Capstone Project

A Service to Parquet

Phu Le, Xiaobin Wu, Yibo Guo

Abstract

The project is intended to provide data services on parquet format storage. The intention is to support developers with easier access to a smaller sized database format. The Parquet is well-known for its lightweight attribution. However, there are not too many easy-to-use SDKs to develop with Parquet. Compare with existing tools, this project offers a set of more accessible Apache Parquet development APIs and therefore gives users more flexibility in taking advantage of this optimized data storing technology.

Approach

Inter-unit Transmission (IUT):
REST API by Oracle Helidon
Data Service Unit (DSU):
Powered by Apache Drill
Storing and Virtualization (S/V):
Oracle Cloud and Amazon AWS

Pipeline:

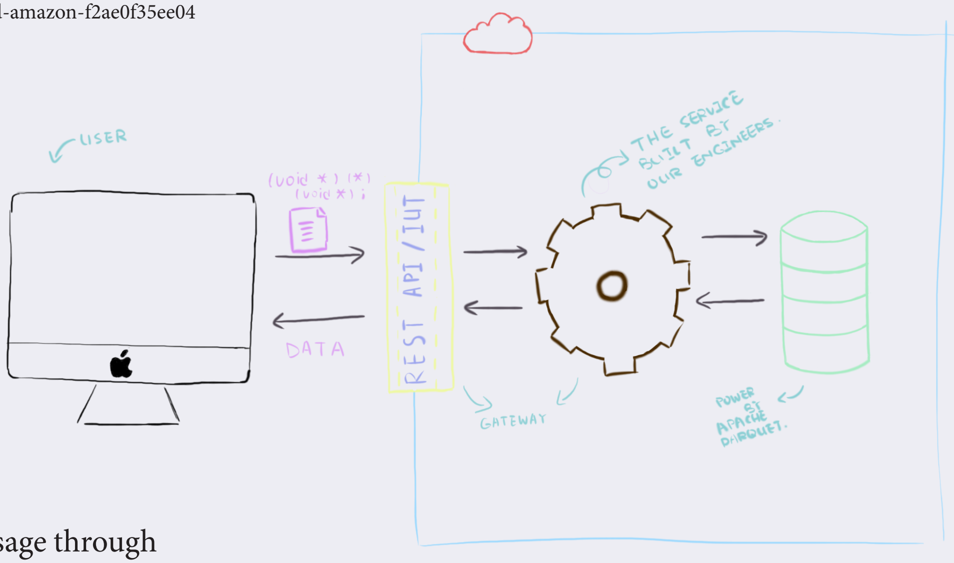
To request service, users need to pass a message through the REST API. The DSU will serve the request after a validation. The service may is a database initiate, filter, search, add, fetch, or format conversion. The service performed is corresponding to the user's request. Unless the user requested to erase it permanently, all changes and modified databases would be stored on the S/V unit.

Overview

Nowadays the storage is dramatically cheaper than years ago. A 4TB disk is as cheap as 320 bucks [1]. However, the volume of data people produced is also growing with the size of storage, which makes data memory is not as cheap as people think. The majority of the current data storage solutions, such as CSV, TSV, and PSV formats, are still not storing data space-sufficient enough. The Parquet storing format is one of those disk-sufficient solutions, which can save approximately 87% of storage space, which is about 99.7% saving on money[2]. This project is conducted under the supervision of Oracle to enables their ability to serve more customers with the current data clusters.

[1] Disk price: Refering price on Amazon as of 5/7/2020

[2] Original data see OpenBridge: https://blog.openbridge.com/how-to-be-a-hero-with-powerful-parquet-google-and-amazon-f2ae0f35ee04



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

Result and Conclution

Result and Conclution TBA Result and Conclution TBA

Credit

Thanks Daniel Langerenken for the sponsorship from Oracle
Thanks ex-member Vishal Vaddadhi and David Hernandez for their effort