

# Package ‘tfio’

October 14, 2022

**Type** Package

**Title** Interface to 'TensorFlow IO'

**Version** 0.4.1

**Description**

Interface to 'TensorFlow IO', Datasets and filesystem extensions maintained by `TensorFlow SIG-IO` <<https://github.com/tensorflow/community/blob/master/sigs/io/CHARTER.md>>.

**License** Apache License 2.0

**URL** <https://github.com/tensorflow/io>

**BugReports** <https://github.com/tensorflow/io/issues>

**SystemRequirements** TensorFlow >= 1.13.0 (<https://www.tensorflow.org/>)  
and TensorFlow IO >= 0.4.0 (<https://github.com/tensorflow/io>)

**Encoding** UTF-8

**LazyData** true

**Depends** R (>= 3.1)

**Imports** reticulate (>= 1.10), tensorflow (>= 1.9), tfdatasets (>= 1.9), forge, magrittr

**RoxygenNote** 7.0.2

**Suggests** testthat, knitr

**VignetteBuilder** knitr

**NeedsCompilation** no

**Author** TensorFlow IO Contributors [aut, cph] (Full list of contributors can be found at <<https://github.com/tensorflow/io/graphs/contributors>>),  
Yuan Tang [aut, cre] (<<https://orcid.org/0000-0001-5243-233X>>),  
TensorFlow Authors [cph],  
Ant Financial [cph],  
RStudio [cph]

**Maintainer** Yuan Tang <[terrytangyuan@gmail.com](mailto:terrytangyuan@gmail.com)>

**Repository** CRAN

**Date/Publication** 2019-12-19 16:50:02 UTC

**R topics documented:**

arrow_feather_dataset . . . . .	2
arrow_stream_dataset . . . . .	3
from_schema . . . . .	4
from_schema.arrow_feather_dataset . . . . .	4
from_schema.arrow_stream_dataset . . . . .	5
ignite_dataset . . . . .	6
kafka_dataset . . . . .	7
kinesis_dataset . . . . .	8
lmbd_dataset . . . . .	9
make_libsvm_dataset . . . . .	9
mnist_image_dataset . . . . .	10
mnist_label_dataset . . . . .	11
parquet_dataset . . . . .	11
pubsub_dataset . . . . .	12
sequence_file_dataset . . . . .	13
tfo . . . . .	13
tiff_dataset . . . . .	14
video_dataset . . . . .	14
webp_dataset . . . . .	15

<b>Index</b>	<b>17</b>
--------------	-----------

---

arrow\_feather\_dataset *Creates a ArrowFeatherDataset.*

---

**Description**

An Arrow Dataset for reading record batches from Arrow feather files. Feather is a light-weight columnar format ideal for simple writing of Pandas DataFrames.

**Usage**

```
arrow_feather_dataset(filenamees, columns, output_types, output_shapes = NULL)
```

**Arguments**

filenamees	A <code>tf.string</code> tensor, list or scalar containing files in Arrow Feather format.
columns	A list of column indices to be used in the Dataset.
output_types	Tensor dtypes of the output tensors.
output_shapes	TensorShapes of the output tensors or NULL to infer partial.

**Examples**

```
## Not run:
dataset <- arrow_feather_dataset(
  list('/path/to/a.feather', '/path/to/b.feather'),
  columns = reticulate::tuple(0L, 1L),
  output_types = reticulate::tuple(tf$int32, tf$float32),
  output_shapes = reticulate::tuple(list(), list()) %>%
  dataset_repeat(1)

sess <- tf$Session()
iterator <- make_iterator_one_shot(dataset)
next_batch <- iterator_get_next(iterator)

until_out_of_range({
  batch <- sess$run(next_batch)
  print(batch)
})

## End(Not run)
```

---

arrow\_stream\_dataset *Creates a ArrowStreamDataset.*

---

**Description**

An Arrow Dataset for reading record batches from an input stream. Currently supported input streams are a socket client or stdin.

**Usage**

```
arrow_stream_dataset(host, columns, output_types, output_shapes = NULL)
```

**Arguments**

host	A <code>tf.string</code> tensor or string defining the input stream. For a socket client, use " <code>&lt;HOST_IP&gt;:&lt;PORT&gt;</code> ", for stdin use "STDIN".
columns	A list of column indices to be used in the Dataset.
output_types	Tensor dtypes of the output tensors.
output_shapes	TensorShapes of the output tensors or NULL to infer partial.

**Examples**

```
## Not run:
dataset <- arrow_stream_dataset(
  host,
  columns = reticulate::tuple(0L, 1L),
  output_types = reticulate::tuple(tf$int32, tf$float32),
```

```

    output_shapes = reticulate::tuple(list(), list()) %>%
    dataset_repeat(1)

    sess <- tf$Session()
    iterator <- make_iterator_one_shot(dataset)
    next_batch <- iterator_get_next(iterator)

    until_out_of_range({
      batch <- sess$run(next_batch)
      print(batch)
    })

    ## End(Not run)

```

---

from\_schema

*Create an Arrow Dataset from the given Arrow schema.*


---

### Description

Infer output types and shapes from the given Arrow schema and create an Arrow Dataset.

### Usage

```
from_schema(object, ...)
```

### Arguments

object	An R object.
...	Optional arguments passed on to implementing methods.

---

from\_schema.arrow\_feather\_dataset

*Create an Arrow Dataset for reading record batches from Arrow feather files, inferring output types and shapes from the given Arrow schema.*


---

### Description

Create an Arrow Dataset for reading record batches from Arrow feather files, inferring output types and shapes from the given Arrow schema.

### Usage

```
## S3 method for class 'arrow_feather_dataset'
from_schema(object, schema, columns = NULL, host = NULL, filenames = NULL, ...)
```

**Arguments**

object	An R object.
schema	Arrow schema defining the record batch data in the stream.
columns	A list of column indices to be used in the Dataset.
host	Not used.
filenames	A <code>tf.string</code> tensor, list or scalar containing files in Arrow Feather format.
...	Optional arguments passed on to implementing methods.

---

from\_schema.arrow\_stream\_dataset

*Create an Arrow Dataset from an input stream, inferring output types and shapes from the given Arrow schema.*

---

**Description**

Create an Arrow Dataset from an input stream, inferring output types and shapes from the given Arrow schema.

**Usage**

```
## S3 method for class 'arrow_stream_dataset'
from_schema(object, schema, columns = NULL, host = NULL, filenames = NULL, ...)
```

**Arguments**

object	An R object.
schema	Arrow schema defining the record batch data in the stream.
columns	A list of column indices to be used in the Dataset.
host	A <code>tf.string</code> tensor or string defining the input stream. For a socket client, use " <code>&lt;HOST_IP&gt;:&lt;PORT&gt;</code> ", for stdin use "STDIN".
filenames	Not used.
...	Optional arguments passed on to implementing methods.

---

ignite_dataset	<i>Create a IgniteDataset.</i>
----------------	--------------------------------

---

### Description

Apache Ignite is a memory-centric distributed database, caching, and processing platform for transactional, analytical, and streaming workloads, delivering in-memory speeds at petabyte scale. This contrib package contains an integration between Apache Ignite and TensorFlow. The integration is based on tf.data from TensorFlow side and Binary Client Protocol from Apache Ignite side. It allows to use Apache Ignite as a datasource for neural network training, inference and all other computations supported by TensorFlow. Ignite Dataset is based on Apache Ignite Binary Client Protocol.

### Usage

```
ignite_dataset(
  cache_name,
  host = "localhost",
  port = 10800,
  local = FALSE,
  part = -1,
  page_size = 100,
  username = NULL,
  password = NULL,
  certfile = NULL,
  keyfile = NULL,
  cert_password = NULL
)
```

### Arguments

cache_name	Cache name to be used as datasource.
host	Apache Ignite Thin Client host to be connected.
port	Apache Ignite Thin Client port to be connected.
local	Local flag that defines to query only local data.
part	Number of partitions to be queried.
page_size	Apache Ignite Thin Client page size.
username	Apache Ignite Thin Client authentication username.
password	Apache Ignite Thin Client authentication password.
certfile	File in PEM format containing the certificate as well as any number of CA certificates needed to establish the certificate's authenticity.
keyfile	File containing the private key (otherwise the private key will be taken from certfile as well).
cert_password	Password to be used if the private key is encrypted and a password is necessary.

**Examples**

```
## Not run:
dataset <- ignite_dataset(
  cache_name = "SQL_PUBLIC_TEST_CACHE", port = 10800) %>%
  dataset_repeat(1)

sess <- tf$Session()
iterator <- make_iterator_one_shot(dataset)
next_batch <- iterator_get_next(iterator)

until_out_of_range({
  batch <- sess$run(next_batch)
  print(batch)
})

## End(Not run)
```

---

kafka_dataset	<i>Creates a KafkaDataset.</i>
---------------	--------------------------------

---

**Description**

Creates a KafkaDataset.

**Usage**

```
kafka_dataset(
  topics,
  servers = "localhost",
  group = "",
  eof = FALSE,
  timeout = 1000
)
```

**Arguments**

topics	A <code>tf.string</code> tensor containing one or more subscriptions, in the format of <code>[topic:partition:offset:length]</code> , by default length is -1 for unlimited.
servers	A list of bootstrap servers.
group	The consumer group id.
eof	If True, the kafka reader will stop on EOF.
timeout	The timeout value for the Kafka Consumer to wait (in millisecond).

**Examples**

```
## Not run:
dataset <- kafka_dataset(
  topics = list("test:0:0:4"), group = "test", eof = TRUE) %>%
  dataset_repeat(1)

sess <- tf$Session()
iterator <- make_iterator_one_shot(dataset)
next_batch <- iterator_get_next(iterator)

until_out_of_range({
  batch <- sess$run(next_batch)
  print(batch)
})

## End(Not run)
```

---

kinesis_dataset	<i>Creates a KinesisDataset.</i>
-----------------	----------------------------------

---

**Description**

Kinesis is a managed service provided by AWS for data streaming. This dataset reads messages from Kinesis with each message presented as a `tf.string`.

**Usage**

```
kinesis_dataset(stream, shard = "", read_indefinitely = TRUE, interval = 1e+05)
```

**Arguments**

<code>stream</code>	A <code>tf.string</code> tensor containing the name of the stream.
<code>shard</code>	A <code>tf.string</code> tensor containing the id of the shard.
<code>read_indefinitely</code>	If <code>True</code> , the Kinesis dataset will keep retry again on EOF after the interval period. If <code>False</code> , then the dataset will stop on EOF. The default value is <code>True</code> .
<code>interval</code>	The interval for the Kinesis Client to wait before it tries to get records again (in millisecond).



---

lmbd_dataset	<i>Create a LMDBDataset.</i>
--------------	------------------------------

---

### Description

This function allows a user to read data from a LMDB file. A lmbd file consists of (key value) pairs sequentially.

### Usage

```
lmbd_dataset(filenamees)
```

### Arguments

filenamees      A `tf.string` tensor containing one or more filenames.

### Examples

```
## Not run:
dataset <- sequence_file_dataset("testdata/data.mdb") %>%
  dataset_repeat(1)

sess <- tf$Session()
iterator <- make_iterator_one_shot(dataset)
next_batch <- iterator_get_next(iterator)

until_out_of_range({
  batch <- sess$run(next_batch)
  print(batch)
})

## End(Not run)
```

---

make_libsvm_dataset	<i>Create a Dataset from LibSVM files.</i>
---------------------	--

---

### Description

Create a Dataset from LibSVM files.

**Usage**

```

make_libsvm_dataset(
    file_names,
    num_features,
    dtype = NULL,
    label_dtype = NULL,
    batch_size = 1,
    compression_type = "",
    buffer_size = NULL,
    num_parallel_parser_calls = NULL,
    drop_final_batch = FALSE,
    prefetch_buffer_size = 0
)

```

**Arguments**

<code>file_names</code>	A <code>tf.string</code> tensor containing one or more filenames.
<code>num_features</code>	The number of features.
<code>dtype</code>	The type of the output feature tensor. Default to <code>tf.float32</code> .
<code>label_dtype</code>	The type of the output label tensor. Default to <code>tf.int64</code> .
<code>batch_size</code>	An integer representing the number of records to combine in a single batch, default 1.
<code>compression_type</code>	A <code>tf.string</code> scalar evaluating to one of "" (no compression), "ZLIB", or "GZIP".
<code>buffer_size</code>	A <code>tf.int64</code> scalar denoting the number of bytes to buffer. A value of 0 results in the default buffering values chosen based on the compression type.
<code>num_parallel_parser_calls</code>	Number of parallel records to parse in parallel. Defaults to an automatic selection.
<code>drop_final_batch</code>	Whether the last batch should be dropped in case its size is smaller than <code>batch_size</code> ; the default behavior is not to drop the smaller batch.
<code>prefetch_buffer_size</code>	An integer specifying the number of feature batches to prefetch for performance improvement. Defaults to auto-tune. Set to 0 to disable prefetching.

---

`mnist_image_dataset`    *Creates a MNISTImageDataset.*

---

**Description**

This creates a dataset for MNIST images.

**Usage**

```
mnist_image_dataset(filenames, compression_type = NULL)
```

**Arguments**

`filenames`      A `tf.string` tensor containing one or more filenames.  
`compression_type`      A `tf.string` scalar evaluating to one of "" (no compression), "ZLIB", or "GZIP".

---

`mnist_label_dataset`      *Creates a MNISTLabelDataset.*

---

**Description**

This creates a dataset for MNIST labels.

**Usage**

```
mnist_label_dataset(filenames, compression_type = NULL)
```

**Arguments**

`filenames`      A `tf.string` tensor containing one or more filenames.  
`compression_type`      A `tf.string` scalar evaluating to one of "" (no compression), "ZLIB", or "GZIP".

---

`parquet_dataset`      *Create a ParquetDataset.*

---

**Description**

This allows a user to read data from a parquet file.

**Usage**

```
parquet_dataset(filenames, columns, output_types)
```

**Arguments**

`filenames`      A 0-D or 1-D `tf.string` tensor containing one or more filenames.  
`columns`      A 0-D or 1-D `tf.int32` tensor containing the columns to extract.  
`output_types`      A tuple of `tf.DType` objects representing the types of the columns returned.

**Examples**

```

## Not run:
dtypes <- tf$python$framework$dtypes
output_types <- reticulate::tuple(
  dtypes$bool, dtypes$int32, dtypes$int64, dtypes$float32, dtypes$float64)
dataset <- parquet_dataset(
  filenames = list("testdata/parquet_cpp_example.parquet"),
  columns = list(0, 1, 2, 4, 5),
  output_types = output_types) %>%
  dataset_repeat(2)

sess <- tf$Session()
iterator <- make_iterator_one_shot(dataset)
next_batch <- iterator_get_next(iterator)

until_out_of_range({
  batch <- sess$run(next_batch)
  print(batch)
})

## End(Not run)

```

---

pubsub_dataset	<i>Creates a PubSubDataset.</i>
----------------	---------------------------------

---

**Description**

This creates a dataset for consuming PubSub messages.

**Usage**

```
pubsub_dataset(subscriptions, server = NULL, eof = FALSE, timeout = 1000)
```

**Arguments**

subscriptions	A <code>tf.string</code> tensor containing one or more subscriptions.
server	The pubsub server.
eof	If True, the pubsub reader will stop on EOF.
timeout	The timeout value for the PubSub to wait (in millisecond).

---

sequence\_file\_dataset *Create a SequenceFileDataset.*

---

### Description

This function allows a user to read data from a hadoop sequence file. A sequence file consists of (key value) pairs sequentially. At the moment, `org.apache.hadoop.io.Text` is the only serialization type being supported, and there is no compression support.

### Usage

```
sequence_file_dataset(filenamees)
```

### Arguments

filenamees      A `tf.string` tensor containing one or more filenames.

### Examples

```
## Not run:
dataset <- sequence_file_dataset("testdata/string.seq") %>%
  dataset_repeat(1)

sess <- tf$Session()
iterator <- make_iterator_one_shot(dataset)
next_batch <- iterator_get_next(iterator)

until_out_of_range({
  batch <- sess$run(next_batch)
  print(batch)
})

## End(Not run)
```

---

tfio

*TensorFlow IO API for R*

---

### Description

This library provides an R interface to the **TensorFlow IO** API that provides datasets and filesystem extensions maintained by SIG-IO.

---

tiff_dataset	<i>Create a TIFFDataset.</i>
--------------	------------------------------

---

### Description

A TIFF Image File Dataset that reads the TIFF file.

### Usage

```
tiff_dataset(filenamees)
```

### Arguments

filenamees      A `tf.string` tensor containing one or more filenames.

### Examples

```
## Not run:
dataset <- tiff_dataset(
  filenames = list("testdata/small.tiff")) %>%
  dataset_repeat(1)

sess <- tf$Session()
iterator <- make_iterator_one_shot(dataset)
next_batch <- iterator_get_next(iterator)

until_out_of_range({
  batch <- sess$run(next_batch)
  print(batch)
})

## End(Not run)
```

---

video_dataset	<i>Create a VideoDataset that reads the video file.</i>
---------------	---

---

### Description

This allows a user to read data from a video file with ffmpeg. The output of VideoDataset is a sequence of (height, weight, 3) tensor in rgb24 format.

### Usage

```
video_dataset(filenamees)
```

**Arguments**

`filenames` A `tf.string` tensor containing one or more filenames.

**Examples**

```
## Not run:
dataset <- video_dataset(
  filenames = list("testdata/small.mp4")) %>%
  dataset_repeat(2)

sess <- tf$Session()
iterator <- make_iterator_one_shot(dataset)
next_batch <- iterator_get_next(iterator)

until_out_of_range({
  batch <- sess$run(next_batch)
  print(batch)
})

## End(Not run)
```

---

webp\_dataset

*Create a WebPDataset.*


---

**Description**

A WebP Image File Dataset that reads the WebP file.

**Usage**

```
webp_dataset(filenames)
```

**Arguments**

`filenames` A `tf.string` tensor containing one or more filenames.

**Examples**

```
## Not run:
dataset <- webp_dataset(
  filenames = list("testdata/sample.webp")) %>%
  dataset_repeat(1)

sess <- tf$Session()
iterator <- make_iterator_one_shot(dataset)
next_batch <- iterator_get_next(iterator)

until_out_of_range({
```

```
    batch <- sess$run(next_batch)
    print(batch)
  })
## End(Not run)
```



# Index

[arrow\\_feather\\_dataset](#), [2](#)  
[arrow\\_stream\\_dataset](#), [3](#)

[from\\_schema](#), [4](#)  
[from\\_schema.arrow\\_feather\\_dataset](#), [4](#)  
[from\\_schema.arrow\\_stream\\_dataset](#), [5](#)

[ignite\\_dataset](#), [6](#)

[kafka\\_dataset](#), [7](#)  
[kinesis\\_dataset](#), [8](#)

[lmbd\\_dataset](#), [9](#)

[make\\_libsvm\\_dataset](#), [9](#)  
[mnist\\_image\\_dataset](#), [10](#)  
[mnist\\_label\\_dataset](#), [11](#)

[parquet\\_dataset](#), [11](#)  
[pubsub\\_dataset](#), [12](#)

[sequence\\_file\\_dataset](#), [13](#)

[tfio](#), [13](#)  
[tiff\\_dataset](#), [14](#)

[video\\_dataset](#), [14](#)

[webp\\_dataset](#), [15](#)