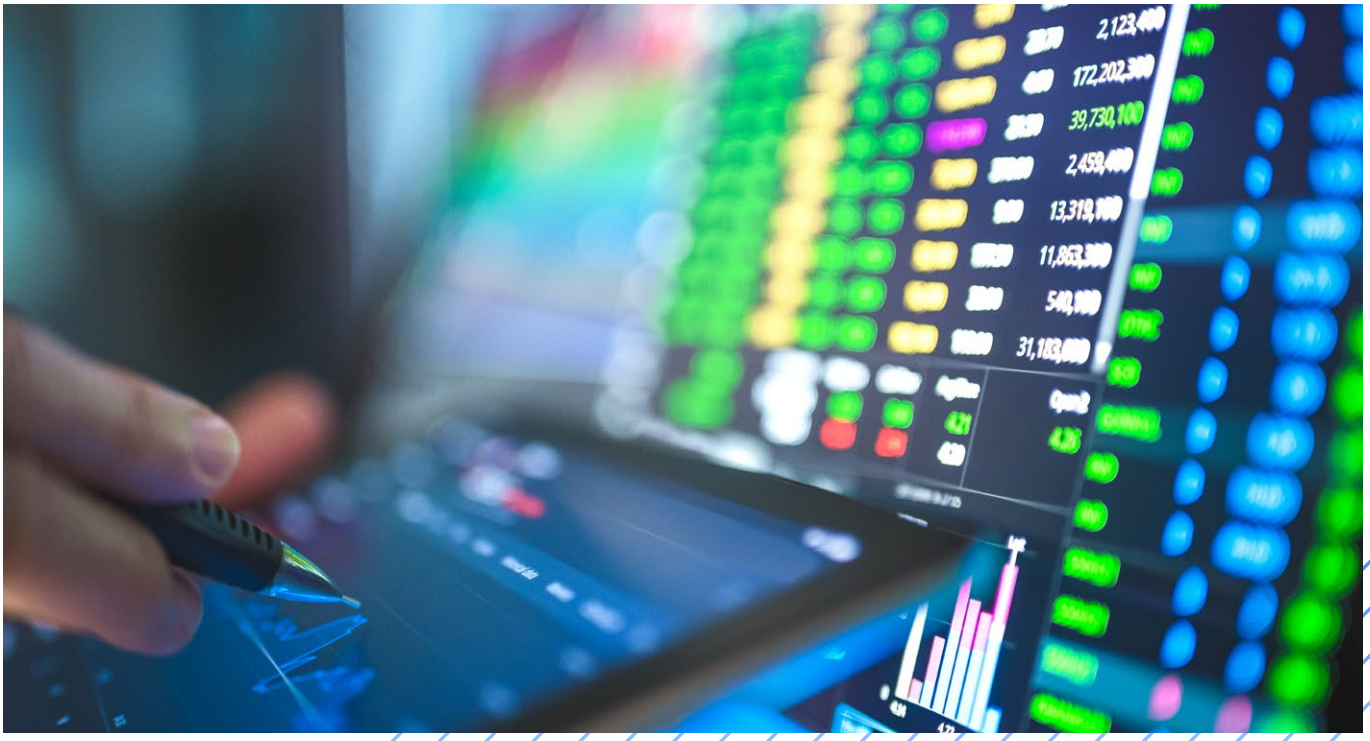
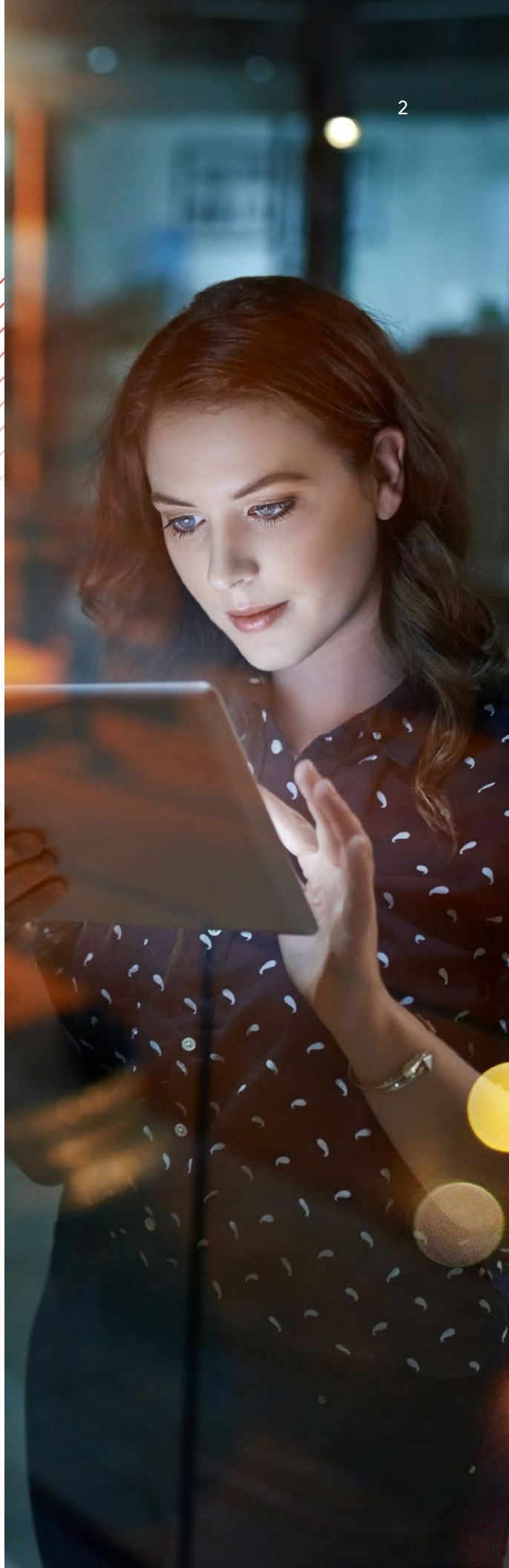


The Next Frontier: Cloud FinOps Powered by Generative AI

Authors: Eric Lam

Contributors and Reviewers: Amy Liu, Anders Geertsen, Bruce Warner, Carolyn Ujcic, Crispin Velez, Daniel Pettibone, David Nguyen, Neama Dadkhahnikoo, Nitin Aggarwal, Pathik Sharma

Google Cloud



Managing cloud costs can be as complex as running a 5-star restaurant. Just as a top restaurant needs a talented manager overseeing operations, and a master chef orchestrating the kitchen, effective budgeting and spending requires the combination of financial discipline and AI-powered insights. FinOps brings discipline to the cloud by aligning technology, finance, and business teams with an overarching operational framework focused on driving cost-effective returns on technology investments, with organizations reducing their cloud spend by as much as 30%. Like a restaurant manager, FinOps promotes accountability, strategy, and optimization. But even the best manager needs a talented chef.

That's where generative AI comes in – providing the expertise to elevate cloud financial management. With its ability to synthesize, analyze, and generate real-time insights from existing cloud financial data, AI is the missing ingredient taking cloud FinOps to the next level. It's akin to having a Michelin Star chef in your kitchen, blending ingredients in innovative ways and tailoring each dish to perfection. With a mix of advanced DataOps and MLOps capabilities in place, AI can sift through your data to uncover savings and enable better spending decisions.

Organizations at the forefront of their cloud FinOps adoption are maximizing cloud value by leveraging predictive insights with their data to unlock previously unrealized benefits including the ability to:

Lower cloud costs

Drive sustainable business value

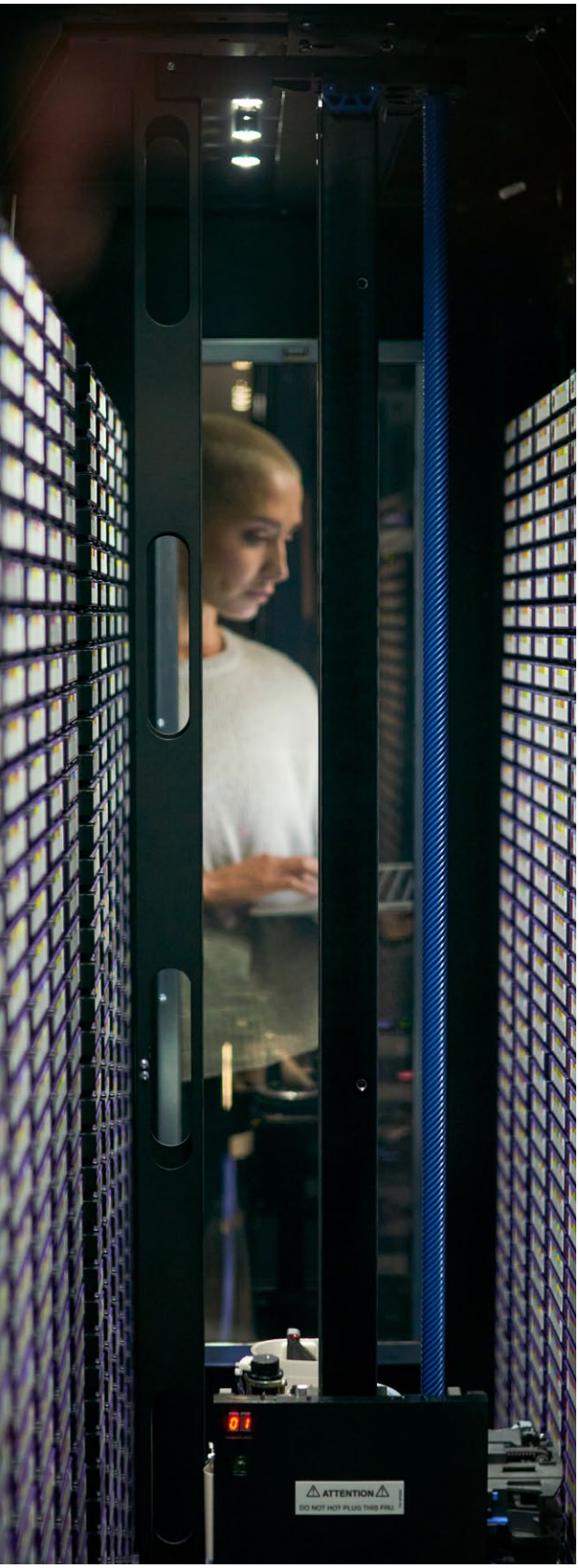
Foster a cost-conscious culture

Improve business agility

Make better business decisions

For years, successfully implementing the cloud FinOps discipline and its practices was an arduous journey filled with mislabeled spreadsheets, siloed data, and human-driven analysis of complex data sets The reliance on manual labor made scaling FinOps practices across the enterprise incredibly difficult. With such a resource-intensive endeavor, requiring significant time and effort to simply take the first step, teams desperately needed help. Fortunately, the rise of generative AI is adding a whole new set of capabilities for the cooks in the kitchen.

Like the master chef with a keen sense of taste, generative AI can analyze mountains of data, and discern ingredients for cost optimization. From the reams of data across your organization, it can create new and original content, such as text, code, or images, to generate tailored reports for individual stakeholders. Furthermore, it can be used for cloud financial data summarization, extraction, and search leading to cost efficiency and new revenue opportunities for the organization. From the cost efficiency perspective, generative AI can automate time-consuming tasks like manually summarizing financial reports, extracting key data points, and searching through large datasets. This frees up employees for higher-value activities and reduces labor costs. In addition, AI can generate reports and analysis automatically, providing businesses with real-time insights into their financial performance. This can improve decision-making and response times. The ability to summarize complex data sets into digestible reports, rapidly search for cost savings opportunities, and continuously extract patterns from noise makes generative AI a powerful tool for automating and accelerating cloud FinOps key capabilities at scale across the enterprise.



Here are some ways that generative AI can be used to accelerate Cloud FinOps adoption at scale:

Cost anomaly detection

AI models like BigQuery ARIMA Plus can identify abnormalities and deviations from normal cloud usage and spending patterns with time series dataset. Over time, the model assists organizations by proactively addressing potential cloud cost-related issues and identifying the root cause of the unusual consumption patterns. These advanced technologies provide FinOps teams with an extra set of eyes to identify causes for cost concern across an organization's cloud real estate.

With a near-real time cost anomaly solution, FinOps teams gain a powerful advantage, including:

- **Proactive cost management:** By identifying and addressing cost anomalies early, organizations can avoid unnecessary spending and save money.
- **Root cause analysis:** The model analyzes the cloud billing data to identify the root cause of cost anomalies, which can help them take steps to prevent them from happening again.
- **Recommended actions:** Based on its diagnosis, the generative AI model can generate recommended actions to address cost anomalies. Teams can take action immediately, rather than hypothesizing potential fixes.

With an AI anomaly detection solution, cloud FinOps team can gain the ability, agility, and foresight to proactively address potential cost-related issues, save money, and optimize their cloud cost management practices in near-real time. As an example, one of the world's leading financial derivatives exchanges deployed an AI anomaly detection solution on Google Cloud Looker business intelligence platform and were able to prevent a runaway cost situation that had compounding implications in a matter of days resulting in significant cost avoidance savings.

Cloud financial forecasting

Predictive analytics is a key area where AI can be used to improve financial forecasting of cloud consumption. Like a seasoned chef, with a well-stocked kitchen exploring new trends, the process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior. Organizations can use historic and current data to forecast trends and behaviors days, weeks, or months into the future with better precision and accuracy.

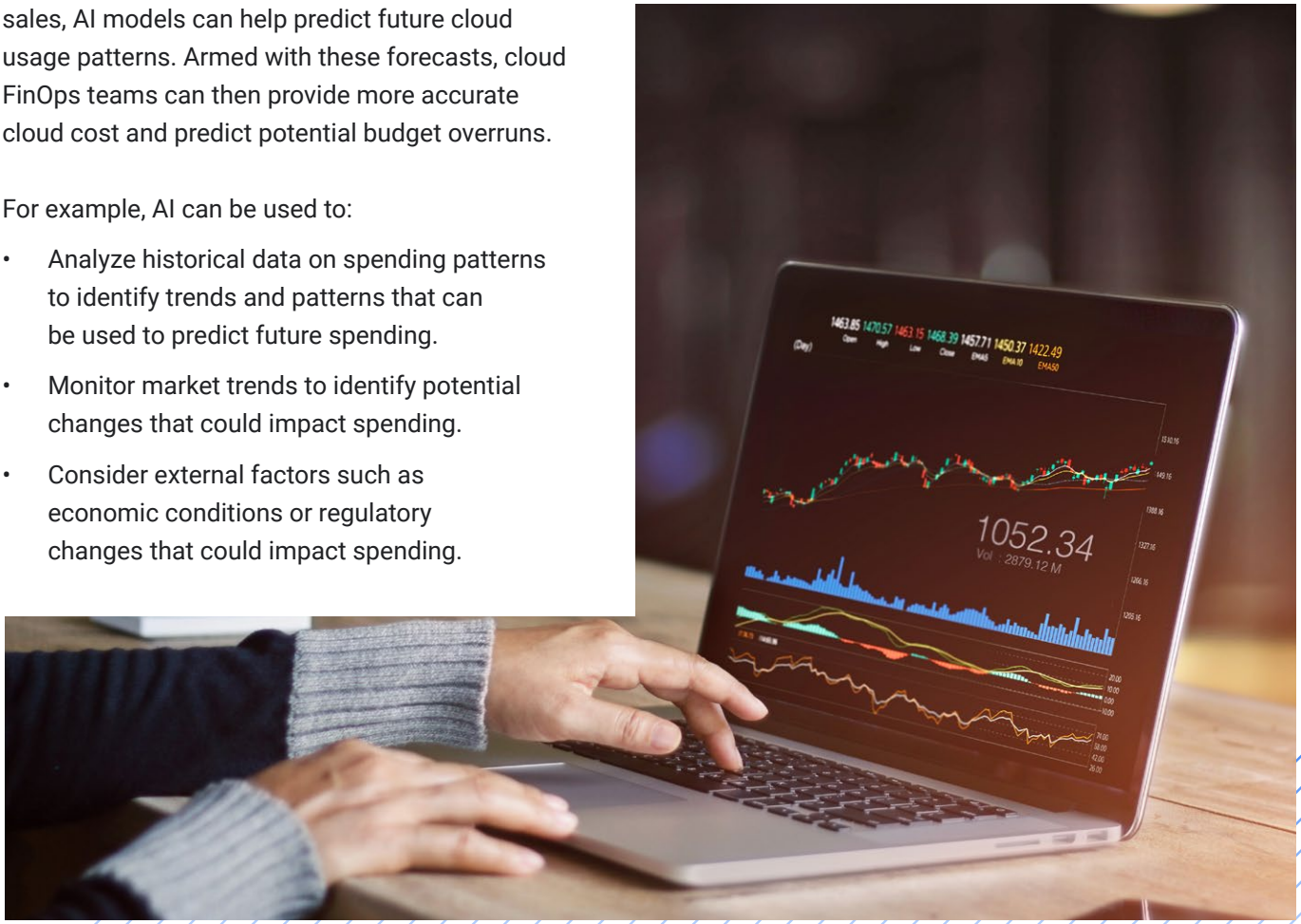
By analyzing historical cloud consumption data, business cost drivers, and external factors such as weather trends impact on a retailer's online sales, AI models can help predict future cloud usage patterns. Armed with these forecasts, cloud FinOps teams can then provide more accurate cloud cost and predict potential budget overruns.

For example, AI can be used to:

- Analyze historical data on spending patterns to identify trends and patterns that can be used to predict future spending.
- Monitor market trends to identify potential changes that could impact spending.
- Consider external factors such as economic conditions or regulatory changes that could impact spending.

The insights generated by advanced predictive analytics can help FinOps teams make data-driven decisions about resource allocation and cost management. For example, by identifying potential cost-saving opportunities, FinOps teams can free up budget for other initiatives or invest in areas that will yield a greater return on investment.

Predictive analytics can potentially revolutionize FinOps. As AI models become more sophisticated, FinOps teams can make even better informed decisions about managing their costs.



Cloud spend optimization

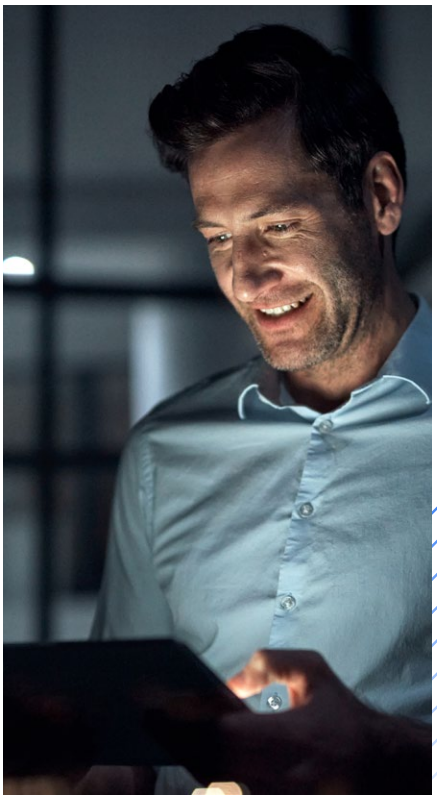
The combination of large language models (LLMs) and generative AI unlocks the ability to derive insights from unstructured text-based data and allow you to extract information to generate new understanding of that data. By leveraging these technologies, we can quickly sift through and make sense of vast amounts of text-based financial documents, like contracts, cloud invoices, and IT spending reports. It's not just about reading the data; LLMs and generative AI allows us to extract crucial information such as dates, prices, and usage patterns.

The ability to analyze and cross-reference financial information in near real-time is invaluable, and leads to actionable recommendations for cost-savings and efficient decision making. By harnessing these state-of-the-art tools and services,we're not only able to gain insights from unstructured data but also optimize financial strategies in a way that was previously impossible. It's a game-changer for financial management, offering a smarter, more informed approach to handling financial data.

LLMs and generative AI are powerful tools that can gain insights from financial data that would otherwise be difficult or impossible to obtain. By leveraging these technologies, cloud FinOps teams can make better financial decisions, save money, and improve their financial performance.

As we delve into the transformative impact of generative AI on financial operations, we recognize its substantial potential to redefine the landscape of cloud FinOps. Google Cloud's hands-on experience in building and delivering generative AI projects has provided us with valuable insights into its strengths, shaping our understanding of its capabilities. The transformation of this framework naturally leads us to consider the reciprocal relationship where FinOps is not just a beneficiary but also a key value enhancer of generative AI workloads. At this pivotal intersection, our discussion shifts from how generative AI can drive innovation in FinOps to how the principles of FinOps can in turn maximize the investments of deploying generative AI solutions at scale.

Cloud FinOps for generative AI



Generative AI can turbocharge your Cloud FinOps adoption and help you hit cloud cost optimization targets ahead of schedule. At the same time, applying FinOps principles and practices allows enterprises to get the most value from their generative AI investments. The combination of AI and FinOps creates a powerful cycle where each one amplifies the strengths of the other, leading to faster time-to-value and lower costs. By bringing them together, organizations can accelerate cloud transformation and make every dollar spent on AI go further.

At Google Cloud, we have developed the Cloud FinOps for Generative AI framework to help customers assess their organization's readiness across people, processes, and technology. This proactive approach uncovers blind spots, prioritizes key areas of focus, and enables rapid deployment of resources to fill potential gaps as they scale.

Cloud FinOps for Generative AI Framework

The Cloud FinOps for Generative AI framework anchors on the following five pillars:

GenAI enablement

Adoption of GenAI models through enablement campaigns and on-demand trainings

Cost allocation

Cost allocation and mapping of GenAI models including shared services cost

Model optimization

Continuous cost optimization of the GenAI models

Pricing model

Understanding of the various pricing construct of GenAI models

Value reporting

Reporting of GenAI cost and value metrics for technology, business, and finance teams to manage and track

Modality

Text | Code | Image | Video | Audio and music | Dialogue

Gen AI enablement

To ensure a successful FinOps integration with your organization, all levels of stakeholders need to understand how their roles and responsibilities are now refocusing to include a fiscal lens. Providing persona-based training curricula, from technical and financial roles all the way to the C-Suite, about how to use generative AI in their FinOps workflows unlocks its full power throughout the organization.

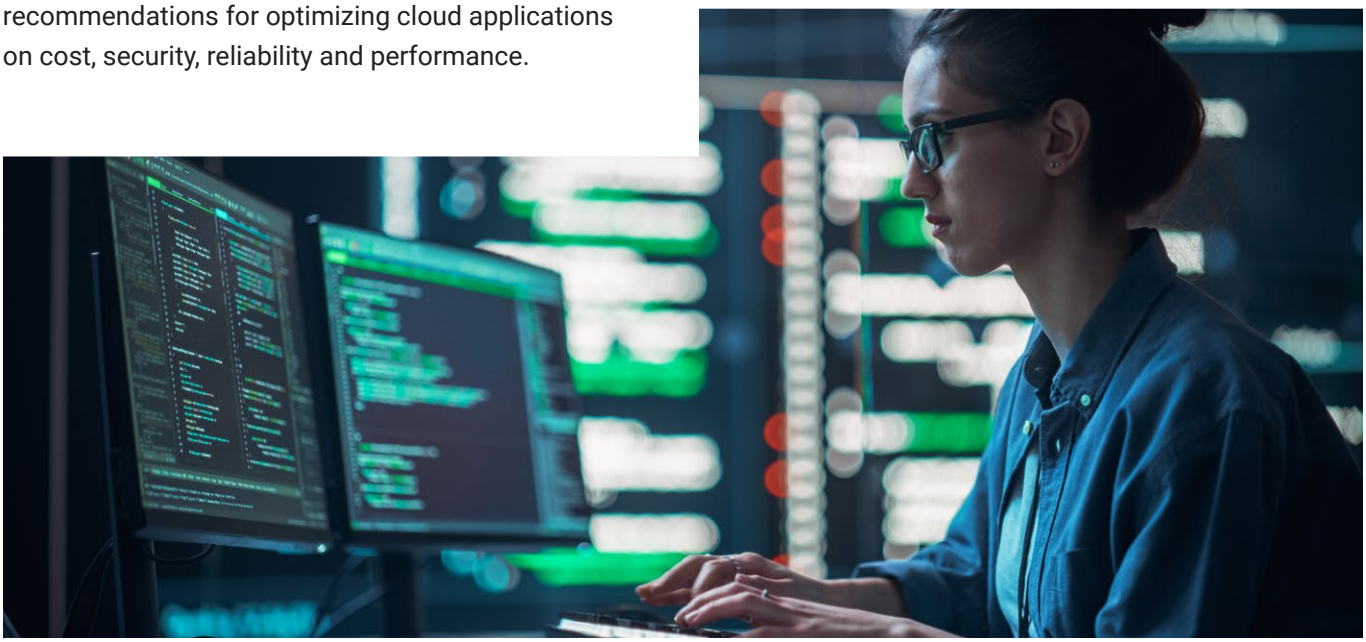
To create a seamless experience for Google Cloud customers, we're building FinOps intelligence directly into our products such as Duet AI for Google Cloud. Duet AI in Google Cloud is an always-on collaborator that offers generative AI-powered assistance to a wide range of Google Cloud users, including developers, data scientists, and operators. A key feature of Duet AI for Google Cloud is assisted operations providing FinOps practitioners a deep well of tactics, best practices, and expertise at their fingertips. For example, Duet AI for Google Cloud can retrieve "How-to" knowledge about Google Cloud infrastructure, deployment and leading practice and provide recommendations for optimizing cloud applications on cost, security, reliability and performance.

This is why a comprehensive training curriculum is essential. The curriculum should be tailored to the specific needs of each persona, from engineers to business and finance leaders.

Engineers

Engineers will need to deeply understand the technical aspects of generative AI. They will need to be able to design and implement generative AI solutions, and troubleshoot and maintain them. The training curriculum for engineers should cover topics such as:

- The business value of generative AI
- Use cases for generative AI
- The ROI of generative AI
- How to prioritize generative AI projects
- Getting the most out of Duet AI for Google Cloud



Business and finance leaders

Business and finance leaders play a crucial role in realizing the full business value of generative AI. They need to be able to identify opportunities where generative AI can be used to improve efficiency, reduce costs, and generate new revenue. The training curriculum for business and finance leaders should cover diverse topics such as:

- The business value of generative AI
- Use cases for generative AI
- The ROI of generative AI
- How to prioritize generative AI projects
- Getting the most out of Duet AI for Google Cloud

We can maximize the impact of generative AI by taking an inclusive approach to training. A robust curriculum tailored to diverse roles and skills empowers everyone across an organization to harness the technology's possibilities. With the right training and generative AI tools, the skills and knowledge of every team member amplify as their productivity increases.

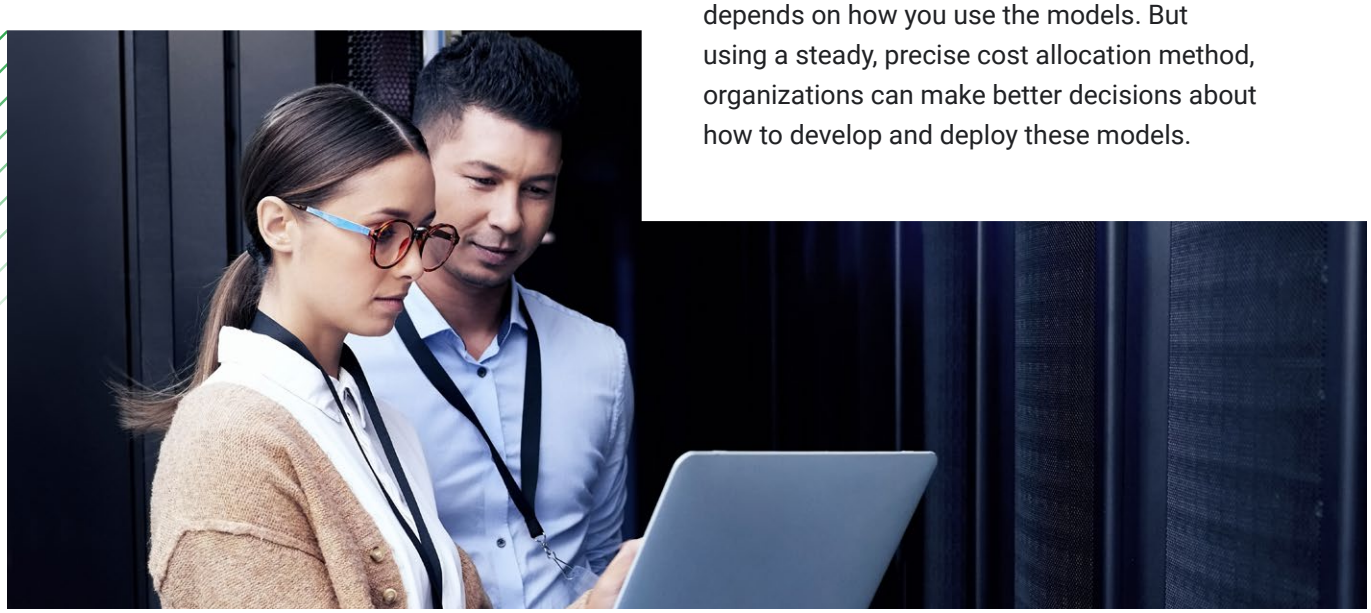
Cost allocation

Precise cost allocation is critical to maximizing the value of generative AI. Understanding the lifetime cost of a model, from service costs such as security and storage, to training, tuning, and monitoring is key to realizing its impact on your bottom line.

The cost of generative AI models can be allocated into a variety of different buckets, including:

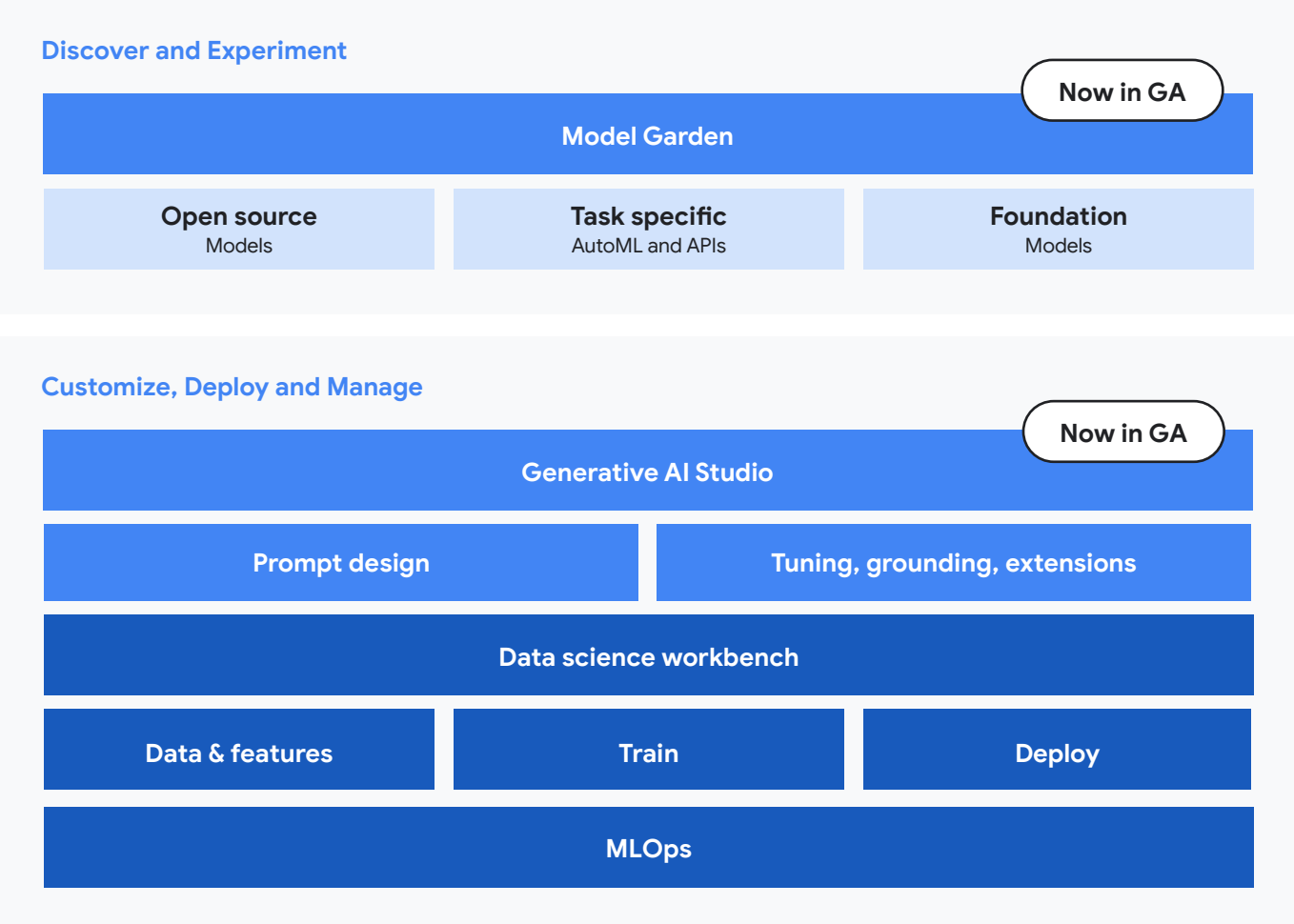
- **Training costs:** These costs include the resources required to train the model, such as TPU and GPUs.
- **Serving and inferencing costs:** These costs include the resources required to run the model in production, such as compute, storage, and networking.
- **Shared services costs:** These costs include the resources that are used by both the serving and training models, such as storage, security, and monitoring.

Allocating generative AI costs helps make smarter choices. The best way to allocate expenses depends on how you use the models. But using a steady, precise cost allocation method, organizations can make better decisions about how to develop and deploy these models.



Model optimization

Choosing the right generative AI model for your use case is critical to ensure efficient use of the services. Once you have the generative AI model deployed, you have to continuously monitor and identify key areas for optimization.



1. **Data & Features:** The process of experimentation allows you to test different model architectures, hyperparameters, and training data to find the best combination for your specific use case. This lets you quickly and cost effectively validate assumptions, confirm hypotheses, and identify the best modeling approach. Once you've identified the right approach, you can scale up to train the full model using more powerful compute resources and accelerators. Experimentation enables rapid iteration at low cost to optimize your model.

2. **Train:** This is the process of feeding data into the model and training it to learn how to generate the desired output.

   Select the appropriate machine type for training models - Typical ML training workloads fit N1 machine types, where you can attach many types of GPUs. Frameworks like TensorFlow and PyTorch benefit from GPU acceleration, while frameworks like scikit-learn and XGboost do not. Hence, it is best to use memory-optimized machines when training large scikit-learn models

   For computationally intensive workloads that require a large number of GPUs to complete training use reduction server, a new Vertex AI feature that optimizes bandwidth and latency of multi-node distributed training on NVIDIA GPUs for synchronous data parallel algorithms. Synchronous data parallelism is the foundation of many widely adopted distributed training frameworks, including TensorFlow's MultiWorkerMirroredStrategy, Horovod, and PyTorch Distributed. By optimizing bandwidth usage and latency of the all-reduce collective operation used by these frameworks, Reduction Server can decrease both the time and cost of large training jobs.

3. **Deploy:** This is the process of making the model available for use in production.

   Use Batch Predictions for offline predictions on large datasets - The Batch Prediction service runs a distributed data processing job at scale for better throughput and lower cost than online prediction. In contrast to online predictions, Google takes care of spinning up the infrastructure, handling the batch prediction, and automatically shutting down the machines. If you need to process large amounts of data quickly and cost-effectively, Batch Prediction is a great option. It is easy to use and can provide significant performance improvements over online prediction.

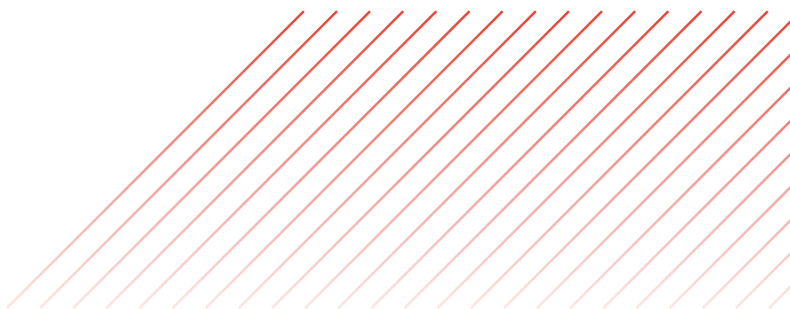
4. **MLOps:** This is the process of automating the steps involved in model development, training, deployment, and monitoring.

Furthermore, by continuously optimizing your generative AI models, you can improve their performance and ensure that they are meeting the needs of your employees, customers, or users.

Here are some additional tips for optimizing your generative AI models:

- **Use the right model architecture:** The type of model architecture you use will depend on the specific task you are trying to accomplish.
- **Tune the hyperparameters:** The hyperparameters of a model are the values that control the learning process. By tuning these hyperparameters, you can improve the performance of your model.
- **Use a large training data set:** The more data you train your model on, the better it will perform.
- **Use a distributed training approach:** A distributed training approach can help you train your model faster.
- **Monitor your model's performance:** It is important to monitor your model's performance to identify any areas that need to be improved.
- **Regularly retrain your model:** As your data evolves, your model will need to be retrained to keep up with the changes.

By following these tips, you can optimize your generative AI models and improve their performance.



Pricing model

In calculating the Total Cost of Ownership (TCO) of a generative AI use case, it is important to first understand the three main types of Large Language Models (LLM) models commonly deployed.

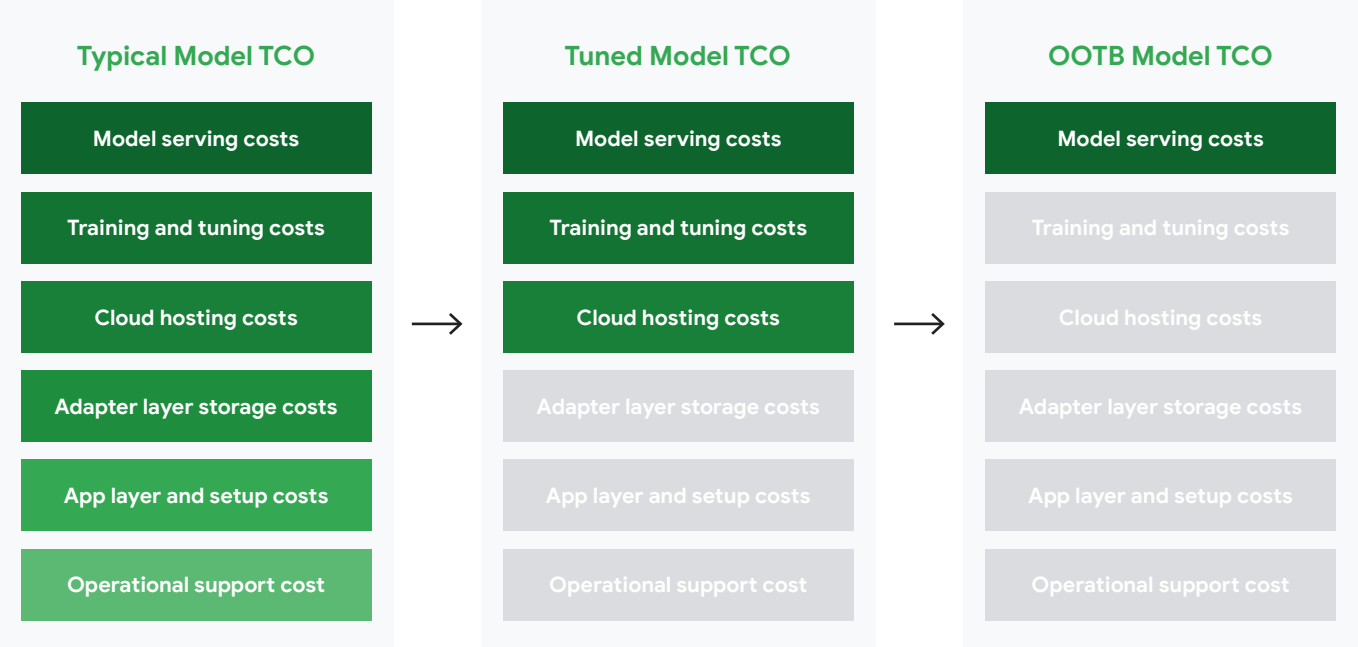


Figure: Understanding the TCO of the three LLM models

The first type of model is the Foundation model (or the typical model). This is where you build, train, and deploy the LLMs from the ground up. As such, you're incurring costs for operational support, application layers, cloud hosting, training and tuning, and the costs for model serving and inference.

The second type of LLM, is a Tuned LLM model, where most enterprises are currently piloting their generative AI use cases. The TCO for the tune model oftentimes includes the model serving costs, the training and tuning costs, cloud hosting costs, and sometimes operational support cost as well.

The third type of model is the Out of the Box model, where users are leveraging the service provider LLM and calling the APIs to analyze and generate new contents. A common example of this is Google's Bard chat service.

There are various pricing constructs for generative AI models, including token-based pricing and characters-based pricing. Understanding how the various models are priced will be essential for conducting a cost-benefit analysis.

In a token-based pricing model, the cost of generating text is based on the number of tokens in the output. A token is a unit of text that can be a word.

In a characters-based pricing model, the cost of generating text is based on the number of characters in the output. A character is a single letter, number, or symbol.

The pricing model used by a generative AI model will depend on the specific model and the provider. For example, the PaLM text bison API model uses a characters-based pricing model. This means that the cost of generating text with the PaLM text bison API model is based on the number of characters in the output.

It is important to understand the pricing model of a generative AI model before using it, so that you can make an informed decision about which model is the best fit for your project's needs

Value reporting

Calculating the Total Cost of Ownership (TCO) for generative AI models is important, but quantifying the business and financial value of these models is equally essential, and even more challenging to do. One effective approach to this is by conducting a very specific pilot use case and iterating through a trial run.

By focusing on a specific use case, you can isolate the variables and more easily measure the impact of the generative AI model. This targeted approach will help you to understand the value that the model is creating for your business and to make more informed decisions about how to use it going forward.

For example, a pilot could test using generative AI for a chatbot that answers customer questions and provides recommendations. Tracking metrics like number of bot interactions, customer satisfaction levels, and impact on retention would reveal how the AI affects customer satisfaction. The focused use case isolates variables to directly measure impact.

This type of pilot use case provides you with valuable data that you can use to quantify the business and financial value of the generative AI model. From there, use this data to make decisions about how to scale the use of the model across your organization.

It is important to note that quantifying the value of generative AI models is an iterative process. As you use the models in more and more ways, you will learn more about their value and be able to make more accurate predictions about their future impact.

Here are some tips for getting started with generative AI for Cloud FinOps:

- **Start with a small pilot project:** This will help you gain hands-on experience with the technology and see the different ways it can benefit your organization.
- **Choose the right generative AI tool for your needs:** There are many different generative AI tools available, so it's important to choose one that is appropriate for your specific needs.
- **Get buy-in from key stakeholders: I**t's important to get buy-in from key stakeholders before you start using generative AI for Cloud FinOps. This ensures greater visibility, resource allocation and advocates.
- **Monitor your results:** It's important to monitor your results and make adjustments as needed.

Google Cloud

The new horizon of cloud FinOps powered by AI technologies

Cloud FinOps is evolving rapidly to meet the complex challenges of today's digital businesses. Generative AI represents the next frontier - enabling a quantum leap in data-driven insights, automation, and optimization. The guidance in this paper shows how to harness AI's potential to transform your FinOps program. Start by proving value in a focused pilot. Then expand thoughtfully to weave AI throughout your cloud operations. Continuously optimize costs. Enable data-driven decision making at scale. Accelerate your FinOps maturity.

This is the future of Cloud FinOps. AI-powered. Insight-driven. Continuously efficient. Google Cloud can help you get there. Our experts can assess your needs, prove out an AI pilot, and chart a course to fully integrate generative intelligence across your cloud.

Don't wait. Contact Google Cloud today to bring the power of AI into your FinOps program. Equip your organization with the cloud financial operations maturity needed to serve your customers and empower your employees in today's demanding digital world.