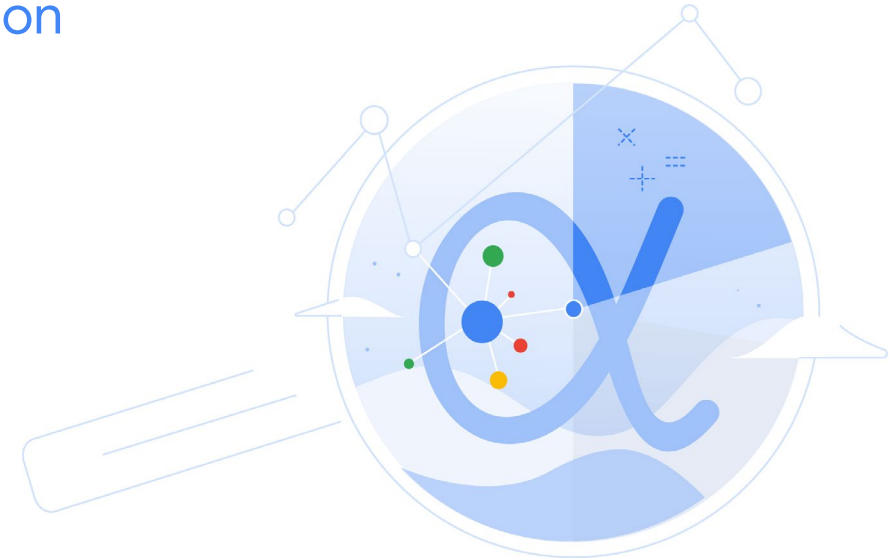
Generating Alpha with Google Cloud

Six Google Cloud tools for signal generation



Google Cloud

Recent years have been challenging for the hedge fund industry. According to the Financial Times, "investors pulled $43bn from hedge funds in 2019[1]" and "more than 4,000 funds were liquidated since 2015[2]," Bloomberg reports. The key driver of this setback was market performance. The S&P 500 returned 31.5% in 2019[3], resulting in a shift from active to passive investment strategies with lower fees and typically higher returns.

Historically, active fund managers have performed in-depth fundamental analysis of company financials leveraging metrics such as P/E ratio, P/EG ratio, earnings per share, and dividends to determine a company's intrinsic value.

Over time, the market has become far more efficient and investment professionals have looked to alternative data to determine a company's value more stringently. This has led to an arms race for alternative data sources, massive computing power to process that data, and Artificial Intelligence (AI) to understand it and predict market behavior. The buy-side is procuring thousands of data sets from alternative data providers, scraping millions of web pages, and turning to Google Cloud to crunch this information. At the same time, sell-side data has exploded with thousands of pieces of research appearing daily – while these investment professionals struggle to interpret these myriads of data sources.

Types of alternative data may include flight trackers, social media posts, credit card transactions, satellite images (crops, slag piles at mines, etc.), foot traffic coordinates, online communities, product pricing, geospatial data, and many more.

Traditionally, firms have leveraged custom-built, self-managed data platforms with custom web scrapers to obtain this data. Yet there's a problem with this approach. The majority of time is spent building and maintaining the platform and applications for the data acquisition, and cleaning the data; only a small portion is focused on the most valuable piece – finding useful insights.

Fortunately, Google Cloud Platform (GCP) has powerful tools that enable engineers, quants, and data scientists to focus on what matters most, driving signal generation and ultimately generating alpha.

4x

Number of alternative data analysts over the last five years.[4]

77%

Of buy-side firms are seeking to or are already using alternative data to inform investment process and strategies.[5]

$901M

Spend on alternative data sets by 2021 – a growth of 19.2% every year.[5]

This has led to an arms race for alternative data sources, massive computing power to process that data, and Artificial Intelligence to understand it.

1 https://www.ft.com/content/55b193a3-2806-4271-9a21-9fe8703d599c
2 https://www.bloomberg.com/news/articles/2019-12-30/hedge-fund-purge-deepens-as-3-trillion-market-retrenches
3 https://seekingalpha.com/article/4322741-s-and-p-500-earnings-growth-in-uptrend
4 Deloitte, "Alternative Data Adoption in Investing and Finance", 2018
5 Aite Group, "Alternative Data in Active Asset Management: A New Source of Alpha?", August 2018

1

Natural Language Processing & Vision APIs

The ability to hear, see, and understand are some of AI's core components. Google has heavily relied on Natural Language Processing (NLP), knowledge graphs, and document understanding to provide its core products to end users.

Here is a snapshot of the most popular tools:

Speech-to-Text

Helps translate earnings calls before they become available via text or documents. Firms can feed this information into their platforms immediately and generate signals.

Translate

Allows firms to cover markets where they don't speak the language. They can translate market-specific data as it becomes available and make important decisions faster.

AutoML NLP entity extraction

Trains Google's NLP models using AutoML NLP on terms that are specific to a firm's own business and the financial industry.

Document AI

Structures sell-side research, news and documents to unlock insights and identify areas of interest through a custom knowledge graph.

TensorFlow Enterprise

Run BERT and ELMo advanced NLP models for news or earnings data and receive a white glove service, including priority bug fixes and security patches. Train a wide variety of NLP models with BERT on TPUs extremely quickly.

NLP is one of the most important functions for investment firms looking to understand news, documents, earnings calls and sell-side research as it's published.

50-70%

Savings in cost and time per integration using APIs.[6]

2

Google Compute Engine for High Performance Computing

Quantitative hedge funds typically have a lot of proprietary code that runs as part of a high performance computing (HPC) cluster with an open source or proprietary job scheduler. The HPC cluster can be used for running monte carlo simulations to understand investment outcomes, value-at-risk (VaR) or for backtesting:

Custom Virtual Machine (VM) types

Allow firms to choose the processor architecture, exact number of cores, and amount of RAM required, giving more operational flexibility.

Preemptible VMs

Are up to 80% cheaper than regular instances[8] and include central processing units (CPUs), graphics processing units (GPUs), and tensor processing units (TPUs) architecture. They last up to 24 hours.

NVIDIA and GCP

Give access to massive parallel computational power and provide up to 1,000 teraflops of mixed precision hardware acceleration performance. NVIDIA Tesla K80, P4, P100, and V100 GPUs are tightly integrated with AI Platform, dramatically reducing the time to train ML models on large datasets from weeks to a few hours.

Live migration

Offers live migration to another host in the event of hardware failure (memory, CPU, network interface cards, disks, power, etc.). It can also host OS and BIOS upgrades, security patches, network and power grid maintenance, and system configuration changes.

An interest rate announcement from the Fed or market volatility can result in large demand for computational power to analyze these changes, and cloud's elastic nature can accommodate this well.

30-40%

Annual savings per buy-side firm on tech costs due to high performance virtualization:

1. Firmwide data center consolidation.
2. Cloud enablement across the technology stack.
3. On-demand and scalable computing models.[8]

7 https://cloud.google.com/preemptible-vms/

8 Celent, Market Trends And Nextgen Invest and Risk-tech For The Buyside: 2020 And Beyond, January 2020

Committed use discounts

Offer **up to 57%** off for most resources like machine types or GPUs, or up to 70% for memory-optimized machine types with one- or three-year commitment.

Batch on Google Kubernetes Engine (GKE)

Brings the functionality and familiarity of a traditional batch job scheduler into a cloud-first world. It allows firms to free their apps from fixed-sized compute clusters by dynamically allocating resources to meet their needs and respond to changes in the market.

Partners, such as NetApp

Offer a simple cloud-native file storage service with performance and advanced data management capabilities, for firms running HPC/HTC clusters with NFS requirements.

3

BigQuery

Google Cloud's serverless, highly scalable, and cost-effective data warehouse is easy to set up and manage, and doesn't require a database administrator. It offers real-time insights from streaming data and has a high-speed in memory business intelligence engine for faster reporting and analysis.

30%

Reduction in time to access and prep data using next generation, scalable big data lakes.[11]

BigQuery Machine Learning (BQML)

To build and operationalize ML models on planet-scale structured or semi-structured data, directly inside BigQuery, using simple SQL code, including logistic regression for financial risk profiling, matrix factorization for factor analysis, and classification algorithms.

BigQuery GIS

Uniquely combines the serverless architecture of BigQuery with native support for geospatial analysis, with support for arbitrary points, lines, polygons, and multi-polygons in WKT, WKB, and GeoJSON formats. For example, this is especially useful for foot traffic data to analyze the economic performance of a group of stores or particular neighbourhood.

BigQuery is NoOps

There is no infrastructure to manage, so no time needed for tuning, managing, and sharding the database.

BigQuery Storage API

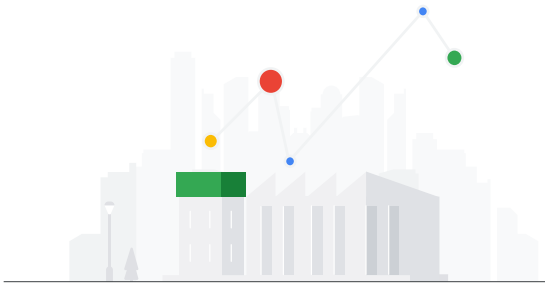
Provides fast access to BigQuery-managed storage and high-read throughput for other technologies such as Cloud Dataproc.

Market data providers

Such as Refinitiv[9] and CME Group[10], now offer access to an increasing number of datasets as they make their data available to customers on BigQuery.

Marketplace

Offers 150+ public datasets available to the general public for free through the BigQuery Public Datasets Program, including GDELT worldwide news and events.



9 https://www.refinitiv.com/en/financial-data/market-data/tick-history

10 https://www.prnewswire.com/news-releases/cme-group-to-offer real-time-market-data-via-google-cloud-platform-300937980.html

11 Celent, Market Trends And Nextgen Invest and Risk-tech For The Buyside: 2020 And Beyond, January 2020

4

Cloud Dataproc

Quantitative hedge funds have depended on the open-source framework Apache Spark or Apache Hadoop to crunch large volumes of alternative data. The promise of Hadoop and Spark also came with a large capital expenditure and heavy operational overhead, as the clusters were difficult to build and manage. Moving these workloads to **Cloud Dataproc** can provide a more cost-effective way to run Spark and Hadoop and, in result, reduce time to insights.

Fast and scalable data processing

Lets firms spin up hundreds of fully managed nodes with thousands of CPUs on Cloud Dataproc in under 90 seconds for each cluster. Specific completed jobs can be deleted to avoid resource contention, wastefulness or operational overhead.

Autoscaling clusters

Estimates the "right" number of clusters and provides auto scaling policies' API to ease accurate worker count estimations.

Custom Images and Initialization Actions

Allow buy-side firms that typically require proprietary packages and images in their clusters to bundle operating system, big data components, and GCP connectors into one package that is deployed on a cluster.

BigQuery Storage API

Integration gives high throughput access to BigQuery via a RPC-based protocol via the BigQuery Storage API. This integration allows firms to read huge volumes of structured data extremely quickly.

Moving Spark and Hadoop workloads to Cloud Dataproc can reduce the operational overhead, costs, and time to insight.

~10x

Reduction in analytics run-time down from 35 hours to 3 hours, while processing >10x more data.[12]

12 Celent, Market Trends And Nextgen Invest and Risk-tech For The Buyside: 2020 And Beyond, January 2020

5

AI Platform Notebooks on Deep Learning VMs

Developers, data scientists, and quants rely heavily on Python and R for working with structured and unstructured data, and JupyterLab is one of the fastest and most efficient ways to do this. **AI Platform Notebooks** offers an integrated JupyterLab environment for investment professionals to get these experiments up and running in minutes.

Here are select features worth mentioning:

Managed JupyterLab

Helps deploy a JupyterLab instance with all firm's libraries pre-installed, without requiring installation or management.

Pre-installed images

Allow instantiating a VM image containing the popular AI frameworks such as TensorFlow, PyTorch, and Scikit-learn on a Google Compute Engine (GCE) instance without worrying about software compatibility. Add GPUs and TPUs for extremely high performance.

AI Hub

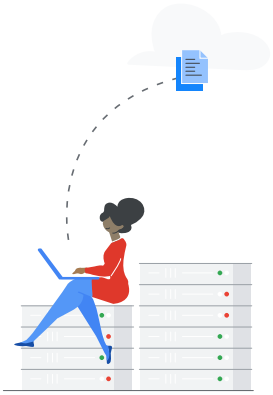
Includes a rich catalog of plug-and-play AI components and supports AI content sharing within a fund to avoid duplicate efforts and foster collaboration.

BigQuery

Integration empowers quants to easily load BigQuery datasets into R and Pandas dataframes, and to work with the data in their preferred language on AI Platform Notebooks.

Deep Learning Containers

Allow firms to bring their preferred container. The containers provide the flexibility to install specific libraries mandated by the fund.



A JupyterLab instance is a **Deep Learning virtual machine** instance with the latest machine learning and data science libraries pre-installed.

Celent predicts that investment functions will be significant beneficiaries of "analytical production factories" enabled by collaborative enterprise data and analytics cloud platforms.[13]

13 Celent, Market Trends And Nextgen Invest and Risk-tech For The Buyside: 2020 And Beyond, January 2020

6

Kubeflow Pipelines

While top investment firms typically have highly skilled quants and data scientists for creating ML models, it can be challenging to operationalize these models, resulting in weeks of wasted valuable time. **Kubeflow Pipelines** are designed with not only data scientists in mind, but also considering software and data engineers, helping firms build and share models and ML workflows within their organizations and across teams, for integration into different parts of the firm.

Model training is just a small part of a typical ML workflow.[15]

End-to-end orchestration

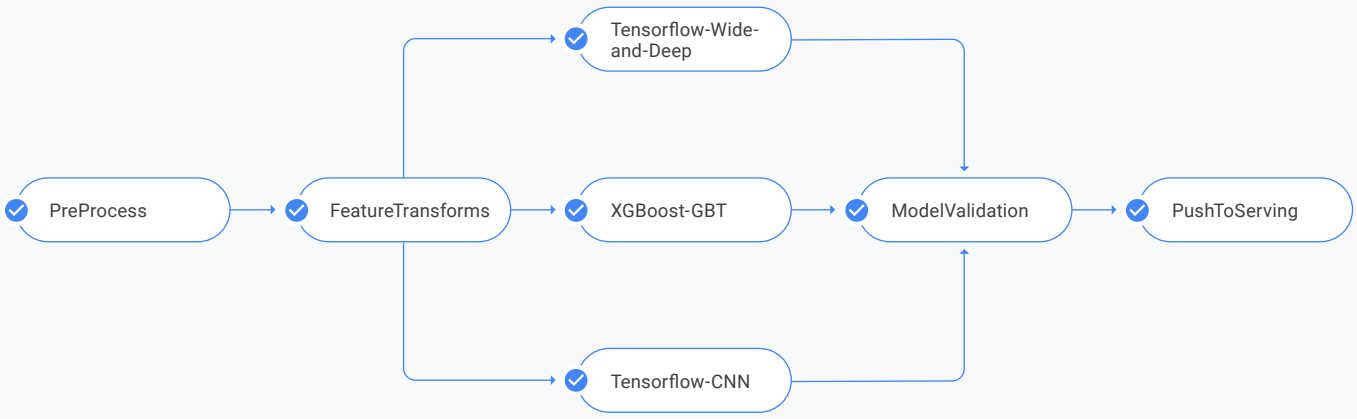
Enables and simplifies the orchestration of ML pipelines within the firm, including all the stages in the below illustration and providing a user interface to monitor them.[14]

Easy experimentation

Allows users to try many ML techniques to identify what works best for their application, supporting multiple trials and experiments.

Reusable pipeline components

Are self-contained sets of code, packaged as container images, that perform a step in the ML workflow. Reusability eliminates the need to rebuild pipelines every time.



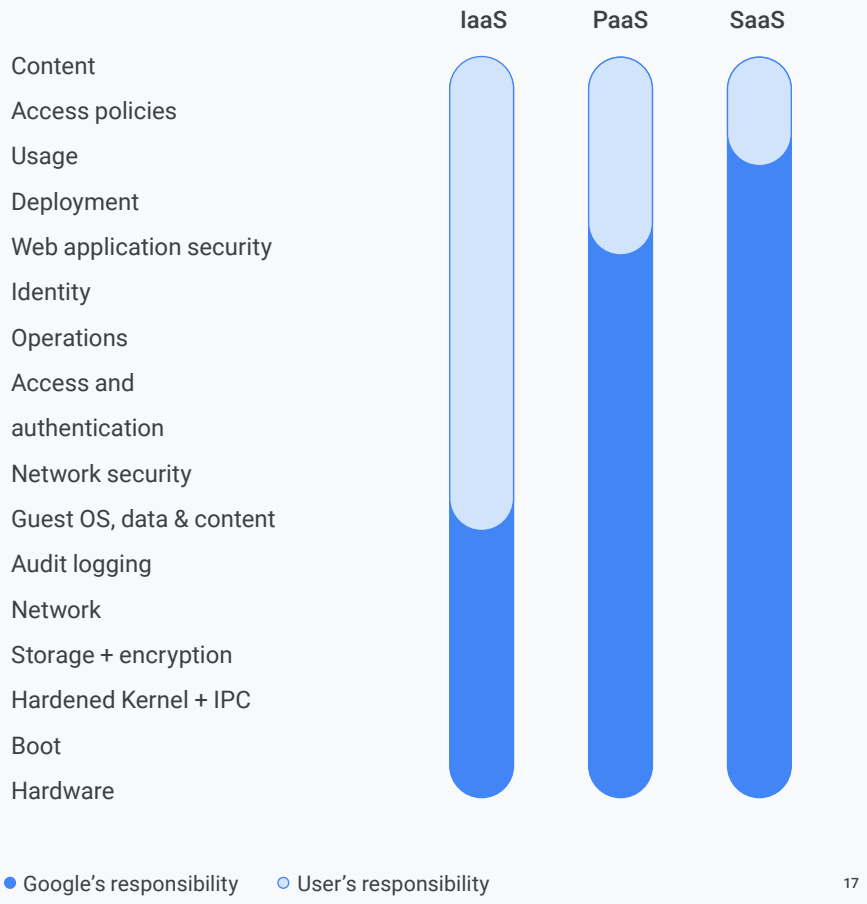
14 NEXT'19: Accelerating ML App development with Kubeflow Pipelines

15 Hidden Technical Debt in Machine Learning Systems

Security & compliance

Data protection and privacy is always a priority for financial services companies — which is why many firms became early adopters of the private cloud. Now, as the global regulatory compliance landscape becomes ever more complex and demanding, Google Cloud is helping these organizations meet today's compliance challenges via a public cloud infrastructure that offers data integrity, portability, and confidentiality.[16]

Forrester Research named Google Cloud a Leader in the Data Security Portfolio Vendors Wave.[18]



17

16 https://cloud.google.com/security/compliance/financial-services/

17 Google Cloud's shared responsibility model

18 The Forrester Wave™: Data Security Portfolio Vendors, Q2 2019

Google Cloud understands that alongside powerful computing tools, buy-side firms require powerful controls to help keep their information secure.

Cloud Security Command Center

Is a comprehensive security management and data risk platform for GCP to help meet compliance requirements as well as prevent, detect, and respond to threats.

Organizational Policy Service

Gives centralized and programmatic control over the firm's cloud resources, and establishes guardrails for firms' development teams to stay within compliance boundaries.

VPC Service Controls

Define a security perimeter around GCP resources such as Cloud Storage buckets, Bigtable instances, and BigQuery datasets to constrain data within a VPC and help mitigate data exfiltration risks. These controls keep sensitive data private and take advantage of the fully managed storage and data processing capabilities of GCP.

Cloud Key Management Service (KMS)

Helps manage encryption keys for funds' cloud services in the same way as managing on-premises. For compliance mandates requiring keys and crypto operations to be performed within a hardware environment, the Cloud KMS integration with cloud-hosted hardware security module (HSM) makes it simple to create a key protected by a FIPS 140-2 Level 3 device.

Cloud Security Scanner

Automatically scans the App Engine, Compute Engine, and GKE apps for common vulnerabilities.

Managed base images

Are base container images that are automatically patched by Google for security vulnerabilities, using the most recent patches available from the project upstream (for example, GitHub).

Container Registry

In-depth vulnerability scanning helps detect vulnerabilities in early stages of the software deployment cycle.

Binary Authorization

Is a deploy-time security control that ensures only trusted container images are deployed on GKE.

Chronicle's backstory

Investigation flows—added to Google Cloud's detection, incident management, and remediation capabilities—can ingest massive amounts of telemetry data and security logs from across the firm. These flows can then index the data, correlate it to known threats, and make it available quickly.

Let's get solving

It's time to free your engineers, quants and data scientists from the burden of building and managing data acquisition platforms and applications. Give them more time to find actionable insights and focus on what's important: driving signal generation and generating alpha.

With Google Cloud's six powerful tools, you have the compute processing power to interpret and understand these myriads of alternative data sources, the AI to predict market behavior, and the ability to uncover market-defining trading signals like never before.

Get in touch with one of our reps today, for a demonstration of how these tools can transform your organization and help you generate alpha with Google Cloud.

With special thanks to our author Colman Madden, Customer Engineer for Google Cloud.

Google Cloud