

Importing NACC Data in Parquet File Format

Last edited: June 25, 2026

Parquet is a columnar storage format optimized for efficient storage and retrieval of large-scale data. It leverages advanced compression and encoding schemes to deliver high performance, particularly when working with complex and large datasets.

Benefits

- Well suited for storing very large datasets
- Extremely fast to filter within a column
- It is compressed and is smaller file size than csv

	CSV	Parquet
Storage layout	Row-based (text)	Columnar (binary)
File Size	Large (uncompressed)	Small (built-in compression)
Human readable	Yes* (open in Notepad/Excel)	No (requires specialized tools)
Schema/Types	No (everything is a string)	Yes (metadata stores data types)

* The UDS csv file is too large to be analyzed in Excel.

Running parquet files in R

1. Direct file access – loads into RAM and uses basic query capability (R/dplyr)

```
# Install the 'arrow' package
install.packages("arrow")

# Load the arrow library into the R session
library("arrow")

# Read the Parquet file into a data frame called 'uds'
uds <- read_parquet("investigator_naccXX.parquet")
```

2. The `open_dataset()` function from the [arrow R package](#) allows you to query Parquet files without loading them into memory immediately. This approach is ideal for datasets larger than your computer's RAM, as it only processes the specific data you need when you call

```
# Load a dataset from the specified file path into a dataset
object 'ds'
ds <- open_dataset("path/to/data")
```

```
# Query using familiar dplyr verbs
uds <- ds %>%
  filter(NACCVNUM == 1) %>%
  select(NACCID, NACCUDSD) %>% # Select variables to keep
  collect() # Only now is data actually read into memory
```

3. Using a database query engine (e.g. DuckDB) – query-on-demand and uses advanced SQL
 - a. [Large Data Work: Intro to parquet files in R](#)

Running parquet files in SAS 9.4

1. SAS 9.4 does not have a native, built-in Parquet engine.
2. [Working with Apache Parquet Files](#)

Running parquet files in SAS Viya

Using SAS Viya (CAS) - It loads and processes the data using in-memory distributed computing.

```
cas mySession sessopts=(caslib="casuser");

/* Define the path to your parquet files */
caslib my_parquet_lib datasource=(srctype="path")
path="/your/file/path/here/";

/* Load into memory */
proc casutil;
  load casdata=" investigator_naccXX.parquet"
  incaslib="uds_parquet_lib"
  casout="uds"
  outcaslib="uds_parquet_lib";
run;
```

Running parquet files in Stata

1. [Import data from parquet files.](#)
<https://www.stata.com/manuals/dimportparquet.pdf>

```
import parquet using investigator_naccXX.parquet
```

2. Parquet files can be imported using Python if an older version of Stata does not support parquet files.

```
python:
import pandas as pd
df = pd.read_parquet("path/to/data/investigator_naccXX.parquet",
columns=["NACCID", "NACCUDSD"])
df.to_stata("uds.dta", write_index=False)
end

use uds.dta, clear
```