

Building a leading observability practice in the era of AI

cisco Live !

Diana Omuoyo

Global Technical Observability Specialist

Cisco Webex App

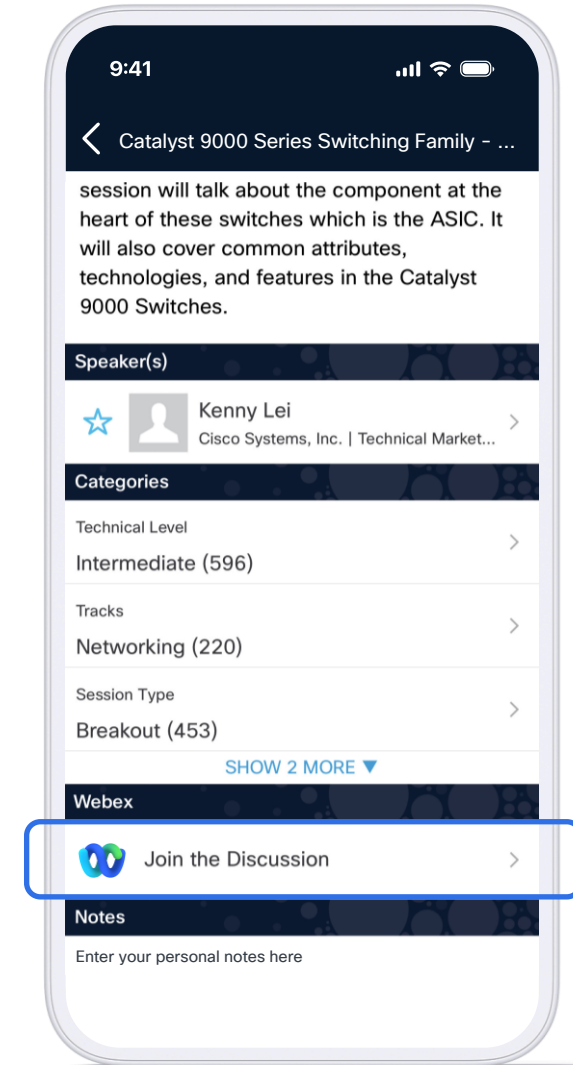
Questions?

Use Cisco Webex App to chat with the speaker after the session

How

- 1 Find this session in the Cisco Live Mobile App
- 2 Click “Join the Discussion”
- 3 Install the Webex App or go directly to the Webex space
- 4 Enter messages/questions in the Webex space

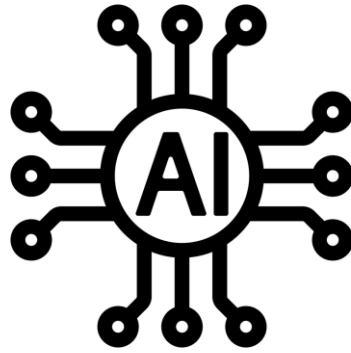
Webex spaces will be moderated by the speaker until 14 November 2025.



<https://ciscolive.ciscoevents.com/ciscolivebot/#BRKOBS-2547>

**~ Fascinated by the Future
of Intelligence ~**





AI Defined

Varied definitions

..the ‘digital experts’ definition of artificial intelligence

“Artificial intelligence (AI) is like *teaching* computers and machines to think, learn, and make decisions the way humans do, but in their own way.”

Internet Search

"AI is when computers *learn to do smart things*, like answering questions, remembering stuff, or recognizing pictures – just like your brain, but made of code." [Chat-GPT]

Chat-GPT

Evolution of AI

Why we are fascinated by it

Early Concepts - 1700s

AI-Boom - 1950s

Modern AI - Late 1990s

Applications evolving to use AI



Predictive Analytics
Machine Learning



Facial Recognition
Deep Learning

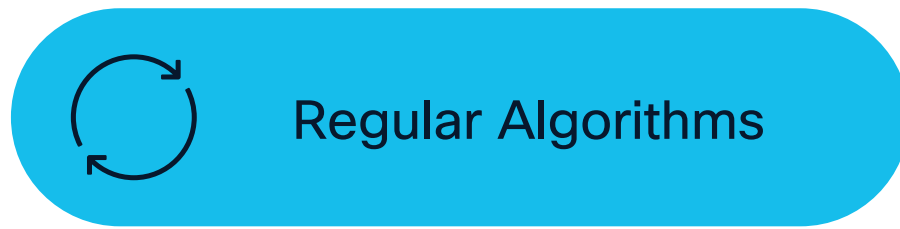
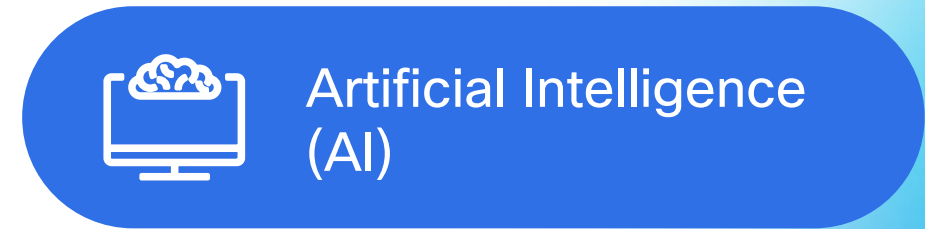


Voice Assistants
Natural Language



ChatBots
Generative AI

How Gen AI different from regular algorithms



A revolution is happening

Productivity is moving
from human centered
to AI powered

Digital ecosystems
are being
transformed by
agentic applications

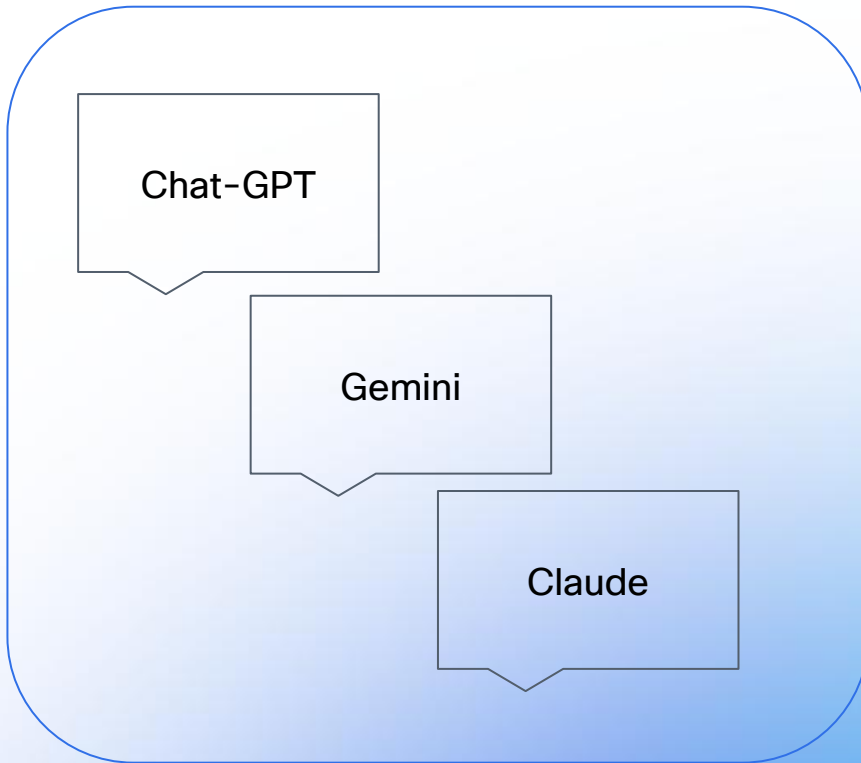
High-quality | scale
data is driving
unprecedented
innovation

Why is this happening now

‘Cambrian explosion’ of LLM models

...different types, sizes & focus areas

Closed Source

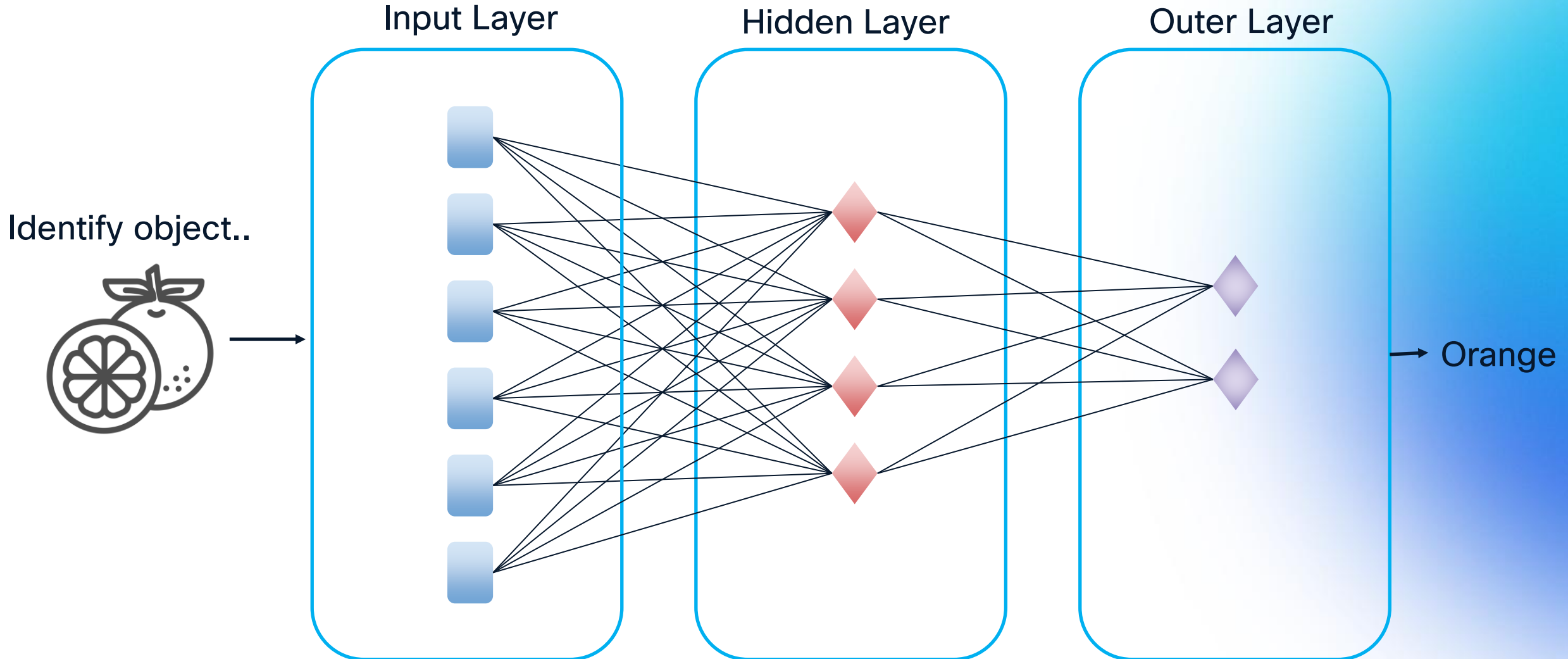


Open Source



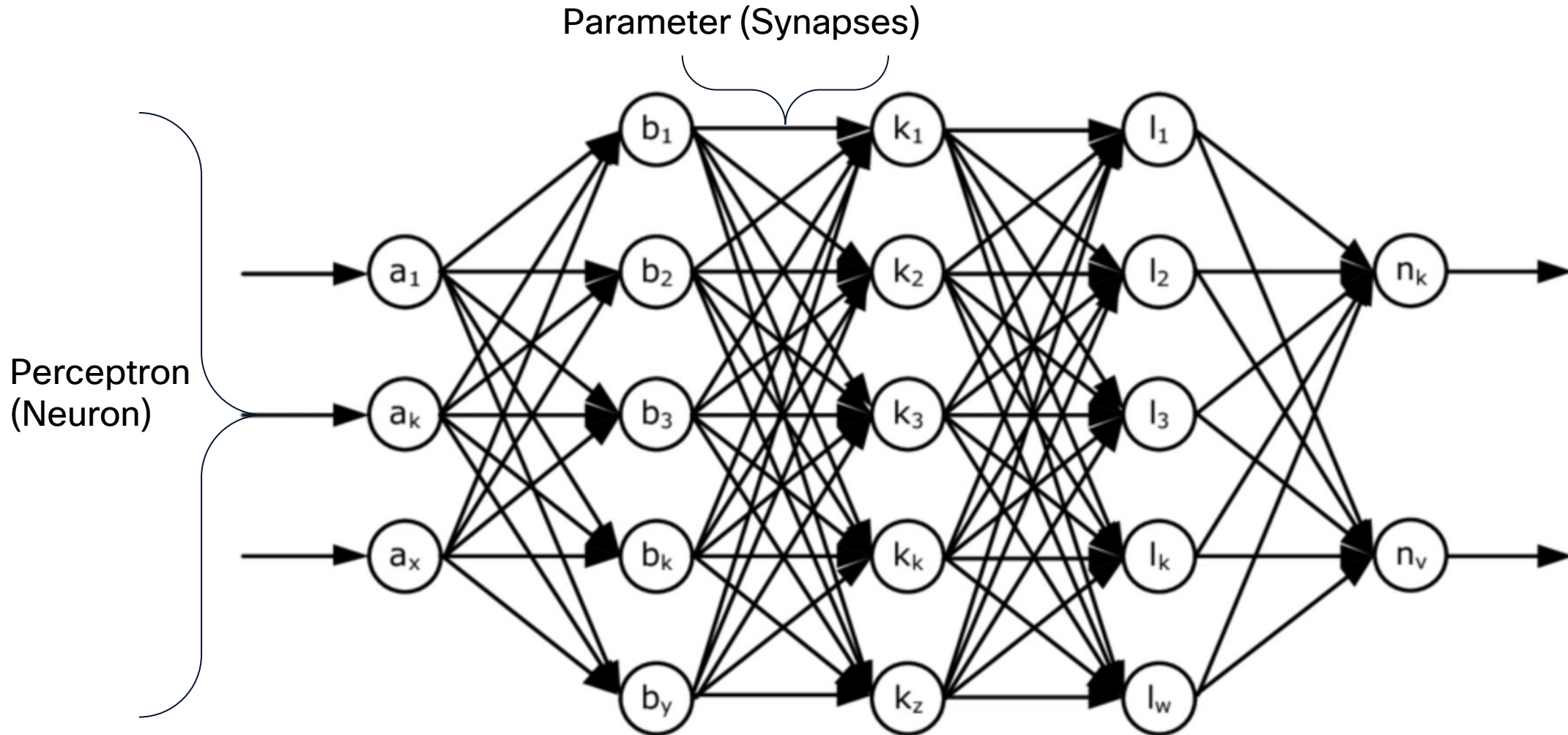
Neural networks advancing pattern identification

...covering large amounts of complex data



Large scale processing capability

...silicone chips – high density, high performance GPUs



Transformer architecture

...attention mechanism

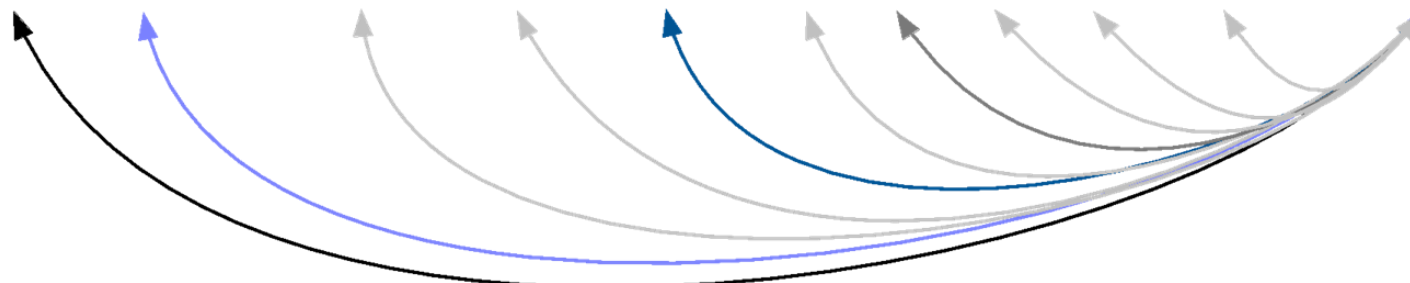
The attention mechanism adds **contextual information** to words in a sentence

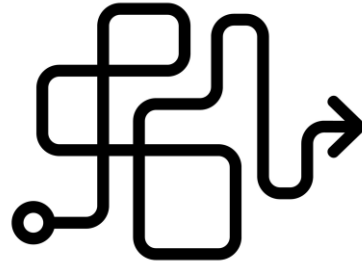
“I swam across the river to get to the other **bank**.”

You have no problem interpreting “bank” in the following sentence:

A machine needs some help...

I swam across the river to get to the other bank.





Areas of complexity along the AI journey

Complexities surrounding AI adoption

...the data challenge



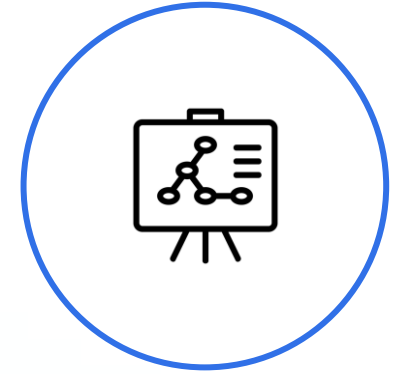
Explosive data growth

By 2026, large enterprises will triple their unstructured data capacity compared to 2023*



Proliferation of data sources and types

Logs, Metrics, Traces, Events
Cloud, On-prem, IOT

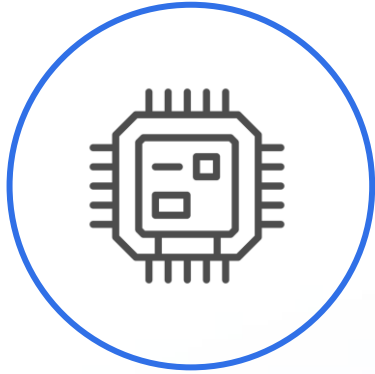


Disparate use cases

Troubleshooting
Threat investigation
Compliance

Complexities surrounding AI adoption

...the infrastructure challenge



Not enough processing
power

Resource constraints
globally



Cost of running AI
datacenters

Procuring, managing,
securing and optimizing

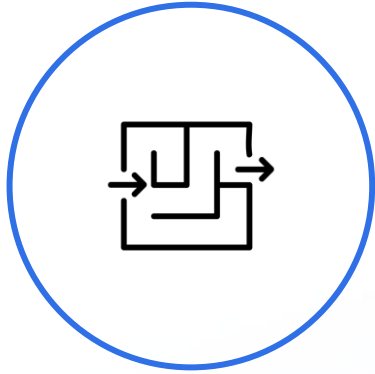


Shifting
investment

Constant disruption & rapid
change

Complexities surrounding AI adoption

...the operational challenge



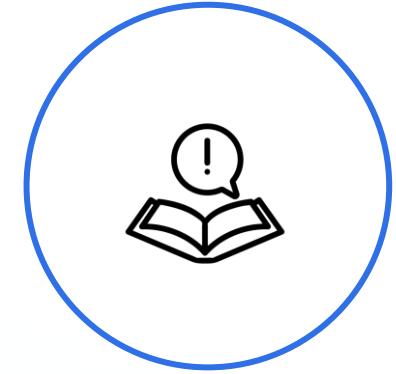
AI predictability

Lack of trust and predictability for AI



Security

Guard rails for new AI powered threats



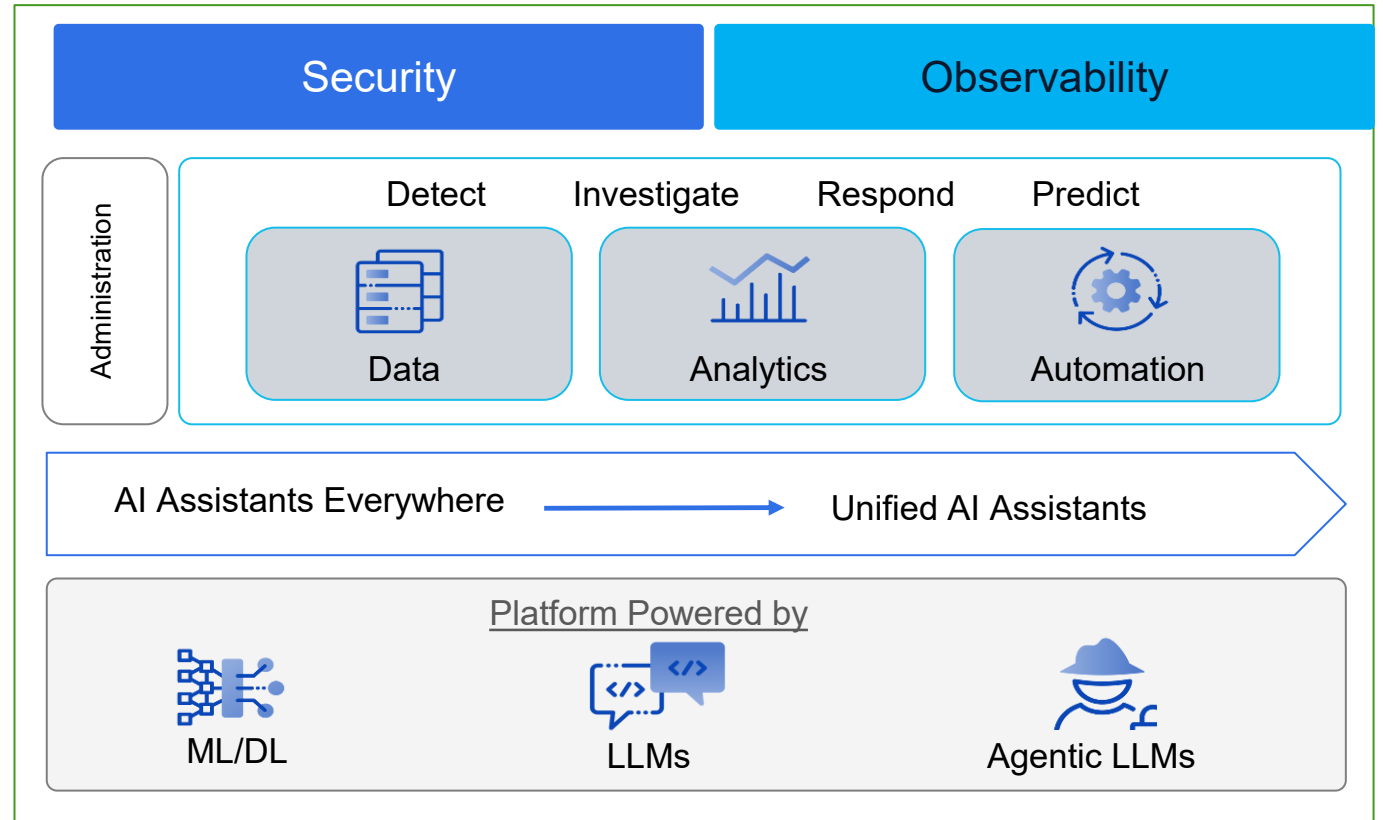
Knowledge Gap

Dynamic & fast incremental changes in AI

Splunk Observability & AI

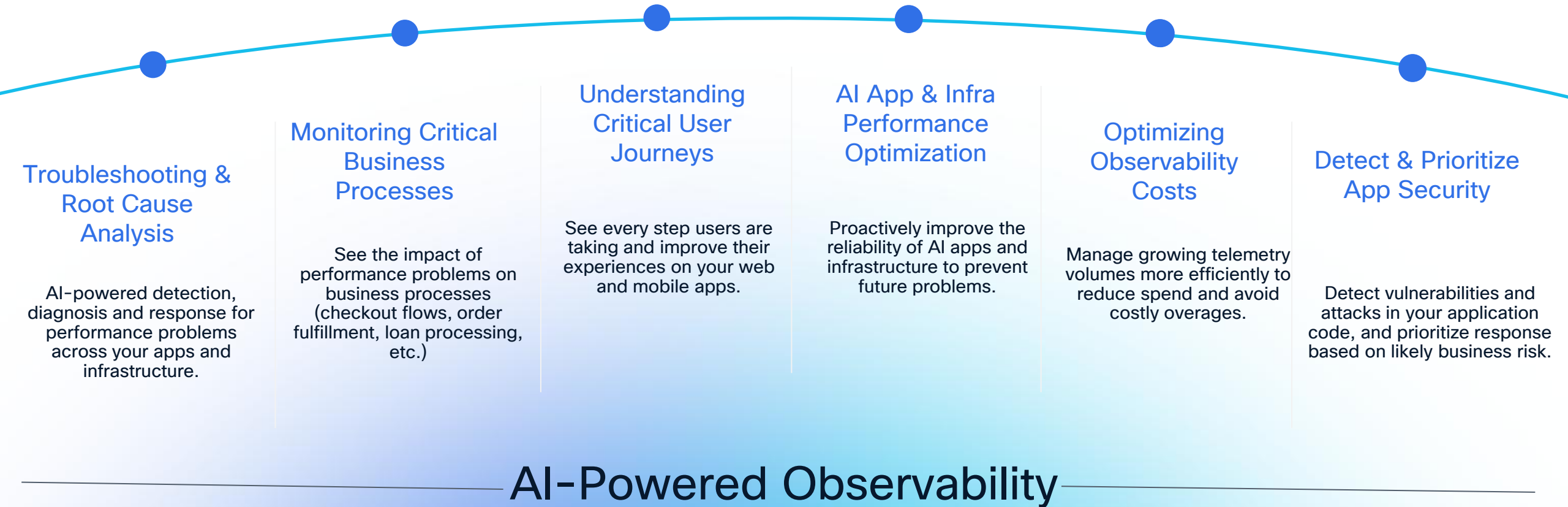
Splunk's foundation of AI services + data

- Generative **AI Assistants** to make your job easier
- Powered by **leading AI technologies** to deliver accuracy and efficiency
- Integral part of the **Splunk portfolio** and Data Platform



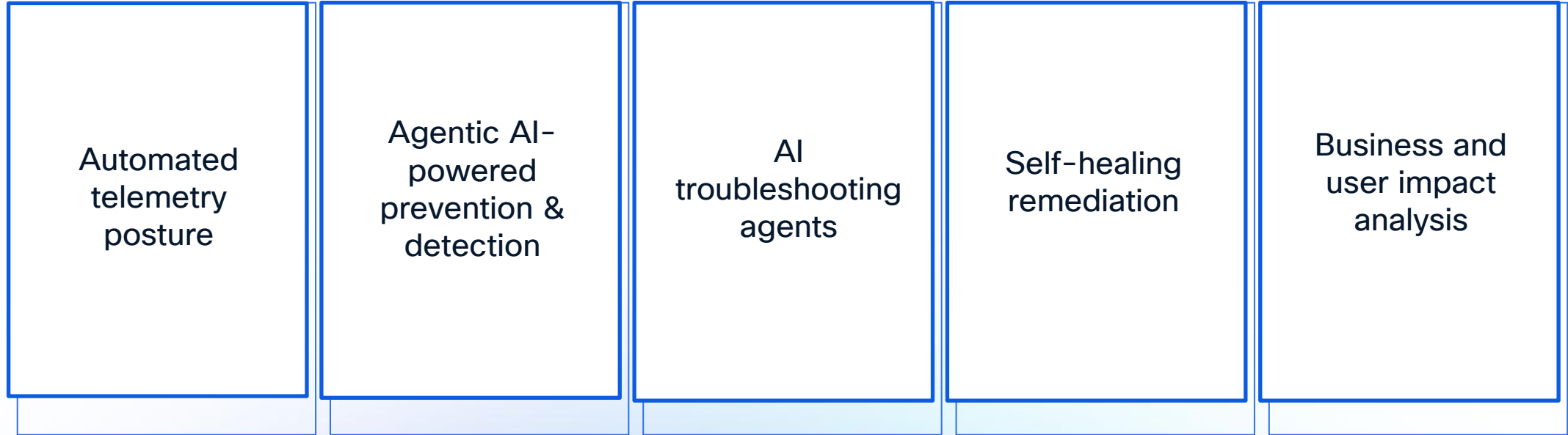
Building an observability practice in the AI era

...foundational critical observability capabilities across your stack



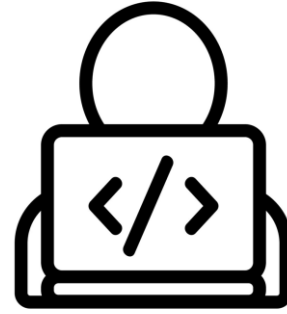
Building an observability practice in the AI era

...a proactive approach to staying ahead with embedded AI



Agentic AI-Powered Observability

AI Assistants

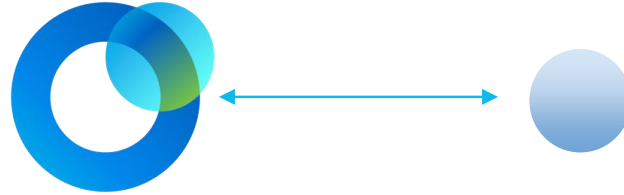


What is an AI Assistant

An **AI assistant** responds to your prompts with information or guidance

AI Assistants have ‘skills’, not features

...domain specific skills



AI Assistant

Domain Portfolio

Natural Language to Query
Translator

“Find the slowest dependency chain for the ‘place-order’ transaction that caused latency degradation for Sydney users.”

System Health Pattern
Summarizer

“Group similar stack traces and show the top three dominant error patterns in auth-service since midnight.”

AI Assistants portfolio across Cisco | Splunk

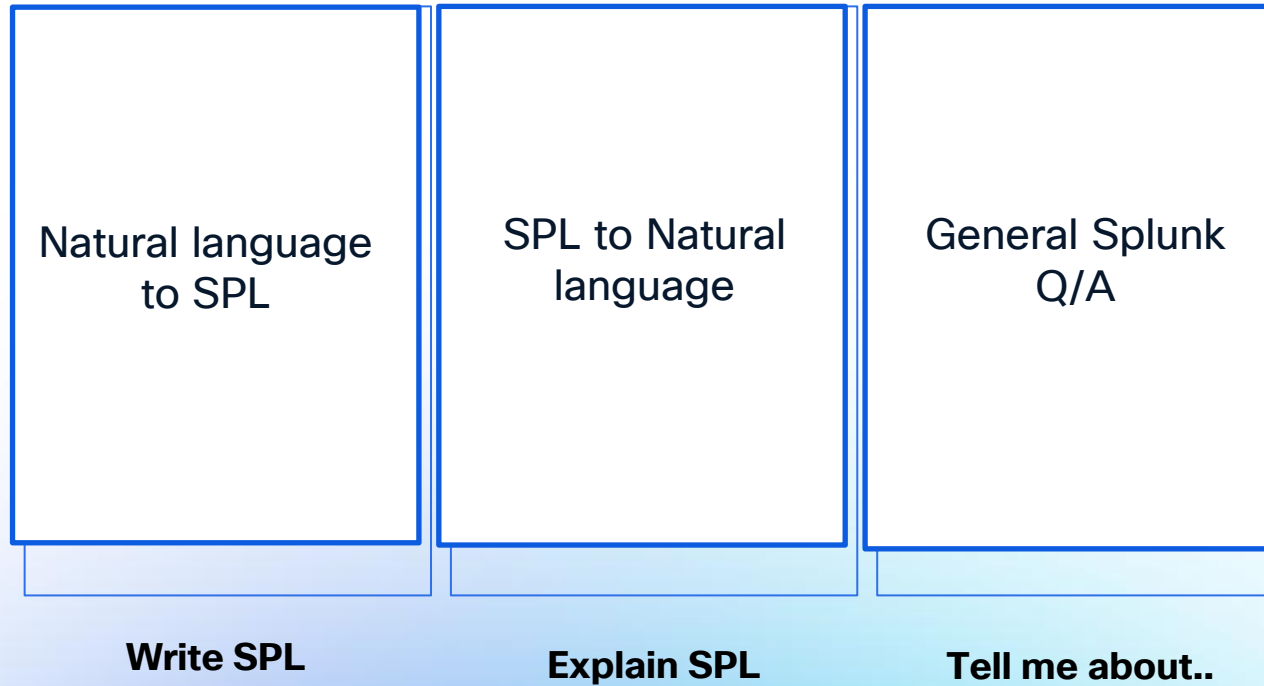
	Security	Firewall, Secure Access, Hypershield, Duo, Identity Intelligence, Splunk Enterprise Security
	Networking	Meraki, Catalyst Center, Catalyst SD-WAN, ThousandEyes, Intersight, Mobility Services
	Observability	Splunk Observability Suite
	Data	Splunk Platform
	Collaboration	Webex Control Hub
	Service Ops	Customer Experience

AI for Splunk Cloud

cisco Live !

Splunk AI Assistant for SPL

...core capabilities



Splunk AI Assistant for SPL

Write or explain SPL with bi-directional translation

Upskill new and advanced Splunk users quickly

Access the full knowledge base for answers in Splunk product

The screenshot displays the Splunk AI Assistant interface. On the left, a sidebar contains 'Chat', 'Usage guidelines', and 'Settings'. The main chat area shows a user query: 'Using data models, get top 10 users with the highest number of failed login attempts over the past week'. The assistant responds with the SPL command: `| tstats count from datamodel=Authentication.action="failure" earliest=Authentication.user | top limit=10 Authentication.user -7d@d latest=now() by`. Below the command, the assistant explains: 'The query retrieves the top 10 users with the highest number of failed authentication attempts in the past 7 days.' and 'The tstats command is used to retrieve statistical information about indexed data, in this case, counting the number of events.' It also specifies that the data is retrieved from the 'Authentication' data model. A feedback section at the bottom left asks for user input. A green button with a paper plane icon is at the bottom right.

Chat

Usage guidelines

Settings

Using data models, get top 10 users with the highest number of failed login attempts over the past week

```
| tstats count from datamodel=Authentication.action="failure" earliest=Authentication.user | top limit=10 Authentication.user -7d@d latest=now() by
```

The query retrieves the top 10 users with the highest number of failed authentication attempts in the past 7 days.

SPL command details

SPL command details

The tstats command is used to retrieve statistical information about indexed data, in this case, counting the number of events.

This specifies that the data should be retrieved from the Authentication data model, which contains structured data related to user authentication events.

Give us feedback!

Want to share your feedback?
Contact us by email at misupport@splunk.com

Splunk AI Assistant is under development and actively improving responses based on your feedback.



Write SPL to determine who made modifications to any AWS security groups



SPL Query:

```
index=botsv4 sourcetype=aws:cloudtrail AND (eventSource="ec2.amazonaws.com" AND  
(eventName="AuthorizeSecurityGroupIngress" OR eventName="AuthorizeSecurityGroupEgress" OR  
eventName="RevokeSecurityGroupIngress" OR eventName="RevokeSecurityGroupEgress" OR eventName="CreateSecurityGroup"  
OR eventName="DeleteSecurityGroup"))
```

Explain SPL

Copy

Open in Search



Output results in table format



BRKOB5-2547

```
eventName="RevokeSecurityGroupIngress" OR eventName="RevokeSecurityGroupEgress" OR eventName="CreateSecurityGroup"  
OR eventName="DeleteSecurityGroup"))
```

[? Explain SPL](#)[Copy](#)[Open in Search](#)

Output the results in table format



SPL Query:

```
index=botsv4 sourcetype=aws:cloudtrail AND (eventSource="ec2.amazonaws.com" AND  
(eventName="AuthorizeSecurityGroupIngress" OR eventName="AuthorizeSecurityGroupEgress" OR  
eventName="RevokeSecurityGroupIngress" OR eventName="RevokeSecurityGroupEgress" OR eventName="CreateSecurityGroup"  
OR eventName="DeleteSecurityGroup"))  
| table _time, eventName, userIdentity.userName, eventName, requestParameters, responseElements
```

[? Explain SPL](#)[Copy](#)[Open in Search](#)

This will output the results in a table format, showing the timestamp, event name, user who made the modification, event name, request parameters, and response elements for each modification to an AWS security group.



> Related content



```
eventName="RevokeSecurityGroupIngress" OR eventName="RevokeSecurityGroupEgress" OR eventName="CreateSecurityGroup"  
OR eventName="DeleteSecurityGroup"))
```

[? Explain SPL](#)[Copy](#)[Open in Search](#)

Output the results in table format



SPL Query:

```
index=botsv4 sourcetype=aws:cloudtrail AND (eventSource="ec2.amazonaws.com" AND  
(eventName="AuthorizeSecurityGroupIngress" OR eventName="AuthorizeSecurityGroupEgress" OR  
eventName="RevokeSecurityGroupIngress" OR eventName="RevokeSecurityGroupEgress" OR eventName="CreateSecurityGroup"  
OR eventName="DeleteSecurityGroup"))  
| table _time, eventName, userIdentity.userName, eventName, requestParameters, responseElements
```

[? Explain SPL](#)[Copy](#)[Open in Search](#)

This will output the results in a table format, showing the timestamp, event name, user who made the modification, event name, request parameters, and response elements for each modification to an AWS security group.



> Related content

What is AuthorizeSecurityGroupIngress?



BRK0BS-2547

Results from GenAI can vary; review for accuracy. [View AI details](#)

This will output the results in a table format, showing the timestamp, event name, user who made the modification, event name, request parameters, and response elements for each modification to AWS security groups.



> Related content



What is AuthorizeSecurityGroupIngress?



AuthorizeSecurityGroupIngress is an AWS API action that allows you to add one or more inbound rules to a security group. In other words, it is used to authorize incoming traffic to a security group.

In the context of the SPL query I provided earlier, **AuthorizeSecurityGroupIngress** is one of the event names that we are searching for in the AWS CloudTrail logs. This event is triggered when someone adds a new inbound rule to a security group, which can potentially allow incoming traffic to reach the resources associated with that security group.

By including **AuthorizeSecurityGroupIngress** in the query, we are trying to capture any modifications made to the security group's inbound rules, which could potentially impact the security posture of the AWS environment.

Here's a brief summary of the four event names we are searching for:

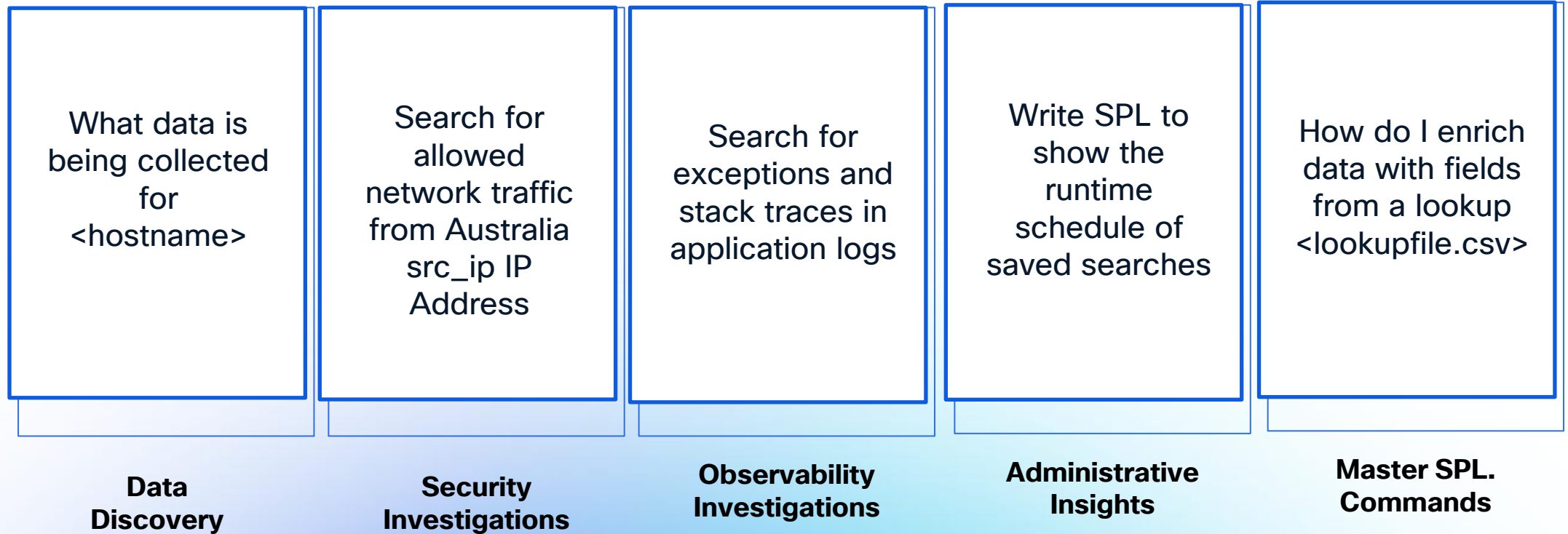
- * **AuthorizeSecurityGroupIngress**: Add inbound rules to a security group
- * **AuthorizeSecurityGroupEgress**: Add outbound rules to a security group
- * **RevokeSecurityGroupIngress**: Remove inbound rules from a security group
- * **RevokeSecurityGroupEgress**: Remove outbound rules from a security group

By monitoring these events, we can get a better understanding of who is making changes to the security groups and what changes are being made.



Splunk AI Assistant for SPL

...use cases

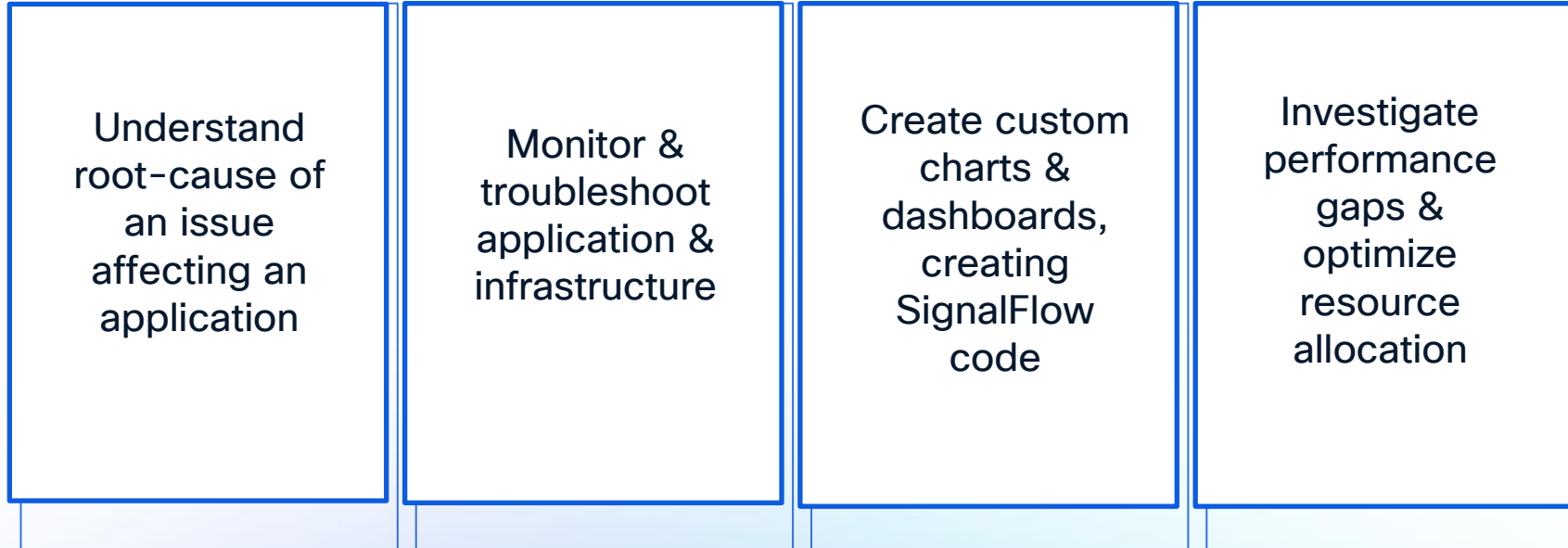


AI Assistant for Splunk Observability

CISCO Live !

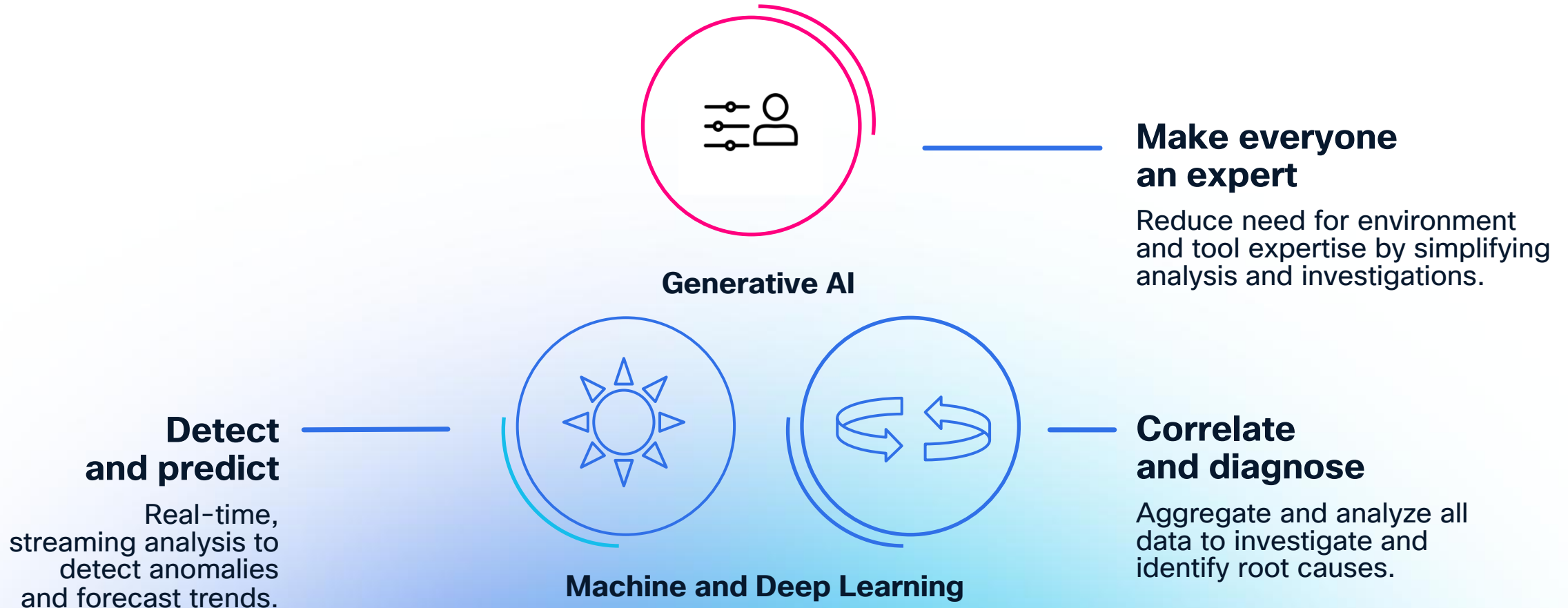
Building an observability practice in the AI era

...how AI Assistant can help



Building an observability practice in the AI era

...machine learning. deep learning | generative AI



Building an observability practice in the AI era

Scenario: On-call SRE paged to investigate a recent application failure

Investigating without AI Assistant (~40 min*)

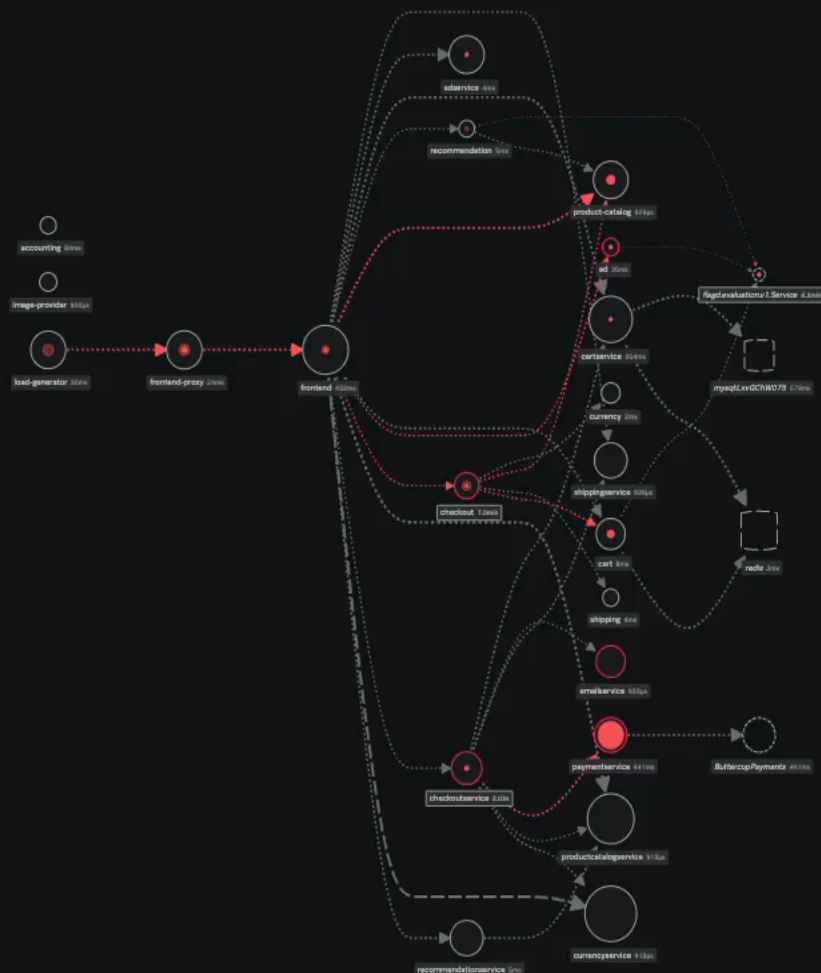
1. **Review alert (2m)**
 - a. Manually review alert to determine investigation starting point
2. **Review service map and dependencies (5m)**
 - a. Leverage tribal knowledge and experience to determine investigation next steps
3. **Review APM traces (10m)**
 - a. Manually review APM service tree for failing services
 - b. Manually review failed trace metadata to further scope the investigation
4. **Review infrastructure metrics (10m)**
 - a. Manually review related infrastructure metrics to further scope the investigation
5. **Review application logs (10m)**
 - a. Manually review related application logs to determine the probable cause of failure

Investigating with AI Assistant (~7 min*)

1. **Review alert (2m)**
 - a. Manually review alert to determine investigation starting point
2. **Ask AI Assistant to investigate (5m)**
 - a. Ask assistant to analyze APM issues using generative AI
 - b. Using conversational chat, ask the assistant to determine the probable cause of the failure

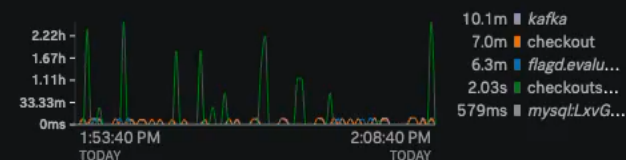
*Results will vary depending on complexity of case

🔄 Refresh



+

Service by Latency (P90)



Tag Spotlight

Ⓢ Logs (0)

Customers already notice dramatic improvements during their troubleshooting

With Application Performance Management (APM), our MTTx has dropped drastically—what used to take 30–60 minutes is now resolved in just **5–10** minutes, thanks to AI!

Fortune 500 industrial manufacturing company

No one can be a subject matter expert on everything...imagine being **able to ask any question** about our metrics, data, and system correlations...

Fintech real-time payment processing company

What used to take engineers 10–30 minutes digging through logs and APM data now takes just **1–2** minutes with AI.

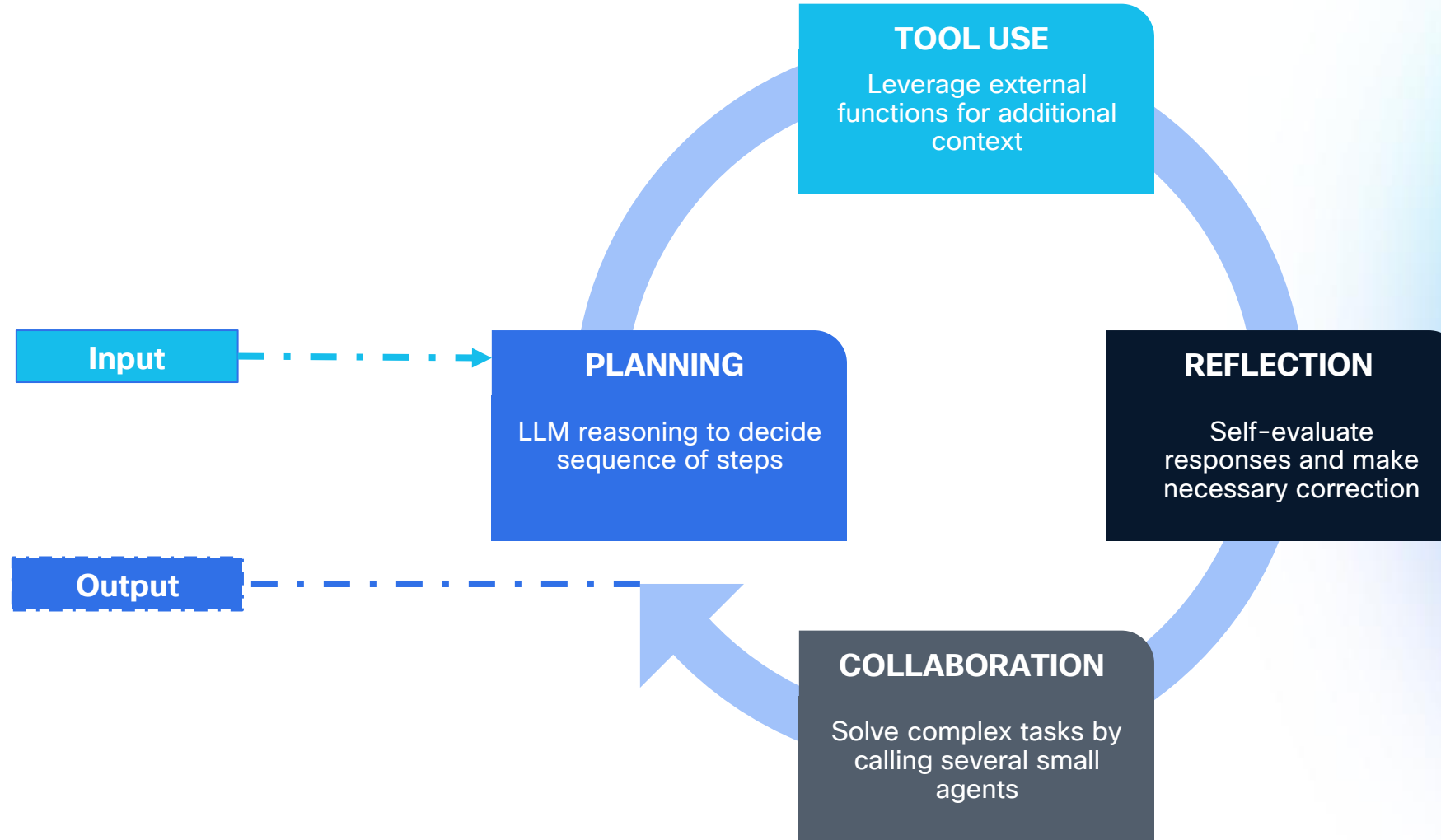
Multinational investment holding company

I no longer have to chase down engineers to find out what's wrong—AI gives me the answers instantly, **saving me 15–30** minutes every time.

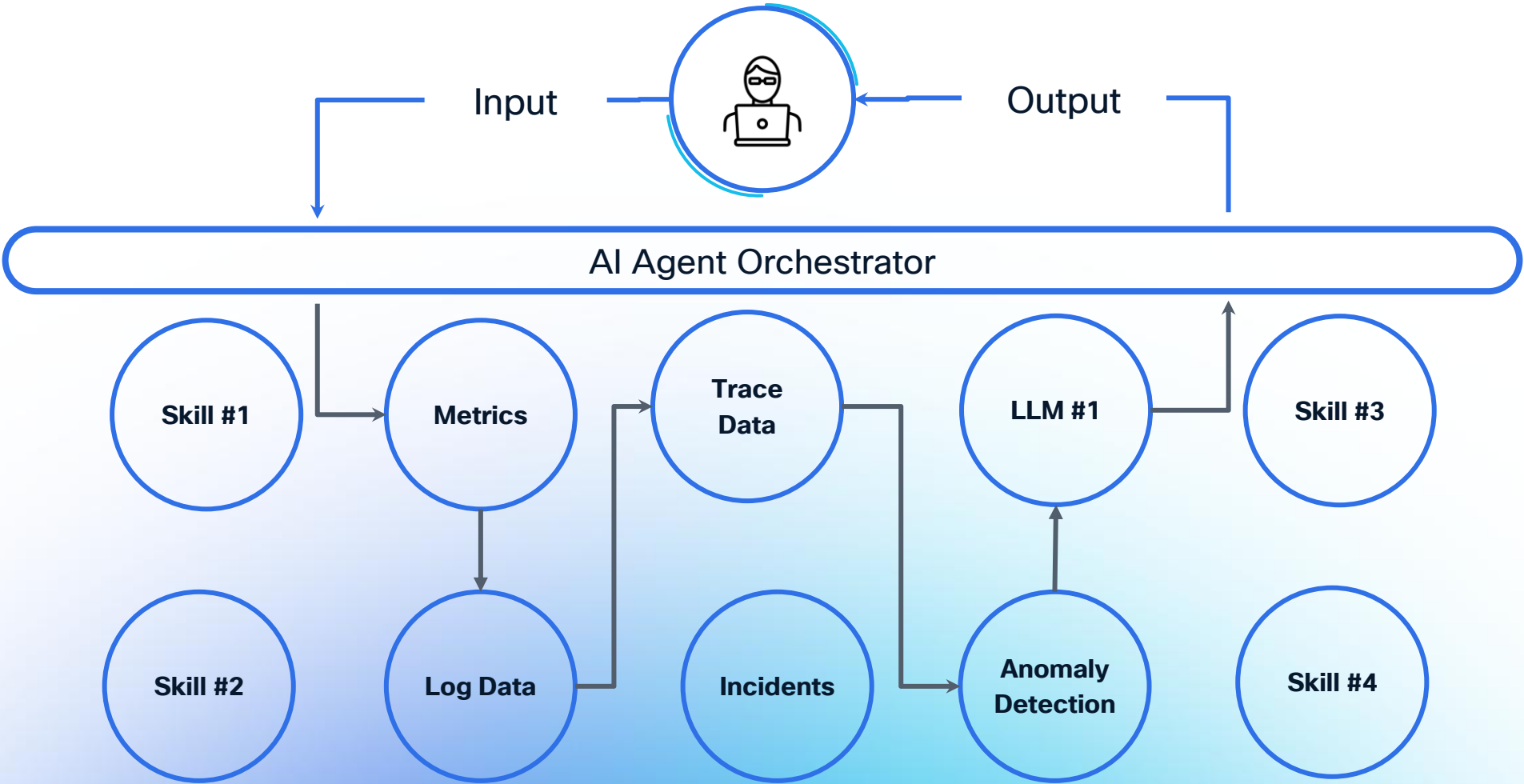
Agentic AI

CISCO Live !

Agentic system autonomy pattern



Agentic system autonomy pattern

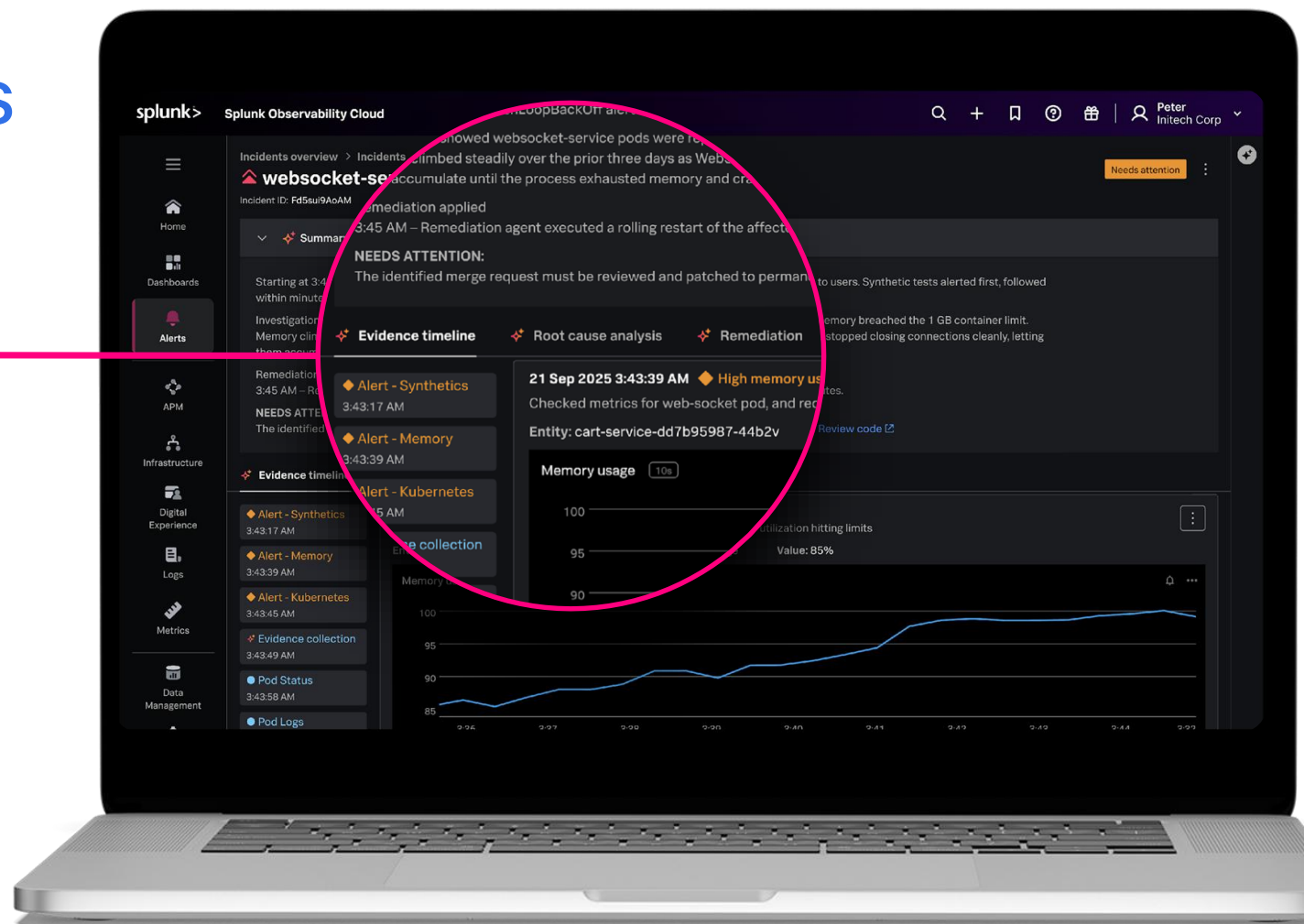


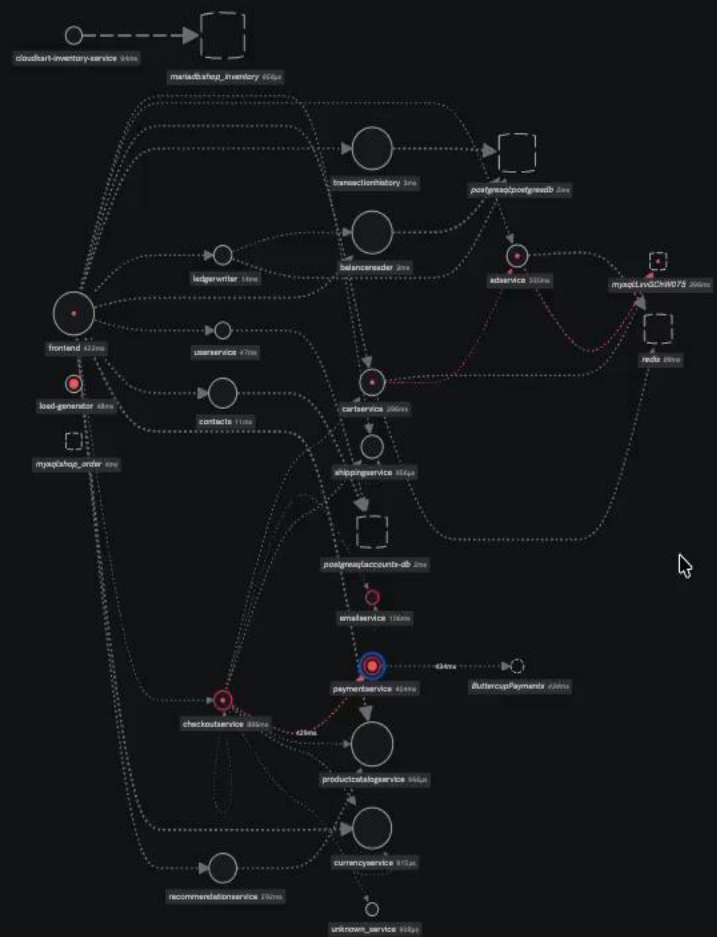
AI Troubleshooting Agents for Splunk Observability

Fix issues faster with automated Root Cause Analysis

Agents have context into your data

True AI native, interactive experience





paymentservice

Breakdown

Add to Filter

Service Requests & Errors

2.14/s Requests
1.9k total

1.15/s Errors
1k total (54%)

1.15/s Root Cause
1k total (54%)

4/s
2/s
0/s

1:31:20 PM TODAY

1:46:20 PM TODAY

Service & Dependency Latency

P50

P90

P99

SERVICE

404ms

This Service

400ms
200ms

1:31:20 PM TODAY

1:46:20 PM TODAY

Dependencies

1 → 1 → 1

Traces

View Trace ID

View Trace ID

Go

Endpoint Performance

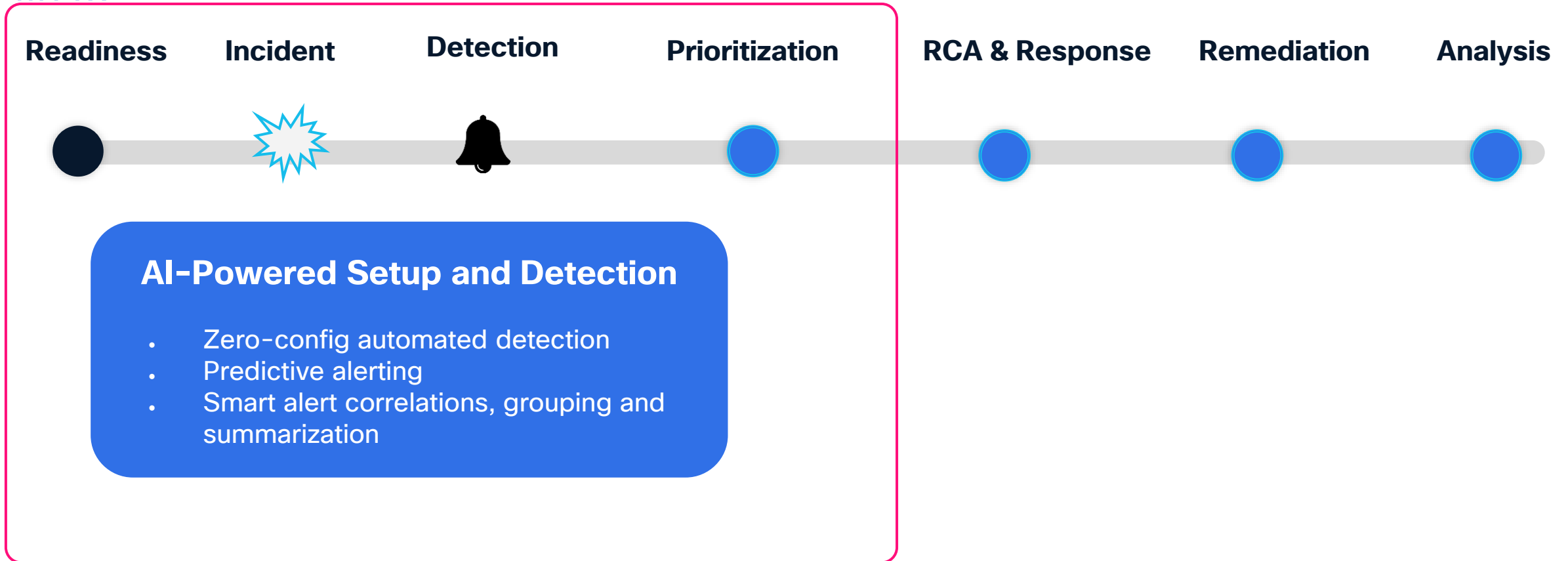
Endpoint	P90
grpc.hipstershop.PaymentService/Charge	434ms
/CreditCardInfo	22ms
grpc.grpc.health.v1.Health/Check	268µs

Tag Spotlight

Our Vision for Agentic AI

Our mission: automate the DevOps incident workflow

Agentic & predictive AI for problem detection, isolation and root cause inference to reduce MTTx



Our mission: automate the DevOps incident workflow

Agentic & predictive AI for problem detection, isolation and root cause inference to reduce MTTx

Readiness

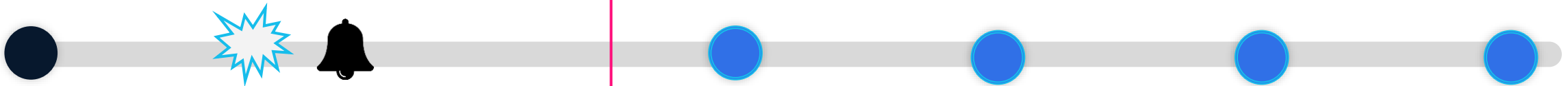
Incident Detection

Prioritization

RCA & Response

Remediation

Analysis



AI-Directed Troubleshooting

- 1-click automated Root Cause Analysis
- AI-driven insights on business and end-user impact
- Incident summarization and recommendation

Our mission: automate the DevOps incident workflow

Agentic & predictive AI for problem detection, isolation and root cause inference to reduce MTTx

Readiness

Incident Detection

Prioritization

RCA & Response

Remediation

Analysis

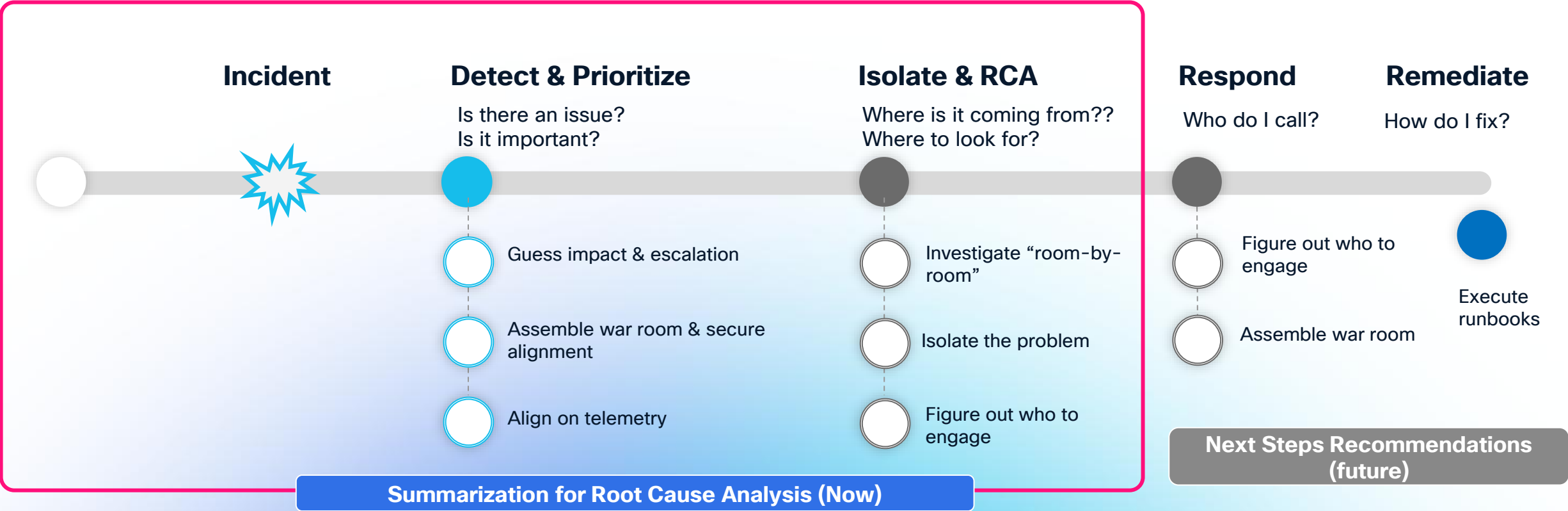


Conversational and In-Context AI

Chat-based AI Assistant for easy, interactive insights
Embedded in-product AI experiences
Slack / Teams AI Agents

Automated troubleshooting workflow

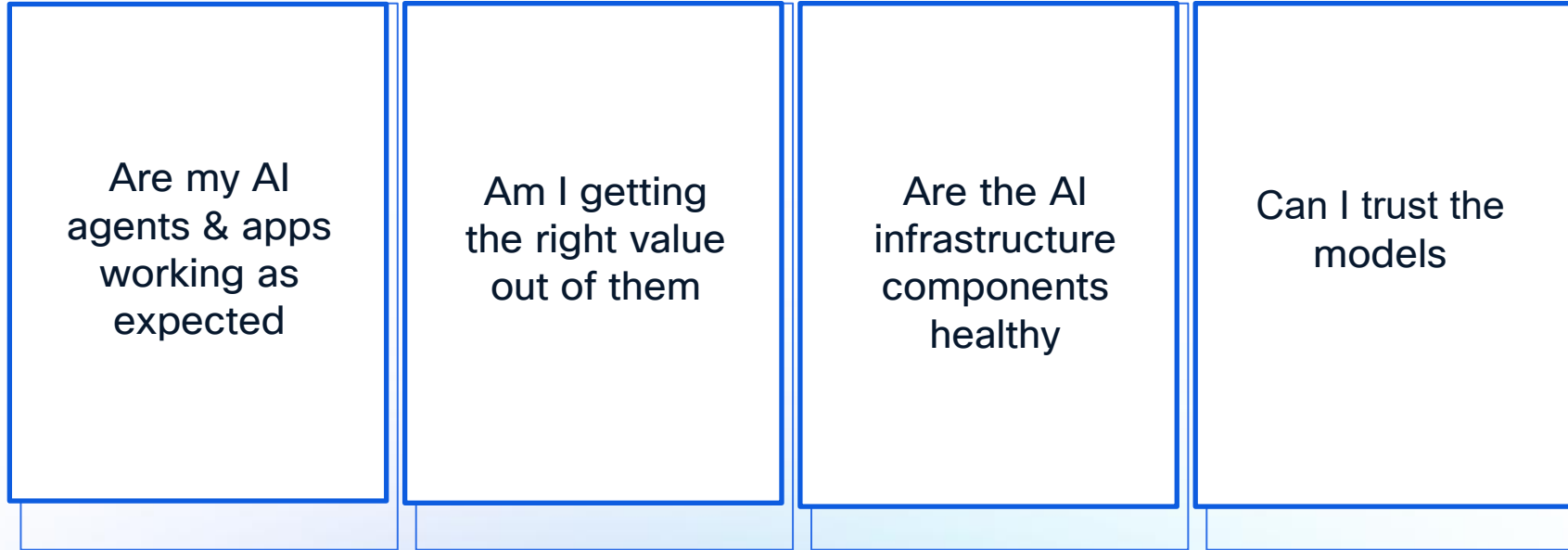
Focus is on building AI-driven problem detection, isolation and root cause inference to reduce MTTx



Observability for AI

The AI stack puts new demands on observability

...we are now asking different questions



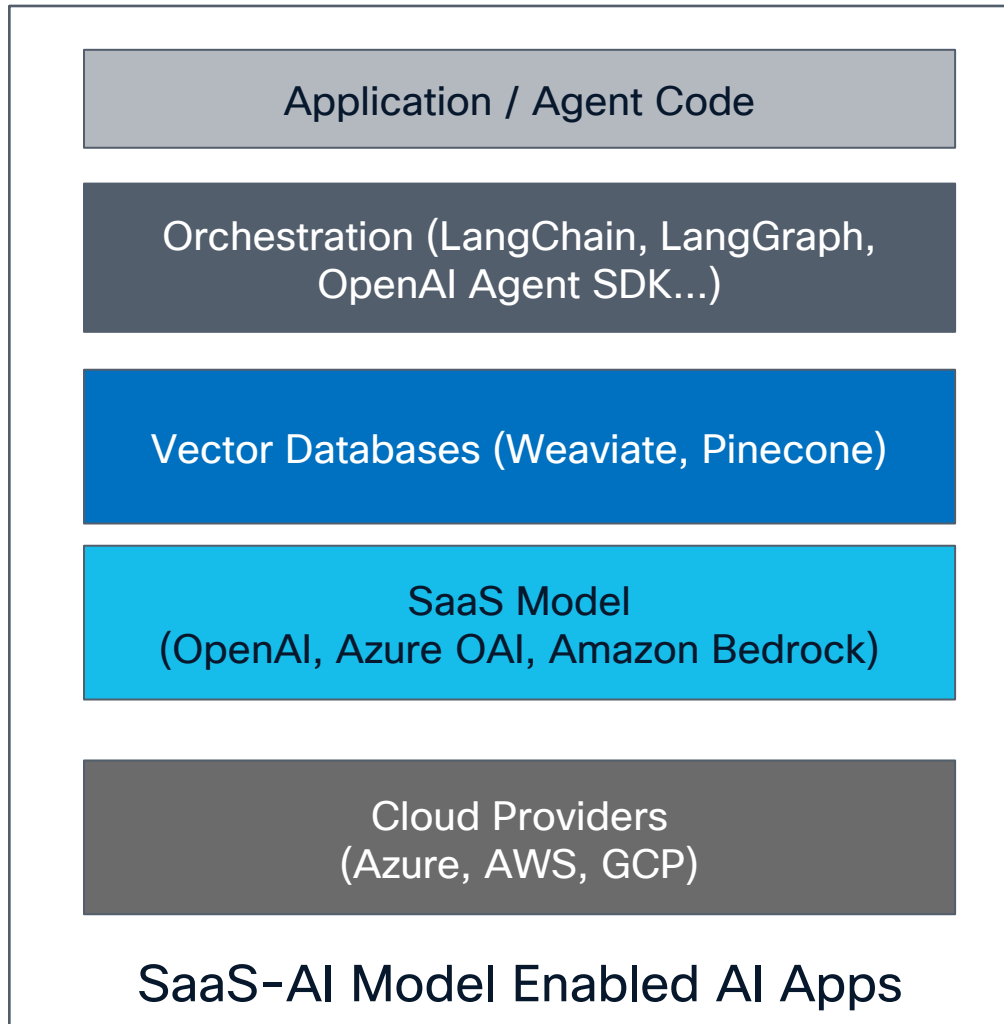
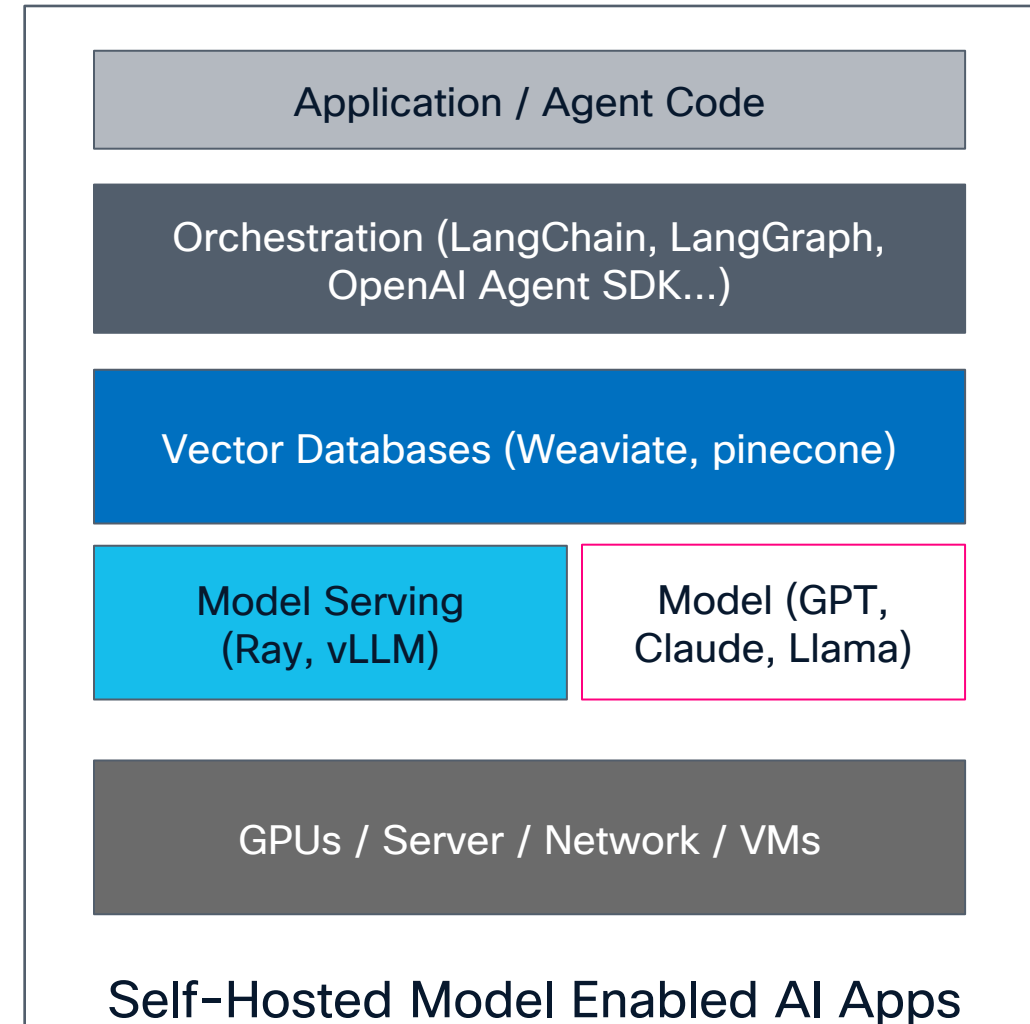
Observability for AI

**“It's a cardinal error to
theorize without data”**

-Sherlock holmes

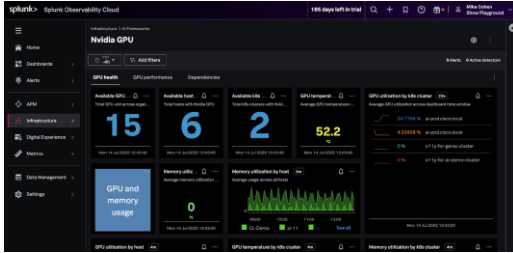
Observability coverage for AI & agentic apps

...monitor & track the health of your AI stack



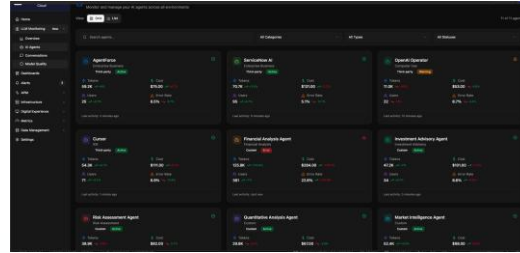
Observability coverage for AI & agentic apps

...monitor & track the health of your AI stack



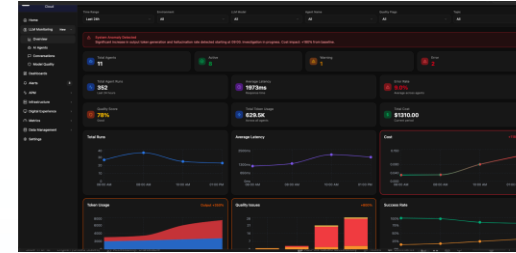
AI infrastructure monitoring

Out-of-the-box views for monitoring the AI infrastructure stack, all the way from the hardware (GPUs) to LLM models, LLM orchestration frameworks, and vector databases



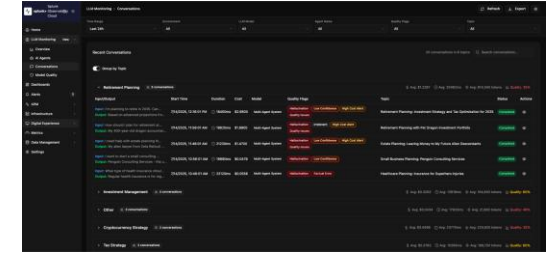
AI Agent observability

Understand agents across the overall environment and at a per agent level, showing key agent metrics around latency, error rate, token usage and cost.



Conversation insights

Code-level integration into LLM frameworks to provide visibility into entire conversations with agents including individual prompts/responses.



Agent quality evaluations

Understand how the agents are faring in production, including prompt level characterizations around hallucinations, confidence levels and costs.

AI Infrastructure

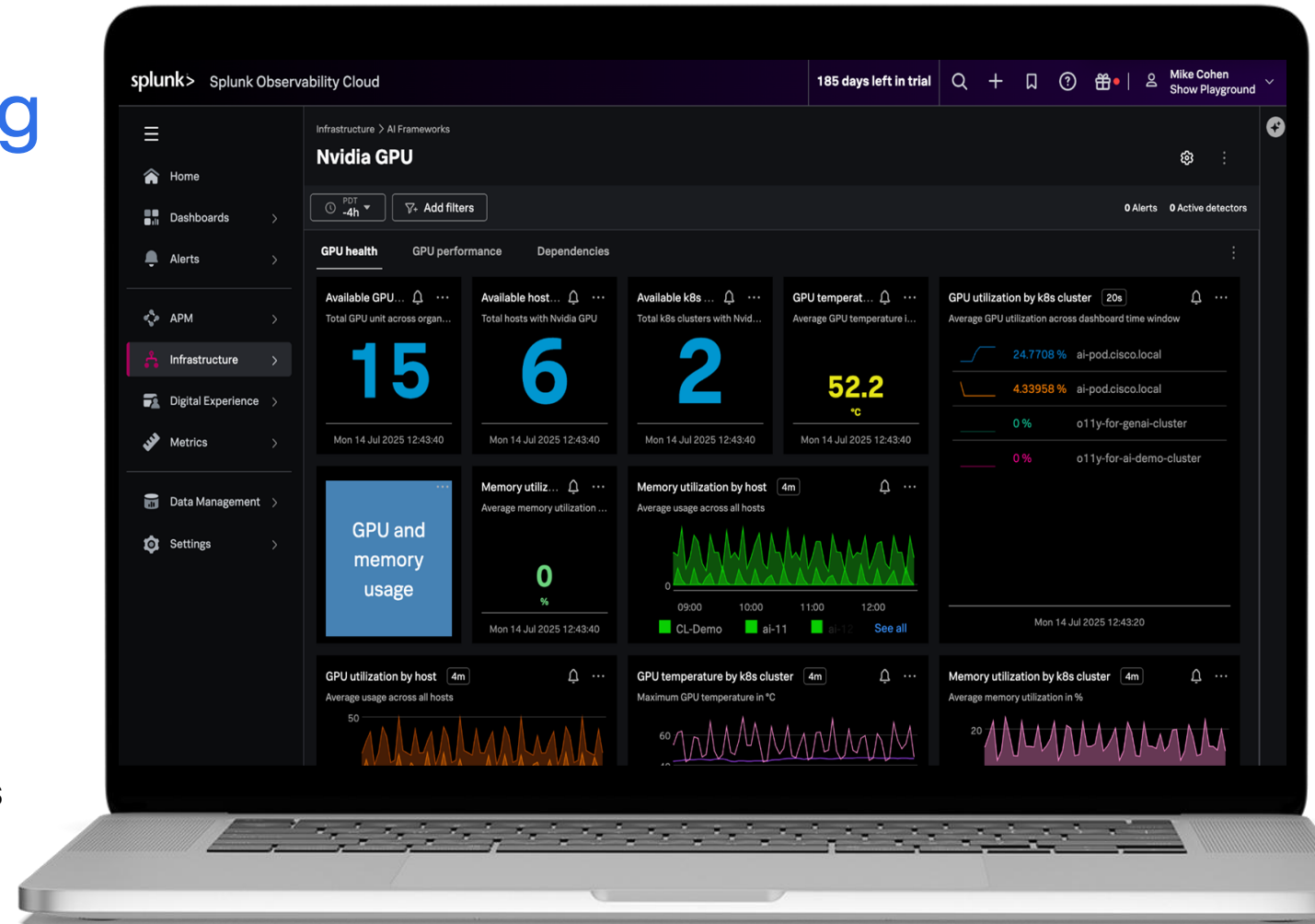
cisco Live !

Observability coverage for AI & agentic apps

...monitor & track the health of your AI stack

AI Infrastructure Monitoring

- **Monitor** health, availability, consumption of the entire AI infrastructure stack
 - GPU metrics now available at host, cluster and pod level
 - Proactive alerting
 - Reactive Troubleshooting
- Component-specific **dashboards and detectors**
 - Frameworks / Agents
 - Model Providers
 - Model Proxies
 - Vector Databases
 - Infrastructure (GPUs)
- **Correlated Troubleshooting** across components and services



Training

Optimization

Inferencing

Cisco AI PODs

Extend to



Cisco Security

AI Defense

Hypershift

Firewall/Nexus® SmartSwitch



Observability



Observability Cloud

Open telemetry extensions



Workload management and OPS



Operations



Intersight® and
Nexus Dashboard

Automation



AI software



NIM operator
NeMo
CUDA

OS/K8S



Ubuntu



Red Hat
OpenShift

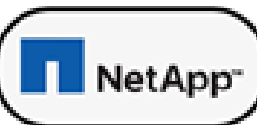
Accelerated compute

Cisco UCS®

High-performance
networking

Cisco Nexus

Extend to
storage platform
ecosystem



Advanced
services
included

Cisco
Customer Experience



BRK0BS-2547



Home



Dashboards



Alerts



APM



Infrastructure



Digital Experience



Application Security

Alpha



Logs



Metrics



Data Management



Settings

Overrides: Filter

Optional

Time

-31d

Chart resolution



Event overlay

Save

Reset

AI Pod Resources



Resources

Nodes

10s



total nodes, with or without GPUs

6

Sun 09 Nov 2025 15:51:20

Current GPU Nodes...



2

Sun 09 Nov 2025 15:51:20

GPU Utilization

10s



AI Pod High GPU Utilization Detector

node3.compact.tenan



worker3.flashstack.lo



GPU Memory Used

1s



GPU High Memory Usage Detector

n/a



GPU Utilization

10s



99

worker3.flashstack.local | 1



0

worker3.flashstack.local | 0



-

worker2.flashstack.local | 1



-

worker2.flashstack.local | 0

Sun 09 Nov 2025 15:51:10

GPU Memory Used

10s



94.3099

worker3.flashstack.local | 1



42.5419

worker3.flashstack.local | 0

Sun 09 Nov 2025 15:51:10

Home

Dashboards

Alerts

APM

Infrastructure

Digital Experience

Application SecurityAlpha

Logs

Metrics

Data Management

Settings

UTC+11
-1h

Add filters

Recently Used

Public Clouds

Containers

AI/ML

ECS (OTel) PROTOTYPE

Hosts

Web Server

Others

Search...

Dashboards vs. Navigators

Network Explorer (2 entities)

AI Frameworks (9 entities)
(Data resolution is 5 minutes.)

AWS Bedrock

2

Active Bedrock Instances

Azure OpenAI

No data found

This can occur if an integration has been recently disabled.

AddRemove navigator

GCP Vertex AI

No data found

This can occur if an integration has been recently disabled.

AddRemove navigator

LangChain

0 Alerts

4

Running services

Nvidia GPU

1 Critical alerts

5

Running GPUs

Nvidia NIM

1

Active models

OpenAI

0 Alerts

3

Active services

Ray Cluster

4 Critical alerts

1

Active services

Weaviate

2

Running instances

ECS Prototype

Hosts overview

My Data Center

Infrastructure > AI Frameworks

Nvidia GPU

UTC+11
-1h

Add filters

1 Alerts1 Active detectors

Overview

Table

Heatmap

GPU health

GPU performance

Dependencies

Available GPUs ...
Total GPU unit across organization

11

Sun 09 Nov 2025 15:35:00

Available hosts ...
Total hosts with Nvidia GPU

8

Sun 09 Nov 2025 15:35:00

Available k8s clust...
Total k8s clusters with Nvidia GPU

5

Sun 09 Nov 2025 15:35:00

GPU temperature ...
Average GPU temperature in °C

23.7
°C

Sun 09 Nov 2025 15:35:00

GPU utilization by k8s cluster 10s
Average GPU utilization across dashboard time window

46.2875 %ai-pod.cisco.local

17.9000 %o11y-inframom-ai

4.92931 %ai-pod.cisco.local

0 %o11y-for-genai-cluster

0 %o11y-for-ai-demo-cluster

0 %tenant2.avatar.local

Sun 09 Nov 2025 15:36:20

GPU and memory usage

Memory utilization ...
Average memory utilization in %

0 %

Sun 09 Nov 2025 15:35:00

Memory utilization by host 1m
Average usage across all hosts

CL-Demo

ip-10-1-123-51.us-west-2.con

See all

GPU utilization by host 1m
Average usage across all hosts

CL-Demo

ip-10-1-123-51.us-west-2.con

See all

GPU temperature by k8s cluster 1m
Maximum GPU temperature in °C

ai-pod.cisco.local

o11y-for-ai-demo-clu

See all

Memory utilization by k8s cluster 1m
Average memory utilization in %

ai-pod.cisco.local

o11y-for-ai-demo-clu

See all

AI Agents & Apps

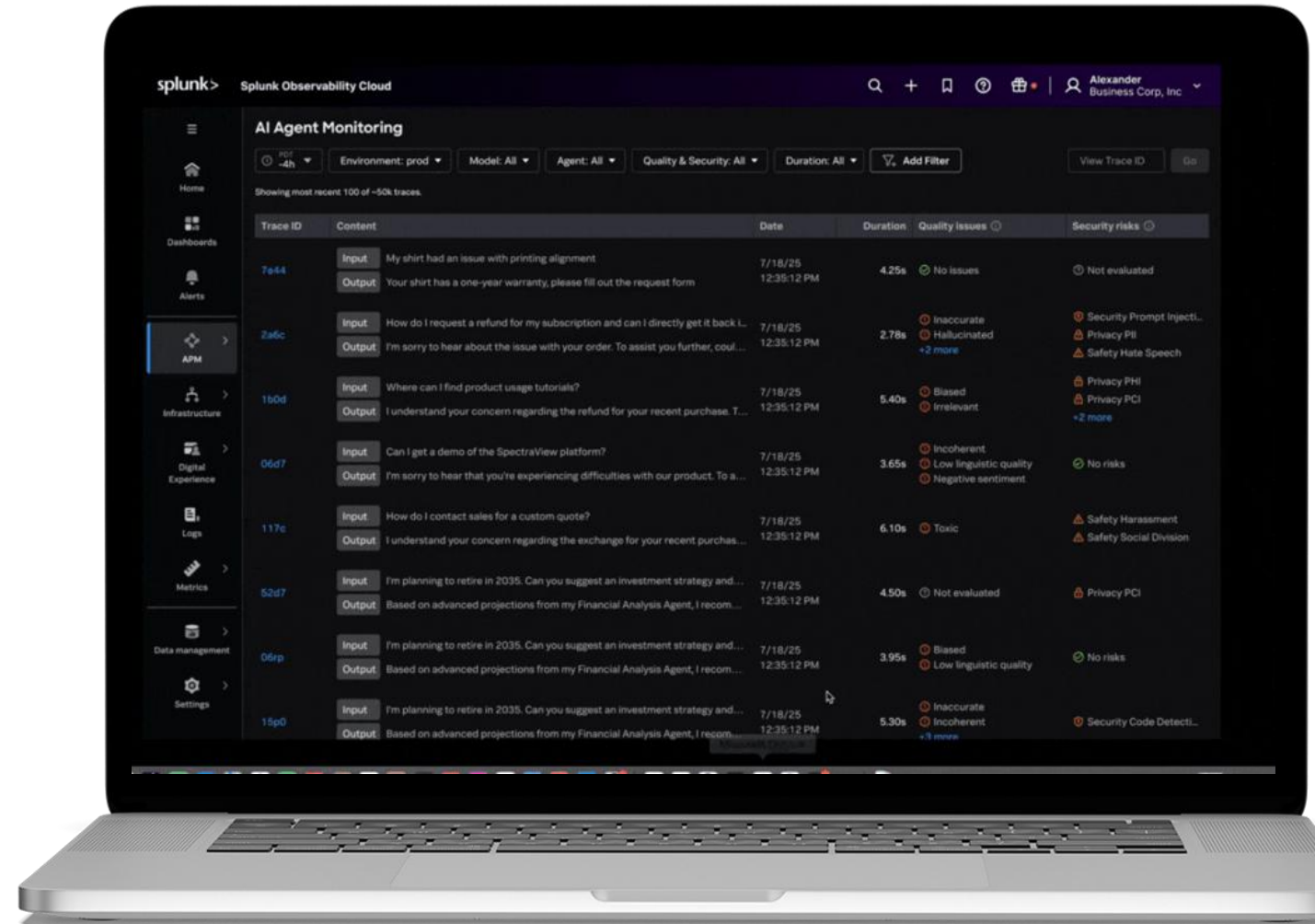
cisco Live !

Observability for AI & agentic apps

...monitor & track the health of your AI stack

AI Agent Performance

- **Trace all interactions** between users, agents and tools in a dedicated view
- **Visualize multi-agent workflows**, to understand inputs, outputs and agent decisions
- **Understand the health, quality and cost** of each agent running
- **Evaluate LLM responses** to assess the effectiveness of applications

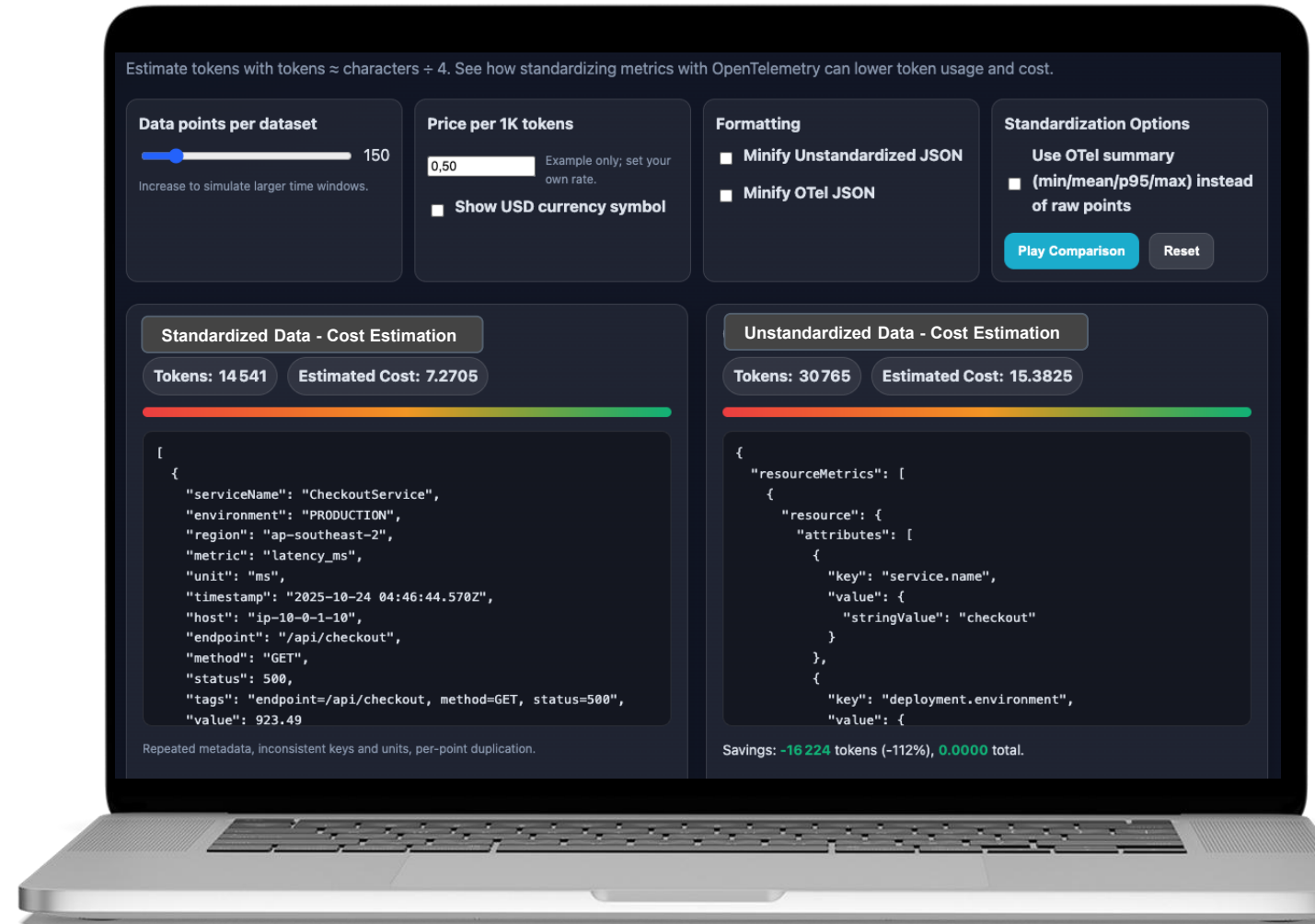


Observability for AI & agentic apps

...monitor & track the health of your AI stack

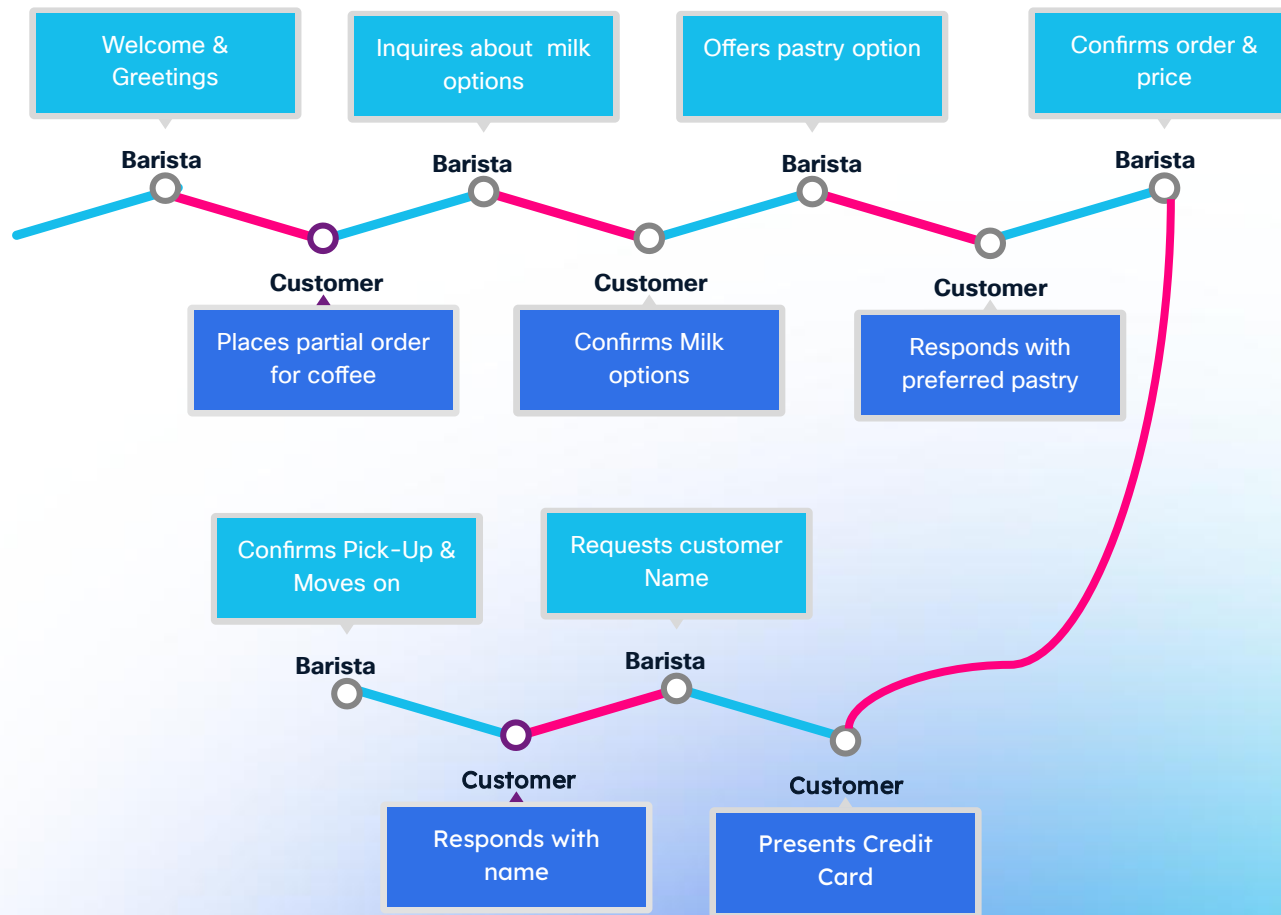
Token Usage

- **Cost Management:** Prevents unexpected expenses by optimizing prompt length
- **Performance Optimization:** Improves speed & efficiency shortening overly verbose inputs and outputs.
- **Resource Allocation:** Provides insights into usage patterns to effectively manage resources and scale LLM applications.
- **Debugging and Traceability:** Offers critical observability for diagnosing issues and understanding LLM behavior in live environments



Monitor & observe AI - token usage

...costly prompts - Scenario 1



Customer approaches looking at the menu.

Barista: "Hi there! What can I get for you today?"

Customer: "Hmm, good morning! I'm ready. Can I get a large latte, please?"

Barista: "Sure thing! And would you like that with regular milk, or a dairy alternative like oat, almond, or soy?"

Customer: "Regular milk, thanks. Could I get an extra shot of espresso in that?"

Barista: "Absolutely, a large latte with an extra shot. Anything else for you today, maybe a pastry or a sandwich?"

Customer: "Oh, actually, yes! Do you have any plain croissants left?"

Barista: "Let me check... Yes, we do! One plain croissant. So, that's a large latte with an extra shot and a plain croissant. Your total comes to \$9.50."

Customer: "Perfect, here's my card."

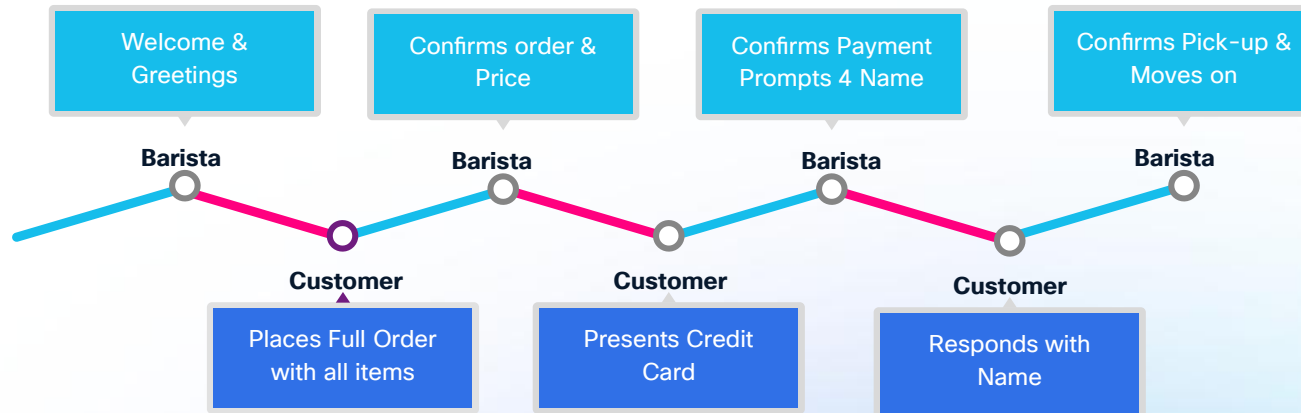
Barista: "Great, payment successful! Your coffee will be ready at the pick-up counter in just a couple of minutes. What name can I put on that for you?"

Customer: "It's Sarah."

Barista: "Got it, Sarah. We'll call you when it's ready!"

Monitor & observe AI - token usage

...costly prompts - Scenario 2



Customer approaches having already decided.

Barista: "Hi there! What can I get for you today?"

Customer: "Good morning! I'd like a large latte with an extra shot of espresso and regular milk, please. And one plain croissant"

Barista: "Alright, so to confirm: a large latte with an extra shot and regular milk, and one plain croissant. Your total comes to \$9.50."

Customer: "Perfect, here's my card."

Barista: "Great, payment successful! Your coffee will be ready at the pick-up counter in just a couple of minutes. What name can I put on that for you?"

Barista: "Your coffee will be ready at the pick-up counter in just a couple of minutes. What name can I put on that for you?"

Customer: "It's Sarah."

-
- Home
- Dashboards
- Alerts
- APM
- Infrastructure
- Digital Experience
- Application Security

Alpha
- Logs
- Metrics
- Data Management
- Settings

UTC+11
-8d

Environment : ad-ecommerce (+60)

Requests

Total count

13.9k

Errors

Total count

24

Rate

0

Tokens

Total count

2.6m

Input count

1.4m

Output count

1.2m

Quality scores

6.9k evaluations

Hallucination

38.0%

Hallucinated

Relevance

92.7%

Relevant

Bias

96.5%

Not biased

Toxicity

98.6%

Non toxic

Sentiment

100.0%

Positive sentiment

Q Search by agent name...

Configure log observer index

Agent ↕	Requests ↕	Errors ↕	Latency (P90) ↕	Tokens ↓	Quality issues ⓘ	Related traces	Related logs
plan_synthesizer	746	0 (0%)	23.33s	575.9k In: 187.2k Out: 388.6k	Hallucinated: 75%		
coordinator	1.2k	24 (1.94%)	16s	401.3k In: 66.7k Out: 334.6k	No issues		
weather_agent	1.2k	0 (0%)	58.79ms	356.4k In: 321.1k Out: 35.3k	Not evaluated		
activity_specialist	2.4k	0 (0%)	7.69s	301.1k In: 171.7k Out: 129.3k	Hallucinated: 53%		
flight_specialist	2.1k	0 (0%)	4.58s	271.7k In: 173.8k Out: 97.8k	Hallucinated: 22%		
hotel_specialist	2.1k	0 (0%)	4.52s	238.1k In: 162.4k Out: 75.7k	Hallucinated: 21%		
formatter-agent-1	135	0 (0%)	14.79s	85.1k In: 57.8k Out: 27.3k	Hallucinated: 100% Biased: 94% + 1 more		

- ☰

Home

Dashboards

Alerts

APM

Infrastructure

Digital Experience

Application SecurityAlpha

Logs

Metrics

Data Management

Settings

Infrastructure > AI Frameworks

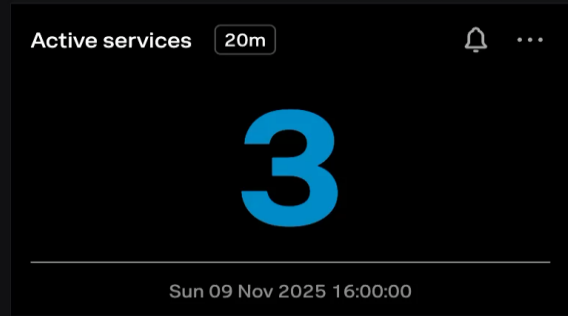
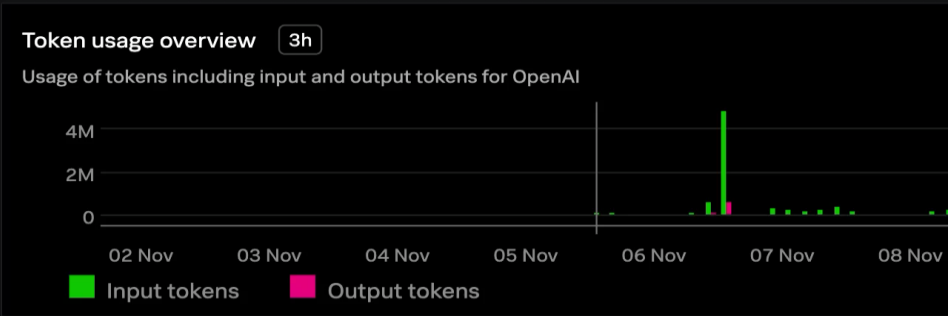
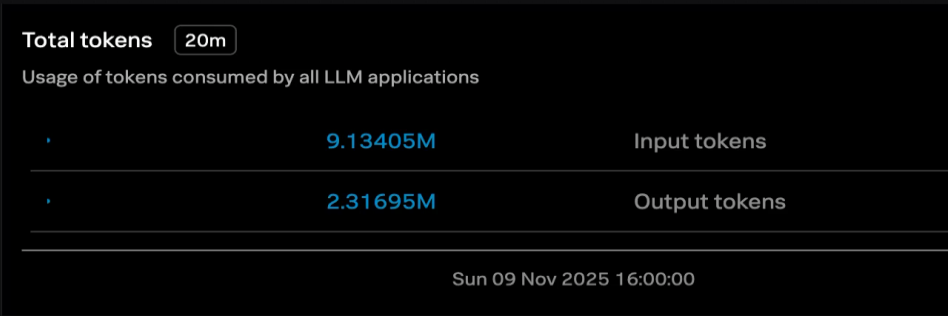
OpenAI

UTC+11
-8d

🔍 Add filters

🔧 Overview📄 Table🔥 Heatmap

OpenAIsDependencies



Total tokens by service20m🔔⋮

Token usage by service

↓ service.name	Input tokens	Output tokens
trip-planner-tl	29,795	31,260
travel-planner-traceloop-raw-admin...	210,763	222,388
travel-planner-traceloop-cyclic-ad...	224,282	222,005

Total tokens by model20m

Token usage by model

gen_ai.request.model	↓ Input tokens	Output tokens
mistralai/Mistral-7B-Instruct-v0.3	642,527	648,807

Conversation Insights & Agent Quality

CISCO Live !

← Agents

plan_synthesizer Alpha

UTC+11 -8d

Environment : ad-ecommerce (+60)

Agent : plan_synthesizer

Configure log observer index

View related traces

View related logs

Performance

Requests

Count

Errors

Count

Latency

p99
p90
p50

Token usage

Token usage

Input
Output
Total

Quality

Quality scores

620 evaluations

Hallucination

75.0%

Hallucinated

Relevance

98.7%

Relevant

Toxicity

100.0%

Non toxic

Bias

100.0%

Not biased

Sentiment

100.0%

Positive sentiment

Quality issues

https://shw-playground.signalfx.com/#/apm/traces?endTime=Now&filters=%7B"traceFilter":%7B"tags":%5B%7B"tag":%7B"sf_environment","operation":%7B"IN","values":%5B"ad-ecommerce","ai-pod","ai-test-env","ai-test-exp-svc","ai-test-val","azure-test","CC-Dev","christianb-lab","datacenter-b01",...

← APM

Trace view

gen_ai.agent.name = plan_synthesizer

gen_ai.o...on.name = *

Add filters

Search...

115 / 115 Spans Loaded

0ms33.

travel-planner-tl: travel_planner_multi_agent.workflow	30.1s
travel-planner-tl: LangGraph.workflow	30.1s
travel-planner-tl: __start__.task	6.9ms
travel-planner-tl: should_continue.task	361.0µs
travel-planner-tl: coordinator.task	10.4s
travel-planner-tl: coordinator_agent.task	10.3s
travel-planner-tl: ChatOpenAI.chat	9.1s
travel-planner-tl: GET	137.6ms
travel-planner-tl: chat gpt-4o-mini	3.6ms
openai:gpt-4o-mini: ChatOpenAI.chat	3.6ms
travel-planner-tl: Ran evaluate()	29.2s
travel-planner-tl: Ran test case	29.2s
travel-planner-tl: Bias	4.1s
travel-planner-tl: Toxicity	4.2s
travel-planner-tl: Answer Relevancy	15.1s
travel-planner-tl: hallucination [GEval]	
travel-planner-tl: sentiment [geval] [GEval]	
travel-planner-tl: Ran evaluate()	
travel-planner-tl: Ran test case	
travel-planner-tl: hallucination [GEval]	
travel-planner-tl: sentiment [geval] [GEval]	
travel-planner-tl: Ran evaluate()	
travel-planner-tl: Ran test case	
travel-planner-tl: Bias	
travel-planner-tl: Toxicity	

Span properties

tural etiquette and must-see attractions. **Next Steps:**

- Confirm travel dates and preferences for the boutique hotel and experiences.

- Begin booking flights and accommodations once details are finalized.

- Prepare a detailed itinerary for the couple, including all bookings and recommendations. T his plan will ensure a memorable and romantic getaway to Paris for the travelers.

Show Less

gen_ai.output.messages ***

[[{"role": "assistant", "parts": [{"type": "text", "content": "***Travel Plan for Paris Trip**\n\n**Travel Dates:** Next month (specific dates to be confirmed)\n\n**Departure City:** Seattle\n\n**Destination:** Paris, France\n\n**Travel Class:** Business-class flights\n\n**Accommodation:** Boutique hotel\n\n**Key Details for Specialist Agents:**\n\n1. **Flights:** - Book business-class flights from Seattle to Paris.\n- Look for direct flights or the best connections to minimize travel time.\n- Ensure flexible cancellation policies in case of changes.\n\n2. **Accommodation:** - Research and recommend boutique hotels in Paris that offer a romantic atmosphere.\n- Focus on areas such as Le Marais, Saint-Germain-des-Prs, or Montmartre for a charming experience.\n- Confirm amenities such as in-room dining, spa services, and proximity to attractions.\n\n3. **Unique Experiences:** - Curate a list of romantic activities, such as:\n- A Seine River dinner cruise.\n- Private guided tours of iconic landmarks (e.g., Eiffel Tower, Louvre).\n- Wine tasting or cooking classes.\n- A visit to a local market followed by a picnic in a park.\n- Consider booking a couple's spa day or a photography session to capture memories.\n\n4. **Additional Considerations:** - Provide information on local transportation options (e.g., metro, taxis).\n- Suggest restaurants for romantic dinners and local cuisine.\n- Include tips on cultural etiquette and must-see attractions.\n\n**Next Steps:**\n- Confirm travel dates and preferences for the boutique hotel and experiences.\n- Begin booking flights and accommodations once details are finalized.\n- Prepare a detailed itinerary for the couple, including all bookings and recommendations. \n\nThis plan will ensure a memorable and romantic getaway to Paris for the travelers."}], "finish_reason": "stop"]}]

Show Less

gen_ai.prompt.0.role ***

system

gen_ai.prompt.1.content ***

What is the best travel plan for a couple visiting Paris in September? The plan should include a boutique

Related Logs

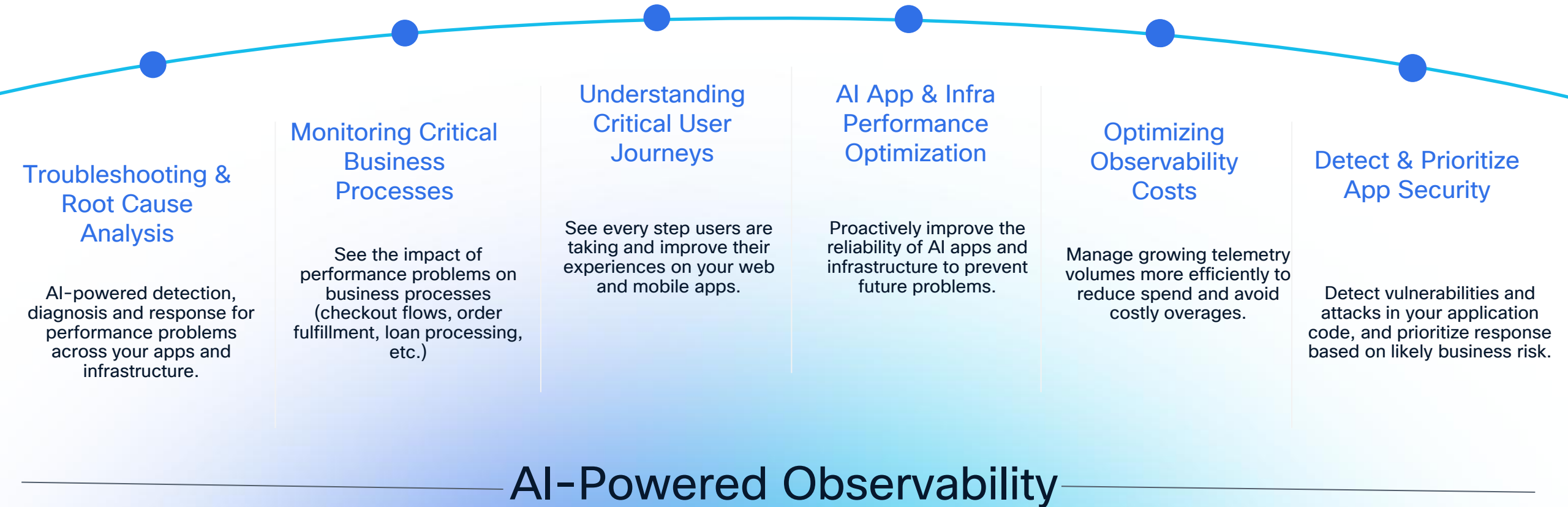
Logs about the entity in view

Logs for trace e396dc0d296a84eee181aec93625c25c in o11y-playground-us1.stg.splunkcloud.com [SWiPE] integration

Global Index Search is enabled. Results include logs from all indexes and may be broad. Review entity selection or refine entity mappings.

Building an observability practice in the AI era

...foundational critical observability capabilities across your stack



**“AI follows the path *we draw*.
We shape its mind, guide its
heart, and drive its purpose”**

Q/A

Complete your session evaluations



Complete a minimum of 4 session surveys and the Overall Event Survey to claim a Cisco Live T-Shirt.



Earn up to 800 points by completing all surveys and climb the Cisco Live Challenge leaderboard.



Level up and earn exclusive prizes!



Complete your surveys in the Cisco Live Events app.

Continue your education



Visit the Cisco Stand for related demos



Book your one-on-one Meet the Expert meeting



Attend the interactive education with Capture the Flag, and Walk-in Labs



Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Thank you

cisco Live !

