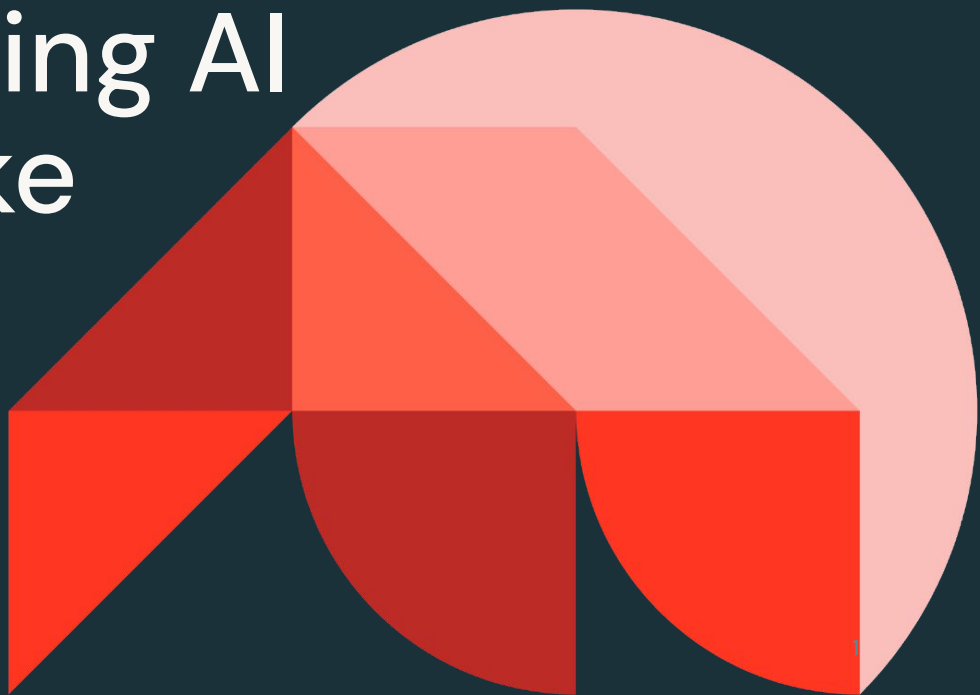




From Data to Decisions: Building AI Agents That Take Action

Maria Zervou



AI is no longer a trend— It's a business imperative

1. *The State of AI in 2024*, [McKinsey & Company](#)

2. From Potential to Profit: Closing the AI Impact Gap, [BCG AI Radar](#)

3. How Generative AI Is Changing The Future Of Work, [Oliver Wyman Forum Generative AI Survey](#)

65%

of organizations now use generative AI in at least one business function, nearly doubling from last year

McKinsey & Company¹

75%

of executives rank AI/GenAI as a top three strategic priority

Boston Consulting Group²

>50%

of employees say they use generative AI weekly at work

Oliver Wyman³





OK...

why
not?

Challenge:

Building and deploying
production-quality Gen
AI solutions

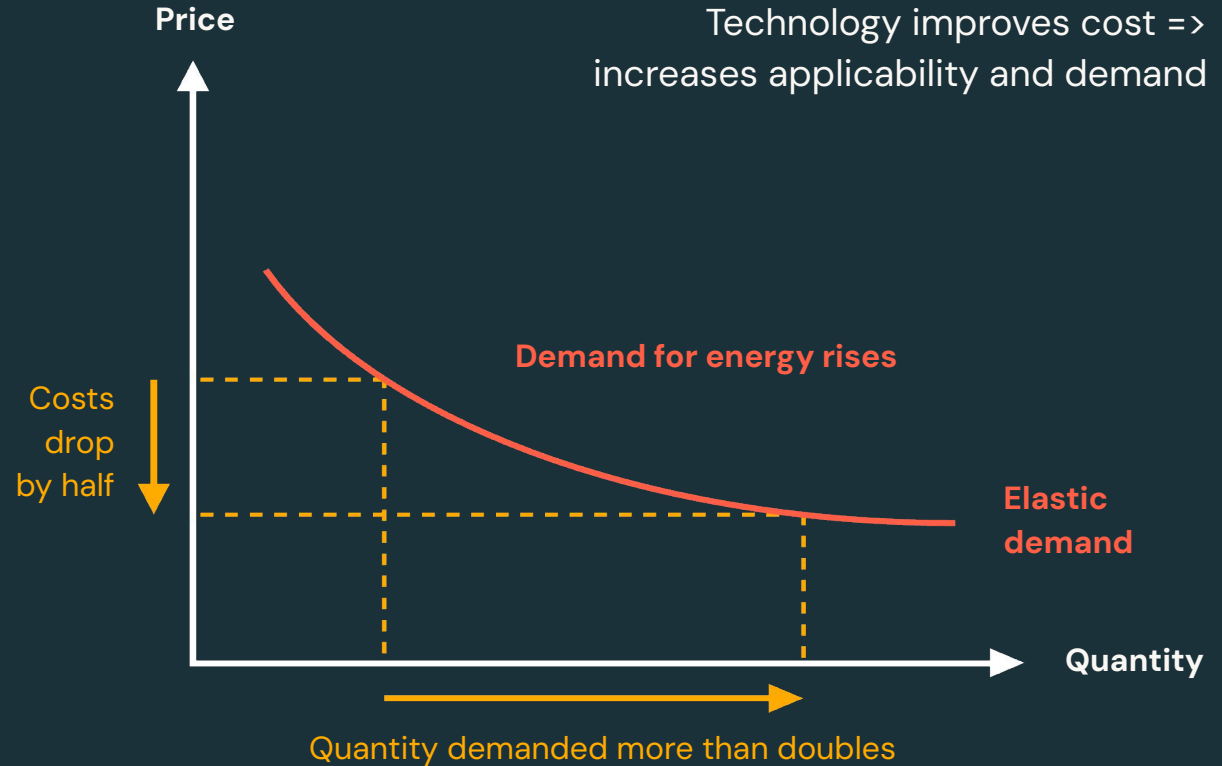
90%

of enterprises *not*
confident going
to production

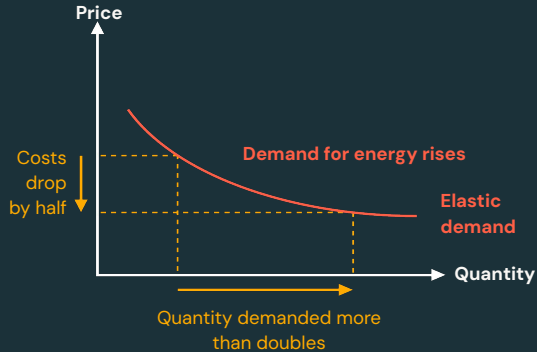


Jevons Paradox

With increased efficiency comes higher consumption



Mosaic's Law



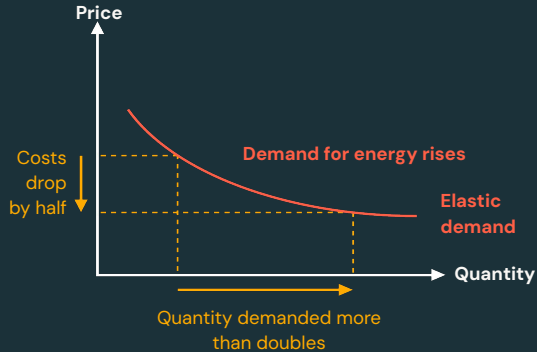
The amount of learning
per dollar will quadruple
every year.

4x YoY for 20 years =
**10 orders of magnitude to
biological parity**

Why haven't cheap
models caused
this disruption?

Jevons Paradox

decreased friction
~~With increased efficiency~~
comes higher consumption



GenAI disrupts the
cost of converting
intention → action

USER INTERFACE

UI

General Intelligence

Consumer models trained on a broad dataset **disconnected** from your business data

vs

Data Intelligence

AI **that reasons** over your enterprise data and is able to solve domain-specific use cases

Difficult to build AI Agents that accurately and securely work on your data



Cannot reason
over your **data**



Agents that lack understanding
of enterprise data will not
respond accurately



Difficult to
evaluate accuracy



Understanding quality and
remediation for your custom use
case is complex

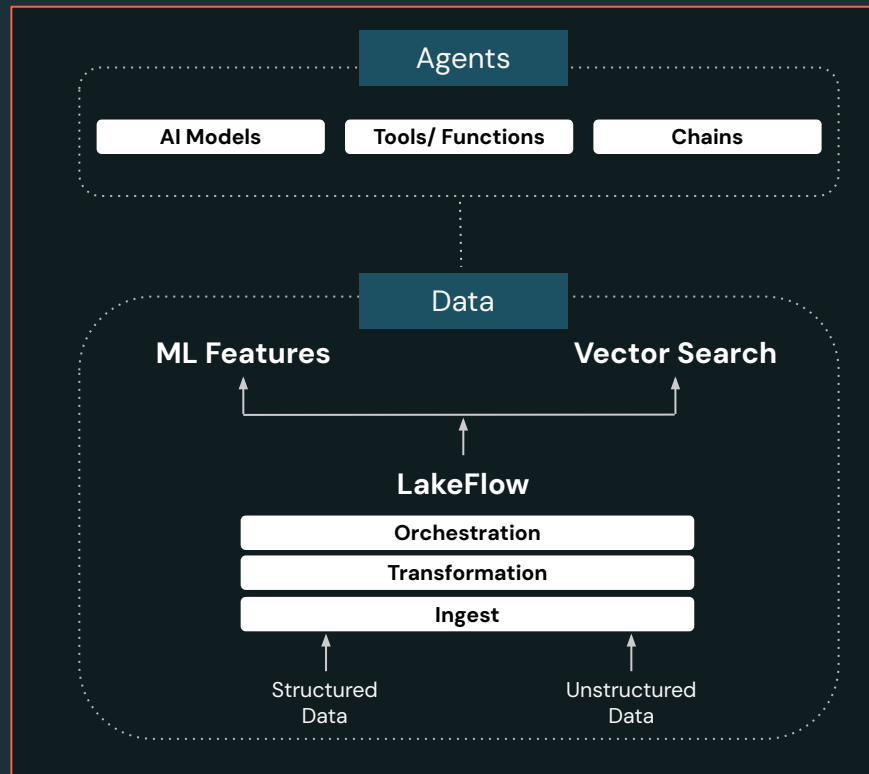


Inability to fully
govern AI apps

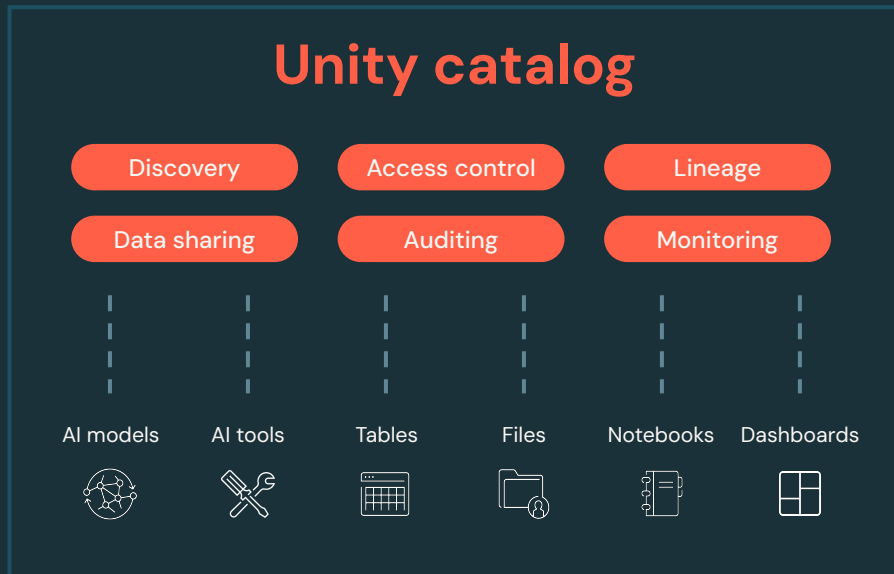


Ensuring access controls,
guardrails, lineage, and cost
controls is disjointed

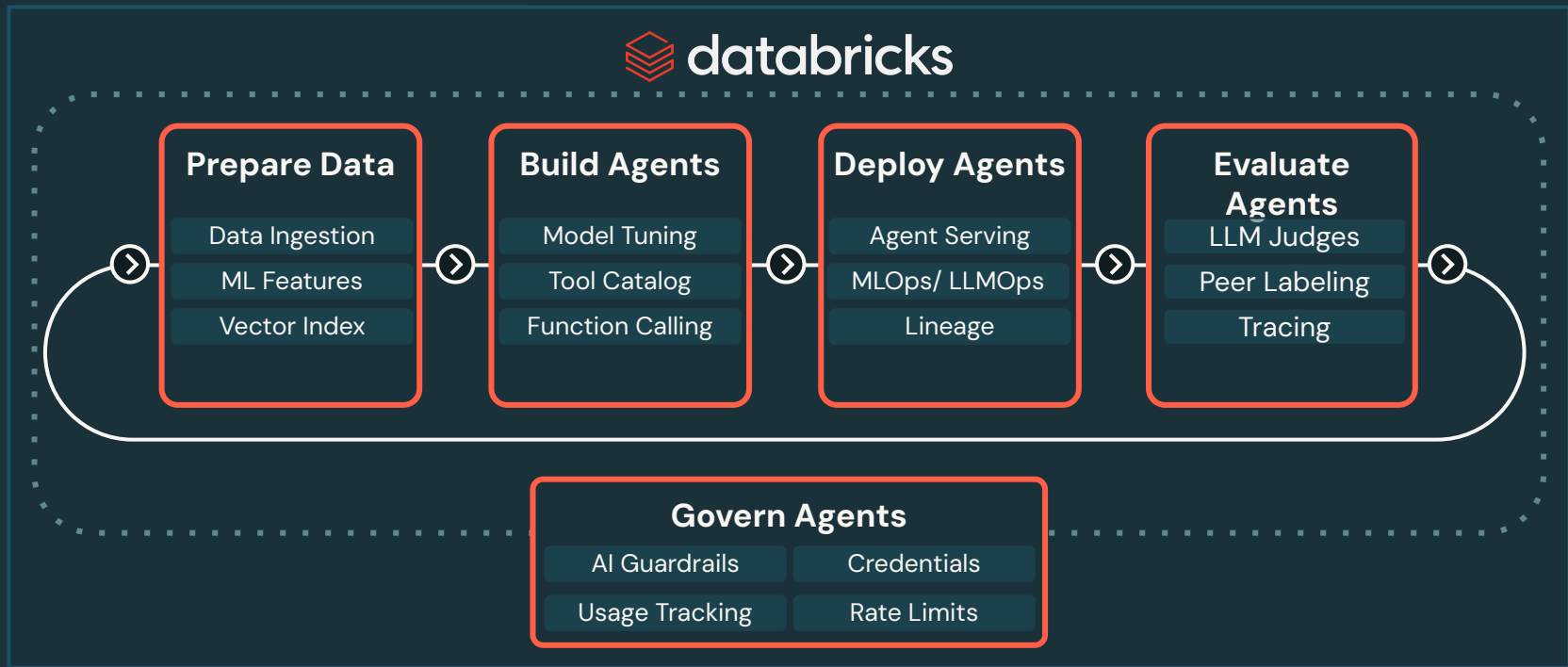
Agents that reason over your data



Governance across data, models and tools



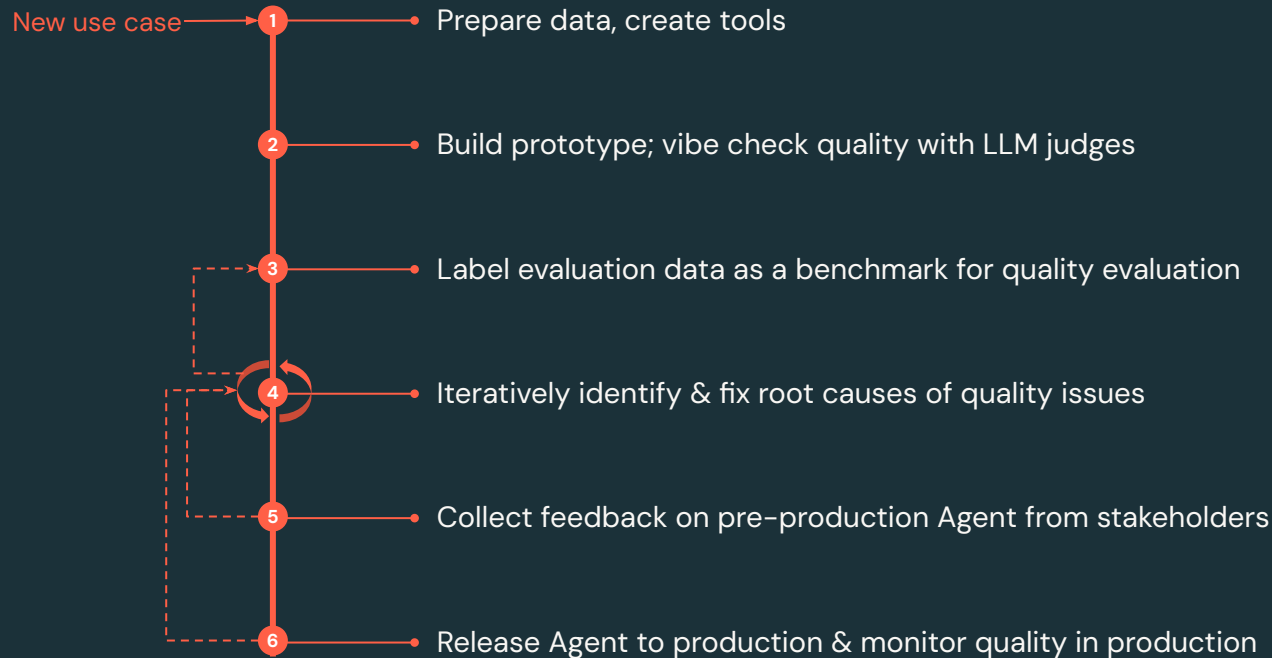
A unified platform to build AI Agents



Agent Lifecycle



Let's walk through the Agent lifecycle

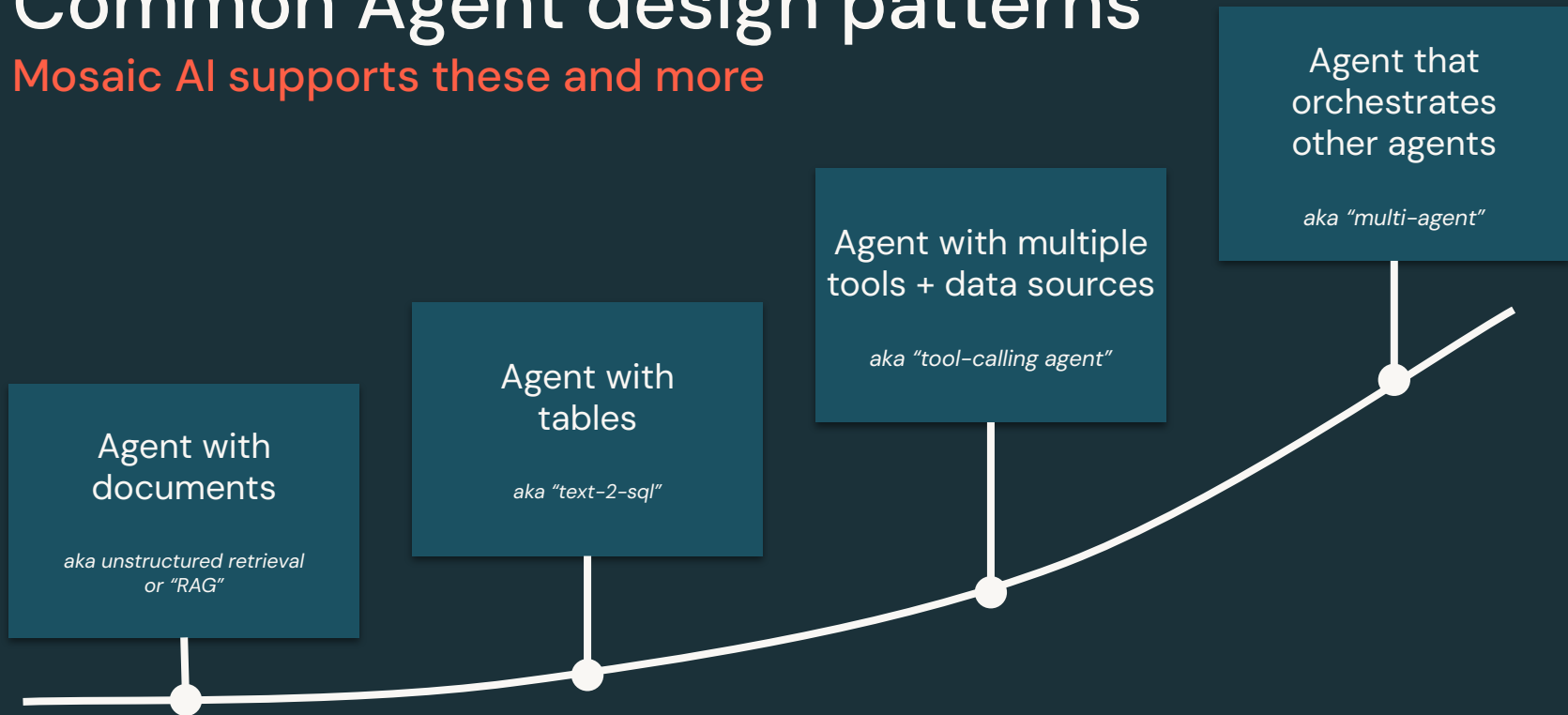


Agents with Tools



Common Agent design patterns

Mosaic AI supports these and more



Increasing Capability, but also Complexity

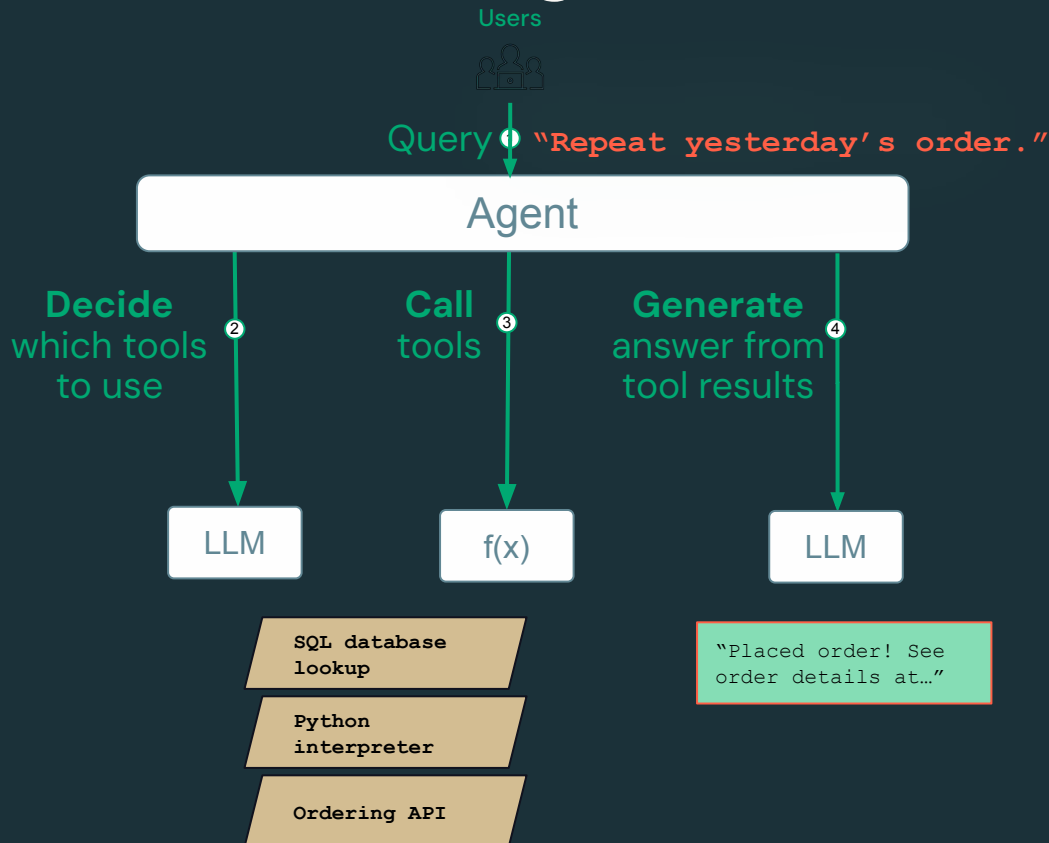
Customer Challenges

How do you **reuse** existing production ready assets **without recreating the wheel?**

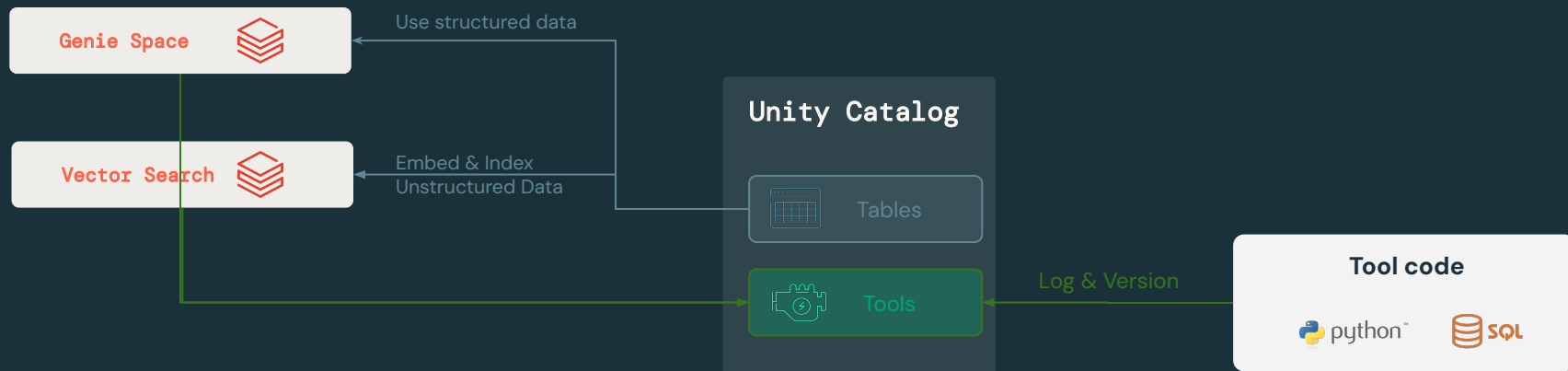


Agents with function/tool calling

Some LLMs can be prompted to call functions or tools, to build complex agents.



Step 1: Prepare data, create tools



Why Tools Matter for AI Agents

Discussion Points

1. What **existing** ML models, APIs, or data sources in your organization could be **exposed as functions** for AI agents?
2. How would you determine which processes should remain **human-led versus agent-led**?
3. What **level of autonomy** would you be comfortable giving to AI agents in your organization?

Agents Iteration

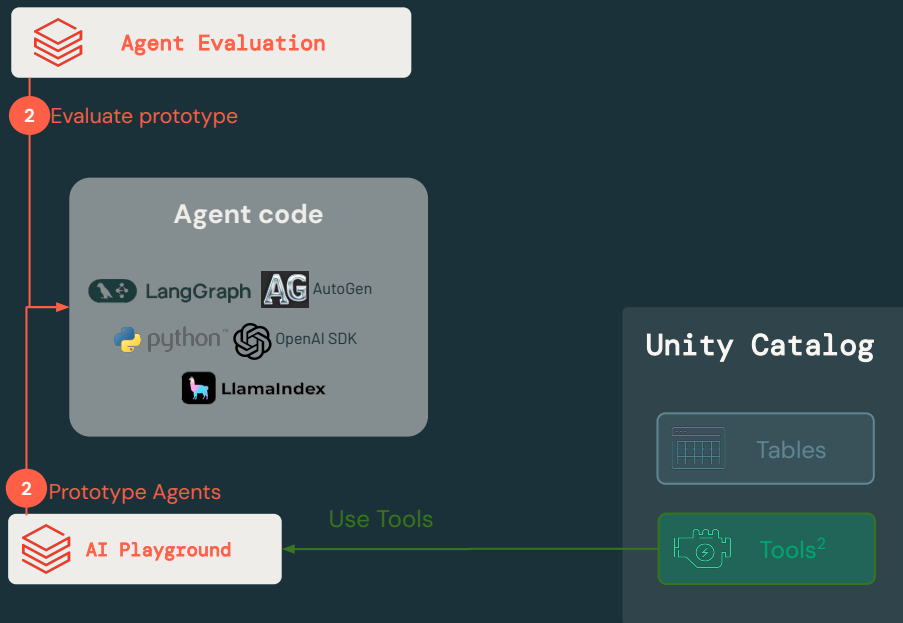


Customer Challenges

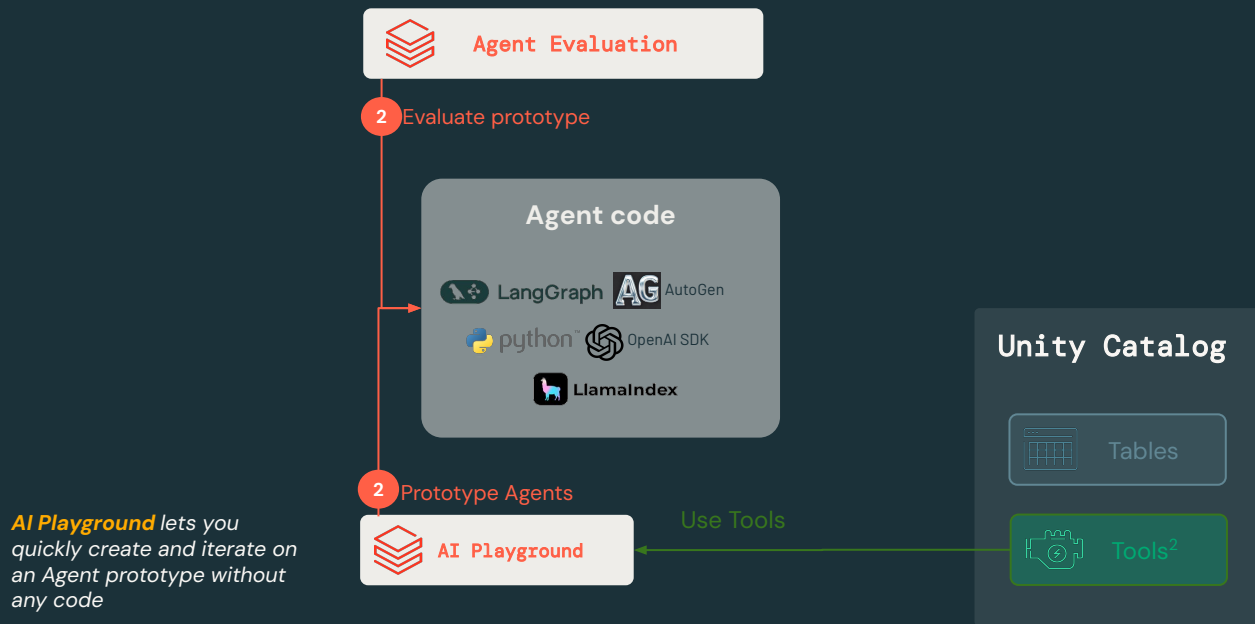
1. How to enable developers of **all skill levels** to **get started quickly**?
2. How to **rapidly iterate** *without needing to* touch code, but yet, still have access to the code when I need more control to improve quality?



Step 2: Rapid prototyping w/ LLM judge quality checks

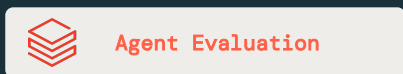


Step 2: Rapid prototyping w/ LLM judge quality checks

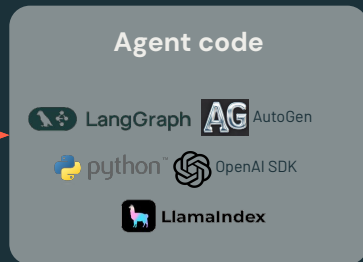


Step 2: Rapid prototyping w/ LLM judge quality checks

Quickly assess initial quality with built-in LLM judges from **Agent Evaluation**



2 Evaluate prototype

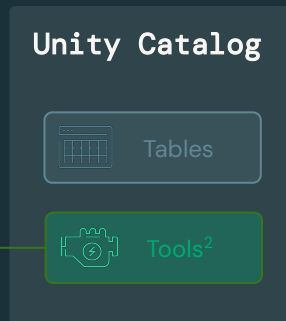


2 Prototype Agents



AI Playground lets you quickly create and iterate on an Agent prototype without any code

Use Tools



**GenAi's biggest
challenge is evaluation!**



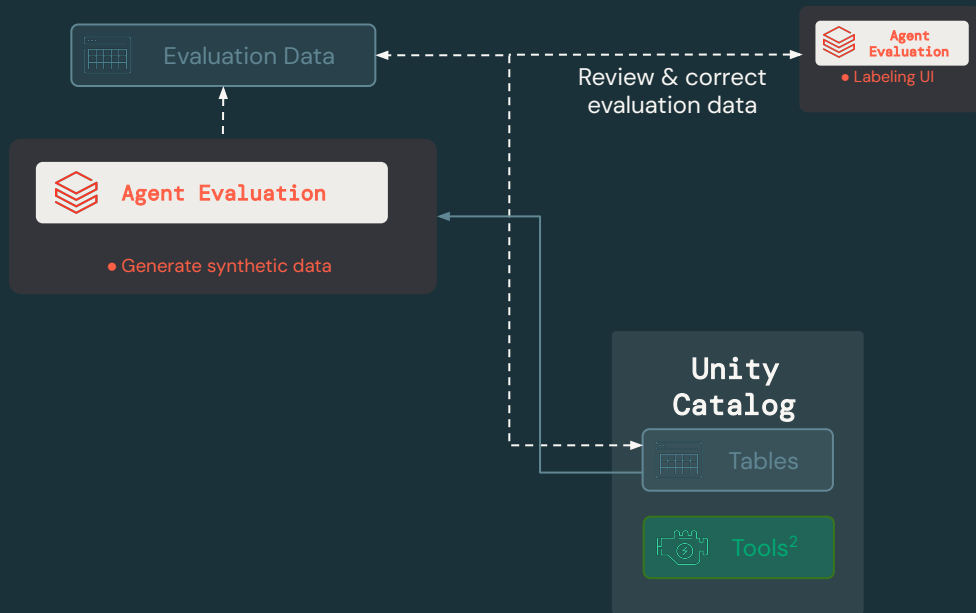
Start with a benchmark!



Customer challenges

1. **Improve quality** with a representative **evaluation dataset** **without SMEs spending months labeling?**
2. **Collect high-quality feedback** from SMEs who are busy and have limited time?

Step 3: Label evaluation dataset



Build your eval dataset

Eval dataset

questions		ground_truth	
21	> What is the requirement for ...	> The requirement for enabling a...	
22	> How can users access the Id...	> Users can access the Ideas Po...	
23	> How can I disable personal a...	> To disable personal access tok...	
24	> How can I set a maximum lif...	> To set a maximum lifetime for ...	
25	> What values do I need to sp...	> In the API call for log delivery, ...	
26	> How long should I expect to ...	> After initial setup or other log ...	
27	> How do I create an IAM role ...	> To create an IAM role for audit ...	



Time for Evaluation



Customer challenges

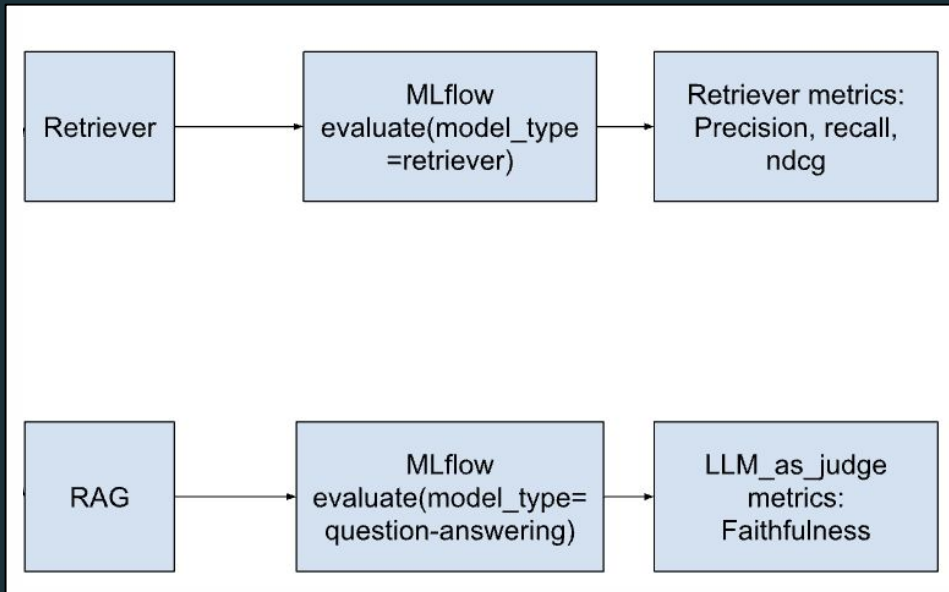
Iteratively identify & fix quality issues

1. What are the **right metrics to evaluate quality**? How do I trust the outputs of these metrics?
2. How do I quickly identify the **root cause of quality problems**?
3. I need to evaluate many ideas – how do I...
 - ...run evaluation quickly so the majority of my time isn't spent waiting?
 - ...track these different versions of my agent?
 - ...quickly debug my Agent's logic?
4. How do I easily **compare 2 versions** of my Agent to understand if my fix worked?
5. What tools do you wish existed for evaluating AI agents?"
Would you use MLflow for AI agent evaluation? What features would make it better?

Agents vs Traditional ML Eval

1. Traditional ML evaluation → Precision, recall, RMSE, F1-score.
2. AI agent evaluation needs additional metrics:
 - ✓ Function-call success rate
 - ✓ Response latency
 - ✓ Correctness of retrieved data
 - ✓ Error handling (does it fail gracefully?)
 - ✓ Action impact (did it take the right step?)

MLFlow for evaluation



Build your eval dataset

Eval dataset

questions	ground_truth
21 > What is the requirement for ...	> The requirement for enabling a...
22 > How can users access the Id...	> Users can access the Ideas Po...
23 > How can I disable personal a...	> To disable personal access tok...
24 > How can I set a maximum lif...	> To set a maximum lifetime for ...
25 > What values do I need to sp...	> In the API call for log delivery, ...
26 > How long should I expect to ...	> After initial setup or other log ...
27 > How do I create an IAM role ...	> To create an IAM role for audit ...

Ask each model to answer the eval questions



baseline model



New model

questions	baseline_answer	FTuned_answer	ground_truth
> What is the requirement for ...	> The requirement for enablin...	> The requirement for enabling a...	> The requirement for enabling a...
> How can users access the Id...	> Users can access the Ideas ...	> Users can access the Ideas Po...	> Users can access the Ideas Po...
> How can I disable personal a...	> To disable personal access ...	> To disable personal access tok...	> To disable personal access tok...
> How can I set a maximum lif...	> To set a maximum lifetime f...	> To set a maximum lifetime for ...	> To set a maximum lifetime for ...
> What values do I need to sp...	> In the API call for log deliver...	> In the API call for log delivery, ...	> In the API call for log delivery, ...
> How long should I expect to ...	> After initial setup or other log...	> After initial setup or other log ...	> After initial setup or other log ...
> How do I create an IAM role ...	> To create an IAM role for au...	> To create an IAM role for audit ...	> To create an IAM role for audit ...

Grade & compare both answers using MLFlow Evaluation



LLM-as-a-Judge metrics with MLFlow

MLFlow.evaluate() API with GenAI metrics

Here is a question with an answer: <question - answer>
You're asked to judge the answer given this expected solution <label>:

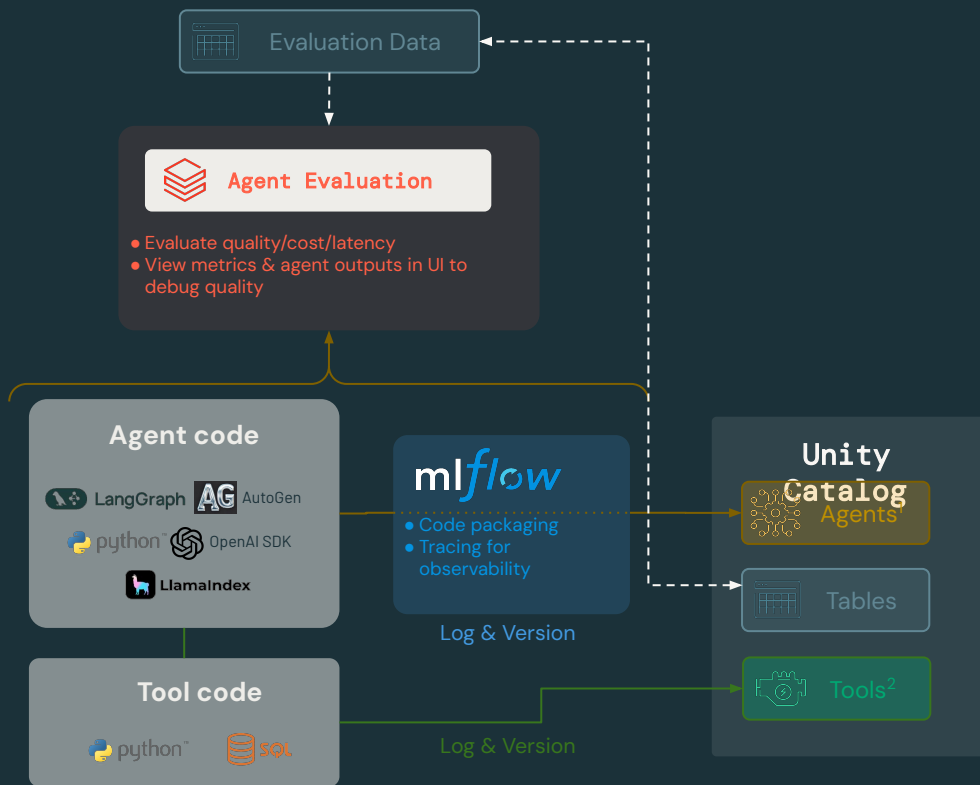
- Is the answer provided correct (note 1-5)?
- Is the answer professional (note 1-5)?
- any custom GenAI metrics



Judge LLM

	Baseline model	New Model
correctness	4.1 (avg)	4.5 (avg)
professionalism	3.5 (avg)	4.2 (avg)
top 10 less correct answers	[...]	[...]

Step 4: Iteratively identify & fix quality issues



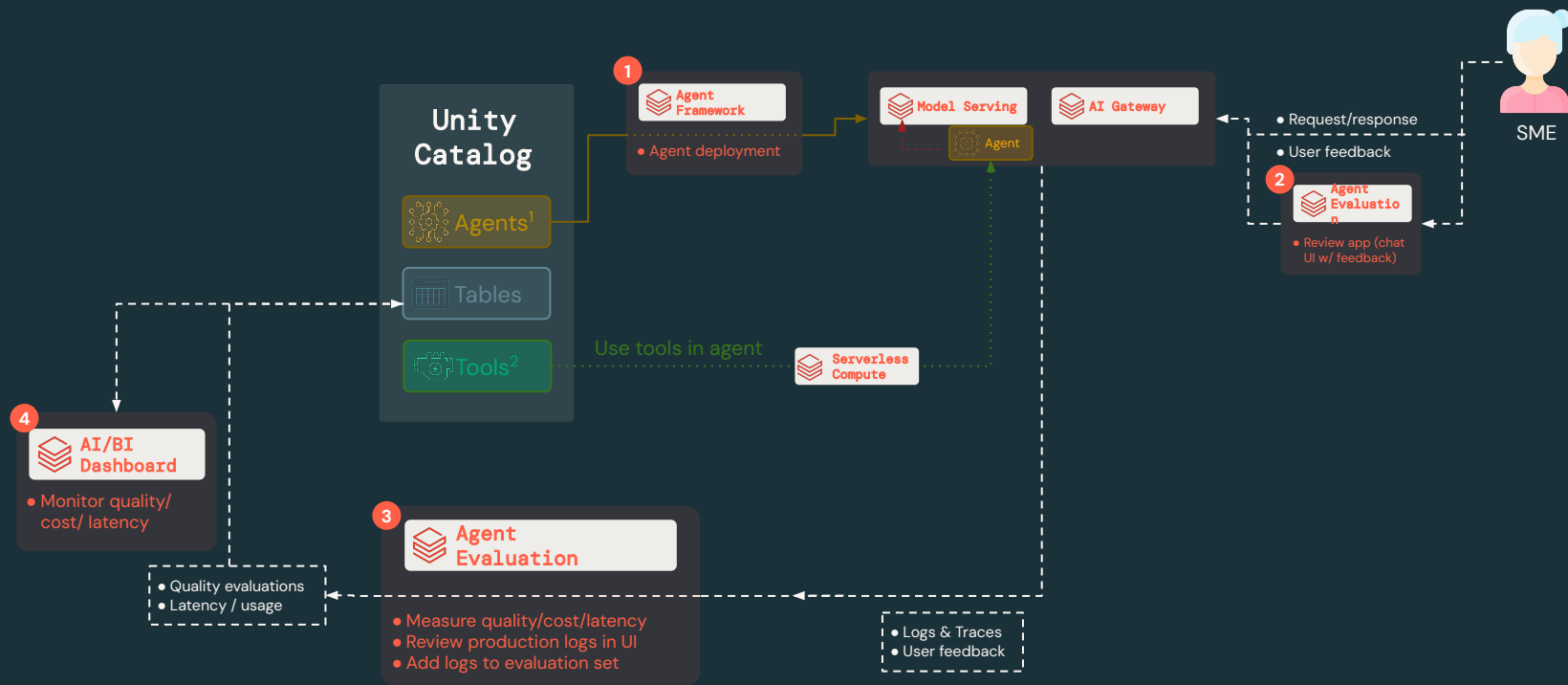
Customer challenges

Pre-production

1. Quickly create a **Chat UI** for **stakeholders** to **test the agent?**
2. **Track** each piece of **feedback?**
3. Understand **100s of pieces** of feedback **at-scale?**



Step 5: Release to pre-production



Time for Monitoring

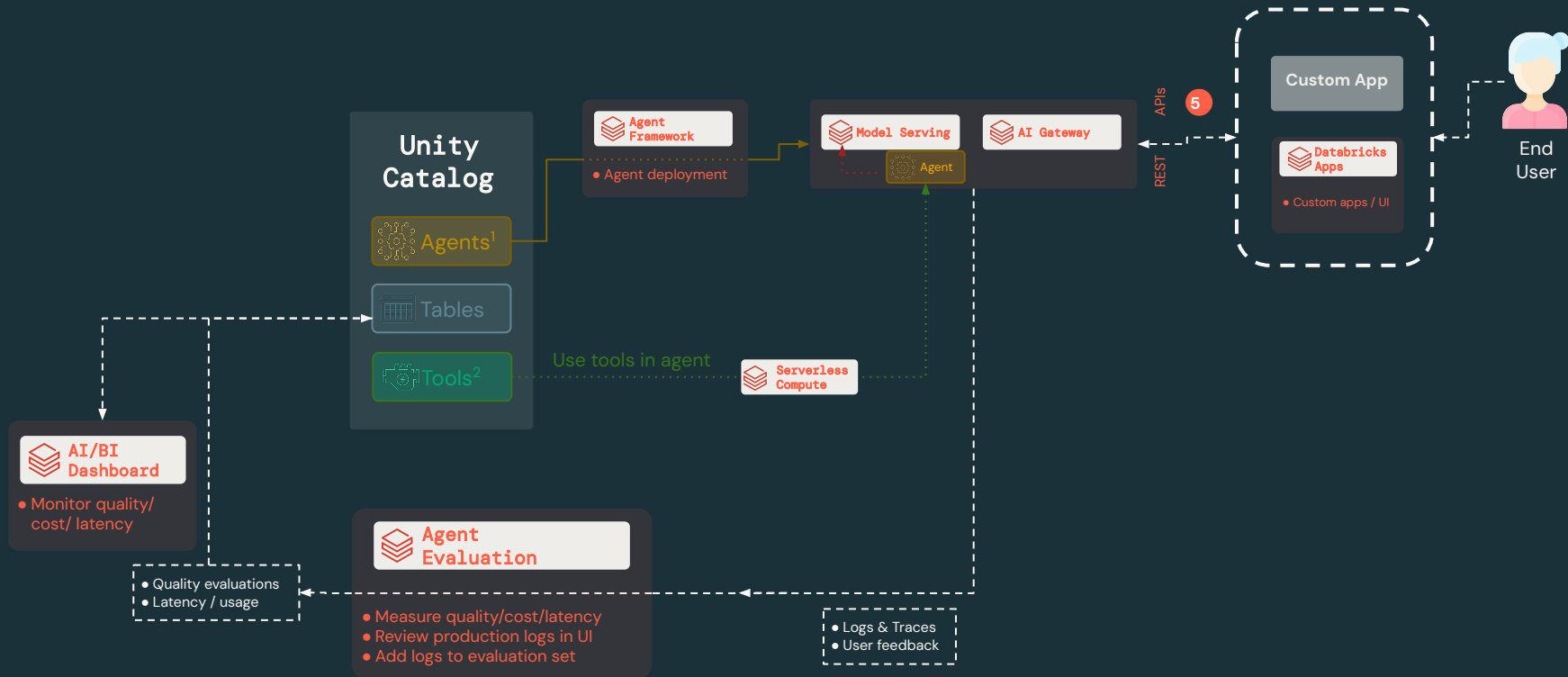


Customer challenges

Deploy to production & monitor quality

1. How do I host my Agent as a **production ready**, scalable service?
2. How do I execute tool code securely and ensure it respects my **governance policies**?
3. How do I enable **telemetry / observability** in development and production?
4. How do I **monitor my Agent's quality** at-scale in production?

Step 6: Release to production

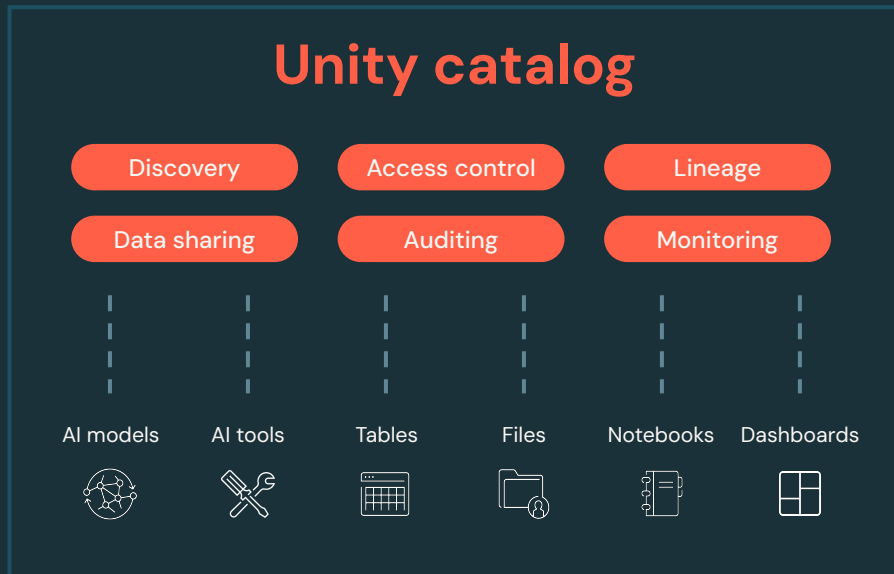


1. Stored as UC Models; 2. Stored as UC Functions

Agent Governance



Governance across data, models and tools



Unity Catalog + AI Agents & Governance

Discussion Points

1. What **governance challenges do you anticipate** when deploying AI agents that can access enterprise data?
2. How do you currently **manage access controls** for data that AI systems might need?
3. What **audit and lineage tracking** capabilities would you need for AI agents in production?

Let's Talk Use Cases



Agent Use Case Identification

Discussion Points

1. Have you or your team built AI agents before? **If yes, what use cases?**
2. What are the **biggest risks or limitations** you see with AI agents?
3. If an AI agent made decisions for your business, what would you need **to trust it?**



How Do I Find the FIRST Use Case?



Agent Use Case Identification

Discussion Points

1. What are some of the biggest pain points your team faces today?
2. Where do you see the most repetitive or time-consuming tasks?
3. What decisions would you want an AI agent to automate or support?
4. Where do you currently lose time or resources due to lack of automation or insight?



Agent Use Case Identification

Best Practices

1. **Customer-Centric Approach:** Start by interviewing stakeholders to uncover pain points.
2. **Mapping Out Processes:** Visualize workflows to identify inefficiencies and automation opportunities.
3. **Data Availability Check:** Before pursuing a use case, ensure data is accessible and of sufficient quality.

Ensuring Solutions Address Real Problems



Ensuring Solutions Address Real Problems

Best Practices

1. **User-Centric Design:** Involve end-users throughout development, not just at the end.
2. **Feedback Loops:** Build mechanisms for continuous feedback collection and monitoring.
3. **Iterative Testing:** Deploy in controlled environments before scaling to production.



Key Takeaways

01

Identify use cases –start with the simplest POC

02

Define success criteria with Agent Evaluation

03

Iterate Fast and Get Feedback

Thank you . . . !

