

Assurance checkup: Monitoring the health of AI agents

Abdiel Hernandez Fentanes

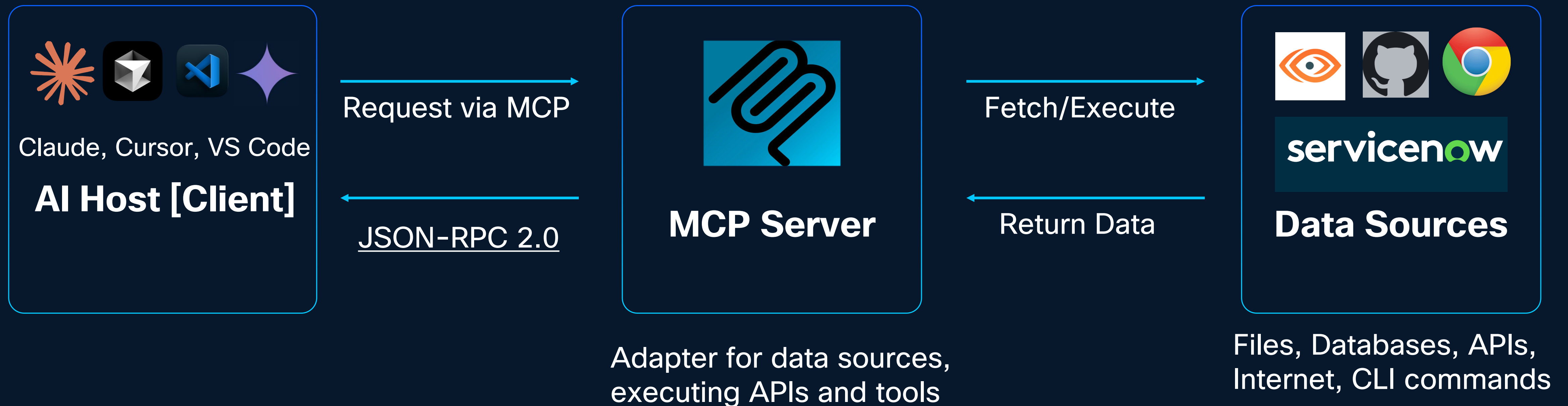


What are AI Agents?

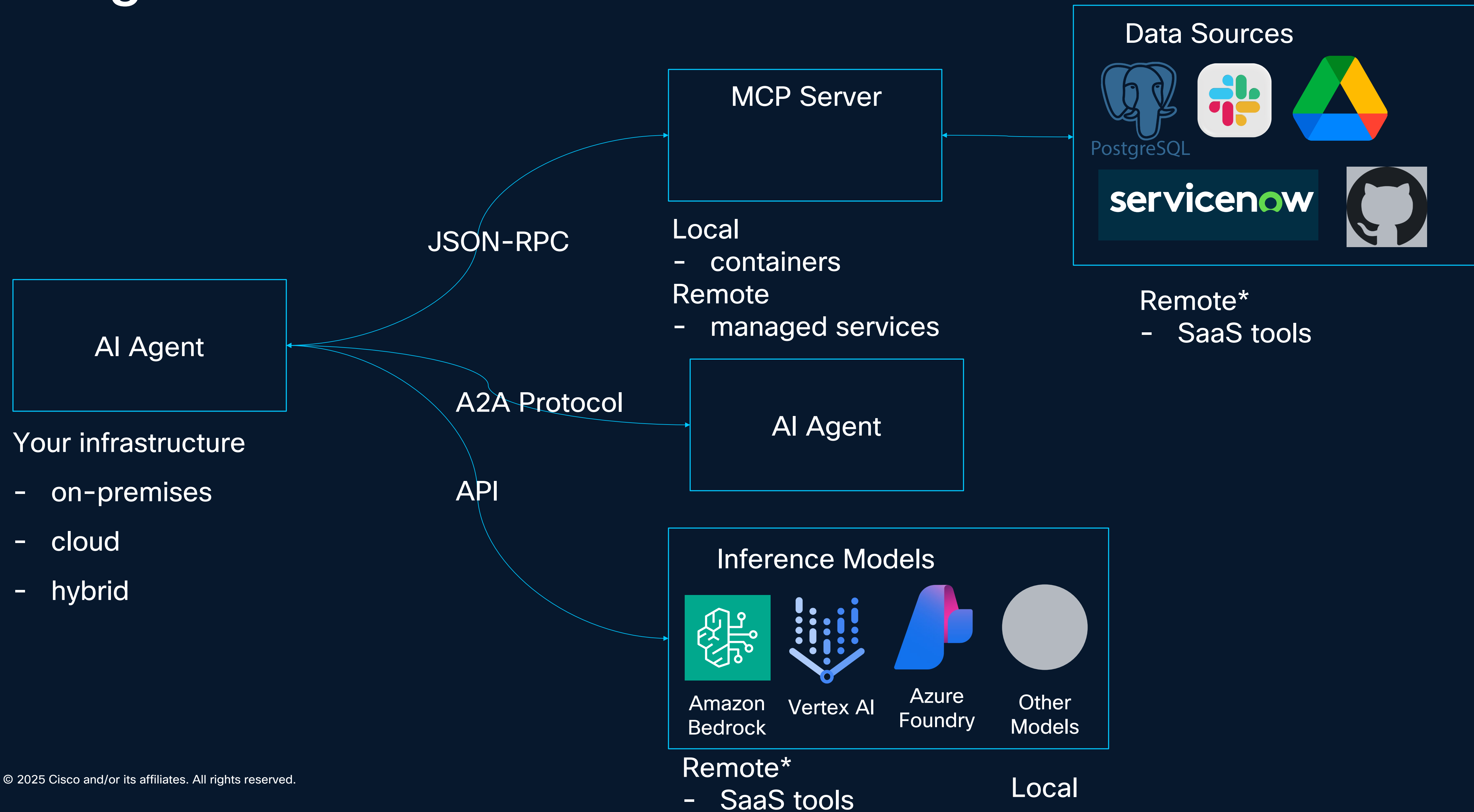
Autonomous and proactive software that has the purpose of performing tasks.

1. They depend on external inference providers.
 - a. Example: OpenAI, Anthropic, AWS Bedrock.
2. They use the Model Context Protocol to access multiple data sources dynamically.
 1. Examples: Querying a database, calling APIs, access document repositories.
3. Performance issues are subtle.
 1. 500ms delay in an MCP server response degrades experience.
 2. Examples: DNS, Model delay, network problem, SaaS, 3rd-party, Data Center delays

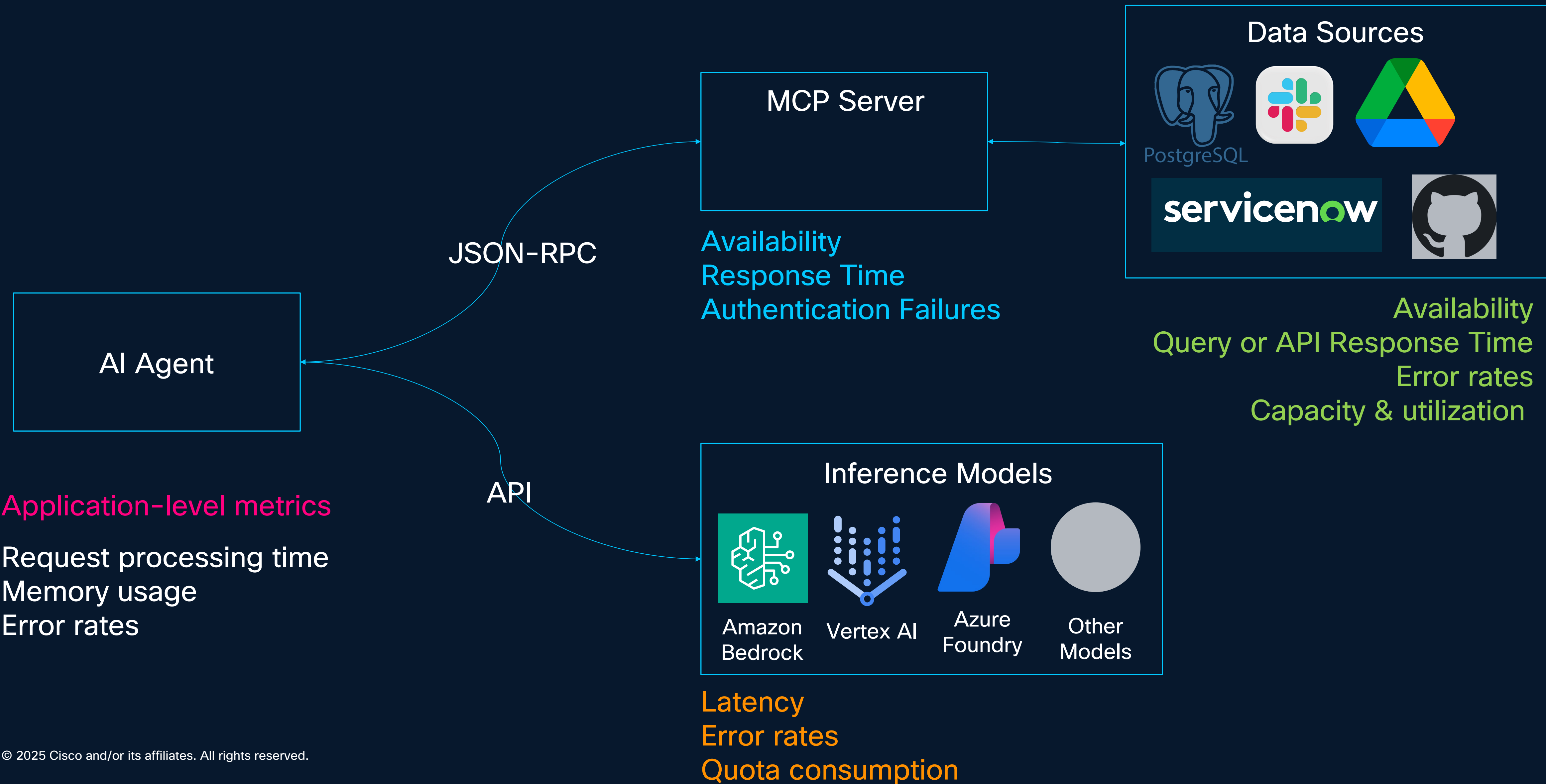
MCP Server Architecture



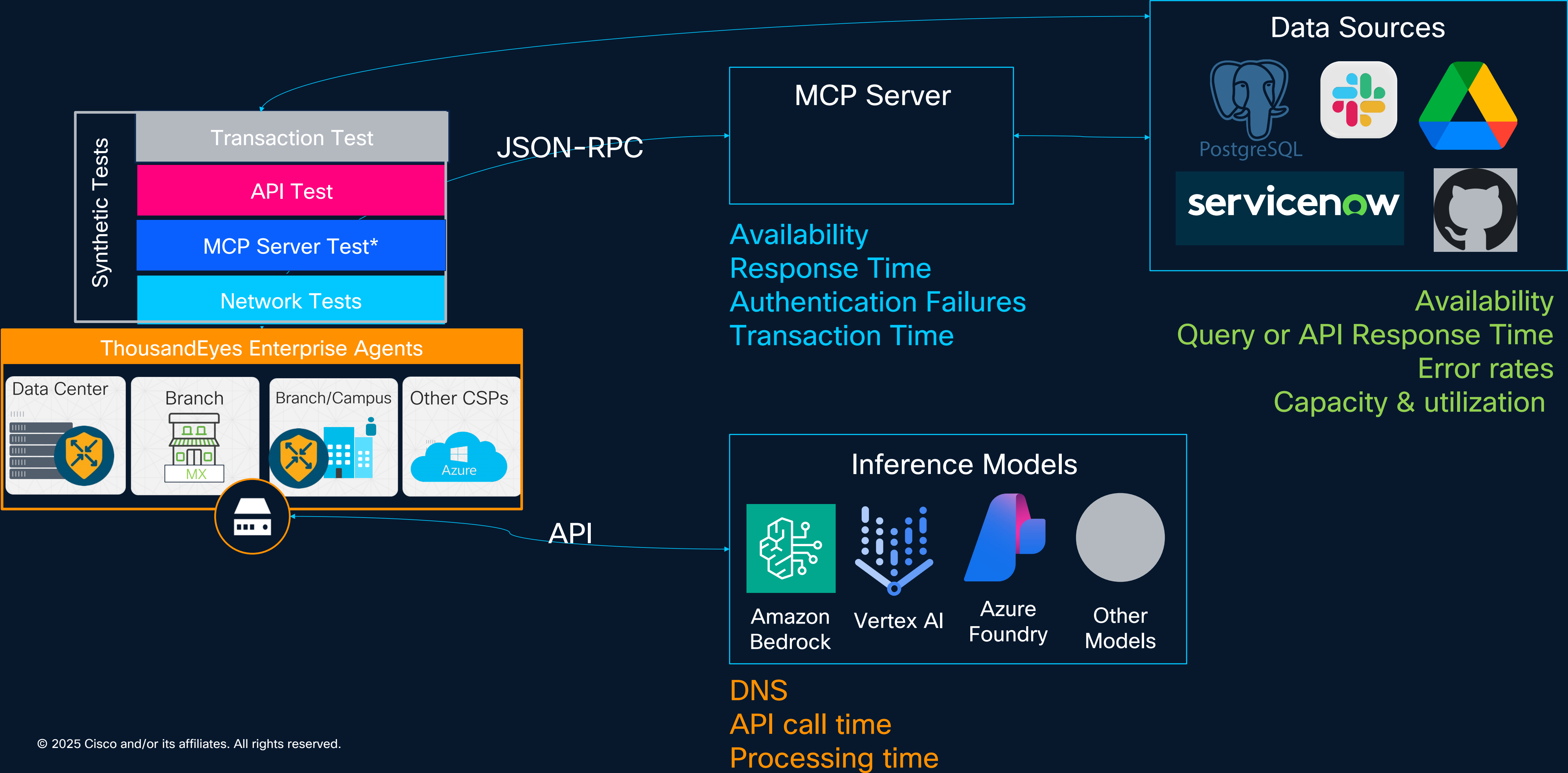
AI Agent Architecture



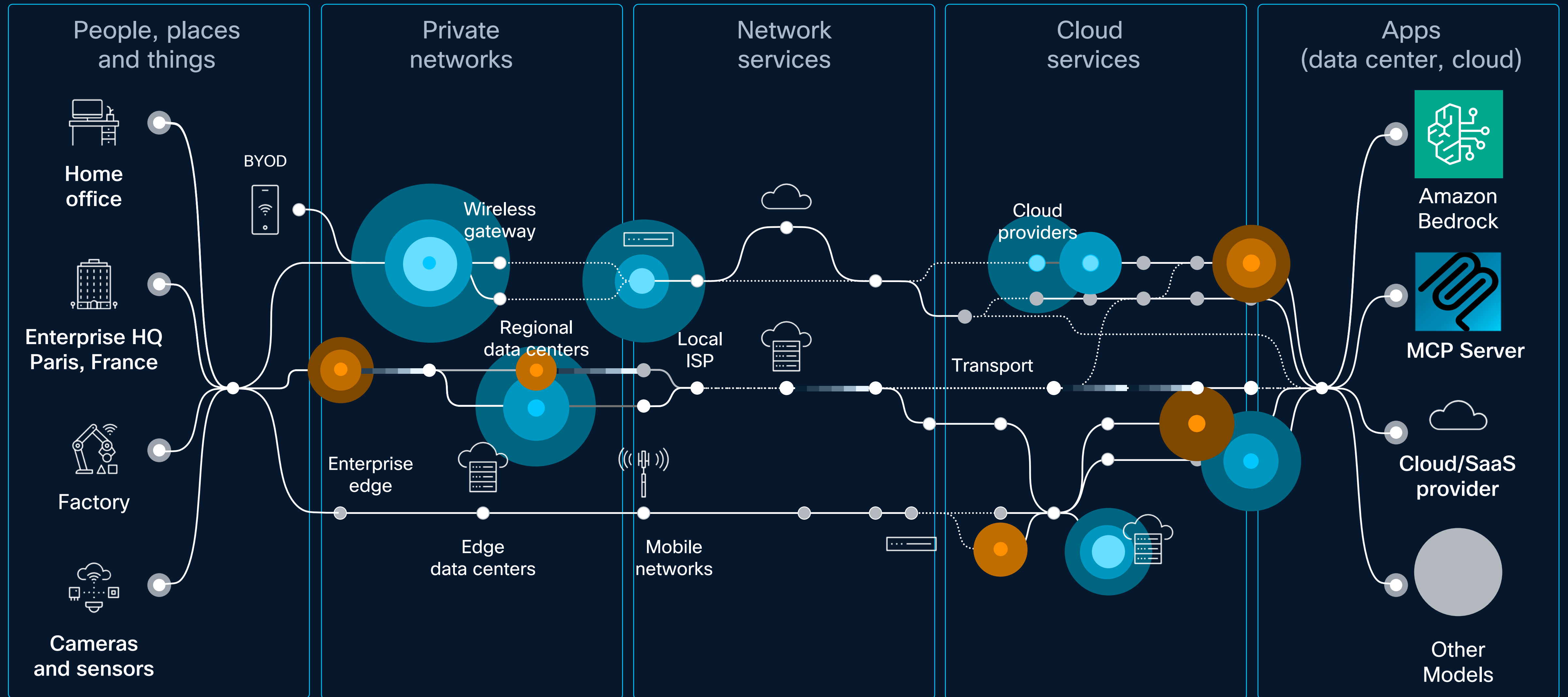
Essential Metrics: What to Monitor and Why It Matters



The Cisco Solution



Better **visibility** across the entire digital supply chain



Assurance for AWS Bedrock

Assurance for AWS Bedrock

- API Server:
 - <https://bedrock-runtime.{REGION}.amazonaws.com/model/{MODEL}/...>
- Supported models:
 - <https://docs.aws.amazon.com/bedrock/latest/userguide/models-supported.html>
 - Note: Some models can be invoked from a region, but inference may take place in other regions, this is “Cross-region inference”, and it affects the way you call the model
- Pricing
 - <https://aws.amazon.com/bedrock/pricing/>
- API Reference:
 - [Bedrock](#)
 - [Bedrock Runtime](#)

AWS Bedrock Test Template for Cloud & Enterprise Agents

Deploy Template - AWS Bedrock [Documentation](#)

1 Select Template 2 **Configure Template** 3 Review Template

Global Settings

- ☒ AWS Bedrock - API Model Inference
- ☒ AWS Bedrock - DNS Trace
- ☒ AWS Bedrock - API Server
- ☒ AWS Bedrock - DNS Nameservers

AWS Bedrock

Global Settings

A template for monitoring the AWS Bedrock API

Enter settings that you want to apply to all tests. You can also configure individual tests from the list on the left.

Which agents should we test from?

0 of 1066 Agents

How often should we run tests that do not consume tokens?

2 minutes

How often should we run tests that do consume tokens?

10 minutes

Enter the name of your AWS region

us-east-1

Enter the Credential name for your AWS Bedrock API key ?

Enter the modelId of the AWS Bedrock model to test

amazon.nova-pro-v1:0

Enter your user prompt

How many of the letter R are in the word Strawberry?

Name your template ?

AWS Bedrock

Deployment Summary

Tests

- API
- DNS Trace
- HTTP Server
- DNS Server

Supporting Resources

Labels (1)

Pre-requisites before deploying the template:

1. Create a Credential in the Credential Repository with your Bedrock API key
(See Appendix for details on where to find/create Bedrock keys)

Deploy the AWS Bedrock template:

1. Navigate to **Network & App Synthetics > Test Settings > Start Monitoring**
2. Under **Start with templates**, select the AWS Bedrock template
3. On the **Configure Template** form:
 1. Select the Agents from which you wish to test Bedrock services
 2. Select the test intervals for the tests which do and do not consume tokens
 3. Provide the name of the region for your Bedrock model
 4. Provide the name of your previously-created Credential
 5. Provide the name of the model you wish to test for model inference
 6. Provide the prompt to use for model inference
 7. Name your template deployment
4. Click **Review Template**, then click the **Deploy Now** button
5. Click **Go to Test Settings**

Post-requisites after deploying the template:

1. On the **Network & App Synthetics > Test Settings** page, select the API Model Inference Test
2. Click the **Next** button, then click the **Update** button

Assurance for AWS Bedrock

4 tests | | | Type: All | Test Labels: AWS Bedrock | Status: Enabled +3 | Agents: All | [Reset Filters](#)

<input type="checkbox"/>	Test Name ↑	Labels	Type ↑	Target ↑	Alerts ↑	Status ↑	
<input type="checkbox"/>	AWS Bedrock - API ...	AWS Bedrock	API	https://bedrock-runtime.us-east-1.amazonaws.com	✓	Enabled	...
<input type="checkbox"/>	AWS Bedrock - API ...	AWS Bedrock	HTTP Server	https://bedrock-runtime.us-east-1.amazonaws.com	✓	Enabled	...
<input type="checkbox"/>	AWS Bedrock - DNS...	AWS Bedrock	DNS Server	bedrock-runtime.us-east-1.amazonaws.com A · UDP	✓	Enabled	...
<input type="checkbox"/>	AWS Bedrock - DNS...	AWS Bedrock	DNS Trace	bedrock-runtime.us-east-1.amazonaws.com A · UDP	✓	Enabled	...

High frequency tests that do not consume tokens and provide baseline visibility

Invoking model inference to ensure availability and measure performance

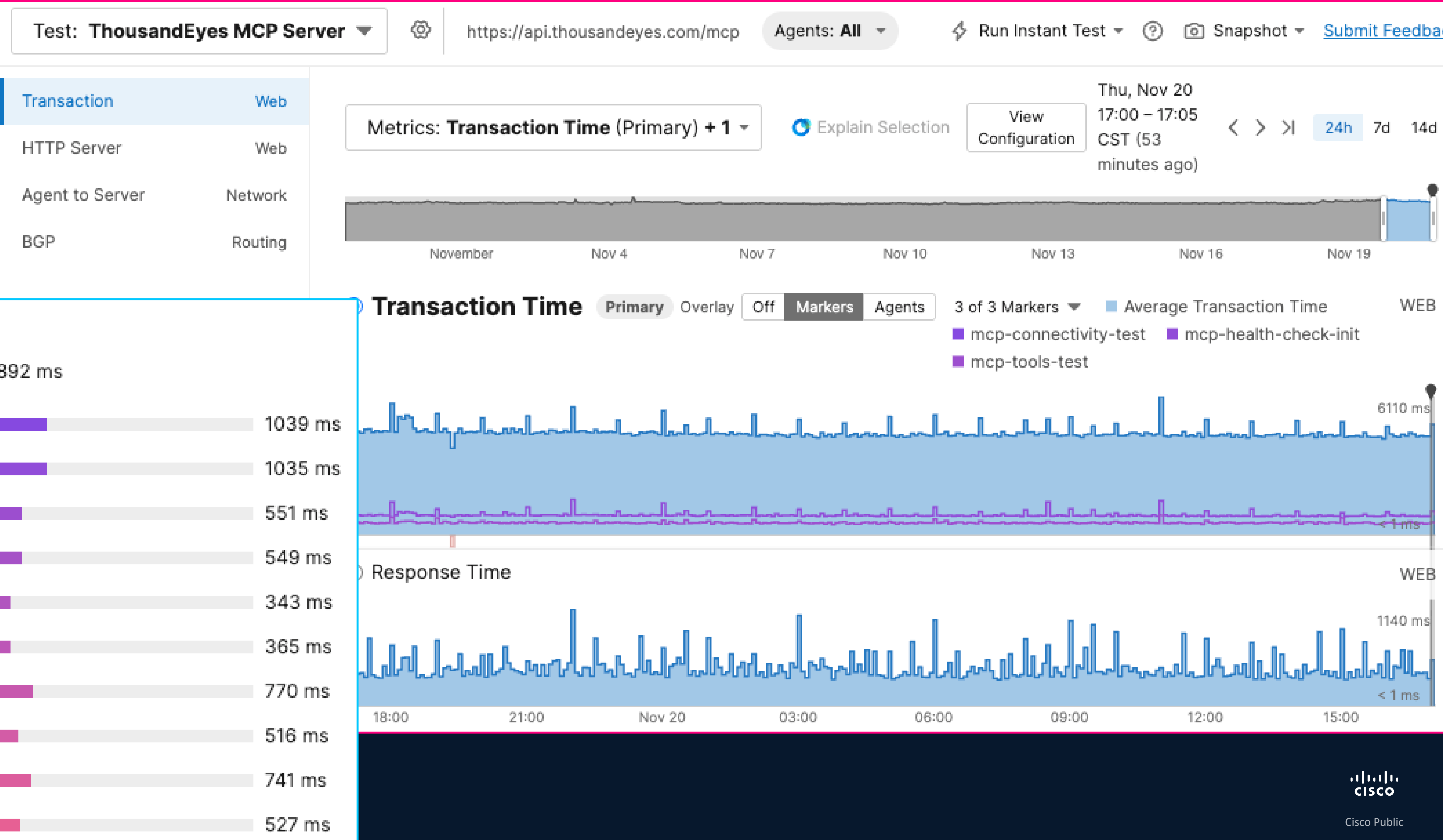
Assurance for AWS Bedrock

API Model Inference

API Steps

1: Converse		Response Code	Assertions	API Call Time	Processing Time
POST	https://bedrock-runtime.us-east-1.amazonaws.com/model/amazon.nova-pro-v1:0/converse	200	2 of 2 passed	1428 ms <div><div></div></div>	41 ms

Assurance for MCP Server



Thank you

