Bridging the gap: Taking agents from prototype to production

Mark Roy | Tech Lead – Agentic AI | AWS
linkedin.com/in/markproy

AI agent production-readiness challenges

Excitement and potential

Prototype Agents

Challenges on the path to production

Performance

Scalability

Security

Governance

Meaningful business value

Production Agents

GETTING AGENTS TO PRODUCTION IS STILL TOO HARD

Securely execute and scale agent code

Remember past interactions & learning

Identity and access controls for all agents and tools

Agentic tool use for executing complex workflows

Discover and connect with custom tools and resources

Understand and audit every interaction

NEW

Amazon Bedrock AgentCore

Deploy and operate highly effective agents securely, at scale using any framework and model

PREVIEW

Amazon Bedrock AgentCore

Deploy and operate highly capable agents securely, at scale using any framework and model

TIME TO VALUE

Build powerful AI agents without infrastructure and ops headaches

FLEXIBLE

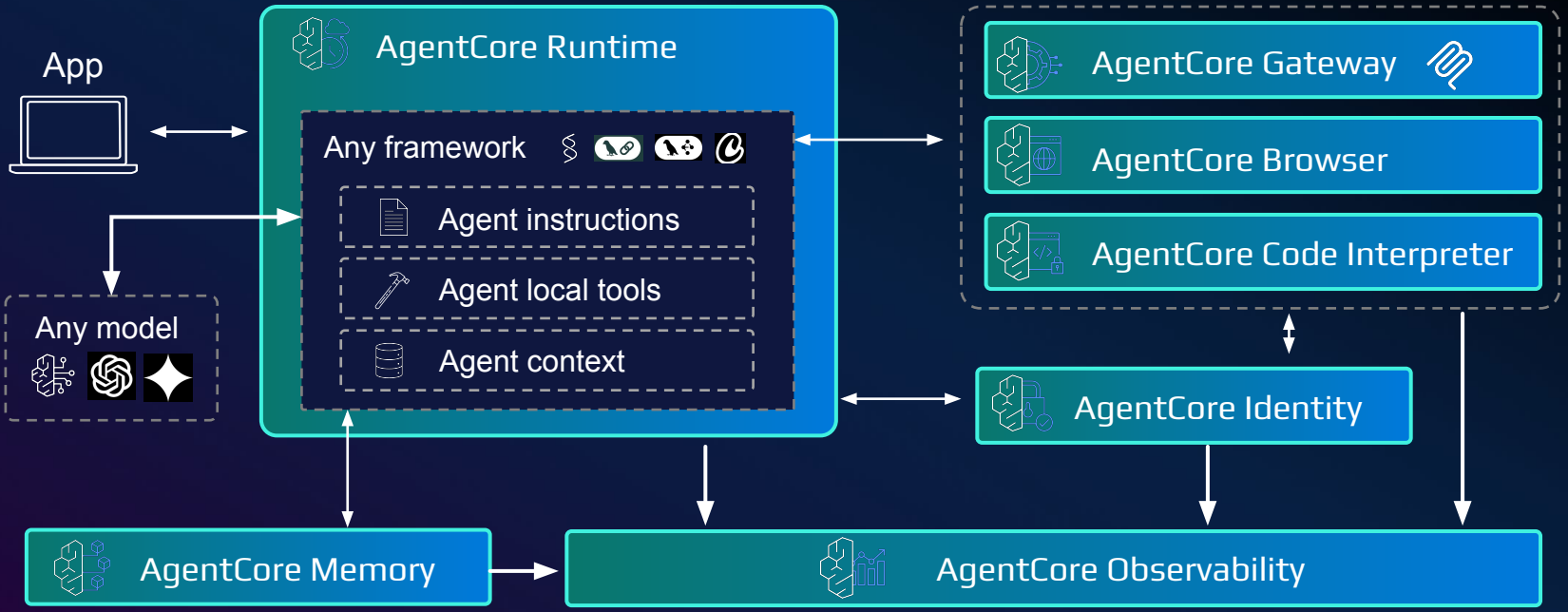
Create agents with any framework or model

TRUSTED

Deploy secure, scalable, and reliable agents your organization can trust

AgentCore for production-ready agents

Any model, any agent framework



Building scalable production-ready agents with AgentCore

AgentCore Runtime

Framework and model independent deployment

- Support for Strands Agents, LangChain, LangGraph, CrewAI and other agentic frameworks
- Use Amazon Bedrock, Amazon SageMaker, OpenAI, Gemini or any other model on your agent

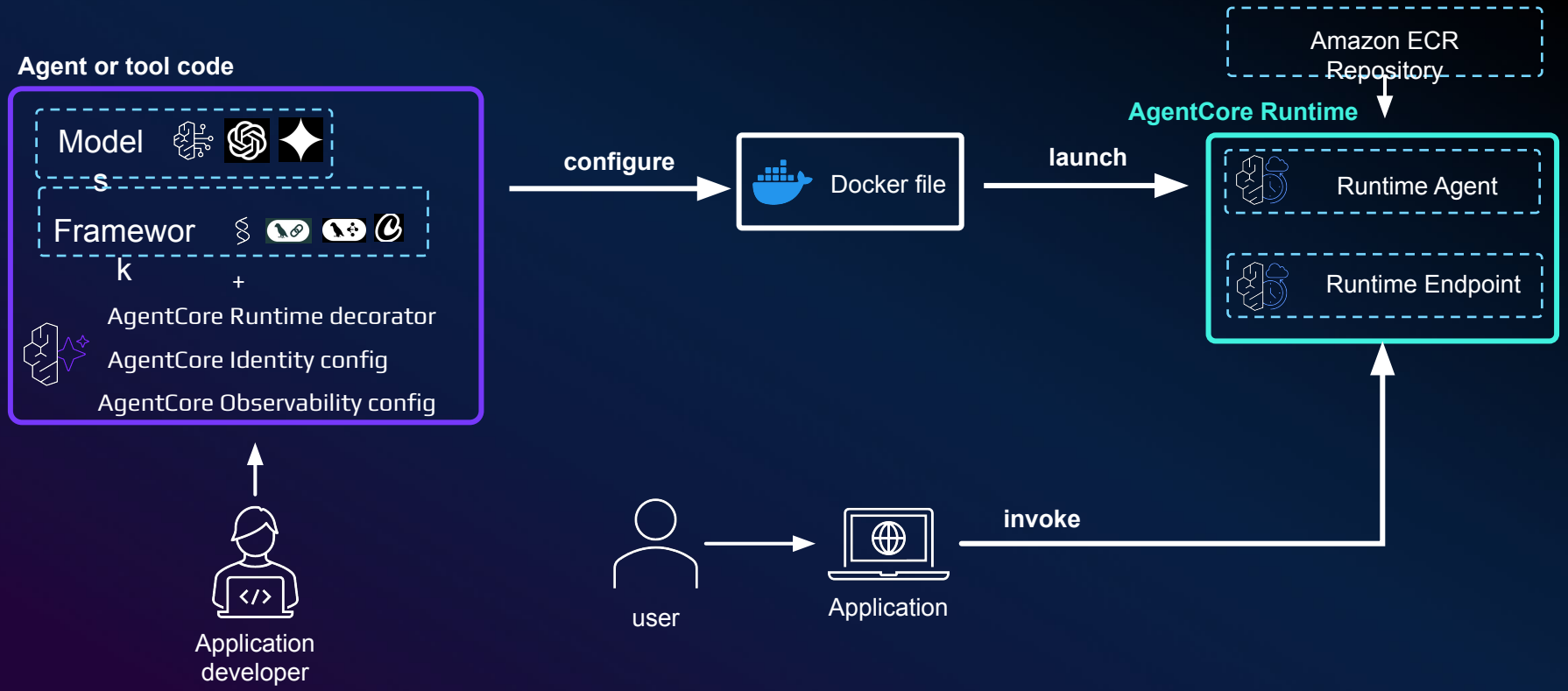
Use case independent

- Handle large payload sizes: text, images, audios, videos and others
- Fast initialization and long running asynchronous workloads
- Host agents and tools

Secure workload with enterprise-grade isolation

- True session isolation with persistent dedicated execution environments
- Built-in authentication and observability capabilities

Secure and scalable runtime for agents and tools



AgentCore Gateway

Simplify tool development and integration

- Transform APIs into agent-ready tools without custom code or infrastructure
- Supports RESTful services via OpenAPI schema, and AWS Lambda functions

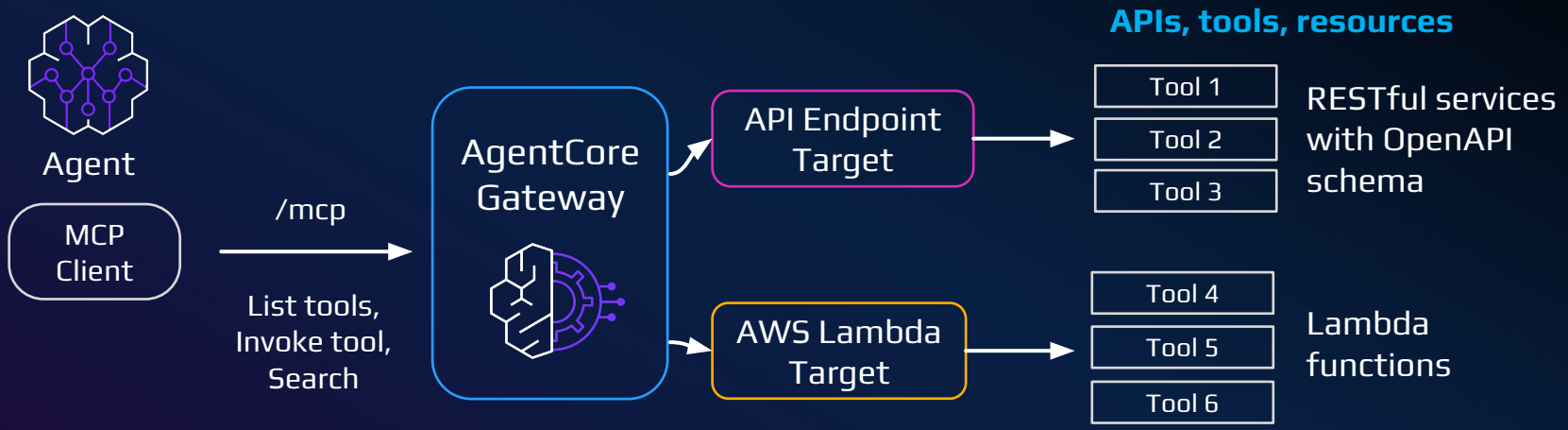
Secure and unified tool access

- Simplified cross-organizational tool sharing with enterprise-grade security
- Built-in inbound and outbound authentication and observability

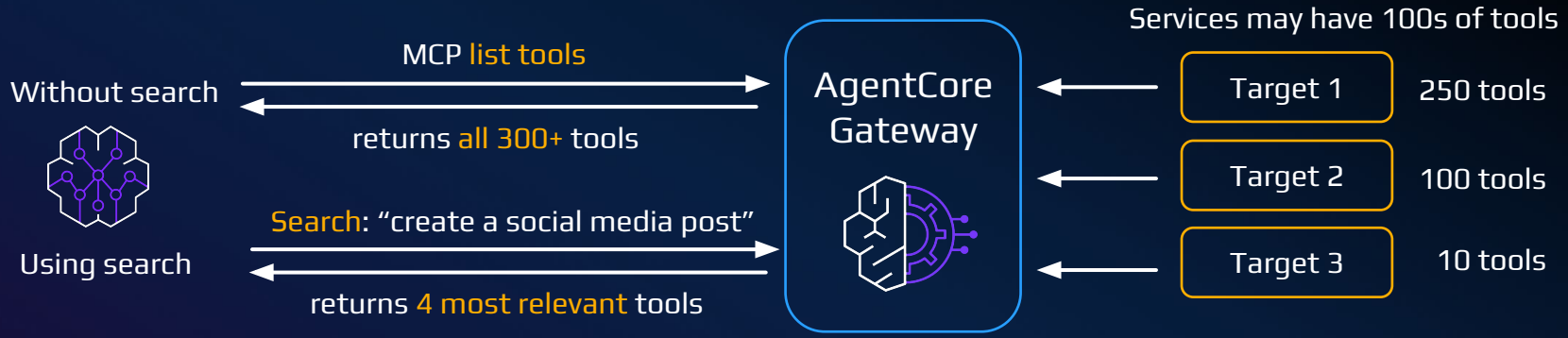
Intelligent tool discovery capabilities

- Expose tools using MCP enabling tool discoverability
- Built-in semantic search matches tools to agent tasks

Unified, secure, agent-ready tools



AgentCore Gateway semantic search



Benefits

- AgentCore Gateway automatically indexes tools, and gives serverless semantic search
- Reduces context passed to the agent’s LLM, improving accuracy, speed, and cost
- Lets agent focus on tools relevant for a given task

AgentCore Memory

Short-term and long term memory management

- Long term memory capabilities from session interactions
- Handle multiple memory strategies for a single conversation

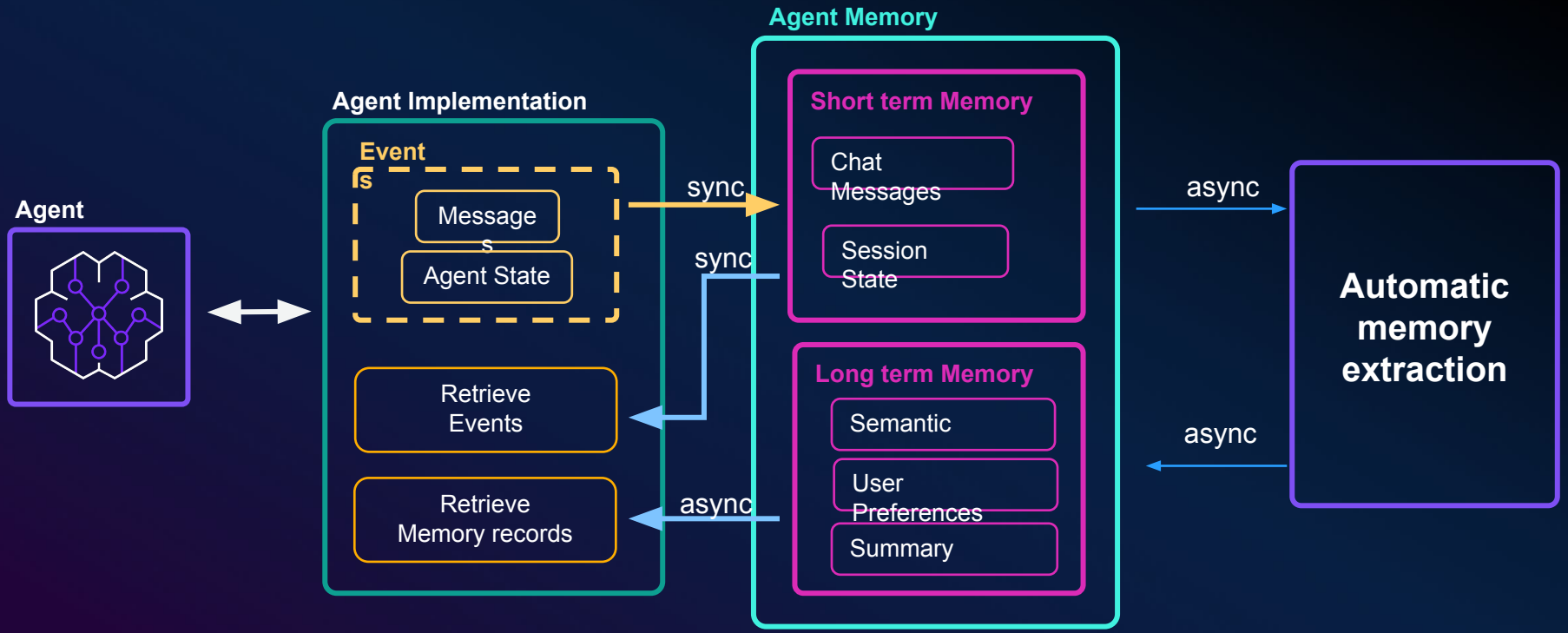
Fully managed and secure

- Abstracts memory infrastructure
- Built-in encryption
- Flexible namespace for memory sharing

Retrieval and extraction of memory

- Multiple memory retrieval patterns – semantic search, filter by namespace
- Built-in and custom memory extraction strategies – summary per session, user preferences, semantic memory

Short-term and long-term memory capabilities



AgentCore Identity

Secure access to agents and tools

- Distinct identities for secure agent operations at scale
- Authentication with enterprise identity providers
- Secure credential management for external service access and integration

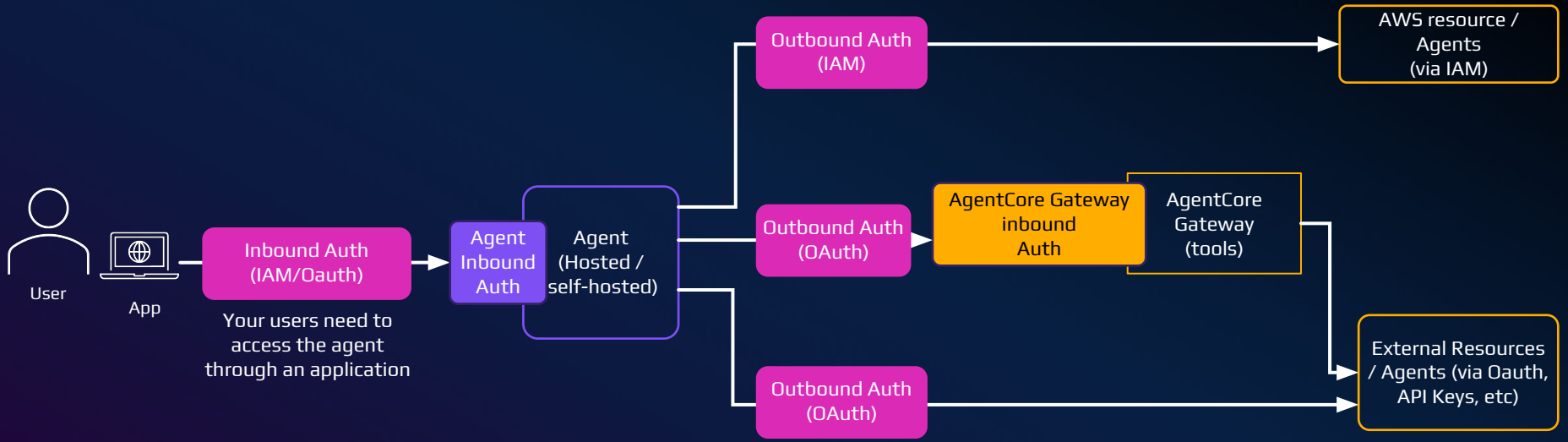
Minimized consent fatigue

- Reduces need for repeated authorization
- Streamlines authentication flows
- Simplifies user experiencer for all agent-powered interactions

Accelerated AI agent development

- Preserves existing identity systems such as Okta, Microsoft Entra ID, or Amazon Cognito
- Inbound and outbound authentication

Inbound and outbound authentication for agents



AgentCore Browser

Serverless and fully managed

- Low latency browser sessions
- Auto-scales from 0 to hundreds of concurrent session

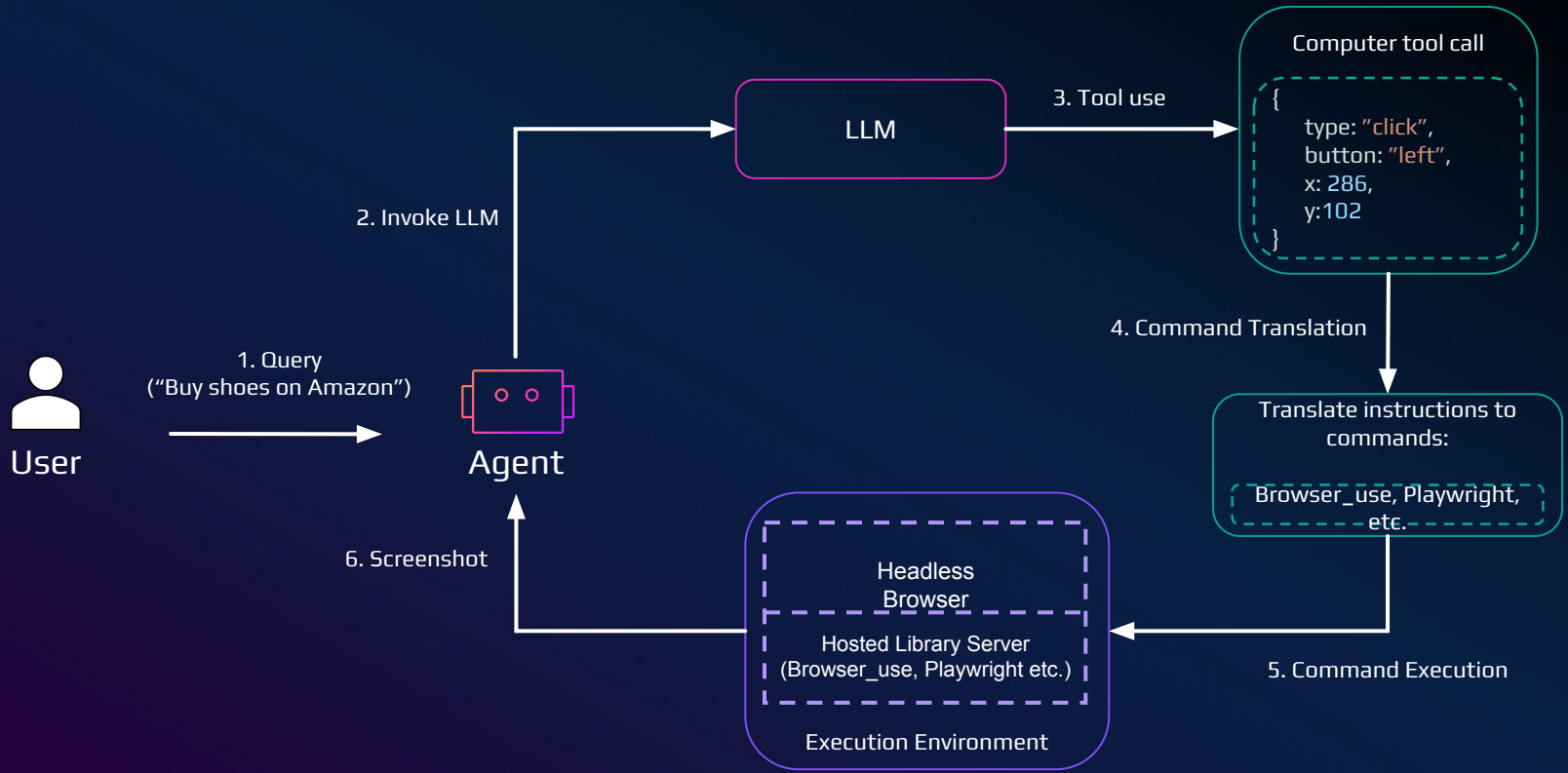
Enterprise-grade security

- Session isolated compute with VM-level isolation per user
- VPC connectivity with configurable network modes
- Secure credential handling

Enterprise Observability

- Live streaming for real-time monitoring
- Session replays for debugging
- Extensive logging of all browser commands to CloudTrail

Web navigation and workflow automation



AgentCore Code Interpreter

Execute Code Securely

- Execute complex workflows and data analysis in isolated sandbox environments
- Access internal data sources securely without exposing sensitive data

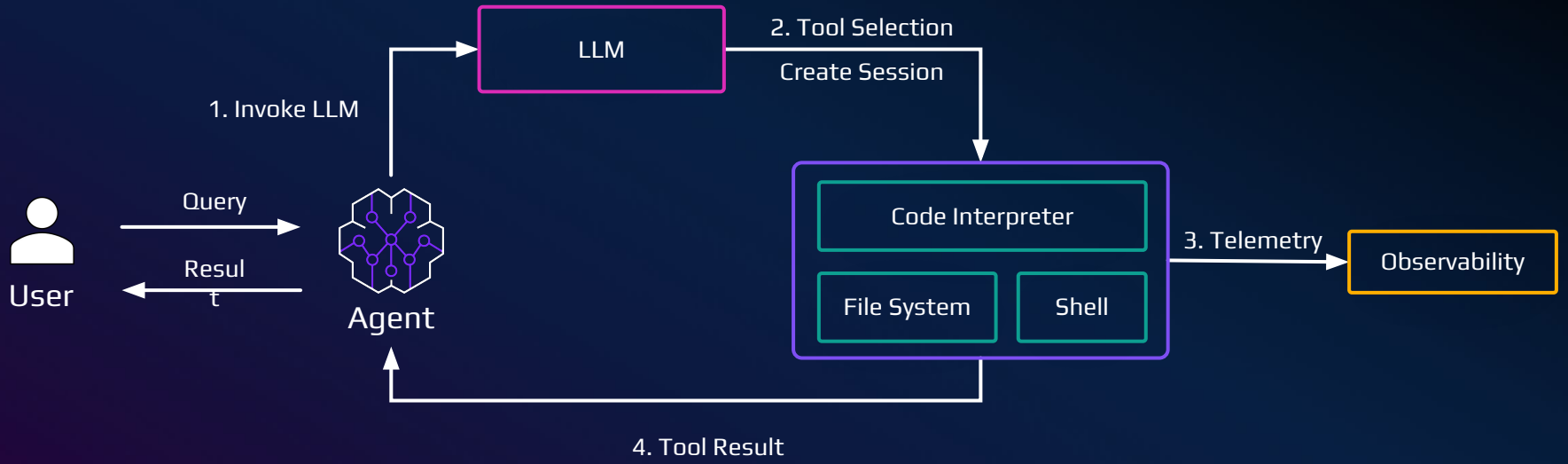
Monitoring and large-scale data processing

- Monitor and troubleshoot code execution with comprehensive observability features
- Process large datasets efficiently using Amazon S3 integration

Ease of use

- Pre-built execution runtimes for JavaScript, TypeScript, and Python with common libraries pre-installed
- Customization to add your own packages

Securely write and execute code in an isolated environment



AgentCore Observability

Full transparency of agent behavior

- Visualize agent workflows with detailed traces in deep dive dashboard
- Comprehensive monitoring for latency, tokens, tools and custom metadata

Enterprise ready security and governance

- Enterprise ready with IAM access controls and PII redaction capabilities

Integrate with 3P Observability tools

- OpenTelemetry (OTEL) compatible for integrating logs, metrics, and traces
- Flexibility to leverage your existing observability stack

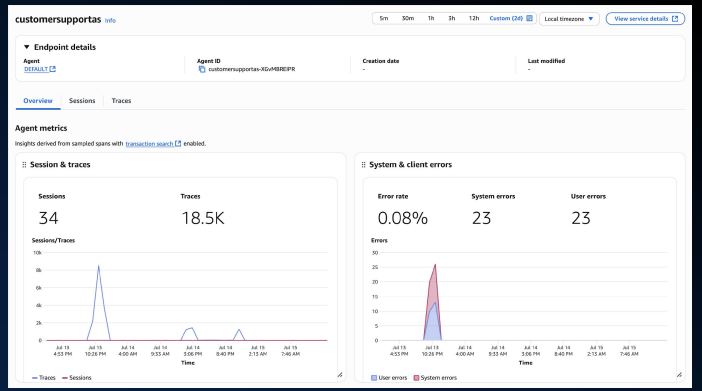
AgentCore Observability

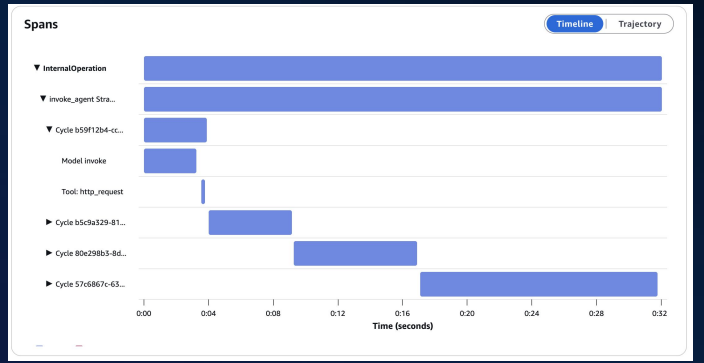
Agent deployed on AgentCore Runtime

Agent running hosted on other technologies

AgentCore Observability

AgentCore Observability dashboards





Closing

Dive deeper on AgentCore

AgentCore samples and tutorials

AgentCore capabilities

AgentCore workshop

Thank You!

Mark Roy | Tech Lead – Agentic AI | AWS
linkedin.com/in/markproy