Different Types of File Formats and Their Comparison

CSV, JSON, Avro, ORC, Parquet

Overview of File Formats

- File formats are used to store and exchange data. Different formats have different features and use cases.

- In this presentation, we will compare the following formats:
- - CSV
- - JSON
- - Avro
- - ORC
- - Parquet

CSV (Comma-Separated Values)

- Features:
- - Plain text format
- - Easy to read and write
- - No support for complex data types
- - Larger file size compared to binary formats

- Use Case: Simple data storage and exchange, especially for tabular data.

JSON (JavaScript Object Notation)

- Features:
- - Text-based format
- - Supports nested structures and complex data types
- - Human-readable and easy to parse
- - Less efficient in storage compared to binary formats

- Use Case: Web APIs, configuration files, and data interchange between systems.

Avro

- Features:
- - Binary format with schema support
- - Compact and efficient
- - Supports rich data types and schema evolution
- - Not human-readable

- Use Case: Data serialization in Hadoop, data exchange in distributed systems.

ORC (Optimized Row Columnar)

- Features:
- - Columnar storage format
- - Highly optimized for read performance
- - Supports complex types and compression
- - Best suited for Hive and Hadoop workloads

- Use Case: Big data processing, especially in Hive for fast query performance.

Parquet

- Features:
- - Columnar storage format
- - Efficient for both storage and query performance
- - Supports nested data structures and compression
- - Widely used in Hadoop, Spark, and other big data ecosystems
- Use Case: Optimized for analytical workloads in big data environments.

Format	Type	Storage Efficiency	Schema Support	Use Case	Read/Write Performance
CSV	Text	Low	No	Simple data exchange	Fast Read/Write
JSON	Text	Medium	Yes	Data interchange	Medium Read/Write
Avro	Binary	High	Yes	Data serialization	Efficient Read/Write
ORC	Binary	High	Yes	Big data processing	Optimized Read
Parquet	Binary	High	Yes	Analytical workloads	Optimized Read/Write

Conclusion

- Each file format has its own strengths and use cases.
- Choose the right format based on your specific requirements:
- - CSV and JSON for simple and human-readable data exchange.
- - Avro, ORC, and Parquet for efficient storage and processing in big data environments.