New tech in town

amazon bedrock

is a fully manage service that lets you use foundation models Claude, LLaMa, Mistral) withouth managing infrastructure

supports “agent capabilities” like RAG, guardrails and tool to use

Agent Core (Guardrails, RAG, Knowledge Bases)

guardralils: polices that filter right and shape LLM output (for safety and formatting)

RAG: allows the retrieval from external knowledge databases or docs and then inject into prompts

Knowlegde bases: Pre-inexed data stores that Bedrock can query for RAG

Agent Core: orchestration layer around the raw models

Amazon SageMaker

is basically a platform to build train and deploying ML models

two layers:

- SageMaker(full platform) –> train/deploy custom models with infra handled
- SageMaker Ai (lower level) –> Raw ML componets (training jobs, endpoints, datasets)

EC2 (Elastic Compute Cloud)

is a raw compute layer - virtual machines (linux/windows servers) on demand

AWS SDK’s

python boto3: allows to call AWS API’s

MCP Server Protocol

MCP = Model Context Protocol is a standardized way for agents/LLM’s to connetc to external tools or services

- you can define toold/functions that agents can call
- you can define resorces and promts (so it can be reuseable)
- any MCP-compliant client (like an LLM or orchestrator) can talk to the servers