



AWS의 클라우드 전환 인사이트

한국 경영정보학회 산학세미나

김웅환 kiwoong@amazon.com

PS Korea SA Head



Agenda

- AWS Intro
- Cloud migration
- Cloud Innovation - GenAI



amazon **aws**

World's Largest Retail Company World's Most Innovative Cloud Provider



- 1994 ○ Founded
- 1995 ○ Amazon.com
- 1998 ○ CDs and DVDs
- 2006 ○ 
- 2007 ○ Kindle
- 2011 ○ Video
- 2012 ○ Groceries
- 2014 ○ Alexa/Echo
- 2015 ○ Bookstores
- 2017 ○ Amazon Go

최초의 웹서비스 형태의
IT 인프라 서비스

Now commonly known as
“Cloud Computing”



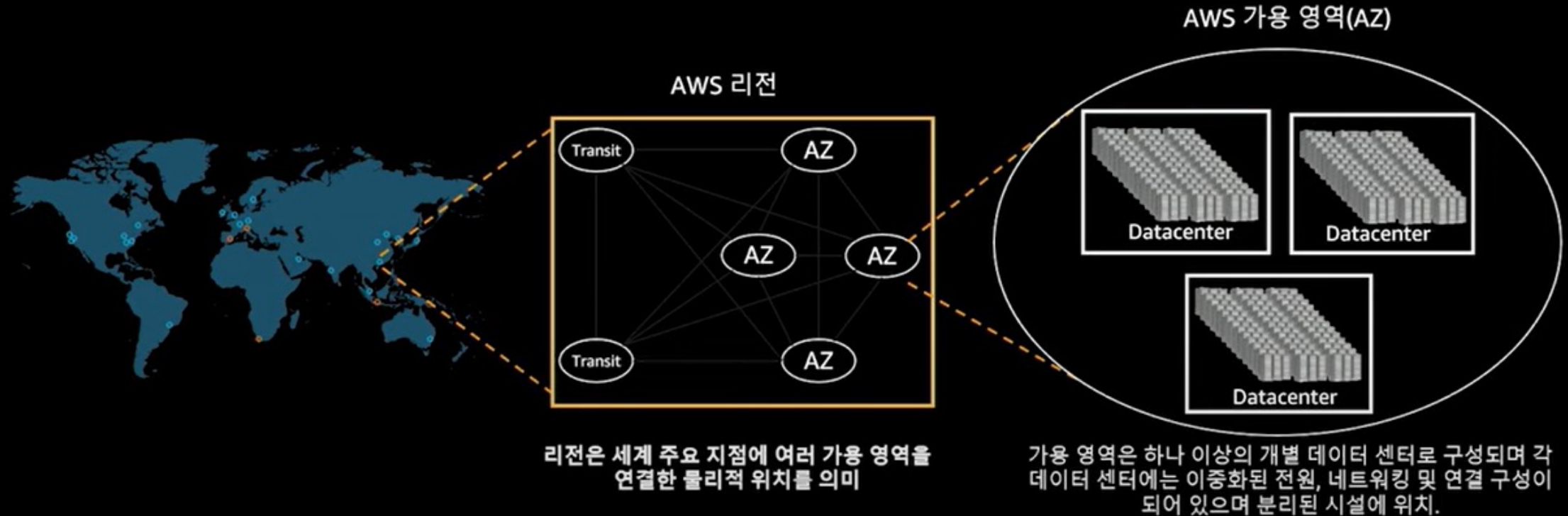
AWS의 글로벌 리전과 가용영역 위치

전 세계 33개 리전과 105개 가용영역으로 풍부하고 방대한 글로벌 가용성 보장



AWS 서울 리전 - 4 AZs

- AWS 리전은 **고가용성, 높은 확장성, 높은 내결함성**을 위해 여러 AZ로 구성
애플리케이션과 데이터는 실시간으로 복제되며 서로 다른 AZ에서 일관성 유지



13년 연속 클라우드 리더로 선정된 AWS

2023년 전략 클라우드 플랫폼 서비스 부문 (Strategic Cloud Platform Services) 매직 쿼드런트

https://aws.amazon.com/ko/resources/analyst-reports/gartner/global-mq-ardm-23-magic-quadrant-for-strategic-cloud-platform-services/?trk=5b366b4c-27fd-4b41-93cb-3195213bc9f6&sc_channel=el

Gartner, Magic Quadrant for Strategic Cloud Platform Services, David Wright, Dennis Smith, Kevin Ji, Miguel Angel Borrega, Alessandro Galimberti, Stephanie Bauman
4 December 2023. Gartner 및 매직 쿼드런트는 미국 및 그 외 지역에서 Gartner, Inc. 및/또는 해당 회사의 등록 상표이며 여기에 따라 이 문서에서 사용됩니다. All rights reserved. Gartner는 연구 간행물에 언급된 어떠한 공급업체, 제품 또는 서비스도 보증하지 않으며, 기술 사용자에게 순위가 높은 공급업체만 선택하도록 권장하지 않습니다. Gartner의 연구 간행물은 Gartner 연구 조사의 의견을 반영하고 있으며, 사실을 진술한 것으로 간주해서는 안 됩니다. Gartner는 정확성 또는 특정 목적 부합성에 대한 보증을 포함하여 이 연구와 관련된 모든 명시적 또는 묵시적 보증을 부인합니다.

전략 클라우드 플랫폼 서비스 부문 Gartner 매직 쿼드런트





Cloud Migration 동기, 유형

마이그레이션 동기



AWS 마이그레이션을 통해 고객이 얻을 수 있는 이득

31%

평균 인프라 비용 절감

7x

더 적은 다운타임

62%

더 높은 IT 스텝 생산성

3x

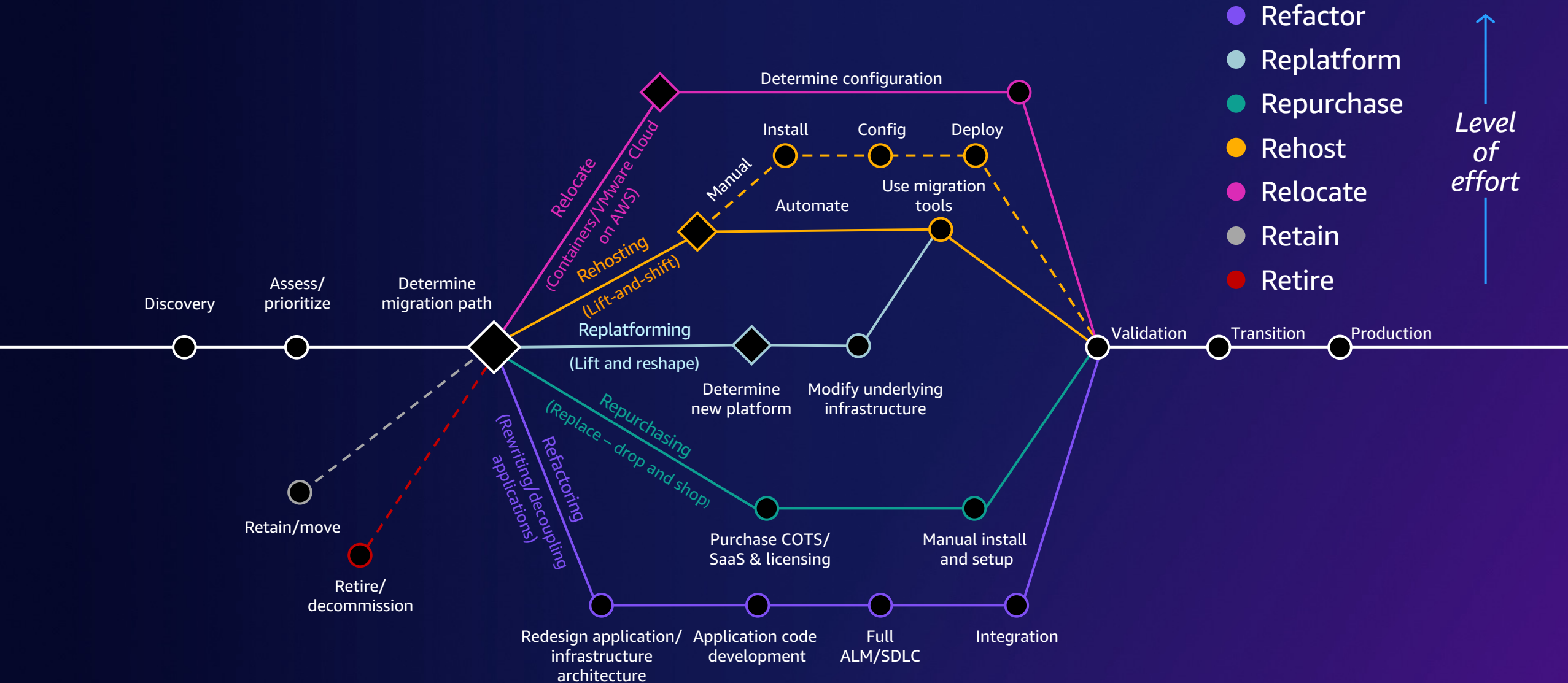
연간 신규 기능 출시

Sources: IDC, AWS Analysis

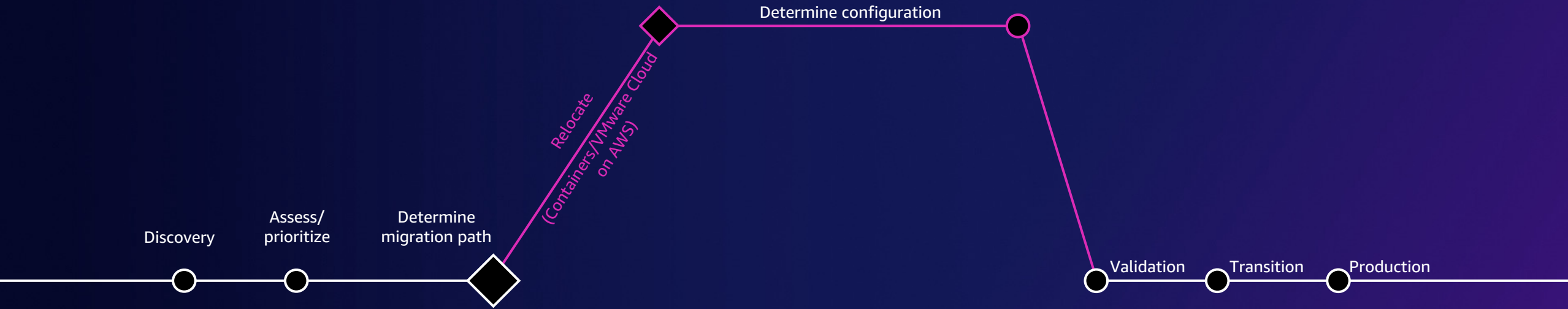


The 7Rs – 마이그레이션 전략

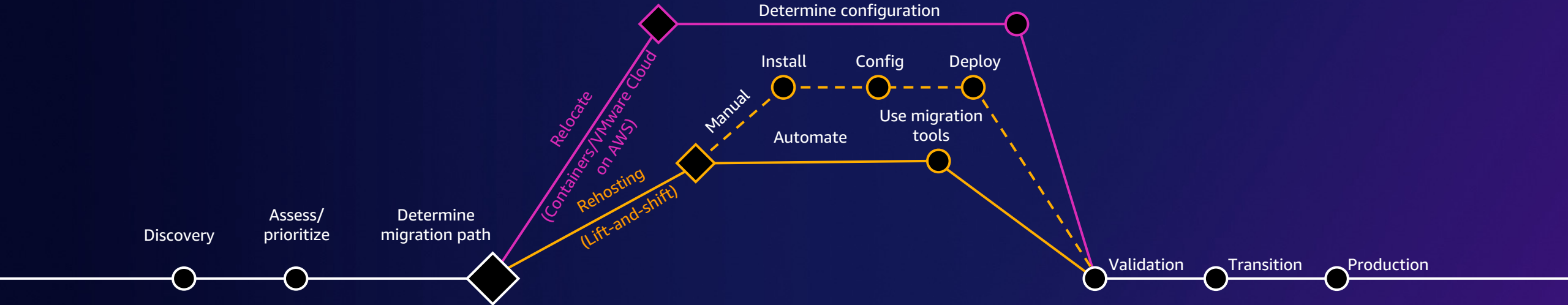
The 7'R's



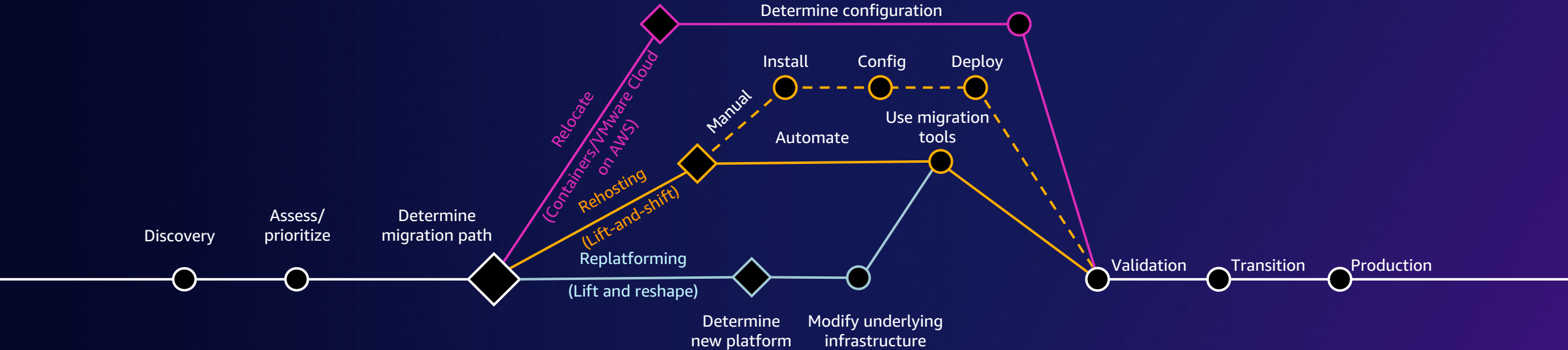
Application migration strategies: **Relocate**



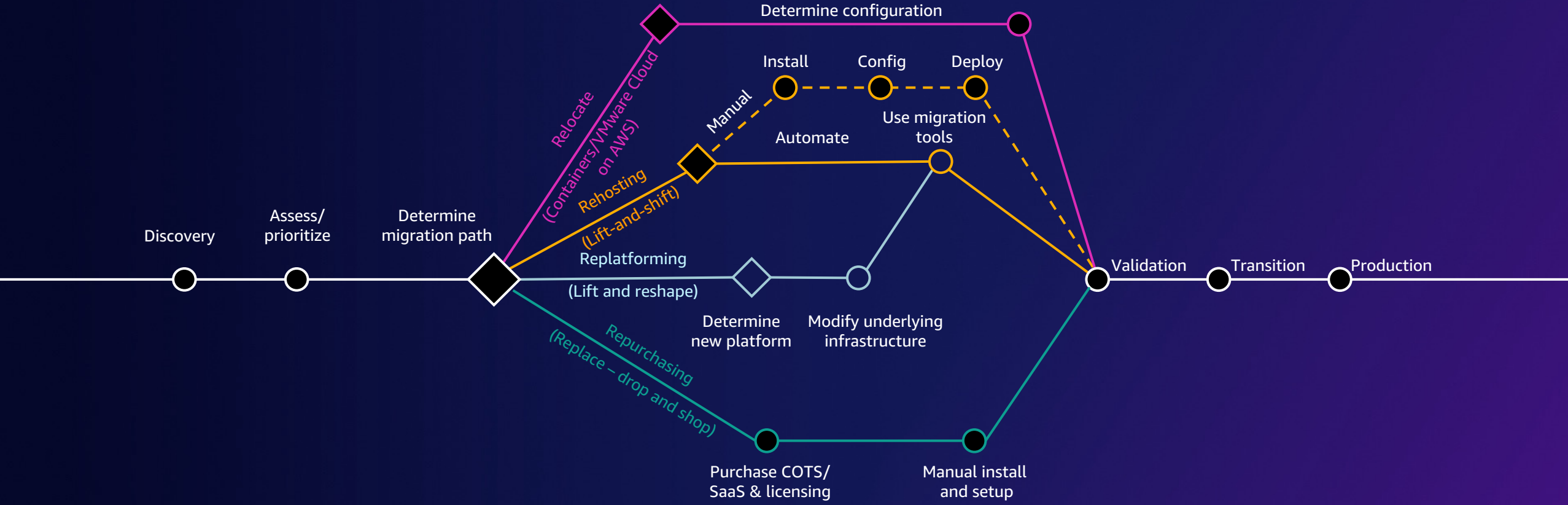
Application migration strategies: **Rehost**



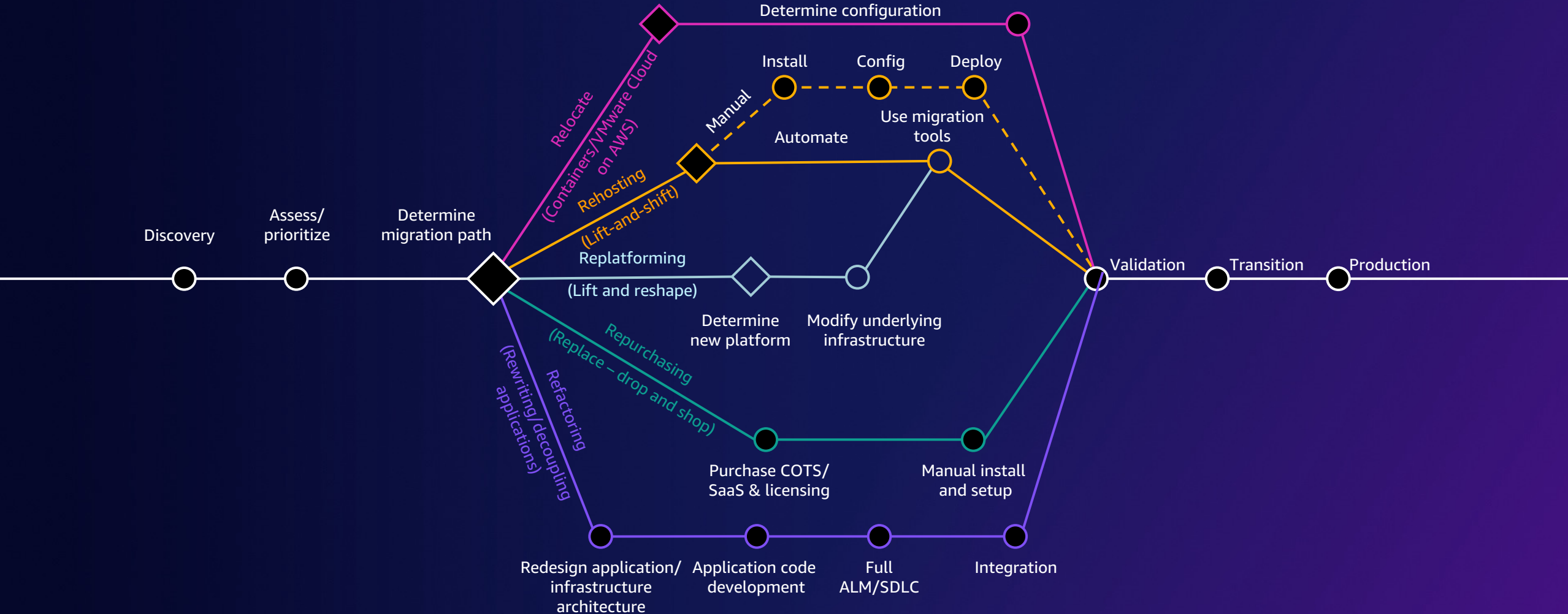
Application migration strategies: **Replatform**



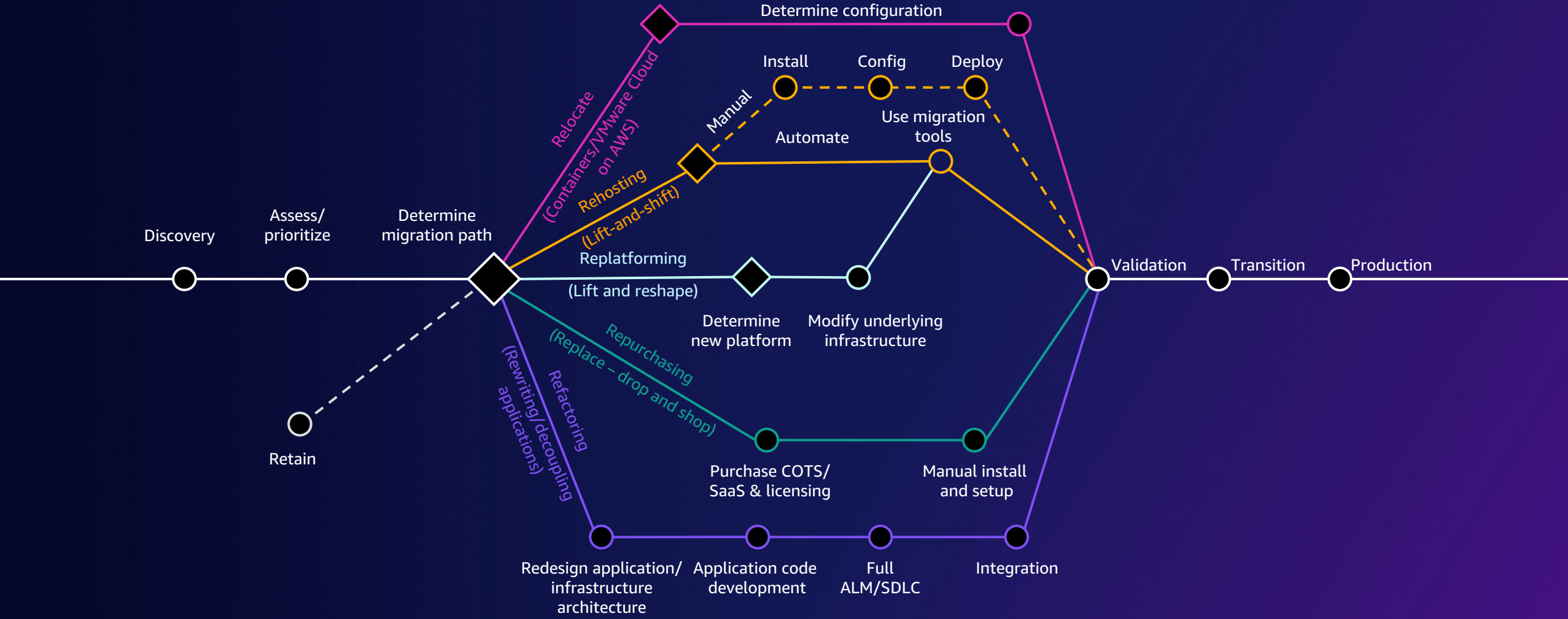
Application migration strategies: Repurchase



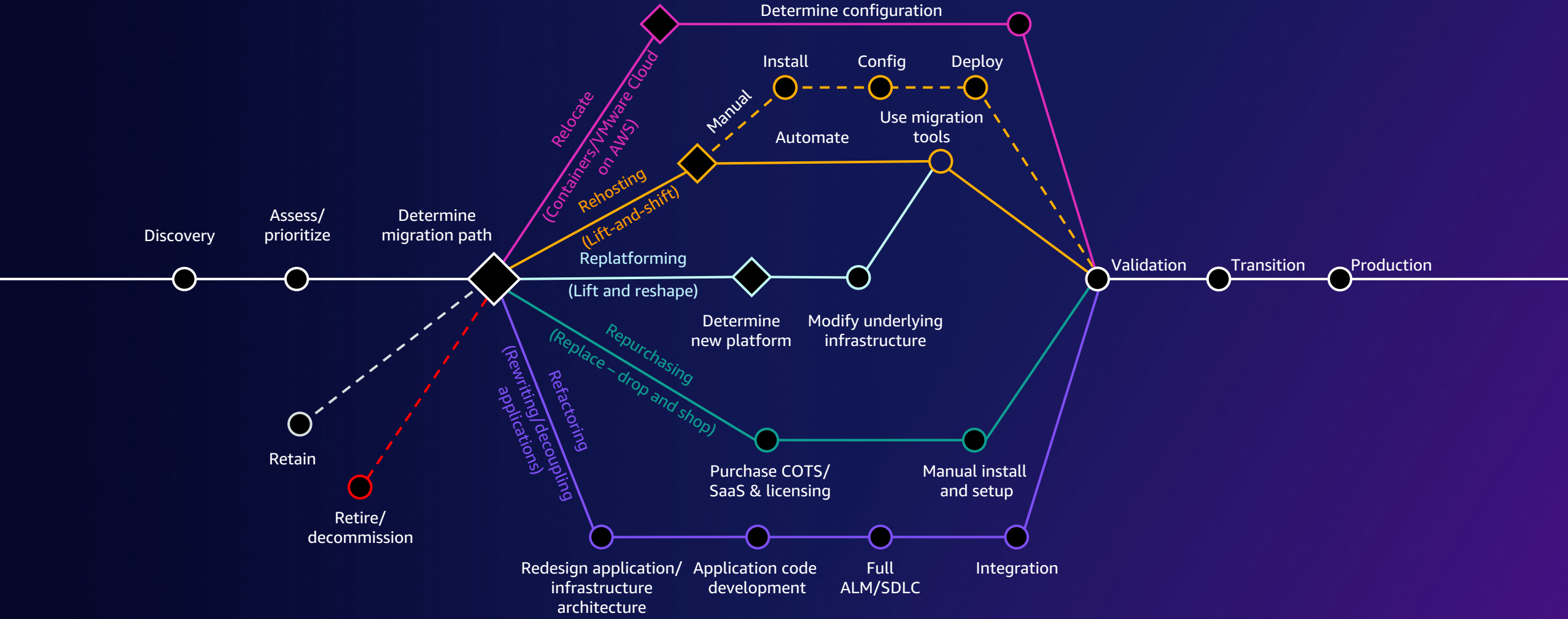
Application migration strategies: Refactoring



Application migration strategies: Retain

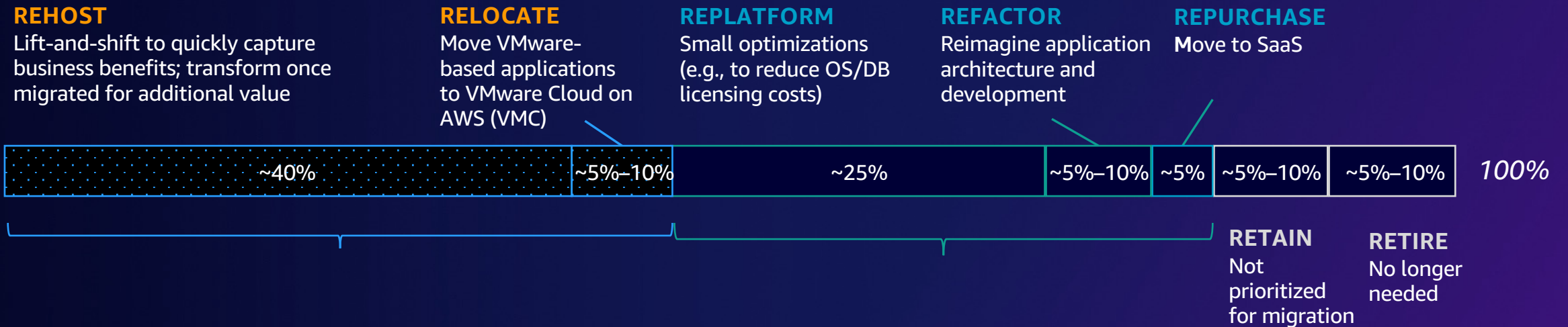


Application migration strategies: **Retire**



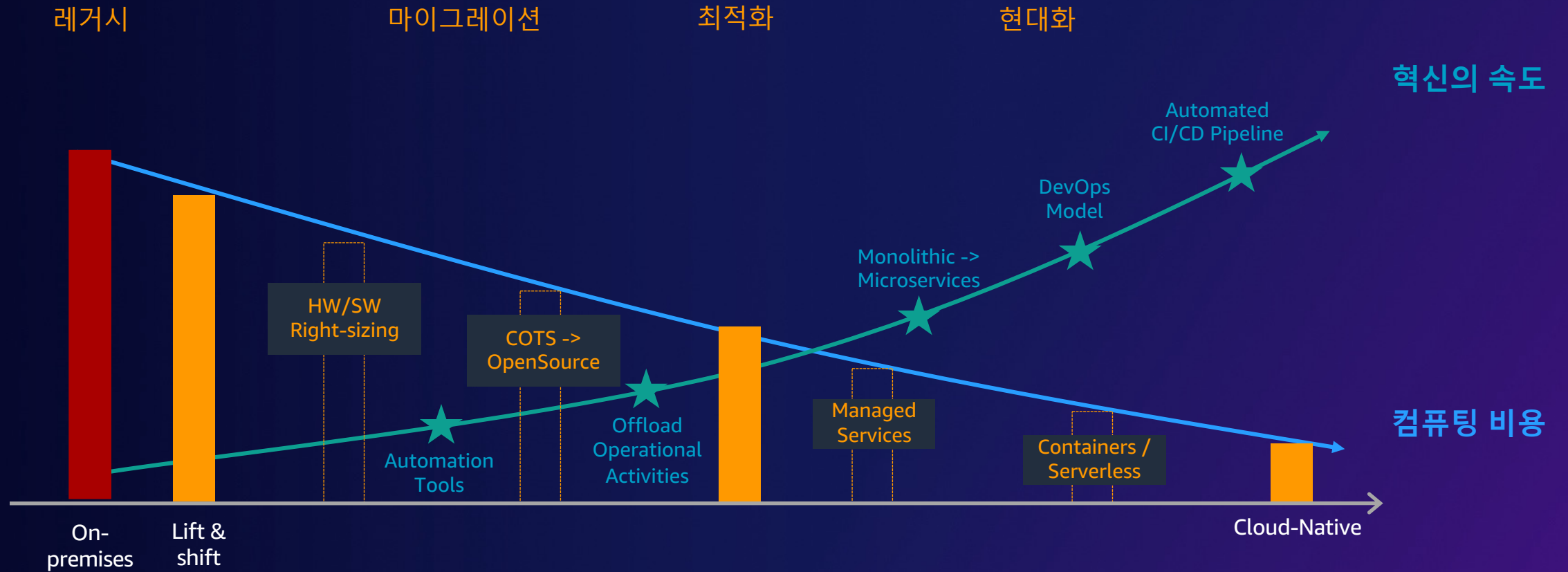
Migration and Modernization Patterns

AVERAGE CUSTOMER ENVIRONMENT, BY MIGRATION PATTERN *(based on AWS experience)*



Full spectrum of patterns is important for transformation – but up to ~50% of typical environment can be *rapidly migrated*, freeing time and budget to focus on modernization

마이그레이션부터 현대화까지

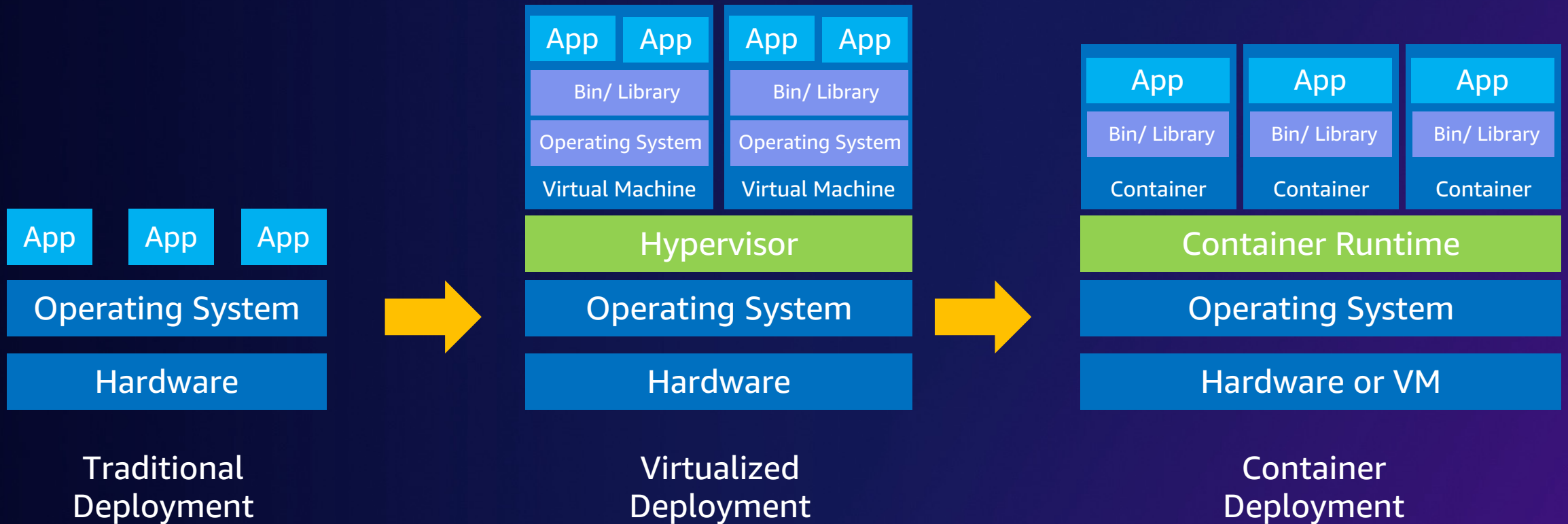


Refactor – Application Architecture

MODERNIZE YOUR APPLICATION TO UTILIZE AWS CLOUD COMPUTING



VM vs, Container



The Beauty of Container, voice from the customers

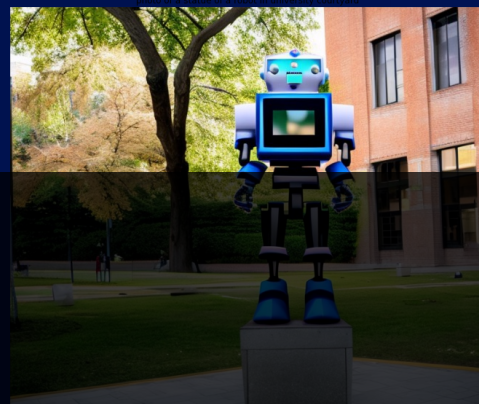
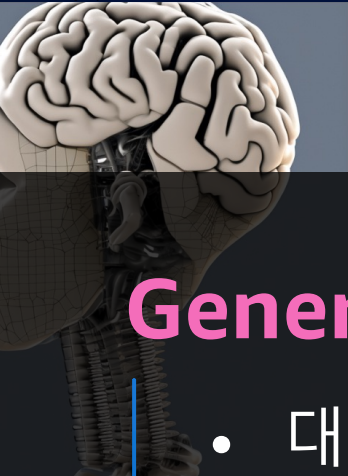
- Easy maintenance:
 - light weight -> it has a room to contain everything. i.e. container image includes configuration, code, server, runtime os.
 - Hence, it is portable so run everywhere such as Unix, Linux, dev, staging, and production.
- Cost effective:
 - Multiple containers run in a VM instance while a VM run one application.
 - Kubernetes orchestrate the containers.

Gen AI

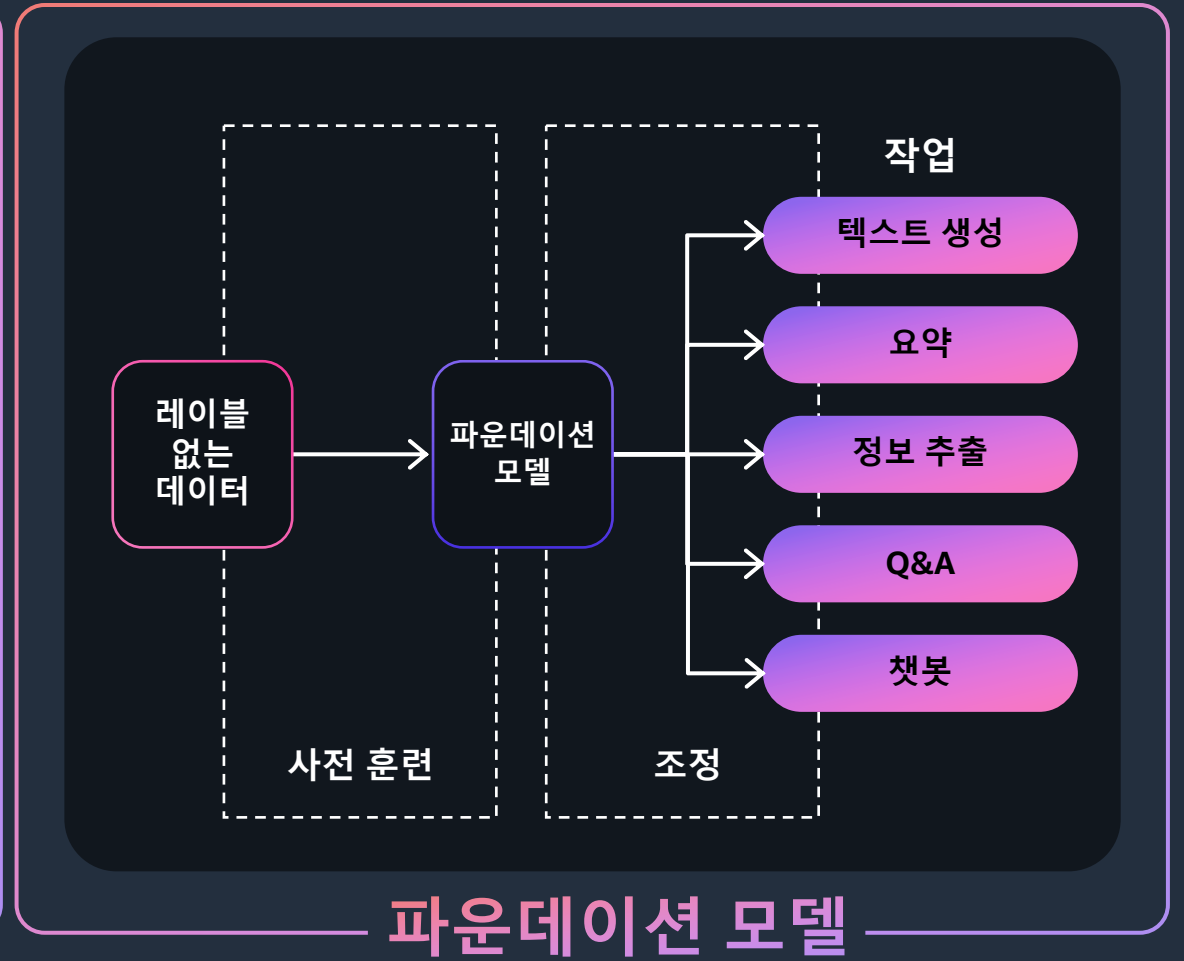


Generative AI (생성형 AI)

- 대화, 스토리, 이미지, 동영상, 음악 등 **콘텐츠**와 **아이디어**를 생성합니다.
- 방대한 데이터 뭉치들에 대해 사전 학습된 대규모 모델을 기반으로 하며, 일반적으로 **Foundation Models(FM)**이라고 합니다.



파운데이션 모델과 다른 기계 학습(ML) 모델의 차이점



생성형 AI 모델 활용

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

- Amazon Q
- Amazon Q in Amazon QuickSight
- Amazon Q in Amazon Connect
- Amazon CodeWhisperer

TOOLS TO BUILD WITH LLMs AND OTHER FMs

- Amazon Bedrock
- Guardrails | Agents | Customization capabilities

생성형 AI 모델 개발

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

- GPUs
- Trainium
- Inferentia
- SageMaker
- EC2 UltraClusters
- EFA
- EC2 Capacity Blocks
- Nitro
- Neuron

Amazon Bedrock FM 지원



Amazon Titan

Text summarization, generation, classification, open-ended Q&A, information extraction, embeddings and search

AI21labs

Jurassic-2

Multilingual LLMs for text generation in Spanish, French, German, Portuguese, Italian, and Dutch

ANTHROPIC

한국어

Claude 3 & Claude Instant

LLM for conversations, question answering, and workflow automation based on research into training honest and responsible AI systems



한국어

New

Command and embed

Text generation model for business applications and embeddings model for search, clustering, or classification in 100+ languages

∞ Meta

New

Llama 3

Fine-tuned models ideal for dialogue use-cases and language tasks

stability.ai

New

Stable Diffusion XL 1.0

Generation of unique, realistic, high-quality images, art, logos, and designs

Now Available in Amazon **Bedrock**

ANTHROPIC

Claude 3.0

Multimodal
업계에서 가장 높은 수준의 **200K**
token context window 지원

모델 사이즈에 따른
Opus, Sonnet, Haiku 제공

엔터프라이즈급 보안/데이터 핸들링
탈옥과 오용에 대한 내성이 10배 증가

Claude 3.0

Claude 3 benchmarks

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge MMLU	86.8% 5-shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning GPQA, Diamond	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math GSM8K	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% MajI@32	86.5% MajI@32
Math problem-solving MATH	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math MGSM	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code HumanEval	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text DROP, F1 score	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations BIG-Bench-Hard	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A ARC-Challenge	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge HellaSwag	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

농심 매운새우깡 400g 대용량 봉지스낵

한번 먹기 시작하면 끝없이 먹게 되는 새우깡!
남녀노소 모두 좋아하는 스낵~

영양정보 총 내용량 400g / 30g 당 145kcal

※ 상품의 포장 디자인은 변경될 수 있습니다.



영양정보		총 내용량 400g 30g 당 145 kcal	
30g 당	1일 영양성분 기준치에 대한 비율	100g 당	
나트륨 220 mg	11%	720 mg	36%
탄수화물 19 g	6%	64 g	20%
당류 2.4 g	2%	8 g	8%
지방 7 g	13%	25 g	46%
트랜스지방 0 g		0 g	
포화지방 2.3 g	15%	8 g	53%
콜레스테롤 5 mg미만	1%	5 mg미만	1%
단백질 2 g	4%	7 g	13%
칼슘 36 mg	5%	121 mg	17%

1일 영양성분 기준치에 대한 비율(%)은 2,000 kcal 기준이므로 개인의 필요 열량에 따라 다를 수 있습니다.

제품명	새우깡	식품유형	과자(유치리제품)
원재료명	위밍프레이크[중국산:소맥분, 옥수수전분, 새우, 맛에이스즈미 분말, 혼합제제(타피오카산화전분, 말트덱스트린)], 미강유(태국산), 팜유(말레이시아산), 새우풍미유, 염미시즈닝 밀, 새우, 대두, 우유 함유		
유통기한	후면 표기일까지	포장재질	플러브로필렌/지퍼·폴리에틸렌
등록번호	197201540013(부산), 2001054954121(구미)		
<p>• 이 제품은 쇠고기, 닭고기, 토마토, 계, 달걀, 오징어, 잣, 땅콩, 계란, 조개류(굴, 전복, 홍합 포함)를 사용한 제품과 같은 제조 시설에서 제조하고 있습니다. • 본 제품은 소비자 기호에 따라 교환, 환불해 드립니다. • 보관상 주의사항: 직사광선을 피하고 서늘하고 건조한 곳에 보관하십시오. 흡습되기 쉬운 제품으로 개봉 후 바로 드십시오. • 간혹 과자에 있는 검은 물집은 새우의 껍질 등이 부분적으로 검게 보이는 것으로 안심하고 드셔도 됩니다. • 부정·불량식품 신고는 국번없이 1399 지파:PE</p>			

Claude 3.0

Claude 3 benchmarks

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge <i>MMLU</i>	86.8% 5-shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning <i>GPQA, Diamond</i>	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math <i>GSM8K</i>	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% MajI@32	86.5% MajI@32
Math problem-solving <i>MATH</i>	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math <i>MGSM</i>	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code <i>HumanEval</i>	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text <i>DROP, F1 score</i>	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations <i>BIG-Bench-Hard</i>	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A <i>ARC-Challenge</i>	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge <i>HellaSwag</i>	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot



Amazon Bedrock > Chat playground

Chat playground Info

A Claude 3 Sonnet v1 | ODT

[Change](#)

이 이미지에 대해 설명해주세요

이 이미지는 농심 매운새우깡 400g 제품의 패키지와 영양성분표를 보여주고 있습니다.

빨간색 패키지 앞면에는 매운 새우 이미지와 함께 "매운새우깡"이라는 제품명이 크게 적혀 있습니다.

패키지 뒷면에는 영양정보 테이블이 있어 1회 제공량 30g 기준 열량, 나트륨, 탄수화물, 당류, 지방, 포화지방, 콜레스테롤, 단백질 등의 영양소 함량을 보여줍니다.

또한 원재료와 유통기한, 제조업체 정보 등이 한글과 영어로 제공되어 있습니다. 패키지 하단에는 재활용 마크와 고객상담실 번호도 있습니다.

Claude 3.0

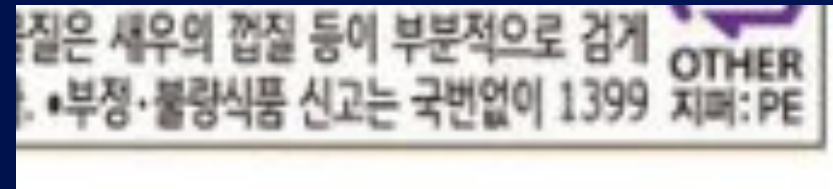
Claude 3 benchmarks

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge <i>MMLU</i>	86.8% 5-shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning <i>GPQA, Diamond</i>	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math <i>GSM8K</i>	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving <i>MATH</i>	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math <i>MGSM</i>	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code <i>HumanEval</i>	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text <i>DROP, F1 score</i>	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations <i>BIG-Bench-Hard</i>	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A <i>ARC-Challenge</i>	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge <i>HellaSwag</i>	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot



이 이미지에서 고객상담실 번호를 알려주세요

농심 매운새우깡 400g 제품 패키지 하단에 고객상담실 번호가 1399라고 명시되어 있습니다.



Fine tune additional models in Amazon Bedrock

ANTHROPIC

COMING SOON

 cohere

 Meta

 amazon

Fine-Tuning – 파이썬 경험이 없어도 몇번의 클릭으로 가능

Amazon Bedrock > Fine-tune > Fine-tune model

Fine-tune a model [Info](#)

Select the foundation model you wish to fine-tune and submit your data location.

Model configuration

Source model
Select a model from a list of models that you can fine-tune.

Select a source model

Fine-tuned model name
Enter a name to identify the new fine-tuned model.

Enter a name here

▼ **Tags - optional**

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

No tags associated with the resource.

Add new tag

You can add up to 50 more tags.

Job configuration

Job name
Enter a name to identify the training job.

Enter a name here

▼ **Tags - optional**

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

No tags associated with the resource.

Input data [Info](#)

S3 location

s3://bucket/path-to-your-data/ View Browse S3

Validation dataset S3 location

s3://bucket/path-to-your-data/ View Browse S3

Hyperparameters [Info](#)

Epochs
The total number of iterations of all the training data in one cycle for training the model.

10

Batch size
The number of samples processed before the model is updated.

1

Learning rate
The step size for incrementing parameters at each iteration.

0.00005

Learning rate warmup steps
Number of iterations over which learning rate is gradually increased to the initial rate specified.

0

Output data [Info](#)

S3 location

s3://bucket/path-to-your-data/ View Browse S3

Service access [Info](#)

The fine-tuning job requires permission to access your S3 buckets.

Choose a method to authorize Bedrock

- Use an existing service role
- Create and use a new service role

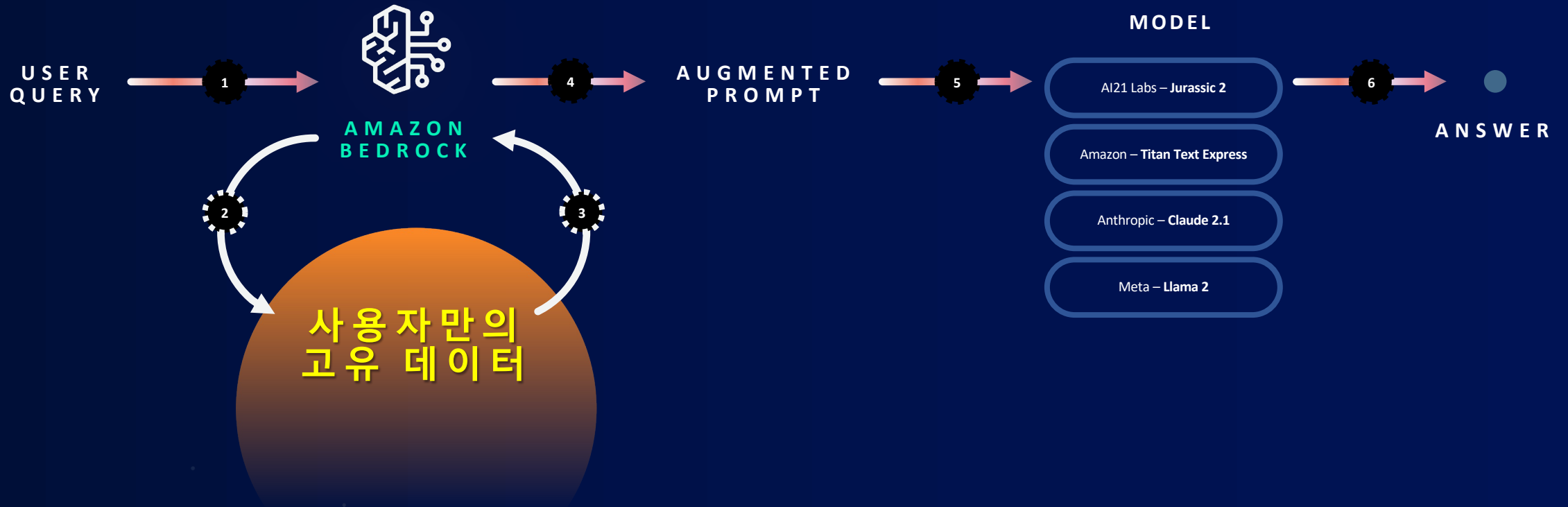
Service role

Choose role

Cancel **Fine-tune model**

검색 증강 생성 (RAG - Retrieval Augmented Generation)

Foundation Model을 **User의 데이터** 소스에 안전하게 연결하여 보다 **관련성 높은 응답 결과** 제공하는 방법



다양한 Vector DB 연계를 통한 RAG 아키텍처 구현

Vector databases for **Amazon Bedrock**



Vector Engine For
Amazon OpenSearch
Serverless



Redis Enterprise
Cloud



Pinecone



Amazon
Aurora

COMING SOON



MongoDB

Responsible AI 정책에 따른 맞춤형 보호 장치 구현

NEW

Guardrails for Amazon Bedrock

책임감 있는 AI 정책으로 생성형 AI 애플리케이션을 보호하세요.

AVAILABLE IN PREVIEW

AI 정책에 따라 유해 콘텐츠 필터링을 간편하게 구성하세요.

모든 FM 또는 상담원에게 가드레일 적용

FM 응답에서 PII 정보 삭제(Coming Soon)

Guardrails for Amazon Bedrock

Amazon Bedrock > Guardrails > antje-banking-assistant > Working Draft

Working draft: antje-banking-assistant

Create version Test

Denied topics (1)

Find versions < 1 > ⚙️

Name	Definition	Example phrases
Investment advice	Investment advice refers to inquiries, guidance, or rec...	2 phrases

Content filters

Prompt filters	Response filters
ON	ON
Hate filter strength for prompts Medium	Hate filter strength for responses High
Insults filter strength for prompts Low	Insults filter strength for responses Medium
Sexual filter strength for prompts Medium	Sexual filter strength for responses Medium
Violence filter strength for prompts High	Violence filter strength for responses High

Blocked messaging

Blocked prompts	Blocked responses
Sorry, I can't comment on that.	Sorry, I can't comment on that.

Test

Working draft

AI Claude Instant v1.2 ODT

Prompt

Should I open a credit card account?

Model response

Here are a few factors to consider when deciding whether to open a credit card account:

- Credit history - If you have no or little credit history, a credit card can help you establish credit. Just be sure to use it responsibly.
- Interest rates - Only open a card if you're confident you can pay off balances in full each month to avoid high interest charges. Introductory 0% rates may help you build credit without interest.

Final response

Here are a few factors to consider when deciding whether to open a credit card account:

- Credit history - If you have no or little credit history, a credit card can help you establish credit. Just be sure to use it responsibly.
- Interest rates - Only open a card if you're confident you can pay off balances in full each month to avoid high interest charges. Introductory 0% rates may help you build credit without interest.

Guardrail check

Passed [View trace >](#)

Run

Edit content filters

Filter strengths for prompts

Filter strength determines the degree of filtering. A higher filter strength increases the likelihood of filtering harmful content from the given category.

Enable filters for prompts

Category	None	Low	Medium	High
Hate	None	Low	Medium	High
Insults	None	Low	Medium	High
Sexual	None	Low	Medium	High
Violence	None	Low	Medium	High

Filter strengths for responses

Filter strength determines the degree of filtering. A higher filter strength increases the likelihood of filtering harmful content from the given category. These filters evaluate and override model responses. They don't modify the model behavior.

Enable filters for responses

Category	None	Low	Medium	High
Hate	None	Low	Medium	High
Insults	None	Low	Medium	High
Sexual	None	Low	Medium	High
Violence	None	Low	Medium	High

Cancel Save Save and exit



Thank you!

정연

ryouc@amazon.com