Saturday, February 6, 2021 11:39 AM

Parquet

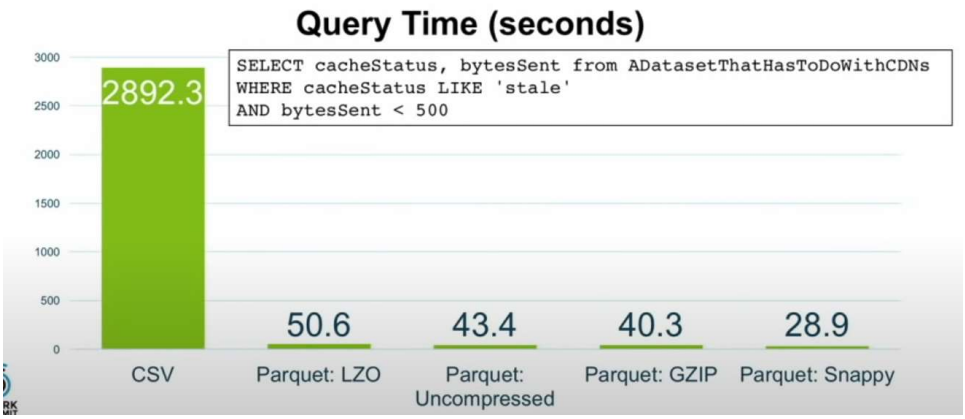
Columnar storage format available to any project in the Hadoop ecosystem, regardless of the choice of data processing framework, data model or programming language.

- Binary format
- API for JVM/Hadoop, C++
- Columnar
- Encoded
- Compressed
- Machine-friendly

Options For Multi-PB Data Lake Storage

	Files	Compressed Files	Databases
Usability	Great!	Great!	OK to BAD (not as easy as a file!)
Administration	None!	None!	LOTS
Spark Integration	Great!	Great!	Varies
Resource Efficiency	BAD (Big storage, heavy I/O)	OK… (Less storage)	BAD (Requires storage AND CPU)
Scalability	Good-ish	Good-ish	BAD (For multi-petabyte!)
CO$$$$T	OK…	OK…	TERRIBLE
QUERY TIME	TERRIBLE	BAD	Good!

CSV vs. Parquet Column Selection Query



CSV vs. Parquet Table Scan Query