# 비즈니스 가치를 극대화하는 생성형 AI

*파일럿을 넘어 생산성 향상으로*



최석재 상무

Lead Architect
Data&AI
한국IBM

© 2024 IBM Corporation

# AI 대한 접근방식에 변화

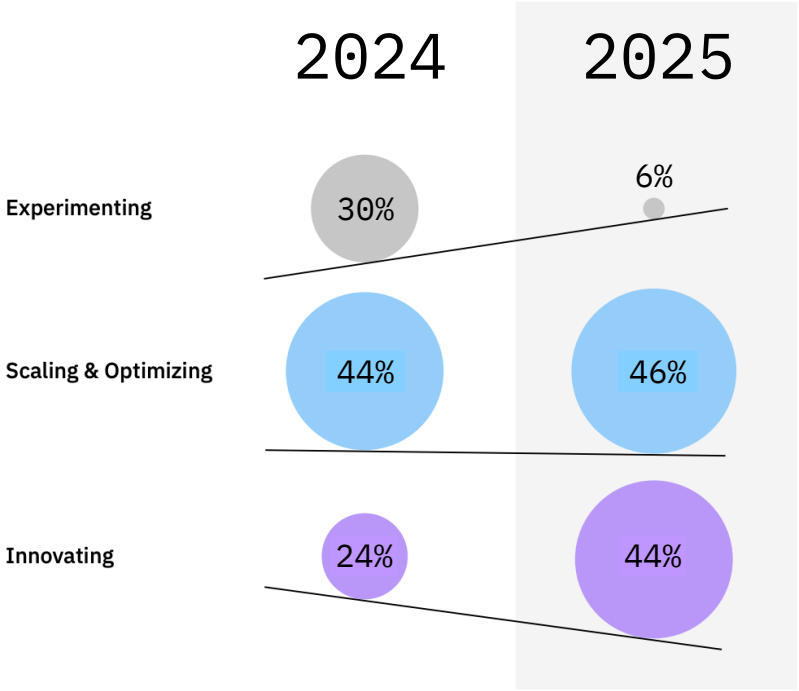
*Leaders are pushing their teams to innovate with AI in 2025*

**2024**

AI를 주로 저위험, 비핵심 기능에서 실험하며 경험을 쌓고 있으며(30%), 새로운 기회를 창출하기 위해 AI를 혁신적으로 활용하는 기업은 24%에 불과

**2025**

하지만 2025년에는 큰 변화가 예상되며, 46%의 기업이 AI를 기존 프로세스 최적화에 활용하고, 44%는 혁신을 위해 사용할 것으로 전망된다. 실험 단계에 머무는 기업은 6%에 불과할 것으로 보임




*Source: 5 Trends for 2025 global pulse survey. Q: Which of the following best describes your organization's approach to adopting AI, today and next year?*
*Note: Sums don't equal 100% because "reevaluating" and "none of the above" were also potential responses.*

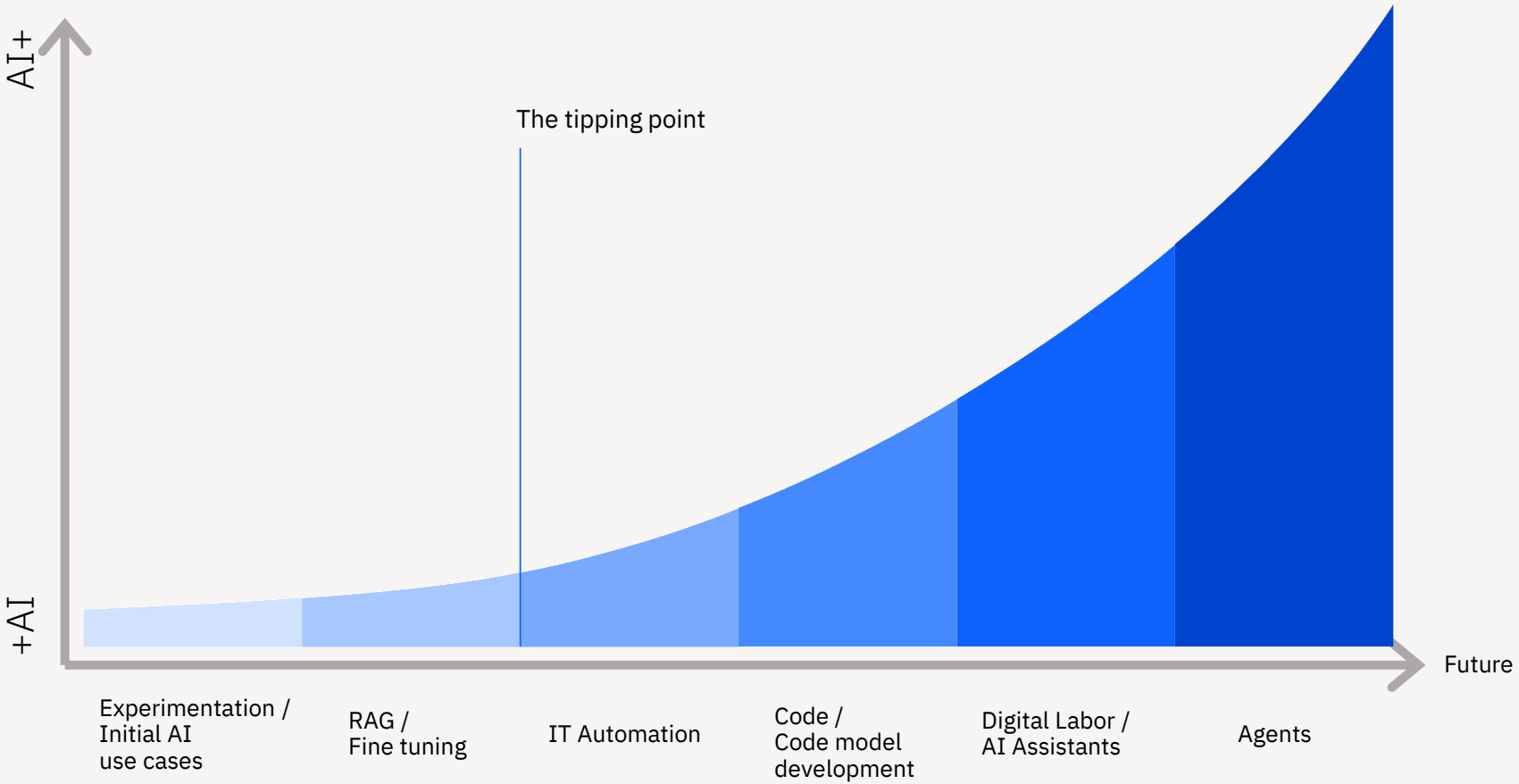
© 2025 IBM Corporation

AI 생산성의
다음 물결



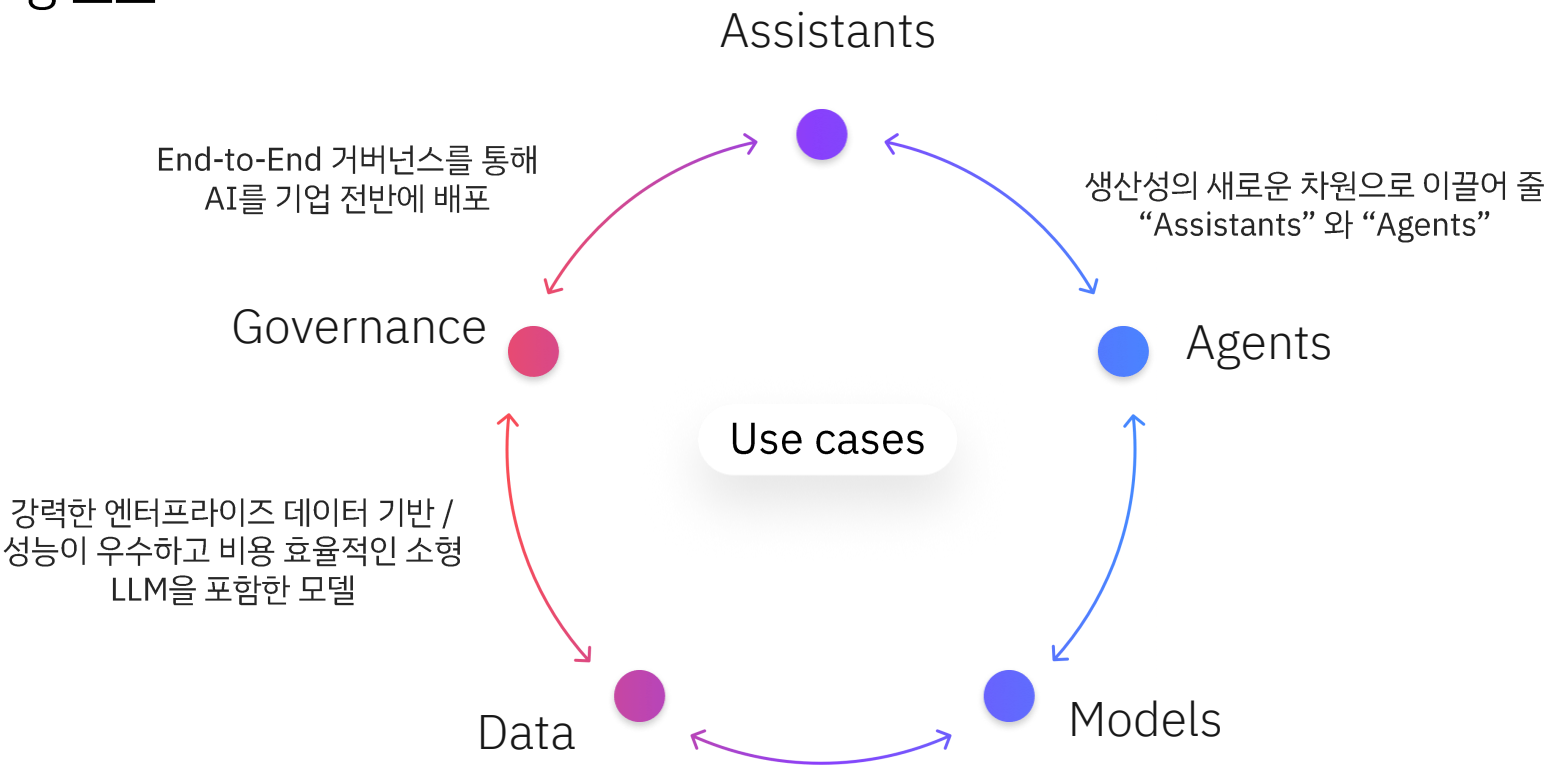

IBM

# 가치 창출 곡선



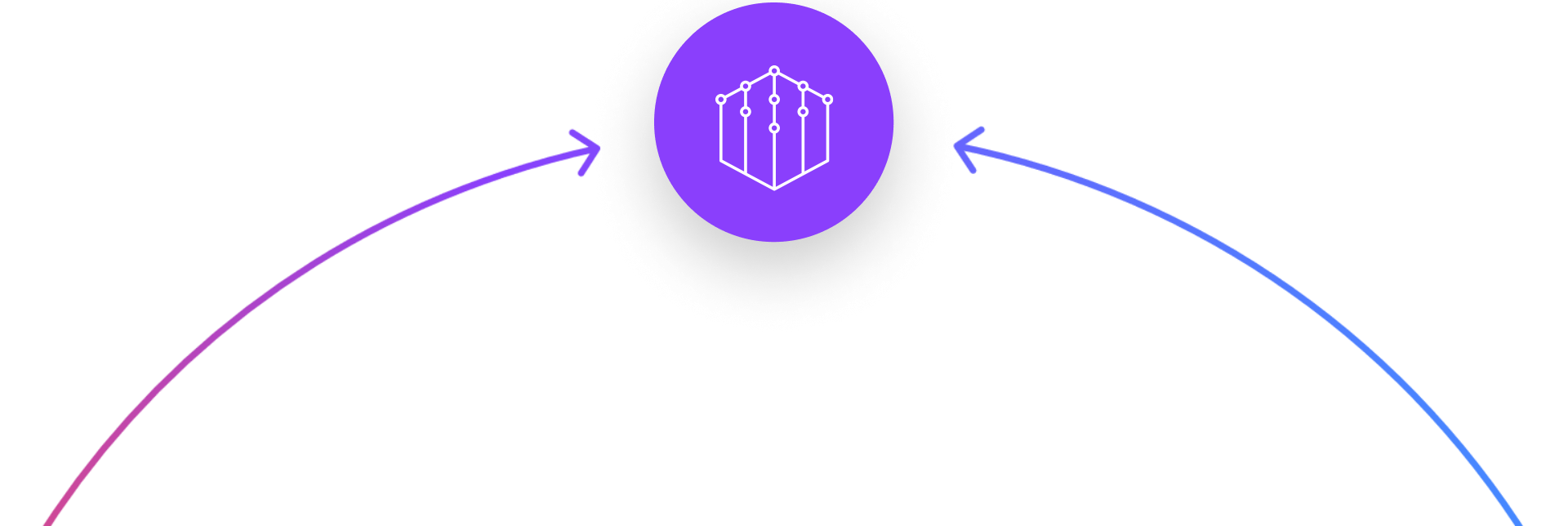

© 2025 IBM Corporation

# 미래를 위한
# AI 구성 요소




IBM

# Assistants



AI assistants를 위한 주요 도메인 및 사용 사례 →

## Human Resources

HR 운영 및 직원 셀프 서비스를 혁신

- 직원 지원
- 인재 채용
- 온보딩
- 성과 관리

## Customer Care

CX를 혁신하고 고객이 셀프 서비스를 할 수 있도록 유도

- 고객 셀프 서비스
- 컨택센터 현대화
- 컨택센터 인사이트
- 에이전트 지원

## Sales

판매 실적을 높이고 리드 평가(Lead Qualification)를 자동화

- 판매 활성화
- 잠재 고객 지원
- 기회 관리

## Procurement

일상 업무 워크플로우를 간소화

- 공급업체 평가
- 소스에서 지불까지
- 조달 통찰력

## Finance

주문 처리, 이행, 송장 발행 및 회비를 신속히 처리

- 신속한 Order to Cash
- 경비 관리
- 급여 대조

## Supply Chain

일상적인 도구를 조율하여 직원 생산성 향상

- 소싱 지원
- 재고 관리
- 지속 가능성

# IBM이 AI를 사용하는 방법

고객 서비스

125,000

분기당 절약되는 시간

watsonx Assistant

디지털 노동(Digital Labor)

50-70%

일부 반복 작업의 자동화

watsonx Orchestrate

IT 운영

70%

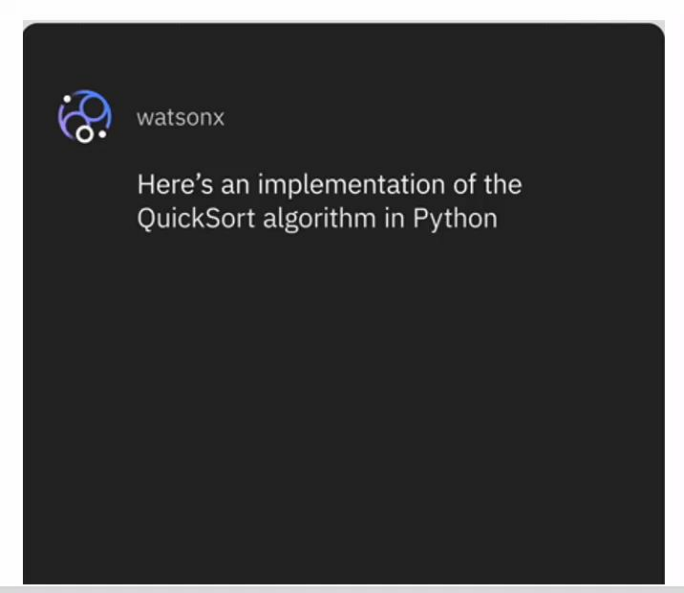
의 Ansible Playbook 콘텐츠가 6배 더 빠르게 생성됨

watsonx Code Assistant

© 2025 IBM Corporation

# IBM watsonx Code Assistant

기술 격차를 해소하고 각 비즈니스 사용 사례에 대한 생산성을 높이기 위한 코드 솔루션을 위한 엔터프라이즈 지원 AI




### AI-powered coding tasks

코드 생성, 설명, 단위 테스트 생성, AI기반 Documentation 등.

### Enterprise Java modernization

Java 엔터프라이즈 애플리케이션을 위한 생성형 AI와 자동화 지원 현대화.

### Chat for Code integrated UI

애플리케이션 및 코드 작성을 위한 자유형 AI 대화형 채팅.

### Enhanced prompt engineering

특정 작업에 맞춰 최적화된 프롬프트가 포함된 사전 구축된 채팅 명령.

### Part of Integrated Development Environments (IDEs)

VS Code와 Eclipse에서 바로 사용할 수 있으며, 워크플로와 완벽하게 통합.

### Trust, transparency, and privacy

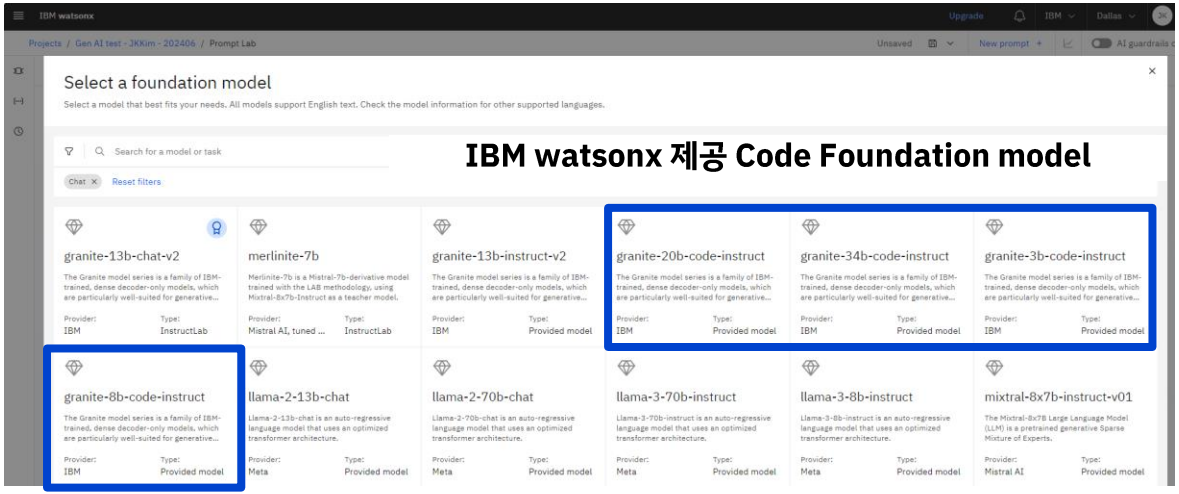
코드 유사성 검사, IP 보호, 클라우드 또는 온프레미스에서 SaaS를 통해 배포.

© 202

# IBM Code Foundation Models

- IBM watsonx.ai 플랫폼에서 아래 4개 Code 파운데이션 모델 제공
  - 🤷🏽‍♀️ **Granite-3b-code**
  - 🧠 **Granite-8b-code**
  - 📈 **Granite-20b-code**
  - ⚡ **Granite-34b-code**
- Python, JavaScript, Java, Go, C++, Rust 등 **167개 코딩 언어로 학습**
- 4조+ 토큰으로 학습

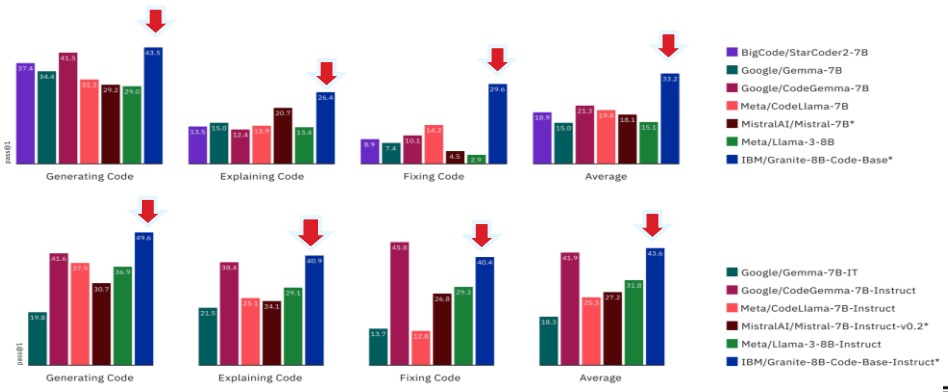
**IBM watsonx 제공 Code Foundation model**




## Key values

1. ***Performance*** *: 타사 모델 대비 다양한 작업 영역에서 우월한 성능(우측 이미지 참조)*
2. ***Safety*** *: IBM만이 할 수 있는 기업용 코드 데이터, 라이선스 허용하는 안전한 데이터로 학습, 문제 시 IBM 배상 (indemnification)*
3. ***Variety*** *: 다양한 코딩 언어 기반 학습으로 많은 코드 관련 use case 수용 /적용 가능*

**IBM Code Foundation model vs. others**




*Apache 2.0 License

© 2025 IBM Corporation



Agents

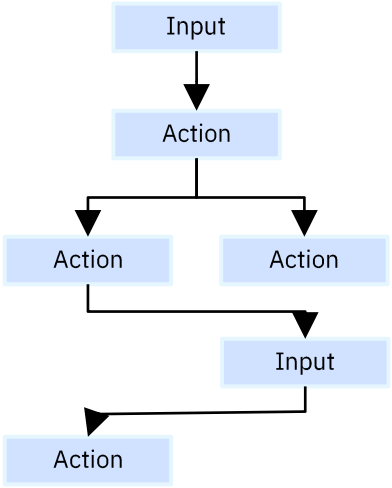
# Agentic AI는 AI 어시스턴트를 위한 새로운 패러다임을 열어줍니다.

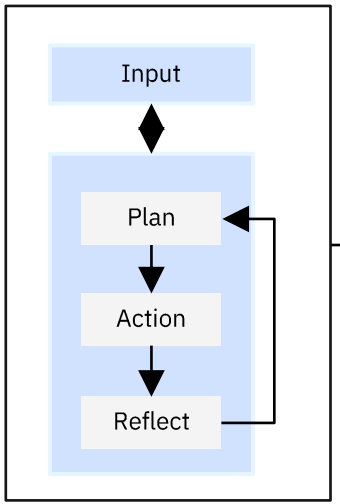
## Traditional assistants → Single-agent assistants → Multi-agent assistants

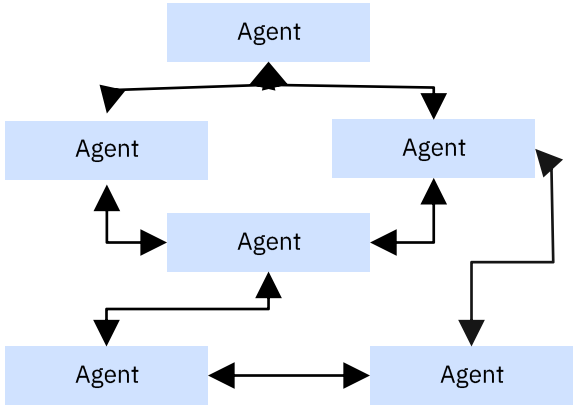
- Rule based (if x, do y)
- 미리 정의된 작업 경로

- Task based (e.g., flight booker)
- 작업 범위가 증가함에 따라 성능 제약
- 제한된 제어

- Domain based (e.g., travel agent)
- 전문 에이전트(플래너, 항공편 예약자, 호텔 예약자 등)가 함께 작업하여 시스템 성능을 개선
- 에이전트가 어떻게 소통하는지에 대한 제어
- 다양한 아키텍처 옵션(네트워크, 계층적, 감독자, 사용자 정의)



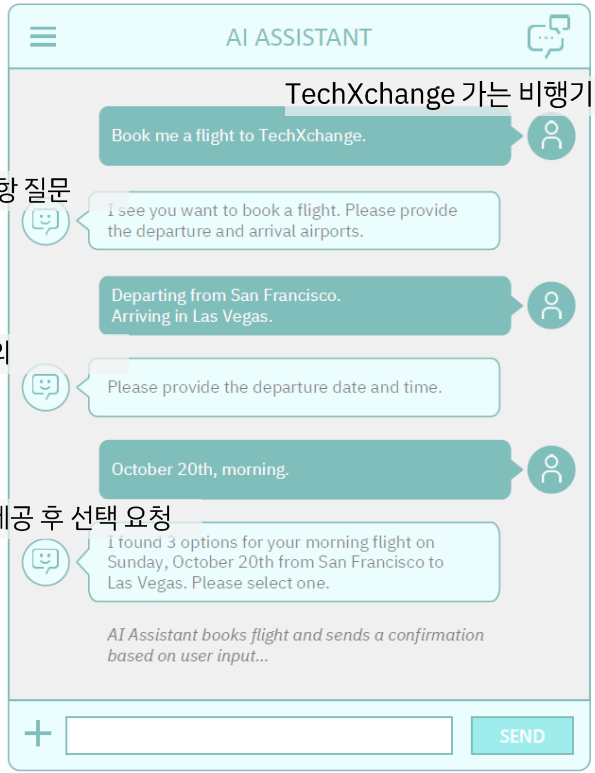




© 2025 IBM Corporation

IBM

# Automated vs. Autonomous

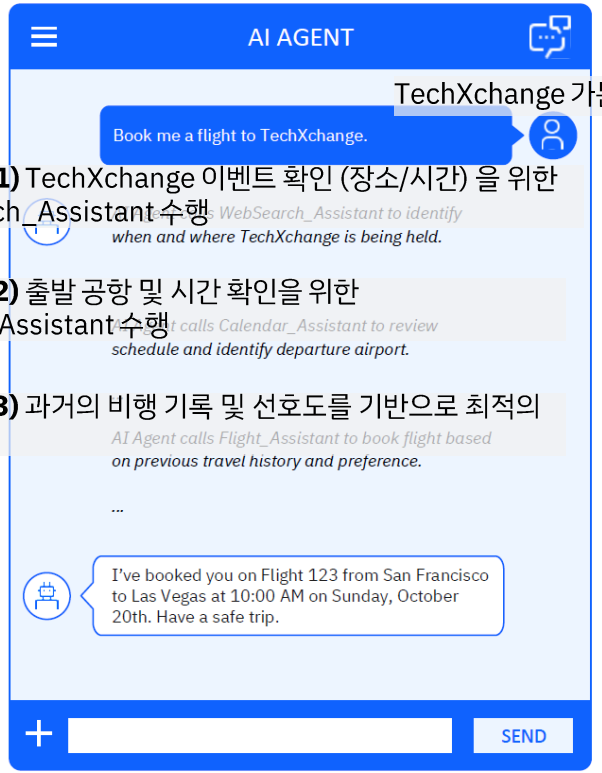



TechXchange 가는 비행기편 예약 요청

**Q1)** 출발, 도착 공항 질문

**Q2)** 출발 시간 질의

**Q3)** 추천 리스트 제공 후 선택 요청




TechXchange 가는 비행기편 예약 요청

**AI Agent 1)** TechXchange 이벤트 확인 (장소/시간) 을 위한 WebSearch_Assistant 수행

**AI Agent 2)** 출발 공항 및 시간 확인을 위한 Calendar_Assistant 수행

**AI Agent 3)** 과거의 비행 기록 및 선호도를 기반으로 최적의 비행편 확인

자동화(**Automated**)된 시스템과 자율(**Autonomous**) 시스템의 주요 차이는 제어 및 의사 결정 수준에 있습니다.

© 2025 IBM Corporation

# 비즈니스 전반에 걸쳐 AI 에이전트를 확장하세요.

**다중 에이전트 오케스트레이션 (Multi-agent orchestration)**

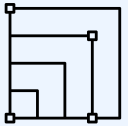
Manage all agents in one place



**watsonx Orchestrate**

**사전 구축 에이전트 (Prebuilt agents)**

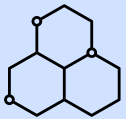
Accelerate AI agent deployment



**watsonx Orchestrate**

**사용자 정의 에이전트 (Custom-built agents)**

Build custom designed agents



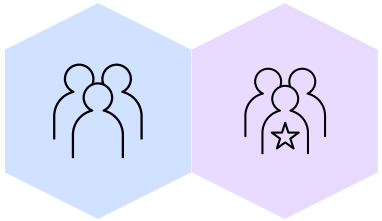
**watsonx.ai**

© 2025 IBM Corporation

# IBM watsonx Orchestrate

기업 전체에 AI 어시스턴트 및 에이전트를 구축, 배포 및 조정합니다.

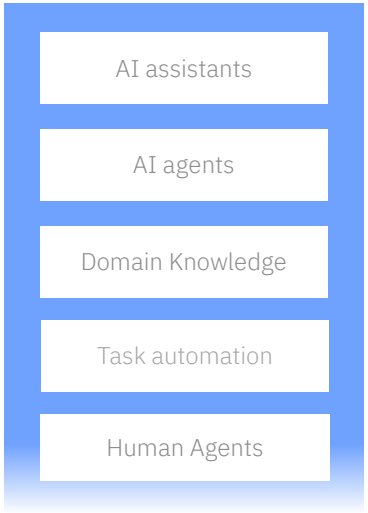
Employees, customers & knowledge workers



Unified AI agent experience

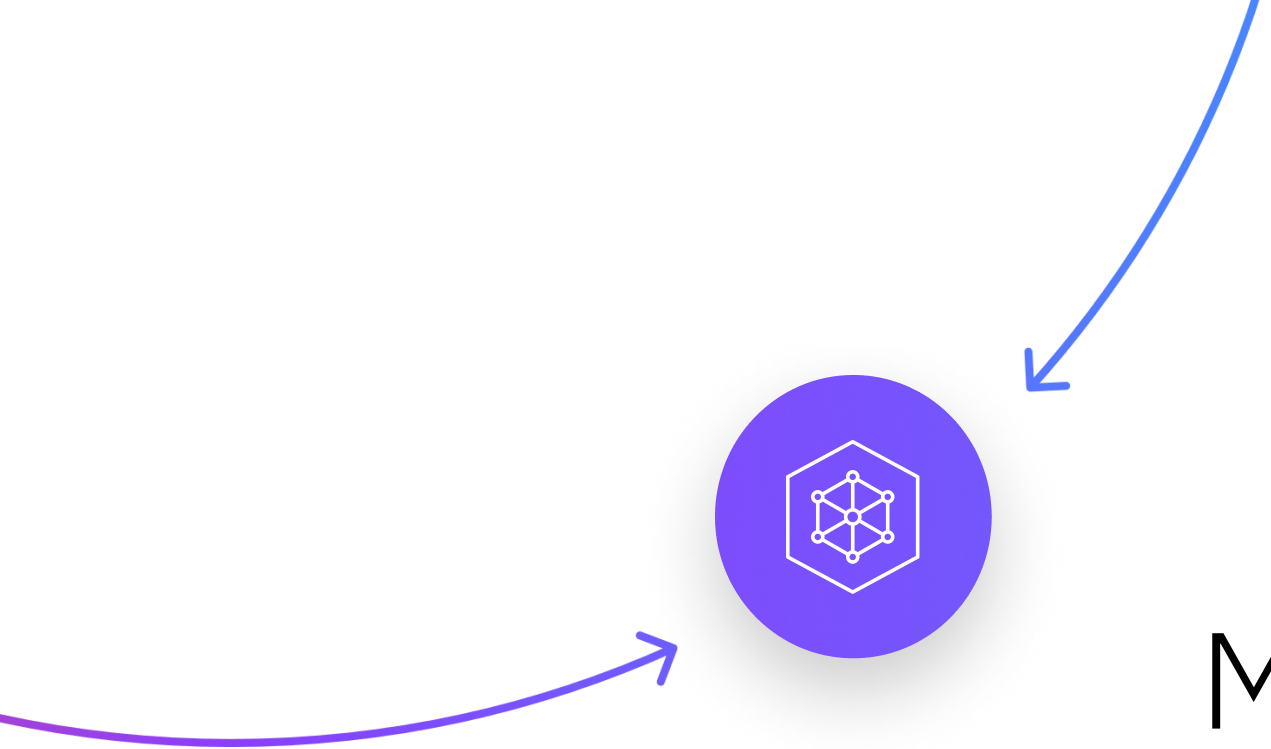


Multi-agent orchestration




workday.
SEISMIC
SAP SuccessFactors
servicenow
salesforce
LangChain
IBM
crewai
tonkean
PEGA
appian
UiPath

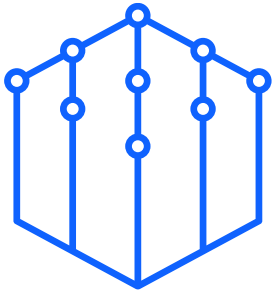
© 2025 IBM Corporation



# Models

# IBM의 접근방식 :
# **fit-for-purpose** models

**추론 비용을**
**최대 42배까지 낮춤**



Your data



The right model

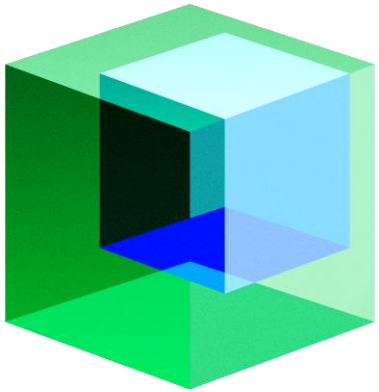


Targeted use case
fine tuning

© 2025 IBM Corporation

# IBM Granite

기업 AI 도입을 가속화하기 위한
개방적이고 성능이 뛰어나며
신뢰할 수 있는 AI 모델 제품군



## Open

- Apache 2.0기반 오픈 소스
- 데이터 및 교육 방법에 대한 투명성
- 사용자 데이터로 customizing

## Performant

- 다양한 범위의 목적에 맞는 모델
- 확장성을 위해 설계

## Trusted

- IP 보증
- 책임감 있고 안전한 AI
- 데이터를 보호하고 위험을 완화하기 위한 가드레일

Granite family of models

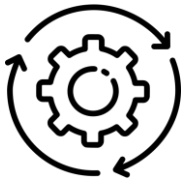


| Large Language Models (LLMs) for enterprise | Inference-efficient Mixture of Experts (MoE) | Guardrail models | Speculative decoding models | Time-series models | Geospatial models |
|---|---|---|---|---|---|
| Granite-3.0-8B-Instruct<br>Granite-3.0-2B-Instruct | Granite-3.0-3B-A800M<br>Granite-3.0-1B-A400M | Granite-Guardian-3.0-8B<br>Granite-Guardian-3.0-2B | Granite-3.0-8B-Instruct-Accelerator | | |

© 2025 IBM Corporation

# IBM의 AutoAI는
# RAG(Retrieval-Augmented Generation) 시스템 개발에 혁명을 일으킵니다.

## RAG 구성 자동화



AutoAI는 최적의 모델 선택, 청킹 및 검색 구성을 자동화하여 RAG 설정을 최적화하고 비즈니스 사용 사례에 대한 최적의 RAG 설정을 찾는 데 시간과 리소스를 절약합니다.

## 종합적인 평가



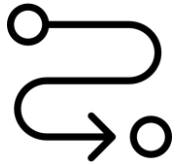
AutoAI는 RAG 지표와 벤치마킹 도구를 사용하여 정확하고 관련성 있는 답변을 보장합니다.

## 하이퍼 파라미터 최적화



AutoAI는 고급 알고리즘을 사용하여 최상의 RAG 구성을 찾아 적용합니다.

## 엔드투엔드 자동화

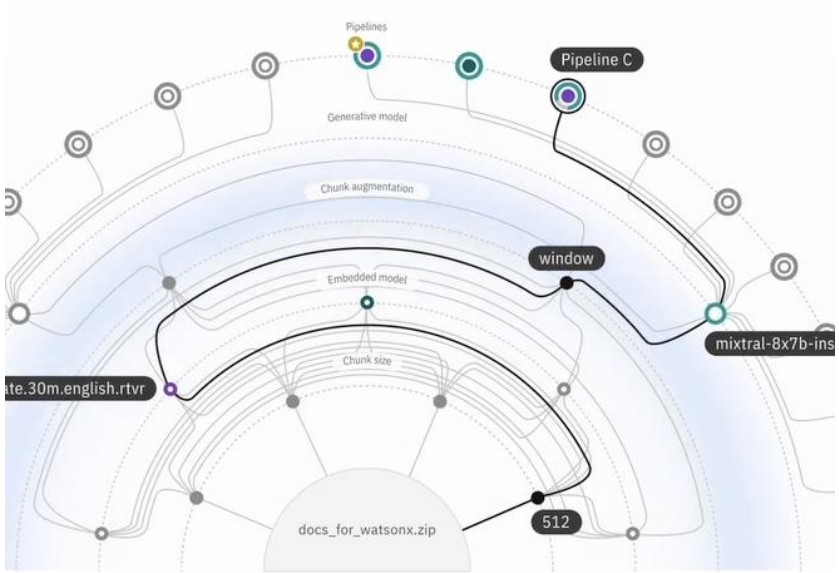


AutoAI는 생성 모델과 정밀한 기초를 결합하여 정확하고 풍부한 응답을 제공하고 RAG 파이프라인 생성과 배포를 간소화합니다.

© 2025 IBM Corporation

IBM

# AutoAI for RAG




## 간소화된 RAG 최적화

AutoAI는 RAG 시스템의 생성 및 최적화를 자동화하여 데이터 품질을 손상시키지 않으면서 개발 시간을 몇 달에서 며칠로 단축합니다.

## 포괄적인 파이프라인 테스트

이 플랫폼은 다양한 파이프라인 구성을 테스트하고, 정확성과 충실성 등의 측정 항목에 따라 순위를 매기고, 더욱 세부화할 수 있는 가장 성능이 좋은 옵션을 식별합니다.

## 원활한 배포

AutoAI는 두 개의 파이프라인으로 배포를 지원합니다. 하나는 데이터 세트 인덱싱용이고 다른 하나는 RAG 추론용이므로 프로덕션에 쉽게 통합할 수 있습니다.

## 사용자 친화적인 실험

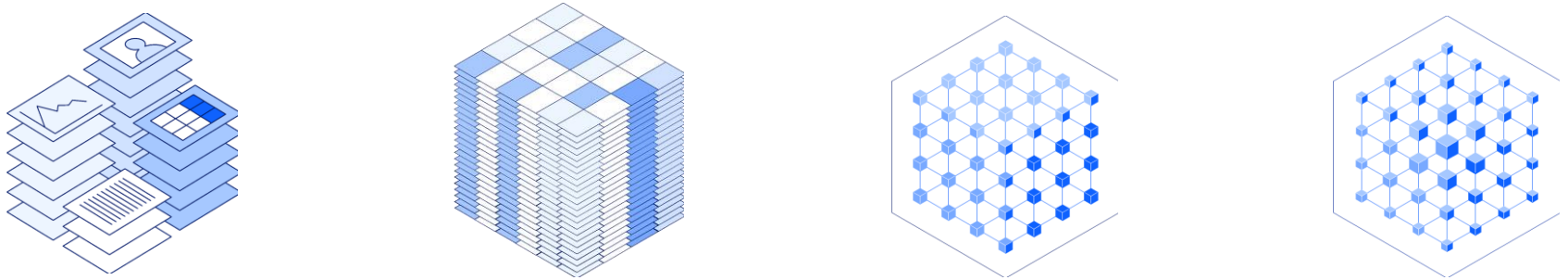
개발자는 직관적인 워크플로 내에서 실험을 빠르게 정의하고, 데이터 세트를 업로드하고, 파이프라인 성능을 테스트할 수 있습니다.

© 2025 IBM Corporation

# 최소한의 비용으로 최첨단 모델 성능 구현
# Technology behind InstructLab (Large-scale Alignment for chatBots)

모델에 기술(Skills)과 지식(Knowledge)을 추가하고, 합성 데이터 생성

맞춤형 모델 생성



(모델 훈련을 위한) 예제 생성

높은 품질의 엄선된 지식 소스와 각 작업별로 사람이 만든 예제(human-generated examples)를 포함하는 분류 체계(taxonomy)를 제공합니다.

티쳐 모델 - Teacher model(s)

티쳐 모델(teacher model)은 분류 체계(taxonomy)에 대한 질문과 답변으로 이루어진 '학습 자료'를 생성합니다.

비평 모델 - Critic model(s)

비평 모델(critic model)은 질문의 정확성과 품질을 검증하고, 합성 데이터(synthetic data)에 금지된 자료를 포함하는지 검사합니다.

학생 모델 - Student model(s)

학생 모델(student model)은 새로운 훈련 방식을 사용하여 앞서 생성된 '학습 자료'에 기반 하여 훈련됩니다.

**InstructLab**을 활용한 LLM 맞춤화

출처: https://arxiv.org/abs/2403.01081

© 2024 IBM Corporation

IBM

# InstructLab으로 비용과 시간 절약

98.5%

Cost savings

35 weeks

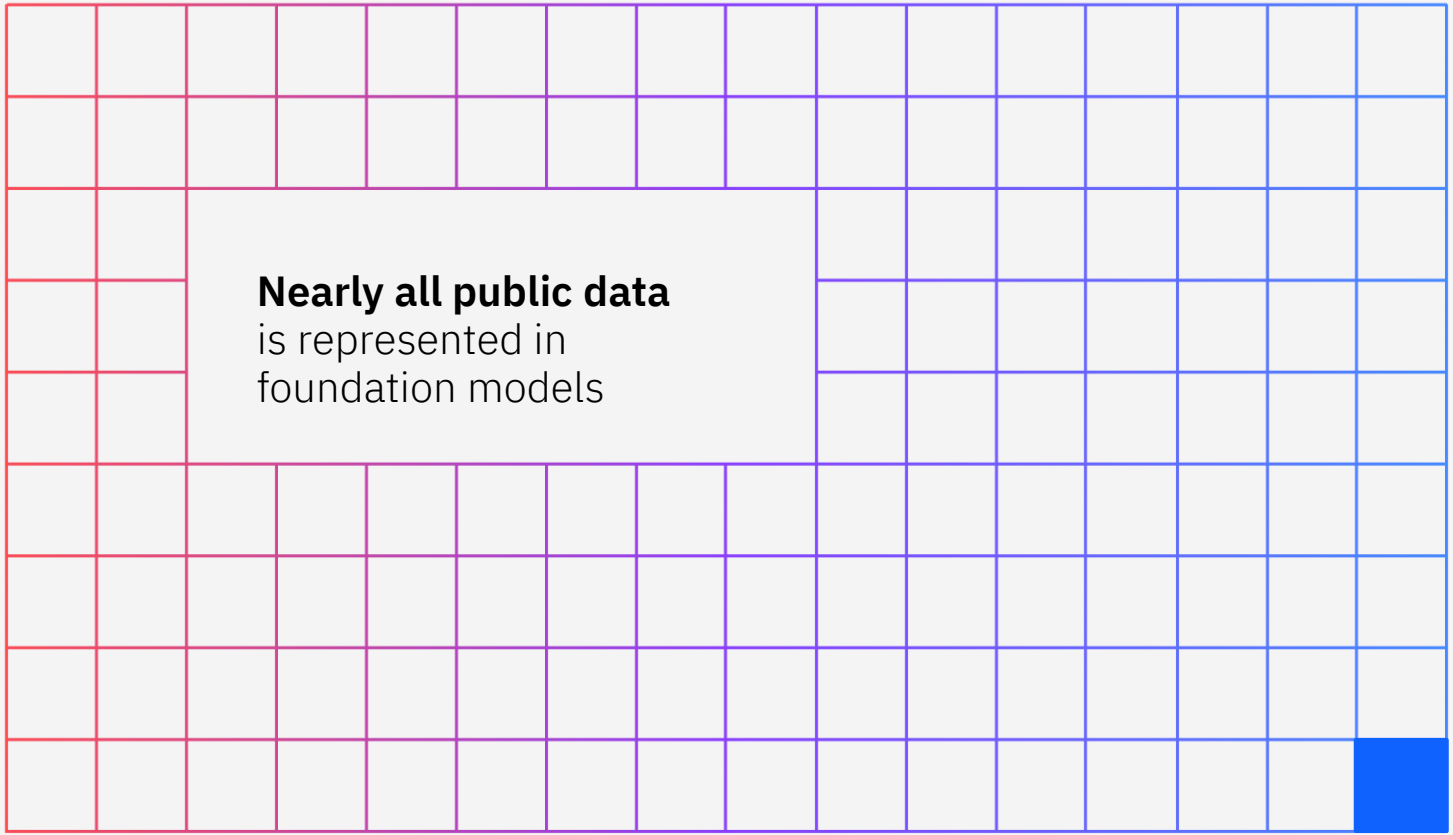
Time savings

© 2025 IBM Corporation



Data

# 데이터는
# 효과적인 AI 전략을 위한 연료




Enterprise data
represents
**less than 1%.**

© 2025 IBM Corporation

# 그러나 신뢰할 수 있는 데이터와 데이터 복잡성에 대한 액세스는 여전히 AI 구현에 대한 최대 장벽으로 남아 있습니다.

## 신뢰할 수 있는 데이터 액세스

**<1/3**

1/3이하의 기술 리더들만이 자사 기업 데이터가 생성형 AI의 효율적인 확장을 지원하는 **접근성**, **품질** 및 **보안** 표준을 충족한다고 믿습니다.

## 증가하는 데이터 복잡성

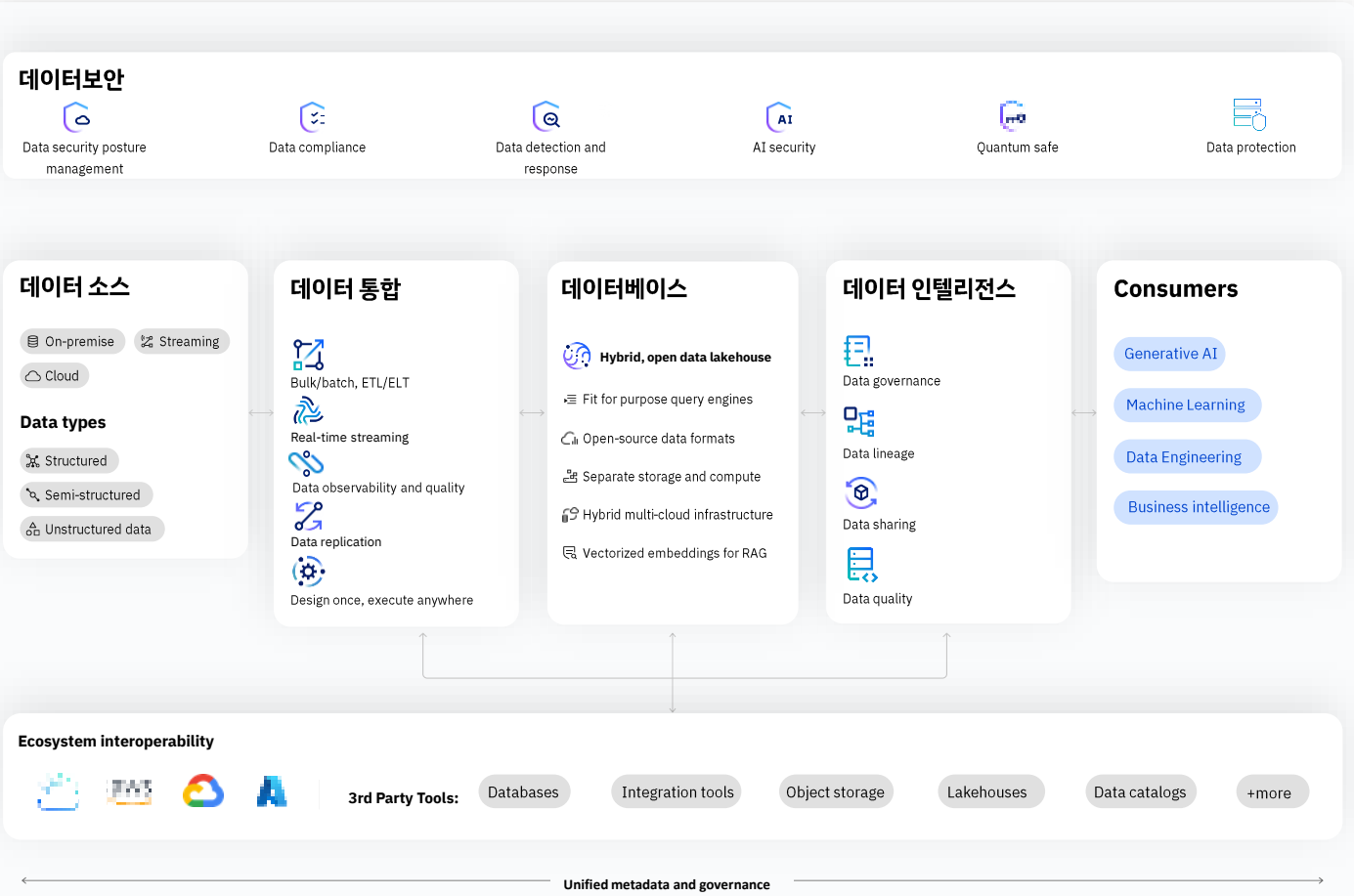
**90%**

모든 기업에서 생성된 데이터의 90%는 **구조화되지 않은 데이터**이며 구조화된 데이터보다 **3배** 빠른 속도로 증가합니다.

Source: IBV *The 2024 CxO Study. 6 blind spots tech leaders must reveal: How to drive growth in the generative AI era*., IDC, UNTAPPED VALUE: What Every Executive Needs to Know About Unstructured Data, Aug 2023

© 2025 IBM Corporation

# 개방형 하이브리드 데이터 아키텍처로 모든 데이터 유형을 통합, 액세스, 관리 및 보호합니다.




Source: IBV *The 2024 CxO Study. 6 blind s*



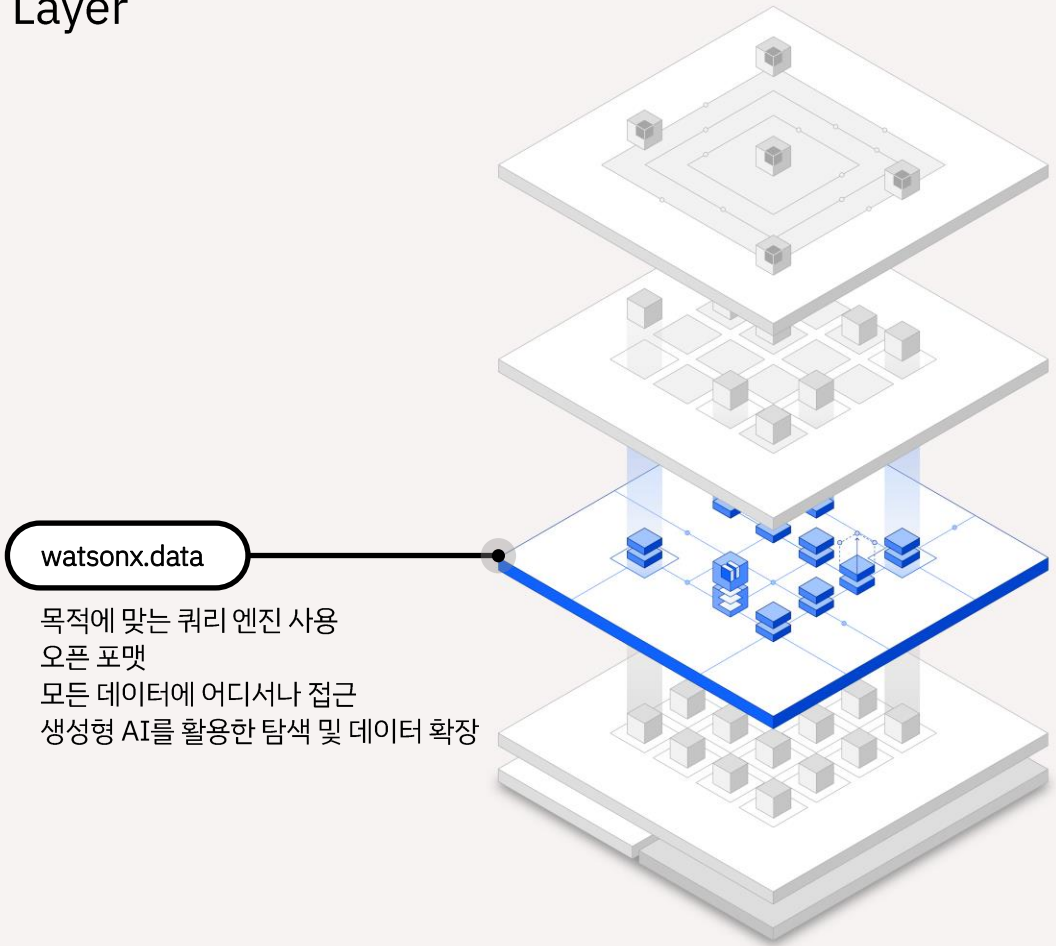
© 2025 IBM Corporation

# AI 워크로드에 적합한 최적의 Data Layer

## Ensure your data is AI ready

AI 활용 및 가속화를 위하여
기업 내 데이터 사일로를
제거하고 모든 데이터를
연결해야 합니다.

***82% 기업이 현재 사일로 문제를 겪고 있음***




Source: VentureBeat

© 2025 IBM Corporation

Governance





# AI 확장의 한계



규정 변경



다양한 이해관계자



부정확한 문서



증가된 위험



다양한 도구 및 데이터



데이터 취약성

© 2025 IBM Corporation

IBM

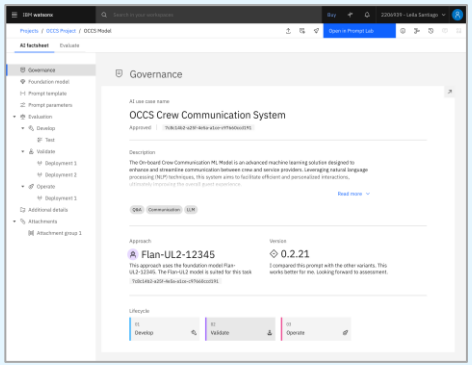
# How to manage the complexity




© 2025 IBM Corporation

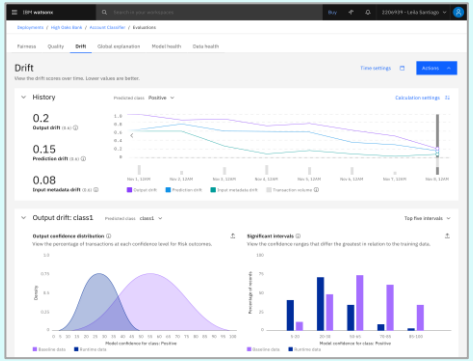
# watsonx.governance – 기업 AI를 위한 Open 거버넌스 플랫폼

## AI Lifecycle Governance
## (AI 생애주기 거버넌스)



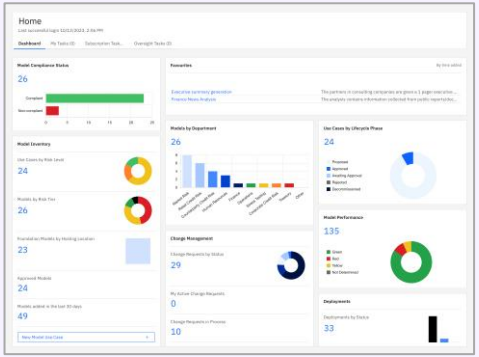

+

## Risk Management
## (전사적 AI 리스크 관리)

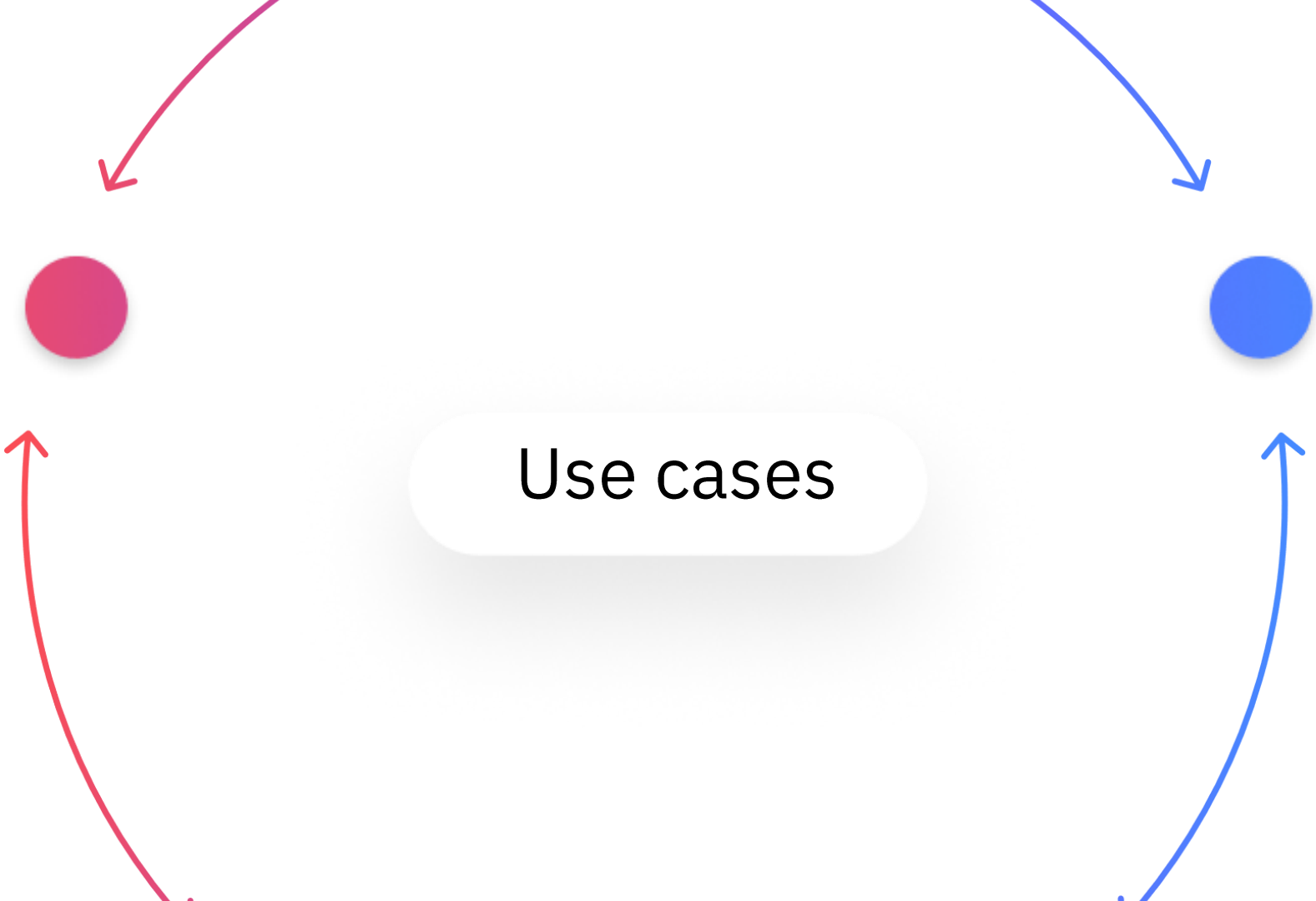



+

## Regulatory Compliance
## (규제 준수)




© 2023 IBM Corporation


Use cases

다양한 영역에 AI 활용 사례들이 있고,

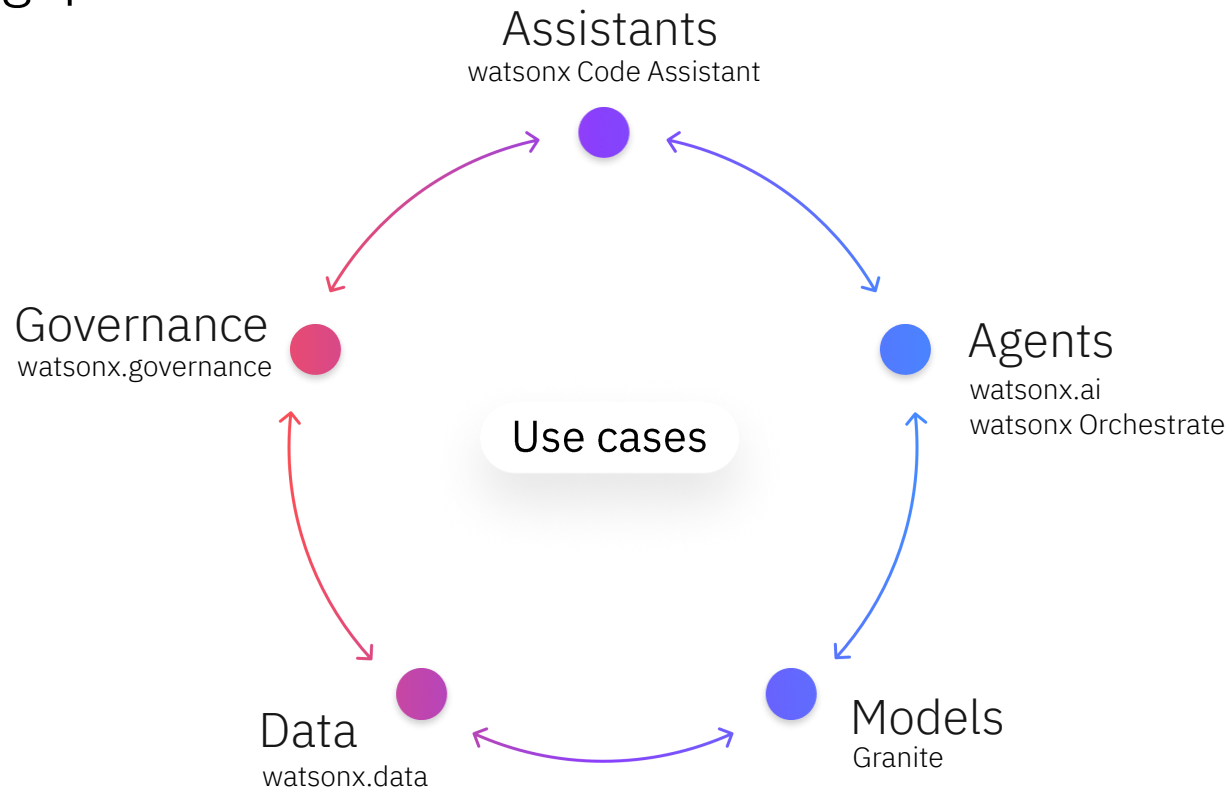
IBM은 광범위한 비즈니스 영역에서 기업 고객과 적극적으로 협업하고 있습니다

| 고객 대면 경험 | HR, 재무 및 공급망 | IT 개발 및 운영 | 핵심 사업 운영 |
|---|---|---|---|
| **고객서비스**<br>고객 문의의 90%를 AI assistant 가 처리<br>vodafone | **HR 자동화**<br>HR 생산성 40% 향상<br>IBM | **앱 현대화, 마이그레이션**<br>60% AI에 의해 자동 생성된 소프트웨어 개발 콘텐츠<br>IBM | **위협관리**<br>잠재적 위협을 8배 더 빠르게 억제<br>TalkTalk |
| **마케팅**<br>2-300만 명의 고객 데이터를 종합하여 고객 조사 수행<br>ARTEFACT | **공급망**<br>송장당 비용 최대 50% 절감<br>Walmart | **IT 자동화**<br>MTTR 50% 감소<br>ExaVault | **자산관리**<br>계획되지 않은 다운타임 43% 감소<br>IDC |
| **콘텐츠 제작**<br>라이브 시청 경험을 비용 효율적으로 확장<br>us open | **Procurement(조달)**<br>회사 비즈니스 관계의 모든 측면에 대한 360도 뷰 확보<br>dun & bradstreet | **AI Ops**<br>12개월 동안 퍼블릭 클라우드 지출에서 $260,000 이상 절감<br>Natura &Co | **제품 개발**<br>5,000개의 테스트 케이스가 몇 분 안에 마이그레이션 및 자동화<br>ContextQA |
| **지식 작업자**<br>분기별 보고서 요약 시간 96% 단축<br>edger .finance | **규정 준수**<br>문제 대응에 소요되는 시간 단축<br>citi | **Cloud FinOps**<br>비즈니스 연속성 저하 없이 클라우드 비용 50% 절감<br>TUI | **Legal**<br>법률 문서를 요약하고 분석하는 데 필요한 시간 90% 단축<br>BG BLENDON GROUP |

© 2024 IBM Corporation

IBM

# 미래를 위한
# AI 구성 요소 활성화




© 2025 IBM Corporation

# IBM Data Portfolio

신뢰할 수 있고 안전한 데이터에 액세스하여
AI 생산성을 향상

## AI Productivity

AI 에이전트/어시스턴트로 작업 수행 방식을 재창조

### AI Assistants

- watsonx Code Assistant™
- watsonx Orchestrate™
- Planning Analytics

## AI/ML Ops

비즈니스를 위해 구축된 AI 모델, 도구 및 거버넌스를 사용하여 애플리케이션의 신뢰와 확장성을 보장하도록 설계

### AI Models

- Granite™
- Meta Llama
- Mistral

### AI Tools

- watsonx.ai™

### AI Governance

- watsonx. governance™

## Data Fabric

모든 비즈니스 데이터를 한데 모아 애플리케이션에서 분석 및 AI를 확장하고 보호하는 동시에 데이터를 시스템에서 이동하는 방식을 최적화

### Databases

- watsonx.data ™

### Data Intelligence

- Data Product Hub
- Knowledge Catalog
- Manta Data Lineage

### Data Integration

- DataStage®
- Databand®
- Streamsets

### Data Security

- Guardium® Data Security Center

## Data Storage

에지, 코어, 클라우드에 걸쳐 데이터 저장

### Software-defined Storage

- Storage Ceph®

IBM에서 제공하는

# 고객 지원 및 체험 프로그램

## Briefing Session
브리핑 세션

신뢰할 수 있는 데이터로 비즈니스 전반에서 AI의 영향력을 확장 및 가속화 할 수 있는 전략과 방법을 데모와 실제 사례를 통해 자세히 알려 드립니다.

- 대상 : CxO, 현업, 데이터 사이언티스트, IT 담당 등
- 소요 시간 : 60분 ~120분
- 진행 방식 : 오프라인 (온라인 진행가능)

## Discovery Workshop
디스커버리 워크샵

watsonx를 활용하여 고객사의 비즈니스 요구사항에 맞는 솔루션을 정의하고, 신속하게 Value를 검증할 수 있도록 IBM 전문 인력이 함께 참여하여 적용 방안을 도출해 보는 워크샵입니다.

- 대상 : CxO, 현업, 데이터 사이언티스트, IT 담당 등
- 소요 시간 : 0.5 ~ 1일
- 진행 방식 : 오프라인

## Hands-On Program
핸즈온 프로그램

기업 내 문서에 대한 수준 높은 인사이트를 얻을 수 있도록 watsonx Discovery를 적용하여 RAG 패턴을 구현하고, watsonx.Assistant 를 이용해서 신속하게 봇을 개발하는 과정을 실습을 통해서 알아봅니다.

- 대상 : 현업, 데이터 사이언티스트, IT 담당 등
- 소요 시간 : 1일 ~일
- 진행 방식 : 오프라인

## Incubation Program
인큐베이션 프로그램

기업이 생성형 AI를 효과적으로 활용할 수 있도록 실전 경험에 중점을 둔 실습 기반 교육이 제공되며, 전문가 멘토링을 통해 고객이 직접 시나리오 기반의 프로토타입을 개발해 보는 프로그램입니다.

- 대상 : 현업, 데이터 사이언티스트, IT 담당 등
- 소요 시간 : 3주 ~ 4주
- 진행 방식 : 오프라인

