

⁺This command is part of [StataNow](#).

[Description](#)
[Options](#)

[Quick start](#)
[Remarks and examples](#)

[Menu](#)
[Stored results](#)

[Syntax](#)
[Also see](#)

Description

`import parquet` reads into memory a Parquet (`.parquet`) file. Before importing data from a Parquet file, `import parquet` with the `describe` option can be used to list columns of a Parquet file and their data types.

Apache Parquet is a data file format that organizes data by columns, and it supports several compression methods for the data to achieve efficient storage. `import parquet` does not support files compressed via methods LZO and LZ4_RAW.

Quick start

Check the contents of Parquet file `mydata.parquet` before importing

```
import parquet using mydata, describe
```

Load data from `mydata.parquet`

```
import parquet using mydata
```

Load only columns `x1` and `x2`

```
import parquet x1 x2 using mydata
```

Load only the first 1,000 rows from `mydata.parquet`

```
import parquet using mydata, rowrange(1:1000)
```

Load only the last 1,000 rows from `mydata.parquet`

```
import parquet using mydata, rowrange(-1000:L)
```

Menu

File > Import > Parquet data (*.parquet)

Syntax

Load a Parquet file

```
import parquet [ using ] filename [ , options ]
```

Load subset of columns from a Parquet file

```
import parquet columnlist using filename [ , options ]
```

Describe contents of a Parquet file

```
import parquet [ using ] filename , describe
```

If *filename* is specified without an extension, .parquet is assumed. If *filename* contains embedded spaces, enclose it in double quotes. *filename* may not be specified as a URL.

columnlist is a list of column names in the Parquet file to be imported.

<i>options</i>	Description
<code>clear</code>	replace data in memory
<code>rowrange([<i>start</i>][<i>:end</i>])</code>	row range of data to load
<code>favormemory</code>	favor conserving memory over speed

collect is allowed with import parquet; see [U] 11.1.10 Prefix commands.

favormemory does not appear in the dialog box.

Options

clear specifies that it is okay to replace the data in memory, even though the current data have not been saved to disk.

rowrange([*start*][*:end*]) specifies a range of rows within the Parquet file to load. *start* and *end* are integer row numbers.

The following option is available with import parquet but is not shown in the dialog box:

favormemory specifies that import parquet favor conserving memory over speed.

Remarks and examples

Remarks are presented under the following headings:

- Introduction
- Mapping Parquet data types to Stata data types
 - Handling integer types
 - Handling float type
 - Handling decimal type
 - Handling string type
 - Handling binary types
 - Handling timestamp type
 - Handling date type
 - Handling dictionary type

Introduction

`import parquet` imports data from a Parquet file into Stata. Apache Parquet is a data file format that organizes data by columns, and it supports several compression methods for the data to achieve efficient storage.

To demonstrate how to import data from a Parquet file into Stata, we begin by copying the `auto` dataset in Parquet format to our current directory.

```
. copy https://www.stata.com/examples/auto.parquet .
```

We can now import the data from `auto.parquet` into Stata by typing the following:

```
. import parquet using auto, clear
(12 vars, 74 obs)
```

We could verify that our data loaded correctly by using [list](#) or [browse](#).

Mapping Parquet data types to Stata data types

`import parquet` reads a Parquet file into an Apache Arrow table and then converts the table to a Stata dataset. Most Parquet data types are supported, except for MAP and LIST. If a data type is not supported, the column will be ignored.

The data types displayed by `import parquet, describe` are the Apache Arrow data types. For example, `utf8` is displayed if a column is interpreted as a UTF-8 string.

Handling integer types

Parquet data type `BOOLEAN` is mapped to Stata type `byte`.

Parquet data types `INT8` and `UINT8` are mapped to Stata type `byte` if the maximum of the column is less than or equal to 100; otherwise, they are mapped to Stata type `int`.

Parquet data types `INT16` and `UINT16` are mapped to Stata type `int` if the maximum of the column is less than or equal to 32,740; otherwise, they are mapped to Stata type `long`.

Parquet data types `INT32` and `UINT32` are mapped to Stata type `long` if the maximum of the column is less than or equal to 2,147,483,620; otherwise, they are mapped to Stata type `double`.

Parquet data types `INT64` and `UINT64` are mapped to Stata type `long` if all numbers in the column are between $-2,147,483,647$ and $2,147,483,620$; otherwise, they are mapped to Stata type `str16` with the number's hex representation.

Note that Stata's `byte` storage type can be used to store values with a maximum of 100, `int` can store values with a maximum of 32,740, and `long` can store values with a maximum of 2,147,483,620. See [\[D\] Data types](#). Therefore, a column with a maximum value exceeding the `byte` limit is mapped to `int`, a column with a maximum value exceeding the `int` limit is mapped to `long`, and so on.

Handling float type

Parquet data type `FLOAT` is mapped to Stata type `float` if the maximum of the column is less than or equal to $1.70141173319 \times 10^{38}$; otherwise, it is mapped to Stata type `double`.

Parquet data type `HALFFLOAT` (or `FLOAT16`), implemented as `FIXED_LEN_BYTE_ARRAY`, is mapped to Stata type `float`.

Parquet data type `DOUBLE` is mapped to Stata type `double`.

Handling decimal type

Parquet data type DECIMAL is mapped to Stata type `str#`, where `#` is 2 plus the precision of the decimal type.

Handling string type

Parquet data type STRING, and equivalently Apache Arrow type UTF8, is mapped to Stata type `str#` if the maximum length of the column is less than or equal to 2,045; otherwise, it is mapped to Stata type `strL`. Any strings with length longer than 2,000,000,000 are ignored.

Handling binary types

Parquet data types BINARY and FIXED_SIZE_BINARY are mapped to Stata type `strL` because the columns with those data types may contain binary 0.

Handling timestamp type

Parquet data type TIMESTAMP is mapped to Stata type `double` with display format `%tc`.

Handling date type

Parquet data type DATE is mapped to Stata type `long` and display format `%td`.

Handling dictionary type

Columns with categorical data are mapped to Apache Arrow type DICTIONARY; these columns are mapped to a Stata numeric type with a value label if the number of items in the dictionary is less than or equal to 65,536 and the maximum length of the items is less than or equal to 32,000.

If the number of items in the dictionary is greater than 65,536 and the maximum length *mlen* of the items is less than 2,045, the column is mapped to a Stata type `str#`, where `#` is *mlen*. Otherwise, the column is mapped to Stata type `strL`.

Stored results

`import parquet` stores the following in `r()`:

Scalars

<code>r(N)</code>	number of observations
<code>r(k)</code>	number of variables

`import parquet, describe` stores the following in `r()`:

Scalars

<code>r(N_rows)</code>	number of rows
<code>r(n_columns)</code>	number of columns

Also see

[D] [import](#) — Overview of importing data into Stata

Stata, Stata Press, Mata, NetCourse, and NetCourseNow are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. StataNow is a trademark of StataCorp LLC. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2025 StataCorp LLC, College Station, TX, USA. All rights reserved.

For suggested citations, see the FAQ on [citing Stata documentation](#).

