

CASE STUDY

Cloud
Data Center



Taking on the World

Descartes Labs is building a digital twin of the world, combining historical and current satellite data. Its technology runs on the Intel® Xeon® Scalable processor, which accelerates the compression of huge data volumes

At a Glance:

- Upgrading to the Intel® Xeon® Scalable processor enabled Descartes Labs to use Intel® Advanced Vector Extensions 512 (Intel® AVX-512)
- By creating a scalable cloud architecture based on Intel® processors, Descartes Lab has been able to build a business with customers in the agriculture, utilities and financial services sectors



Google Cloud Platform



Descartes
Labs

Descartes Labs helps companies to get business insight from huge volumes of satellite and geographic data, using a combination of Software as a Service and custom development. Handling petabytes of data, compression is hugely important for packaging the data in usefully sized files and for driving down storage costs. By upgrading to the latest generation Intel® processor, provided in the Google Cloud Platform*, Descartes Labs was able to accelerate its compression.

Challenge

- Enable the storage and processing of petabytes of satellite and geographic data
- Create an architecture that scales across storage, compute and networking so it can ingest huge volumes of data regularly as data volumes increase
- Drive down the cost of the architecture

Solution

- Descartes Labs chose Google Cloud Platform* for its linear scalability across storage, compute and networking
- The company uses preemptible VMs to drive down costs
- The 96-core Intel® Xeon® Scalable processor is used to deliver the performance required, including for compression
- Intel® VTune™ Amplifier is used to help identify performance bottlenecks and fine-tune the code

Results

- Descartes now counts Cargill and DARPA among its customers, which also include businesses in the agriculture, financial services and utilities industries

Making Sense of Huge Volumes of Satellite Data

Over the last few decades, satellites have shrunk dramatically. Whereas early satellites were the size of a small bus and weighed a ton, today's CubeSats are closer in size to a smartphone, weighing no more than 1.3kg per unit. Costs have dropped from around USD 100 million to around USD 65,000¹. The commercial

space industry is working hard to find more affordable ways to launch and recover rockets too, further driving down the cost of launching satellites and acquiring data from space. Within five years, the boom in private satellites could be giving us continuous updates covering the whole planet, every 20 minutes.

For businesses, more timely satellite data represents a unique opportunity to understand and forecast change, both environmental and economic. For example: Infrastructure can be measured as it is built; crop yields can be predicted based on imagery of farmland worldwide; and solar capacity can be measured to inform decisions in the energy industry.

Making sense of the huge volumes of satellite data is a big challenge, though. The Landsat 8 satellite captures 3.1 trillion pixels per color band (red, green, blue), totaling 70 trillion pixels between 2013 and 2017. That's 320 terabytes of data, captured by just one satellite¹. For a more complete picture, data from different satellites can be combined, but that presents challenges of its own because the data is unlikely to be consistently aligned and formatted.

Solution Details

Descartes Labs is building a digital twin of the world by applying machine learning to satellite imagery and other massive data sets, such as weather data, pricing and customer data. The solution is based in the cloud, which means it can scale storage for the massive data sets, and

scale compute capability to enable analysis results and data to be returned more quickly.

The Descartes Labs data refinery offers geographic data including the entire library of satellite data from the NASA Landsat and ESA Sentinel missions, the entire Airbus OneAtlas* catalog, and NOAA's Global Surface Summary of the Day weather dataset. The data has been combined and cleaned, so it is ready for machine learning analysis.

Customers with experience of machine learning can build their own applications and access Descartes Labs' data using an application programming interface (API). Data available includes imagery and vector data describing features such as county boundaries. Using a short Python program, it's possible to build applications that scale to thousands of processor cores in the cloud, enabling the huge volumes of data to be processed quickly. Customers can request geographic data covering a particular region and time period and receive the data back as imagery or a CSV file suitable for analysis in a spreadsheet.

Customers without the experience to write their own solutions can work with the team at Descartes Labs, who can combine customer data sets with Descartes Labs' own geographic data, then build a machine learning model, and execute it on a subscription basis, with new data being continuously added.

"We were extremely impressed with how GCP scaled linearly across multiple components, not just compute, but how well the network, cloud storage, and Google Cloud PubSub* [used for messaging] all scaled linearly," said Tim Kelton, co-founder

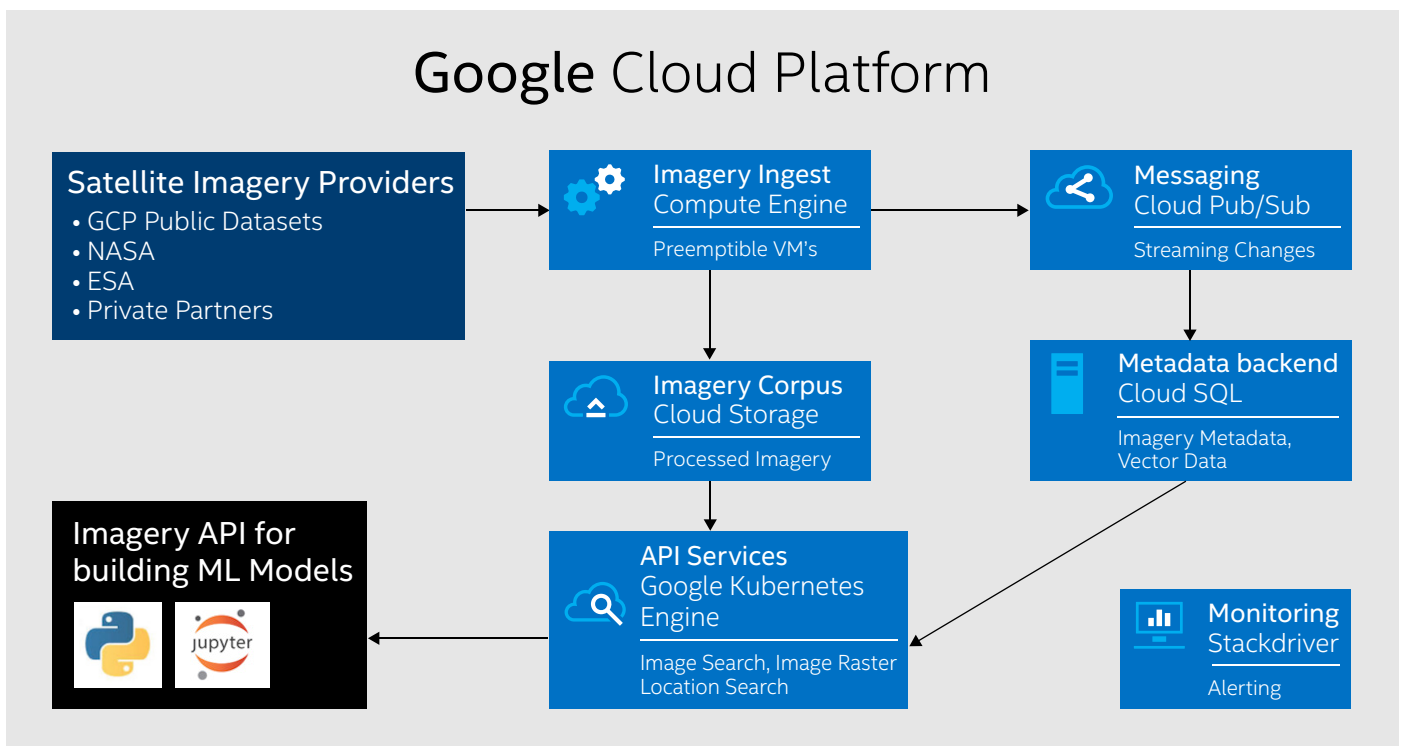


Figure 1. Descartes Labs ingests satellite data from multiple sources, and writes vector data into a database as images are analyzed.

Technical Components of Solution

- **Google Cloud Platform***. To store the huge volumes of data it handles and to enable highly scalable compute capabilities, Descartes Labs uses Google Cloud Platform for both compute and storage
- **Intel® Xeon® Scalable processor**. The latest generation Intel® processor increases performance, compared to the Intel® Xeon® processor E5 v3 family, which the company was using previously. In particular, the introduction of Intel® Advanced Vector Extensions 512 (Intel® AVX-512) accelerated compression operations, which are essential for optimizing storage costs and packaging data in usefully sized volumes

and head of cloud operations at Descartes Labs. “When we began, we were just a few people above a pizza shop in New Mexico, with no physical servers. One of the first things we did was to clean and calibrate 43 years of satellite imagery from NASA, and using GCP we scaled that to 30,000 cores in the cloud.”

Those 43 years of NASA imagery amounted to 1 petabyte. Processing that volume of data could be a weekly requirement within five years, Descartes Labs estimates, so its use of historical data is not only important for analyzing changes over time but also for testing the scalability of the cloud environment.

Descartes Labs uses preemptible virtual machines (VMs), which are VMs that Google may withdraw at any time, and which will be available for no more than 24 hours. They are offered at a substantial discount, and have helped Descartes Labs to drive down its costs. The processing pipeline is an embarrassingly parallel problem, which means it can be easily divided up and distributed across multiple cores. Descartes Labs uses a Python queue called Celery* to manage tasks, and ensure they are all completed. Redis Stackdriver* is used for monitoring. Both the queuing and monitoring applications run on non-preemptible VMs to ensure continuity across the application.

As images are analyzed, information is captured and written to a PostGIS database for geospatial queries, using Google Cloud Pub/Sub for messaging. Google Kubernetes Engine* is used for managing and isolating the workloads of different customers.

Intel® Processors Power the Cloud

The Descartes Labs solution runs today on the 96-core Intel® Xeon® Scalable processor, provided through Google Cloud Platform. The Intel Xeon Scalable processor introduces Intel® Advanced Vector Extensions 512 (Intel® AVX-512), doubling the amount of data that can be processed simultaneously

using a single instruction, compared to the previous generation processor. “We chose the Intel Xeon Scalable platform for its performance,” said Kelton. “We found that we could recompile our code without needing to make any code changes to take advantage of Intel AVX-512.”

The Google Compute Engine running on the Intel Xeon Scalable processor is used to ingest the processing pipelines, where compression is one of the requirements, and for the Software as a Service platform where models are executed against imagery (which requires imagery expansion). The software is Descartes Labs' proprietary stack, written in C, C++ and Python. Customers executing models on the platform often use libraries from the Python machine learning stack such as Numpy, SciPy, SciKit-Learn, TensorFlow and Keras.

Given the data volumes Descartes Labs is working with, compression is essential to minimize storage cost and to deliver data in usefully sized files. A satellite might capture 15 bands of light, for example, but a particular use case might only require the infrared band. The solution needs to be able to provide just the data required, in a compressed file for ease of use.

The machine learning models can require 1000 iterations to train. “We see improved performance on the Intel Xeon Scalable processor, compared to the Intel Xeon processor E5 v3 family we had used previously,” said Kelton. “I love it when I can get an answer faster, or reduce my billed processing time. That’s pretty amazing! I’ll take either one of those!”

While most of the company’s developers work at the level of the algorithm, coding in C, C++ and Python, one of the engineers is engaged in performance tuning. “We used Intel® VTune™ Amplifier to help optimize the early stages of image preprocessing,” said Kelton. “It helped us to see where our code was spending too much time on a particular operation, so we could debug and fine-tune the details that we couldn’t see in a regular integrated development environment (IDE). Intel makes some of the best tools because they understand the back end architecture and what’s going on in the processor.”

Intel has helped Descartes Labs with advice on isolating workloads in a multitenant environment, and Descartes Labs is exploring the open source project Kata Containers for container security, which Intel contributed to, and the Intel® Distribution for Python, which is tuned to optimize performance on Intel® processors.

Winning New Business

Descartes has secured new business from customers in the agriculture, energy, and financial services sectors, among others. “Previously, one company might own 70 percent of production, transportation, and the supply chain for a particular commodity,” said Kelton. “They could trade in the market with greater insight than anyone else. Now, using

satellite imagery, there's more transparency there. We're starting to see more opportunities for disruption."

For the grain trader Cargill, Descartes Labs combined Cargill's data sets with their own to create a model that improved on both companies' previous models for forecasting corn production in the United States.

The Defense Advanced Research Projects Agency (DARPA) in the US has commissioned Descartes Labs to build cloud infrastructure for its Geospatial Cloud Analytics program, which will integrate up to 75 different types of data. For that phase, Descartes Labs will help organizations to build sample projects on top of the new infrastructure. Potential

Spotlight on Descartes Labs

Founded by a team from Los Alamos National Laboratory in 2014, Descartes Labs is building a digital twin of the world. Through its API and its custom services, it helps companies to use huge volumes of geographic data to inform business decisions. Its customers include Cargill and DARPA, and come from sectors including agriculture, financial services and utilities.

www.descarteslabs.com

Learn More

- [Intel® Xeon® Scalable processors](#)
- [Intel® VTune™ Amplifier](#)
- [Google Cloud Platform*](#)
- [How tech startup Descartes Labs is helping Cargill unlock the future of data](#)

applications include detecting illegal fishing and monitoring the construction of fracking sites.

Lessons Learned

- By building a close relationship with its cloud provider, Descartes Labs has had an opportunity to get early access to technologies, including the Intel® Xeon® Scalable processor, and a chance to help shape Google's own innovations
- Upgrading to the Intel Xeon Scalable processor and recompiling software to take advantage of new processor features can deliver significant performance improvements, depending on the workload
- The use of preemptible VMs can drive down costs significantly. The processing pipeline used in Descartes Labs' workloads can be easily distributed across VMs, and the company has built a queue system to account for the possibility that a VM will be withdrawn at short notice

Find the solution that is right for your organization. Contact your Intel representative or visit intel.com/cloud



¹ How Computers See the Earth: A ML Approach to Understanding Satellite Imagery (Cloud Next '18), Kyle Story, Descartes Labs. <https://www.youtube.com/watch?v=5PNnPagENxQ>

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com/CSP

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Intel, the Intel logo, and Xeon are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.