



Cut down infrastructure costs by 20% **with ML-based capacity planning**

- Your ultimate guide to smarter infrastructure management

Table of contents

■	Introduction	3
■	How accurate capacity planning prevents downtime and drives cost savings	5
■	Achieving precise capacity predictions with ML-powered multivariate forecasting	11
■	Creating tailored resource projections that align with business realities	16
■	Streamlining cloud resource selection with ML-powered cloud cost analytics	19
■	Conclusion	22
■	About ManageEngine Analytics Plus	23

Introduction

As technology evolves, innovative solutions like AI, the Metaverse, and other cutting-edge advancements are finding increasing application in the business world. This has led to exponential growth in IT infrastructure spending, reflecting organizations' growing reliance on digital transformation initiatives. **According to Gartner®, worldwide IT spending is expected to reach^[1]**

\$5.74 trillion in 2025

driven primarily by cloud-based services and data center modernization to support AI and generative AI technologies.

However, this rapid growth comes with a downside. Inefficient resource utilization, over provisioning, and a lack of proactive capacity planning often lead to suboptimal returns on infrastructure investments. As organizations scale their digital operations, such inefficiencies can create significant cost drains.

IT infrastructure now accounts for a significant share of IT budgets, yet effectively managing it remains a critical challenge for IT teams. In today's turbulent economic climate, organizations can no longer afford to overlook these inefficiencies. IT leaders are under pressure to optimize infrastructure usage, improve resource efficiency, and accelerate ROI while maintaining scalability and productivity.

The role of capacity planning

Effective capacity planning is pivotal in bridging this gap. It enables businesses to balance growth with cost control, ensuring sustained operations and streamlined services. Beyond cost savings, it reveals hidden inefficiencies, empowers leaders to address bottlenecks early, and aligns IT investments with business goals.

However, traditional capacity planning methods, which rely heavily on historical data and reactive approaches, fail to account for the complexities of modern IT environments, such as:

- **Dynamic workloads:** Seasonal fluctuations and business expansions can lead to demand surges that historical data alone cannot anticipate.
- **Complex hybrid environments:** The integration of on-premises, cloud, and edge computing resources complicates capacity planning, requiring more sophisticated approaches.
- **Over-provisioning and underutilization:** Without precise forecasting, organizations risk allocating too many or too few resources, leading to increased costs or performance issues.

Addressing these gaps in capacity planning is essential for enhancing IT operational efficiency and reducing unnecessary infrastructure cost drains.

Embrace the power of ML-driven capacity planning

To address these gaps, organizations must shift to ML-powered capacity planning, which leverages predictive analytics, multivariate forecasting, and other advanced analytics capabilities to analyze historical usage trends, workload patterns, business growth plans, and even external factors like market conditions. By optimizing resource usage based on these insights, ML-driven capacity planning delivers significant benefits.

Unlike traditional methods, the ML-based approach helps predict demand, adjust resources in real-time, and eliminate inefficiencies with unparalleled precision. This can enhance agility, facilitate optimum operational efficiency, and reduce infrastructure costs by 20% on average (based on data gathered by customers who have used our **ROI calculator**)^[2].

This e-book will explore how ML-based capacity planning revolutionizes infrastructure management, empowering businesses to reduce costs, maximize ROI, and stay ahead in an ever-evolving digital landscape.

01 | How accurate capacity planning prevents downtime and drives cost savings

ML-powered analytics can deliver a modern, data-driven solution to enhance the precision and efficiency of capacity planning. By accurately forecasting resource needs, this approach minimizes unexpected downtime and its associated costs. Beyond operational reliability, the cost savings achieved through ML-driven capacity planning can be significant.

The sections ahead will highlight scenarios and opportunities within IT infrastructure management where ML-based capacity planning acts as a game-changer.

The best starting point in this journey is identifying the most common IT infrastructure cost drains and then implementing targeted strategies to address and optimize these areas. Unplanned IT downtime is a significant financial drain for organizations, often stemming from infrastructure-related capacity constraints. Research consistently highlights the heavy toll downtime has on revenue, productivity, and reputation.

According to a **2024 report by Splunk^[3]**, unplanned downtime costs Global 2000 companies a staggering \$400 billion per year, amounting to around 9% of their total profits. The severity of this problem is further underscored by the **Uptime Institute's 2022 Outage Analysis^[4]**, which revealed that over 60% of downtime results in at least \$100,000 in total losses—a substantial increase from 39% in 2019.

These numbers reaffirm the importance of promptly identifying and addressing capacity shortages before they escalate into failures, resulting in application or service disruptions and leading to significant financial and productivity losses.

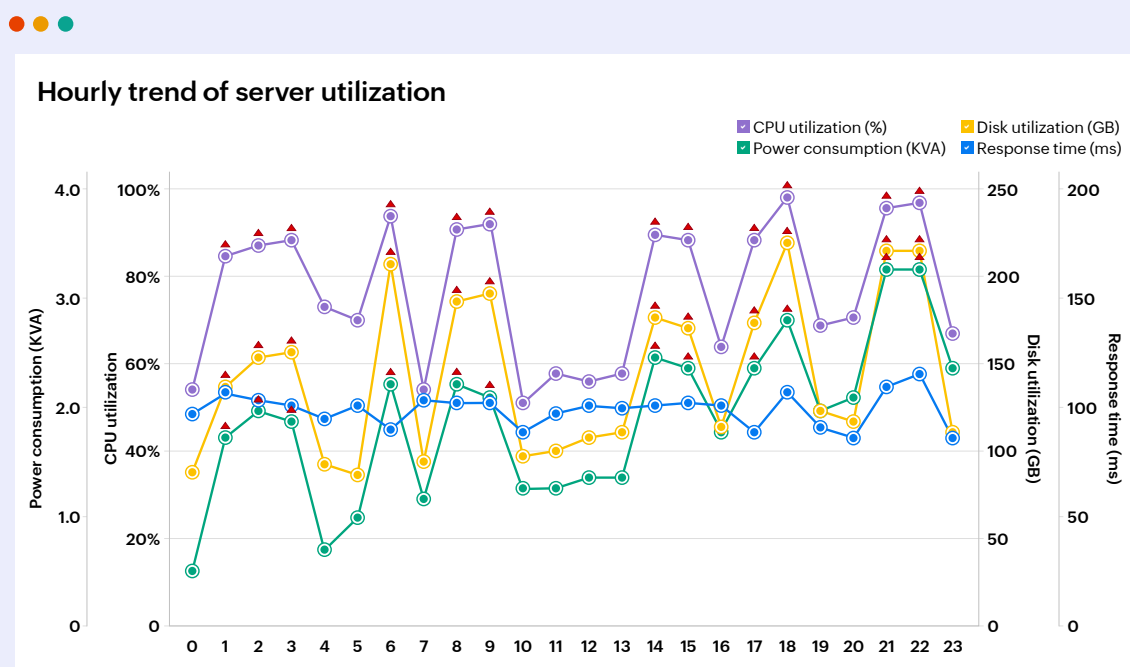
ML-based analytics excels in this area, offering organizations the ability to:

- **Track critical indicators:** Metrics such as CPU utilization, disk usage, and application response time can signal impending capacity overloads.
- **Prevent failures proactively:** By analyzing historical trends and current conditions, organizations can predict when additional resources are required, mitigating the risk of downtime.

Many IT teams rely on static or dynamic threshold-based alerts to monitor these capacity indicators. However, these methods often fail to adapt to evolving usage patterns and unique business conditions. The result?

- False alarms leading to unnecessary resource allocation.
- Missed warnings on critical anomalous deviations that do not breach preset thresholds, resulting in capacity overloads and subsequent downtime.
- Alert fatigue: Excessive notifications that desensitize technicians, causing them to overlook high-priority issues.

Leveraging machine learning-based analytics can provide a more effective solution to capacity planning challenges. By implementing automated anomaly detection powered by ML, IT teams can configure alerts that trigger when a metric or indicator deviates significantly from the average observed value. Anomaly-based alerts can dynamically adjust the alert conditions based on real-world scenarios and evolving business requirements, offering a more responsive approach. These accurate warnings enable IT teams to allocate additional resources preemptively, preventing overload and ensuring uninterrupted service availability.



The analysis illustrates hourly trends in utilization metrics of an application server hosting an enterprise service desk application. ML-based algorithms identify anomalous usage spikes at specific intervals, signaling capacity overloads even when static thresholds are not breached.

Traditional threshold-based systems often struggle to distinguish between normal fluctuations and true anomalies. This can lead to overlooked spikes in utilization—direct signs of capacity overload that soon result in application downtime. However, automated anomaly detection can provide an early warning, allowing NOC teams to allocate additional resources before infrastructure breaks down. This helps prevent application lag and reduce costly downtime incidents.

Pinpointing the root causes of infrastructure incidents

While automated anomaly detection enables organizations to adopt a more efficient infrastructure maintenance strategy, additional measures can further ensure sustained IT and business operations. Beyond identifying anomalous usage and implementing efficient maintenance, IT managers must also identify troublesome IT resources that encounter frequent downtimes and work to provision stable alternatives.

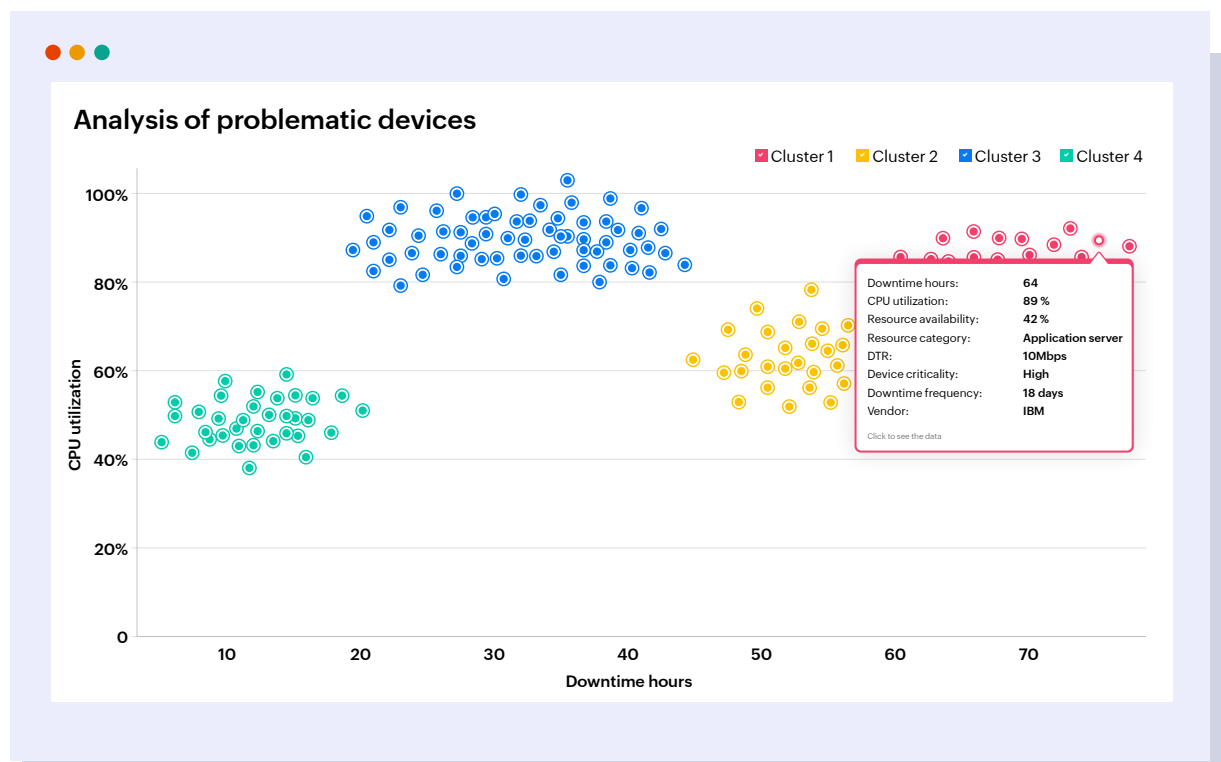
Infrastructure-related downtime cannot be completely avoided in enterprise IT, but the goal of optimized, ML-based capacity planning is to minimize business disruptions caused by IT outages. By rigorously investigating the root cause of infrastructure incidents, organizations can develop a comprehensive strategy to maintain reliable, high-performing resources and services.

Avoiding further investment in unreliable assets and replacing them with stable alternatives enables organizations to reduce recurring issues. ML-powered root cause analysis can detect troublesome resources based on their likelihood and frequency of outages and also uncover the most prominent factors leading to their outage.

However, for organizations with extensive, hybrid IT environments, identifying and isolating troublesome devices can be a daunting task. Analyzing each resource in isolation rarely provides actionable insights into overarching patterns or systemic issues. ML-based clustering solves this by categorizing devices into groups based on behavioral patterns and failure indicators.

Clustering reveals critical insights into the why, how, and what of infrastructure downtime, i.e., why resources fail, how they behave leading up to failure, and what common characteristics they share. This information simplifies remediation by offering a holistic view of problematic resources and enabling NOC teams to focus their efforts effectively.

The below visualization demonstrates how this works:



The analysis above groups over 200 resources from different categories into four clusters based on key attributes relevant to capacity planning.

Cluster 1: Replace immediately— Represents high-priority devices with frequent and long down times.

Cluster 2: Avoid future purchases— Includes devices with moderate usage but high-frequency down times. These can be used until end of life but should be avoided in future purchase plans.

Cluster 3: Monitor and evaluate— These are resources nearing utilization limits or exhibiting early warning signs. These require careful monitoring to prevent issues.

Cluster 4: Reliable resources— These are the ideal resources with high availability, low downtime, and efficient utilization.

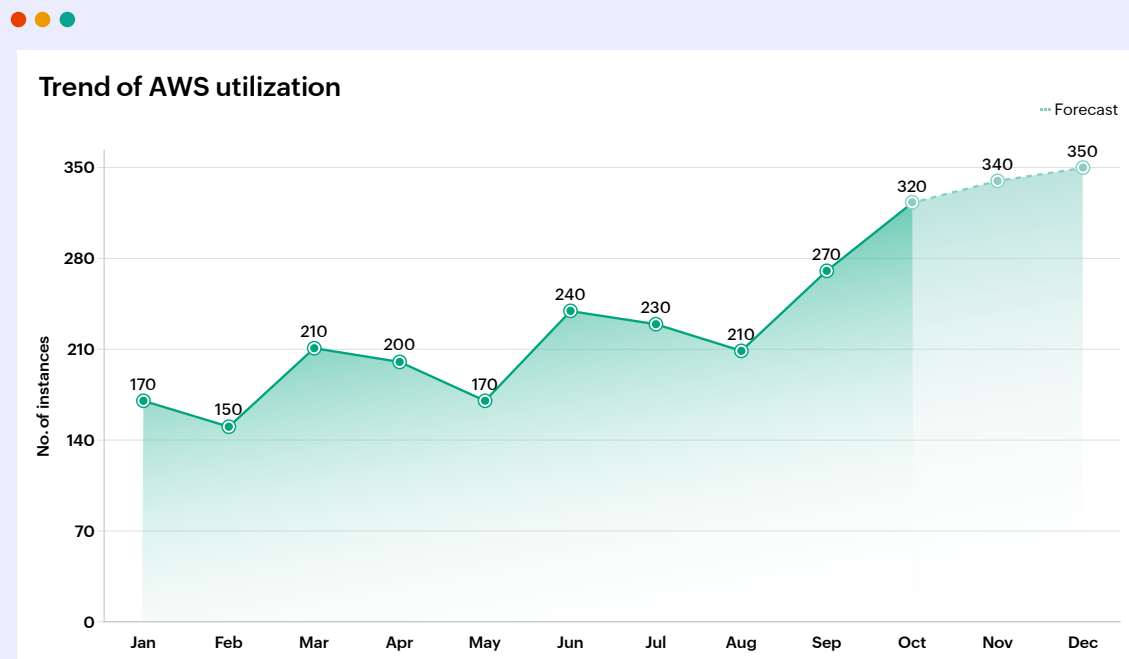
By adopting advanced ML-based techniques like anomaly detection and clustering, IT teams can proactively prevent downtime, reduce costs associated with frequent repairs or replacements, improve productivity, and maintain operational and business continuity. This strategic approach not only enhances infrastructure efficiency but also empowers businesses to scale confidently, ensuring that IT investments yield maximum returns.

02 | Achieving precise capacity predictions with ML-powered multivariate forecasting

As enterprise IT environments grow increasingly dynamic and adopt hybrid modes of operation, traditional methods of capacity planning are becoming ineffective and outdated. The conventional approach to resource capacity planning typically involves three key steps:

- **Understanding current resource utilization:** Assessing how existing resources are consumed.
- **Anticipating future needs:** Projecting future capacity requirements based on current consumption trends and planned business growth.
- **Aligning capacity with growth:** Correlating anticipated demands with business objectives to create a resource plan that ensures operational readiness.

In this, forecasting future inventory consumption is a critical step. By leveraging standard forecasting capabilities offered by analytics solutions, NOC teams can predict capacity requirements based on historical consumption patterns.



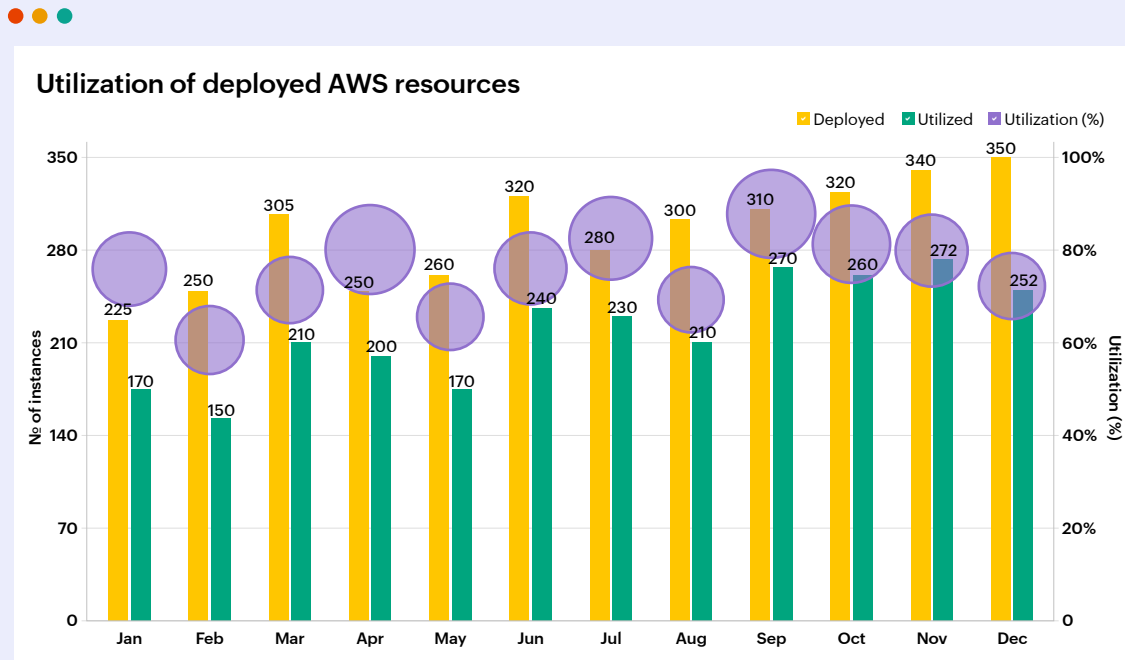
The above univariate forecasting approach, where values are predicted based on historical trends, can offer a basic estimate as a starting point for capacity planning in small-scale operations. However, as the scale of infrastructure grows, it often falls short in accounting for the nuanced and constantly evolving demands of today's complex IT environments, which combine cloud and on-premises infrastructures.

Limitations of univariate resource demand forecasting

Univariate forecasting, which relies solely on past usage trends, comes with significant drawbacks:

- **Lack of contextual insights:** These models fail to account for factors like seasonal usage spikes, emerging technologies, and business growth strategies.
- **Static projections:** These models are deployed under the assumption that historical trends will hold true in all scenarios, which is rarely the case in evolving IT ecosystems.

- **Missed interdependencies:** Hybrid infrastructures often exhibit complex interrelationships between different systems. Ignoring these can lead to over-provisioning or under-provisioning.



The above visualization clearly outlines the limitations with univariate forecasting. The analysis tracks the utilization data of instances and compares it to the actual number of allocated resources, including those that were deployed using forecasts from the previous analysis. Here, we can observe that the number of instances deployed in October, November, and December, based on the univariate forecast values, were critically underutilized.

This shows that forecasts relying solely on past usage trends may not be sufficient to provide accurate capacity projections that reflect actual resource requirements. These shortcomings can result in:

- Unanticipated capacity overloads.
- Increased costs from unnecessary resources.
- Performance bottlenecks, outages, and operational inefficiencies.

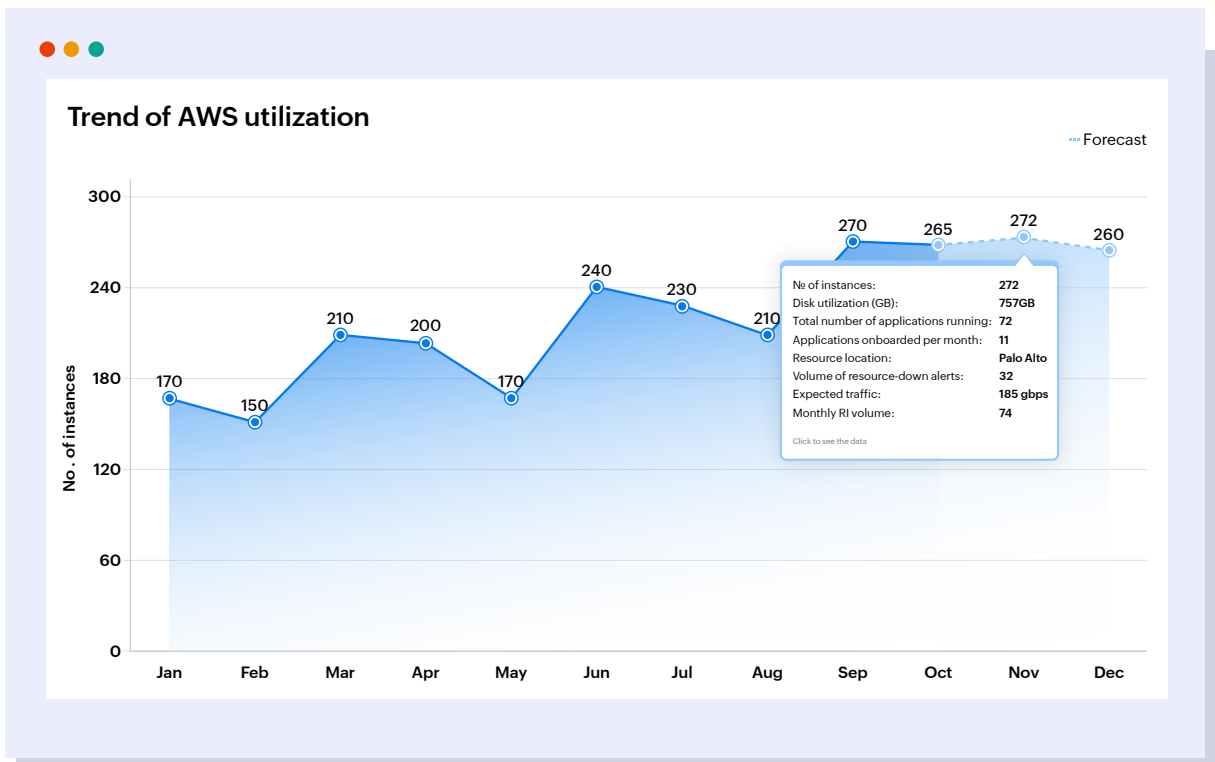
To address these challenges, organizations must adopt a forecasting methodology that factors in the multifaceted nature of hybrid infrastructures and aligns capacity planning with changes in other influencing variables spread across the IT infrastructure.

● **Introducing multivariate forecasting: A game-changer for capacity prediction**

Multivariate forecasting, powered by machine learning, transforms the capacity planning process by incorporating multiple variables from across the IT landscape.

Unlike univariate models, the multivariate forecasting engine offered by Analytics Plus can account for diverse factors from across the IT landscape such as traffic patterns, application load, and more that can directly impact resource consumption.

By considering interdependencies and broader business strategies, multivariate forecasting enables more precise resource allocation tailored to support operational efficiency and growth objectives.



This analysis highlights how multivariate forecasting aligns closely with real-world demands, reducing the divergence seen with traditional forecasting methods.

The visualization illustrates the monthly trend of AWS resource utilization using multivariate forecasting to predict future demand. With multivariate forecasting, the initial univariate AWS utilization forecast is further refined by incorporating diverse influential parameters, such as disk utilization, number of applications running, expected traffic and more. This multivariate forecasting approach results in predicted instance values for the last quarter that more accurately reflect actual consumption.

Multivariate forecasting minimizes the risk of resource over-provisioning while ensuring capacity is optimized to handle fluctuating needs effectively. This approach enhances forecasting accuracy and can eventually translate into substantial cost savings by avoiding spending on unused and underutilized resources.

Creating tailored resource projections that align with business realities

As seen in the previous section, multivariate forecasting introduces a transformative approach to capacity planning. By leveraging an adaptive forecasting engine, it enhances both the accuracy and the efficiency of capacity predictions, setting a new benchmark for effective infrastructure and resource management.

With the business landscape evolving rapidly, organizations find themselves at varying stages of digital transformation and technology maturity, navigating diverse operational environments. In such complex scenarios, IT teams are forced to expand their capacity planning beyond just standard indicators and readily available KPIs derived from IT monitoring tools. A one-size-fits-all approach often falls short in addressing the nuanced demands of dynamic markets and evolving organizational priorities. A robust and effective capacity plan must also account for tailored operational landscapes, strategic changes, objectives, and unique market conditions.

Standard out-of-the-box models and analyses provided by BI tools, while useful for general scenarios, often lack the flexibility to anticipate capacity demands shaped by specific business events such as product launches, promotional campaigns, or major infrastructure upgrades. These events demand highly customized forecasts that incorporate the complexities of an organization's operations, market trends, and long-term goals.

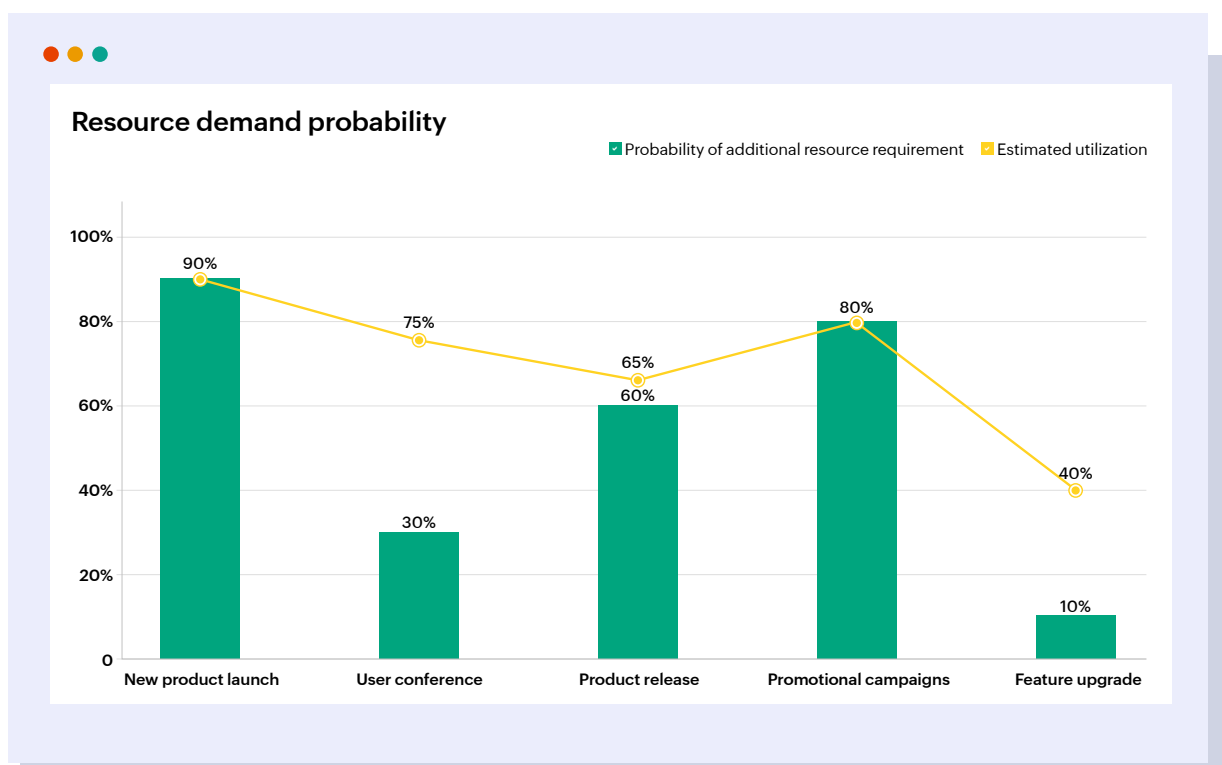
However, crafting such bespoke ML models traditionally demands specialized AI/ML expertise, considerable development time, and high financial investment—factors that make its adoption difficult for many organizations.

Enter no-code ML models

Advancements in AI/ML technologies have introduced no-code ML model building capabilities. Advanced analytics platforms like ManageEngine Analytics Plus harness these capabilities to empower IT and NOC teams to create custom models without writing a single line of code.

No-code ML models can be built in minutes, allowing IT teams to generate insights without requiring extensive AI/ML expertise. These models analyze historical resource usage during past events in the organization, help with understanding the influence of common events or initiatives on resource consumption, and generate predictions tailored to the organization's unique needs.

With this generated model, as shown in the analysis below, IT leaders can seamlessly predict the probability of additional capacity requirements, the criticality of those needs, and the expected utilization levels for specific business events or market scenarios.



The analysis showcases how custom ML models can analyze unique, upcoming strategic events to estimate the likelihood of additional resource need and it's estimated utilization.

Armed with these tailored insights, IT teams can:

- **Preempt downtime:** Allocate buffer resources in advance to avoid performance bottlenecks during critical events.
- **Leverage vendor discounts:** Plan cost-effective bulk resource procurement by taking advantage of reserved instances or committed use discounts.
- **Enhance operational efficiency:** Streamline workflows and prevent disruptions by aligning capacity with projected demand.

By integrating tailored, no-code ML models into capacity planning, organizations can tailor their resource forecasts with organization-specific initiatives and operating conditions. Such tailored insights reduce unplanned spending, minimize cost drains, and avoid business disruptions due to resource shortages.

Streamlining cloud resource selection with ML-powered cloud cost analytics

The unmatched flexibility and scalability of cloud resources has made them the backbone of modern IT operations.

Despite the widespread adoption of cloud computing, IT and business leaders are concerned about the return on their cloud investments. Hidden cost drains in cloud infrastructure prevent organizations from fully benefiting from the potential of cloud infrastructure.

The delay in achieving an ROI and the presence of unidentified cost drains have made cloud cost optimization a critical component in capacity planning.

Improper capacity planning and cloud cost management often result in avoidable issues, including:

- **Over provisioning:** Paying for resources that go underutilized.
- **Idle resources:** Keeping unused resources running unnecessarily.
- **Poor vendor fit:** Selecting resources or vendors that do not align with operational needs.
- **Higher data transfer costs:** Overspending due to unplanned or excessive data movement.
- **Scaling errors:** Failing to scale efficiently, leading to cost spikes during demand surges.

One of the primary causes of cloud cost leakage is the lack of vendor-resource alignment. Organizations often face difficulties in identifying the right vendor, pricing plans, or resource type tailored to their needs, missing out on cost-saving opportunities like provider-specific discounts and reserved instances.

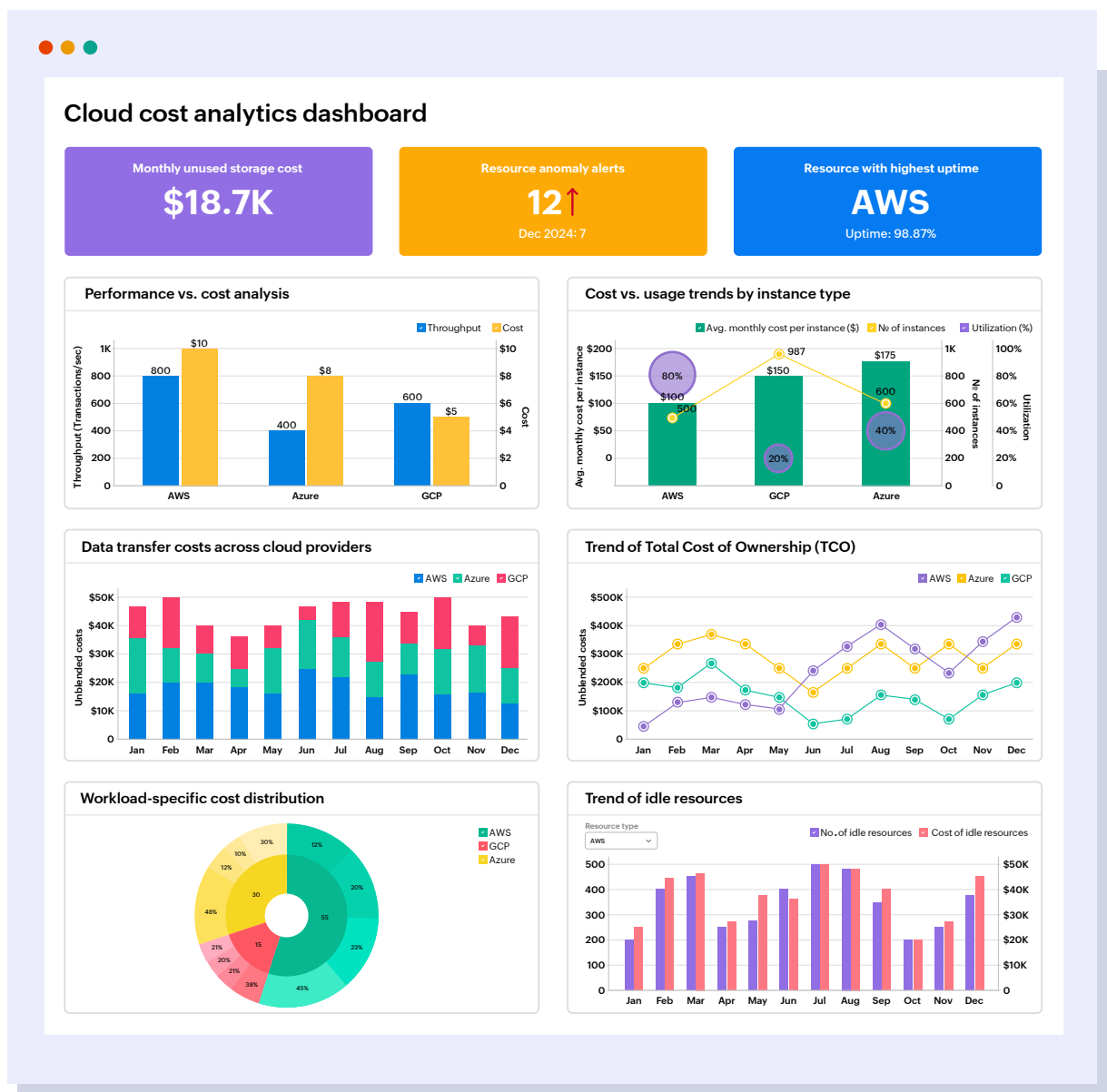
Common cloud vendor management challenges plaguing IT teams

- **Complexity in cloud offerings:** Navigating the countless cloud vendors, service types, and pricing models can make selecting the right resource a daunting task.
- **Lack of real-time insights can break the budget:** Most organizations base their cloud purchase decisions on static or historic data alone. With cloud cost analytics typically conducted just once a year, future purchase decisions often fail to account for real-time usage or consumption patterns. This static approach can leave IT teams unprepared for unexpected usage spikes, severely impacting their budget, operational efficiency, and business growth.
- **Procurement pitfalls:** Due to inadequate need-usage analysis, organizations frequently procure unsuitable resource volumes. Furthermore, since vendors charge based on allocated instance attributes like RAM or CPUs rather than actual utilization, IT teams are often forced to pay for much more than their actual usage needs, resulting in large volumes of idle or wasted resources.

ML-driven cloud cost analytics continuously monitor and analyze cloud expenditures and usage in real-time, providing actionable insights to improve cloud capacity planning. By consolidating cloud spending patterns, resource utilization, and performance metrics into a single dashboard, NOC teams are empowered to optimize their cloud strategy and choose the right resources and vendors for their unique operational needs.

These advanced, ML-powered analytics dashboards help overcome the common gaps in traditional cloud cost analytics, such as siloed data, time-consuming manual analysis, and a lack of actionable insights. With a comprehensive view of cloud resource consumption and expenditure, IT teams are empowered to make informed decisions.

The cloud cost analytics dashboard consolidates key insights on resource utilization, idle resources, cost trends, and cost drivers across providers and services. This ensures optimized cloud infrastructure management and enables data-driven decision making across the organization.



The dashboard above provides IT teams with the comprehensive visibility and insights needed to optimize cloud resource selection. By analyzing multiple facets, including cost-efficiency, workload suitability, vendor performance, historical costs, and projected trends, the dashboard helps teams identify the most suitable cloud resources.

This dashboard serves as an ultimate decision intelligence tool, giving IT teams an in-depth understanding of their cloud infrastructure's performance, costs, and resource utilization. It helps reduce wastage, lower total cost of ownership (TCO), and optimize utilization by selecting resources tailored to their operational needs and budget constraints.

Cloud cost analytics offers a unified platform to monitor, analyze, and optimize cloud spending and resource utilization. This enables proactive cloud capacity planning and management through real-time insights from a single window. Organizations can leverage ML-powered cloud cost analytics to reduce wastage, save valuable dollars, enhance operational efficiency, and align cloud strategies with their business goals.

Conclusion

The introduction of ML-powered and analytics-driven capacity planning is a game-changer for NOC teams striving to effectively manage IT infrastructure costs while enhancing operational efficiency. By addressing inefficiencies in traditional capacity planning methods and harnessing the power of advanced ML-driven analytics, organizations can cut infrastructure costs while ensuring scalability and agility in an ever-evolving IT landscape.

About

ManageEngine Analytics Plus is an IT analytics and decision intelligence solution designed to provide organizations with a unified view of their IT operations, correlate interdependencies and derive meaningful insights. It breaks down data silos by consolidating both on-premises and cloud infrastructure KPIs. Analytics Plus measures the efficiency of network operations, tracks the responsiveness and availability of business applications, evaluates technician performance, assesses the progress of processes and flags security anomalies. This comprehensive analysis is achieved by connecting to all IT software that forms the backbone of an IT infrastructure. These consolidated insights enable organizations to make data-driven decisions that enhance operational efficiency and drive business success.

For more information about Analytics Plus,
visit: www.manageengine.com/analytics-plus/



Reference

1. www.gartner.com/en/newsroom/press-releases/2024-10-23-gartner-forecasts-world
2. www.manageengine.com/analytics-plus/roi-calculator.html
3. www.splunk.com/en_us/newsroom/press-releases/2024/conf24-splunk-report-shows
4. uptimeinstitute.com/about-ui/press-releases/2022-outage-analysis-finds-downtime-c



© ManageEngine, a division of Zoho Corporation