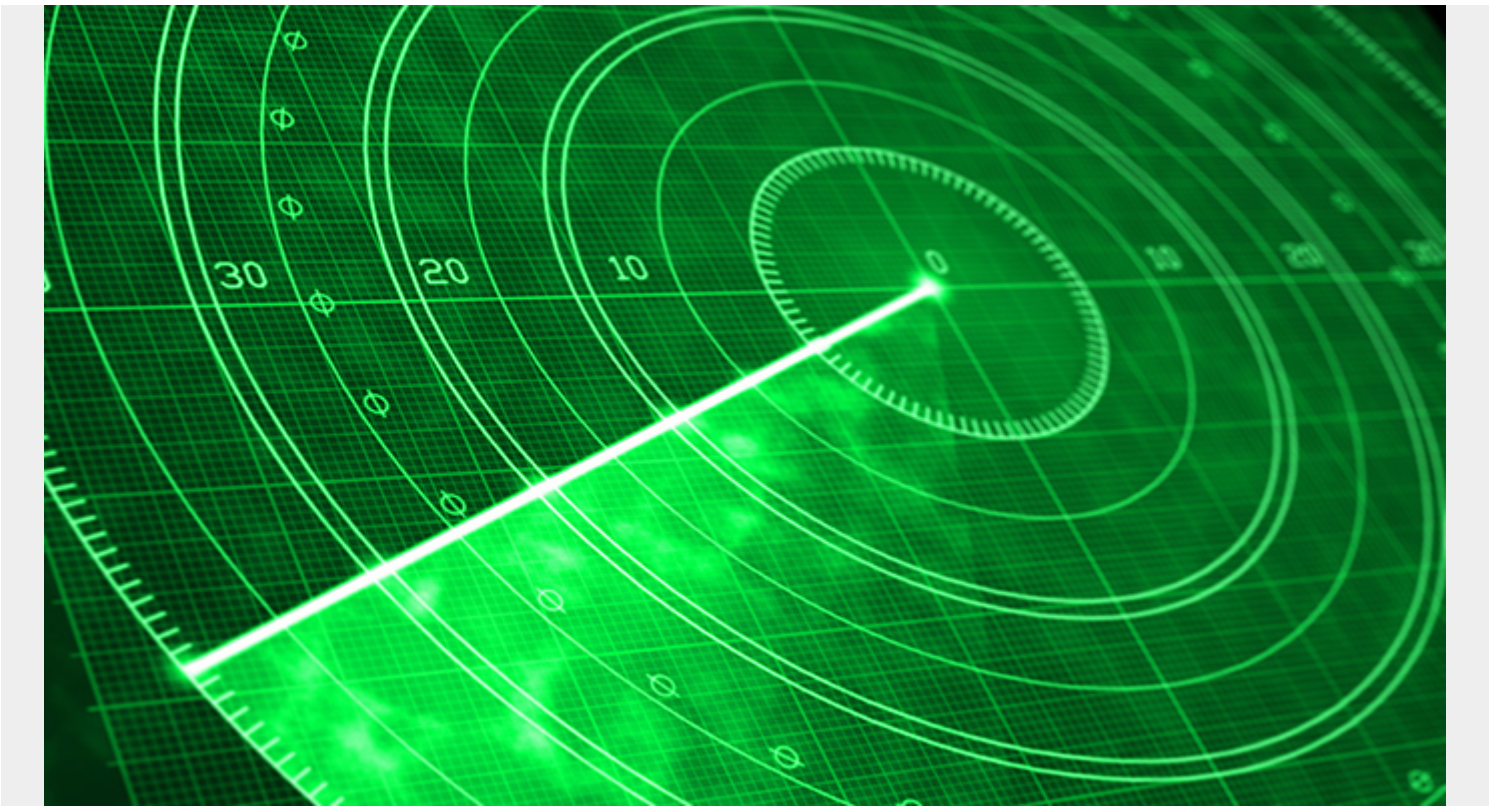
BMC HELIXGPT: OBSERVABILITY AND AIOPS WITH PREDICTIVE, CAUSAL, AND GENERATIVE AI



A recent survey reports that 65 percent respondents say that their organizations are regularly using generative artificial intelligence (AI), and predicting it will lead to a disruptive change in their industries. AI has been curating content in the shadows and providing suggestions for decades; examples include everyday systems like search engines and media streaming applications. It wasn't until early 2023 with the introduction of ChatGPT that AI became mainstream, bringing a conversational experience where AI responds instantaneously and generates contextual and human-like response on any topic to our fingertips. This is having a disruptive impact at all levels, from learning in classrooms to investment decisions in corporate boardrooms.

R&D investments in generative AI accelerated in the last year, uncovering new and compelling uses cases for IT operations (ITOps). According to IDC research spending on generative AI will jump from $16 billion this year to $143 billion in four years.

In this blog, I want to focus on generative AI use cases along with intelligent automation recommendations that make ITOps more efficient and cost-effective. Generative AI-based learning and the resulting conversational experience is transforming how IT uses actionable and contextual insights from their operations and service management data. However, generative AI-based outcomes in addition to inputs from telemetry, ticket, and knowledgebase data also require domain-specific context to respond with a high level of confidence.

Figure 1 below summarizes how observability and service management converge across a

composite (predictive, causal, and generative) AI pipeline to solve complex IT problems. The dynamic service model acts as input (domain-specific context) to the pipeline and carries context forward to aid in noise reduction, root cause analysis, and intelligent recommendations. The context helps put a boundary around the problem and impacted assets, so the AI algorithms predict and generate accurate responses.

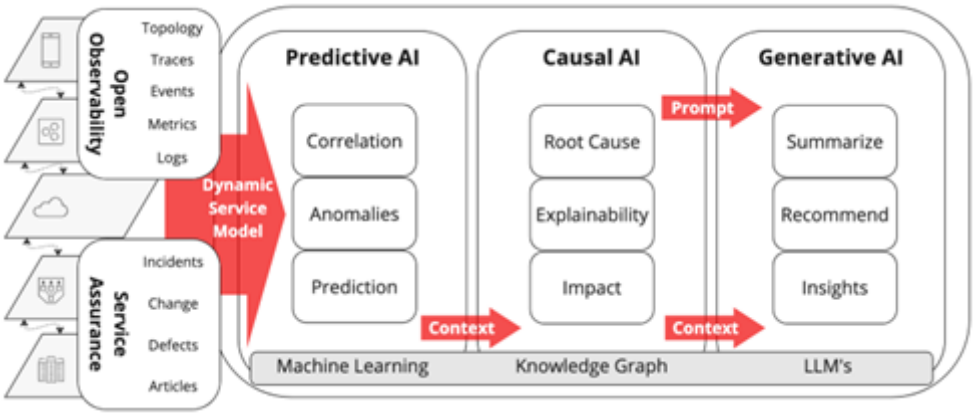


Figure 1: Composite AI with predictive, causal, and generative AI

Some history

The need for AIOps arose in the early 2000s for organizations struggling with maintaining an acceptable customer experience when dealing with a large, complex infrastructure. This resulted in a need for intelligent systems and automation to reduce mean time to repair (MTTR) or better predict or eliminate customer impacting outages.

AIOps solutions in early 2010 used AI and machine learning (ML) algorithms for anomaly detection and event correlation. These solutions applied predictive AI techniques and were effective in finding outliers and reducing event noise, but they proved hard to operationalize in large, complex environments. The biggest gap was their inability to encapsulate assets in a service model and perform contextual root cause isolation. This gap was difficult to fill for these early AIOps solutions as it not only required a connected and reconciled topology from the cloud to mainframe and application to network, but also a bidirectional integration with service management solutions to reduce service outages related to change.

Input to the AI for domain context

Data is the foundation of AI. The context of the input data is even more important to explain a problem or situation in completeness and high accuracy.

Dynamic service model

The dynamic service model is the foundational building block of a business service. It provides the domain-specific knowledge as input to the AI pipeline; this helps build cause-and-effect relationships and weed out nonsense correlations and hallucinations. This knowledge helps with contextual noise reduction, root cause isolation, and generative AI insights and recommendations.

The service model represents an end-to-end service view represented by configuration items (CIs) and their relationships, the model is automatically built and dynamically maintained. Figure 2 below

shows an example for a dynamic service model, depicting a service spanning a distributed, hybrid, multi-cloud landscape connecting topology from the cloud to the mainframe and application to the network. The reconciled topology comes from discovery and monitoring tools for the application, infrastructure, and network. This service model breaks down organizational silos and is commonly used for:

1. Noise reduction, root cause isolation, and change-impact monitoring in ITOps.
2. Incident, change, and asset management in service management.
3. Bidirectional incident, change, and root cause correlation and automation in a new ServiceOps approach that bridges service management and operations

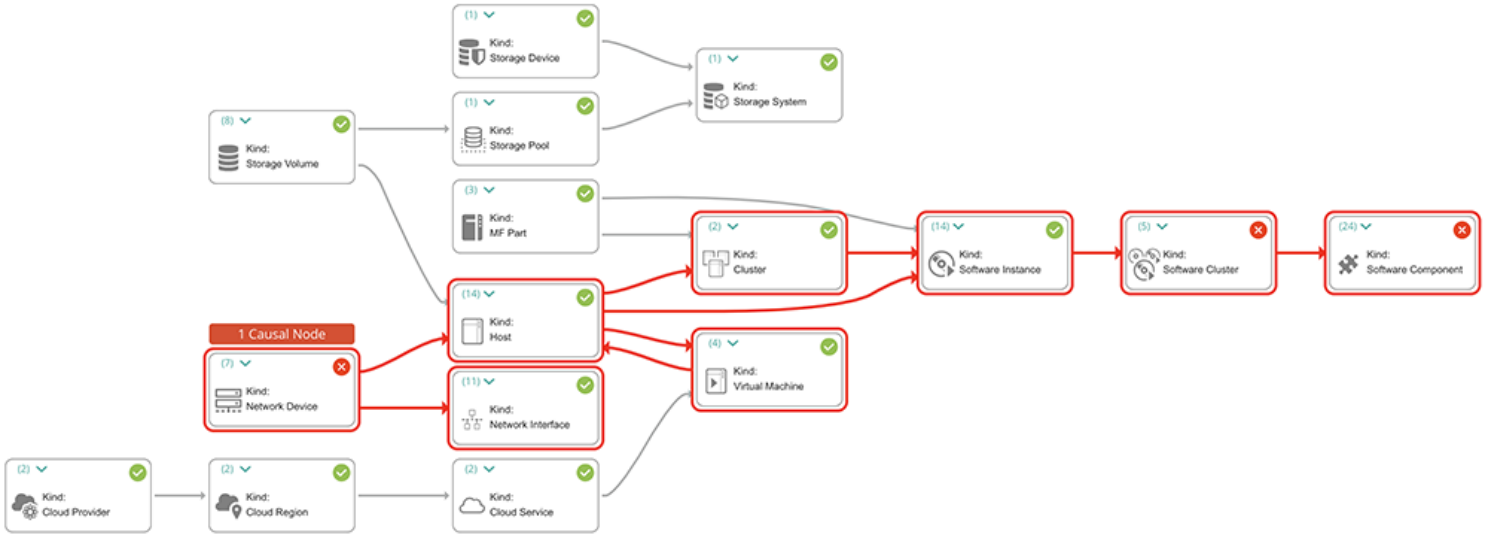


Figure 2. Dynamic service model for a distributed, hybrid, multi-cloud business service

Composite AI pipeline—Derive insights for intelligent automation

The dynamic service model provides contextual input to the composite AI pipeline. The pipeline applies predictive, causal, and generative AI techniques to help predict service disruptions, determine the root cause to expedite the triage process, and derive insights to recommend best actions to automate with confidence.

Predictive AI

Predictive AI is used to learn from historical metric trends to predict future impact. Correlation-based algorithms are also used to uncover relationships in data and see common trends and patterns to help with noise reduction and proactive problem analysis.

A prediction-based methodology approach predicts service level agreement (SLA) breaches and can detect abnormalities in telemetry data. SLA breaches impact service key performance indicators (KPIs) and impact critical business services. Proactively detecting and fixing potential service or end-user impact helps prevent service disruptions, resulting in improving customer experience.

Anomaly detection is also used to alert unusual patterns in the data that violate the normal baseline. Observability data inundates the systems with a high volume of telemetry data, making it very difficult to assess and set manual thresholds. Prediction-based anomaly detection eliminates the need to set manual metric thresholds and will find outliers.

Event, ticket, and log data correlation is becoming more important for predicting trends and uncovering underlying causes. Correlation also helps with noise reduction by clustering data based on natural language processing (NLP) algorithms and correlating insights with service impact.

Causal AI

Causal AI is used to determine root cause, and this would not be possible without the domain context that comes from the dynamic service model. The dynamic service model in Figure 2 shows how an impacted network device labeled as the "causal node" is impacting the upstream application.

When production issues unfold, observability events are mapped as cause-and-effect relationships across the dimensions of time and topology. This results in a graph depicting a causal chain of events as shown in Figure 3, with the impacted layers on the left, causal chain of observability events across time on the right, and root cause event in red. In this example, it is a network device.

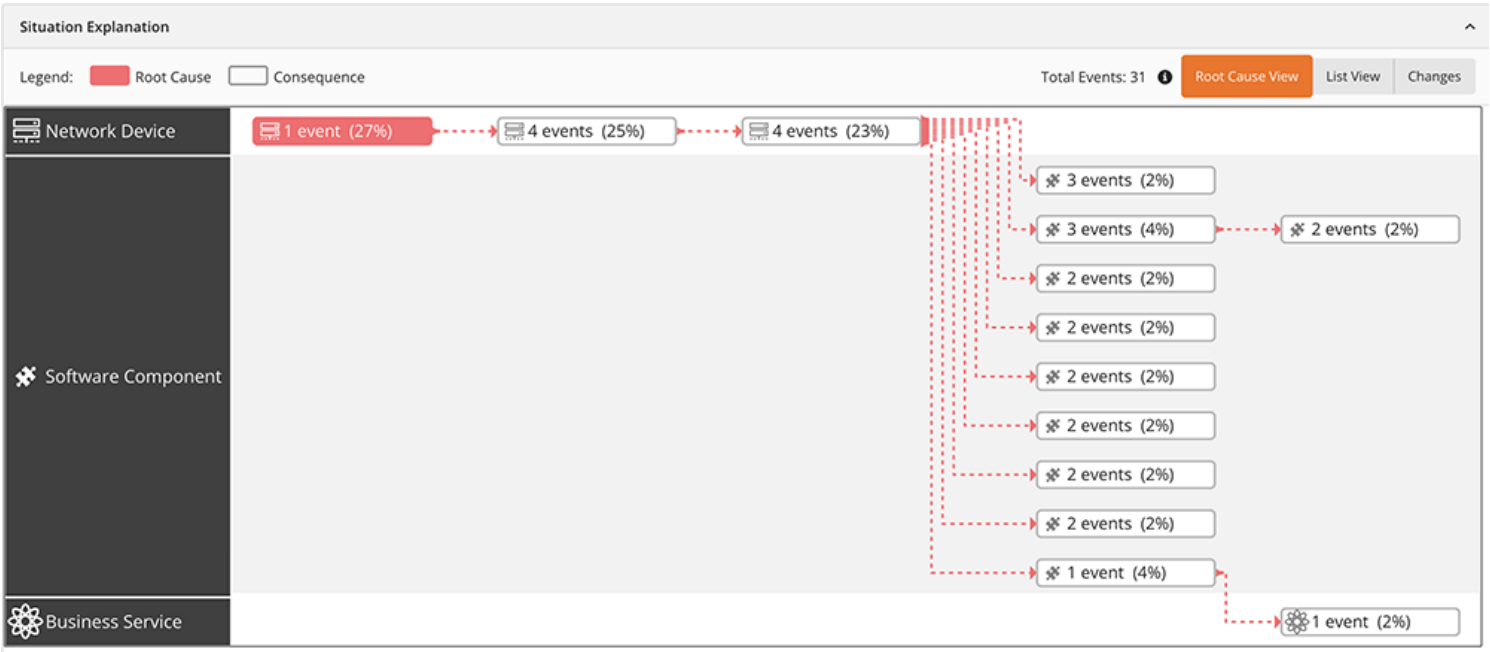


Figure 3: Causal graph of monitoring events and impacted layers

The causal graph is akin to a fingerprint and is stored in a knowledge graph for historical analysis. Causal AI helps to:

1. Automate root cause CI isolation, eliminating the need for manual L1 triage or a war room scenario. The issue in this example is automatically assigned to the network subject matter expert (SME), reducing help desk fatigue.
2. Correlate ITSM change requests to the root cause CI. If a rollback is required, the impactful change request is correlated with the root cause CI, henceforth ignoring the tens or hundreds of non-impactful change requests.
3. Identify business service or end-user impact. A flapping switch is only interesting if the end user is, or is predicted to be, impacted. This helps prioritize and focus expensive engineering resources on impactful issues.

Generative AI

The dynamic service model, coupled with predictive and causal AI outputs, provides context for

generative AI to execute the last mile and produce human-readable insights to improve service resiliency and avoid future disruptions.

Generative AI is used to train large language models (LLMs). The LLM comes pre-trained with domain knowledge on how to solve common operational issues and can optionally be further fine-tuned with customer data in a secure-tenant model to help improve the accuracy of the response.

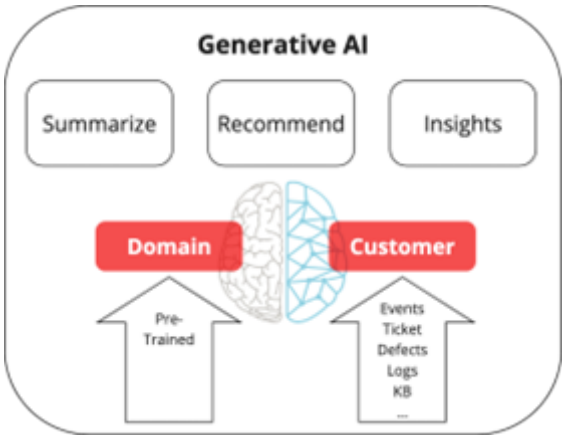


Figure 4: Generative AI uses pretrained domain and fine-tuned LLMs with customer data

The trained LLM can be relied upon to:

1. Generate a human-readable problem summary of the ongoing issue based on the identified root cause, causal graph, and end-user impact to expedite the triage process and reduce MTTR.
2. Recommend the automated best action to remediate based on past similar situations, tickets, defects, logs, and knowledgebase data to better avoid service disruptions.
3. Converse and derive contextual insights from telemetry, incident, and change data, which helps break down silos and surface change-related risk and outages.

Path to self-remediation

The composite AI pipeline applies AI techniques to input necessary domain context, derive actionable insights from domain-specific contextual intelligence, and use enhanced decision-making to automate with confidence, as shown in Figure 5.

Figure 5: AI pipeline applying composite AI techniques to

automate decision-making

The goal for operations team is to achieve zero-touch operations, where the AI detects, triages, and remediates with minimal human intervention. In my previous blog, I talked about the 7 core principles for AIOps to achieve zero-touch operations.