

# Hands-On Analyzing Parquet Files, using "parquet-cli"

Esilv 2024

Exercise : Analyzing Parquet Files,  
using "parquet-cli"

# Analysing Parquet File, using "parquet-cli"

The screenshot shows the GitHub repository for `parquet-cli` within the `parquet-java` project. The browser address bar shows `github.com/apache/parquet-java/tree/master/parquet-cli`. The repository page includes a sidebar with a file tree, a commit history table, and a view of the `README.md` file.

**Files**

- master
- Go to file
- .github
- .mvn
- dev
- doc
- parquet-arrow
- parquet-avro
- parquet-benchmarks
- parquet-cli
  - src
  - README.md
  - pom.xml
- parquet-column
- parquet-common
- parquet-encoding
- parquet-format-structures
- parquet-generator
- parquet-hadoop-bundle
- parquet-hadoop

**parquet-java / parquet-cli**

Commit history:

Name	Last commit message	Last commit date
..		
src	<a href="#">GH-2976</a> : Parquet CLI compression commands should accept lowercase com...	4 months ago
README.md	<a href="#">GH-2952</a> : Add maven wrapper ( <a href="#">#2953</a> )	4 months ago
pom.xml	MINOR: bump version to 1.16.0-SNAPSHOT ( <a href="#">#3097</a> )	4 days ago

**README.md**

### Building

You can build this project using maven:

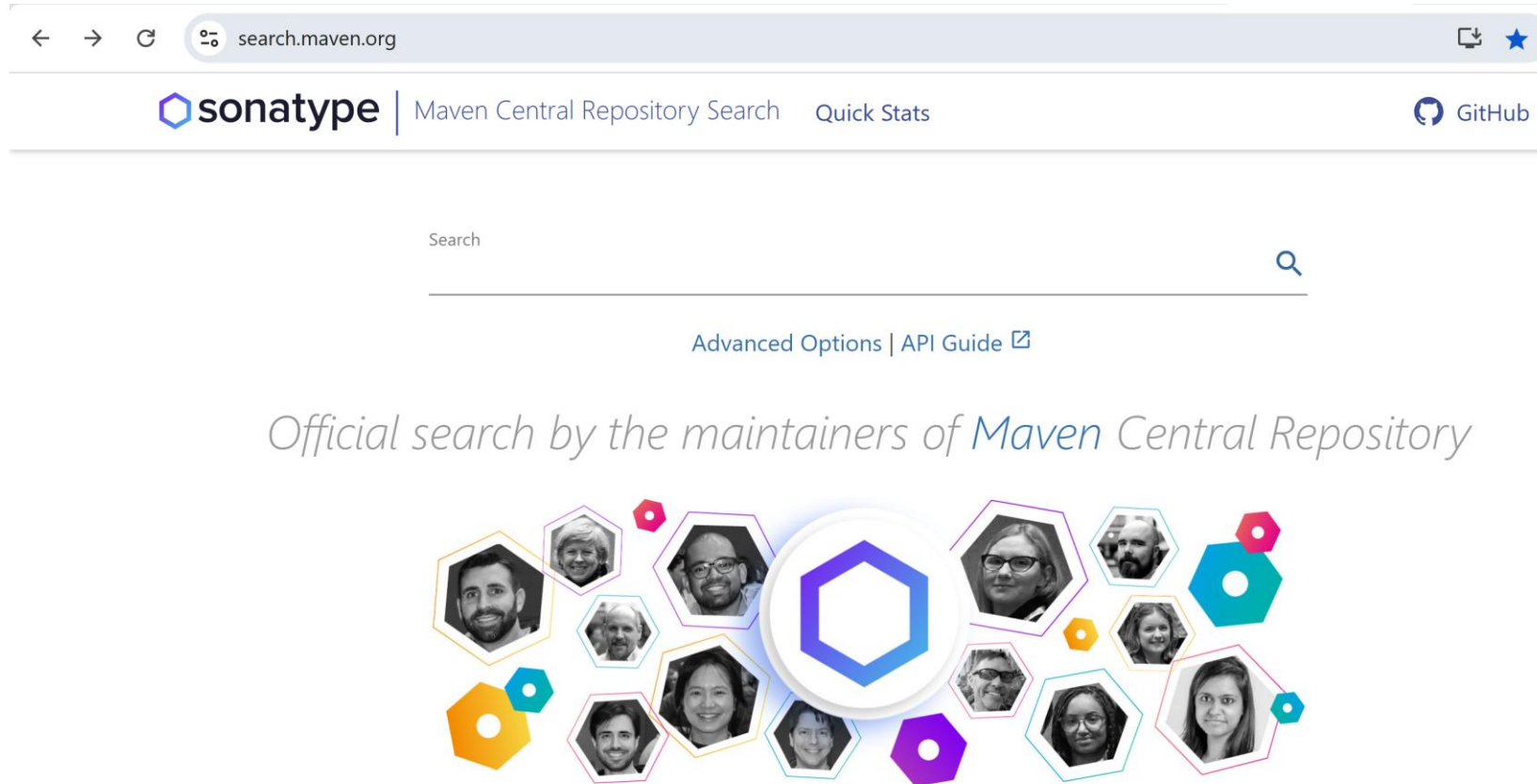
```
./mvnw clean install -DskipTests
```

### Running

The build produces a shaded Jar that can be run using the `hadoop` command:



```
hadoop jar parquet-cli-1.12.3-runtime.jar org.apache.parquet.cli.Main
```

# Search jar in maven repo: <https://search.maven.org>



# type search "a:parquet-cli" g:org.apache.parquet

← → ↻ 🔍 search.maven.org 📄 ★

 Maven Central Repository Search [Quick Stats](#)  [GitHub](#)

Search ✕ 🔍

**a:parquet-cli**

<b>org.apache.parquet</b>		Updated 28-Nov-2024
Artifact ID	parquet-cli	
Latest Version	1.15.0	
<b>com.gojek.parquet</b>		Updated 31-May-2021
Artifact ID	parquet-cli	
Latest Version	1.11.9	
<b>com.intel.qat</b>		Updated 29-May-2020
Artifact ID	parquet-cli	
Latest Version	1.10.0	

# Downloading "parquet-cli" runtime.jar for same version 1.13.0 (same as spark/jars/\*.jar)

search.maven.org/artifact/org.apache.parquet/parquet-cli/1.13.0/jar

sonatype | Maven Central Repository Search | Quick Stats | GitHub

org.apache.parquet:parquet-cli:1.13.0

org.apache.parquet:parquet-cli: 1.13.0 View on OSS Index Browse Downloads

**Apache Parquet Command-line**  
Home page <https://parquet.apache.org>

**org.apache.parquet:parquet-cli**  
1.13.0

```
<!--
~ Licensed to the Apache Software Foundation (ASF) under one
~ or more contributor license agreements. See the NOTICE file
~ distributed with this work for additional information
~ regarding copyright ownership. The ASF licenses this file
~ to you under the Apache License, Version 2.0 (the
~ "License"); you may not use this file except in compliance
~ with the License. You may obtain a copy of the License at
~
~ http://www.apache.org/licenses/LICENSE-2.0
~
~ Unless required by applicable law or agreed to in writing,
~ software distributed under the License is distributed on an
```

**Apache Maven**  
[maven.apache.org](https://maven.apache.org)

```
<dependency>
<groupId>org.apache.parquet</groupId>
<artifactId>parquet-cli</artifactId>
<version>1.13.0</version>
</dependency>
```

**Gradle Groovy DSL**  
[gradle.org](https://gradle.org)

```
implementation 'org.apache.parquet:parquet-cli:1.13.0'
```

**Gradle Kotlin DSL**

Available files:

- cyclonedx.json
- cyclonedx.xml
- jar
- javadoc.jar
- pom
- runtime.jar
- sources.jar
- tests.jar

# Launching parquet-cli

**java -jar parquet-cli-runtime.jar help**

(using shaded jar, otherwise may need to add hadoop classpath)

```
C:\apps\spark>java -jar parquet-cli-1.15.0-SNAPSHOT-runtime.jar help

Usage: parquet [options] [command] [command options]

Options:
  -v, --verbose, --debug
                        Print extra debugging information

Commands:
  help
                        Retrieves details on the functions of other commands
  meta
                        Print a Parquet file's metadata
  pages
                        Print page summaries for a Parquet file
  dictionary
                        Print dictionaries for a Parquet column
  check-stats
                        Check Parquet files for corrupt page and column stats (PARQUET-251)
  schema
                        Print the Avro schema for a file
  csv-schema
                        Build a schema from a CSV data sample
  convert-csv
```

# parquet-cli meta [1/2]

**java -jar parquet-cli-runtime.jar meta file.parquet**

```
C:\apps\spark>java -jar parquet-cli-1.15.0-SNAPSHOT-runtime.jar meta C:\apps\spark\spark-warehouse\db1.db\addr\part-c000.snappy.parquet

File path: C:\apps\spark\spark-warehouse\db1.db\addr\part-c000.snappy.parquet
Created by: parquet-mr version 1.13.1 (build db4183109d5b734ec5930d870cdae161e408ddba)
Properties:
    org.apache.spark.version: 3.5.0
    org.apache.spark.sql.parquet.row.metadata: {"type":"struct","fields":[{"name":"uid_adresse","type":"string","nullable":true,"metadata":{}}, {"name":"commune_insee","type":"string","nullable":true,"metadata":{}}, {"name":"commune_deleguee_insee","type":"string","nullable":true,"metadata":{}}, {"name":"commune_deleguee_nom","type":"string","nullable":true,"metadata":{}}, {"name":"voie_nom","type":"string","nullable":true,"metadata":{}}, {"name":"lieudit_complement_nom","type":"string","nullable":true,"metadata":{}}, {"name":"numero","type":"string","nullable":true,"metadata":{}}, {"name":"suffixe","type":"string","nullable":true,"metadata":{}}, {"name":"position","type":"string","nullable":true,"metadata":{}}, {"name":"y","type":"double","nullable":true,"metadata":{}}, {"name":"x","type":"double","nullable":true,"metadata":{}}, {"name":"lat","type":"double","nullable":true,"metadata":{}}, {"name":"cad_parcelle","type":"string","nullable":true,"metadata":{}}, {"name":"source","type":"string","nullable":true,"metadata":{}}, {"name":"certification_commune","type":"string","nullable":true,"metadata":{}}, {"name":"date_der_maj","type":"date","nullable":false,"metadata":{}}]}
Schema:
message spark_schema {
  optional binary uid_adresse (STRING);
  optional binary cle_interop (STRING);
  optional binary commune_insee (STRING);
  optional binary commune_nom (STRING);
  optional int32 commune_deleguee_insee;
  optional binary commune_deleguee_nom (STRING);
  optional binary voie_nom (STRING);
  optional binary lieudit_complement_nom (STRING);
  optional int32 numero;
  optional binary suffixe (STRING);
  optional binary position (STRING);
  optional double y;
  optional double x;
  optional double lat;
  optional binary cad_parcelle (STRING);
  optional binary source (STRING);
  optional binary certification_commune (STRING);
  optional date date_der_maj;
```



# parquet-cli meta [2/2]

Row group 0: count: 219962 71,92 B records start: 4 total(compressed): 15,086 MB total(uncompressed):33,177 MB

	type	encodings	count	avg size	nulls	min / max
uid_adresse	BINARY	S _ R_ F	219962	34,79 B	47399	" @a:00003b17-45be-48c1-aa..." / "Argence"
cle_interop	BINARY	S _	219962	5,47 B	0	"01001_0005_00026" / "54"
commune_insee	BINARY	S _ R	219962	0,01 B	1	"01001" / "01350"
commune_nom	BINARY	S _ R	219962	0,02 B	0	"Ambléon" / "Évosges"
commune_deleguee_insee	INT32	S _ R	219962	0,02 B	202278	"1015" / "1442"
commune_deleguee_nom	BINARY	S _ R	219962	0,02 B	202277	"6518380.18" / "Étрез"
voie_nom	BINARY	S _ R	219962	1,44 B	0	""le château"" / "Îlot Grammont"
lieudit_complement_nom	BINARY	S _ R	219962	0,09 B	197279	""le château"" / "Étрез"
numero	INT32	S _ R	219962	1,45 B	1	"0" / "99999"
suffixe	BINARY	S _ R	219962	0,11 B	206335	"1" / "z"
position	BINARY	S _ R	219962	0,25 B	15376	"bâtiment" / "service technique"
x	DOUBLE	S _	219962	5,06 B	226	"-0.0" / "943078.82"
y	DOUBLE	S _	219962	5,37 B	227	"6505390.74" / "6603083.59"
long	DOUBLE	S _	219962	7,99 B	227	"4.730471" / "6.163086"
lat	DOUBLE	S _	219962	7,64 B	227	"45.617249" / "46.506743"
cad_parcelles	BINARY	S _ R	219962	1,87 B	167172	"010010000A1033" / "01380000_E0291"
source	BINARY	S _ R	219962	0,06 B	2266	"arcep" / "inconnue"
certification_commune	INT32	S _ R	219962	0,03 B	1	"0" / "1"
date_der_maj	INT32	S _ R	219962	0,23 B	0	"1901-01-01" / "2024-12-05"

Row group 1: count: 215938 72,72 B records start: 15819187 total(compressed): 14,976 MB total(uncompressed):33,489 MB

	type	encodings	count	avg size	nulls	min / max
uid_adresse	BINARY	S _	215938	36,79 B	39793	" @a:00003774-58fc-46a7-b1..." / " @v:ffe769c6-a8t
cle_interop	BINARY	S _	215938	5,19 B	0	"01350_0005_00707" / "02546_rcnpc4_00054"
commune_insee	BINARY	S _ R	215938	0,02 B	0	"01350" / "02546"
commune_nom	BINARY	S _ R	215938	0,04 B	0	"Abbécourt" / "Évergnicourt"
commune_deleguee_insee	INT32	S _	215938	0,01 B	208894	"1059" / "2811"
commune_deleguee_nom	BINARY	S _	215938	0,03 B	208894	"Anizy-le-Château" / "Virieu-le-Petit"

# parquet-cli meta | grep "Row group"

```
java -jar parquet-cli-runtime.jar meta file.parquet > meta.txt
```

## grep "Row group" meta.txt | head

```
Row group 0: count: 219962 71,92 B records start: 4 total(compressed): 15,086 MB total(uncompressed):33,177 MB
Row group 1: count: 215938 72,72 B records start: 15819187 total(compressed): 14,976 MB total(uncompressed):33,489 MB
Row group 2: count: 213992 73,68 B records start: 31523114 total(compressed): 15,036 MB total(uncompressed):33,821 MB
Row group 3: count: 233199 70,29 B records start: 47289138 total(compressed): 15,631 MB total(uncompressed):34,232 MB
Row group 4: count: 231558 69,04 B records start: 63679653 total(compressed): 15,246 MB total(uncompressed):33,383 MB
Row group 5: count: 219962 74,89 B records start: 79665824 total(compressed): 15,710 MB total(uncompressed):34,531 MB
Row group 6: count: 218409 73,91 B records start: 96138854 total(compressed): 15,394 MB total(uncompressed):33,957 MB
Row group 7: count: 199907 75,02 B records start: 112281114 total(compressed): 14,301 MB total(uncompressed):31,914 MB
Row group 8: count: 218409 69,87 B records start: 127277278 total(compressed): 14,552 MB total(uncompressed):34,174 MB
Row group 9: count: 272568 65,80 B records start: 142536439 total(compressed): 17,104 MB total(uncompressed):35,616 MB
```

## grep "Row group" meta.txt | tail

```
Row group 108: count: 252150 64,55 B records start: 1719335249 total(compressed): 15,522 MB total(uncompressed):34,027 MB
Row group 109: count: 215240 71,68 B records start: 1735611436 total(compressed): 14,715 MB total(uncompressed):34,014 MB
Row group 110: count: 219962 68,48 B records start: 1751040857 total(compressed): 14,365 MB total(uncompressed):32,978 MB
Row group 111: count: 219962 68,18 B records start: 1766103960 total(compressed): 14,302 MB total(uncompressed):32,914 MB
Row group 112: count: 234887 64,52 B records start: 1781100645 total(compressed): 14,454 MB total(uncompressed):33,355 MB
Row group 113: count: 236624 66,44 B records start: 1796256273 total(compressed): 14,993 MB total(uncompressed):37,495 MB
Row group 114: count: 234063 65,95 B records start: 1811977242 total(compressed): 14,722 MB total(uncompressed):33,626 MB
Row group 115: count: 213992 73,42 B records start: 1827414634 total(compressed): 14,983 MB total(uncompressed):34,526 MB
Row group 116: count: 230100 69,58 B records start: 1843125866 total(compressed): 15,269 MB total(uncompressed):36,039 MB
Row group 117: count: 152925 67,36 B records start: 1859136956 total(compressed): 9,823 MB total(uncompressed):26,622 MB
```

# parquet-cli column-size

**java -jar parquet-cli-runtime.jar column-size file.parquet**

```
C:\apps\spark>java -jar parquet-cli-1.15.0-SNAPSHOT-runtime.jar column-size C:\apps\spark\spa
commune_deleguee_insee-> Size In Bytes: 431641 Size In Ratio: 2.3089352E-4
commune_insee-> Size In Bytes: 322609 Size In Ratio: 1.725701E-4
numero-> Size In Bytes: 26951190 Size In Ratio: 0.014416738
voie_nom-> Size In Bytes: 35412828 Size In Ratio: 0.01894304
uid_adresse-> Size In Bytes: 882820344 Size In Ratio: 0.4722385
date_der_maj-> Size In Bytes: 7501158 Size In Ratio: 0.0040125214
cle_interop-> Size In Bytes: 139359234 Size In Ratio: 0.07454608
source-> Size In Bytes: 1142012 Size In Ratio: 6.108854E-4
long-> Size In Bytes: 213866608 Size In Ratio: 0.11440159
commune_deleguee_nom-> Size In Bytes: 836453 Size In Ratio: 4.4743568E-4
suffixe-> Size In Bytes: 3721170 Size In Ratio: 0.0019905292
cad_parcelles-> Size In Bytes: 66126109 Size In Ratio: 0.035372198
certification_commune-> Size In Bytes: 673339 Size In Ratio: 3.601827E-4
x-> Size In Bytes: 133896306 Size In Ratio: 0.07162385
lieudit_complement_nom-> Size In Bytes: 2886263 Size In Ratio: 0.0015439206
y-> Size In Bytes: 143779166 Size In Ratio: 0.07691039
commune_nom-> Size In Bytes: 549711 Size In Ratio: 2.9405157E-4
position-> Size In Bytes: 4716951 Size In Ratio: 0.0025231927
lat-> Size In Bytes: 204444396 Size In Ratio: 0.109361455
```

# Column Sizes ... sorting

cat column-size

| sed 's/-> Size In Bytes//g' | sed 's/ Size In Ratio//g' | sed 's/: /;/g' > column-size.csv

Column	Size	Ratio	Size in Mo
uid_adresse	882820344	47,22%	841,9
long	213866608	11,44%	204,0
lat	204444396	10,94%	195,0
y	143779166	7,69%	137,1
cle_interop	139359234	7,45%	132,9
x	133896306	7,16%	127,7
cad_parcelles	66126109	3,54%	63,1
voie_nom	35412828	1,89%	33,8
numero	26951190	1,44%	25,7
date_der_maj	7501158	0,40%	7,2
position	4716951	0,25%	4,5
suffixe	3721170	0,20%	3,5
lieudit_complement_nom	2886263	0,15%	2,8
source	1142012	0,06%	1,1
commune_deleguee_nom	836453	0,04%	0,8
certification_commune	673339	0,04%	0,6
commune_nom	549711	0,03%	0,5
commune_deleguee_insee	431641	0,02%	0,4
commune_insee	322609	0,02%	0,3
total	1869437488	100,00%	1782,8

# Parquet Column Sizes Summary

zipcode ("commun\_insee") : 0.3 Mo  
+  
city name ("commun\_nom") : 0.5 Mo  
+  
street name ("voie\_nom") : 33.8 Mo  
+  
numero : 25.7 Mo

=> 60 Mega (3% of file)

+ longitude, latitude : 204 Mo + 195 Mo  
=> 459 Mo

most of the column size is

uid\_address : 840 Mo (47.2%)  
+  
x, y ... redundant with longitude,latitude  
+  
cle\_interop : 132 Mo (7.4%)

... we will see parquet "column-pruning" in action : skipping unnecessary columns !