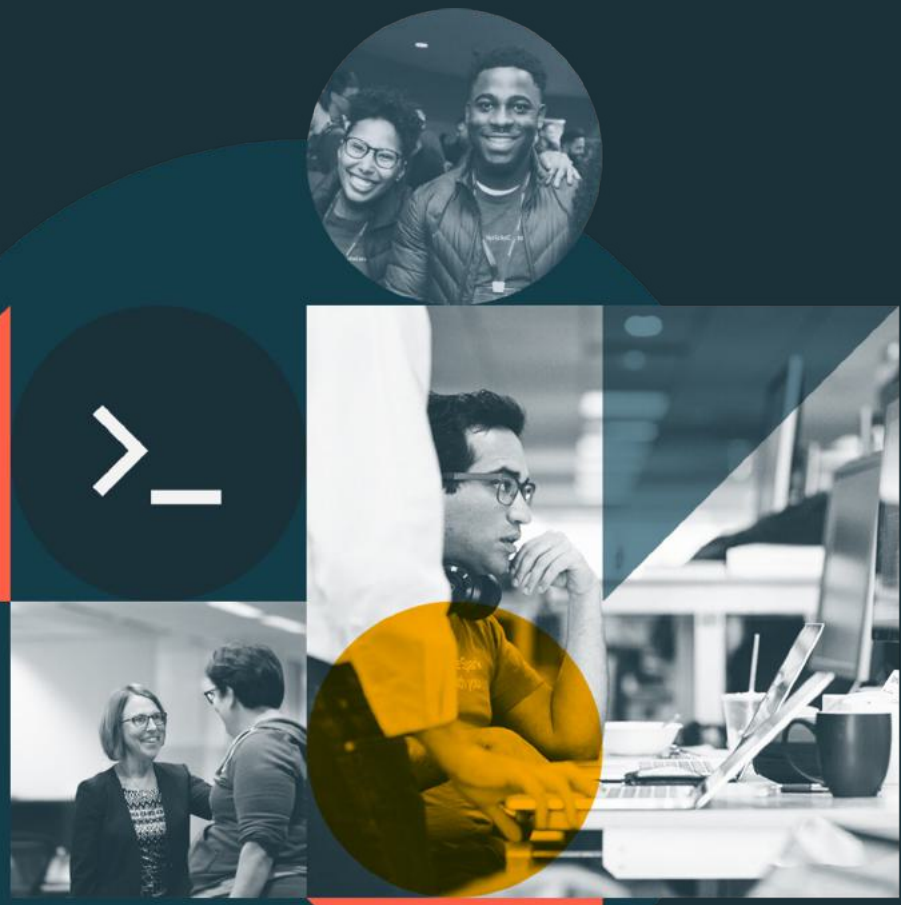
Databricks & AWS Zero to GenAI 워크샵
(Databricks GenAI 소개)



Databricks Korea
Ph.D. Insung Lee
2024.08.07

©2024 Databricks Inc. — All rights reserved

Databricks 소개

©2024 Databricks Inc. — All rights reserved

Databricks Data Intelligence Platform 소개

Databricks Data Intelligence Platform이란?

- 대규모 엔터프라이즈급 데이터의 **수집, 저장, 처리, 공유, 분석을 통합**하는 데이터 플랫폼
- 사용자의 퍼블릭 클라우드 환경에 배포되어 사용자 계정의 저장소 및 보안과 통합
- Amazon Web Services, Microsoft Azure, Google Cloud Platform에 배포 지원

Databricks와 Open Technologies

- Databricks는 **오픈소스 및 개방형 표준**을 지향
- Databricks주도로 설립된 오픈소스 : Apache Spark, Delta Lake, MLflow, Redash
- Databricks는 개방형 기술을 보다 최적화하여 성능향상과 사용편의성 증대를 위한 Workflows, Unity Catalog, Delta Live Tables, Photon등의 독점적 기능을 제공

Databricks와 생성형 AI

- 대규모 GPU-Farm기반의 플랫폼을 제공하여 대규모 분산/병렬 학습 기능 제공
- Langchain, LlamaIndex 등 다양한 LLM 라이브러리를 지원하여 GenAI앱 개발 가속화

< Databricks 논리 아키텍처 >

©2024 Databricks Inc. — All rights reserved

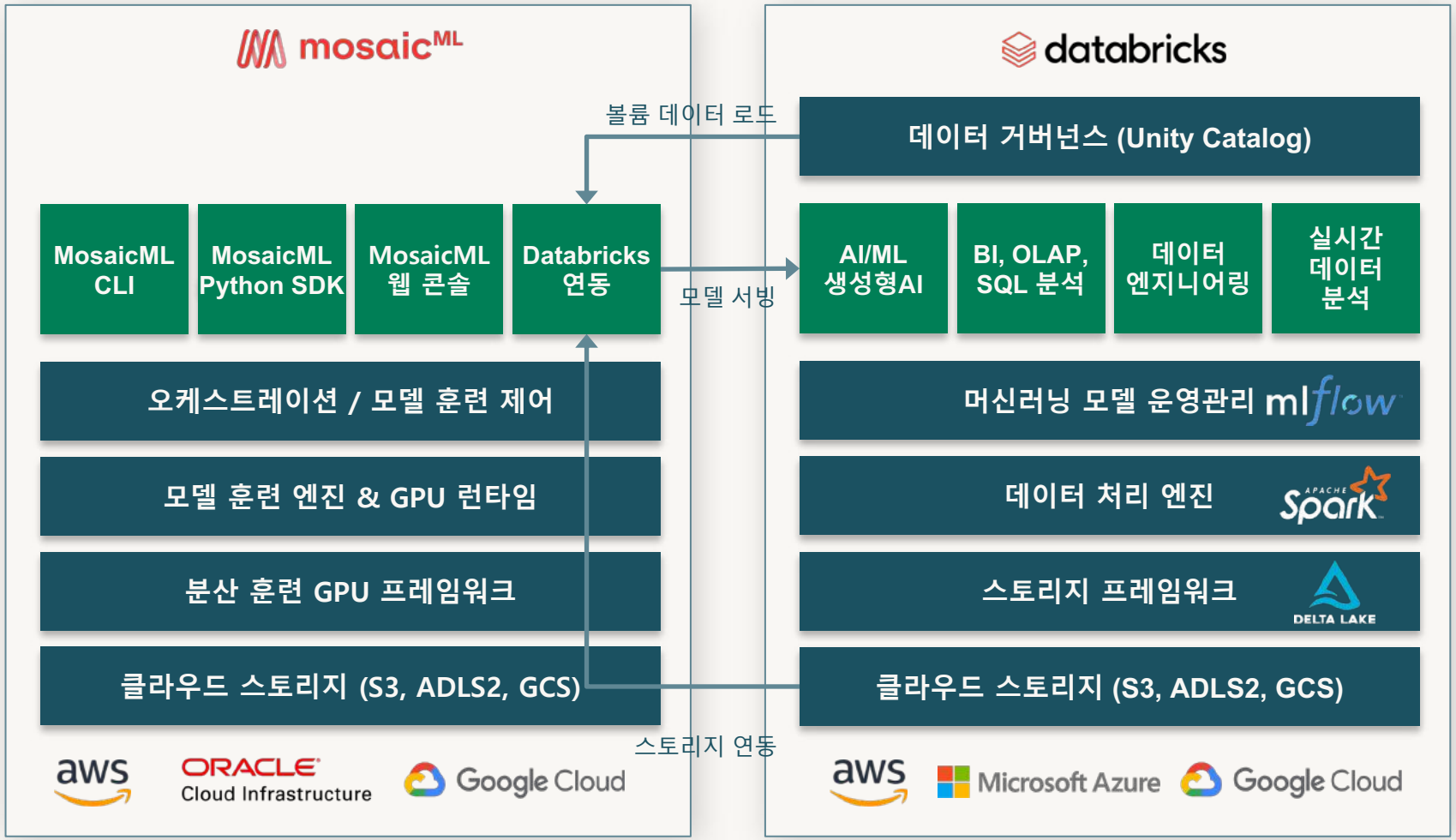
생성형 AI를 위한 Databricks

Databricks MosaicML이란?

- 생성형AI 분석 모델을 생성하기위한 **대규모 GPU 분산처리 및 병렬처리**를 위한 전문 모델 훈련 플랫폼
- Foundation Model 또는 Fine-Tune Model 생성시 대규모 GPU 자원을 안정적이며 신뢰할 수 있는 수준에서 사용가능 하도록 **분산훈련용 프레임워크**를 제공
- 모델에 사용되는 데이터와 생성된 모델의 거버넌스와 보안을 유지할 수 있도록 Databricks DI 플랫폼과 연계하여 안전한 생성형AI 모델 운영환경을 지원

< Databricks MosaicML 플랫폼 논리 아키텍처 >

< Databricks DI 플랫폼 논리 아키텍처 >



©2024 Databricks Inc. — All rights reserved

Databricks Data Intelligence Platform의 사용주체별 용도

데이터 엔지니어

데이터의 수집, 저장, 처리 작업을 수행하기위해 데이터 파이프인을 구축하고 데이터의 보안과 활용에대한 규정을 관리하는 역할

적용 업무

- RDBMS 또는 DW의 데이터를 Databricks로 적재하여 **레이크하우스 시스템** 구축
- 비정형 파일을 Databricks 저장소로 적재하여 **머신러닝/딥러닝 분석 시스템** 구축
- 배치 업무와 스트리밍 업무를 Databricks 플랫폼으로 통합하여 플랫폼 단일화
- ETL / CDC / ML 워크로드를 단일 작업으로 연속성 있게 구성하여 워크플로우 통합

데이터 분석가

정제된 데이터를 활용하여 OLAP 분석 모델을 생성하고 BI도구, 시각화도구, SQL을 통해 데이터의 차원을 회전하며 데이터를 분석하는 역할

적용 업무

- RDBMS, DW, File, NoSQL등의 소스 데이터가 통합된 레이크하우스 시스템에서 SQL을 통한 **다차원 분석 시스템** 구축
- Tableau, Power BI, Microstrategy등 **외부 BI 솔루션을 연계한 분석 시스템** 구축
- 자체 대시보드 기능을 통해 업무별 대시보드를 생성하고 공유하는 환경 구성
- DW 시스템의 경험과 동일한 SQL개발 환경

데이터 사이언티스트

정제된 데이터를 활용하여 특징변수를 관리하고 추론통계, 머신러닝, 딥러닝 모형을 작성하여 데이터를 분석하는 역할

적용 업무

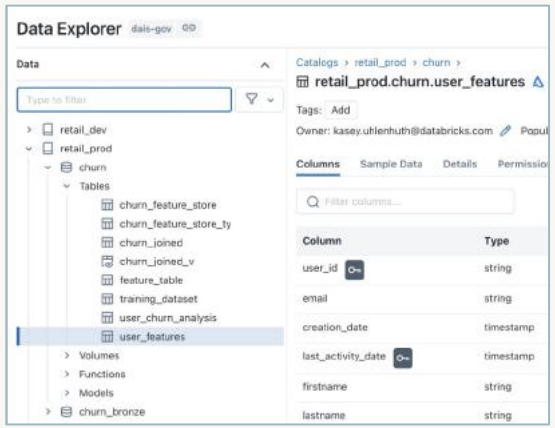
- 과거 데이터로부터 미래의 상황을 예측하는 **머신러닝 예측 분석 시스템** 구축
- 비정형 데이터로부터 상황을 인식하고, 분류하는 **딥러닝 예측 분석 시스템** 구축
- GPU를 활용하여 텍스트로부터 텍스트 또는 이미지를 생성하는 **생성형 AI 시스템** 구축
- 시티즌 데이터사이언티스트를 위한 AutoML 기반의 대중화된 머신러닝 분석플랫폼 구축

©2024 Databricks Inc. — All rights reserved

Databricks의 데이터 엔지니어링

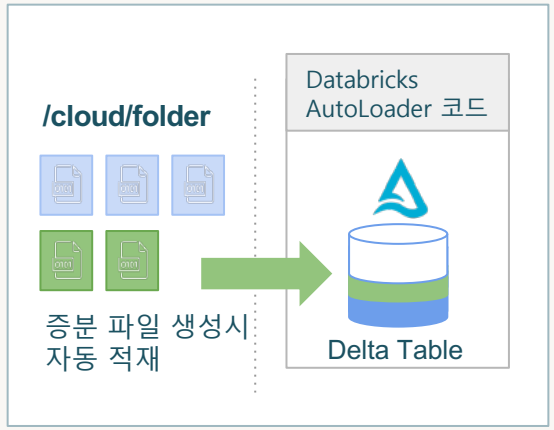
데이터 거버넌스 기능

- 카탈로그, 스키마 구조 지원
- 데이터 객체들에 대한 권한관리
- 데이터간 계보를 자동으로 생성
- 데이터 객체들의 감사이력 조회
- 인가된 객체들의 검색 및 탐색



데이터 추출 기능

- 스토리지의 데이터를 직접 쿼리할 수 있는 기능 제공
- 스토리지에 새 파일이 도착하면 자동으로 테이블로 적재
- 변경적재 기능 지원

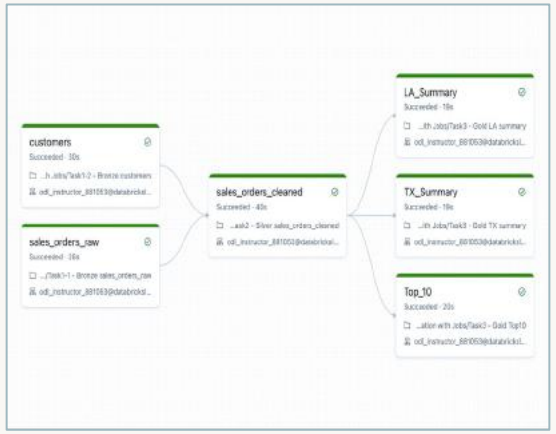


데이터 ETL/CDC 기능

- DAG 데이터 파이프라인을 생성하여 데이터 흐름 제어
- 배치 또는 실시간 스트리밍 데이터 처리 작업 지원
- 데이터 파이프라인내에서 데이터 품질을 검증하는 기능 제공

작업 워크플로우 기능

- 다양한 형태의 작업을 단일 워크플로우로 오케스트레이션
- ETL, CDC, ML 등 성격이 다른 작업들을 단일 작업으로 생성
- 작업 실패시 알림 기능 제공
- Job Cluster 활용으로 비용 최적화

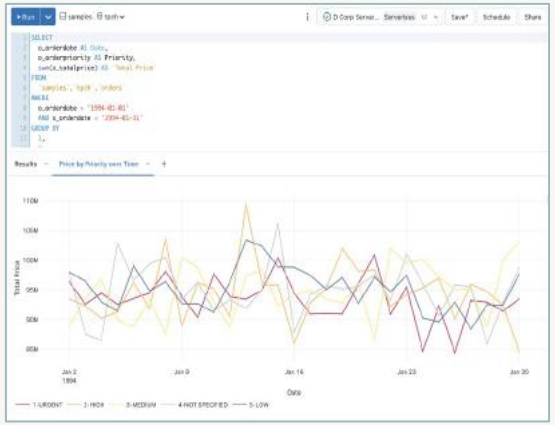


©2024 Databricks Inc. — All rights reserved

Databricks의 데이터 분석

SQL 분석 기능

- SQL 전용 개발환경 제공
- 쿼리 결과를 바로 시각화 가능
- 쿼리 편집기 또는 노트북에서 위젯 기능 제공
- 쿼리를 스케줄링하여 수행가능



대시보드 기능

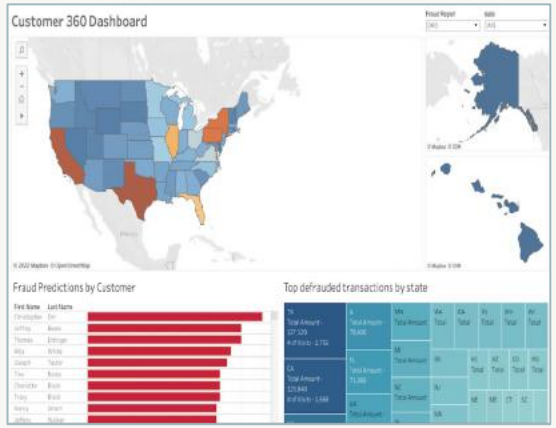
- 바차트, 라인차트, 파이차트 등 다양한 시각화 유형 제공
- 웹기반 Drag&Drop으로 대시보드 생성 가능
- 대시보드를 스케줄링하여 업데이트 및 공유 기능 제공

외부 앱의 API 연동 기능

- SDK 또는 RestAPI를 통해서 작성된 쿼리를 수행
- 서비스 주체를 생성하여 Oauth 인증기반 M2M 통신 지원
- JDBC/ODBC를 통해서 외부앱에서 데이터 처리

외부 BI도구 연계 기능

- Amazon QuickSight, Tableau, Power BI, Microstrategy 등 다양한 외부 BI 도구 연동 지원
- 외부 BI 도구용 드라이버 제공
- 라이브 커넥션 지원

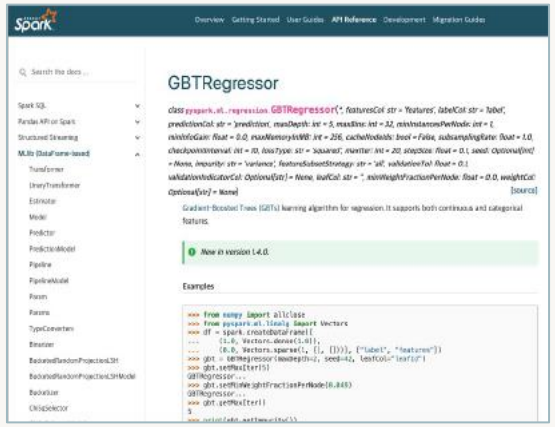


©2024 Databricks Inc. — All rights reserved

Databricks의 데이터 사이언스

분산 학습 기능

- Spark MLlib 분산 학습 제공
- Horovod, Ray로 분산학습 지원
- Petastorm으로 Tensorflow, Pytorch로 데이터 변환
- Hyperopt로 매개변수 분산 튜닝

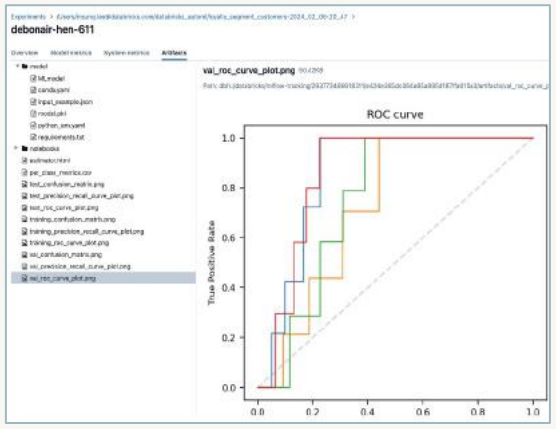


오픈소스 프레임워크 지원 기능

- 오픈소스 ML 프레임워크 지원
- 오픈소스 DL 프레임워크 지원
- Hugging Face, LangChain , DeepSpeed 등 생성형 AI 프레임워크 지원

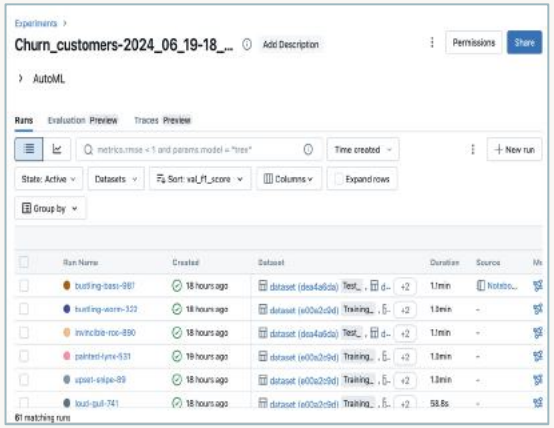
모델 생명주기 관리 기능

- 실험 및 실행 추적 기능 제공
- 모델 관리 및 배포기능 제공
- ML코드 및 모델 패키징, 재배포
- 모델의 거버넌스 관리
- 모델의 추론기능 제공



AutoML 기능

- 분류, 회귀, 예측분석 제공
- GUI로 자동화 파이프라인 생성
- 하이퍼파라미터 자동 튜닝
- 모델, 파라미터, 지표 자동 로깅
- 생성된 챔피언 모델의 코드 제공



©2024 Databricks Inc. — All rights reserved

GenAI 고객 사례

©2024 Databricks Inc. — All rights reserved

edmunds

edmunds는 미국의 자동차 정보를 제공하는 온라인 서비스로 Databricks GenAI 기술을 활용하여 자동차 리뷰정보에 대한 자동분류체계를 확립

Challenge

Edmunds 서비스에 매일 300개가 넘는 자동차 및 딜러에 대한 리뷰가 등록되고 이를 검토, 분류, 승인하는데 수일이 소요. 사이트에 리뷰가 승인되어 게시되는데 최대 48~72시간이 소요.

Solution

Databricks GenAI 기술을 활용하여 리뷰의 자동 조정 모델을 개발. 이 모델을 활용하여 매일 수백개의 리뷰의 구문을 자동으로 분석하고 분류하여 온라인에 게시할 수 있도록 진화.

Impact

3–5 시간

리뷰 자동 조정으로 주당 절약되는 시간

2명

300개 이상의 리뷰를 평가하는데 필요한 중재자 수

©2024 Databricks Inc. — All rights reserved

UK Power Networks

UK Power Networks는 영국의 전력 배전망 운영사로 800만명의 고객을 대상으로 하는 고객응대 업무를 Databricks GenAI 기술을 활용하여 자동화 및 간소화

Challenge

UKPN의 고객 서비스 팀은 많은 양의 이메일을 처리하는 데 어려움을 겪었습니다. 사람의 수작업으로 인해 오류가 많이 발생했으며 서비스 수준 합의기준 (SLA)을 충족시키지 못하는 경우가 많이 발생.

Solution

UKPN은 이메일 분류 및 요약 작업을 자동화하는 가상 에이전트를 Databricks 플랫폼내에서 GenAI기능을 이용하여 개발 및 배포 하였으며 이로인해 상담원의 고객응대 업무를 간소화.

Impact

10x

고객 이메일 응대 시간 감소

300–400

하루 이메일 처리량

©2022 Databricks Inc. — All rights reserved

databricks

LIPPERT

레저용 차량, 선박 및 주택제조에 엔지니어링 부품을 제공하는 미국의 Lippert사는 Databricks의 GenAI기술을 활용하여 상담원 지원 시스템을 구축하여 상담원의 업무부하를 감소시키며 상담 효율성을 향상

Challenge

특수한 산업의 고객지원센터 운영은 상담원 이탈에 따른 신규채용의 어려움과 신규 상담원의 숙련도를 향상 시키기 위한 시간이 많이 소요.

Solution

상담원이 고객 문의에 대한 관련된 정보를 확인해 볼 수 있도록 Databricks 플랫폼을 기반으로 RAG 솔루션을 개발. 제품 메뉴얼, 제품관련 YouTube 동영상, 과거 지원사례 등에서 컨텐츠를 수집하여 RAG 구성.

Impact

효율성 향상

고객 지원 환경과 지원 시간을 개선하여 상담원의 업무 처리 속도 향상

©2024 Databricks Inc. — All rights reserved

databricks

refuel

Try Refuel Cloud

refuel은 LLM 기술을 기반으로 재무, 인사, 법률, 전자상거래 등 도메인 데이터에대한 레이블링을 자동화해주는 서비스를 제공하는 기업으로 자체 LLM을 Databricks GenAI 기반으로 개발

Challenge

고객의 데이터 중 레이블링이 필요한 데이터에 레이블을 생성하여 고품질의 데이터 세트를 구축할 수 있도록 Refuel LLM과 Refuel LLM Playground를 개발.

Solution

Databricks와 함께 Mosaic AI를 사용하여 오픈 소스 LLM인 Llama를 파인튜닝하여 모델을 개발했고, 사람이 메뉴얼작업을 하는 방식보다 보다 나은 성과를 달성.

Impact

86%

커스텀 LLM을 사용하여 데이터 레이블 품질 개선 비율

3일

Mosaic AI 플랫폼을 사용하여 50B 모델을 처음부터 생성하는 데 걸리는 일수

databricks

replit

온라인 통합 개발환경을 서비스하는 미국의 Replit사는 자사의 고스트라이터 모델을 Databricks 플랫폼을 사용하여 처음부터 사전학습하여 개발하여 개발자들에게 보다 지능화된 IDE환경을 제공

Challenge

Replit의 온라인 통합 개발환경(IDE) 시스템에 LLM 기술을 접목하려 하였으나 기존 LLM을 튜닝하여 사용하기에는 너무 많은 오류가 발생.

Solution

자체 LLM모델을 Pre-training 부터 다시 시작하여 2.7B 파라미터 모델을 직접 생성하였고, replit의 고스트라이터 모델을 다시 생성.

Impact

3일

LLM 트레이닝에 걸리던 몇달간의 시간을 단축

1일

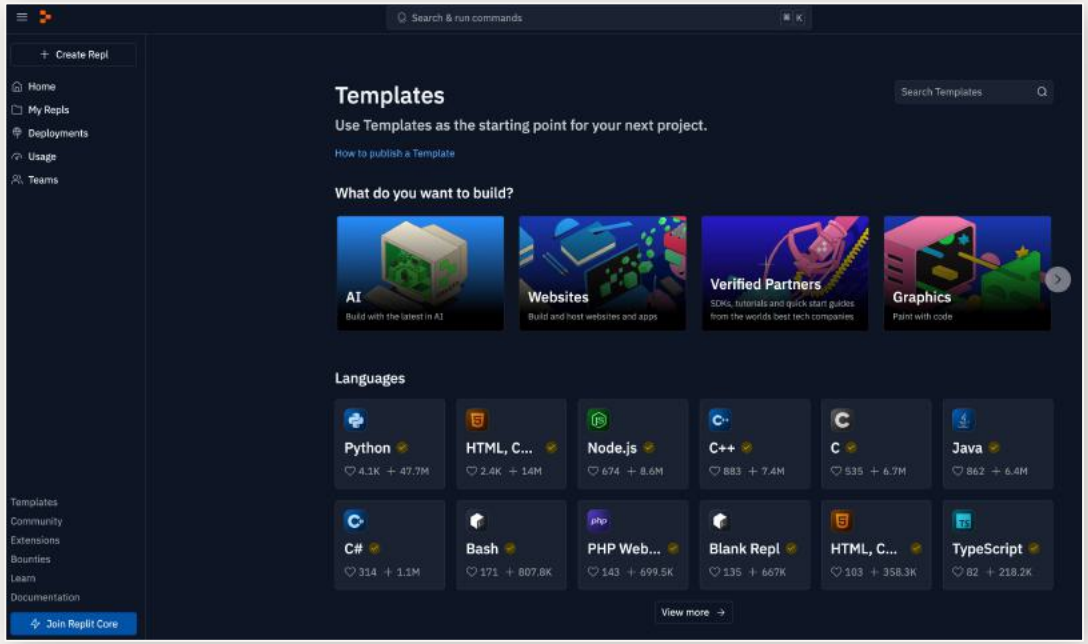
모델 배포에 걸리는 시간

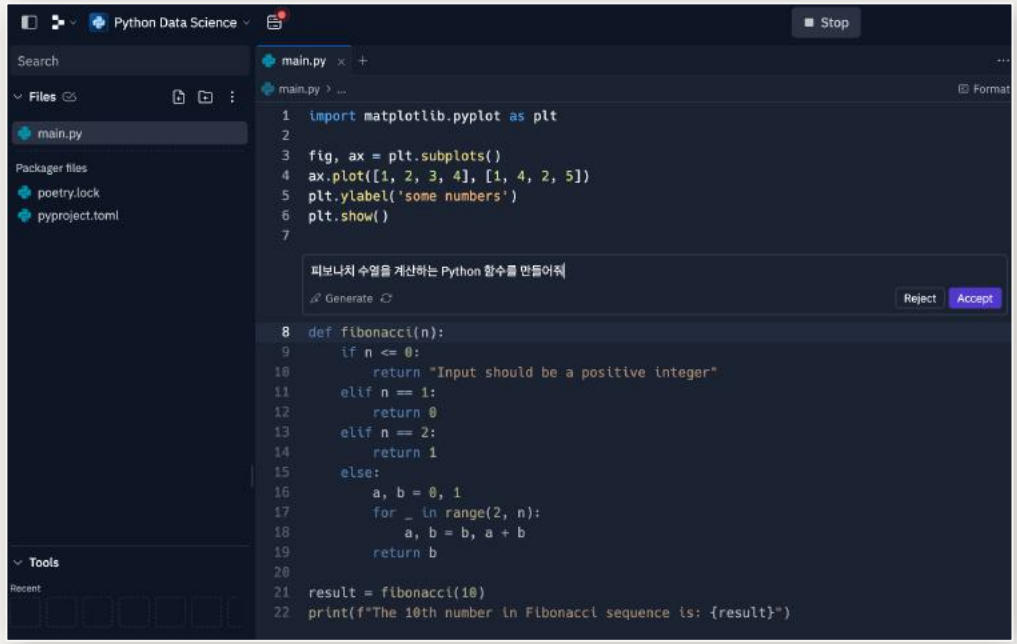
©2024 Databricks Inc. — All rights reserved

고객 사례 – Replit

Replit의 LLM 모델 개발 전략

- 2500만명의 개발자가 replit 서비스를 사용중이며, 소프트웨어 개발 수요 증가대비 진입장벽을 낮추기 위해 GenAI 기술을 도입
- 소프트웨어 개발 및 배포시간을 단축 시키기위해 코드생성에 특화된 비용 효율적인 고스트라이터 LLM을 개발
- LLM 모델을 가장 효율적으로 학습시킬 수 있는 도구로 Databricks Data Intelligence Platform + Hugging Face 조합을 선택





https://www.databricks.com/customers/replit
https://replit.com/

©2024 Databricks Inc. — All rights reserved

Replit은 왜 자체 모델을 개발 하였는가?

커스텀 모델의 필요성 인지

- GPT-4와 같은 범용 모델이 생성하는 코드는 플랫폼별 특성, 용어, 컨텍스트를 고려하지 않기 때문에 사용자 환경에 적합하지 않을 수 있음
- Javascript React(JSX)나 Typescript React(TSX)와 같은 특정 프로그래밍 언어에 대한 학습량이 부족할 수 있음

서비스 의존성 감소

- 특정 서비스 제공업체에만 의존하는 것보다는 목적별로 전문화된 서비스를 이용하는 것이 서비스 품질 관점에서 이점
- 특정 서비스 제공업체에만 의존할 경우 서비스 잠금현상(Lock-in)이 발생할 수 있음
- Replit에서는 다양한 모델들을 오픈소스로 공개할 계획을 전제로 모델 개발

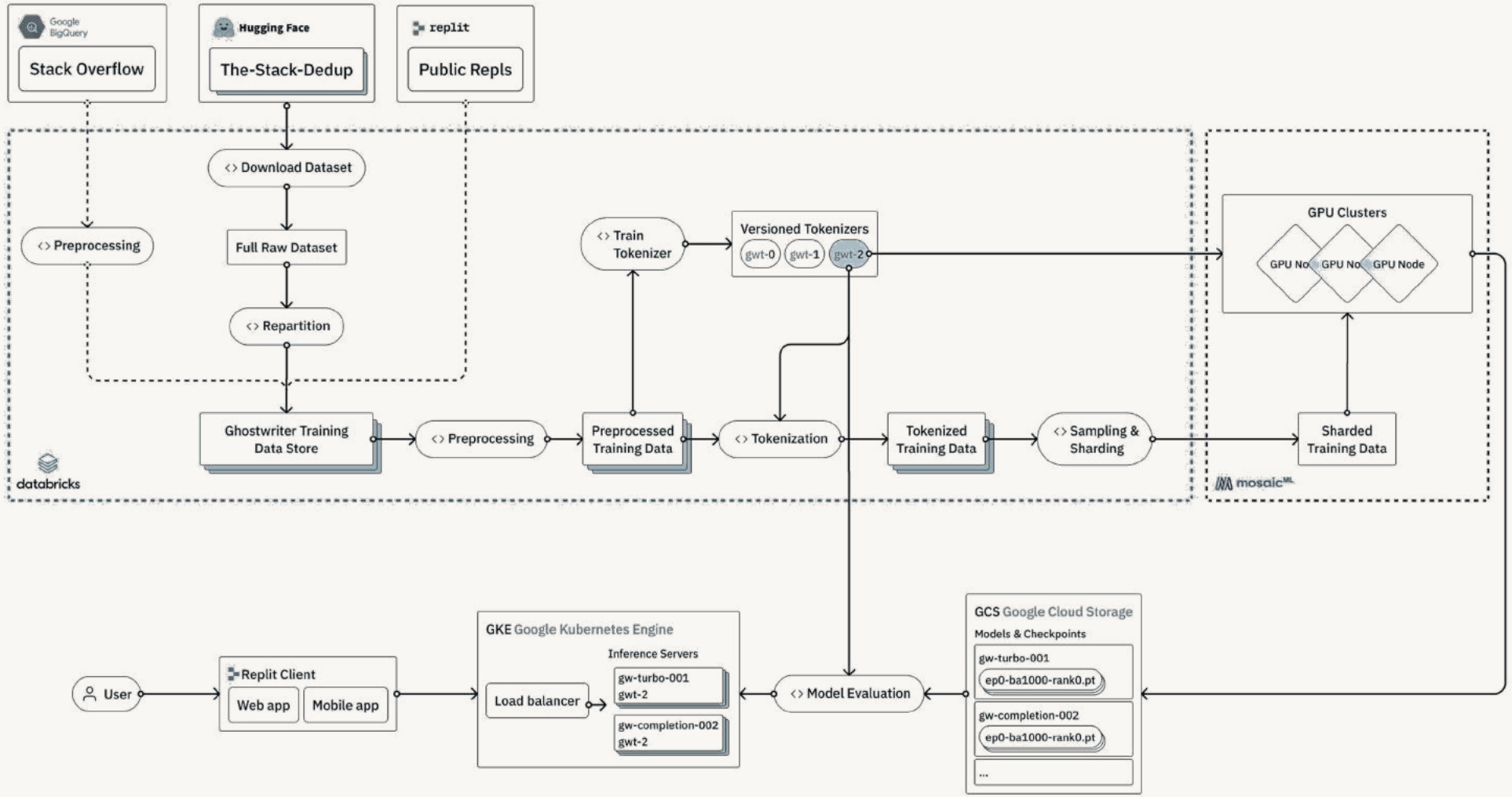
비용 효율성 증대

- 개발 도상국에서의 개발자도 실리콘 밸리에서의 개발자와 동일한 수준의 AI에 접근할 수 있어야 함을 전제로 낮은 가격대를 책정
- 소프트웨어 개발을 전문으로 하는 스타트업들에게 저렴한 AI 서비스를 제공하기 위해 노력

https://blog.replit.com/llm-training

©2024 Databricks Inc. — All rights reserved

Replit – 데이터 파이프라인



https://blog.replit.com/llm-training

©2024 Databricks Inc. — All rights reserved

Replit – 데이터 수집

LLM 모델 학습에 사용된 데이터 소스

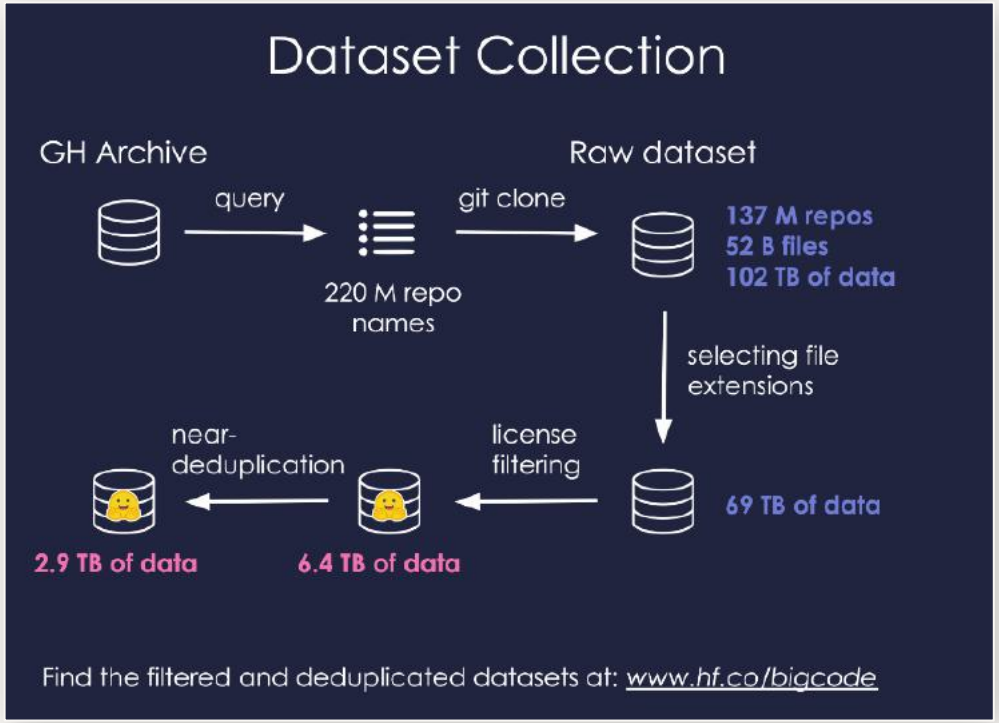
- The Stack 데이터세트를 시작으로 Stack Overflow, Replit 자체 데이터 등 다양한 프로그래밍 코드를 학습
- 특히 The Stack은 Bigcode 프로젝트의 하위 프로젝트로 358개 프로그래밍 언어로 구성된 6.4TB의 데이터세트

LLM 모델 학습에 사용된 라이브러리

- Hugging Face에서 제공하는 Transformer 위주의 라이브러리를 사용
- 토크나이저, 모델 인퍼런스, 코드 평가 라이브러리를 사용

Hugging Face 라이브러리 활용의 문제

- Hugging Face는 LLM 모델 학습을 위한 유용한 라이브러리를 제공하지만, 대규모 데이터세트를 대규모 GPU-farm에서 분산처리하기 위한 라이브러리 제공은 부족



< The Stack이 데이터를 수집하는 방법 >

https://www.bigcode-project.org/docs/about/the-stack/

©2024 Databricks Inc. — All rights reserved

Replit – 데이터 처리

대용량 데이터 처리를 위한 분산 프로세스

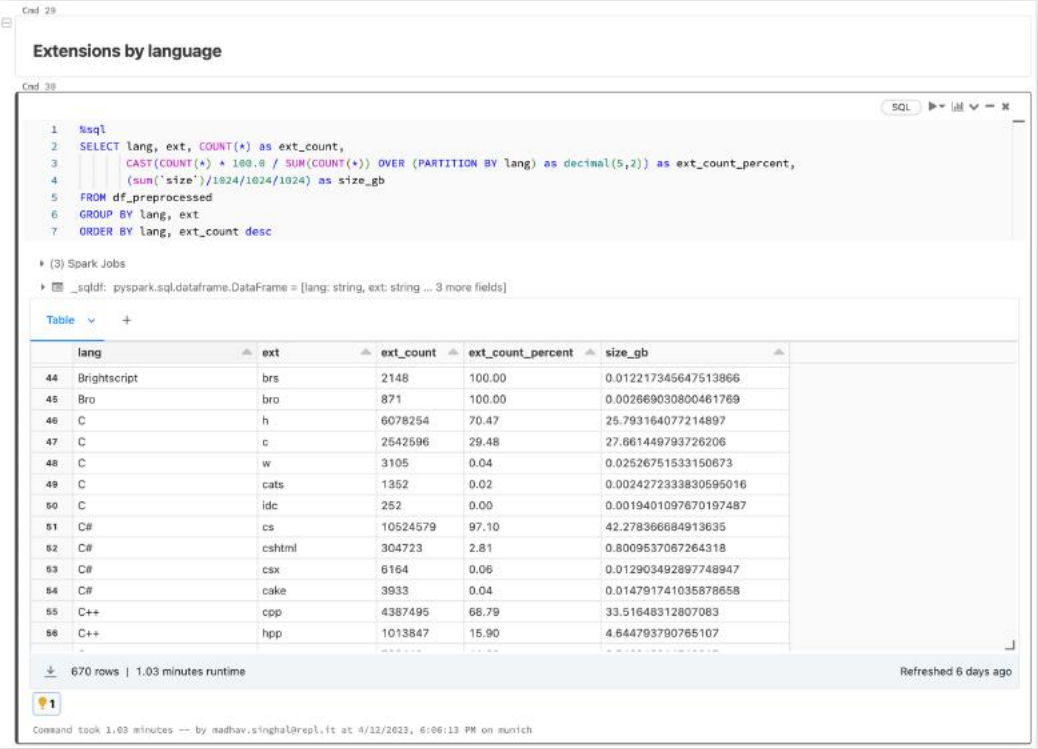
- 대량의 코드 데이터를 처리하기위해 분산/병렬 데이터 처리 프로세스가 필요로 됨
- Databricks Data Intelligence Platform의 Spark엔진과 Workflow를 통해 반복적으로 수집되는 대량의 데이터를 고속으로 처리

대용량 데이터 전처리

- Kocetkov et al.(2022)논문의 내용을 바탕으로 Stack, Stack Overflow, Replit 데이터를 병합하여 중복을 제거
- 이메일, IP주소, 암호키, 인증키 등 모든 개인 식별 정보를 익명화
- 머신이 작성한 코드를 식별하여 제거하기위해 휴리스틱 식별 로직 적용
- 구문분석기를 사용하여 컴파일되지 않는 코드들을 제거

데이터 전처리 품질 관리

- 전처리된 코드 데이터세트는 데이터 파이프라인상의 프로세스 문제나, 데이터의 불일치를 추적하기위해 데이터 품질 추적을 위한 ML모델을 개발
- 품질 추적을 위한 ML은 Mlflow를 통해 추적 및 재현



< Databricks의 SparkSQL을 활용한 데이터 분산 처리 >

https://blog.replit.com/llm-training

©2024 Databricks Inc. — All rights reserved

Replit – 모델학습

사용자 지정 어휘학습 및 토큰화

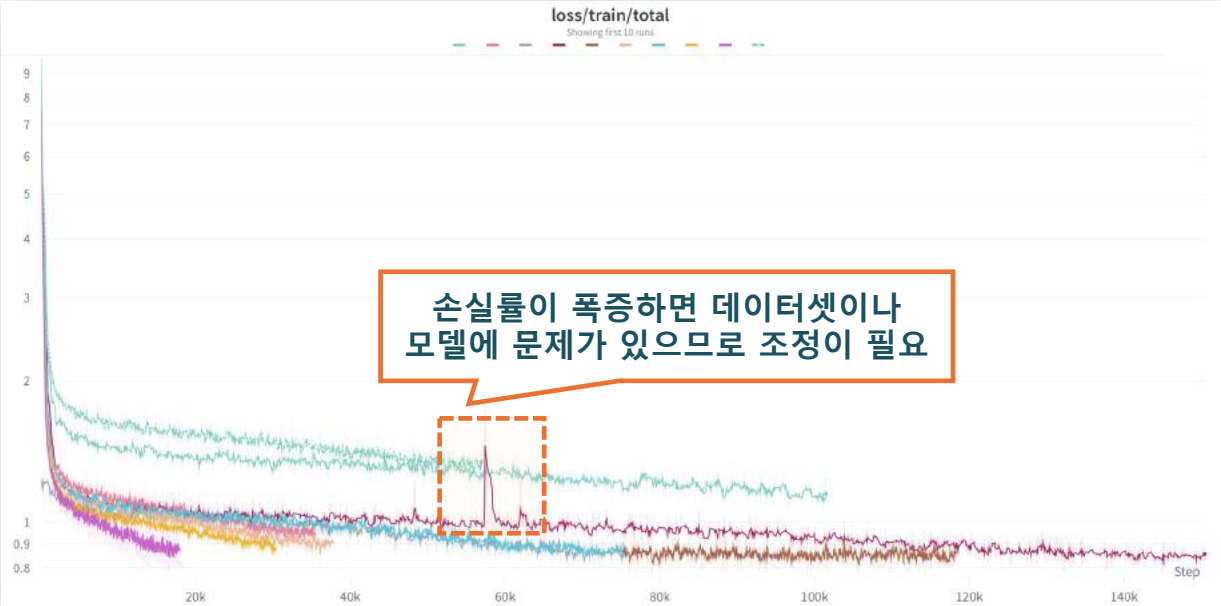
- 데이터를 무작위 샘플링하여 사용자 지정 어휘를 학습하여 모델이 코드 컨텐츠를 더 잘 이해할 수 있도록 작업
- 어휘 크기, 특수 토큰, 센티널 토큰등을 고려하여 토큰화 한 뒤 결정론적 데이터 포멧인 **MDS(Mosaic Data Shard)**의 학습데이터로 생성

모델 학습용 플랫폼으로 Databricks Mosaic AI를 선정

- 멀티 클라우드에서 GPU인스턴스를 활용 가능
- **Composer 라이브러리**를 사용하여 사전 튜닝된 다양한 학습 파라미터를 사용 가능
- GPU 노드 장애로부터 복구하는 **내결함성** 제공

모델 학습

- 모델 파라미터 수를 결정하기위해 모델 크기, 컨텍스트 윈도우, 추론시간, 메모리 사용량 등을 고려
- Replit은 코드 컴플리션에 응답이 매우 빨라야 한다고 판단하였기 때문에 추론이 가장 빠른 모델을 개발



< Databricks Mosiac AI를 사용한 모델 학습진행 상황 모니터링 >

https://blog.replit.com/llm-training

©2024 Databricks Inc. — All rights reserved

Replit – 모델 평가

모델 평가

- 모델을 평가하기 위해 Chen et al.(2021)논문내용을 바탕으로 HumanEval 프레임워크의 변형을 사용
- 이 모델을 사용하여 자연어/코드 쌍으로 만들어진 Python 코드 블록을 생성
- 생성된 함수에 대해 테스트 케이스를 실행하여 생성된 코드 블록이 예상대로 작동하는지 확인
- 프롬프트가 k개의 코드를 생성하여 정답을 맞춘 코드블럭이 있으면 Pass로 측정하는 Pass@K 평가 수행

```
runs.summary["latest_code_eval"]
```

	task_id	completion_id	passed	result	generation	canonical_solution	reference
1	4	0	True	passed	def mean_absolute_deviation(List[float]) -> float:	mean = sum(numbers) / len(numbers) return sum(abs(x - mean)	def check(candidate): assert abs(candidate([1.0, 2.0,
2	2	0	False	failed:	def truncate_number(number: float) -> float:	return number % 1.0	def check(candidate): assert candidate(3.5) == 0.5
3	0	0	False	failed:	def has_close_elements(numb List[float], threshold: float)	for idx, elem in enumerate(numbers): for idx2, elem2 in	def check(candidate): assert candidate([1.0, 2.0, 3.9, 4.0, 5.0, 2.2], 0.3)
4	7	0	True	passed	def filter_by_substring(strings: List[str], substring: str) ->	return [x for x in strings if substring in x]	def check(candidate): assert candidate([], 'john') == [] assert candidate(['xxx',

< Pass@K 평가 수행 결과 >

https://blog.replit.com/llm-training

©2024 Databricks Inc. — All rights reserved

Replit – 운영환경 배포

고속 추론을 위한 모델 배포

- Databricks Mosaic AI에서 개발한 모델을 추론서버로 배포
- 고스트라이터, 코드 컴플리션 기능은 고속의 추론이 필요
- Transformer 기반 신경망 추론을 위한 NVIDIA의 FasterTransformer 라이브러리 사용
- 구성이 간단하고 추론이 빠른 NVIDIA의 Triton 서버 사용
- 위 조합으로 Transformer 모델과 GPU 사이에 고도로 최적화된 레이어를 제공하고 대규모 모델을 초고속으로 추론

탄력적 추론환경 구성

- Kubernetes 인프라를 사용하여 워크로드에 따라 자동으로 추론환경이 증/감 될 수 있도록 구성
- 추론환경을 지속적으로 모니터링

< 추론 환경 모니터링 >

https://blog.replit.com/llm-training

©2024 Databricks Inc. — All rights reserved

Databricks Mosaic AI

©2024 Databricks Inc. — All rights reserved

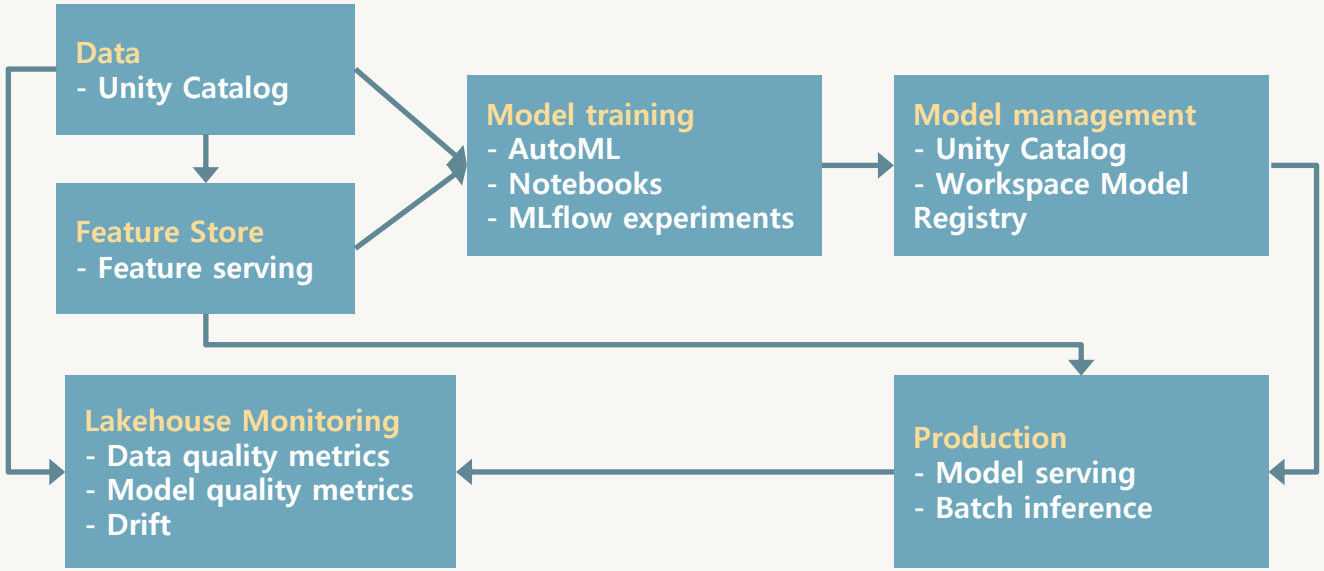
Databricks Machine Learning 개요

Databricks Mosaic AI

- Databricks Mosaic AI는 Databricks AI/ML과 MosaicML을 통합한 개념
- ML(Machine Learning), DL(Deep Learning), GenAI(Generative AI) 모든 워크로드를 단일 플랫폼에서 개발
- Databricks Mosaic AI 기능은 ML워크로드를 개발하고 ML워크플로우를 구축하기위한 다양한 기능을 제공

모델 개발을 위한 Databricks Mosaic AI의 특장점

- 데이터세트, 피처, 모델에 대한 Unity Catalog 거버넌스 지원
- Lakehouse Monitoring을 통한 데이터 모니터링 지원
- 피처 엔지니어링 지원
- 모델 생명주기 관리
- 모델 서빙 지원
- 워크플로우 생성 지원
- Git 통합 지원



< Databricks기반의 ML모델 개발 및 배포 프로세스 >

©2024 Databricks Inc. — All rights reserved

딥러닝과 생성형AI를 위한 Databricks Runtime ML

Deep Learning 개발을 위한 Databricks ML Runtime

- Tensorflow, Keras, PyTorch 딥러닝 라이브러리 지원
- 딥러닝 라이브러리의 MLflow 통합 지원
- Petastorm, Hyperopt, Horovod, Ray 분산처리 라이브러리 지원
- GPU라이브러리가 사전 구성된 GPU 런타임 제공
- 딥러닝 모델 서빙을 위한 GPU 엔드포인트 지원(지정 리전)

LLM등 GenAI 개발을 위한 Databricks ML Runtime

- GenAI 모델 개발을위해 Hugging Face Transformers, LangChain 라이브러리 지원
- Transformers 파이프라인, 모델, 프로세싱 컴포넌트를 위한 MLflow 통합 지원
- Hugging Face, DeepSpeed 오픈소스 도구를 사용하여 사용자 데이터세트로 훈련 가능
- Databricks가 관리형으로 제공하는 파운데이션 모델 API를 제공(DBRX, Llama3, BGE 등)
- 타사 서비스에서 제공하는 외부 모델을 등록 가능(OpenAI, Bedrock, Anthropic 등)

©2024 Databricks Inc. — All rights reserved

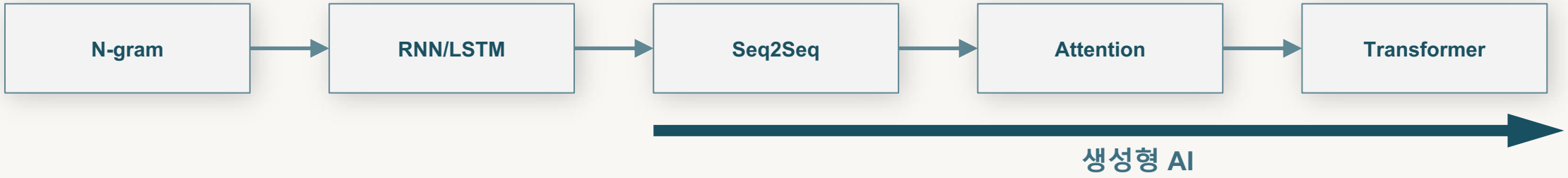
거대 언어 모델
(Large Language Model)

©2024 Databricks Inc. — All rights reserved

언어모델의 발전과정

언어 모델의 역사

- N-gram : 특정 단어 뒤에 등장할 단어를 단순하게 빈도를 계산하여 추정하는 매우 원시적인 언어모델
- RNN/LSTM : 이전 상태값과 현재 상태값을 바탕으로 특정 단어 뒤에 등장할 단어를 확률적으로 계산
- Seq2Seq : RNN을 인코더와 디코더로 분리하여 순차적 번역에 활용
- Attention : Seq2Seq에 Attention 모델을 더하여 이전의 모든 단어를 한번에 인식하여 중요도를 식별
- Transformer : Attention에 LSTM모듈을 제거하여 문장내 단어의 중요도를 저비용으로 고효율의 번역에 활용

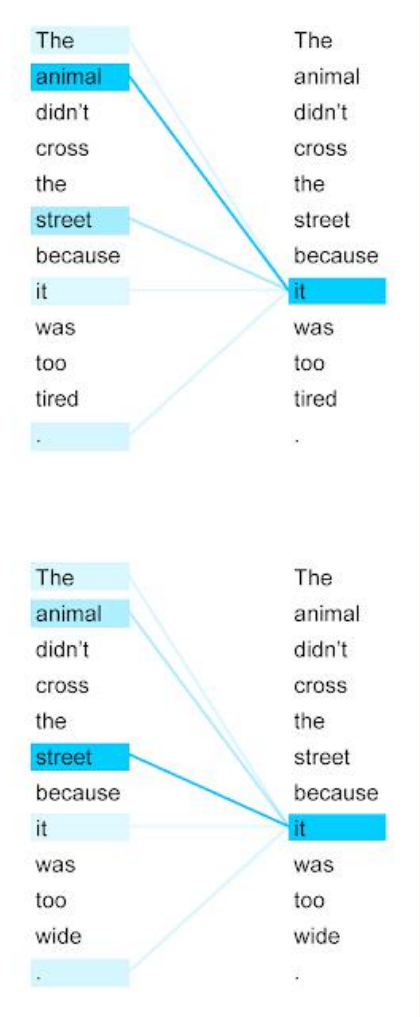


©2024 Databricks Inc. — All rights reserved

Transformer : 언어 이해를위한 Neural Network

The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

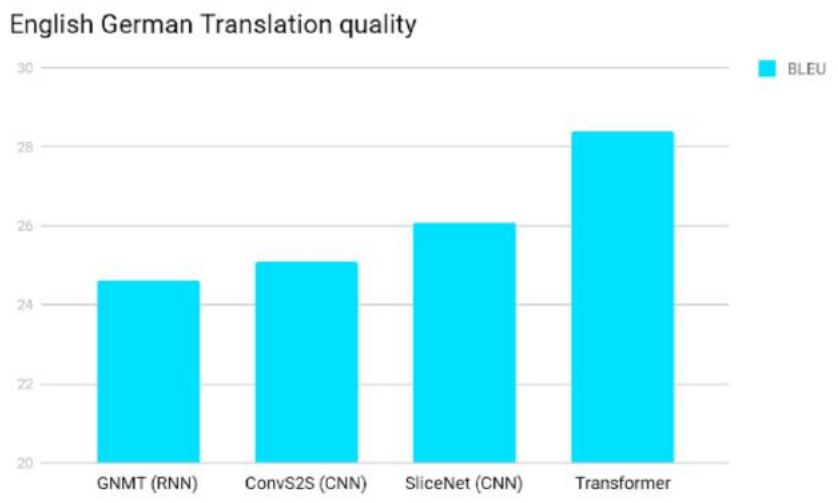
The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.



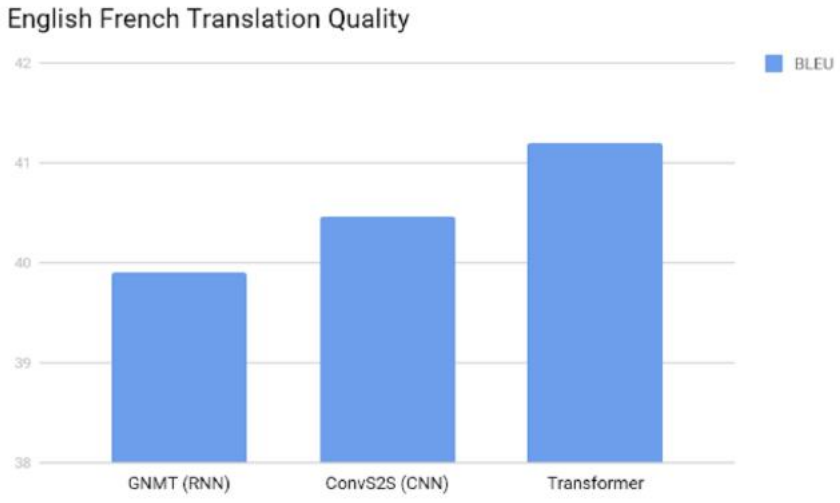
https://blog.research.google/2017/08/transformer-novel-neural-network.html?m=1

©2024 Databricks Inc. — All rights reserved

Transformer : 언어 이해를위한 Neural Network



BLEU scores (higher is better) of single models on the standard WMT newstest2014 English to German translation benchmark.



BLEU scores (higher is better) of single models on the standard WMT newstest2014 English to French translation benchmark.

https://blog.research.google/2017/08/transformer-novel-neural-network.html?m=1

©2024 Databricks Inc. — All rights reserved

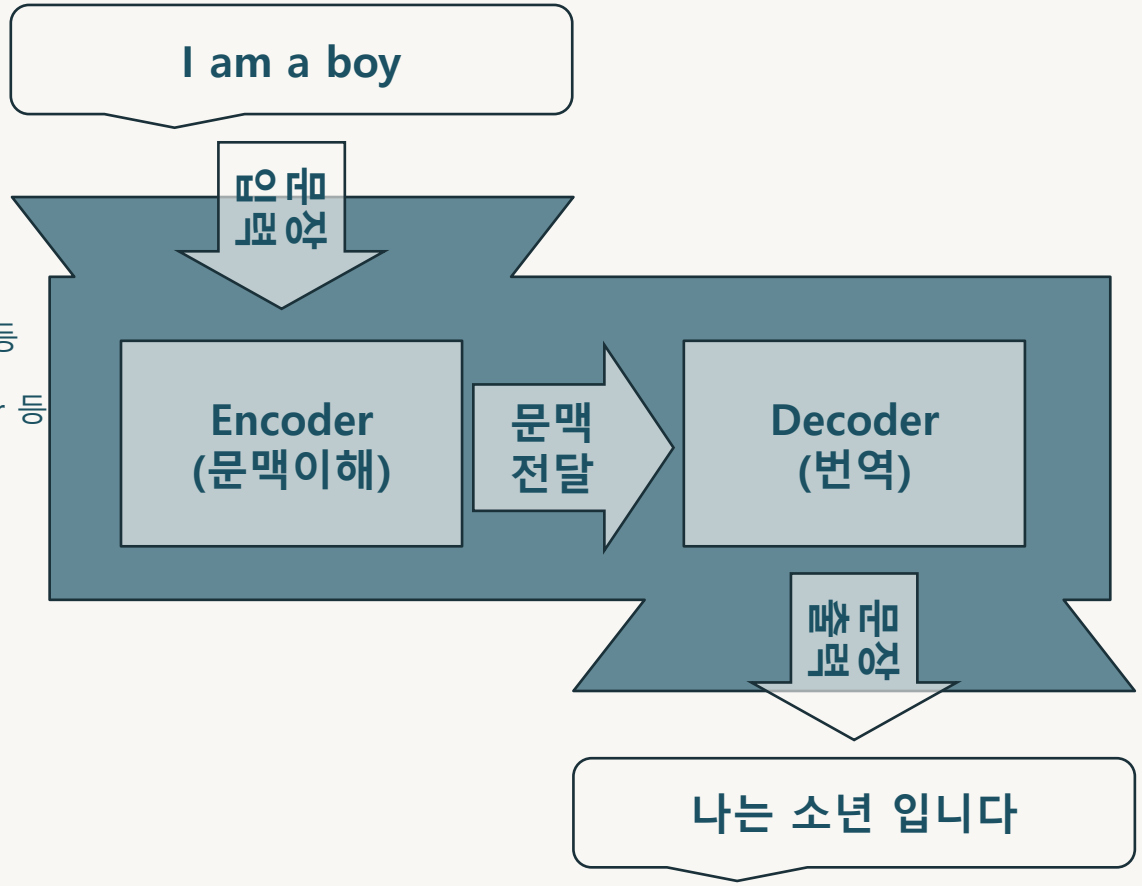
Transformer : 언어 이해를위한 Neural Network

Transformer의 인코더와 디코더

- Encoder : 주로 자연어 이해를 담당
- Decoder : 주로 자연어 생성을 담당

인코더와 디코더를 사용하는 언어모델

- 인코더만 갖고있는 언어 모델 : Titan, BERT, roBERTa, ALBERT 등
- 디코더만 갖고있는 언어 모델 : Claude, Llama, Jurassic, GShard, GPT-3 등
- 인코더 / 디코더를 모두갖는 언어 모델 : seq2seq, T5, BART, Transformer 등

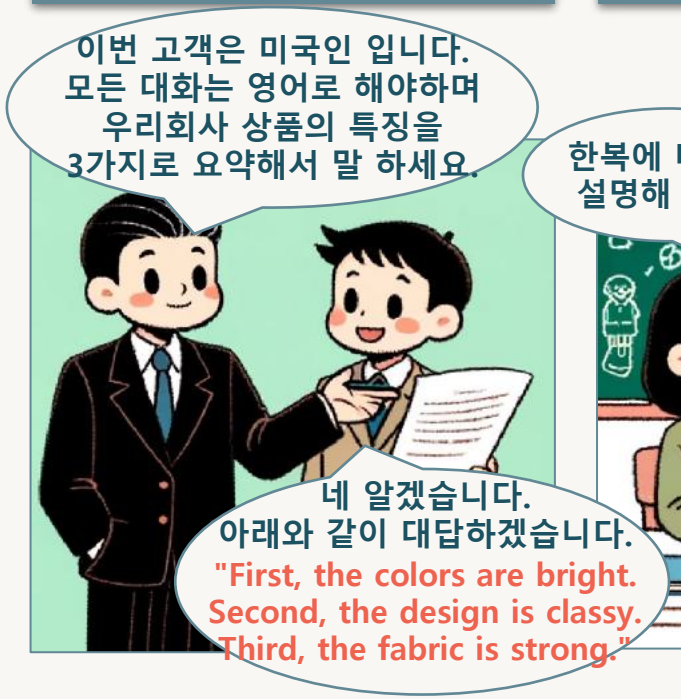


< Encoder / Decoder 구조 >

©2024 Databricks Inc. — All rights reserved

인간이 언어를 습득하는 과정

상황을 설정하여 대화하기



- 지시자가 상황을 설명하고 **주어진 조건내에서 답변**할 수 있도록 지시
- 상황을 설명하여 보다 더 구체적인 답변을 할 수 있도록 가이드를 제공

자료를 토대로 대화하기

- 문서나 책을 참고해가며 오픈북으로 대화
- **자신의 지식 외에 외부자료를 참고**하며 기 학습된 문장 구사능력과 지식을 활용하여 주제 특화된 대화를 진행

지식 학습 하기

- 교사가 지식을 전수
- **문제와 답을 학습**할 수 있도록 학습하여 유사한 문제를 제시하였을때 학습된 지식을 바탕으로 답변할 수 있는 능력을 배양

언어 배우기

- 부모의 말, 동화책, TV등 다양한 매체를 통해 사실 유무와 관계없이 **문법을 습득**
- 질문의 문맥은 이해할 수 있으나 지식이 부족하여 습득한 문장내에서만 답변

©2024 Databricks Inc. — All rights reserved

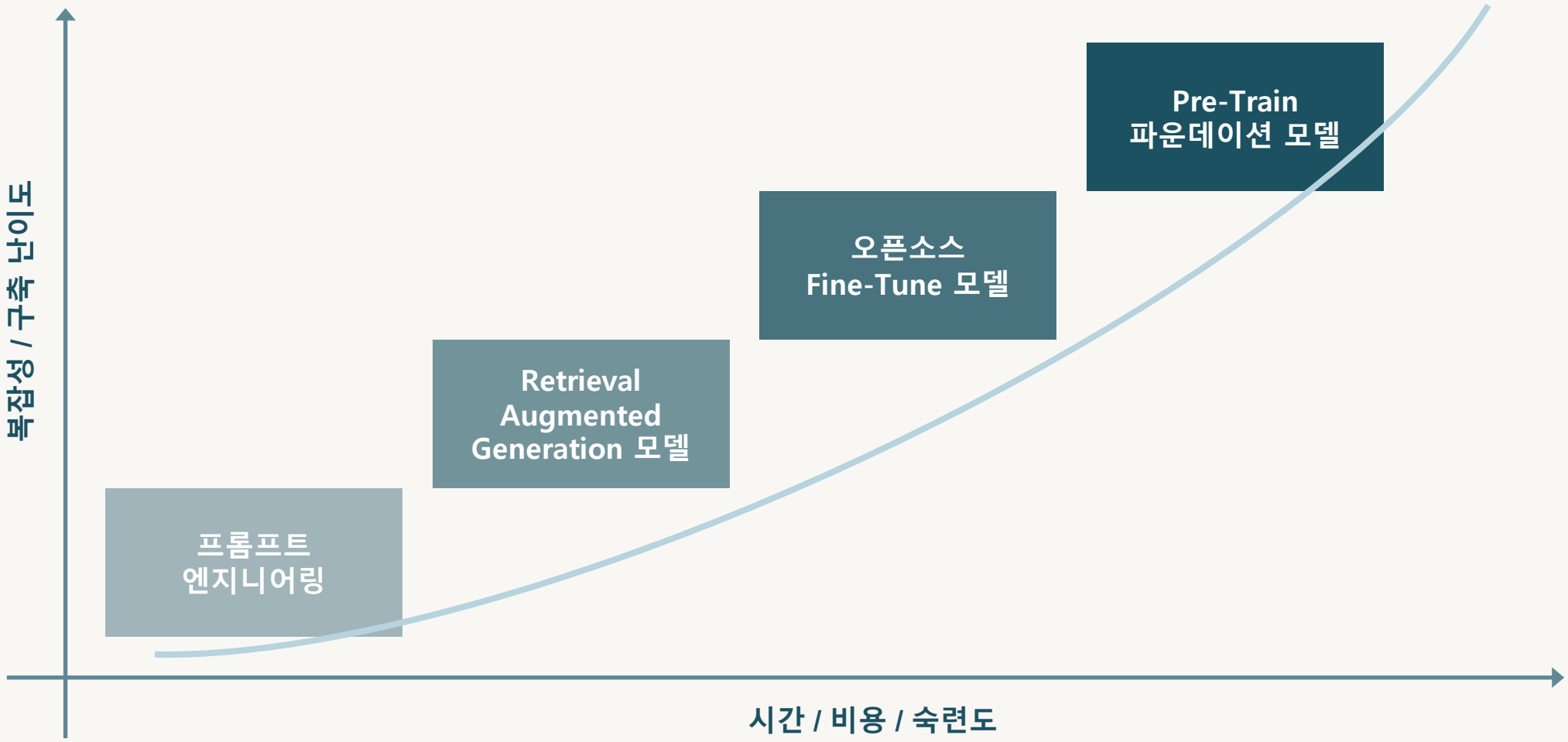
인간이 언어를 습득하는 과정

상황을 설정하여 대화하기	자료를 토대로 대화하기	지식 학습 하기	언어 배우기

프롬프트 엔지니어링 (Prompt Engineering)	검색 증강 생성 (Retrieval-Augmented Generation)	미세 조정 (Fine-tuning)	사전 학습 (Pre-training)

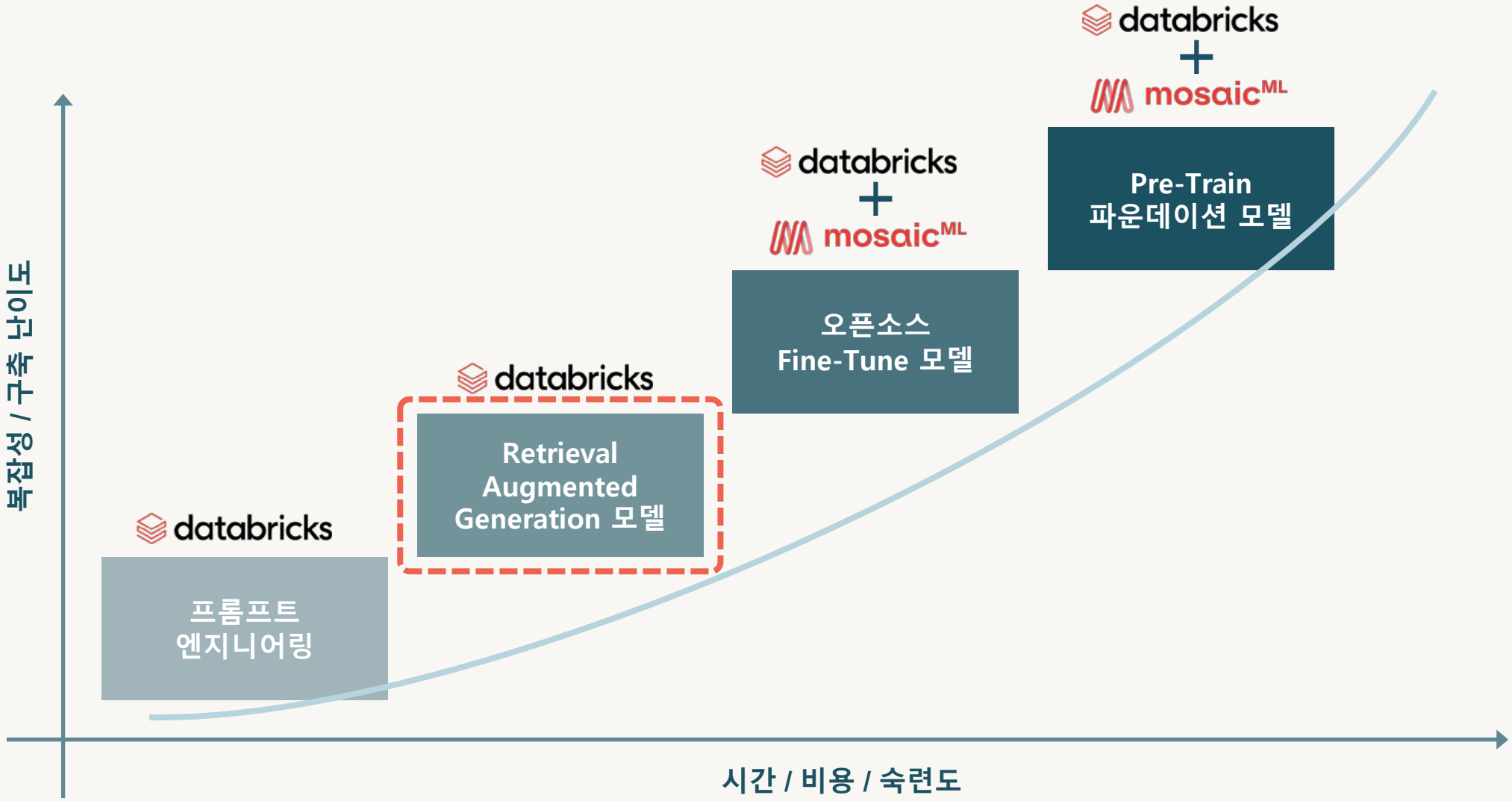
©2024 Databricks Inc. — All rights reserved

생성형 AI의 성숙도 곡선



©2024 Databricks Inc. — All rights reserved

생성형 AI의 성숙도 곡선



©2024 Databricks Inc. — All rights reserved

생성형 AI의 적용방법별 구분

적용방법	정의	주요 사례	데이터 요구사항	학습 시간	장점	고려사항
Prompt engineering	LLM동작을 안내하는 전문화된 프롬프트 생성	민감하지 않은 내용을 다루는 챗봇	없음	없음	빠르고, 비용 효율적이며, 교육이 필요 없음	Fine-tuning에 비해 제어력 부족
RAG (Retrieval Augmented Generation)	LLM을 외부 지식과 결합	내부 문서 검색 챗봇	벡터 데이터베이스	보통 (VectorSearch에 임베딩하는 시간)	동적으로 업데이트되는 컨텍스트, 향상된 정확도	프롬프트 길이와 추론 계산이 크게 증가하여 LLM 추론 비용 증가
Fine-tuning	도메인 특화된 특정 데이터세트를 사전 학습된 LLM에 적용	도메인 특화된 응답, 요약, 번역 시스템	수백만 개의 도메인별 원시 데이터 토큰 / 수천 개의 QA 데이터세트	보통 ~ 장시간 (데이터 크기에 따라 다름)	세분화된 제어, 높은 전문성	레이블링된 데이터가 필요하며 높은 학습비용 발생
Pre-training	LLM 모델을 처음부터 학습	기반이 되는 언어모델 생성	수십억에서 수조 개의 토큰 규모의 코퍼스 데이터세트	장시간 (수일~수주)	특정 요구사항에 맞춘 극대화된 모델	리소스 집약적이며 매우 높은 학습비용 발생

©2024 Databricks Inc. — All rights reserved

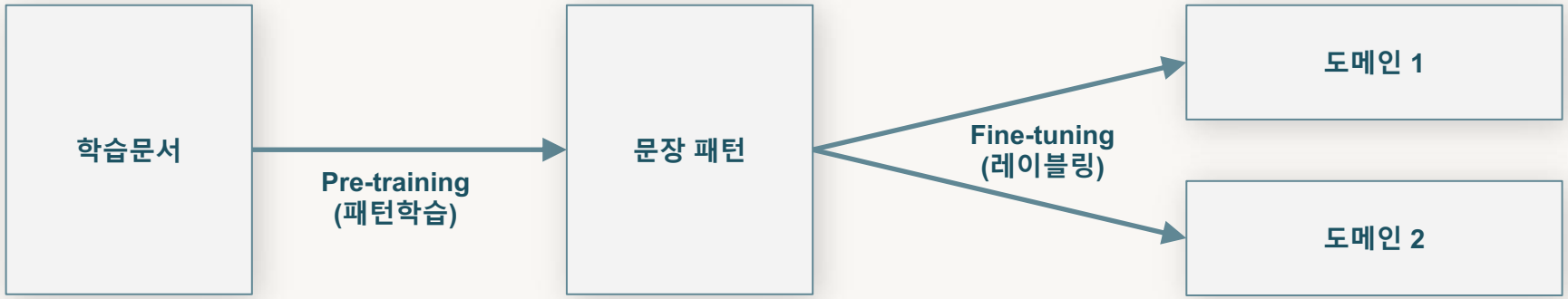
LLM이 사용하는 학습방법

비지도 사전학습 (Unsupervised Pre-training)

- 딥러닝 기법 중 **정답지가 없는(Unlabeled)** 자연어 데이터를 모델에 학습 시키는 기법
- 데이터에서 자연어의 패턴을 학습시키는 과정

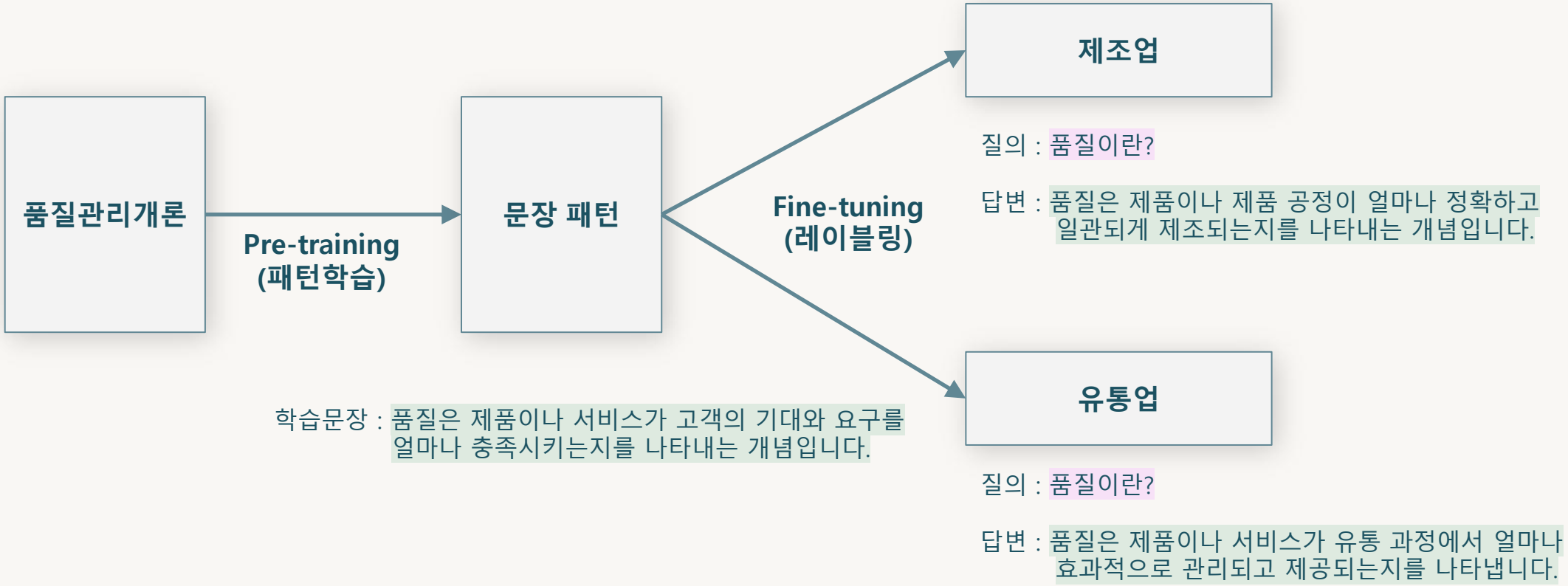
지도 미세조정 (Supervised Fine-tuning)

- 딥러닝 기법 중 사전에 **정답지를 미리 알고(Labeled)** 사전학습 모델에 정답을 학습 시키는 기법
- 주어진 사전학습 모델이 특정 주제영역(도메인)에 어떻게 정답을 출력 해야하는지 학습시키는 과정



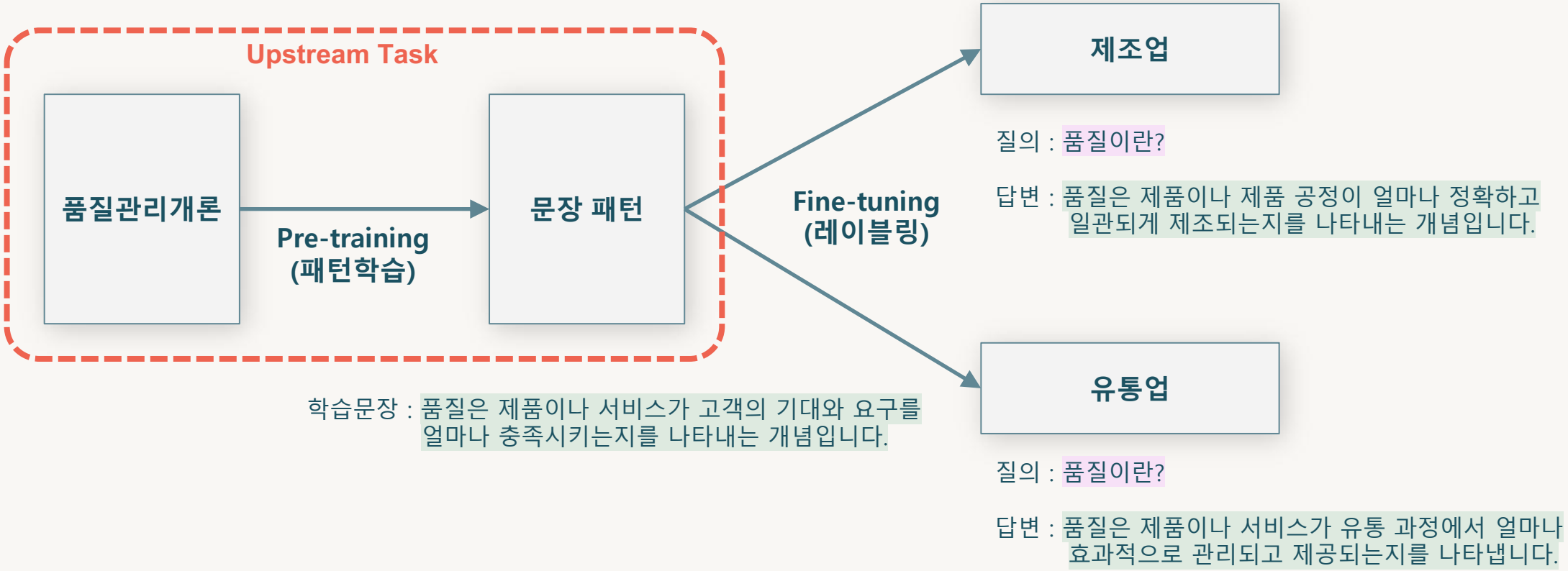
©2024 Databricks Inc. — All rights reserved

LLM이 사용하는 학습방법



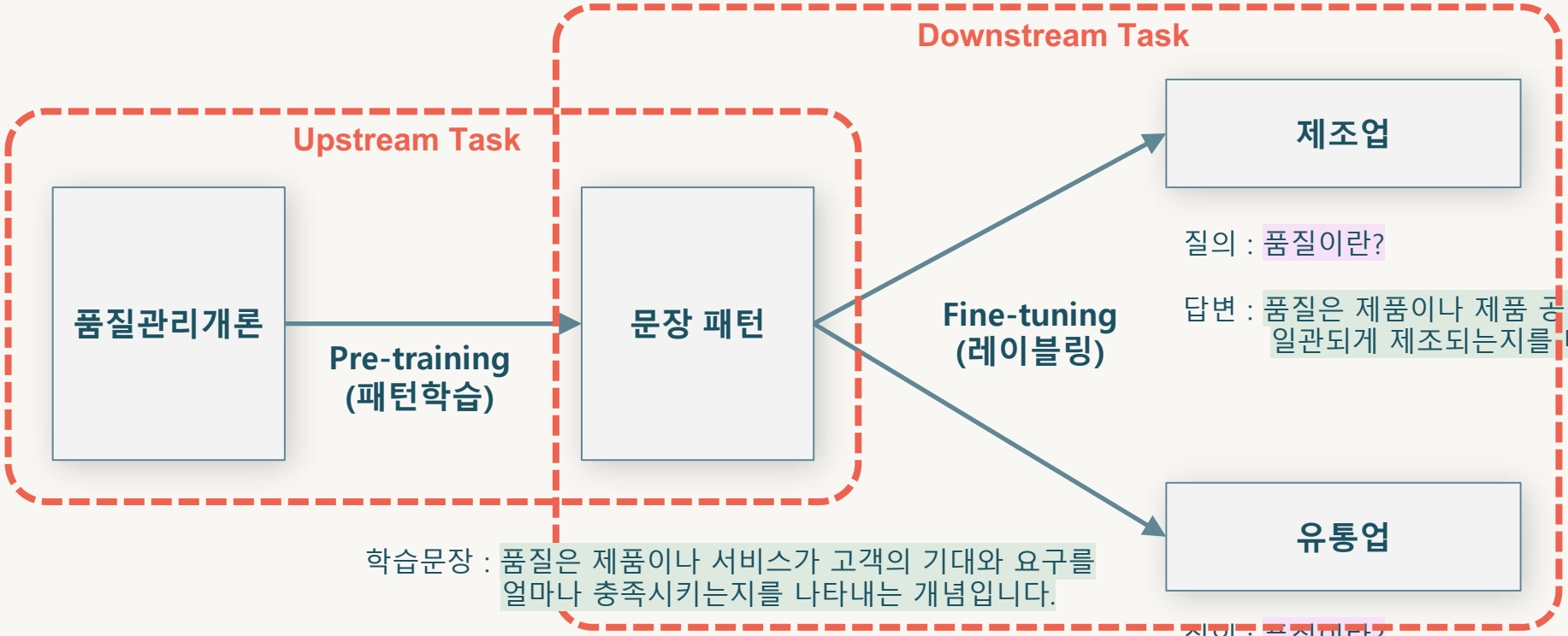
©2024 Databricks Inc. — All rights reserved

LLM이 사용하는 학습방법



©2024 Databricks Inc. — All rights reserved

LLM이 사용하는 학습방법



학습문장 : 품질은 제품이나 서비스가 고객의 기대와 요구를 얼마나 충족시키는지를 나타내는 개념입니다.

질의 : 품질이란?

답변 : 품질은 제품이나 제품 공정이 얼마나 정확하고 일관되게 제조되는지를 나타내는 개념입니다.

질의 : 품질이란?

답변 : 품질은 제품이나 서비스가 유통 과정에서 얼마나 효과적으로 관리되고 제공되는지를 나타냅니다.

©2024 Databricks Inc. — All rights reserved

GPT가 사용한 평가데이터

비지도 사전학습(Unsupervised Pre-training)

- BooksCorpus Dataset (7000권의 책)
- 1B word language model benchmark(10억개 단어)

지도 미세조정(Supervised Fine-tuning)

- Natural language inference : 2가지 명제의 관계를 파악하는 데이터
  - SNLI, MultiNLI, Question NLI, RTE, SciTail
- Question Answering : 질문이 주어졌을때 답을하는 데이터
  - Story Cloze, RACE
- Sentence similarity : 두 문장간의 유사도 데이터
  - MSR Paraphrase Corpus, Quora Question Pairs, STS Benchmark
- Classification : 감성분석, 주어진 문장이 문법적으로 정확한지를 판단
  - Stanford Sentiment Treebank-2, CoLA

Radford et al.(2017), Improving Language Understanding by Generative Pre-Training

©2024 Databricks Inc. — All rights reserved

허깅 페이스의 스코어링

허깅페이스 리더보드

- 허깅 페이스는 각 LLM 모델들이 얼마나 뛰어난 성능을 지니고 있는지에따라 점수를 부여하고 순위를 매긴다
- 허깅 페이스는 AI 오류(Hallucination)을 얼마나 잘 회피하는지에 따른 4가지 벤치마크 모델 평가방법을 제시한다
- Pre-trained, Fine-tuned , Instruction-tuned, RL-tuned model의 4가지 모델이 제출 가능하다

허깅 페이스의 4가지 스코어링 지표

- 추론문제(AI2 Reasoning Challenge)
- 상식추론(Common Sense Inference)
- 문맥이해(Contextual Understanding)
- 사실정확도(Factual Accuracy)

T	Model	Average ↑	ARC	HellaSwag	MMLU	TruthfulQA
◆	garage-bAInd/Platypus2-70B-instruct	73.13	71.84	87.94	70.48	62.26
◆	upstage/Llama-2-70b-instruct-v2	72.95	71.08	87.89	70.58	62.25
◆	deepnight-research/llama-2-70B-inst	72.95	71.08	87.89	70.58	62.25
◆	psmathur/model_007	72.72	71.08	87.65	69.04	63.12
◆	psmathur/orca_mini_v3_70b	72.64	71.25	87.85	70.18	61.27
○	MayaPH/GodziLLa2-70B	72.59	71.42	87.53	69.88	61.54
◆	psmathur/model_009	72.36	71.59	87.7	69.43	60.72
◆	upstage/Llama-2-70b-instruct	72.29	70.9	87.48	69.8	60.97
◆	garage-bAInd/Camel-Platypus2-70B	71.7	71.08	87.6	70.04	58.09
○	garage-bAInd/Camel-Platypus2-70B	71.66	71.25	87.61	69.65	58.15
○	v2ray/LLaMA-2-Wizard-70B-QLoRA	71.5	67.58	87.52	69.11	61.79
◆	quantumaikr/llama-2-70b-fb16-orca-chat-10k	71.48	68.09	87.07	69.21	61.56
◆	v2ray/LLaMA-2-Wizard-70B-QLoRA	71.48	67.75	87.5	68.96	61.72

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

©2024 Databricks Inc. — All rights reserved

허깅 페이스의 스코어링

벤치마크

- 벤치마크에 사용된 데이터는 모두 영어 데이터로 모델의 품질이 한국어에도 적용된다는 보장은 없음

평가 벤치마크	데이터 내용	언어	사용 데이터세트
추론문제(AI2 Reasoning Challenge)	초등학교 과학과목 질문 세트	영어	ARC
상식추론(Common Sense Inference)	사람은 쉽게 풀수있는 상식문제 세트	영어	HellaSwag
문맥이해(Contextual Understanding)	초등수학, 미국역사, 컴퓨터과학, 법률등 57개 과제의 자연어 세트	영어	MMLU
사실정확도(Factual Accuracy)	거짓을 판단하는 테스트 세트	영어	TruthfulQA

- 모델의 평가 지표 : BLEU, GLUE, MAUVE 등 120여개 (huggingface.co/metrics)

언어 모델에 평가지표(Metric)가 필요한 이유?

- AI기술을 기반으로하는 거대 언어 모델이, 인간의 언어능력을 어느 수준까지 모방할 수 있는지를 측정하기위함
- 인간도 언어를 이해하고 활용하는 방법이 여럿 있듯이, LLM도 그 수준을 평가하기위한 다양한 방법이 존재

©2024 Databricks Inc. — All rights reserved

언어모델의 평가지표

BLEU 평가지표

- 번역의 품질을 평가 하기위한 평가 알고리즘
- 기계번역의 결과가 인간의 번역결과에 가까울수록 높은 점부를 부여
- 문장의 문법 정확성은 고려되지 않음

BLEU 입력값

- 예측(predictions) : 기계가 번역한 문장
- 정답(reference) : 인간이 번역한 문장

```python

from nltk.translate.bleu_score import sentence_bleu

reference = [["그는", "도시", "공원에서", "달리기를", "한다"]]
predictions = ['그녀는', "공원에서", "달리기를", "하였다"]

bleu_score = sentence_bleu(reference, predictions, weights=(1, 0, 0, 0))

print(bleu_score)

```

```
0.38940039153570244
```

BLEU 출력값

- 유사도 : 0과 1사이의 실수로 1에 가까울 수록 번역의 품질이 사람의 품질과 유사하다

©2024 Databricks Inc. — All rights reserved

언어모델의 평가지표

GLUE 평가지표

- 일반 언어를 이해하는 9가지 능력을 측정하는 벤치마크 도구*

데이터세트	태스크	데이터 수(학습/평가)	평가지표	레이블
CoLA	문장이 문법적, 언어학적으로 수용 가능한지 분류하는 이진분류 문제	8.5k / 1k	매튜 상관계수	acceptable / not acceptable
SST-2	영화 리뷰데이터의 긍정과 부정을 분류하는 이진분류 문제	67k / 1.8k	정확도	positive / negative
MRPC	주어진 두개의 문장이 동일한 의미의 문장인지를 분류하는 이진분류 문제	3.7k / 1.7k	가중조화평균, 정확도	same / not same
STS-B	뉴스 데이터에서 쌍으로 이루어진 두개의 문장을 비교하여 유사도를 5가지로 분류하는 다중분류 문제	7k / 1.4k	피어슨 상관계수, 스피어만 상관계수	1 ~ 5 (similarity score)
QQP	주어진 두개의 질문이 동일한 의미의 질문인지를 분류하는 이진분류 문제	364k / 391k	가중조화평균, 정확도	same / not same
MNLI	전제와 가설의 쌍으로 이루어진 문장의 모순, 함의, 중립으로 분류하는 다중분류 문제	393k / 20k	일치 정확도, 불일치 정확도	contradiction / entailment / neutral
QNLI	질문과 문답의 쌍으로 이루어진 문장에서 비함의, 함의로 분류하는 이진분류 문제	105k / 5.4k	정확도	not entailment / entailment
RTE	전제와 가설 쌍으로 이루어진 문장의 비함의, 함의로 분류하는 이진분류 문제	2.5k / 3k	정확도	not entailment / entailment
WNLI	지시 대명사가 포함된 문장에서 대명사가 지시하는 대상을 선택하는 추론 문제	634 / 146	정확도	-

* Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

©2024 Databricks Inc. — All rights reserved

LLM모델의 종류

©2024 Databricks Inc. — All rights reserved

밀집모델과 전문가혼합모델

통합 전문가

전문가 그룹

©2024 Databricks Inc. — All rights reserved

밀집모델과 전문가혼합모델

통합 전문가

우리회사의 지난해 대비 올해 매출실적을 비교했을때 내년 매출을 증대하기위한 전략은 무엇이 있을까요?

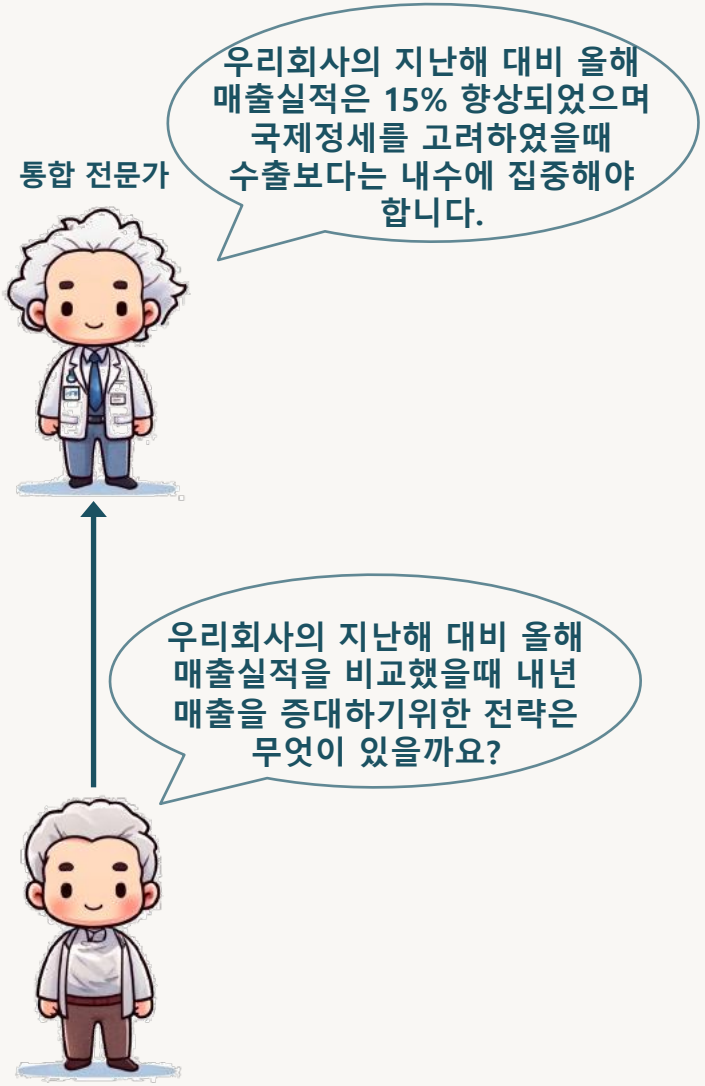
전문가 그룹

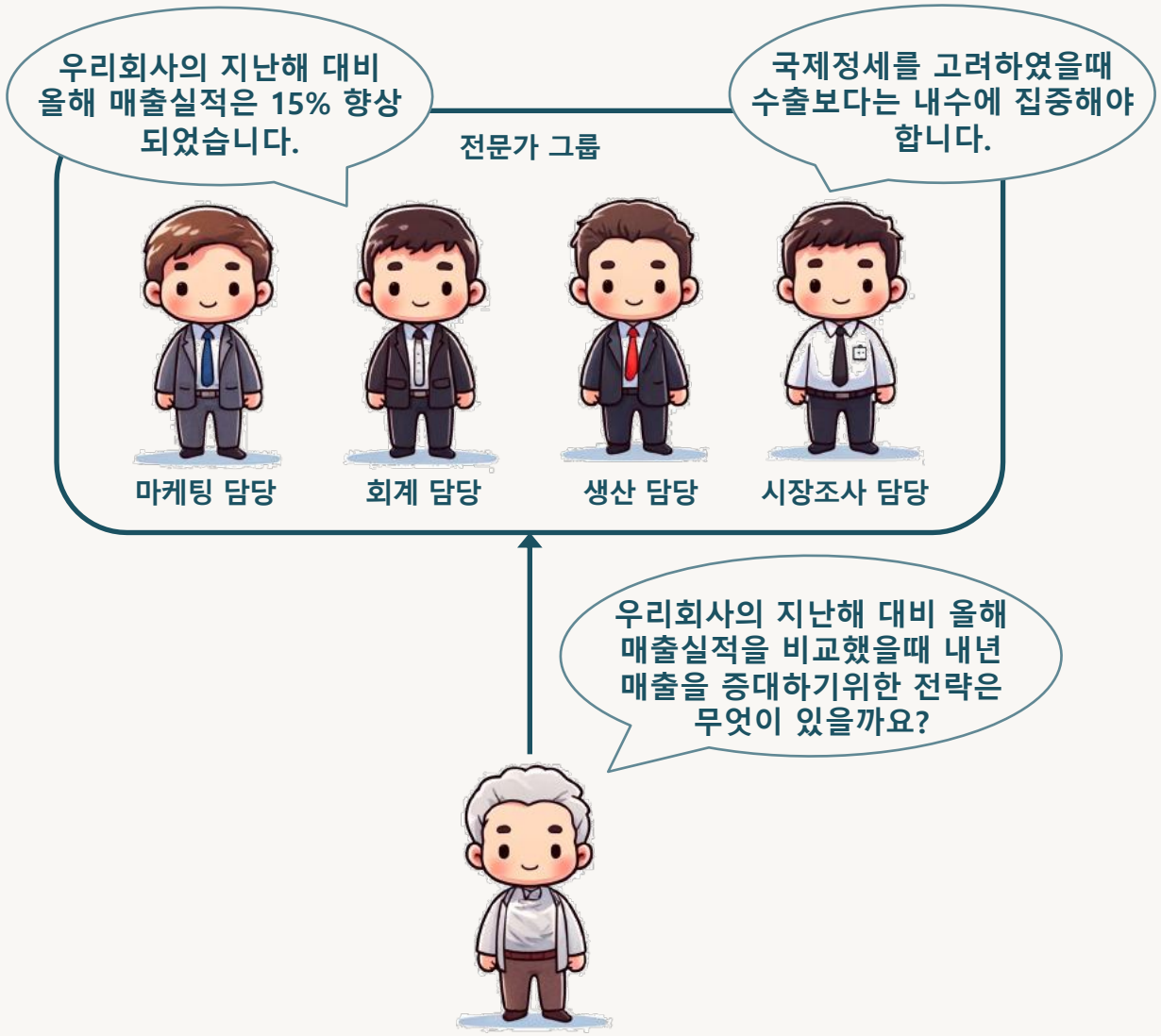
마케팅 담당 회계 담당 생산 담당 시장조사 담당

우리회사의 지난해 대비 올해 매출실적을 비교했을때 내년 매출을 증대하기위한 전략은 무엇이 있을까요?

©2024 Databricks Inc. — All rights reserved

밀집모델과 전문가혼합모델





©2024 Databricks Inc. — All rights reserved

밀집형모델과 전문가혼합형모델 특징 비교

밀집형(Dense) LLM모델

- 고밀도의 단일 네트워크를 사용하는 모델
- 7B, 13B, 70B 등의 파라미터 모델을 사용
- 모든 파라미터가 자연어 생성에 활성화됨
- 문맥의 이해깊이가 높음
- 토큰당 처리되는 시간이 느림
- 모델 구현이 쉬움
- 출력의 일관성을 유지
- 멀티모달 모델 구현에 불리
- GPT 3.5, Llama, MPT, Mistral 등이 대표적 모델

전문가 혼합형(Mixtral of Experts) LLM모델

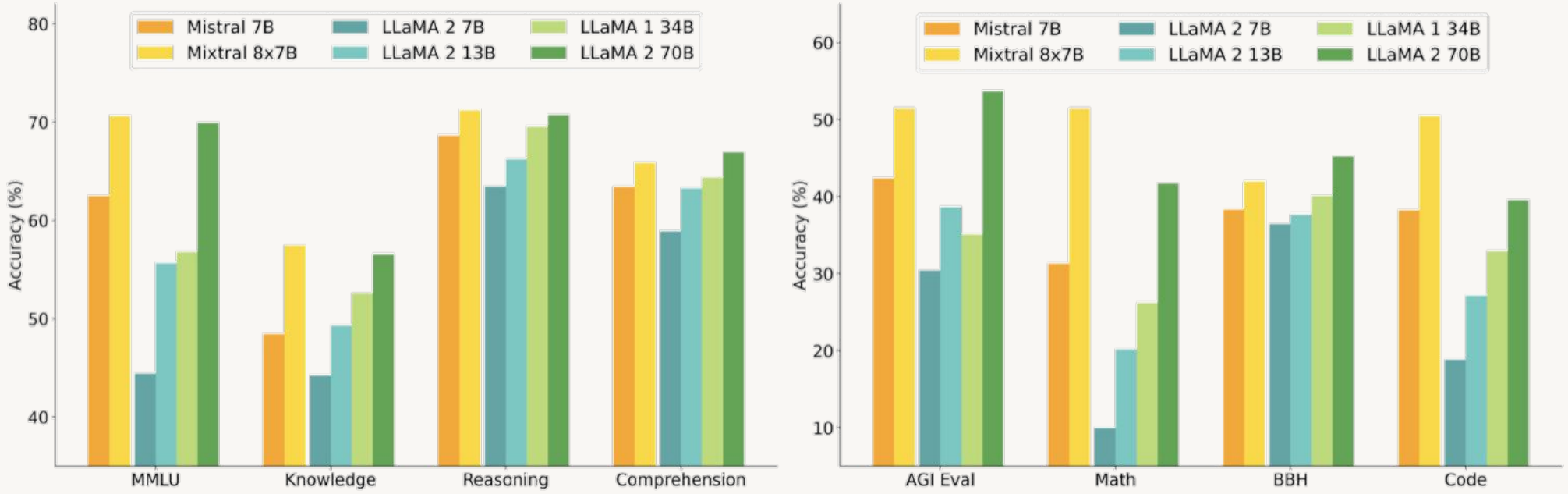
- 희소밀도의 여러 네트워크를 혼합한 모델
- 7B X 2, 7B X 8 등의 파라미터 모델을 사용
- 일부 선택된 전문가 모델의 파라미터들만 활성화됨
- 문맥의 이해깊이가 높음
- 토큰당 처리되는 시간이 빠름
- 라우팅 로직이나 전문가관리가 필요하여 모델 구현이 어려움
- 출력의 일관성을 유지하지 못할 수 있음
- 멀티모달 모델 구현에 유리
- GPT 4, Mixtral, Gemini, DBRX 등이 대표적 모델

©2024 Databricks Inc. — All rights reserved

밀집형모델과 전문가혼합형모델 성능비교

Mistral AI사의 성능비교 연구

- 7B X 8 MoE 모델은 70B Dense 모델의 성능과 유사하거나 상회
- 7B X 8 MoE 모델은 13B Dense 모델의 속도와 유사하거나 상회

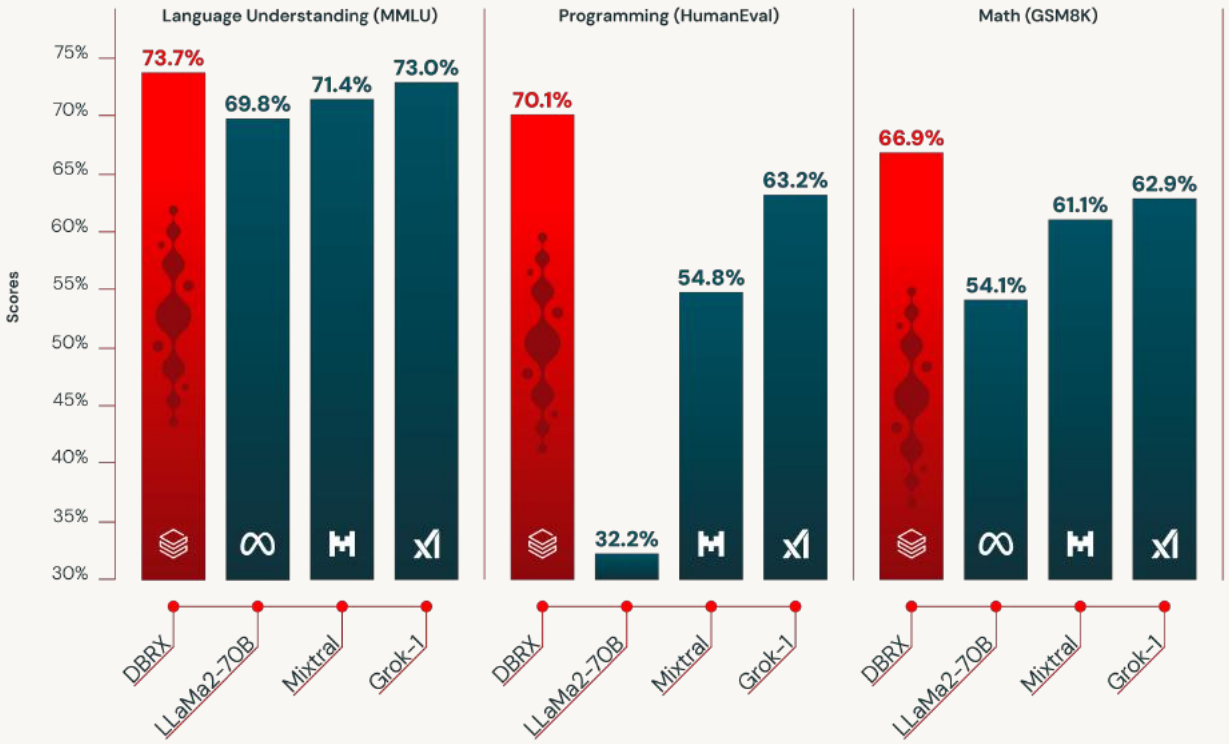


Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... & Sayed, W. E. (2024). Mixtral of experts. arXiv preprint arXiv:2401.04088.

©2024 Databricks Inc. — All rights reserved

밀집형모델과 전문가혼합형모델 성능비교

Databricks의 성능비교 연구



< 벤치마킹 결과비교 >

Model	DBRX Instruct	Mixtral Instruct	GPT-3.5 Turbo (API)	GPT-4 Turbo (API)
Answer in Beginning Third of Context	45.1%	41.3%	37.3%	49.3%
Answer in Middle Third of Context	45.3%	42.7%	37.3%	49.0%
Answer in Last Third of Context	48.0%	44.4%	37.0%	50.9%
2K Context	59.1%	64.6%	36.3%	69.3%
4K Context	65.1%	59.9%	35.9%	63.5%
8K Context	59.5%	55.3%	45.0%	61.5%
16K Context	27.0%	20.1%	31.7%	26.0%
32K Context	19.9%	14.0%	—	28.5%

< Long-Context Task 품질비교 >

https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm

©2024 Databricks Inc. — All rights reserved

밀집형모델과 전문가혼합형모델 성능비교

Databricks의 성능비교 연구

Model	DBRX Instruct	Mixtral Instruct	Mixtral Base	LLaMA2-70B Chat	LLaMA2-70B Base	Grok-1
Open LLM Leaderboard (Avg of next 6 rows)	74.5%	72.7%	68.4%	62.4%	67.9%	-
ARC-challenge 25-shot	68.9%	70.1%	66.4%	64.6%	67.3%	-
HellaSwag 10-shot	89.0%	87.6%	86.5%	85.9%	87.3%	-
MMLU 5-shot	73.7%	71.4%	71.9%	63.9%	69.8%	73.0%
Truthful QA 0-shot	66.9%	65.0%	46.8%	52.8%	44.9%	-
WinoGrande 5-shot	81.8%	81.1%	81.7%	80.5%	83.7%	-
GSM8k CoT 5-shot maj@1	66.9%	61.1%	57.6%	26.7%	54.1%	62.9% (8-shot)
Gauntlet v0.3 (Avg of 30+ diverse tasks)	66.8%	60.7%	56.8%	52.8%	56.4%	-
HumanEval 0-Shot, pass@1 (Programming)	70.1%	54.8%	40.2%	32.2%	31.0%	63.2%

https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm

©2024 Databricks Inc. — All rights reserved

전문가혼합형모델 적용 예시

생성형AI 적용 범위

- 내부 직원을 대상으로하는 지식검색, 문서요약, 자연어 기반 리포트 생성, 홍보 이미지 생성 등
- 외부 사용자를 대상으로하는 상품문의, FAQ, 자연어 기반 서비스 조합, 자연어 기반 상품 검색 등

생성형AI 활용 예시

예시	설명	필요 기술
개인화 마케팅	마케팅 이메일을 고객의 취향에 맞게 디자인 문구등을 개인화하여 생성한뒤 전달	Image AI
가상 피팅룸	자신의 전신 사진에 상품을 정교하게 렌더링하여 가상 피팅룸 서비스를 제공	Image AI
챗봇/고객 서비스 자동화	제품 정보, 실시간 주문정보를 자연어 챗봇 서비스로 제공	LLM AI
배송 및 설치일정 예약	상품의 배송, 설치일정, 옵션을 자연어 기반으로 예약	LLM AI
보안 위협, 해킹 징후 파악	서비스내의 커뮤니케이션을 분석하여 피싱시도를 포착	LLM AI
매장안내 도우미	매장내 제품의 위치확인을 자연어 기반으로 탐색	LLM AI
재고 및 수요 플래닝	기존 정형 데이터와 더불어 소비자 심리, 행동, 경쟁사 정보를 분석하여 재고 및 수요 플래닝	LLM AI
제품 개발	시장 동향 및 소비자 요구를 파악하여 새로운 제품 디자인을 자동으로 생성	Image AI
예측모델 개발	수치적 분석 보고서 해석을 분석가에 의지하지 않고 자연어로 해석을 생성	LLM AI

https://www.databricks.com/blog/2023/04/13/retail-age-generative-ai.html

©2024 Databricks Inc. — All rights reserved

사전학습(Pre-training)을 위한 Databricks의 기능

©2024 Databricks Inc. — All rights reserved

생성형AI 개발의 난관

생성형AI 모델 개발시 대규모 데이터 학습에 대한 고난도의 스킬이 필요로됨

- RAG모델 구현시 백터DB 사용의 이질감과 난해함
- 분산처리 + 병렬처리 모델 학습 구현의 어려움
- 분산처리 + 병렬처리시 상황재현 또는 디버깅의 어려움
- 분산처리 + 병렬처리시 샘플링의 어려움
- 이상의 기술문제 해결을 하지 못한다면 싱글노드 병렬처리에만 의존해야 하고 결국 모델 학습에 매우 많은 비용과 시간이 소요

어떻게 해결 할 수 있을까?

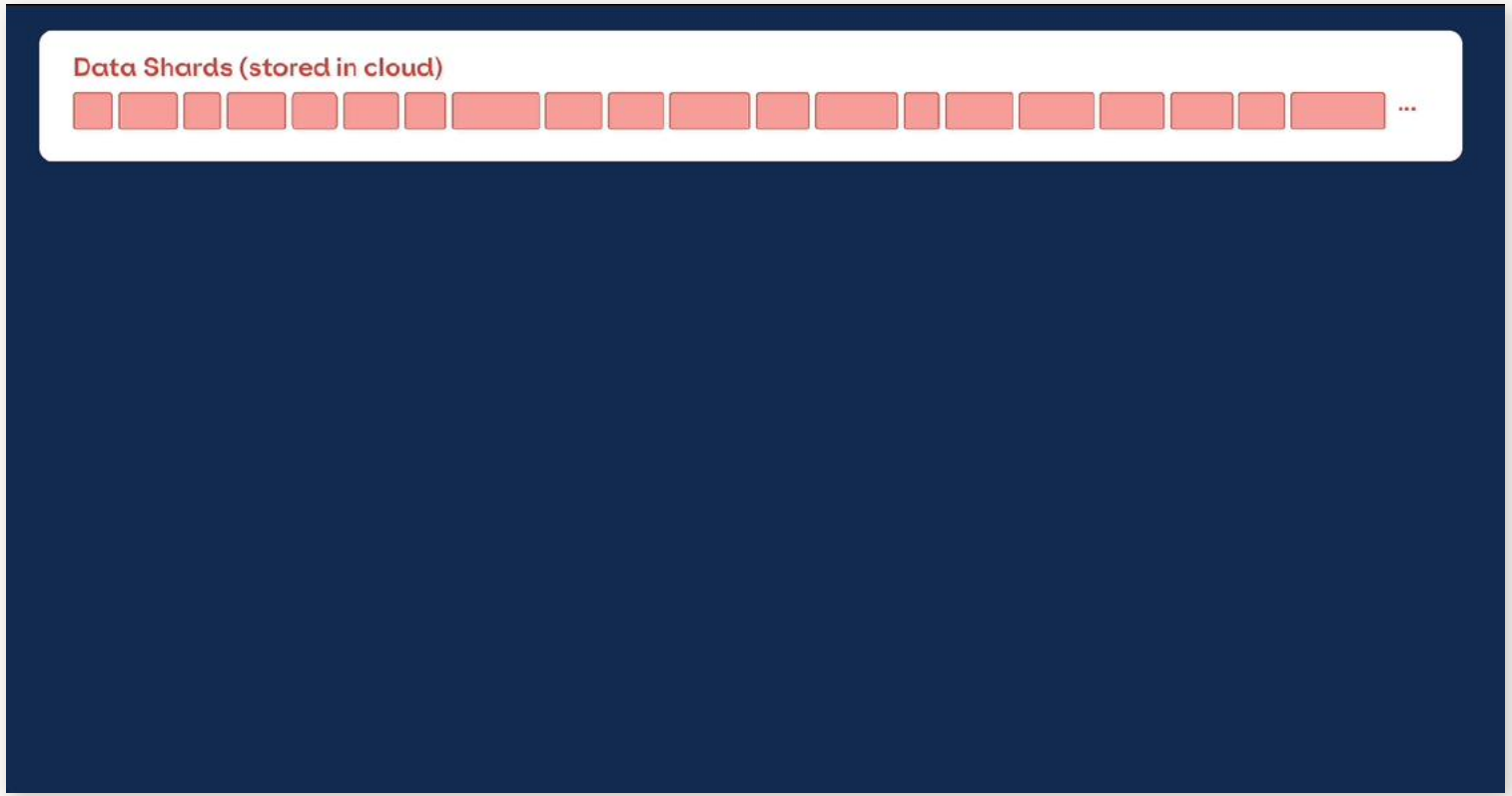
- 백터DB의 활용 -> 테이블과 싱크가능한 직관적인 백터DB의 도입 } databricks
- 학습 데이터 준비 -> 인메모리 분산처리 기술을 통해 대규모의 데이터를 딥러닝 학습 데이터 셋으로 준비
- 모델 학습 -> 간소화되고 자동화된 분산/병렬처리 도구들을 통해 손쉽게 모델 학습
- 모델 디버깅 -> 모델 학습간 상황을 재현하여 불필요한 비용과 시간을 절약 } mosaicML
- 모델 성능향상 -> 자동화된 샘플링 기법적용으로 데이터 편향성을 제거하고 샘플링 우려없이 모델 생성

©2024 Databricks Inc. — All rights reserved

MosaicML StreamingDataset

StreamingDataset이란?

- 클라우드 스토리지에 저장된 대규모 데이터 세트에 대해 빠르고, 비용효율적고, 확장가능한 다중노드 분산 모델학습을 제공하기위한 새로운 데이터 포멧
- StreamingDataset은 이미지, 텍스트, 비디오 및 멀티모달 데이터를 포함한 모든 데이터 유형과 호환
- 주요 클라우드 공급자의 클라우드 스토리지(AWS , OCI , GCS , Azure , Databricks, Cloudflare R2 , Coreweave , Backblaze)를 지원
- PyTorch의 IterableDataset 클래스를 대체하여 사용할 수 있도록 설계하여 기존 모델학습 워크플로에 손쉽게 통합



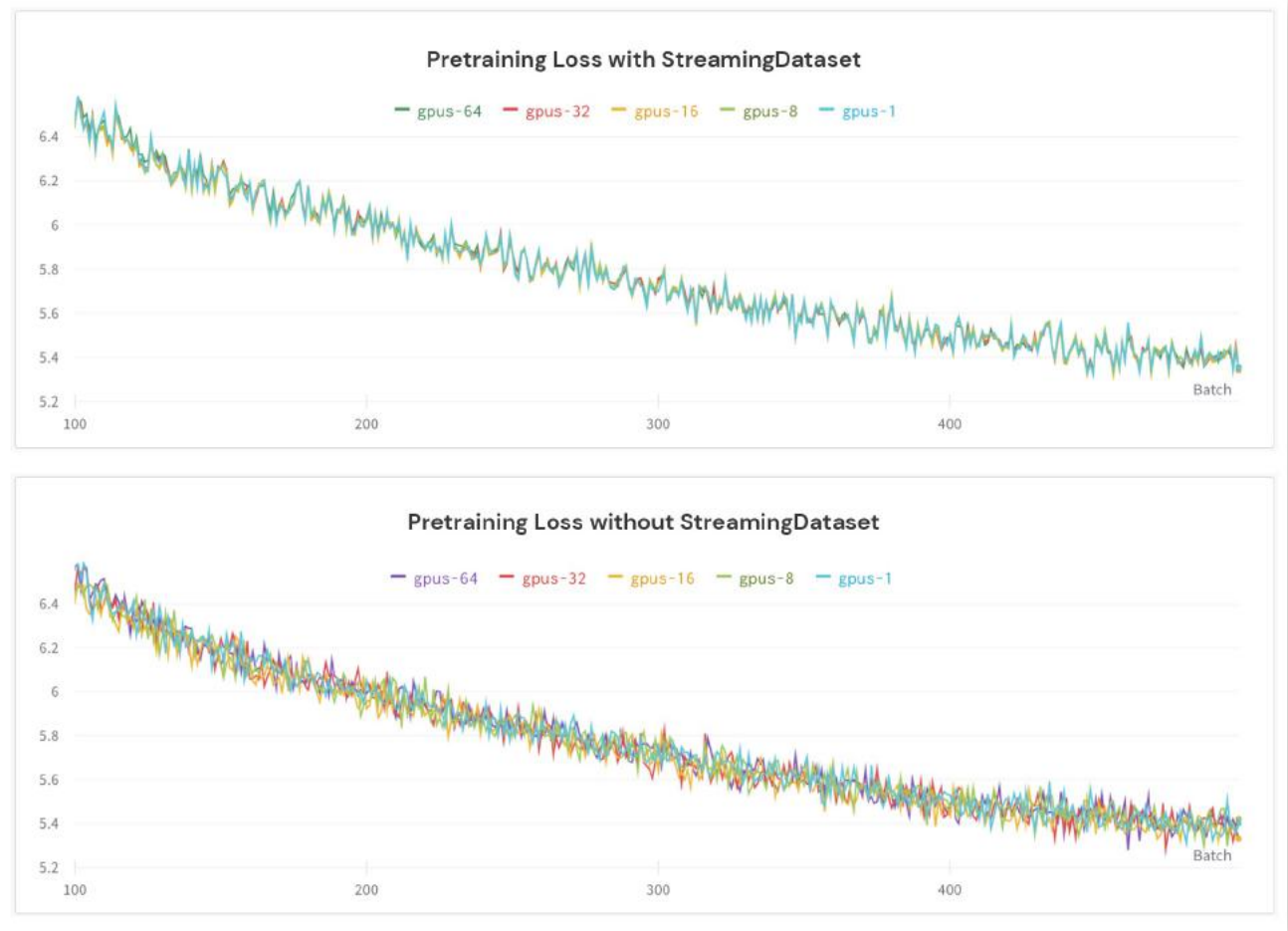
< StreamingDataset이 처리되는 메커니즘 >

©2024 Databricks Inc. — All rights reserved

MosaicML StreamingDataset

StreamingDataset의 작업 일관성

- 결정론적 샘플순서로 작업 일관성 유지
- StreamingDataset을 사용하면 GPU수, GPU노드, CPU수에 관계없이 배치 샘플의 순서가 동일하게 적용
- 데이터 학습 상황이나 로스 스파이크 상황의 재현과 디버그 가능
- 예를들면, 64개의 GPU에서 학습된 체크포인트를 로드하여 8개의 GPU에서 디버그 가능



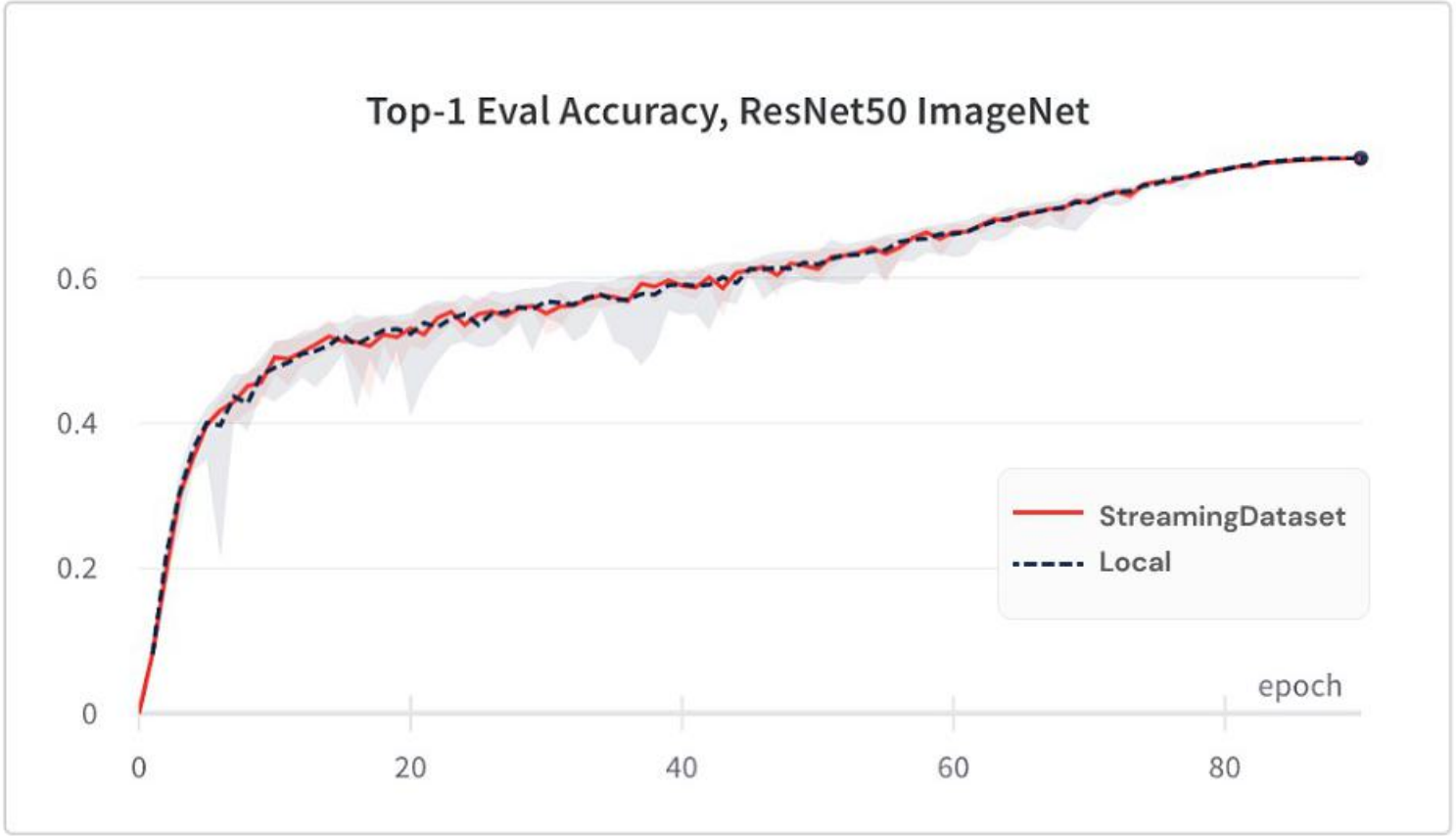
< StreamingDataset 사용 유무에 따른 GPU수량별 Loss율 일관성 비교 >

©2024 Databricks Inc. — All rights reserved

MosaicML StreamingDataset

StreamingDataset의 성능

- StreamingDataset을 사용하는 경우 로컬 스토리지보다 상대적으로 느린 클라우드 스토리지를 사용함에도 불구하고 스트리밍동안의 자체 셔플링 알고리즘이 적용되어 데이터 편향성을 완화하여 성능을 유지
- 로컬 디스크를 사용한 로컬 셔플링기법과 StreamingDataset의 스트리밍 셔플링기법의 모델 성능의 차이가 거의 없으며 에포크의 수가 많아질수록 두 환경간의 성능은 수렴됨



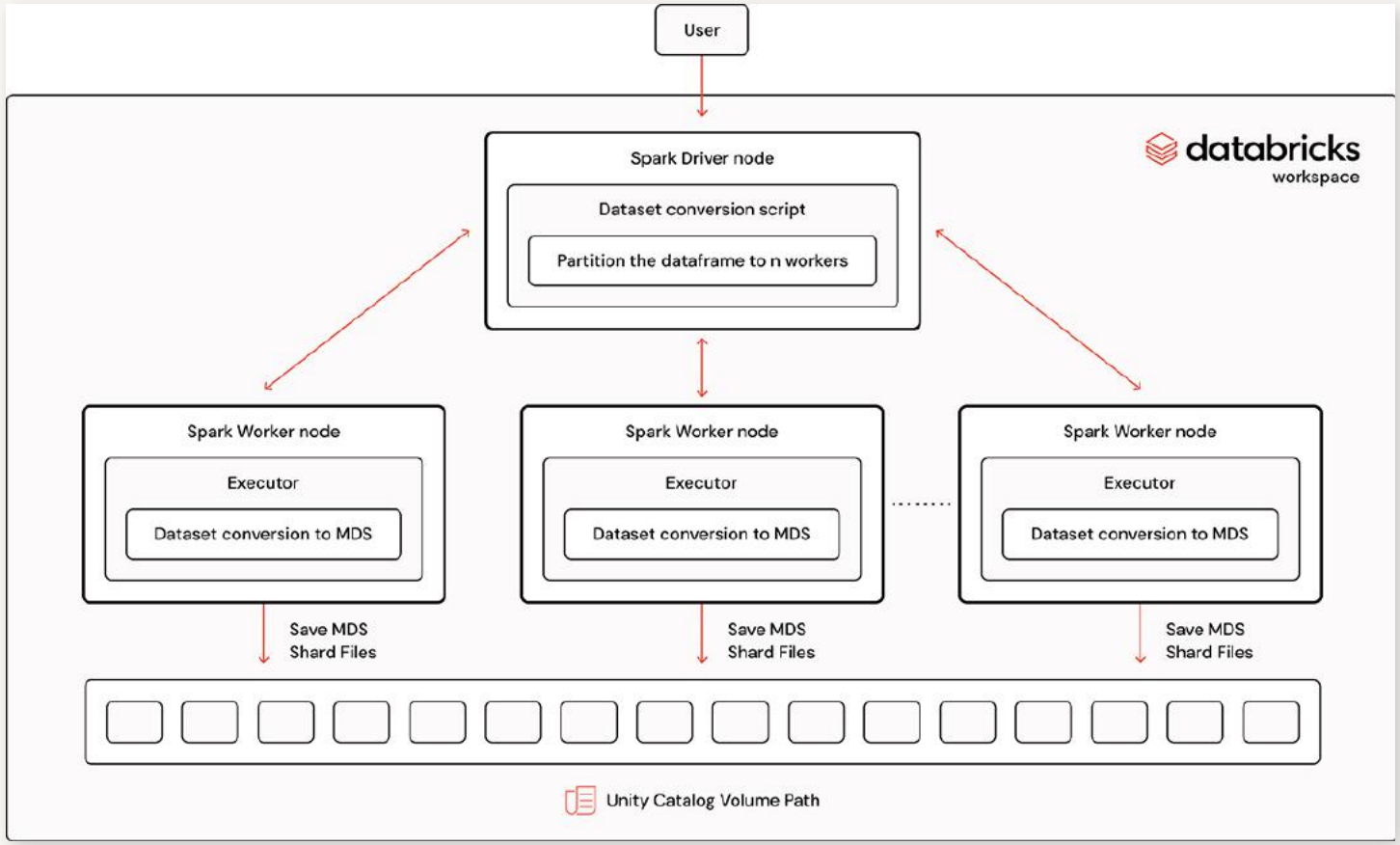
< 로컬 디스크를 사용하는 경우와 StreamingDataset을 경우의 모델성능 비교 >

©2024 Databricks Inc. — All rights reserved

Databricks DIP + MosaicML

StreamingDataset MDS 생성

- Databricks의 Spark 클러스터를 사용하여 원시 데이터(CSV, JSON, Parquet 등)을 MDS(Mosaic Data Shard) 형태로 변환
- Spark의 인메모리 컴퓨팅, 분산처리, 확장성 기능을을 활용하여 고속으로 변환
- 중복제거, 토큰화와같은 데이터 전처리 작업을 함께 진행
- 변환된 MDS 파일은은 Databricks Unity Catalog Volume 저장소에 저장하여 거버넌스 체계하에서 관리



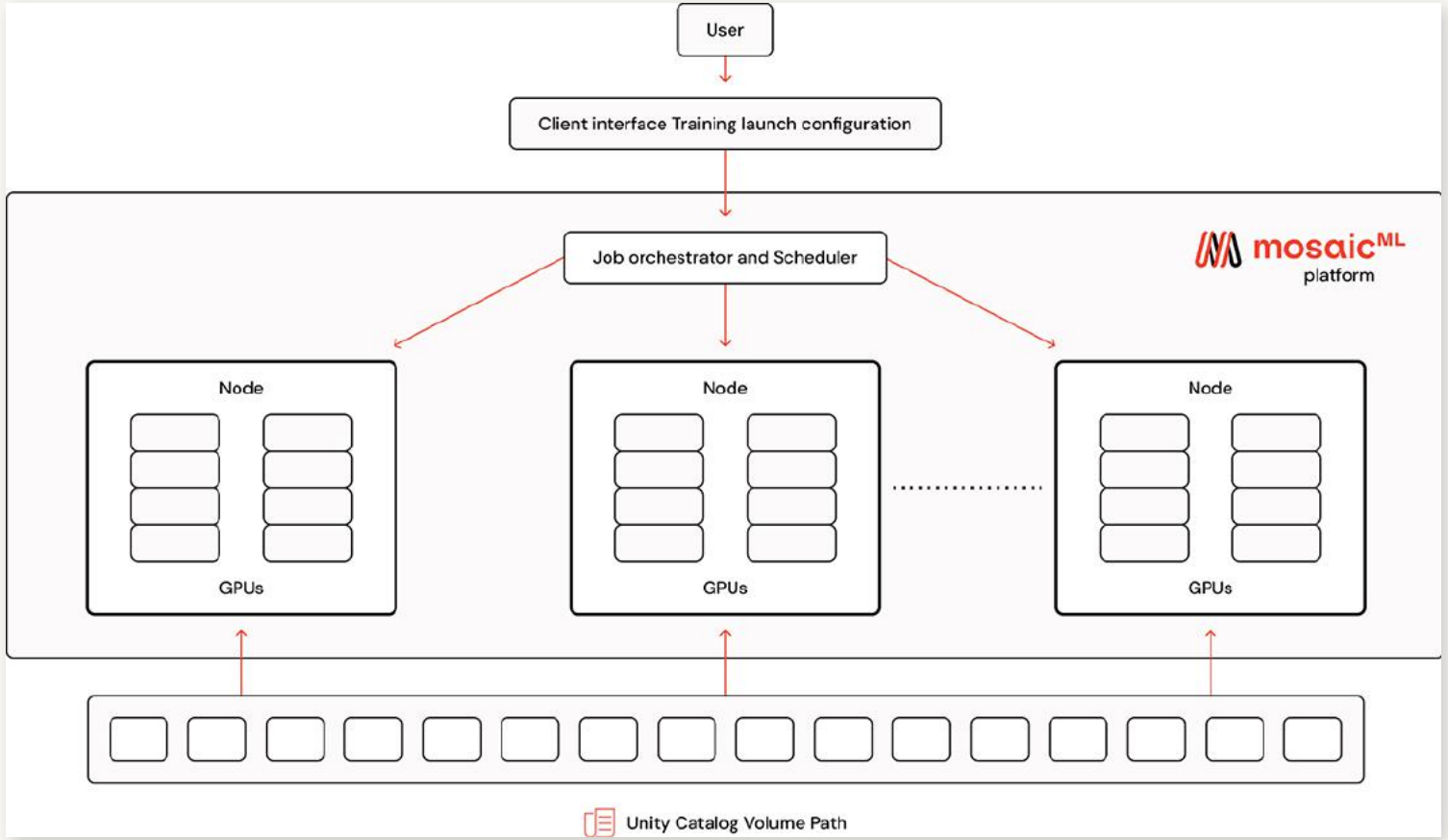
< 원시 소스의 데이터 전처리 및 MDS 포멧변환 구조 >

©2024 Databricks Inc. — All rights reserved

Databricks DIP + MosaicML

StreamingDataset MDS 학습

- MosaicML 플랫폼에서 Databricks DI 플랫폼의 Volume에 접근하기위해 Databricks PAT(Personal Access Token)을 등록하여 접근
- GPU 노드 로드밸런싱을위해 MDS 데이터세트를 노드수로 균등하게 나눔
- Databricks Unity Catalog Volume 저장소에 저장되어있는 MDS파일을 MosaicML의 GPU 클러스터로 스트리밍
- 스트리밍된 MDS파일을 사용하여 모델학습



< Databricks Volume에서 MDS 데이터를 로드하여 모델을 학습하는 구조 >

©2024 Databricks Inc. — All rights reserved

MosaicML Composer

개발 효율성 향상

- GenAI 모델개발에 있어 GPU의 분산처리, 병렬처리 복잡성을 자동으로 해결. composer 라이브러리를 추가하고 코드 몇 줄만 추가하여 이러한 복잡한 메서드를 간단하게 구현가능.

```
from composer import Trainer, algorithms

trainer = Trainer(
    model=your_model,
    algorithms=[
      algorithms.BlurPool(),
      algorithms.LabelSmoothing(alpha=0.1),
      algorithms.ChannelsLast()
    ]
)
```

모델 학습을 위한 All-in-One 패키지

- Composer에는 표준 모델 및 데이터 세트, HuggingFace, Weights & Biases, DeepSpeed 및 TIMM과의 통합, 산업용 모델을 생성하기위한 강력한 분산 모델학습, 스트리밍 데이터 로더 등이 함께 제공.

```
import composer.functional as cf
import torchvision

models = torchvision.models.resnet18()
cf.apply_blurpool(model)
cf.apply_squeeze_excite(model, min_channels=16)

for x, y in train_loader:
  x, y = cf.cutmix(x, y, alpha=1, num_classes=10)
  output = model(x)
  # <-- your training loop here -->
```

코드수준으로 도구 제공

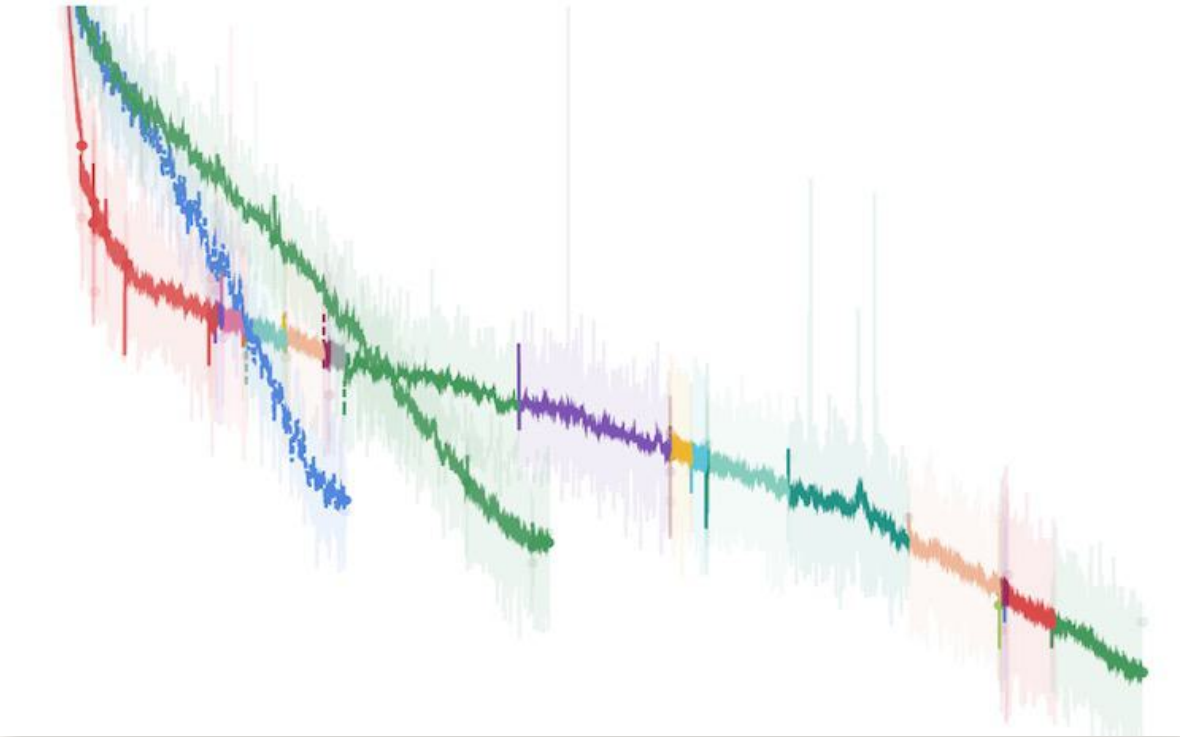
- 기존 모델학습 방식과 동일한 경험을 제공하기위해 코드 수준의 기능적 인터페이스를 제공하여 보다 효율적인 GenAI 구현이 가능. 커널을 지속적으로 출시함에 따라 코드베이스에서 Composer를 사용하여 최신 상태를 유지.

```
from composer import Trainer, models

trainer = Trainer(
  model=models.TIMM(
    model_name="resnet50",
    drop_rate=0.5,
    pretrained=True,
  ),
  algorithms=[...]
)
```

©2024 Databricks Inc. — All rights reserved

MosaicML 하드웨어 장애 대응

GPU는 1,000 GPU일당 1건의 하드웨어 장애 발생



하드웨어 장애 문제

- 512개의 H100을 사용하면 연간 예상되는 GPU 장애 건수는 186건
- 이를 주간으로 환산할 경우 주당 3.56회
- 대부분의 GPU 장애는 5~10시간 동안 전체 실행을 중단함

MosaicML의 솔루션

- MosaicML의 소프트웨어 기술은 작업을 실시간으로 모니터링하여 장애를 감지
- 실패가 발생하면 실행이 일시 중지되었다가 체크포인트에서 자동으로 재개

* MosaicML을 통해 512개의 GPU에서 학습한 결과. 각 색상은 GPU 오류로 인한 재시작을 나타냄

©2024 Databricks Inc. — All rights reserved

미세조정(Fine-tuning)을 위한 Databricks의 기능

©2024 Databricks Inc. — All rights reserved

Continued Pre-training이란?

Pre-training과 Continued Pre-training의 차이

- Pre-training은 LLM의 기반 모델을 처음부터 생성하는 작업으로 대량이 데이터세트와 자원이 소요됨
- Continued Pre-training은 Pre-training 모델의 가중치를 변경하여 기반 모델을 생성하는 작업으로 비교적 소량의 데이터 세트와 자원이 소요됨
- Databricks는 Pre-training과 Continued Pre-training을 손쉽게 진행할 수 있는 API를 제공

Pre-training과 Continued Pre-training의 비교

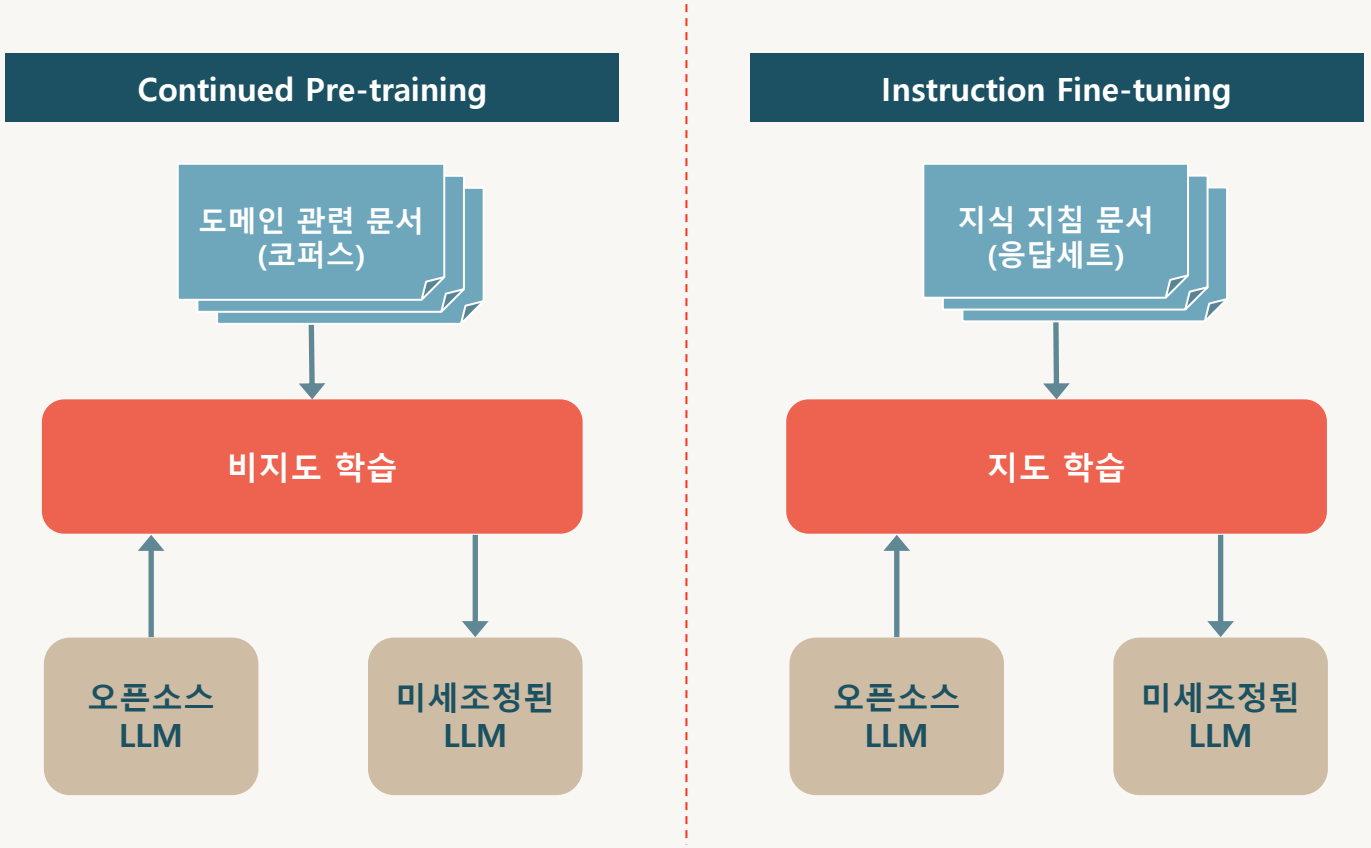
구 분	Pre-training	Continued Pre-training
작업 구분	사전학습	파인튜닝
학습 시작 가중치	랜덤 가중치	기존 Pre-training 모델의 가중치
학습시간	몇달	몇시간 ~ 며칠
자원 사용량	수십 ~ 수백 GPU 자원 사용	1대 ~ 수십 GPU 자원 사용
학습 비용	매우 높음	보통
학습 토큰	1조 ~ 10조 토큰	1백만 ~ 10억 토큰
결과물	기반 모델	도메인 특화된 튜닝 모델

©2024 Databricks Inc. — All rights reserved

Fine-tuning의 종류

CPT와 IFT의 비교

- CPT(Continued Pre-training)은 레이블이 없는 데이터로 기존 가중치를 조정하는 비지도 학습 방법
- IFT(Instruction Fine-tuning)은 레이블이 있는 데이터로 모델의 가중치를 조정하는 지도 학습 방법
- 도메인 특화된 LLM을 만들기위해 PT -> CPT -> IFT 순으로 훈련을 진행
- CPT나 IFT는 PT에 비하여 비용이 적게 들기 때문에 베이스 모델 품질이 적합하다면 적극 활용이 필요
- 하지만, PT -> CPT -> IFT 순으로 진행한 작업결과에도 LLM 품질이나 성능이 충족되지 못하는 경우 Pre-training을 다시 검토해 보는 것이 필요



©2024 Databricks Inc. — All rights reserved

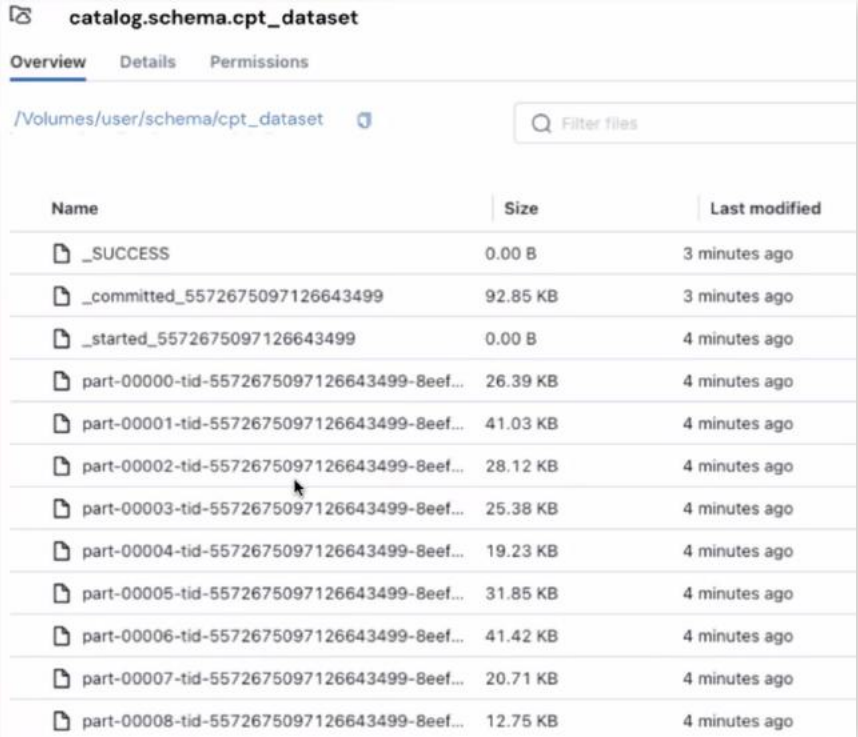
Fine-tuning 고려사항

Continued Pre-training의 고려사항

- 데이터 파일은 Databricks Unity Catalog Volume내에 위치해야 함
- 오직 txt 파일만 학습파일로 사용가능
- 학습파일은 노이즈가 없는 고품질의 데이터여야 함
- 중복과 데이터 정제작업이 선행 되어야 함
- Instruction Fine-tuning과의 관련성을 고려해야 함

Instruction Fine-tuning의 고려사항

- 데이터 파일은 Databricks Unity Catalog Volume내에 위치해야 함
- 고품질의 명령/응답 데이터셋을 준비해야 함
- 명령/응답 데이터셋의 양이 확보되어야 함(최소 만건 단위)
- 명령/응답 데이터셋의 문장의 길이는 다양하게 존재해야 함
- 범용 용도의 LLM이 목표인지, 도메인 특화 LLM이 특화인지를 결정해야 함



< Unity Catalog Volume내 위치한 학습 파일들 >

©2024 Databricks Inc. — All rights reserved

API 기반의 Fine-tuning

파인튜닝 API Python

파일 편집 보기 실행 도움말 마지막 편집: 지금 New cell UI: 설정

모두 실행 Shared 15.x 스케줄 공유

2024. 4. 19. (4초) 6: EngineerCodeSnippetName

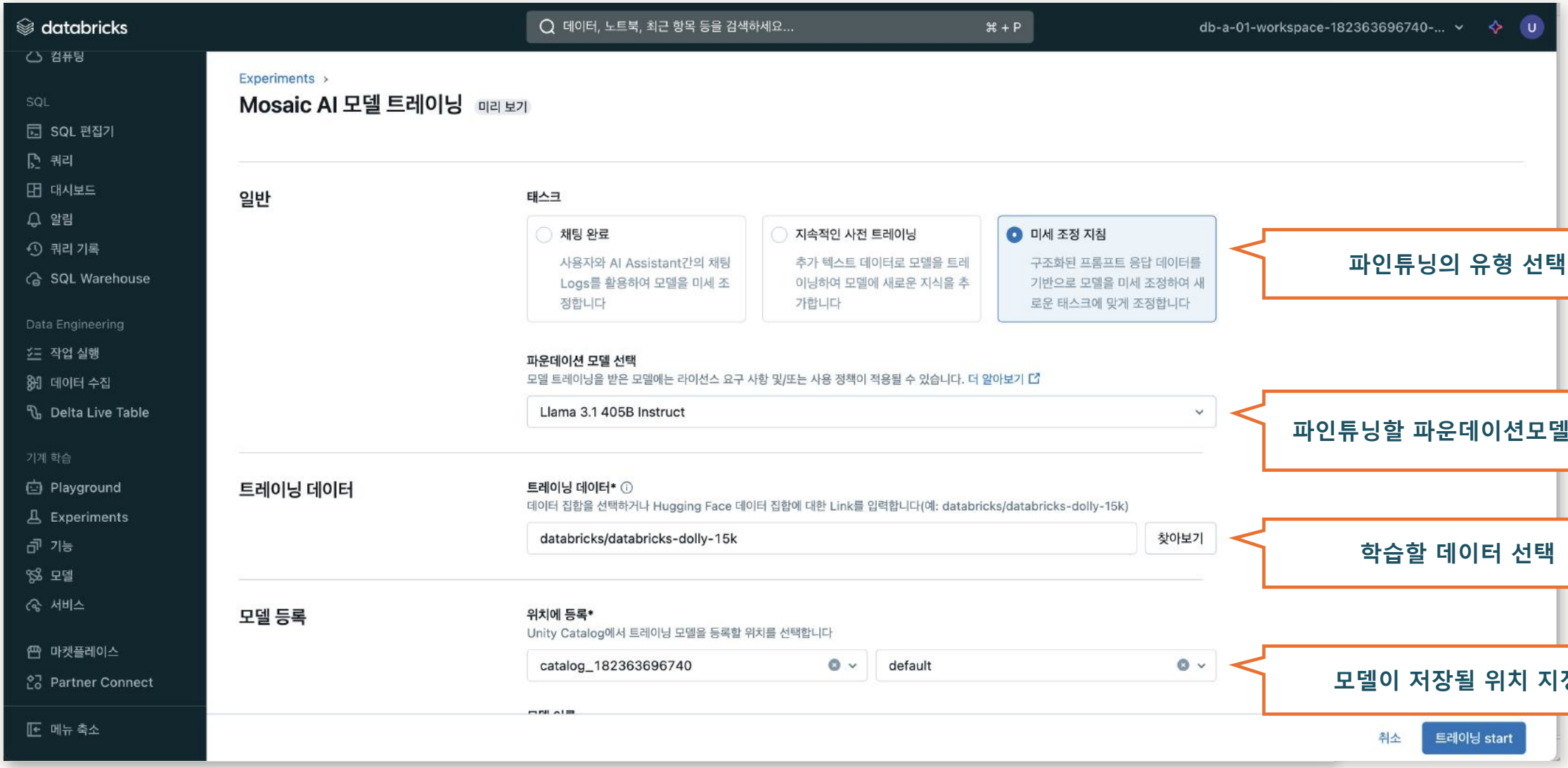
```python
model = 'mosaicml/mpt-7b'
train_data_path = '/Volumes/main/all-hands-ft-demo-feb/data/ns-t2s.jsonl'
register_to = 'main.nancyhung'
training_duration = '2ba'
learning_rate = '5e-7'

for task_type in ['INSTRUCTION_FINETUNE', 'CHAT_COMPLETION']:
  run = ft.create(
    model=model,
    train_data_path=train_data_path,
    task_type=task_type,
    register_to=register_to,
    training_duration=training_duration,
    learning_rate=learning_rate,
  )
  print(task_type, run)
```

Python SDK를 통해 Fine-Tuning API를 호출하여 학습

```
d_at=datetime.datetime(2024, 4, 18, 18, 27, 57, 657000, tzinfo=datetime.timezone.utc), started_at=None, estimated_end_t
ime=None, formatted_eta=None, completed_at=None, submitted_config=FinetuneConfig(model='mosaicml/mpt-7b', train_data_pa
th='dbfs:/Volumes/main/all-hands-ft-demo-feb/data/ns-t2s.jsonl', save_folder='dbfs:/databricks/mlflow-tracking/{mlflow_
experiment_id}/{mlflow_run_id}/artifacts', task_type='INSTRUCTION_FINETUNE', eval_data_path=None, eval_prompts=[], cust
om_weights_path=None, training_duration='2ba', experiment_tracker={'mlflow': {'experiment_path': '/Users/nancy.hung@dat
abricks.com/ift-mpt-7b-q6t1st', 'tracking_uri': 'databricks', 'model_registry_path': 'main.nancyhung'}}, learning_rate=
5e-07, context_length=None, data_prep_config=None, disable_credentials_check=None), events=[], _required_properties=('i
d', 'name', 'status', 'createdByEmail', 'createdAt', 'updatedAt'))
```

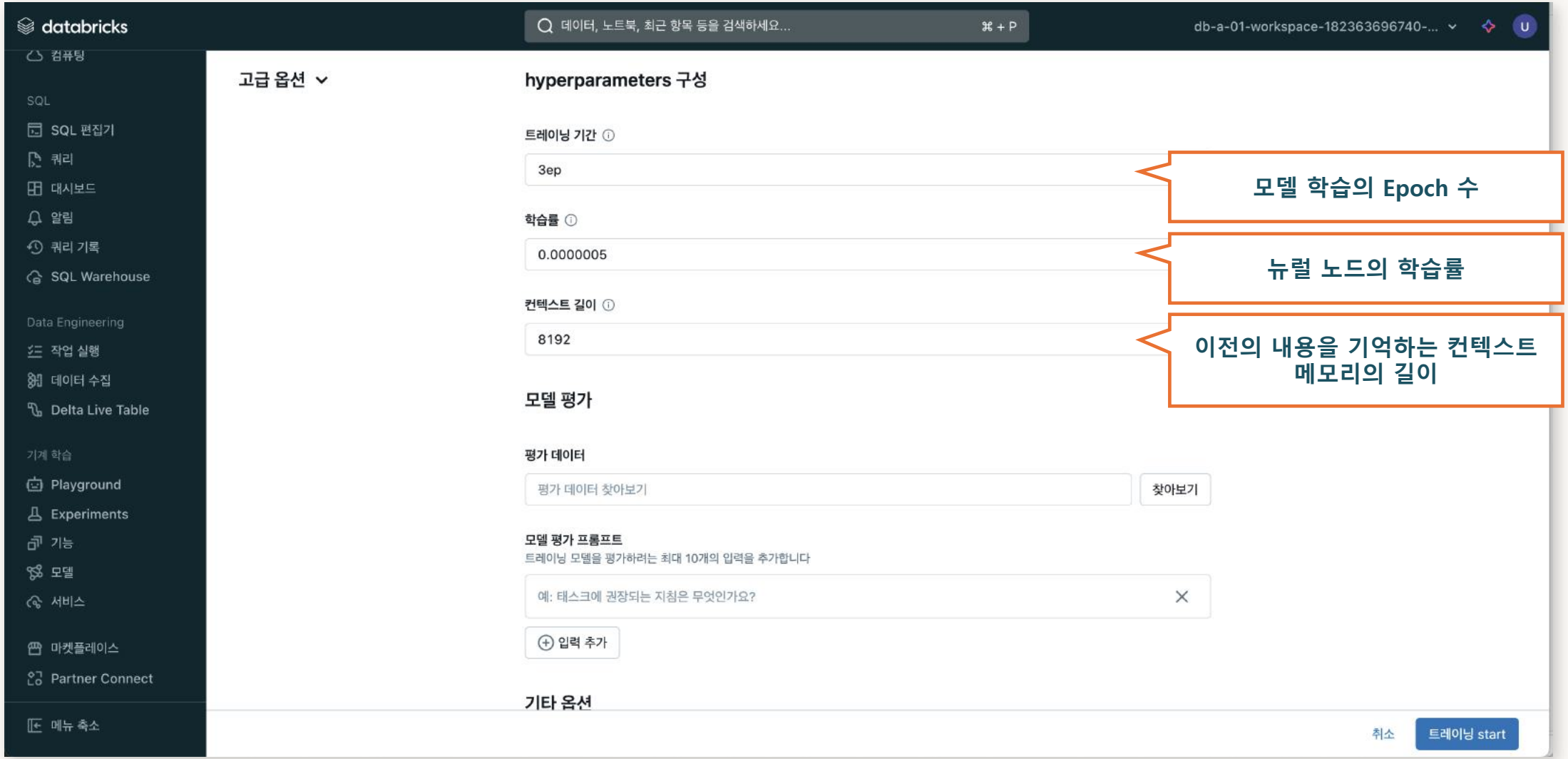
©2024 Databricks Inc. — All rights reserved

GUI 기반의 Fine-tuning



©2024 Databricks Inc. — All rights reserved

GUI 기반의 Fine-tuning



모델 학습의 Epoch 수

뉴럴 노드의 학습률

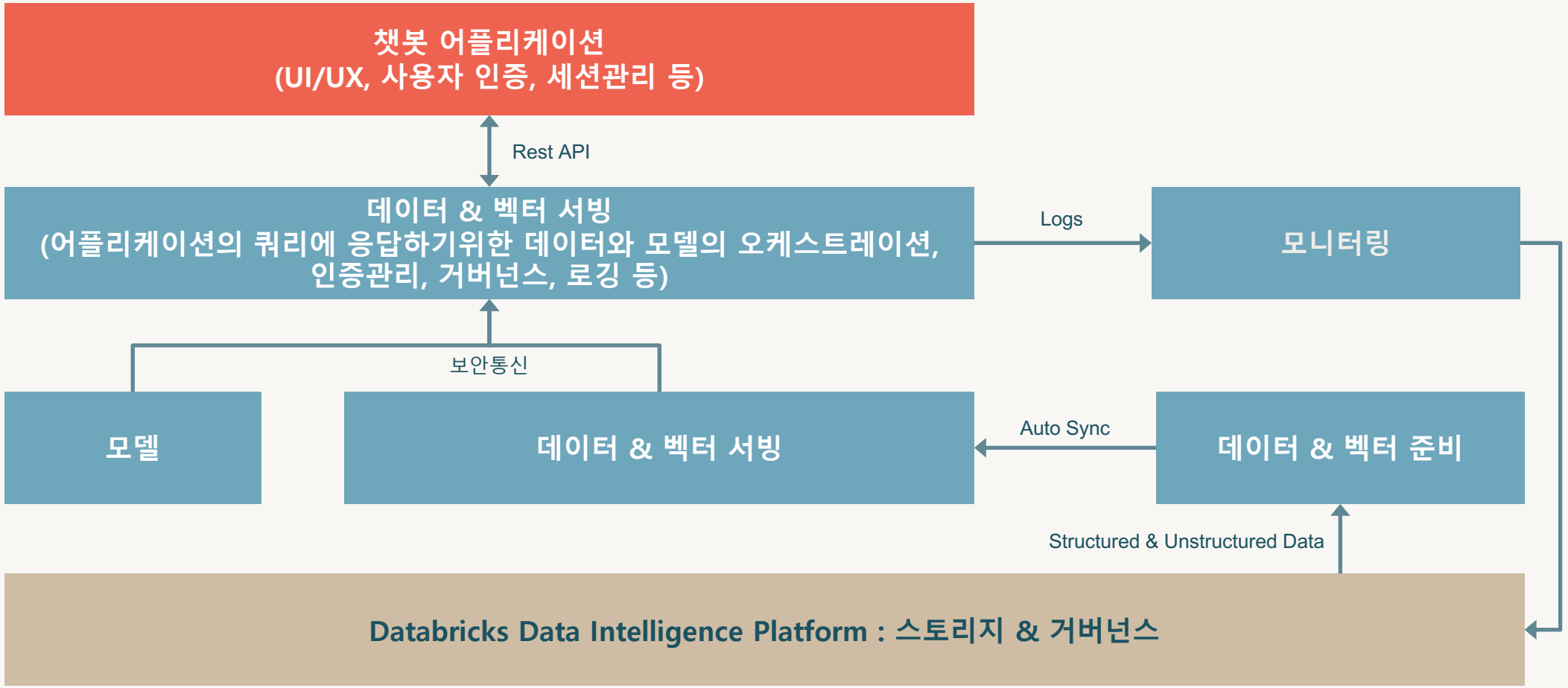
이전의 내용을 기억하는 컨텍스트 메모리의 길이

©2024 Databricks Inc. — All rights reserved

검색 증강 생성 (Retrieval-Augmented Generation) Databricks 기능

©2024 Databricks Inc. — All rights reserved

Databricks 생성형 AI의 구현구조

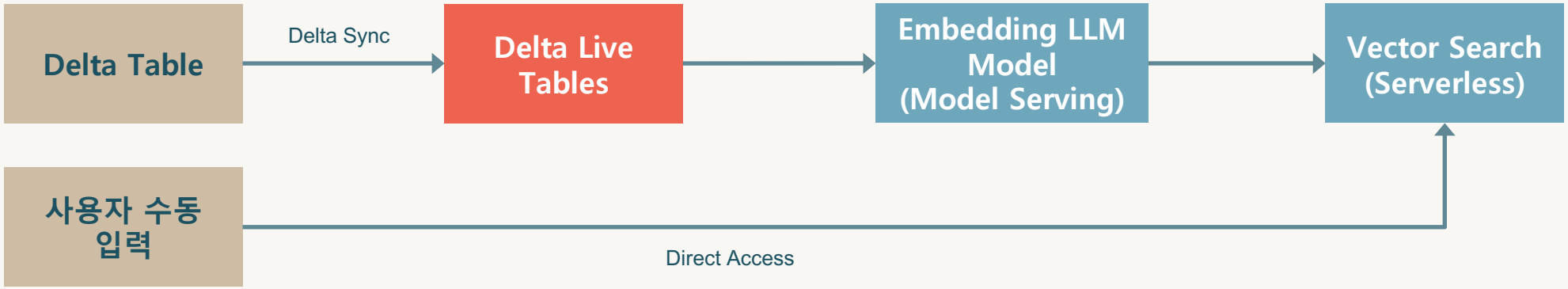


©2024 Databricks Inc. — All rights reserved

Databricks Vector Search

Databricks Vector Search

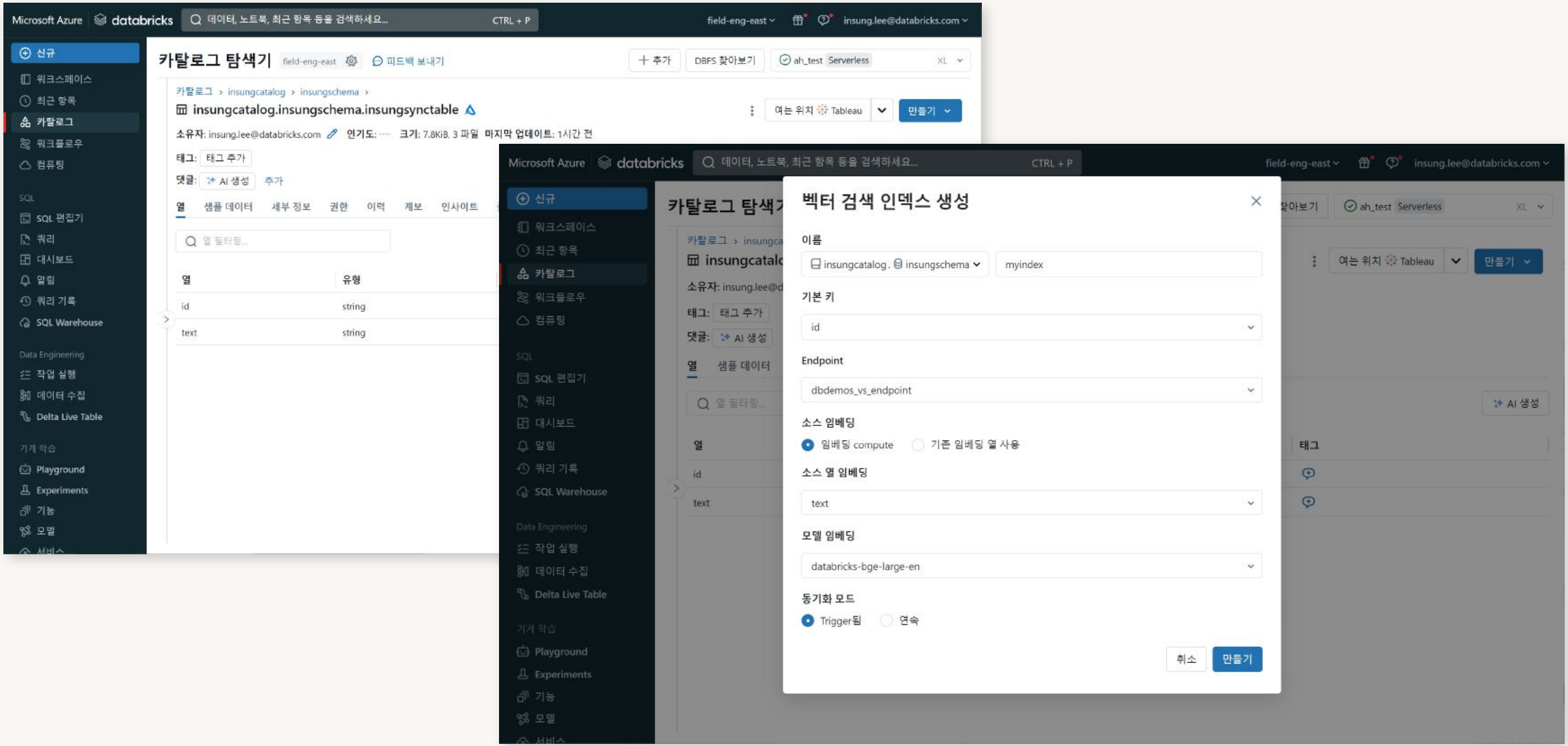
- Databricks Vector Search란, Databricks Data Intelligence Platform에서 제공하는 고속 Vector DB 서비스
- 일반적인 Vector DB들이 Data를 입력할때 SDK/API만을 사용해야해서 데이터의 효과적인 관리가 어려웠으며, 데이터 조회시에도 입력데이터가 이미 벡터화되어서 값들의 관리가 어려웠음
- Databricks Vector Search는 **Delta Table로 데이터를 입력한 뒤 데이터를 Vector Store에 Sync**하는 방식으로 테이블만 관리해주면 됨
- 데이터를 임베딩할때는 Local 또는 External의 선호하는 LLM 임베딩 모델을 사용하여 텍스트 임베딩



< Databricks Vector Search의 임베딩 방식 >

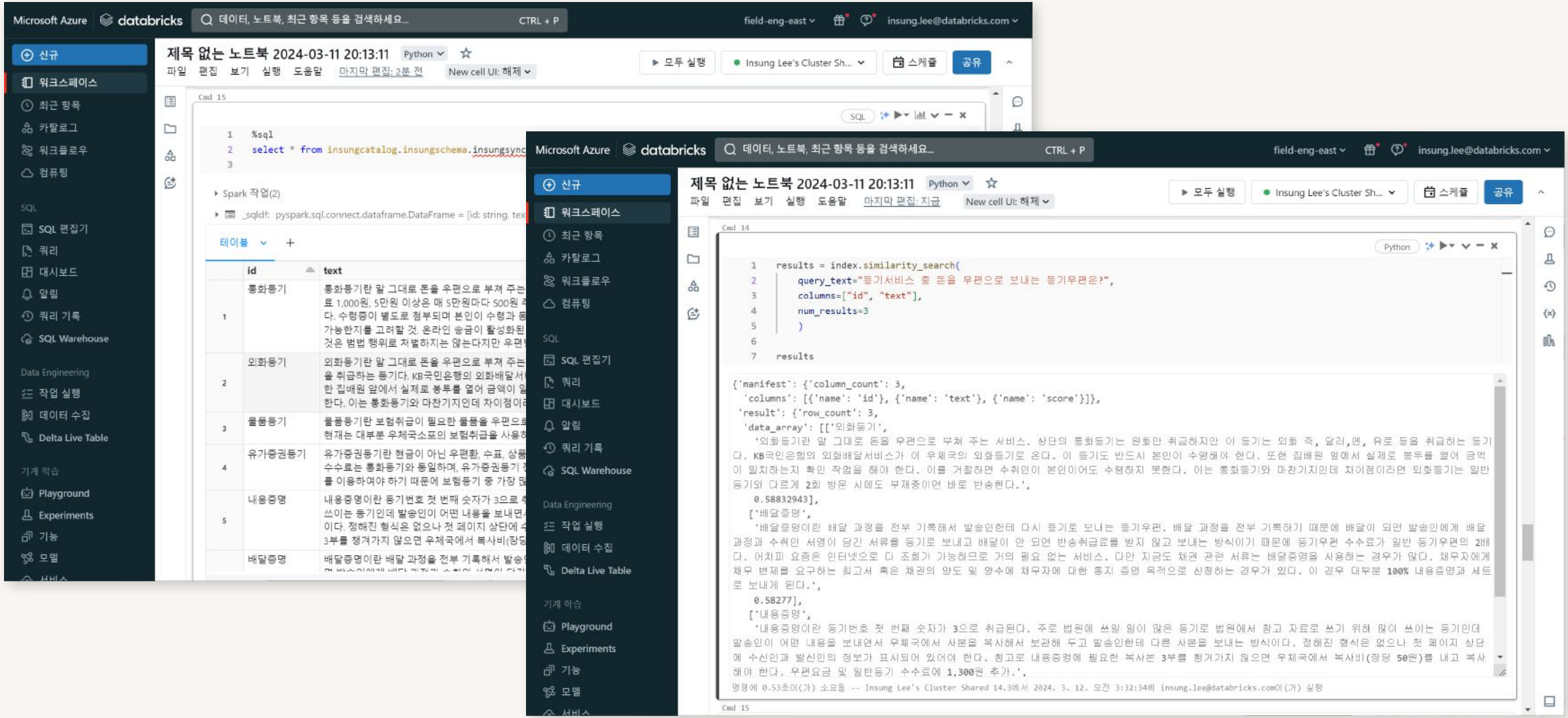
©2024 Databricks Inc. — All rights reserved

Databricks Vector Search



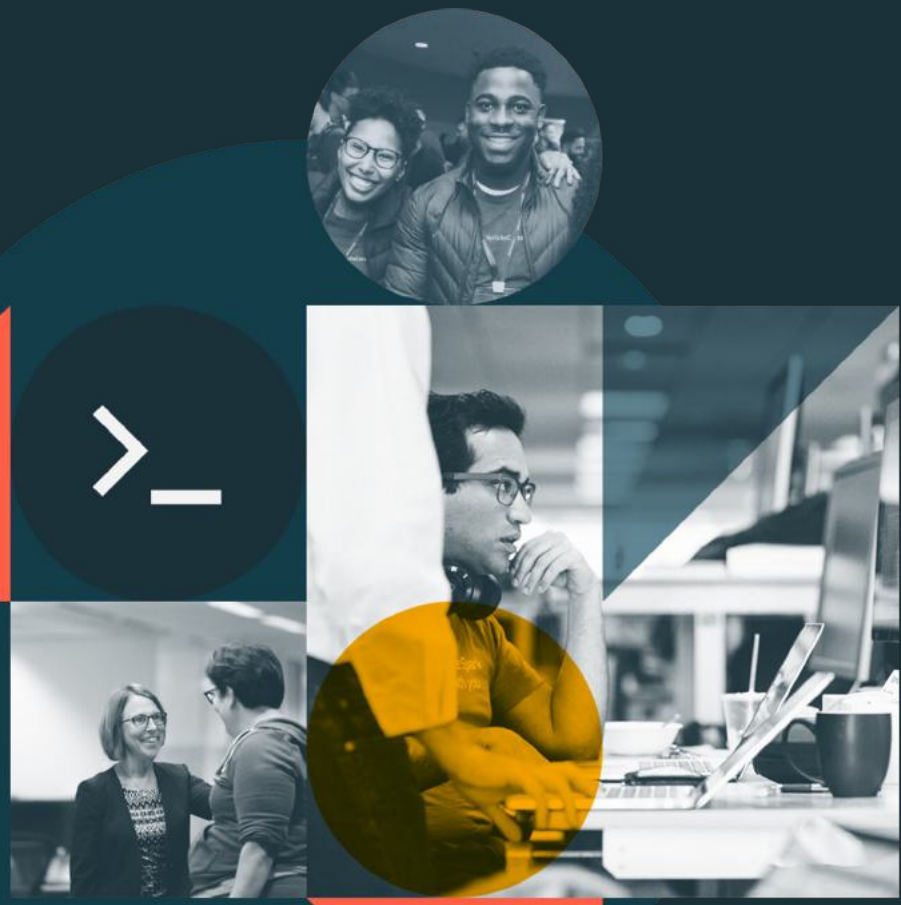
©2024 Databricks Inc. — All rights reserved

Databricks Vector Search



©2024 Databricks Inc. — All rights reserved

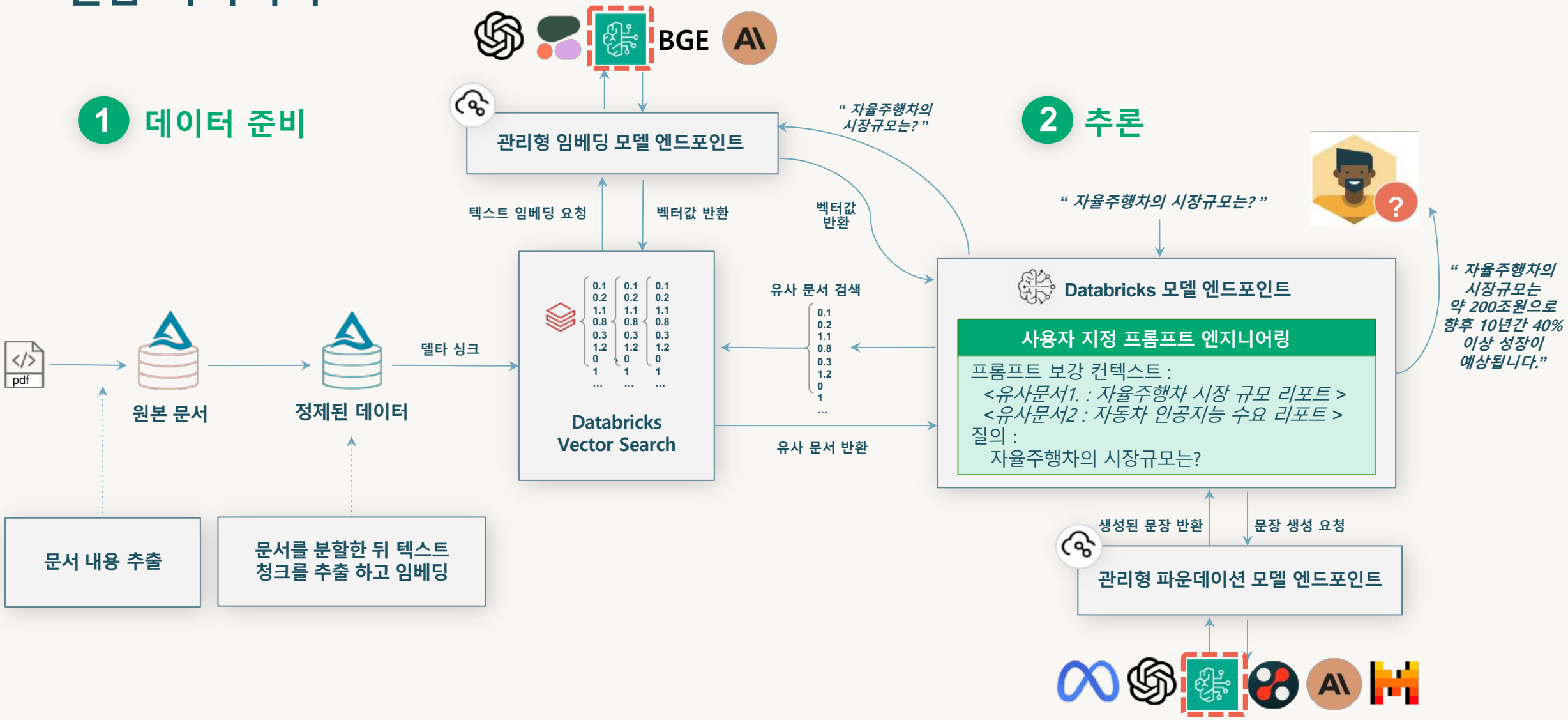
Databricks & AWS Zero to GenAI 워크샵
(RAG 실습)



Databricks Korea
Ph.D. Insung Lee
2024.08.07

©2024 Databricks Inc. — All rights reserved

실습 아키텍처



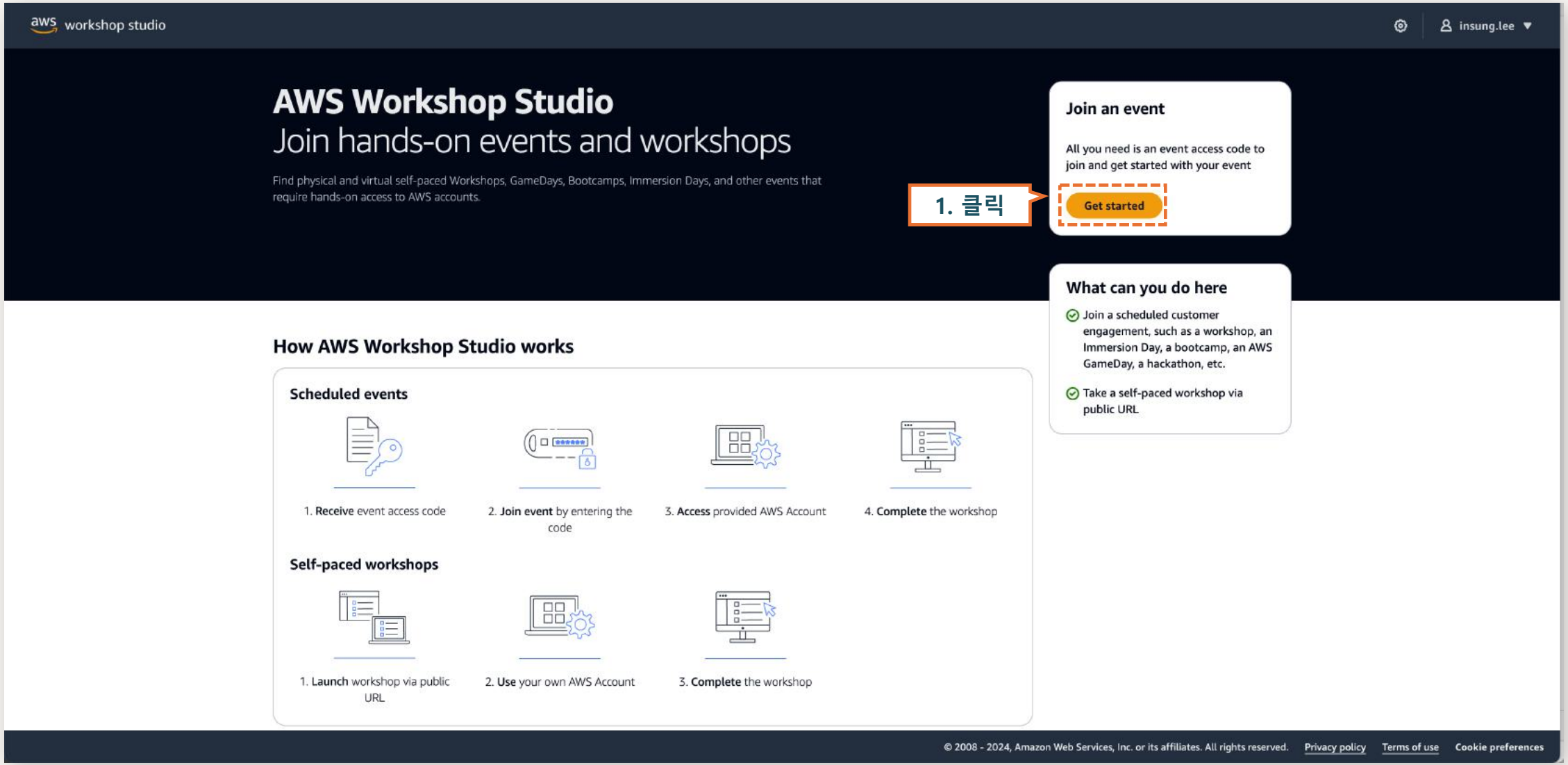
©2024 Databricks Inc. — All rights reserved

실 습 1 :

AWS Workshop Studio에서 Databricks Workspace 접속

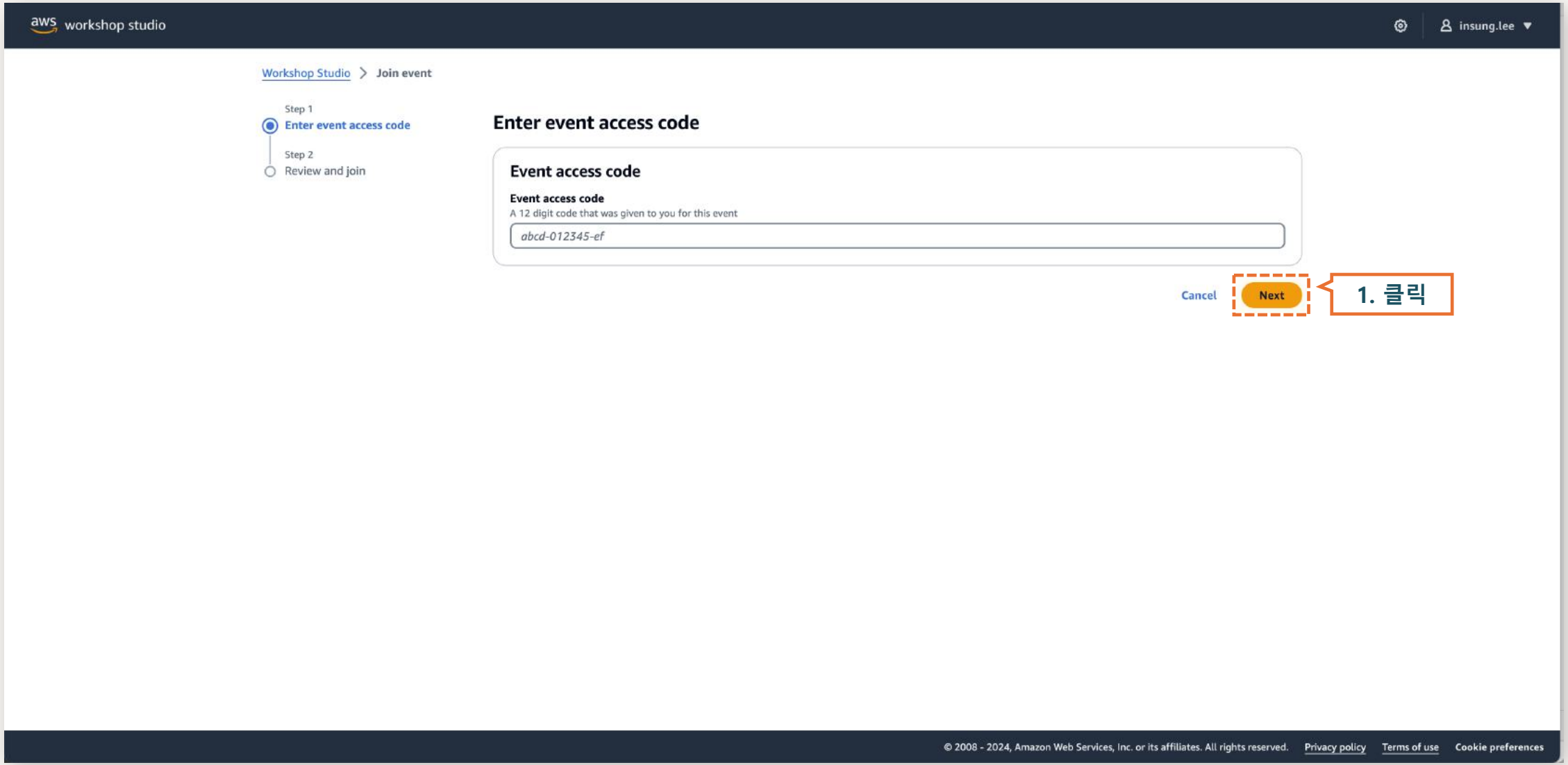
©2024 Databricks Inc. — All rights reserved

AWS Workshop Studio 접속



©2024 Databricks Inc. — All rights reserved

AWS Workshop Studio 접속



©2024 Databricks Inc. — All rights reserved

AWS Workshop Studio 접속

aws workshop studio

insung.lee

Workshop Studio > Join event

Step 1
Enter event access code

Step 2
Review and join

Review and join

Event details

Name	Start time	Duration	Level
Databricks GenAI workshop	8/05/2024 04:02 PM	48 hours	200

Description
Databricks GenAI workshop

Terms and Conditions

Read and accept before joining the event

Read and accept before joining the event:

1. By using AWS Workshop Studio for the relevant event, You agree to the AWS Event Terms and Conditions, the AWS Responsible AI Policy, and the AWS Acceptable Use Policy.
2. If You are under 18 years old, you may participate in the relevant event using AWS Workshop Studio: (a) if You are at least the minimum age below based on the country or region in which You reside, and (b) with the involvement of a parent, guardian, or educator.

Country or region	Minimum age
All countries or regions not listed below (including the United States, Brazil, the United Kingdom, and India)	13
Canada, China, Republic of Korea (South Korea)	14
Australia	15
Japan, Switzerland, Vietnam, and countries in the European Economic Area	16
Azerbaijan, Bolivia, Colombia, Dominican Republic, Indonesia, Lebanon, Malaysia, Mexico, Montenegro, Nepal, Philippines, Thailand, Turkey, and countries in Africa	18

1. 아래로 스크롤

© 2008 - 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy policy Terms of use Cookie preferences

AWS Workshop Studio 접속

aws workshop studio

insung.lee

Australia	15
Japan, Switzerland, Vietnam, and countries in the European Economic Area	16
Azerbaijan, Bolivia, Colombia, Dominican Republic, Indonesia, Lebanon, Malaysia, Mexico, Montenegro, Nepal, Philippines, Thailand, Turkey, and countries in Africa	18

3. You acknowledge and agree that You are using an AWS-owned account that You will only be able to access during the relevant event. You have no ownership rights over this AWS-owned account.
4. During the relevant event, while using this AWS-owned account, You will not use, import, input, or introduce any data, dataset, or other material that contains personal data, financial information, or any other data or materials that may be subject to laws and regulations (such as the General Data Protection Regulation or The Health Insurance Portability and Accountability Act of 1996).
5. If You find residual resources or materials in this AWS-owned account, You will notify your Event Operator immediately.
6. AWS, its affiliates, and any entities or persons acting on AWS's behalf reserves the right to terminate this AWS-owned account and to delete its contents at any time, without any notice to You.
7. During the relevant event, while using this AWS-owned account, You will not process or run any operation on any data other than test datasets or lab materials that have been approved by AWS.
8. You will not copy, import, export or otherwise create derivative works of materials provided by AWS for use outside of the relevant event.
9. AWS, its affiliates, and any entities or persons acting on AWS's behalf have no obligation to enable the transmission of Your materials through AWS Workshop Studio, and may, in their discretion, edit, block, refuse to post, or remove Your materials at any time, without notice to You.
10. If You access and use a service and/or third-party models that have their own terms during the relevant event, while in the AWS-owned account, You agree to review those terms and comply with them during the event.
11. If You are an AWS Partner using AWS Workshop Studio as part of Your participation in the AWS Partner Network Program, Your use of AWS Workshop Studio is governed by these terms, the AWS Partner Network Terms and Conditions, and the AWS Customer Agreement or other agreement with us governing your use of AWS Services.
12. Your use of AWS Workshop Studio will comply with these terms and all applicable laws. If You fail to comply with any of these terms, Your access to AWS Workshop Studio may be immediately terminated, without notice to You.

1. 체크

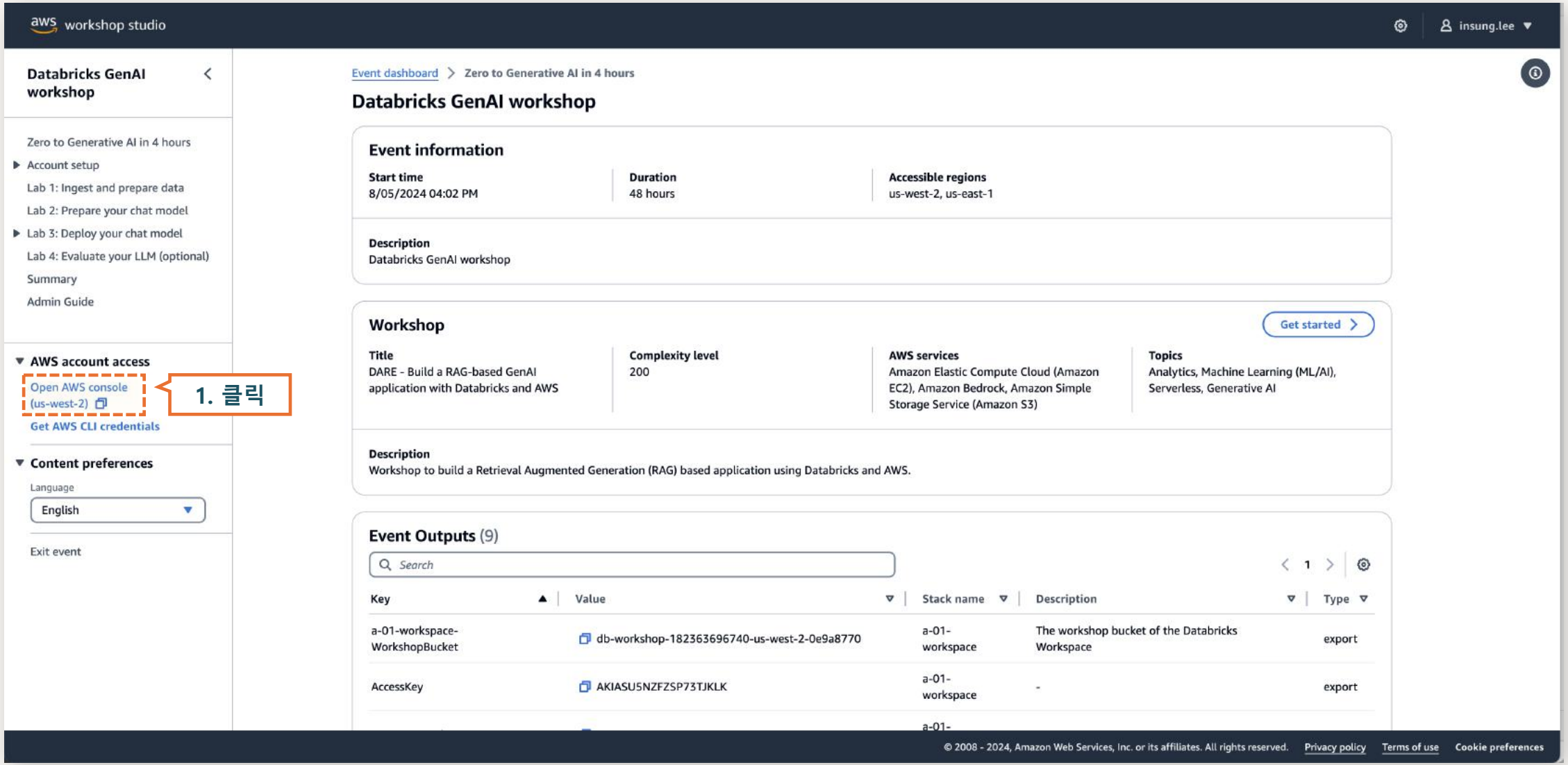
☑ I agree with the Terms and Conditions

Cancel | Previous | Join event

2. 클릭

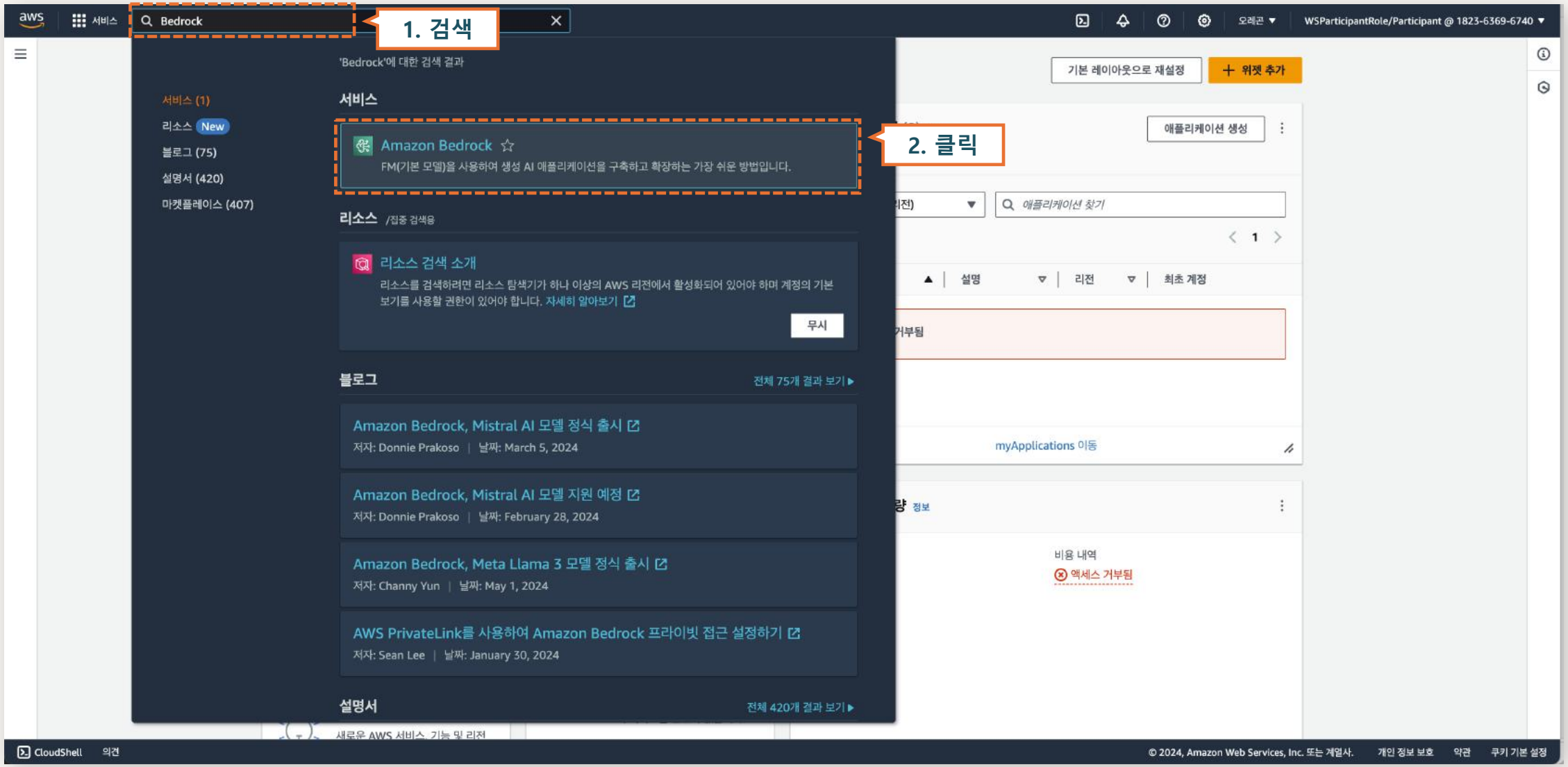
© 2008 - 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy policy | Terms of use | Cookie preferences

AWS console 접속



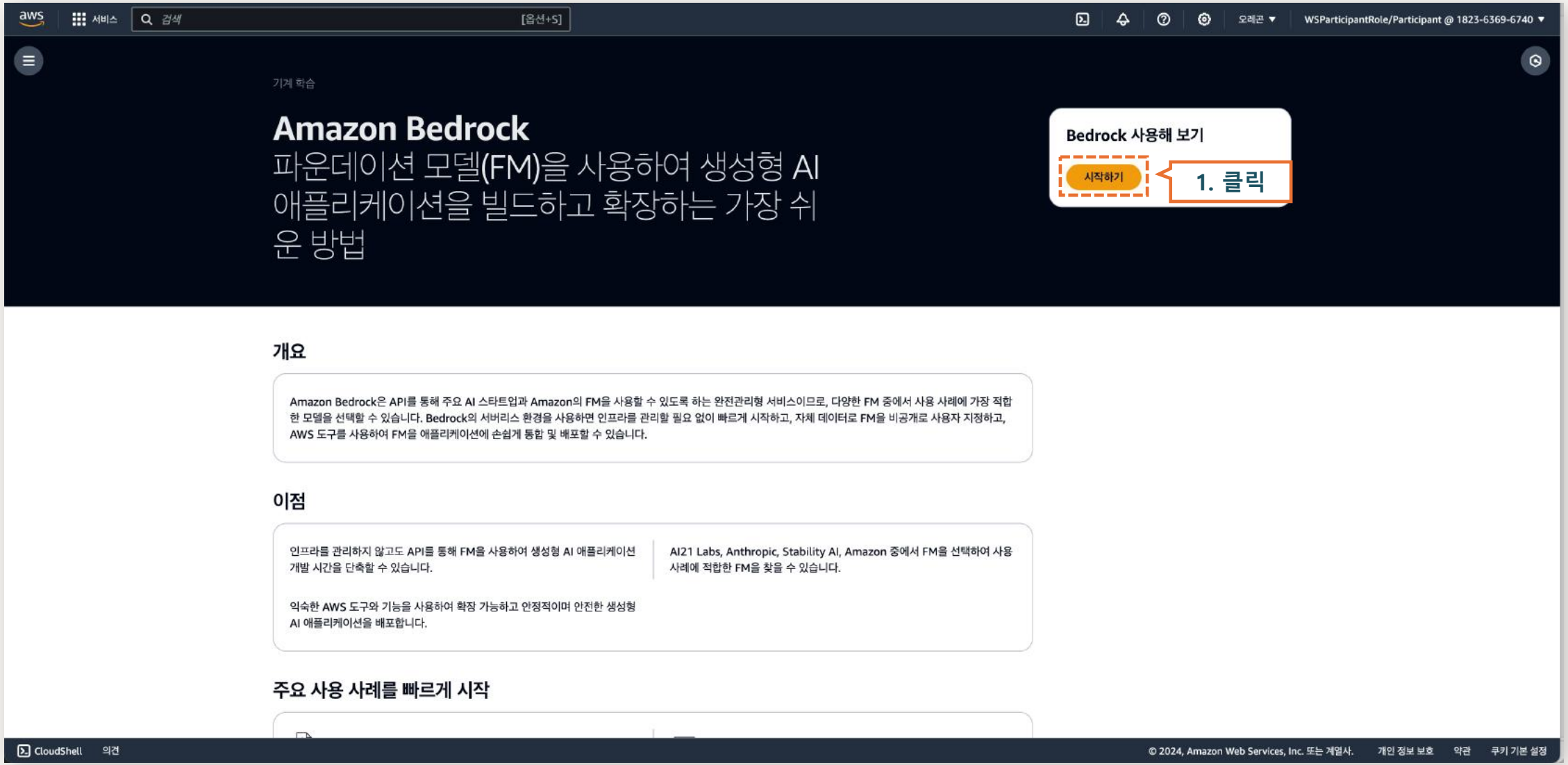
©2024 Databricks Inc. — All rights reserved

AWS console 접속



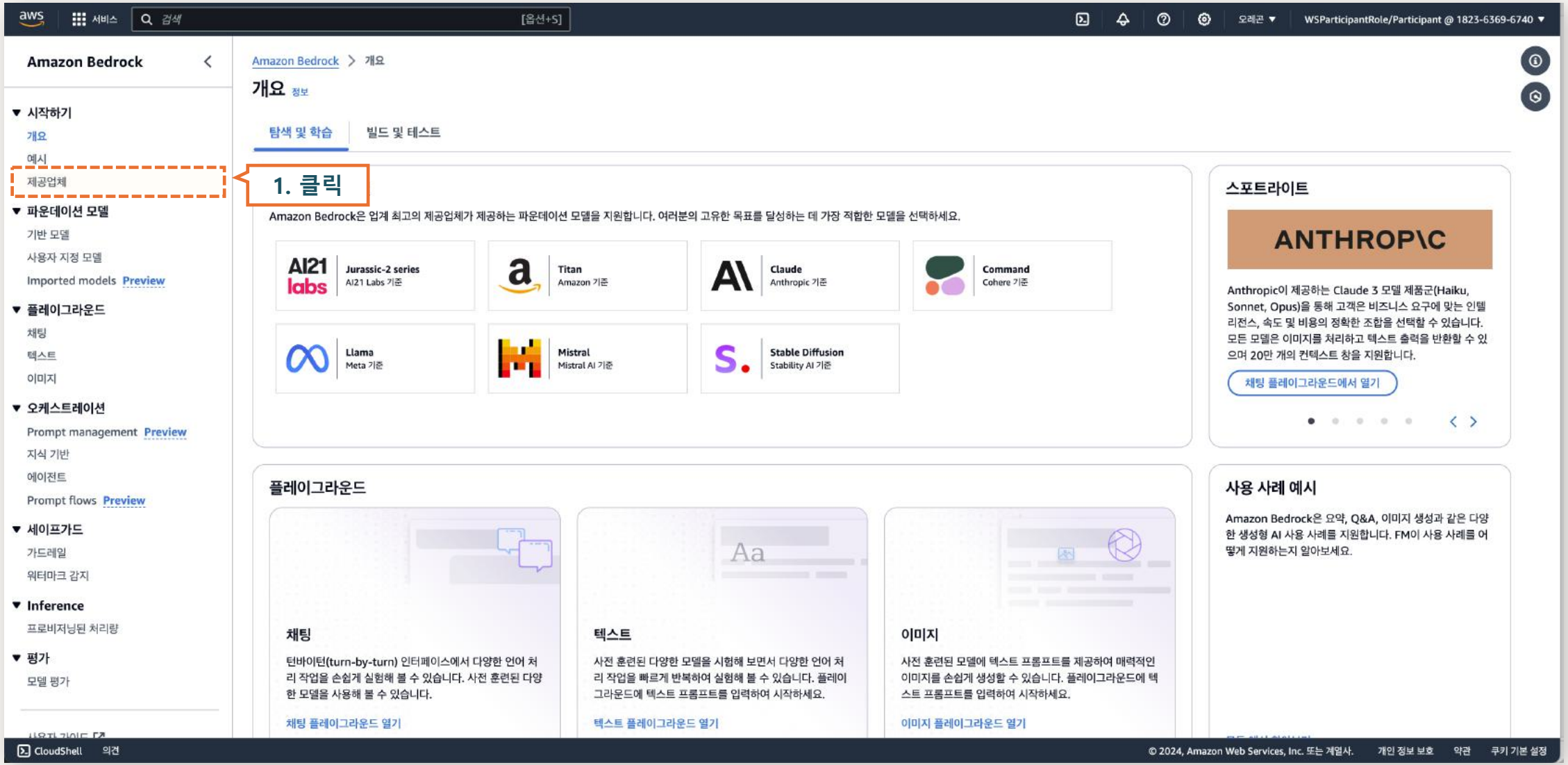
©2024 Databricks Inc. — All rights reserved

AWS console에서 Bedrock 모델 권한 부여



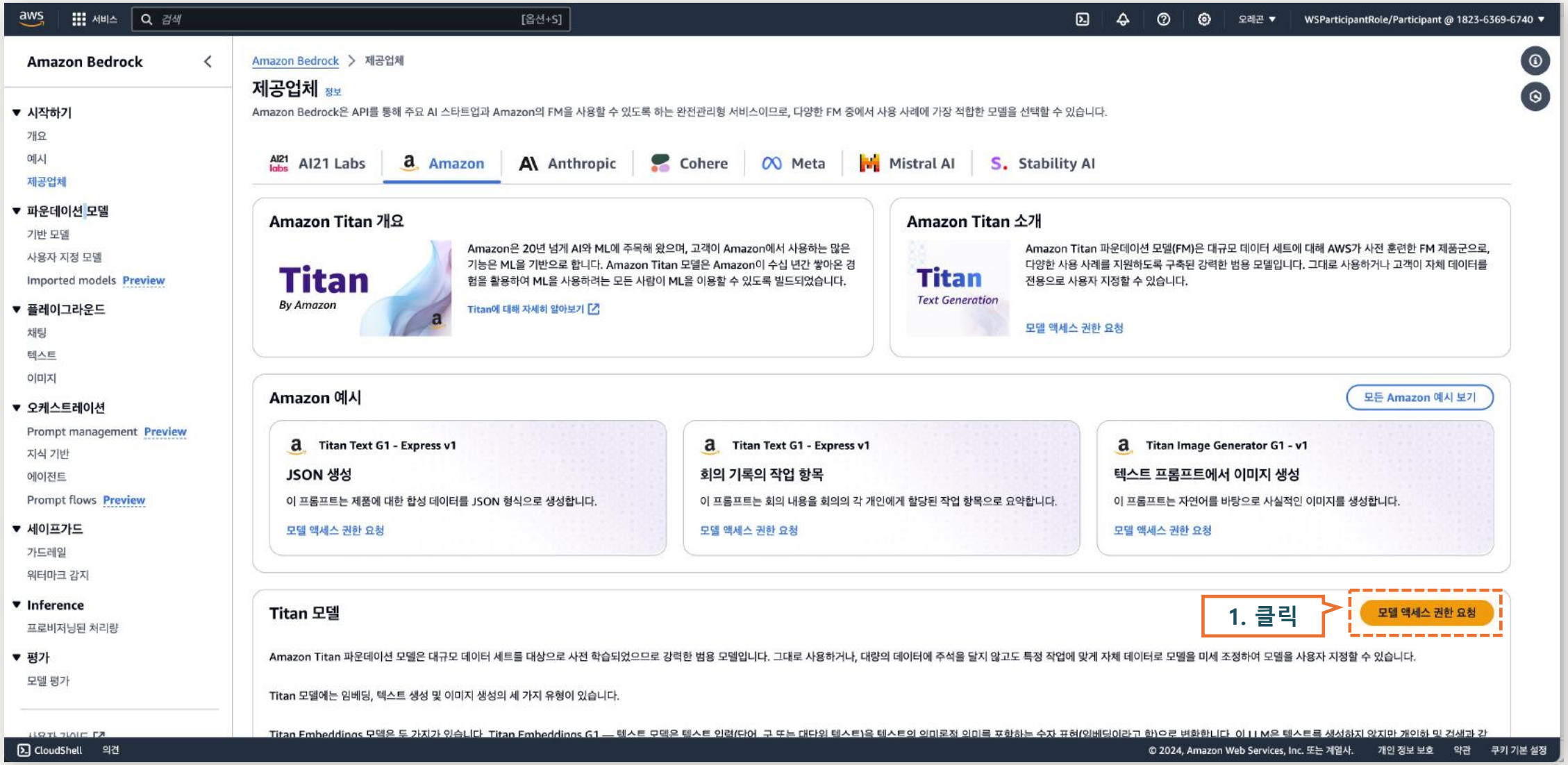
©2024 Databricks Inc. — All rights reserved

AWS console에서 Bedrock 모델 권한 부여



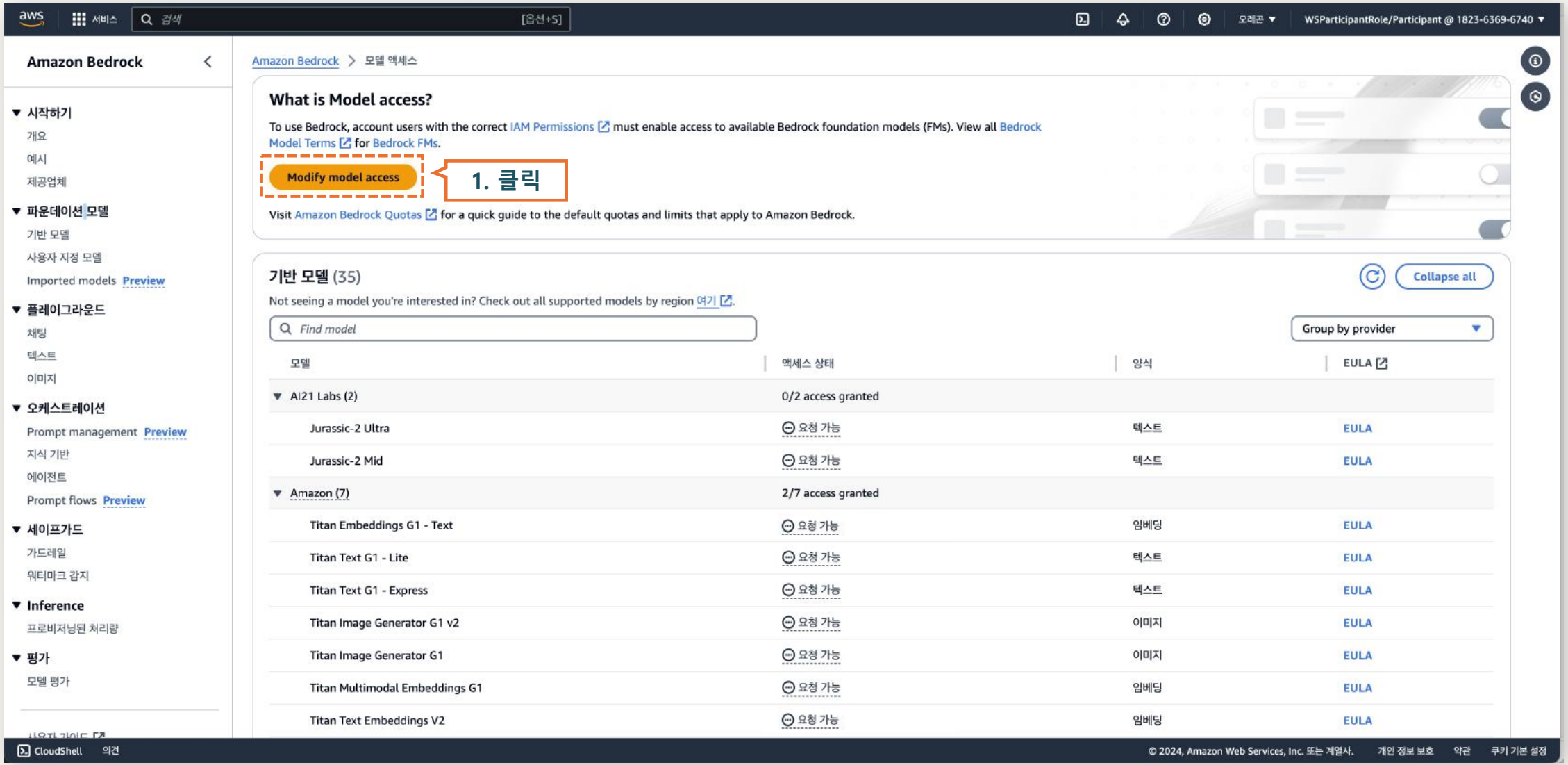
©2024 Databricks Inc. — All rights reserved

AWS console에서 Bedrock 모델 권한 부여



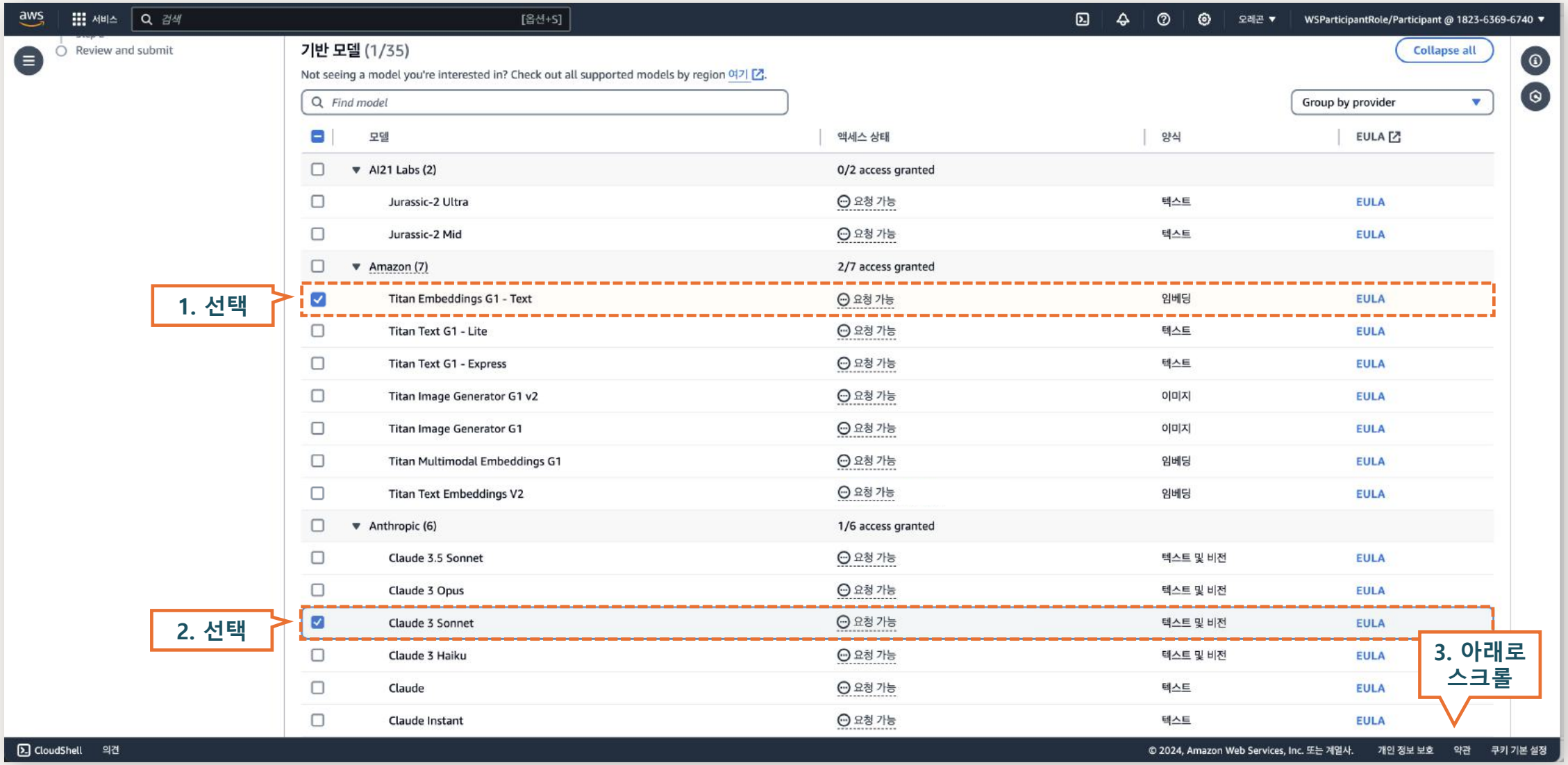
©2024 Databricks Inc. — All rights reserved

AWS console에서 Bedrock 모델 권한 부여



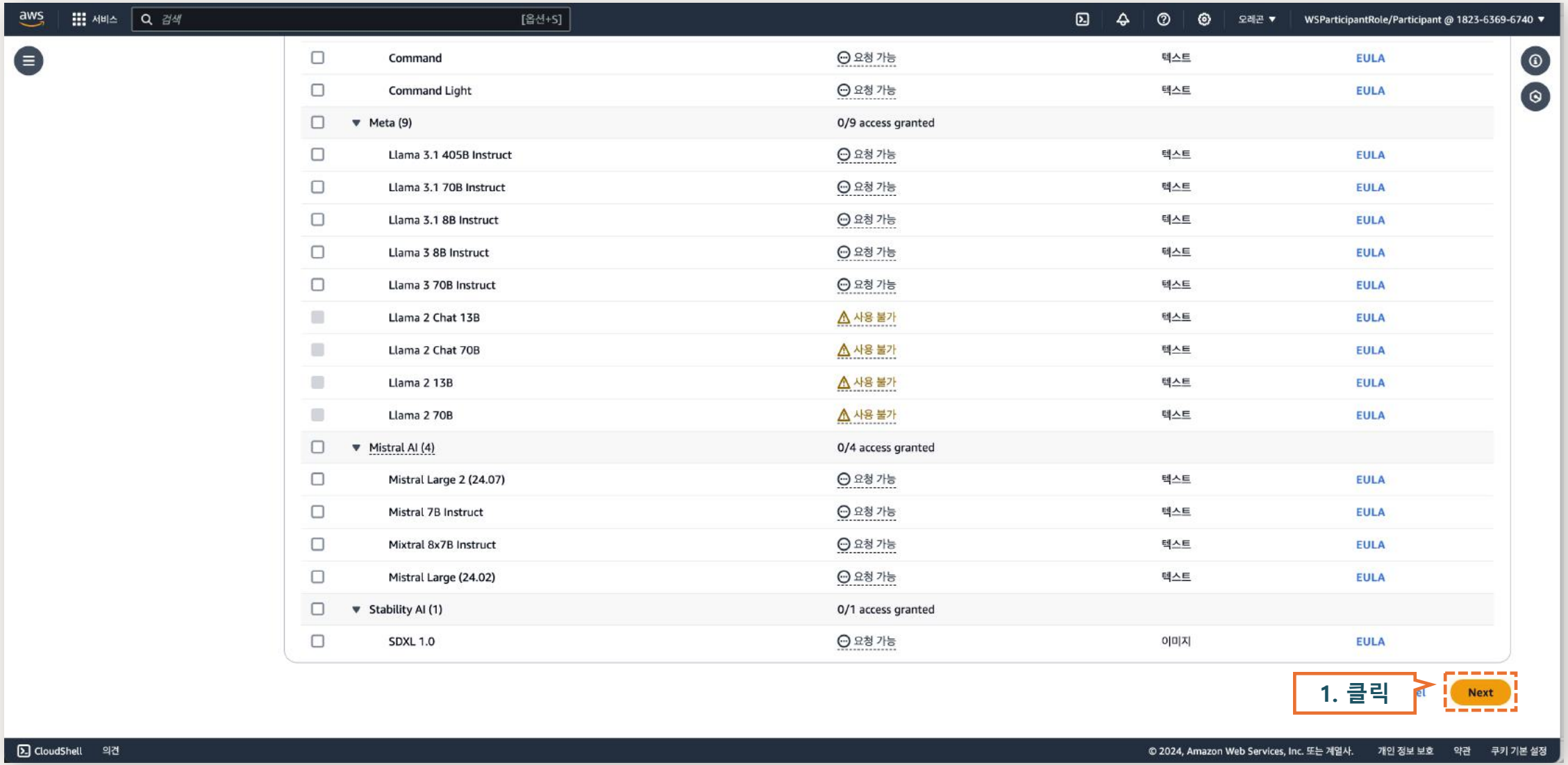
©2024 Databricks Inc. — All rights reserved

AWS console에서 Bedrock 모델 권한 부여



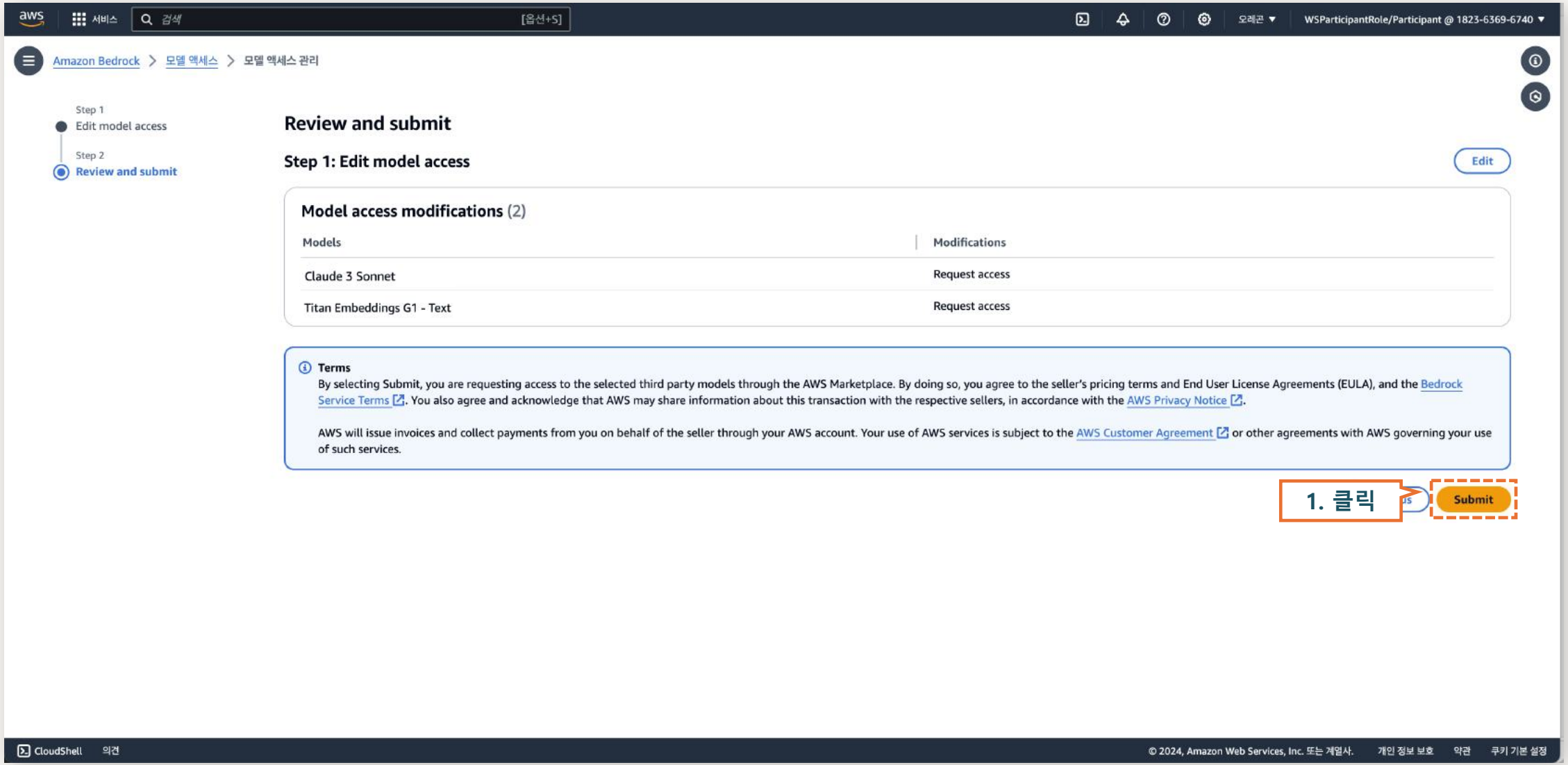
©2024 Databricks Inc. — All rights reserved

AWS console에서 Bedrock 모델 권한 부여



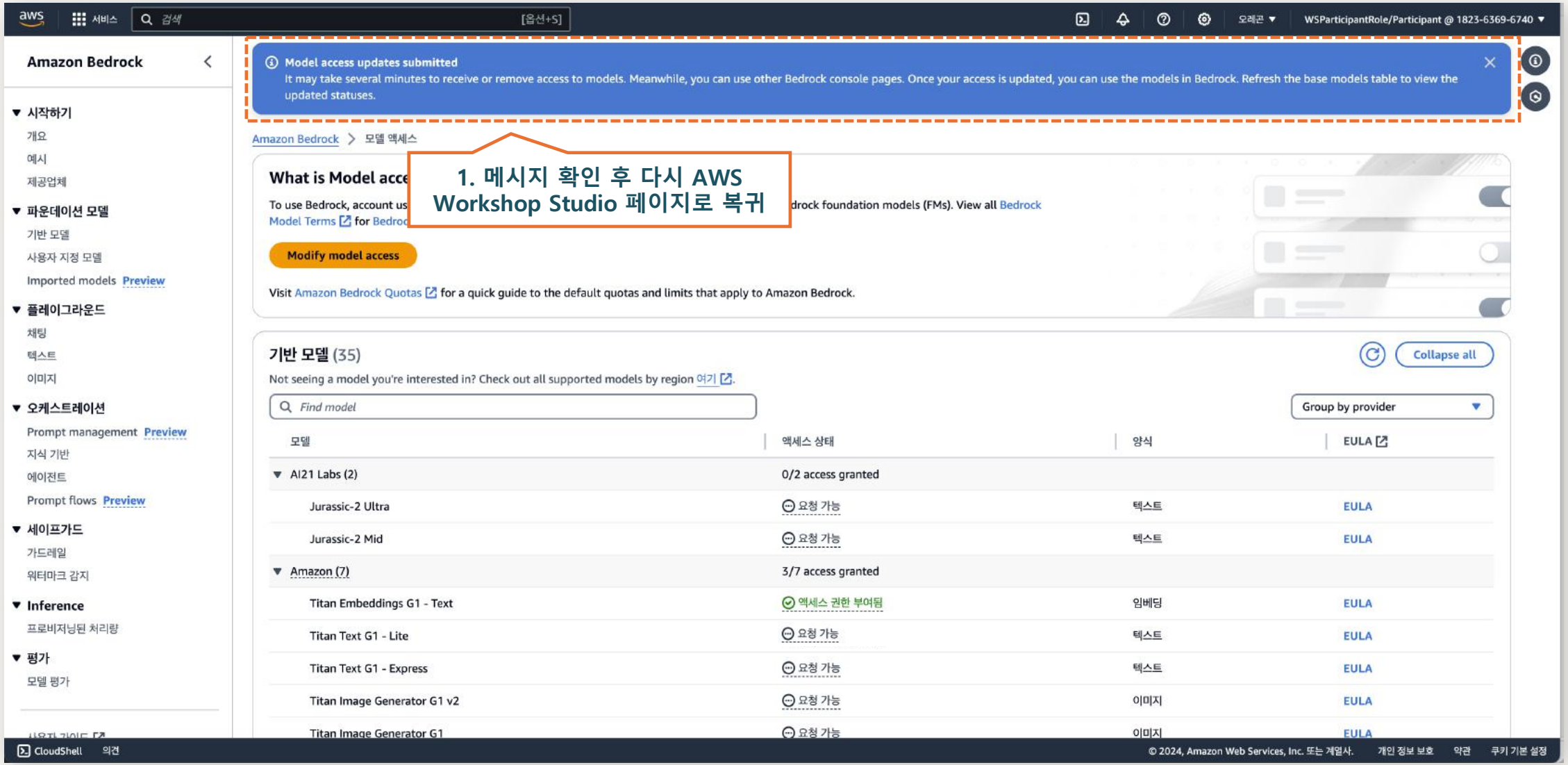
©2024 Databricks Inc. — All rights reserved

AWS console에서 Bedrock 모델 권한 부여



©2024 Databricks Inc. — All rights reserved

AWS console에서 Bedrock 모델 권한 부여



©2024 Databricks Inc. — All rights reserved

AWS Workshop Studio에서 Databricks Workspace 접속

aws workshop studio

insung.lee

Databricks GenAI workshop

- Zero to Generative AI in 4 hours
- Account setup
- Lab 1: Ingest and prepare data
- Lab 2: Prepare your chat model
- Lab 3: Deploy your chat model
- Lab 4: Evaluate your LLM (optional)
- Summary
- Admin Guide

AWS account access

Open AWS console (us-west-2)

Get AWS CLI credentials

Content preferences

Language

English

Exit event

Event dashboard > Zero to Generative AI in 4 hours

Databricks GenAI workshop

Event information

Start time	Duration	Accessible regions
8/05/2024 04:02 PM	48 hours	us-west-2, us-east-1

Description
Databricks GenAI workshop

Workshop

Get started

Title	Complexity level	AWS services	Topics
DARE - Build a RAG-based GenAI application with Databricks and AWS	200	Amazon Elastic Compute Cloud (Amazon EC2), Amazon Bedrock, Amazon Simple Storage Service (Amazon S3)	Analytics, Machine Learning (ML/AI), Serverless, Generative AI

Description
Workshop to build a Retrieval Augmented Generation (RAG) based application using Databricks and AWS.

Event Outputs (9)

Search

1

Key	Value	Stack name	Description	Type
a-01-workspace-WorkshopBucket	db-workshop-182363696740-us-west-2-0e9a8770	a-01-workspace	The workshop bucket of the Databricks Workspace	export
AccessKey	AKIASU5NZFZSP73TJKLK	a-01-workspace	-	export
		a-01-		

1. 아래로 스크롤

© 2008 - 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy policy Terms of use Cookie preferences

AWS Workshop Studio에서 Databricks Workspace 접속

aws workshop studio

insung.lee

Databricks GenAI workshop

- Zero to Generative AI in 4 hours
- Account setup
- Lab 1: Ingest and prepare data
- Lab 2: Prepare your chat model
- Lab 3: Deploy your chat model
- Lab 4: Evaluate your LLM (optional)
- Summary
- Admin Guide

AWS account access

Open AWS console (us-west-2)

Get AWS CLI credentials

Content preferences

Language: English

Exit event

Storage Service (Amazon S3)

Description

Workshop to build a Retrieval Augmented Generation (RAG) based application using Databricks and AWS.

Event Outputs (9)

Search

Key	Value	Stack name	Description	Type
a-01-workspace-WorkshopBucket	db-workshop-182363696740-us-west-2-0e9a8770	a-01-workspace	The workshop bucket of the Databricks Workspace	export
AccessKey	AKIASU5NZFZSP73TJKLK	a-01-workspace	-	export
AwsAccountId	182363696740	a-01-workspace	AWS account ID	output
DatabricksCatalog	catalog_182363696740	a-01-workspace	Catalog to use for Labs	output
DatabricksPassword	D@taisC00l	a-01-workspace	Password for your databricks workspace	output
DatabricksUserName	u+182363696740@awsbricks.com	a-01-workspace	Username for your databricks workspace	output
DatabricksWorkspaceURL	https://apj-aws-lab-182363696740.cloud.databricks.com	01-orkspace	The name of the Databricks workspace deployment	output
InstanceAccessURL	http://35.94.229.19:8080	a-01-workspace	The public IP address of the instance	output
SecretAccessKey	z3Pjpit4TV9VOuxnr24kdSTgs9WPrQ7j0T0WFnA6	a-01-workspace	-	export

1. 클릭

© 2008 - 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy policy Terms of use Cookie preferences

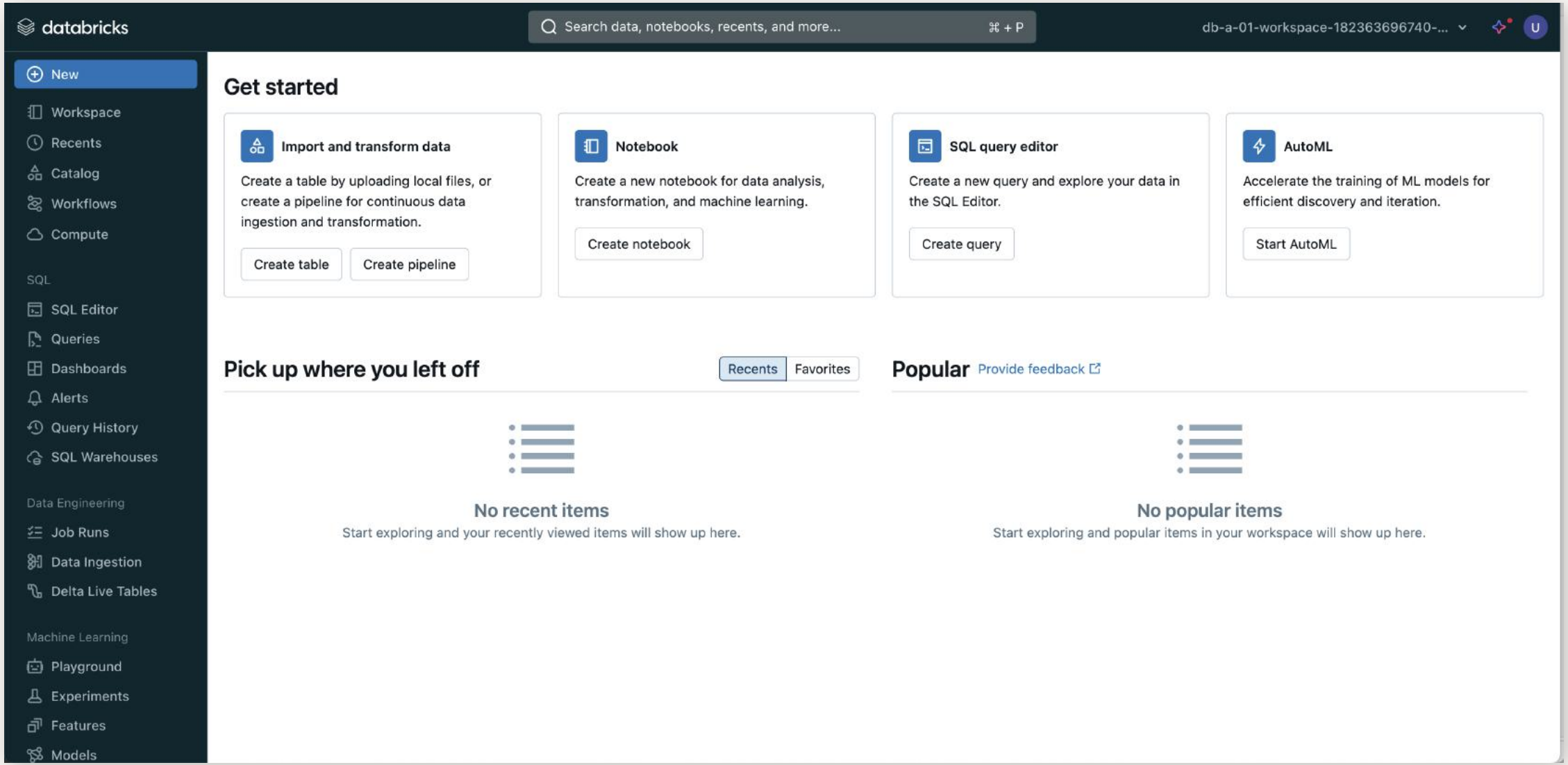
AWS Workshop Studio에서 Databricks Workspace 접속

1. 앞 페이지에서 확인한 DatabricksUserName / DatabricksPassword 입력

2. 클릭

©2024 Databricks Inc. — All rights reserved

AWS Workshop Studio에서 Databricks Workspace 접속

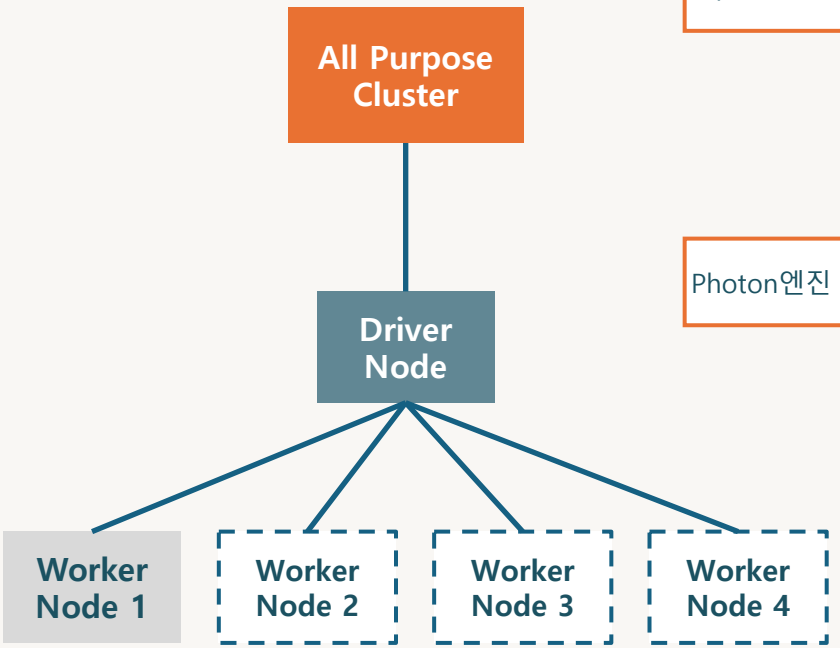


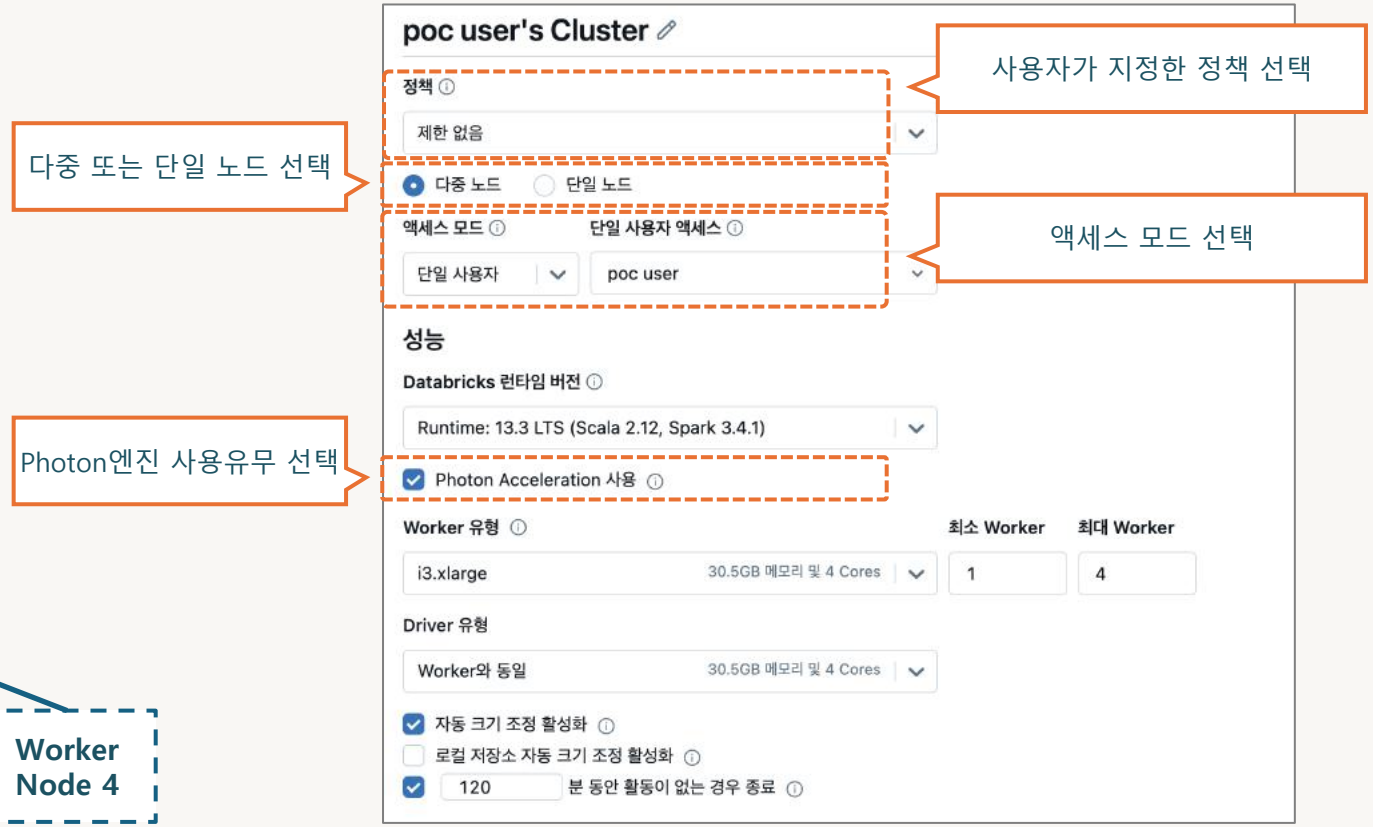
©2024 Databricks Inc. — All rights reserved

실습 컴퓨팅 환경과 스키마 구조

©2024 Databricks Inc. — All rights reserved

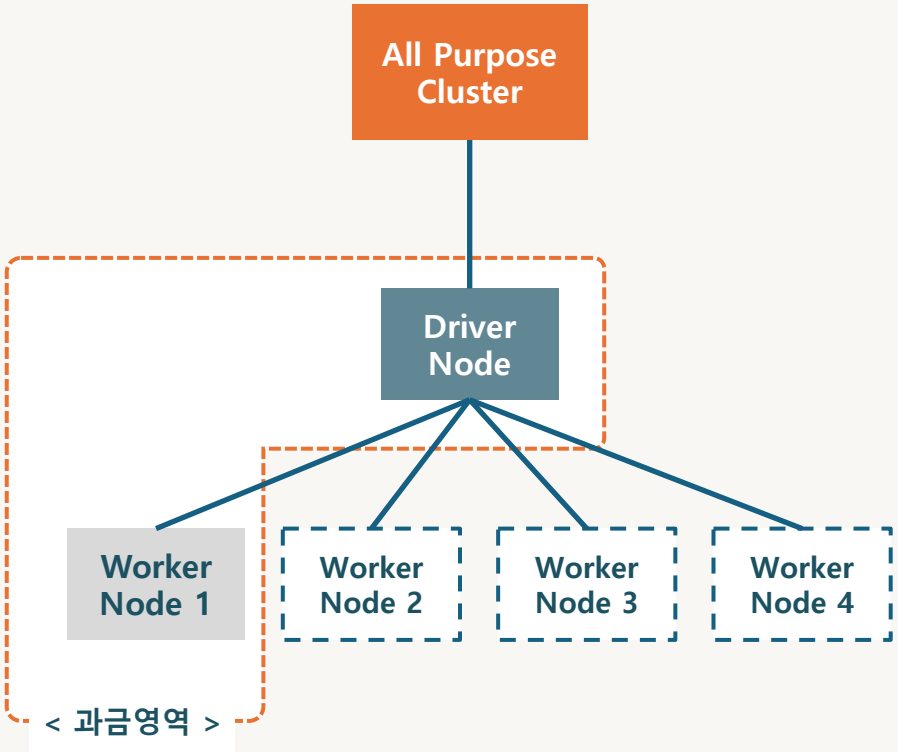
Databricks 컴퓨팅 Cluster의 동작 구조





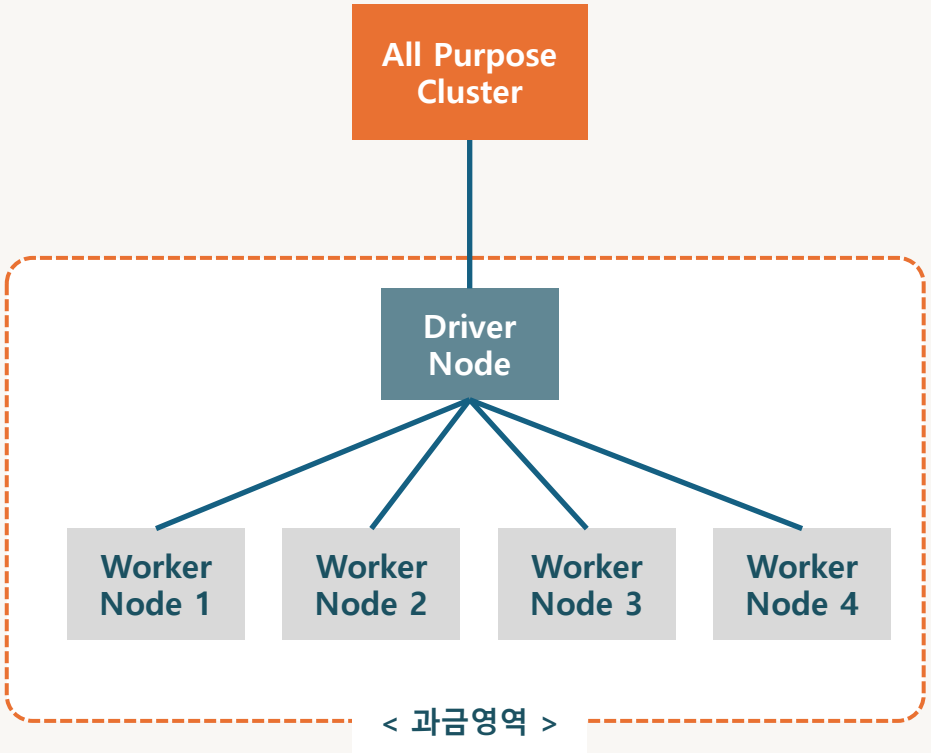
©2024 Databricks Inc. — All rights reserved

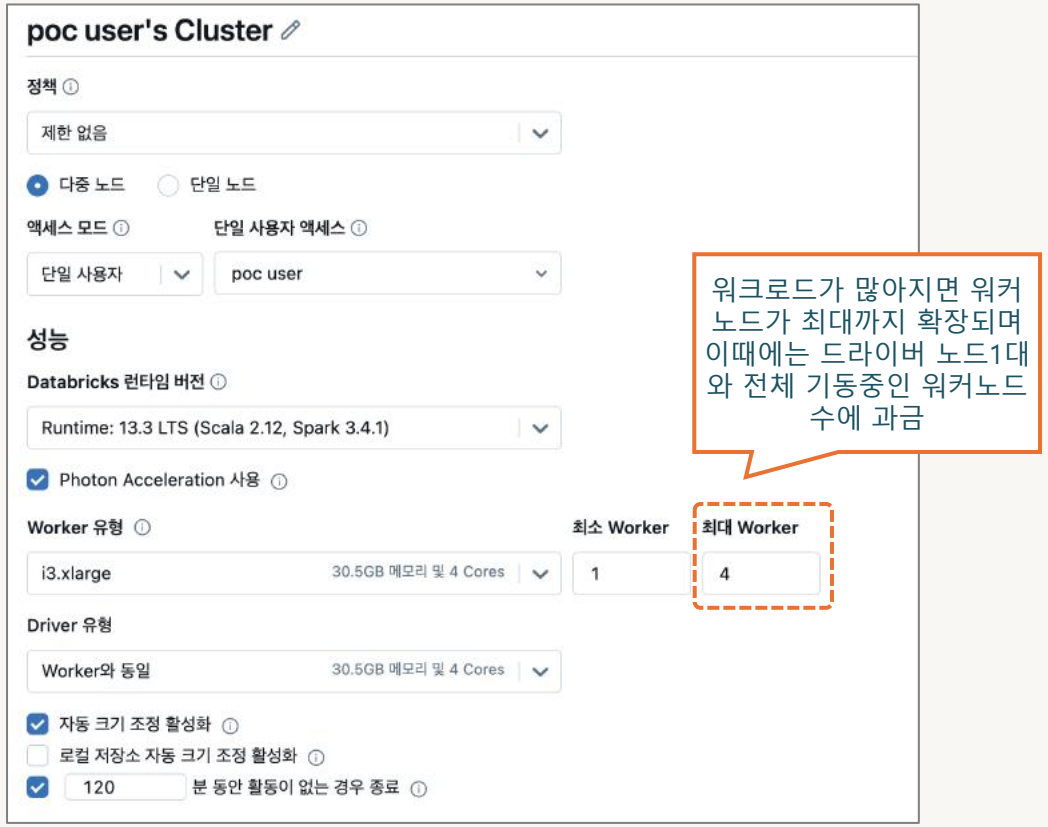
Databricks 컴퓨팅 Cluster의 동작 구조



©2024 Databricks Inc. — All rights reserved

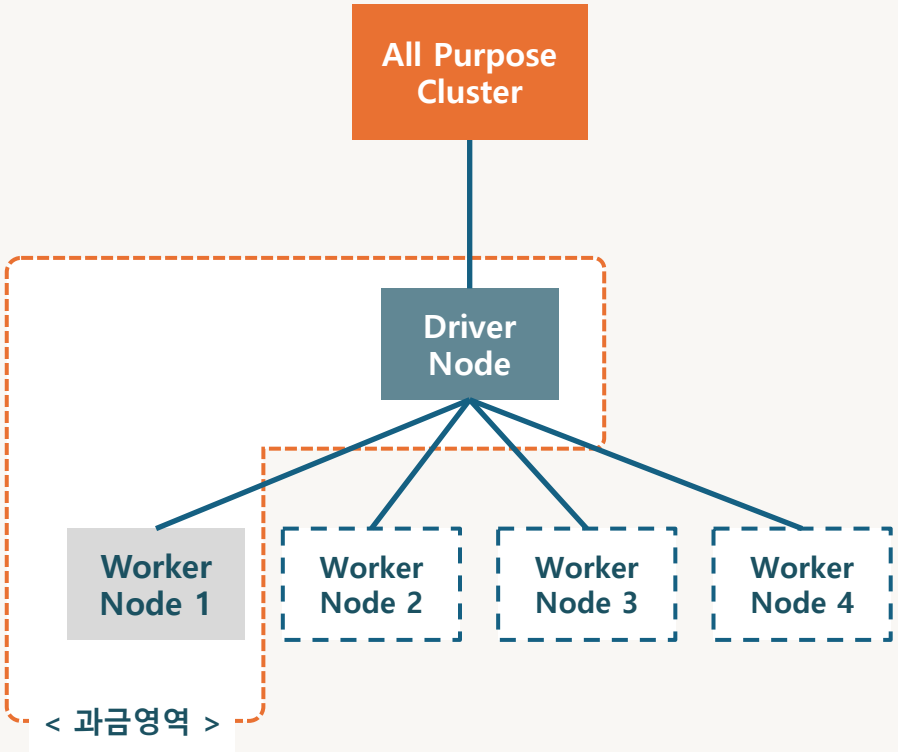
Databricks 컴퓨팅 Cluster의 동작 구조

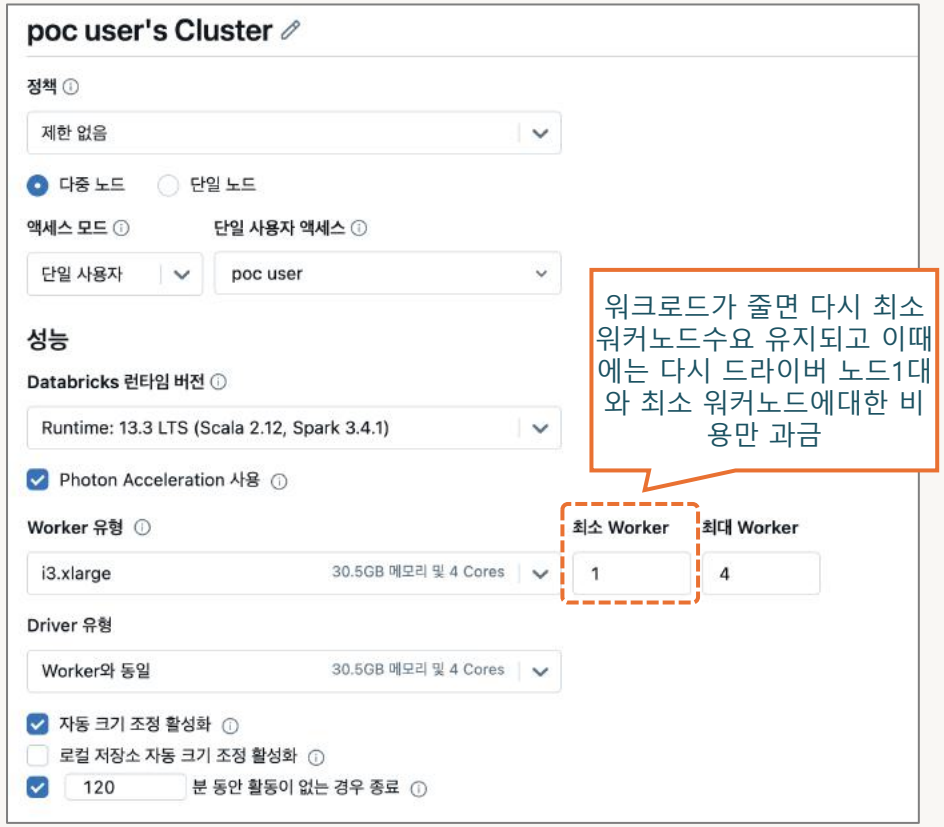




©2024 Databricks Inc. — All rights reserved

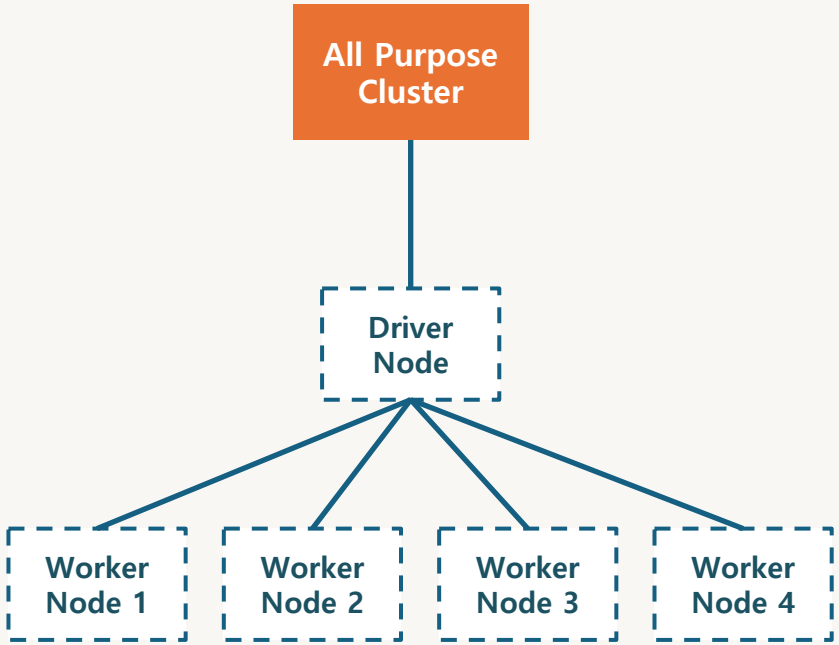
Databricks 컴퓨팅 Cluster의 동작 구조

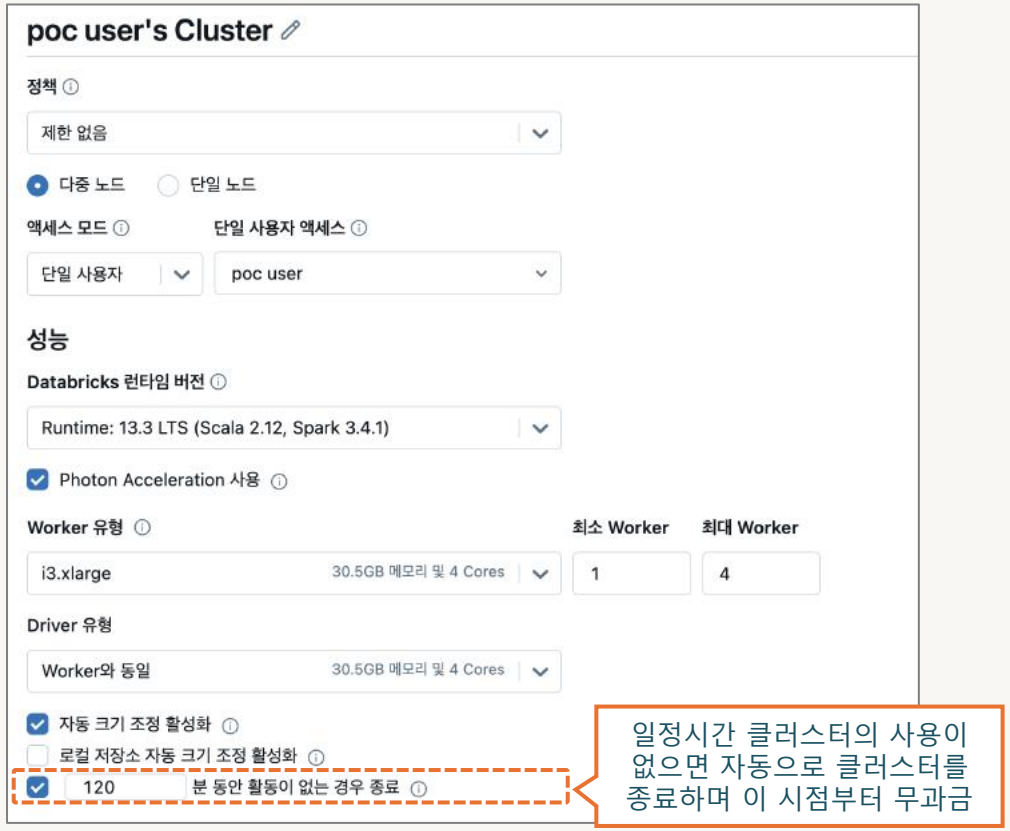




©2024 Databricks Inc. — All rights reserved

Databricks 컴퓨팅 Cluster의 동작 구조



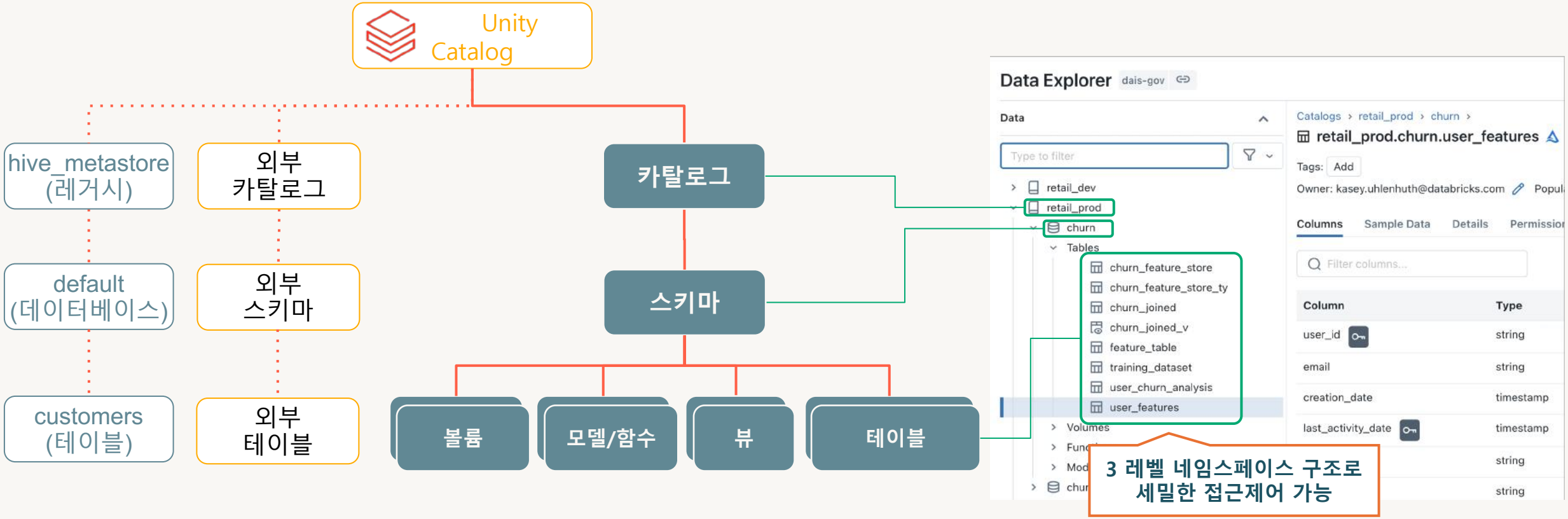


©2024 Databricks Inc. — All rights reserved

Databricks Unity Catalog 개요

Databricks Unity Catalog란?

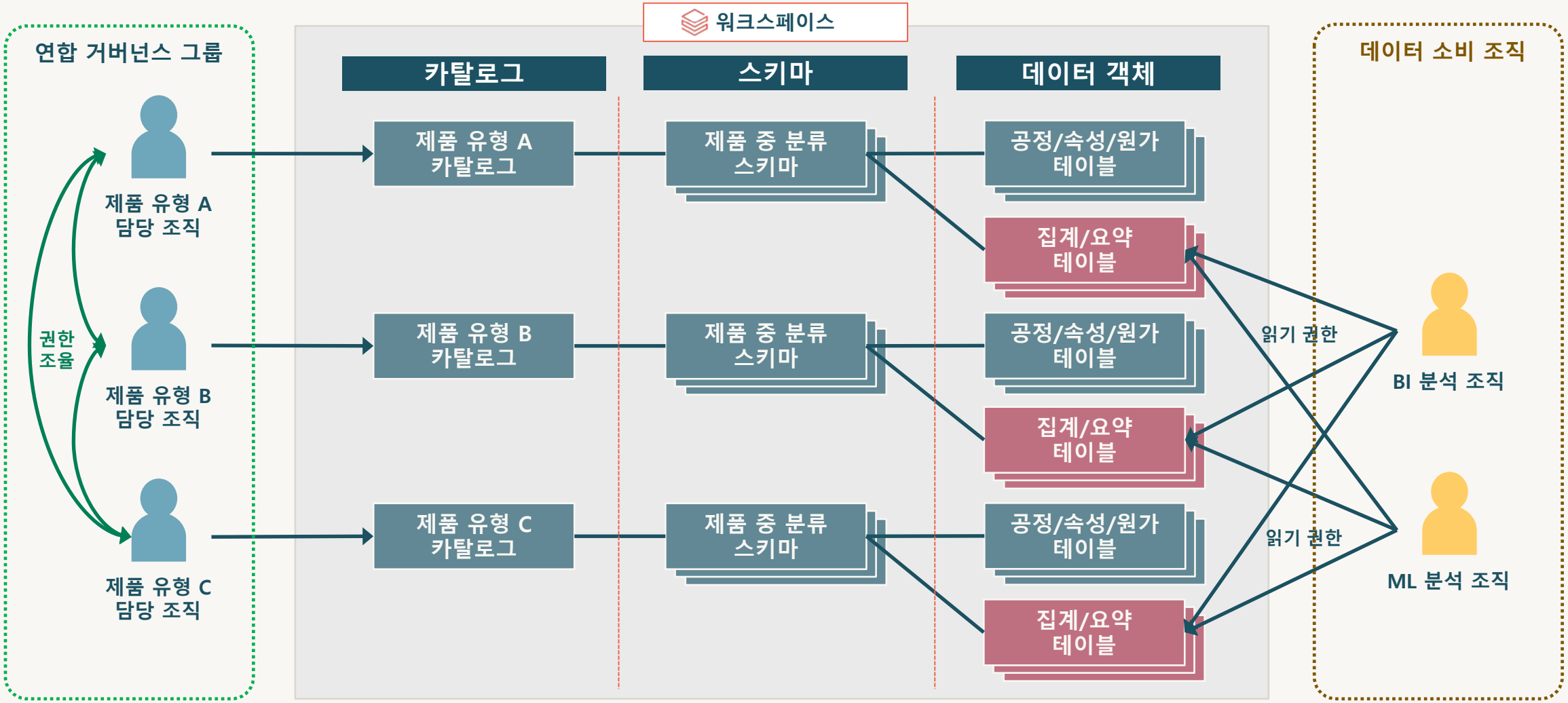
- Unity Catalog는 Databricks 플랫폼 내/외부의 데이터 및 AI객체들을 관리하는 통합 거버넌스 솔루션



SELECT * FROM mycatalog.myschema.mytable; --<catalog>.<database>.<table>

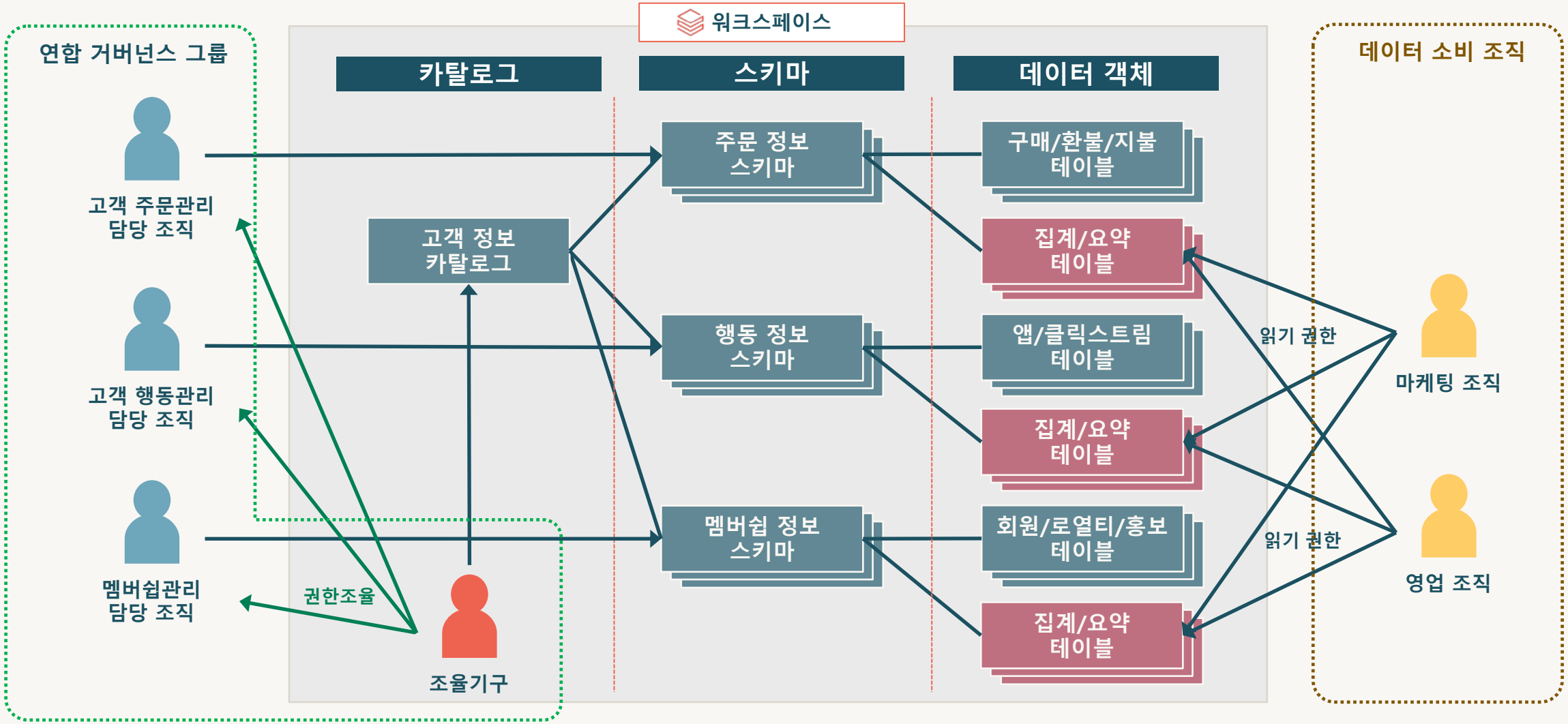
©2024 Databricks Inc. — All rights reserved

제조 A사의 카탈로그/스키마 구조



©2024 Databricks Inc. — All rights reserved

유통 B사의 카탈로그/스키마 구조

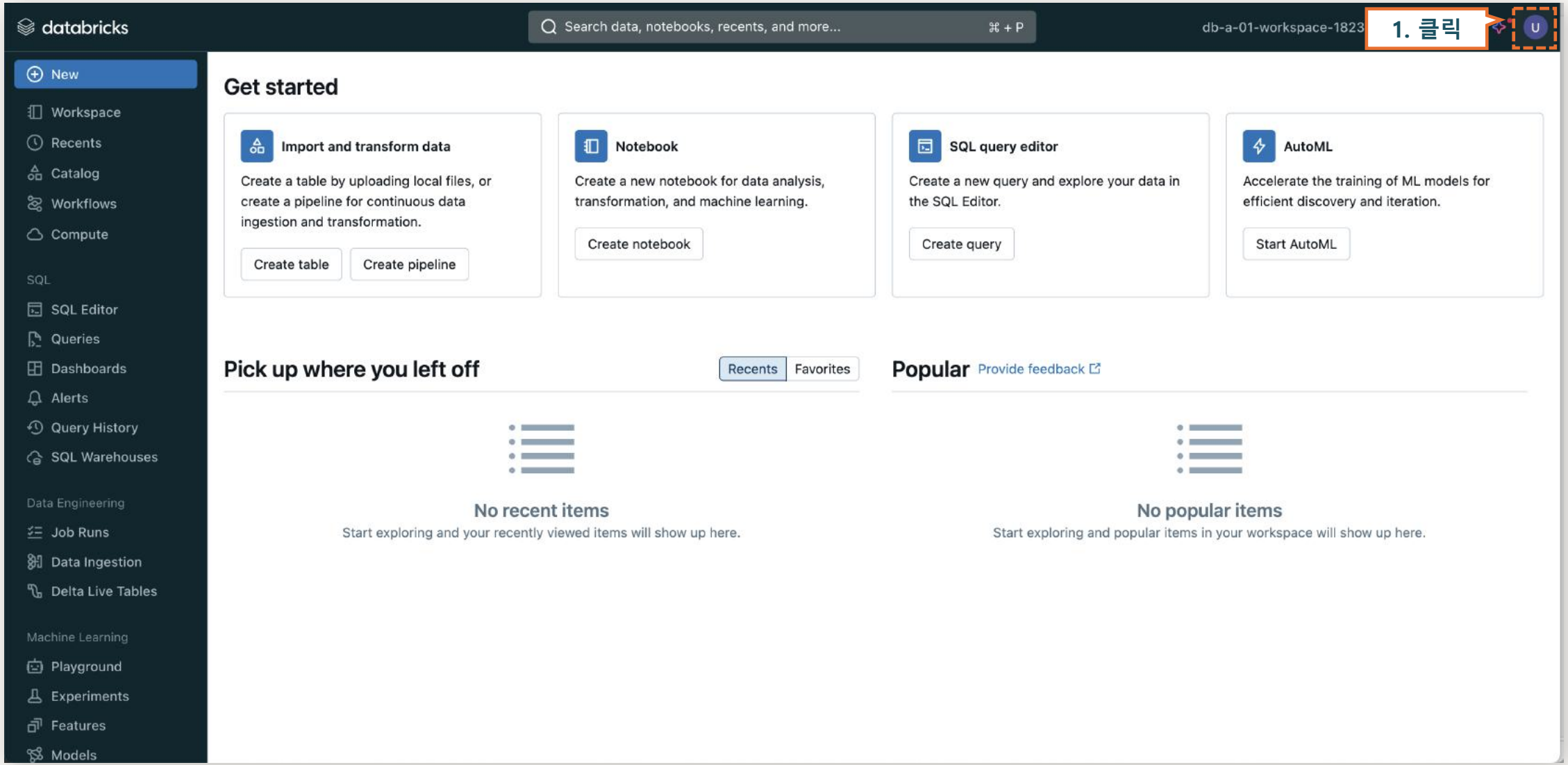


©2024 Databricks Inc. — All rights reserved

실 습 2 :
실습환경 셋업

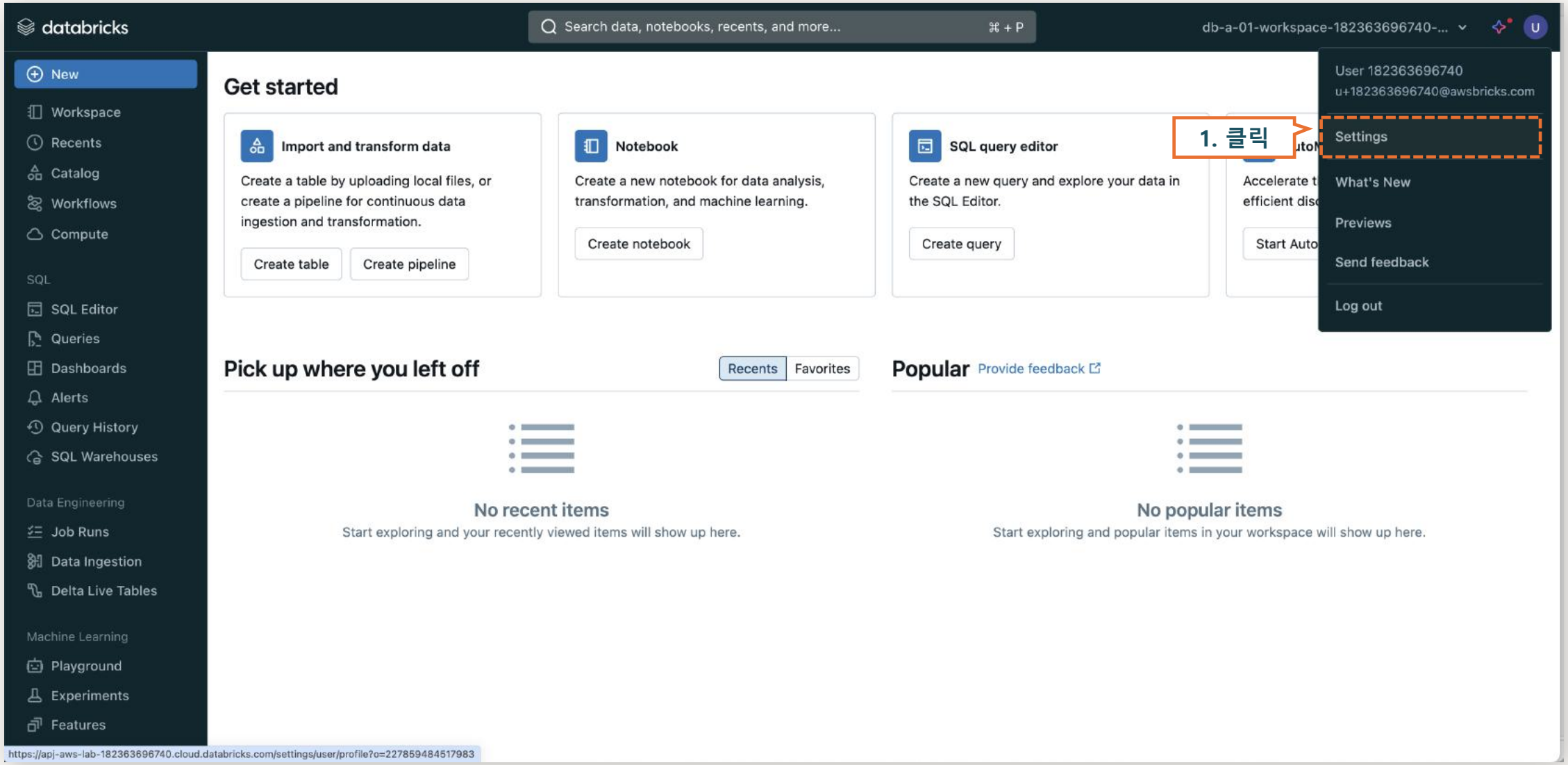
©2024 Databricks Inc. — All rights reserved

Workspace 메뉴 언어 변경



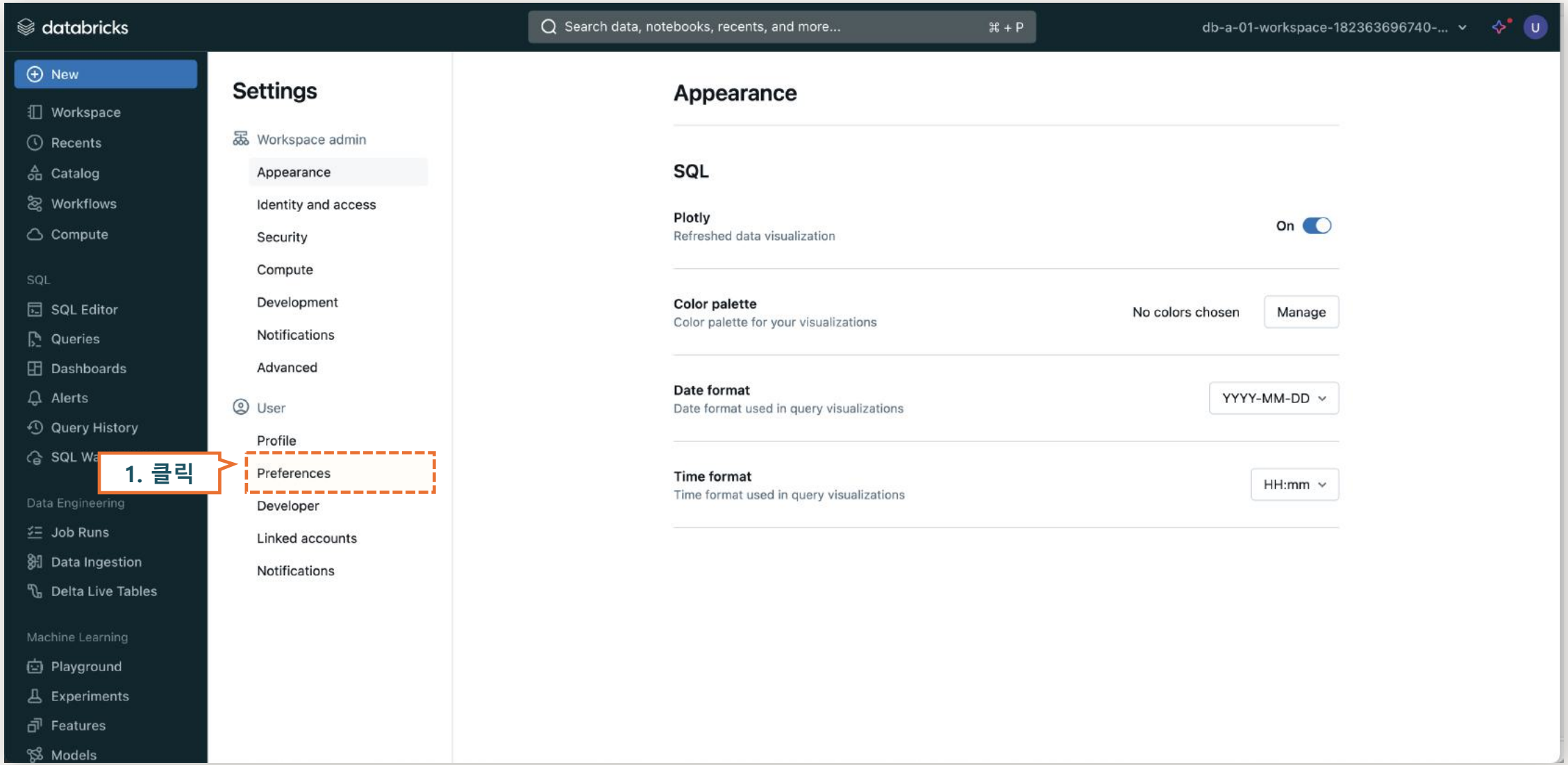
©2024 Databricks Inc. — All rights reserved

Workspace 메뉴 언어 변경



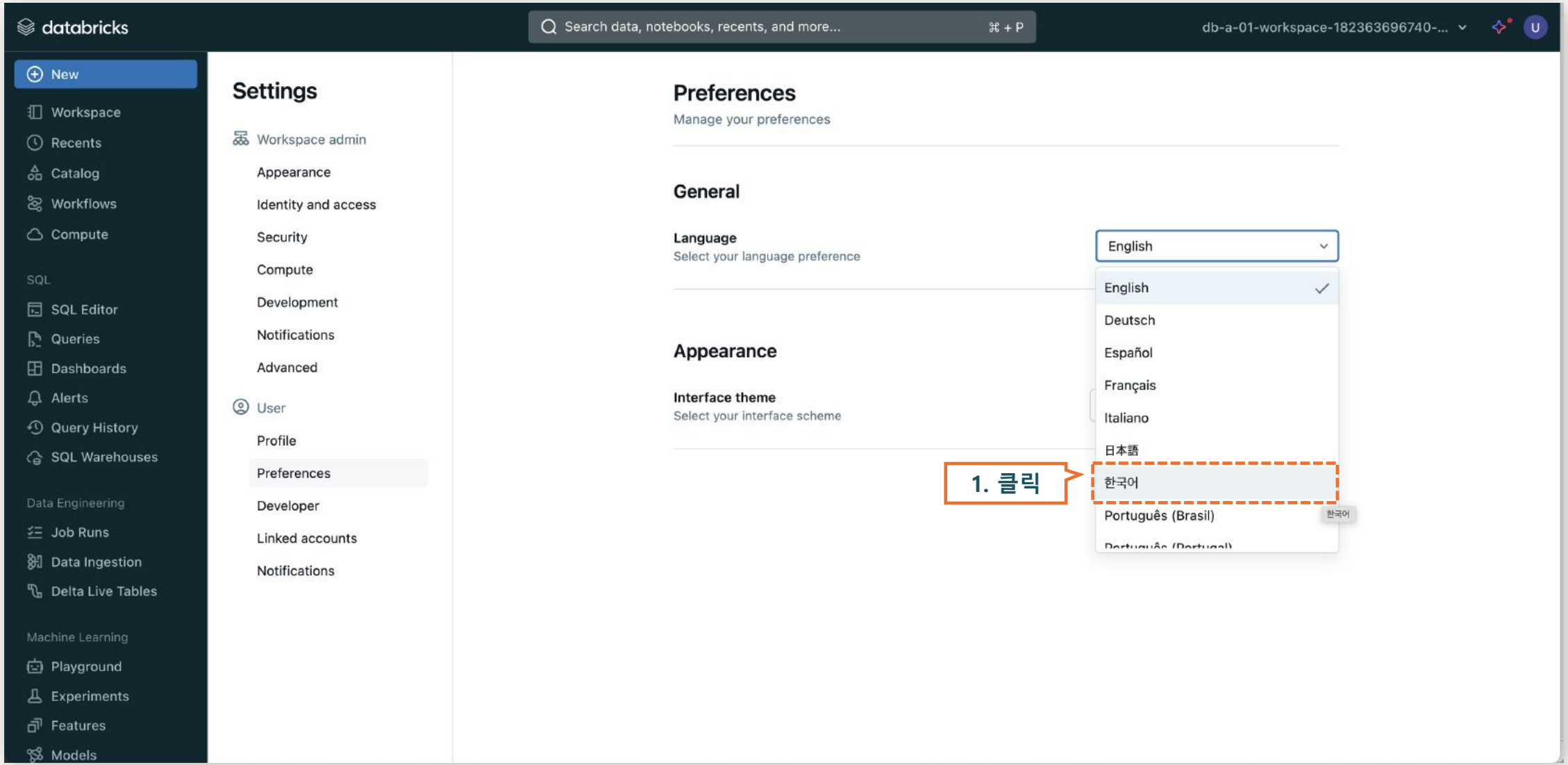
©2024 Databricks Inc. — All rights reserved

Workspace 메뉴 언어 변경



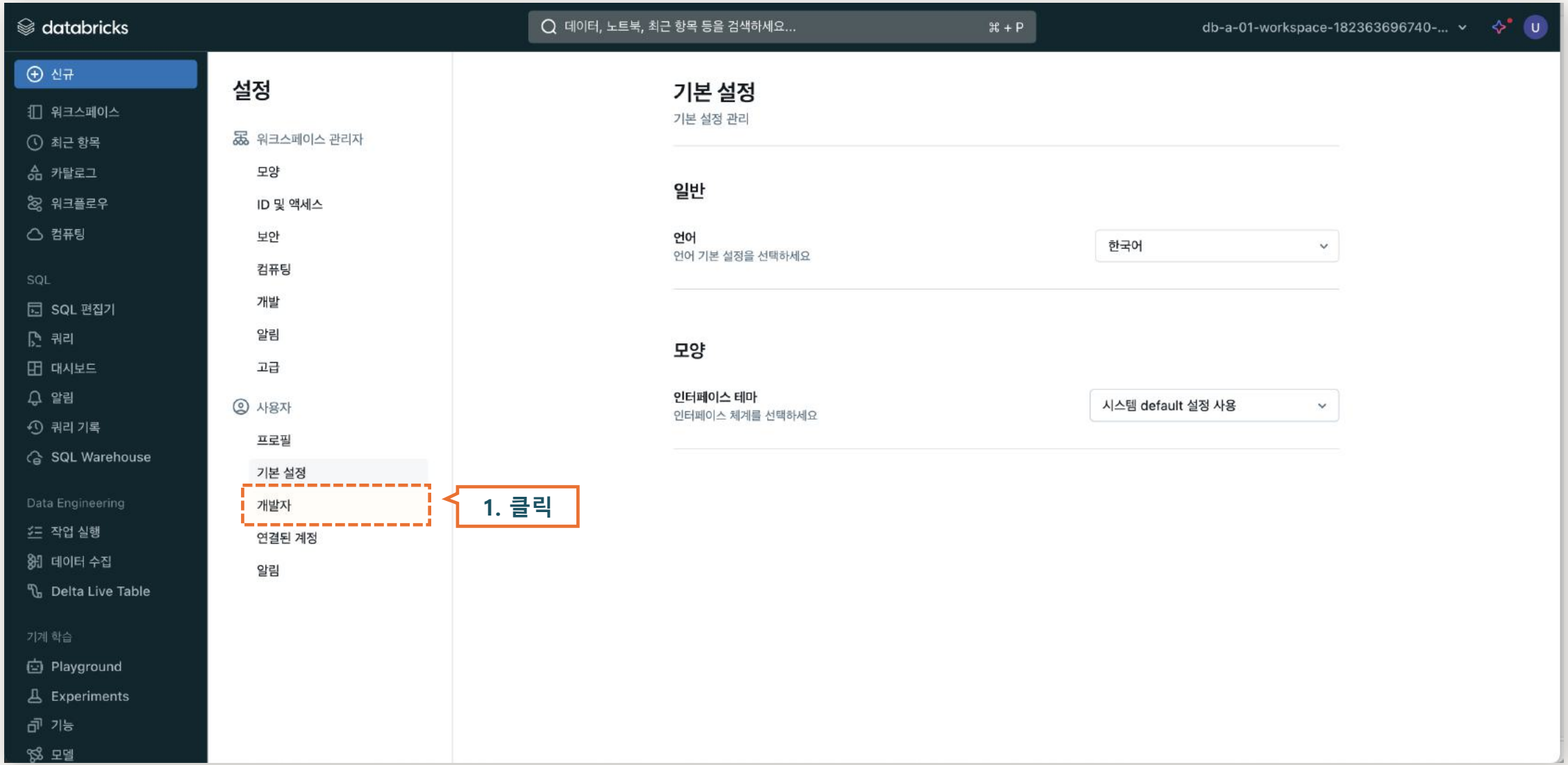
©2024 Databricks Inc. — All rights reserved

Workspace 메뉴 언어 변경



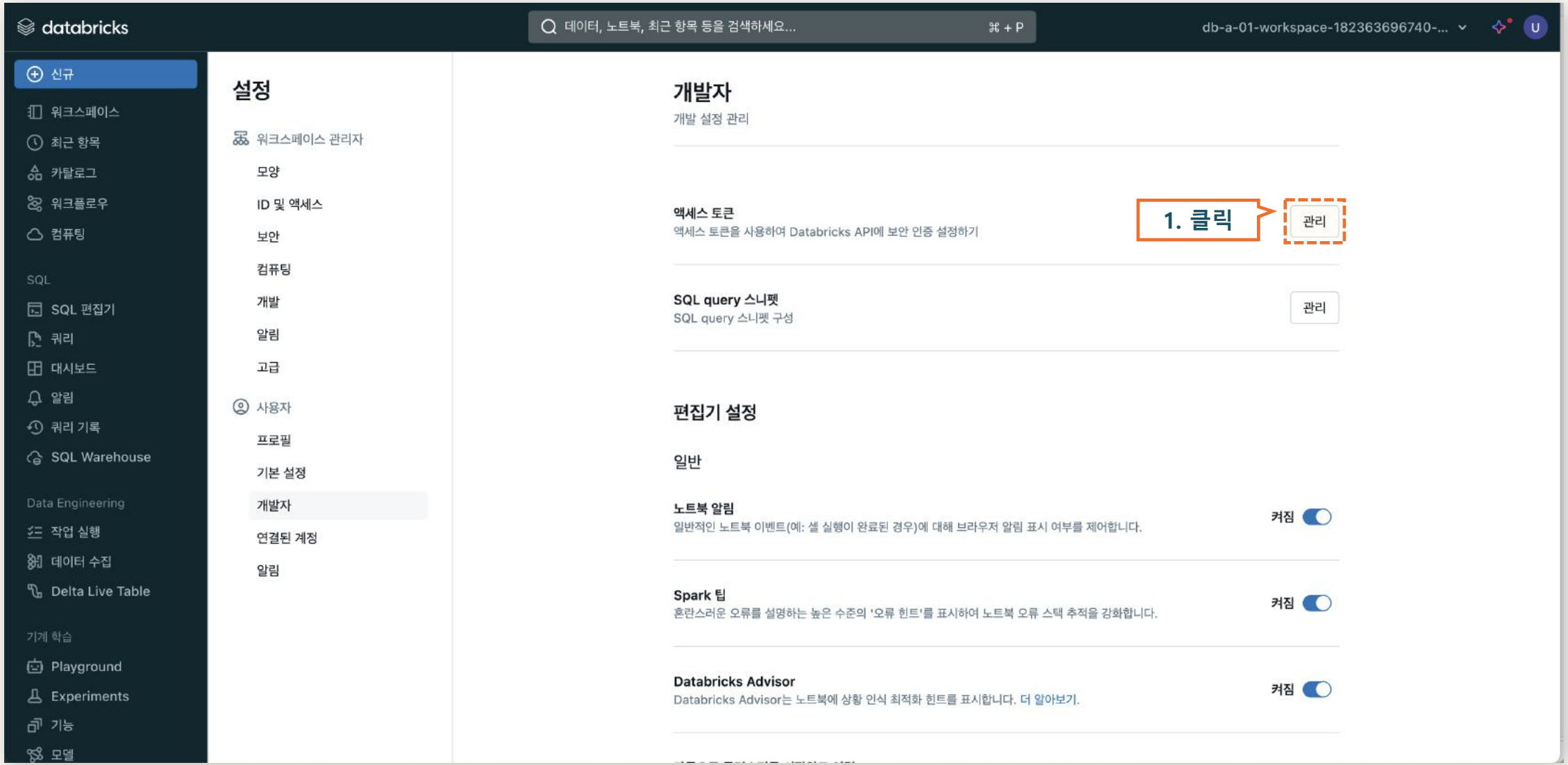
©2024 Databricks Inc. — All rights reserved

Databricks PAT(Personal Access Token) 발행



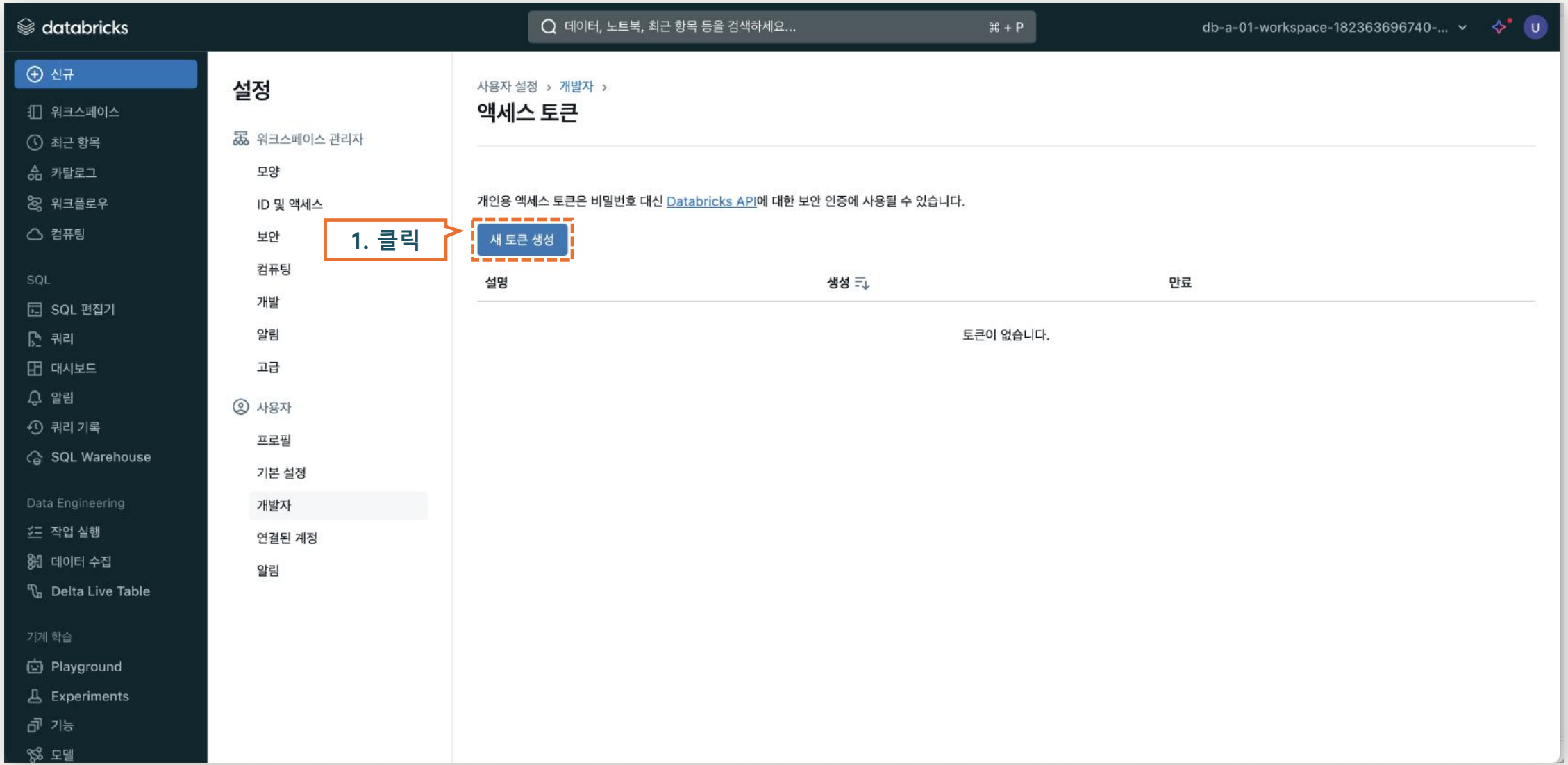
©2024 Databricks Inc. — All rights reserved

Databricks PAT(Personal Access Token) 발행



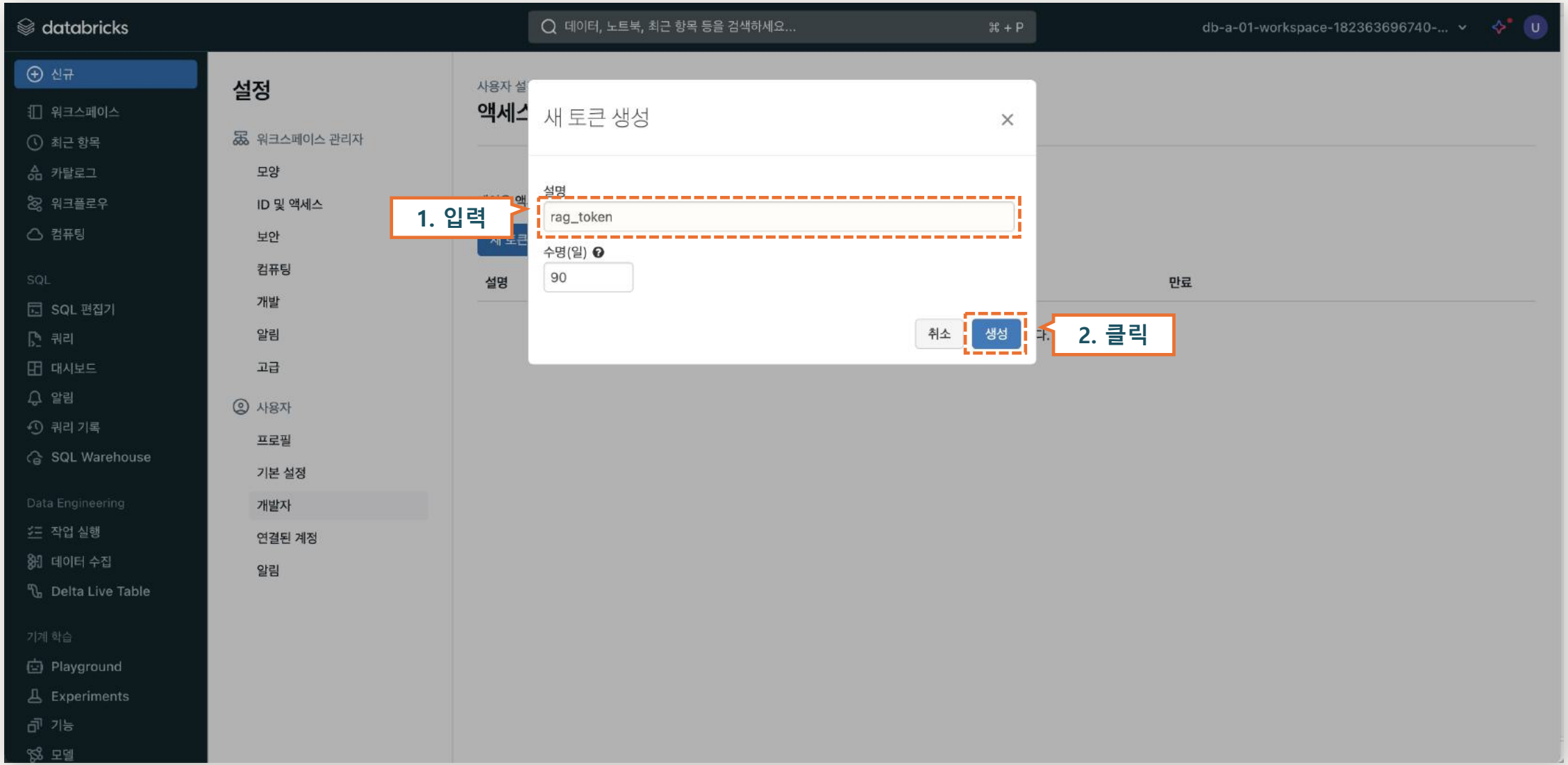
©2024 Databricks Inc. — All rights reserved

Databricks PAT(Personal Access Token) 발행



©2024 Databricks Inc. — All rights reserved

Databricks PAT(Personal Access Token) 발행

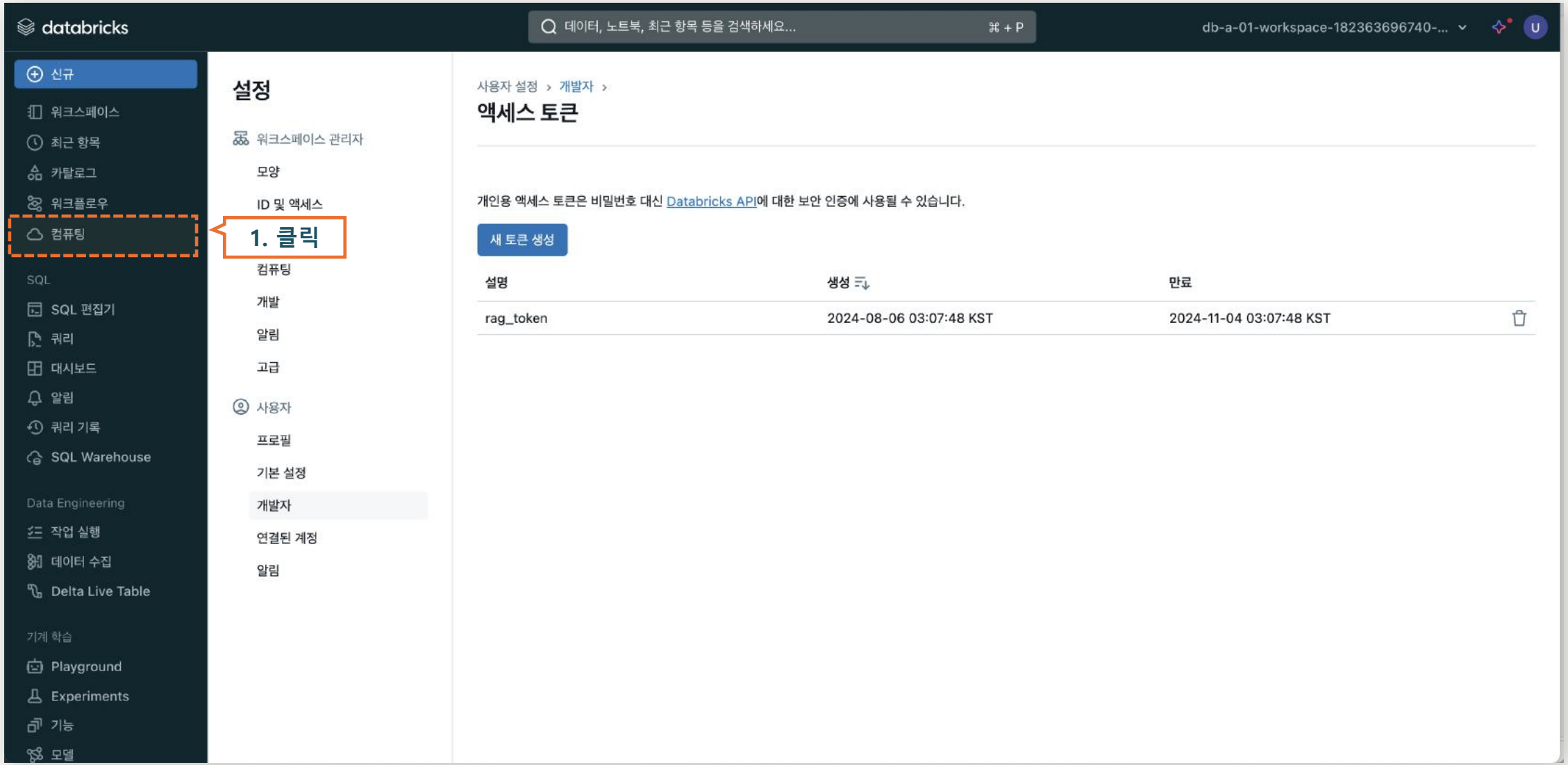


©2024 Databricks Inc. — All rights reserved

Databricks PAT(Personal Access Token) 발행

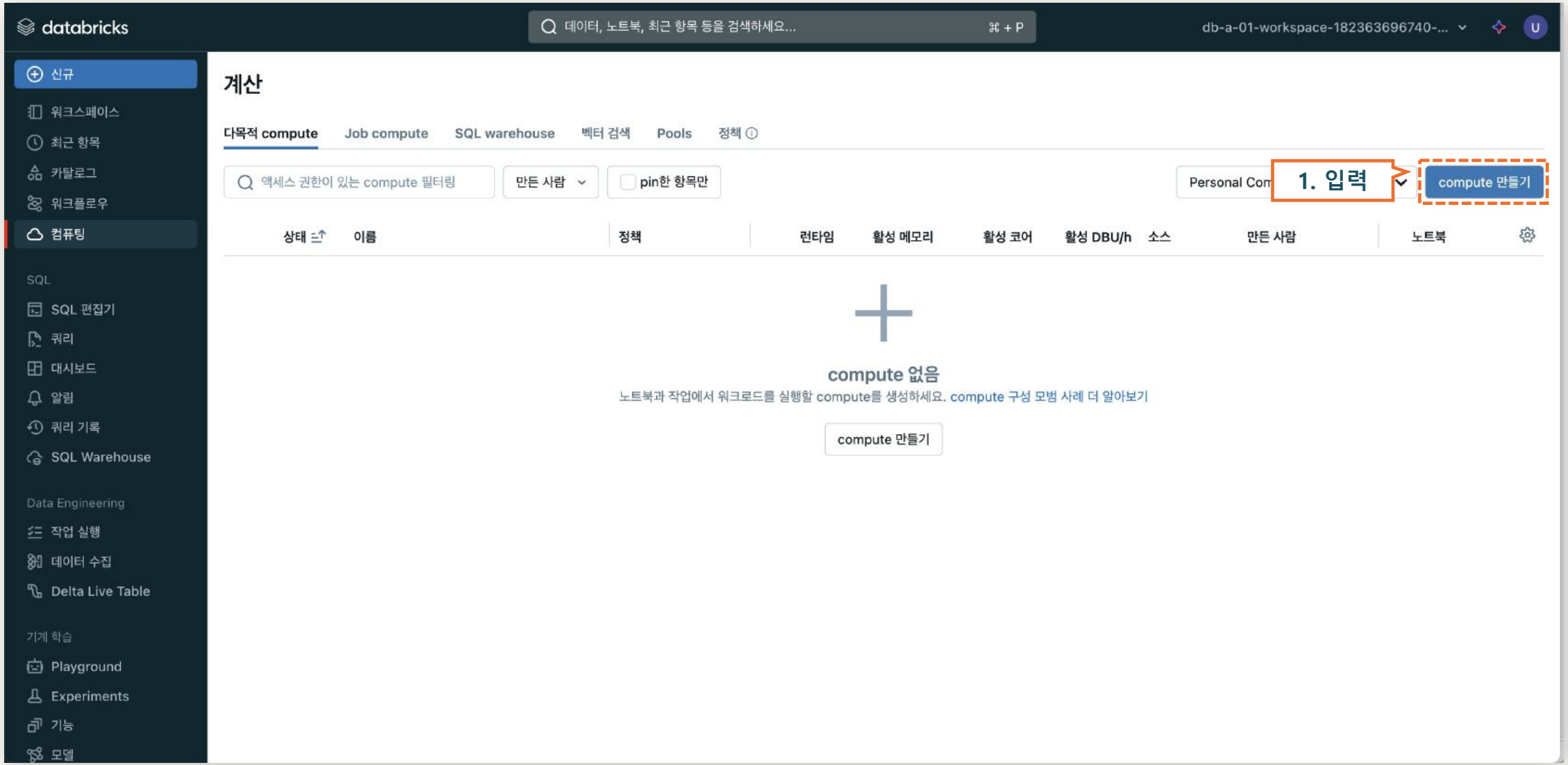
©2024 Databricks Inc. — All rights reserved

컴퓨팅 클러스터 생성



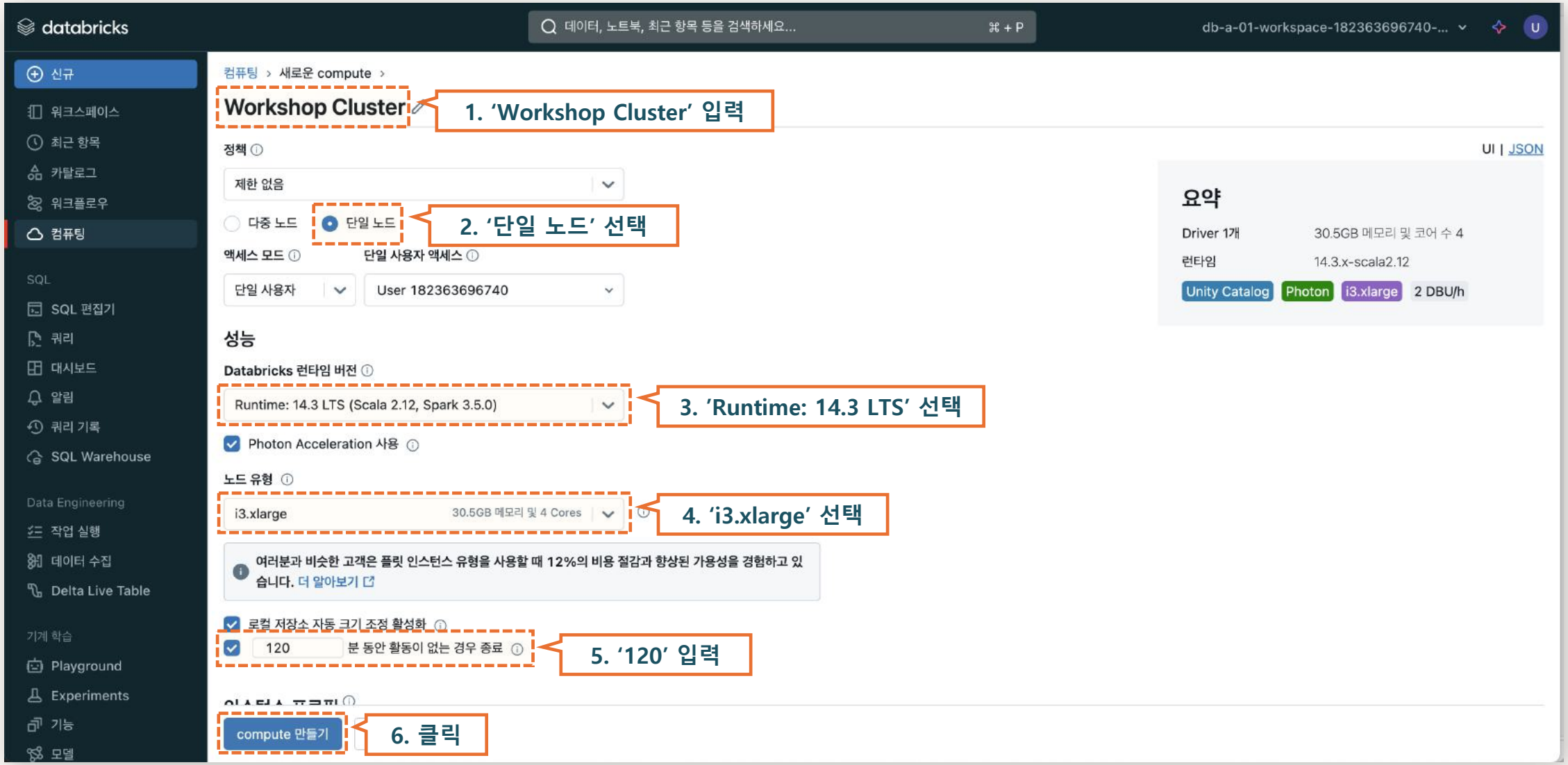
©2024 Databricks Inc. — All rights reserved

컴퓨팅 클러스터 생성



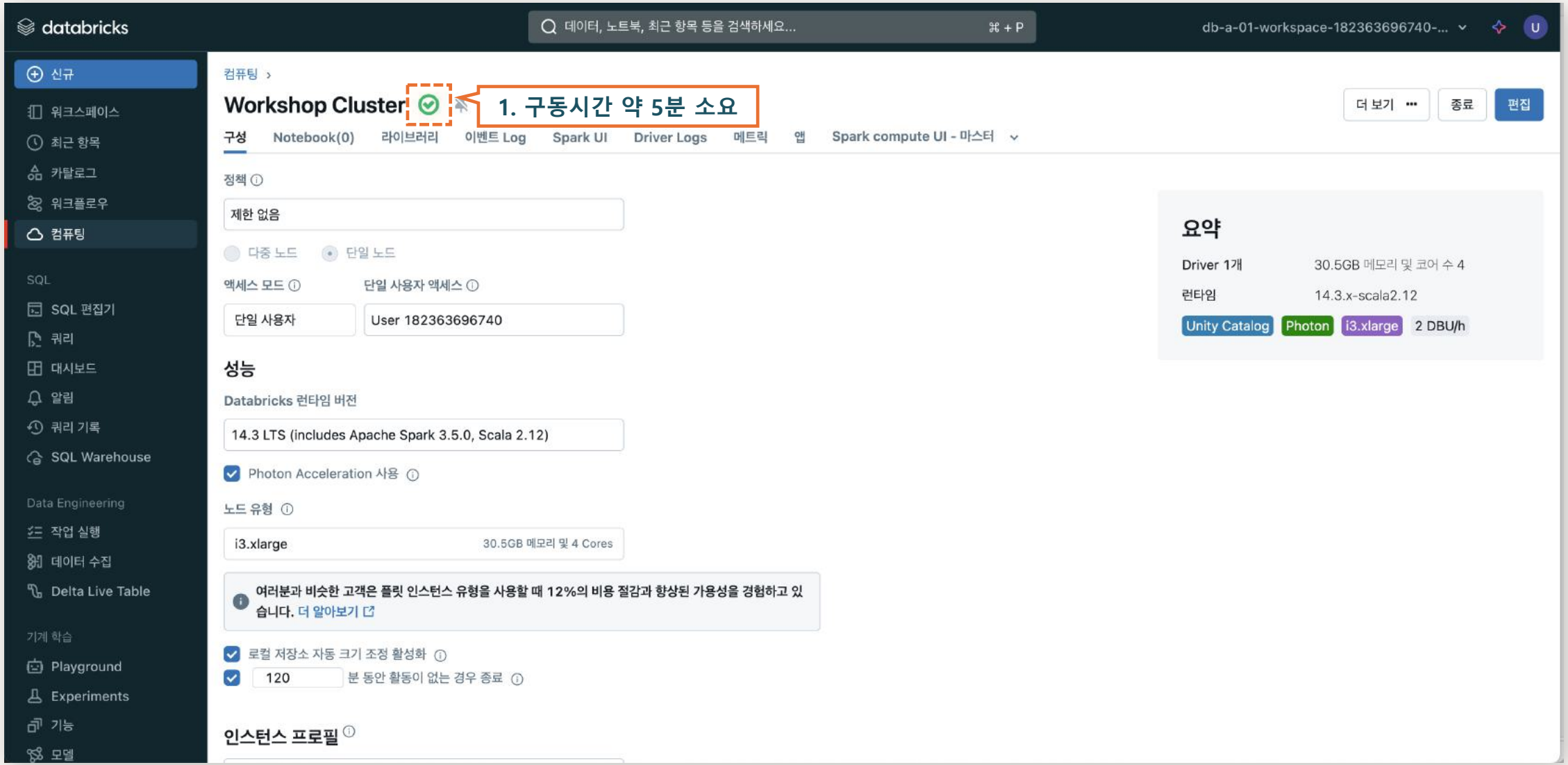
©2024 Databricks Inc. — All rights reserved

컴퓨팅 클러스터 생성



©2024 Databricks Inc. — All rights reserved

컴퓨팅 클러스터 생성

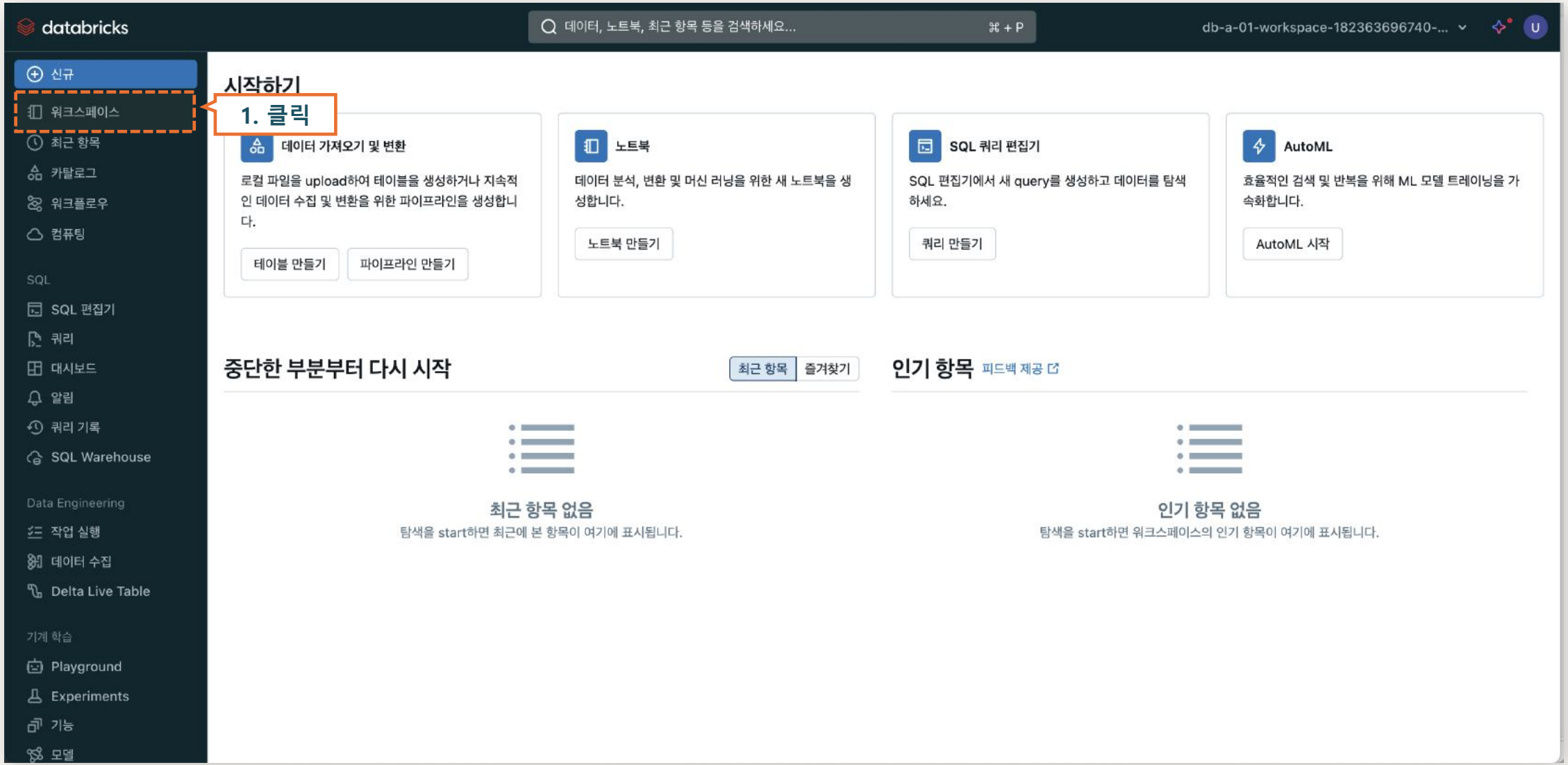


©2024 Databricks Inc. — All rights reserved

실 습 3 :
실습용 노트북 가져오기

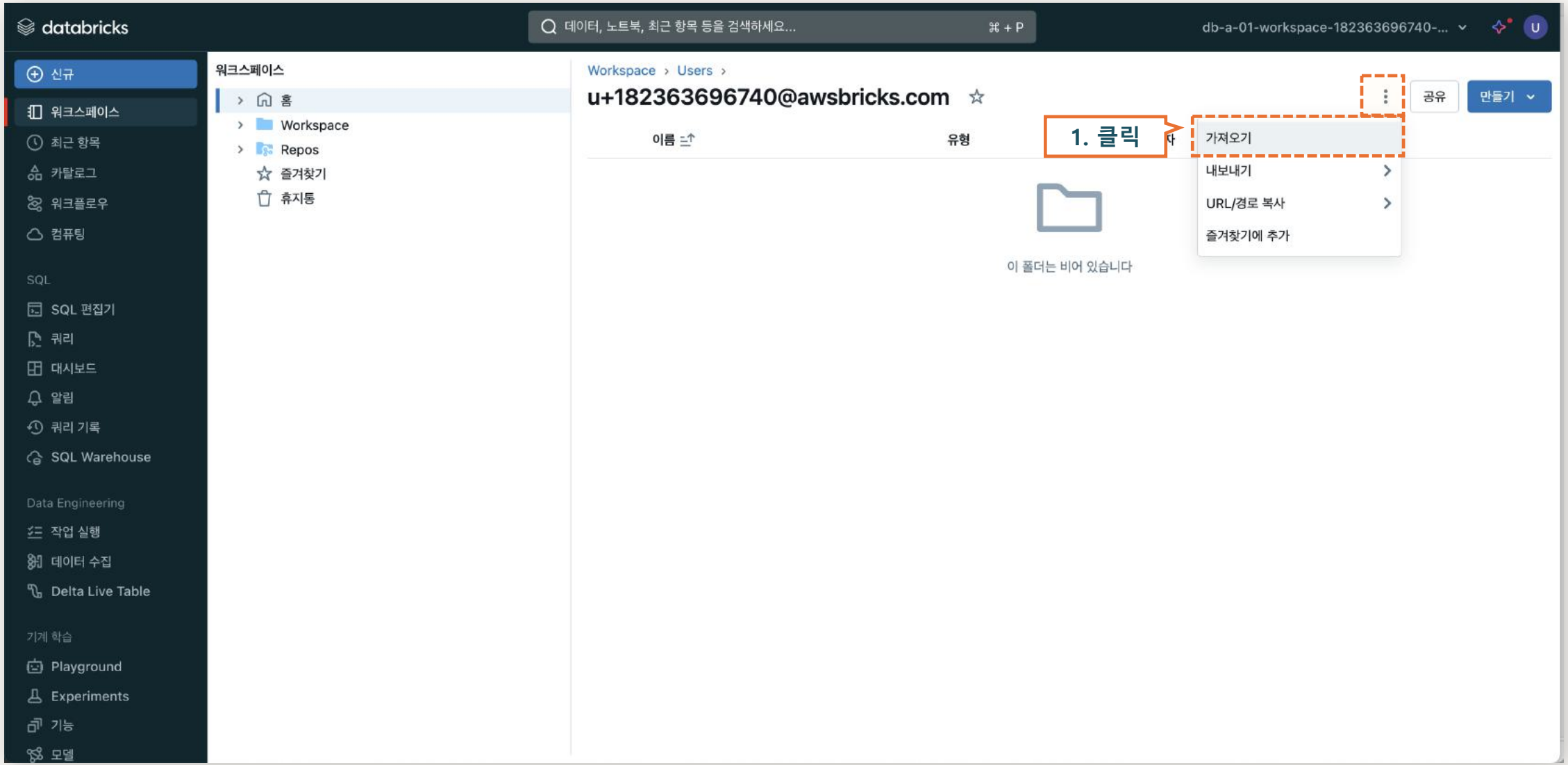
©2024 Databricks Inc. — All rights reserved

실습용 노트북 가져오기



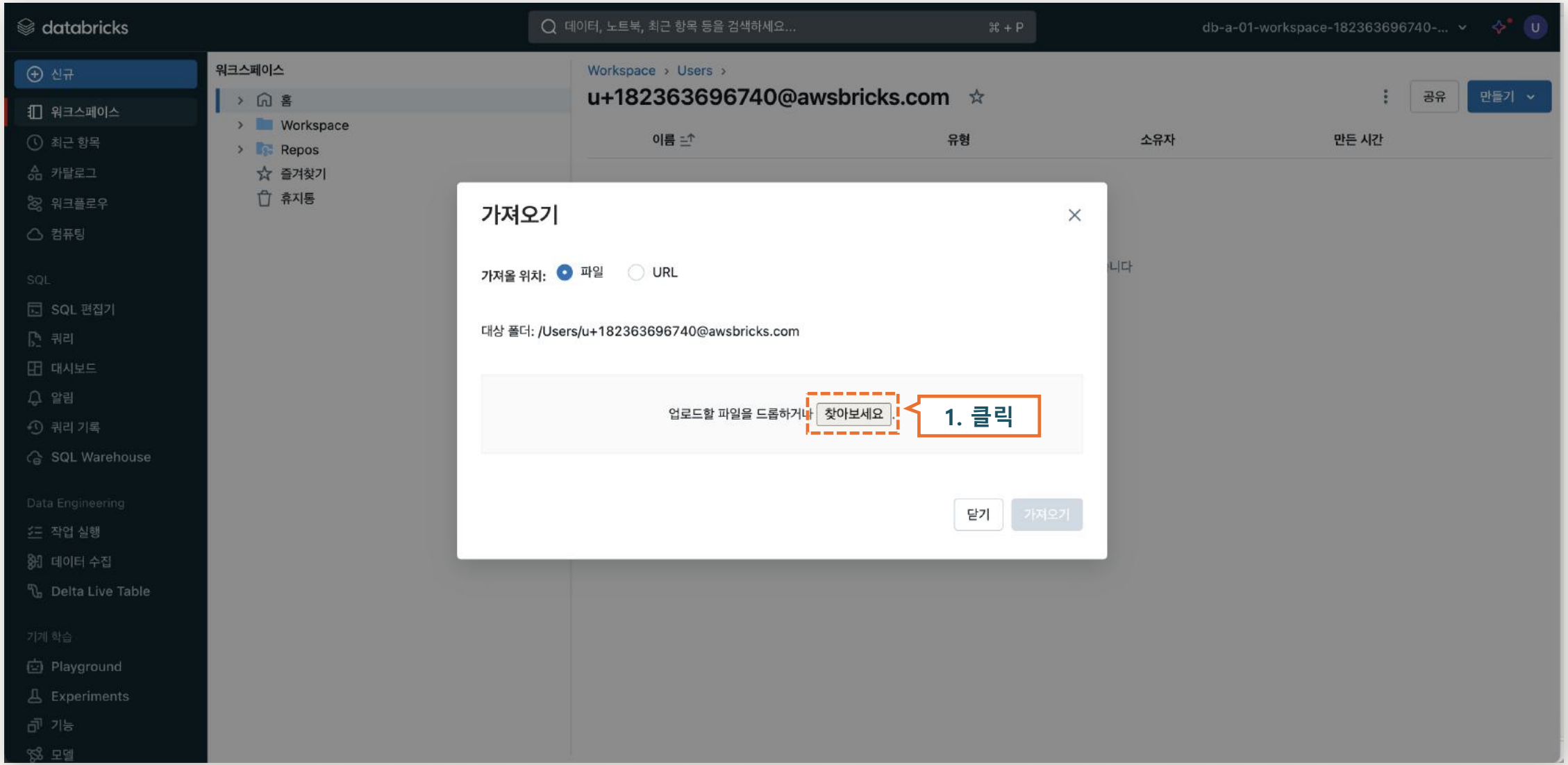
©2024 Databricks Inc. — All rights reserved

실습용 노트북 가져오기



©2024 Databricks Inc. — All rights reserved

실습용 노트북 가져오기



©2024 Databricks Inc. — All rights reserved

실습용 노트북 가져오기

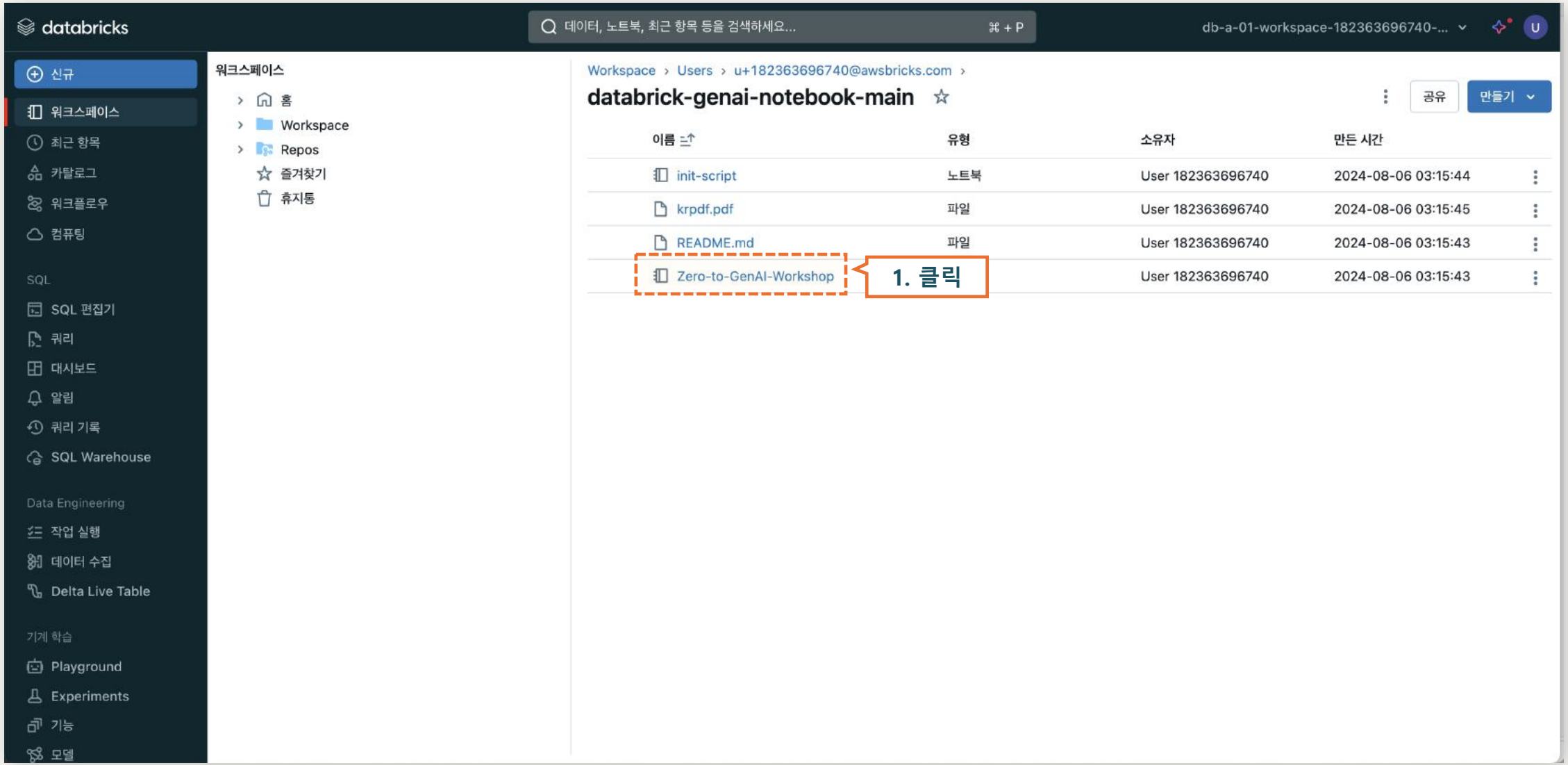
https://github.com/jinuland/databrick-genai-notebook/archive/refs/heads/main.zip
에서 다운로드받은 실습 파일 선택

©2024 Databricks Inc. — All rights reserved

실습용 노트북 가져오기

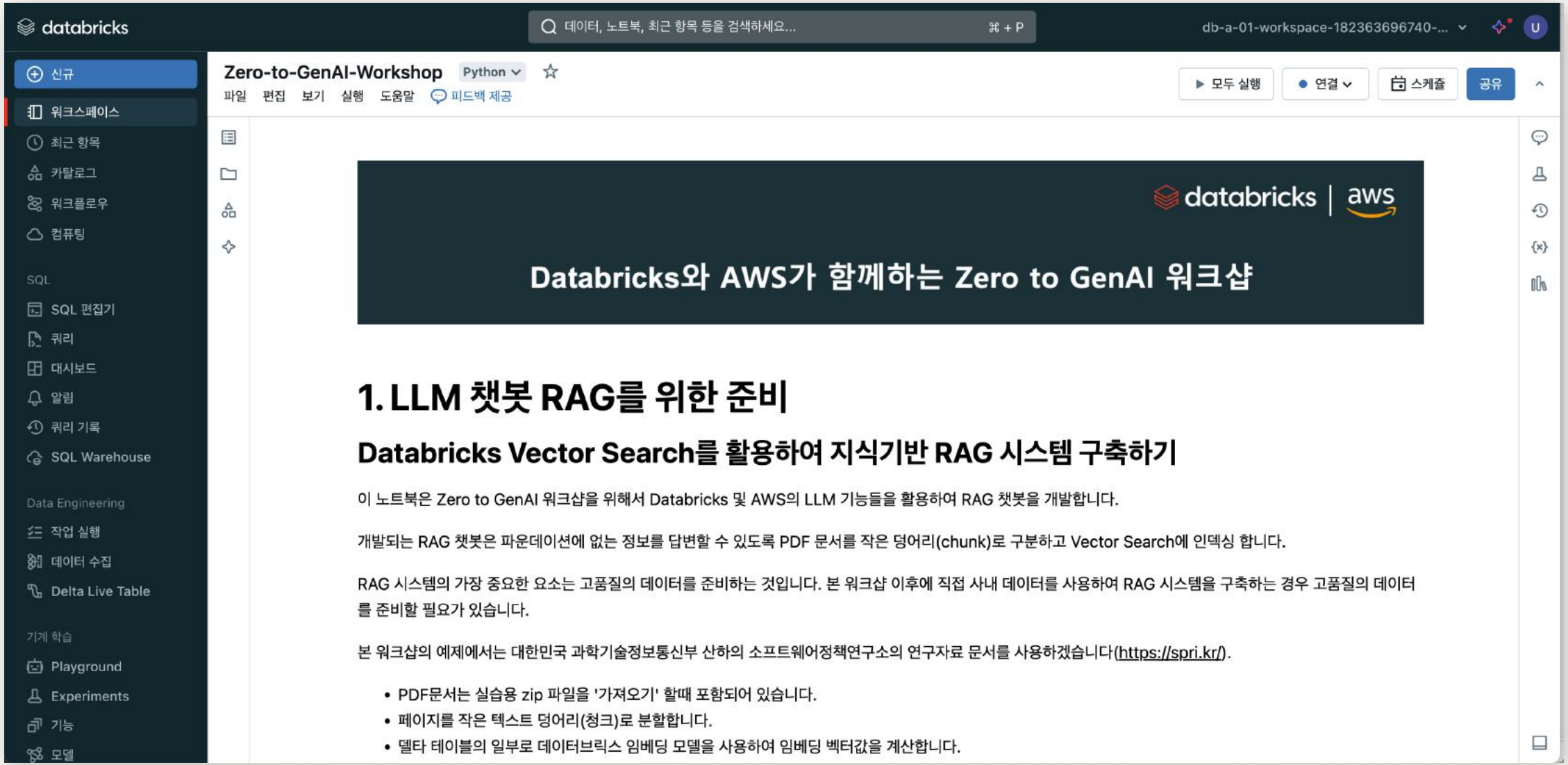
©2024 Databricks Inc. — All rights reserved

실습용 노트북 가져오기



©2024 Databricks Inc. — All rights reserved

실습용 노트북 가져오기



©2024 Databricks Inc. — All rights reserved

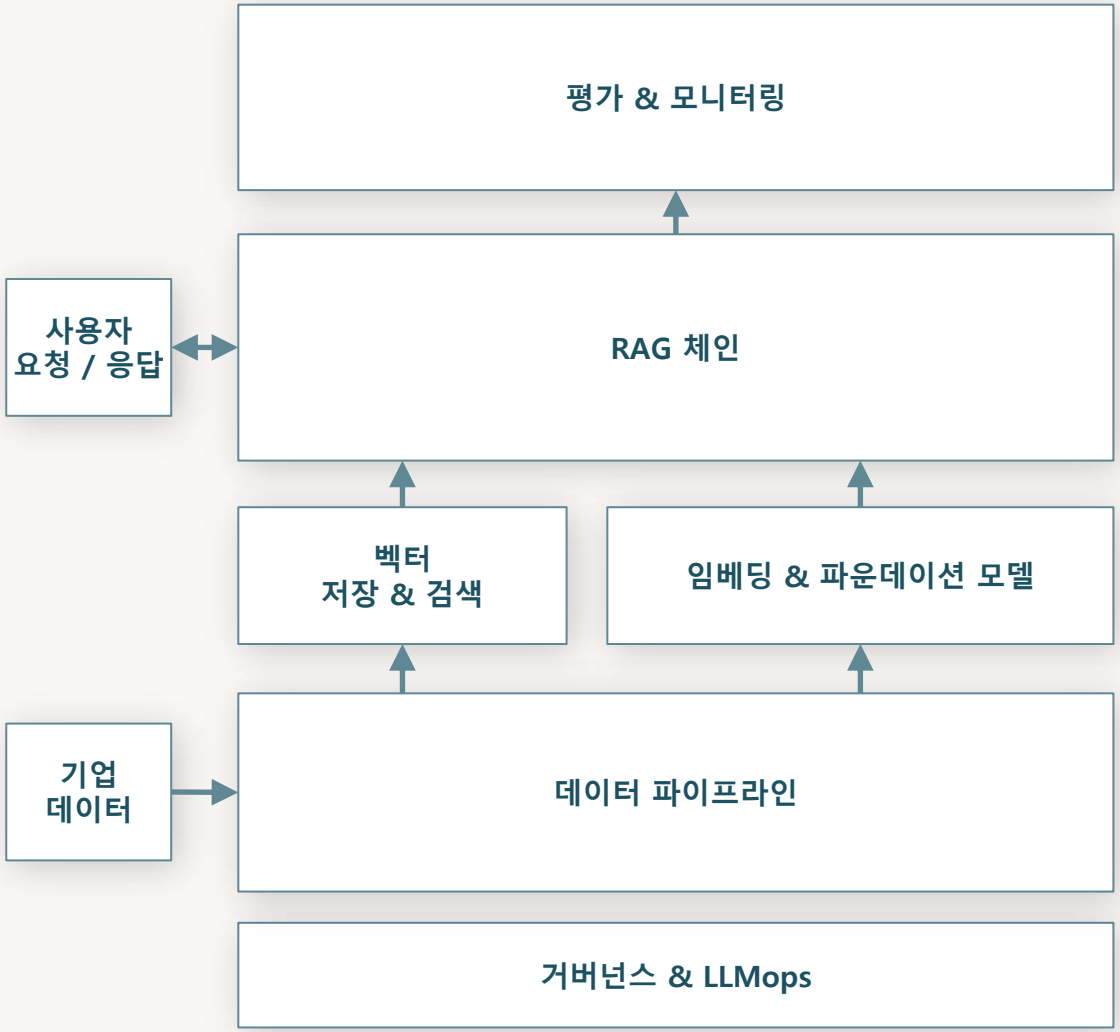
RAG 구현 전략

©2024 Databricks Inc. — All rights reserved

RAG 기본사항

RAG 애플리케이션 구현을 위한 핵심 구성요소

- 데이터 파이프라인 : 데이터 파이프라인을 사용하여 PDF와 같은 비정형 문서를 검색에 적합한 형식으로 변환
- RAG 체인 : 사용자의 질문과 저장된 벡터 데이터를 기반으로 아래와 같이 언어처리 체인을 호출
  ① 사용자의 질문을 이해
  ② 지원 데이터를 검색
  ③ LLM을 호출하여 사용자의 질문과 지원 데이터를 기반으로 응답을 생성
- 평가 & 모니터링: RAG 애플리케이션을 평가하여 품질/비용/대기 시간을 파악하고 비즈니스 요구 사항을 충족하는지 확인
- 거버넌스 및 LLMOps : 데이터 계보 및 거버넌스(액세스 제어)를 포함하여 각 구성 요소의 수명 주기를 추적하고 관리



< RAG 시스템 구조 >

©2024 Databricks Inc. — All rights reserved

청크란?

청크(Chunk)란?

- 검색 증강 생성(RAG) 시스템에서 지식정보를 벡터 데이터베이스에 저장하기위해 지식정보를 분할하는 단위
- 청크는 RAG시스템의 **효율성과 성능**에 영향을 미치는 주요한 매개변수 중 하나

청크의 크기(Chunk Size)

- 일반적으로 청크는 512 ~ 2048 토큰의 크기를 사용
- 청크의 크기는 LLM에 지식정보를 제공하는 정보의 양이기 때문에 그 크기를 결정하는 것이 중요
- 청크의 크기가 크면 많은 정보량을 LLM에 제공할 수 있으나 컨텍스트의 크기가 커지기 때문에 응답의 생성시간이 지연
- 청크의 크기가 작으면 컨텍스트의 크기가 작아지기 때문에 응답의 생성시간이 짧아지지만 정보량이 작아 응답의 관련성이 저하
- 따라서 청크의 크기는 필수 정보의 손실을 줄이면서 사용자에게 적절한 응답시간을 제공할 수 있도록 균형을 맞추는 것이 필요

https://www.llamaindex.ai/blog/evaluating-the-ideal-chunk-size-for-a-rag-system-using-llamaindex-6207e5d3fec5

©2024 Databricks Inc. — All rights reserved

청크 크기결정

청크의 평가방법

- 충실도 평가기(Faithfulness Evaluator) : 응답이 소스노드와 얼마나 일치하는지를 평가(응답이 허위가 아닌지 측정)
- 관련성 평가기(Relevancy Evaluator) : 질의가 응답 + 소스노드와 얼마나 일치하는지를 평가(응답이 질의와 관련이 있는지 측정)
- 응답시간 평가기(Response Evaluator) : 질의와 응답 시간간의 차이를 평가

LlamaIndex사가 측정한 청크의 크기별 성능

청크 크기	평균 응답시간(초)	평균 충실도	평균 관련도
128	1.55	0.85	0.78
256	1.57	0.90	0.78
512	1.68	0.85	0.85
1024	1.68	0.93	0.90
2048	1.72	0.90	0.89

청크의 최적 크기는 언어별, 도메인별로 차이가 있을 수 있으므로 직접 측정하여 결정할 필요가 있음

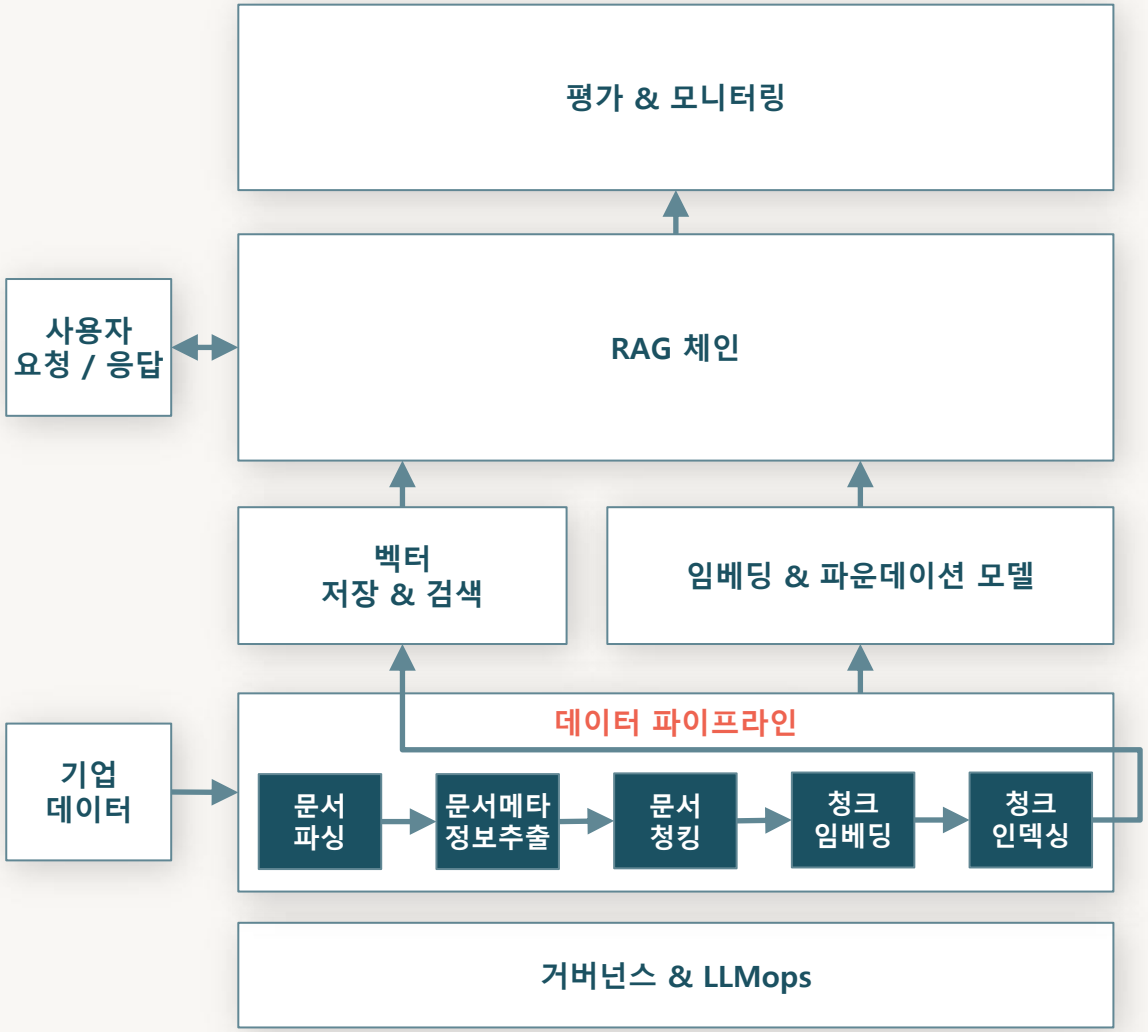
https://www.llamaindex.ai/blog/evaluating-the-ideal-chunk-size-for-a-rag-system-using-llamaindex-6207e5d3fec5

©2024 Databricks Inc. — All rights reserved

데이터 파이프라인

구조화되지 않은 소스 데이터를 사용자 질의와의 관련성에 따라 쿼리할 수 있는 개별 청크로 변환

- 문서 파싱: PDF 추출기나 OCR 기술등을 활용하여 비정형 문서에서 텍스트, 표 등을 추출
- 문서 메타정보 추출 : 목적에 따라 문서명, 페이지 번호, URL 또는 기타 정보와 같은 문서 메타데이터를 추출
- 문서 청킹 : 파싱된 문서가 임베딩 모델과 LLM의 컨텍스트 창에 들어갈 수 있도록 하기 위해 파싱된 문서를 더 작고 개별적인 청크로 구분
- 청크 임베딩 : 임베딩 모델을 사용하여 청크를 수치형 벡터의 목록으로 변환하여 문장내 각 토큰의 의미를 캡슐화
- 청크 인덱싱 : 벡터화된 청크 정보를 벡터 데이터를 저장하는 특수한 데이터베이스에 저장하여 사용자의 질의에 가장 유사한 청크 벡터를 찾고 원본 텍스트의 청크를 반환받을 수 있도록 색인



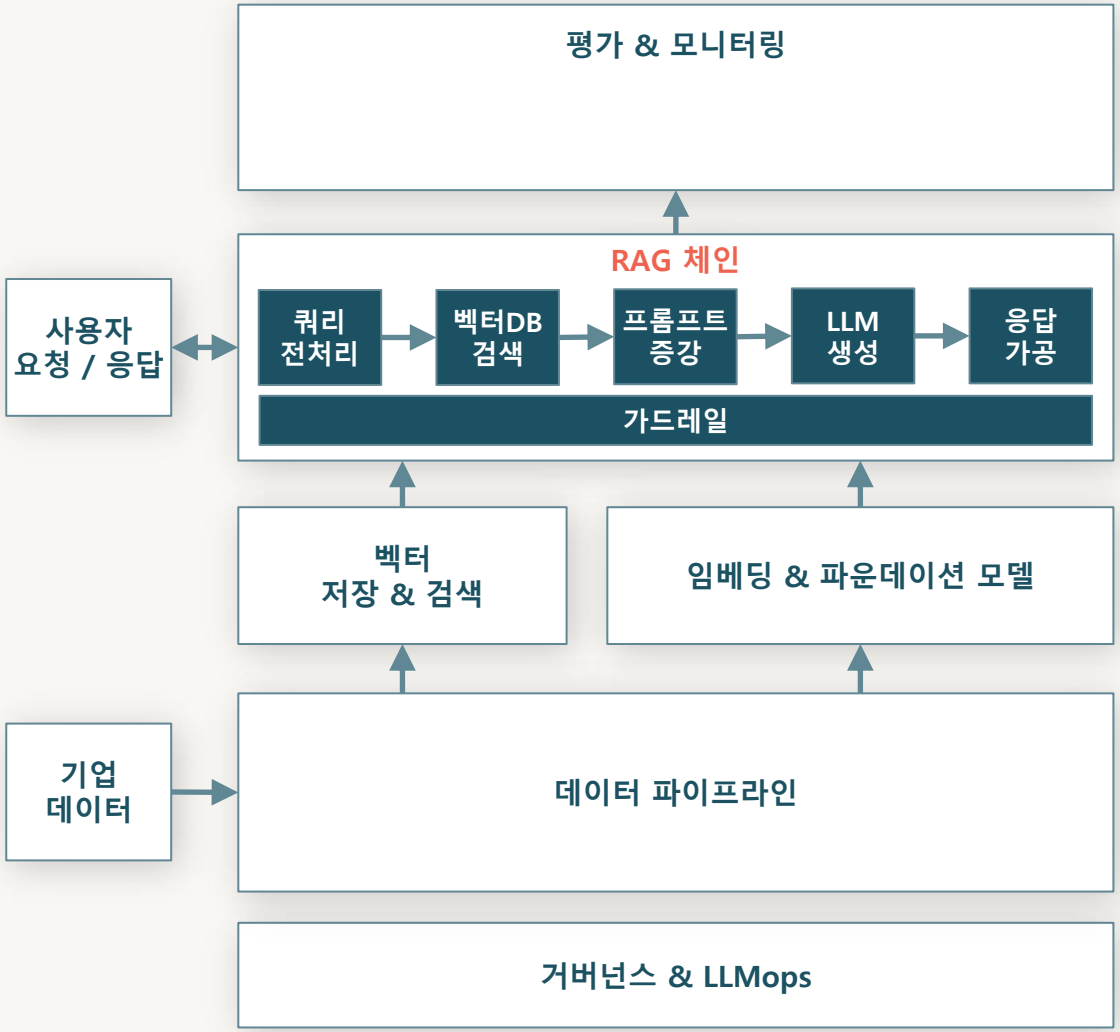
< RAG 시스템 구조 >

©2024 Databricks Inc. — All rights reserved

RAG 체인

RAG 애플리케이션의 프로세스

- 쿼리 전처리 : 쿼리 템플릿 작성, LLM을 사용하여 쿼리 재작성, 키워드 추출 등 벡터DB 쿼리에 더 적합하도록 사용자 쿼리를 전처리
- 벡터DB 검색 : 쿼리를 임베딩 후 벡터DB에서 의미론적 유사성이 높은 관련 정보를 검색
- 프롬프트 증강 : LLM에 전송되는 프롬프트는 벡터DB의 검색 결과에서 반환된 정보를 사용하여 사용자 쿼리를 증강하여 프롬프트 템플릿을 생성(프롬프트 엔지니어링)
- LLM 생성 : LLM은 사용자의 질의와 벡터DB에서 검색된 결과 데이터를 포함하는 증강된 프롬프트를 입력으로 받아 추가 컨텍스트에 기반한 자연어 응답을 생성
- 응답 가공 : LLM의 응답은 추가적인 비즈니스 로직을 적용하거나, 인용문을 추가하거나, 사전 정의된 규칙이나 제약 조건에 따라 생성된 텍스트를 정제하기 위해 추가로 사후 처리를 진행



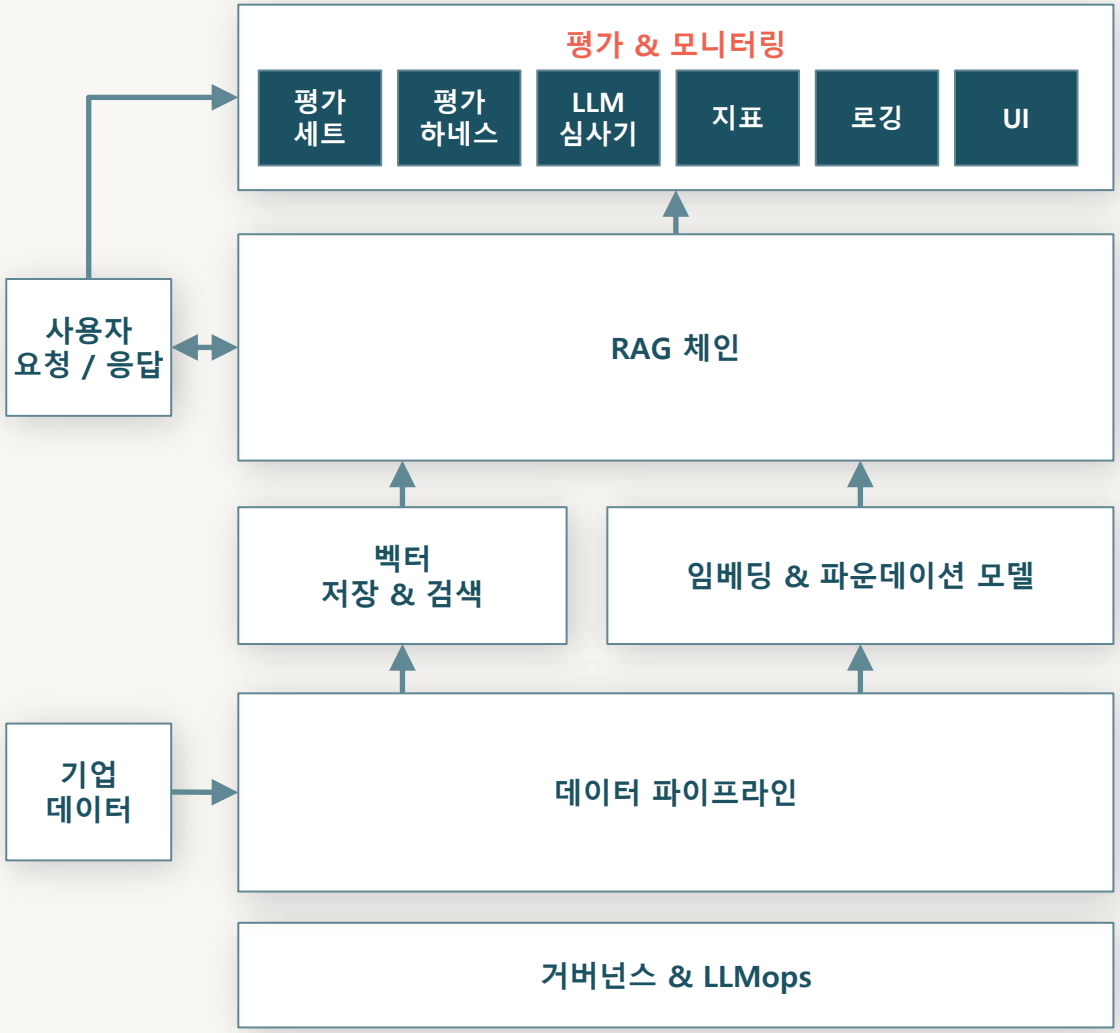
< RAG 시스템 구조 >

©2024 Databricks Inc. — All rights reserved

평가 및 모니터링

평가와 모니터링을 위한 구성요소

- 평가 세트 : RAG 애플리케이션을 엄격하게 평가하려면 애플리케이션의 의도된 용도를 대표하는 평가 쿼리 세트가 필요
- 지표 정의 : RAG 애플리케이션에 따라 중요한 측정항목에는 응답 정확도, 대기 시간, 비용 또는 주요 이해관계자의 평가가 포함될 수 있으며 각 구성 요소가 서로 상호 작용하는 방식 및 전체 시스템을 측정하는 측정 항목이 필요
- LLM 심사기(LLM Judges) : LLM 응답의 개방형 특성을 고려할 때 출력이 올바른지 판단하기 위한 수작업 평가로는 한계가 있기 때문에 다양한 LLM을 사용하여 평가 규모를 확장
- 평가 하네스(Evaluation Harness) : 평가를 효율화하기 위한 도구를 개발하여 평가를 자동화 및 병렬화 하여 대량의 평가 쿼리 세트를 짧은 시간내에 수행하고 지표를 관리
- 운영환경 추적 로깅 : 운영 환경의 대량 쿼리에 대한 모니터링
- 이해관계자 지향 UI(Stakeholder-facing UI) : 평가 도구만으로는 도메인 정확성을 판단하기 어렵기 때문에 도메인 전문가들이 직접 평가할 수 있도록 UI를 제공하여 피드백을 수집



< RAG 시스템 구조 >

©2024 Databricks Inc. — All rights reserved

평가 및 모니터링

RAG시스템의 평가방법

- RAG시스템은 전통적인 머신러닝 애플리케이션과 달리 품질에 영향을 미치는 구성요소가 복잡하게 구성되어있음
- 하나의 구성요소를 조절하면 연쇄적으로 영향을 미칠 수 있기 때문에 구성요소들을 개별적으로 평가 및 모니터링 할 필요가 있음

LlamaIndex사가 측정한 청크의 크기별 성능

구 분		전통적 머신러닝	생성형 AI
지표	구성요소 지표	- 컴포넌트 지표 : 입력값과 출력값에 대한 평가지표가 존재(Feature Drift, Precision/Recall, 응답시간 외 다수)	- 입력값과 출력값에 대한 평가지표가 존재(예. Precision@K, nDCG, toxicity, 응답시간 외 다수) - 추천 시스템을 평가하는 평가지표를 차용하는 경우가 많음
	복합지표	-	- GenAI내 Retriever, VectorDB, Generator등 여러 컴포넌트간의 입력값과 출력값이 얼마나 충실하게 정보를 생산하고 교환하여 상호작용 하는지를 평가(예. 충실도(faithfulness))
	전체지표	-	- 시스템의 전체 입력값 및 출력값을 평가(예. 답변 정확성, 응답시간)
평가		- 결정론적으로 '일치' 또는 '불일치'를 판단	- 비결정론적 '일치' 또는 '불일치'를 판단 (예. 10개중 3개는 매우일치, 3개는 일치, 4개는 불일치) - 더 높은 수준의 일치율을 확보하기 위해서는 인간의 피드백이 필요

©2024 Databricks Inc. — All rights reserved

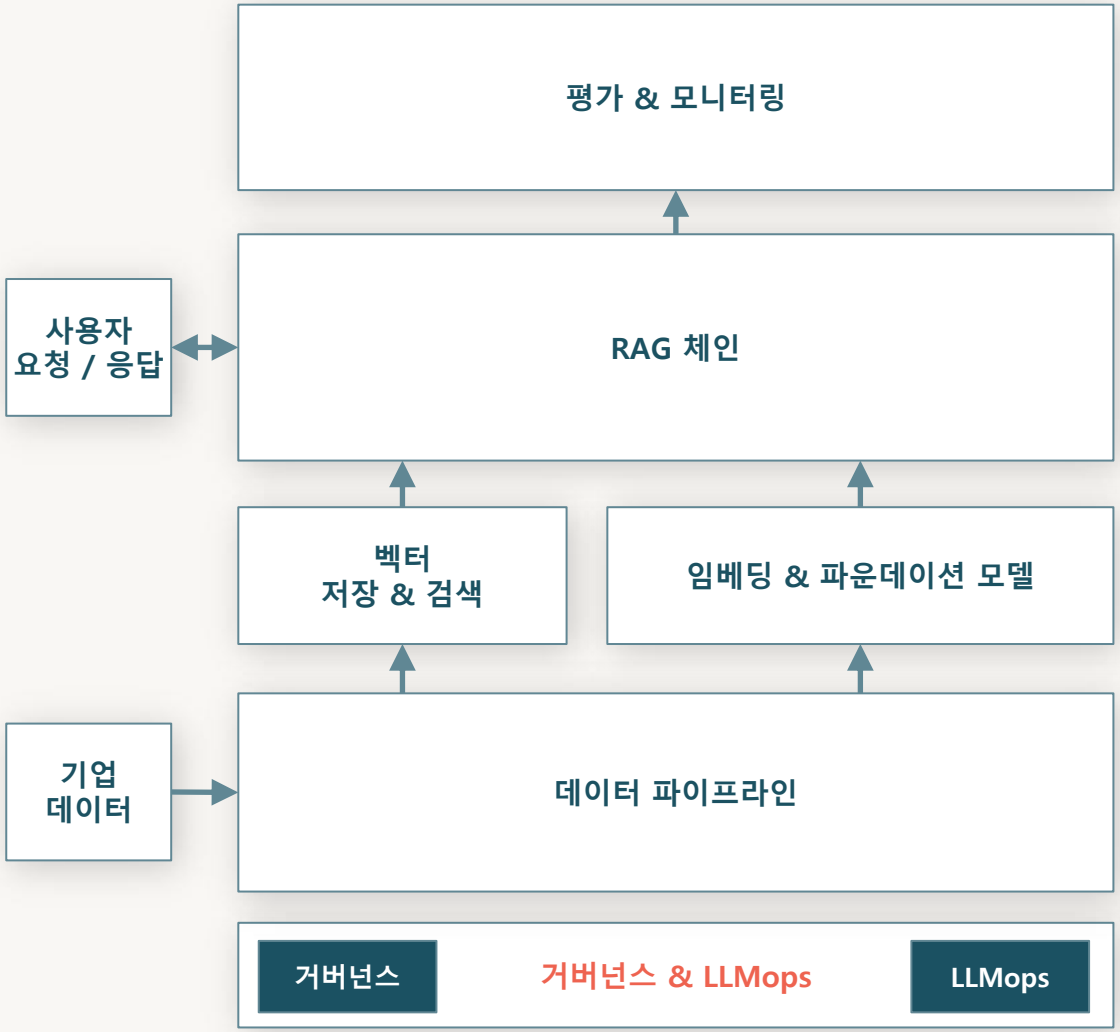
거버넌스와 LLMops

거버넌스

- 적절한 액세스 제어 및 로깅을 통해 엔터프라이즈 데이터와 AI 자산을 보호하려면 거버넌스가 필요

LLMops

- 생성형AI 시스템의 성능, 안정성 및 장기적인 효율성을 향상시키기 위해 데이터, 코드 및 모델을 관리하기 위한 프로세스 및 자동화 필요



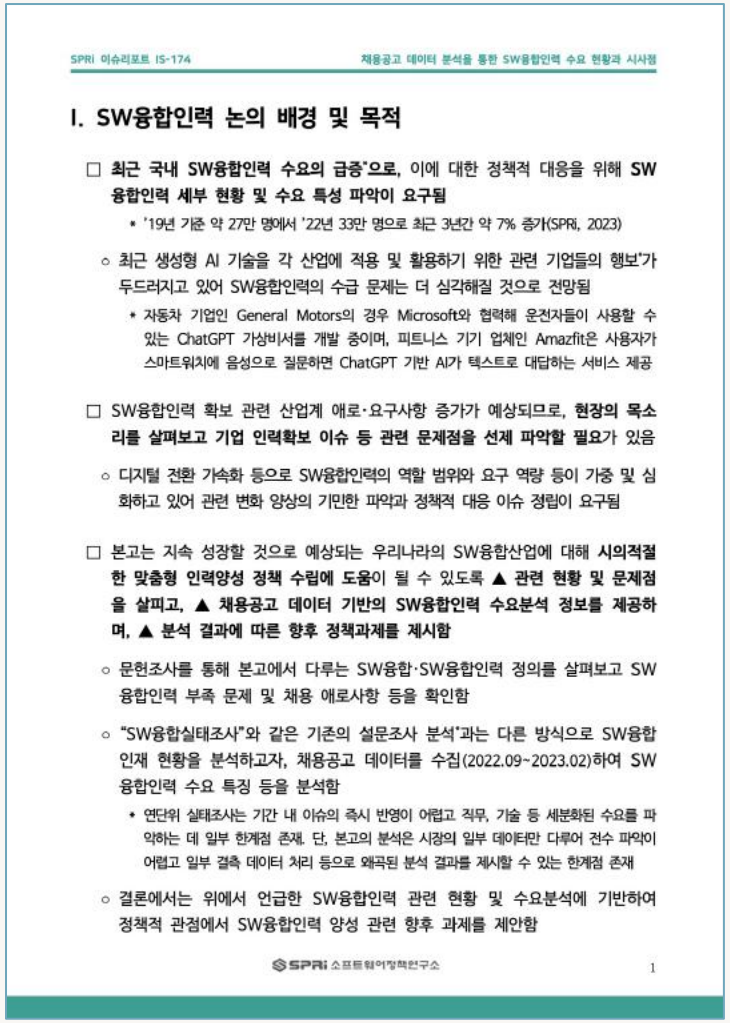
< RAG 시스템 구조 >

©2024 Databricks Inc. — All rights reserved

문서 데이터 전처리

©2024 Databricks Inc. — All rights reserved

문서 데이터 전처리



SPRi 이슈리포트 IS-174 채용공고 데이터 분석을 통한 SW융합인력 수요 현황과 시사점

I. SW융합인력 논의 배경 및 목적

□ **최근 국내 SW융합인력 수요의 급증*으로,** 이에 대한 정책적 대응을 위해 **SW융합인력 세부 현황 및 수요 특성 파악이 요구됨**

* '19년 기준 약 27만 명에서 '22년 33만 명으로 최근 3년간 약 7% 증가(SPRi, 2023)

○ 최근 생성형 AI 기술을 각 산업에 적용 및 활용하기 위한 관련 기업들의 행보*가 두드러지고 있어 SW융합인력의 수급 문제는 더 심각해질 것으로 전망됨

* 자동차 기업인 General Motors의 경우 Microsoft와 협력해 운전자들이 사용할 수 있는 ChatGPT 가상비서를 개발 중이며, 피트니스 기기 업체인 Amazfit은 사용자가 스마트워치에 음성으로 질문하면 ChatGPT 기반 AI가 텍스트로 대답하는 서비스 제공

□ SW융합인력 확보 관련 산업계 애로·요구사항 증가가 예상되므로, **현장의 목소리를 살펴보고 기업 인력확보 이슈 등 관련 문제점을 선제 파악할 필요**가 있음

○ 디지털 전환 가속화 등으로 SW융합인력의 역할 범위와 요구 역량 등이 가중 및 심화하고 있어 관련 변화 양상의 기민한 파악과 정책적 대응 이슈 정립이 요구됨

□ 본고는 지속 성장할 것으로 예상되는 우리나라의 SW융합산업에 대해 **시의적절한 맞춤형 인력양성 정책 수립에 도움**이 될 수 있도록 **▲ 관련 현황 및 문제점을 살피고, ▲ 채용공고 데이터 기반의 SW융합인력 수요분석 정보를 제공하며, ▲ 분석 결과에 따른 향후 정책과제를 제시함**

○ 문헌조사를 통해 본고에서 다루는 SW융합·SW융합인력 정의를 살펴보고 SW융합인력 부족 문제 및 채용 애로사항 등을 확인함

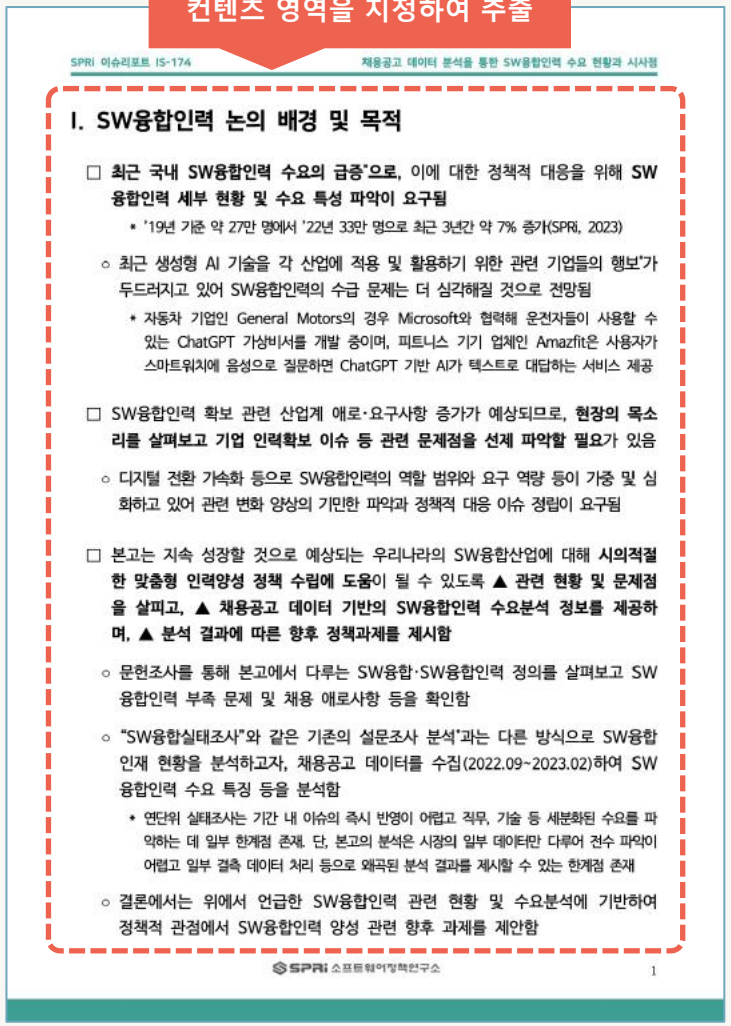
○ "SW융합실태조사"와 같은 기존의 설문조사 분석*과는 다른 방식으로 SW융합인재 현황을 분석하고자, 채용공고 데이터를 수집(2022.09~2023.02)하여 SW융합인력 수요 특징 등을 분석함

* 연단위 실태조사는 기간 내 이슈의 즉시 반영이 어렵고 직무, 기술 등 세분화된 수요를 파악하는 데 일부 한계점 존재. 단, 본고의 분석은 시장의 일부 데이터만 다루어 전수 파악이 어렵고 일부 결측 데이터 처리 등으로 왜곡된 분석 결과를 제시할 수 있는 한계점 존재

○ 결론에서는 위에서 언급한 SW융합인력 관련 현황 및 수요분석에 기반하여 정책적 관점에서 SW융합인력 양성 관련 향후 과제를 제안함

SPRi 소프트웨어정책연구소 1

컨텐츠 영역을 지정하여 추출



SPRi 이슈리포트 IS-174 채용공고 데이터 분석을 통한 SW융합인력 수요 현황과 시사점

I. SW융합인력 논의 배경 및 목적

□ **최근 국내 SW융합인력 수요의 급증*으로,** 이에 대한 정책적 대응을 위해 **SW융합인력 세부 현황 및 수요 특성 파악이 요구됨**

* '19년 기준 약 27만 명에서 '22년 33만 명으로 최근 3년간 약 7% 증가(SPRi, 2023)

○ 최근 생성형 AI 기술을 각 산업에 적용 및 활용하기 위한 관련 기업들의 행보*가 두드러지고 있어 SW융합인력의 수급 문제는 더 심각해질 것으로 전망됨

* 자동차 기업인 General Motors의 경우 Microsoft와 협력해 운전자들이 사용할 수 있는 ChatGPT 가상비서를 개발 중이며, 피트니스 기기 업체인 Amazfit은 사용자가 스마트워치에 음성으로 질문하면 ChatGPT 기반 AI가 텍스트로 대답하는 서비스 제공

□ SW융합인력 확보 관련 산업계 애로·요구사항 증가가 예상되므로, **현장의 목소리를 살펴보고 기업 인력확보 이슈 등 관련 문제점을 선제 파악할 필요**가 있음

○ 디지털 전환 가속화 등으로 SW융합인력의 역할 범위와 요구 역량 등이 가중 및 심화하고 있어 관련 변화 양상의 기민한 파악과 정책적 대응 이슈 정립이 요구됨

□ 본고는 지속 성장할 것으로 예상되는 우리나라의 SW융합산업에 대해 **시의적절한 맞춤형 인력양성 정책 수립에 도움**이 될 수 있도록 **▲ 관련 현황 및 문제점을 살피고, ▲ 채용공고 데이터 기반의 SW융합인력 수요분석 정보를 제공하며, ▲ 분석 결과에 따른 향후 정책과제를 제시함**

○ 문헌조사를 통해 본고에서 다루는 SW융합·SW융합인력 정의를 살펴보고 SW융합인력 부족 문제 및 채용 애로사항 등을 확인함

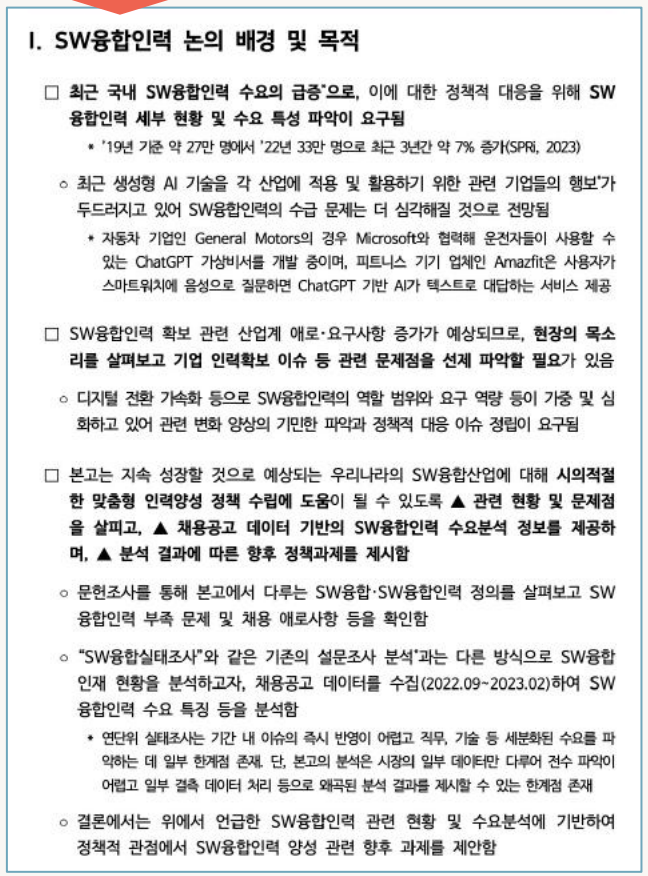
○ "SW융합실태조사"와 같은 기존의 설문조사 분석*과는 다른 방식으로 SW융합인재 현황을 분석하고자, 채용공고 데이터를 수집(2022.09~2023.02)하여 SW융합인력 수요 특징 등을 분석함

* 연단위 실태조사는 기간 내 이슈의 즉시 반영이 어렵고 직무, 기술 등 세분화된 수요를 파악하는 데 일부 한계점 존재. 단, 본고의 분석은 시장의 일부 데이터만 다루어 전수 파악이 어렵고 일부 결측 데이터 처리 등으로 왜곡된 분석 결과를 제시할 수 있는 한계점 존재

○ 결론에서는 위에서 언급한 SW융합인력 관련 현황 및 수요분석에 기반하여 정책적 관점에서 SW융합인력 양성 관련 향후 과제를 제안함

SPRi 소프트웨어정책연구소 1

문서의 머리말 꼬리말등 주요 컨텐츠 영역 외에 존재하는 내용 제거하고 컨텐츠 추출



I. SW융합인력 논의 배경 및 목적

□ **최근 국내 SW융합인력 수요의 급증*으로,** 이에 대한 정책적 대응을 위해 **SW융합인력 세부 현황 및 수요 특성 파악이 요구됨**

* '19년 기준 약 27만 명에서 '22년 33만 명으로 최근 3년간 약 7% 증가(SPRi, 2023)

○ 최근 생성형 AI 기술을 각 산업에 적용 및 활용하기 위한 관련 기업들의 행보*가 두드러지고 있어 SW융합인력의 수급 문제는 더 심각해질 것으로 전망됨

* 자동차 기업인 General Motors의 경우 Microsoft와 협력해 운전자들이 사용할 수 있는 ChatGPT 가상비서를 개발 중이며, 피트니스 기기 업체인 Amazfit은 사용자가 스마트워치에 음성으로 질문하면 ChatGPT 기반 AI가 텍스트로 대답하는 서비스 제공

□ SW융합인력 확보 관련 산업계 애로·요구사항 증가가 예상되므로, **현장의 목소리를 살펴보고 기업 인력확보 이슈 등 관련 문제점을 선제 파악할 필요**가 있음

○ 디지털 전환 가속화 등으로 SW융합인력의 역할 범위와 요구 역량 등이 가중 및 심화하고 있어 관련 변화 양상의 기민한 파악과 정책적 대응 이슈 정립이 요구됨

□ 본고는 지속 성장할 것으로 예상되는 우리나라의 SW융합산업에 대해 **시의적절한 맞춤형 인력양성 정책 수립에 도움**이 될 수 있도록 **▲ 관련 현황 및 문제점을 살피고, ▲ 채용공고 데이터 기반의 SW융합인력 수요분석 정보를 제공하며, ▲ 분석 결과에 따른 향후 정책과제를 제시함**

○ 문헌조사를 통해 본고에서 다루는 SW융합·SW융합인력 정의를 살펴보고 SW융합인력 부족 문제 및 채용 애로사항 등을 확인함

○ "SW융합실태조사"와 같은 기존의 설문조사 분석*과는 다른 방식으로 SW융합인재 현황을 분석하고자, 채용공고 데이터를 수집(2022.09~2023.02)하여 SW융합인력 수요 특징 등을 분석함

* 연단위 실태조사는 기간 내 이슈의 즉시 반영이 어렵고 직무, 기술 등 세분화된 수요를 파악하는 데 일부 한계점 존재. 단, 본고의 분석은 시장의 일부 데이터만 다루어 전수 파악이 어렵고 일부 결측 데이터 처리 등으로 왜곡된 분석 결과를 제시할 수 있는 한계점 존재

○ 결론에서는 위에서 언급한 SW융합인력 관련 현황 및 수요분석에 기반하여 정책적 관점에서 SW융합인력 양성 관련 향후 과제를 제안함

©2024 Databricks Inc. — All rights reserved

문서 데이터 전처리

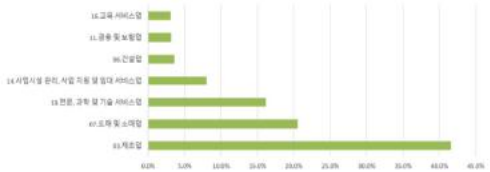
□ **(산업)** 채용공고 건수를 SW산업과 SW융합산업(SW산업 외)으로 구분하여 집계한 결과, **SW산업에서 44.5%, SW융합산업에서 45.5%**로 나타남

○ SW산업 채용공고 수는 소분류 기준 "소프트웨어 개발 및 공급업", "컴퓨터 프로그래밍, 시스템 통합 및 관리업", "기타 정보 서비스업[5]" 순으로, "기타 정보 서비스업"은 특히 수요가 늘어나는 특성을 보임

* 이하 " " 표시한 부분은 10차 한국표준산업분류 용어를 사용

○ SW융합인재 공고 건수는 제조(41.6%), 도소매(20.5%), 과학·기술 서비스(16.1%), 사업 관리·지원, 건설, 금융, 교육 등의 순으로 많았고, 제조, 도소매 산업이 대부분의 SW융합 산업 수요를 차지함

[그림 1] SW융합산업별 SW융합 인재 채용공고 비중



* 전체 SW융합 인력은 33.2만 명 중, 산업별 인력은 광업/제조업 31.0%, 도소매 5.4%, 과학·기술 서비스 15.7% 등임(SPRi, 2023), 한국표준산업분류(10차) 대분류 기준

○ "제조업" 중 전자·전기 관련 부문을 제외하면 의료, 자동차 순으로 SW융합인력 수요가 많음

- 의료 부문[6]은 헬스케어 IT시장, 특히 인공지능 서비스[7]가 빠르게 성장할

5) 세세분류 "데이터베이스 및 온라인 정보 제공업"은 소분류 "기타 정보 서비스업"에 포함되어 버킷플레이스(오늘의집), 아놀자, 티맵모빌리티 등 온라인 플랫폼, 혁토파이낸셜 등 간편결제 포함되어 수요가 눈에 띄게 늘어나는 것으로 분석됨

6) 헬당 변화 앱, 헬스케어 플랫폼, 병의원 의료 기기, 치과용 기기 등의 사업기업이 AI, 클라우드, 풀스택, ERP, MCU 등의 역량을 가진 개발자 채용하는 것으로 분석됨

7) 중증질환·소아 희귀질환 등 진단을 위한 의료 AI을 개발하고, 다양한 서비스 제공을 위해 2025년부터 건강정보 마이데이

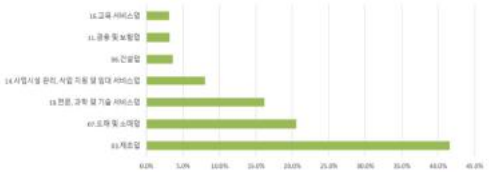
□ **(산업)** 채용공고 건수를 SW산업과 SW융합산업(SW산업 외)으로 구분하여 집계한 결과, **SW산업에서 44.5%, SW융합산업에서 45.5%**로 나타남

○ SW산업 채용공고 수는 소분류 기준 "소프트웨어 개발 및 공급업", "컴퓨터 프로그래밍, 시스템 통합 및 관리업", "기타 정보 서비스업[5]" 순으로, "기타 정보 서비스업"은 특히 수요가 늘어나는 특성을 보임

* 이하 " " 표시한 부분은 10차 한국표준산업분류 용어를 사용

○ SW융합인재 공고 건수는 제조(41.6%), 도소매(20.5%), 과학·기술 서비스(16.1%), 사업 관리·지원, 건설, 금융, 교육 등의 순으로 많았고, 제조, 도소매 산업이 대부분의 SW융합 산업 수요를 차지함

[그림 1] SW융합산업별 SW융합 인재 채용공고 비중



* 전체 SW융합 인력은 33.2만 명 중, 산업별 인력은 광업/제조업 31.0%, 도소매 5.4%, 과학·기술 서비스 15.7% 등임(SPRi, 2023), 한국표준산업분류(10차) 대분류 기준

○ "제조업" 중 SW융합인력 수요가

- 의료 부 르게 성장할

5) 세세분류 "데이터베이스 및 온라인 정보 제공업"은 소분류 "기타 정보 서비스업"에 포함되어 버킷플레이스(오늘의집), 아놀자, 티맵모빌리티 등 온라인 플랫폼, 혁토파이낸셜 등 간편결제 포함되어 수요가 눈에 띄게 늘어나는 것으로 분석됨

6) 헬당 변화 앱, 헬스케어 플랫폼, 병의원 의료 기기, 치과용 기기 등의 사업기업이 AI, 클라우드, 풀스택, ERP, MCU 등의 역량을 가진 개발자 채용하는 것으로 분석됨

7) 중증질환·소아 희귀질환 등 진단을 위한 의료 AI을 개발하고, 다양한 서비스 제공을 위해 2025년부터 건강정보 마이데이

본 실습에서는 각주의 내용이 문맥을 이해하는데 도움이 되지 않는다고 판단하여 삭제

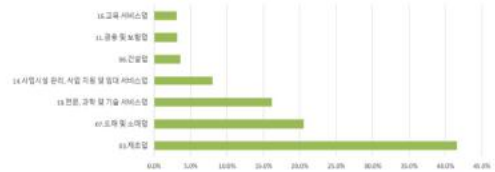
□ **(산업)** 채용공고 건수를 SW산업과 SW융합산업(SW산업 외)으로 구분하여 집계한 결과, **SW산업에서 44.5%, SW융합산업에서 45.5%**로 나타남

○ SW산업 채용공고 수는 소분류 기준 "소프트웨어 개발 및 공급업", "컴퓨터 프로그래밍, 시스템 통합 및 관리업", "기타 정보 서비스업[5]" 순으로, "기타 정보 서비스업"은 특히 수요가 늘어나는 특성을 보임

* 이하 " " 표시한 부분은 10차 한국표준산업분류 용어를 사용

○ SW융합인재 공고 건수는 제조(41.6%), 도소매(20.5%), 과학·기술 서비스(16.1%), 사업 관리·지원, 건설, 금융, 교육 등의 순으로 많았고, 제조, 도소매 산업이 대부분의 SW융합 산업 수요를 차지함

[그림 1] SW융합산업별 SW융합 인재 채용공고 비중



* 전체 SW융합 인력은 33.2만 명 중, 산업별 인력은 광업/제조업 31.0%, 도소매 5.4%, 과학·기술 서비스 15.7% 등임(SPRi, 2023), 한국표준산업분류(10차) 대분류 기준

○ "제조업" 중 전자·전기 관련 부문을 제외하면 의료, 자동차 순으로 SW융합인력 수요가 많음

- 의료 부문[6]은 헬스케어 IT시장, 특히 인공지능 서비스[7]가 빠르게 성장할

모든 페이지의 각주를 정규표현식으로 제거한 뒤 주요 컨텐츠만 추출

©2024 Databricks Inc. — All rights reserved

문서 데이터 전처리

Ⅱ. SW융합인력 현황 및 이슈

1. SW융합인력 정의

□ 본고에서 다루는 SW융합과 SW융합인력에 대한 정의는 「소프트웨어진흥법」제2조와 "SW융합실태조사"의 정의를 차용함

○ 「소프트웨어진흥법」 제2조 6에서는 소프트웨어융합에 대해 "SW와 다른 분야 간 기술 또는 서비스의 결합이나 복합을 통하여 새로운 사회적·시장적 가치를 창출하는 창의적이고 혁신적인 활동 및 현상"으로 정의

○ "SW융합실태조사"에서는 SW융합인력에 대해 **한국표준산업분류에서 SW산업을 제외한 산업에서 SW융합 활동이 있는 기업에 재직하는 SW인력**으로 정의(SPRi, 2022)

2. SW융합인력 현황

□ **국내 SW융합인력은 약 33만 명 수준**으로(SPRi, 2023), 향후 SW융합인력 규모는 **지속적으로 증가할 것으로 전망**됨

○ 국내 SW융합인력은 '19년 기준, 272,841명 → '22년 331,984명으로 최근 3년간 6.8% 증가한 바 있음

□ 코로나-19 이후 산업 전반의 디지털 전환이 가속화되면서, 신산업 분야뿐 아니라, **전통산업 분야에서도 SW인력의 수요가 급증**하고 있음

○ (자동차) 소프트웨어중심자동차(Software Defined Vehicle, SDV) 전환 기조로, 산업 전반에 SW전문인력 확보의 필요성이 급증하고 있으며(대한경제, 2023.02.14.), 해당 분야 인력 수요는 2030년까지 연평균 4% 증가할 것으로 예측됨(정보통신뉴스, 2022.03.13.)

○ (헬스케어) 산업의 높은 성장세로 인하여 디지털 헬스케어 인력이 향후 10년간 연평균 4.7% 증가할 것으로 전망되고 있음(한국산업기술진흥원, 2022)

- 디지털 헬스케어 기업 대상 설문조사에 따르면, 디지털 헬스케어 참여기업들은 우선적 채용할 전공분야로 "컴퓨터 활용 분야"를 1순위로 뽑은 비중이 약 38.5%로 가장 높았음(한국산업기술진흥원, 2022)

문서의 컨텐츠를 글머리를 기준으로 대/중/소분류로 구분

정규표현식을 통해 대/중/소분류를 마크다운 문서로 변환하여 추출

. SW융합인력 현황 및 이슈

SW융합인력 정의

본고에서 다루는 SW융합과 SW융합인력에 대한 정의는 「소프트웨어진흥법」제2조와 "SW융합실태조사"의 정의를 차용함

○ 「소프트웨어진흥법」 제2조 6에서는 소프트웨어융합에 대해 "SW와 다른 분야 간 기술 또는 서비스의 결합이나 복합을 통하여 새로운 사회적·시장적 가치를 창출하는 창의적이고 혁신적인 활동 및 현상"으로 정의

○ "SW융합실태조사"에서는 SW융합인력에 대해 **한국표준산업분류에서 SW산업을 제외한 산업에서 SW융합 활동이 있는 기업에 재직하는 SW인력**으로 정의(SPRi, 2022)

SW융합인력 현황

국내 SW융합인력은 약 33만 명 수준으로(SPRi, 2023), 향후 SW융합인력 규모는 지속적으로 증가할 것으로 전망됨

○ 국내 SW융합인력은 '19년 기준, 272,841명 → '22년 331,984명으로 최근 3년간 6.8% 증가한 바 있음

코로나-19 이후 산업 전반의 디지털 전환이 가속화되면서, 신산업 분야뿐 아니라, 전통산업 분야에서도 SW인력의 수요가 급증하고 있음

○ (자동차) 소프트웨어중심자동차(Software Defined Vehicle, SDV) 전환 기조로, 산업 전반에 SW전문인력 확보의 필요성이 급증하고 있으며(대한경제, 2023.02.14.), 해당 분야 인력 수요는 2030년까지 연평균 4% 증가할 것으로 예측됨(정보통신뉴스, 2022.03.13.)

○ (헬스케어) 산업의 높은 성장세로 인하여 디지털 헬스케어 인력이 향후 10년간 연평균 4.7% 증가할 것으로 전망되고 있음(한국산업기술진흥원, 2022)

- 디지털 헬스케어 기업 대상 설문조사에 따르면, 디지털 헬스케어 참여기업들은 우선적 채용할 전공분야로 "컴퓨터 활용 분야"를 1순위로 뽑은 비중이 약 38.5%로 가장 높았음(한국산업기술진흥원, 2022)

©2024 Databricks Inc. — All rights reserved

문서 데이터 전처리

. SW융합인력 현황 및 이슈

SW융합인력 정의

본고에서 다루는 SW융합과 SW융합인력에 대한 정의는 「소프트웨어진흥법」제 2조와 "SW융합실태조사"의 정의를 차용함

- 「소프트웨어진흥법」 제2조 6에서는 소프트웨어융합에 대해 "SW와 다른 분야 간 기술 또는 서비스의 결합이나 복합을 통하여 새로운 사회적·시장적 가치를 창출하는 창의적이고 혁신적인 활동 및 현상"으로 정의
- "SW융합실태조사"에서는 SW융합인력에 대해 **한국표준산업분류에서 SW산업을 제외한 산업에서 SW융합 활동이 있는 기업에 재직하는 SW인력**으로 정의(SPRi, 2022)

SW융합인력 현황

국내 SW융합인력은 약 33만 명 수준으로(SPRi, 2023), 향후 SW융합인력 규모는 **지속적으로 증가할 것으로 전망됨**

- 국내 SW융합인력은 '19년 기준, 272,841명 → '22년 331,984명으로 최근 3년간 6.8% 증가한 바 있음

코로나-19 이후 산업 전반의 디지털 전환이 가속화되면서, 신산업 분야뿐 아니라, **전통산업 분야에서도 SW인력의 수요가 급증**하고 있음

- (자동차) 소프트웨어중심자동차(Software Defined Vehicle, SDV) 전환 기조로, 산업 전반에 SW전문인력 확보의 필요성이 급증하고 있으며(대한경제, 2023.02.14.), 해당 분야 인력 수요는 2030년까지 연평균 4% 증가할 것으로 예측됨(정보통신뉴스, 2022.03.13.)
- (헬스케어) 산업의 높은 성장세로 인하여 디지털 헬스케어 인력이 향후 10년간 연평균 4.7% 증가할 것으로 전망되고 있음(한국산업기술진흥원, 2022)
  - 디지털 헬스케어 기업 대상 설문조사에 따르면, 디지털 헬스케어 참여기업들은 우선적 채용할 전공분야로 "컴퓨터 활용 분야"를 1순위로 뽑은 비중이 약 38.5%로 가장 높았음(한국산업기술진흥원, 2022)

괄호안에 포함되어 있는 인용구문은 문맥을 이해하는데 도움이 되지 않는다고 판단하여 삭제

. SW융합인력 현황 및 이슈

SW융합인력 정의

본고에서 다루는 SW융합과 SW융합인력에 대한 정의는 「소프트웨어진흥법」제 2조와 "SW융합실태조사"의 정의를 차용함

- 「소프트웨어진흥법」 제2조 6에서는 소프트웨어융합에 대해 "SW와 다른 분야 간 기술 또는 서비스의 결합이나 복합을 통하여 새로운 사회적·시장적 가치를 창출하는 창의적이고 혁신적인 활동 및 현상"으로 정의
- "SW융합실태조사"에서는 SW융합인력에 대해 **한국표준산업분류에서 SW산업을 제외한 산업에서 SW융합 활동이 있는 기업에 재직하는 SW인력**으로 정의(SPRi, 2022)

SW융합인력 현황

국내 SW융합인력은 약 33만 명 수준으로~~(SPRi, 2023)~~, 향후 SW융합인력 규모는 **지속적으로 증가할 것으로 전망됨**

- 국내 SW융합인력은 '19년 기준, 272,841명 → '22년 331,984명으로 최근 3년간 6.8% 증가한 바 있음

코로나-19 이후 산업 전반의 디지털 전환이 가속화되면서, 신산업 분야뿐 아니라, **전통산업 분야에서도 SW인력의 수요가 급증**하고 있음

- (자동차) 소프트웨어중심자동차(Software Defined Vehicle, SDV) 전환 기조로, 산업 전반에 SW전문인력 확보의 필요성이 급증하고 있으며~~(대한경제, 2023.02.14.)~~, 해당 분야 인력 수요는 2030년까지 연평균 4% 증가할 것으로 예측됨~~(정보통신뉴스, 2022.03.13.)~~
- (헬스케어) 산업의 높은 성장세로 인하여 디지털 헬스케어 인력이 향후 10년간 연평균 4.7% 증가할 것으로 전망되고 있음~~(한국산업기술진흥원, 2022)~~
  - 디지털 헬스케어 기업 대상 설문조사에 따르면, 디지털 헬스케어 참여기업들은 우선적 채용할 전공분야로 "컴퓨터 활용 분야"를 1순위로 뽑은 비중이 약 38.5%로 가장 높았음~~(한국산업기술진흥원, 2022)~~

괄호안의 인용구문을 모두 삭제 후 추출

. SW융합인력 현황 및 이슈

SW융합인력 정의

본고에서 다루는 SW융합과 SW융합인력에 대한 정의는 「소프트웨어진흥법」제 2조와 "SW융합실태조사"의 정의를 차용함

- 「소프트웨어진흥법」 제2조 6에서는 소프트웨어융합에 대해 "SW와 다른 분야 간 기술 또는 서비스의 결합이나 복합을 통하여 새로운 사회적·시장적 가치를 창출하는 창의적이고 혁신적인 활동 및 현상"으로 정의
- "SW융합실태조사"에서는 SW융합인력에 대해 **한국표준산업분류에서 SW산업을 제외한 산업에서 SW융합 활동이 있는 기업에 재직하는 SW인력**으로 정의(SPRi, 2022)

SW융합인력 현황

국내 SW융합인력은 약 33만 명 수준으로 향후 SW융합인력 규모는 **지속적으로 증가할 것으로 전망됨**

- 국내 SW융합인력은 '19년 기준, 272,841명 → '22년 331,984명으로 최근 3년간 6.8% 증가한 바 있음

코로나-19 이후 산업 전반의 디지털 전환이 가속화되면서, 신산업 분야뿐 아니라, **전통산업 분야에서도 SW인력의 수요가 급증**하고 있음

- (자동차) 소프트웨어중심자동차(Software Defined Vehicle, SDV) 전환 기조로, 산업 전반에 SW전문인력 확보의 필요성이 급증하고 있으며 , 해당 분야 인력 수요는 2030년까지 연평균 4% 증가할 것으로 예측됨
- (헬스케어) 산업의 높은 성장세로 인하여 디지털 헬스케어 인력이 향후 10년간 연평균 4.7% 증가할 것으로 전망되고 있음
  - 디지털 헬스케어 기업 대상 설문조사에 따르면, 디지털 헬스케어 참여기업들은 우선적 채용할 전공분야로 "컴퓨터 활용 분야"를 1순위로 뽑은 비중이 약 38.5%로 가장 높았음

©2024 Databricks Inc. — All rights reserved

문서 데이터 전처리

[표 2] 산업별 기술 스택[16] 수요 Top 7

순위	SW융합산업	SW산업
1	프로그램 언어	프로그램 언어
2	DB 설계, 개발, 관리	DB 설계, 개발, 관리
3	운영체계	운영체계
4	ERP (전사적 자원 관리)	클라우드 컴퓨팅(AWS)
5	펨웨어	개발플랫폼(Node.js, Docker, Spring Boot)
6	클라우드 컴퓨팅(AWS)	딥러닝
7	품질관리	품질관리

○ 학력별로도 프로그램 언어 수요가 많았으며, 학력별로 요구하는 프로그램 언어의 종류가 수요의 수가 다름

[표 3] 학력별 기술 스택 수요 Top 10

순위	고등학교 졸업	전문대학교 졸업	대학교 졸업	석사	박사
1	프로그램언어(C#/C++/JavaScript/C/Python/Java)	프로그램언어(Java/C++/C#/JavaScript/Python)	프로그램언어(C++/Java/Python/C/C#/JavaScript)	프로그램언어(C++/C/MatLab/TensorFloiw/R)	프로그램언어(C++/Java/Python/C/C#/TensorFloiw)
2	데이터(Oracle)	운영체제 운영 관련 기술	운영체제 운영 관련 기술	인공지능	인공지능
3	운영체제 운영 관련 기술	Firmware	Firmware	운영체제 운영 관련 기술	회로 설계(PCB)
4	전사적 자원관리	전사적 자원관리	전사적 자원관리	Firmware	운영체제(RTOS)
5	품질관리(QA)	데이터(Oracle)	데이터(Oracle)	MCU	전사적 자원관리
6	회로설계(PADS)	MCU	MCU	FPGA	Firmware
7	Firmware	품질관리(QA)	인공지능	전사적 자원관리	품질관리(QA)
8	생산관리시스템	회로 설계(PCB)	FPGA	데이터(MySQL)	무선통신 설계(RF)
9	MCU	MCU	MCU	품질관리(QA)	MCU
10	인공지능	인공지능	회로 설계(PCB)	회로 설계(PCB)	FPGA

※ 데이터 (Oracle/MSSQL/SQL/MySQL), 운영체제 (Linux, Embedded Linux, UNIX, iOS), 전사적 자원관리 (ERP), 생산관리시스템 (MES), MCU (Microcontroller Unit), 인공지능 (DeepLearning), FPGA (field- programmable gate array)

- 인공지능 분야는 딥러닝에 대한 석·박사 수요가 많으며, 데이터 역량은 상업적 데이터 관리 도구 사용이 가능한 대졸 이하 수요가 많고, 기업 및 공장 관리를 위한 도구(ERP, MES 등)는 전 학력에서 고르게 수요가 있음

불필요한 참고문구를 제거하고 표를 식별

[표 2] 산업별 기술 스택[16] 수요 Top 7

순위	SW융합산업	SW산업
1	프로그램 언어	프로그램 언어
2	DB 설계, 개발, 관리	DB 설계, 개발, 관리
3	운영체계	운영체계
4	ERP (전사적 자원 관리)	클라우드 컴퓨팅(AWS)
5	펨웨어	개발플랫폼(Node.js, Docker, Spring Boot)
6	클라우드 컴퓨팅(AWS)	딥러닝
7	품질관리	품질관리

○ 학력별로도 프로그램 언어 수요가 많았으며, 학력별로 요구하는 프로그램 언어의 종류가 수요의 수가 다름

[표 3] 학력별 기술 스택 수요 Top 10

순위	고등학교 졸업	전문대학교 졸업	대학교 졸업	석사	박사
1	프로그램언어(C#/C++/JavaScript/C/Python/Java)	프로그램언어(Java/C++/C#/JavaScript/Python)	프로그램언어(C++/Java/Python/C/C#/JavaScript)	프로그램언어(C++/C/MatLab/TensorFloiw/R)	프로그램언어(C++/Java/Python/C/C#/TensorFloiw)
2	데이터(Oracle)	운영체제 운영 관련 기술	운영체제 운영 관련 기술	인공지능	인공지능
3	운영체제 운영 관련 기술	Firmware	Firmware	운영체제 운영 관련 기술	회로 설계(PCB)
4	전사적 자원관리	전사적 자원관리	전사적 자원관리	Firmware	운영체제(RTOS)
5	품질관리(QA)	데이터(Oracle)	데이터(Oracle)	MCU	전사적 자원관리
6	회로설계(PADS)	MCU	MCU	FPGA	Firmware
7	Firmware	품질관리(QA)	인공지능	전사적 자원관리	품질관리(QA)
8	생산관리시스템	회로 설계(PCB)	FPGA	데이터(MySQL)	무선통신 설계(RF)
9	MCU	MCU	MCU	품질관리(QA)	MCU
10	인공지능	인공지능	회로 설계(PCB)	회로 설계(PCB)	FPGA

~~※ 데이터 (Oracle/MSSQL/SQL/MySQL), 운영체제 (Linux, Embedded Linux, UNIX, iOS), 전사적 자원관리 (ERP), 생산관리시스템 (MES), MCU (Microcontroller Unit), 인공지능 (DeepLearning), FPGA (field- programmable gate array)~~

- 인공지능 분야는 딥러닝에 대한 석·박사 수요가 많으며, 데이터 역량은 상업적 데이터 관리 도구 사용이 가능한 대졸 이하 수요가 많고, 기업 및 공장 관리를 위한 도구(ERP, MES 등)는 전 학력에서 고르게 수요가 있음

식별된 표를 마크다운 문서로 변환하여 추출

표 2 산업별 기술 스택 수요 Top 7

순위	SW융합산업	SW산업
1	프로그램 언어	프로그램 언어
2	DB 설계, 개발, 관리	DB 설계, 개발, 관리
3	운영체계	운영체계
4	ERP (전사적 자원 관리)	클라우드 컴퓨팅(AWS)
5	펨웨어	개발플랫폼
6	클라우드 컴퓨팅(AWS)	딥러닝
7	품질관리	품질관리

- 학력별로도 프로그램 언어 수요가 많았으며, 학력별로 요구하는 프로그램 언어의 종류가 수요의 수가 다름

표 3 학력별 기술 스택 수요 Top 10

순위	고등학교 졸업	전문대학교 졸업	대학교 졸업	석사	박사
1	프로그램언어(C# /C++/JavaScript /C/Pyth [illegible] 로그램언어(Java/C++/C#/Jav aScript/Python)	프로그램언어(C+ +/Java/Python/C /C#/JavaScript)	프로그램언어(C ++/C/MatLab/T ensorFloiw/R)	프로그램언어(C+ +/Java/Python/C/ C#/TensorFloiw)
2	데이터(Oracle)	운영체제 운영 관련 기술	운영체제 운영 관련 기술	인공지능	인공지능
3	운영체제 운영 관련 기술	Firmware	Firmware	운영체제 운영 관련 기술	회로 설계(PCB)
4	전사적 자원관리	전사적 자원관리	전사적 자원관리	Firmware	운영체제(RTOS)
5	품질관리(QA)	데이터(Oracle)	데이터(Oracle)	MCU	전사적 자원관리
6	회로설계(PADS)	MCU	MCU	FPGA	Firmware
7	Firmware	품질관리(QA)	인공지능	전사적 자원관리	품질관리(QA)
8	생산관리시스템	회로 설계(PCB)	FPGA	데이터(MySQL)	[illegible]
9	MCU	MCU	MCU	품질관리(QA)	MCU
10	인공지능	인공지능	회로 설계(PCB)	회로 설계(PCB)	FPGA

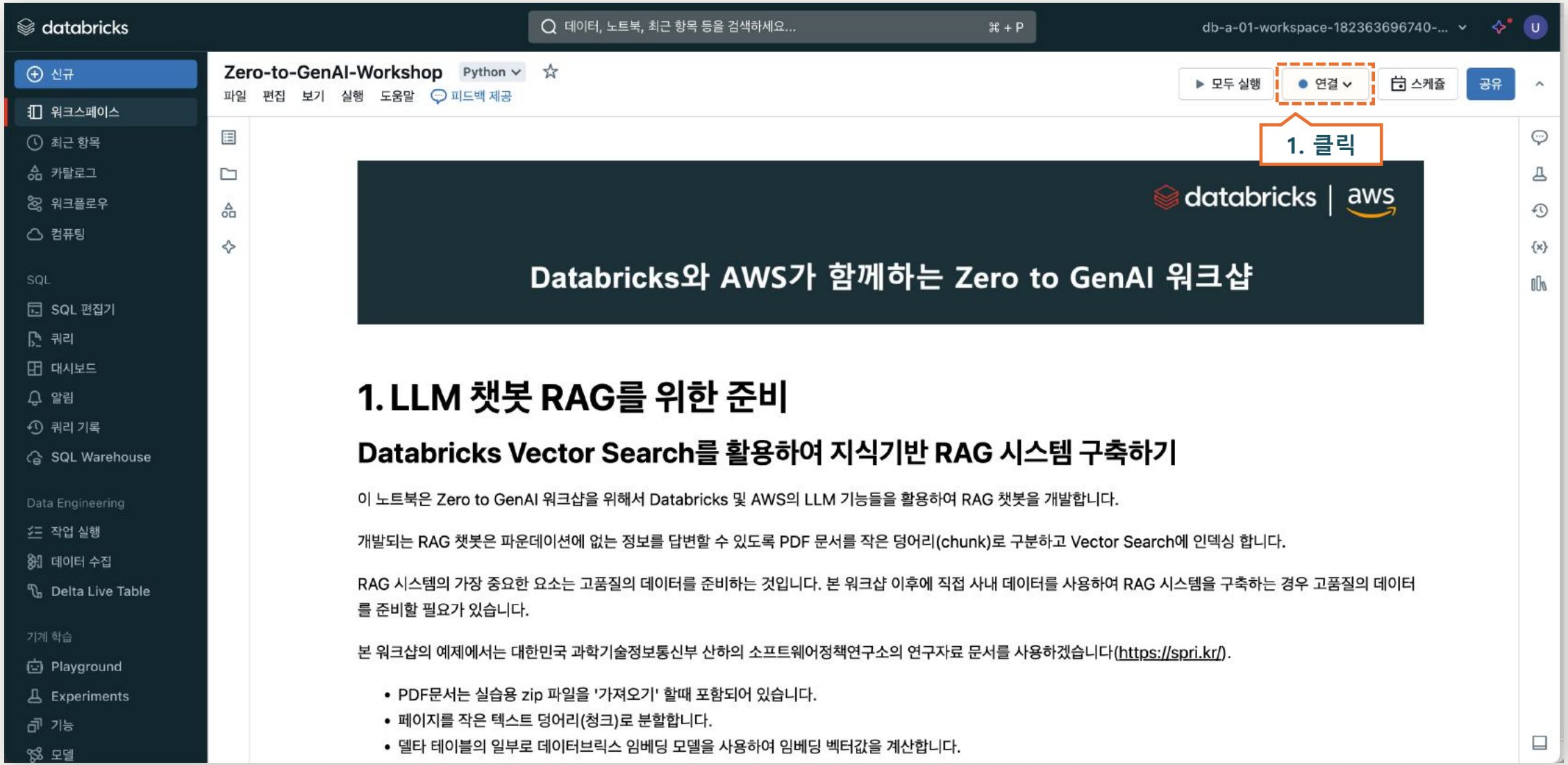
인공지능 분야는 딥러닝에 대한 석·박사 수요가 많으며, 데이터 역량은 상업적 데이터 관리 도구 사용이 가능한 대졸 이하 수요가 많고, 기업 및 공장 관리를 위한 도구는 전 학력에서 고르게 수요가 있음

©2024 Databricks Inc. — All rights reserved

실 습 4 :
실습용 노트북 실행하기

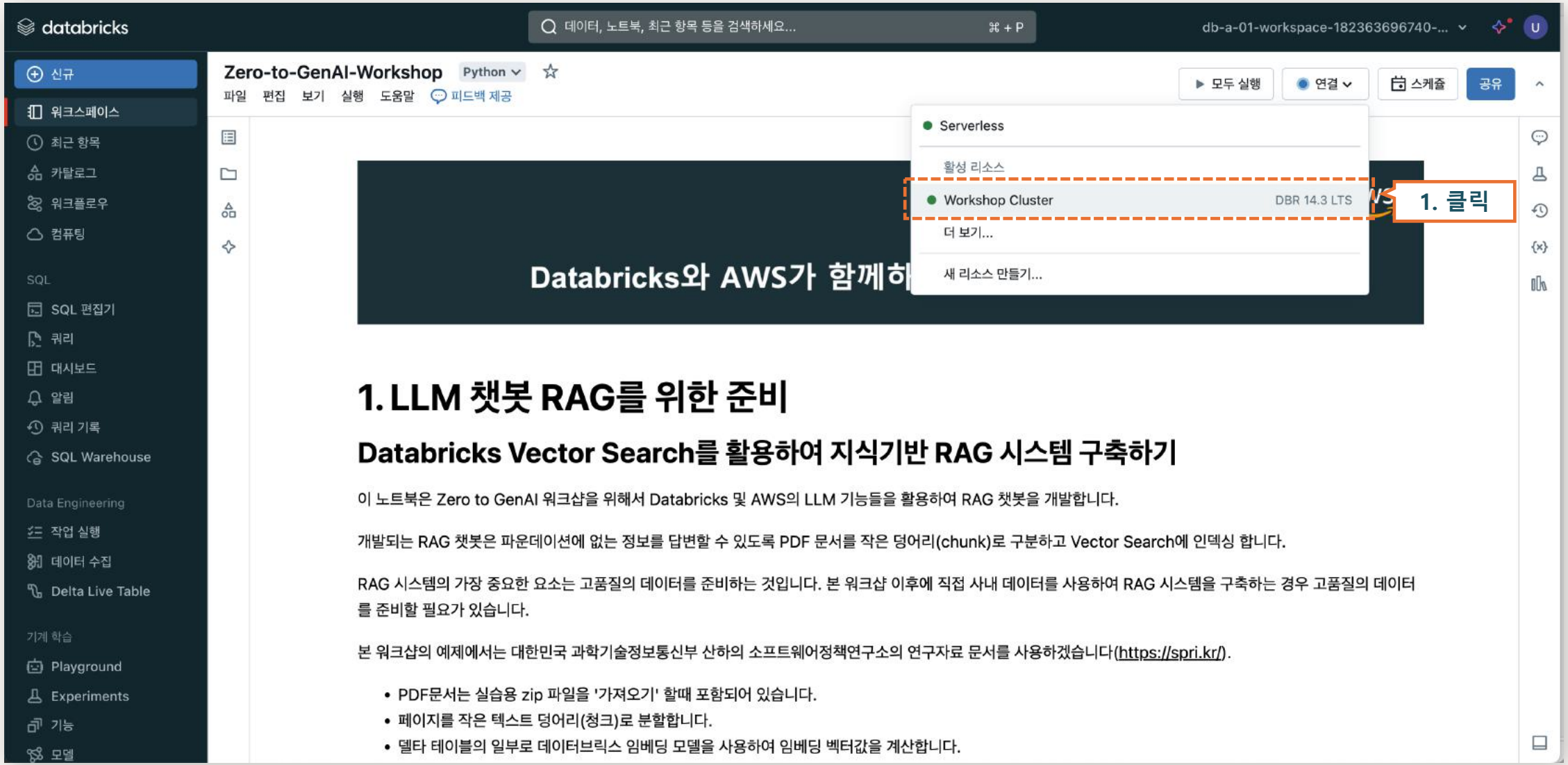
©2024 Databricks Inc. — All rights reserved

실습용 노트북 실행하기



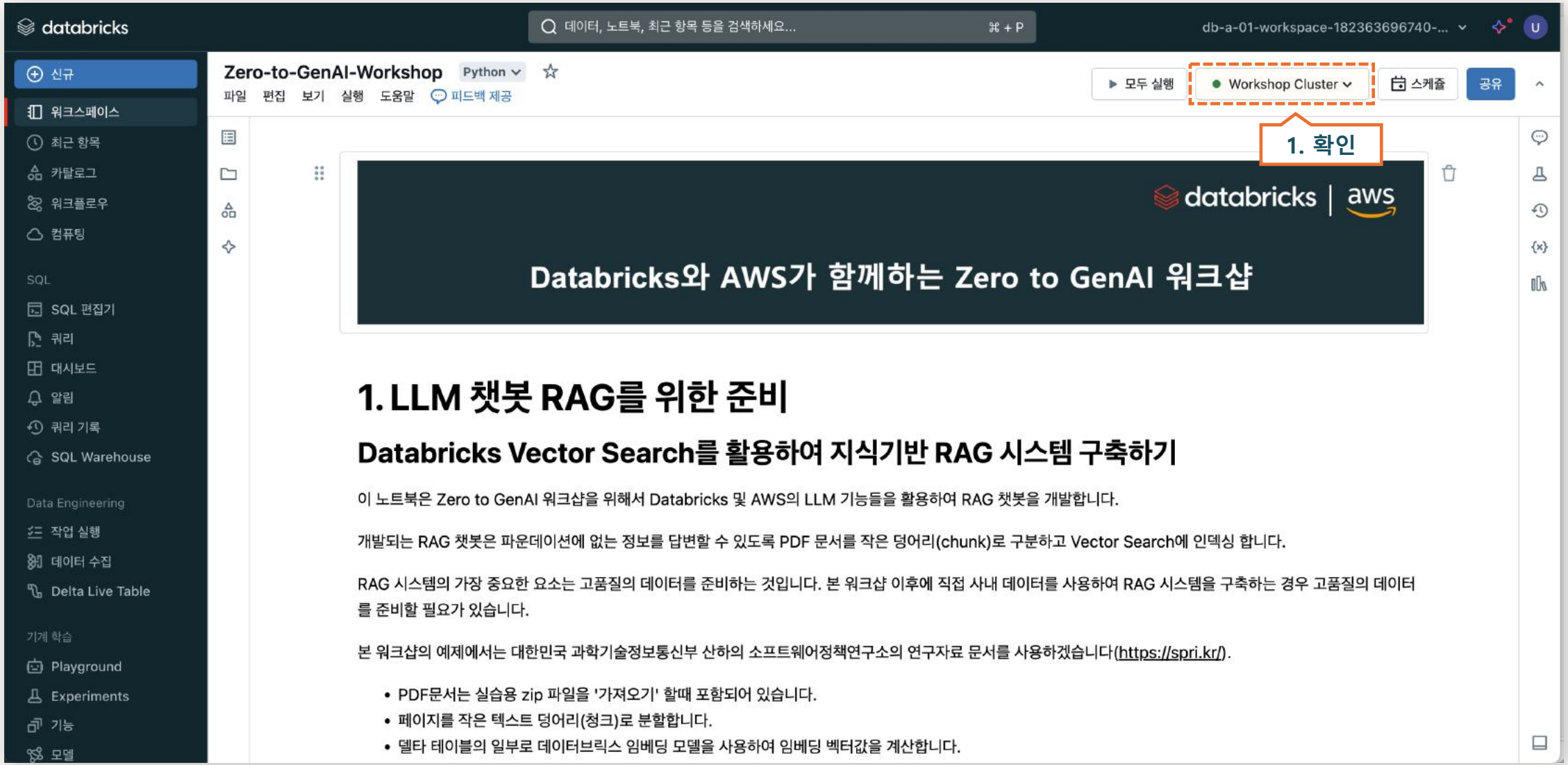
©2024 Databricks Inc. — All rights reserved

실습용 노트북 실행하기



©2024 Databricks Inc. — All rights reserved

실습용 노트북 실행하기



©2024 Databricks Inc. — All rights reserved

실습용 노트북 실행하기

databricks

데이터, 노트북, 최근 항목 등을 검색하세요... ⌘ + P

db-a-01-workspace-182363696740-...

신규
워크스페이스
최근 항목
카탈로그
워크플로우
컴퓨팅

SQL
SQL 편집기
쿼리
대시보드
알림
쿼리 기록
SQL Warehouse

Data Engineering
작업 실행
데이터 수집
Delta Live Table

기계 학습
Playground
Experiments
기능
모델

Zero-to-GenAI-Workshop Python

파일 편집 보기 실행 도움말 피드백 제공

모두 실행 | Workshop Cluster | 스케줄 | 공유

1. 클릭

6: RAG 구현이 필요로 되는 Python 패키지 설치 Python

```
!pip3 install -qqqq --upgrade pip
!pip3 install -qqqq mlflow==2.10.1 lxml==4.9.3 transformers==4.34.0 langchain==0.1.20 beautifulsoup4==4.12.2 pymupdf4llm==0.0.10
aiohttp==3.10.0 gradio==4.0.0 #3.50.2
!pip3 install -qqqq dbtunnel[gradio] databricks-vectorsearch==0.22 databricks-sdk databricks databricks-genai-inference
dbutils.library.restartPython()
```

+ 코드 + 텍스트

👇 사용자의 개별 환경에 따라 아래 정보를 업데이트 해주세요. 👇

✓ 22시간 전 (<1초) 8: 개별 환경 변수 설정

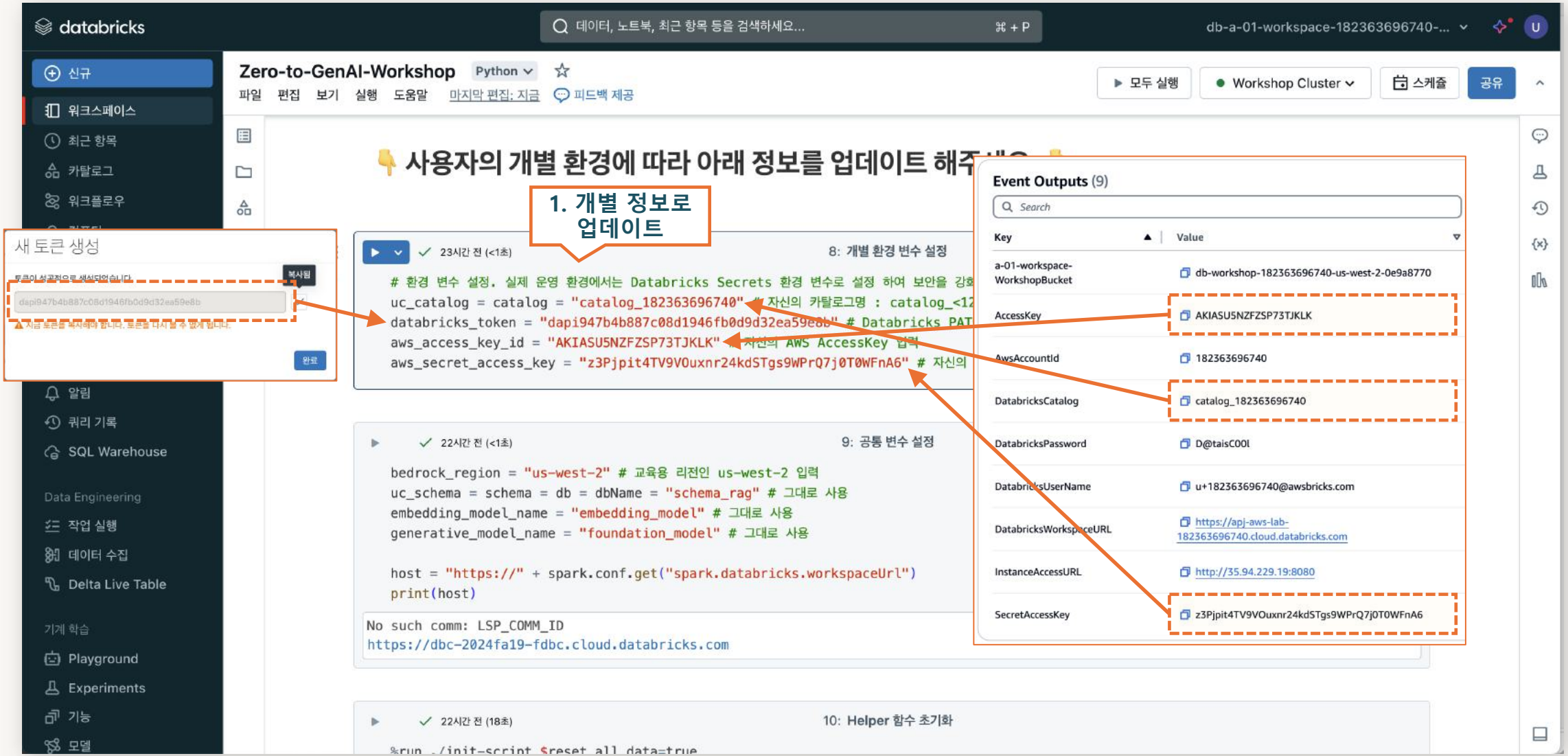
```
# 환경 변수 설정. 실제 운영 환경에서는 Databricks Secrets 환경 변수로 설정 하여 보안을 강화해야합니다.
uc_catalog = catalog = "<자신의 카탈로그명>" # 자신의 카탈로그명 : catalog_<12자리_AWS_Account_ID>
databricks_token = "<Databricks 토큰>" # Databricks PAT에서 발급받은 dapi로 시작하는 토큰
aws_access_key_id = "<자신의 AWS AccessKey>" # 자신의 AWS AccessKey 입력
aws_secret_access_key = "<자신의 AWS SecretAccessKey>" # 자신의 AWS SecretAccessKey 입력
```

✓ 22시간 전 (<1초) 9: 공통 변수 설정

```
bedrock_region = "us-west-2" # 교육용 리전인 us-west-2 입력
uc_schema = schema = db = dbName = "schema_rag" # 그대로 사용
embedding_model_name = "embedding_model" # 그대로 사용
generative_model_name = "foundation_model" # 그대로 사용
```

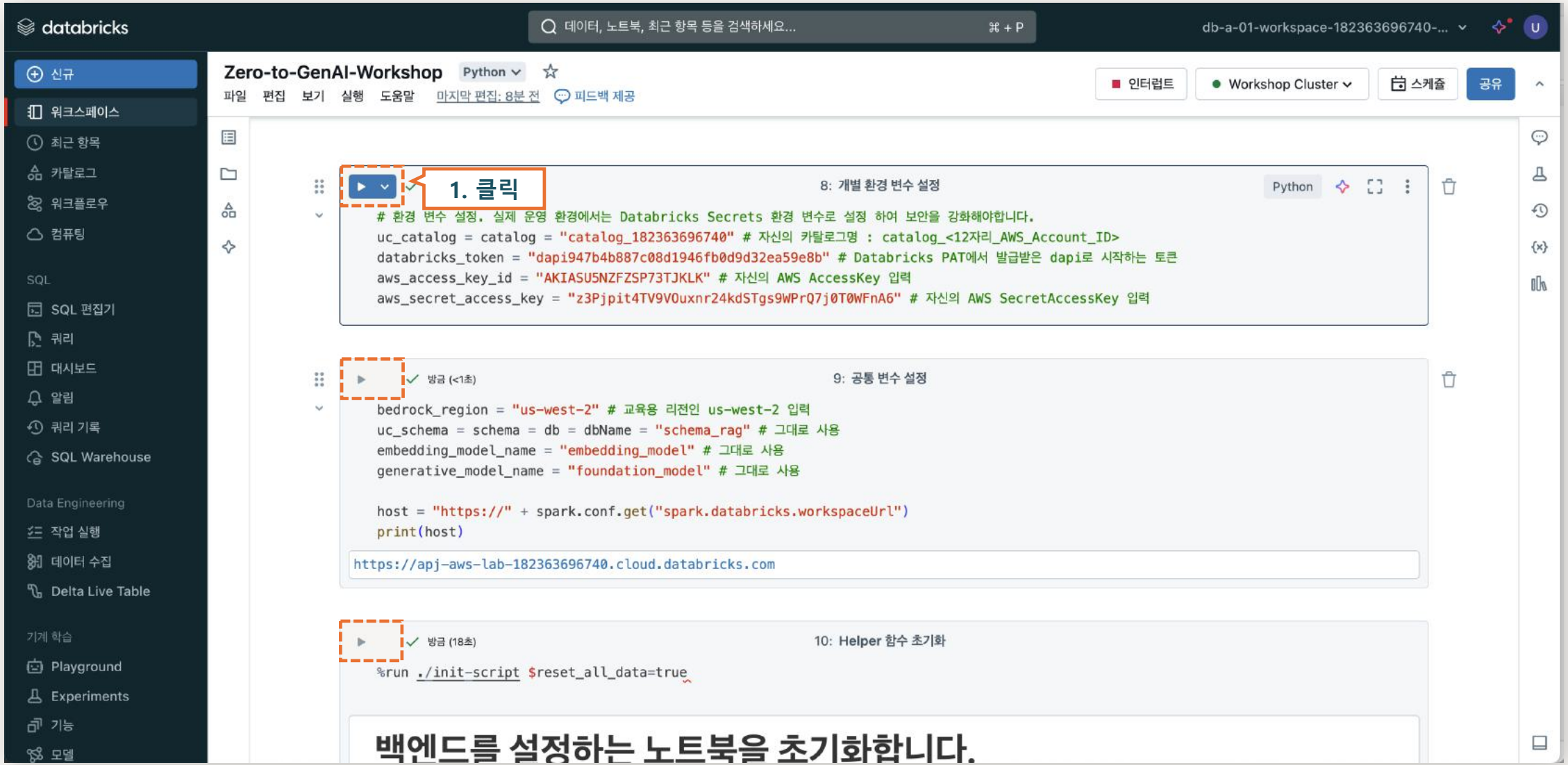
©2024 Databricks Inc. — All rights reserved

실습용 노트북 실행하기



©2024 Databricks Inc. — All rights reserved

실습용 노트북 실행하기



©2024 Databricks Inc. — All rights reserved

실습용 노트북 실행하기

databricks

데이터, 노트북, 최근 항목 등을 검색하세요... ⌘ + P

db-a-01-workspace-182363696740-...

신규
워크스페이스
최근 항목
카탈로그
워크플로우
컴퓨팅

SQL
SQL 편집기
쿼리
대시보드
알림
쿼리 기록
SQL Warehouse

Data Engineering
작업 실행
데이터 수집
Delta Live Table

기계 학습
Playground
Experiments
기능
모델

Zero-to-GenAI-Workshop Python

파일 편집 보기 실행 도움말 마지막 편집: 21분 전 피드백 제공

인터럽트 Workshop Cluster 스케줄 공유

1. 클릭

19: 생성 파운데이션 모델 엔드포인트 생성

19번 셀까지 실행

```
import mlflow.deployments
deploy_client = mlflow.deployments.get_deploy_client("databricks")

if generative_model_name in [endpoints['name'] for endpoints in deploy_client.list_endpoints()]:
    #deploy_client.delete_endpoint(generative_model_name)
    print("동일한 모델의 엔드포인트가 이미 존재합니다.")

# claude-3-5-sonnet-20240620-v1:0, claude-3-sonnet-20240229-v1:0
if generative_model_name not in [endpoints['name'] for endpoints in deploy_client.list_endpoints()]:
    deploy_client.create_endpoint(
        name=generative_model_name,
        config={
            "served_entities": [
                {
                    "external_model": {
                        "name": "claude-3-sonnet-20240229-v1:0",
                        "provider": "amazon-bedrock",
                        "task": "llm/v1/chat",
                        "amazon_bedrock_config": {
                            "aws_region": f"{bedrock_region}",
                            "aws_access_key_id": "{{secrets/zero-to-scope/access_key}}",
                            "aws_secret_access_key": "{{secrets/zero-to-scope/secret_access_key}}",
                            "bedrock_provider": "anthropic",
                        },
                    }
                }
            ]
        },
    )
```

©2024 Databricks Inc. — All rights reserved

샘플링 디코딩

©2024 Databricks Inc. — All rights reserved

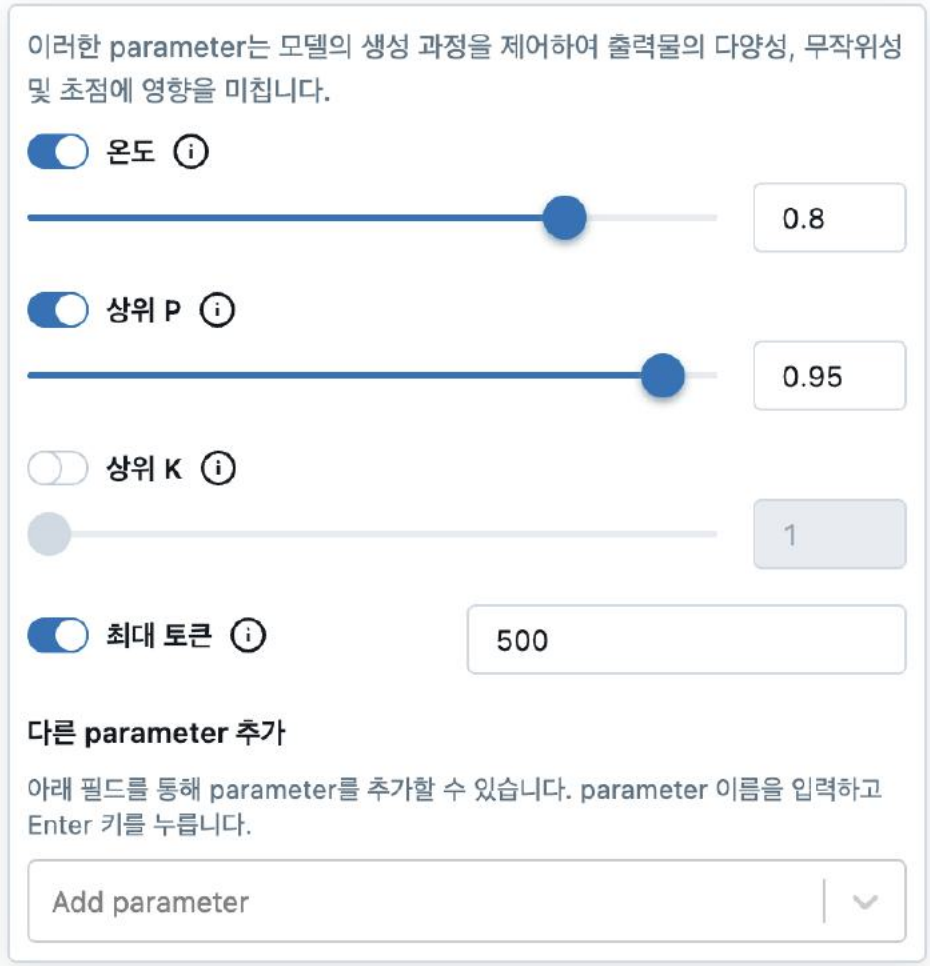
Decoding & Stopping 기준

샘플링 디코딩(Sampling Decoding) 전략

- 샘플링 디코딩은 LLM 추론에서 텍스트 생성의 **가변성과 무작위성**을 제어하기위해 모델의 예측분포에서 가장 가능성이 높은 토큰만 샘플링하는 기술
- 샘플링 디코딩은 할루시네이션을 줄이기위한 기법이 아닌 출력의 복잡도를 제어하여 **다양성과 창의성을 제어**하기위한 기술이지만 가변성이 높아지면 할루시네이션이 높아질 가능성이 있음
- Top P와 Top K는 목적이 동일하기 때문에 둘 중 하나만 선택하는 것이 토큰 선택의 복잡성을 낮출 수 있음.

중단기준(Stopping Criteria) 전략

- 언어모델이 생성하는 토큰의 최대 수를 제어



©2024 Databricks Inc. — All rights reserved

샘플링 디코딩 파라미터

온도 (Temperature)

- 토큰이 선택될 확률 스코어(Probability Score)를 계산하여 확률분포를 생성할 때, **토큰 간의 차이를 얼마나 증폭**시킬지에 대한 파라미터
- Temperature값이 낮아지면, 선택될 가능성이 높은 토큰의 확률 스코어는 선택될 가능성이 낮은 토큰의 확률 스코어보다 점수가 높아진다
- Temperature값이 낮아지면, 프롬프트에 있는 토큰과 유사도가 높은 토큰이 선택될 확률이 높아진다
- 값이 높을 수록 무작위성, 가변성이 높아짐
- 0.0(최소 가변성) ~ 2.0(최대 가변성) 범위의 부동소수점

온도 예시

- 만약 언어모델이 ***"나는 식당에 가서"*** 문장을 생성하고 그 다음 토큰 예측을 위한 확률분포를 생성해야 한다면,

확률 스코어의 합은 1.0

토큰	확률 스코어(Probability score)
식사를	0.4
주문을	0.3
메뉴를	0.15
예약을	0.1
환불을	0.05

< 온도가 낮은 경우 >

토큰	확률 스코어(Probability score)
식사를	0.3
주문을	0.25
메뉴를	0.2
예약을	0.15
환불을	0.1

< 온도가 높은 경우 >

확률 스코어의 합은 1.0

©2024 Databricks Inc. — All rights reserved

샘플링 디코딩 파라미터

Top P (Nucleus Sampling)

- 다음 토큰을 선택할 때, 토큰이 선택될 확률 스코어(Probability Score)의 **상위 누적 합이 P임계값 까지만 선택**될 수 있도록 제한하는 파라미터
- 값이 높을 수록 무작위성, 가변성이 높아짐
- Top P는 토큰 선택의 확률분포에 영향을 받기 때문에 온도(Temperature)값에 따라 선택되는 토큰의 범위가 달라짐
- 0.0(1개 선택) ~ 1.0(모두 선택) 범위의 부동소수점

Top P 예시

- 만약 언어모델이 ***"나는 식당에 가서"*** 문장을 생성하고 그 다음 토큰 예측을 위해 생성된 확률분포에서 가능성 높은 토큰을 선택을 해야 하고, Top P가 **0.7**로 설정 되었다면,

0.4 + 0.3 = 0.7
-> 2개 선택

토큰	확률 스코어(Probability score)
식사를	0.4
주문을	0.3
메뉴를	0.15
예약을	0.1
환불을	0.05

< 온도가 낮은 경우 >

토큰	확률 스코어(Probability score)
식사를	0.3
주문을	0.25
메뉴를	0.2
예약을	0.15
환불을	0.1

< 온도가 높은 경우 >

0.3 + 0.25 + 0.2
= 0.75
-> 3개 선택
(0.7이 될때까지 선택하기 때문에 0.7이 넘을 수 있음)

©2024 Databricks Inc. — All rights reserved

샘플링 디코딩 파라미터

Top K

- 다음 토큰을 선택할 때, 토큰이 선택될 확률 스코어(Probability Score)의 **상위 K개의 개수 까지만** 선택될 수 있도록 제한하는 파라미터
- 값이 높을 수록 무작위성, 가변성이 높아짐
- Top K는 토큰 선택의 확률분포를 고려하지 않기 때문에 온도(Temperature)값에 따라 선택되는 토큰의 범위가 크게 달라지지 않음
- 1(1개 선택)~100(100개 선택) 범위의 정수

Top K 예시

- 만약 언어모델이 ***"나는 식당에 가서"*** 문장을 생성하고 그 다음 토큰 예측을 위해 생성된 확률분포에서 가능성 높은 토큰을 선택을 해야하고, Top K가 **3개**로 설정 되었다면,

상위 3개 선택

토큰	확률 스코어(Probability score)
식사를	0.4
주문을	0.3
메뉴를	0.15
예약을	0.1
환불을	0.05

< 온도가 낮은 경우 >

토큰	확률 스코어(Probability score)
식사를	0.3
주문을	0.25
메뉴를	0.2
예약을	0.15
환불을	0.1

< 온도가 높은 경우 >

상위 3개 선택

©2024 Databricks Inc. — All rights reserved

실 습 5 :
생성한 LLM 모델 테스트

©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트

databricks

데이터, 노트북, 최근 항목 등을 검색하세요... ⌘ + P

db-a-01-workspace-182363696740-...

컴퓨팅

SQL

- SQL 편집기
- 쿼리
- 대시보드
- 알림
- 쿼리 기록
- SQL Warehouse

Data Engineering

- 작업 실행
- 데이터 수집
- Delta Live Table

기계 학습

- Playground
- Experiments
- 기능
- 모델
- 서비스
- 마켓플레이스
- Partner Connect

메뉴 축소

Zero-to-GenAI-Workshop Python

파일 편집 보기 실행 도움말 마지막 편집: 21분 전 피드백 제공

인터럽트 Workshop Cluster 스케줄 공유

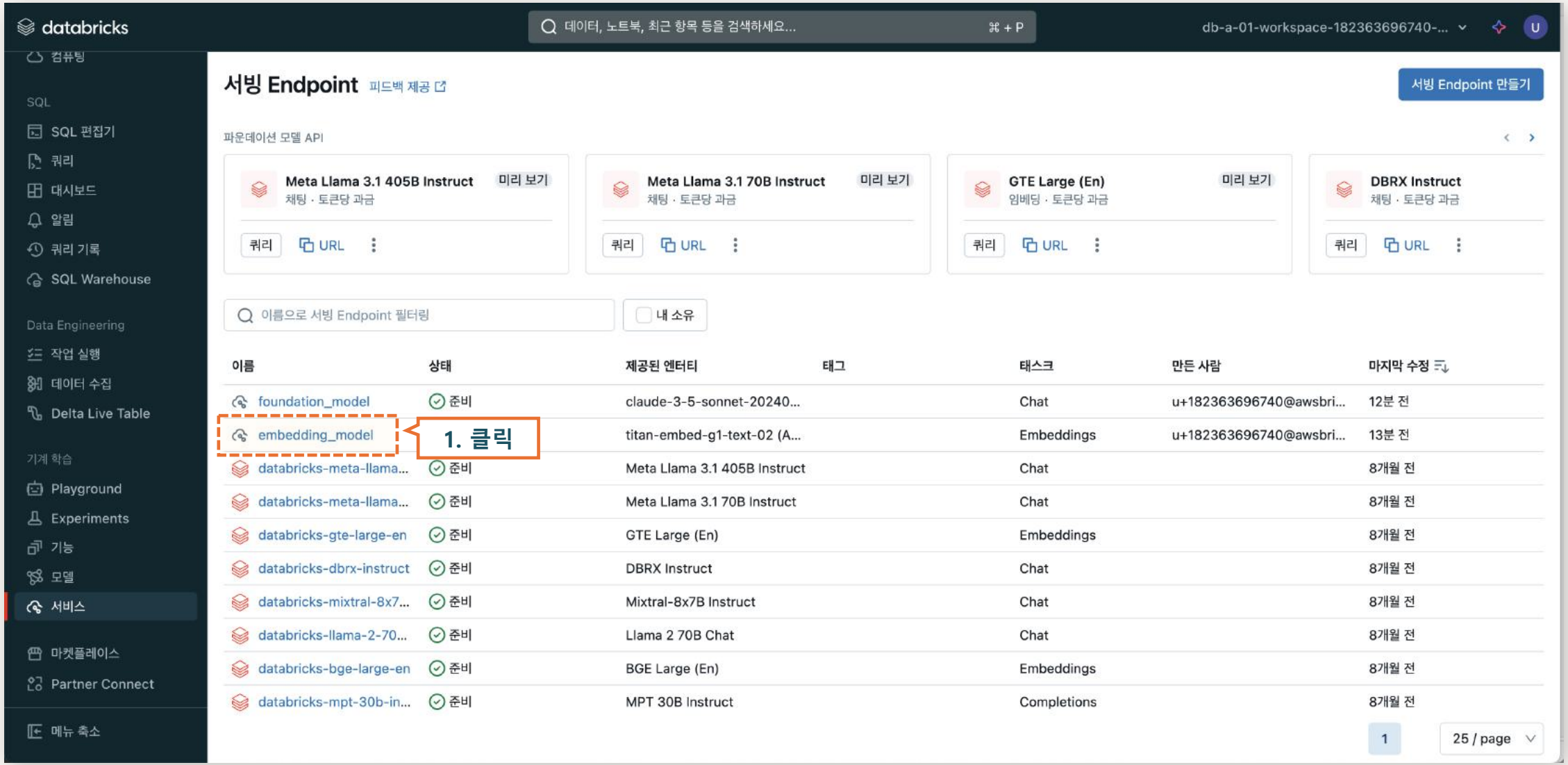
오전 03:46 (1초) 19: 생성 파운데이션 모델 엔드포인트 생성

```
import mlflow.deployments
deploy_client = mlflow.deployments.get_deploy_client("databricks")

if generative_model_name in [endpoints['name'] for endpoints in deploy_client.list_endpoints()]:
    #deploy_client.delete_endpoint(generative_model_name)
    print("동일한 모델의 엔드포인트가 이미 존재합니다.")

# claude-3-5-sonnet-20240620-v1:0, claude-3-sonnet-20240229-v1:0
if generative_model_name not in [endpoints['name'] for endpoints in deploy_client.list_endpoints()]:
    deploy_client.create_endpoint(
        name=generative_model_name,
        config={
            "served_entities": [
                {
                    "external_model": {
                        "name": "claude-3-sonnet-20240229-v1:0",
                        "provider": "amazon-bedrock",
                        "task": "llm/v1/chat",
                        "amazon_bedrock_config": {
                            "aws_region": f"{bedrock_region}",
                            "aws_access_key_id": "{{secrets/zero-to-scope/access_key}}",
                            "aws_secret_access_key": "{{secrets/zero-to-scope/secret_access_key}}",
                            "bedrock_provider": "anthropic",
                        },
                    }
                }
            ]
        },
    )
```

1. 새창에서 열기
(Windows : Ctrl + 클릭
Mac : Cmd + 클릭)

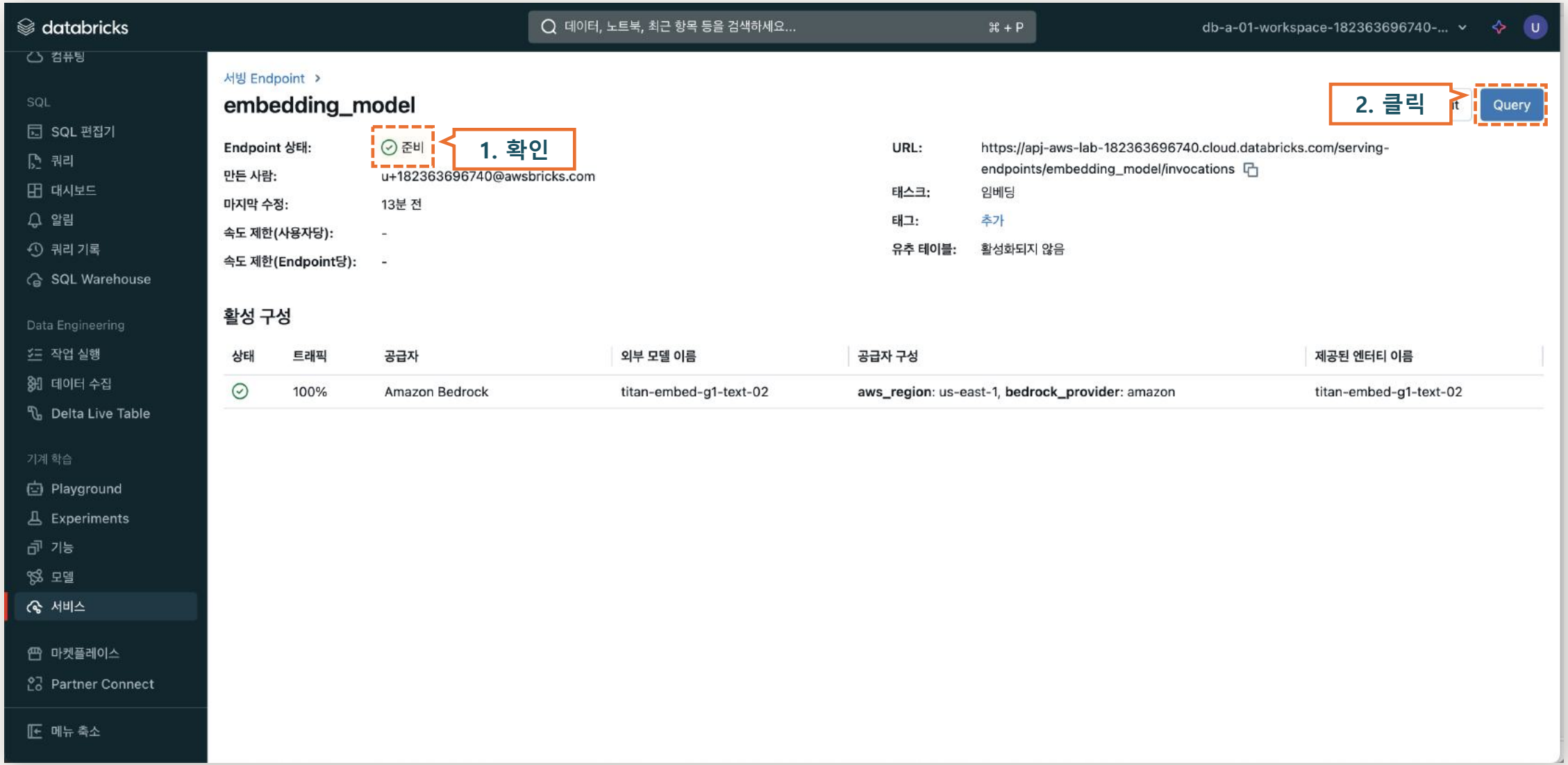
©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트



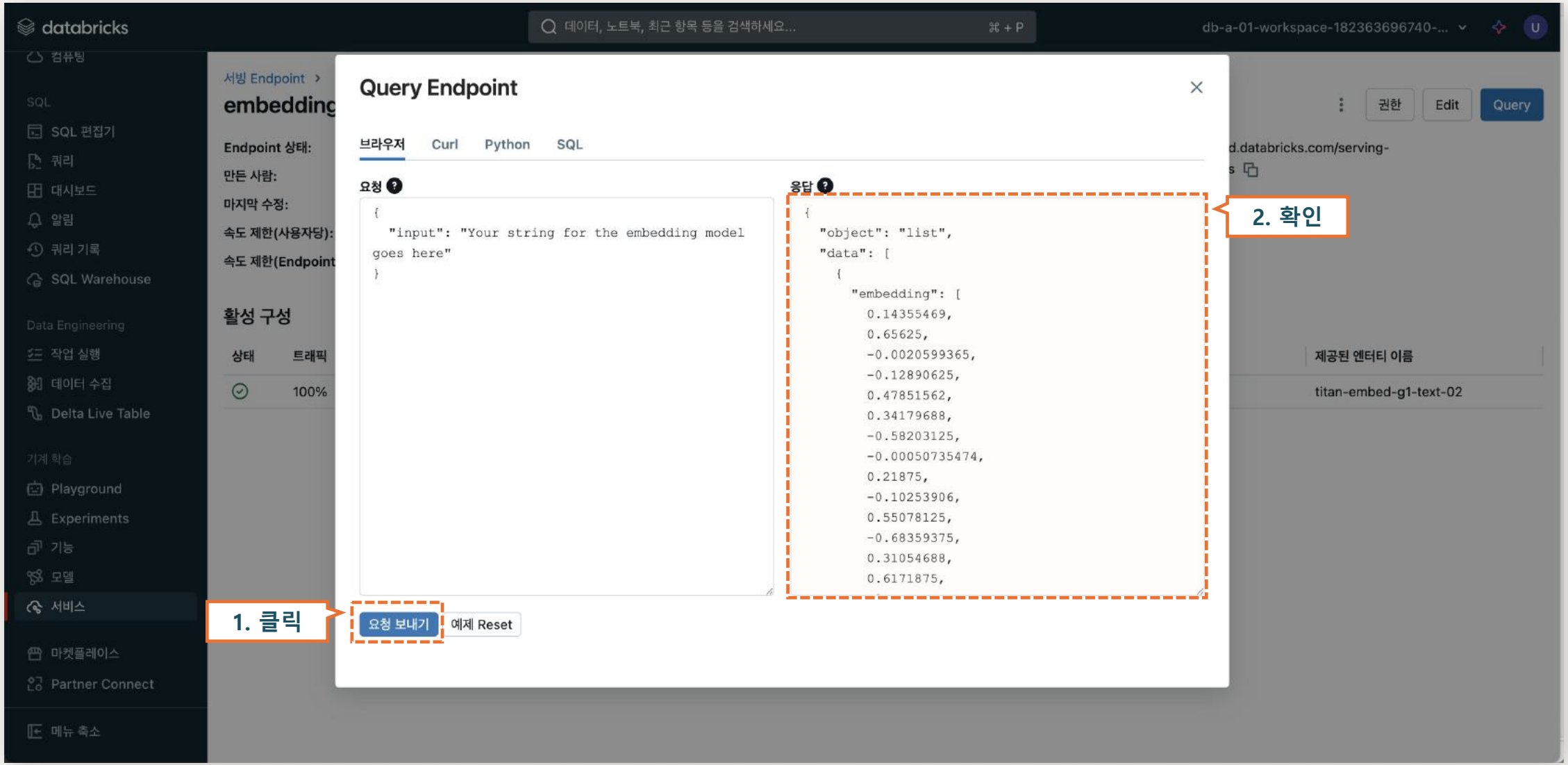
©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트



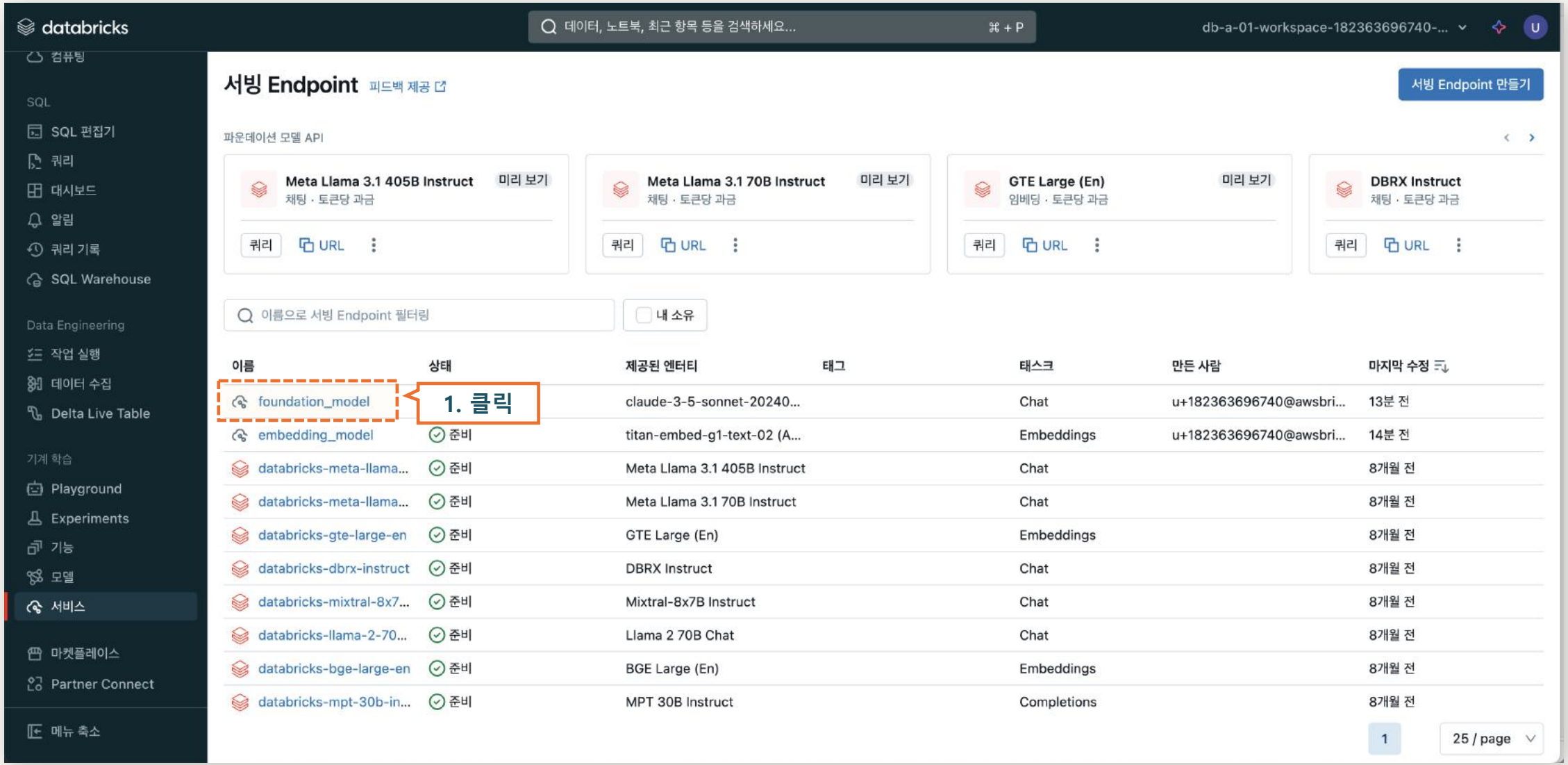
©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트



©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트

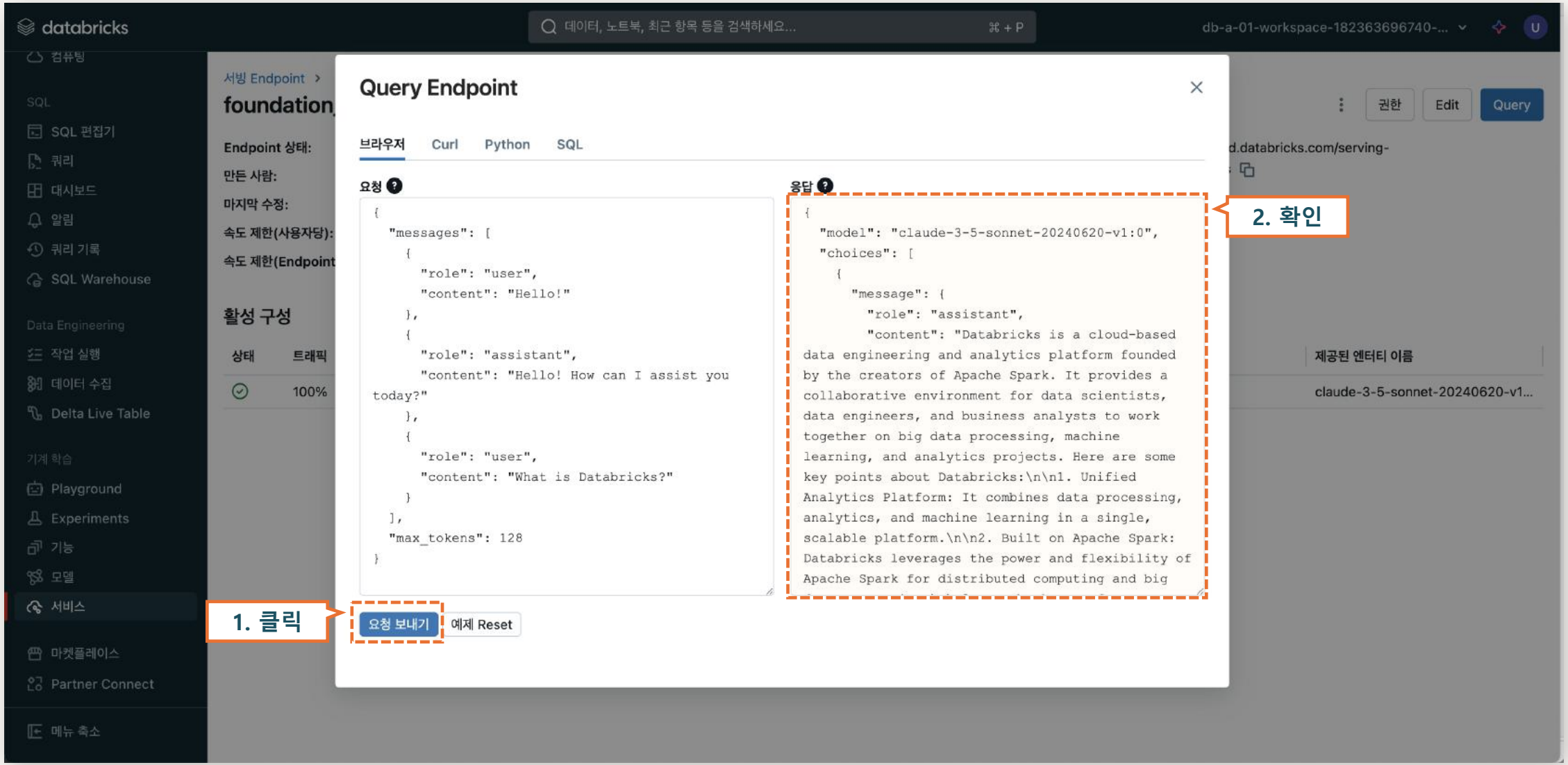


©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트

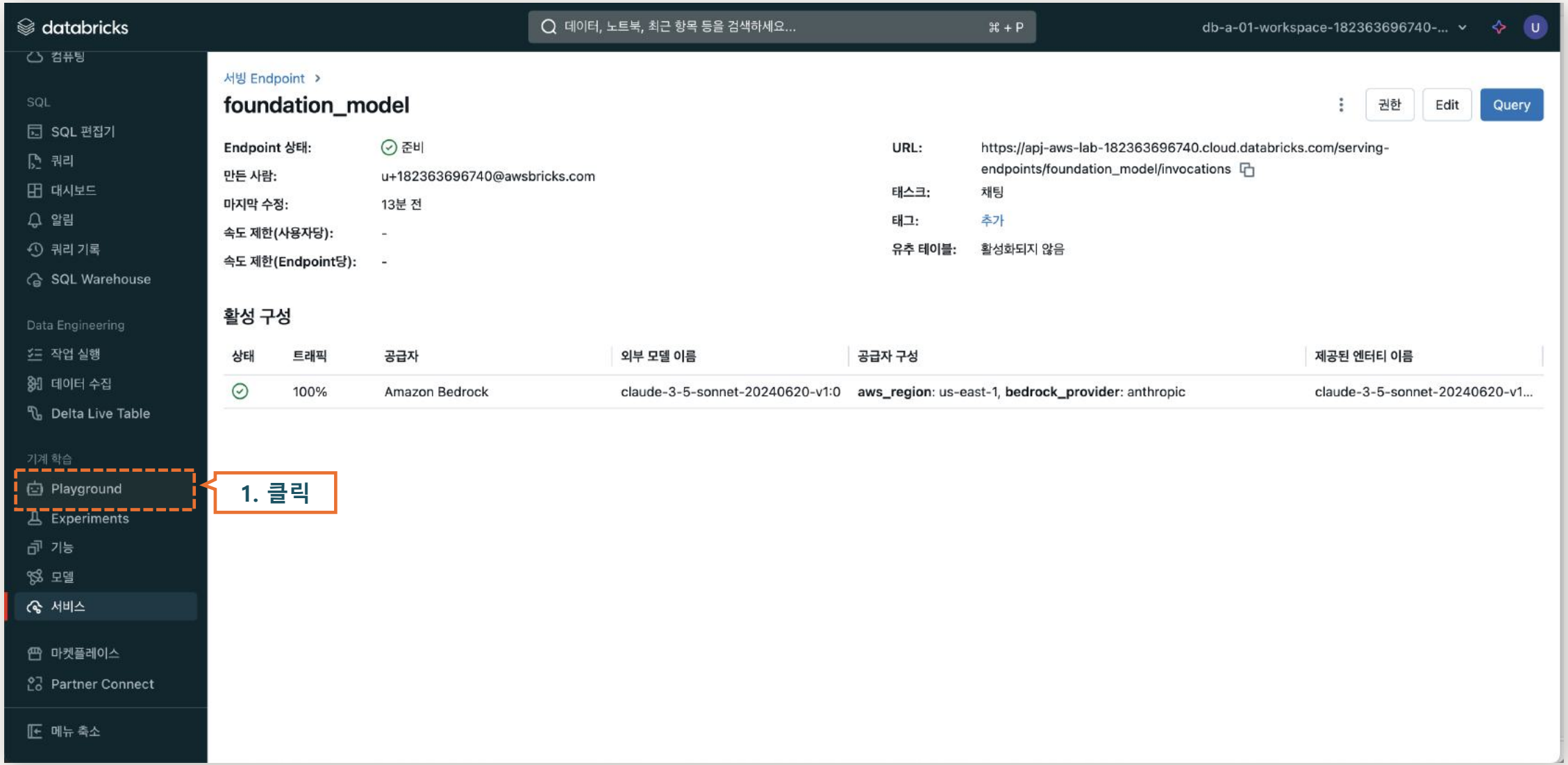
©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트



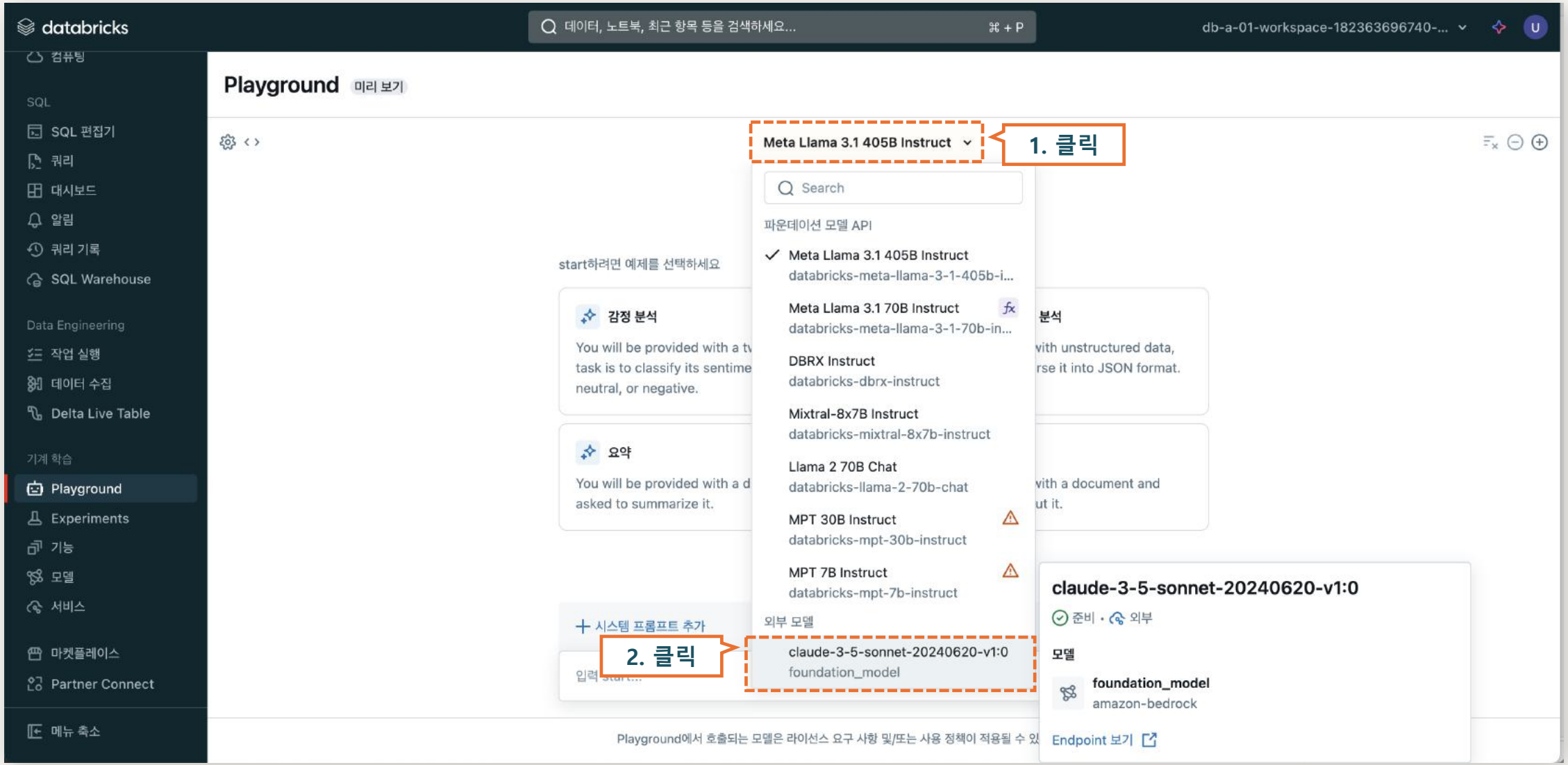
©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트



©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트

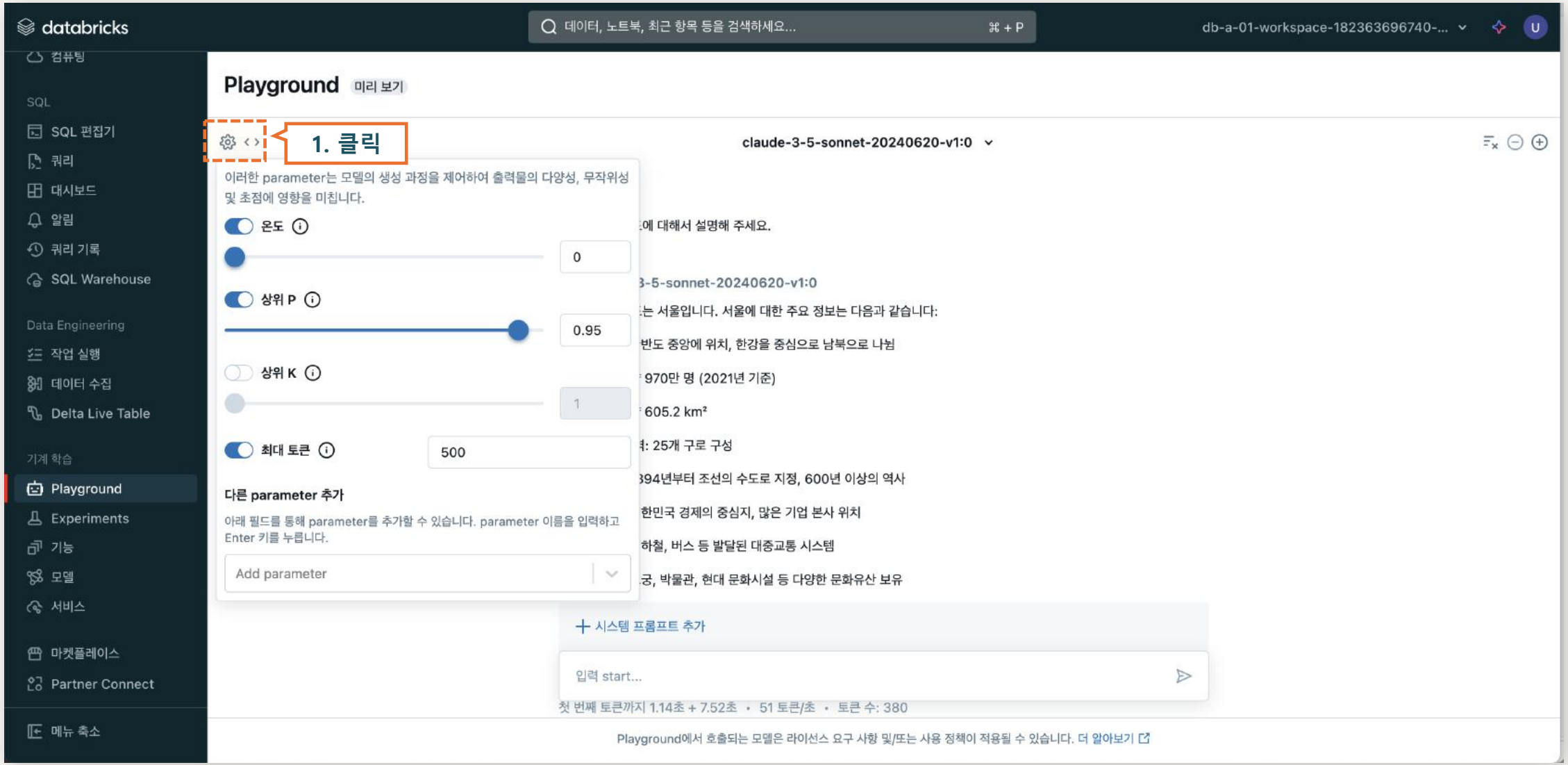


©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트

©2024 Databricks Inc. — All rights reserved

생성한 LLM 모델 테스트



©2024 Databricks Inc. — All rights reserved

벡터 임베딩

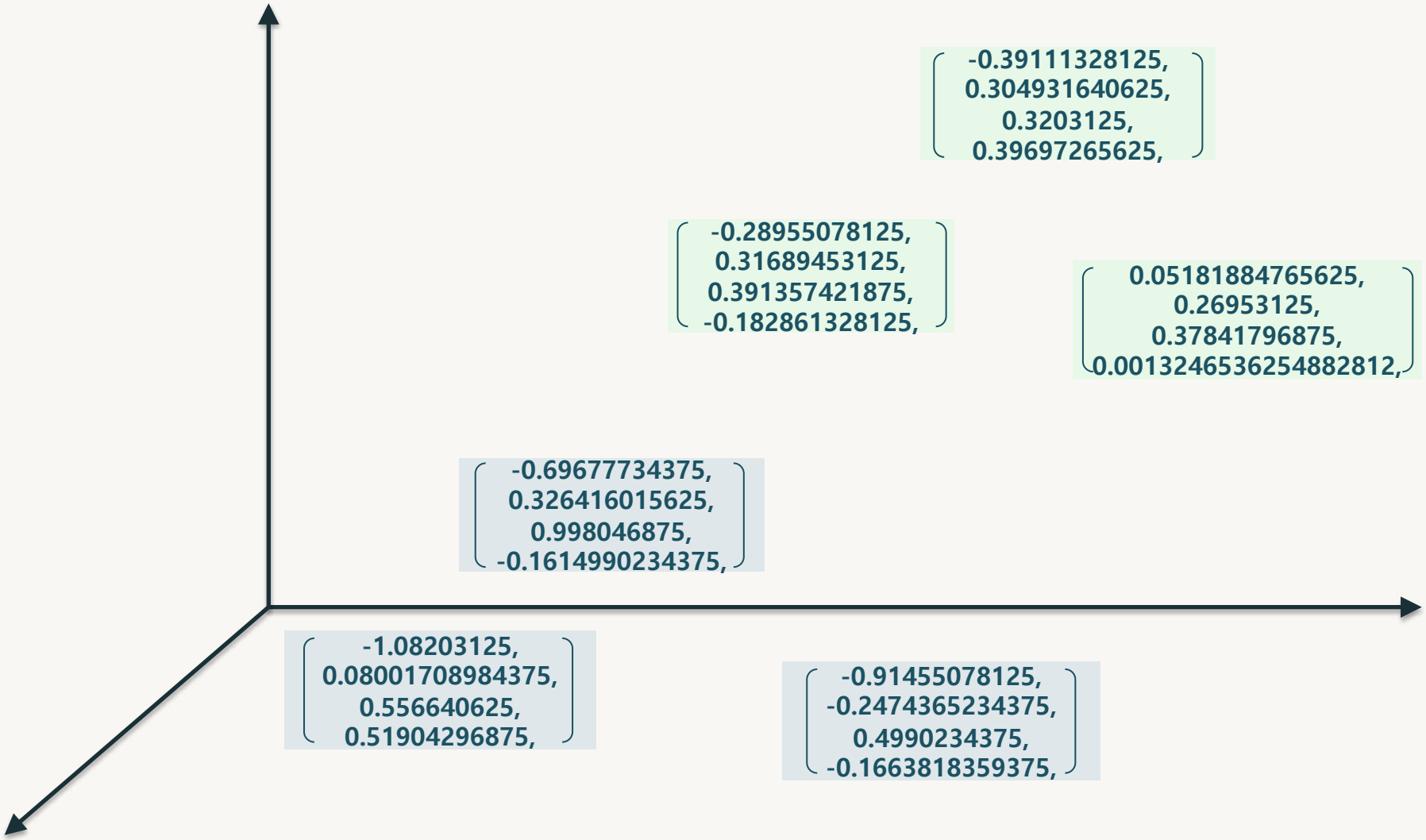
©2024 Databricks Inc. — All rights reserved

문장의 벡터 임베딩

“수요분석에 기반하여 정책적 관점에서 SW융합인력 양성 관련 향후 과제를 제안”	Embedding Model	-0.39111328125, 0.304931640625, 0.3203125, 0.39697265625,
“채용공고 데이터를 수집하여 SW융합인력 수요 특징 등을 분석함”		-0.28955078125, 0.31689453125, 0.391357421875, -0.182861328125,
“채용공고 데이터 기반의 SW융합인력 수요분석 정보를 제공”		0.05181884765625, 0.26953125, 0.37841796875, 0.0013246536254882812,
“자동차 부문은 SW인력 수요 증가와 부족 현상이 심화됨”		-0.69677734375, 0.326416015625, 0.998046875, -0.1614990234375,
“자율주행차가 완전 자율주행 수준을 의미하는 자율주행 레벨4 개발에 주력”		-1.08203125, 0.08001708984375, 0.556640625, 0.51904296875,
“자율주행차 시장 규모가 2025년 1,549억 달러, 2035년 1조 달러로 연평균 40% 이상 성장예상”		-0.91455078125, -0.2474365234375, 0.4990234375, -0.1663818359375,

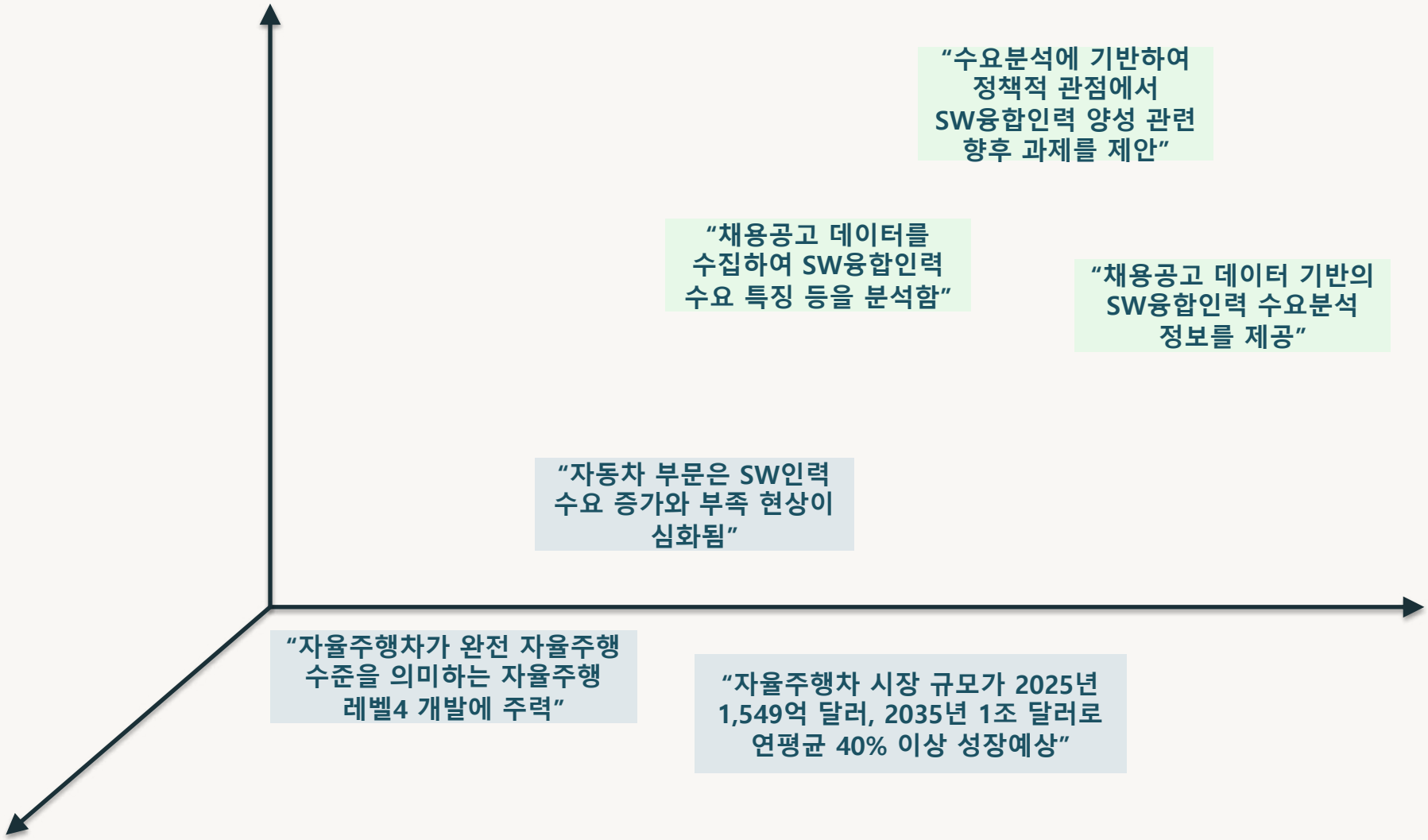
©2024 Databricks Inc. — All rights reserved

문장의 벡터 임베딩



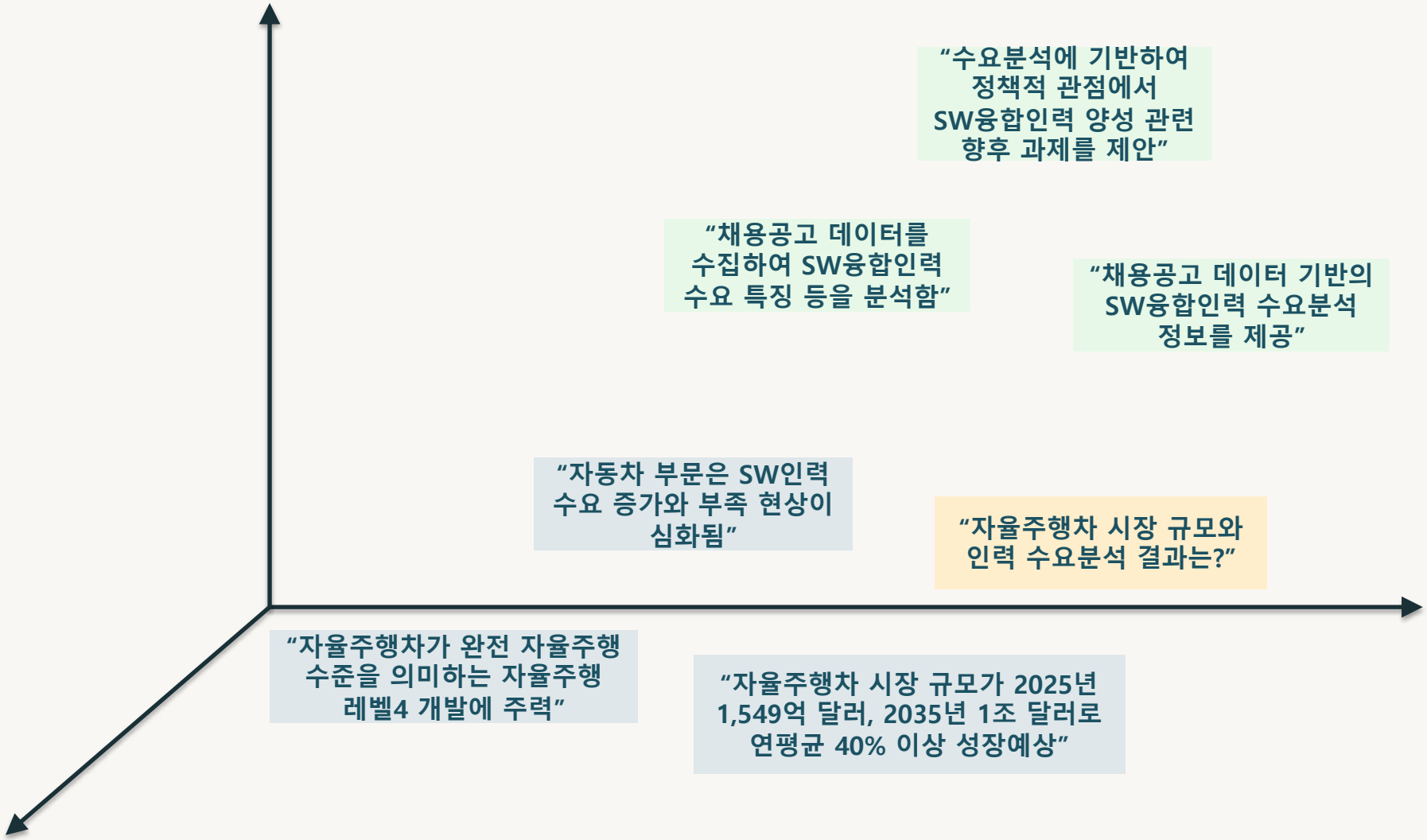
©2024 Databricks Inc. — All rights reserved

문장의 벡터 임베딩



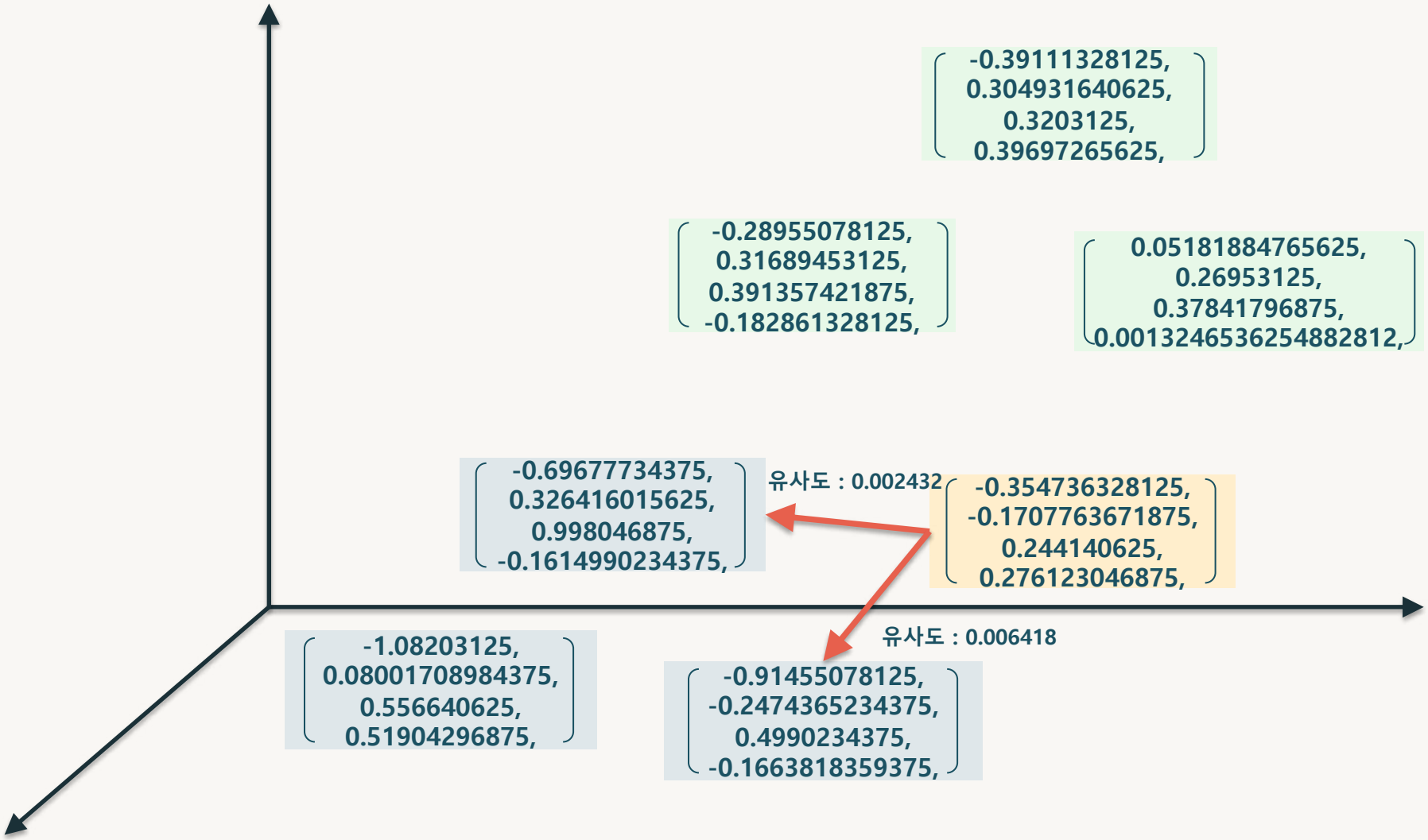
©2024 Databricks Inc. — All rights reserved

문장의 벡터 임베딩



©2024 Databricks Inc. — All rights reserved

문장의 벡터 임베딩

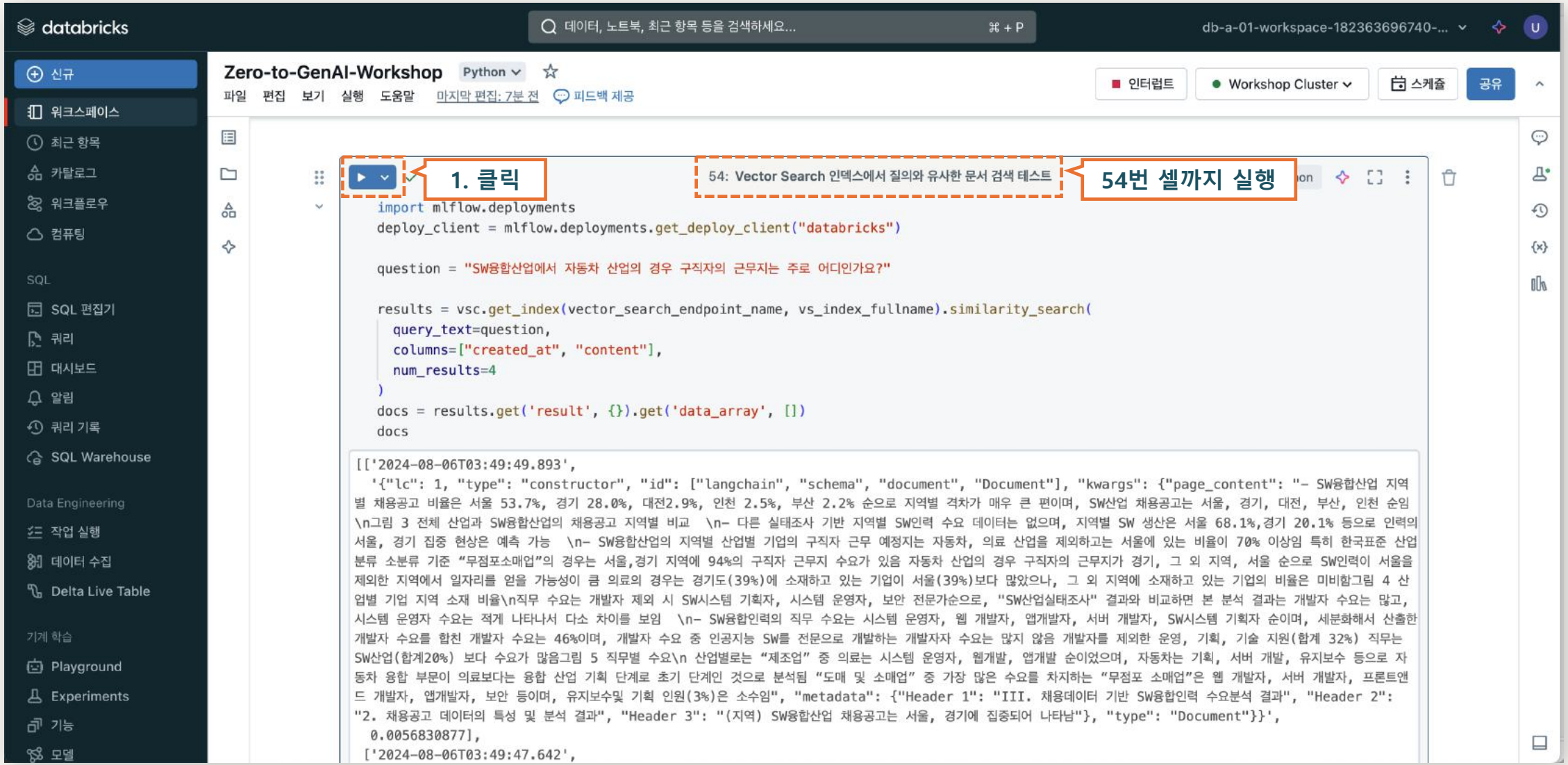


©2024 Databricks Inc. — All rights reserved

실 습 6 :
Databricks Vector Search 인덱스 생성

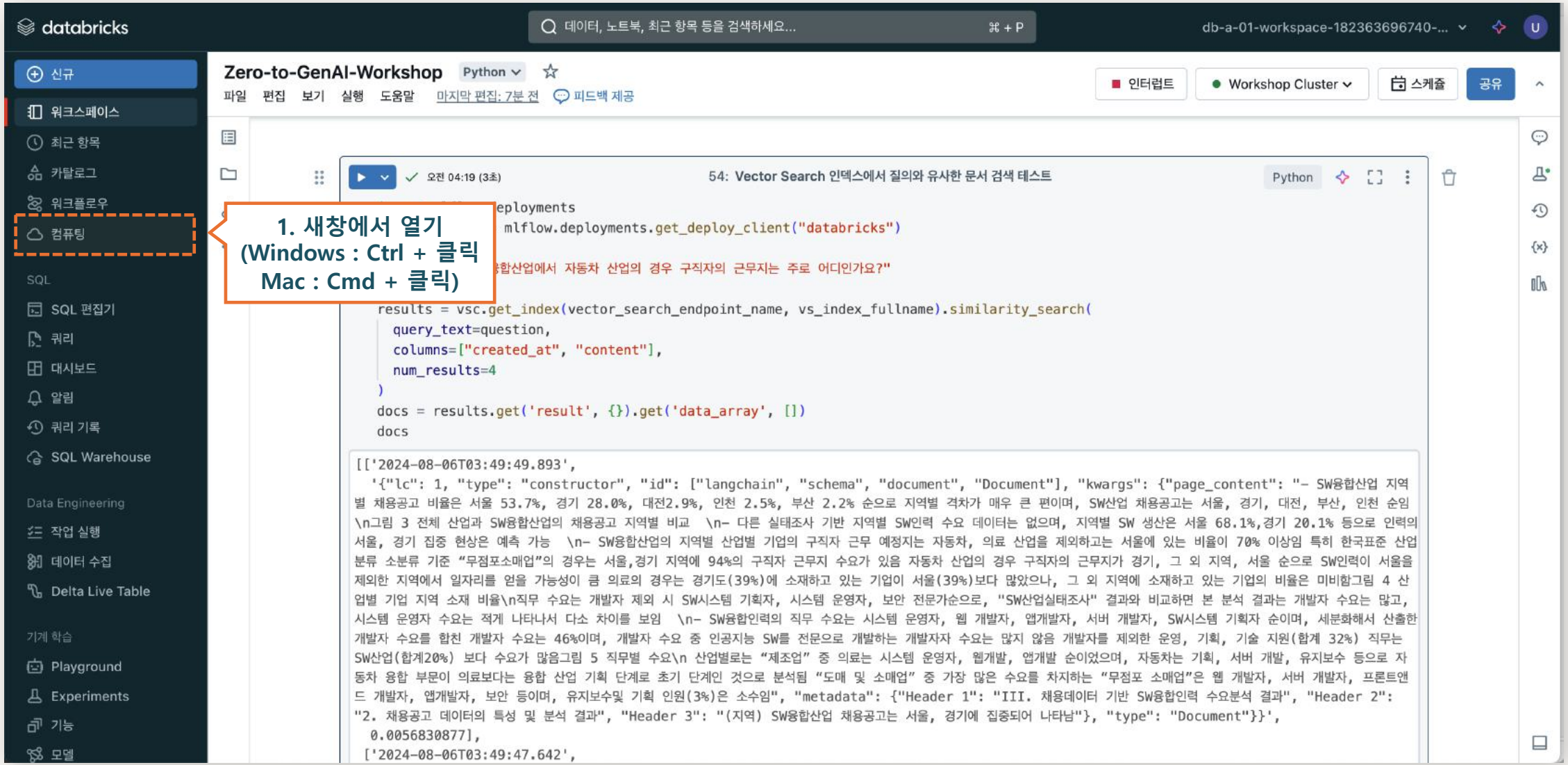
©2024 Databricks Inc. — All rights reserved

Databricks Vector Search 인덱스 생성



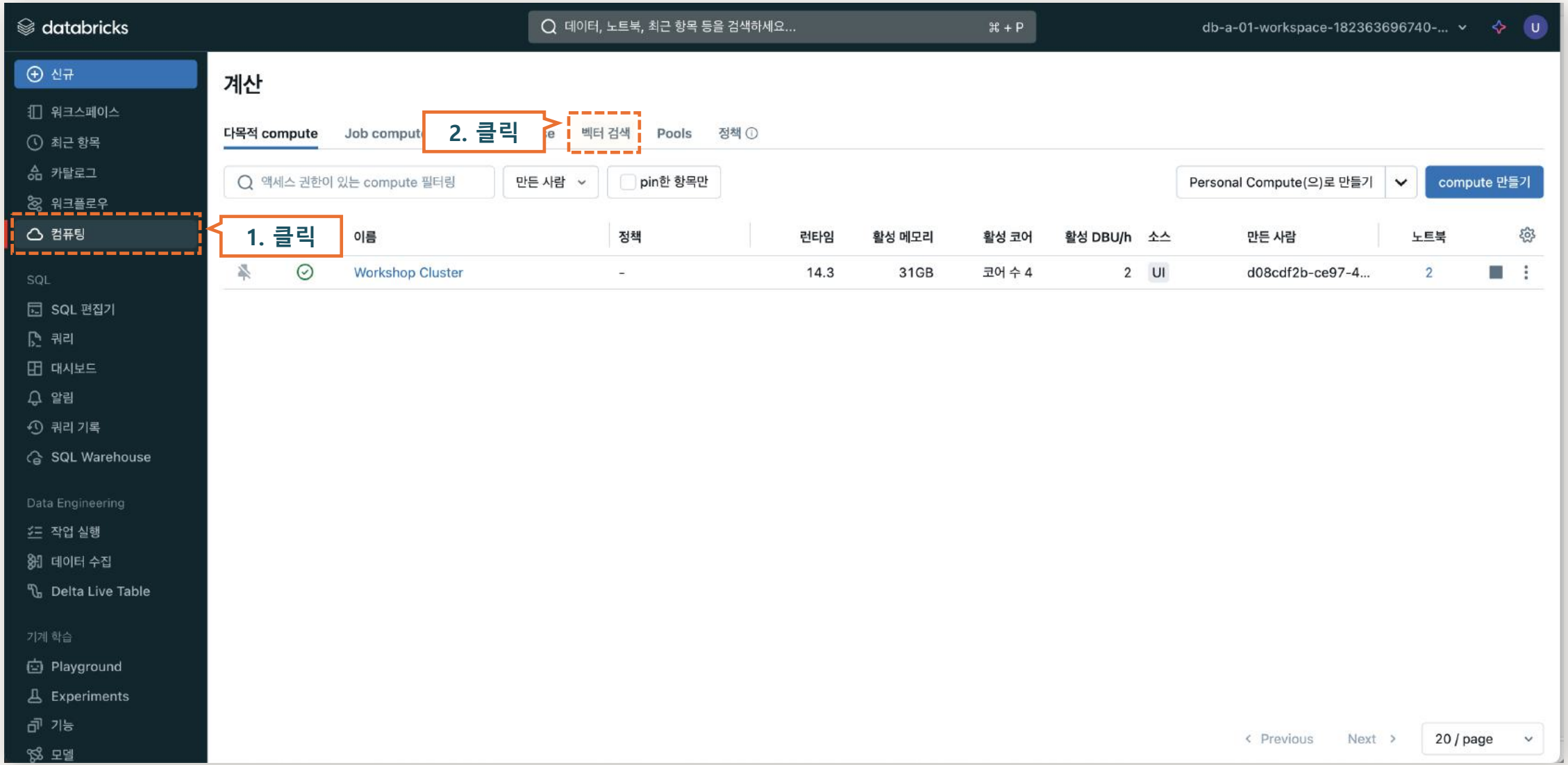
©2024 Databricks Inc. — All rights reserved

Databricks Vector Search 인덱스 생성



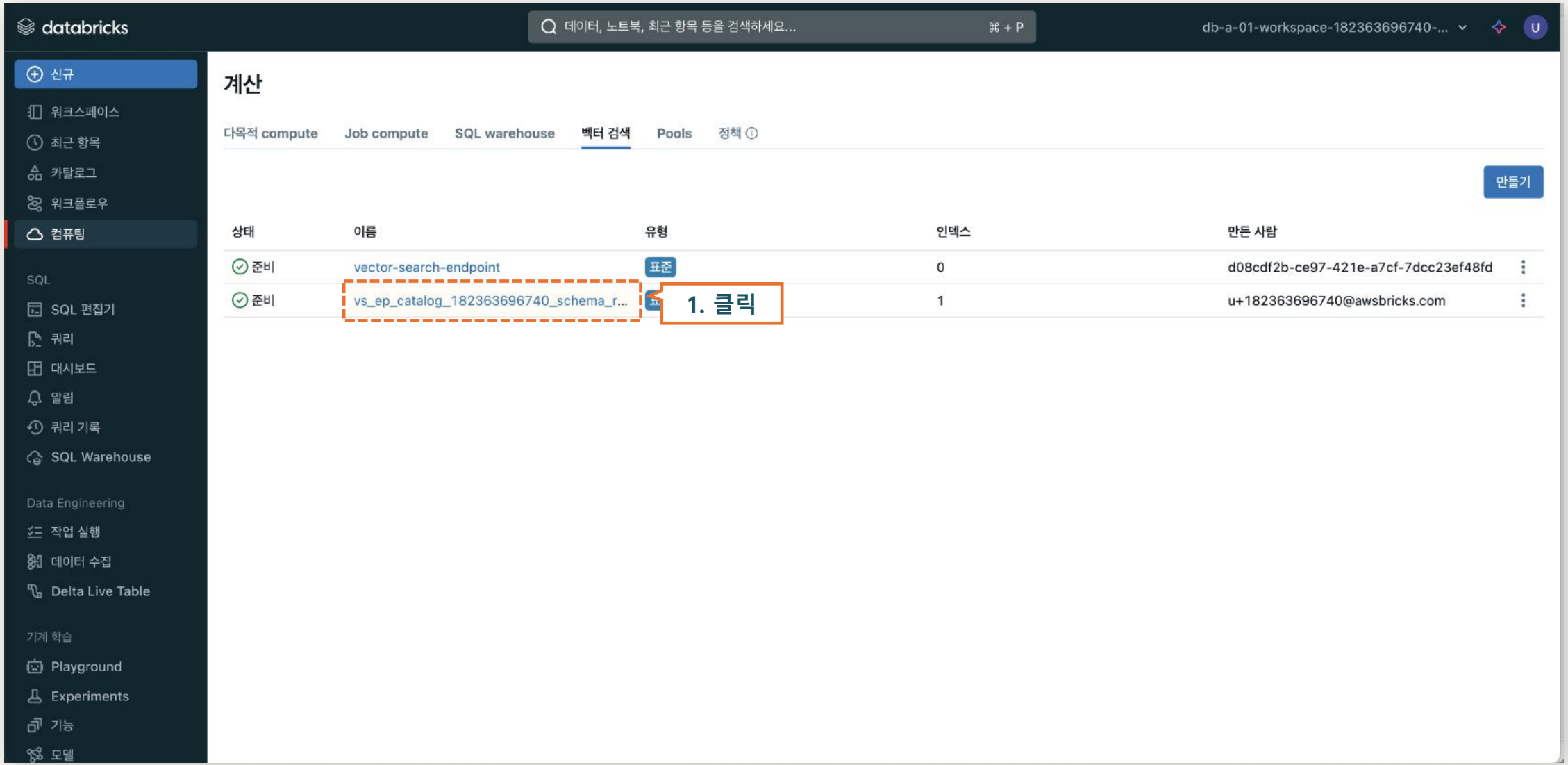
©2024 Databricks Inc. — All rights reserved

Databricks Vector Search 인덱스 생성



©2024 Databricks Inc. — All rights reserved

Databricks Vector Search 인덱스 생성

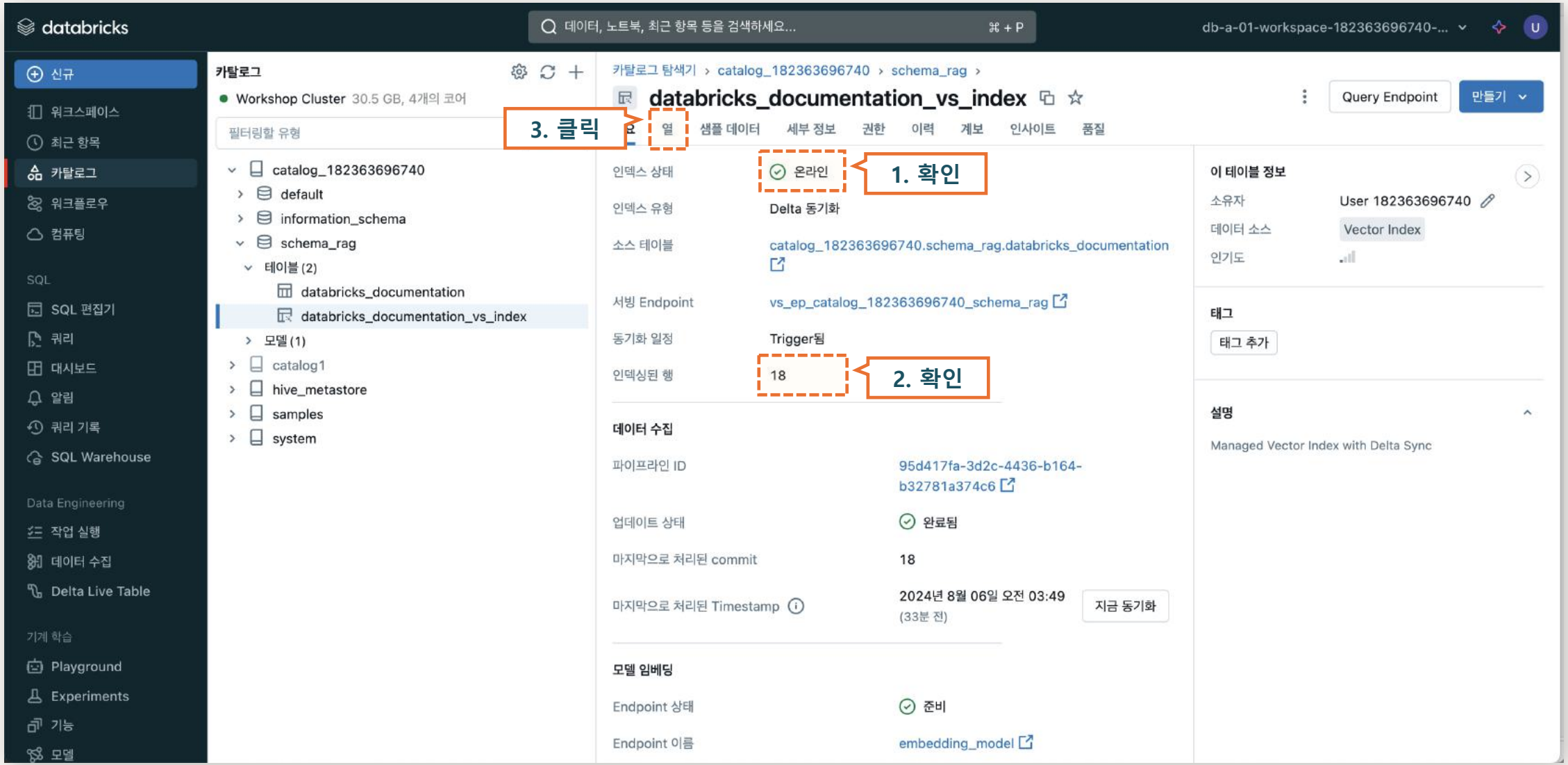


©2024 Databricks Inc. — All rights reserved

Databricks Vector Search 인덱스 생성

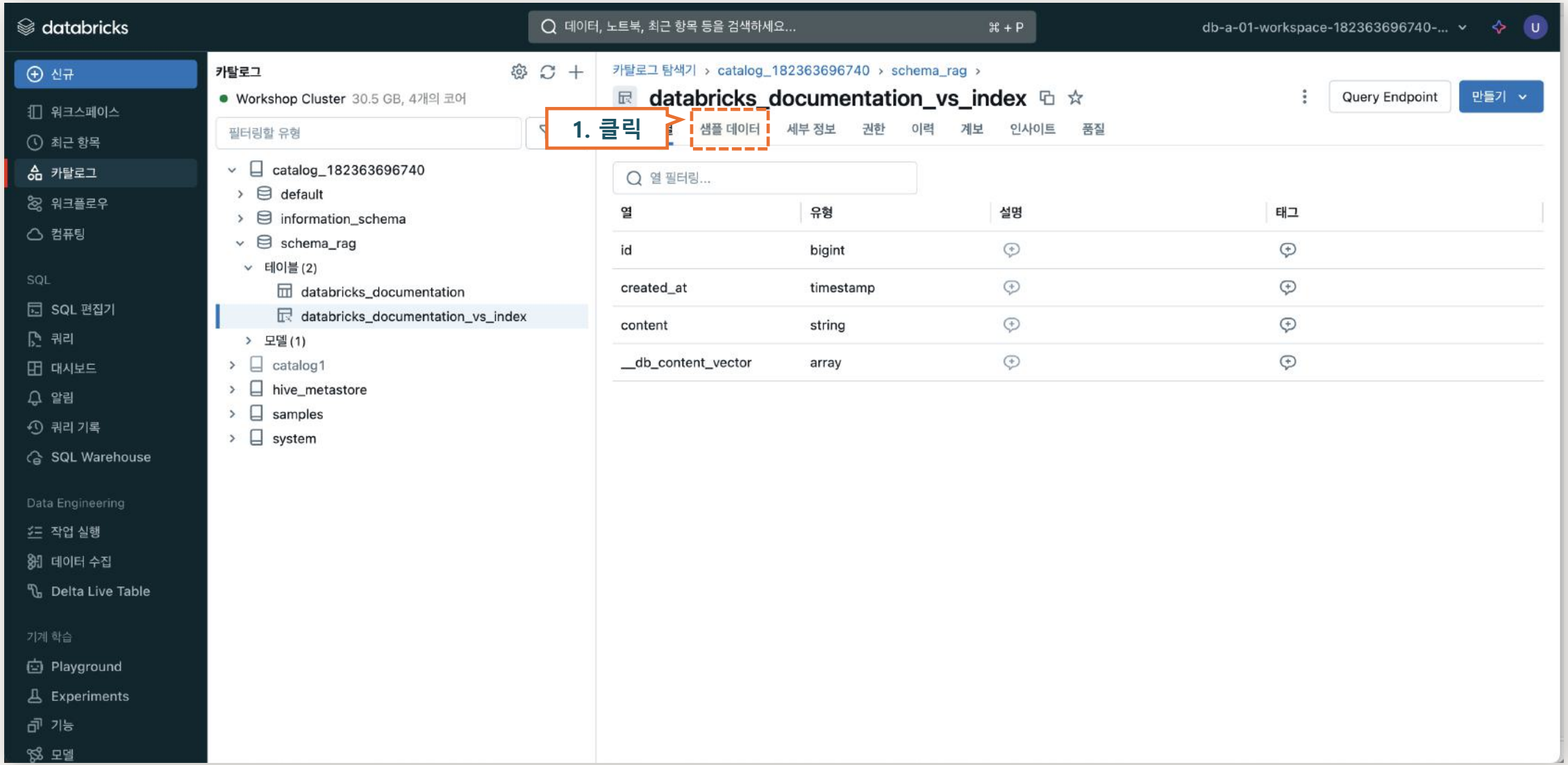
©2024 Databricks Inc. — All rights reserved

Databricks Vector Search 인덱스 생성



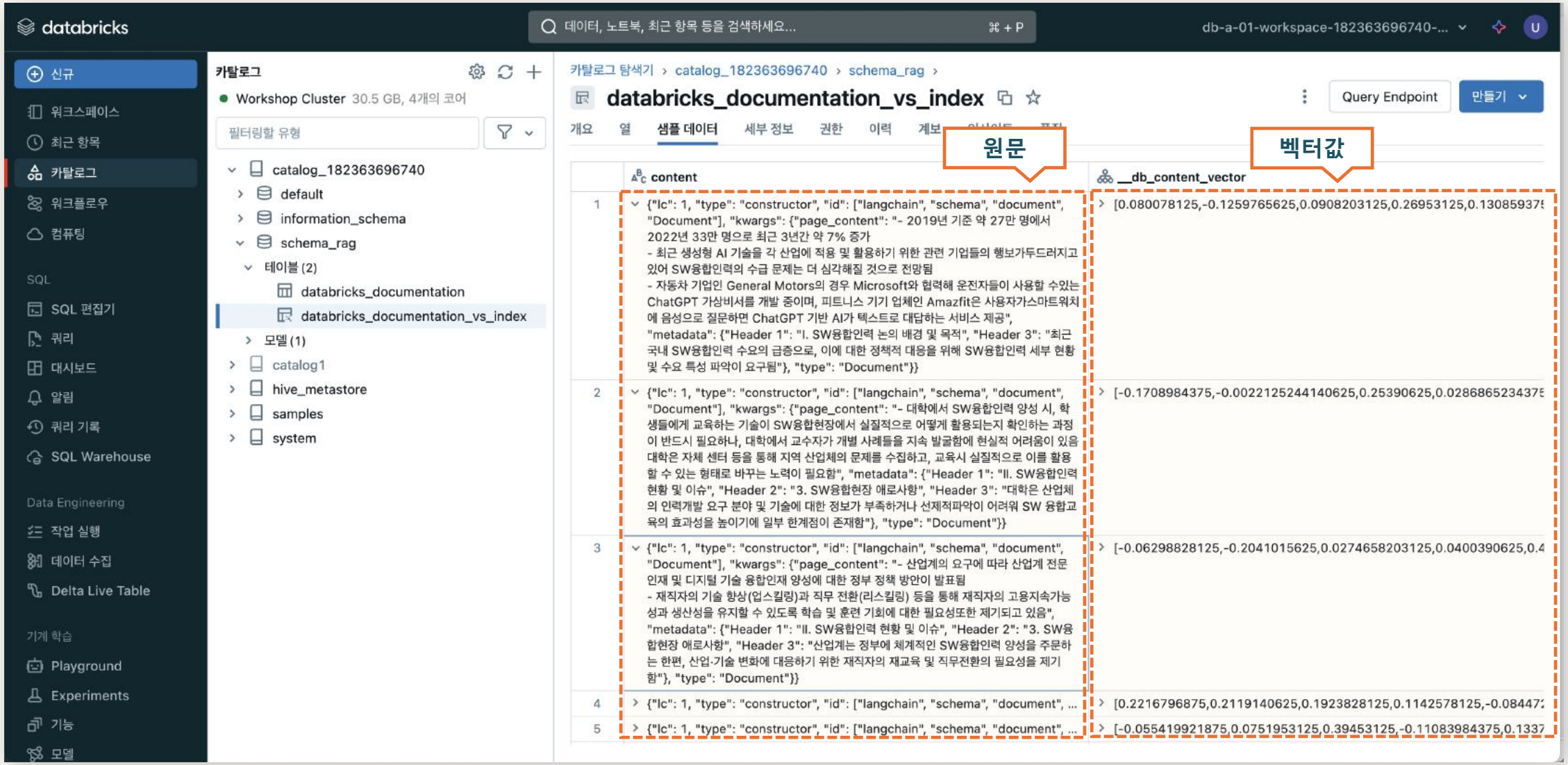
©2024 Databricks Inc. — All rights reserved

Databricks Vector Search 인덱스 생성



©2024 Databricks Inc. — All rights reserved

Databricks Vector Search 인덱스 생성



©2024 Databricks Inc. — All rights reserved

실 습 7 :
RAG 체인을 엔드포인트로 배포

©2024 Databricks Inc. — All rights reserved

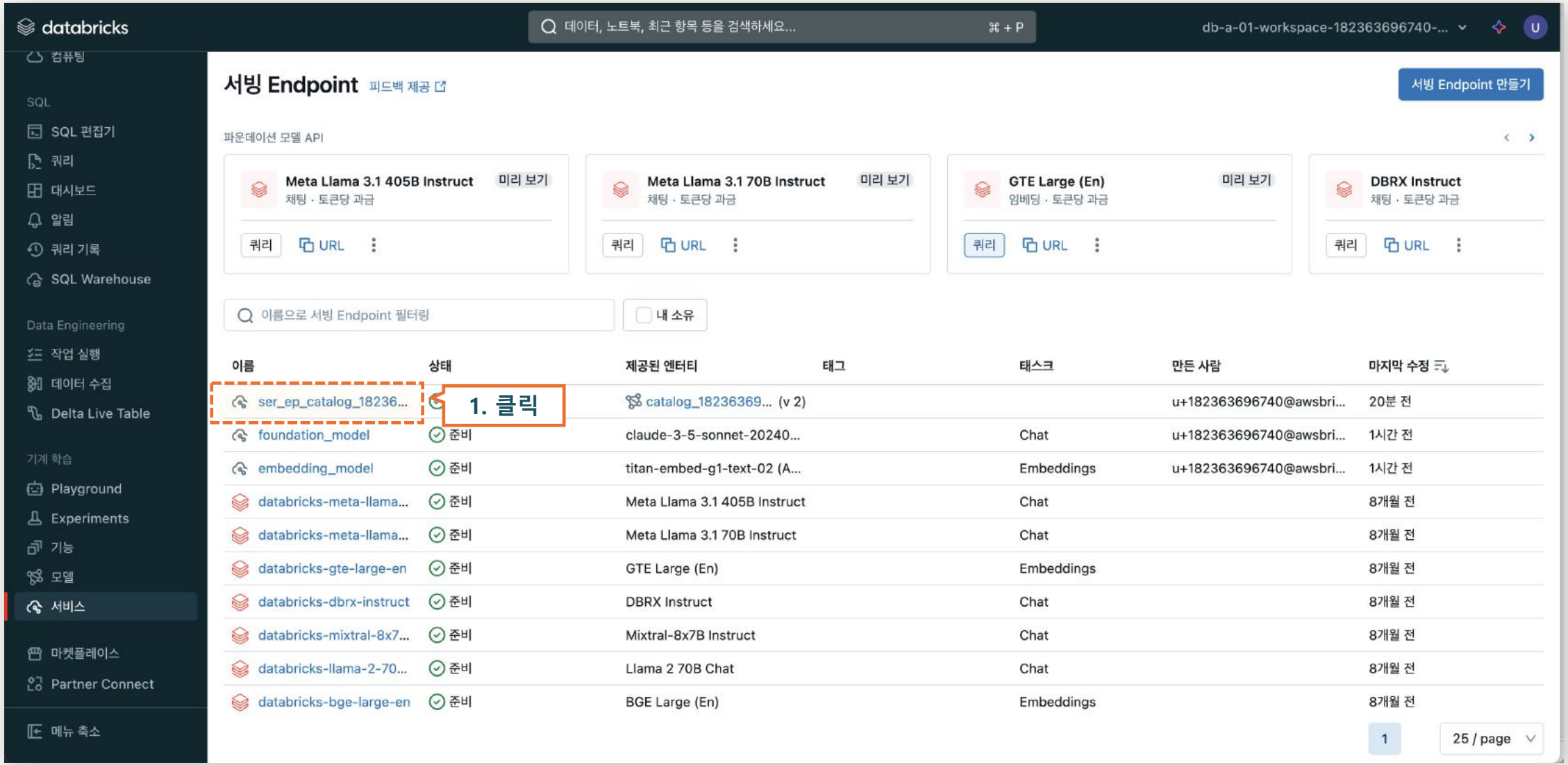
RAG 체인을 엔드포인트로 배포

©2024 Databricks Inc. — All rights reserved

RAG 체인을 엔드포인트로 배포

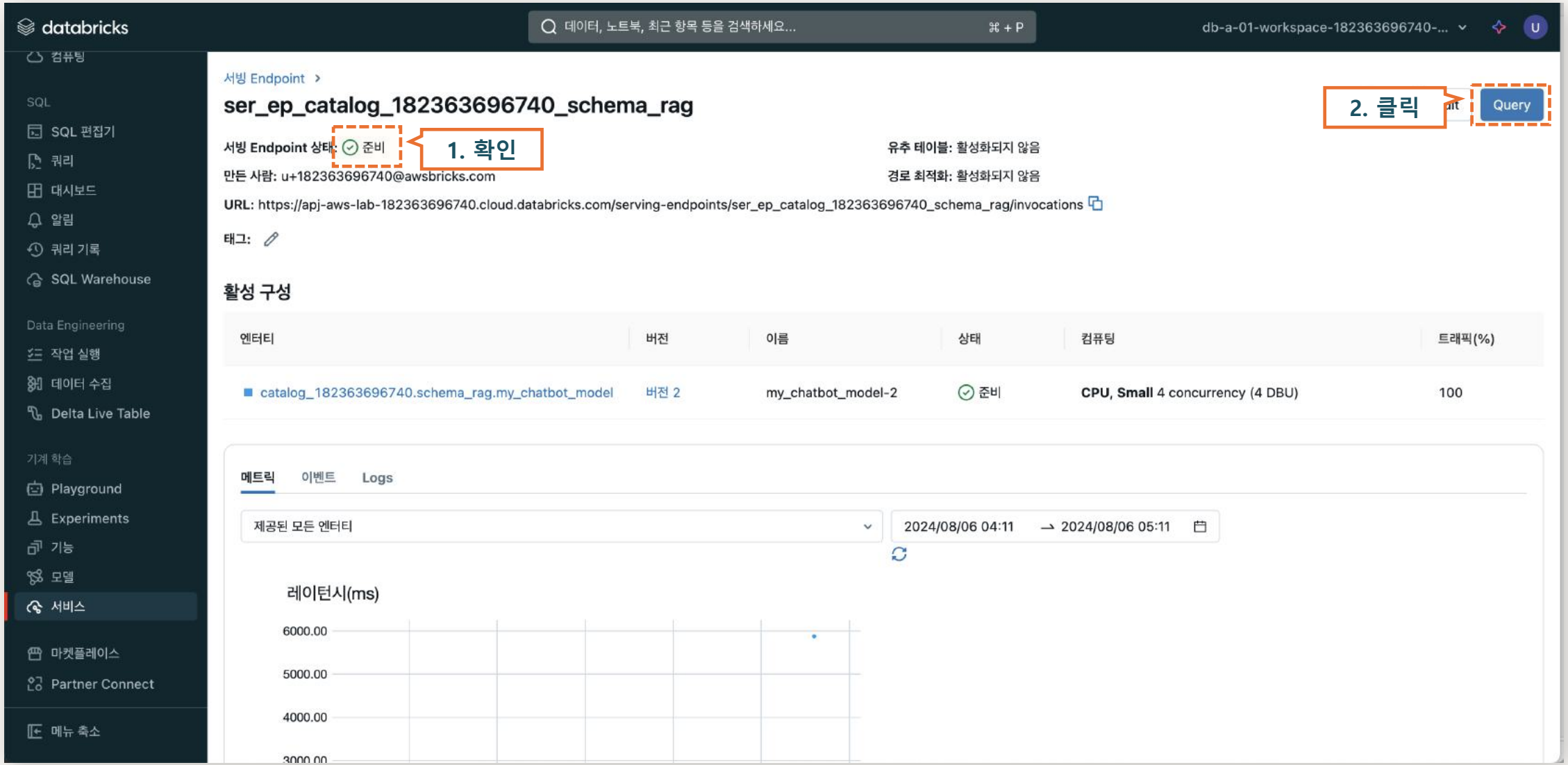
©2024 Databricks Inc. — All rights reserved

RAG 체인을 엔드포인트로 배포



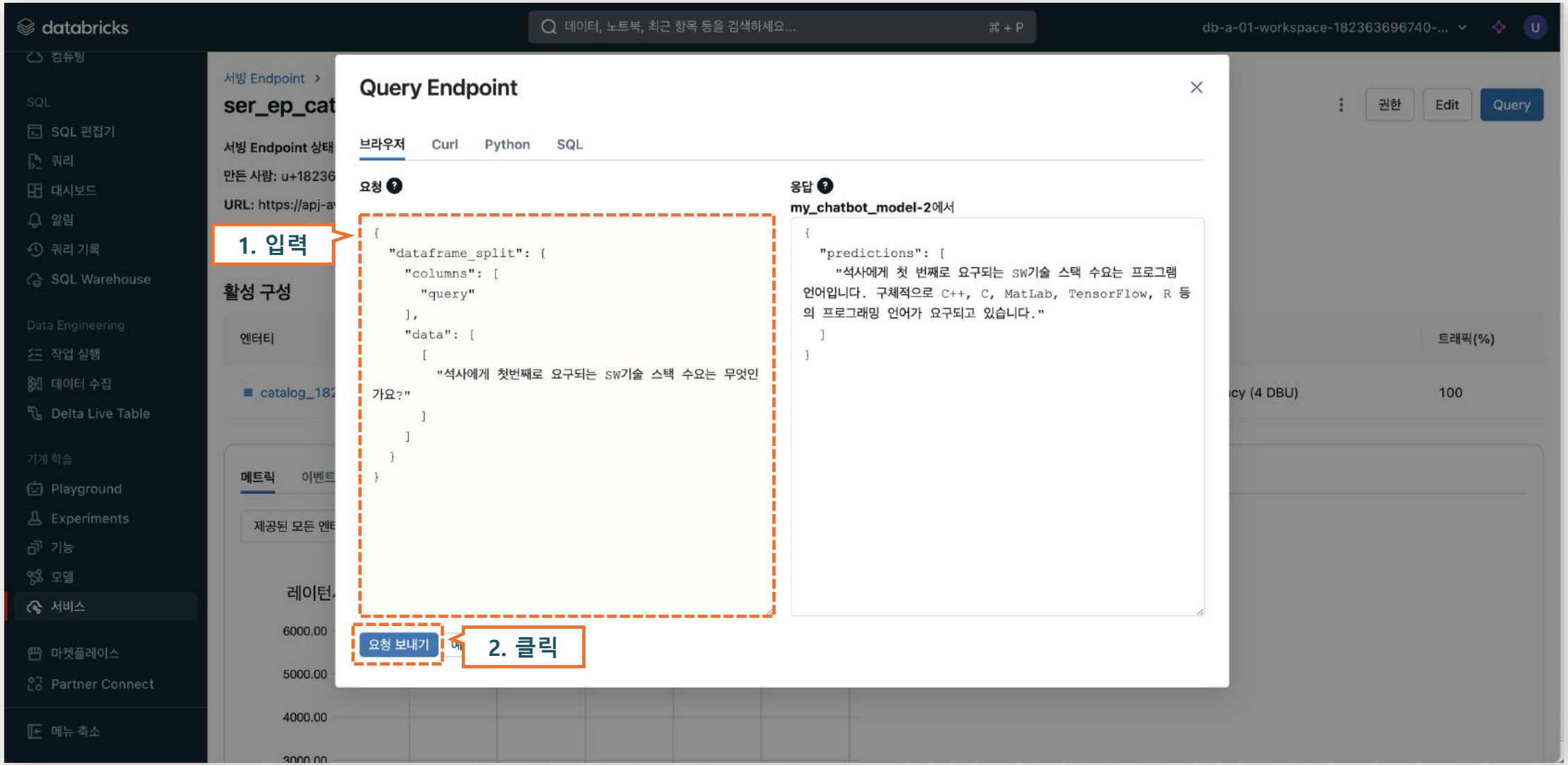
©2024 Databricks Inc. — All rights reserved

RAG 체인을 엔드포인트로 배포



©2024 Databricks Inc. — All rights reserved

RAG 체인을 엔드포인트로 배포

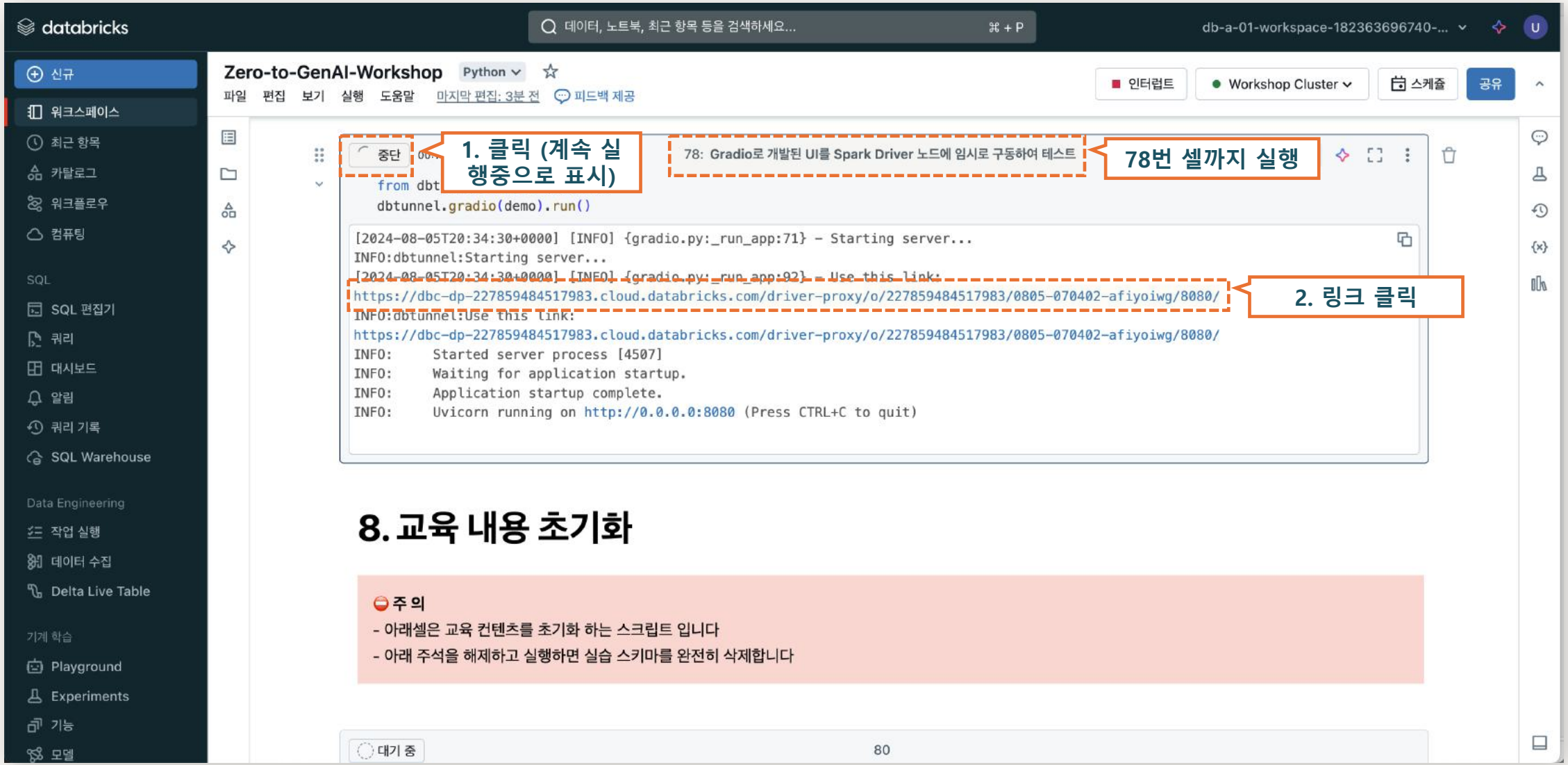


©2024 Databricks Inc. — All rights reserved

실 습 8 :
Gradio를 활용한 RAG 챗봇 UI 테스트

©2024 Databricks Inc. — All rights reserved

Gradio를 활용한 RAG 챗봇 UI 테스트



©2024 Databricks Inc. — All rights reserved

Gradio를 활용한 RAG 챗봇 UI 테스트

©2024 Databricks Inc. — All rights reserved

Gradio를 활용한 RAG 챗봇 UI 테스트

©2024 Databricks Inc. — All rights reserved

Gradio를 활용한 RAG 챗봇 UI 테스트

©2024 Databricks Inc. — All rights reserved

GenAI 한계와 윤리적 문제

©2024 Databricks Inc. — All rights reserved

LLM은 완벽한가?

LLM의 성능

- LLM 모델들의 정확도(accuracy)는 모델의 파라미터 크기가 클수록 높아짐
- LLM 모델들의 정확도는 파라미터가 70B와 7B X 8에서도 80% 수준
- 특히 영어가 아닌 언어인 경우의 정확도는 더욱 낮아짐

Model	Active Params	MMLU	HellaS	WinoG	PIQA	Arc-e	Arc-c	NQ	TriQA	HumanE	MBPP	Math	GSM8K
LLaMA 2 7B	7B	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	17.5%	56.6%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	13B	55.6%	80.7%	72.9%	80.8%	75.2%	48.8%	16.7%	64.0%	18.9%	35.4%	6.0%	34.3%
LLaMA 1 33B	33B	56.8%	83.7%	76.2%	82.2%	79.6%	54.4%	24.1%	68.5%	25.0%	40.9%	8.4%	44.1%
LLaMA 2 70B	70B	69.9%	**85.4%**	**80.4%**	82.6%	79.9%	56.5%	25.4%	**73.0%**	29.3%	49.8%	13.8%	69.6%
Mistral 7B	7B	62.5%	81.0%	74.2%	82.2%	80.5%	54.9%	23.2%	62.5%	26.2%	50.2%	12.7%	50.0%
Mixtral 8x7B	13B	**70.6%**	84.4%	77.2%	**83.6%**	**83.1%**	**59.7%**	**30.6%**	71.5%	**40.2%**	**60.7%**	**28.4%**	**74.4%**

< Llama, Mistral, Mixtral 모델의 벤치마크별 성능 비교 >

Model	Active Params	French			German			Spanish			Italian		
		Arc-c	HellaS	MMLU	Arc-c	HellaS	MMLU	Arc-c	HellaS	MMLU	Arc-c	HellaS	MMLU
LLaMA 1 33B	33B	39.3%	68.1%	49.9%	41.1%	63.3%	48.7%	45.7%	69.8%	52.3%	42.9%	65.4%	49.0%
LLaMA 2 70B	70B	49.9%	72.5%	64.3%	47.3%	68.7%	64.2%	50.5%	74.5%	66.0%	49.4%	70.9%	65.1%
Mixtral 8x7B	13B	**58.2%**	**77.4%**	**70.9%**	**54.3%**	**73.0%**	**71.5%**	**55.4%**	**77.6%**	**72.5%**	**52.8%**	**75.1%**	**70.9%**

< Llama, Mixtral 모델의 영어 외 언어 성능 비교 >

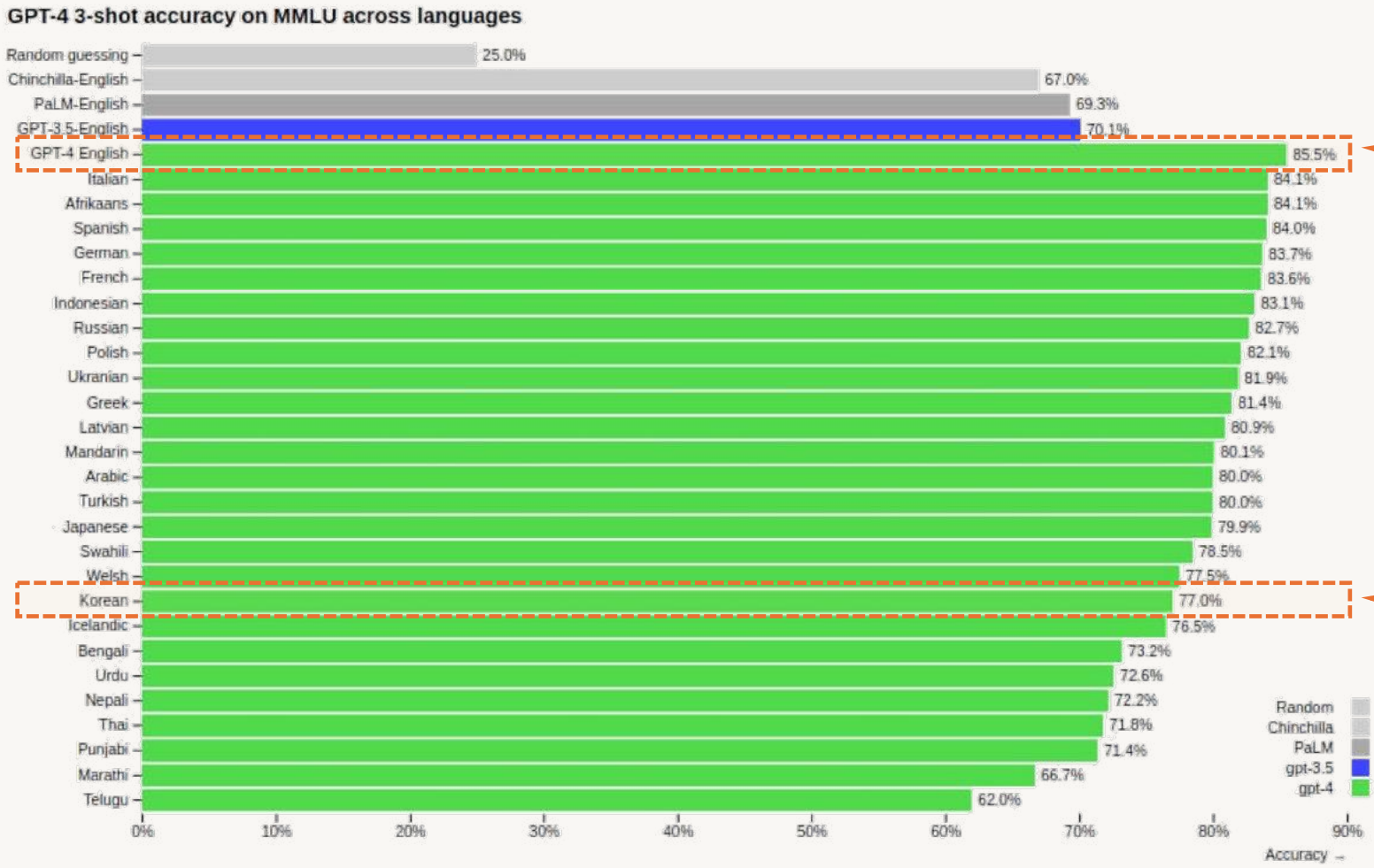
	LLaMA 2 70B	GPT-3.5	Mixtral 8x7B
MMLU (MCQ in 57 subjects)	69.9%	70.0%	**70.6%**
HellaSwag (10-shot)	**87.1%**	85.5%	86.7%
ARC Challenge (25-shot)	85.1%	85.2%	**85.8%**
WinoGrande (5-shot)	**83.2%**	81.6%	81.2%
MBPP (pass@1)	49.8%	52.2%	**60.7%**
GSM-8K (5-shot)	53.6%	57.1%	**58.4%**
MT Bench (for Instruct Models)	6.86	**8.32**	8.30

< Llama, GPT-3.5, Mixtral 모델의 벤치마크별 성능 비교 >

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... & Sayed, W. E. (2024). Mixtral of experts. arXiv preprint arXiv:2401.04088.

©2024 Databricks Inc. — All rights reserved

현세대 LLM의 한글지원은 어느정도인가?



영어를 사용했을 때의 정확도 : 85.5%

한국어를 사용했을 때의 정확도 : 77.0%

https://blog.tryamigo.com/gpt-4-vs-gpt-3-how-much-better-is-gpt-4/

©2024 Databricks Inc. — All rights reserved

GenAI의 반윤리적 사용

그 외 사례

- 미 국방부가 미사일공격을 받아 격추됐다는 가짜 AI이미지에 속아 미국 주가가 폭락한 사례
- 영국 에너지회사 CEO가 음성복제에 속아서 돈을 송금한 사례
- 영국에서 자녀 양육권 소속에서 가짜 녹취록이 재판부에 제출된적이 있다
- 여자 사진을 올리면 누드로 만들어주는 사이트가 개설 8개월만에 3800만 사람들이 방문
- 연예인 얼굴로만드는 딥페이크 포르노 사이트
- 딥페이크 기술로 프랑스 축구선수 킬리안 음바페가 이강인을 지지하는 영상을 올려 1000만 조회수 기록

©2024 Databricks Inc. — All rights reserved

인공지능 윤리기준

2020 과학기술정보통신부 인공지능 윤리기준

- 인간의 종중심주의 또는 이기주의를 표방하지 않으면서 인간성을 최고가치로 여김
- 인공지능 시대 바람직한 인공지능 개발,활용 방향을 제시하기 위함

3대 기본원칙
1. 인간의 존엄성 원칙
2. 사회의 공공선 원칙
3. 기술의 합목적성 원칙

10대 핵심요건
1. 인권 보장
2. 프라이버시 보호
3. 다양성 존중
4. 침해금지
5. 공공성
6. 연대성
7. 데이터 관리
8. 책임성
9. 안전성
10. 투명성

©2024 Databricks Inc. — All rights reserved

인공지능 윤리기준

2023 한국정보통신정책연구원 인공지능 윤리기준 실천을 위한 자율점검표

- 챗봇, 작문, 영상분야의 인공지능 윤리기준 자율점검표 제시
- 인공지능 윤리기준의 10대 핵심요건별로 35개의 점검항목을 제시

핵심요건	인권보장	프라이버시 보호	다양성 존중	침해금지	공공성	연대성	데이터 관리	책임성	안정성	투명성
챗봇	5	2	5	4	3	3	3	4	3	3
작문	5	5	6	5	4	3	3	7	3	4
영상	5	6	6	5	3	4	4	7	4	5

자율점검표 예시

- 챗봇이 허위정보나 잘못된 정보를 제공하여 인간의 생명, 신체, 정신 또는 재산에 피해를 발생시킬 우려가 있는지를 사전에 검토하고, 이를 예방하기 위한 조치를 취했는가?
- 작문용 AI의 사용 중 발생할 수 있는 위험(표절, 오정보 생성, 허위정보 생성, 과의존 등)을 사전에 평가하고, 이를 완화하기 위해 노력하고 있는가?
- AI영상시스템이 해당 사용자 이외에 다른 제3의 개인 또는 집단에 피해나 손해를 야기하는 방식으로 작동하거나 산출물을 제시한 경우, 이를 신속히 사용자 에게 고지하고 그러한 작동을 중단 및 산출물을 파기하거나 접근 제한하는 조치 등을 통하여 제3자의 피해나 손해를 방지하기 위한 노력을 하였는가?

©2024 Databricks Inc. — All rights reserved

인공지능 윤리기준

2023 한국정보통신기술협회 신뢰할 수 있는 인공지능 개발 안내서

- 신뢰할 수 있는 인공지능 개발을 위한 일반, 의료, 공공사회, 자율주행분야 안내서 제공
- 데이터 오염, 모델 편향등 인공지능이 오류를 범할 가능성을 최소화하기위한 개발 안내서

개발안내 예시

- 보호변수의 선정 이유 및 반영 여부를 설명하였는가? -> 편향적 모델 생성 방지
- 레이블링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가? -> 데이터 품질 확보
- 데이터 공격에 대한 방어 수단을 강구하였는가? -> 의도적 데이터 변질 방지
- 데이터 중독, 회피 등 공격에 대한 방어 대책을 마련하였는가? -> 의도적 모델 변질 방지
- 데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가? -> 무의식적 편향 방지
- 인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가? -> 사람에의한 편향 방지
- 활성화된 오픈소스 라이브러리를 사용하였는가? -> 신뢰할 수 있는 도구의 사용
- 모델 추출 공격에 대한 방어 방안을 수립하였는가? -> 모델 복제 방지

©2024 Databricks Inc. — All rights reserved

감사합니다

©2024 Databricks Inc. — All rights reserved

databricks

모두를 위한 데이터 인텔리전스 온라인 컨퍼런스 2024!

일시: 2024년 9월 5일 목요일 | 오전 10시 - 오후 5시

설문 참여 이벤트!

설문에 참여해주신 분들께 소정의 선물을 드립니다!

궁금하신 사항은 koreamarketing@databricks.com로 문의해 주세요.