



Parquet : un nouveau format au fort potentiel

Éric Mauvière, Icem7



Parquet : un format né en 2014, en vogue en France depuis 2023



En 2023, l'Insee frappe fort avec 2 bases détaillées du recensement sur **data.gouv.fr** : elles ont 20 à 25 millions de lignes, 70 colonnes et pèsent 500 Mo au format parquet.

Recensement de la population - Fichiers détail : Logements ordinaires

Description

Producteur



Institut National de la
Statistique et des Etudes
Economiques (Insee) ✓

Service public

2 FICHIERS PRINCIPAUX

fd-logemt-2021.parquet

Mis à jour le 12 février 2025 — parquet (501.1Mo) — 43

↓ PARQUET

fd-logemt-2020.parquet

Mis à jour le 12 février 2025 — parquet (481.0Mo) — 985

↓ PARQUET

[» Recensement de la population - fichier détail logements](#)



Parquet : un format compact, 2 à 4 fois plus léger qu'un CSV zippé

Service statistique ministériel de la sécurité intérieure... + Suivre ...

1 872 abonnés
2 sem. •

[OPEN DATA]
Le SSMSI expérimente un nouveau format #opensource de mise à disposition de 2 de ses jeux de données en #opendata, le format parquet. ...voir plus



The poster features the SSMSI logo at the top, followed by the title 'Open Data' in large blue letters. Below the title, it reads 'Nouveau format Parquet pour les données de la délinquance enregistrée par la police et gendarmerie nationales (2016-2022)'. The central illustration shows a man in a blue shirt pointing at a large digital screen displaying a bar chart and a pie chart. To the left of the screen is a clock icon, and below it are some books and a small potted plant. At the bottom center is the French flag.

Les statistiques communales de la **délinquance enregistrée** sont diffusées non seulement au format CSV mais aussi en **Parquet** (depuis février 2024).

- 3,5 millions de lignes
- CSV : 400 Mo (40 Mo en csv.gz)
- Parquet : **11 Mo**



Parquet : un format tabulaire typé : des colonnes numériques, caractères...



	A-Z CODGEO_2024 ▼	123 annee ▼	A-Z classe ▼	A-Z unité.de.compte ▼	123 fait ▼
1	01004	16	Coups et blessures volontaires	victime	59
2	01004	16	Coups et blessures volontaires intrafamiliaux	victime	21
3	01004	16	Autres coups et blessures volontaires	victime	38
4	01004	16	Violences sexuelles	victime	13
5	01004	16	Vols sans violence contre des personnes	victime entendue	133
6	01004	16	Cambriolages de logement	infraction	72
7	01004	16	Vols de véhicules	véhicule	49
8	01004	16	Vols dans les véhicules	véhicule	46
9	01004	16	Vols d'accessoires sur véhicules	véhicule	22
10	01004	16	Destructions et dégradations volontaires	infraction	116
11	01004	16	Trafic de stupéfiants	Mis en cause	44

Les données extraites
d'un format Parquet
sont proprement **typées**
et donc **immédiatement**
réutilisables.



Requêttable en ligne, avec un simple navigateur...

localhost:4213

DuckDB

Search

Notebooks +

Faits de délinqu...

Attached databases +

memory

main

Faits de délinquance - base communale pluri-annuelle

Run memory

```
1 FROM 'https://static.data.gouv.fr/resources/bases-statistiques
2 SELECT CODGEO_2024, '20' || annee annee, classe, faits
3 WHERE faits > 0 ;
```

Query results

	CODGEO_2024	annee	classe	faits
i	01004	2016	Coups et blessures volontaires	59
	01004	2016	Coups et blessures volontaires intrafamiliaux	21
	01004	2016	Autres coups et blessures volontaires	38
	01004	2016	Violences sexuelles	13
	01004	2016	Vols sans violence contre des personnes	133
	01004	2016	Cambriolages de logement	72
	01005	2016	Cambriolages de logement	12
	01005	2016	Destructions et dégradations volontaires	5
	01007	2016	Cambriolages de logement	10
	01007	2016	Violences sexuelles	1

Filter 312 586 Rows

Sans **nul besoin** de logiciel lourd
ou de serveur, il est facile
d'interroger un fichier Parquet
local, ou même **distant**.

Par exemple avec DuckDB
et en SQL.



... ou utilisé dans un pipeline scripté

Python

```
import duckdb

# Connexion à DuckDB (en mémoire)
con = duckdb.connect()

# Créer une relation à partir du Parquet distant
rel = con.read_parquet(
    'https://static.data.gouv.fr/resources/commune2025/20250321-104820/commune.parquet'
)

# Appliquer filtre et sélection via l'API relationnelle
rel_filtered = (
    rel.filter("INSEE_DEP = '34'")
    .project("INSEE_COM, NOM, POPULATION")
)

# Matérialiser en DataFrame pandas
df = rel_filtered.to_df()
```

Si vous préférez votre langage habituel de script et/ou avez besoin de construire une chaîne avancée de traitements et de visualisations, **tous les grands langages (Python, R...)** savent lire et écrire du Parquet.

R

```
library(duckdb)

con = dbConnect(duckdb())
dbExecute(con, "SET autoloading_extensions = 1")

df = tbl(con, "'https://static.data.gouv.fr/resources/commune2025/20250321-104820/commune.parquet'") |>
  filter(INSEE_DEP == '34') |>
  select(INSEE_COM, NOM) |>
  collect()
```



Premiers avantages de Parquet



C'est un format de fichier **compact** et **portable**, qui n'est pas lié à un logiciel particulier.

Les colonnes sont déjà **typées**, cela facilite les calculs et la compression des données.

On peut **travailler sur des fichiers en ligne** directement, sans recopie ni devoir en passer par une base de données.



Cela marche aussi avec la géographie Hello Geoparquet !



Admin Express de l'IGN, couche COMMUNE :

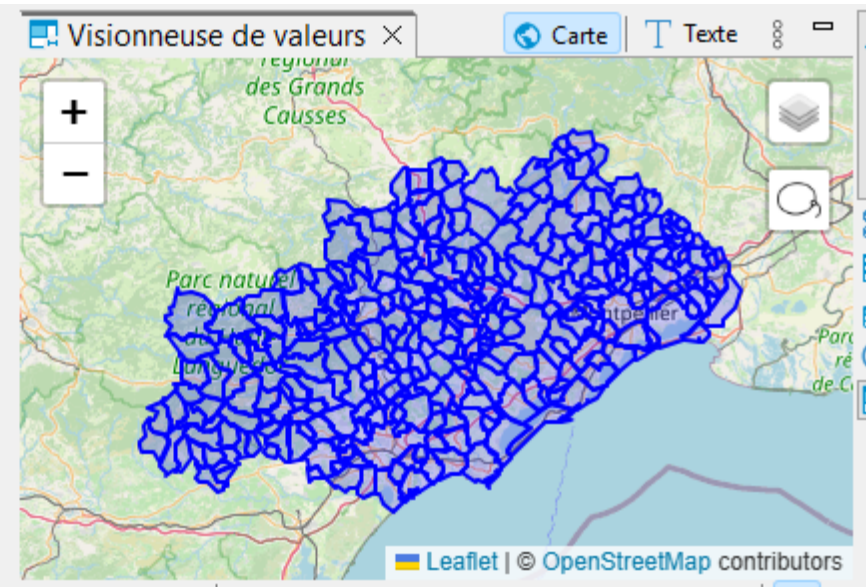
- Shapefile : 420 Mo
- Geopackage : 430 Mo
- FlatGeobuf : 410 Mo
- GeoJSON : 1 100 Mo
- **GeoParquet 1.1 : 270 Mo**



Extraction d'un département en ligne

```
FROM 'https://static.data.gouv.fr/resources/commune2025/20250321-104820/commune.parquet'  
SELECT insee_com, nom, population, geom  
WHERE insee_dep = '34';
```

	A-Z INSEE_COM	A-Z NOM	123 POPULATION	geom
1	34071	Ceilhes-et-Rocozeles	274	POLYGON ((3.148800596556373 43.8100...
2	34233	Roqueredonde	212	POLYGON ((3.151446264822092 43.8142...
3	34231	Romiguières	21	POLYGON ((3.22270193120059 43.81764...
4	34230	Les Rives	152	POLYGON ((3.2866645066025 43.829742...
5	34064	Le Caylar	465	POLYGON ((3.343324794287086 43.8544...
6	34196	Pégairolles-de-l'Escalette	154	POLYGON ((3.32335445044465 43.77574...
7	34253	Saint-Félix-de-l'Héras	33	POLYGON ((3.286555009765466 43.8483...
8	34132	Lauroux	202	POLYGON ((3.248053445389567 43.8140...
9	34212	Poujols	181	POLYGON ((3.299701421455338 43.7850...
10	34304	Soubès	943	POLYGON ((3.36817176583715 43.80457...
11	34251	Saint-Étienne-de-Gourq	522	POLYGON ((3.388983815844709 43.8161...



48 Mo extraits par des « range requests » http d'un fichier de 275 Mo

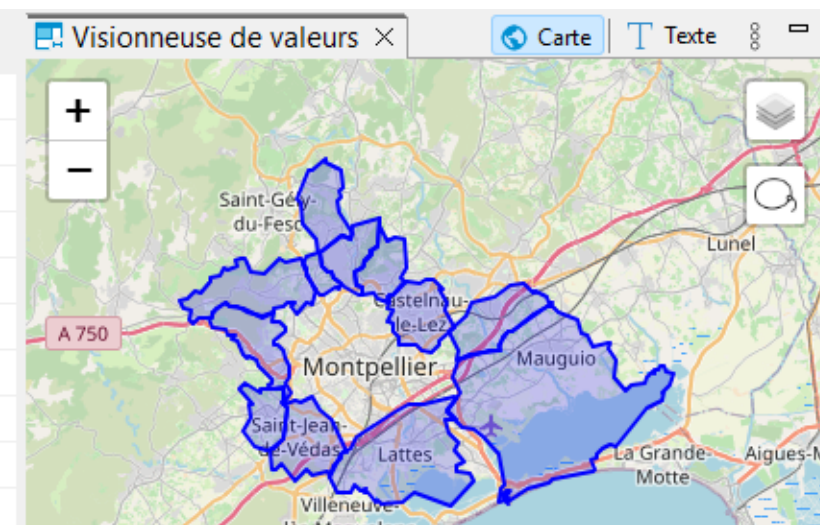


Requêtage géographique fin sans base de données spatiale

```
WITH t1 AS (
  FROM 'https://static.data.gouv.fr/resources/commune2025/20250321-104820/commune.parquet'
  SELECT insee_com, nom, population, geom,
  FIRST(geom) OVER (ORDER BY IF(insee_com = '34172', 0, 1)) geom_montpellier
  WHERE bbox.xmin BETWEEN 3.02 AND 4.49 -- Hérault
  AND bbox.ymin BETWEEN 43.17 AND 43.98
) FROM t1
SELECT * EXCLUDE(geom_montpellier)
WHERE ST_Touches(geom, geom_montpellier);
```

Communes jouxtant Montpellier

	A-Z INSEE_COM	A-Z NOM	123 POPULATION	geom
1	34240	Saint-Aunès	4 136	POLYGON ((3.933067201823484 43.6
2	34077	Clapiers	5 811	POLYGON ((3.883263110011164 43.6
3	34057	Castelnau-le-Lez	24 888	POLYGON ((3.889472099063758 43.6
4	34169	Montferrier-sur-Lez	4 054	POLYGON ((3.883263110011164 43.6
5	34247	Saint-Clément-de-Rivière	5 058	POLYGON ((3.832587971960665 43.7
6	34116	Grabels	8 973	POLYGON ((3.761402135747402 43.6
7	34123	Juvignac	12 783	POLYGON ((3.816935445903341 43.6
8	34134	Lavérune	3 312	POLYGON ((3.783361093539908 43.5
9	34270	Saint-Jean-de-Védas	12 570	POLYGON ((3.816256136190438 43.5
10	34129	Lattes	17 544	POLYGON ((3.847643470144744 43.5



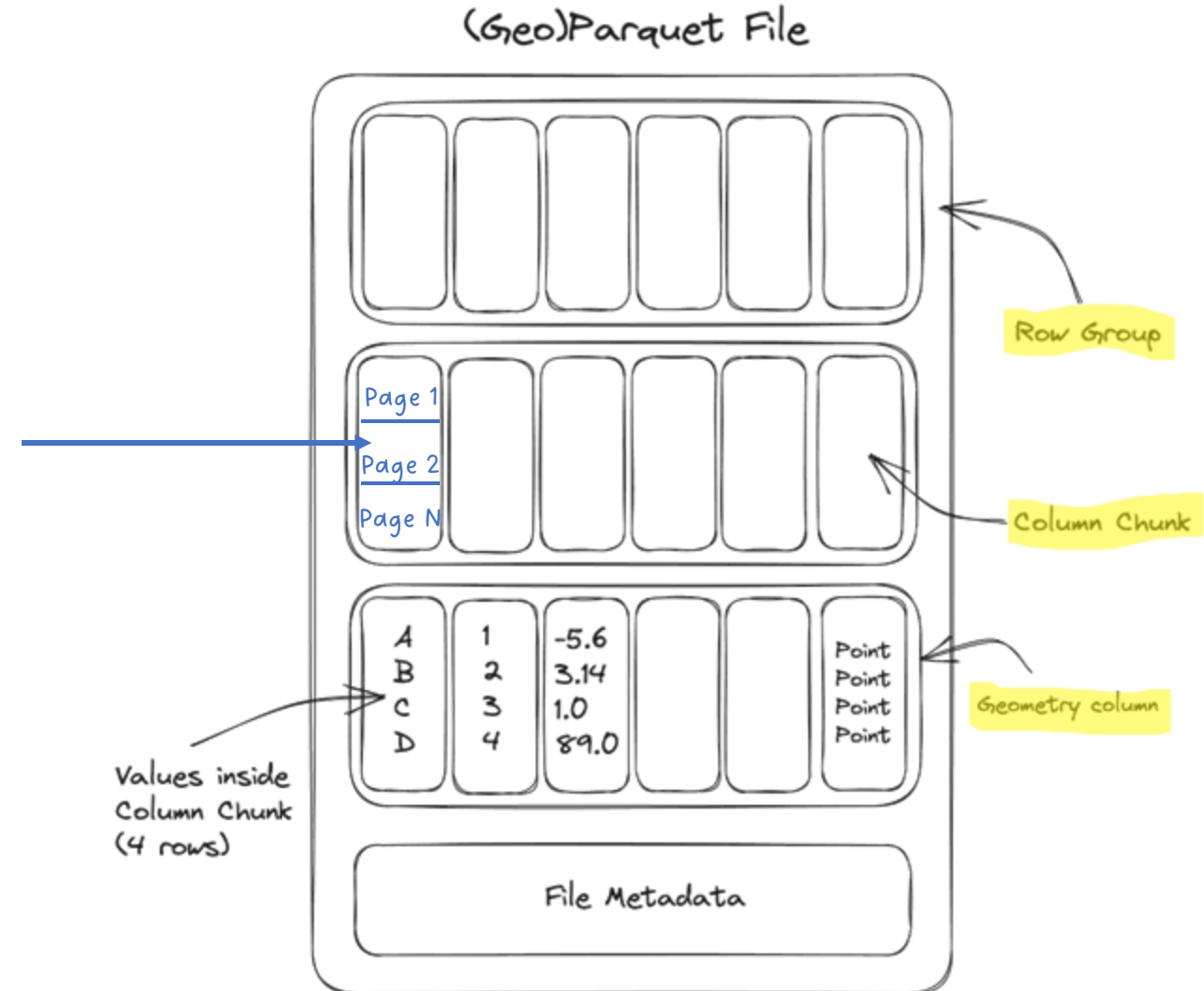


Un fichier parquet sépare les colonnes et se structure en groupes de lignes

D'où la dénomination **Parquet**.

Un « **column chunk** » est lui-même découpé en **pages**.

Des **métadonnées** sont stockées dans chaque **row-group**, et même pour chaque **page**.



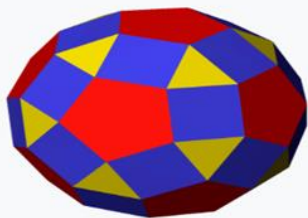


Parquet est soutenu par l'OGC...



ENCODINGS

Data Encoding Standards provide the rules for structuring and organising geospatial data for use in a given context. They apply to binary formats (e.g.: GeoTIFF), as well as human-readable formats (e.g.: JSON).



JSON-FG

If you are using GeoJSON, but your data is not in WGS84, be sure to check this encoding. OGC Features and Geometries JSON extends GeoJSON, in order to overcome some of its limitations.



GeoParquet

Columnar geospatial format, based on Apache Parquet. It enables interoperability between cloud data warehouses.



Cloud Optimized GeoTIFF (COG)

If you work with imagery and grid coverage data on the web, you could check out this format; backwards compatible with GeoTIFF.

25 JANVIER 2023

Contact : info@ogc.org

GeoParquet ajoute des types géospatiaux à Apache Parquet pour servir de format de données vectorielles natif cloud efficace.

Depuis 2023, un groupe de travail OGC (GeoParquet SWG) travaille à en faire un **standard officiel pour l'échange de données géospatiales tabulaires.**



... et par une multitude d'industriels ou de projets libres

Who is involved in GeoParquet?



FOURSQUARE





MERCI DE VOTRE ATTENTION

