Formato Parquet

El formato Parquet es un formato de almacenamiento columnar diseñado para ser eficiente en términos de espacio y rendimiento, especialmente en el contexto de grandes volúmenes de datos y procesamiento distribuido. Aquí tienes un resumen de sus características y un ejemplo real:

Características del Formato Parquet

1. **Almacenamiento Columnar**: Los datos se almacenan por columnas en lugar de filas, lo que permite una lectura más eficiente de datos específicos.
2. **Compresión Eficiente**: Parquet utiliza compresión columnar, lo que reduce significativamente el tamaño del archivo y mejora el rendimiento de las consultas.
3. **Autodescriptivo**: Incluye metadatos que describen el esquema y la estructura de los datos, facilitando su interpretación y uso.
4. **Soporte para Esquemas Complejos**: Puede manejar datos anidados y estructuras complejas, lo que es útil para análisis avanzados.
5. **Compatibilidad**: Es compatible con muchas herramientas de procesamiento de datos como Apache Spark, Apache Hive, y Amazon Athena

Ventajas del Formato Parquet

- **Eficiencia en el Acceso a Datos**: La compresión columnar permite una lectura selectiva eficiente, acelerando las operaciones de filtrado y agregación.
- **Reducción de Costes de Almacenamiento**: Los archivos Parquet ocupan menos espacio en disco gracias a la compresión.
- **Optimización para Consultas**: Ideal para consultas analíticas que requieren acceso a columnas específicas

Representación simplificada de la estructura de un archivo Parquet.

Los archivos Parquet están diseñados para ser eficientes y optimizados para consultas rápidas, por lo que su estructura interna es bastante compleja. Aquí tienes una vista simplificada:

Estructura de un Archivo Parquet

1. **Header**: Contiene información sobre el archivo, como la versión y los metadatos.
2. **Row Groups**: Los datos se dividen en grupos de filas, cada uno de los cuales contiene:
    - **Column Chunks**: Cada columna se almacena por separado dentro de un grupo de filas.
    - **Page Headers**: Metadatos sobre las páginas de datos dentro de cada columna.
    - **Data Pages**: Las páginas de datos reales que contienen los valores de las columnas.

3. **Footer**: Contiene metadatos globales sobre el archivo, como el esquema de la tabla y estadísticas.

Ejemplo Simplificado

```
Header
  - Versión: 1
  - Metadatos: { "creado_por": "Apache Parquet" }

Row Group 1
  - Column Chunk 1: [1, 2, 3, ..., 1000]
  - Column Chunk 2: ["Ana", "Luis", "Marta", ...]
  - Column Chunk 3: [30, 28, 35, ...]

Row Group 2
  - Column Chunk 1: [1001, 1002, 1003, ..., 2000]
  - Column Chunk 2: ["Carlos", "Eva", "Juan", ...]
  - Column Chunk 3: [40, 25, 29, ...]

Footer
  - Esquema: { "columnas": ["id", "nombre", "edad"] }
  - Estadísticas: { "num_filas_totales": 2000, "tamaño_total":
"20MB" }
```

Este es un ejemplo muy simplificado y los archivos Parquet reales contienen mucha más información y están comprimidos para optimizar el almacenamiento y la velocidad de acceso[1][2].

[1]: datos.gob.es

[2]: elmundodelosdatos.com

References

[1] Dominando Apache Spark (VIII): El formato Parquet - El mundo de los datos

[2] ¿Por qué deberías de usar ficheros Parquet si procesas muchos datos?