UKCrawl: Code Notes and Snippets

The parquet format has several benefits over alternatives like CSV; parquet supports file compression, allowing for smaller file sizes, and data types are preserved, meaning the ‘all\_postcodes’ list column is parsed correctly.

Most major programming languages have libraries that can read parquet files. The following examples show how this data may be read.

In Python there are a number of libraries that can read parquet files, including pandas, polars, and duckdb:

import pandas as pd  
import polars as pl  
import duckdb  
  
df = pd.read\_parquet("ukcrawl-2019.parquet")  
df = pl.read\_parquet("ukcrawl-2019.parquet")  
df = duckdb.read\_parquet("ukcrawl-2019.parquet")

In R you can use the arrow library (https://arrow.apache.org/docs/r/):

library(arrow)  
  
# reads into a 'tibble'  
df <- read\_parquet("ukcrawl-2019.parquet")