

AWS State, Local, and Education Learning Days

Washington, D.C.



BREAKOUT SESSION

Data foundations in the era of Generative AI

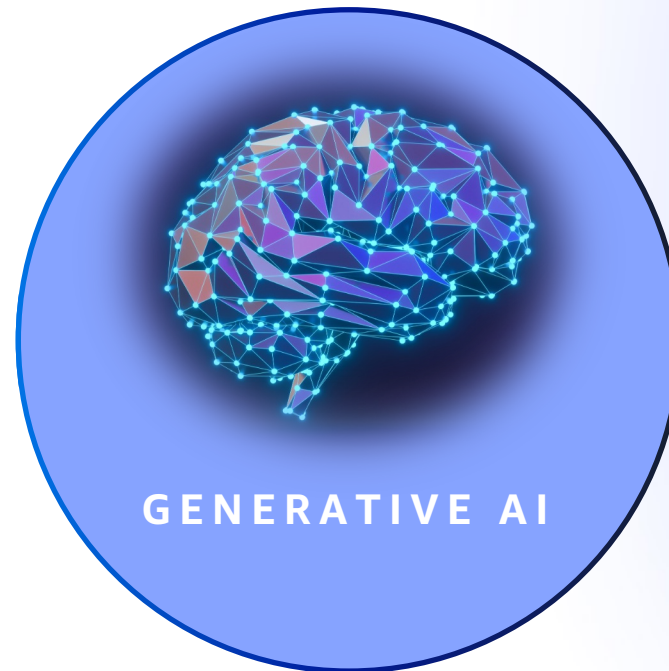
ANKIT BATTAWAR (he/him)

Solution Architect
Amazon Web Services
battawar@amazon.com

GOVIND PALANISAMY (he/him)

Sr Solution Architect
Amazon Web Services
govpal@amazon.com

Innovation can transform industries

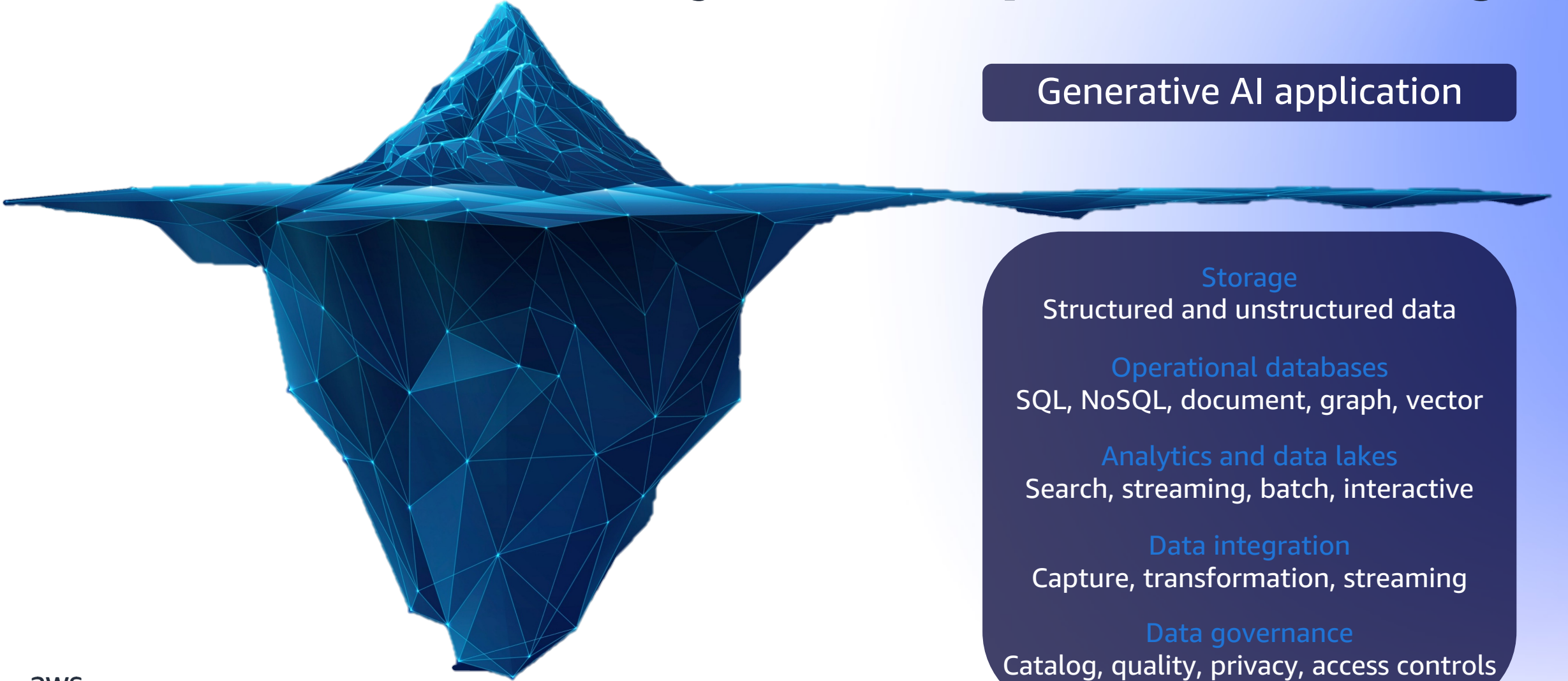


Generative AI application



Generative AI application

But business value is just the tip of the iceberg



Generative AI application

Storage

Structured and unstructured data

Operational databases

SQL, NoSQL, document, graph, vector

Analytics and data lakes

Search, streaming, batch, interactive

Data integration

Capture, transformation, streaming

Data governance

Catalog, quality, privacy, access controls

What every CEO should know about generative AI

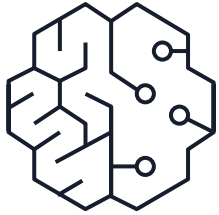
May 12, 2023 | Article

“... the lifeblood of generative AI is **fluid access to data** honed for a specific business context or problem. Companies that have not yet found ways to effectively harmonize and provide **ready access to their data** will be unable to fine-tune generative AI to unlock more of its potentially transformative uses . . . A **clear data and infrastructure strategy** anchored on the **business value** and competitive advantage derived from generative AI will be critical.”

Source: McKinsey & Company, “What every CEO should know about generative AI,” May 2023: <http://tinyurl.com/4v28a4f9>.



Your data is the differentiator



**Generic
generative AI**



**Generative AI that knows your
organization and your end user**

Automated budgeting and resource allocation

The objective

Analyze spending data to identify areas where resources can be allocated more efficiently

The data

 Enterprise Systems and Databases

 Internal guides

 Spreadsheets



More relevant conversational search

The objective

Provide answers to employees based on enterprise data

The data

 Slack

 Internal guides

 SharePoint



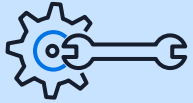
Use an
existing FM

OR

Train a
new FM?

Customizing foundation models

1



Retrieval Augmented Generation (RAG)

Guide pre-trained models with private, domain-specific contextual data

For example, virtual agents with limited domain-specific requirements

2

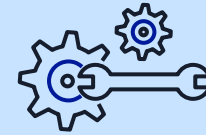


Fine-tune a pre-trained model

Specialized knowledge for specific tasks with labeled examples

For example, domain-intensive knowledge agents

3



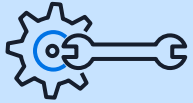
Continued pre-training

Generalized and specialized knowledge for your domain with unlabeled data

For example, deep domain-specific applications trained on specific data

Customizing foundation models

1



Retrieval Augmented Generation (RAG)

Guide pre-trained models with private, domain-specific contextual data

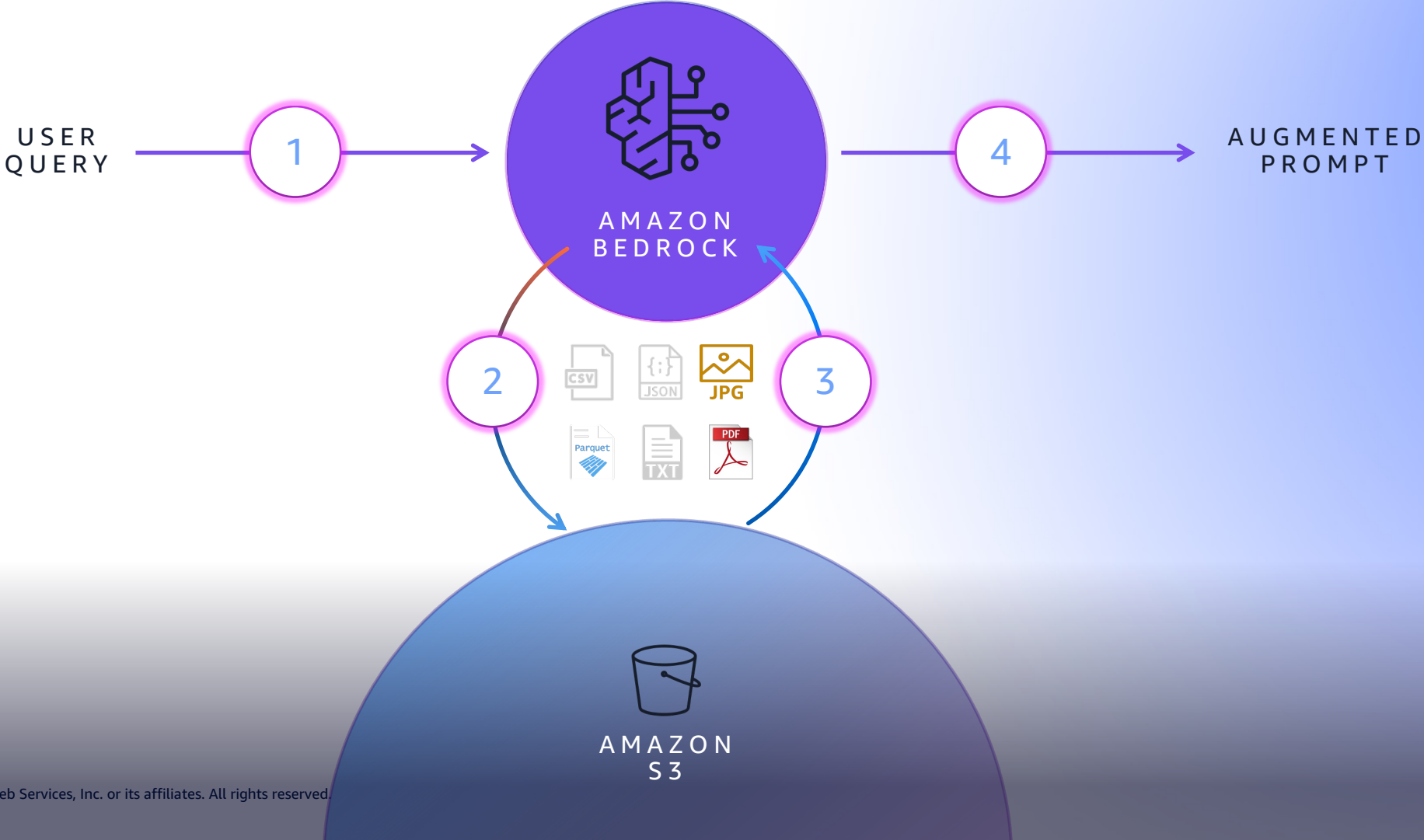
For example, virtual agents with limited domain-specific requirements

Retrieval augmented generation (RAG)

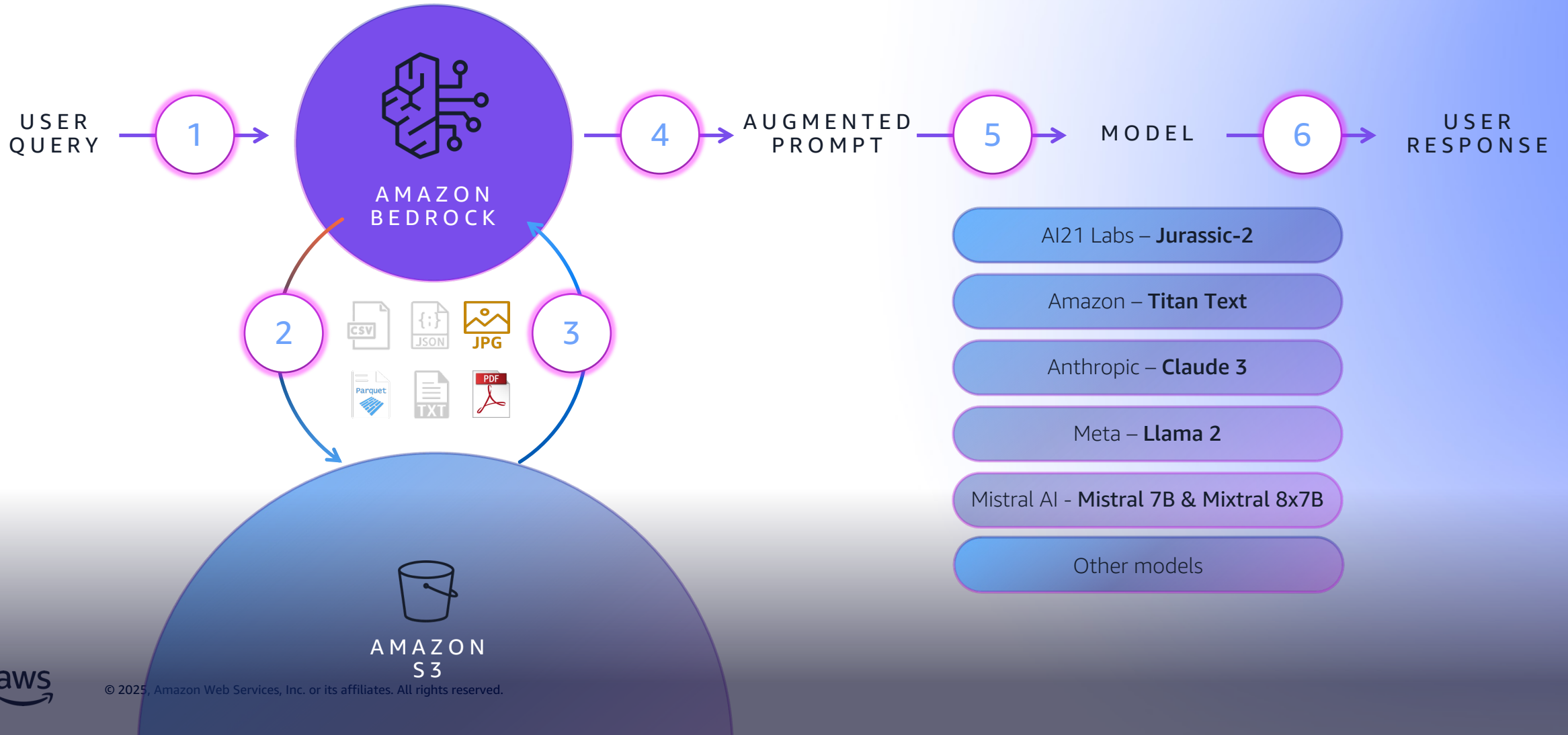


AMAZON
S3

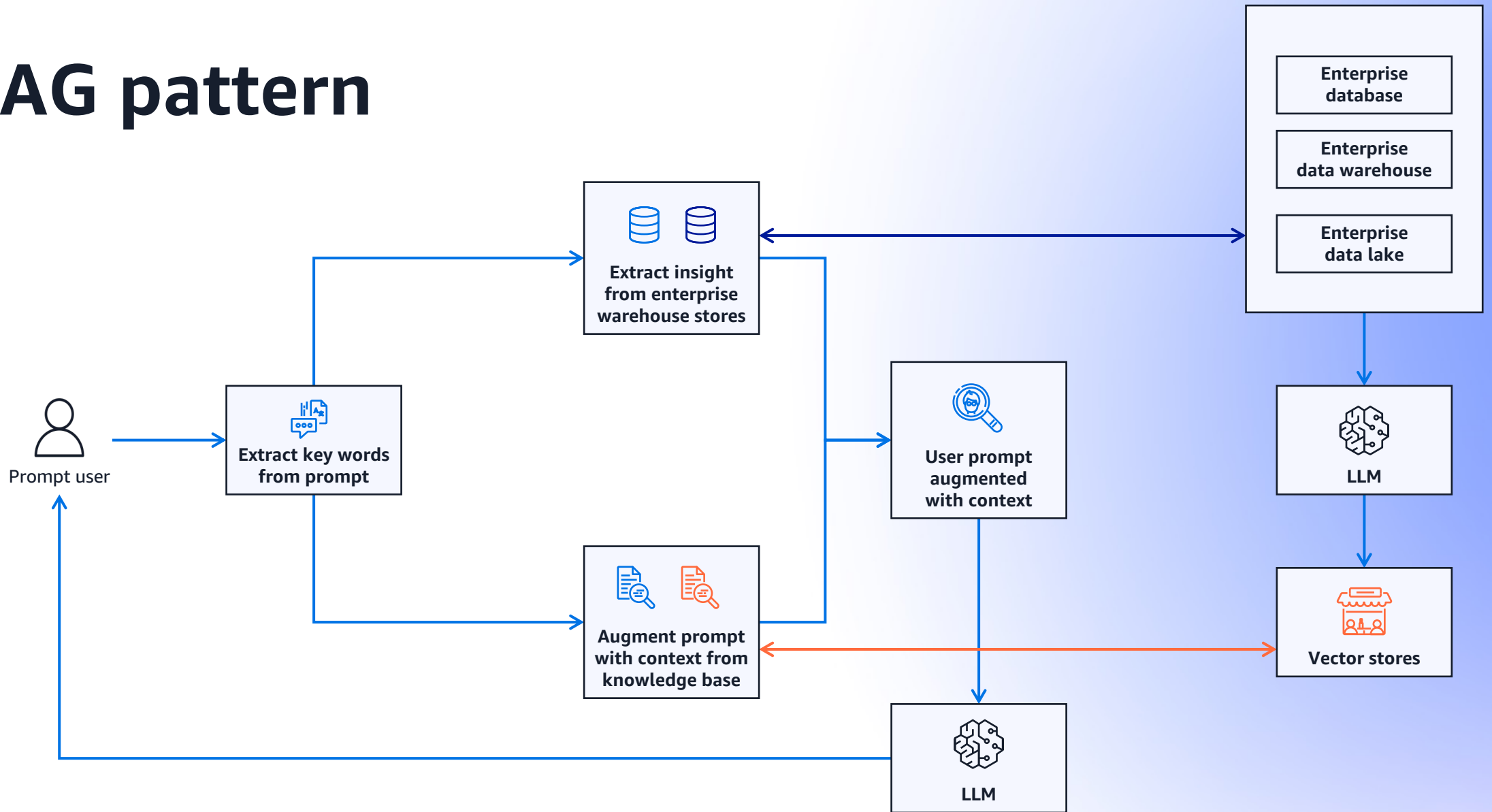
Retrieval augmented generation (RAG)



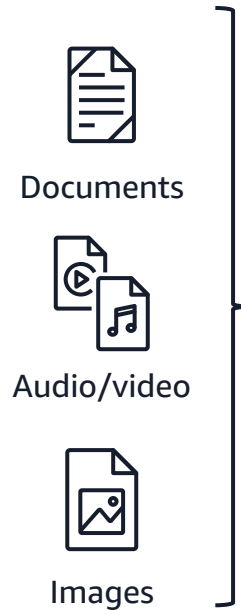
Retrieval augmented generation (RAG)



A RAG pattern



What are vector embeddings?



Semantic elements:

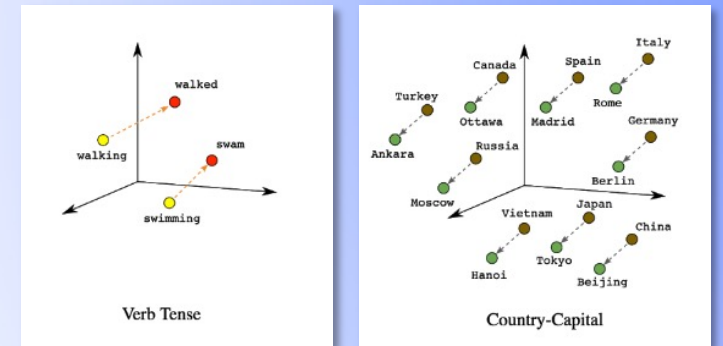
- Words, phrases
- Paragraphs, documents
- Scenes, song sections
- Faces, detected picture elements
- And more



0.35 0.1 0 0.9 001.0 00 0001.0 0 0...


0.35 0.1 0 0.8 001.0 00 0001.0 0 0...


0.15 0.1 0 0.7 001.0 00 0001.0 0 0...




3D simplified representation. Embeddings can have thousands of dimensions. Source: <https://daleonai.com/embeddings-explained>

Embeddings: When vector elements are semantic, used in generative AI


Amazon
OpenSearch Service 

Amazon
OpenSearch Serverless 


Amazon Aurora
PostgreSQL 


Amazon RDS for
PostgreSQL 

Enabling vector search across AWS services

 Amazon
DocumentDB

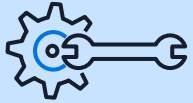
 Amazon DynamoDB
via zero ETL

 Amazon MemoryDB
for Redis

 Amazon Neptune

Customizing foundation models

1



Retrieval Augmented Generation (RAG)

Guide pre-trained models with private, domain-specific contextual data

For example, virtual agents with limited domain-specific requirements

2

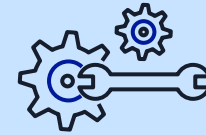


Fine-tune a pre-trained model

Specialized knowledge for specific tasks with labeled examples

For example, domain-intensive knowledge agents

3



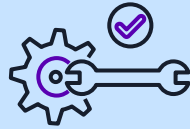
Continued pre-training

Generalized and specialized knowledge for your domain with unlabeled data

For example, deep domain-specific applications trained on specific data

Customizing foundation models

2

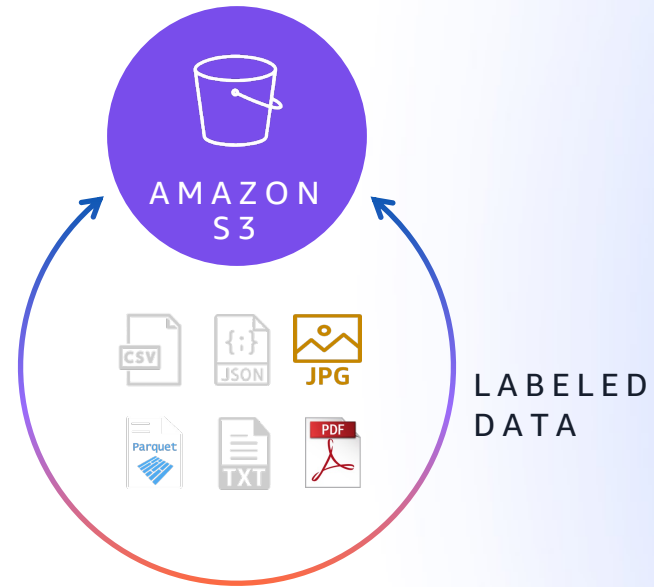


Fine-tune a pre-trained model

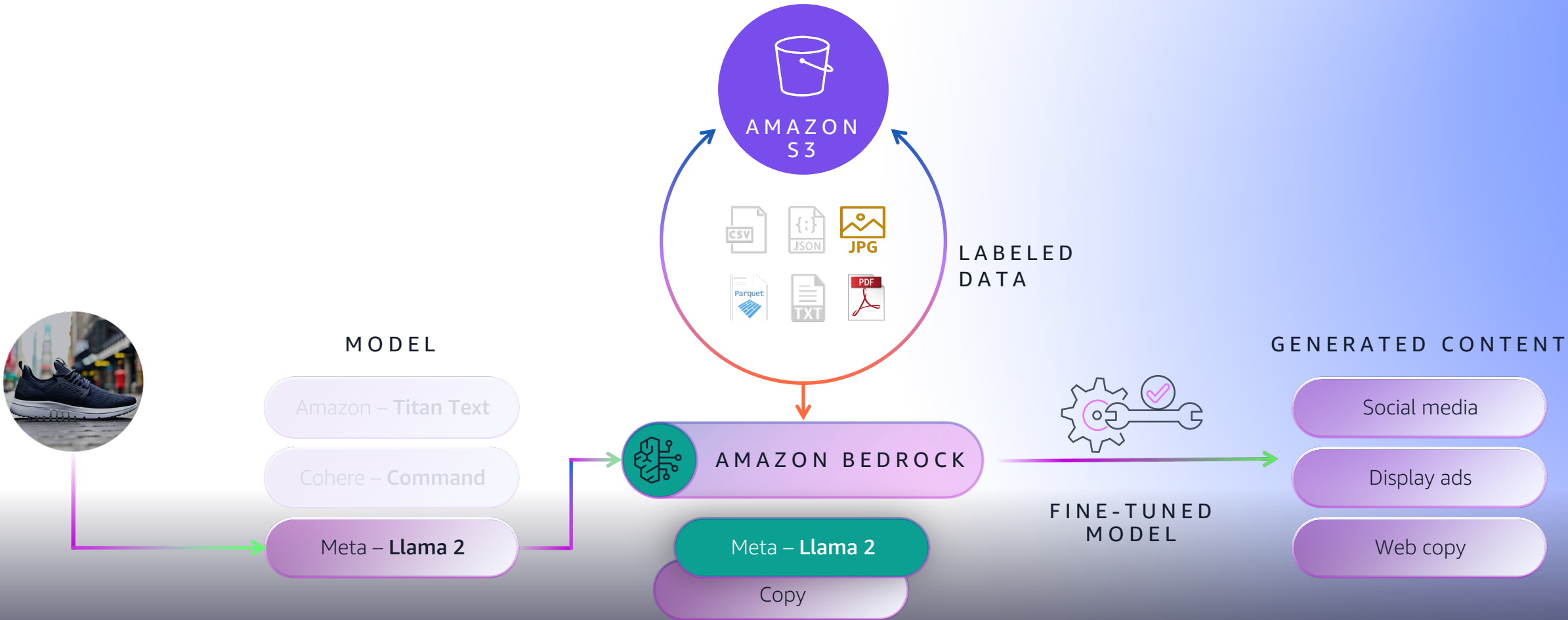
Specialized knowledge for specific tasks with labeled examples

For example, domain-intensive knowledge agents

Fine-tuning your FMs

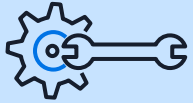


Fine-tuning your FMs



Customizing foundation models

1

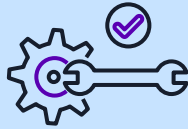


Retrieval Augmented Generation (RAG)

Guide pre-trained models with private, domain-specific contextual data

For example, virtual agents with limited domain-specific requirements

2

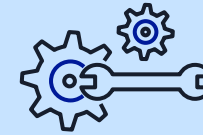


Fine-tune a pre-trained model

Specialized knowledge for specific tasks with labeled examples

For example, domain-intensive knowledge agents

3



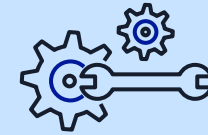
Continued pre-training

Generalized and specialized knowledge for your domain with unlabeled data

For example, deep domain-specific applications trained on specific data

Customizing foundation models

3

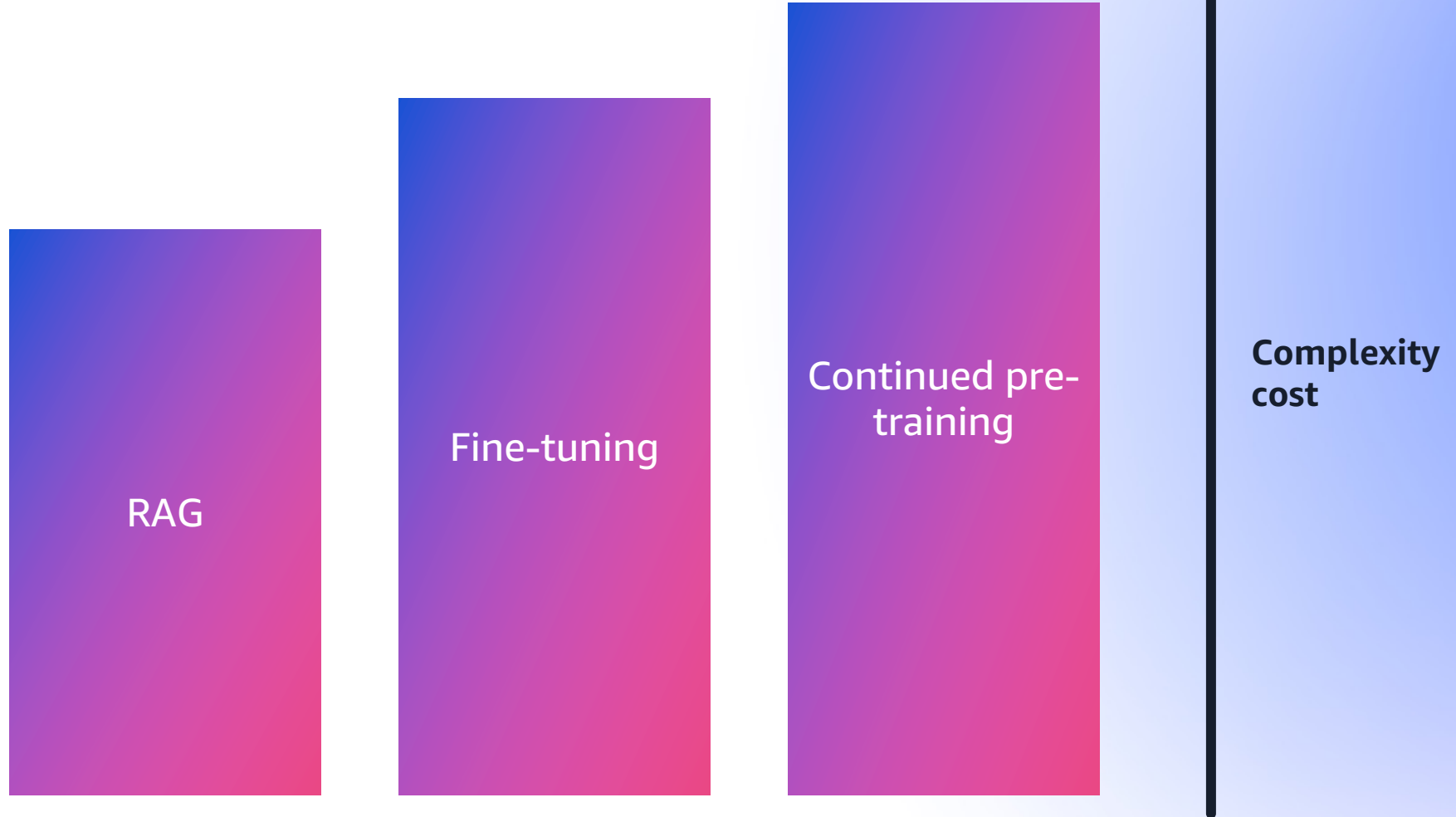


Continued pre-training

Generalized and specialized knowledge for your domain with unlabeled data

For example, deep domain-specific applications trained on specific data

Your journey to customizing FMs



AWS data services for generative AI



Generative AI and AI/ML Stack

APPLICATIONS TO BOOST PRODUCTIVITY



Amazon Q
Business



Amazon Q
Developer



Amazon Q in
QuickSight



Amazon Q in
Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs



Amazon Bedrock

Guardrails | Agents | Studio

INFRASTRUCTURE TO BUILD AND TRAIN AI MODELS



GPUs



Trainium



Inferentia



SageMaker

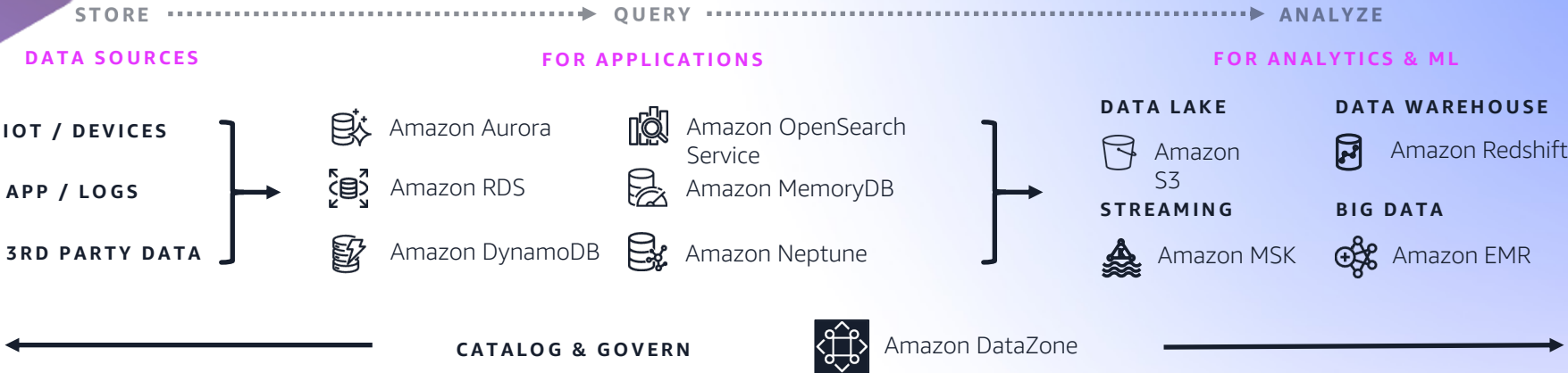
Overlay of data on the generative AI/ML stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

DATA THAT SUSTAINS YOUR INNOVATION FLYWHEEL



AWS is an excellent place to build a data strategy to fuel your generative AI applications



Comprehensive

Comprehensive set of services for storing and querying structured, unstructured, and vector data



Integrated

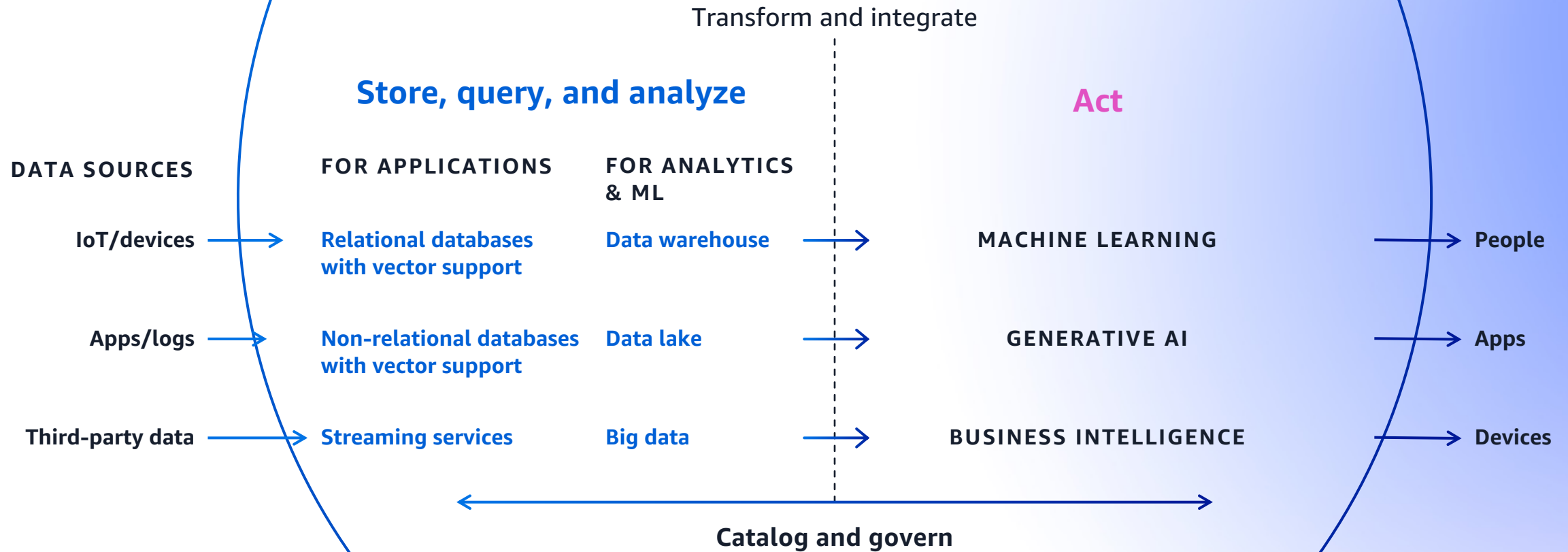
Choices for integrating data including zero ETL so you can easily connect to all your data



Governed

End-to-end data governance capabilities, responsible AI, and regulating user interactions with LLMs

A comprehensive set of services for your data foundation



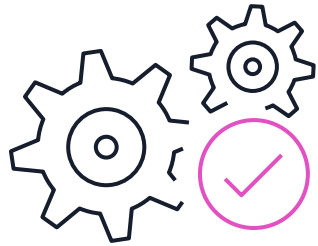
Simplify data integration

Zero-ETL: Eliminate requirement for creating data pipelines with data in place integration

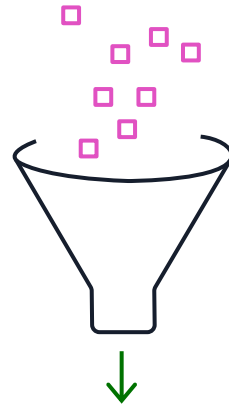
Zero-ETL integrations



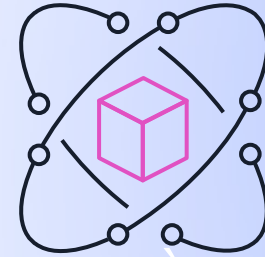
Simplified data pipelines enable faster time to market



Direct service integrations to **eliminate ETL**

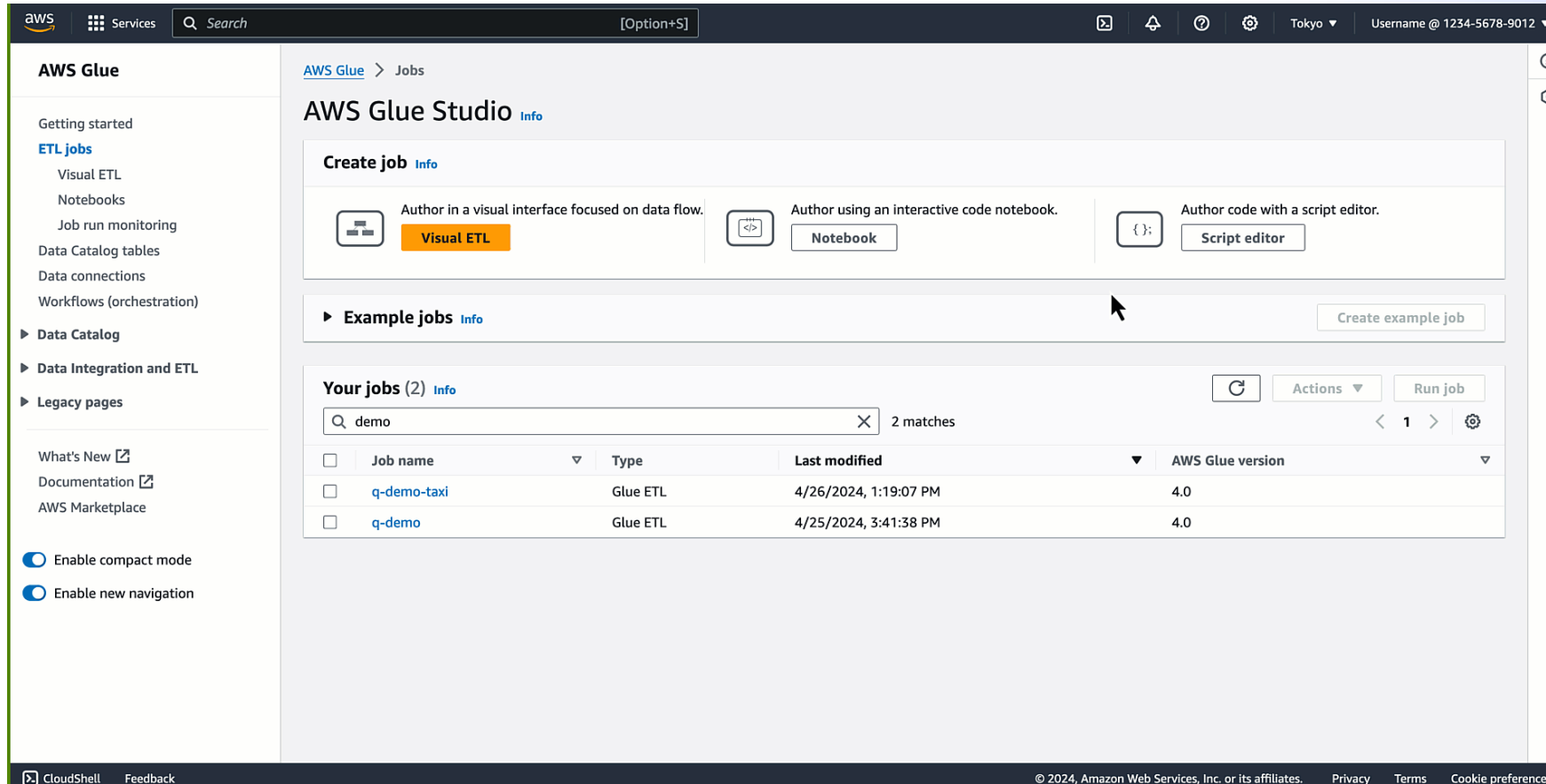


AWS Glue for value-add data transformations and more



Connectors to **hundreds of data sources** and **services for partner and third-party data**

Amazon Q Data Integration in AWS Glue



The screenshot displays the AWS Glue Studio interface. The left sidebar contains navigation options such as 'Getting started', 'ETL jobs', 'Data Catalog', and 'Legacy pages'. The main content area is titled 'AWS Glue Studio' and features a 'Create job' section with three options: 'Visual ETL' (author in a visual interface), 'Notebook' (author using an interactive code notebook), and 'Script editor' (author code with a script editor). Below this is an 'Example jobs' section with a 'Create example job' button. The 'Your jobs (2)' section shows a search for 'demo' with 2 matches, displaying a table of jobs.

<input type="checkbox"/>	Job name	Type	Last modified	AWS Glue version
<input type="checkbox"/>	q-demo-taxi	Glue ETL	4/26/2024, 1:19:07 PM	4.0
<input type="checkbox"/>	q-demo	Glue ETL	4/25/2024, 3:41:38 PM	4.0

- Tell Amazon Q Developer what you need in English, it will return a complete job
- It can generate complex data integration jobs
- Helps troubleshoot jobs
- Available in: Amazon Q chat and AWS Glue Studio notebook

Amazon Q generative SQL in Amazon Redshift Query Editor

Count the number of schools
in Alameda County that
have less than 100 test takers



Natural output

```
SELECT COUNT(*)  
FROM schools  
WHERE district = 'Alameda'  
AND num_test_takers < 100
```

Generative SQL recommendation

```
SELECT COUNT(*)  
FROM schools a  
JOIN satscores ss  
ON s.CDSCode=ss.cds  
WHERE s.County = 'Alameda'  
AND ss.NumTstTskr < 100
```

Personalized SQL recommendations
Personalized to your database, tables, and schema

The next generation of Amazon SageMaker

Collaborate and build faster with a single data and AI development environment

Develop and scale AI use cases with the broadest set of tools

Reduce data silos with an open lakehouse to unify all your data

Meet your enterprise security needs with built-in data and AI governance





Amazon SageMaker

Unified Studio

COMING SOON

COMING SOON

COMING SOON

SQL Analytics

Amazon Redshift

Data Processing

Amazon EMR
AWS Glue
Amazon Athena

Model Development

Amazon SageMaker AI

Gen AI App Development

Amazon Bedrock

Streaming

Amazon MSK
Amazon Kinesis

Business Intelligence

Amazon QuickSight

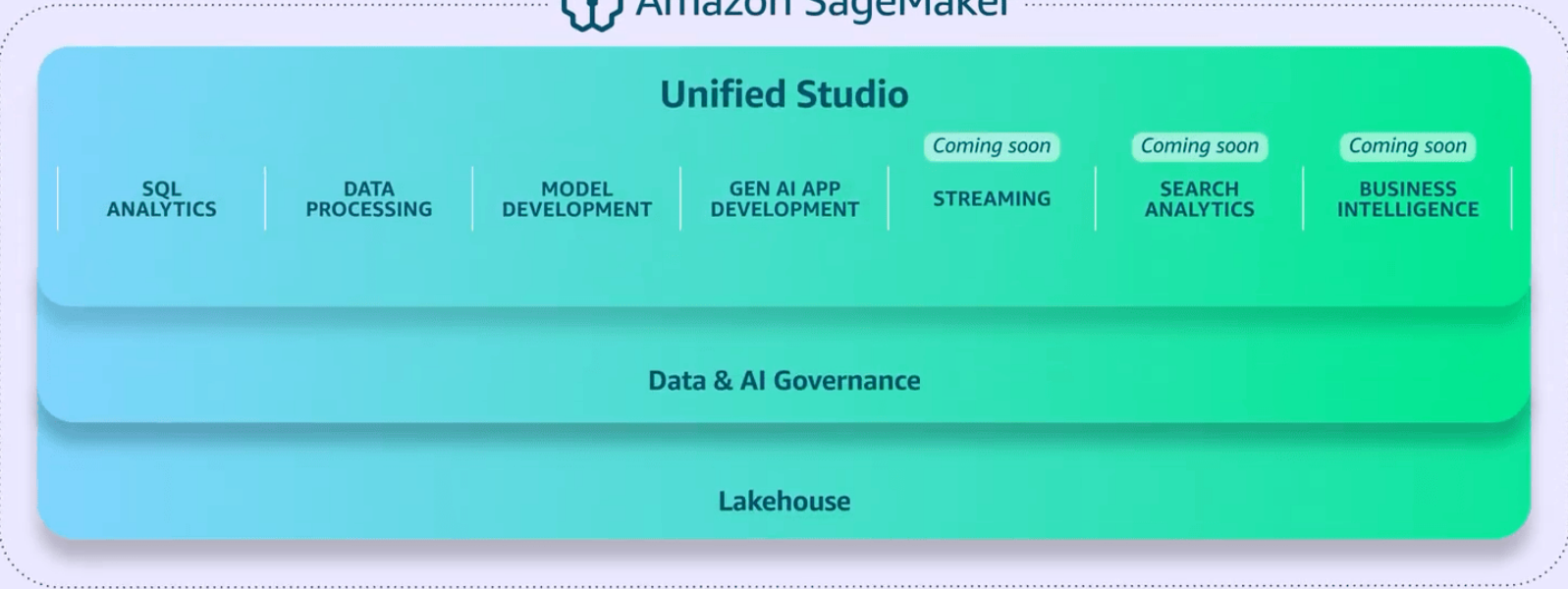
Search Analytics

Amazon OpenSearch Service

Lakehouse

Data & AI Governance

Amazon SageMaker Unified Studio



We are here to help

Need peer-level executive guidance?

AWS Data Strategy

- Mental models and strategies based on the firsthand experience of former CXOs
- Get a peer-level sounding board and sparring partner

Inspire and accelerate your data transformation

Want to build a data vision and strategy?

AWS Data-Driven Everything

- Create an organizational vision for innovation with data to drive business outcomes
- Define the first pilot, learn, and build

Jump-start the data flywheel

Want to modernize your data foundation?

AWS re:Imagine Data

- Define the migration and modernization strategy for a future data foundation
- Lower cost, increase capacity, and unlock business access

Migrate and modernize data

AWS Generative AI Innovation Center





Thank you!

ANKIT BATTAWAR (he/him)

Solution Architect
Amazon Web Services
battawar@amazon.com

GOVIND PALANISAMY (he/him)

Sr Solution Architect
Amazon Web Services
govpal@amazon.com

Please complete the survey
for this session



Data and Analytics Track

Data foundations in the age of
Generative AI