1000 Clusters, 1 Smarter Brain

The Next Evolution of AIOps at Salesforce

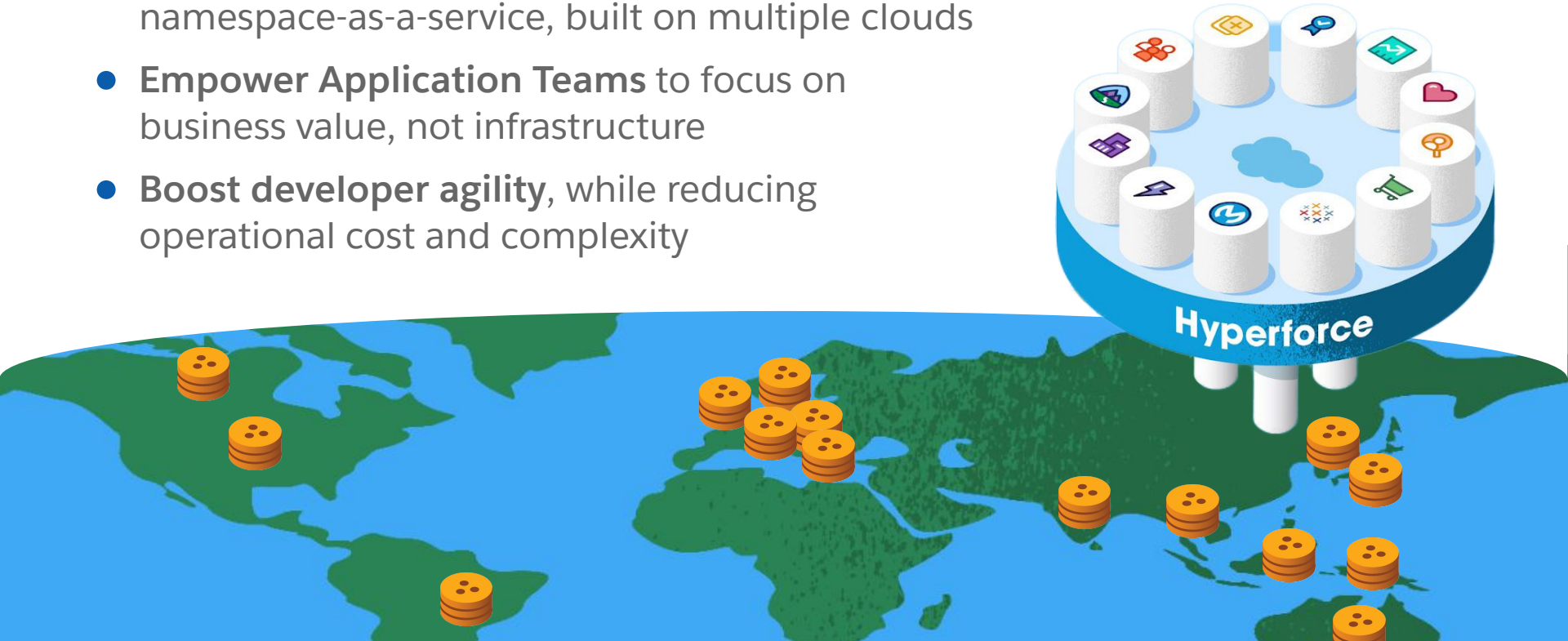
Anubhav Dhoot
SVP, Engineering

Srikanth Rajan
VP, Engineering

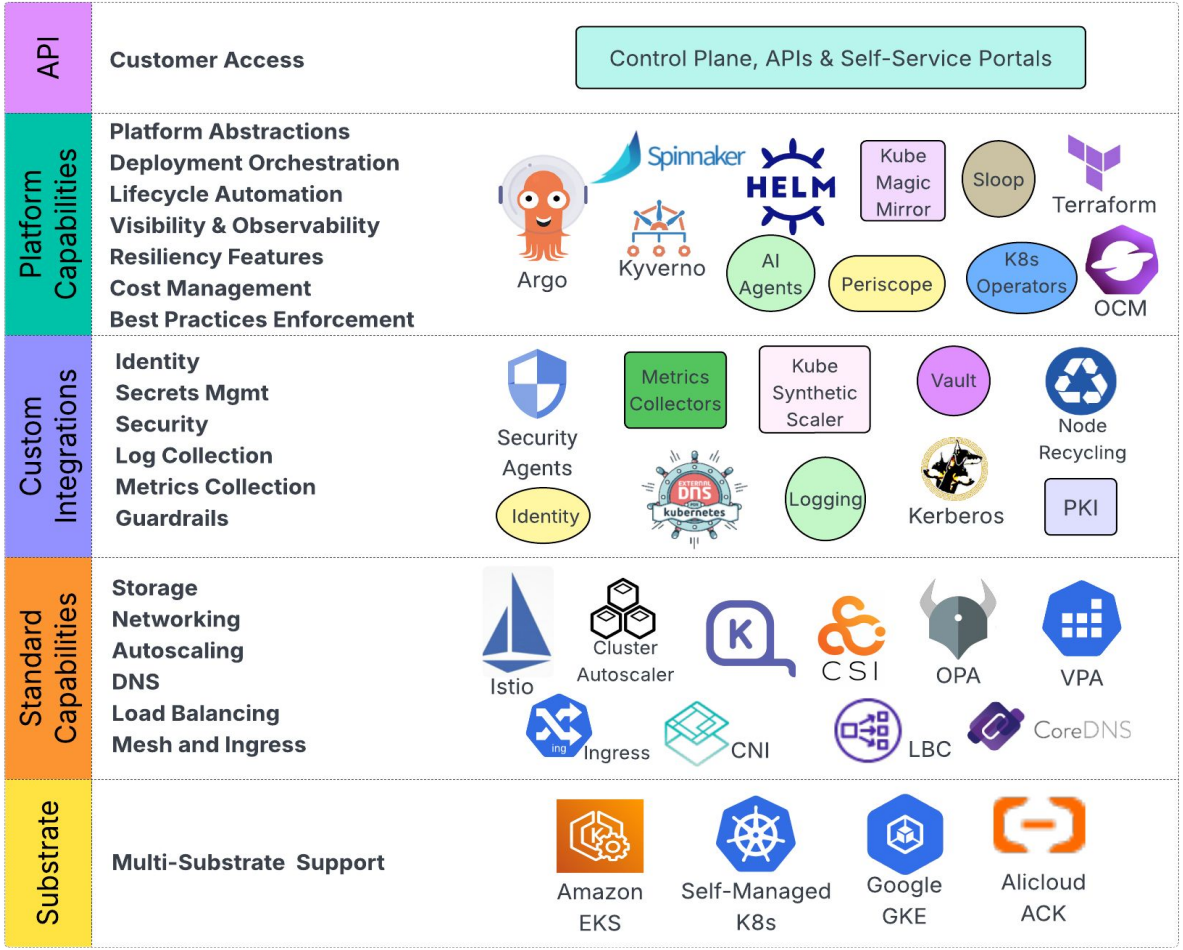


Intro: Hyperforce Kubernetes Platform

- **Managed Kubernetes Platform** providing namespace-as-a-service, built on multiple clouds
- **Empower Application Teams** to focus on business value, not infrastructure
- **Boost developer agility**, while reducing operational cost and complexity



Tech Stack


API
Customer Access
Control Plane, APIs & Self-Service Portals
Platform Capabilities
Platform Abstractions
Deployment Orchestration
Lifecycle Automation
Visibility & Observability
Resiliency Features
Cost Management
Best Practices Enforcement
Argo
Spinnaker
Kyverno
HELM
AI Agents
Kube Magic Mirror
Periscope
Sloop
K8s Operators
Terraform
OCM
Custom Integrations
Identity
Secrets Mgmt
Security
Log Collection
Metrics Collection
Guardrails
Security Agents
Identity
Metrics Collectors
Kube Synthetic Scaler
Logging
Vault
Kerberos
Node Recycling
PKI
Standard Capabilities
Storage
Networking
Autoscaling
DNS
Load Balancing
Mesh and Ingress
Istio
Cluster Autoscaler
CSI
OPA
VPA
Ingress
CNI
LBC
CoreDNS
Substrate
Multi-Substrate Support
Amazon EKS
Self-Managed K8s
Google GKE
Alicloud ACK



Operational Scale



50+
Operators and Integrations

Millions
Pods

Thousands
Compute Nodes

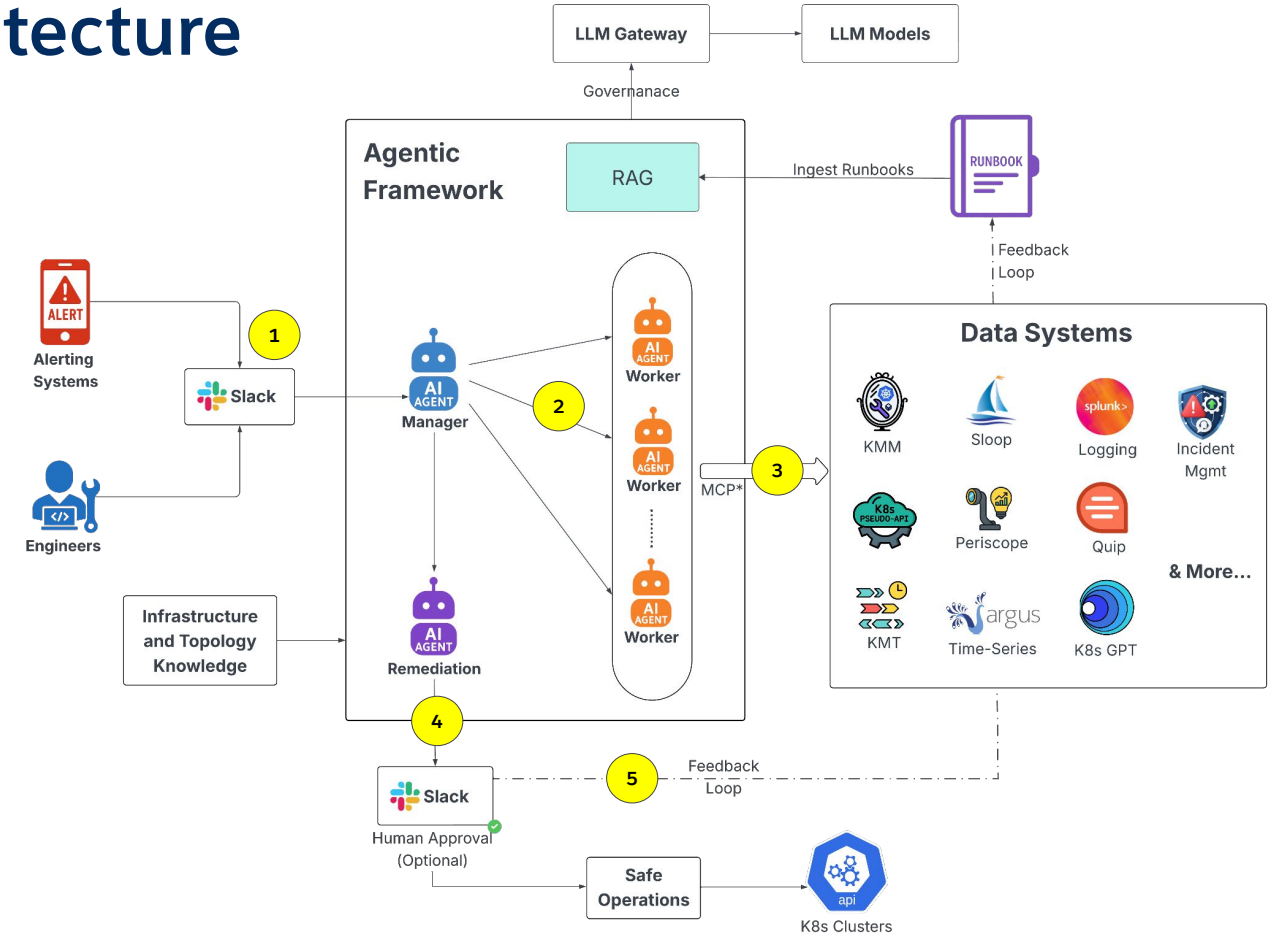
Multi-Substrate
AWS, GCP, Ali, Private DC

2,000+
K8s Clusters

250+
Monitoring Plugins

1000+
Support Hours

Agentic Self-Healing Architecture



Agents at Scale

1 Hypothesis

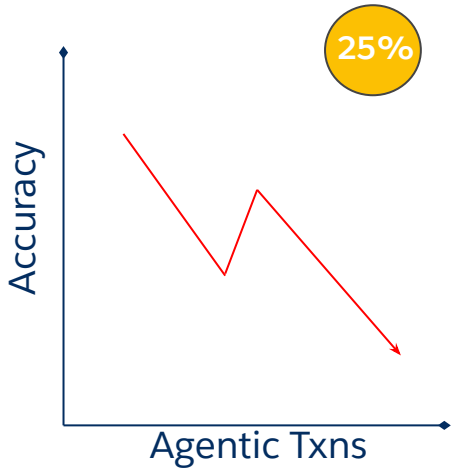
Lots of Data

+

High Volume Agentic Txns

= ?

2 Reality



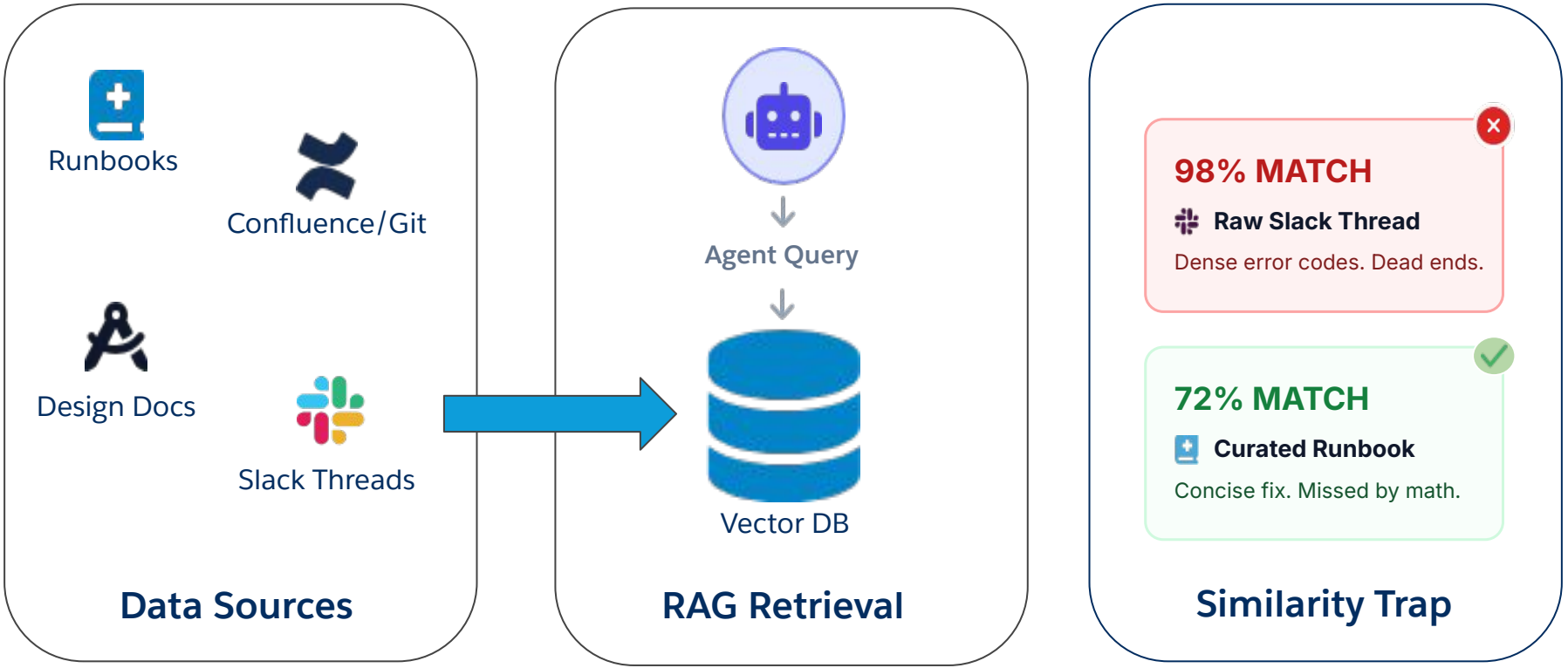
3 Bottlenecks

Knowledge Management

Evaluating Agentic Output

Agent Design

Knowledge Mgmt: 101

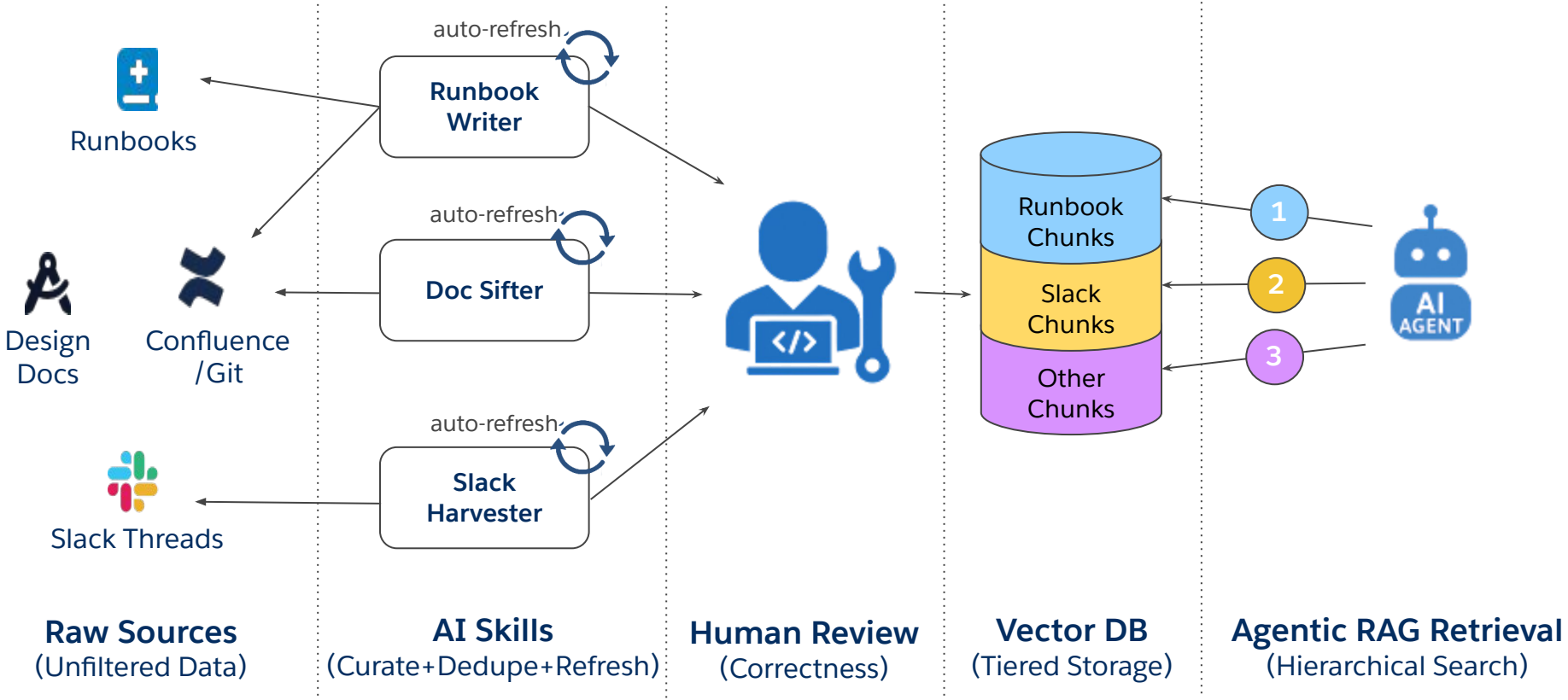


Slack Snippets: (Knowledge Issues)

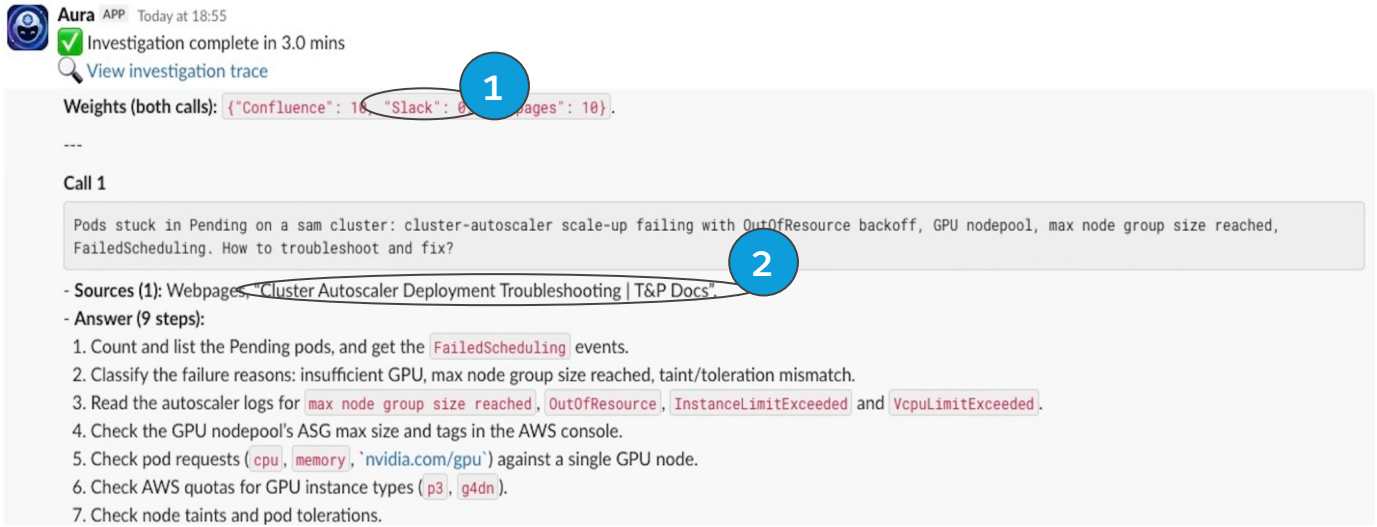


1. Slack thread has much lower weights compared to other sources
2. But it still wins the similarity race when querying RAG

Continuously Curated Knowledge Base

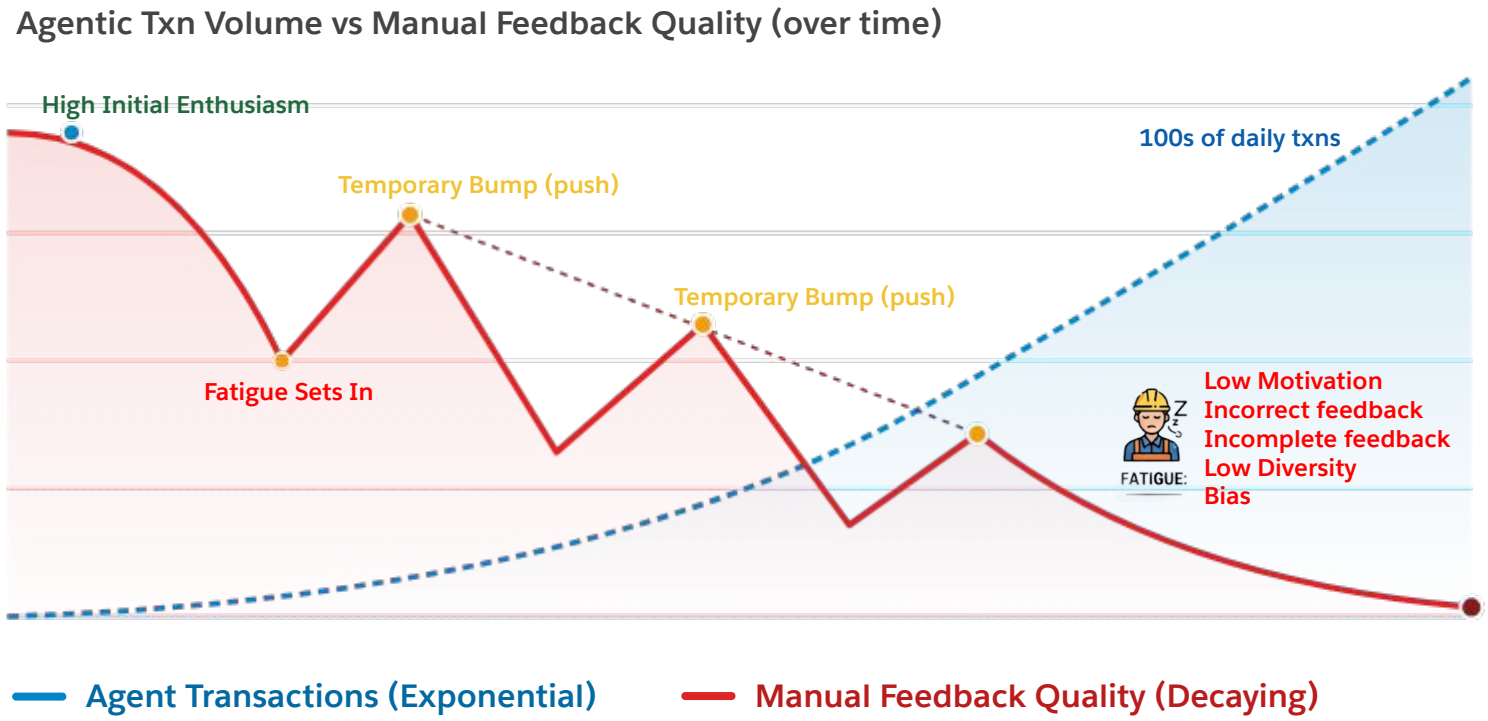


Slack Snippets: (Knowledge Improvements)



1. Weights are dynamically adjusted in the hierarchical RAG queries.
2. As a result, High quality runbooks are retrieved first

Evaluating Agentic Output (bottlenecks)



Scaling Agentic Output Evals

LLM as Judge Rubric

1. **Aggregate RCA Quality**
   Aggregate Quality Score, rolled up from all evals
2. **Tool Limitation Honesty**
   Admits what its tools can't do, instead of bluffing.
3. **Evidence Grounding**
   Claims grounded in tool & data evidence
4. **Root Cause Correctness**
   Identifies the true cause, not just the symptom.
5. **Completeness & Actionability**
   Covers everything needed, in steps you can act on.

? Is it sufficient?

Stratified Human Feedback Sampling

Sampling Formula

$n_0 = Z^2 p(1 - p) / e^2$

(95% confidence)

88 samples
Required sample size

1,000 txns
Total transaction volume

Final Calibration



LLM as Judge
Automated Score

≈



Sampled Human
Feedback Score

Slack Snippets: (Eval Scores)

Feedback Funnel [redacted]

- 123 feedback requests → 83 received (67.5%) → 40 not received (32.5%)
- 83 Feedbacks received → *83 actually rated (68 correct (81.9% correct)*, 15 needs-improvement, 0 wrong)

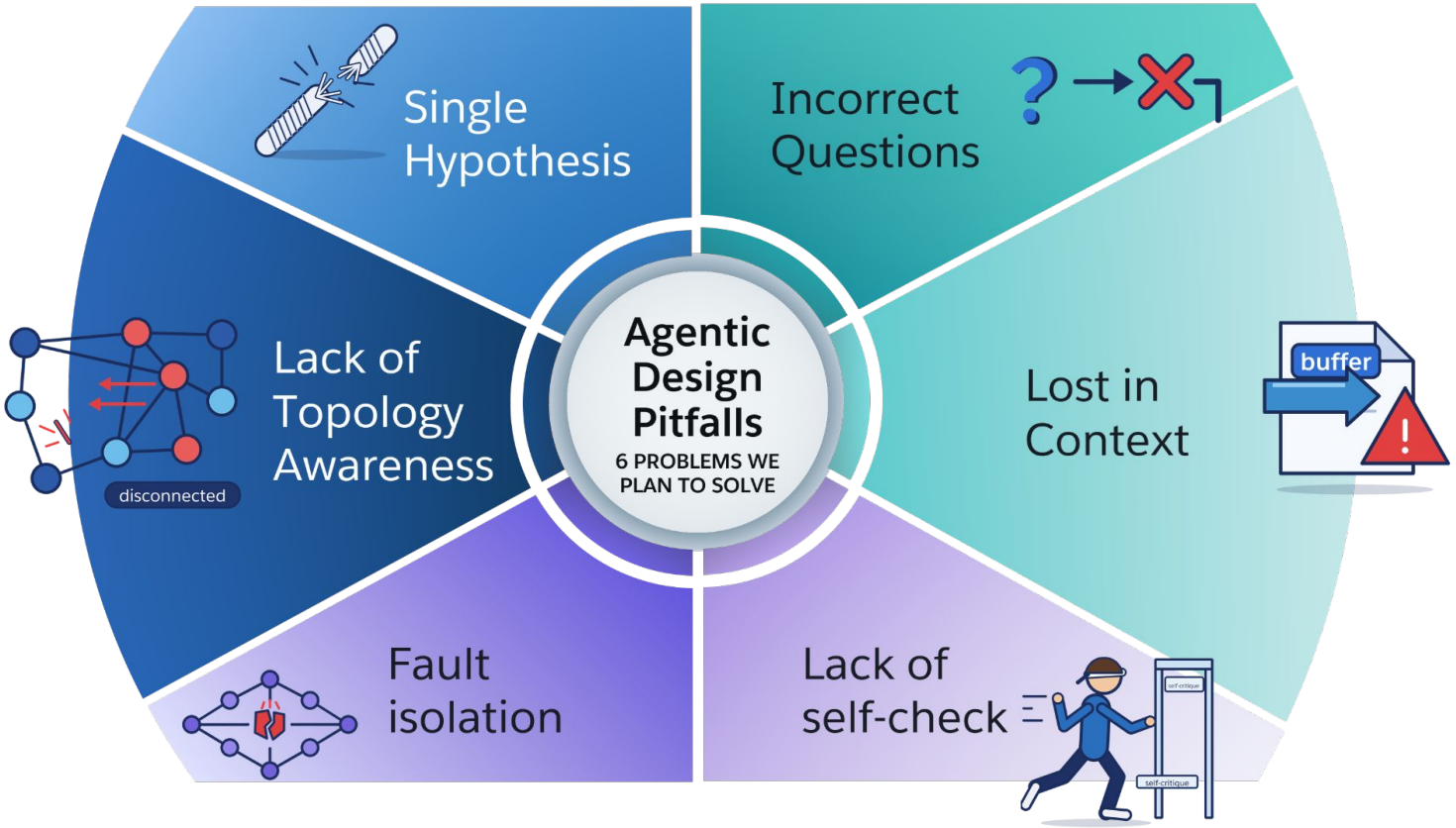
✅ *Feedback Received nearly doubled again: 83 out of 123 requested (67.5%, up from 65.2% last week and 32.0% the week before).* 🎉🎉
➡️ *Pending feedbacks List: Pending feedback*

Touchstone Eval — RCA Quality (83 threads scored, 83 rated)

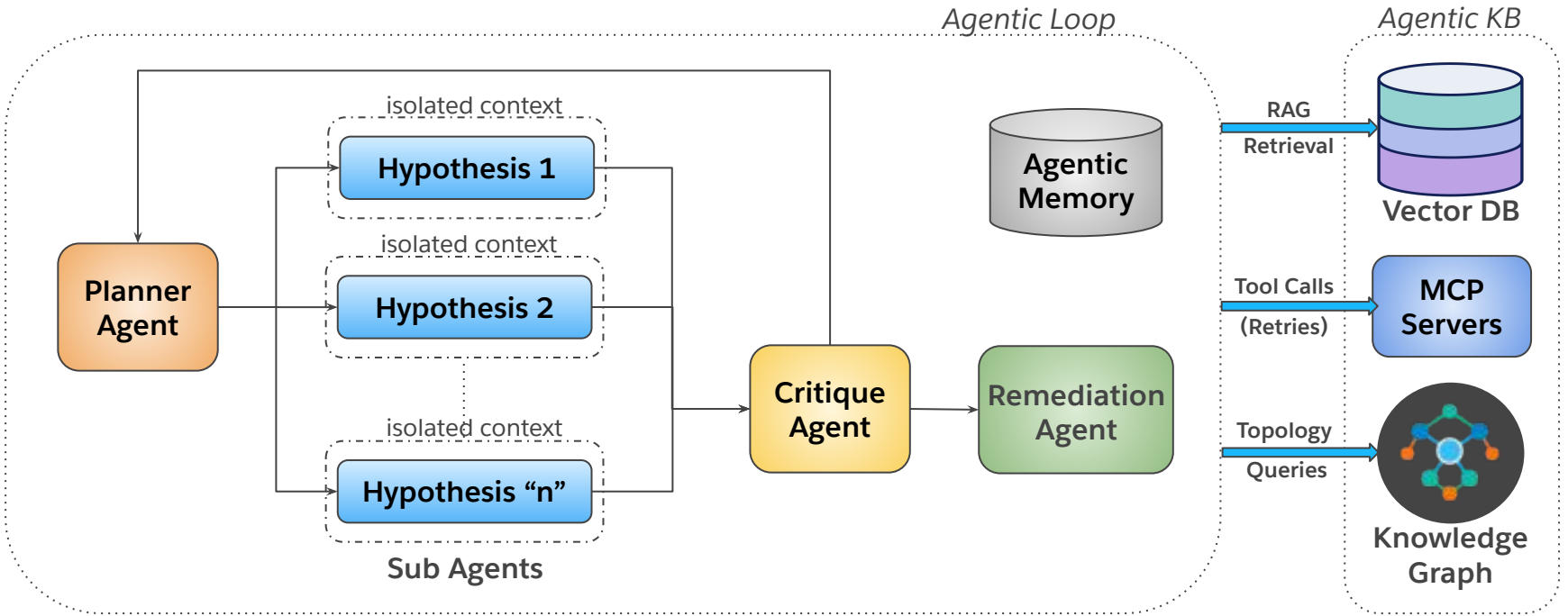
1. Full eval rubric report: Eval Report
2. Aggregate RCA Quality: 3.43/5 ✅ (3.29 → 3.54 → 3.10 → 3.47 → 3.43, steady)
3. Tool-Limitation Honesty 4.55 ✅ (2.22 → 2.25 → 2.67 → 4.62 → 4.55, still the highest rubric)
4. Evidence Grounding 3.75 ✅ (down from 3.89)
5. Root-Cause Correctness 3.28 ✅ (up from 3.20)
6. Completeness 3.06 ✅ (down from 3.22)
7. Actionable Fix 2.95 ❌ (2.96 last week, still the only failing rubric)

Human feedback rated correct for 81.9% of threads rated (68/83), down from 97.8%

Agentic Design Pitfalls



Agentic Design: Evolution



Planner Agent
Prompt Validation
Prompt Enrichment
Dynamic Planner
KG-driven Hypothesis

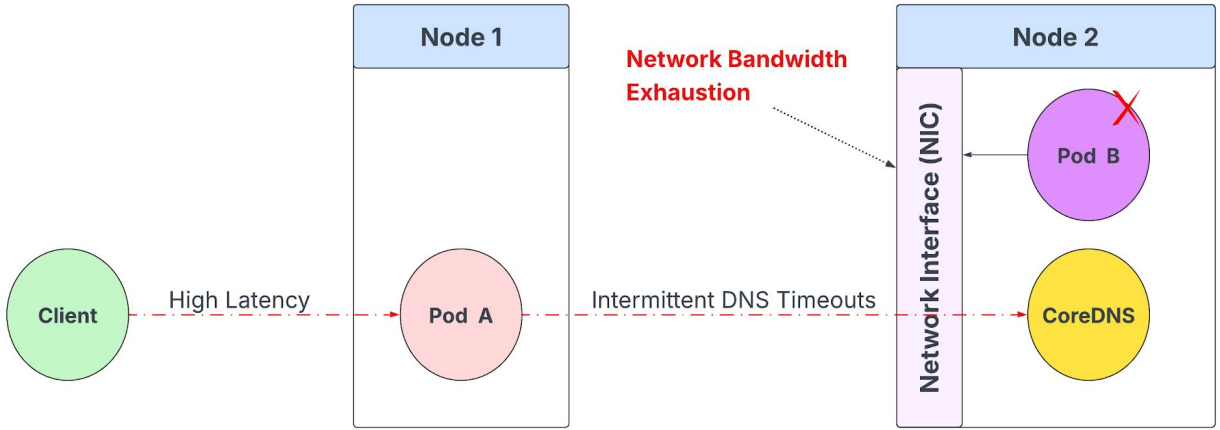
Sub Agents
Parallel Hypothesis Investigations
Isolated Context Mgmt

Critique Agent
Validate Results
Authorize Remediations
Trigger Replanning Loop

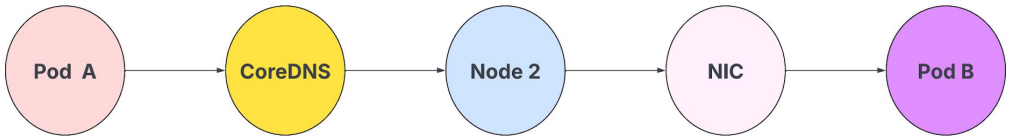
Remediation Agent
Execute Remediations (safely with approvals)

Knowledge Base
MCP Query for Live Data
RAG Retrieval
Dependency Graph
Network Flow Graph

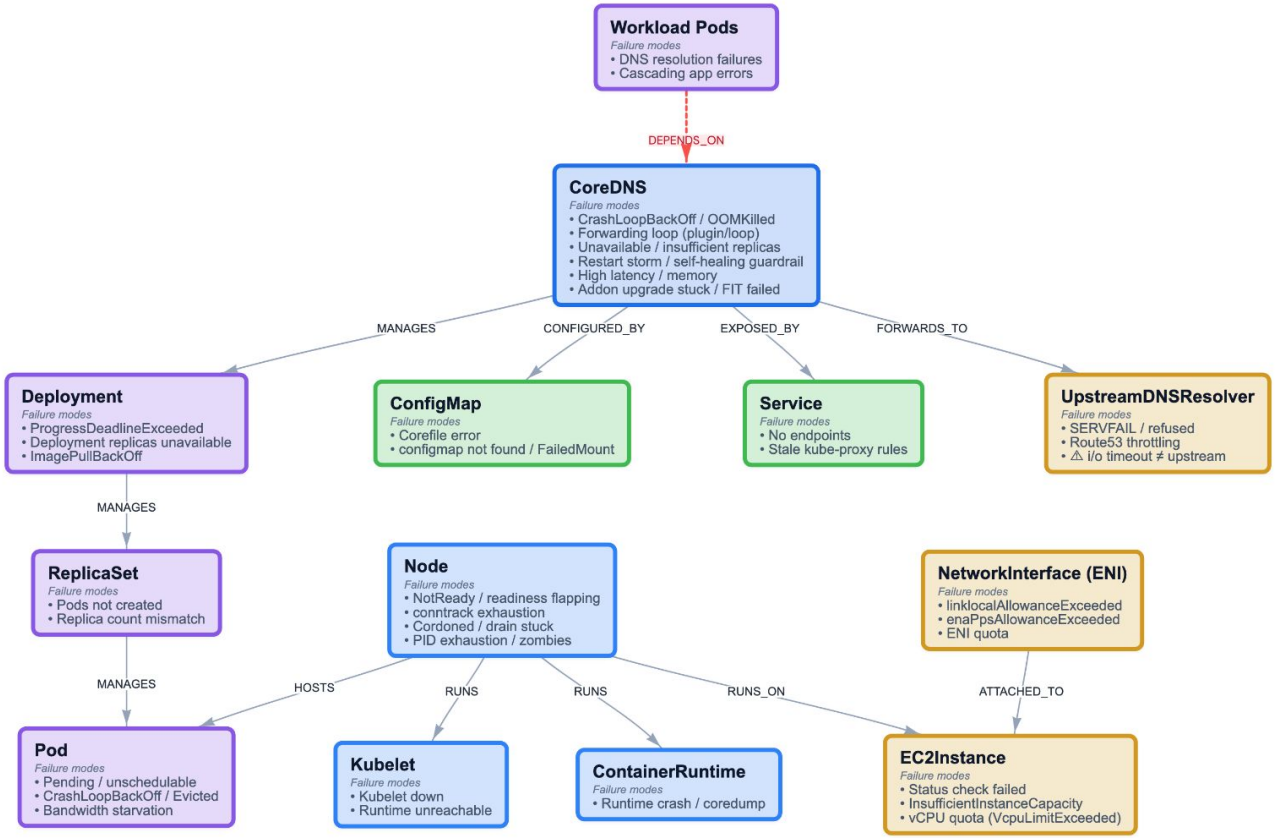
Connecting the Dots



Dependency graph



Dependency Knowledge Graph



Slack references for the Agent runs

[redacted] Yesterday at 2:13 PM

@Aura, I am seeing intermittant failures with name resolutions in my workloads running on [redacted] cluster. Can you investigate why the DNS failures are occuring?

Aura APP Yesterday at 2:20 PM

✅ Investigation complete in 3.7 mins

🔍 View investigation trace

Root cause (confirmed, score 0.80)

The backup worker pod `sam-system/backup-worker-29878440-b2pn4` (container `s3-sync`) on node [redacted] sent about 1.9 TB of outbound traffic in roughly 2.5 hours.

- Its eth0 transmit counter rose from about 1.2 MB at 07:05 UTC to 1.94 TB at 09:35, then flattened. (*h3, confidence 0.80*)
- This used up the node's `ens5` outbound bandwidth allowance. `bw_out_allowance_exceeded` was 0 until about 07:10, then climbed in steps to 162.3M by about 09:00. (*h2, confidence 0.75*)
- As a result, the co-located CoreDNS pod's upstream lookups to [redacted] timed out.
- The other CoreDNS nodes ([redacted], [redacted], [redacted] stayed at 0. Note that `20-211` has data only up to about 08:00.

Culprit pod

`sam-system/backup-worker-29878440-b2pn4` . Per h3, no other pod on the node came close; the next largest sent about 1.3 GB.

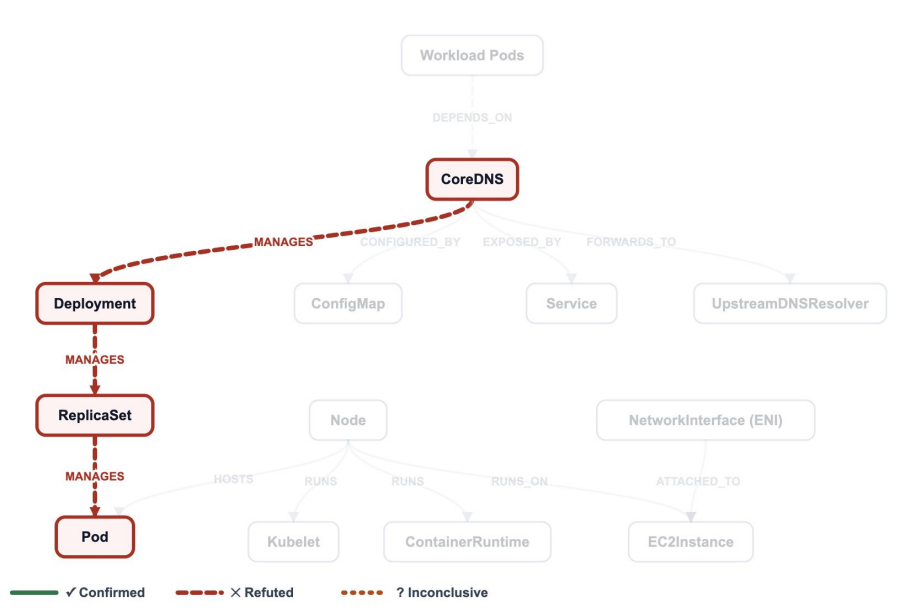
Impacted pods

`kube-system/coredns-6d8b4b6cc4-nc2kb` on [redacted] It logged all 119,997 [redacted] `A: read udp` [redacted] `->` [redacted] `i/o timeout` errors, in the 08:30 UTC bucket. The other six CoreDNS pods logged no [redacted] errors (about 1.9k log lines each). The record does not name any other impacted pods.

How the agent walks the knowledge graph (1)

First pass: the planner turned graph paths into four hypotheses, tested in parallel. Verdicts shown as reported.

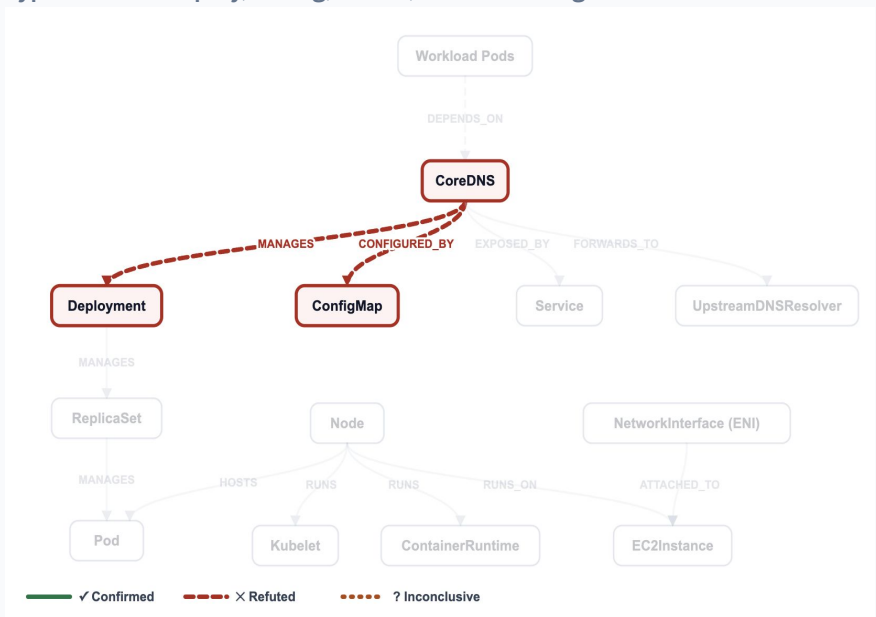
Hypothesis 1: CPU Throttling and Pressure



✕ **Refuted:** Confidence 0.88

Finding: CPU throttling was zero on every CoreDNS pod. Memory sat far below the limit.

Hypothesis 2: Deploy, config, secret, manual change



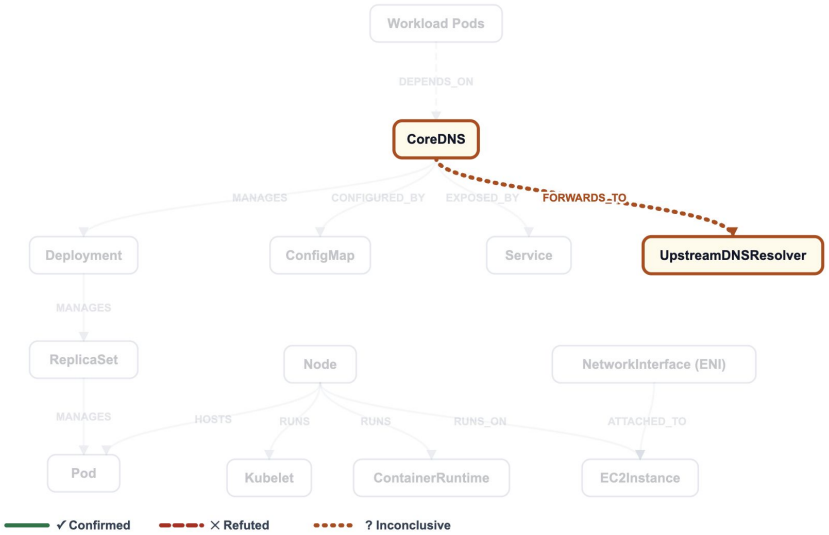
✕ **Refuted:** **Confidence 0.55**

Finding: No deploy, config, secret or manual change found. The CoreFile was well-formed.

How the agent walks the knowledge graph (2)

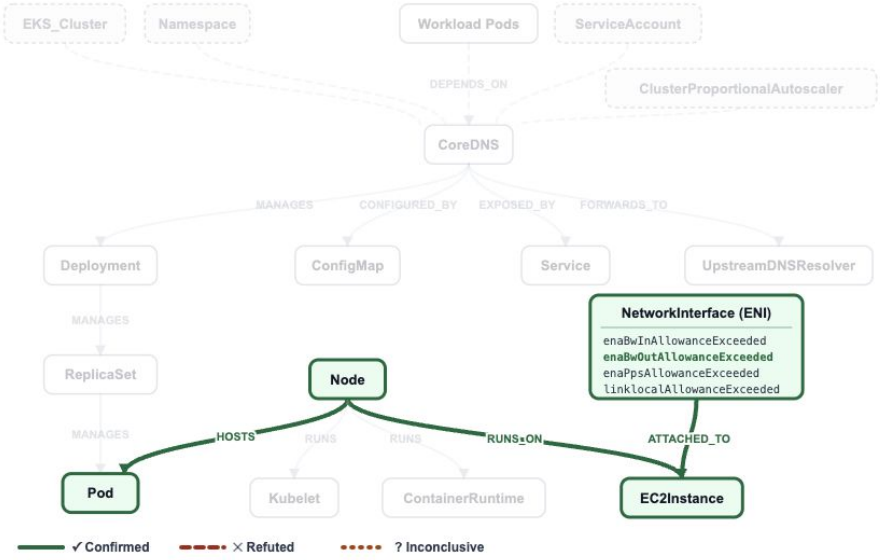
First pass: the planner turned graph paths into four hypotheses, tested in parallel. Verdicts shown as reported.

Hypothesis 3: Upstream resolver throttling



? Inconclusive: Confidence 0.40
Finding: The upstream kept answering other queries, and its health-check failures stayed flat. Two graph lookups confirmed relation to CoreDNS.

Hypothesis 4: Node level noisy neighbour problems



Confirmed: Confidence 0.80
Finding: Another pod on the same node running network-intensive traffic

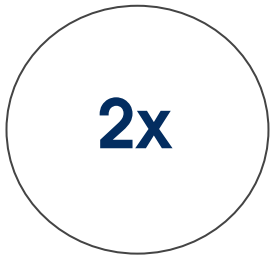
Impact



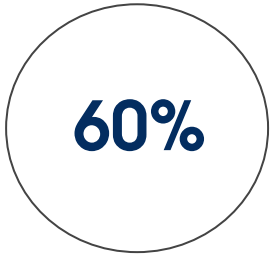
RCA Accuracy



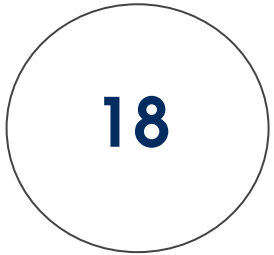
Reduced Human Toil



Platform Scale



Troubleshooting Time



Teams

What's Next?

Path to fully autonomous troubleshooting agents



Building a Ecosystem of Agents

Cross-agent collaboration via A2A protocols, integrations with internal agents and vendor agents like AWS DevOps Agent



Zero Touch Knowledge & True Self-Learning

Continuous autonomous synthesis of post-mortems and runbooks from resolution telemetry without human curation fatigue.



Advanced Reasoning Capabilities

Agents using First-Principles deductive reasoning to solve novel and complex performance problems not seen before



Cost efficiency at Scale

Ensure the agentic solutions stay economical as the volume of agentic transactions continues to grow

Thank You!



We would appreciate your feedback!

1000 Clusters, 1 Smarter Brain: The Next Evolution of AIOps at Salesforce