AWS State, Local, and Education Learning Days

New York City

aws Learning Days
State, Local, and Education

BREAKOUT SESSION

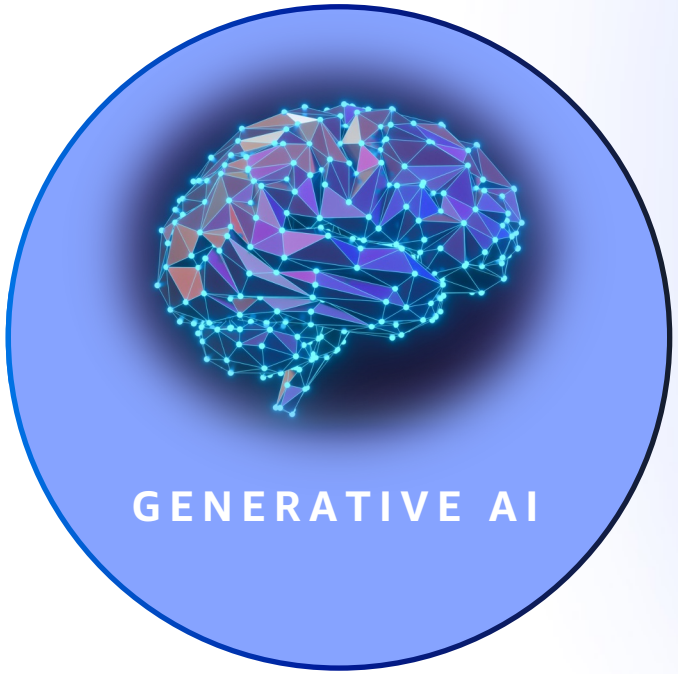
Data foundations in the era of Generative AI

Arianna Burgman (she/her)

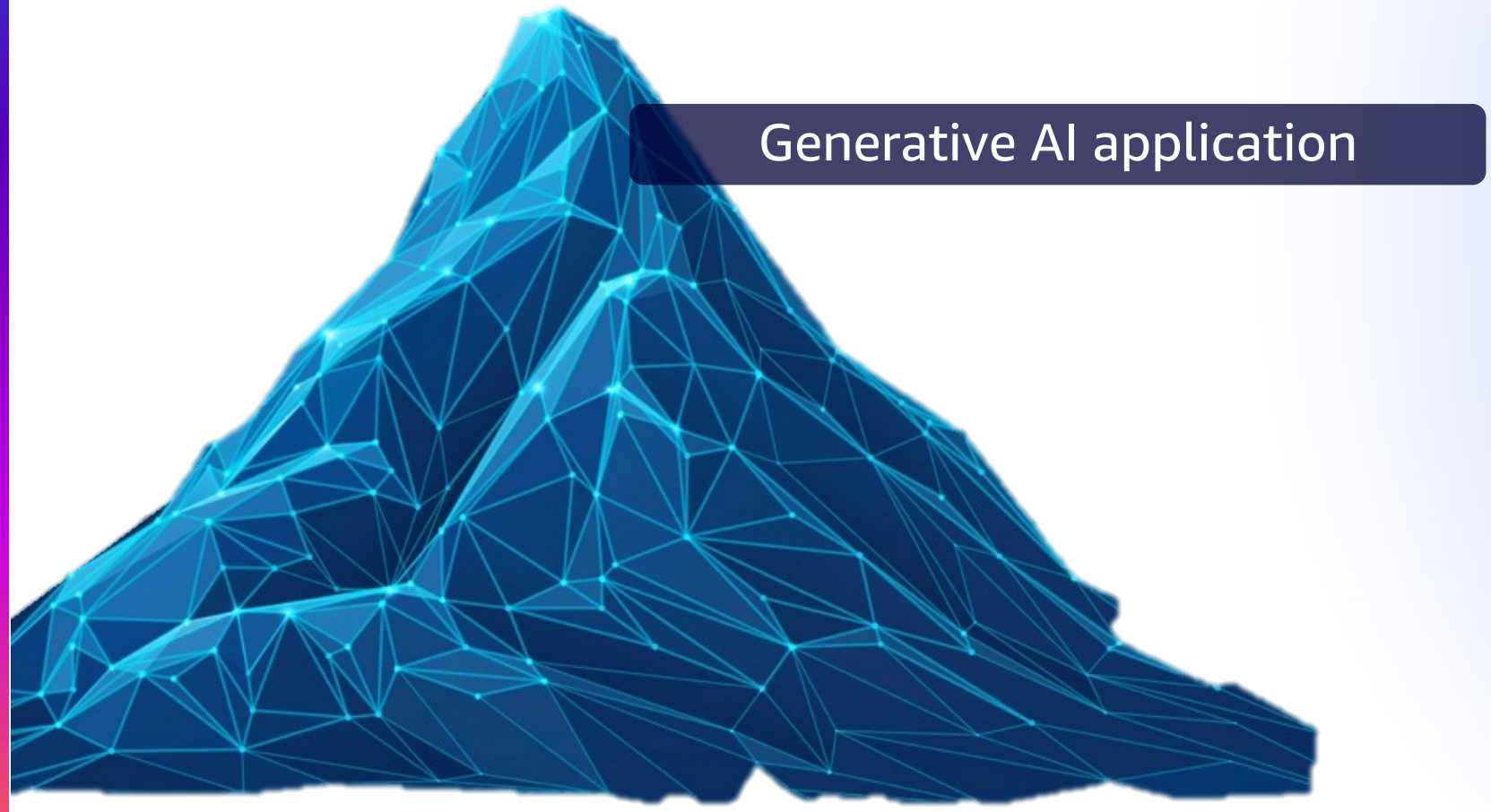
Solution Architect
Amazon Web Services
burgmaa@amazon.com

aws Learning Days
State, Local, and Education

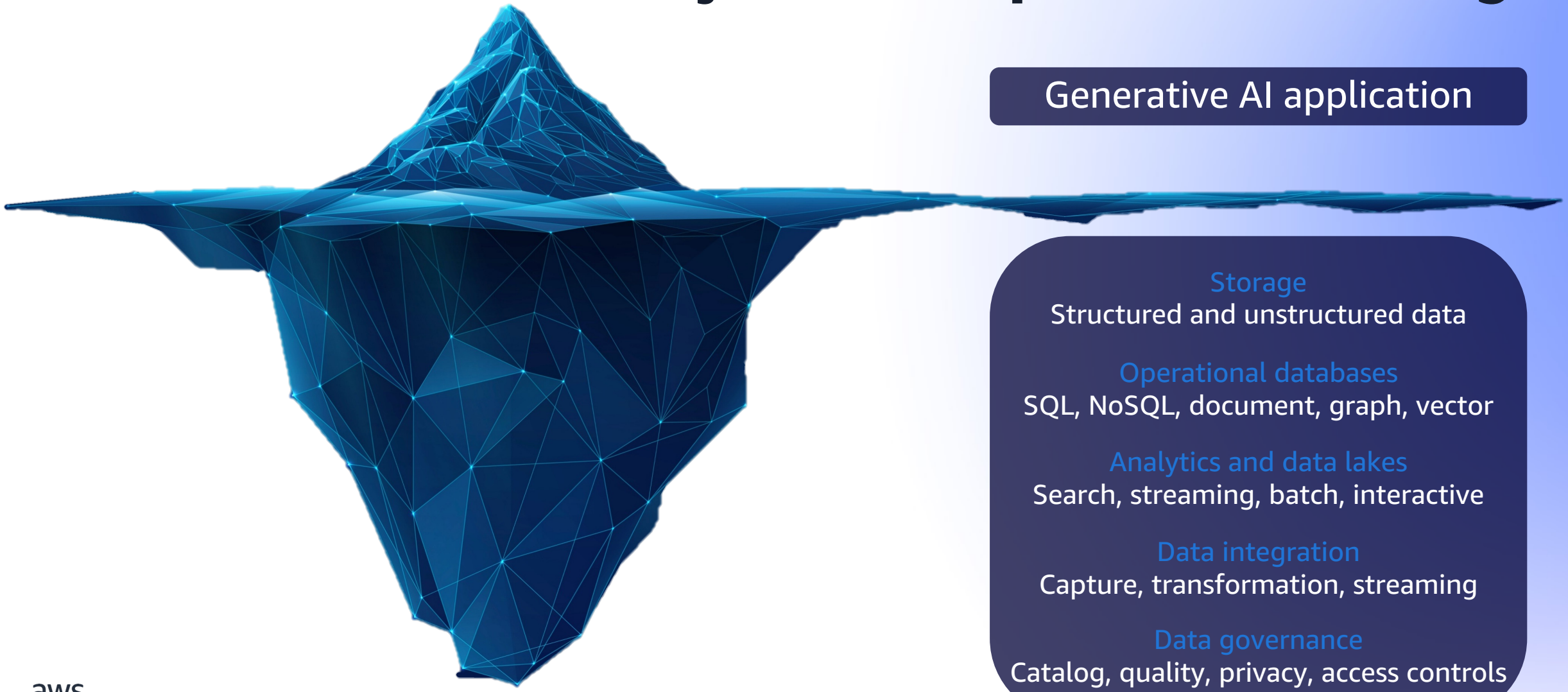
Innovation can transform industries



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

More relevant conversational search

The objective

Provide answers to employees based on enterprise data

The data

- Slack
- Internal guides
- SharePoint

Automated budgeting and resource allocation

The objective

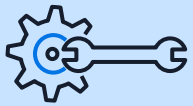
Analyze spending data to identify areas where resources can be allocated more efficiently

The data

- Enterprise Systems and Databases
- Internal guides
- Spreadsheets

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

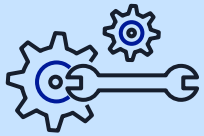
2

Fine-tune a pre-trained model

Specialized knowledge for specific tasks with labeled examples

For example, domain-intensive knowledge agents

3



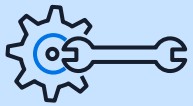
Continued pre-training

Generalized and specialized knowledge for your domain with unlabeled data

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

Customizing foundation models

1

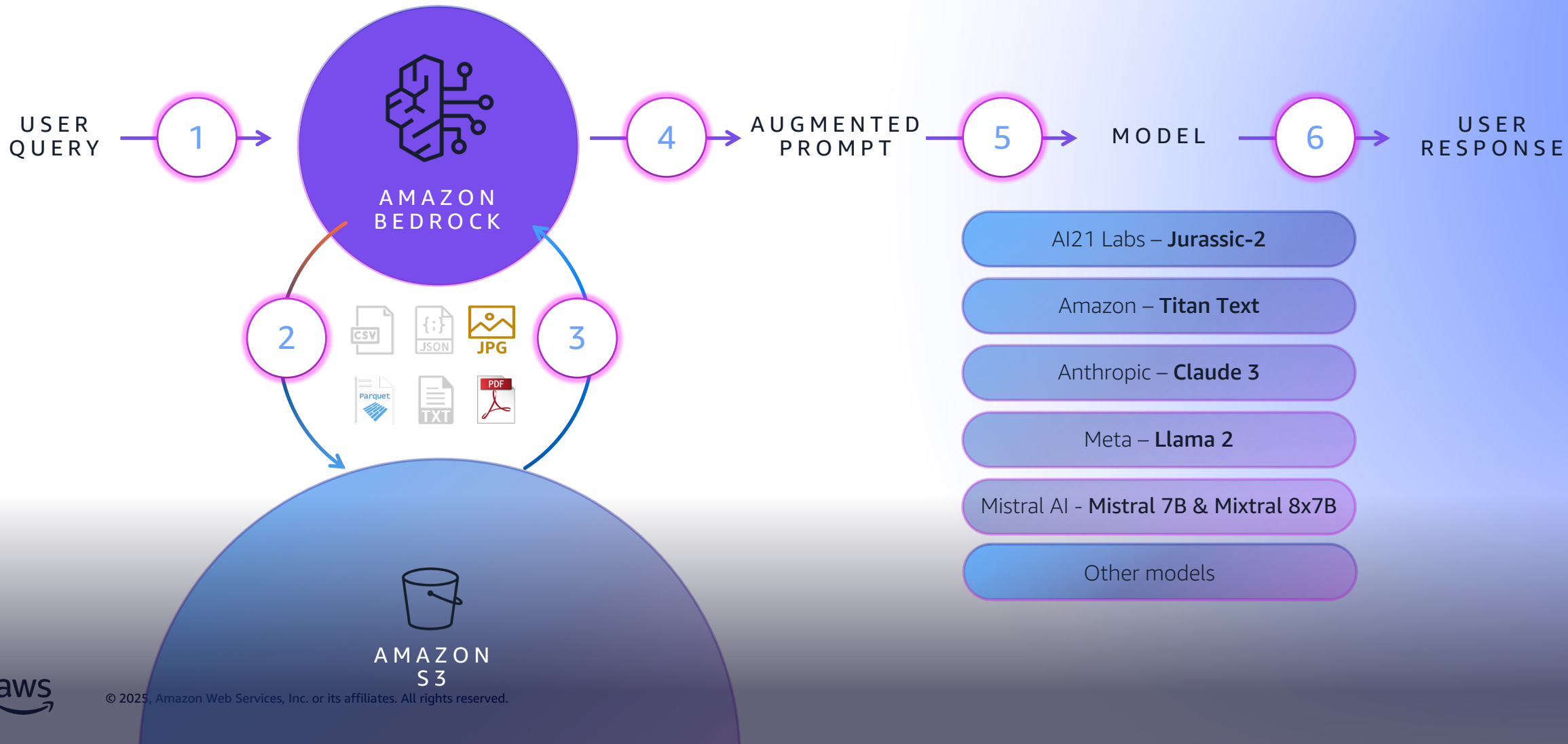


Retrieval Augmented Generation (RAG)

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

For example, virtual agents with limited domain-specific requirements

Retrieval augmented generation (RAG)



What are vector embeddings?

Source domain-specific data → Tokenization → Vectorization → Store in vector data store → Perform semantic similarity search → Include semantically similar context in prompt

Documents

Audio/video

Images

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

Enabling vector search across AWS services

Amazon OpenSearch Service

Amazon OpenSearch Serverless

Amazon Aurora PostgreSQL

Amazon RDS for PostgreSQL

Amazon DocumentDB

Amazon DynamoDB via zero ETL

Amazon MemoryDB for Redis

Amazon Neptune

Customizing foundation models

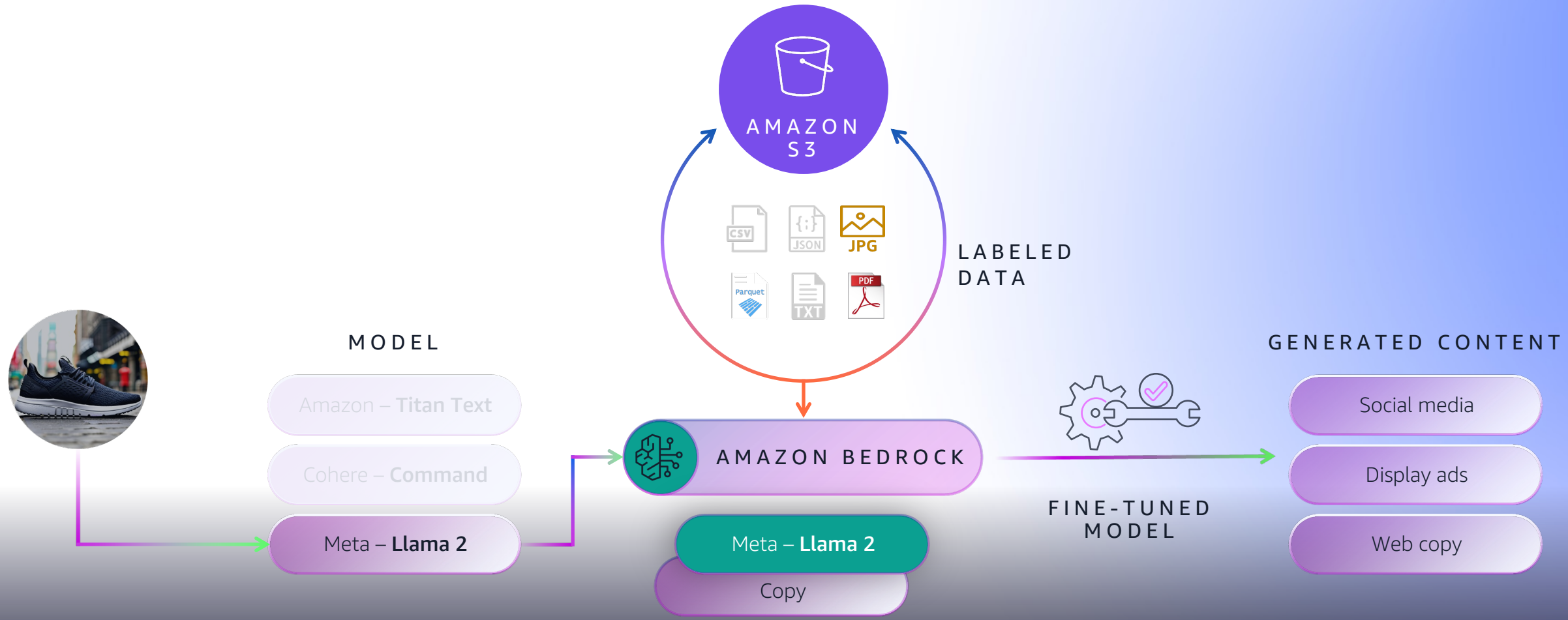
2

Fine-tune a pre-trained model

Specialized knowledge for specific tasks with labeled examples

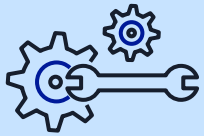
For example, domain-intensive knowledge agents

Fine-tuning your FMs



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

STORE → QUERY → ANALYZE

DATA SOURCES

- IOT / DEVICES
- APP / LOGS
- 3RD PARTY DATA

FOR APPLICATIONS

- Amazon Aurora
- Amazon RDS
- Amazon DynamoDB
- Amazon OpenSearch Service
- Amazon MemoryDB
- Amazon Neptune

FOR ANALYTICS & ML

DATA LAKE
- Amazon S3

DATA WAREHOUSE
- Amazon Redshift

STREAMING
- Amazon MSK

BIG DATA
- Amazon EMR

CATALOG & GOVERN

Amazon DataZone

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

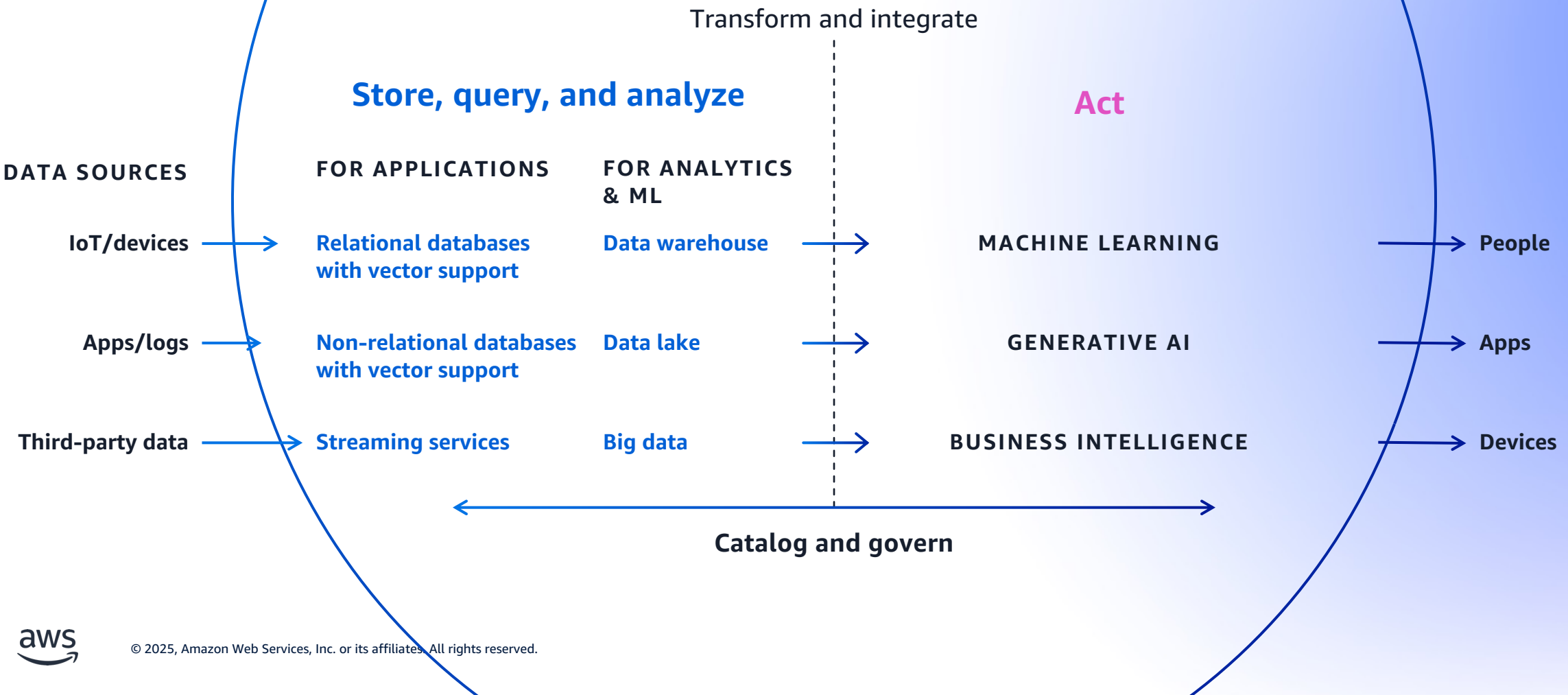
Integrated

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

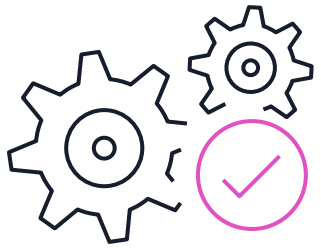
Governed

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

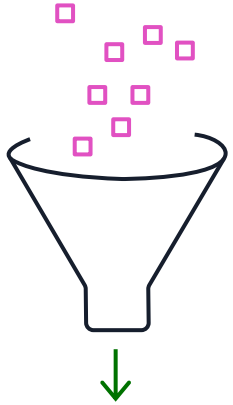
A comprehensive set of services for your data foundation



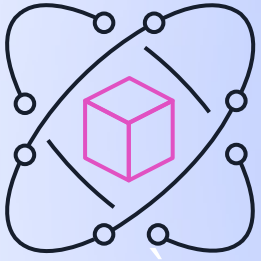
Simplified data pipelines enable faster time to market



Direct service integrations to **eliminate ETL**



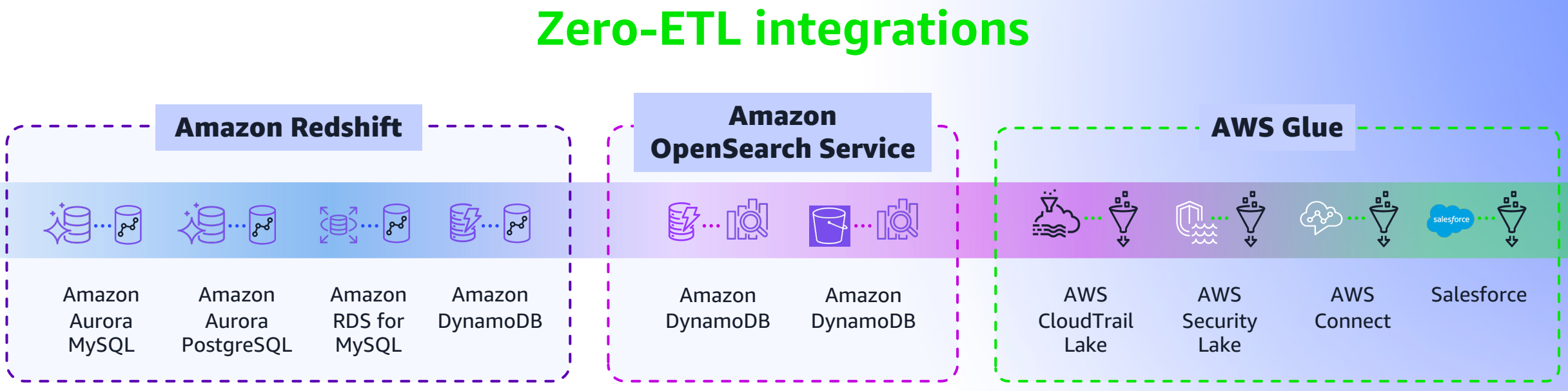
AWS Glue for value-add data transformations and more



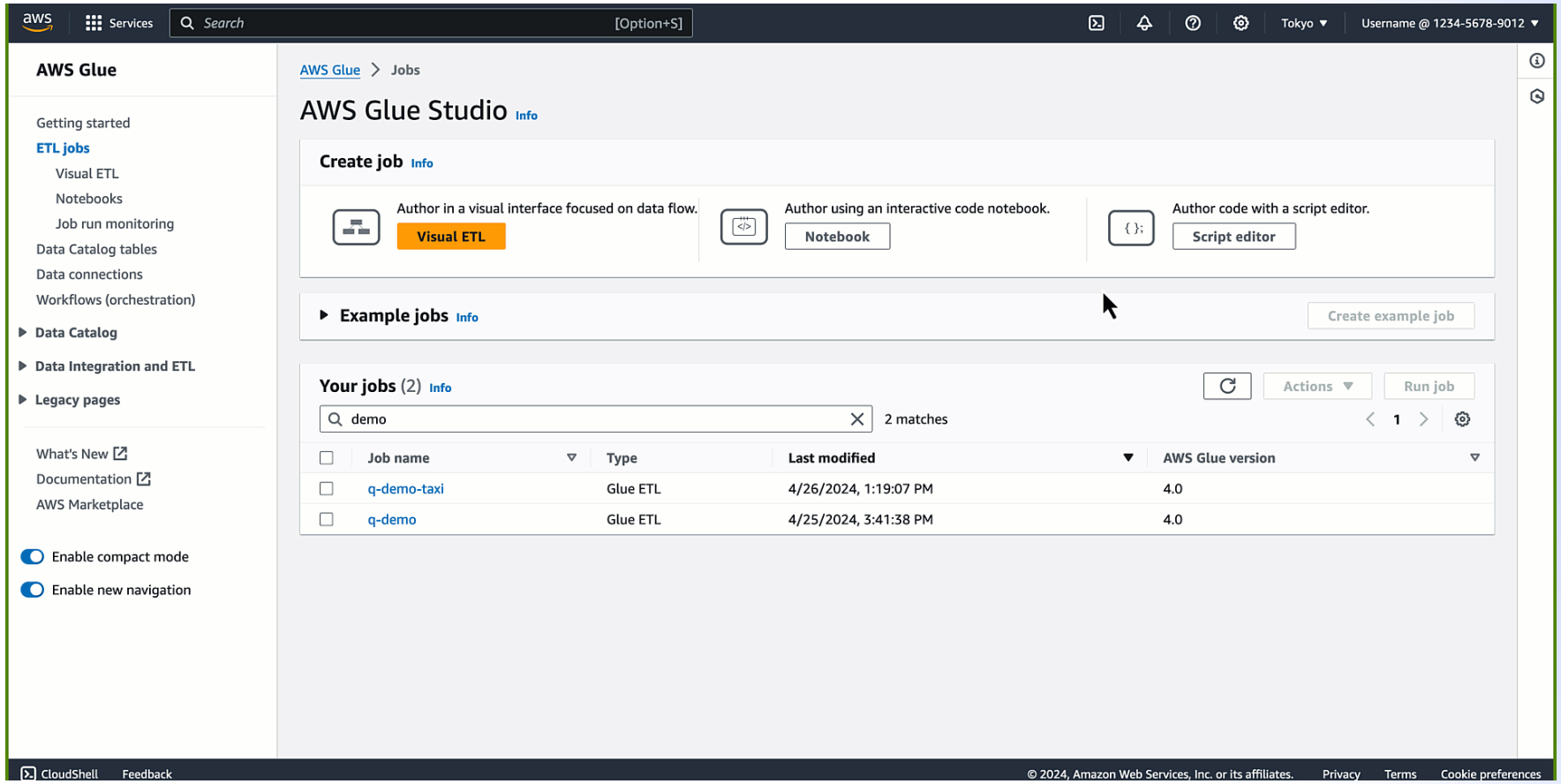
Connectors to **hundreds of data sources** and **services for partner and third-party data**

Simplify data integration

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



Amazon Q Data Integration in AWS Glue



- 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.NumTstTakr < 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

SQL Analytics
Amazon Redshift

Data Processing
Amazon EMR
AWS Glue
Amazon Athena

Model Development
Amazon SageMaker AI

Gen AI App Development
Amazon Bedrock

COMING SOON

Streaming
Amazon MSK
Amazon Kinesis

COMING SOON

Business Intelligence
Amazon QuickSight

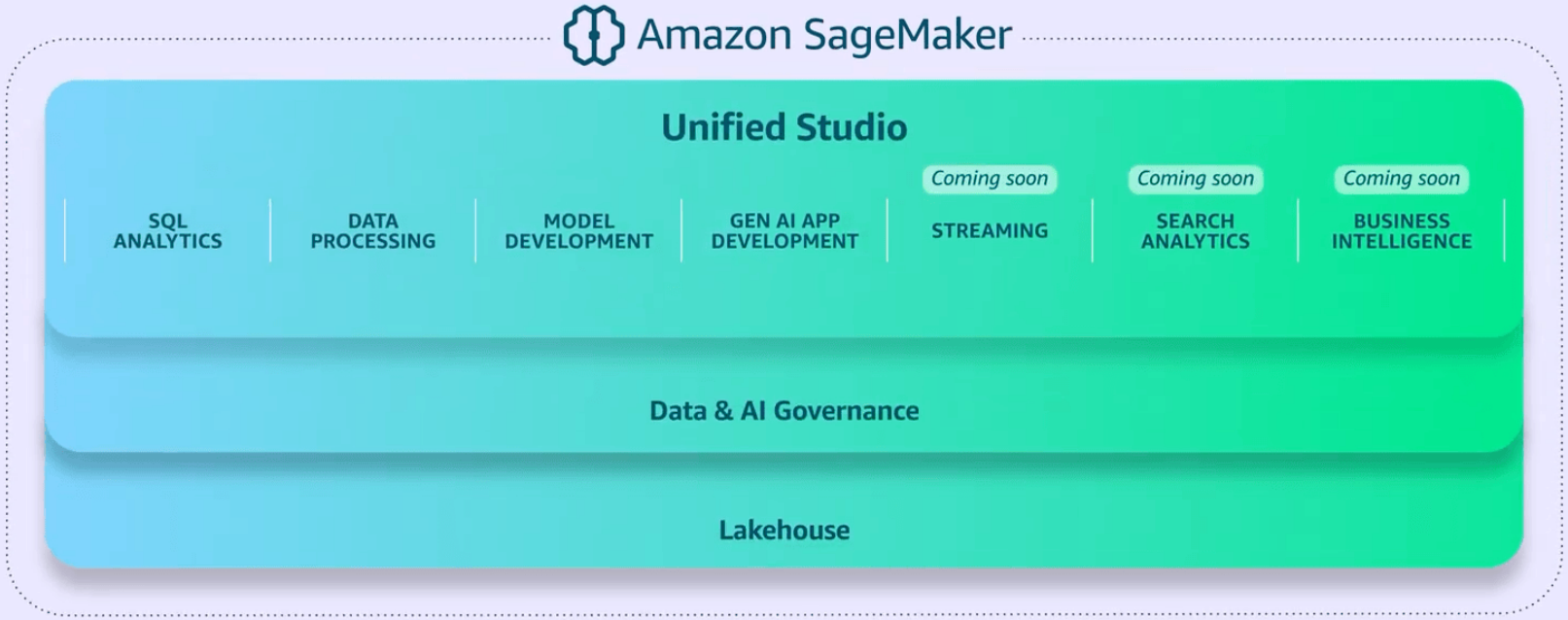
COMING SOON

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!

Please complete the survey for this session

Arianna Burgman (she/her)

Solution Architect
AWS
burgmaa@amazon.com

Track: Data and Analytics Track

Session: Data foundations in the age of Generative AI

aws Learning Days
State, Local, and Education