data(rows = 100:101, cols = "Species")
b$distinct(b$rownames, "Species")

# Classification task using this backend
task = mlr3::TaskClassif$new(id = "iris_tibble", backend = b, target = "Species")
print(task)
task$head()

# Create a temporary SQLite database
con = DBI::dbConnect(RSQLite::SQLite(), ":memory:")
dplyr::copy_to(con, data)
tbl = dplyr::tbl(con, "data")

# Define a backend on a subset of the database
```

```
tbl = dplyr::select_at(tbl, setdiff(colnames(tbl), "Sepal.Width")) # do not use column "Sepal.Width"
tbl = dplyr::filter(tbl, row_id %in% 1:120) # Use only first 120 rows
b = DataBackendDplyr$new(tbl, primary_key = "row_id")
print(b)

# Query distinct values
b$distinct(b$rownames, "Species")

# Query number of missing values
b$missings(b$rownames, b$colnames)

# Note that SQLite does not support factors, column Species has been converted to character
lapply(b$head(), class)

# Cleanup
rm(tbl)
DBI::dbDisconnect(con)
```

---

DataBackendDuckDB

*DataBackend for DuckDB*


---

## Description

A `mlr3::DataBackend` for **duckdb**. Can be easily constructed with `as_duckdb_backend()`.

## Super class

`mlr3::DataBackend` -> `DataBackendDuckDB`

## Public fields

`levels` (named `list()`)

List (named with column names) of factor levels as `character()`. Used to auto-convert character columns to factor variables.

`connector` (`function()`)

Function which is called to re-connect in case the connection became invalid.

`table` (`character(1)`)

Data base table or view to operate on.

## Active bindings

`table_info` (`data.frame()`)

Data frame as returned by `pragma table_info()`.

`rownames` (`integer()`)

Returns vector of all distinct row identifiers, i.e. the contents of the primary key column.

`colnames` (`character()`)

Returns vector of all column names, including the primary key column.



`nrow` (integer(1))  
Number of rows (observations).

`ncol` (integer(1))  
Number of columns (variables), including the primary key column.

`valid` (logical(1))  
Returns NA if the data does not inherit from "tbl\_sql" (i.e., it is not a real SQL data base).  
Returns the result of `DBI::dbIsValid()` otherwise.

## Methods

### Public methods:

- `DataBackendDuckDB$new()`
- `DataBackendDuckDB$finalize()`
- `DataBackendDuckDB$data()`
- `DataBackendDuckDB$head()`
- `DataBackendDuckDB$distinct()`
- `DataBackendDuckDB$missings()`

**Method** `new()`: Creates a backend for a `duckdb::duckdb()` database.

#### Usage:

```
DataBackendDuckDB$new(
  data,
  table,
  primary_key,
  strings_as_factors = TRUE,
  connector = NULL
)
```

#### Arguments:

`data` (connection)

A connection created with `DBI::dbConnect()`. If constructed manually (and not via the helper function `as_duckdb_backend()`), make sure that there exists an (unique) index for the key column.

`table` (character(1))

Table or view to operate on.

`primary_key` (character(1))

Name of the primary key column.

`strings_as_factors` (logical(1) || character())

Either a character vector of column names to convert to factors, or a single logical flag: if FALSE, no column will be converted, if TRUE all string columns (except the primary key). For conversion, the backend is queried for distinct values of the respective columns on construction and their levels are stored in `$levels`.

`connector` (function())\cr If not NULL, a function which re-connects to the database in case the connection has become invalid. Database connections can become invalid due to timeouts or if the backend is serialized to the file system and then de-serialized again. This round trip is often performed for parallelization, e.g. to send the objects to remote workers. `DBI::dbIsValid()` is called to validate the connection. The function must return just the

connection, not a `dplyr::tbl()` object! Note that this function is serialized together with the backend, including possible sensitive information such as login credentials. These can be retrieved from the stored `mlr3::DataBackend/mlr3::Task`. To protect your credentials, it is recommended to use the **secret** package.

**Method** `finalize()`: Finalizer which disconnects from the database. This is called during garbage collection of the instance.

*Usage:*

```
DataBackendDuckDB$finalize()
```

*Returns:* `logical(1)`, the return value of `DBI::dbDisconnect()`.

**Method** `data()`: Returns a slice of the data.

The rows must be addressed as vector of primary key values, columns must be referred to via column names. Queries for rows with no matching row id and queries for columns with no matching column name are silently ignored. Rows are guaranteed to be returned in the same order as rows, columns may be returned in an arbitrary order. Duplicated row ids result in duplicated rows, duplicated column names lead to an exception.

*Usage:*

```
DataBackendDuckDB$data(rows, cols, data_format = "data.table")
```

*Arguments:*

`rows` `integer()`

Row indices.

`cols` `character()`

Column names.

`data_format` (`character(1)`)

Desired data format, e.g. "data.table" or "Matrix".

**Method** `head()`: Retrieve the first n rows.

*Usage:*

```
DataBackendDuckDB$head(n = 6L)
```

*Arguments:*

`n` (`integer(1)`)

Number of rows.

*Returns:* `data.table::data.table()` of the first n rows.

**Method** `distinct()`: Returns a named list of vectors of distinct values for each column specified. If `na_rm` is `TRUE`, missing values are removed from the returned vectors of distinct values. Non-existing rows and columns are silently ignored.

*Usage:*

```
DataBackendDuckDB$distinct(rows, cols, na_rm = TRUE)
```

*Arguments:*

`rows` `integer()`

Row indices.

`cols` `character()`

Column names.

na\_rm logical(1)

Whether to remove NAs or not.

*Returns:* Named list() of distinct values.

**Method** missings(): Returns the number of missing values per column in the specified slice of data. Non-existing rows and columns are silently ignored.

*Usage:*

```
DataBackendDuckDB$missings(rows, cols)
```

*Arguments:*

rows integer()

Row indices.

cols character()

Column names.

*Returns:* Total of missing values per column (named numeric()).

#### See Also

<https://duckdb.org/>

# Index

`as_data_backend()`, 2, 3  
`as_duckdb_backend`, 2  
`as_duckdb_backend()`, 2, 8, 9  
`as_sqlite_backend`, 3  
`as_sqlite_backend()`, 2

`data.table::data.table()`, 6, 10  
`DataBackend`, 2–4  
`DataBackendDataTable`, 2, 3, 5  
`DataBackendDplyr`, 3, 4, 4, 5  
`DataBackendDuckDB`, 2, 3, 8  
`DBI::dbConnect()`, 9  
`DBI::dbDisconnect()`, 6, 10  
`DBI::dbIsValid()`, 5, 6, 9  
`dplyr::filter()`, 6  
`dplyr::select()`, 6  
`dplyr::tbl()`, 4–6, 10  
`duckdb::duckdb()`, 9

`mlr3::as_data_backend()`, 5  
`mlr3::DataBackend`, 2–4, 6, 8, 10  
`mlr3::Task`, 4, 6, 10  
`mlr3db` (`mlr3db-package`), 2  
`mlr3db-package`, 2

`R_user_dir()`, 3, 4

`Task`, 3, 4  
`tempdir()`, 3, 4  
`tibbles`, 4, 5