



Generative AI - *Data is your Differentiator*

Kumar Lakshmipathi
Generative AI Lead for Energy
Amazon Web Services

Generative AI is Transformative



NEW EXPERIENCES

Create new innovative and engaging ways of interacting with your customers and employees



PRODUCTIVITY

Radically improve productivity across all lines of business



INSIGHTS

Extract insights and clear answers from all your corporate information, enabling faster and better decisions



CREATIVITY

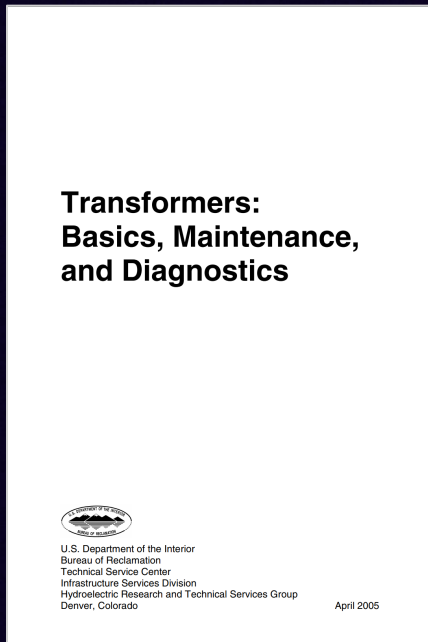
Create new content and ideas, including conversations, stories, images, videos, and music

Gen AI for Maintenance

Gen AI for Maintenance - Chatbot

Example

Step 1: Upload document to platform



256 Page PDF guide

Step 2: Ask it questions

I've found some burned paper near a transformer. Is this a problem?

How do I perform a DP test?

Step 3: Get the answer

It is unclear whether the burned paper outside the transformer is a problem without further information and testing. The document provides information on how to perform a DP test...

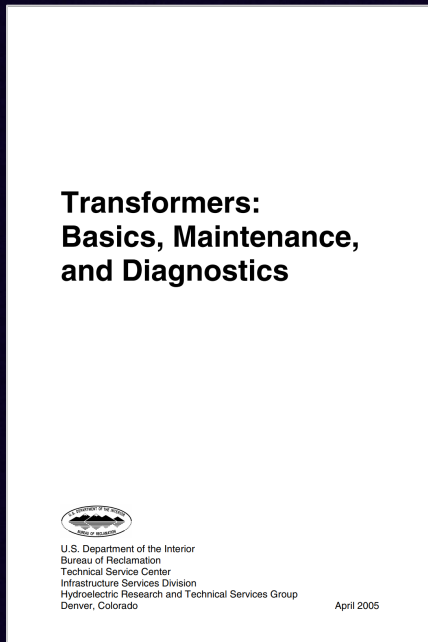
See: pages 138, 162, 197, 216

To perform a DP test, remove a sample of the paper insulation from a convenient location near the top of the center phase

See: page 243

Gen AI for Maintenance - IoT

Step 1: Upload document to platform



256 Page PDF guide

Step 2: Pass sensor data

```
{ "sensorData": {  
  "timestamp": 1632578627,  
  "motorId": "a1b2c3",  
  "rpm": 2500,  
  "temperature": 150,  
  "voltage": 12.5,  
  "amperage": 25,  
} }
```

Step 3: Get the Analysis

RPM of 2500 is higher than the rated speed of 1445 rpm

See: page 243

Gen AI for Maintenance - Images

Model: Claude 3 Sonnet



KL do you see any rust or cracks or freeze on the pipes in the image?

Yes, upon closer inspection, there appears to be a crack or fracture visible on one of the pipe segments or valve components in the image.

Copy Retry Like Comment

Gen AI for Maintenance - Video



What time did the maintenance trucks arrive?

The trucks arrived late and only got here at 6:45 PM

Gen AI for Maintenance - Agents



Monitor my RS Pro 3 Motor and text the on-call person if there is a problem. Get their confirmation before ordering parts, scheduling maintenance or announcing downtime

ok

A screenshot of a web browser window displaying a Gen AI agent interface. The browser has a single tab with a search bar. The page content includes a header with instructions, followed by two columns of actions: Assistant Actions and Utility Actions. The Assistant Actions column lists 'Find Motor Specs' and 'Find On-Call Team'. The Utility Actions column lists 'Send Email' and 'Order Part'. Each action specifies its input and output types.

Instructions: Your task is to detect anomalies in...

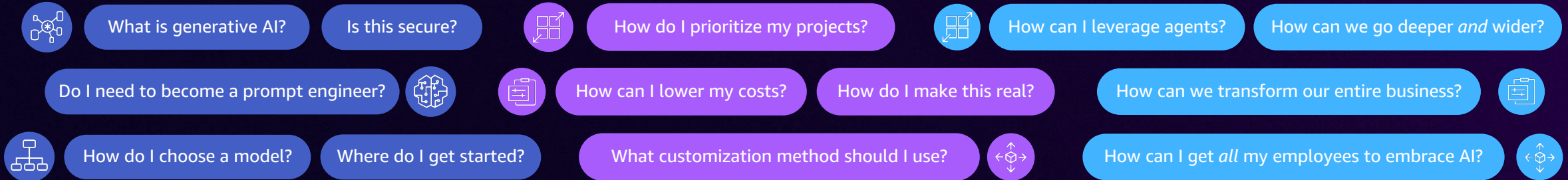
ASSISTANT ACTIONS	UTILITY ACTIONS
<u>Find Motor Specs</u> in: equipment Id out: PDFs	<u>Send Email</u> in: names out: email
<u>Find On-Call Team</u> in: time out: names	<u>Order Part</u> in: partID out: status

Gen AI questions over the years

2023
The year of POCs

2024
The year of Production

2025
The year of Business Value





Generative AI Application



Generative AI
Application

Data Foundation

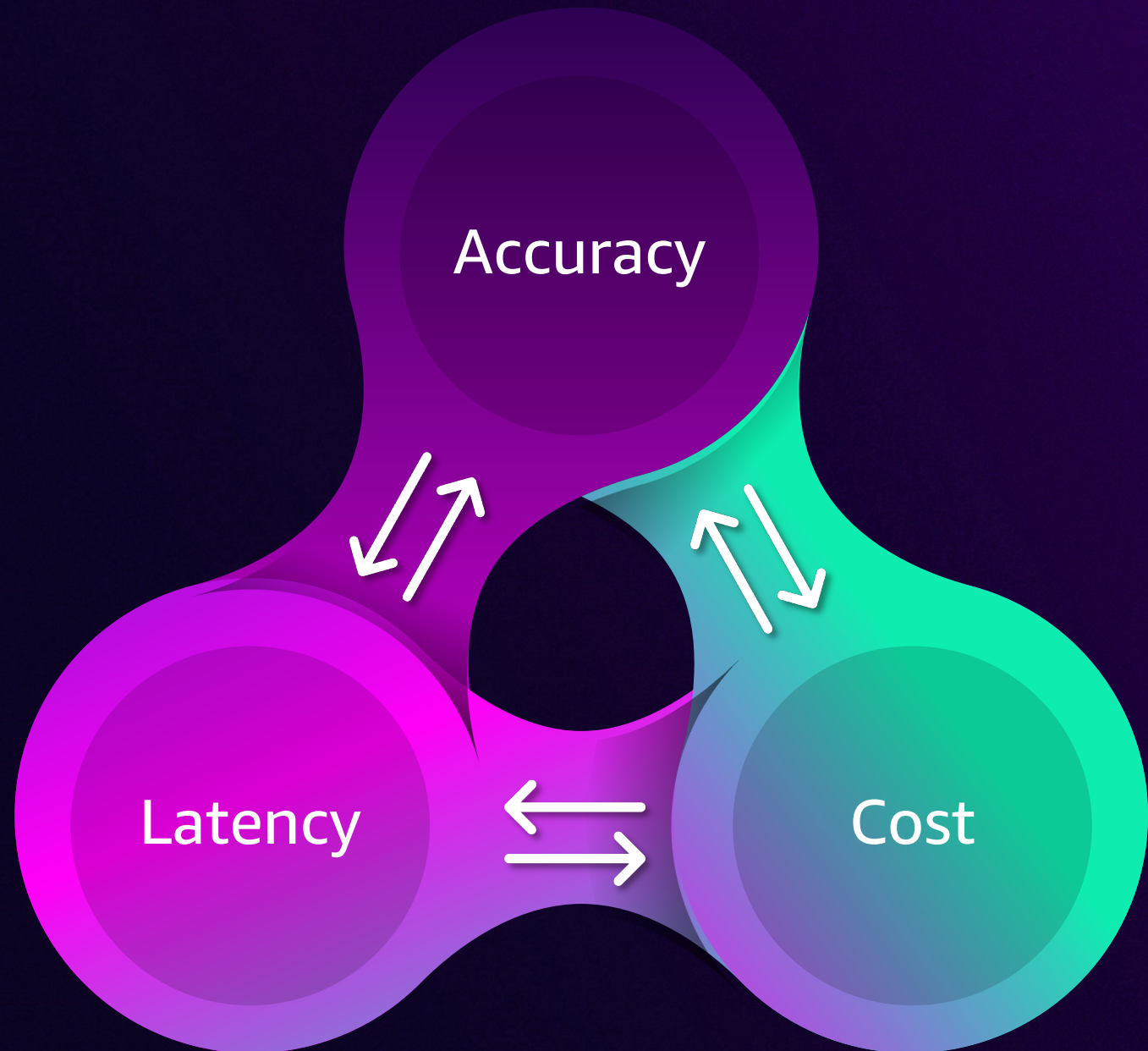
STORAGE

GOVERNANCE
& COMPLIANCE

DATABASES,
ANALYTICS,
& DATA LAKES




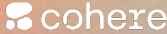
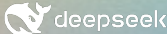





DATA
INTEGRATION

**Yes, model
selection is
important**



Amazon Bedrock

BROAD CHOICE OF MODELS

 Effective reasoning & rapid analysis for long context windows JAMBA	 Frontier intelligence & industry leading price performance, high-quality image & video generation AMAZON NOVA	 Advanced reasoning & coding capabilities, including computer use skills CLAUDE	 Multimodal search & advanced retrieval powering multilingual knowledge agents COMMAND EMBED RERANK	 Advanced reasoning, comprehension, scientific & mathematical analysis, & code generation DEEPSEEK-R1	 High-quality video generation from text & images LUMA RAY 2	 Advanced image & language reasoning LLAMA	 Knowledge summarization, expert agents, & code completion MISTRAL MIXTRAL	 Software engineering AI for large enterprises MALIBU POINT	 High-quality AI image generation, easily deployable at scale STABLE DIFFUSION STABLE IMAGE
---	---	--	---	--	---	---	--	---	---

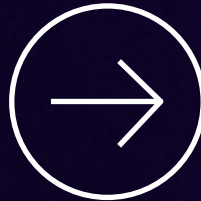
Coming soon

Amazon Bedrock Marketplace enables developers to discover, test, and use over 100 popular, emerging, and specialized foundation models (FMs) alongside the current selection of industry-leading models in Amazon Bedrock.

But your data is your *differentiator*



Generic
generative AI



Generative AI that
knows your business
and your customers

How to provide your data to gen AI applications?

RAG

Guide foundational models by prompting with contextually relevant data (RAG)

Fine-tuning a Foundation Model

Fine-tune a foundation model using your curated, labeled data

Training your own purpose-built LLM

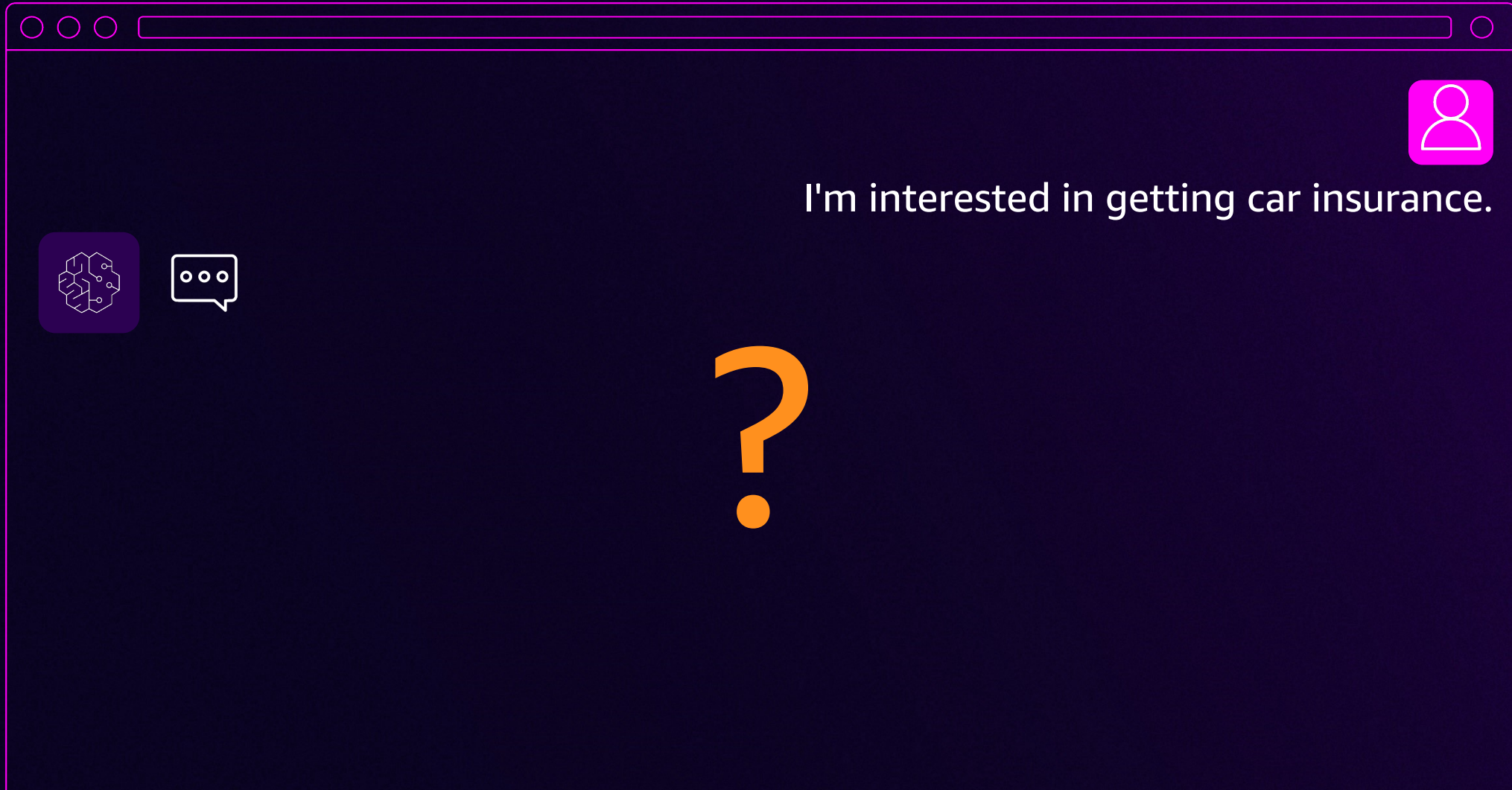
Train a foundation model using your curated, specialized data

Pat needs Car Insurance

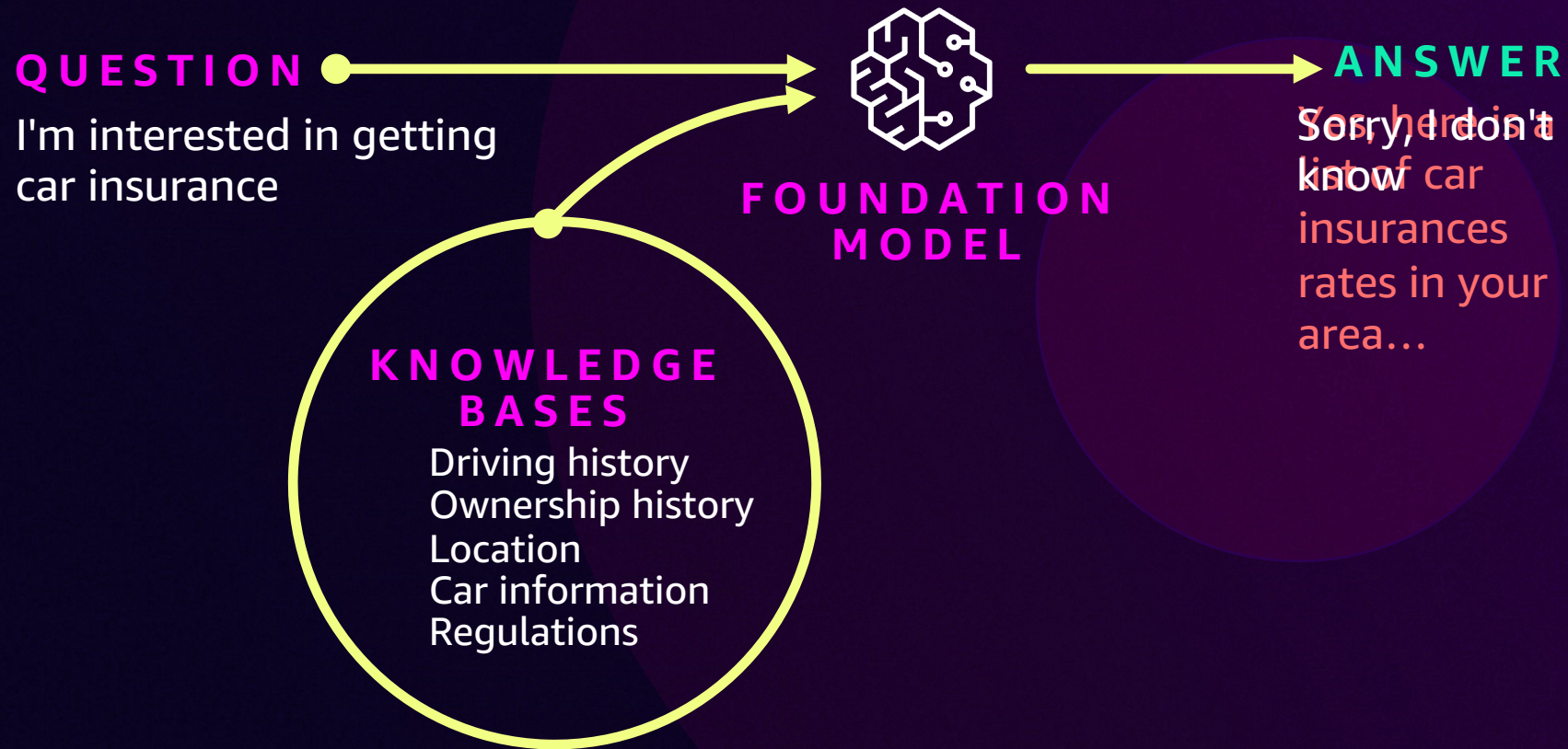
- Pat lives in Nevada and bought a new car!
- This is the 3rd car Pat has owned; Pat sold his previous cars
- **Pat needs car insurance**
- Pat wants to compare different rates and coverages



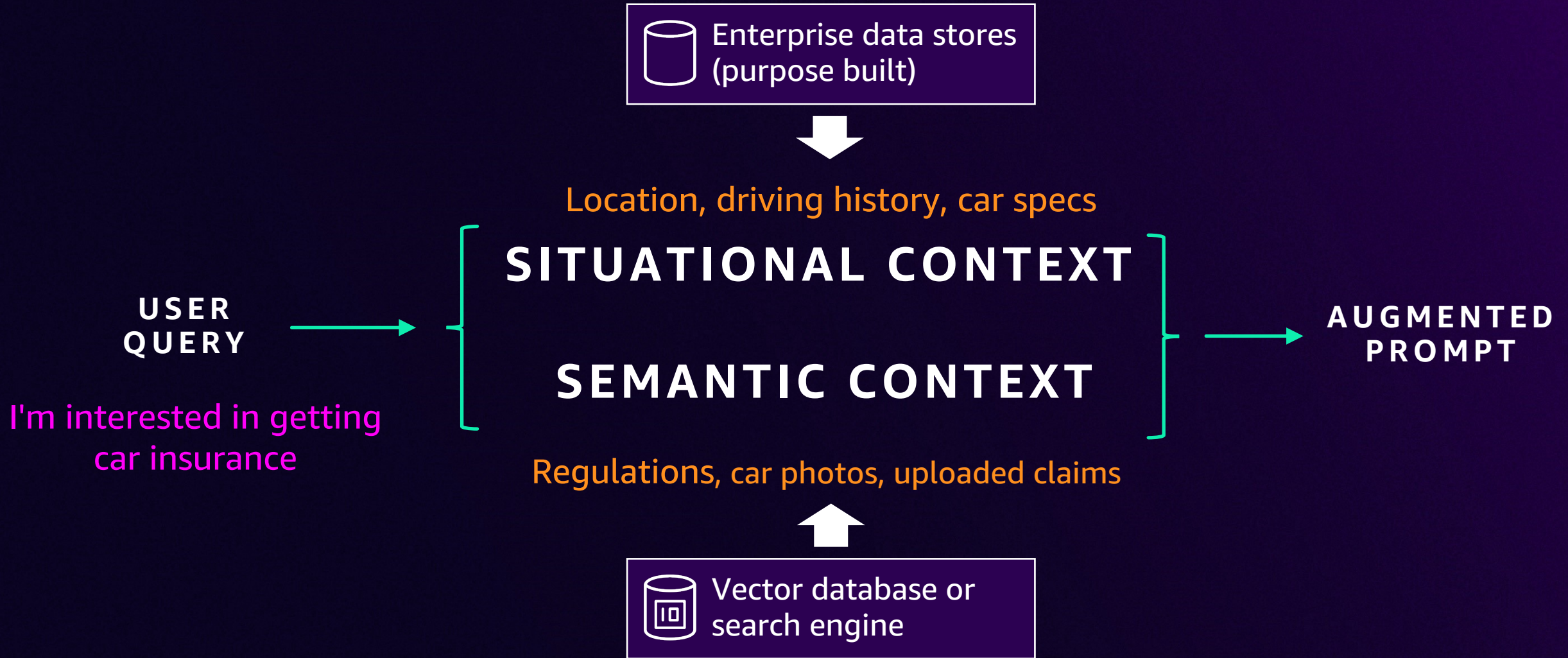
Pat needs car insurance



Retrieval Augmented Generation (RAG)



What type of data do you need for RAG?



Designing an augmented prompt

1. Instructions for the foundation model
2. Situational context
3. Semantic context
4. User input

Building an augmented prompt

User input

"I'm interested in getting car insurance."

Prompt engineering

You are a conversation agent for the HorizonGuard insurance marketplace. Responses to questions are written to be helpful and inform in an unbiased manner. Ask human to confirm responses.

Instructions for the model

Building an augmented prompt

User input

"I'm interested in getting car insurance."

Prompt engineering

You are a conversation agent for the HorizonGuard insurance marketplace. Responses to questions are written to be helpful and inform in an unbiased manner. Ask human to confirm responses.

The human is Pat, lives at 123 Noname St, Anytown, NV, USA and owns a 2024 SUV. Household members include 1 adult and 1 child of legal driving age.

Instructions for the model

Situational context

Building an augmented prompt

User input

"I'm interested in getting car insurance."

Prompt engineering

You are a conversation agent for the HorizonGuard insurance marketplace. Responses to questions are written to be helpful and inform in a unbiased manner. Ask human to confirm responses.

The human is Pat, lives at 123 Noname St, Anytown, NV, USA and owns a 2024 SUV. Household members include 1 adult and 1 child of legal driving age.

Car insurance is one of the most used types of insurance. Most states require purchase of some kind of coverage to legally drive. [...]. Reference: <https://www.....> [...]

Instructions for the model

Situational context

Semantic context

Building an augmented prompt

User input

Prompt engineering

"I'm interested in getting car insurance."

You are a conversation agent for the HorizonGuard insurance marketplace. Responses to questions are written to be helpful and inform in a unbiased manner. Ask human to confirm responses.

The human is Pat, lives at 123 Noname St, Anytown, NV, USA and owns a 2024 SUV. Household members include 1 adult and 1 child of legal driving age.

Auto insurance is one of the most used types of personal insurance. Most states require that you purchase some kind of insurance coverage to drive legally in the state. [...]. Reference: <https://content.naic.org/consumer/auto-insurance.htm>

[...]

I'm interested in getting car insurance.

Instructions for the model

Situational context

Semantic context

User prompt/question

Building an augmented prompt

User input

"I'm interested in getting car insurance."

Prompt augmentation

The human is Pat, lives at 123 Noname St, Anytown, NV, USA and owns a 2024 SUV. Household members include 1 adult and 1 child of legal driving age.

Auto insurance is one of the most used types of personal insurance. Most states require that you purchase some kind of insurance coverage to drive legally in the state. [...]. Reference: <https://content.naic.org/consumer/auto-insurance.htm>

[...]

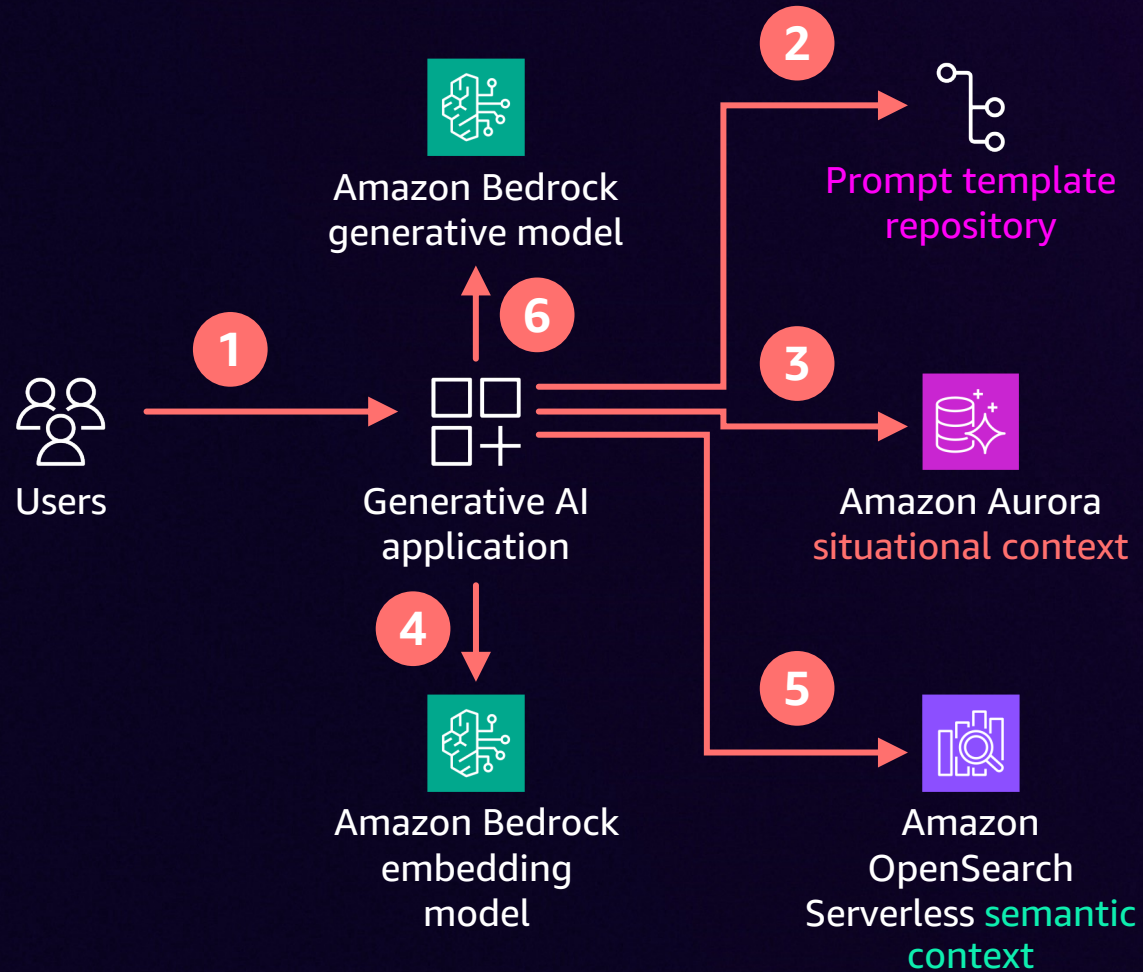
```
SELECT
  addresses.*, cars.*, dependents.*
FROM cars
JOIN addresses ON
  addresses.user_id = $1
JOIN dependents ON
  dependents.user_id = $1
WHERE cars.user_id = $1;
```

Situational context

Semantic context

```
POST /insurance-index/_search
{
  "size": 10,
  "query": {
    "knn": {
      "location": {
        "vector": [... ],
        "k": 10
      }
    }
  }
}
```


Basic Architecture



- 1** User asks question
- 2** Retrieve **prompt template** (optional, if not baked into app)
- 3** Retrieve **situational context** from enterprise database
- 4** Use embedding model to create vector for question
- 5** Perform vector search and retrieve **semantic context**
- 6** Submit augmented prompt to generative model; get response

Impact of augmented prompt on the experience

User input

"I'm interested in getting car insurance."

Prompt engineering

You are a conversation agent for the HorizonGuard insurance marketplace. Responses to questions are written to be helpful and inform in a unbiased manner. Ask human to confirm responses.

The human is Pat, lives at 123 Noname St, Anytown, NV, USA and owns a 2024 SUV. Household members include 1 adult and 1 child of legal driving age.

Auto insurance is one of the most used types of personal insurance. Most states require that you purchase some kind of insurance coverage to drive legally in the state. [...]. Reference: <https://content.naic.org/consumer/auto-insurance.htm>

[...]

I'm interested in getting car insurance.

Foundation model response

Good morning, Pat, I'd be happy to help you with car insurance. I have your car details on record. You drive a 2024 SUV and live at 123 Noname St, Anytown, NV, USA. Is that correct?

Impact of augmented prompt on the experience

User input

"Yes, that's correct."

Prompt engineering

You are a conversation agent for the HorizonGuard insurance marketplace. Responses to questions are written to be helpful and inform in a unbiased manner. Ask human to confirm responses.

The human is Pat, lives at 123 Noname St, Anytown, NV, USA and owns a 2024 SUV. Household members include 1 adult and 1 child of legal driving age.

Auto insurance is one of the most used types of personal insurance. Most states require that you purchase some kind of insurance coverage to drive legally in the state. [...]. Reference: <https://content.naic.org/consumer/auto-insurance.htm>

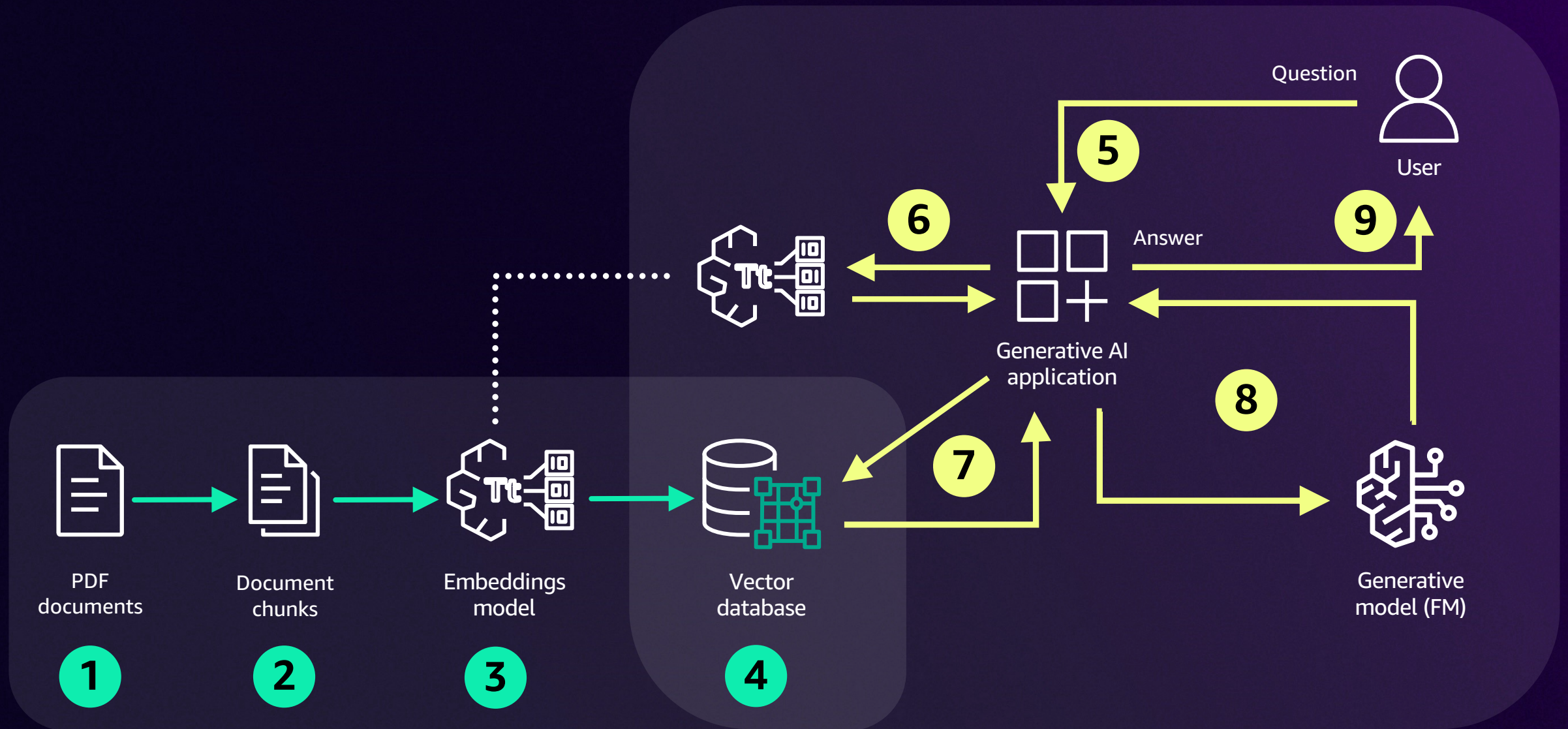
[...]

I'm interested in getting car insurance.

Foundation model response

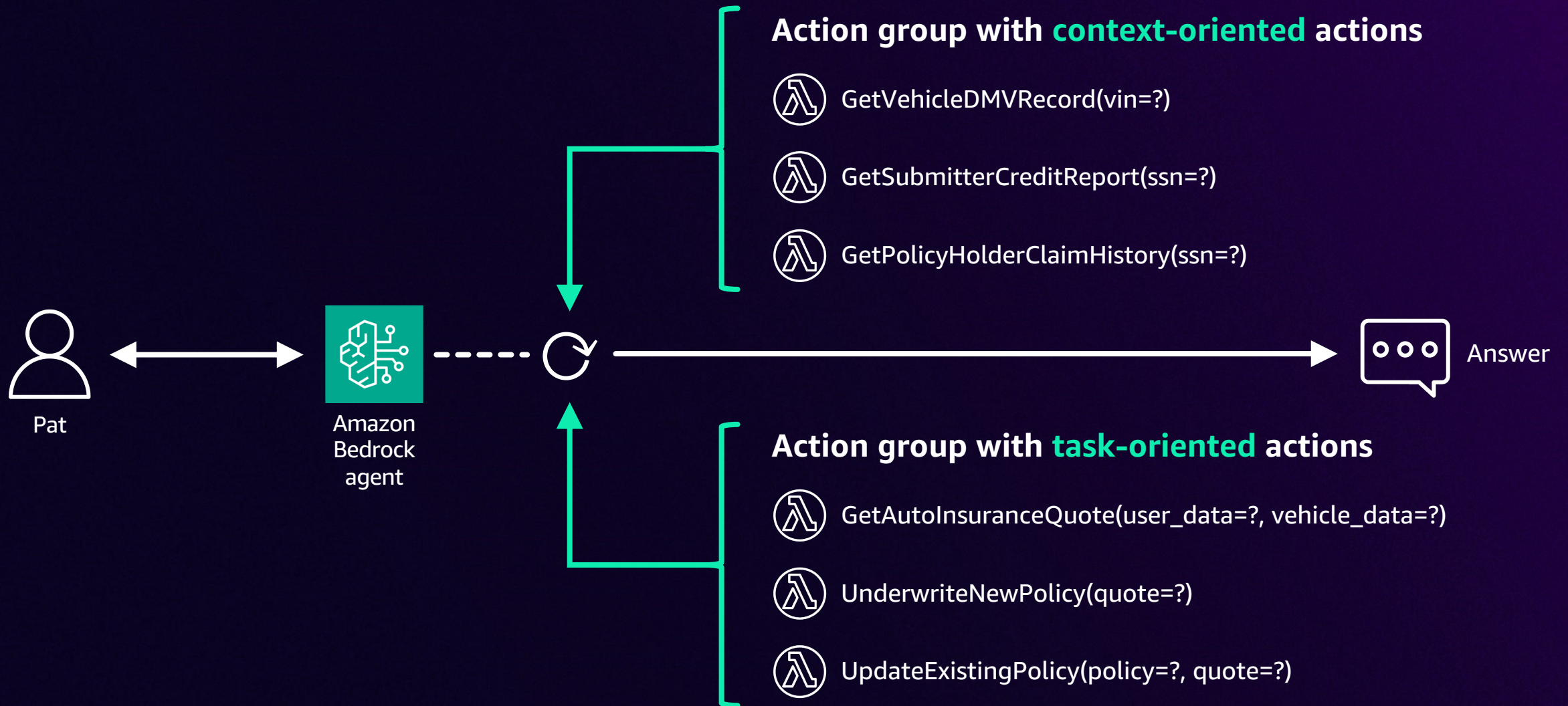
Car insurance in your area costs between \$200 and \$250 per month. Do you want more information?

RAG - Adding Semantic Context



What about Agents?

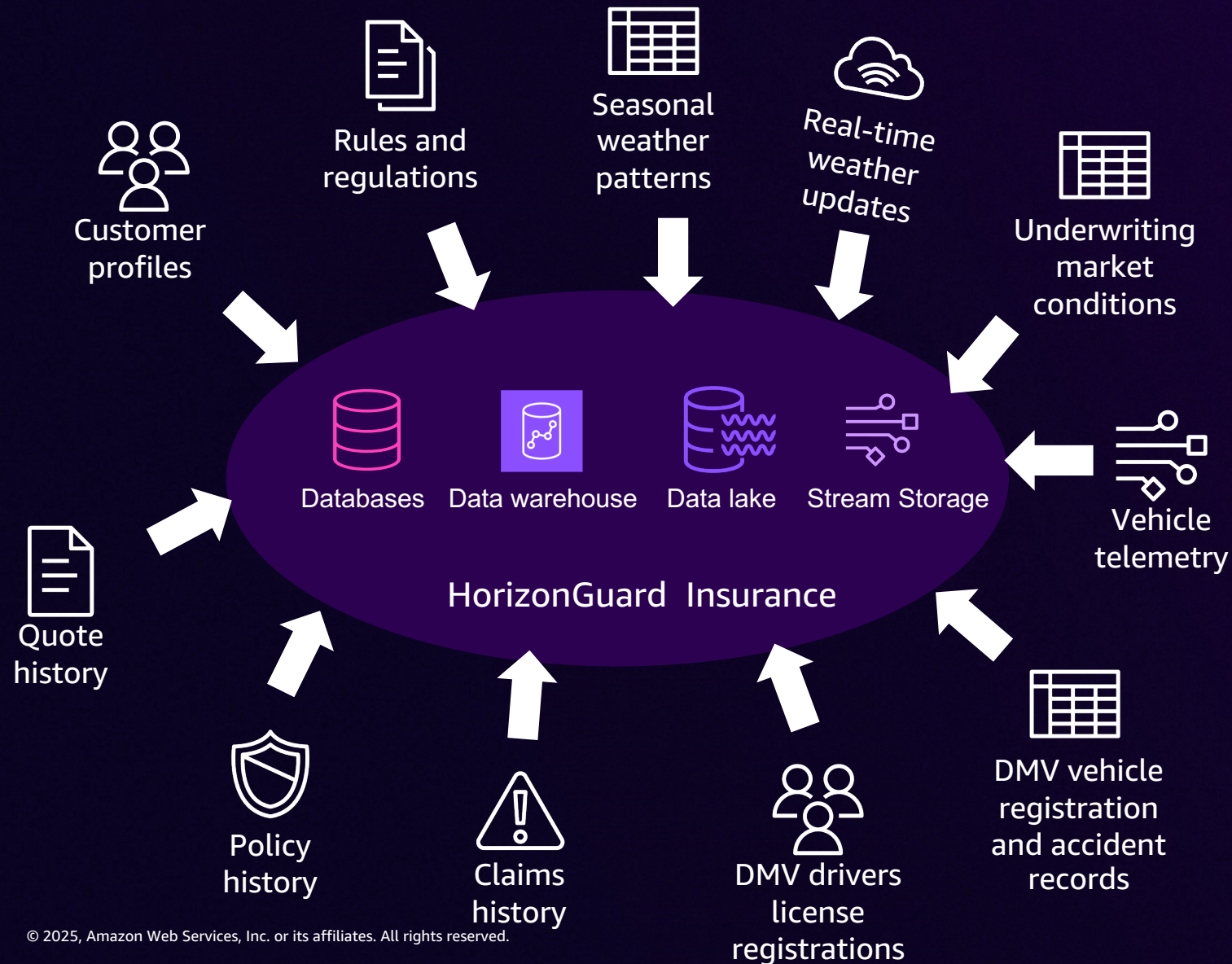
Continuing the Car Insurance example



Establishing a GenAI Data Foundation [RAG Centric]



Understand where your RAG Context comes from



Challenges:

- Data silos
- Lineage and currency
- Data quality
- PII
- Identifier resolution
- Access controls

Types of Source Data



Structured

Existing database

Defined meaning,
schema, and query
patterns



Semi-structured

May not be under
management

JSON, XML,
STRUCT



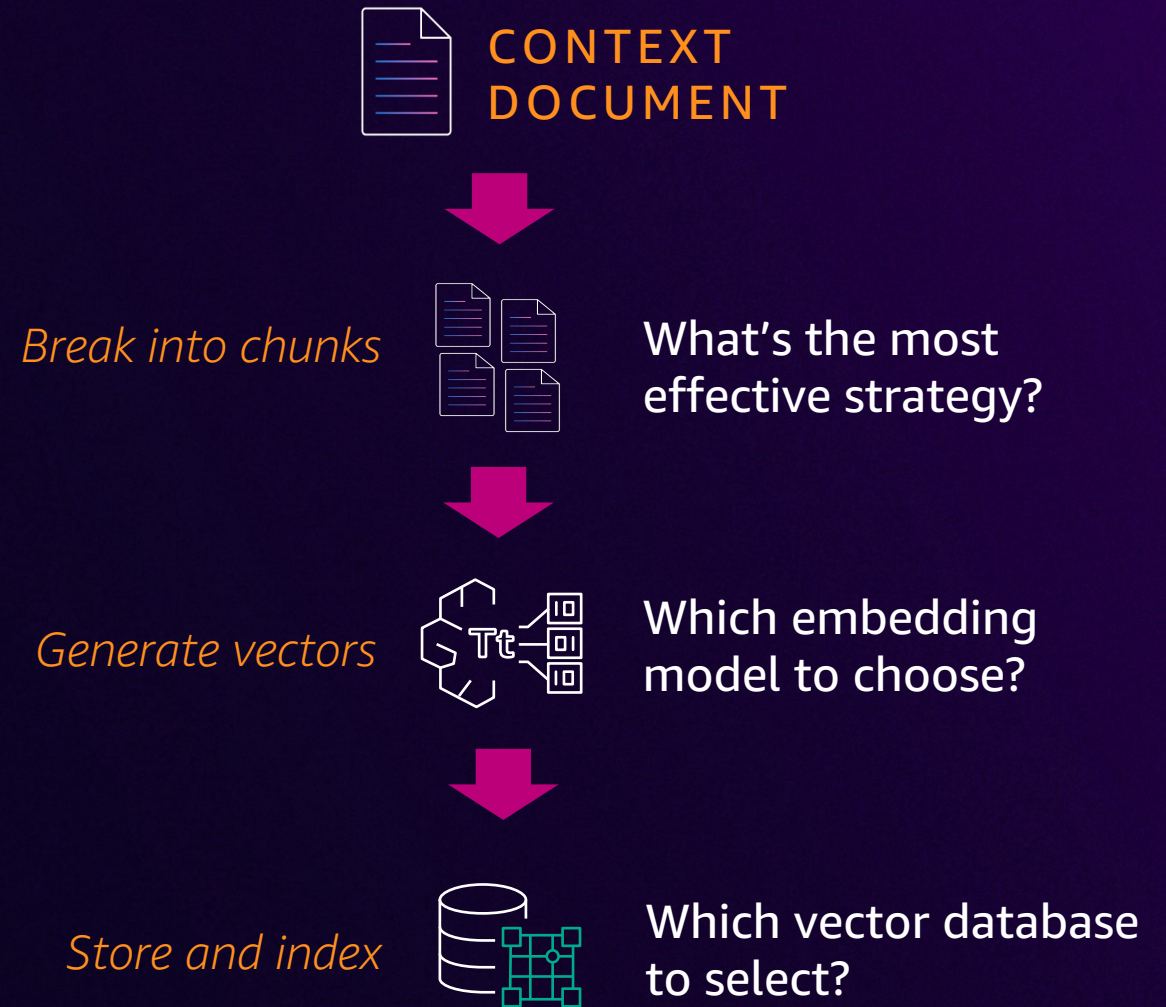
Unstructured

Amazon S3, content
management systems

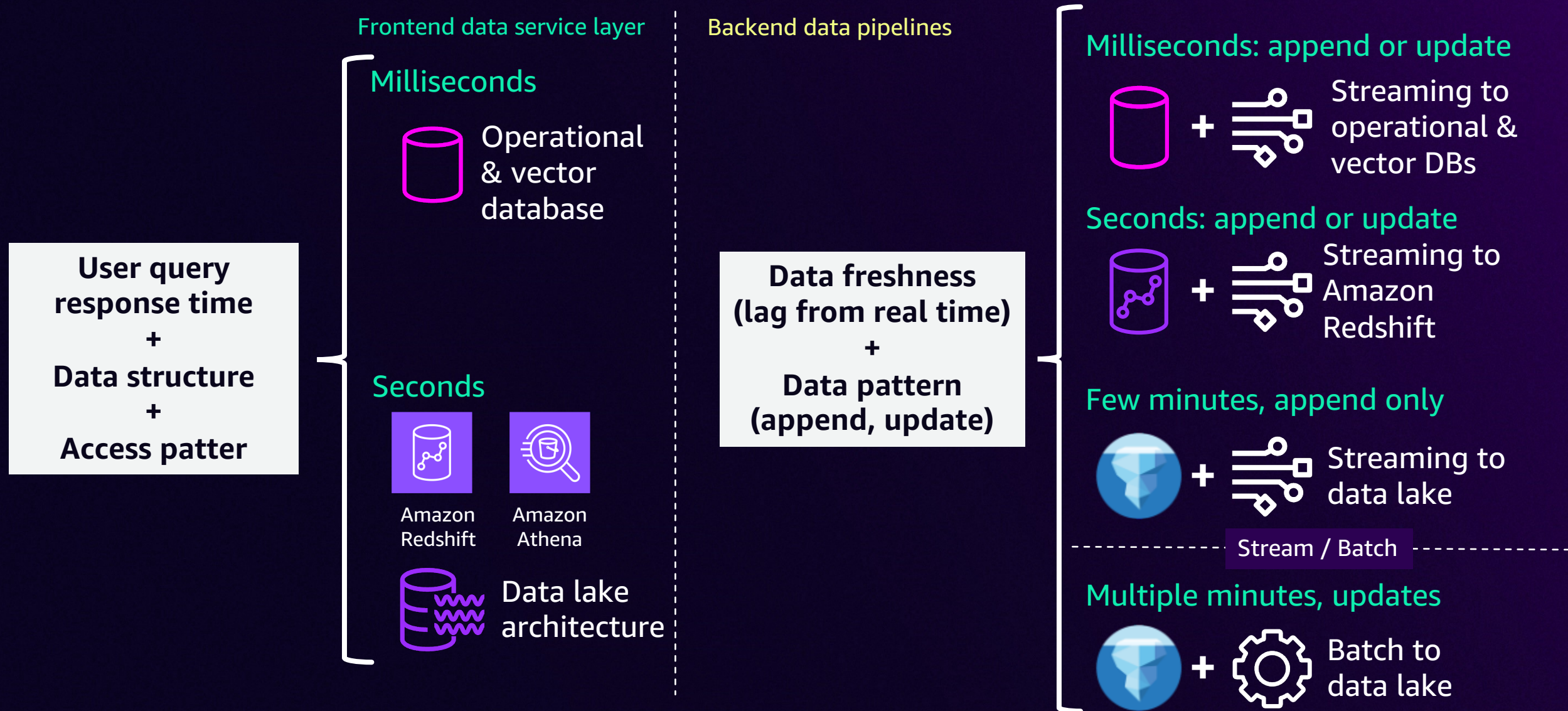
May not be labeled

Enable Unstructured Data for RAG Applications

- Data transform is a must
- Understand data quality
- Define chunking strategy
- Generate embeddings
- Create a pipeline



Choose the right data architecture for RAG



Address Data Quality and Lineage

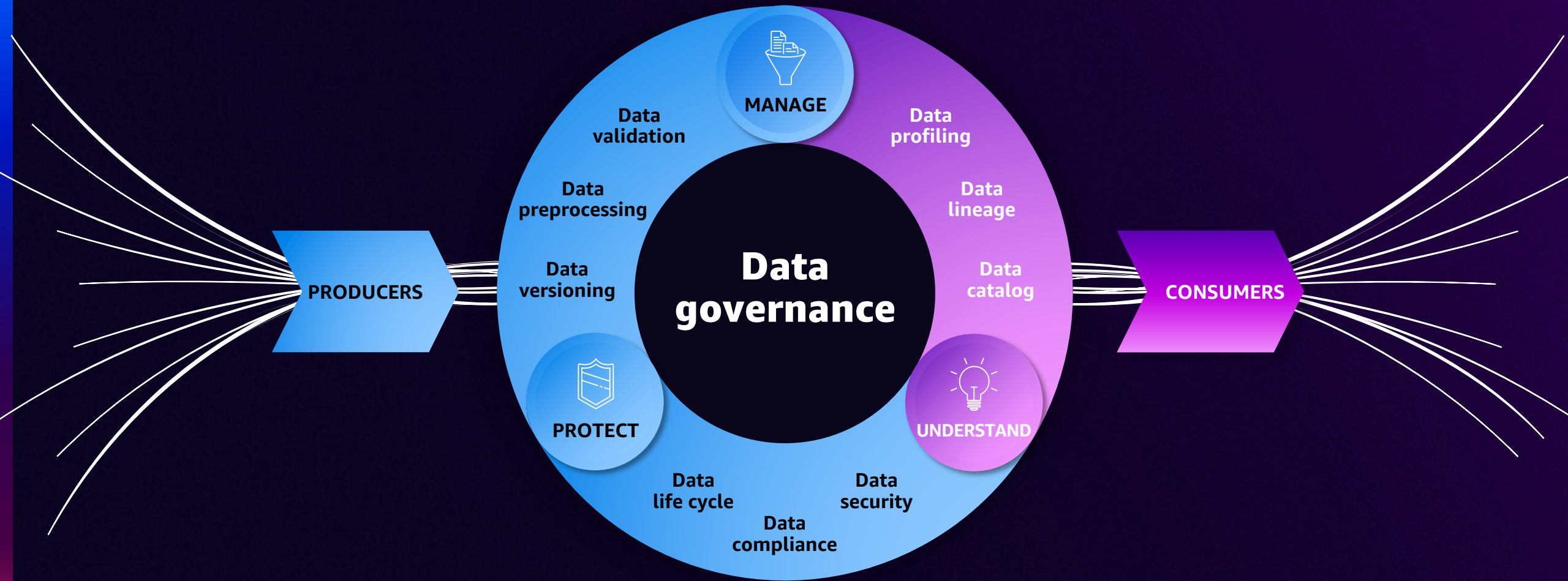
TECHNICAL CHALLENGES

- Data format
(e.g. email, phone number, policy ID)
- Data values
(e.g. valid ranges, completeness)
- Lack of trust and transparency
(e.g. Where did this data come from?)

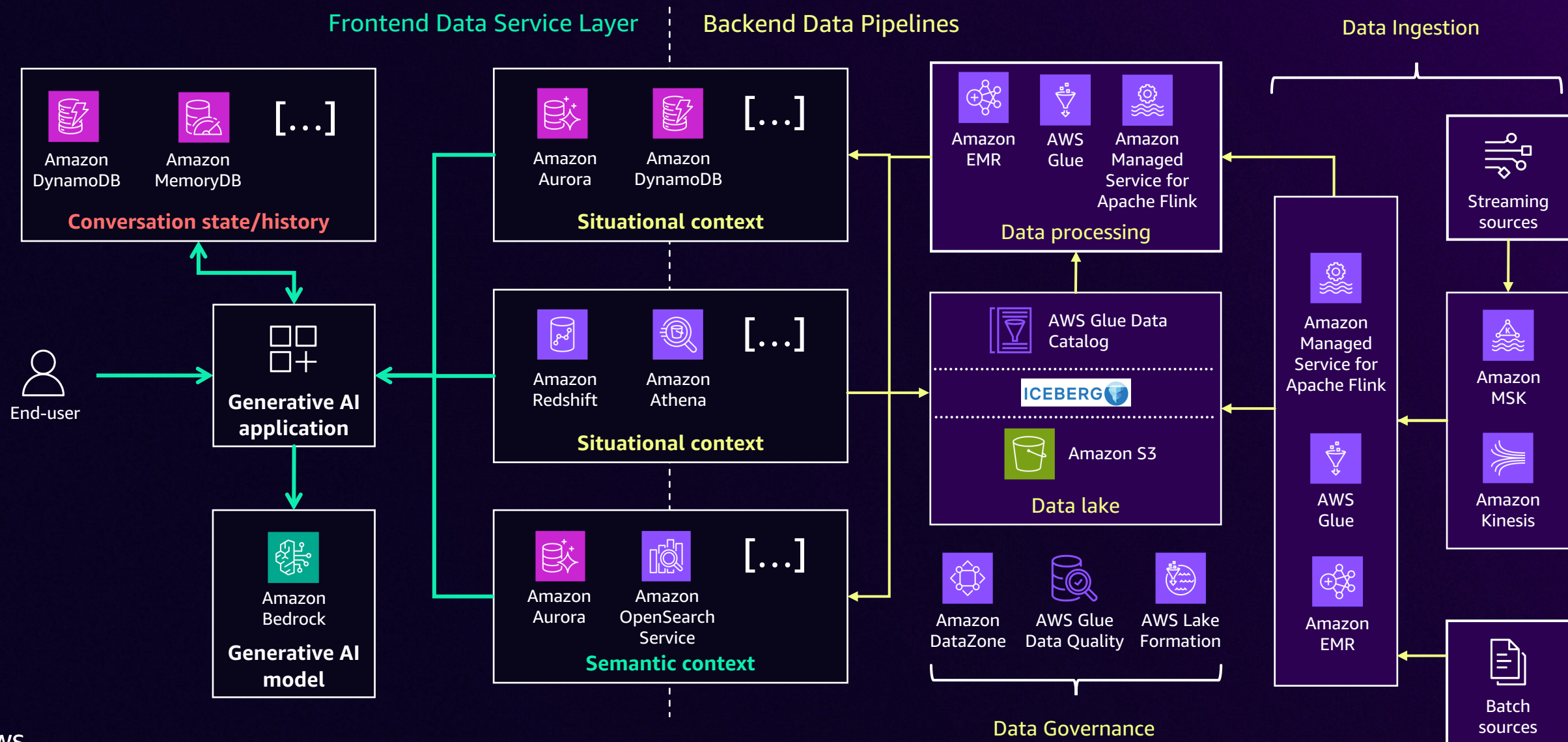
BUSINESS CHALLENGES

- Meeting business rules
(e.g. claim dates, policy limits)
- Regulatory compliance
(e.g. minimum liability insurance limits)
- Building transparency and trust

Establish Data Governance



Putting it all together



Adopt Big Data Best Practices



Build decoupled systems

- Data → Store → Process → Store → Analyze → Answers

Use the right tool for the job

- Data structure, latency, throughput, access patterns

Leverage managed and serverless services

- Scalable/elastic, available, reliable, secure, no/low admin

Use log-centric design patterns

- Immutable logs (in a data lake), and materialized views

Be cost-conscious

- Big Data ≠ big cost

Get cracking with Gen AI

01

Get your Data
Foundation
Right

02

Empower your teams with
tools and trainings

03

Get started on
high value use
cases



Rethink work

A full-page background image showing a person in a red jacket and blue backpack standing on a dark, rocky outcrop in the foreground. They are looking up at a massive, jagged mountain peak covered in snow and ice. The mountain's peak is illuminated by a warm, golden light, possibly from the sun low on the horizon, creating a strong contrast with the cool blue tones of the sky and the shadows on the mountain. The sky is filled with soft, wispy clouds. The overall mood is one of awe and ambition.

Think big.



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Thank you!

Kumar Lakshmipathi
laksku@amazon.com

