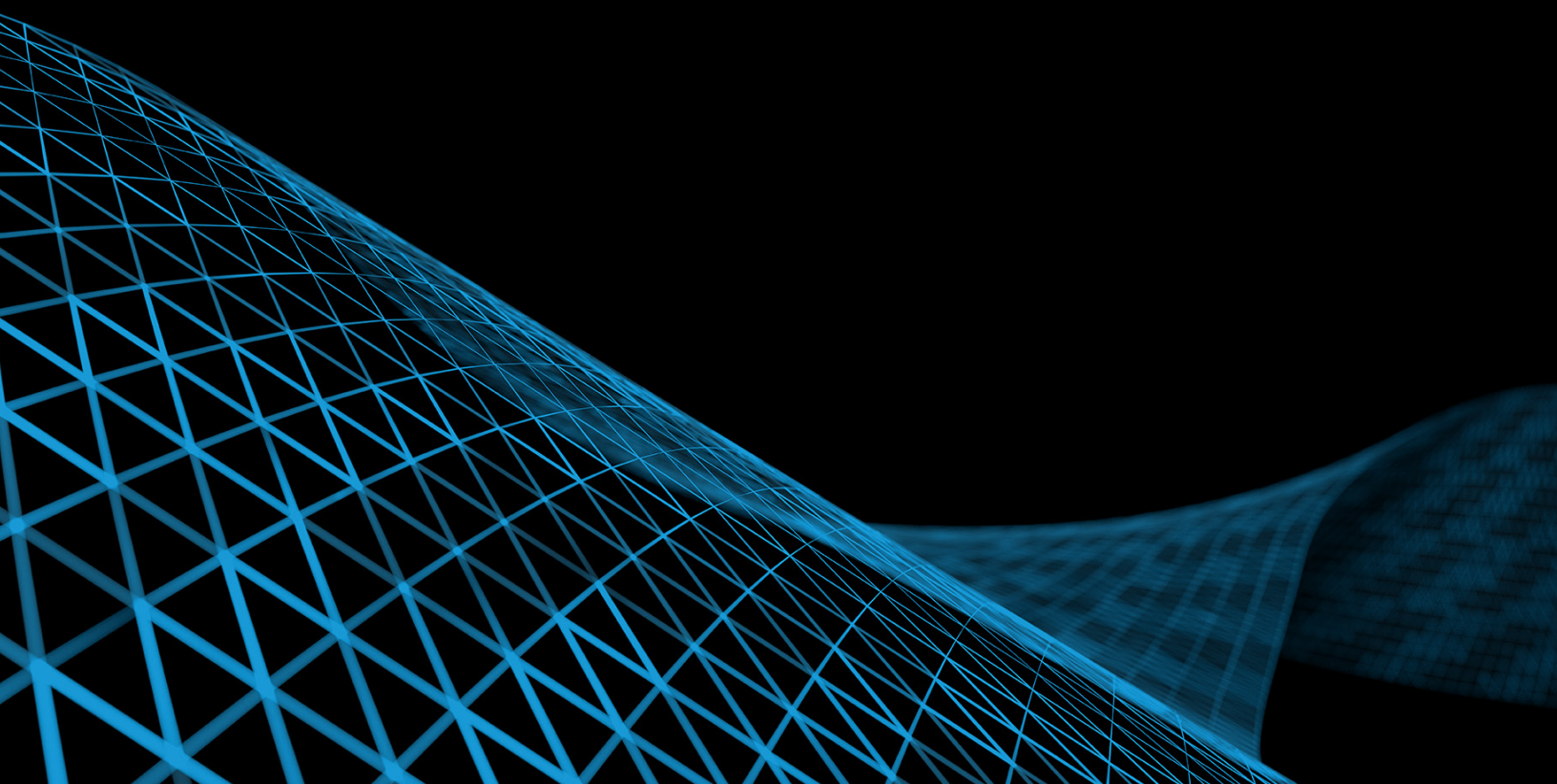
TechTarget | Custom Media
E-Book

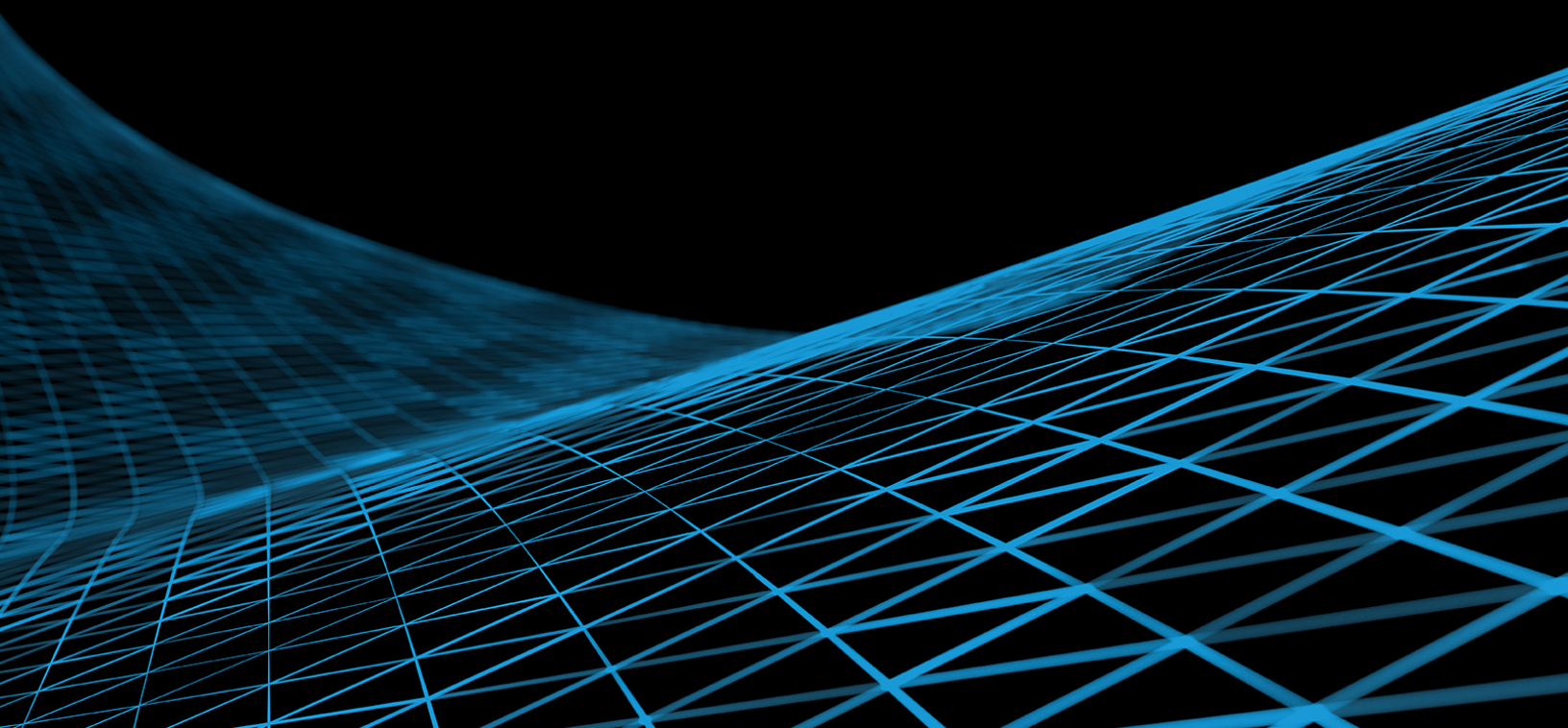
AI가 데이터 센터에서 엣지, 클라우드에 이르기까지 모든 것을 변화시키는 방식

AI의 혁신적인 영향력은 아무리 강조해도 지나치지 않습니다. AI는 모든 업계, 지역, 사용 사례에 걸쳐 클라우드와 엣지 컴퓨팅이 기업에 가치를 제공하는 방식을 본질적으로 변화시키고 있습니다. 이 e-Book은 분산형 클라우드에서 점점 더 중요해지고 있는 AI의 역할에 대한 실질적인 관점을 제공합니다.

Akamai



목차



```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


소개

이 e-Book이 중요한 이유

IT 의사 결정권자, 최고 경영진, 사업부 이해관계자 또는 이사회 구성원이라면 AI가 기업에 미치는 영향에 대해 읽고, 이야기하고, 생각하는 데 점점 더 많은 시간을 할애하고 있을 것이며, 개인적으로 또한 업무에서 AI의 영향을 받고 있을 것입니다.

AI는 이전의 대부분의 유망 기술보다 훨씬 더 빨리 시장에 안착했습니다. AI는 최근 몇 년 동안 초기 시장의 기대가 실제로 정당화되었으며 가시적이고 관찰 가능한 결과가 나타난 몇 안 되는 중요한 기술 중 하나입니다. (이런 사실을 입증하는 수치도 곧 소개할 예정입니다.)

그러나 이 e-Book은 AI의 성장과 잠재력을 현실 세계에 대입해 보여주는 데 그치지 않고 AI를 바라보는 중요한 관점을 제공합니다. 바로 사용자와 가까운 곳에, 분산 클라우드에 있는 AI입니다. 우리는 사용자 주변과 클라우드에 있는 데이터의 중요성에 대해 많이 들어 왔습니다. 이제 분산형 AI 추론과 분산형 클라우드 AI의 시대가 열렸습니다.

그래서 제목을 이 e-Book이 중요한 이유라고 붙였습니다. 의사 결정권자, 인플루언서, 게이트키퍼가 이 e-Book을 읽어야 하는 이유를 4가지로 정리했습니다.

1. **엣지 컴퓨팅과 클라우드 컴퓨팅은 AI 혁신과 가치 창출의 새로운 영역입니다.** 기업은 데이터 센터 외부에서 AI 기능을 생성하고 배포함으로써 지난 10년 동안 광범위한 IT 시장이 걸어온 것과 동일한 길을 따라 새로운 인사이트, 비즈니스 모델, 솔루션의 혜택을 누릴 수 있습니다.
2. **AI는 규제와 거버넌스 면에서 거대한 존재입니다.** AI가 주로 온프레미스 샌드박스에 배치되었을 때 “책임감 있는 AI” 사용과 실행이 까다롭다고 생각했다면, 사용자 주변과 분산형 클라우드에서 AI의 컴플라이언스, 데이터 보호, 보안 문제를 고려해 보세요. 기업은 이 영역을 제대로 파악해야 합니다. 그렇지 않으면 많은 문제에 부딪힐 리스크가 있습니다.
3. **AI가 변화하는 인력 패턴과 관행에 미치는 영향은 예상보다 훨씬 클 것입니다.** 많은 반복적인 IT 기능이 인간 노동자에서 AI로 급격하게 이동할 것이라는 데는 의심의 여지가 없지만, AI를 데이터 센터 밖으로 이동시키면 IT와 비즈니스 노동자 모두에게 흥미롭고 영감을 주는 기회가 될 것입니다.
4. **낭비할 시간이 없습니다.** AI 관련 이니셔티브는 모든 기업의 전략적 우선순위 목록에서 가장 높은 순위를 차지하고 있습니다. 새로운 사용 사례를 테스트하고 AI가 할 수 있는 일에 대해 더 많이 배우기 위해 온프레미스 AI를 충분히 활용하지 않는 기업은 의심할 여지 없이 뒤처지고 있을 것입니다. 그러나 사용자 주변과 분산형 클라우드에 있는 AI로 이동하는 것은 새로운 기회를 의미합니다.

이 e-Book이 유용한 또 다른 이유는 마지막 부분에서 기업이 분산형 클라우드의 AI를 최대한 활용할 수 있도록 실행 가능한 조언을 제공한다는 것입니다.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh

1장

AI 혁명과 그것이 시장 역학에 가져온 변화

AI는 실제로 수십 년 동안 존재해 왔는데, 현재의 시대를 "AI 혁명"이라고 부르는 것이 타당할까요? 아니면 AI는 단순히 이전의 형태와 기능에서 더 빠른 속도로 진화하고 있는 것일까요? 이것은 수사적인 질문이 아닙니다. 업계에서는 마치 하룻밤 사이에 기술적 혁명이 닥친 것처럼 이야기하는 경향이 있습니다. 실험실에서 이미 실험을 해 본 기술 전문가들 외에는 아무도 알아채지 못했다고 생각합니다.

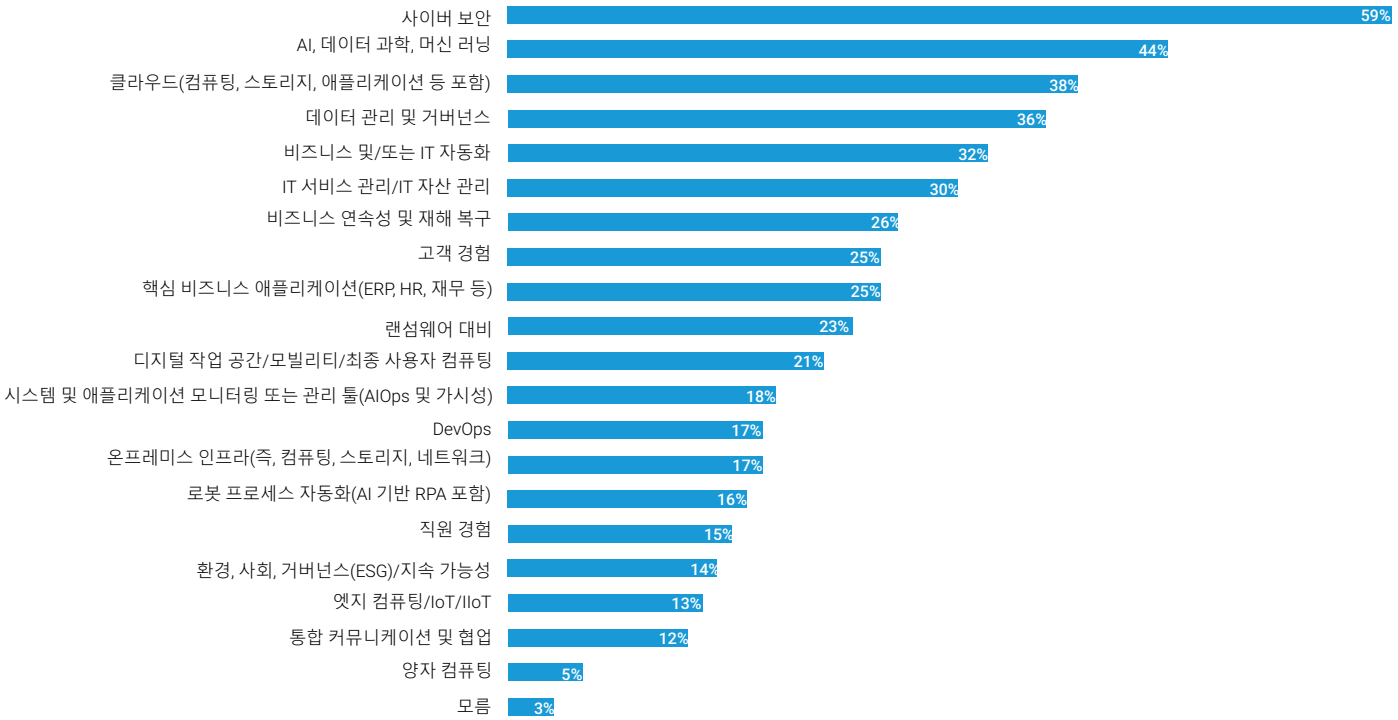
그러나 현재를 AI 혁명이라고 부르는 것은 여러 가지 이유로 타당합니다. 첫째, AI 시장의 성장과 기업 리더들이 생각하는 AI 기술의 중요성이 놀라울 정도로 커졌다는 것은 부인할 수 없는 사실입니다. AI의 성장률과 현재 시장 영향력은 개인용 컴퓨터, 근거리 통신망, 클라우드 컴퓨팅, IoT 등 IT 업계 발전의 이정표로 여겨져 온 기술적 움직임을 훨씬 능가합니다.

다음의 데이터 포인트는 AI 혁명의 영향, 중요성, 잠재력을 보여줍니다.

- AI 관련 하드웨어, 소프트웨어, 서비스에 대한 전 세계 지출이 2024년 말까지 6천억 달러를 넘어섰고 2032년에는 2조 7천억 달러를 넘으며 연평균 성장률이 20% 이상 급증할 것으로 예상됩니다.[1]
- 비즈니스 리더의 97%가 기업이 AI 투자에 대해 긍정적인 수익을 얻고 있다고 말했습니다.[2]
- AI를 기술 우선순위 2위 안에 넣은 CEO의 비율이 단 1년 만에 4%에서 24%로 증가했습니다.[3]
- 아래 그림에서 볼 수 있듯이 지난 2년 동안 미국 기업의 거의 절반이 AI가 기업의 미래에 훨씬 더 중요하다고 답했으며, 기업 이니셔티브로 AI보다 상위에 있는 것은 사이버 보안이 유일합니다.[4]

그림 1. 중요한 기술 이니셔티브로 자리매김한 보안, AI, 클라우드, 데이터 관리

지난 2년 동안 다음의 광범위한 기술 이니셔티브 중 어느 것이 기업의 미래에 훨씬 더 중요해졌습니까? (응답자 비율, N=1351, 복수 응답 가능)



1 출처: Fortune Business Insights, "Artificial Intelligence market size ... 2024-2032," 2024년 12월.
2 출처: Lizzie McWilliams, "Artificial intelligence investments set to remain strong in 2025, but senior leaders recognize emerging risks," EY.com, 2024년 12월 10일.
3 출처: LinkedIn, Gartner's 2024 CEO survey reveals AI as top strategic priority, 2024년 5월.
4 출처: Enterprise Strategy Group Complete Survey Results, *2025 Technology Spending Intentions Survey*, 2024년 12월.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

실제로, AI 도입과 자본 투자의 급증은 IT 환경에서 가장 중요하고 빠르게 성장하는 분야 중 하나인 클라우드 컴퓨팅의 성장률을 능가했습니다. AI의 성장이 클라우드의 성장을 능가하는 이유는 매우 간단합니다. 클라우드 컴퓨팅은 IT 서비스를 개발하고 제공하는 데 있어 강력하고 활용도가 높은 방법이지만, AI는 우리가 일하고, 생활하고, 상호 작용하는 방식을 근본적으로 변화시키고 있습니다.

또한, AI의 급속한 도입으로 인해 기존의 클라우드 컴퓨팅 아키텍처의 한계가 드러났습니다. 예를 들어, 중앙집중식 클라우드 모델은 AI의 실시간, 데이터 집약적 특성에 따른 높은 성능 요구사항에 대응하지 못해 어려움을 겪고 있습니다. 본질적으로, AI는 성능, 비용, 인력 기술, 보안 및 데이터 거버넌스와 같은 문제와 관련해 클라우드 아키텍처에 대한 기대치를 극적으로 변화시켰습니다.

AI 시장의 급격한 성장과 시장 영향력을 주도하는 요인은 무엇일까요? 주요 요인은 지난 1년 동안 기업들이 AI 모델을 훨씬 더 분별력 있고 정교하게 사용했기 때문입니다. AI 모델의 유용성과 사용 편의성이 급속히 증가하고 많은 기업들이 경험한 인상적인 ROI로 인해 AI 사용 방법에 대한 초기 사고를 강화하고 AI 모델의 장점을 활용하는 방법에 대한 완전히 새로운 사고 방식을 여는 중요한 사용 사례가 등장했습니다.

다음과 같은 예가 있습니다.

- **비즈니스 예측**은 항상 성공적인 기업의 필수적인 부분이었으며, AI는 그 기능을 더욱 정확하고, 더 정확하고, 더 시기적절하고, 더 실행 가능하게 만들었습니다. 최신 AI 모델이 기업의 자체 데이터로 강화되어 있기 때문에 예측 AI 모델과 같은 툴을 사용해 미래의 트렌드를 예측하는 것은 기업이 빅 데이터 분석을 도입했던 10년 전보다 훨씬 더 발전된 기술입니다. 이를 통해 데이터 과학자뿐만 아니라 비즈니스 경영진도 기존 모델보다 더 빠르고 정확하게 방대한 규모의 데이터를 면밀하게 조사할 수 있습니다.
- **문서 요약**은 과거에는 사무직 근로자의 영역이었지만, 생성형 AI(GenAI)는 이 게임의 룰을 바꾸었습니다. 긴 문서(난해하고 고도의 기술적 내용이 가득한 문서 포함)를 검토, 분석, 요약하는 작업이 훨씬 더 짧은 시간 안에 자동으로 이루어집니다. 이는 더 나은 인사이트를 이끌어낼 뿐 아니라 IT 전문가와 비즈니스 전문가가 더 혁신적이고 전략적인 방식으로 업무를 수행할 수 있도록 함으로써 생산성을 높이고 직원 경험을 개선합니다.
- 고객 관계 관리에서 **고객 손실 모델링**은 이제 중요한 요소로 자리잡았습니다. 기업이 고객 손실을 예측하고, 고객이 어떻게, 언제, 왜 다른 곳으로 옮기는지 또는 고객 매출을 늘릴 수 있는 이유를 파악하는 데 도움이 됩니다.
- **AI 기반 챗봇**은 일상적인 요청에 대해 지능적인 실시간 셀프 서비스를 가능하게 함으로써 고객 서비스, IT 서비스 관리, 인력 기능을 변화시키고 있습니다. 챗봇은 코드 생성 및 새로운 기능 모델링과 같은 고도의 기술적 엔지니어링 및 컴퓨터 과학 활용 사례에도 사용됩니다.

이러한 활용 사례는 사이버 보안 위협 탐지, 데이터 분석 및 비즈니스 인텔리전스, 경쟁 분석, 신속한 프로토타이핑, 컴플라이언스 관리 등과 같은 많은 다른 AI 애플리케이션, 워크플로우, 활용 사례를 창출하는 데 도움이 되었습니다.

물론 AI만큼 추진력과 가능성을 가진 새로운 기술도 개발, 배치, 관리에 어려움이 없는 것은 아닙니다. 그러한 어려움에는 데이터의 낮은 품질, 올바른 인프라 구축, 대규모 언어 모델(LLM)을 구동하는 데 필요한 고성능 보장 등과 같은 기술적 문제가 있을 수 있습니다. 또한 데이터 거버넌스, 책임 있는 AI 사용 보장, 점점 커지고 있는 AI 기술 격차 해소 등과 같이 비즈니스 또는 리스크와 관련된 것일 수도 있습니다.

그러나 지금까지 기업들은 이러한 과제를 이해하고 있으며, 적절한 리소스와 최고 경영진 및 이사회의 지원을 통해 이러한 과제를 해결하고 있습니다. 다음 장에서는 AI를 과학 프로젝트 이상의 것으로 만드는 주요 발전 사항 중 하나인 AI의 초점을 모델 학습에서 추론으로 전환하는 것에 대해 설명하겠습니다.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh

2장

AI 패러다임의 변화: 학습에서 추론으로

AI 모델 학습은 시장 발전에 중요한 요소였으며, 기업이 AI 이니셔티브에 처음 집중할 때 반드시 중점을 두는 부분이었습니다. 기업은 데이터의 낮은 품질(또는 적절한 유형과 양의 데이터가 부족함)이 AI 모델 개발에 영향을 미친다는 사실을 일찍 깨달았습니다. Informa TechTarget의 Enterprise Strategy Group의 조사에 따르면, 기업의 37%가 GenAI를 구축할 때 가장 큰 도전 과제로 데이터 품질 문제를 꼽았으며, 이는 AI 기술 격차에 이어 두 번째로 많이 언급된 과제였습니다.[5]

기업이 자체 데이터를 활용하는 것이 데이터 품질 문제를 해결하고, 더 정확하고 맥락에 맞게 인식하며 반응성이 뛰어난 LLM을 만드는 가장 좋은 방법이라는 것을 인식했기 때문에, 이제는 적절한 AI 모델을 구축할 수 있게 되었습니다. 이러한 접근 방식은 데이터 편향과 부정확성 또는 맞춤화 기능의 부족을 야기할 수 있는 써드파티 데이터를 기반으로 학습된 상용 기성 AI 모델을 구입하는 것보다 훨씬 낫습니다. 대부분의 기업에서 표준 관행은 중앙집중식 클라우드 환경에서 LLM 학습에 집중하는 것이었고, 이는 LLM 개발의 초기 단계에 효과적이었습니다.

AI 모델 학습은 품질과 무결성 면에서 성숙하고 안정화되었습니다. 이미 기업들은 자체 구축한 교육 모델에 더 적은 리소스를 할당하고 있으며, 특히 AI 튜닝과 최적화로 인해 컴퓨팅 집약적이며 GPU 중심의 인프라에 대한 필요성이 줄어들고 있습니다.

Enterprise Strategy Group의 AI, 애널리틱스, 비즈니스 인텔리전스 담당 이사 마이크 리오니(Mike Leone)는 학습에서 추론으로의 전환에 대해 독립적인 관점을 제시했습니다.

그는 "사전 학습된 모델과 실시간 AI 애플리케이션이 더욱 성숙해지고 널리 보급됨에 따라, 추론은 이제 AI의 약속된 가치를 효율적이고 비용 효율적으로 전달하는 데 핵심적인 역할을 하고 있습니다." 라며 “이러한 변화는 업계 전반에 걸쳐 확장 가능한 배포를 가능하게 하고 있으며 하드웨어와 소프트웨어 스택에서 지속적으로 이루어지는 발전이 이 변화를 뒷받침하고 있습니다. 벤더사는 AI 인프라의 적정 규모와 최적화, 모델 효율성 개선, 지속적인 운영 비용 관리에 집중하고 있습니다. 물론 추론으로 AI 가용성이 널리 확대됨에 따라 확장성, 개인 정보 보호 문제, 규제 심사와 같은 도전 과제가 발생하고 있습니다.” 라고 말했습니다.

AI 모델 학습에서 추론 기반 AI 애플리케이션과 고영향 사용 사례로의 전환이 활발하게 진행되고 있습니다. IBM의 AI 플랫폼 담당 부사장 Armand Ruiz는 이를 간결하고 강력하게 표현했습니다. “제 생각에 추론은 AI 워크로드를 지배할 것입니다. … 일단 모델이 학습을 제대로 하고 나면 이상적으로는 더 이상 학습할 필요가 없습니다. 그러나 추론은 계속 진행됩니다.”[6]

시장 전환이 일어날 때는 자본이 투자되는 방식과 장소에 불균형적인 변동이 불가피하게 발생합니다. 예를 들어, 한 추산에 따르면 AI 실리콘 칩에 대한 지출의 약 15%가 AI 학습에 사용되고, 45%는 데이터 센터 기반 AI 추론에, 나머지 40%는 엣지 기반 추론에 사용된다고 합니다.[7]

5 출처: Enterprise Strategy Group Complete Survey Results, “*The State of the Generative AI Market: Widespread Transformation Continues*,” 2024년 9월.
6 출처: LinkedIn, Armand Ruiz, 2024년 12월.
7 출처: Jonathan Goldberg, “The AI chip market landscape: Choose your battles carefully,” TechSpot.com, 2023년 5월 30일.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


다른 리서치에서도 추론에 대한 비슷한 경향을 지적하고 있으며 Google의 한 연구는 이러한 트렌드를 다른 방식으로 강조했습니다. Google은 머신 러닝을 중심으로 한 에너지 소비의 60%가 학습에 사용되는 것이 아니라 추론에 사용된다고 밝혔습니다.[8]

당연히 AI 모델 학습에서 AI 추론으로의 전환은 AI 인프라와 클라우드 아키텍처에 중요한 영향을 미칩니다. 이러한 변화는 AI 애플리케이션의 개발, 배포, 관리에 사용할 적합한 클라우드 환경과 아키텍처를 찾는 기업에 특히 중요합니다.

기존의 중앙집중식 클라우드 모델은 많은 비즈니스 워크로드에 여전히 잘 작동하지만, AI는 대규모 AI 성능 요구를 충족하기 위해 클라우드 아키텍처에 많은 것을 요구합니다. 이러한 문제는 쿼리와 응답 사이에 눈에 띄는 시간 차이가 있는 지연 시간 문제와 같은 형태로 나타날 가능성이 높습니다.

컴퓨팅 인프라 또한 기업이 AI 모델 학습에서 AI 추론으로 전환할 때 고려해야 할 핵심적인 문제입니다. AI 중심 GPU는 높은 연산 성능과 메모리 대역폭을 제공하기 때문에 LLM을 학습할 수 있는 능력을 인정받고 있습니다. 이에 비해 추론은 저지연 시간과 효율적인 리소스 사용에 초점을 맞춘 다른 인프라 사고 방식을 필요로 합니다.

다음 장에서는 추론 리소스를 사용자 주변과 분산형 클라우드 아키텍처에 배치함으로써 기업이 AI 추론의 성능 요구를 충족할 수 있도록 지원하는 방법을 자세히 살펴보겠습니다.

8 출처: “The AI boom could use a shocking amount of electricity,” Scientific American, 2023년 10월 13일.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

3장

분산형 AI 추론의 부상 및 분산형 클라우드

이전 장에서 설명한 바와 같이, 기존의 중앙집중식 클라우드 아키텍처는 획기적이고 혁신적인 애플리케이션을 만드는 데 AI 사용을 최적화하려는 기업에 문제가 되는 한계와 도전과제를 발생시킵니다. 지연 시간, 네트워크 대역폭 제한, 확장에 따른 성능 보장, AI 요구사항이 모델 학습에서 추론으로 이동함에 따라 발생하는 관리 복잡성과 같은 문제를 모두 고려하고 계획해야 합니다.

최신 및 차세대 AI 추론 애플리케이션은 데이터가 있는 곳에서 추론을 개발하고, 배포하고, 활용할 수 있는 능력이 필요합니다. 사용자 주변이나 민첩하고, 확장 가능하고, 유연한 분산형 클라우드 아키텍처의 중요성이 점차 커지고 있습니다.

더 깊이 들어가기 전에 몇 가지 명확한 정의가 도움이 될 것입니다.

- 분산형 AI 추론은 중앙 집중식 데이터 센터(클라우드)와 인간과 물리적 사물(엣지)에 더 가까운 클라우드 외부의 디바이스에 걸쳐 있는 AI 워크플로우를 만드는 패러다임입니다. 이는 사람들이 클라우드 AI라고 부르기 시작한 AI 애플리케이션이 전적으로 클라우드에서 개발되고 실행되는 일반적인 관행과는 대조됩니다. 또한 사람들이 데스크톱에서 AI 알고리즘을 제작한 다음, 데스크톱이나 특수 하드웨어에 배포해 작업(예: 수표 번호 읽기)을 수행하는 이전의 AI 개발 접근 방식과도 다릅니다.

클라우드 서비스의 분산은 기업이 애플리케이션 성능 및 응답 시간, 엣지 컴퓨팅, 규제 의무 또는 기타 요구사항에 대한 목표를 달성하는 데 도움이 될 수 있습니다.

- 클라우드, 보안, 콘텐츠 전송 플랫폼을 제공하는 선도 기업인 Akamai는 분산형 클라우드를 "기업이 여러 지리적 위치에 퍼블릭 클라우드 인프라를 활용하는 클라우드 컴퓨팅의 한 형태로서, 운영, 관리, 업데이트가 단일 퍼블릭 클라우드 서비스 사업자에 의해 중앙에서 관리되는 클라우드 컴퓨팅"이라고 정의합니다. 클라우드 서비스를 분산시키면 기업이 최종 사용자 및 디바이스와 가까운 위치에서 클라우드 서비스를 제공해 애플리케이션 성능 및 응답 시간, 엣지 컴퓨팅, 규제 의무화 또는 충족할 수 있는 기타 요구사항에 대한 목표를 달성하는 데 유용할 수 있습니다. 사물인터넷(IoT) 네트워크, 인공 지능, 방대한 양의 데이터를 실시간으로 처리해야 하는 기타 사용 사례가 증가하면서 분산형 클라우드 컴퓨팅의 활용이 탄력을 받고 있습니다.

분산형 AI 추론과 분산형 클라우드는 모두 AI 추론의 개발과 활용에 필수적인 요소입니다. 왜냐하면 이 두 가지 요소는 앞서 언급한 고성능, 대규모 성능, 컴퓨팅, 저장 공간, 네트워킹 리소스의 효율적인 사용, 그리고 대용량 데이터 세트와 관련된 지연 시간 문제를 극복하기 위한 현대적이고 효율적이며 안전하고 신뢰할 수 있는 방법을 제시하기 때문입니다.

실제 생활에서는 어떻게 작동할까요? 한 회사는 개인 맞춤화를 고객 경험의 핵심 요소로 삼고자 했고, 모든 종류의 디바이스와 데이터 소스에서 데이터를 수집할 수 있어야 했습니다. 이 회사는 LLM 기반 AI 챗봇을 사용하기로 결정했고, LLM 기반 AI 챗봇은 현지화된 데이터를 사용자 주변의 위치로 분산시켰습니다. 이로 인해 지연 시간이 줄어들고 응답 시간이 개선되어 성능이 크게 향상되고 더 빠르고 개인화된 고객 지원이 가능해졌습니다.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

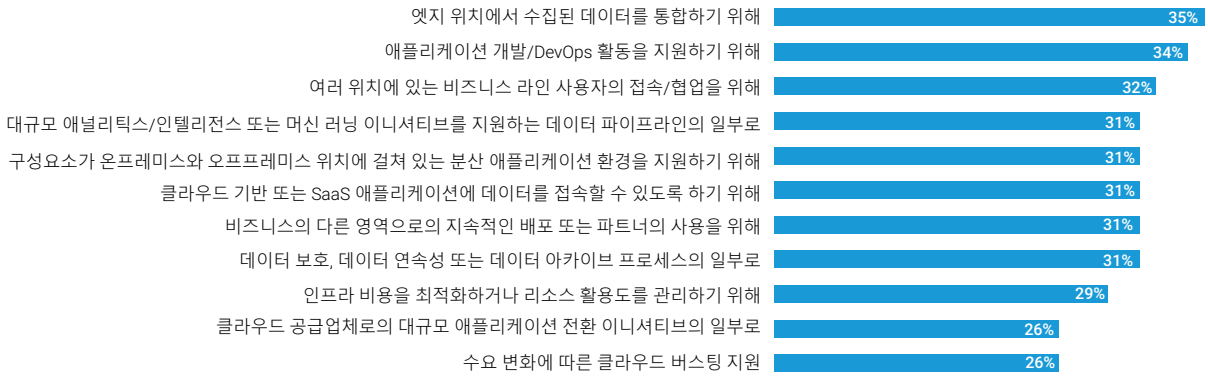

이 접근 방식은 AI 추론 애플리케이션의 성장 기회를 활용하고자 하는 기업에서 점점 더 일반화되고 있습니다. 컴퓨팅 인프라의 한계와 지연 시간 문제로 인해 발생하는 성능 문제를 해결하기 위해, AI 워크로드는 분산형 AI 추론과 분산형 클라우드 아키텍처를 통해 더 많은 최종 사용자가 활용하는 데이터 소스에 더 가까이 이동하고 있습니다.

예컨대, 이 모델은 리테일, 통신, 미디어 등 시설이 분산되어 있고 지리적으로 분산된 업계에서 널리 사용됩니다. 이러한 접근 방식은 복잡한 AI 추론 요구사항과 관련된 지연 시간과 대규모 성능 문제를 극복할 수 있기 때문에 AI 추론용 표준 아키텍처 모델로 사용되는 사례가 늘고 있습니다.

AI 추론을 위한 분산형 클라우드 모델로의 전환을 주도하는 주요 요인은 데이터 센터와 클라우드 서비스 사업자(CSP) 간에 데이터를 이동하기 위해 엣지를 사용해야 한다는 기업의 요구가 증가하고 있다는 것입니다. Enterprise Strategy Group은 데이터 센터와 CSP 인프라 간에 정기적으로 데이터를 이동하는 기업의 35%가 엣지 위치에서 수집된 데이터를 통합하기 위해 그렇게 하고 있으며, 이것이 데이터 이동의 가장 큰 동기라고 밝혔습니다.[9]

그림 2. 엣지, 데이터 센터와 퍼블릭 클라우드 서비스 간에 데이터를 이동하는 데 가장 큰 영향을 미치는 요인

귀하는 데이터 센터와 퍼블릭 클라우드 서비스 간에 정기적으로 데이터를 이동한다고 응답했습니다. 어떤 목적으로 데이터 센터와 퍼블릭 클라우드 서비스 간에 데이터를 이동합니까? (응답자 비율, N=170, 복수 응답 가능)



9 출처: Enterprise Strategy Group Research Report, *Distributed Cloud Series: The State of Infrastructure Modernization Across the Distributed Cloud*, 2023년 11월.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

4장

분산형 AI 추론이 미래인 이유

엣지 위치의 데이터를 활용해 AI 추론을 구축하고 배포하는 것은 기업이 추론을 최대한 활용하는 데 핵심적인 요소입니다. 스마트폰, 특수 목적의 휴대용 디바이스, 디지털 사이니지, IoT 디바이스 등 다양한 엣지 디바이스에서 생성되고 저장되는 데이터가 훨씬 더 많아짐에 따라, 기업은 AI 추론에 필수적인 실시간 데이터 처리를 달성하기 위해 엣지 사고 방식을 도입해야 합니다.

AI 추론 요구사항에 부합하기 위한 분산형 AI 추론으로의 전환은 빠르고 심오했습니다. 그리고 리서치에 따르면 사용자와 가까운 추론 시스템에 대한 지출은 2024년 270억 달러에서 2032년에는 거의 2700억 달러로 10배 증가할 것으로 예상됩니다.[10]

분산형 AI 추론은 기업이 AI 추론 애플리케이션을 개발하고 배포하는 데 도움이 되는 여러 가지 중요한 고려 사항, 즉 디바이스 네이티브 AI 처리, 엣지 디바이스에 최적화된 통합 보안 프레임워크, 낮은 에너지 소비, 짧은 지연 시간 등을 특징으로 합니다. 이러한 기능은 규모에 따른 더 빠른 성능, 더 정확하고 심층적인 데이터 인사이트, 대역폭 투자 절감으로 이어집니다. 이로써 지리적으로 분산된 기업의 구성원들은 AI 추론을 사용해 데이터가 생성, 저장, 처리, 접속되는 위치에 훨씬 더 근접한 곳에서 더 스마트한 실시간 의사 결정을 내릴 수 있습니다.

Bloomfield는 분산형 AI 추론을 사용해 필수적인 운영을 개선하고 데이터 관리를 간소화한 미래 지향적인 기업의 한 예입니다. 분산된 스마트 카메라에서 수집되는 방대한 양의 데이터를 효율적, 경제적이면서도 안전하게 관리해야 하는 경우가 많았습니다. 농부들이 네트워크 접속이 제한된 환경에서 작업하는 경우가 많기 때문입니다. Bloomfield는 분산형 AI 추론 아키텍처 모델을 사용해 AI 추론을 활용해 데이터를 구성하고, 저장하고, 표시하는 최선의 방법을 결정했습니다. 그 결과, 농작물 건강 관리가 간소화되고, 생산량이 향상되었고, 온프레미스 데이터 센터에서 멀리 떨어진 곳에서도 더 빠르고 정확한 의사 결정을 내릴 수 있게 되었습니다.

분산형 AI 추론은 다양한 분산된 위치에서 AI 추론 애플리케이션을 지원할 수 있는 능력을 갖추고 있기 때문에 다음과 같은 사용 사례에 적합합니다.

- **병상 의학과 원격 건강 관리 애플리케이션:** 이동 중인 의사가 스마트폰, IoT 디바이스, 기타 가볍지만 강력한 휴대용 디바이스를 사용해 의사 결정을 내립니다. 분산형 AI 추론은 또한 주입 펌프, 심박 조정기, 심장 모니터, 약물 애널리틱스 같은 헬스케어 분야에서 스마트 디바이스를 사용해 실시간으로 의사 결정을 내리는 데 이상적입니다.
- **스마트 도시:** 엣지 시스템이 교통 흐름을 제어하고 공공 시설 시스템의 동작을 모니터링하는 것부터 유권자 등록을 간소화하고 공공 안전을 개선하는 것까지 모든 것을 수행합니다.
- **무인 차량:** 스마트 카메라와 온보드 센서 데이터를 사용해 경로 옵션을 평가하고, 연료 사용량을 모니터링 및 개선하고, GPS 변경 사항을 기록하고, 법 집행 기관 및 공공 안전 기관과 소통합니다.
- **리테일:** IP 비디오 감시, 재고 관리, 도난 방지, 고객의 매장 내 이동 패턴 등 다양한 AI 추론 애플리케이션을 지원합니다.

기업들이 실시간 질문에 대한 응답 지연 시간을 줄일 방법을 모색함에 따라, 분산형 AI 추론은 보다 효율적이고, 정보에 입각하고, 보다 정확한 비즈니스 의사 결정의 원천을 위한 중요한 요건이 될 것입니다. AI 애플리케이션은 더 이상 멀리 떨어진 데이터 센터에 데이터를 보내거나 요청할 필요가 없습니다. 대신 전 세계에 분산형 클라우드에서 데이터를 처리하고 공유할 것입니다.

한 주요 광고 기술 회사의 엔지니어링 책임자는 "AI 모델을 최적화하는 데 많은 시간을 할애했습니다. 이제 추론을 미루면 안 된다는 것을 깨달았습니다." 라고 말했습니다.

기업은 더 이상 AI 추론을 비즈니스 전략의 주요 구성요소로 도입할지 말지를 결정하는 사치를 누리지 못합니다. 기업은 긴박감을 느끼고 분산형 AI 추론을 추론에 사용해야 합니다. 많은 경쟁사들이 이미 사용자 주변에서 AI 추론을 사용하기 위해 큰 진전을 이루고 있기 때문입니다.

10 출처: Fortune Business Insights, "Edge AI Market Size ... 2024-2032," 2024년 12월.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


5장

전략적 고찰: AI, 엣지, 분산형 클라우드의 미래와 그 의미

검증된 기술 파트너를 평가하고 선택하고 협력하는 것은 기업의 성공적인 AI 추론 이니셔티브의 기본 요소이지만 달성하기는 쉽지 않습니다.

AI 인프라, 클라우드 아키텍처, 추론 사용 사례, 애플리케이션 최적화 및 완벽한 배포와 관련해 상당한 AI 기술 격차가 존재한다는 사실은 부인할 수 없기 때문입니다.

그러나 적합한 후보가 부족하기 때문에 까다롭기도 합니다. 위에서 언급한 바와 같이, 성공적인 AI 추론 노력은 기술 노하우, 전문 지식, 보안, 거버넌스, 리스크 관리에 대한 이해, 구축과 지속적인 관리에 대한 철저한 세부 사항에 대한 관심을 완벽하게 결합해야 합니다.

기업은 코어에서 엣지, 클라우드에 이르기까지 물리적 기업과 가상 기업 전체에 걸쳐 AI 추론을 개발하고 배포하기 위해서 이전에 이를 수행한 경험이 있는 공급업체와 협력해야 합니다. 성공적인 AI 프로젝트의 기술적 요소와 비즈니스적 요소를 모두 이해하는 파트너를 확보하는 것은 필수적입니다.

클라우드 컴퓨팅, 보안, 콘텐츠 전송을 위한 정교한 솔루션을 제공하는 선도 기업인 Akamai는 모델 학습에서 추론에 이르기까지 AI 이니셔티브를 구축하고 관리하는 데 있어 확고한 실적을 보유하고 있습니다. 복잡하고 컴퓨팅 집약적인 워크로드에 대한 실무 경험을 바탕으로 다양한 사용 사례, 업계, 지역에 걸쳐 광범위한 고객과 협력할 수 있습니다.

Akamai는 기업이 AI 추론 요구사항을 원활하게 해결할 수 있도록 돕기 위해 일관성 있고, 신뢰할 수 있고, 안전한 방식으로 분산 클라우드 아키텍처를 성공적으로 배포해 기업이 최종 사용자와 가까운 곳에서 AI 워크로드를 처리할 수 있도록 했습니다. 이는 컴퓨팅 집약적인 AI 추론 워크로드에 필수적인 지연 시간을 줄이고 규모에 맞는 고성능을 보장하는 데 매우 중요합니다.

Akamai의 분산형 클라우드 아키텍처의 주요 장점 중 하나는 개발자가 개방형 클라우드 표준을 기반으로 하는 클라우드 네이티브 프레임워크에서 플랫폼에 구애받지 않는 개발에 집중할 수 있다는 점입니다.

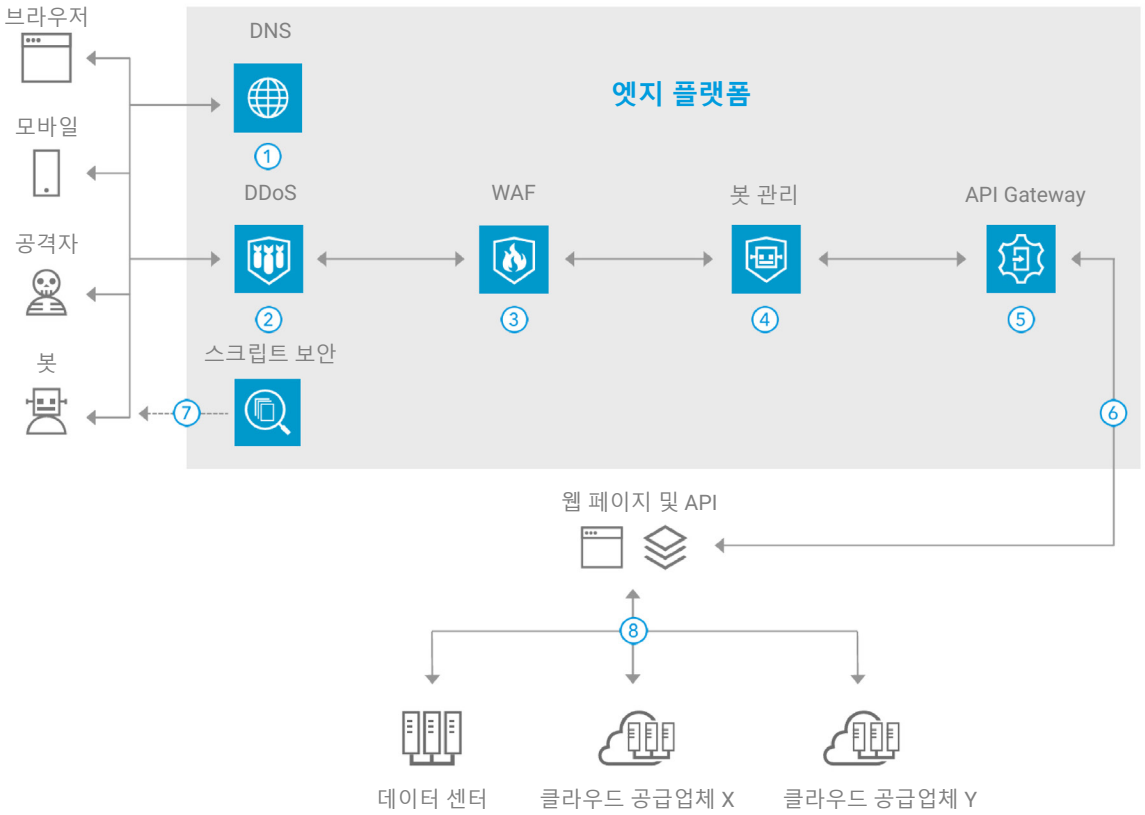
Akamai는 대규모 고성능을 달성하기 위해 오픈 소스 소프트웨어와 통합된 다양한 캐싱 기술을 사용해 충분한 처리량을 보장하고 지연 시간을 제한합니다. 일례로, 분산형 클라우드 아키텍처를 사용하는 많은 Akamai 고객들은 매우 중요하고 인기 있는 추론 활용 사례인 고객 대면 챗봇의 응답 시간이 50% 단축되었다고 보고했습니다. 이러한 종류의 거의 실시간 응답을 통해 리테일, 헬스케어, 금융 서비스, 첨단 기술 등 다양한 업계의 기업은 고유한 요구사항에 더 잘 맞춰진 더 빠른 고객 서비스를 제공할 수 있습니다.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

또한, Akamai는 기업이 빠르게 시작해 높은 경제적 가치와 운영 가치를 달성할 수 있도록 하는 AI 참조 아키텍처(아래 그림 3 참조)를 제공합니다. AI 참조 아키텍처는 익숙한 브라우저 세트와 사용자 정의 디바이스를 통해 기업이 핵심 데이터 센터에서 엣지와 클라우드에 이르기까지 엔터프라이즈 AI 솔루션을 배포할 수 있다는 점에서 포괄적입니다. 이는 개발자, IT 전문가, 인프라 관리자, 클라우드 설계자, 비즈니스 라인 이해관계자 등 기업 내 모든 당사자에게 이익이 됩니다.

그림 3.

기술의 급속한 발전과 AI 워크로드의 복잡성과 가치의 증가로 인해 기업은 경쟁에서 앞서기 위해 신속하고 단호하게 움직여야 합니다. 특히 Akamai와 같은 선도 기업이 계획하고, 개발하고, 배포하고, 관리하는 분산형 클라우드 아키텍처와 분산형 AI 추론을 수용함으로써 AI 추론을 최대한 활용해야 합니다.



package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

결론

다음으로 기업이 해야 할 일

AI가 IT 우수성과 비즈니스 혁신을 위한 게임의 룰을 바꾸었다고들 합니다. 그러나 기업의 AI 투자 가치를 최적화하는 열쇠는 AI 모델 학습에서 AI 추론으로 빠르게 전환하는 것입니다.

기업은 AI 추론이 제공하는 모든 장점을 최대한 활용하기 위해 업무가 이루어지고 핵심 인력이 시스템으로 놀라운 일을 할 수 있는 곳 가까이에서 추론 애플리케이션과 워크로드를 개발하고 배치해야 합니다. 즉, 중앙집중식 클라우드 컴퓨팅 모델에서 AI가 주도하는 분산형 클라우드 모델로 전환해야 합니다.

이 e-Book의 다른 부분에서 언급했듯이, 기업이 고려해야 할 많은 문제와 이해하고 극복해야 할 여러 가지 과제가 있습니다. 아직 AI 이니셔티브를 전략적 우선순위로 삼지 않은 기업에게는 부담스러울 수 있지만, 계속 앞서 나가기 위해 취할 수 있고 취해야 하는 중요한 단계가 있습니다.

Enterprise Strategy Group의 리오니는 "AI를 따라잡기 위해 노력하는 기업은 기술을 쫓는 데 뒤처질까 봐 두려워해서는 안 됩니다." 라며 “AI가 해결할 수 있는 구체적인 비즈니스 문제를 파악하고, 기존 데이터 인프라와 내부 기술을 평가하고, 부족한 부분을 채울 수 있는 핵심 파트너를 파악하는 등의 단계를 통해 올바른 기반을 구축하는 데 초점을 맞춰야 합니다. 기존 양질의 데이터를 사용해 명확한 가치를 입증할 수 있는 영향력이 큰 파일럿 프로젝트와 사용 사례를 하나 또는 두 개 파악해야 합니다.

측정할 대상과 측정 방법을 아는 것은 모든 성공적인 AI 추론 이니셔티브의 기본입니다.

그리고 이 모든 과정이 진행되는 동안, 내부 인재를 개발하고 책임감 있는 사용을 보장하기 위한 거버넌스 프레임워크를 구축해야 합니다. 핵심은 성급한 구축을 피하고, 기초를 다지고, 이상적으로는 지속적인 역량을 구축하는 전략적 단계를 따르는 것입니다.” 라고 강조했습니다.

이를 위해서는 네 가지 사항을 최대한 빨리 이해하고 실천해야 합니다.

1. **사용 사례 선택에 주의를 기울이세요.** 이것은 AI 추론을 적절하게 배포할 수 있는 능력에 대한 자신감을 구축해야 하기 때문에 필수적인 첫 번째 단계입니다. 또한 초기 성공을 통해 비즈니스 이해관계자와 재무 승인자 사이에서 자신감을 구축하는 것도 중요합니다.
2. **인프라 선택이 중요합니다.** AI 추론은 컴퓨팅 집약적이기 때문에 핵심 시스템에 시장을 선도하는 GPU를 사용하기로 손쉽게 결정할 수 있습니다. 그러나 GPU와 다른 AI 인프라에 대해 훌륭한 선택지가 많이 있습니다. 시간을 들여서 제대로 선택하세요(더욱 가능성을 높이려면 이전에 해본 경험이 있는 파트너의 경험과 전문성에 의존하세요).
3. **경영진의 지지를 얻으세요.** 최고 경영진과 이사회가 모두 관심을 갖는 것에 집중하세요. 기회는 무엇이고, 잠재적 리스크와 어떻게 균형을 이룰 수 있나요? 프로젝트와 사용 사례가 고객 경험 개선, 시장 기회 파악 및 활용 가속, 보다 정확한 베타 테스트 프로세스 구축과 같은 핵심 비즈니스 목표와 일치하는지 확인하세요. 이 작업을 하는 동안, IT, 사업 분야, 이사회 등의 그룹에서 높은 수준의 지지자를 파악하고 모집하세요.
4. **성공을 정의하고 전달하세요.** 측정할 대상과 측정 방법을 아는 것은 모든 성공적인 AI 추론 이니셔티브의 기본입니다. 초기 AI 추론 활용 사례에서 쉽게 얻을 수 있는 성과를 거둔 후에는, 현실적이고, 적절하고, “큰 성과”를 거두기에 충분한 견고성을 갖춘 지표를 선택하거나 파트너와 협력해 맞춤형 지표를 개발하세요.

이 콘텐츠는 Akamai의 의뢰로 TechTarget Inc.에서 제작했습니다.