Formato ORC

El formato ORC (Optimized Row Columnar) es un formato de almacenamiento columnar utilizado principalmente en el ecosistema de Hadoop. Fue desarrollado originalmente por Apache Hive en 2013 para mejorar la eficiencia del almacenamiento y acelerar las consultas en Hive

- **Almacenamiento Columnar**: Los datos se almacenan en columnas, lo que permite lecturas y compresiones más eficientes.
- **Autodescriptivo**: Cada archivo ORC contiene metadatos que describen su contenido, facilitando la lectura y el procesamiento de los datos.
- **Optimización para Lecturas en Streaming**: Diseñado para grandes lecturas en streaming, lo que mejora el rendimiento de las consultas.
- **Compatibilidad**: Es compatible con varias herramientas de procesamiento de datos como Apache Spark, Apache Flink y Apache Hadoop

Estructura

1. **Header**: Contiene información sobre el archivo, como la versión y los metadatos.
2. **Stripe**: Los datos se dividen en "stripes" (franjas), cada una de las cuales contiene:
    - **Index**: Índices para las filas en el stripe.
    - **Data**: Los datos reales, almacenados en columnas.
    - **Footer**: Metadatos sobre el stripe, como el número de filas y la posición de los datos.
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 Hive" }

Stripe 1
  - Index: { "fila_inicio": 0, "fila_fin": 999 }
  - Data:
    - Columna 1: [1, 2, 3, ..., 1000]
    - Columna 2: ["Ana", "Luis", "Marta", ...]
    - Columna 3: [30, 28, 35, ...]
  - Footer: { "num_filas": 1000, "tamaño": "10MB" }

Stripe 2
  - Index: { "fila_inicio": 1000, "fila_fin": 1999 }
  - Data:
    - Columna 1: [1001, 1002, 1003, ..., 2000]
    - Columna 2: ["Carlos", "Eva", "Juan", ...]
    - Columna 3: [40, 25, 29, ...]
  - Footer: { "num_filas": 1000, "tamaño": "10MB" }

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