



Part 1. AI/ML을 위한 클라우드 활용

2025. 10. 18 (토)

Yoonseok Heo (허 윤석)
nlp.ysheo419@gmail.com



LLM Agent

● 새로운 정보에 대해서 LLM은 어떻게 답변을 해줄 수 있을까?

- 질문 : 2025년 5월 8일에 NBS에서 발표한 대통령 후보 선호도 조사에서 1위한 후보는?



연합뉴스 PiCK · 1시간 전 · 네이버뉴스

:

이재명 43% 선두...한덕수 23%·김문수 12%[NBS]

차기 대통령으로 더불어민주당 이재명 대선후보가 적합하다는 응답이 43%라는 여론 조사 결과가 8일 나왔다.... 전국지표조사(NBS) 결과에 따르면 이같이 집계됐다. 무 소속 한덕수 예비후보의 차기 대통령 적합도는 23%를 기록했다. 국민의힘 김문수 ...



[속보] 이재명 43%·한덕수 23%·김문수 12%·이준석 5%[NBS] 매일경제 PiCK · 1시간 전 · 네이버뉴스

이재명 43% 선두...한덕수 23% · 김문수 12% [NBS] SBS PiCK · 58분 전 · 네이버뉴스

이재명 43%, 한덕수 23%, 김문수 12% [NBS] 경기일보 PiCK · 1시간 전 · 네이버뉴스

NBS "이재명 43%...한덕수 23%·김문수 12%" 연합뉴스TV PiCK · 44분 전 · 네이버뉴스

LLM Agent

● 새로운 정보에 대해서 LLM은 어떻게 답변을 해줄 수 있을까?

- 질문 : 2025년 5월 8일에 NBS에서 발표한 대통령 후보 선호도 조사에서 1위한 후보는?
- Deepseek-v3 모델의 답변 결과

USER

2025년 5월 8일에 NBS에서 발표한 대한민국 대통령 후보 선호도 조사에서 1위한 후보는? 한국어로 답변해줘.

ASSISTANT

2025년 5월 8일 현재, NBS에서 발표한 대한민국 대통령 후보 선호도 조사 결과는 아직 공개되지 않았습니다. 해당 날짜가 미래이므로 실제 조사 데이터가 존재하지 않기 때문입니다.

만약 가상의 시나리오나 특정 조건(예: 특정 정당 또는 인물을 상정한 설문)을 알고 계시다면 추가 정보를 주시면 답변을 조정해 드릴 수 있습니다.

현재 시점(2024년)에서 가장 최근의 여론조사나 관련 정보를 원하신다면 알려주세요!

→ Deepseek-v3 모델을 학습할 당시에는 2025년 5월 8일 뉴스 기사에 대한 데이터가 없음.
안배운 것은 당연히 모르는 것.

LLM Agent

● 새로운 정보에 대해서 LLM은 어떻게 답변을 해줄 수 있을까?

- 질문 : 2025년 5월 8일에 NBS에서 발표한 대통령 후보 적합도 조사에서 1위한 후보는?
- ChatGPT-4o 모델의 답변 결과

2025년 5월 8일에 NBS에서 발표한 대한민국 대통령 후보 적합도 조사에서 1위한 후보는? 한국어로 답변해줘.

웹 찾아보는 중

2025년 5월 8일에 발표된 전국지표조사(NBS)에 따르면, 차기 대통령 후보 적합도 조사에서 이재명 더불어민주당 후보가 43%의 지지율로 1위를 차지했습니다. 이어 한덕수 무소속 후보가 23%, 김문수 국민의힘 후보가 12%, 이준석 개혁신당 후보가 5%를 기록했습니다. 이 조사는 엠브레인퍼블릭, 케이스탯리서치, 코리아리서치, 한국리서치 등 4개 여론조사 기관이 공동으로 5월 5일부터 7일까지 전국 만 18세 이상 남녀 1,000명을 대상으로 실시하였으며, 표본오차는 95% 신뢰수준에서 ±3.1%포인트입니다.

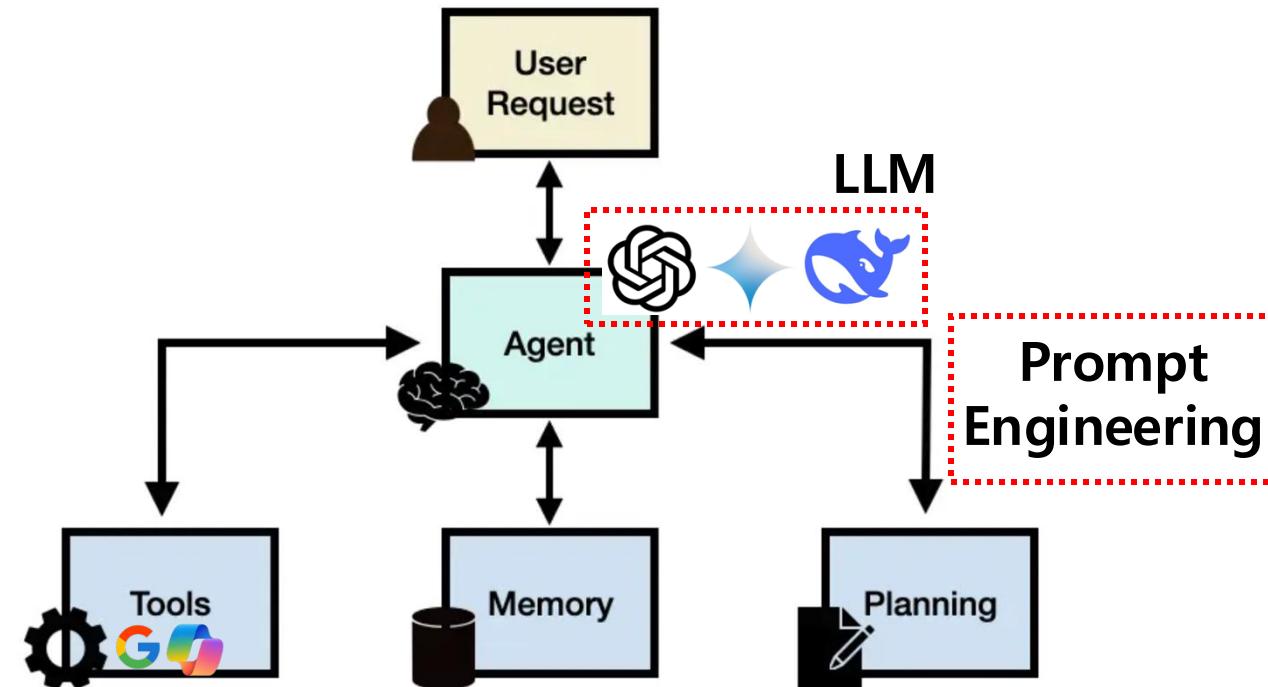
한겨레 +4 동아일보 +2

출처

→ 웹 검색 결과를 반영한 답변 생성을 통해 LLM이 학습하지 못한 지식에 대한 정확한 답변 생성

LLM Agent

- 도구 사용(Tool -Use) 및 메모리와 같은 핵심 모듈과 결합된 LLM을 통해 복잡한 작업을 수행할 수 있는 LLM 애플리케이션
 - 여기서 LLM은 작업이나 사용자 요청을 완료하는 데 필요한 작업 흐름을 제어하는 주요 컨트롤러 또는 '두뇌' 역할
 - LLM 에이전트는 계획, 메모리, 도구와 같은 다양한 핵심 모듈이 필요할 수 있음



LLM Agent

● Memory 예시: ChatGPT

- 이전 대화 내용을 별도의 메모리에 저장
→ 향후 새로운 대화 세션에서 사용자 맞춤형 답변 생성 가능

설정

일반

맞춤형 지침

x

개인 맞춤 설정

말하기

데이터 제어

빌더 프로필

연결된 앱

보안

메모리

x

ChatGPT는 세부 내용과 선호 사항을 파악하여 사용자님께 맞는 응답을 찾고자 하므로, 채팅을 하면 할수록 더 유용해집니다. 자세히 알아보기

ChatGPT가 어떤 것을 기억하는지 알고 싶거나 ChatGPT에게 새로운 것을 가르치고 싶다면, ChatGPT와 채팅하세요.

- “나는 간결한 답변이 좋아. 기억해 둬.”
- “얼마 전에 강아지를 입양했어!”
- “나에 대해 어떤 걸 기억하고 있어?”
- “최근에 하던 프로젝트를 어디까지 했더라?”

관리

ChatGPT 메모리 지우기

B You

I live with a golden retriever named Ellie and a Maine Coon cat named Teddy

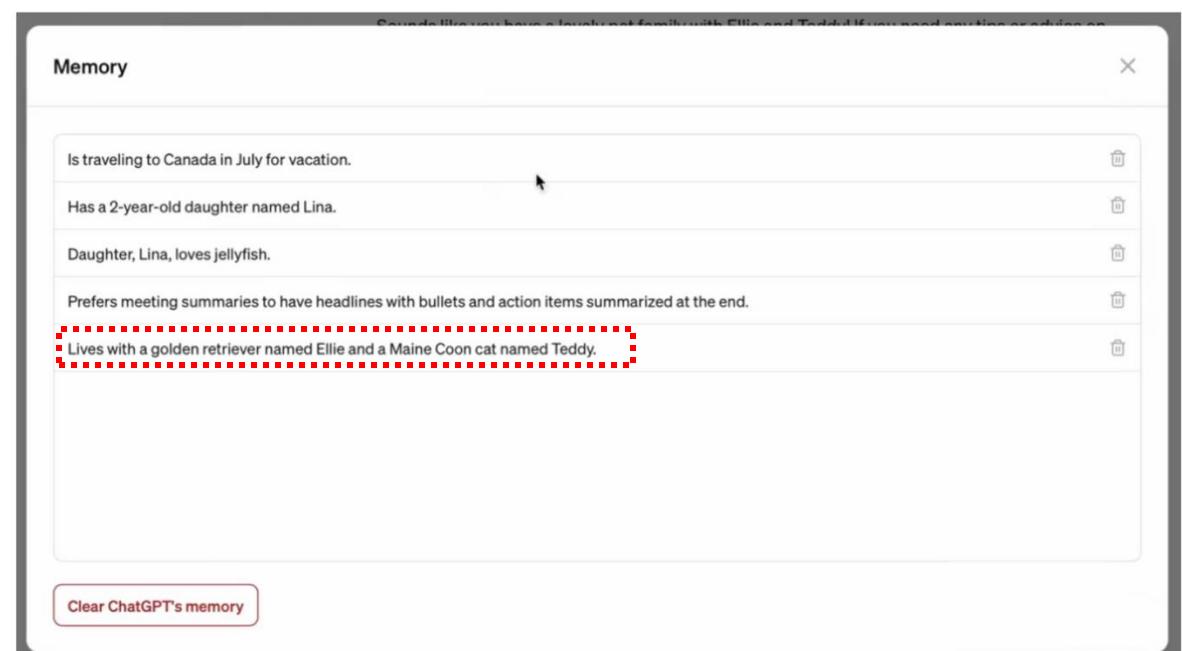
ChatGPT

Memory updated

Lives with a golden retriever named Ellie and a Maine Coon cat named Teddy.

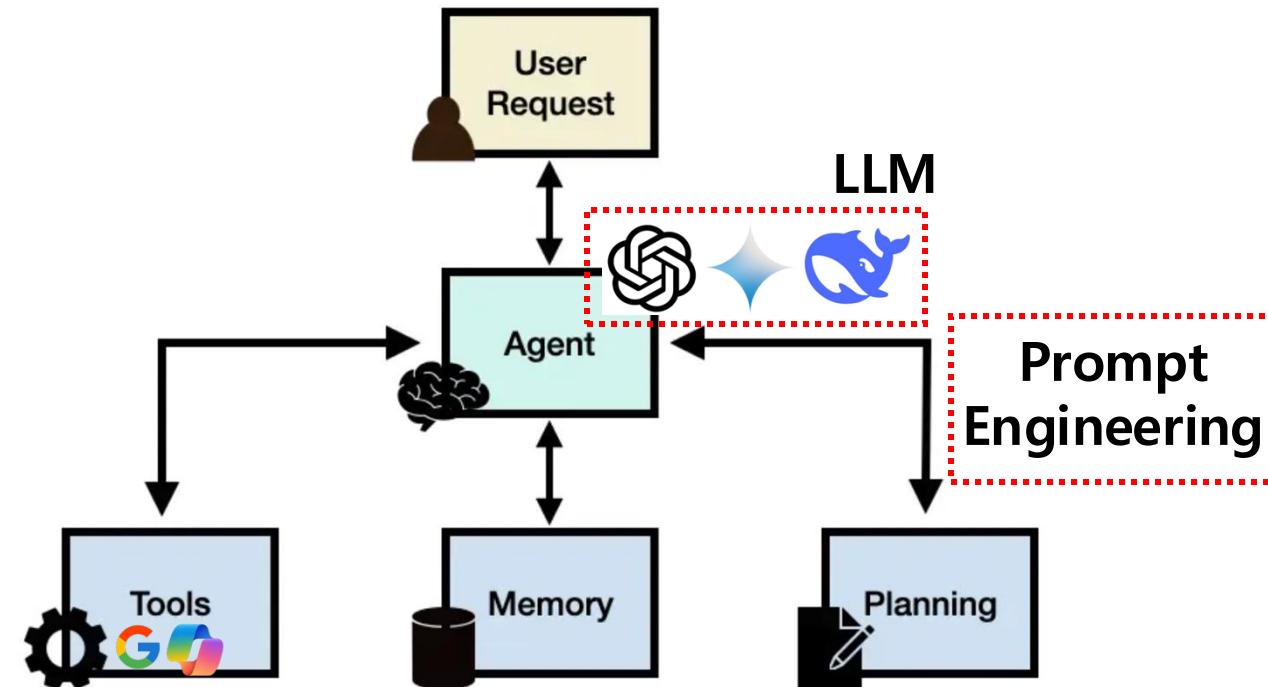
Manage memories

ny tips or advice on



LLM Agent

- 도구 사용(Tool -Use) 및 메모리와 같은 핵심 모듈과 결합된 LLM을 통해 복잡한 작업을 수행할 수 있는 LLM 애플리케이션
 - 여기서 LLM은 작업이나 사용자 요청을 완료하는 데 필요한 작업 흐름을 제어하는 주요 컨트롤러 또는 '두뇌' 역할
 - LLM 에이전트는 계획, 메모리, 도구와 같은 다양한 핵심 모듈이 필요할 수 있음



From Google to ChatGPT: How LLMs Changed Everything

- Just 3 years ago, this was how we found answers



Source: Generated by ChatGPT

Just Google it!

From Google to ChatGPT: How LLMs Changed Everything



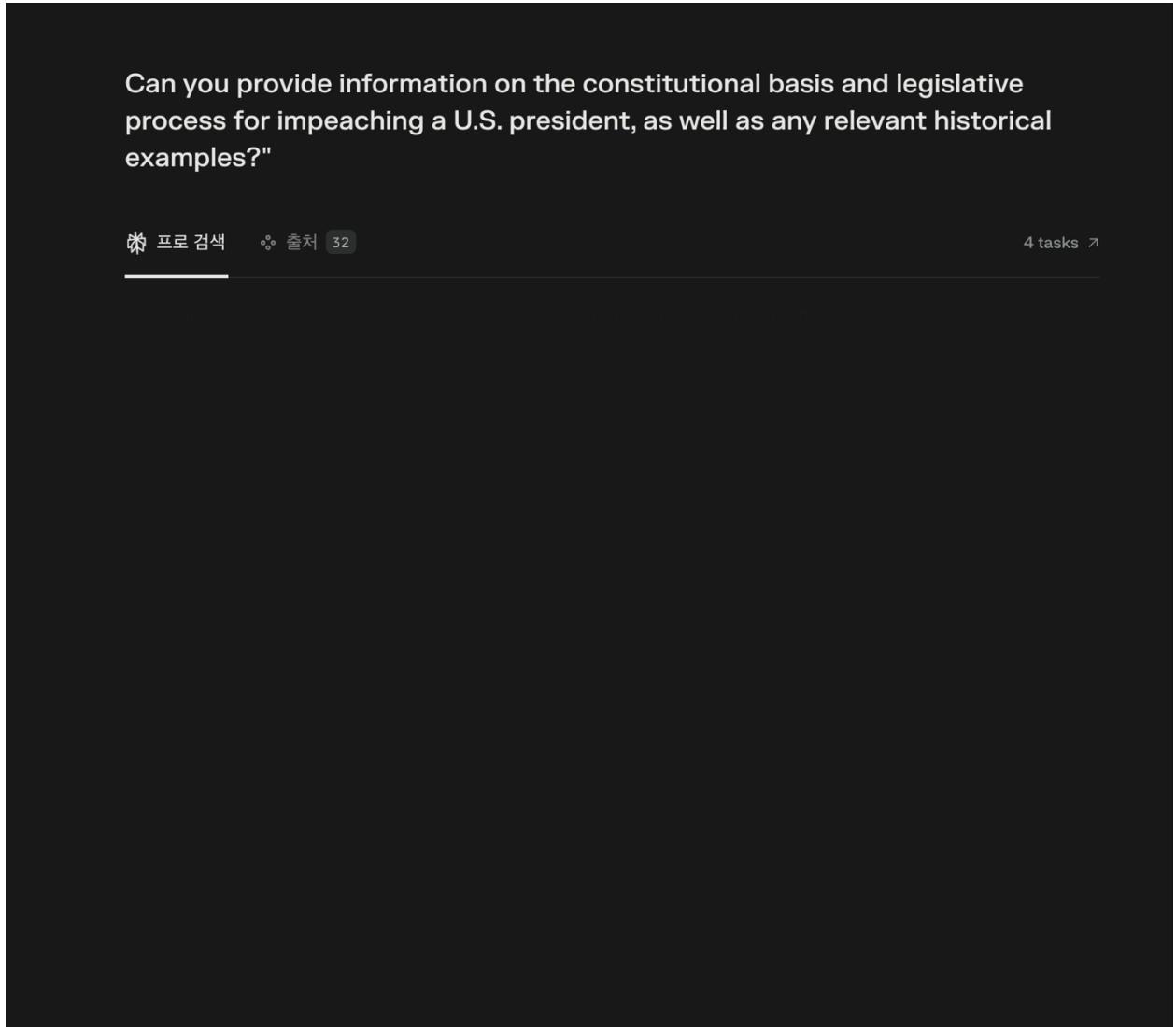
Source: Generated by ChatGPT

No more googling!

Asking AI

Beyond ChatGPT: LLM-based application (1)

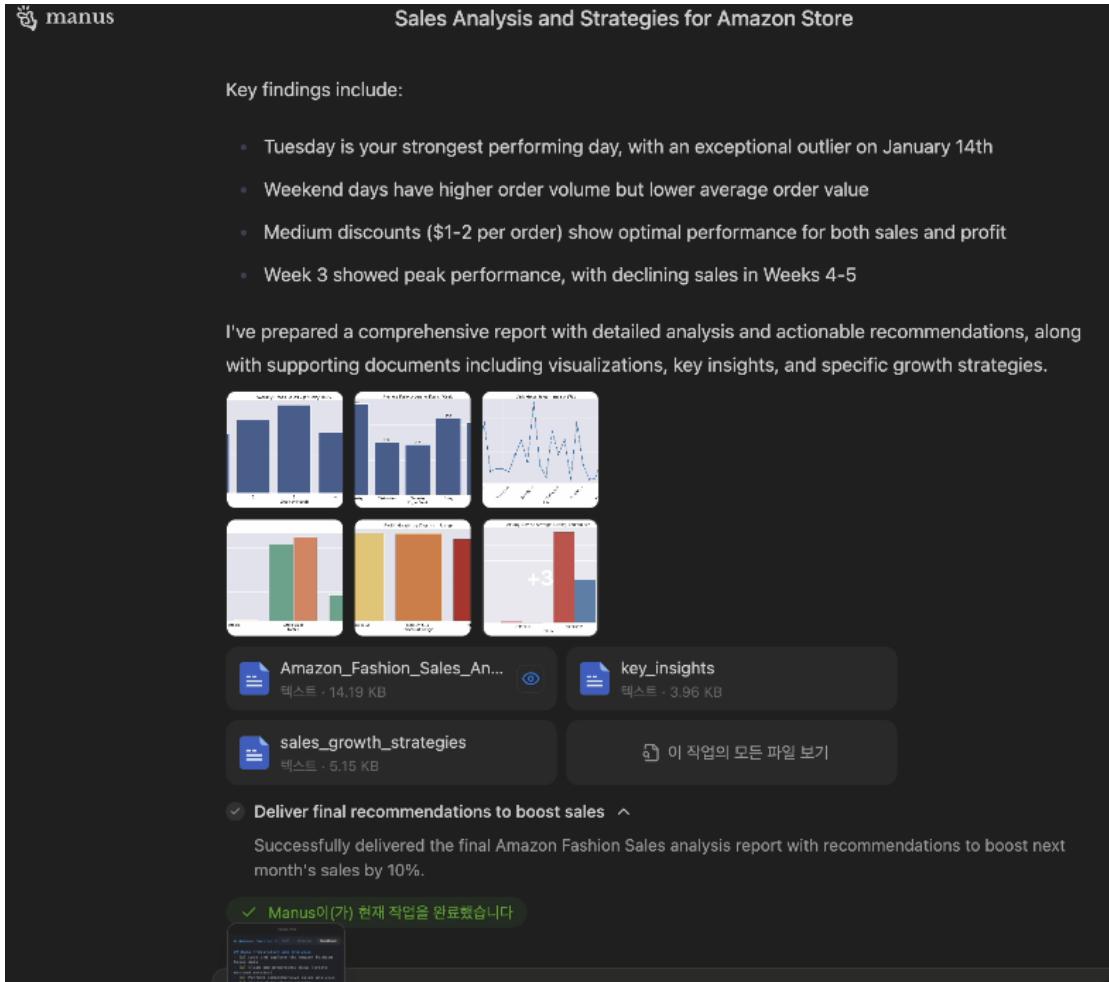
- Perplexity AI
 - Incorporating **Search tools** with ChatGPT
→ More trustworthy results



Source: Perplexity AI

Beyond ChatGPT: LLM-based application (2)

- Automatic Sales Report Generation: Manus AI
 - From data collection to writing analytic report



The screenshot shows the Manus AI platform interface. On the left, there's a sidebar with a user icon, a search bar, and a navigation menu. The main area displays a "Sales Analysis and Strategies for Amazon Store" dashboard. It features a section titled "Key findings include:" with a bulleted list of insights about sales performance, including days of the week, order volume, and discounts. Below this is a section for "actionable recommendations" with a link to a file named "Amazon_Fashion_Sales_An...". At the bottom, there are sections for "Deliver final recommendations to boost sales" and a success message: "Manus이(가) 현재 작업을 완료했습니다". On the right side of the dashboard, there are several small thumbnail images of different data visualizations like bar charts and line graphs.

Sales Analysis and Strategies for Amazon Store

Key findings include:

- Tuesday is your strongest performing day, with an exceptional outlier on January 14th
- Weekend days have higher order volume but lower average order value
- Medium discounts (\$1-2 per order) show optimal performance for both sales and profit
- Week 3 showed peak performance, with declining sales in Weeks 4-5

I've prepared a comprehensive report with detailed analysis and actionable recommendations, along with supporting documents including visualizations, key insights, and specific growth strategies.

Amazon_Fashion_Sales_Analysis_Report

Amazon Fashion Sales Analysis and Growth Strategy

January 2025 Performance Review and 10% Growth Plan

Executive Summary

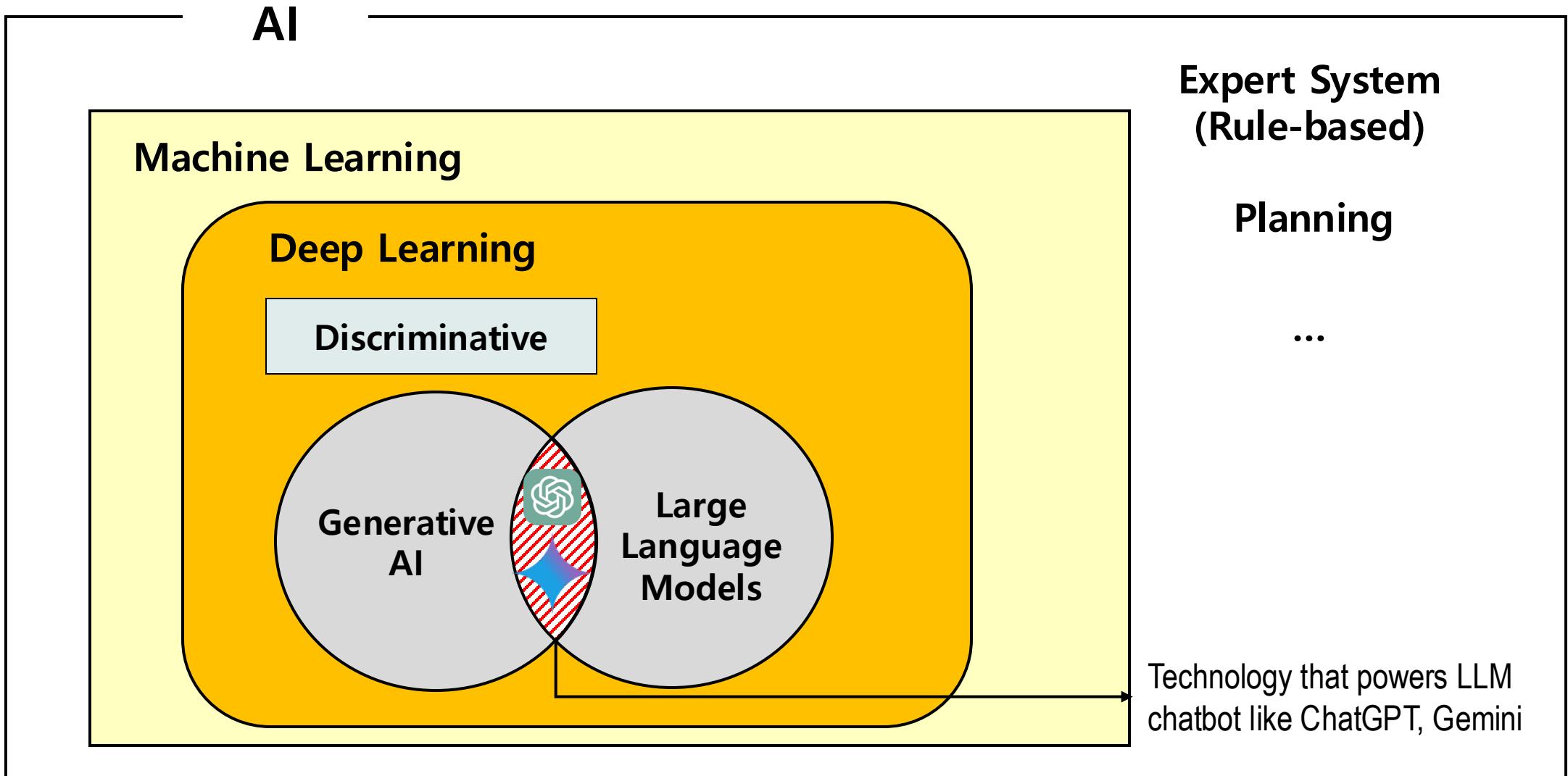
This comprehensive analysis examines the January 2025 sales data from your Amazon Fashion store to identify patterns, trends, and opportunities for growth. The analysis reveals several key insights, including day-of-week performance variations, the impact of discount strategies, and customer purchasing behaviors. Based on these findings, we've developed a multi-faceted strategy projected to increase sales by 14%, well above the requested 10% target. The strategy focuses on optimizing day-specific promotions, enhancing customer purchase behavior, refining pricing and promotion tactics, and improving inventory management. This report provides both detailed analysis and actionable recommendations with implementation timelines to achieve sustainable growth.

Data Analysis Methodology

The analysis began with a thorough exploration of the January 2025 sales data, which included 31 days of daily metrics covering orders, customers, units sold, and various financial indicators. The data was preprocessed to create derived metrics such as average order value, units per order, profit margins, and day-of-week features. Comprehensive statistical analysis was performed to identify patterns, correlations, and anomalies in the data. Multiple visualizations were created to illustrate key trends and relationships, providing a foundation for data-driven strategy development.

AI vs Deep Learning vs Machine Learning vs Gen AI

- AI Hierarchy



최종 프로젝트: 금융 서비스를 위한 Agentic AI

Week 1: 예측 모델의 서비스화 (MLOps Foundation)

- 주요 역량: 클라우드 기반 모델 개발·배포·API화
- 결과물: 🚀 AI 대출 심사 API



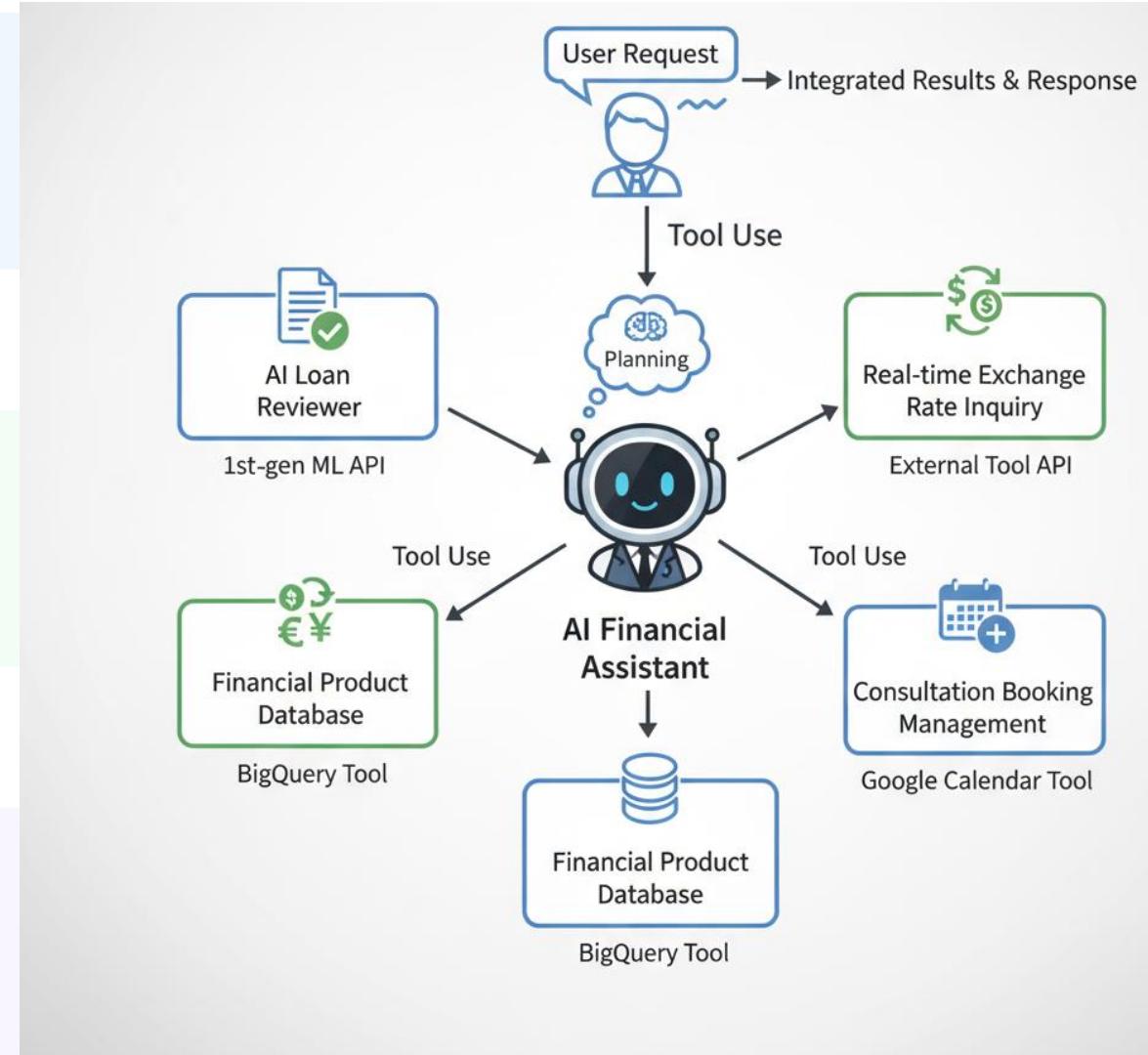
Week 2: 에이전트 기반 워크플로우 구축 (Agentic Workflow)

- 주요 역량: LLM 에이전트 설계·A2A(Agent-to-Agent) 협업 구현
- 결과물: 💬 대화형 금융 상담 챗봇



Week 3: 외부 시스템 연동 및 자동화 (System Integration)

- 주요 역량: 멀티-툴 활용·실제 데이터 연동 기반 업무 자동화
- 결과물: 💡 AI 개인 금융 비서



1차시 교육 목표

- 클라우드 기반 대출 심사 예측 모델 API 구축
- 학습 목표
 - Vertex AI 활용 MLOps 핵심 파이프라인 경험 (개발→저장→등록→배포)
 - 금융 데이터 기반 신용 리스크 평가 머신러닝 모델 구현
 - 범용 호출 가능한 표준 REST API 형태 모델 서비스화

1차시 교육 커리큘럼 상세

● Cloud 상에서 ML 모델의 API화

세션 1

Vertex AI 개요 및 환경 설정

세션 2

Vertex AI Workbench: 클라우드 통합 개발 환경

세션 3

모델 학습(Training): 데이터 기반 패턴 학습·모델 생성

세션 4

모델 등록(Registration): GCS 연동·모델 버전 관리

세션 5

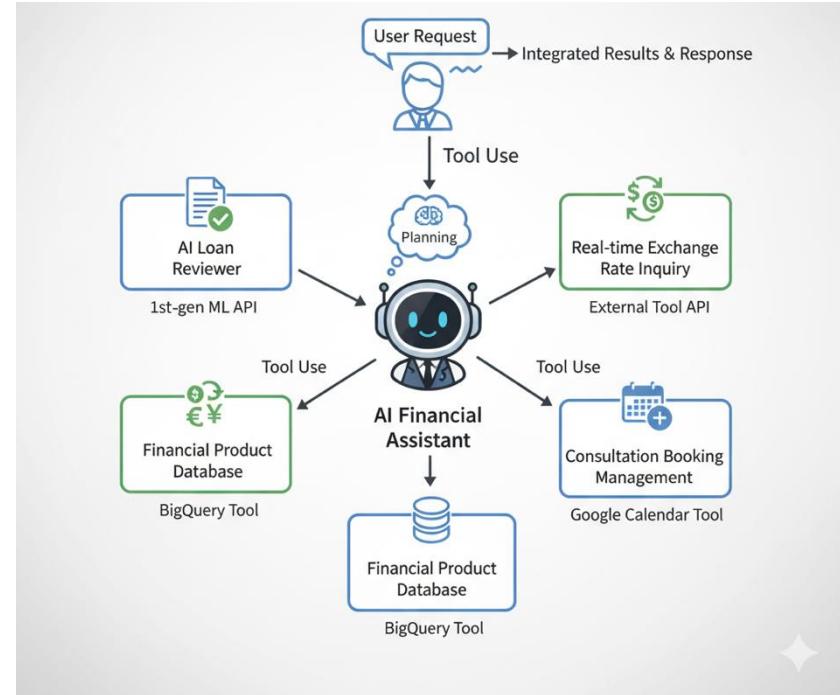
모델 배포(Deployment): Endpoint 생성 통한 서비스화

세션 6

API 테스트: SDK 이용 Endpoint 호출·결과 검증

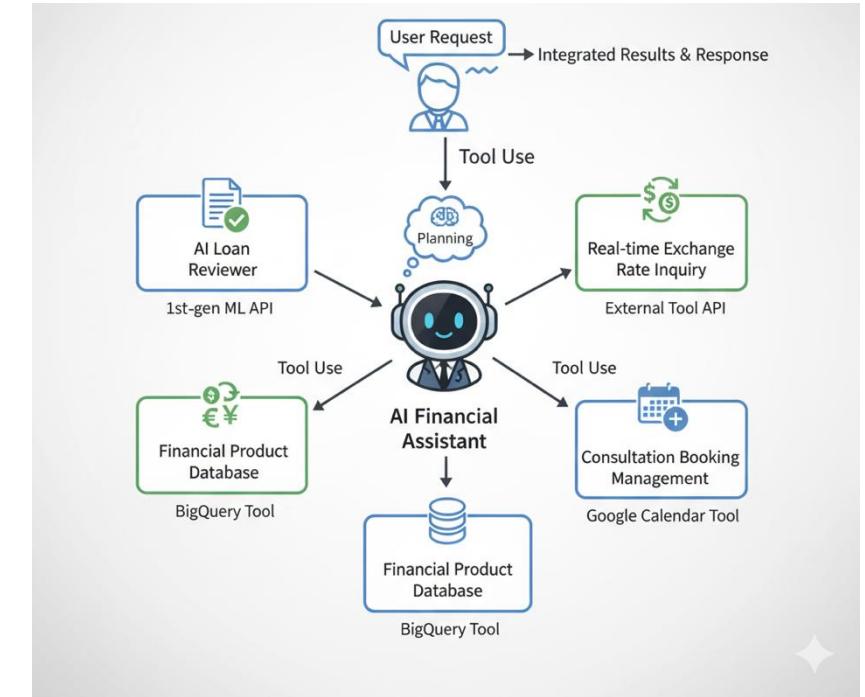
2차시 교육 목표

- 지능형 에이전트 설계
 - A2A(Agent-to-Agent) Collaboration
 - 1차시에서 확보한 API 한계
 - 데이터 기반 의사 결정 가능
 - 대화 기반 자율적 작업 수행 불가
 - 2차시 실습 목표
 - 1차시에서 확보한 대출 심사 예측 API를
 - Gemini 모델 기반 "매니저 Agent" 설



3차시 교육 목표

- MCP를 통한 시스템 통합 및 자동화
 - 2차시 교육에서 확보한 Agent 시스템에 다중 기능 및 업무 수행 능력 부여
 - 연동 시스템
 - BigQuery: 고객 데이터 조회
 - 외부 금융 API: 실시간 정보 분석
 - Google Calendar API: 상담 예약 자동화
 - 실제 비즈니스 가치 창출하는 'Task Automation' 구현



3차시 교육 목표

Getting Started

지금부터 Vertex AI 플랫폼에서 실습을 시작하겠습니다

(<https://cloud.google.com/?hl=ko>)

과금 설정 막는 방법

● 상단 메세지

무료 체험판 상태: ₩423,000.56 크레딧 및 91일이 남았습니다. 일반 계정을 활성화하여 Google Cloud의 모든 기능에 무제한 액세스하세요. 남아 있는 크레딧을 사용한 다음, 사용한 만큼만 비용을 지불하면 됩니다.

- 활성화 버튼 클릭 → 무료 크레딧을 다 사용하면 결제 카드로 과금이 자동으로 시작됨
- 활성화 시 혜택

- GPU를 자유롭게 사용 가능
 - 자유롭게 := 여러 종류
 - 주의: 사용량이 많다는 의미가 아님. 사용량이 많아지면 크레딧이 소모됨.
- 무료 크레딧 사용 후에도 서비스가 유지 됨

주의!!!! 본인의 책임 하에 과금 되지 않게 항상 인스턴스 관리를 잘 하시기 바랍니다.

과금 설정 막는 방법

● 막는 방법 1: 예산 알림 설정

- 설정한 사용 금액에 도달하면 이메일로 경고를 보내줌
- Step 1: Google Cloud Console에서 왼쪽 메뉴의 **'결제'**로 이동합니다.
- Step 2: 왼쪽 메뉴에서 **'예산 및 알림'**을 선택합니다.
- Step 3: '예산 만들기' 버튼을 클릭합니다.
- Step 4: 이름을 지정합니다 (예: '무료 크레딧 한도 알림').
- Step 5: 적용 대상은 기본값(모든 프로젝트, 모든 서비스)으로 둡니다.

과금 설정 막는 방법

● 막는 방법 1: 예산 알림 설정

- Step 6: 금액에서 '예산 유형'을 '지정된 금액'으로 설정하고,
목표 금액을 남은 크레딧보다 약간 적게 입력 (예: ₩400,000)
- Step 7: 작업 설정이 매우 중요
 - '실제 비용이 예산의 n%일 때'를 기준으로 알림을 설정
 - 50%, 90%, 그리고 가장 중요한 100%에 체크
 - 100%는 설정한 ₩420,000을 모두 사용했다는 의미
 - '이메일 알림 관리'를 통해 알림을 받을 이메일 주소를 지정
- **효과:** 이제 크레딧을 50%, 90%, 100% 소진할 때마다 경고 이메일을 받게 됨. 하지만 이것만으로는 과금이 멈추지 않음

- 범위
- |
- 금액
- |
- ③ 작업

Set alert threshold rules

실제 지출액이나 예상 지출액이 예산의 일정 비율이나 지정된 금액을 초과하면 이메일 알림을 보냅니다. [자세히 알아보기](#)

예산 비율 1 *	금액 1 * ₩ 190000	트리거 기준 1 실제 지출
50 %	트리거 기준 2 실제 지출	?
예산 비율 2 *	금액 2 * ₩ 285000	트리거 기준 3 실제 지출
75 %	트리거 기준 4 실제 지출	?
예산 비율 3 *	금액 3 * ₩ 342000	?
90 %	트리거 기준 4 실제 지출	?
예산 비율 4 *	금액 4 * ₩ 361000	?
95 %	+ 기준 추가	

알림 관리

- 결제 관리자 및 사용자에게 이메일 알림
- 프로젝트 소유자에게 이메일 알림 ?
- 이 예산에 모니터링 이메일 알림 연결

프로젝트 1개와 최대 5개의 모니터링 이메일 알림 채널을 선택하세요.
- Pub/Sub 주제를 이 예산에 연결

프로젝트 및 Pub/Sub 주제를 선택하세요. 이 예산을 볼 수 있는 모든 사용자가 프로젝트 ID 및 주제 이름도 볼 수 있습니다. [도메인 제한 공유](#) 를 사용 설정한 조직에 속한 경우 Pub/Sub 주제를 추가하지 못할 수도 있습니다.

과금 설정 막는 방법

● 막는 방법 2: 본인이 설정한 최대 사용액 알람이 오면 수동으로 결제 해제

- 1단계 알림을 받고 자동화 설정이 불안하다면, 이메일을 확인하는 즉시 수동으로 조치 가능
- 실행 중인 리소스 중지/삭제
 - 가장 많은 비용을 유발하는 VM 인스턴스(컴퓨터), 데이터베이스 등을 직접 중지하거나 삭제
- 프로젝트에서 결제 계정 연결 해제: 이게 가장 확실한 수동 조치입니다.
 - '결제' 메뉴에서 **'계정 관리'**로 이동합니다.
 - 과금을 막고 싶은 프로젝트 오른쪽에 있는 점 3개 메뉴(:)를 클릭합니다.
 - '결제 사용 중지'를 선택합니다.

➔ 프로젝트와 결제 계정의 연결이 끊어지므로, 해당 프로젝트에서는 더 이상 어떠한 비용도 발생하지 않음 (단, 서비스도 모두 중지됨)

과금 설정 막는 방법

● 막는 방법 2: 본인이 설정한 최대 사용액 알람이 오면 수동으로 결제 해제

The screenshot shows the Azure Billing Management interface. On the left, there's a sidebar with various navigation options like '결제 계정', '비용 관리', '예산 및 알림', etc. At the bottom of the sidebar, the '계정 관리' button is highlighted with a red box. In the main content area, it shows a summary for a billing account with ID 019017-A1C8ED-AC946F, categorized under '직접'. Below that, it lists '이 결제 계정에 연결된 프로젝트' (Connected projects) with one entry: 'My First Project' (Project ID: sigma-nimbus-475403-v9). A context menu is open over this project entry, also enclosed in a red box. The menu items are: '결제 사용 중지' (Stop payment), '결제 변경' (Change payment), and '결제 잠금' (Lock payment).

과금 설정 막는 방법

- 막는 방법 3: 한도액 도달 시 자동으로 결제 중지 (가장 이상적임)
 - 스크립트를 통해 한도액 도달 시 관리자 허가 이후부터 과금 되도록 시스템 설정
 - https://cloud.google.com/billing/docs/how-to/disable-billing-with-notifications?hl=ko#functions_cap_billing_dependencies-python

VM Instance를 만들면 과금의 시작!

- VM Instance를 사용하지 않을 때에는 반드시 중지버튼 클릭하기!
- 활성화 상태

인스턴스 이름 ↑	영역	자동 업그레이드	버전	머신 유형
my-instance JupyterLab 열기	asia-northeast3-a	-	M134	Efficient Instance: vCPU 4 개, 16GB RAM

- 사용하지 않는 인스턴스는 반드시 중지!

X 인스턴스 1개 선택됨 ▶ 시작 ⚡ 재설정 ■ 중지 ━ 삭제

인스턴스 실행 일정

View: 인스턴스 사용자 관리 노트북 관리형 노트북

■ 이제 Vertex AI Workbench에서 JupyterLab 4를 사용할 수 있습니다.

인스턴스 이름 ↑	영역
my-instance JupyterLab 열기	asia-northeast3-a

Workbench 인스턴스에는 JupyterLab 3이 사전 설치되어 있으며 GPU 지원 머신러닝 프레임워크로 구성되어 있습니다. [자세히 알아보기](#)

≡ 필터

인스턴스 이름 ↑	영역	자동 업그레이드	버전
my-instance	asia-northeast3-a	-	M134

Vertex AI Workbench: 엔터프라이즈급 ML 개발 환경

- Google Cloud Vertex AI 플랫폼에 통합되어 End-to-End 머신러닝 워크플로우를 지원하는 관리형 JupyterLab 환경 → 진정한 MLOps 플랫폼

구분	Vertex AI Workbench (관리형/사용자 관리형)	Colab Enterprise
통합 및 확장성	Vertex AI와의 완벽한 통합 (AutoML, Feature Store, Model Registry, Experiments, Endpoint 등) 제공.	개별 노트북 실행 환경 제공에 집중 (Vertex AI와의 통합은 별도 코딩 필요)
인프라 제어 및 커스터마이징	높은 제어력: VM 탑재, OS 이미지, 네트워크 구성(VPC-SC), 보안 정책 등 인프라를 직접 선택 및 관리 가능.	제한적 제어력: 런타임 환경은 Cloud Console에서 설정된 템플릿에 의존하며, 인프라 수준의 커스터마이징은 어려움.
활용 시나리오	프로덕션 ML 파이프라인 개발, 보안 요구사항이 높은 환경, 커스텀 인프라 요구되는 대규모 프로젝트.	탐색적 데이터 분석(EDA), 개념 증명(PoC), 소규모 모델 개발 및 교육 목적.

Vertex AI Workbench: 엔터프라이즈급 ML 개발 환경

- Google Cloud Vertex AI 플랫폼에 통합되어 End-to-End 머신러닝 워크플로우를 지원하는 관리형 JupyterLab 환경 → 진정한 MLOps 플랫폼

구분	Vertex AI Workbench (관리형/사용자 관리형)	Colab Enterprise
보안 및 규정 준수	엔터프라이즈 보안 강화: VPC 서비스 제어(VPC-SC) 및 Private IP 전용 모드를 통해 데이터 격리 및 엄격한 규정 준수 환경 구축에 용이.	Google Drive 기반 접근 사용자 자격증명 기반의 접근 방식이며, 대규모 데이터 격리 및 복잡한 네트워크 보안 요구사항 충족에 한계가 있음.
지속적인 실행 (Scheduled Execution)	직접적인 스케줄링 및 파이프라인 연동: Notebook Executor를 통한 간편 실행 예약 및 Vertex AI Pipelines으로의 전환이 자연스러움.	스케줄링 기능 존재: 간편 예약 실행은 가능하나, 장기적이고 복잡한 워크플로우 제어는 Vertex AI Pipelines 으로 별도 구성 필요.

Vertex AI Workbench – Instance 만들기

- 하나의 인스턴스 만들기: 가상 컴퓨터 환경 (Virtual Machine, VM)

The screenshot shows the Vertex AI Workbench interface. On the left sidebar, under the 'Model Development' section, the 'Workbench' option is highlighted with a red box. The main area is titled 'Workbench' and shows the 'Instances' tab selected (also highlighted with a red box). A message indicates that JupyterLab 4 is available. Below the message is a search bar and filter options. A large icon of a computer monitor with a cloud inside is displayed. At the bottom right, there is a blue button labeled 'Create New' (also highlighted with a red box).

1

2

3

새 인스턴스

4 이름 *

영문자로 시작해야 합니다. 이어서 최대 47자(영문 기준)의 소문자, 숫자 또는 하이픈(-)이 와야 하며 하이픈으로 끝나면 안 됩니다.

5 리전 * 영역 *

Apache Spark 및 BigQuery 커널 사용 설정

이 프로젝트의 네트워크 공유 네트워크

네트워크

서브네트워크 *

인스턴스 속성

머신 유형	e2-standard-4
데이터 디스크	균형 있는 영구 디스크 100GB
권한	Compute Engine 기본 서비스 계정
예상 비용 <small>(?)</small>	US\$186.00/월, US\$0.25/시간

6 고급 옵션 취소 만들기

Vertex AI Workbench – Instance 만들기

● 고급 옵션에서 단계별 상세 설정

• 세부정보

• 환경

• 머신 유형

• 디스크

• 네트워킹

• IAM 및 보안

• 시스템 상태

이름 * my-instance
영문자로 시작해야 합니다. 이어서 최대 47자(영문 기준)의 소문자, 숫자 또는 하이픈(-)이 와야 하며 하이픈으로 끝나면 안 됩니다.

리전 * asia-northeast3 (Seoul) ▾ ⓘ 영역 * asia-northeast3-a ▾ ⓘ

Apache Spark 및 BigQuery 커널 사용 설정

라벨

+ 라벨 추가

네트워크 태그

네트워크 태그

Workbench 리소스에 태그를 할당합니다. [자세히 알아보기](#) ↗

계속

Vertex AI Workbench – Instance 만들기

● 고급 옵션에서 단계별 상세 설정

← 인스턴스 만들기

세부정보

- 환경
- 머신 유형
- 디스크
- 네트워킹
- IAM 및 보안
- 시스템 상태

환경

모든 환경은 기본적으로 JupyterLab 3을 사용하며 최신 NVIDIA GPU 및 Intel 라이브러리와 드라이버가 설치되어 있습니다. 대신 이전 버전을 지정할 수 있습니다. [자세히 알아보기](#)

JupyterLab 버전

JupyterLab 3.x

JupyterLab 4.x [신규](#)

커스텀 컨테이너 사용

버전

최신 버전 사용

이전 버전 사용

특정 버전에 대한 자세한 내용은 [Vertex AI Workbench 출시 노트](#)를 참조하세요.

시작 후 스크립트

Path to post-startup script [찾아보기](#)

인스턴스 부팅 후 자동으로 실행되는 Cloud Storage 스크립트의 경로

메타데이터

data-disk-uri, framework, notebooks-api, notebooks-api-version, nvidia-driver-gcs-path, proxy-url, restriction, shutdown-script, title, version 등의 일부 메타데이터 키는 시스템 사용 전용으로 예약되어 있습니다. 이러한 변수 이름을 아래에 사용하면 시스템 값으로 덮어쓰게 됩니다.

[+ 메타데이터 추가](#)

[뒤로](#) [계속](#)

1

Vertex AI Workbench -

● 고급 옵션에서 단계별

[← 인스턴스 만들기](#)

- 세부정보
- 환경
- 머신 유형

- 디스크
- 네트워킹
- IAM 및 보안
- 시스템 상태

2

1

머신 유형

범용 GPU

일반적인 작업 부하에 적합한 머신 유형이며 가격 및 유연성을 위해 최적화되었습니다.

Series	설명	vCPUs	Memory
<input checked="" type="radio"/> E2	저렴한 비용, 일상적인 컴퓨팅 처리	2 - 32	4~128GB
<input type="radio"/> N2	균형을 이룬 가격과 성능	2 - 128	4~864GB
<input type="radio"/> N2D	균형을 이룬 가격과 성능	2 - 224	4~896GB
<input type="radio"/> N1	균형을 이룬 가격과 성능	2 - 96	3.6~624GB

머신 유형 — e2-standard-4(vCPU 4개, 코어 2개, 메모리 16GB)



vCPU
4

Memory
-

▼ CPU 플랫폼 및 GPU

예약

사용 안함

▼ (?)

이 노트북을 만들 때 기존 Compute Engine 예약을 사용하세요.

보안 VM

가장 안전한 구성을 위해 모든 설정을 사용 설정하세요. [자세히 알아보기](#)

보안 부팅

부팅 수준 및 커널 수준의 멀웨어로부터 보호하기 위해 모든 부팅 구성요소의 디지털 서명을 확인합니다. GPU가 있는 인스턴스에서는 지원되지 않습니다.

vTPM(Virtual Trusted Platform Module)

게스트 VM 사전 부팅 및 부팅 무결성을 검증하고 키 생성 및 보호를 제공합니다.

무결성 모니터링

Stackdriver를 사용하여 런타임 부팅 무결성을 모니터링하고 검증합니다. vTPM이 필요합니다.

유형 상태 종료

유형 상태 종료 사용 설정

종료 전 비활성 시간(분) *

30

범위: 10분~1일(1,440분)

뒤로

계속

Pricing summary

월 US\$186.00 예상

시간당 약 \$0.255

사용한 만큼만 비용 지불: 선불 비용 없이 초당 청구

네트워킹 비용도 적용됩니다. [자세히 알아보기](#)

▼ 세부정보

Vertex AI Workbench – Instance 만들기

● 고급 옵션에서 단계별 상세 설정

[← 인스턴스 만들기](#)

세부정보
 환경
 머신 유형
 디스크
 네트워킹
 IAM 및 보안
 시스템 상태

네트워킹

이 인스턴스를 사용하려면 인터넷 액세스가 가능해야 합니다. 다음 중 하나가 선택되어 있는지 확인하세요. [자세히 알아보기](#)

- 외부 IP 주소 할당
- 인터넷에 액세스할 수 있는 네트워크를 선택합니다.
- [비공개 Google 액세스](#)를 사용 설정합니다.

이 프로젝트의 네트워크
 공유 네트워크

네트워크 ━ default

서브네트워크 * ━ default(10.178.0.0/20)

외부 IP 주소 할당
인스턴스가 인터넷 또는 다른 VPC 네트워크와 통신할 수 있도록 합니다.

프록시 액세스 허용
프록시 액세스가 사용 중지된 경우 SSH를 통해 인스턴스에 액세스해야 합니다. [자세히 알아보기](#)

[뒤로](#) [계속](#) 1

Vertex AI Workbench – Instance 만들기

● 고급 옵션에서 단계별 상세 설정

← 인스턴스 만들기

✓ 세부정보
✓ 환경
✓ 머신 유형
✓ 디스크
✓ 네트워킹
• IAM 및 보안
• 시스템 상태

IAM 및 보안

인스턴스의 JupyterLab 인터페이스를 사용할 수 있는 사용자를 결정합니다. 인스턴스를 만든 후에는 변경할 수 없습니다. [자세히 알아보기](#)

1 서비스 계정
iam.serviceAccounts.actAs가 있는 사용자는 누구나 인스턴스 계정에 액세스할 수 있습니다.

단일 사용자
사용자 1명으로 액세스 제한

Colab은 로그인한 단일 유저만 가능

VM에서 기본 Compute Engine 서비스 계정을 사용하여

기본 Compute Engine 서비스 계정을 변경했다면 충분한 API 권한을 보유하고 있는지 확인하세요. [자세히 알아보기](#)

보안 옵션

인스턴스에 대한 루트 액세스
 nbconvert
다른 파일 형식으로 노트북을 내보내고 다운로드합니다.
 파일 다운로드
JupyterLab에서 다운로드 허용
 터미널 액세스
JupyterLab에서 셀 명령어 실행

2

뒤로 **계속**

Vertex AI Workbench – Instance 만들기

● 고급 옵션에서 단계별 상세 설정

[← 인스턴스 만들기](#)

<input checked="" type="checkbox"/> 세부정보	<p>시스템 상태</p> <p><input type="checkbox"/> 환경 자동 업그레이드 실행 중인 인스턴스를 자동으로 업그레이드합니다. 자세히 알아보기</p> <p>Reporting</p> <p><input checked="" type="checkbox"/> 시스템 상태 보고 핵심 서비스의 상태를 확인합니다. 자세히 알아보기</p> <p><input type="checkbox"/> Cloud Monitoring에 커스텀 측정항목 보고 시스템 상태 및 JupyterLab 측정항목을 수집합니다. 자세히 알아보기</p> <p><input type="checkbox"/> Cloud Monitoring 설치 시스템 및 애플리케이션 측정항목(디스크, CPU, 네트워크, 프로세스)을 보고합니다. 자세히 알아보기</p> <p><input checked="" type="checkbox"/> 필수 Google 도메인의 DNS 상태 보고 프록시 등록 및 상태 이벤트 보고를 위해 도메인의 DNS 상태를 확인합니다.</p>
<input checked="" type="checkbox"/> 환경	
<input checked="" type="checkbox"/> 머신 유형	
<input checked="" type="checkbox"/> 디스크	
<input checked="" type="checkbox"/> 네트워킹	
<input checked="" type="checkbox"/> IAM 및 보안	
<input type="radio"/> 시스템 상태	

[뒤로](#)

1
[만들기](#) 취소

만들기 클릭 후 시간 다수 소요

Vertex AI Workbench – Instance 만들기

- Instance 생성 완료 후 JupyterLab 열기

View: 인스턴스 사용자 관리 노트북 관리형 노트북

ⓘ 이제 Vertex AI Workbench에서 JupyterLab 4를 사용할 수 있습니다. 닫기

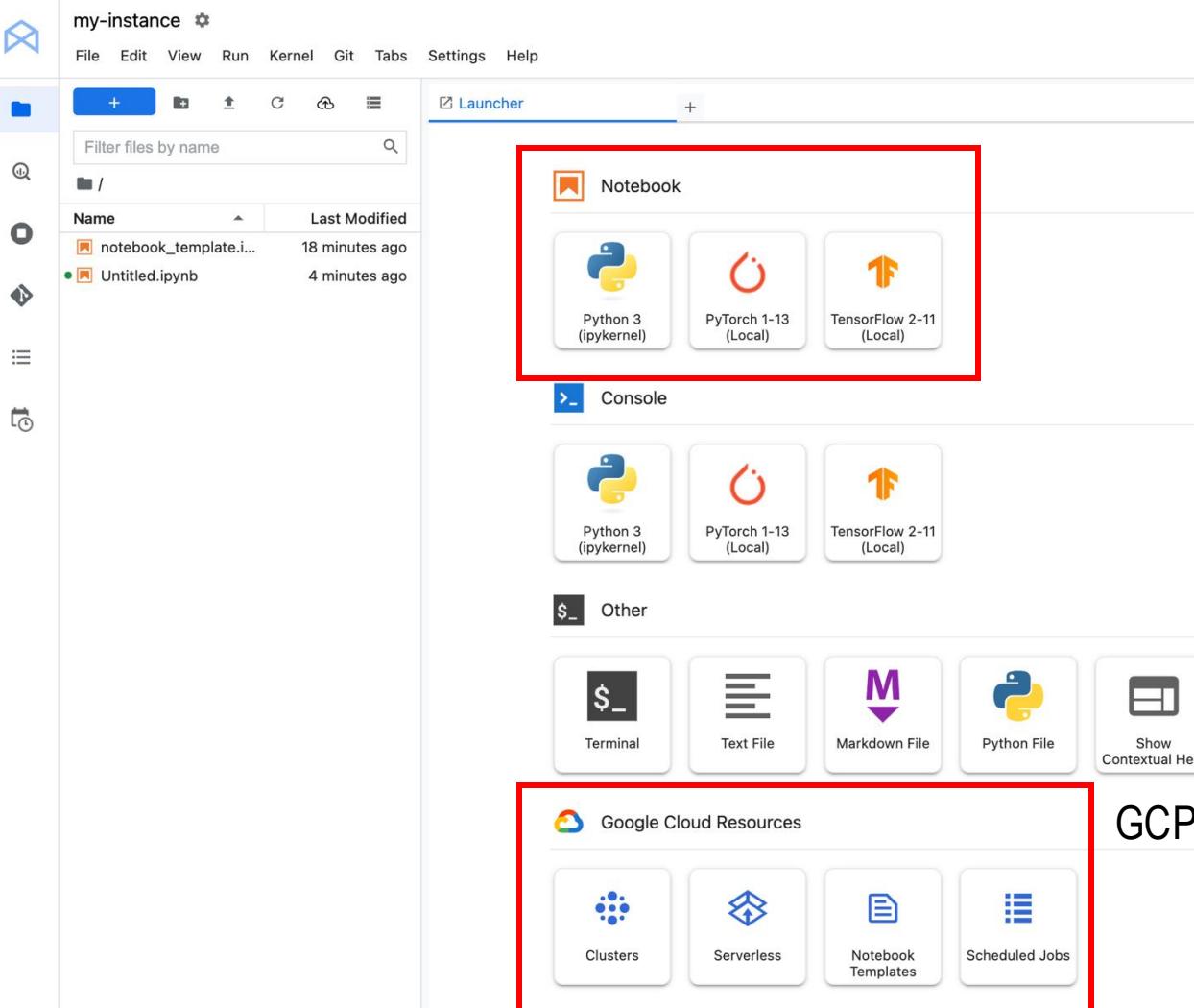
Workbench 인스턴스에는 JupyterLab 3이 사전 설치되어 있으며 GPU 지원 머신러닝 프레임워크로 구성되어 있습니다. [자세히 알아보기](#)

필터	인스턴스 이름	영역	자동 업그레이드	버전	머신 유형	GPU	소유자	생성됨	라벨
<input type="checkbox"/> <input checked="" type="radio"/> my-instance JupyterLab 열기	my-instance	asia-northeast3-a	—	—	Efficient Instance: vCPU 4 개, 16GB RAM	없음	672193963995-compute@developer.gserviceaccount.com	2025. 10. 18. 오전 3:57:29	consumer-p... :

정상적으로
생성된 인스턴스 1

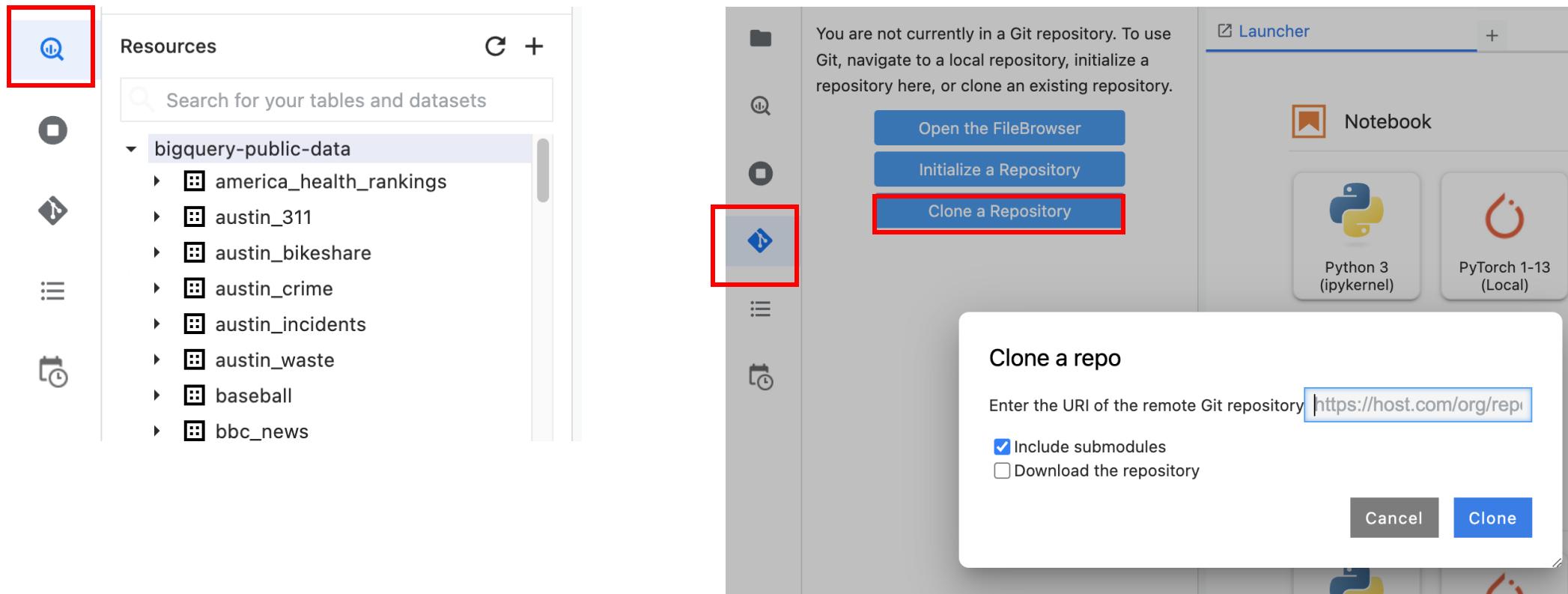
Vertex AI Workbench – Jupyter Lab 살펴보기

- JupyterLab은 GCP의 모든 리소스를 포함한 핵심 ML 프레임워크들을 다양한 형태(e.g. jupyter notebook, 터미널 등)로 하나의 환경에서 모두 사용 가능하게 함



Vertex AI Workbench – Jupyter Lab 살펴보기

- (강점) BigQuery, Git Repo, Bucket를 모두 하나의 가상환경에서 엑세스 가능
→ 모델 학습 시 데이터 가공 파이프라인 구축에 매우 용이함 (구현 및 시스템 운영 측면)
- BigQuery / Git 추가

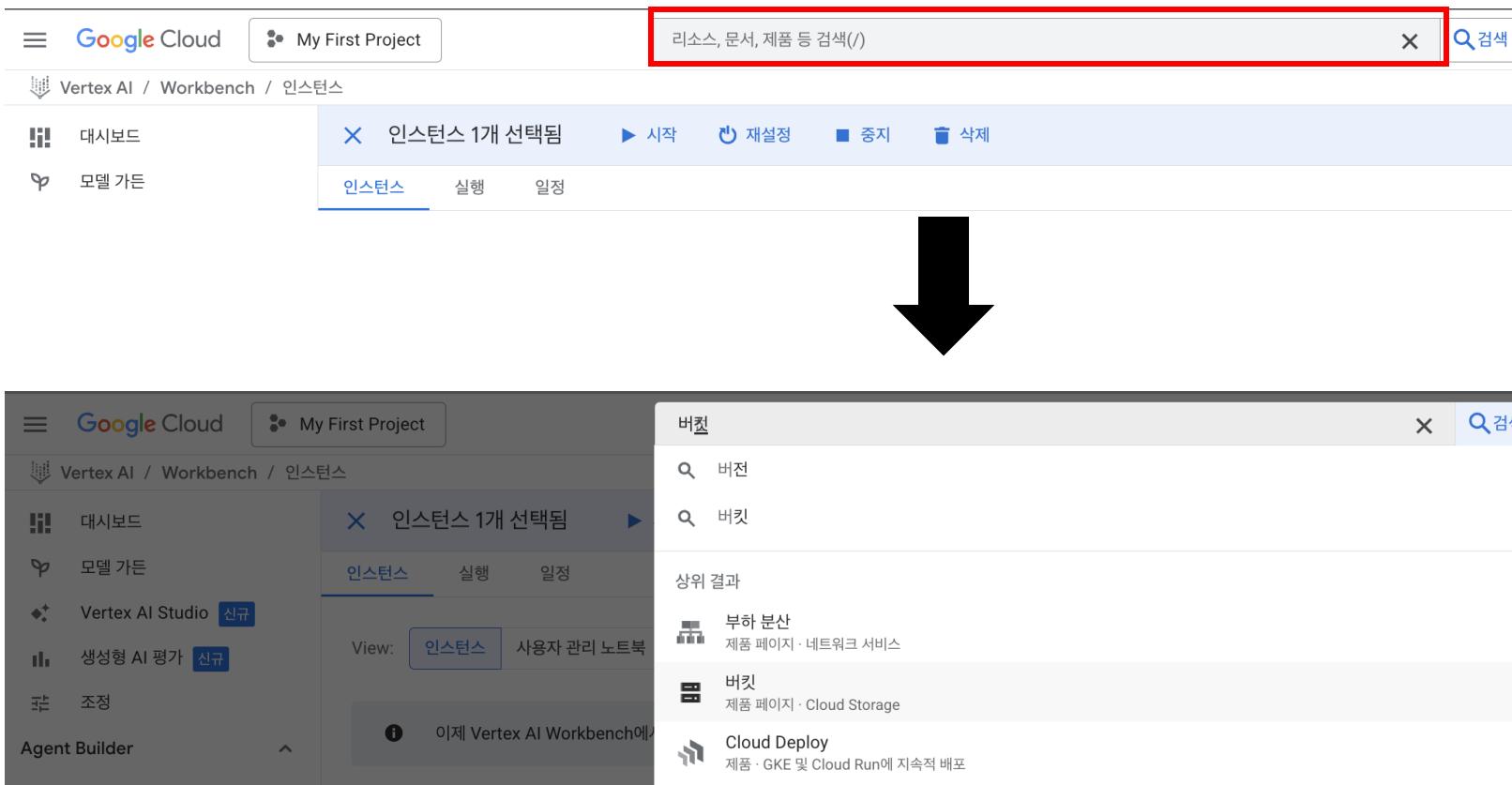


Git repository 주소에 아래 링크 넣고 Clone

<https://github.com/GoogleCloudPlatform/vertex-ai-samples.git>

Vertex AI Workbench – 버킷 만들기

- GCP Console 접속 후, 상단 검색창을 통해 '버킷' 서비스로 이동



Vertex AI Workbench – 버켓 만들기

- 버켓 이름(ID) 설정 : 이름은 전세계에서 unique 해야함
 - 이름 설정 규칙: [인스턴스 이름]-[본인이름약자]-banking-data
 - 예시: my_instance-ysh-banking-data

The screenshot shows the 'Bucket 만들기' (Bucket Creation) step of the Vertex AI Workbench wizard. The left sidebar shows 'Cloud Storage' navigation with 'Bucket' selected. The main area has three sections:

- Bucket Name:** 'my_instance-ysh2-banking-data' (highlighted with a red box). Below it is a note: 'Tip: 민감한 정보를 포함하면 안 됩니다.'
- Next Step:** A blue '계속' (Continue) button highlighted with a red box.
- Storage Location Selection:** A note: '선택사항에 따라 데이터의 지리적 위치가 정의되고 비용, 성능, 가용성이 영향을 받습니다. 나중에 변경할 수 없습니다.' A link: '자세히 알아보기'.
 - Location Type:** 'Multi-region' (highlighted with a red box).
 - Region Selection:** A dropdown menu showing 'asia (아시아의 여러 리전)' (highlighted with a red box).
 - Storage Transfer Service:** An unchecked checkbox with a note: 'Storage Transfer Service를 통해 버킷 간 복제 추가'.
 - Other Options:** Radio buttons for 'Dual-region' and 'Region', and a note: '리전 2곳에서 고가용성 및 짧은 지연 시간'.

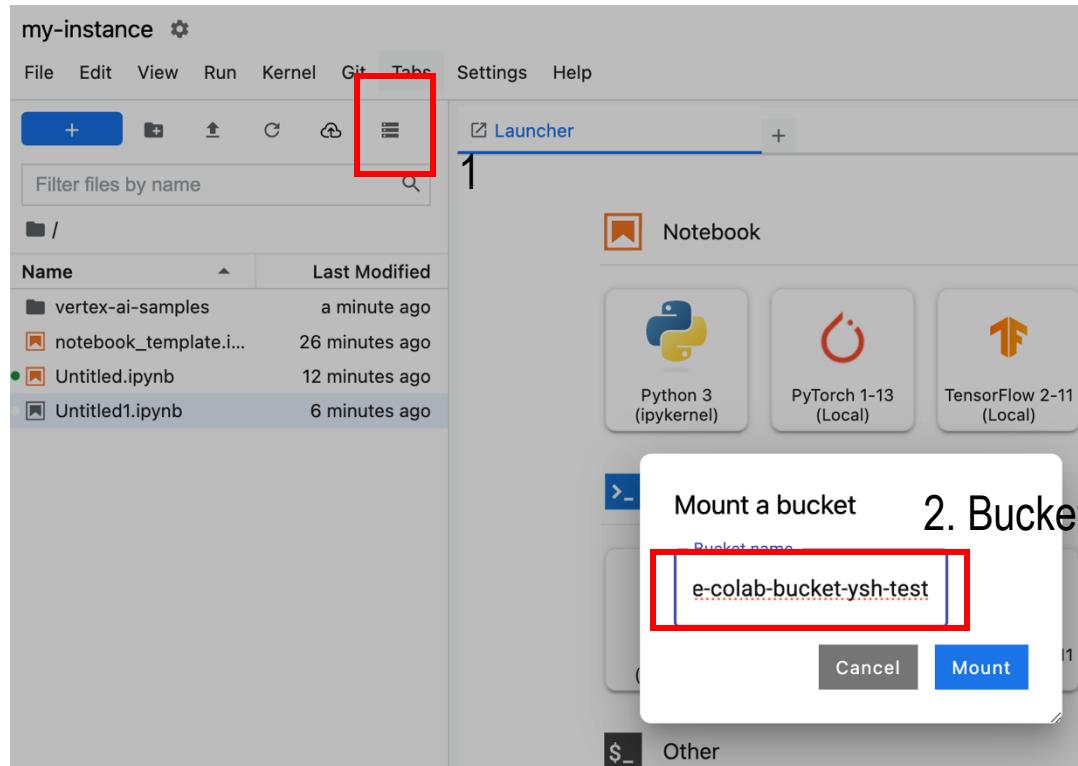
Right Panel:

- Data Storage Method Selection:** A note: '데이터 저장 방법 선택'. Options: '기본 스토리지 클래스: Standard', '계층적 네임스페이스: 사용 중지됨', 'Anywhere Cache: 사용 중지됨'.
- Object Access Method Selection:** A note: '객체 액세스를 제어하는 방식 선택'. Options: '공개 액세스 방지: 사용', '액세스 제어: 군일'.
- Object Data Protection Method Selection:** A note: '객체 데이터를 보호하는 방법 선택'. Options: '소프트 삭제 정책: 기본값', '객체 버전 관리: 사용 중지됨', '버킷 보관 정책: 사용 중지됨', '객체 보관: 사용 중지됨', '암호화 유형: Google 관리'.

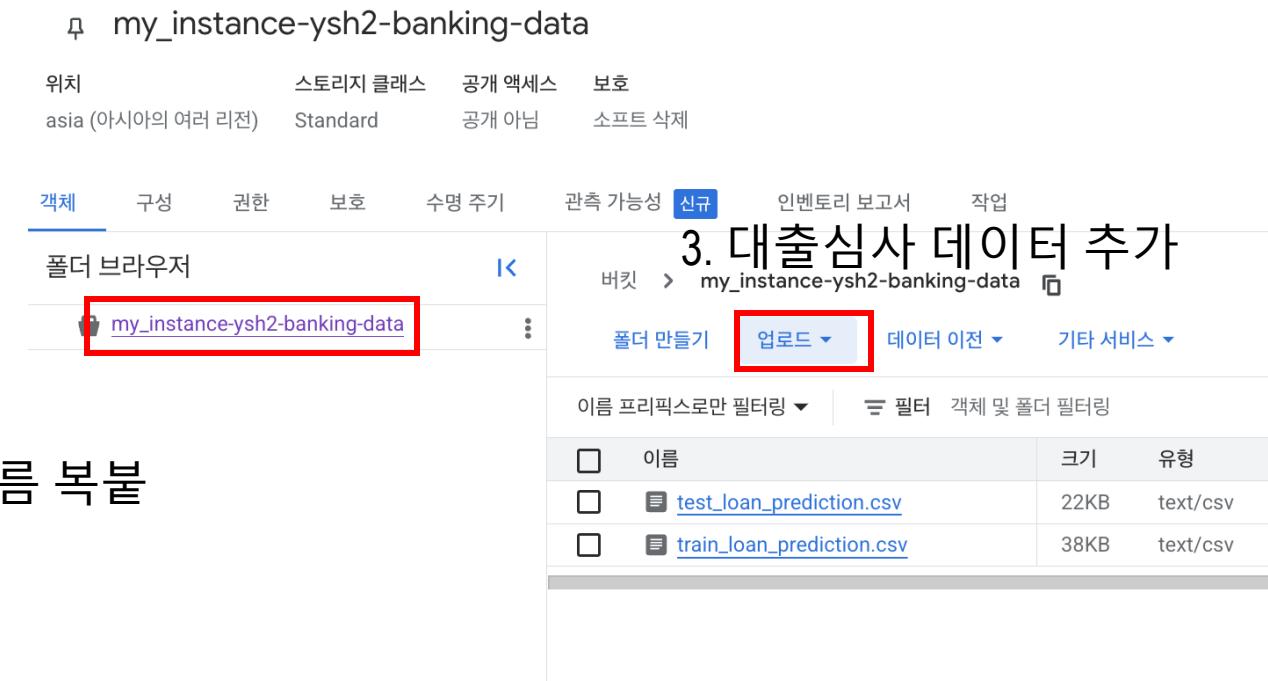
Bottom Buttons: 'Marketplace' (disabled), '출시 노트' (disabled), a large blue '만들기' (Create) button highlighted with a red box, and a '취소' (Cancel) button.

Vertex AI Workbench – Jupyter Lab 살펴보기

- (강점) BigQuery, Git Repo, Bucket를 모두 하나의 가상환경에서 엑세스 가능
→ 모델 학습 시 데이터 가공 파이프라인 구축에 매우 용이함 (구현 및 시스템 운영 측면)
- Bucket 설정 예시
 - 버킷 이름이 다른건 무시해주세요. 직전에 만든 버킷 이름을 복붙할것!



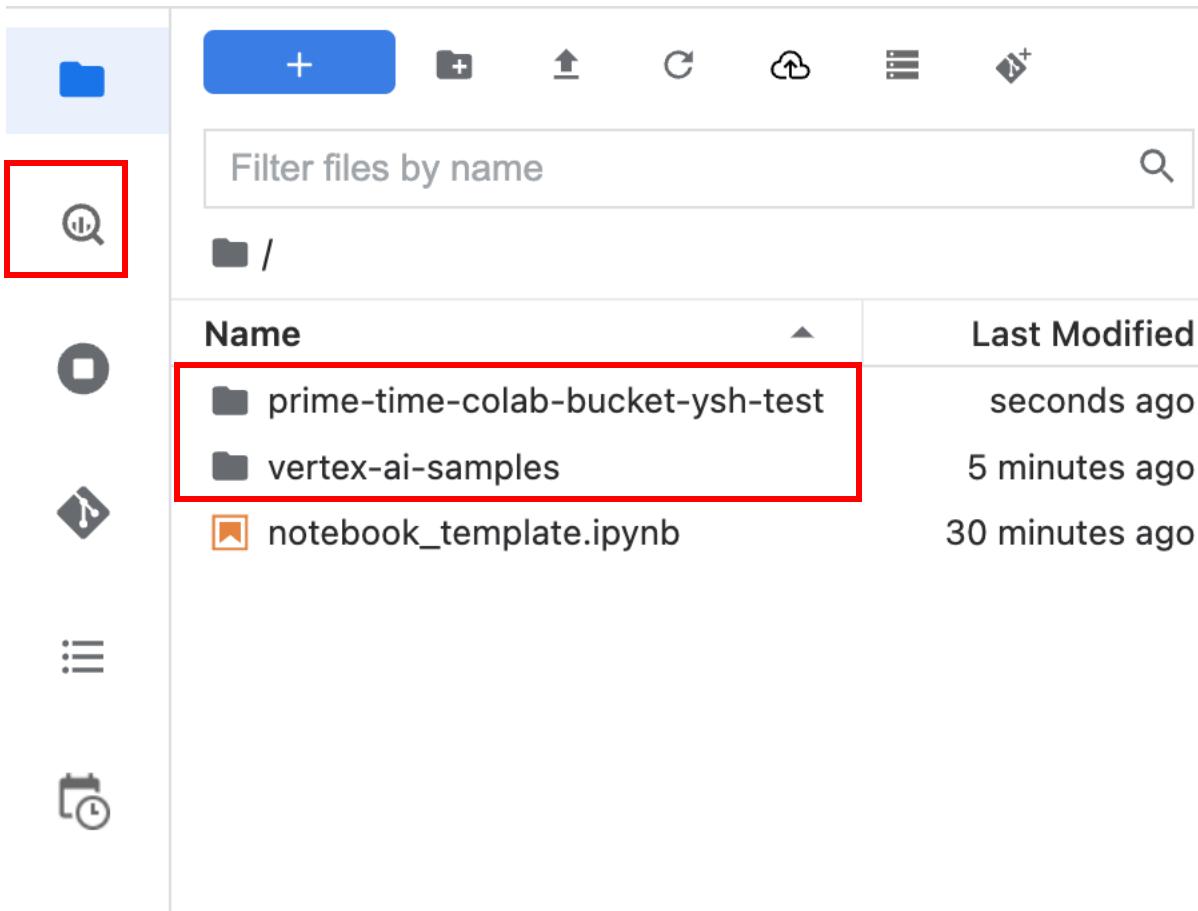
2. Bucket 이름 복붙



The screenshot shows the Google Cloud Storage console for the bucket 'my_instance-ysh2-banking-data'. The bucket details are shown at the top: location 'asia (아시아의 여러 리전)', storage class 'Standard', public access '공개 아님', and encryption '소프트 삭제'. Below this, the 'Objects' tab is selected, showing two files: 'test_loan_prediction.csv' (22KB, text/csv) and 'train_loan_prediction.csv' (38KB, text/csv). A red box highlights the bucket name 'my_instance-ysh2-banking-data' in the list of objects. Another red box highlights the 'Upload' button in the top right of the object list area. The text '3. 대출심사 데이터 추가' is overlaid on the right side of the screen.

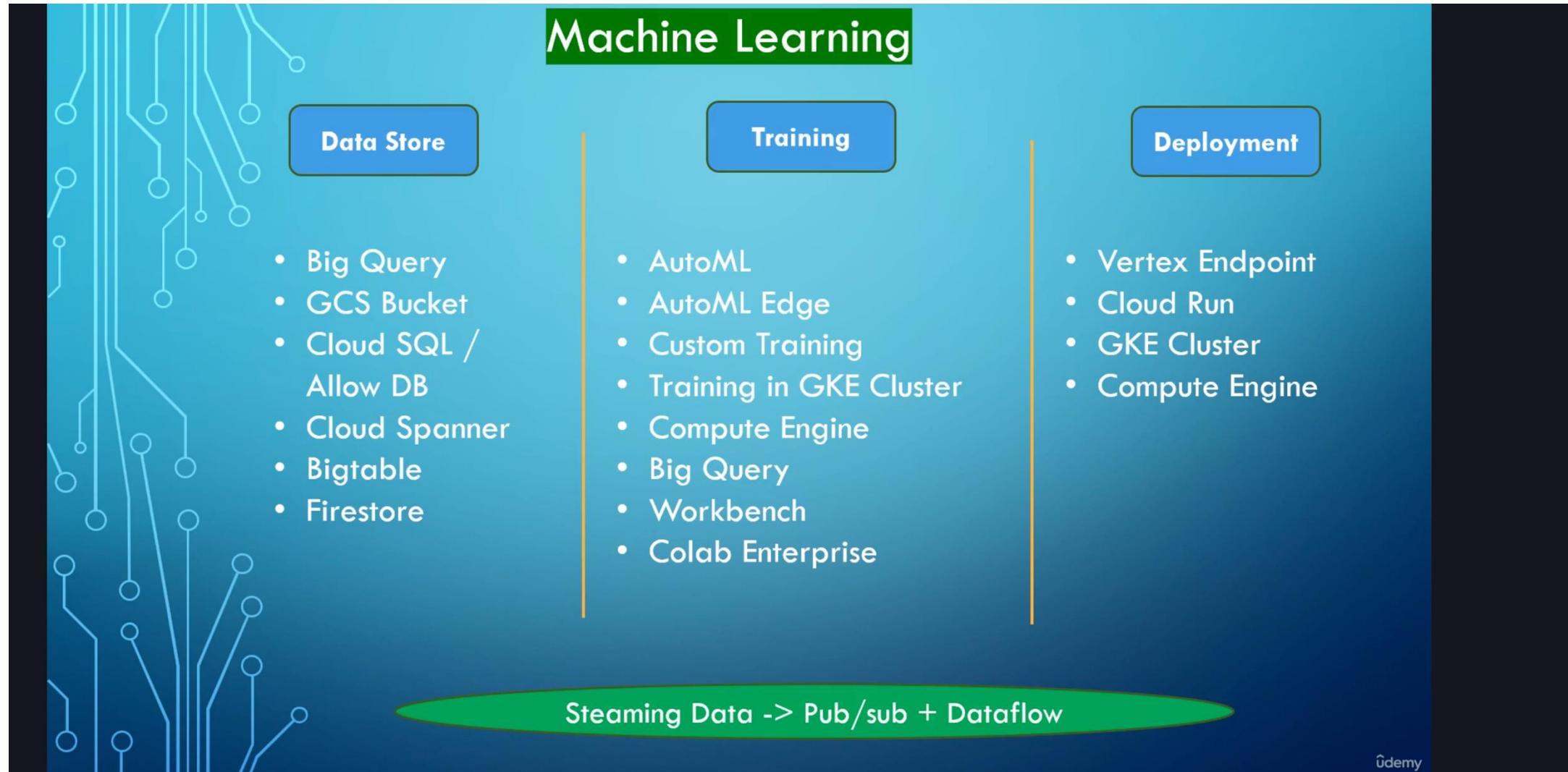
Vertex AI Workbench – Jupyter Lab 살펴보기

- (강점) BigQuery, Git Repo, Bucket를 모두 하나의 가상환경에서 엑세스 가능
→ 모델 학습 시 데이터 가공 파이프라인 구축에 매우 용이함 (구현 및 시스템 운영 측면)
- 모든 데이터 및 깃헙 연동 후 모습: 최상단 폴더 하위에 데이터들이 모두 위치



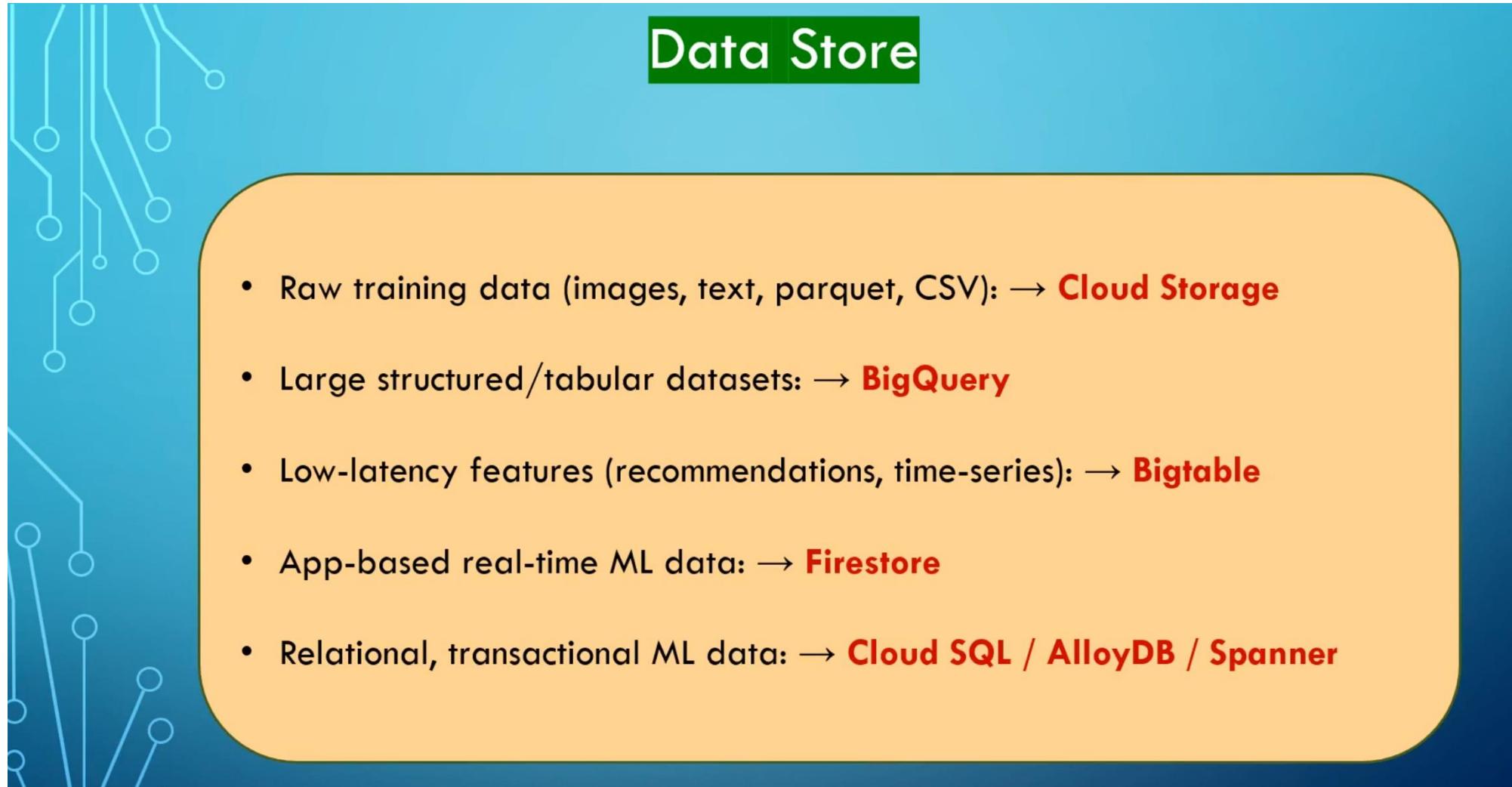
Appendix: Machine Learning

- Whole Process



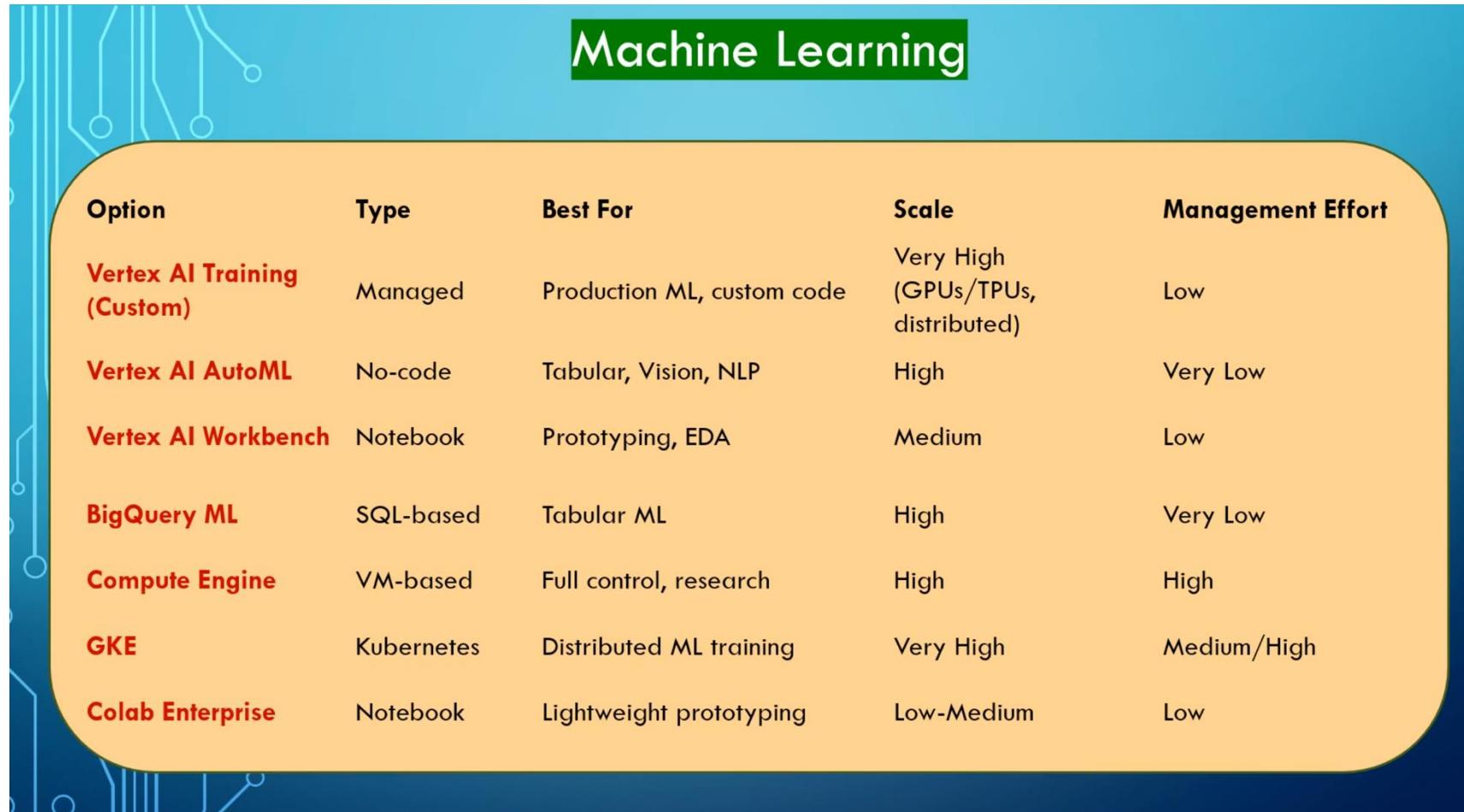
Appendix: Data Store

- Data stores



Appendix: Machine Learning in Vertex AI

- Machine Learning in Vertex AI → Scalable



Machine Learning

Option	Type	Best For	Scale	Management Effort
Vertex AI Training (Custom)	Managed	Production ML, custom code	Very High (GPUs/TPUs, distributed)	Low
Vertex AI AutoML	No-code	Tabular, Vision, NLP	High	Very Low
Vertex AI Workbench	Notebook	Prototyping, EDA	Medium	Low
BigQuery ML	SQL-based	Tabular ML	High	Very Low
Compute Engine	VM-based	Full control, research	High	High
GKE	Kubernetes	Distributed ML training	Very High	Medium/High
Colab Enterprise	Notebook	Lightweight prototyping	Low-Medium	Low

Appendix: Deployment in Vertex AI

- Deployment Option

Option	Type	Best For	Scaling	Management Effort
Vertex AI	Managed	Online & batch ML	Auto	Very low
Cloud Run	Serverless containers	Lightweight APIs	Auto	Low
GKE	Kubernetes	Complex ML serving	Manual/auto	Medium/High
Compute Engine	VMs	Legacy/custom ML	Manual	High