Les nouveaux formats de données : parquet et geoparquet

Webinaire

7 février 2024

Éric Mauvière, icem7

Programme de l'intervention

1 – Parquet comparé aux formats classiques et aux bases de données

2 – Démos live

3 – Parquet, comment ça marche ?

4 – Parquet bouscule les usages établis

Annexe : bibliographie

Quel est ce nouveau format que teste la statistique publique ?

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**

⟫ SSMSI - data.gouv.fr

L’Insee a montré la voie courant 2023…

Avec deux bases de données détaillées du recensement sur **data.gouv.fr :**
20-25 millions de lignes, 500 Mo au format parquet.

Guide d’utilisation des données du recensement de la population au format Parquet

Un post de blog pour accompagner la mise à disposition des données détaillées du recensement au format Parquet.

PYTHON R PARQUET

AUTHORS
Antoine Palazzolo
Lino Galiana

PUBLISHED
October 23, 2023

⟫ Guide recensement parquet Insee

… et s'en est fait fièrement l'écho à l'étranger

We tried Parquet for bulk data dissemination…

Detailed results of the Census requested with DuckDB

Soapbox Session | Workshop of the HLG-MOS 2023

Insee

⟫ Is Apache Parquet the future of OpenData? Romain Lesur, 2023

Parquet as a default file format for dissemination?

- Parquet can handle any size of data (small, large and big)
- Adopted as INSEE's internal default file format, replacing SAS format (SAS7BDAT)
- Our experiment for bulk data dissemination was a great success!

Is it the future of open data?

Parquet comparé aux formats classiques

et aux bases de données

Les formats de fichiers traditionnels

```
1  COMMUNE;ARM;DCRAN;ACHLR;AGEMEN8;AGEREVQ;ANEMC;
2  01001;ZZZZZ;01001;1;80;085;5;0;7;7;16;49;ZZ;3;
3  01001;ZZZZZ;01001;4;55;060;5;0;5;5;11;01;16;1;
4  01001;ZZZZZ;01001;5;25;035;3;0;6;6;13;69;16;2;
5  01001;ZZZZZ;01001;1;65;065;5;0;7;7;11;01;ZZ;1;
6  01001;ZZZZZ;01001;1;40;045;3;0;2;2;15;01;21;1;
7  01001;ZZZZZ;01001;1;40;015;3;0;5;2;14;01;16;1;
8  01001;ZZZZZ;01001;1;40;015;3;0;6;2;12;01;16;1;
9  01001;ZZZZZ;01001;5;25;005;3;0;8;6;ZZ;69;ZZ;2;
10 01001;ZZZZZ;01001;5;25;010;3;0;8;4;ZZ;01;ZZ;1;
```

CSV, XLS(X), JSON, XML, GeoJSON, Shapefile…

Avantages

- faciles à **produire** ;
- se consultent relativement facilement, dans un éditeur, un **tableur**, un **logiciel SIG** ;
- certains sont des **standards recommandés** par les autorités (ex. : RGI),
  parmi eux, XML et JSON savent gérer des structures imbriquées.

Inconvénients

- formats parfois fragiles (encodage, **délimiteurs**) ;
- pauvres en métadonnées (types de colonnes absents, projection ± définie) ;
- **verbeux** (JSON, XML) ;
- compliqués à manier quand le **nombre de lignes** ou de colonnes devient important.

Les bases de données

PostgreSQL, PostGIS, Oracle, MySQL, MSSQL, SAS…

Avantages

- des **métadonnées** riches (colonnes typées, statistiques) ;
- gros volumes de données, accès concurrents en lecture/**écriture** ;
- avec l'**indexation**, possibilité de ne lire que les données pertinentes.

Inconvénients

- complexité de configuration et maintenance d'un **serveur** ;
- protocoles de **transfert client-serveur** (sérialisation/désérialisation) lourds ;
- Formats de données souvent **propriétaires et/ou dépendants** du moteur de requête ad hoc.

Parquet propose une synthèse

- C'est un format de fichier **portable**, indépendant de tout outil de lecture/écriture.
- Les colonnes sont **typées**, cela facilite les calculs.
- Il est « **orienté colonnes** », optimisé pour la lecture sélective d'informations.
- Parquet est un format **ouvert**, géré sous forme d'un projet Apache.
- GeoParquet est candidat à l'OGC pour devenir le **standard d'encodage des données géo-vectorielles** en ligne (« *cloud native vector data* »).
- Parquet est auto-documenté : des **statistiques détaillées** renseignent des groupes de lignes et, à l'intérieur, des « pages » de données.

Parquet a 10 ans, en est à sa version 2

Julien le Dem
co-créateur de Parquet

- Il résulte d'une initiative conjointe de **Twitter** et Cloudera, dès 2012, devient un **projet Apache** indépendant et libre en 2014.
- Puis utilisé par Uber, Netflix, Facebook, LinkedIn, et **tout le monde du big data**.
- Devient davantage grand public en 2022, notamment poussé par **DuckDB**.
- En France, il est dès 2023 **valorisé et promu par l'Insee**, et à sa suite par d'autres services de la statistique publique.

#2

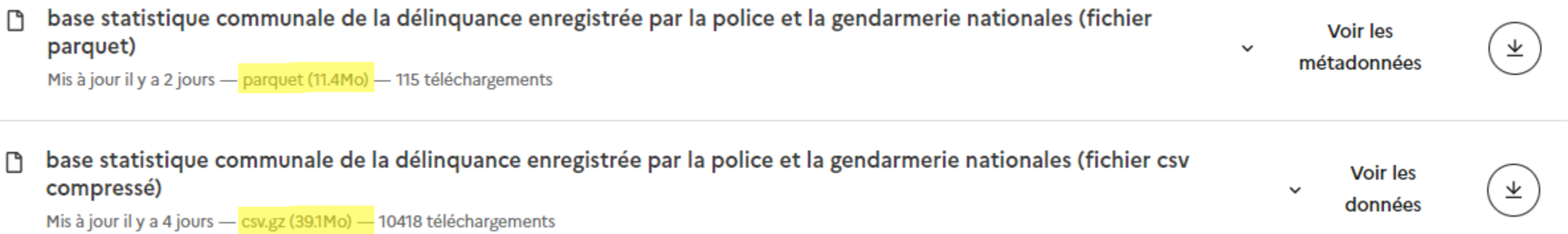
Deux démos :

1 – Délinquance enregistrée au sein de Metz Métropole

2 – Étab. Sirene proches de la place Kléber à Strasbourg

Démo 1 : délinquance enregistrée au sein de Metz Métropole

3,5 millions de lignes, 11 Mo au format parquet.



⟫ Bases de la délinquance enregistrée - SSMSI - data.gouv.fr

Deux outils légers et gratuits pour jouer avec Parquet

DBeaver : éditeur SQL
qui se connecte à une variété
de bases de données.

dbeaver.io

DuckDB : moteur SQL
ultra-léger et ultra-rapide.

duckdb.org

DuckDB

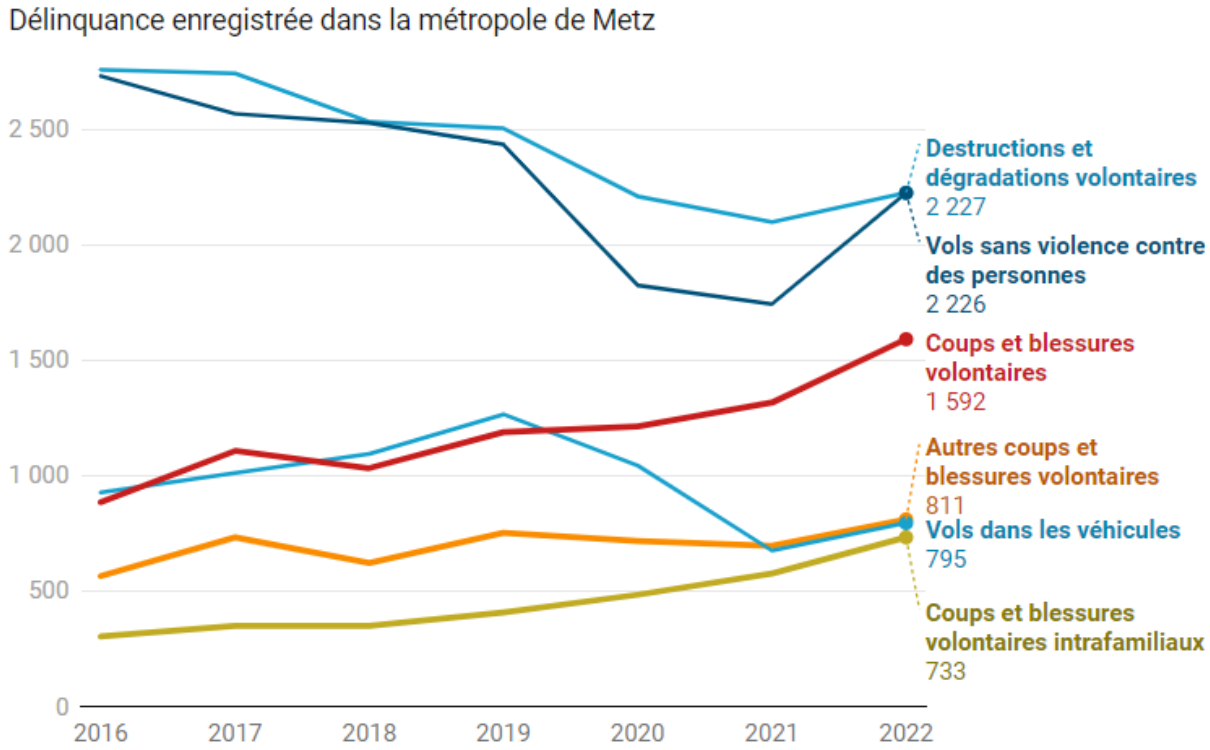
Démo 1 : requête SQL sur 2 fichiers parquet en ligne

```sql
CREATE OR REPLACE VIEW basecom_delinq AS
FROM read_parquet('https://static.data.gouv.fr/resources/bases-statistiques-communale-et-departementale-de-la-delinquance-enregistree-par-la-police-et-la-gendarmerie-nationales/'
|| '20240214-113324/base-communale-2023-07-17.parquet') ;


CREATE OR REPLACE VIEW com_epci_metz_metro AS
FROM 'https://static.data.gouv.fr/resources/communes-2023-format-parquet/20240122-085355/communes2023.parquet'
SELECT codgeo AS codgeo_2023
WHERE epci = '200039865' ; -- Metz Métropole


WITH agg_metz_metro AS (
  FROM basecom_delinq
  JOIN com_epci_metz_metro USING (codgeo_2023)
  SELECT '20' || annee AS annee, classe, sum(faits) AS faits
  GROUP BY ALL
)
PIVOT agg_metz_metro ON annee USING(sum(faits))
ORDER BY "2022" DESC
LIMIT 6 ;
```

Testez dans votre navigateur

Démo 1 : résultats

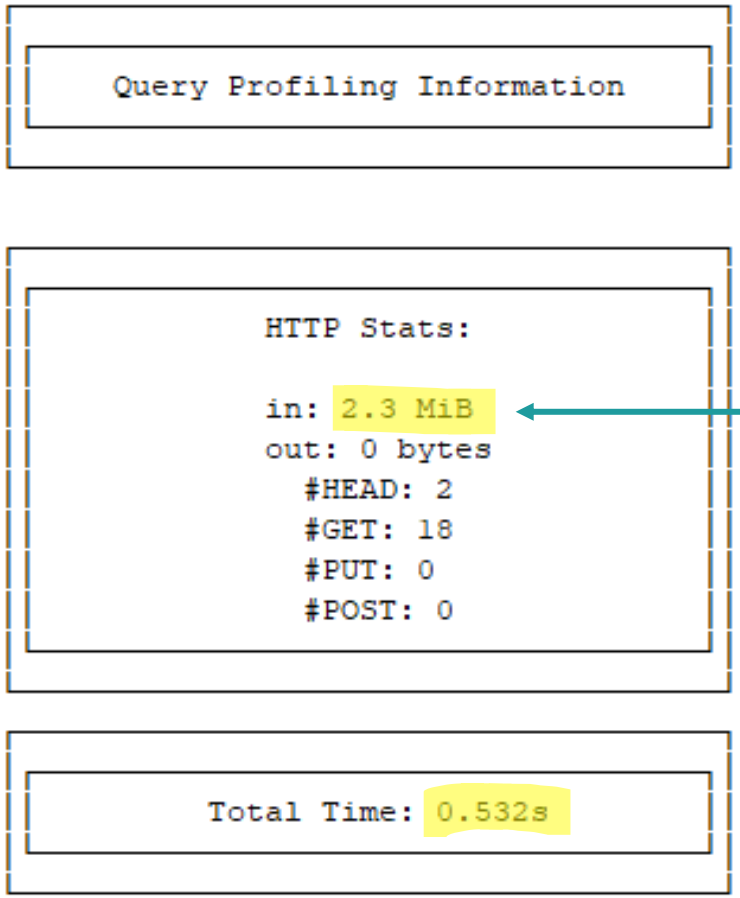
classe	2016	2017	2018	2019	2020	2021	2022
Destructions et dégradations volontaires	2 760	2 744	2 534	2 506	2 211	2 100	2 227
Vols sans violence contre des personnes	2 733	2 569	2 530	2 436	1 826	1 745	2 226
Coups et blessures volontaires	885	1 108	1 033	1 189	1 214	1 317	1 592
Autres coups et blessures volontaires	565	733	622	752	717	696	811
Vols dans les véhicules	928	1 012	1 095	1 266	1 045	676	795
Coups et blessures volontaires intrafamiliaux	303	349	349	408	484	576	733



La requête n’a chargé qu’une petite partie des fichiers détails

Les métadonnées du format parquet permettent d’**optimiser** le plan d’exécution des requêtes.

Elles indiquent comment scanner (et donc télécharger) **seulement les blocs pertinents** du fichier.



2 Mo lus sur 11 Mo

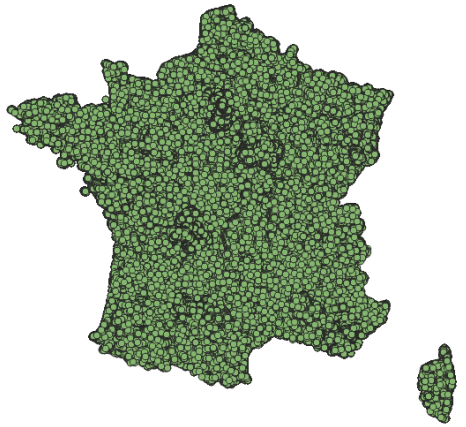
Démo 2 : extrait Sirene autour de la place Kléber à Strasbourg

Sirene géolocalisé par Étalab

15 millions d'établissements

Format	Poids
CSV	4 500 Mo
CSV gzippé	1 200 Mo
GeoParquet	**900 Mo**

⟫ Sirene géolocalisé - Etalab 2024



Filaire de circulation Eurométropole de Strasbourg

53 000 tronçons

Format	Poids
GeoJSON	85 Mo
Shapefile	50 Mo
Geopackage	24 Mo
GeoParquet	**5 Mo**

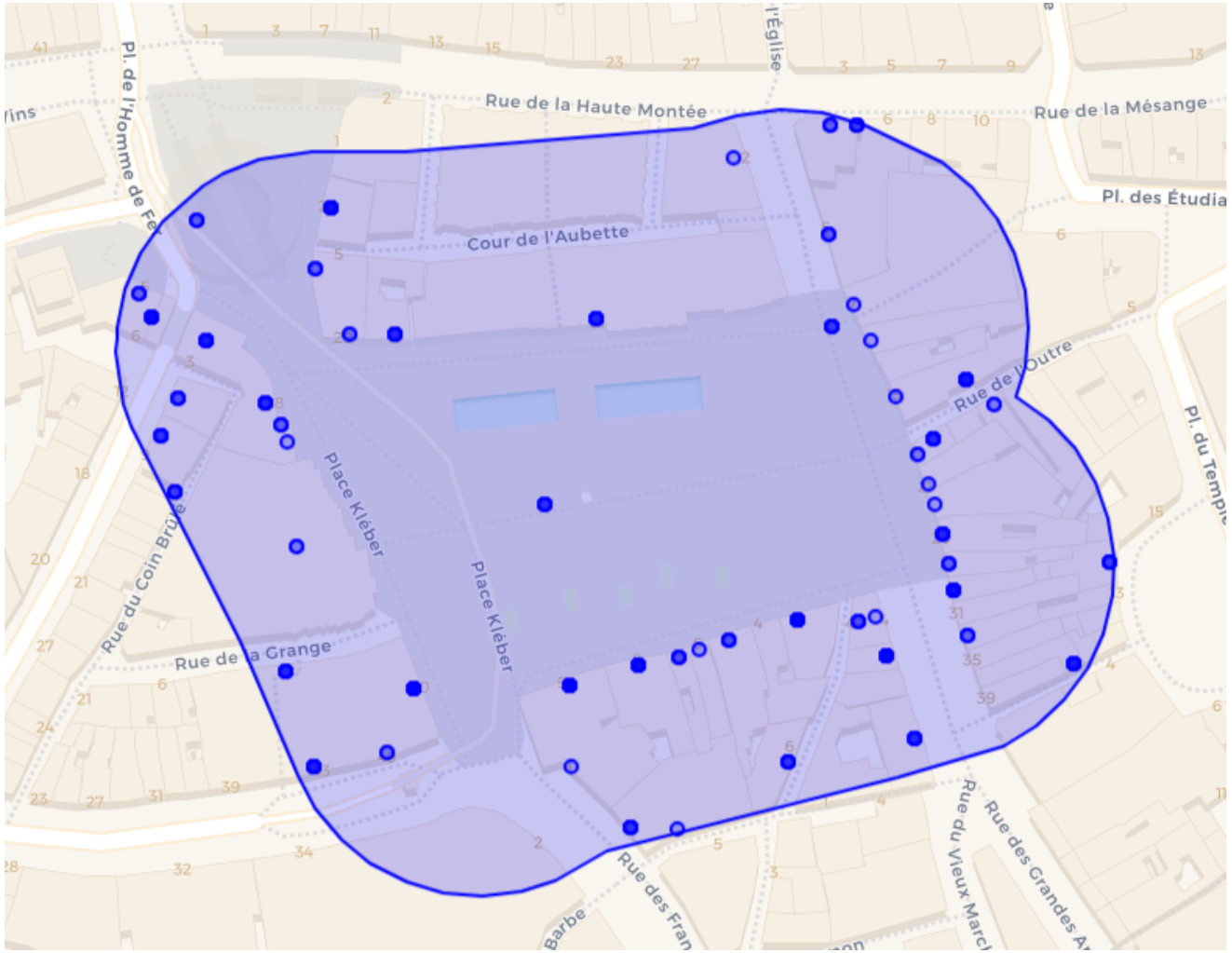
⟫ Filaire de circulation Eurométropole de Strasbourg

Démo 2 : requête SQL sur 2 fichiers geoparquet en ligne

```
LOAD spatial ;

CREATE OR REPLACE VIEW sirenegeo AS
FROM 'https://static.data.gouv.fr/resources/sirene-geolocalise-parquet/20240107-143656/sirene2024-
geo.parquet' ;

CREATE OR REPLACE VIEW filaire_strasbourg AS
FROM 'https://static.data.gouv.fr/resources/filaire-de-voies-strasbourg-format-parquet/20240303-
140908/filaire-de-circulation-strasbourg-geo.parquet' ;

CREATE OR REPLACE TABLE buffer_kleber AS
FROM filaire_strasbourg
SELECT ST_Union_Agg(geometry.ST_geomFromWkb())
        .ST_Transform('EPSG:4326','EPSG:2154',true) -- reprojection en référentiel Lambert93
        .ST_Buffer(50) -- buffer de 50 mètres
        .ST_Transform('EPSG:2154','EPSG:4326',true) AS geometry
WHERE nom_commune_droit = 'Strasbourg' AND nom_voie_droit = 'PLACE KLEBER' ;

FROM sirenegeo
JOIN buffer_kleber ON ST_Within(ST_GeomFromWKB(sirenegeo.geometry), buffer_kleber.geometry)
SELECT siret, denominationUniteLegale AS nom, activitePrincipaleEtablissement AS NIV5,
geo_adresse, sirenegeo.geometry.ST_geomFromWkb() AS geometry
WHERE codeCommuneEtablissement = '67482' ;
```

Testez dans votre navigateur

400 établissements actifs Sirene à moins de 50 m de la place Kleber




```
HTTP Stats:

in: 12.9 MiB
out: 0 bytes
  #HEAD: 1
  #GET: 46
  #PUT: 0
  #POST: 0
```

```
Total Time: 0.989s
```

siret	nom	NIV5
89889300300019	1ERE AVENUE	68.31Z
88517639600418	1MONDE9	47.72A
85390952100017	3D DECORATION	43.34Z
48967873000016	45 PROD	59.12Z
44122586900045	ABAX	70.22Z
42462578800098	ADONIS	46.19A
52880545000012	AEQUALIS	46.90Z
41400963900033	AFIM	68.31Z
81998146500019	AGENT COMMERCIAL AGENCY	46.19B
53066364000015	AGORA COMMUNICATION	70.21Z

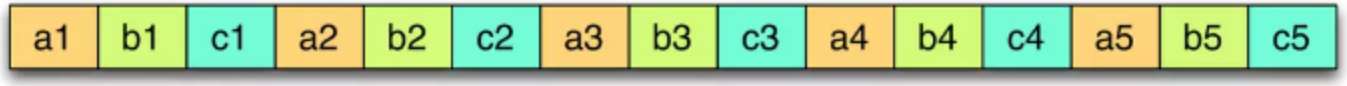
Parquet, comment ça marche ?

C’est un format « orienté colonnes »

Une table logique

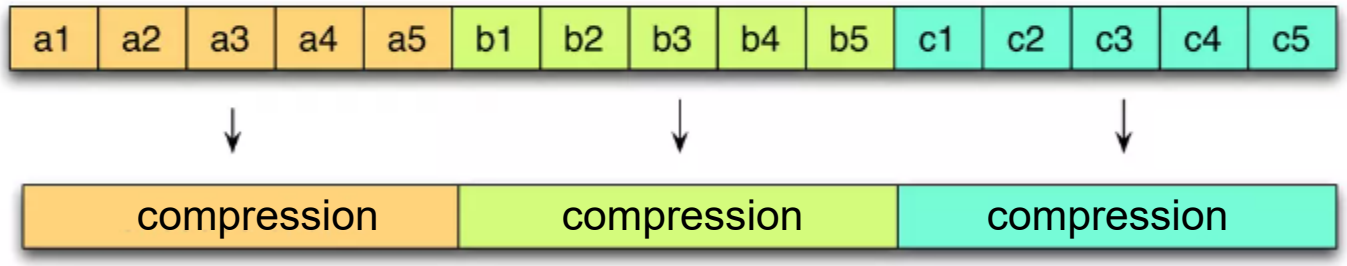
a	b	c
a1	b1	c1
a2	b2	c2
a3	b3	c3
a4	b4	c4
a5	b5	c5

Stockage physique en lignes



→ Insertion de nouveaux enregistrements aisée.

Stockage physique en colonnes



→ Encodage et compression ++

Il devient plus facile de lire les seules colonnes utiles dans une requête.

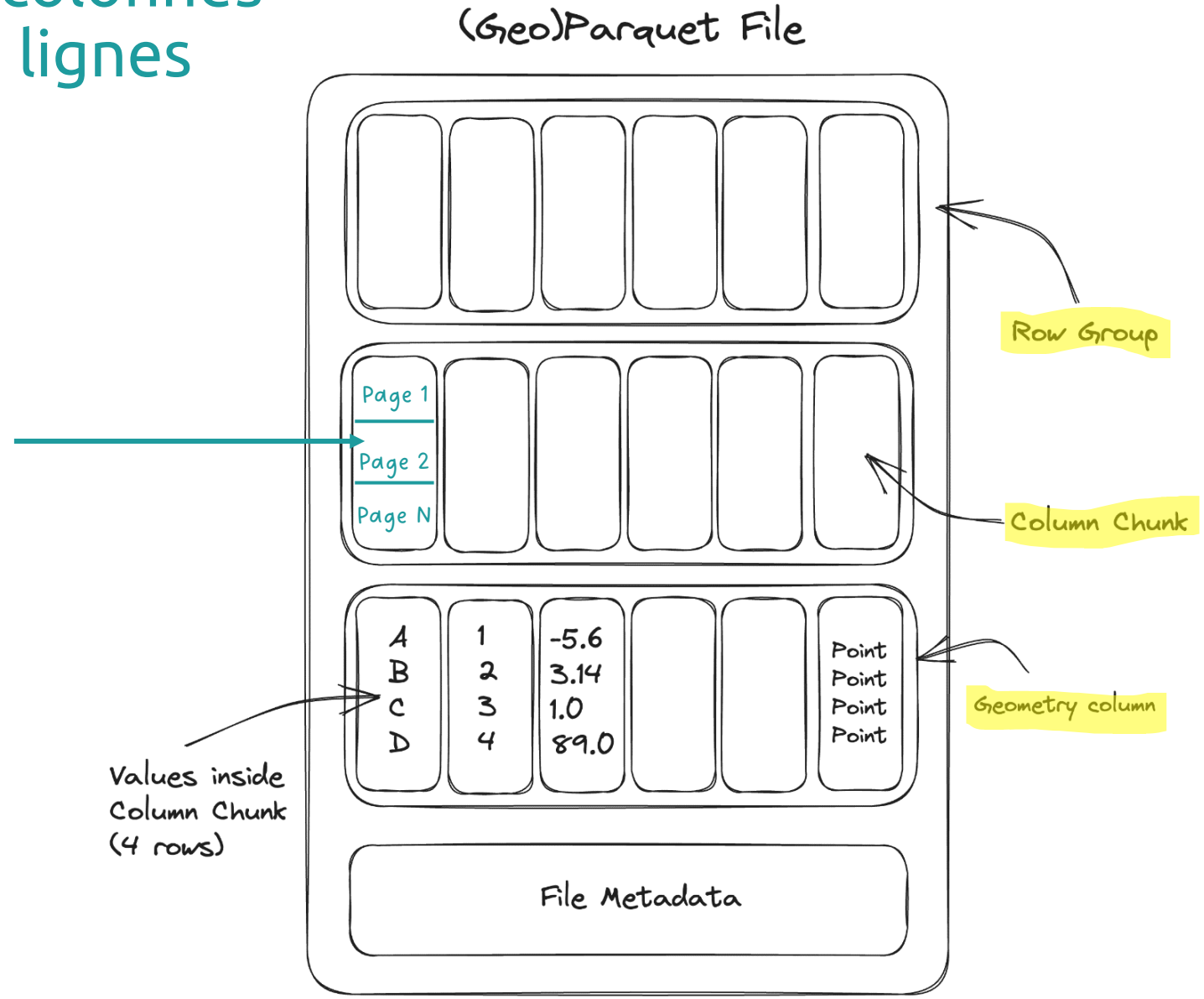
⟫ How to use Parquet

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 ».



» Guidelines for GeoParquet

Lire et écrire du parquet : facile

- DuckDB : FROM / COPY TO
- R : package arrow : read_parquet() / write_parquet(), ou le package duckdb
- Python : pyarrow ou polars ou duckdb

Pour des exemples : voir le tutoriel de l'Insee sur les bases du recensement

Les plus de Parquet

Il est **compressé**, sans que cela nuise à la rapidité de lecture.

Il supporte des **types complexes**, géométriques par exemple,
et des structures hiérarchiques.

Il est conçu pour que les données se chargent en mémoire très rapidement (même objectif que le format **Arrow**, dont les concepteurs sont proches les uns des autres).

On peut le lire et l'interroger avec des **outils gratuits open-source ultralégers** comme DuckDB, et même dans le navigateur.

Parquet bouscule les usages établis

Parquet concurrence les formats plats et les bases de données

CSV, (Geo)JSON ou XML
ne sont plus **les seuls standards** ouverts.

Parquet séduit par sa **compacité** et sa **maniabilité**.

Pour faire de l'analyse, il permet bien souvent
de se passer d'une base de données.

Il est en voie d'adoption par la communauté SIG.
QGIS le supporte.

Il lui manque encore d'être reconnu
par les tableurs courants (Excel, Sheets, Calc).

Parquet devrait remplacer le format CSV

Éric Mauvière - 29 décembre 2022

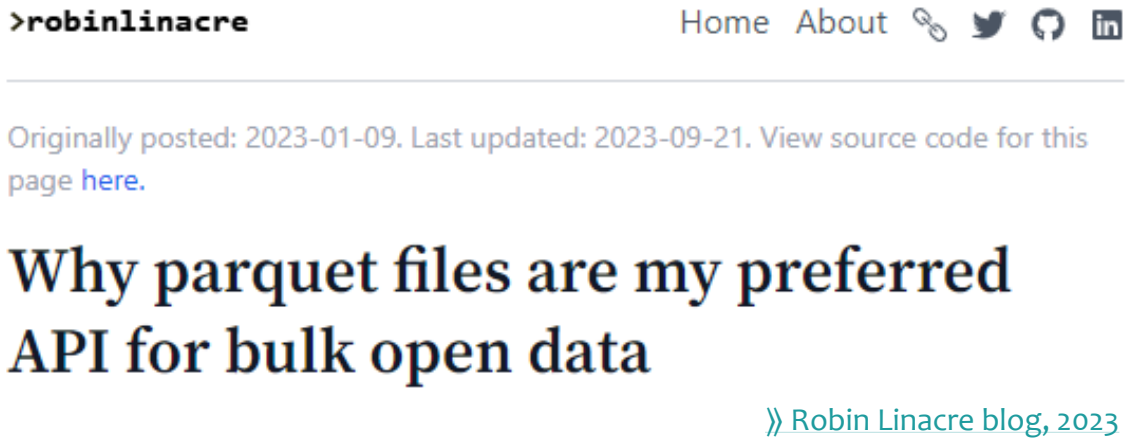
⟫ Blog Icem7, 2022

Parquet concurrence certaines API

Un fichier Parquet bien préparé peut s'interroger **en ligne**.

Le **navigateur** est aussi un requêteur très efficace, et bénéficie d'un **cache**.

SQL est une **API universelle**.



⟫ Robin Linacre blog, 2023

Pour de petites bases (typiquement les nomenclatures), plus besoin de gérer des quotas, des architectures complexes : **économie ++**.

Merci pour votre attention !

Site et blog : https://www.icem7.fr/

Twitter : @ericmauviere

LinkedIn : www.linkedin.com/in/ericmauviere

Merci

Bibliographie

Parquet file format

Guide d'utilisation des données du recensement de la population au format Parquet, Insee 2023

Why parquet files are my preferred API for bulk open data, Robin Linacre, 2023

Parquet devrait remplacer le format CSV, blog icem7, 2022

The birth of Parquet, Julien le Dem, 2024

Fichiers au format Parquet sur data.gouv.fr

OGC to form new GeoParquet standard working group

Quels formats pour quelles données ? Insee, Courrier des statistiques N9, 2023

GeoParquet.org

Utilitaire pour travailler avec GeoParquet : gpq

3 explorations bluffantes avec DuckDB – Croiser les requêtes spatiales, blog icem7, 2023

TadViewer : un utilitaire pour manipuler des fichiers Parquet