Enhance your investment research with Google Cloud

Four keys to discovering valuable insights at scale
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>03</td>
</tr>
<tr>
<td>Four keys to improve your investment research</td>
<td>04</td>
</tr>
<tr>
<td>1. Data acquisition, analytics, and discovery</td>
<td>06</td>
</tr>
<tr>
<td>2. Burst compute workloads</td>
<td>14</td>
</tr>
<tr>
<td>3. ML and model deployment</td>
<td>17</td>
</tr>
<tr>
<td>4. Natural Language and Document AI</td>
<td>20</td>
</tr>
<tr>
<td>Next steps</td>
<td>23</td>
</tr>
</tbody>
</table>
Introduction

Investment management is a heavily data-driven industry, and highly competitive. The initial challenge comes from the volume of data and building robust systems that can accommodate the scale and performance needed.

Investment teams require new datasets to be onboarded regularly and tested to determine if they can provide insights in a constantly changing world. This needs to be done in an agile fashion in order to generate maximum insights from the data, and ultimately maximum returns.

Extracting insights or signals presents the next challenge, and providing research teams an environment where they can discover the datasets and easily experiment with new tools and techniques to apply to their research process - is where the value lies.

In this paper, we build on the ideas presented in *Generating Alpha with Google Cloud* and describe how the key components of Google Cloud can reshape the way investment managers organize their data and their processes.

Check out our first whitepaper, *Generating Alpha with Google Cloud*, where we talked about some of the tools that quantitative investment firms can use for research and signal generation.
Four keys to improve your investment research

An investment research process often starts with onboarding data such as market prices, macroeconomic activity, earnings reports, sell side research, company fundamentals, news, as well as numerous alternative datasets, via a data pipeline.

Using this data, various models perform compute-intensive calculations and portfolio simulations that need to ramp up resources quickly, and perform reliably throughout their life cycles. This may include the deployment orchestration and the execution of specialty models that use techniques such as Natural Language Processing (NLP), and Document AI. Each one of these steps play a part in the process of unlocking insights from data.

Firms bringing investment research workflows onto Google Cloud can consider a reference architecture, covering four distinct areas, as a guide for their own research deployments.

“Differentiated investment strategies require new types of information sources and new ways to process that information. And that, of course, relies heavily on having access to reliable and scalable storage, computational, and ML/AI resources. In particular, quantitative strategies can benefit from the computational platforms and embedded AI/ML capabilities the cloud can offer.”

David Easthope, Senior Analyst, Market Structure and Technology, Greenwich Associates

Data acquisition, analytics and discovery
Empowering data engineering teams to ingest, discover, and analyze various datasets with minimal extract, transform, and load (ETL) toil.

Burst compute workloads
Taking advantage of compute-optimized and preemptible machine types.

ML and model deployment
Tools that enable data scientists and analysts to build and operate ML models at scale.

Natural Language & Document AI
Tools to help you derive insights from disparate sources, for example: news, social media, job postings, earnings calls, or research reports.

Key to the wielding of these key components is the use of Terraform's Infrastructure as Code (IaC) (IaC) methods to declaratively define, provision and manage resources. Additionally, IaC is an important mechanism through which operations and compliance divisions can encapsulate a firm’s networking and security policies to automate the governance of cloud resources. The Cloud Foundations Toolkit provides a series of Terraform reference templates based on Google Cloud best practices.

Figure 1. Investment research platform reference architecture

Let’s examine each component in more detail.
Data acquisition

Among the most fundamental parts of a data acquisition strategy is determining which datasets are necessary to test your investment thesis.

For those with large datasets to load, ingestion can be simplified with the Storage Transfer Service, which enables you to manage large-scale data transfers easily, securely, and efficiently.
Storage Transfer service for on-premises data is a scalable, reliable, and managed service that allows you to move your data without investing in engineering teams or buying transfer solutions. To use it, install a Docker container containing the on-premises agent for Linux on your data center's servers, and the Storage Transfer service for on-premises data coordinates the agents to transfer data securely to Google Cloud Storage (GCS) buckets.

For smaller transfers of on-premises data, you can use gsutil, which can be especially useful in the following scenarios:

• Your transfers need to be executed on an as-needed basis or using command-line tooling.
• You're transferring either a small number of files or very large files.
• You're consuming the output of a program, or streaming output directly to GCS.
• You need to watch a directory with a moderate number of files and sync any updates with very low latencies.

BigQuery users similarly have many options. Among these is setting up a cloud function to trigger when a file lands on GCS. The cloud function can then perform any requisite pre-processing or normalization before loading the data into BigQuery tables.
A lower-code alternative to cloud functions are Dataflow templates: configurable, pre-built Apache Beam pipelines that encapsulate common tasks, like compressing a set of files in a storage bucket or loading data into BigQuery from a Pub/Sub topic. Templated pipelines can be invoked from the Google Cloud Console, using the gcloud CLI, or by REST API calls. Classic templates are staged as execution graphs within cloud storage while Flex Templates package the pipeline as a Docker image and stage these images on your project’s Container Registry.

When choosing data encodings, consider that Apache Avro is one of the fastest encodings to load into BigQuery as the data can be read in parallel even if the data blocks are compressed.

Figure 4. Loading data - file formats

Figure 5. Moving data from external source systems to BigQuery
Managing a large number of commercial datasets can contribute a high percentage of your data estate’s toil. One option recently made available to assist Google Cloud customers in this area is Crux Informatics, who offers a solution for delivering datasets from over 125 data providers and 41 other public data sources. Data provisioned via Crux enables Google Cloud clients to consume data that has already undergone validation (e.g., schema change), standardization (e.g., data consistency) and enrichment (e.g., record-level operational metadata tagging).

One of the benefits of BigQuery is that data publishers can load data once and easily share it with many data consumers. Data providers can take advantage of this by deploying access control through BigQuery authorized views. This means the data stays in place and the vendor simply grants the customer access to the dataset using identity and access management (IAM). This alleviates some of the complex ETL and data validation processes teams undertake and ensures consistency with vendor data. This leads to higher data quality and a more agile data engineering operation.

BigQuery also has more than 100 public datasets that can be accessed via the Marketplace. Google Cloud Public Datasets simplify the process of getting started with analysis, as you don’t need to search and onboard large data files or find licensing terms. Public datasets hosted in BigQuery provide users with free access of up to 1 TB/month in queries.

Pub/Sub provides a foundation for asynchronous processing and real-time data ingestion for event-driven systems. Pub/Sub is a global messaging service that decouples event-producing services from event-processing services and provides durable message storage alongside real-time delivery - allowing for high throughput and consistent performance at scale.

For example, CME Group utilizes Pub/Sub to distribute its Smart Stream real-time market data product on Google Cloud. It grants its licensed customers access to Pub/Sub topics for real-time market data delivery, and its customer’ applications simply consume data from that topic.

“Our partnership with Google Cloud removes unnecessary friction in the data supply chain and is a win-win for suppliers and consumers alike. Our ready-made pipelines and value-add services enable fast delivery and customization of data into your BigQuery environment.”

Philip Brittan, Founder of Crux Informatics

“‘Our clients around the world increasingly are looking for quality real-time data within the cloud. With Google Cloud’s Pub/Sub, it will be easier for our clients to access the data they need from anywhere with an internet connection.”

Trey Berre, Global Head of Data Services, CME Group
Some helpful tips for managing data acquisition in BigQuery include:

- Extract, load, and transform (ELT) is often preferred over ETL, as you can simply leverage the distributed processing capabilities of BigQuery to transform your data with SQL at runtime (schema-on-read vs. schema-on-write).
- Use compressed data types, as they will load much quicker (e.g., Avro, Parquet, ORC, instead of CSV or JSON).

**Data discovery**

Data Catalog can help you organize business and technical metadata related to your datasets as a fully managed service and provide access to this metadata through a search interface. Once a dataset has been discovered, users can query it in BigQuery using traditional methods, such as SQL/Python, or leverage BigQuery’s Data QnA (alpha) capability to write natural language questions against the dataset.

**Analytics**

Now that we have covered various ways to acquire and onboard data into BigQuery, let’s discuss ways you can query and analyze that data. A good example of this is the Refinitiv Tick History dataset which offers the full breadth and depth of Refinitiv’s historical tick data. The multi-petabyte dataset covers over-the-counter (OTC) and exchange-traded instruments from a large selection of asset classes in more than 500 venues dating back to 1996.

“Combining Google Cloud’s machine learning tools with Refinitiv’s Tick History data in BigQuery is a step-change for customers looking to develop new trading models, interpret trade patterns or comply with regulations. The cloud is changing the paradigm for the financial community, enabling less time and money spent managing data and more time innovating and getting answers from data that gives a competitive edge.”

Catalina Vazquez, Proposition Director, Tick History, Refinitiv
BigQuery can achieve sublinear scaling on such massive datasets, as the storage layer gets more efficient after adding more data, especially highly repeated data like pricing time series.

Figure 7: Example of encoding semistructured data using definition and repetition levels

```protobuf
document {  
  required int64 DocId  
  optional group Links {  
    repeated int64 Backward;  
    repeated int64 Forward;  
  }  
  repeated group Name {  
    repeated group Language {  
      required string Code;  
      optional string Country;  
    }  
    optional string Url;  
  }  
}
```

Figure 6. Refinitiv’s Tick History on Google Cloud solution architecture (source: Refinitiv’s fact sheet)
Here is an example of a query we ran on BigQuery for calculating hourly high/low bars on [Refinitiv Tick History Level 1 and 2 Data](#) for a five-year period:

```sql
SELECT
  RIC,
  TIMESTAMP_TRUNC(date_time, HOUR) AS time_group,
  MAX(L2_BidPrice) AS max_bid,
  MIN(L2_BidPrice) AS min_bid
FROM 'dbd-sdlc-prod.LSE_NORMALISEDLL2.LSE_NORMALISEDLL2'
WHERE
  Date_Time BETWEEN '2015-01-01 00:00:00' AND '2020-01-01 00:00:00'
GROUP BY
  RIC,
  time_group
```

The query aggregated 2.6TB / 77.9 billion records in ~60 seconds. Similar queries can take many hours to run on some traditional database systems.

Looker can then be used to create dashboards on top of the Refinitiv Tick History datasets as demonstrated below here. As an enterprise platform for business intelligence, data applications, and embedded analytics, Looker can help create an interface for portfolio managers to interact with and visualize the data. This can be embedded into existing research portals.

> Looker’s ML and the data modeling layer provided an opportunity to get hands-on and technical in our implementation. Its clean, crisp, and concise view of our data and the different looks and dashboards we’ve built have clearly communicated the point that we wanted to drive home about our information security program.”

Colleen Valentine, Head of Information Security, Governance and Compliance Group, Nasdaq

Figure 8. Sample Looker dashboard on Tick History data
Additionally, as ESG and climate risk become increasingly important topics in portfolio management, firms are looking to geospatial platforms to monitor risk from hurricanes, flooding, wildfire, and global warming. BigQuery GIS has geography data types and functions to analyze spatial data and provide valuable insights to portfolio managers. For example, Google Cloud’s partner CARTO offers geospatial solutions and capabilities via CARTO BigQuery Tiler that allows you to visualize very large location datasets.

Some helpful tips for analytics in BigQuery include:

- Use nested or repeated structures within BigQuery for denormalizing data. This leads to faster query times and leverages BigQuery’s low storage cost. This can be particularly useful for portfolio data.
- Use partitioned tables to reduce table scans, query times, and cost (number of bytes read).
- Use clustering on partitioned tables for additional performance improvements.
- Use point-in-time queries to access data bitemporally.
- Use google-cloud-bigquery Python client library to retrieve BigQuery data as a pandas dataframe, or to load a pandas dataframe to a BigQuery table.

For streaming analytics workloads, you can write code that reads from Pub/Sub and calls the BigQuery streaming API directly, or you can use Dataflow with the Apache Beam SDK to set up a streaming pipeline. These pipelines can be custom or employ one of the several no-code Dataflow templates available within the Cloud Console. Additionally, features like Dataflow SQL enable you to perform streaming analytics using SQL against a Pub/Sub topic - while joining the stream with existing BigQuery datasets.
Portfolio managers, researchers, and data engineers require ready access to high-performance compute (HPC) capabilities to perform backtesting, portfolio simulations, run risk calculations, and deliver other useful insights. In this section, we will review the HPC components that can enhance or make up your investment research platform.

Figure 9. An architecture for quantitative research on Google Cloud
We are a wholesaler of risk. We need to be measuring, factoring, analyzing, and monitoring risk, and it consumes a lot of compute power. As we thought about scaling our own business and how much more intense that risk management activity could become in the world, Google Cloud was an obvious choice.”

George Lee, Co-Chief Information Officer, Goldman Sachs

Google Compute Engine VMs boot in seconds, have standardized high-performance configurations, and deliver consistent performance. Custom machine types allow you to right-size your workload; and you can also leverage preemptible VMs at an 80% discount. Our compute-optimized (C2) VMs offer 3.8GHz sustained all-core turbo and are Non-Uniform Memory Access (NUMA) aware for added performance. Additional performance optimizations can be found by using placement groups, which offer rack-level co-location of VMs to reduce inter-node communications latency in your HPC cluster.

Google Cloud’s high-performance private network is a global backbone that underpins all of Google Cloud’s resources providing for optimized throughput and lower latency than public networks.

There are many storage options, including local NVMe SSDs for ephemeral storage, persistent disk for high-performance replicated block storage, and GCS for object storage.

Cloud Filestore provides highly available, durable, POSIX-compliant shared storage. Filestore High Scale is a managed, elastic, scale-out file service for any customer requiring high capacity and performance in their in-cloud and hybrid environments, supporting ~480k max IOPS and 64k max connections.

We also recently announced the Public Preview of a CentOS 7-based Virtual Machine (VM) image optimized for high performance computing (HPC) workloads, with a focus on tightly-coupled MPI workloads. The HPC VM image is available at no additional cost via the Google Cloud Marketplace.

There are many third-party options as well, including SchedMD, Slurm, HTCondor, and Univa Grid Engine. These third party offerings provide enterprise capabilities such as elastic scaling based on queue depth, dynamically bursting compute from on premise to cloud, and federating workloads between on-premise and cloud clusters.
Apache Spark is a popular framework in the investment management world and used for processing large amounts of data in parallel. On-premise Hadoop clusters can have a large amount of resource contention, and users are limited to the resources that are available at a given time. Cloud Dataproc provides fully managed Apache Hadoop, Spark, and Presto clusters, and can deploy a new cluster with thousands of cores in under 90 seconds, process data, and then shut down. Clusters can auto-scale to the required amount of cores and memory required by the job and leverage preemptible VMs to do so at a low cost. The BigQuery Storage API offers high-performance throughput between BigQuery storage and Dataproc clusters.

Dask has become quite a popular framework in the investment management community as it is written in python on top of numpy, pandas and other popular python libraries, and can add scale and performance efficiencies to these workloads.

We recently announced Dask support for Dataproc, Google Cloud’s fully managed Apache Hadoop and Apache Spark service, via a new Dask initialization action. With this Dataproc initialization action we’ve made it even easier for researchers to get Dask up and running on a Dataproc cluster. Additionally, you get access to the full set of features offered by the Dataproc service, including Autoscaling, Jupyter component and component gateway for submitting jobs via a Jupyter Notebook.

Containerized workloads, including web scraping applications are leveraged often in investment research teams. Google Kubernetes Engine (GKE) or Cloud Run can be used to run and auto scale these workloads and deploy nodes in different regions on our global network. GKE is a managed Kubernetes service with four-way auto-scaling and multi-cluster support. Cloud Run allows you to do this on a serverless platform without having to manage a cluster.
Quantitative researchers scour vast amounts of market and alternative data sources searching for signals and correlations. ML Engineers have the challenge of taking the research output (whitepaper, Jupyter notebook) from the researchers and moving it into production. Google Cloud provides managed JupyterLab Notebook instances as a fully managed solution for data science and ML that allows quantitative researchers to perform research and an MLOps framework for ML engineers to deploy, monitor, and maintain research in production.

Figure 10. AI Platform component overview
First, prepare and store your datasets in Cloud Storage. Then use the built-in Data Labeling Service to label your training data for classification, object detection, entity extraction, and other objectives for image, video, tabular, and text data. Data in financial firms is often confidential, and the data labelling service can be used to farm out data labelling tasks to internal teams.

You can build no-code ML models with AutoML's UI, or use your own code written in Notebooks, a managed Jupyter Notebook service. Use the latest open-source deep learning frameworks on Deep Learning VM Image or Deep Learning Containers. Then, train your models with a fully managed training service. An outline of how this process works for custom coded models is:

1. Create a Python application that trains your model.
2. Put training and verification data in Cloud Storage, Cloud Bigtable, or another Google Cloud storage service.
3. Package the application and transfer it to a Cloud Storage bucket (automated when you use gcloud command line to run a training job).
4. The AI Platform Training service sets up resources for your job, allocating one or more VMs, called training instances. Each training instance is set up by:
   - Applying the standard machine image for the version of AI Platform Training your job uses.
   - Loading your application package and installing it with pip.
   - Installing any additional packages that you specify as dependencies.
5. The training service runs your application, passing through any command-line arguments you specify when you create the training job.
6. You can get information about your running job through Cloud Logging or an API request.

Refinitiv Tick History on Google Cloud provides superfast querying and analytics with ML-ready data output that can quickly be deployed into ML models or to create and train new ones. There is no need to extract, batch or break down data into sizeable compute chunks for the ML models – all queries and analytics can be run straight from the source data in the BigQuery engine across days, weeks, months or even years of exchange data for your ML workflows to learn and provide close to instantaneous insights.”

Tim Anderson, Solutions Business Director, Refinitiv
When the training job succeeds or encounters an unrecoverable error, AI Platform Training halts all job processes and cleans up the resources. Validate your model with AI Explanations and What-If Tool, which can help you understand your model's outputs, verify model behavior, identify bias, and find ways to improve your model and training data. Take model tuning a step further using Vizier, a black-box optimization service, to tune hyperparameters and optimize your model's performance.

Deploy your models at scale to get predictions in the cloud with Prediction, which serves your models for either online or batch prediction requests.

Finally, manage your models, experiments, and end-to-end workflows with Pipelines by applying MLOps best practices and incorporating a cross-functional operating model for model approval. Pipelines’s continuous performance evaluation helps you monitor your model.

These capabilities reduce the toil of operationalizing ML models and free your quants and data scientists to focus on the most differentiating activities.
Thousands of financial filings, news articles, and sell-side research reports are generated every day, and it’s difficult for humans alone to process this volume of information. Our mission at Google is to organize the world’s data and make it available and useful. We have leveraged NLP heavily to achieve this mission, and now, we provide these models through our Natural Language API, AutoML Natural Language, and Document AI APIs. We also open-sourced a technique for NLP pre-training called Bidirectional Encoder Representations from Transformers, or BERT, in 2018, which enables anyone in the world to train their own state-of-the-art question answering system (or a variety of other models) in about 30 minutes on a single Cloud TPU, or in a few hours using a single GPU.

In addition to identifying entities and events, Google’s Translation API and BERT architecture can also be used to solve an important security master linking problem. Let’s take an example written in Korean: "아이폰 SE2, 3백99달러에 출시" (translated to English via Google Translate as: "iPhone SE2 launches for $399").

When processing this text, Accern’s Entity Classification model used a combination of original and English translated text to detect “iPhone” as a product. It then associates it to Apple Inc. and its unique identifiers, and produces the following output, which can also be accessed via an API:

Ted Merz, Global Head of News Product, Bloomberg

More and more news and information that was critical for financial markets was breaking in local languages in other countries. Within our news product, we aggregate news from hundreds of thousands of different sources. We use Google Translate to make it accessible in the language that’s useful to our customers at a moment’s notice.”
Accern's No-Code AI Platform, using Google's BERT architecture, helps our global clients make sense of the unstructured financial content from 100+ languages, including news, filings and internal documents. The platform also gives users the ability to extract key financial entities/themes using out-of-box or custom taxonomies and link them back to security master and unique identifiers such as OpenFIGI or LEI.”

Anshul Pandey, Co-Founder and CTO, Accern
This is provided as a JSON response, which enables you to systematically identify and react to events as they occur. Sentiment analysis, entity analysis, entity sentiment analysis, content classification, and syntactical analysis is included in the response. AutoML Entity Extraction (EE) models can be trained to recognize customized or industry-specific terminology as well.

**Document AI APIs** can be used to process complex documents and extract form and table data from financial filings. There are a number of templates available for specific form types and more are added on a regular basis. For example, in SEC form 13F filings, you can easily extract the table data from the report and return them as JSON documents to be processed downstream.

Figure 11. Sample SEC Form 13F (source: SEC edgar DB)
Next steps

There are plenty of emerging technologies, tools, and approaches available to aid investment firms today. At Google Cloud, we can help you access, organize, and utilize these four essential components to make your investment research fast and reliable.

To learn more or explore how Google Cloud fits into your investment research platform architecture, get in touch with your account representative or visit our website today.

With special thanks to the authors Colman Madden, Principal Architect and Sal Sferrazza, Solution Architect, as well as contributors Petra Kass, Head of Marketing, Capital Markets, and Joe Zagorski, Copy Editor at Google Cloud.