A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

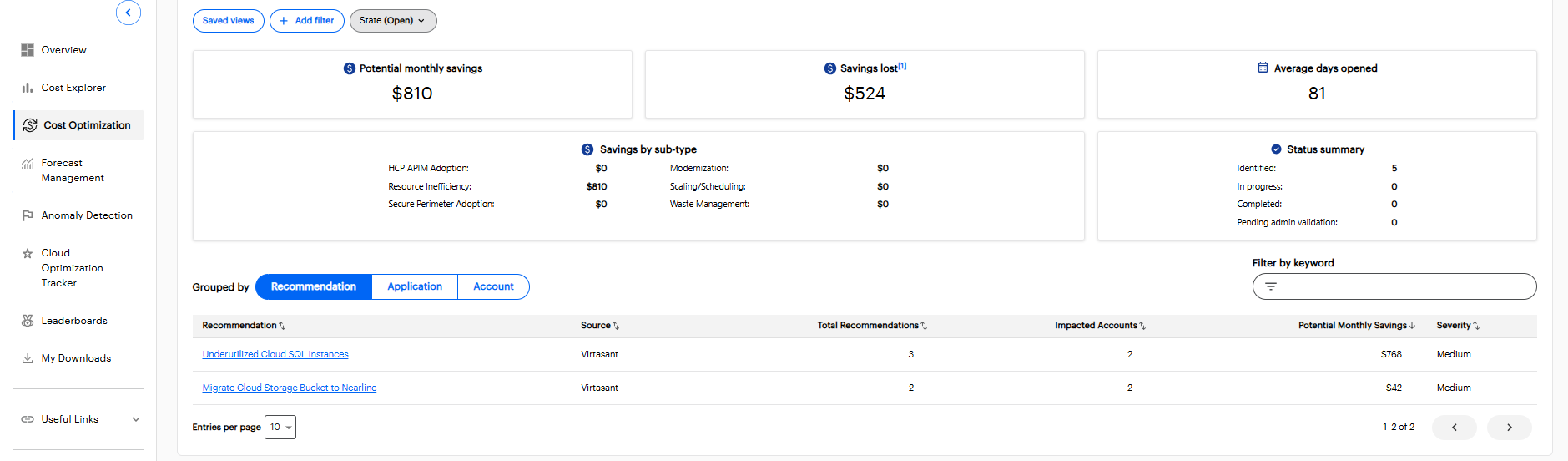
AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.



A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Overview**

Public cloud forecast management is a critical practice for businesses leveraging cloud services. It involves analyzing historical usage data, market trends, and business growth projections to accurately predict future cloud resource needs. By leveraging sophisticated forecasting tools and techniques, organizations can optimize their cloud spending, avoid over-provisioning, and ensure they have the necessary resources to meet their evolving demands.

Accurate forecasts allow businesses to proactively adjust their cloud infrastructure, scale resources efficiently, and minimize costs. Furthermore, by forecasting resource requirements, organizations can improve their cloud cost management strategies, identify potential bottlenecks, and ensure smooth business operations. Effective public cloud forecast management is essential for achieving optimal cloud performance, cost-efficiency, and scalability.

**What does the forecasting process involve**

The public cloud forecasting process combines data analysis, predictive modeling, and human expertise to anticipate future resource needs. It also involves analyzing trends and seasonality, incorporating known future roadmaps and business initiatives, potential growth, technology upgrades, and emerging business needs. Continuous monitoring and feedback mechanisms are critical for refining the forecast and ensuring resource optimization, cost control, and a seamless cloud experience. Ultimately, public cloud forecasting aims to strike a delicate balance between providing sufficient capacity to meet future demands while avoiding unnecessary wastage.

**What are the benefits of providing an accurate Public Cloud forecast**

Providing an accurate public cloud forecast offers numerous benefits. A precise public cloud forecast empowers organizations to achieve cost efficiency, maintain optimal performance, and drive business growth.

* Secure larger discounts on resources through reservations purchases.
* Improves capacity planning, enabling organizations to do bigger commits and procure deeper discounts.
* Facilitates informed decision-making, allowing organizations to make strategic choices about cloud adoption, resource allocation, and cost management.

A precise public cloud forecast empowers organizations to achieve cost efficiency, maintain optimal performance, and drive business growth.

**What are the common issues while Forecasting**

Forecasting public cloud spend is a challenging endeavor, fraught with several common issues of Over/Under Forecasting.

**What is over-forecasting**

Over-forecasting in Public Cloud is a common issue that can lead to unnecessary expenses and budget misalignments. This can happen due to factors like inaccurate historical data analysis, overly optimistic projections, or lack of optimization strategies. The consequences of over forecasting include increased cloud costs, inefficient resource utilization and loosing deeper discounts.

To mitigate over-forecasting, teams must implement robust forecasting methodologies, leverage historical data effectively, and adopt a proactive approach to resource optimization and demand management.

**What is under-forecasting**

Under-forecasting in Public Cloud occurs when organizations underestimate their future cloud resource needs. This can lead to several negative consequences, unexpected costs from exceeding usage limits, and disruptions caused by constrained budgets. It can be caused by factors such as inaccurate workload projections, insufficient data analysis, and a lack of understanding of cloud resource consumption patterns.

To mitigate under-forecasting, teams must implement a robust forecasting methodology that incorporates historical data, future business growth projections, and dynamic workload analysis.

**How is it different to forecast Public Cloud spend Vs Shared Services**

Forecasting for applications on Public Cloud or Public Cloud shared services is one and the same. There are few additional parameters to think about when its shared services depending on the service. For instance Data Platforms - Public Cloud. This Shared service comprises of Clusters as a service and Entity Builder as a service. Public Cloud spend for Clusters as a service depends on the size of the clusters provisioned. While Entity Builder as a service spend is dependent on the number of canonicals and the complexity of the canonicals.

**Use CSP Pricing Calculators to Calculate Accurate Forecast**

**Note**

**Use**[**Azure's cost calculator**](https://azure.microsoft.com/en-us/pricing/#explore-cost)**and take off 42% because of UHG Price Improvements**

**Use**[**AWS's cost calculator**](https://calculator.aws/#/)**and take off 30% because of UHG Price Improvements**

**Use**[**GCP's cost calculator**](https://cloud.google.com/products/calculator)**and take off 24% because of UHG Price Improvements**

**Use**[**Mongo DB's pricing**](https://www.mongodb.com/pricing)**and take off 14.5% because of UHG Price Improvements**

**UHG’s way to forecast Public Cloud**

**FinOps Role:**

Public Cloud FinOps team owns Public Cloud forecasts. We play a crucial role in forecasting by combining financial expertise with technical understanding. We analyze historical cloud usage data, identifying patterns and trends to predict future consumption. Through collaboration with application teams and finance stakeholders, we develop accurate forecasts for various cloud services (Azure, Aws, GCP, MongoDB, etc.). These forecasts are used to establish budgets, optimize resource allocation, and negotiate favorable contracts with cloud providers. By leveraging data-driven insights, we enable UHG to manage cloud costs effectively, driving accountability, and achieve our business objective to migrate 75% of our infrastructure to Cloud.

**Forecast Cycles:**

Forecast cycles at UHG occur quarterly, but teams are encouraged to update forecasts whenever there are changes in the roadmap or potential cloud cost optimizations that may impact future consumption. The forecast cycles include:

1. 2+10F: This cycle involves 2 months of actual cloud usage data and 10 months of forecasted data. It helps determine if the forecasts align with the budgets set during the previous 8+4 cycle.
2. 5+7F: This cycle involves 5 months of actuals and 7 months of forecasted data. It is used to analyze the rate card and suggest possible increases or decreases for the next year.
3. 8+4F: This cycle involves 8 months of actuals and 4 months of forecasted data. It is utilized to set official budgets for application teams for the next year.
4. 10+2F: This cycle is optional and decided by the finance team. It typically occurs when there is volatility in forecasts for specific services. It considers 10 months of actuals and 2 months of forecasted data.

**Forecast Accuracy**

The aim of UHG's forecasting process is to achieve 93% forecast accuracy. Teams are encouraged to use [**Cost Explorer**](https://costmanagement.optum.com/cost-explorer) to analyze and understand their current spend in detail, enabling them to make informed decisions about future spend and potential savings. Teams can also leverage Predictions generated by Finops team as a baseline to come up with their forecasts.

**Potential Monthly Savings**

We request every application to leverage our Cost Optimization capabilities and think about potential savings that team can do during the current year and next year. App teams should incorporate Cost Optimizations into their roadmap and based on the timing, tweak the forecasts to reflect the savings. Learn more about Cost Optimization [**here**](https://docs.hcp.uhg.com/cost-management/cost-optimization-overview).

**Finance Collaboration**

We collaborate with finance partners like OT Finance and Segment Relationship Managers (SRM’s) throughout the year. We partner during each forecast cycle to come up with an accurate forecast for Public Cloud, and engage Application teams and SRM’s every month to address variances between forecasts and actual spend.

**Usage vs. Forecast Reporting**

Usage vs. forecast reporting is used to capture reasons for usage anomalies, and the Public Cloud FinOps teams work closely with application teams to update forecasts accordingly. Outlier app details are shared with SRMs and finance leaders to ensure app teams can course correct their forecasts. Also see [**Anomaly Detection Overview**](https://docs.hcp.uhg.com/cost-management/anomaly-detection-overview)

**AI Cost Estimation Best Practices**

AI/ML engineers often prioritize model performance without fully considering the cost implications of their design and deployment choices. This can lead to inefficient use of cloud resources and unexpected expenses. To build scalable and cost‑effective AI systems, engineers need to integrate cost awareness into every stage of the AI lifecycle—from development to deployment and monitoring.

**Introduction**

This guide provides FinOps practitioners, developers, and engineering managers with a comprehensive framework to **forecast and optimize AI service costs** in cloud environments. It focuses on **text‑based large language models (LLMs)** and outlines deployment models, cost drivers, forecasting strategies, and best practices.

**AI Deployment Models**

Today enterprises deploying AI offerings typically use three options to meet requirements across quality, privacy, accuracy, performance, and costs. Understand where your use case fits in:

| **S No** | **Model** | **Description** | **Pros** | **Cons** |
| --- | --- | --- | --- | --- |
| 1 | Third‑Party Closed Source | Managed services (e.g., OpenAI, Google, Microsoft) | Easy setup, high quality | Expensive, vendor lock‑in |
| 2 | DIY on Cloud | Build in‑house using cloud tools (e.g., Bedrock, Vertex, Azure AI) | Full control, IP ownership | High complexity, longer time‑to‑market |

**Key Cost Drivers**

**API Usage (Tokens In/Out)**

* **What it means:** Most AI services (like OpenAI, Azure OpenAI, or Claude) charge based on the number of tokens processed—both input (prompt) and output (response).
* **Example:** If your app sends a 500‑word prompt (~750 tokens) and receives a 300‑word response (~450 tokens), you’re billed for ~1,200 tokens. At $0.03 per 1,000 tokens, that’s **$0.036 per request**.

**Data Storage and Transfer**

* **What it means:** Storing training data, embeddings, logs, and intermediate results incurs storage and I/O costs.
* **Example:** Storing 1 TB of training data on AWS S3 Standard costs ~$23/month. Frequent access or transfer to compute nodes adds to the bill.

**Model Training and Fine‑Tuning**

* **What it means:** Training or fine‑tuning models consumes compute and storage resources.
* **Example:** Fine‑tuning a model on 100,000 labeled examples might take 10 hours on a high‑end GPU, costing $30/hour = **$300**.

**Model Serving**

* **What it includes:** Running the model in production to respond to user queries or process data.
* **Example:** Hosting a sentiment analysis model on AWS SageMaker to process 10,000 customer reviews per hour.

**Cloud Provider Pricing Models**

* **What it means:** Discounts for reserved capacity vs. on‑demand pricing.
* **Example:** GPT‑4 on‑demand at $0.06 / 1K tokens vs. reserved capacity at $0.03 / 1K tokens if fully utilized.

**Estimation Techniques**

**Token‑based cost estimation for API usage**

Multiply expected usage (e.g., tokens) by unit cost.

**Example:**

* 1M requests/month × 1,200 tokens/request × $0.03 / 1K tokens = **$36k/month**.
* After application of Enterprise discount in Azure (42%) the cost is ~**$21k/month**.

**Benchmarking and capacity planning**

* **Benchmarking** measures performance characteristics of an AI model—throughput (requests per second), latency (response time), and token usage—under realistic workloads.
* **Capacity Planning** estimates and provisions the necessary infrastructure (e.g., GPU, memory, throughput units) to meet demand while optimizing cost and performance.

**Worked example:**

1. **Benchmarking findings**
   * Each request uses ~1,200 tokens.
   * The model can handle ~300,000 tokens per minute.
   * Average latency is 500 ms per request.
2. **Expected traffic**
   * Peak: 10,000 requests/hour for 10 hours  
     → 1,200 tokens/request × 10,000 req/hr × 10 hr = **120,000,000 tokens**
   * Non‑peak: 2,000 requests/hour for 14 hours  
     → 1,200 tokens/request × 2,000 req/hr × 14 hr = **33,600,000 tokens** (≈ 37M in rounded estimates)
3. **Daily total**
   * **~157,000,000 tokens/day** (depending on rounding for non‑peak)

**Guidance**

* **Token‑based estimation** is good for cost forecasting.
* **Latency** considerations are essential for **performance planning**.
* Balance token usage and latency requirements to meet both cost and performance goals.

**Fine‑tuning and quantization impact analysis**

**Fine‑tuning** adapts a pre‑trained model to a specific task/dataset. It can improve accuracy and relevance but adds training and storage cost.

**Quantization** reduces parameter precision (e.g., 32‑bit → 8‑bit), shrinking model size and compute needs and reducing inference costs.

**Model Deployment Example: Llama 2 13B for Sentiment Analysis**

| **Aspect** | **Without Quantization** | **With Quantization (e.g., q4)** |
| --- | --- | --- |
| Model Size | ~26 GB | ~6.5 GB |
| Required GPU | 96 GB instance | 24 GB instance |
| Monthly Hosting Cost | ~$2,500 | ~$500 |
| Cost Savings | — | Over **80% reduction** |

**Fine‑Tuning Scenario**

* Dataset: 10,000 labeled customer reviews
* Training Cost: **$300** (one‑time upfront)

**Impact on Performance and Cost**

* Improves model quality and reduces prompt “few‑shot” instructions
* Reduces token usage per request
* Lower token usage → reduced API or compute cost per request

**Summary of Impact**

* **Fine‑Tuning:** Improves quality/efficiency with initial upfront investment
* **Quantization:** Significant ongoing runtime savings; enables lower‑cost infrastructure
* **Combined:** Better accuracy, performance, and cost‑effectiveness

**Use of Cost Calculators**

* **Azure Pricing Calculator:** https://azure.microsoft.com/en-us/pricing/calculator/

**Forecasting Framework**

**Step‑by‑Step Guide for Engineers**

1. Define the AI workload and use case (e.g., chatbot, summarization, code generation).
2. Choose the appropriate deployment model based on control, cost, and performance needs.
3. Identify key cost drivers relevant to the workload. and break down the cost components: compute, storage, data transfer, and API usage.
4. Estimate token usage or infrastructure requirements.
5. Calculate the forecast cost.
6. Apply provider pricing models to each cost component (e.g., instance hours × rate).
7. Aggregate costs and apply **Enterprise discounts**:
   * https://docs.hcp.uhg.com/cost-management/enterprise-discounts
8. Utilize **App Cost Management** for:
   * **Forecasting:** Enter/update/track forecasted application spend for Public Cloud and Data Platform → https://costmanagement.optum.com/forecast-management
   * **Cost Explorer (Monitoring):** Visualize, understand, and manage AI costs over time → https://costmanagement.optum.com/cost-explorer
   * **Optimization:** Identify mismanaged AI resources, eliminate waste, right‑size → https://costmanagement.optum.com/cost-optimizations
   * **Anomaly Detection:** Find forecast and cost anomalies at the application level → https://costmanagement.optum.com/anomaly-detection

[Get access to your application in App Cost Management - Cost Management | HCP Docs](https://docs.hcp.uhg.com/cost-management/get-access-to-the-cost-management-app)