Análisis Detallado del Formato de Datos Columnar Apache Parquet

Introducción

Apache Parquet es un formato de datos autodescriptivo que incorpora el esquema o estructura dentro de los propios datos. El formato de archivo aprovecha un modelo de trituración y ensamblaje de registros, que se originó en Google. El resultado es un archivo optimizado para el rendimiento de las consultas y la minimización de la E/S.

En concreto, tiene las siguientes características:

- **Apache Parquet está orientado a columnas** y diseñado para proporcionar un almacenamiento eficiente de datos en columnas (bloques, grupos de filas, trozos de columnas...) en comparación con el almacenamiento basado en filas, como CSV.
- **Apache Parquet se construye desde cero** utilizando el algoritmo de trituración y ensamblaje de Google
- Los archivos Parquet se diseñaron pensando en **estructuras de datos anidadas complejas**.
- Apache Parquet está diseñado para soportar **esquemas de compresión y codificación** muy eficientes (véase Google Snappy).
- **Apache Parquet permite reducir los costes de almacenamiento** de los archivos de datos y maximiza la eficacia de la consulta de datos con tecnologías sin servidor como Amazon Athena, Redshift Spectrum, BigQuery y Azure Data Lakes.
- Con licencia de la fundación de software Apache y disponible para cualquier proyecto.
- Admite tipos de datos conocidos, metadatos de archivos y codificación automática de diccionarios.

Aplicaciones en Google y Amazon

Empresas líderes como Google y Amazon aprovechan las ventajas de Parquet para optimizar la gestión de datos a gran escala. Google utiliza Parquet en conjunto con herramientas como BigQuery para consultas analíticas eficientes. Amazon, por su parte, integra Parquet en servicios como Amazon S3 y Amazon Redshift, mejorando la velocidad y eficacia de las operaciones de análisis de datos.

Beneficios y Desafíos

Entre los beneficios clave de Parquet se encuentran su eficiencia en el almacenamiento y procesamiento, así como su interoperabilidad con diversas herramientas y plataformas. Sin embargo, algunos desafíos pueden surgir en cuanto a la complejidad de su implementación y la necesidad de gestionar adecuadamente los esquemas de datos. A pesar de estos desafíos, Parquet sigue siendo una opción poderosa para optimizar el manejo de grandes conjuntos de datos en entornos de big data.

Bibliografía

- https://parquet.apache.org
- https://parquet.apache.org
- https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-parquet?hl=es-419