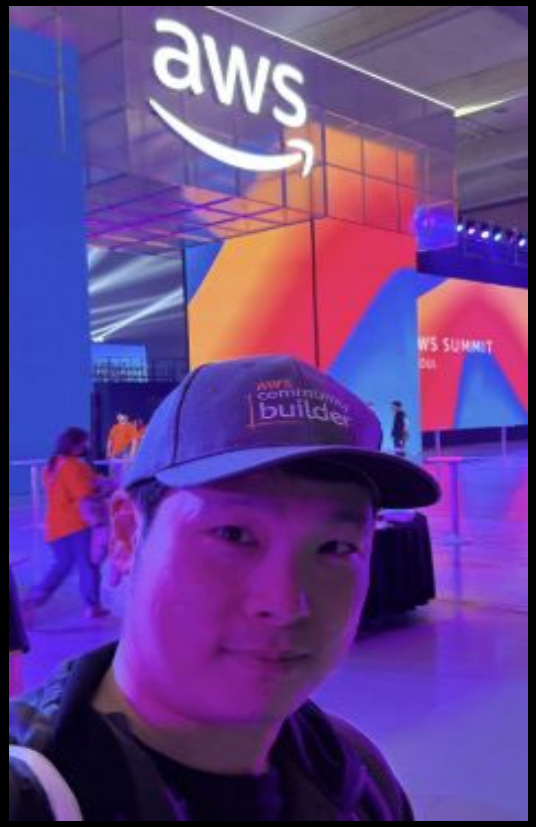
Private한 환경에서 Amazon Bedrock 구성하기

연사자 소개



김동일

- Serverless Community Builder (2025~)
- AWS KRUG Architecture 소모임 오거나이저 (2024~)
- 생활가전 전문 기업 백엔드 개발 및 DevOps
- 데옵린이's 기술 노트 블로그 운영

블로그

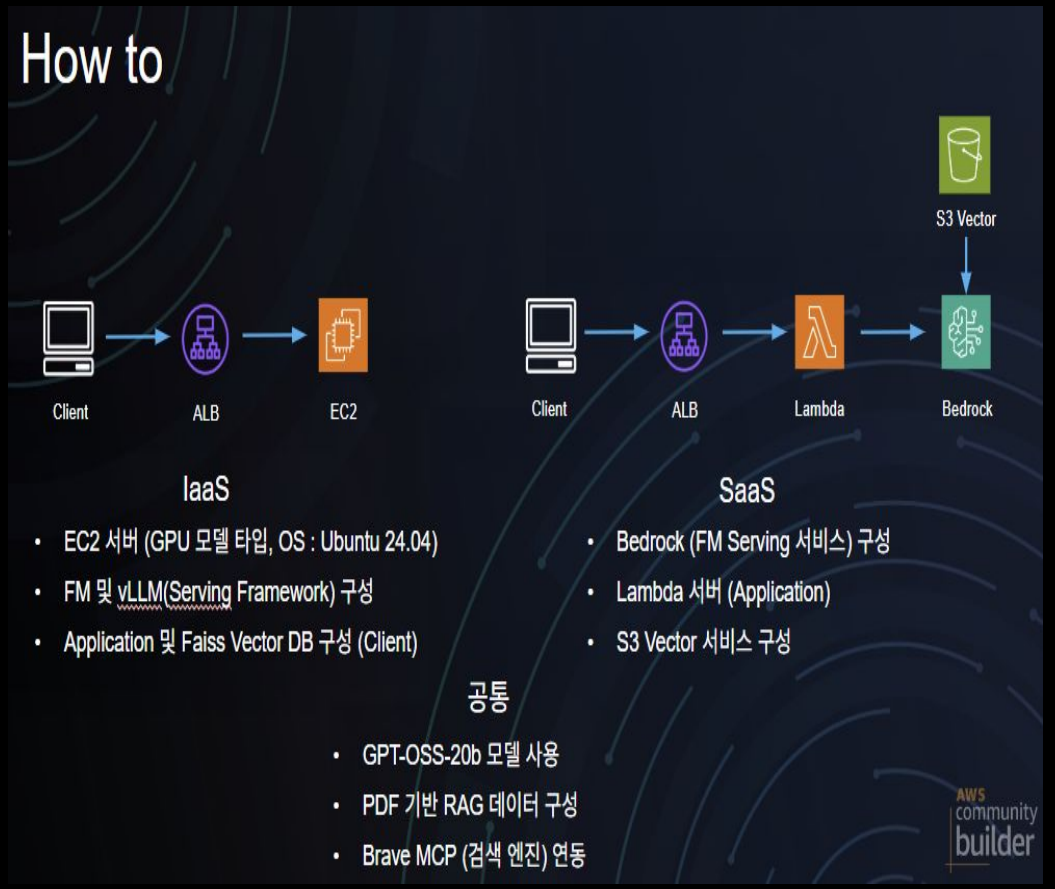
링크드인

Contents

- 시작하기 전에
- Amazon Bedrock이란
- 구성해 봅시다.
    - Case 1
    - Case 2
    - Case 3
- 결과
- AI를 잘 사용하기 위한 Tips

시작하기 전에...

Previous



Previous

```
🧠 Agent sending input to model: AWS Summit 중, aws와함께생성형AI로비즈니스혁신하기 세션 내용 요약해줘
Score: 0.8105762004852295
Content: © 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.
AWS와 함께
생성형 AI로 비즈니스혁신하기
양 승 도
D 1 T 1 S 1
솔루션즈 아키텍트 리더
AWS...
Source: {'producer': 'Microsoft® PowerPoint® Microsoft 365용', 'creator': 'Microsoft® PowerPoint® Microsoft 365용', 'creationdate': '2024-05-30T10:11:55+09:00', 'moddate': '2024-05-30T10:11:55+09:00', 'source
: '/mnt/d/personal_study/aws/2024/summit/d1t1s1-aws와함께생성형AI로비즈니스혁신하기.pdf', 'total_pages': 43, 'page': 0, 'page_label': '1'}
```

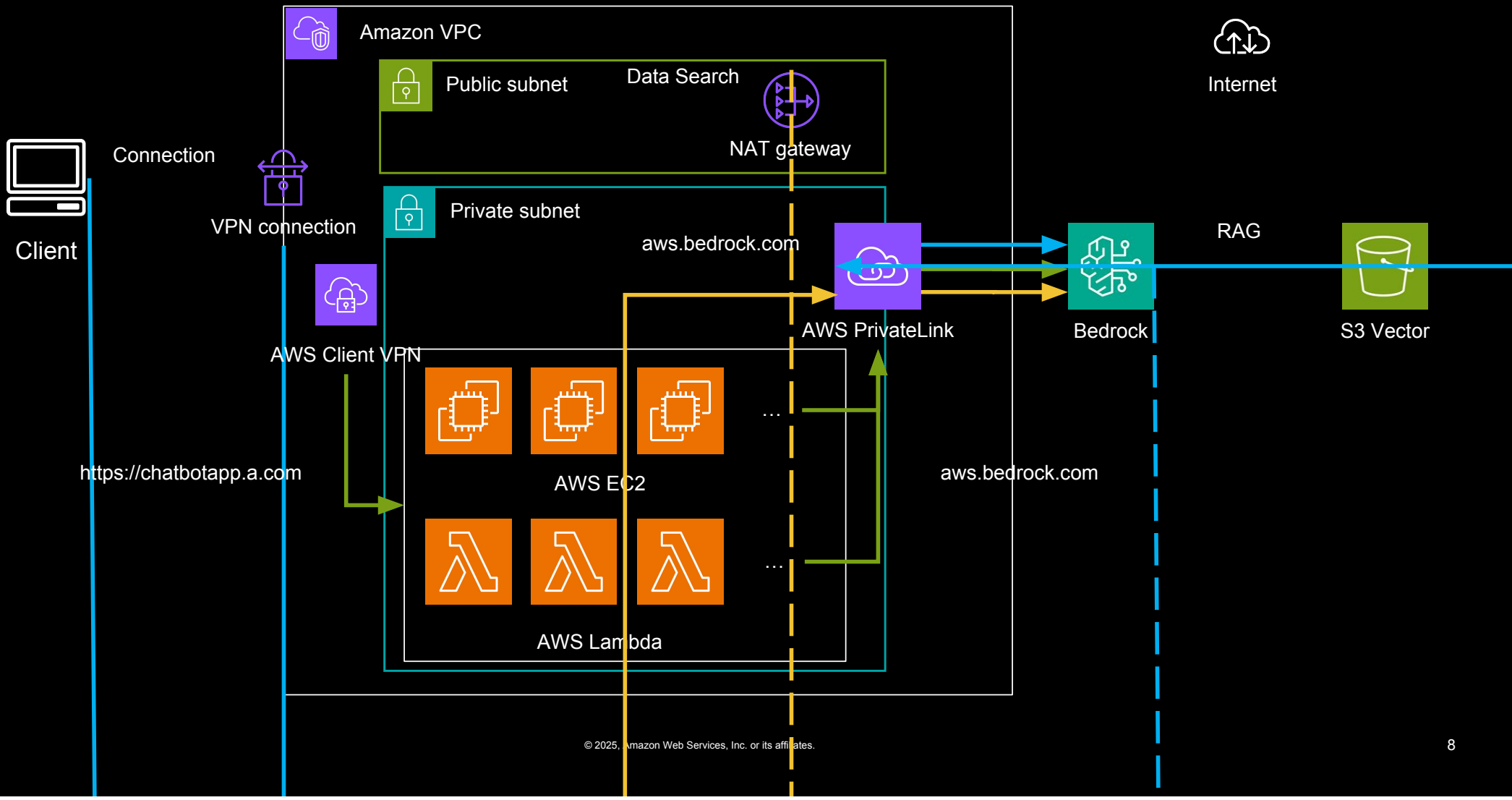
	IaaS	SaaS
리소스	EC2	Lambda + Amazon Bedrock
RAG	Faiss Vector DB, Elasticsearch, Postgresql 등	Opensearch Serverless, Aurora(Postgresql), S3 Vector
MCP 연계	가능	가능
장점	사용자 정의 리소스 구축 가능(폐쇄망, 하이브리드 등) 파인튜닝을 통한 CustomModel 생성 가능	비용 최적화(호출 건수/Token 당 비용 등) 상대적으로 쉬운 구축
단점	시간당 비용이 절대적으로 비쌈(GPU 모델은 특히…) 구축 난이도가 상대적으로 높음	Managed 서비스 간의 제약 사항 존재 파인튜닝을 통한 CustomModel 생성이 제한적

방향성을 설정하고, 최적의 아키텍처를 구축

AWS community builder

Serverless를 Private하게 구성할 수 있나요?

네. 가능합니다.



Amazon Bedrock 이란

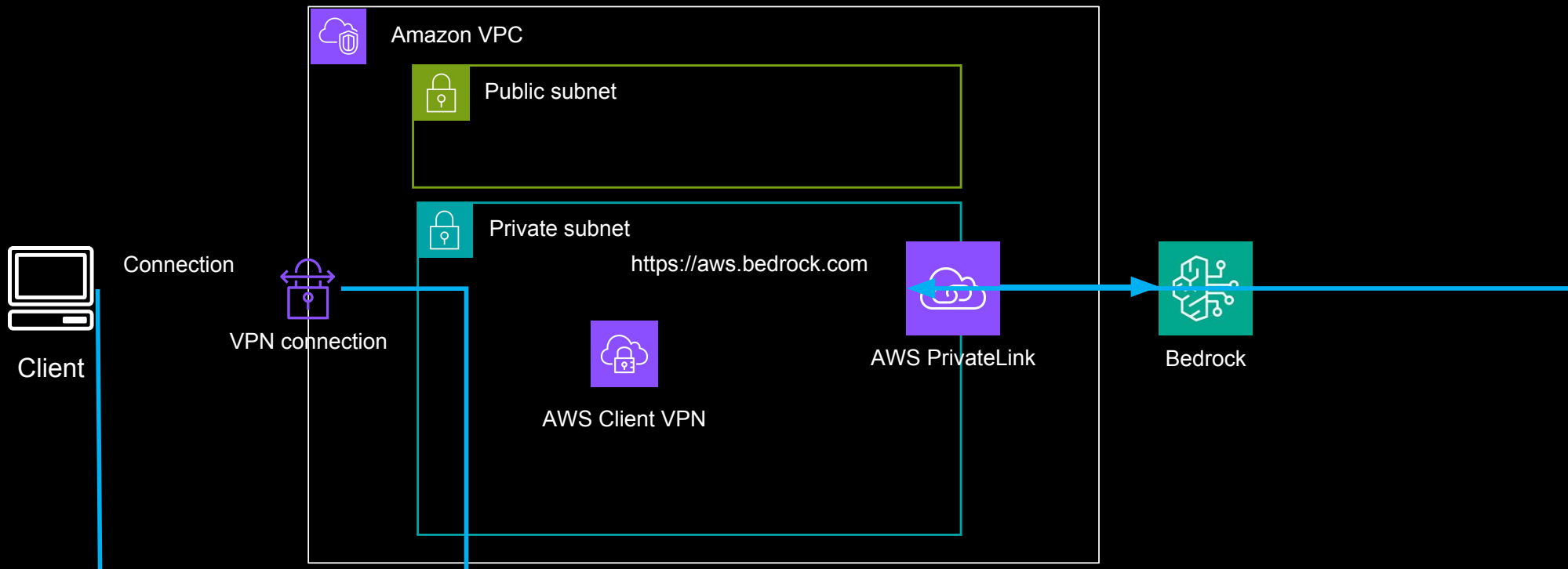
Amazon Bedrock 이란

- Amazon Web Services(AWS)에서 제공하는 완전 관리형 생성형 AI 서비스
- 파운데이션 모델(FM)을 사용하여 생성형 AI 애플리케이션을 구축하고 확장할 수 있도록 지원
- 주요 특징
  - SLM부터 고성능 LLM까지 다양한 모델 지원 : Amazon, Anthropic, Meta, Stability AI, Google, OpenAI 등 전세계 유명한 모델을 통합 API로 제공
    (한국에서 만든 FM도 있습니다.^^)
  - 서버리스(Serverless) 기반 서비스 제공 : 별도의 인프라를 관리할 필요가 없으며, 사용한 만큼만 비용을 지불(Token 당), 운영 효율성이 높음.
  - 엔터프라이즈급 보안 및 프라이버시: 데이터가 기본 모델 학습에 사용되지 않도록 보호, AWS의 강력한 보안 인프라를 통해 데이터를 안전하게 관리
  - 에이전트(Agents) 및 지식 기반(Knowledge Bases) 기능 제공
    - Agents: 사용자의 의도를 이해하고 시스템 API를 호출하여 복잡한 작업을 자동으로 수행하는 AI 에이전트를 구축
    - Knowledge Bases: 자체적인 내부 데이터를 연결하여 최신 정보에 기반한 정확한 답변을 생성하는 RAG(검색 증강 생성) 기능을 쉽게 구현
  - 사용자 지정 모델 사용 지원 (Customization): 미세 조정(Fine-tuning) 및 지속적인 사전 학습 기능을 통해 모델 최적화 가능
  - API Key 제공 : 단기/장기 API키를 지원하여 목적에 맞는 API Key 제공

이 외에 강력한 기능들이 많습니다.(특히 Agent 관련한 기능들...)

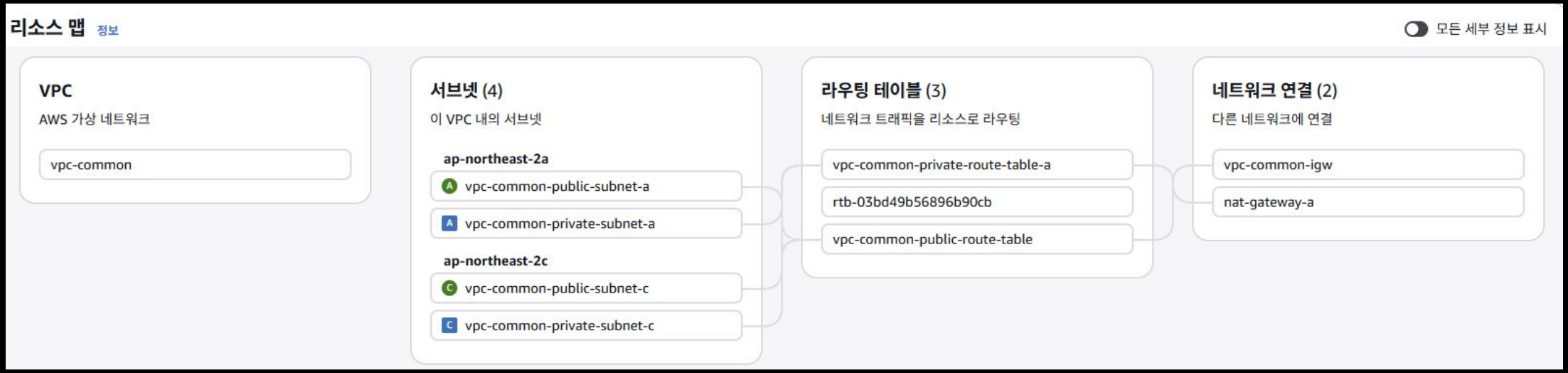
구성해 봅시다.

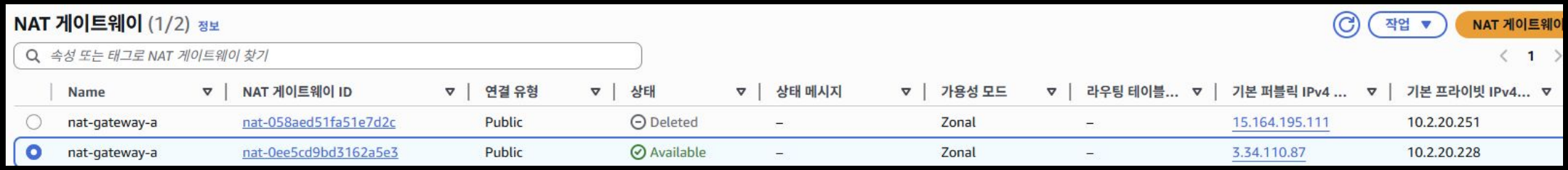
Case 1



- VPC 구성(Public/Private Subnet, NAT Gateway)
- VPN Connection Client VPN, PrivateLink 구성
- Bedrock (FM Serving 서비스) 구성
- Local PC에서 Application 구성

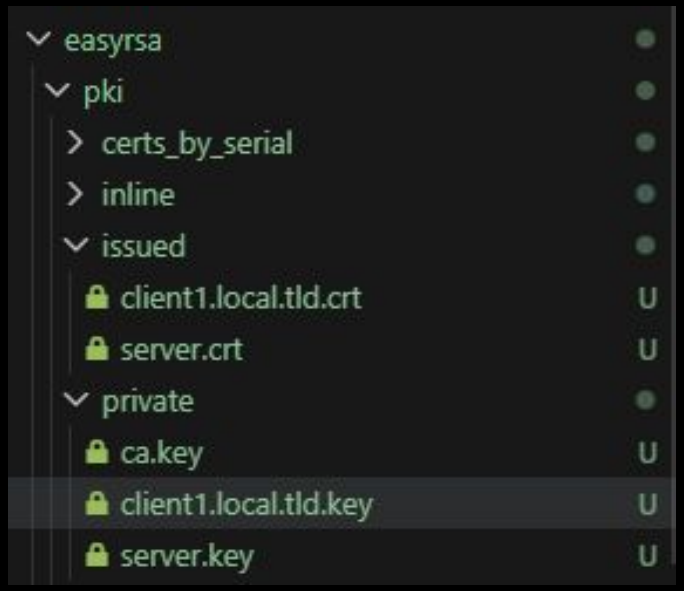
VPC 구성





VPN Connection 구성 - 인증서 구성

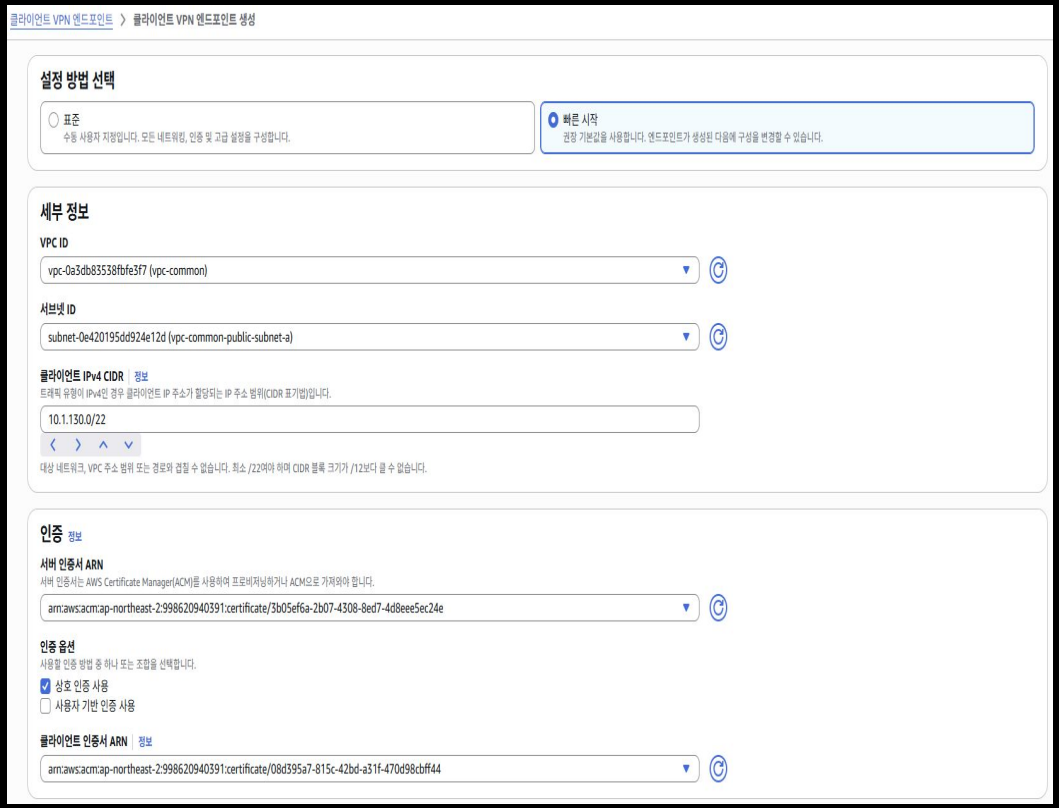
- EasyRSA를 통한 서버/클라이언트 인증서 생성

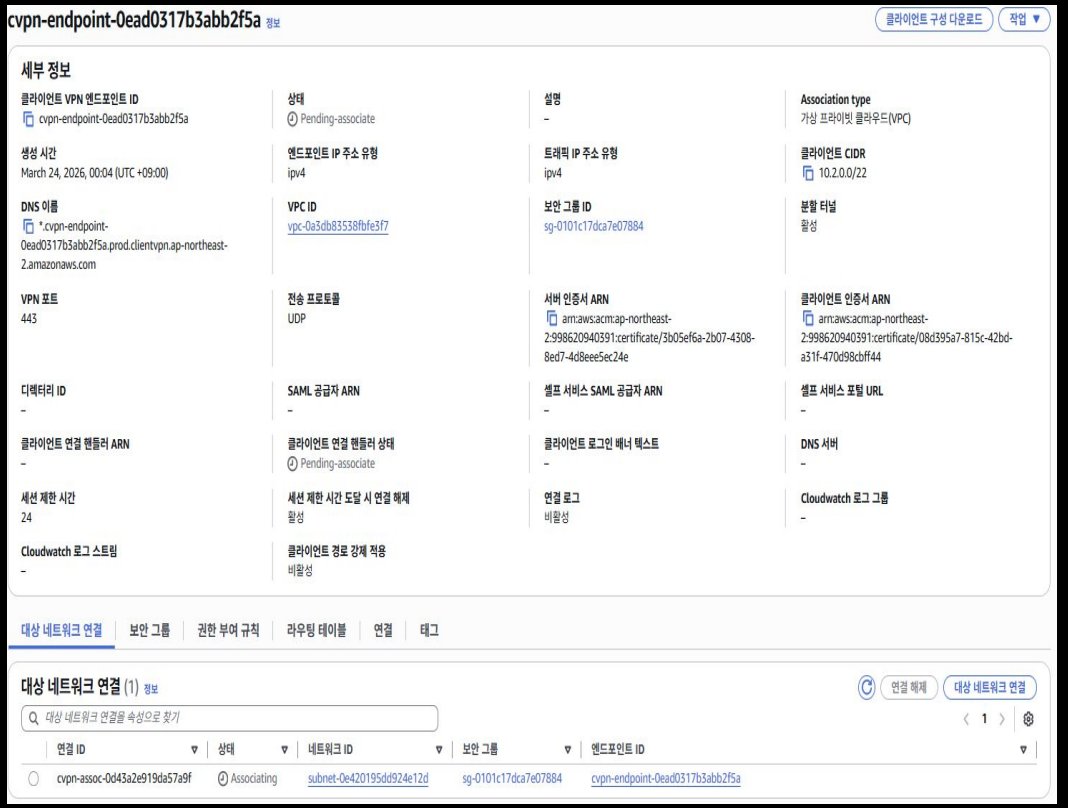


인증서 ID	도메인 이름	유형	상태	사용 중	갱신 자격	키 알고리즘
08d395a7-815c-42bd-a31f-470d98cbff44	client1.local.tld	가져옴	발급됨	아니요	부적격	RSA 2048
3b05ef6a-2b07-4308-8ed7-4d8eee5ec24e	aws-server	가져옴	발급됨	아니요	부적격	RSA 2048

VPN Connection 구성 - 클라이언트 VPN 엔드포인트

- 서버/클라이언트 인증서 및 클라이언트 IP4 CIDR 설정 후 클라이언트 VPN 엔드포인트 생성 (클라이언트 연결에 약 15분 내외 소요)
- 클라이언트 구성 다운로드

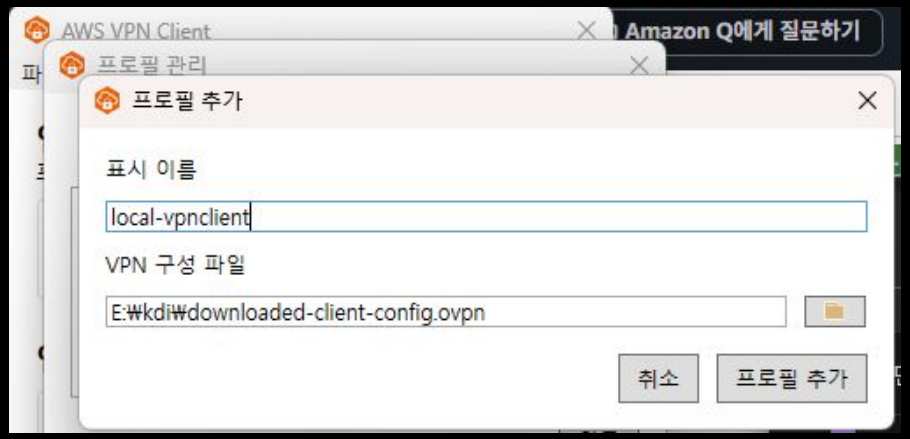


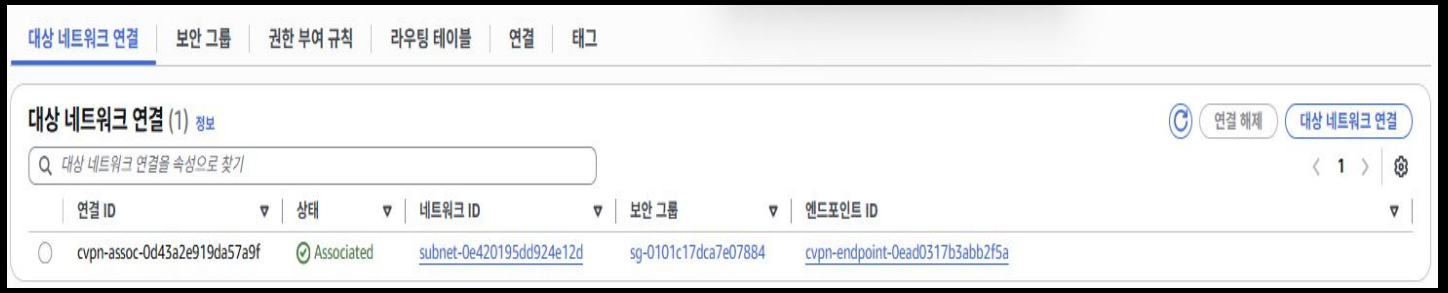


VPN Connection 구성 - 클라이언트 VPN 연결

- 다운로드 받은 VPN 구성 파일을 열어 클라이언트 인증서의 cert와 key 추가
- AWS VPN Client 프로그램에 구성파일에 대한 프로필 추가 및 연결

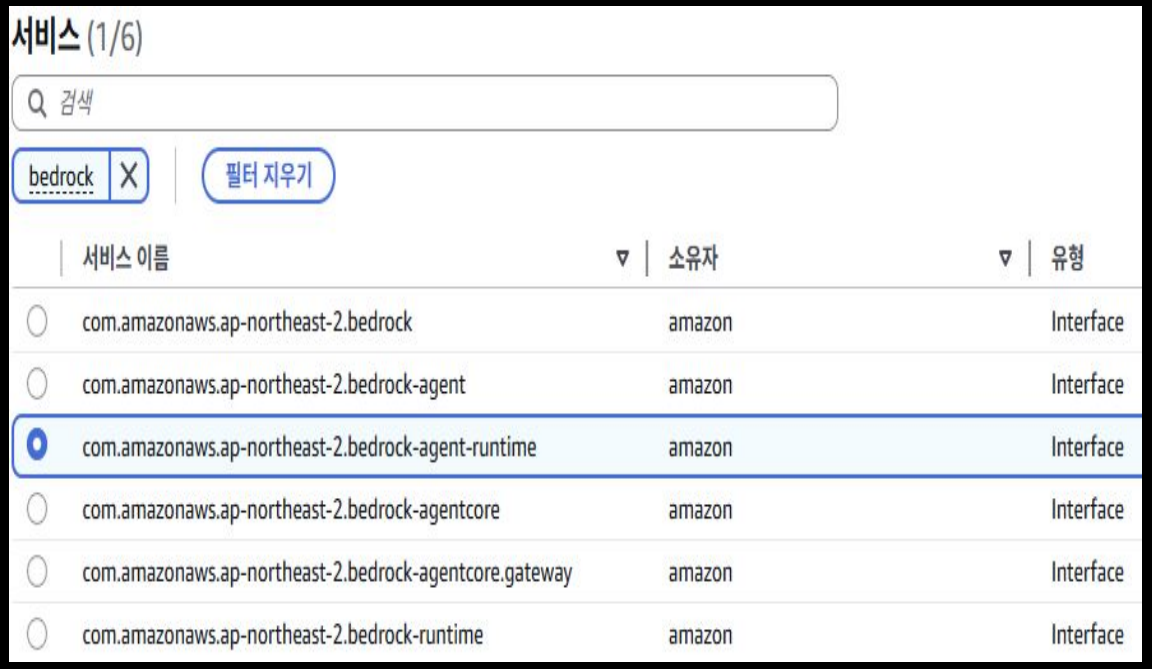
```
<cert>
-----BEGIN CERTIFICATE-----
MIIDZDCCAkygAwIBAgIRALgrI8OEyuBQuLbg9wrHpucwDQYJKoZIhvcNAQELBQAw
GDEWMBQGA1UEAwwNbG9jYWwta2RpMDkxMzAeFw0yNjAzMjMxNDM4MzFaFw0yODA2
MjUxNDM4MzFaMBwxGjAYBgNVBAMMEWNsaWVudDEubG9jYWwudGxkMIIBIjANBgkq
hkiG9w0BAQEFAAOCAQ8AMIIBCgKCAQEAI/f2m/pTMDRwyjZ74QJL3tix2xsLmDTm
wjbq7AoH7VugNuYe4+Acwky6v0RURPeihEI5yZ+5njLVhLtPIMnay/uctDRz7pDJ
A2lFisFJjaOnqAXprv3+e2Yh+BwskbFliavJZkAQr/m42H6vDqex/hNt/7utRwxq
QymeW/clWfFTL7Az+sLzYWJ4kEprG58sRmGyQEHWzyrq855kwj1s8VqsZxQOSiqs
Mk5nNc9rGlck5cZBOTPecfNZoQircvkAft+9JBPLWdHLDvH18ONLzdn+dYmX0V7p
hSRQV2BGss8H/VfEBpFxHPk0zs5DWHlLeqvBqAcX3mZtOktAcTn7TQIDAQABo4Gk
MIGhMAkGA1UdEwQCMAAwHQYDVR0OBBYEFCMm4NO5tLUenHF+yiPBwwRPqqimMFMG
A1UdIwRMMEqAFCiQbQeLeaM6e1AxO7wVm42wqSwXoRykGjAYMRYwFAYDVQQDDA1s
b2NhbC1rZGkwOTEzghQnLeKWT2mpqozwq+8wz7iI+Ve2HDATBgNVHSUEDDAKBggr
BgEFBQcDAjALBgNVHQ8EBAMCB4AwDQYJKoZIhvcNAQELBQADggEBAJPLLlqBSo1x
Zv7dogk212GxdJtob9T8u36XjRNoN1cevzUBGpwZrUsQ/cGA+VnWmHcOxwsx7aes
4qkN69OVGIWHn6X7jjAz2PKD3m47yOCE9LXeUzufYgxNJh+Vsu0UKqElfAccD1JE
RDyHNi5/WQcMbB3N5igtyV6HOHk91c3WapPQS022gT3KPvJ2D9GXM/Ev/1Mb7gdA
V9zjDfkCgJg91Q6gCJNcoU8ZcXTvbIol/f165INxTm8CSiJRJ827lWJlZsmp9LQX
Plc0rTPIzTWBx050qVqxCMIiN+kxBfkqqPKw5PutFPCrEOn5GNZafMiGAnY3SltV
GIJ393AzHR0=
-----END CERTIFICATE-----
</cert>

<key>
-----BEGIN PRIVATE KEY-----
MIIEvgIBADANBgkqhkiG9w0BAQEFAASCBKgwggSkAgEAAoIBAQCX9/ab+IMwNHDK
NnvhAkve2LHbGwuYNObCNursCgftW6A25h7J4BzCTLq/RFRE96KEQjnJn7meMtWE
u08gydrL+5y0NHPukMkDaUWKwUmNo6eoBemu/f57ZiH4HCyRsWWJq8lmQBCv+bjY
fq8Op7H+E23/u61HDGpDKZ5b9whZ8VMvsDP6wvNhaPiQSmsbnyxGYbJAQdbPKurz
nmTCPWwFWqxnFA5KKqwyTmc1z2sYhyTlxkE5M95x81mhCKty+QB+370kE8tZ0csO
8fXw40vN2f51iZfRXumFJFBXYEayzwf9V8QGkXEc+TTOzkNYeUt6q8GoBxfeZm06
S0BxOftNAgMBAAECggEACOFo5BsH/vZli++MpCy77ckYdZ8aOZoHzRwj9MiuR2jW
Dzc54WOgkpskw5YU3ZXiBZwQH+/pSy0ZzKCcgmRBWK+6q84v4yZLVNtzQKVEbNY+
LyOvUPZPR3tDq2nMEfR9pEKpaYP/ZaUwb9hRVwgmLtyITkwzl0YKtKOLOUERmB3a
yueOX9Aj54X2t+YzT+mBABmUIBRjGmE57qF5y35ubyxR1kZVanPRQYU5U4txJSj4
OYbj6iuda1gmWkdxnBHMDoyRDuAX8BGTBBm9m9lz17QZEkNoLA1iaKrRAWkFNhC4
4VQVQzgTzbH/4d1O8wKJEftlUDbpzuAD+1TYCsTYEQKBgQDFyg8mnVwitokSFhfj
QvOLUkdz2NQo3xzOG6n5wLqi1SRoz/U+ezTrYGOCbc5K+bjmTFcv6MpFmnGbfPSw
JdpNBqAfiQ7BcMEjzzKDQD7UCnhEduJf68Kv3HIjGniTFtTVwcaazEt6Hm76BQha
c70ZQ5SbX51SALpW2p4gmmc5kQKBgQDEsainZsJLd7rvfm5GqjVrRR0Ni9/464mr
86bISk6I18rX4T17HQCzrdbZs1CKnU1kQAVW3MS7/QTc6UBS97tG+USa6z2JCff/
n42Z2u9BMMLRP9hnaOguOZKiOPnopSgBZiITO2udlulov8RZ8NGDrngJ23DF/z2A
c06Ul+En/QKBgQCYLbcqhrh1HOyptcX7XW/1KgE/6PyT1dZb3n0RZHA1CLlSy25t
```





PrivateLink 구성

- VPC -> VPC Endpoint 에서 PrivateLink 구성
- PrivateLink 서비스는 {prefix}.bedrock-runtime, {prefix}.bedrock-agent-runtime Interface Endpoint 선택
  (Agentic AI 구성을 원할 경우, {prefix}.bedrock-agentcore 고려)
- 사전 구성된 Private Subnet ID를 가용영역으로 선택



Bedrock 구성

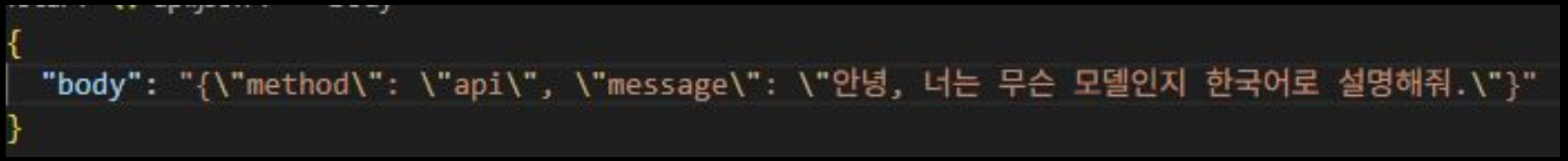
- FM 모델 :
  - 리전별로 사용 가능한 모델 범위가 다르고, 최대 사용 가능한 Token 수가 정해져 있으므로 팀 상황에 맞는 모델 선택
  - 대부분의 모델이 Serverless로 구성되어 있으며, Token 당 가격이 책정되어 있음(1,000 Token 당 In/Out 비용)
- Bedrock API Endpoint 호출 : 아래 3가지 방법 중 하나 선택
  - AWS SDK를 이용한 Endpoint 호출
  - Strends Agent(Framework)를 이용한 Endpoint 호출
  - 단기/장기 API 키를 이용한 Endpoint 호출
- Amazon Nova Lite 호출

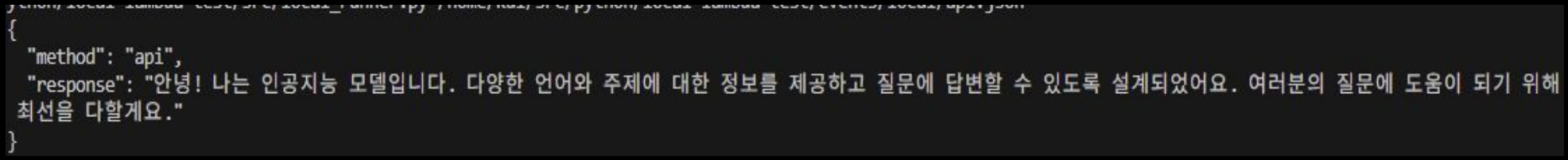
Region:

Asia Pacific (Seoul)

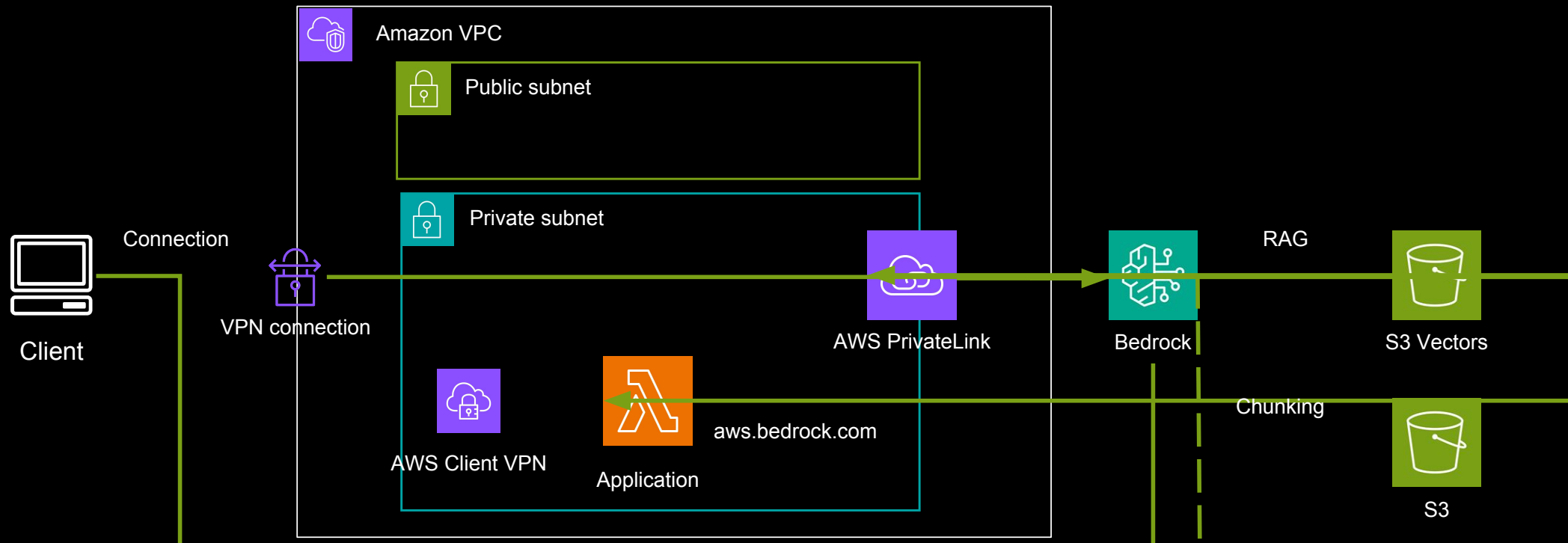
Anthropic models	Price per 1M input tokens	Price per 1M output tokens	Price per 1M input tokens (batch)	Price per 1M output tokens (batch)	Price per 1M input tokens (5m cache write)	Price per 1M input tokens (1h cache write)	Price per 1M input tokens (cache read)
Claude Sonnet 4.6	$3.00	$15.00	$1.50	$7.50	$3.75	$6.00	$0.30
Claude Sonnet 4.6 - Long Context	$6.00	$22.50	$3.00	$11.25	$7.50	$12.00	$0.60
Claude Opus 4.6	$5.00	$25.00	$2.50	$12.50	$6.25	$10.00	$0.50
Claude Opus 4.6 - Long Context	$10.00	$37.50	$5.00	$18.75	$12.50	$20.00	$1.00
Claude Opus 4.5	$5.00	$25.00	$2.50	$12.50	$6.25	$10.00	$0.50
Claude Haiku 4.5	$1.00	$5.00	$0.50	$2.50	$1.25	$2.00	$0.10
Claude Sonnet 4.5	$3.00	$15.00	$1.50	$7.50	$3.75	$6.00	$0.30
Claude Sonnet 4.5 - Long Context	$6.00	$22.50	$3.00	$11.25	$7.50	$12.00	$0.60

Local에서 Bedrock 호출





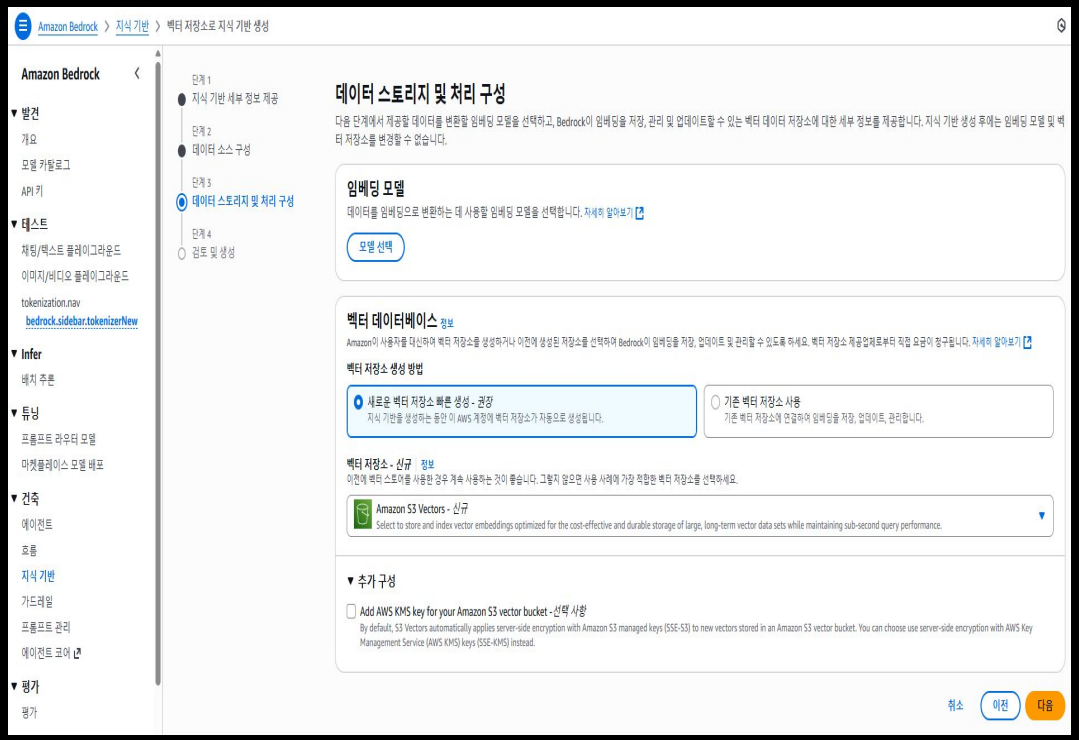
Case 2

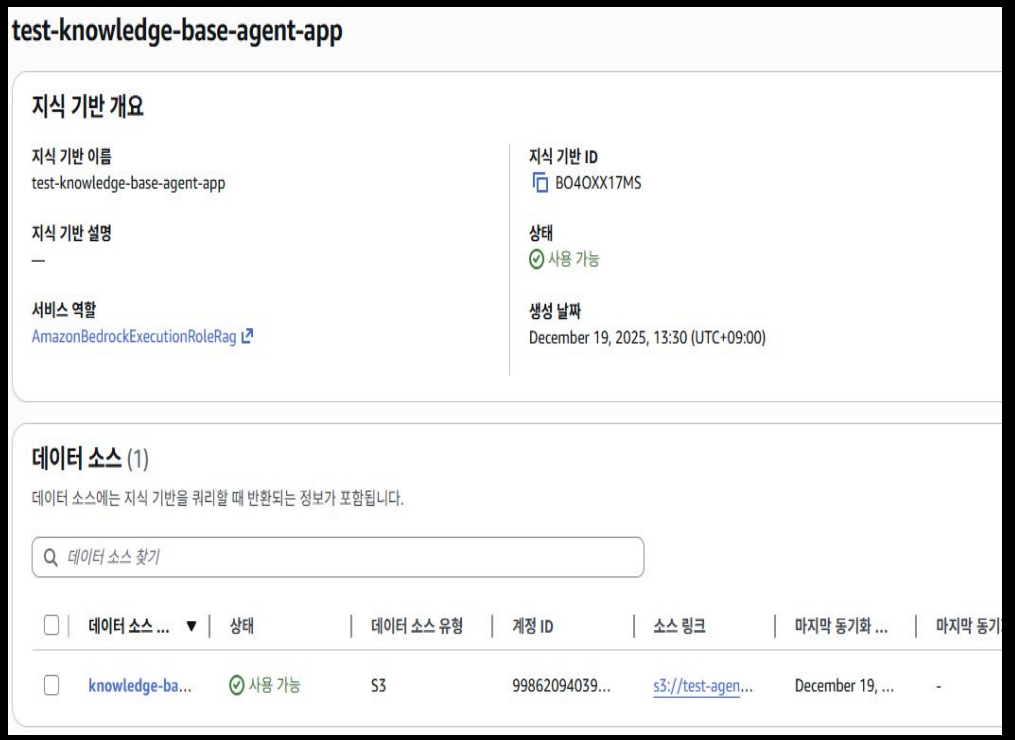


- VPC 구성(Public/Private Subnet, NAT Gateway)
- VPN Connection, PrivateLink 구성
- Bedrock (FM Serving 서비스) 구성
- S3 Vector(RAG) 구성
- Lambda에 Application 구성

RAG Chunking - S3 Vectors

- Vector 데이터를 객체 단위로 저장하고 조회하는 객체 스토리지
- 25년 리인벤트에서 처음 공개되었으며, 벡터 저장 및 쿼리 총 비용을 최대 90%까지 절감 가능
- Chunking할 RAW 파일을 S3에 업로드 후, Bedrock 지식기반에서 Vector Data 구성





Application 구성

1. Lambda 공통 Layer 추가 (strands-agent 관련 라이브러리)
2. 생성된 RAG의 모델ID 추가
3. RAG 호출 함수 생성

```python
def _build_message(message: str) -> str:
    search_result = brave_search.search(message)
    rag_result = bedrock_rag.retrieve(message)

    if not search_result and not rag_result:
        return message

    parts = []
    if rag_result:
        parts.append(f"[Knowledge Base]\n{rag_result}")
    # if search_result:
    #     parts.append(f"[Search Results]\n{search_result}")
    parts.append(f"[User Question]\n{message}")

    return "Based on the following context, answer the user's question.\n\n" + "\n\n".join(parts)


def lambda_handler(event, context):
    try:
        body = json.loads(event.get("body", "{}")) if isinstance(event.get("body"), str) else event
        method = body.get("method", "boto3")
        message = body.get("message", "")

        if not message:
            return _response(400, {"error": "message is required"})

        if method not in ENDPOINTS:
            return _response(400, {"error": f"Invalid method. Choose from: {list(ENDPOINTS.keys())}"})

        enriched = _build_message(message)
        print(enriched)
        endpoint = importlib.import_module(ENDPOINTS[method])
        answer = endpoint.invoke(enriched)

        return _response(200, {"method": method, "response": answer})

    except Exception as e:
        return _response(500, {"error": str(e)})
```

RAG 구성 결과

```
{
  "body": "{\"method\": \"strands\", \"message\": \"LLM 및 기타 FM을 활용하는 애플리케이션 관련 문서가 있으면 내용을 요약해줘.\"}"
}
```

```
[Knowledge Base]
[Score: 0.6032051399592643]
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.
<figure>
<figure_type>Diagram</figure_type>
LLM 및 기타 FM을 활용하는 애플리케이션
</figure>
<figure>
<figure_type>Diagram</figure_type>
LLM 및 기타 FM을 사용하여 구축할 수 있는 도구
| AWS 서비스 | 설명 |
| - | - |
| Amazon Bedrock | 생성형 AI 모델을 구축하고 호스팅하는 서비스 |
| Trainium | 훈련을 위한 고성능 인스턴스 |
| Inferentia | 추론을 위한 고성능 인스턴스 |
| SageMaker | 머신 러닝 모델을 훈련 및 배포하는 서비스 |
</figure>
<figure>
<figure_type>Diagram</figure_type>
FM 훈련 및 추론을 위한 인프라
| AWS 서비스 | 설명 |
| - | - |
| EC2 용량 블록 | 훈련 및 추론을 위한 고성능 인스턴스 |
| Neuron | 추론을 위한 가속화 인스턴스 |
| EC2 UltraClusters | 대규모 훈련을 위한 클러스터 인스턴스 |
| EFA | 고성능 인스턴스 간 통신을 위한 네트워킹 서비스 |
| Nitro | 고성능 인스턴스의 가상화 및 관리를 위한 서비스 |
| Neuron | 추론을 위한 가속화 인스턴스 |
</figure>
<figure>
<figure_type>Diagram</figure_type>
생성형 AI 스택
| AWS 서비스 | 설명 |
| - | - |
| Amazon Bedrock | 생성형 AI 모델을 구축하고 호스팅하는 서비스 |
| 가드레일 에이전트 | 사용자 최적화 기능을 제공하는 서비스 |
</figure>


[User Question]
LLM 및 기타 FM을 활용하는 애플리케이션 관련 문서가 있으면 내용을 요약해줘.
문서는 LLM(Large Language Models) 및 기타 기계 학습 모델을 활용하여 애플리케이션을 구축하고, 훈련, 추론 인프라, 그리고 생성형 AI 스택에 관한 내용을 다룹니다.

- **LLM 및 기타 FM을 활용하는 애플리케이션**
  - 다양한 애플리케이션에서 LLM 및 기타 기계 학습 모델을 활용하는 방법을 소개합니다.
```

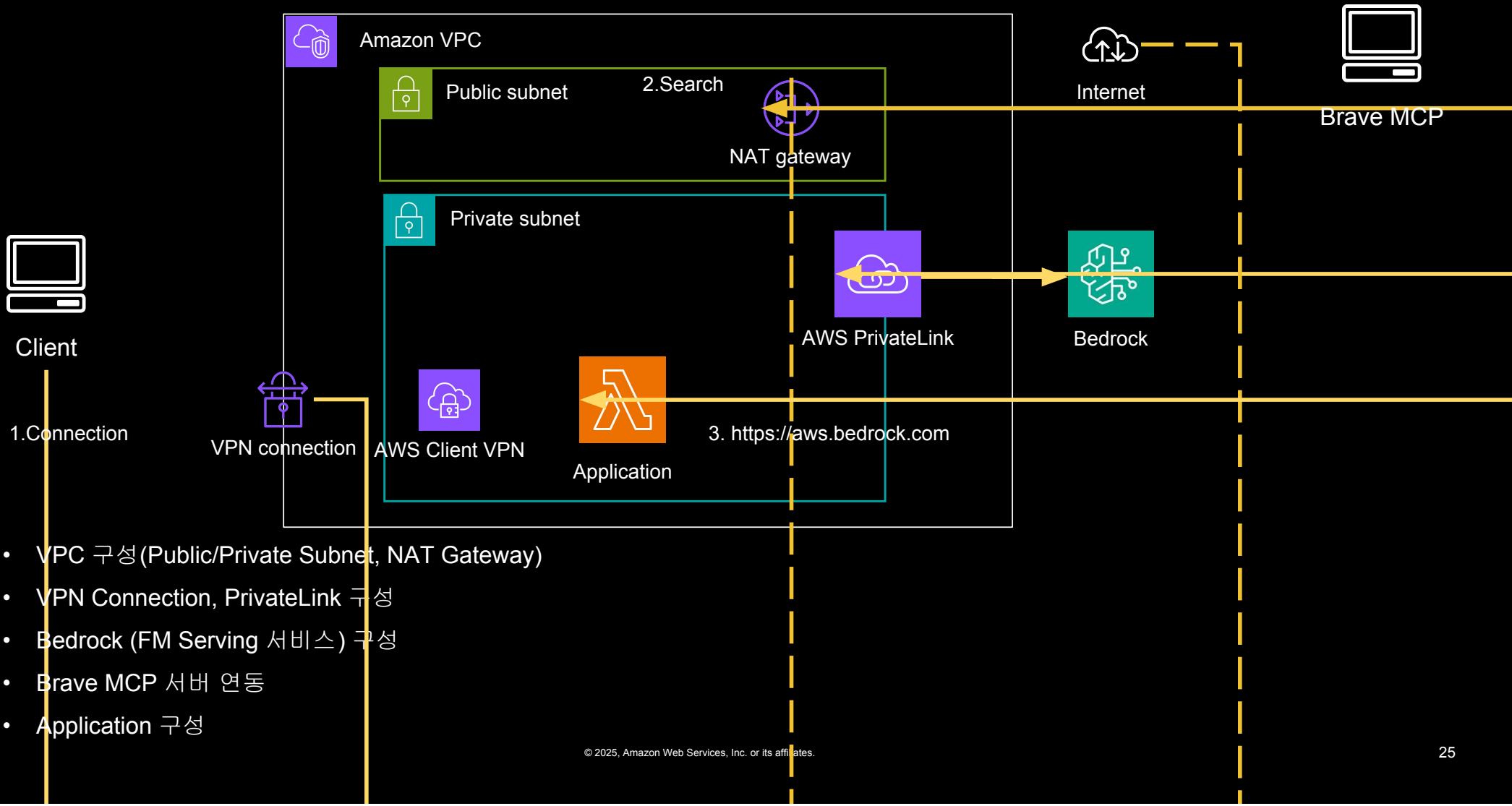
세부 설명

RAG 데이터

응답 결과

```
[Knowledge Base]
[Score: 0.6032051399592643]
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.
<figure>
<figure_type>Diagram</figure_type>
LLM 및 기타 FM을 활용하는 애플리케이션
</figure>
<figure>
<figure_type>Diagram</figure_type>
LLM 및 기타 FM을 사용하여 구축할 수 있는 도구
| AWS 서비스 | 설명 |
| - | - |
| Amazon Bedrock | 생성형 AI 모델을 구축하고 호스팅하는 서비스 |
| Trainium | 훈련을 위한 고성능 인스턴스 |
| Inferentia | 추론을 위한 고성능 인스턴스 |
| SageMaker | 머신 러닝 모델을 훈련 및 배포하는 서비스 |
</figure>
<figure>
<figure_type>Diagram</figure_type>
FM 훈련 및 추론을 위한 인프라
| AWS 서비스 | 설명 |
| - | - |
| EC2 용량 블록 | 훈련 및 추론을 위한 고성능 인스턴스 |
| Neuron | 추론을 위한 가속화 인스턴스 |
| EC2 UltraClusters | 대규모 훈련을 위한 클러스터 인스턴스 |
| EFA | 고성능 인스턴스 간 통신을 위한 네트워킹 서비스 |
| Nitro | 고성능 인스턴스의 가상화 및 관리를 위한 서비스 |
| Neuron | 추론을 위한 가속화 인스턴스 |
</figure>
<figure>
<figure_type>Diagram</figure_type>
생성형 AI 스택
| AWS 서비스 | 설명 |
| | - |
| Amazon Bedrock | 생성형 AI 모델을 구축하고 호스팅하는 서비스 |
| 가드레일 에이전트 | 사용자 최적화 기능을 제공하는 서비스 |
</figure>

[User Question]
LLM 및 기타 FM을 활용하는 애플리케이션 관련 문서가 있으면 내용을 요약해줘.
문서는 LLM(Large Language Models) 및 기타 기계 학습 모델을 활용하여 애플리케이션을 구축하고, 훈련, 추론 인프라, 그리고 생성형 AI 스택에 관한 내용을 다룹니다.

- **LLM 및 기타 FM을 활용하는 애플리케이션**
  - 다양한 애플리케이션에서 LLM 및 기타 기계 학습 모델을 활용하는 방법을 소개합니다.
```

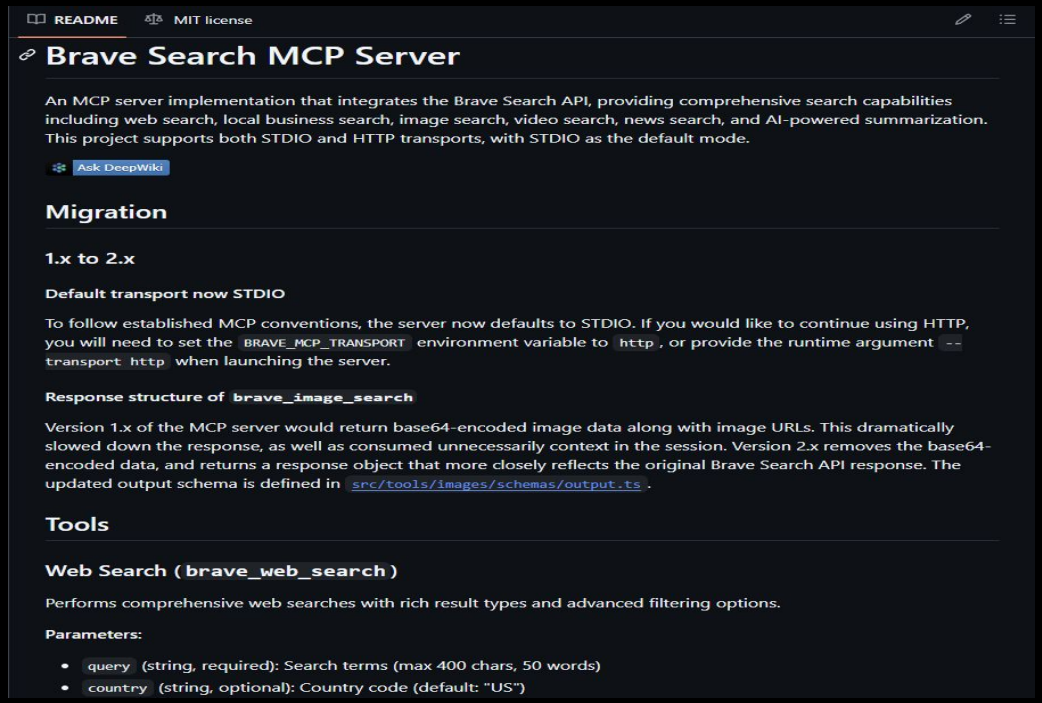
Case 3

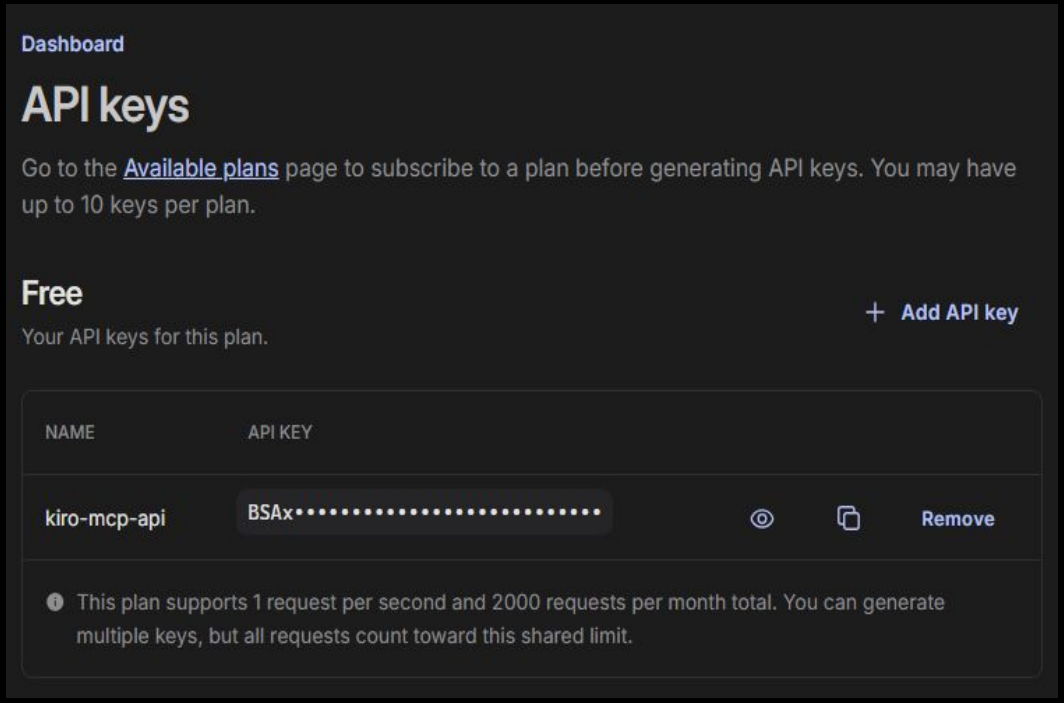


- VPC 구성(Public/Private Subnet, NAT Gateway)
- VPN Connection, PrivateLink 구성
- Bedrock (FM Serving 서비스) 구성
- Brave MCP 서버 연동
- Application 구성

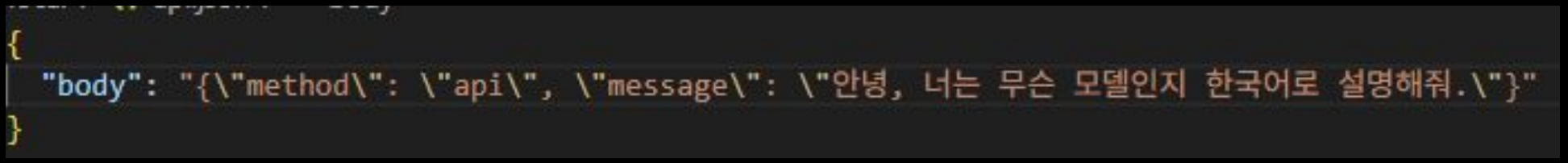
MCP 연동 - Brave Browser

- Brave Software, Inc.에서 개발한 크로미엄 웹 브라우저에 기반 오픈 소스 웹 브라우저
- Brave API를 통한 이미지/비디오/텍스트 기반 검색 엔진 기능을 MCP로 제공
- Brave API 회원 가입 및 API 키 발급(https://brave.com/ko/search/api/, 매월 무료 사용 가능)
- 발급받은 API 키를 어플리케이션에 추가





MCP 연동 결과



```
{
  "body": "{\"method\": \"api\", \"message\": \"안녕, 너는 무슨 모델인지 한국어로 설명해줘.\"}"
}
```

```
Based on the following search results, answer the user's question.

[Search Results]
[1] Korean ↔ English Translator with EXAMPLES | Collins
FREE Translations available in more than 100 languages including Spanish, English, French, German, Chinese, etc
URL: https://www.collinsdictionary.com/us/translator/korean-english

[2] Google 번역
무료로 제공되는 Google의 서비스는 영어와 100가지 이상의 다른 언어로 단어, 구문, 웹페이지를 즉시 번역합니다.
URL: https://translate.google.com/?hl=ko&sl=en&tl=ko&op=translate

[3] Is this correct, 저는 한국어를 조금 말 해요? - Quora
Answer: Yes and no! :) It is not grammatically wrong, but it is semantically incorrect. In other words, there is a confusion over the way 말 is used. Short ans
wer - the correct (i.e. the most natural) way to say this is: &gt; 저는 한국어를 조금 해요. Long answer - I assume that you&#x27;re directly translating ...
URL: https://www.quora.com/Is-this-correct-%EC%A0%80%EB%8A%94-%ED%95%9C%EA%B5%AD%EC%96%B4%EB%A5%BC-%EC%A1%B0%EA%B8%88-%EB%A7%90-%ED%95%B4%EC%9A%94

[User Question]
안녕, 너는 무슨 모델인지 한국어로 설명해줘.
{
  "method": "api",
  "response": "안녕하세요, 저는 [모델의 이름]입니다. 저는 여러분의 질문에 답변하고 다양한 정보를 제공할 수 있는 AI 어시스턴트입니다. 한국어로 설명하자면, 저는
[모델의 이름]이라는 AI 어시스턴트로, 사용자의 질문에 답변하고 정보를 제공하는 역할을 합니다. 더 궁금한 점이 있으시면 언제든지 물어보세요."
}
```

세부 설명

MCP 데이터

응답 결과

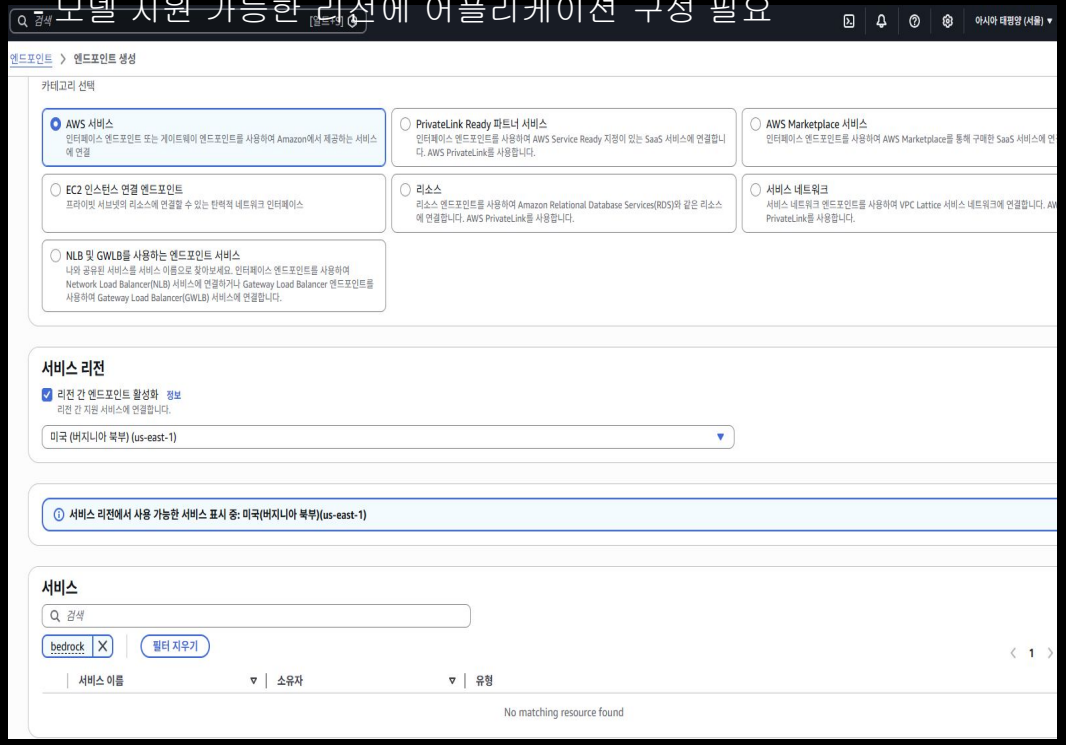
```
Based on the following search results, answer the user's question.

[Search Results]
[1] Korean ↔ English Translator with EXAMPLES | Collins
FREE Translations available in more than 100 languages including Spanish, English, French, German, Chinese, etc
URL: https://www.collinsdictionary.com/us/translator/korean-english

[2] Google 번역
무료로 제공되는 Google의 서비스는 영어와 100가지 이상의 다른 언어로 단어, 구문, 웹페이지를 즉시 번역합니다.
URL: https://translate.google.com/?hl=ko&sl=en&tl=ko&op=translate

[3] Is this correct, 저는 한국어를 조금 말 해요? - Quora
Answer: Yes and no! :) It is not grammatically wrong, but it is semantically incorrect. In other words, there is a confusion over the way 말 is used. Short ans
wer - the correct (i.e. the most natural) way to say this is: &gt; 저는 한국어를 조금 해요. Long answer - I assume that you&#x27;re directly translating ...
URL: https://www.quora.com/Is-this-correct-%EC%A0%80%EB%8A%94-%ED%95%9C%EA%B5%AD%EC%96%B4%EB%A5%BC-%EC%A1%B0%EA%B8%88-%EB%A7%90-%ED%95%B4%EC%9A%94

[User Question]
안녕. 너는 무슨 모델인지 한국어로 설명해줘.
```

```
{
  "method": "api",
  "response": "안녕하세요, 저는 [모델의 이름]입니다. 저는 여러분의 질문에 답변하고 다양한 정보를 제공할 수 있는 AI 어시스턴트입니다. 한국어로 설명하자면, 저는
[모델의 이름]이라는 AI 어시스턴트로, 사용자의 질문에 답변하고 정보를 제공하는 역할을 합니다. 더 궁금한 점이 있으시면 언제든지 물어보세요."
}
```

고려 사항

고려 사항

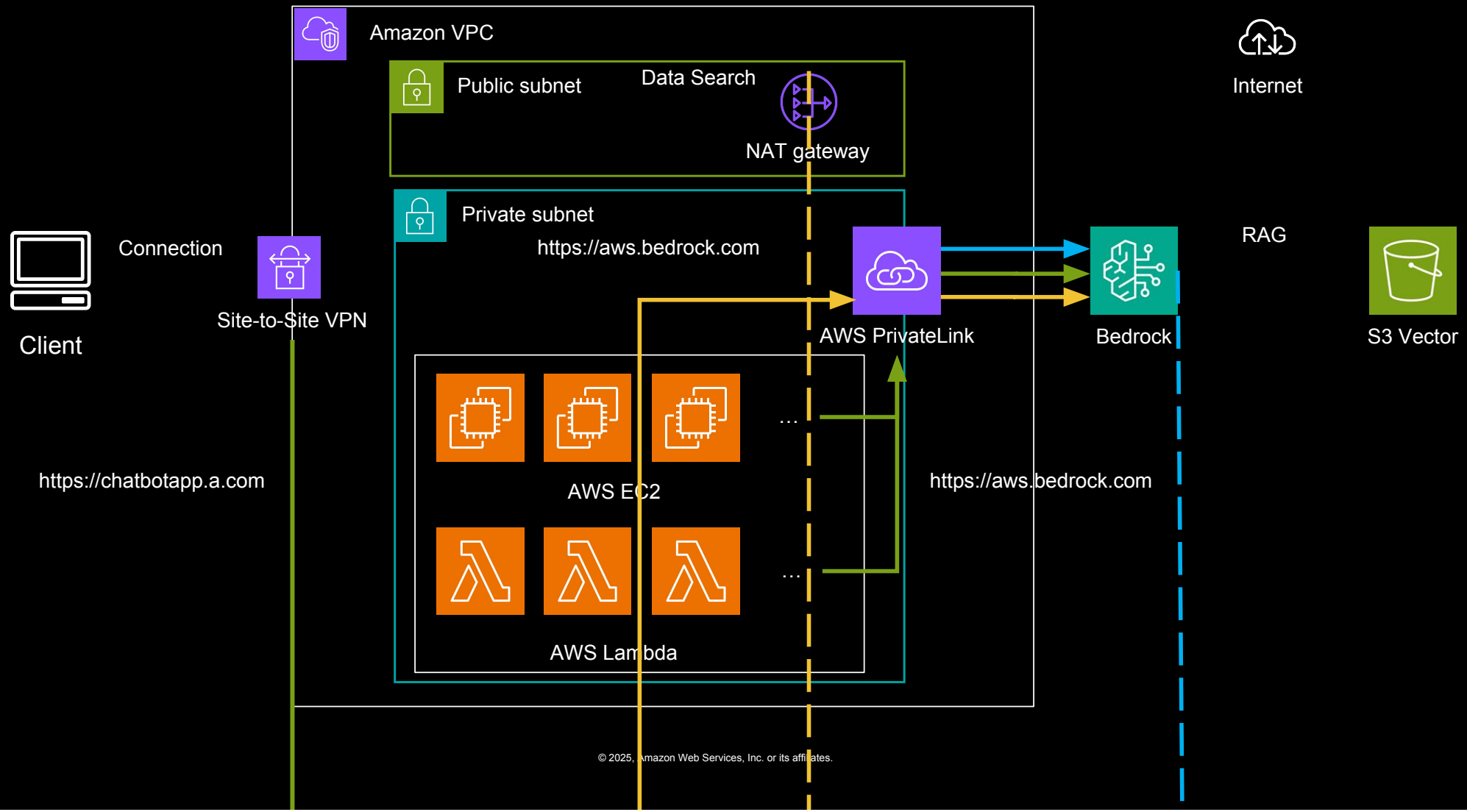
1. PrivateLink
- bedrock interface Endpoint는 리전간 엔드포인트 연결을 지원하지 않음
  (서울 리전 <-x-> 도쿄리전)
  모델 지원 가능한 리전에 어플리케이션 구성 필요



2. Bedrock 지식 기반
- 이미지와 텍스트가 함께 있는 문서는 AWS에서 제공하는 Titan 모델로
  구성 불가 (데이터 동기화 시 에러 발생)
  text & vision chunking 을 지원하는 모델로 RAG 구성

AI를 잘 사용하기 위한 Tips

회사에 VPN 장비가 있다면 Site-to-Site VPN를 활용하세요.

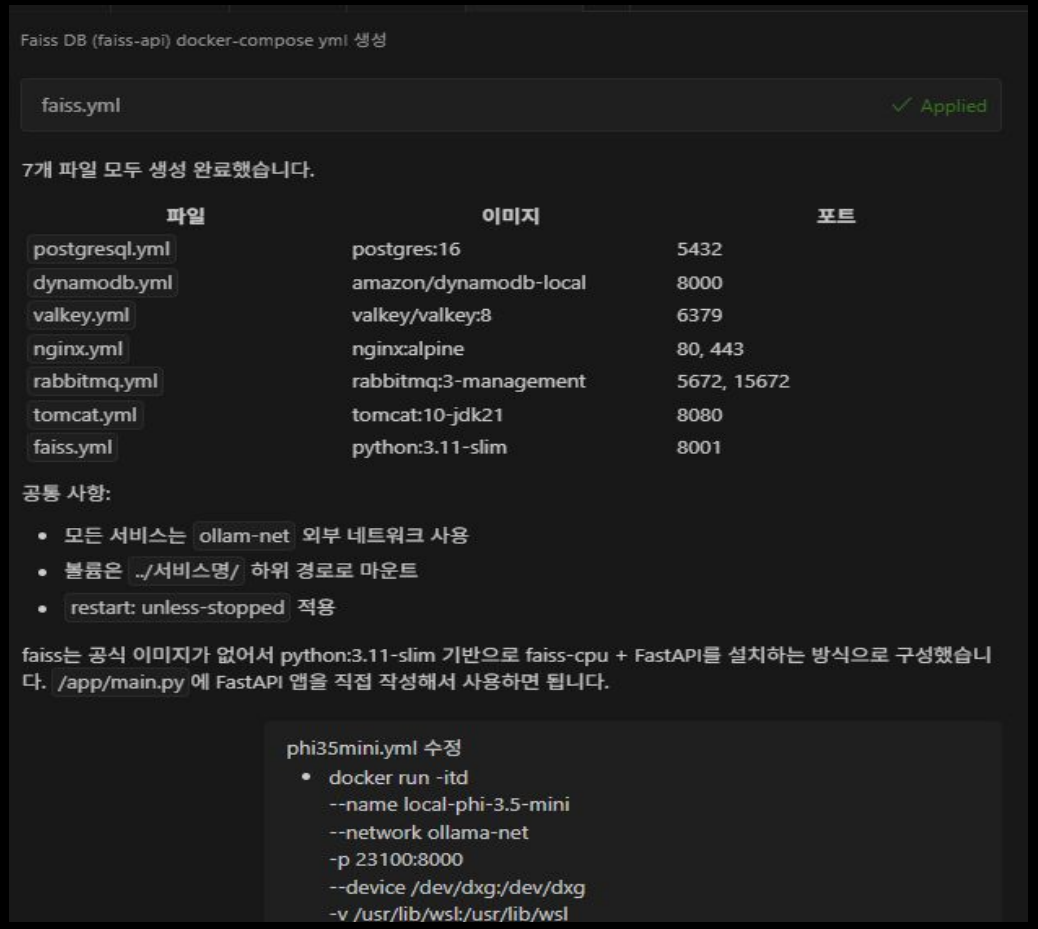


코드 작성 및 유지보수 목적 : Amazon Q Developer를 활용하세요.

Amazon Q Developer는 소프트웨어 개발 수명 주기(SDLC)를 지원하는 AWS의 생성형 AI 기반 대화형 어시스턴트

- 개발자가 코드를 더 빠르고 안전하게 작성, 테스트, 디버깅 제공
- 코드 마이그레이션(업그레이드)을 빠른 시간에 전환 가능
- Anthropic사의 Claude FM 제공
- Chatbot모드 / Agent 모드 제공
- AWS Builder 회원 가입 시, 매월 일정 사용량에 대한 Free Tire 제공

	Free Tier: Advanced capabilities at zero cost	Pro Tier: Expanded limits $19/mo. per user
요약	• 월별 제한된 에이전트 요청 • 최신 Claude 모델에 액세스 • IDE 또는 CLI에서 사용	프리 티어에 포함된 모든 기능에 다음이 포함됩니다. • 에이전트 요청 한도 증가 • Java 및 .NET 애플리케이션 변환에 대한 제한 증가 • 관리자 대시보드 및 제어를 통한 ID 센터 지원 • IP 보상



감사합니다.