Unlock On-Prem AI with Customer-Managed GPUs for GenAI & AITK Acceleration

PLA1525

September 2025

.conf25

splunk>

Forward-looking statements

This presentation may contain forward-looking statements regarding future events, plans or the expected financial performance of our company, including our expectations regarding our products, technology, strategy, customers, markets, acquisitions and investments. These statements reflect management's current expectations, estimates and assumptions based on the information currently available to us. These forward-looking statements are not guarantees of future performance and involve significant risks, uncertainties and other factors that may cause our actual results, performance or achievements to be materially different from results, performance or achievements expressed or implied by the forward-looking statements contained in this presentation.

For additional information about factors that could cause actual results to differ materially from those described in the forward-looking statements made in this presentation, please refer to our periodic reports and other filings with the SEC, including the risk factors identified in our most recent quarterly reports on Form 10-Q and annual reports on Form 10-K, copies of which may be obtained by visiting the Splunk Investor Relations website at www.investors.splunk.com or the SEC's website at www.sec.gov. The forward-looking statements made in this presentation are made as of the time and date of this presentation. If reviewed after the initial presentation, even if made available by us, on our website or otherwise, it may not contain current or accurate information. We disclaim any obligation to update or revise any forward-looking statement based on new information, future events or otherwise, except as required by applicable law.

In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. We undertake no obligation either to develop the features or functionalities described, in beta or in preview (used interchangeably), or to include any such feature or functionality in a future release.

splunk> .conf25

Speaker

Nick Ma

Senior Staff Engineering Product Manager, AI
Splunk

Shang Cai

Senior Staff Machine Learning Engineer
Splunk

Challenge, AI and Deployments

Challenges

SPL Creation & Insights

Security: Detect → Triage → Fix

Observability: Detect → Triage → Fix

Challenge, AI and Deployments

Challenges

SPL Creation & Insights

Security: Detect → Triage → Fix

Observability: Detect → Triage → Fix

AI

Splunk AI Assistant for SPL (SAIA)

Security AI Assistant (SECA)

AITK - hosted LLM

AITK - predictive AI

Observability AI Assistant

Challenge, AI and Deployments

Challenges

- SPL Creation & Insights
- Security: Detect → Triage → Fix
- Observability: Detect → Triage → Fix

AI

- Splunk AI Assistant for SPL (SAIA)
- Security AI Assistant (SECA)
- AITK - hosted LLM
- AITK - predictive AI
- Observability AI Assistant

Deployments

- Splunk Cloud → Cloud AI Deployment
- Splunk Enterprise → Hybrid AI (Cloud-Connected)
- Splunk Enterprise → Fully On-Prem AI (AI Tier)

Challenge, AI and Deployments

Challenges

SPL Creation & Insights

Security: Detect → Triage → Fix

Observability: Detect → Triage → Fix

AI

Splunk AI Assistant for SPL (SAIA)

Security AI Assistant (SECA)

AITK - hosted LLM

AITK - predictive AI

Observability AI Assistant

Deployments

Splunk Cloud → Cloud AI Deployment

Splunk Enterprise → Hybrid AI (Cloud-Connected)

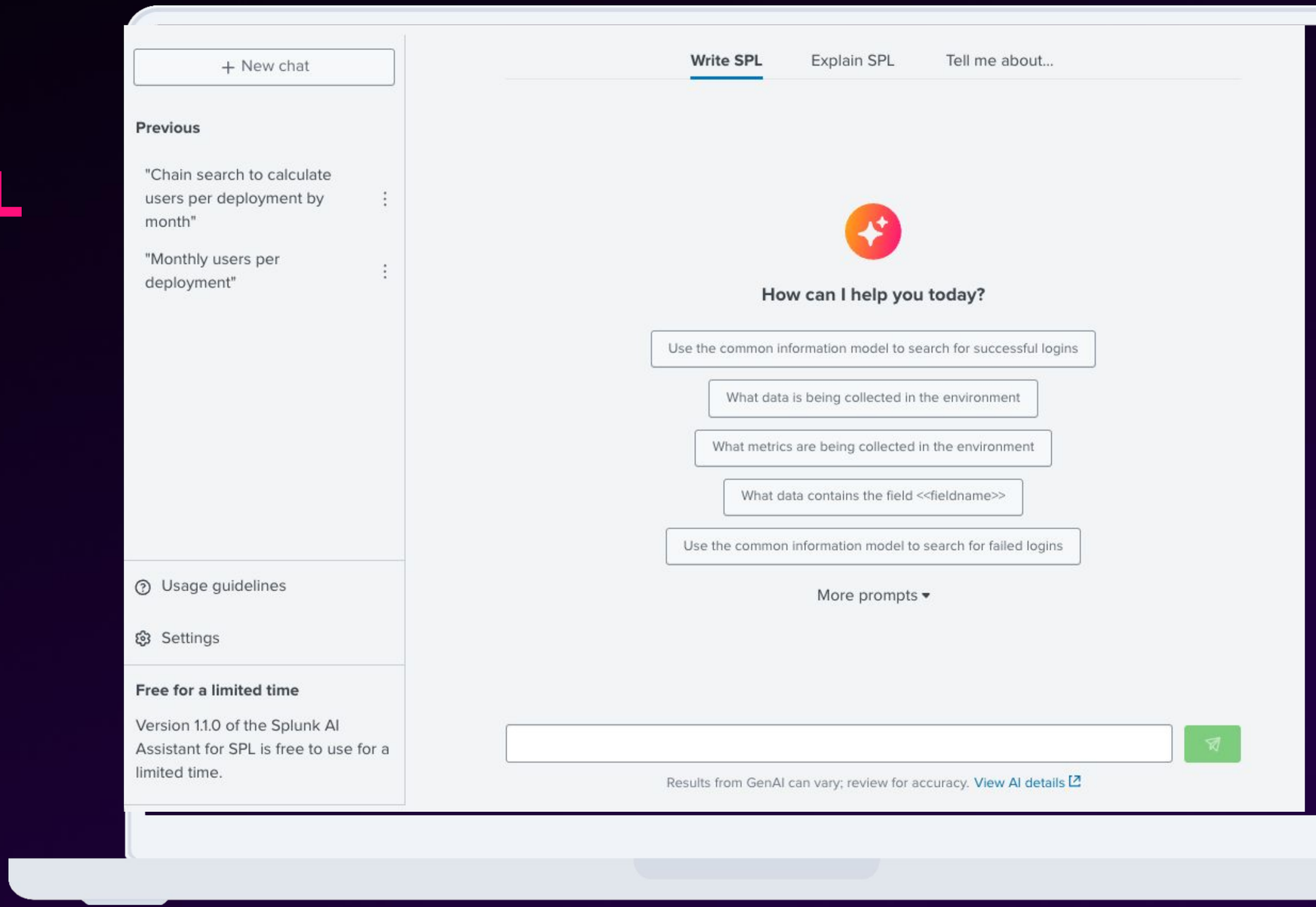
Splunk Enterprise → Fully On-Prem AI (AI Tier)

Splunk AI Assistant

Splunk AI Assistant for SPL

GA in Splunk Cloud and Splunk Enterprise

- Get your job done faster using natural language
- Chat with your data to drill down to deeper insights
- Accelerate your learning journey to become the expert

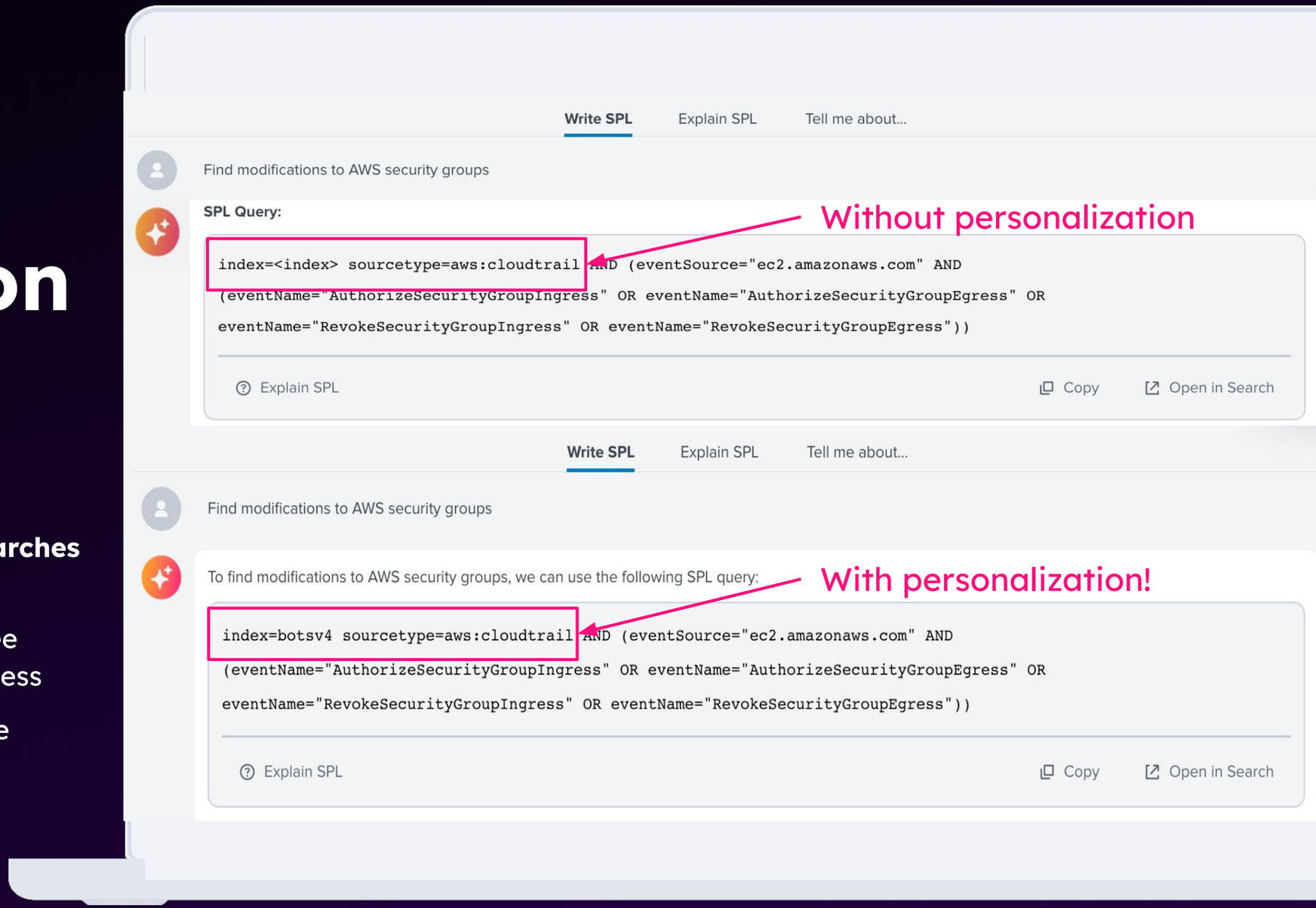


Recently Released

Personalization

For Write SPL — GA in v1.1.0

- Generated SPL now considers **index**, **sourcetype**, **field names** and **past searches** when generating results
- Fully supports RBAC - Users will not see indexes to which they do not have access
- Requires opt-in via SAIA settings page



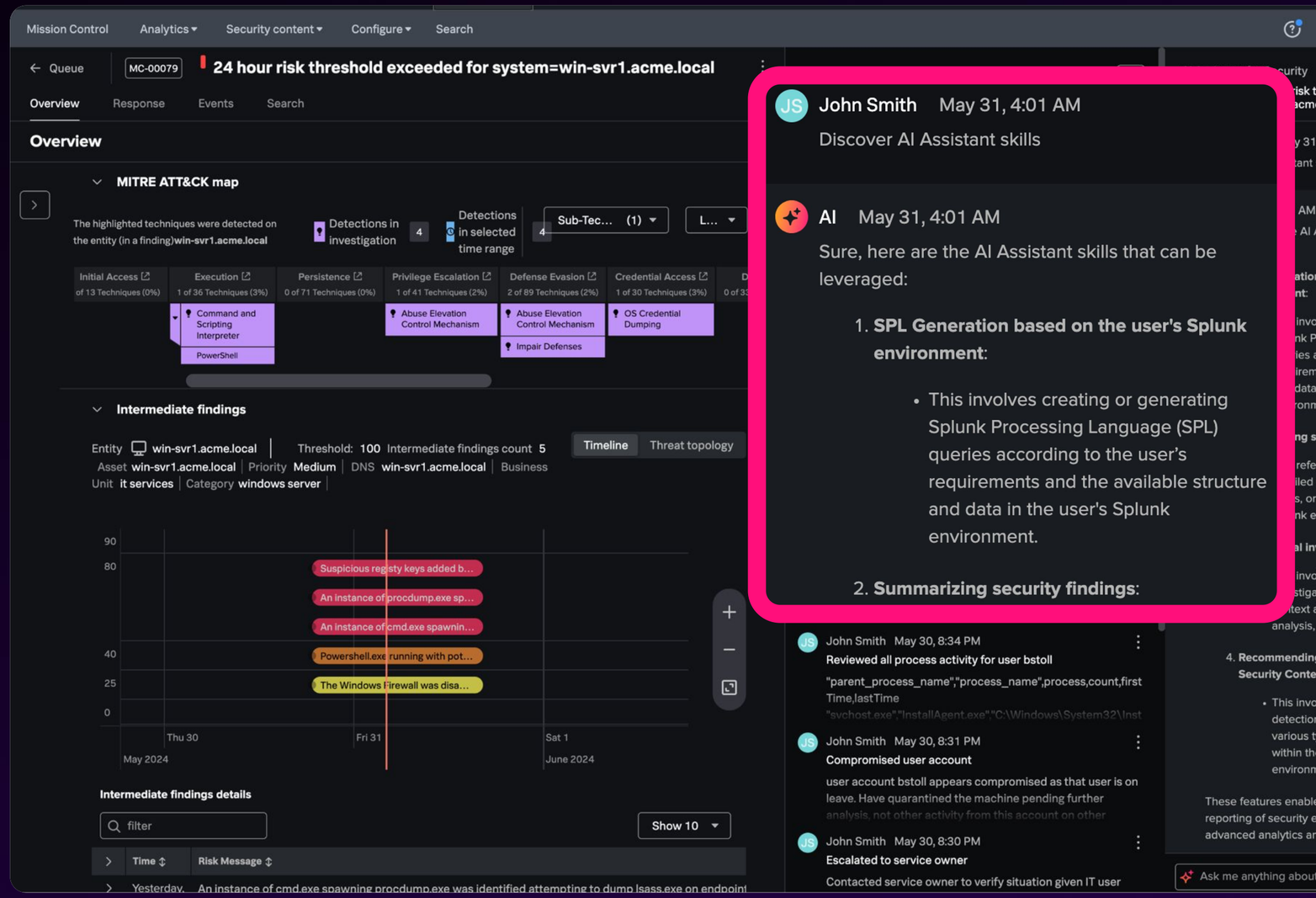
AI Assistant for SPL Demo Video

Preview

AI Assistant in Enterprise Security

Guided Enterprise Security workflows

- Answer analyst questions to guide daily workflows.
- Use natural language queries to investigate more quickly.
- Save time while addressing threats more rapidly.



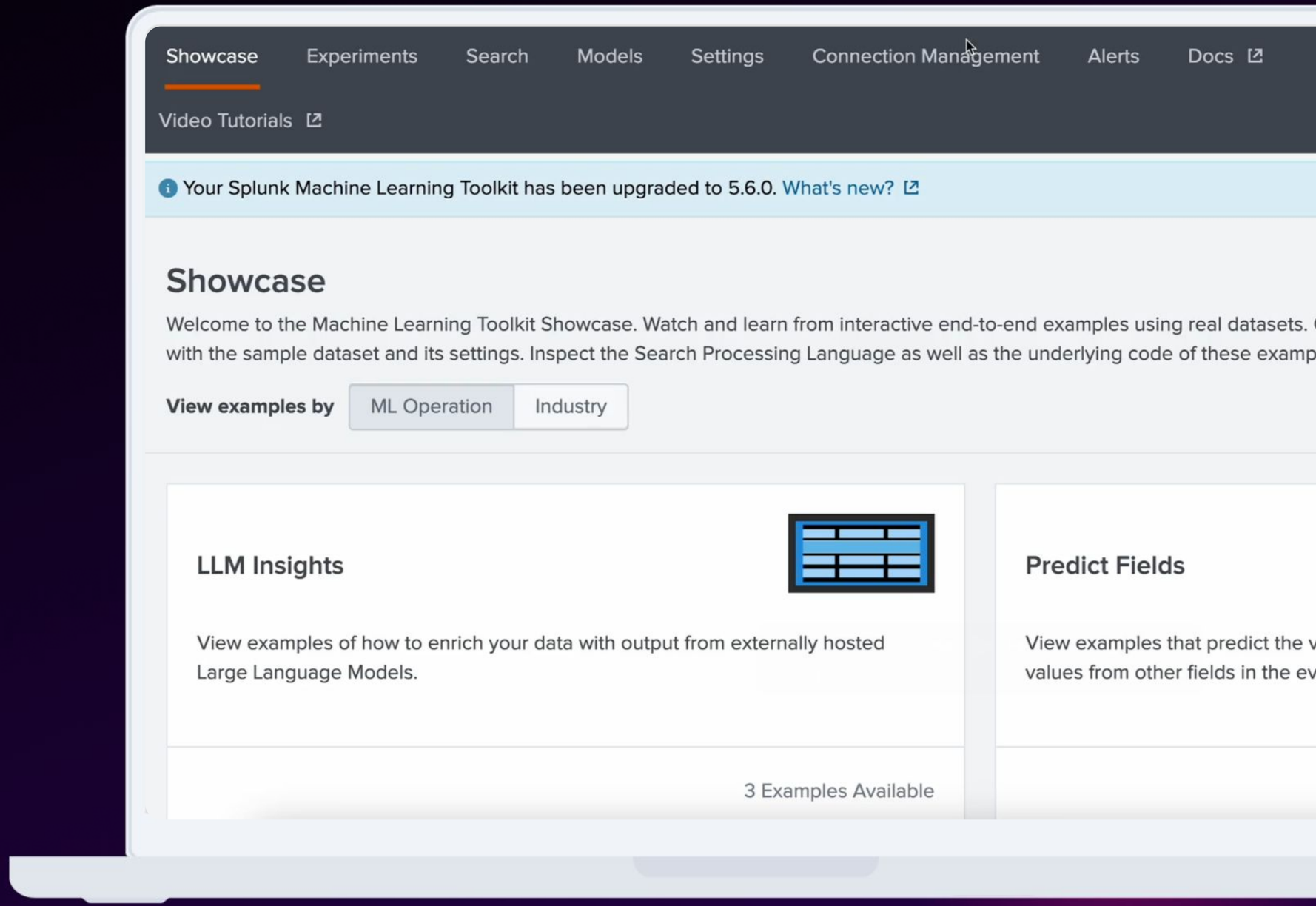
AI Assistant in Enterprise Security is currently in preview. UI shown is for illustration; not final product.

AI Assistant in Security Demo Video

AI Toolkit (Known as MLTK)

An App for Splunk Platform and Splunk Enterprise Security

- Apply machine learning to your data to discover hidden meaningful patterns in your data to inform faster, smarter decisions
- Get step-by-step guided workflows to build and deploy models for common business challenges
- Spot red flags with anomaly detection



What’s new in AITK 5.6?

LLM Integrations

Connect to external LLMs from providers such as OpenAI, AWS Bedrock, Gemini, Anthropic, Azure hosted OpenAI, Groq or Ollama

AI command

Utilise external LLMs during your search processes with the new ai command

ONNX apply enhancements

Ability to generate more sophisticated insights from your data

Splunk Enterprise -> On-prem AI: Deployment Offering

Customer Pain #1 : On-Prem Customers Locked Out of SPL AI Assistant

Context:

- Splunk launched the SPL AI Assistant in mid of 2024, but **only for Splunk Cloud customers.**

Pain Point: Accessibility of AI for Splunk Enterprise users

- Frustration from **air-gapped, regulated, or hybrid** deployments
- Customers see **Cloud peers gaining productivity advantages**
- Pressure on admins and execs: *“When will we get this?”*

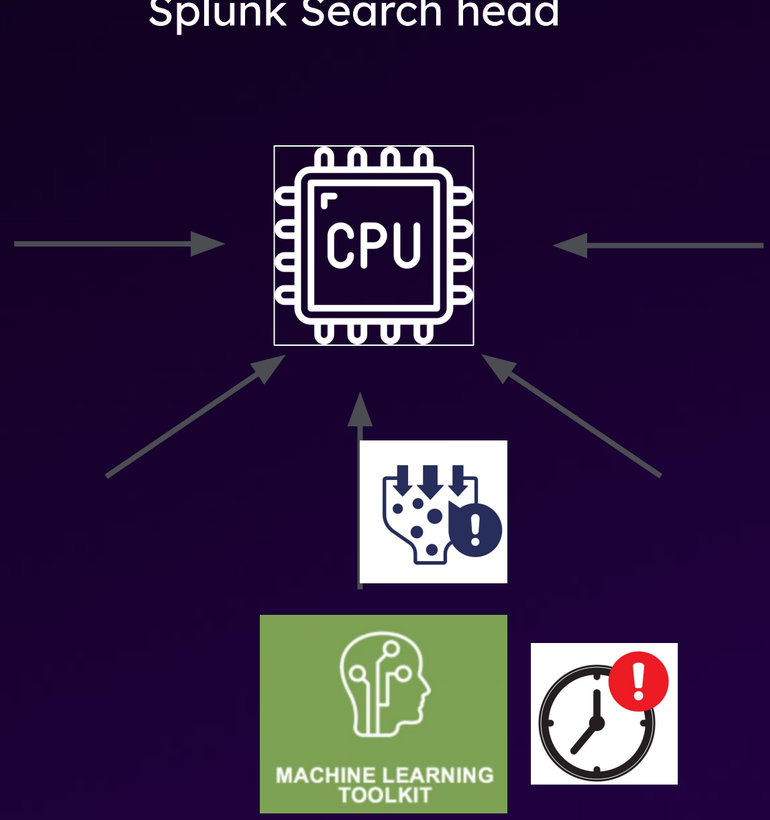
splunk>cloud™

splunk>enterprise

Customer Pain #2 : AITK Struggles with Scalability and Performance

Pain Point: Scalability and Performance

- Search Heads are already saturated with **core Splunk workloads (search, indexing, apps).**
- CPU limits make it hard to scale for:
  - **High-cardinality** use cases
  - **Heavy compute** operations
- Users face slow performance or extremely long wait times when analyzing large datasets.
- Customers need GPU acceleration to unlock advanced ML and GenAI at scale.



Recently Released

Splunk AI Assistant for SPL, Cloud Connected solution, GA

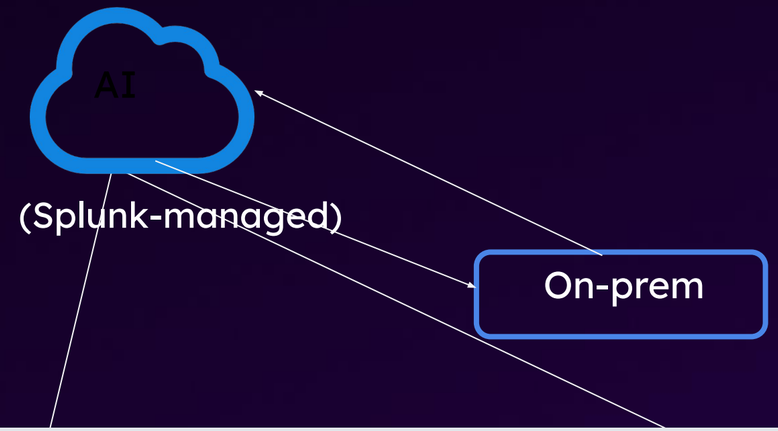
Cloud Connected leverages Splunk-managed AI services in the cloud, with on-prem environments connecting to it

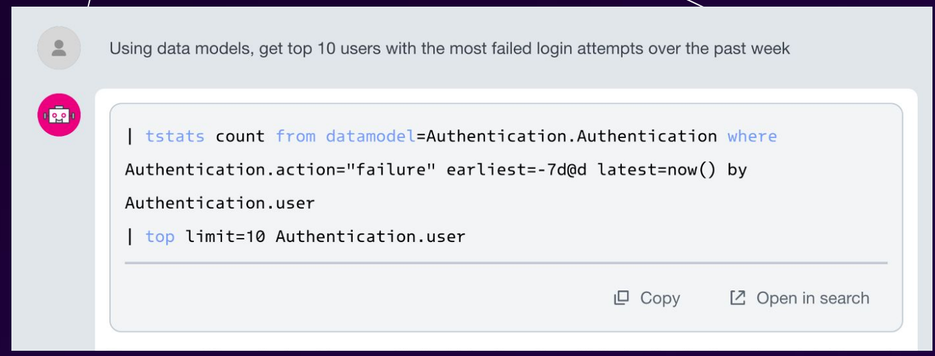
GPU-Free AI for On-Prem Splunk

Same Secure, Compliant Services as SAIA Cloud

Natural Language SPL Generation & Explanation and Q&A

Personalization





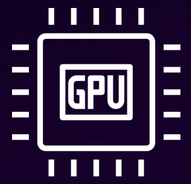
Next

Splunk AI Assistant for SPL, AI Tier, Preview

AI Tier deploys the AI service as an additional tier—similar to the Search and Indexing tiers—leveraging **customer-managed GPUs** and delivered via the **Splunk Operator for Kubernetes (SOK).**

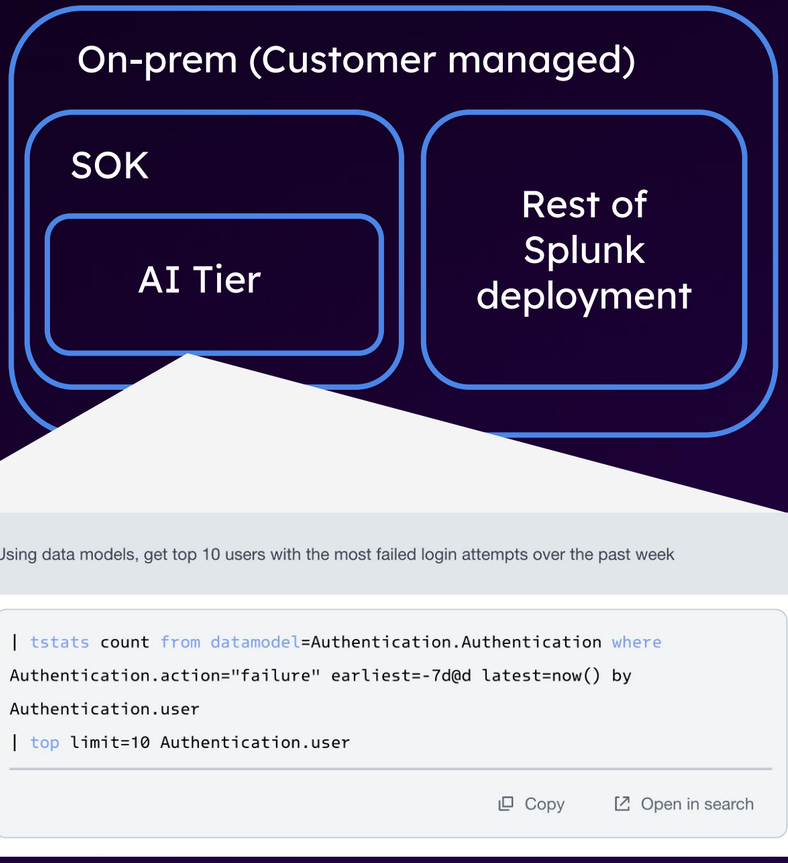
Fully on-prem AI

Strict Data Residency & Compliance



Customer-Managed GPUs

Kubernetes-Native Deployment via SOK



Comparisons of Deployment Options

Customer Requirements	“CLOUD CONNECTED”	“AI TIER W/ SOK”
Legal/Compliance Requirements	**Data remains within Splunk's cloud environment**, ensuring compliance	**All data and models reside in the customer's environment** Supports running in public cloud, private cloud, bare metal
GPU Management & Infrastructure	Splunk managed • Customers do not need to procure or manage GPUs	Customer managed • **Customers must acquire GPUs and manage their scalability and performance** • **Runs on k8s**
Frequency of updates	Frequency based on Splunk App release frequency	Follows SOK release cadence
Feature Availability	Follow cloud version environment, regions, languages • AWS is supported • HIPAA is supported • Azure, GCP, Fedramp not supported No Splunk Enterprise Version requirement	Splunk Enterprise Version requirement: - SAIA depend on 10.2

AI tier with SOK

Why Customers choose Fully On-Prem AI

Compliance /Security

Data stays 100% inside customer-control led boundaries (incl. air gapped environment)

Control

Choose hardware, k8s distro, and deployment topology

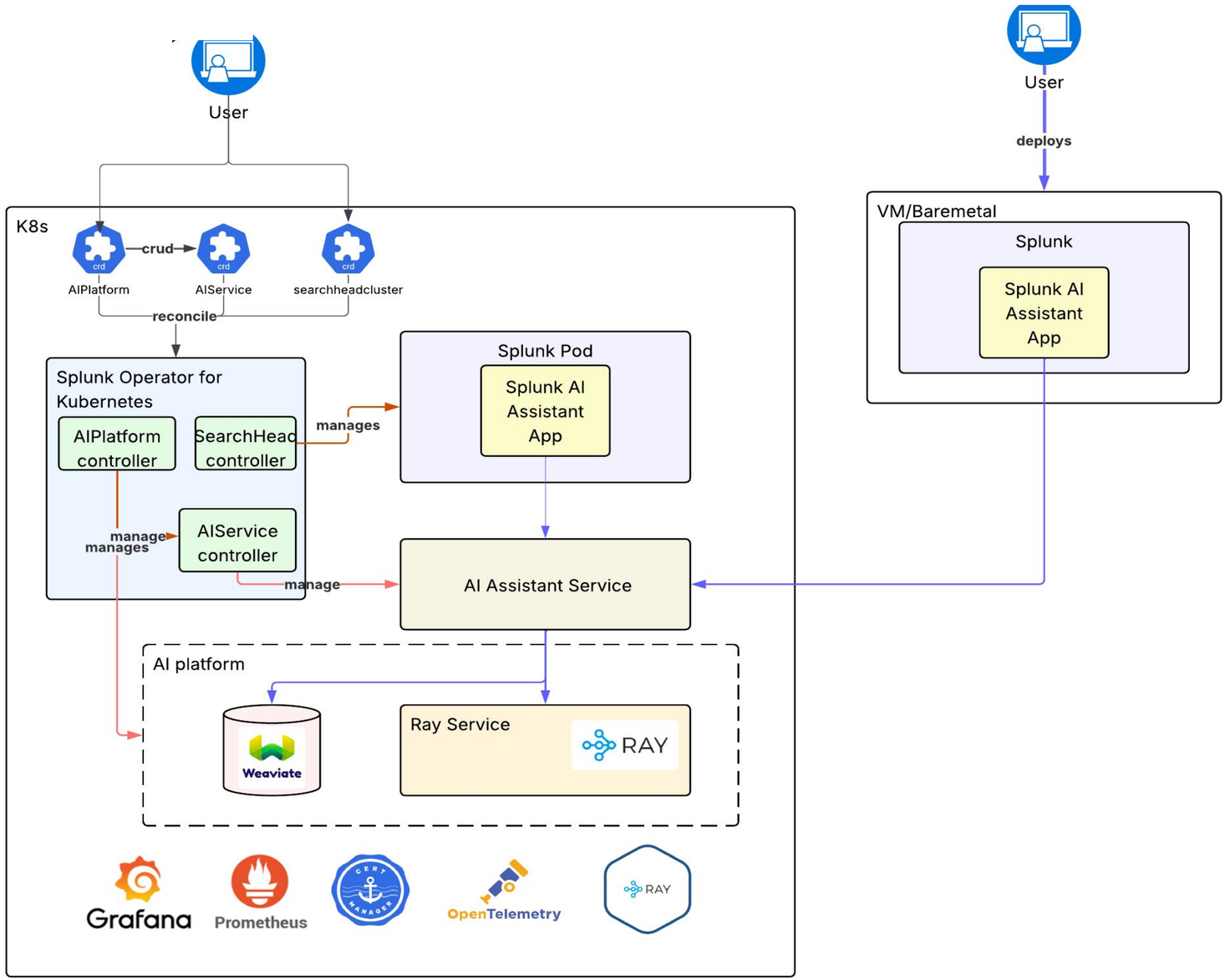
Existing GPUs

Leverage existing GPU/infra investments

Existing integration

Aligns with on-prem change-management and monitoring tools

Solution Deep Dive



Prerequisite #1: GPU requirement for AI tier

***Minimum Specifications* (all required)**

GPU: **8 Nvidia A10G Cores** with 4 co-located in the same instance

CPU: CPU memory requirement > 120GB

Disk: Disk space requirement > 400GB

Recommended Specifications (all required)

GPU: **4 Nvidia L40S Cores** with 2 co-located in the same instance

CPU: CPU memory requirement > 120GB

Disk: Disk space requirement > 400GB

Alternative GPU specifications - if customer has other GPU specifications not listed below, please contact Splunk team

GPU Type	**Total number of cores** required [1]	Number of cores required per instance/ **Co-location requirements** [2]	FYI for those on BYOL, **BYO Cloud Provider Equivalent**
Nvidia A10G	8	4	AWS: g5.12xlarge
Nvidia L40S	4	2	AWS: g6e.12xlarge
Nvidia A100 (40GB)	4	4	AWS: p4d.24xlarge Azure: NDasrA100_v4
Nvidia A100 (80GB)	4	2	AWS: p4de.24xlarge Azure: NDm_A100_v4
Nvidia H100	4	2	AWS: p5.48xlarge Azure: ND-H100-v5
Nvidia H200	2	1	AWS: p5e.48xlarge, p5en.48xlarge Azure: ND-H200-v5

Notes:

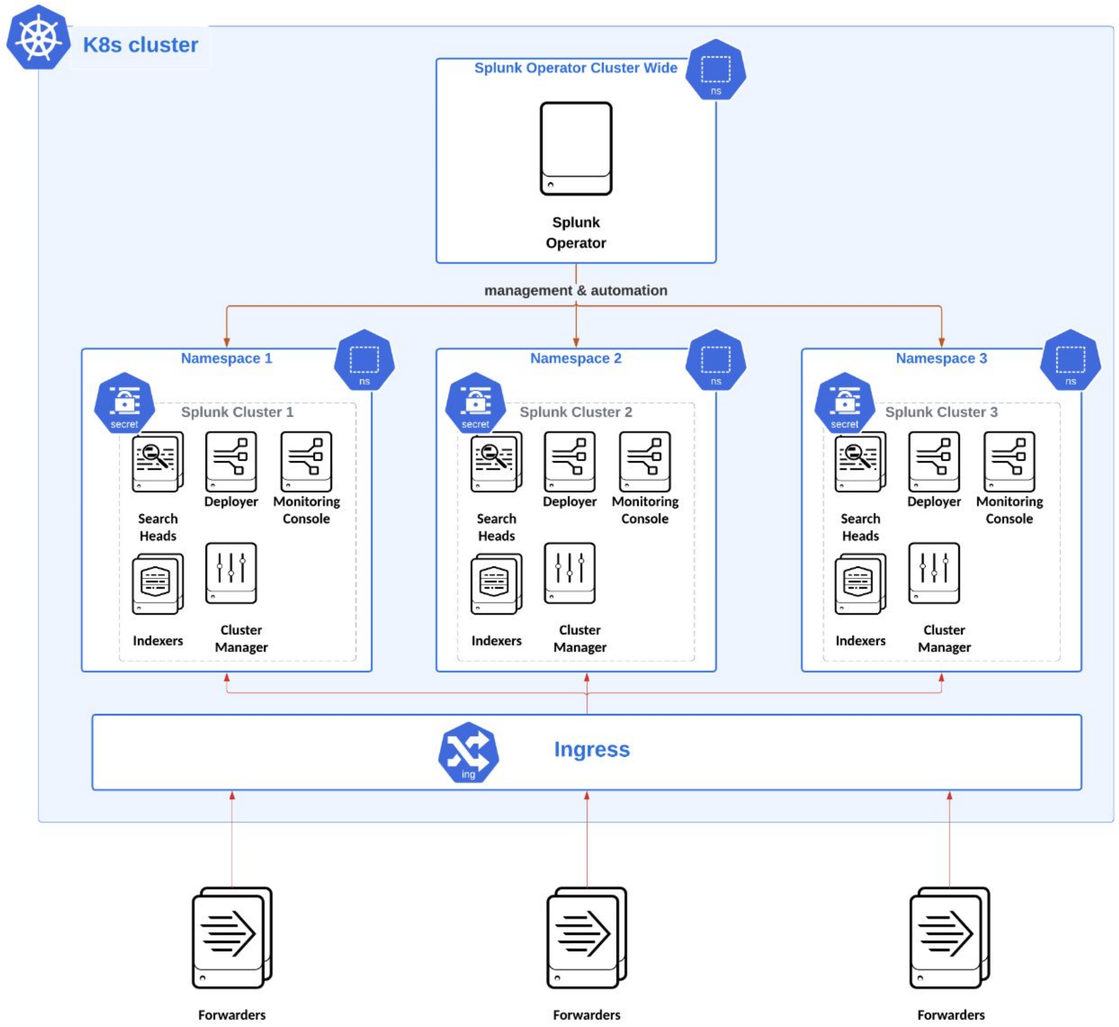
1. Total number of cores is the number of cores required to serve all models required by current AI Assistants

2. Number of cores required per node is the minimum number of GPUs on the same instance required to serve 1 copy of the biggest LLM models we use

Prerequisite #2: SOK

What is SOK?

- Kubernetes-native solution
- Deploying, managing, and scaling CMP in containerized environments
- Vendor-agnostic: on-prem, AWS, GCP, Azure, and more



Why Splunk Operator for Kubernetes (SOK)

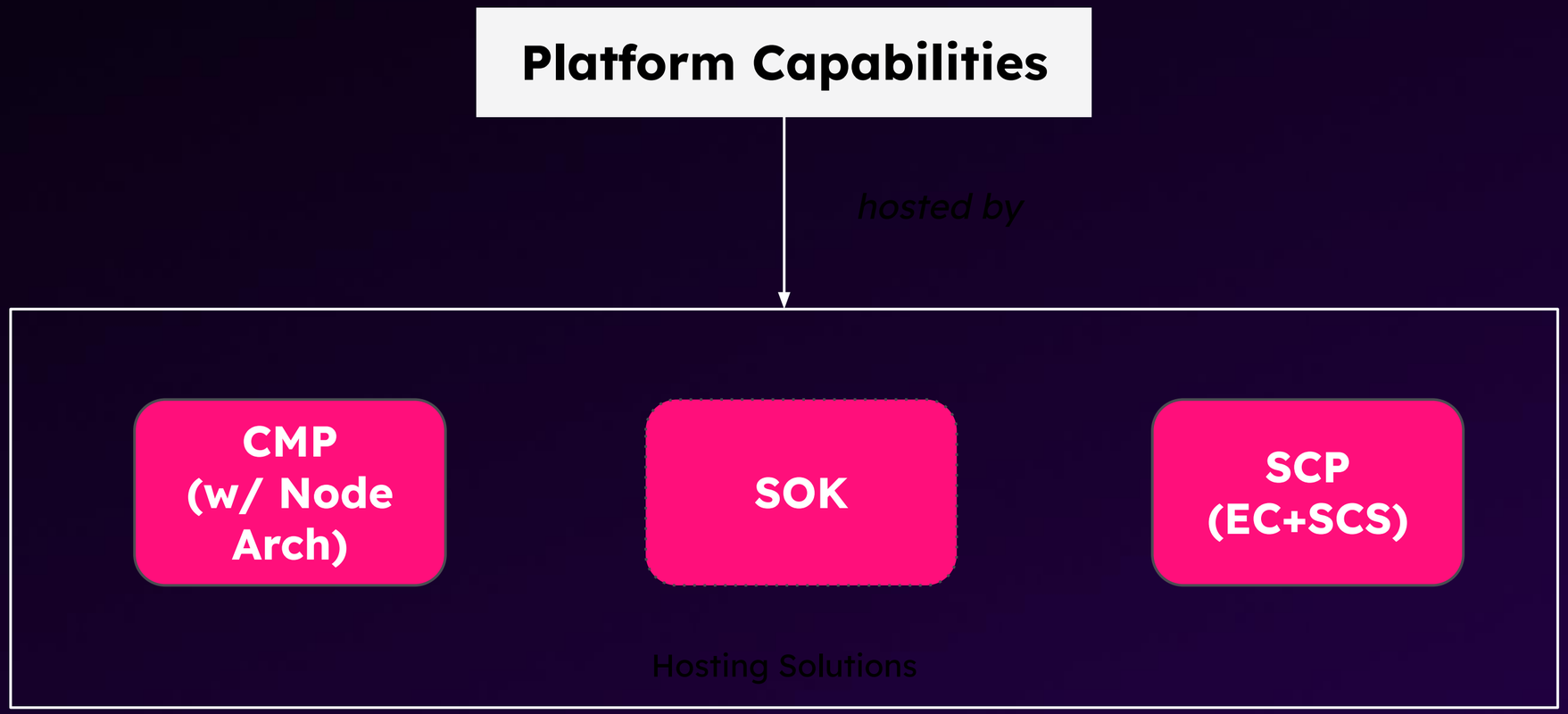
The Admin Reality

- Brittle scripts for scaling and upgrades
- Config-drift across dev / test / prod clusters
- Tedious patching & routine maintenance windows
- Hardware under-utilisation and cost creep
- Pressure to adopt containers *without rewriting everything*

What SOK Delivers

- Declarative deployments, upgrades & rollback via CRDs
- Splunk-validated topologies (SVA templates out-of-the-box)
- Hands-off lifecycle automation – pods self-heal, SmartStore mounts auto-attach
- GPU & resource-aware scheduling for AI Tier workloads
- Free, open-source & community-supported

SOK Is a Strategic Investment for the Splunk Platform



Prerequisite #3: Splunk Enterprise 10.2 and AI Assistant App

Release Phase	Enterprise version requirements	Splunk AI Assistant for SPL
GA	10.2	Splunkbase
Preview	Private build of 10.0	Private build from VOC portal
Public preview (TBD)	10.2 public preview program	TBD

Install flow for Splunk AI Assistant for SPL on AI Tier

Hardware level	GPU ready
Software - Infrastructure	AI Tier with SOK • AI Tier CR -> triggers Operator reconciliation • Operator installs required dependencies (Ray, Weaviate, GPU Nodes) • Auto-scale based on load & GPU requirements • Secure authentication via Splunk
Software - Application	SAIA App • Download and Install SAIA App • Follow the instructions to choose fully on-prem solutions and connect SAIA App to AI Tier

Call to action

Join the Splunk AI Assistant through AI tier with SOK, Preview

1. **Confirm eligibility** – you already have GPU readily available
2. **Visit the VOC Preview Portal**: https://voc.splunk.com
3. **Select "Splunk AI Assistant for SPL through AI tier with SOK" Preview**
4. **Complete the 3-minute survey** – tell us more about your data.
5. **Accept the electronic EULA** for preview.
6. **Follow instructions to onboard**

- **Preview starts Oct 15, 2025.**
- **Registration has opened!**

Interested or have questions? Email

- Nick Ma, AI PM, at qiumma@cisco.com
- Jahnavi Reddy, AI PM, at jahreddy@cisco.com

Key takeaway:

- Fully on-prem Splunk Enterprise? **AI Tier** is your path to GenAI.
- Use AI Tier when **Cloud-Connected** isn’t an option.
- **Prereqs (3):** Customer-managed **GPUs**, **SOK**,
  Splunk Enterprise 10.2+/Preview is 10.0 private build
- **Preview opens October 15 —** enroll to participate.

Q&A