Technical Case Study

Optimal Architectures and Methods for AI Pattern Mining Algorithms

January 2022

DASK | kuberre systems

Sponsored By Coiled

The Client

An emerging hedge fund built upon quantitative investment strategies that requires a robust data science infrastructure to identify and execute portfolio strategies. More specifically, the firm is running and improving upon a pattern mining algorithm. The firm had invested in an a massively parallel “on-premise” hardware/software platform to support this effort and is now looking for alternatives that address the problems defined below.

Our Role

Kuberre was brought in to optimize data science operations by evaluating existing platforms, data sources and processes. Our mandate was to identify unnecessary costs, address system/usability limitations and deliver improvements that leave the operation in a future-ready state.

Problem Definition

After a full evaluation of the costs, processes, strengths, limitations and usability of the current data science infrastructure it was determined that a change was required. The challenge was to find an optimal alternative to the existing, massively parallel system currently supporting a pattern matching algorithm. The current IBM NPS was selected because it offered the advantage of having the data processed as close to computing as possible. This was especially helpful as the algorithm has many sequential steps with interdependencies as well as the need to leverage temporary tables that are created/used in subsequent steps. While powerful, the existing platform presents the following challenges that needed to be solved for:

- The hardware/software itself is aging and imposing limitations in the form of rigid restrictions regarding the software and processes that can be run on the hardware.
- Annual maintenance and support fees were too costly
- Fully leveraging the power of the platform requires special technical skillsets that are increasingly difficult to source in the current talent market. A move to a 100% Python approach is required… enabling the use of popular open-source libraries including scikit learn, XGBoost and Pandas while making talent onboarding far easier.
- Database choice limitations need to be removed because they prevent the team from isolating themselves from database dependencies
- The data science team wants to abstract SQL calls as a means to avoid creating temporary tables during algorithmic runs without sacrificing control of database management, production vs. research resource conflicts, etc.

Options Evaluated

The Kuberre team partnered with the clients' data science team to assemble an exhaustive list of options and concepts that may address the challenges outlined in the Problem Definition. Of those considered, these three approaches rose to the short list as most feasible.

Snowflake UDTF

A must-look framework as the data is close to compute power. While initially attractive, this approach was ruled out because it currently supports only JavaScript or Java (with some constraints) UDTF's.

Apache Spark

An obvious route given its support for Python based UDF's. However, this option had to be dismissed because it still left us with three problems that could not be solved for, including:

- How would we optimize the creation of temporary tables?
- How would we abstract the SQL from the end users effectively?

- How would we integrate native Python libraries such as scikit learn, SGBoost and NumPy arrays easily to make the transition from existing implementations faster?

Dask UDTF's

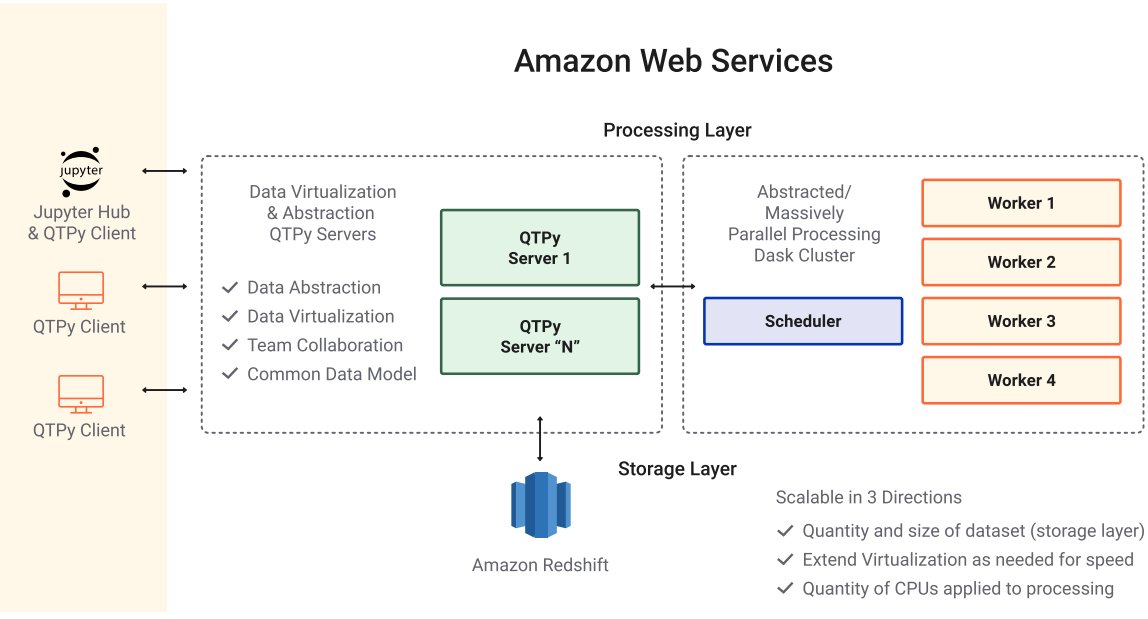
Of the three leading options considered, Dask emerged as the most likely approach as it is implemented in Python and can support all of the native Python libraries including Pandas, scikit learn, etc. If this choice was to hold up as a clear winner, we would still have to satisfy two critical requirements…

- How do we optimize creation of temporary tables?
- How do we abstract SQL from the end users?

Overview of the Solution Delivered

The team selected the Dask approach because it satisfied the requirement of a being capable of massively parallel processing and the use of any customer UDTF's in Python. Kuberre's QTPy product was implemented to enhance the Dask framework to limit instances in which data would travel back and forth from the database layer and to meet SQL abstraction requirement.

Pattern Mining Using QTPy + Dask on AWS



Retired Infrastructure

"On-Premise" massively parallel hardware with data as close as possible to compute.

Dedicated Vendor Hardware

- 126 physical CPUs
- 1.45TB Memory

Simulations per week = 10
Cost per run = ~$290.00/run
Annual cost to maintain device ~$150,000.00 from vendor
Internal cost to operate and maintain

New Infrastructure with "Fully Optimized Performance" ***(97% Reduction in Cost)***

AWS Environment

- 7- c5.24xlarge (192GB/96 vCPUs) providing 672 simultaneous processes
- 1 – c5.9xlarge (72GB/36 vCPUs) orchestration server
- Kuberre QTPy Software

Simulations per week = 10
Cost per run = ~$4.60/run
Annual cost = ~$2,400.00/year

How did Kuberre QTPy Enable the Improvements?

Quantitative Techniques in Python, or QTPy (pronounced Cutie-Pie) is unique in that it can be implemented to leverage the kind of cloud computing configurations outlined above with immense freedom when it comes to the programing libraries used. In this case, QTPy was used to:

- Abstract users from the SQL layer through a common data model with a simple set of APIs used to abstract sources (type of databases, APIs, scripting, etc.)
- Openness to importing additional programing libraries in Python. In this case scikit learn, SGBoost, NumPy and Pandas.
- Customizable virtualization to have datasets available with compute in memory as desired to increase performance
- Store, share and improve upon prior algorithm code in future work as a team
- Mapping of Issuer, Instrument and Tradable Instrument identifiers across markets, vendors and benchmarks making the data set far easier to leverage
- Freedom to be deployed on-premises or in any popular cloud service provider environments

Customer Satisfaction

> "Kuberre is our trusted technology partner – they help transform our vision into reality by combining their cutting-edge software solutions, including QTPy, with open-source tools such as Dask, to deliver ideal solutions. This partnership let us build an enhanced infrastructure on AWS that supports our vision with its ability to scale, adapt to changes easily, and empower our team to continue innovating. It's nice to focus on the desired outcomes rather than how something is done or who can do it."
>
> Howard Getson, CEO Capitalogix

More about QTPy

QPTY is Kuberre's Data Science Optimization technology and it serves as a critical part of the company's Enterprise Data Management story. While our EDM delivers a comprehensive solution for security mastering, data lineage, reporting, accessibility, process management, data normalization and controls for currencies/corporate actions, QTPy adds additional value to data science and AI teams in the form of capabilities including:

Abstraction of data structures, databases, sources and other variations that make accessing and using data sets challenging

Virtualization of data sets when their size or complexity make traditional approaches too time consuming and clumsy

Teamwork Enhancement improvements through re-use and sharing of prior models and code

Mapping of Complex Identifiers and Relationships across vendors/markets accounting for issuer, instrument and tradable instrument levels of complexity

To learn more about what Kuberre Systems can do for your data infrastructure please visit our website at www.kuberresystems.com , email us at info@kuberresystems.com and we will work with you to layout the optimum solution for your business.

About Kuberre

Kuberre Systems has been servicing capital markets firms for 20 years with a complete set of Enterprise Data Management, Data as a Service and Data Science/Analytics optimization solutions. As a team of data science/enterprise data management practitioners and visionaries, Kuberre delivers clients a fully modernized approach to managing and enabling data intensive operations with superior care for data governance, operational simplicity and flexible delivery/support options. Kuberre is unique in its ability to understand and solve the real-time challenges facing capital market firms with solutions architected for today's realities. For more information about Kuberre Systems and to learn more about our latest successes, please visit our website, our LinkedIn page or via email.

About Dask

Dask is an open source Python project that is a general purpose distributed computing framework for large scale data workloads. Dask is the #1 big data Pyhon-native platform for distributed computing. Learn more about Dask by visiting dask.org

About Coiled

Coiled scales Python to the cloud for data professionals. Based on Dask, the leading Python-native solution for distributed computing, Coiled has hosted more than 2.7B+ tasks for data professionals, scientists, and researchers around the globe including Capital One, Anthem Health, and the Air Force to solve challenges in business, research, and science. Coiled is a remote-first company with the best and brightest working from around the globe. Founded by the initial author of Dask, Coiled is on a mission to create a platform that gives Data Scientists the power of the cloud and machine learning, freeing them from today's limitations so they can solve important problems. Learn more about Coiled by visiting Coiled.io