UC SANTA CRUZ
Baskin Engineering

Capstone Project

A Service to Parquet

Phu Le, Xiaobin Wu, Yibo Guo

ORACLE

Abstract

This project is intended to provide a cloud service that provides a variety of tools that can aid Oracle clients in storing and processing their data. Switching to Parquet is made easier with the file conversion library and data processing tools like filtering, searching, and sorting can transform that same data and store it on the cloud. Users can access this service on the Oracle cloud and will easily be able to streamline their data analytics process.

Approach

Service: The core of the service provides a way for a client to communicate with the server. This was accomplished with Helidon, an Oracle open-sourced library.

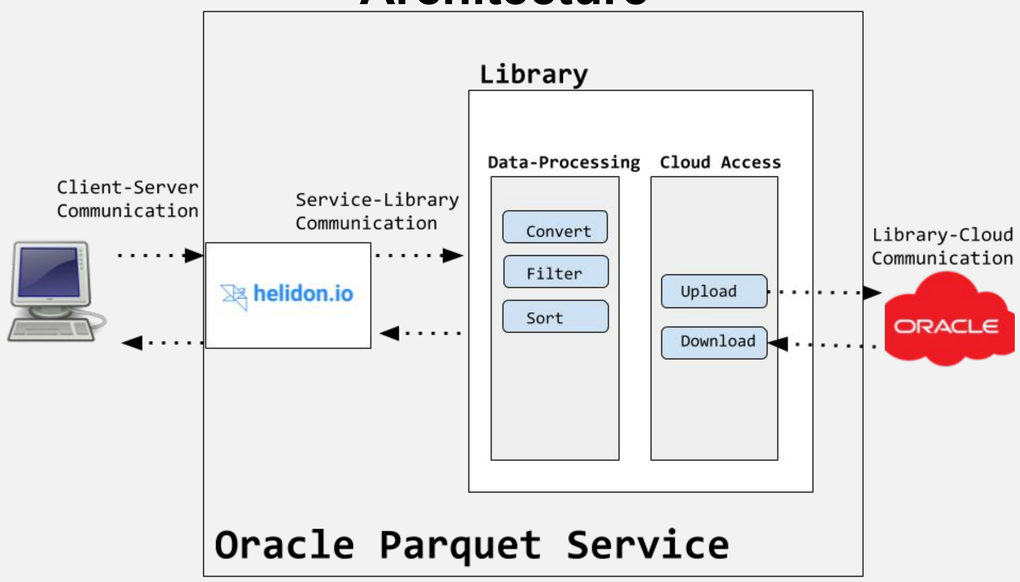
Library: After communicating with the server, the Helidon service calls on the library portion that provides all the functionality, including data-processing operations and cloud access. This required using the Apache Parquet library and the Oracle Cloud Infastructure SDK.

Overview

As data sets increase, storage and data analytics require optimizations. When Oracle customers store data, analyzing the data involves extracting, transforming, and loading the data, called the ETL process. While this process is costly and inefficient, Oracle's solution revolves around a file format called Apache Parquet. Designed as a columnar storage, Parquet files can offer efficient and complex data processing. However, current solutions handling Parquet are not-so-lightweight. Oracle hopes to create a cloud service for Parquet that is both lightweight and easy-to-use in order to streamline the ETL process.

Dataset	Size on Amazon S3	Query Run time	Data Scanned	Cost
Data stored as CSV files	1 TB	236 seconds	1.15 TB	$5.75
Data stored in Apache Parquet format*	130 GB	6.78 seconds	2.51 GB	$0.01
Savings / Speedup	87% less with Parquet	34x faster	99% less data scanned	99.7% savings

Architecture



Acknowledgments

Daniel Langerenken – Vishal Vaddadhi – David Hernandez

Results

Clients who use our service have access to a variety of tools, including file conversion, row filtering, column filtering, and cloud access. Compared to current solutions, our service makes data processing simpler. With our time constraints, we used Apache Drill as a shortcut to implement some functionality. Additionally, we did not reach our intial goal of deploying our service on the cloud.

Benchmark

Benchmarking our native solution with the Drill implementation, we found that our native solution performed better than Drill's. We also see that that improved performance is more noticeable for smaller files.