Hewlett Packard
Enterprise

Business white paper

HPE InfoSight

Artificial Intelligence for your Hybrid Cloud World

Table of contents

Introduction

Managing infrastructure has always brought with it frustration, headaches, and wasted time. That's because IT professionals have to spend their days, nights, and weekends dealing with problems that are disruptive to their applications and businesses and manually tune their infrastructure. And, the challenges increase as the number of applications and reliance on infrastructure continue to grow.

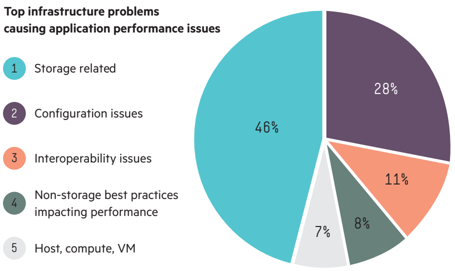
Luckily, there is a better way. HPE InfoSight is Artificial Intelligence (AI) that predicts and prevents problems across the infrastructure stack and ensures optimal performance and efficient resource use.

In this paper, we will explore how **HPE InfoSight** with its recommendation engine paves the way for autonomous infrastructure, so IT can focus its efforts on creating value for the business.

Making the case for autonomous

Every business leader is aware of Digital Transformation. However, underpinning this is the need for infrastructure to consistently and reliably deliver data to its applications. Businesses simply cannot afford any disruptions or delays or the degree of high-touch resources that are needed today.

Next-generation storage platforms, such as **HPE Storage's enterprise flash arrays**, continue to drive up storage and application performance specs. However, fast storage alone cannot ensure reliable, nondisruptive access to data or eliminate the manual attention needed. The complexity of infrastructure inevitably affects businesses and the people who administer it.


Top infrastructure problems causing application performance issues

As much as IT works to move their business ahead, infrastructure continues to hold them back. The result is an endless cycle of break-fix-tune-repeat.

Traditional monitoring and support are no longer good enough

IT has always relied on monitoring tools to troubleshoot their environment. Unfortunately, this has meant staff spending dozens of hours mining log files and interpreting graphs, all in an effort to gain some insight into the cause of a disruption so it could be resolved.

When troubleshooting gets too difficult, IT turns to vendors for support. However, across the industry, support means time-consuming, multi-tiered escalations.

As infrastructure becomes increasingly vital to an organization's bottom line, this model will not suffice. It's no longer acceptable to find out about a disruption after it's occurred. Organizations need a solution that can transform how infrastructure is managed and supported—a solution that can predict problems before they occur.

Making infrastructure work for you is tiresome

Constantly ensuring the optimal performance for every application is fraught with arduous manual intervention. For ever-changing workloads, **fine-tuning** infrastructure requires specialized resources that often involve time-consuming trial and error. Overprovisioning is an easy way out but at a cost of paying for more than what's needed. Even when business requirements don't change, there can be **missed opportunities** to improve performance with existing resources. Perhaps moving an application from an AFA to a hybrid or resizing a volume would make a difference. However, not knowing is a tremendous opportunity cost missed.

Ideally, IT gets recommendations that tell them what to and when to do it to optimize performance and available resources.

AI sees beyond limits

Being human, we can see the present and remember a little bit about the past, which is the same as the tools IT administrators deploy to manage their environments. However, removing the burden of managing infrastructure requires having the foresight to predict problems before they occur—and to have deep intelligence on the underlining workloads and resources to know how the environment can be optimized. Traditional tools fall short for these reasons:

- **Inability to learn from others:** Analytics that simply report on local system metrics offer limited value because the behavior of thousands of peer systems cannot be used to aid in the detection and diagnosis of developing issues. In contrast, a global approach to data collection and analysis can pool observations from an immense variety of workloads. This allows for rare events identified at one site to be preemptively avoided at another and for more common events to be detected earlier and with greater accuracy.

- **Analytics confined to infrastructure silos:** Problems disrupting applications can occur anywhere in the infrastructure stack. Tools providing system status per device only tell part of the story. However, cross-stack analytics that correlate across the multiple layers including applications, compute, virtualization, databases, networks, and storage can tell the full story.
- **Missing domain expertise:** Predictive modeling requires deep domain experience—understanding all the operating, environmental, and telemetry parameters within each system in the infrastructure stack. General-purpose analytics can only go so deep. However, pairing domain experts with AI can enable machine-learning algorithms to identify causation from historical events to predict the most complex and damaging problems.
- **Inability to act:** The ideal state is autonomous operation without the need for human intervention. This requires not only knowing what changes need to be made to avoid a problem or improve the environment but to be able to carry them out on behalf of the administrator. Achieving this level of automation requires a proven history of automated recommendations that provide the necessary level of trust and confidence.

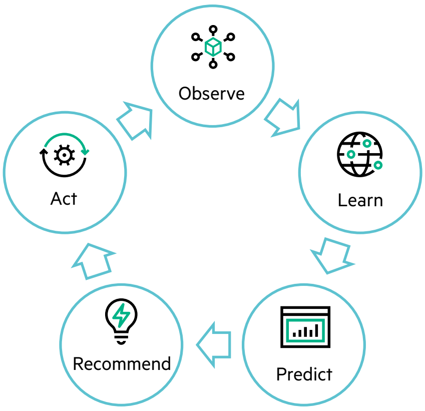


Figure 1. AI for infrastructure framework

Infrastructure powered by AI can overcome these limitations through the following framework:

1. **Observe:** Through simultaneous monitoring of all systems in an installed base, AI develops a steady-state understanding of the ideal operating environment for every workload and application. Then, abnormal behavior can be identified through recognition of the underlining I/O patterns and configurations of each environment.
2. **Learn:** Deep system telemetry coupled with global connectivity creates a foundation of data that exploits the experiences of every system connected. Machine learning in the cloud rapidly accelerates the knowledge and global learning of the AI.
3. **Predict:** For any new problem detected, AI can learn to predict the issue and use pattern-matching algorithms to determine if any other system in the installed base will be susceptible. Additionally, application performance can be modeled and tuned for new infrastructure based on historical configurations and workload patterns.
4. **Recommend:** Based on the predictive analytics, the AI determines the appropriate recommendation needed to improve and ensure the ideal environment. Recommendations are system operational decisions that free IT and eliminate the guesswork in managing their infrastructure.
5. **Act:** Through mutual trust between the infrastructure and AI, recommendations can be applied automatically on behalf of the IT administrators. When automation is not available, specific recommendations can be delivered through support case automation.

AI can watch over your infrastructure, continuously learn from a global installed base, and apply what it learns to predict and prevent issues and eliminate the guesswork in managing infrastructure. AI can make infrastructure smarter and more reliable.

Benefits of HPE InfoSight:

86% of issues automatically predicted and resolved[1]

99.9999% of proven availability[2]

79% reduction in IT storage OPEX[3]

85% less time spent on storage issues[4]

HPE InfoSight: AI for your Hybrid Cloud World

HPE InfoSight was founded on the belief that infrastructure management and support need to evolve. Rather than deal with unexpected problems and reactive vendor support, AI should make infrastructure smart enough to anticipate issues before they occur and resolve them without human intervention. It's only in this self-healing model when businesses can most efficiently utilize their resources to drive their business ahead.

HPE InfoSight is an AI platform that uses intelligence to make infrastructure autonomous. Built on a unique approach to data collection and analysis, HPE InfoSight collects and analyzes millions of sensor data points every second from our globally connected installed base. This sensor data provides comprehensive measurements of the operation and state of each system, subsystem, and surrounding IT infrastructure. It learns from this data to drive its **predictive analytics** and **recommendation engines**, resulting in significant impact for our customers.

Global learning

AI and machine learning requires massive amounts of data beyond the limited logs and metrics of traditional hardware platforms. The HPE storage platforms, which use **Intel® Xeon® processors and SSDs**, were architected with deep diagnostic sensors. As HPE InfoSight has been collecting this data since 2010, its breadth of telemetry creates an architectural advantage.

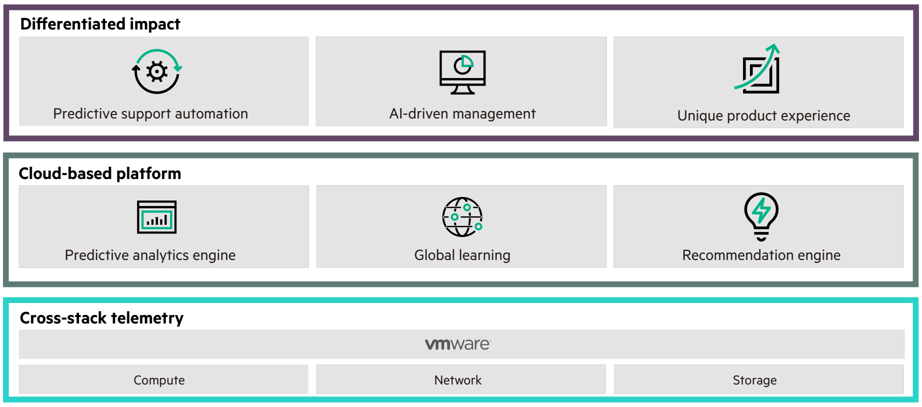


Figure 2. HPE InfoSight platform

Predictive analytics engine

Seeing ahead to eliminate disruptions and to get IT ahead.

HPE InfoSight offers predictive analytics that extend across the infrastructure lifecycle—from planning to expanding.

- **For planning:** It rightsizes new infrastructure by anticipating performance and resources needed based on different applications seen in our installed base. Through telemetry from deployed systems, HPE InfoSight continuously refines its machine-learned models for better sizing accuracy.
- **Once arrays are deployed:** The predictive analytics transforms the product and support experience. HPE InfoSight constantly looks for leading indicators of problems and automatically resolves them before customers even realized there was an issue. If HPE InfoSight detects a new issue, it learns to predict the issue and prevent other systems in the installed base from seeing the same problem.
- **Completing the lifecycle:** HPE InfoSight accurately predicts future capacity, performance, and bandwidth needs based on historical use and autoregressive and Monte Carlo simulations.

[1] "**Redefining the standard for system availability**," 2017

[2] "**HPE Get 6-Nines Guarantee**," 2017

[3, 4] "**Assessing the financial impact of HPE InfoSight predictive analytics**," 2017

Predictive analytics go beyond storage

The predictive capabilities of HPE InfoSight go beyond storage.

For example, HPE InfoSight predicted and prevented a catastrophic All-Paths-Down situation for HPE Nimble Storage customers due to a potential issue with a network VIC card in the host. Leveraging HPE InfoSight, HPE Nimble Storage support engineers determined that the Fibre Channel recovery mechanism might fail due to a double abort issue within the card. HPE InfoSight used a signature pattern-matching algorithm to identify 100 customers susceptible to this issue and applied a workaround that prevented the issue.

As demonstrated with **HPE Nimble Storage**, HPE InfoSight automatically predicts and resolves 86% of issues. This translates to a 79% reduction in IT operational expenses, 85% less time spent on storage issues, and over 99.9999% of proven availability across the HPE Nimble Storage installed base.

The rise of driverless cars

Recommendation engines are being leveraged across multiple industries to automate and optimize everything from online shopping carts to business operations. One area where they are making a dramatic impact is in enabling self-driving cars. Recommendations tell a driverless car how fast it can drive, when to brake, and how to avoid collisions. The right recommendation at the right time is the difference between avoiding an accident and being in one.

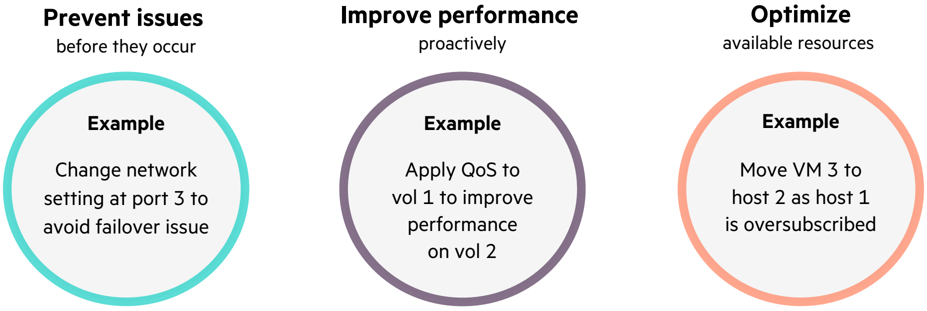
Recommendation engine

Making infrastructure effortless to manage.

For infrastructure to be autonomous, HPE InfoSight needs the ability to not only see what's ahead to predict issues but also to dynamically make intelligent recommendations and decisions that improve and optimize each environment proactively. It needs to be application-aware to serve the right recommendation at the right time without impacting other applications.

Through a recommendation engine, HPE InfoSight builds off its predictive capabilities to automatically tell IT how to prevent issues, proactively improve performance, and optimize resources. The engine advises based on the experience learned from its knowledge base.

As many of the more difficult responsibilities in infrastructure management are related to system performance, we'll take a closer look at what the recommendation engine does for performance management.



Prevent issues before they occur

Example

Change network setting at port 3 to avoid failover issue

Improve performance proactively

Example

Apply QoS to vol 1 to improve performance on vol 2

Optimize available resources

Example

Move VM 3 to host 2 as host 1 is oversubscribed

Figure 3. Benefits of HPE InfoSight recommendation engine

AI performance recommendations

The reality today is that ensuring optimal performance is a time-consuming and costly ordeal. First, it's reactive as application-impacting problems come up unexpectedly. Second, the analytics from other tools aren't smart enough to understand why a problem came up and how to resolve it. Third, it's painful as there's too much manual tuning and guesswork involved.

Through advanced machine learning, the recommendation engine in HPE InfoSight identifies opportunities to improve performance based on I/O workload patterns, accurately determines variables that have the highest impact, and proactively provides the right recommendation to improve performance. The recommendation engine takes away the guesswork and optimizes performance and resources.

Architecting the recommendation engine

In this section, we will take a closer look into the recommendation engine. We'll show the design methodology as well as the architecture.

Figure 4 represents the problem space of potential infrastructure problems on a bar chart with the type of problems and frequency labeled respectively. Problems generally fall within two categories—those that are **simple and common**, as marked in grey, and those that are **complex and unique**, as marked in blue—forming a Pareto distribution. It's important to note the pain curve correlated with the type of problems.

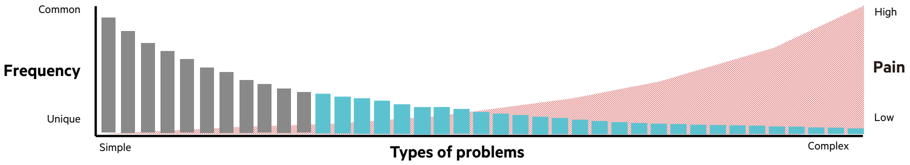


Figure 4. Problem space correlated with the relative pain index

Simple and common problems, such as failed disk drives are more frequent, but only account for a small percentage of the pain inflicted on IT administrators. The frequency of these problems makes them easier to predict and resolve with an automated solution. However, the reality of IT environments is that problems can vary widely, and it's the issues that are complex and unique—the ones that come up unexpectedly and require numerous people and resources to resolve—that cause the most pain and suffering.

Businesses need to have the entire range of problems from the most basic to the most complex and multifaceted, predicted, and auto-resolved. Simple problems can be identified by looking at only a few pieces of data qualitatively with hard-coded rules to trigger events and alarms. While other vendors may claim they provide recommendations, much of their proficiency is generally limited to addressing issues that lie in the simple and common side of the distribution.

For complex and unique problems, the number of variables and the level of quantitative precision required to make a diagnostic determination increases almost exponentially. As problems become more complex, hard-coding rules involving numerous quantitative variables become error-prone and inefficient. Even the most talented of human experts struggle with challenges that go beyond the simple threshold behavior for quantitative problems (for example, this problem should trigger when sensor X climbs above threshold Y). And all too often, even those solutions are derived from anecdotal experience rather than rigorous analysis.

HPE InfoSight recommendation engine goes beyond the simple and common problems to identify and prevent the complex and unique issues. It's with AI and machine learning that we can address the long tail and provide recommendations to avoid business disruptions.

Design methodology for the AI performance recommendations

Constructing a robust recommendation engine for performance requires answering some key questions.

Sensors alone lack context
Analogous to a biometric screening, there's no definitive way to know if a systolic blood pressure measure of 133 mmHg is cause for worry any more than 121 mmHg without much more context and data on the human body. Therefore, the diagnosis and recommendation for values between 120 to 139 mmHg is the same and based on presumptions rather than scientific reasoning. Subsequently, it's incorrect to assume that an average read latency of 10 ms has any more noticeable impact on performance than 5 ms, as this metric alone lacks the full context.

Question 1: Are performance metrics actually an accurate indicator of an unoptimized system or potential problem?

Sensors collect real-time measurements of their environment with the purpose of detecting events or changes. Typically, IT administrators rely on the value of these sensors (that is, read latency, write latency, IOPS, throughput, and others) to determine if behavior is problematic. However, this approach is flawed, as sensors alone lack the full context to determine if their values are truly indicative of impact on the application and end customer experience.

Different workloads and applications have differing performance characteristics and sensitivity to the end customer experience. For instance, large block operations like backup jobs are naturally more latent, but less response time sensitive as a transactional workload. Naively assuming higher latency means there are problems resulting in false positives and wasted time chasing down the wrong events—a fundamental problem to event management.

Design approach

Determining how much latency is truly impactful depends on the sensitivity of the underlining application. Leveraging global system telemetry in HPE InfoSight, we have developed machine-learned models of typical performance to more precisely identify events that actually matter to users. We have validated these models using customer case data that reflect the **potential impact** also referred to as the **latency severity score** that can negatively affect performance.

Outcome

As shown in Figure 5, HPE InfoSight understands the true impact of latency and provides a severity index within a defined time frame as color coded in orange and accompanying numerical value (1 to 10). Darker shades of orange indicate higher potential impact to latency.

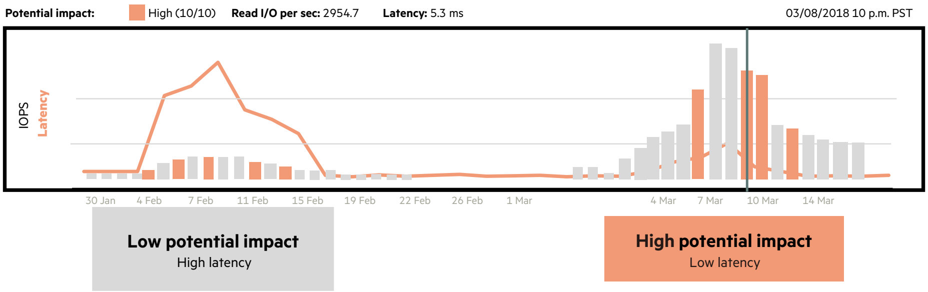


Figure 5. Historical IOPS activity with potential impact reflected in orange color and a numerical latency severity score

This visualization filters the noise and allows IT administrators to focus only on the events that matter. The net result is the elimination of false positives and the ability to discern when performance can be improved.

Question 2: Based on workloads running on the system, what factors can be affecting the application performance and to what degree?

Now that we know if and when sensor measurements are indicative of an unoptimized system, the next step is to determine the cause of it.

Traditionally, IT administrators go through a series of trial-and-error exercises to resolve a performance problem, hoping that something works and keeps the problem away. However, this guesswork is time consuming and often does not resolve the issue indefinitely, if at all.

Why machine learning?
Machine learning is ideal for problems that require the examination of multiple quantitative variables simultaneously and require signatures that lack a concise qualitative description. Human-curated rules are poorly suited for solving these problems in the same way that it would be implausible for a human to handwrite code that determine if a matrix of pixels matches a specific person's face.

Design approach

To ensure our system is capable of recognizing issues across the problem space with high precision, we integrate the analysis from the models described in Question 1 with two types of machine-learning models: **expert trained** and **globally trained**. Expert-trained models are trained and validated against specific examples of rare events that have been labeled by our support engineers. Globally trained models are trained and validated against our installed-base telemetry to recognize uncommon problems by looking for expected correlations with latency or when a system is underperforming relative to expectations.

This hybrid approach ensures HPE InfoSight is able to address the long tail of complex and unique issues.

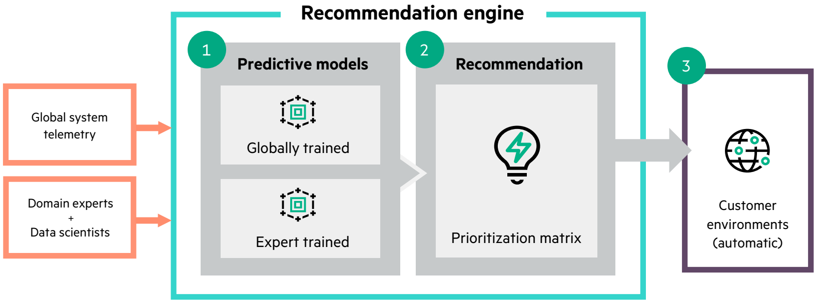


Figure 6. Architecture for the recommendation engine

Multivariate analysis
Expert trained classifiers have been useful, for example, for identifying instances of **SSD bandwidth saturation:** an uncommon event in which high degree of I/O throughput is being directed to SSDs. This scenario is interesting because we have determined that looking at any one SSD metric (for example, latency, queue depth, IOPS, MB/s, proportion of recent milliseconds spent active, and such) is wholly insufficient to accurately determine whether the SSD is causing upstream performance issues. Instead, multiple samples of these metrics need to be examined simultaneously. Looking at one metric alone produces a heuristic that either produces a large number of false positives (low precision) or fails to identify a large proportion of the problematic events (low recall). In order to produce a model that could simultaneously achieve both high precision and recall: a multivariate machine-learned model was needed. Because of the quantitative complexity of the problem, our machine-learned classifier was able to recognize this issue much more effectively than any of the human-written heuristics that preceded it.

Expert-trained models

Our expert-trained models are classifiers that use human-labeled instances of a problem that have been observed in the field. Through the support process, our data scientists train the classifiers to recognize new instances of those events in the field without human intervention and with high level of accuracy. The augmentation of telemetry with human labels ensures the system will make correct diagnosis and recommendations in the face of uncommon events.

Globally trained models

Expert-trained models work well for identifying conditions that are discretely true or false but are not well suited for problems that can have multiple root causes co-occurring to varying degrees. If several distinct contributors to latency are detected on a particular system, it is important to have a consistent way of determining which of these is most responsible for the detected problem. In this situation, it is implausible for a human expert to produce training examples of sufficient quantity. Instead, we train models against our global install-based telemetry to quantify how different sources of latency contribute (often nonlinearly) to produce any observed latency. With this model, we can identify which issues are most important to resolve first. The breadth and richness of our telemetry allows us to generate very comprehensive diagnostic models that would be impossible to train otherwise.

Outcome

Our hybrid machine-learning approach continuously improves the accuracy of the fault monitoring system and its breadth of coverage, minimizing unknown problems. The net result is accurate, root cause diagnostics for every system in our installed base.

Question 3: What is the right recommendation to improve performance?

From the output of Question 1 and 2, HPE InfoSight can determine if there is opportunity to improve the customer environment. The approach to Question 3 results in HPE InfoSight telling IT administrators automatically what they should do to improve the situation.

Design approach

The easiest, but most inefficient, recommendation is to advise the need to upgrade hardware—to simply tell customers that resources are beyond physical limits and larger hardware is required. In contrast, HPE InfoSight provides a much richer set of recommendations, including but not limited to QoS limits, software updates, workload changes, configuration changes, and hardware upgrades. HPE InfoSight is aware of the applications, resources, and preferences (for example, time of day and days of week critical, sensitivity to latency) for every system. And it uses this understanding to prioritize the recommendations.

Details provided with the recommendation inform those portions of the users' workload consuming a saturated resource (for example, the volumes using the most storage CPU when the array is CPU bound). These details are critical because they are required to enable the user to decide between workload-based remediation (that is, throttling volume activity or otherwise attenuating the volumes' requirements) or hardware remediation (that is, adding hardware to the system to extend the system's capabilities and alleviate the resource bottleneck).

Outcome

Before HPE InfoSight, IT administrators had to deal with the pain of managing storage performance. It was a reactive process and involved countless time interpreting graphs and logs and tuning infrastructure manually.

Because of the recommendation engine, customers simply don't have to worry about performance anymore. HPE InfoSight informs IT if there is an opportunity to improve performance and tells them what they should do. They can run their storage systems hard, consolidate multiple applications, and never worry about the applications slowing down due to infrastructure. They know their system is running in an optimal state.

In summary, the recommendations generated by HPE InfoSight are:

- **Automatic:** Available at any time to every customer across the globe
- **Preemptive:** Sees ahead of bottlenecks before they can impact businesses
- **Extensive:** Using machine learning to predict the long tail of complex and unique problems
- **Prescriptive:** Beyond hardware upgrades and includes specific operational changes

Paving the path for autonomous infrastructure

Businesses today need to ensure uninterrupted access to data for all their growing applications. However, this is increasingly difficult with the complexity of infrastructure and demands placed on limited resources. CIOs can no longer afford to be held back by their infrastructure.

Our vision is nothing short of an autonomous infrastructure that no longer needs constant attention, manual tuning, and reactive troubleshooting. This is an infrastructure that self-manages, self-heals, and self-optimizes. While this may seem far from reality, businesses with infrastructure powered by **HPE InfoSight** can see this vision realized sooner than later. And the key is AI.

As the industry's most experienced AI, HPE InfoSight has fundamentally changed how infrastructure is managed and supported. Through cloud-based machine learning, it is predicting and preventing problems while delivering the optimal performance and availability from the infrastructure it supports. And with almost a decade of experience and learning, HPE InfoSight continues to be more sophisticated and proficient.

Building upon these predictive capabilities, the recommendation engine in HPE InfoSight gets us even closer to an autonomous infrastructure. Instead of reacting to problems or trying to figure out how to best manage resources, HPE InfoSight sees ahead and tells customers exactly what to do to avoid issues and to improve environments. These recommendations are making intelligent decisions today that in the future can be applied automatically on behalf of our customers.

Learn more at
hpe.com/info/infosight.html

Share now

Get updates

a00044051ENW, November 2018, Rev. 2