ODSC East 2023 Tutorial: Introduction to Apache Arrow and Apache Parquet, using python and pyarrow

Abstract:

Apache Arrow https://arrow.apache.org/ is a language-independent columnar memory format for flat and hierarchical data, organized for efficient analytic operations on modern hardware like CPUs and GPUs. The Arrow memory format supports zero-copy reads for lightning-fast data access without serialization overhead.

After completing this workshop, you will understand the basics of Apache Arrow and Apache Parquet, how to load data to/from pyarrow arrays, csv and parquet files, and how to use pyarrow to quickly perform analytic operations such as filtering, aggregation, joining and sorting. In addition, you will also experience the benefits of the open Arrow ecosystem and see how Arrow allows fast and efficient interoperability with pandas, pol.rs, DataFusion, DuckDB and other technologies that support the Arrow memory format.

Prerequisites

Please have the following installed on your computer:

1. Python3: https://www.python.org/downloads/
2. The following python packages: pyarrow pandas datafusion ipython duckdb

For example, install via

```
pip install pyarrow pandas datafusion ipython duckdb
```