
BEAM
SUMMIT

Apache Beams' Role in Agentic AI Era

BEAM SUMMIT NYC 2026

Building the Data Backbone for Autonomous AI Agents

Agenda

01 Agentic AI Overview

02 Scaling LLM Inference and Agent Execution

03 Data Freshness (The Context Gap)

04 Memory Management

05 Real-time Guardrails (Governance)

06 Summary

Agentic AI Overview

From Scripted...

Simple, predefined automation workflows with limited adaptability.

...To Autonomous

Agents that can plan, use tools, and execute complex tasks at scale.

Autonomous AI Agents

1. Planning & Reasoning

Decomposes complex enterprise objectives into manageable steps. Employs reflection and reasoning loops (e.g., ReAct, Chain-of-Thought) to analyze intermediate outcomes before taking action.

2. Memory Systems

Coordinates short-term task sensory buffers with deep long-term storage vector layers, ensuring persistent context is maintained cleanly across multi-turn sessions.

3. Tool Utilization

Connects raw LLM engines directly to external operational environments, like triggering SQL queries, executing code scripts in sandboxes, and querying internal API networks.

Autonomous AI Agent - Challenges

Data Freshness

Agents rely on up-to-date information. Stale data leads to grounding errors and flawed decision-making. High latency in data retrieval pipelines further compounds this issue, making real-time responses difficult.

The Context Layer

Managing state is critical but complex. Maintaining context for long-running tasks introduces latency constraints and risks **context window congestion**, forcing difficult trade-offs between history and performance.

Token Cost

Autonomous loops are **highly iterative**. A single prompt can trigger multiple tool calls, vector searches, and model re-queries, leading to unpredictable and potentially exponential token consumption and costs.

So the Question is:

How does Apache Beam fit within the Autonomous AI Agent world?

Scalable LLM and Agent Inference

Run Inference API

High throughput ML inference for LLMs and other self-hosted models. Local vs Remote options.

Dynamic Intelligent Batching

Automatically batches inputs based on throughput for optimal LLM efficiency

Bulk Inference & Evals

Process large scale context data or generate responses for millions of requests.

Run Agent Transforms

Orchestrates multi-step agent actions directly within the data pipeline.

Data Freshness (The Context Gap)

Documents

Databases

Web Content

Apache Beam Pipeline

Building the knowledge ingestion pipelines that ground agents in factual data. Continuous Enrichment!

Data Preprocessing: Chunking documents.

MLTransform: Generating embeddings.

Real-time Enrichment: Fetching context.

Memory Management

Short-Term Memory with Business Logic Wrapper

Stateful DoFns maintain context across multi-step autonomous workflows.

Session State

Persistent sessions track user-agent interactions over time.

Resource Optimization

Shared handlers minimize overhead during continuous execution.
Cross-Process Weight Sharing and vLLM Singleton Infrastructure.

Real-time Guardrails (Governance)

Validation Layer (The Inline Gatekeeper)

LLM get struck in a loop, so ensure the repetitive behavior is detected and rectified.

Ensure that an agent's planned action or generated response meets safety and compliance standards before it is executed or sent to a user.

Evaluation Layer (Automated Auditor)

Continuous safety checks and prompt classification. Beam pipelines categorize incoming prompts and agent tool calls to analyze and understand agent behavior at massive scale.

Observability Layer

Multi-model ensemble and cascade patterns for reliable monitoring. Orchestrate complex, multi-model pipelines where the output of one model (e.g., image captioner) feeds into another (e.g., LLM reasoner).

Summary! Bring it all together...

Scalable Agent Inference/Evals

High throughput ML inference for LLMs and other self-hosted models

Data Freshness

Automatically batches inputs based on throughput for optimal LLM efficiency

Memory

Process large scale context data or generate responses for millions of users.

Governance

real-time guardrails to ensure agent actions are safe and compliant.

Ashwin Sampathkumar

Linkedin

QUESTIONS?