IT WORLD CIO

Market Pulse

AI 열풍을 뒷받침하는 클라우드 "유연성이 핵심 가치"

2024년 국내 클라우드 컴퓨팅 및 AI 현황과 전망

2024년 현재 IT 업계의 핫 이슈는 단연 AI이다. 챗GPT로 촉발된 생성형 AI에 대한 관심이 AI 전반으로 확산됐고, 올해에도 계속될 것으로 보인다. 클라우드 분야 역시 예외는 아니어서 주요 클라우드 서비스 업체 모두 클라우드 기반 AI 서비스를 전면에 내세우고 있다. 이와 함께 불확실한 경제 상황과 함께 클라우드 비용은 계속 과제로 남아 있다. 클라우드 자원에 대한 정확한 관리의 중요성이 강조되고, 한편으로는 낮아진 인프라 솔루션의 가격 때문에 자체 인프라의 가치가 재조명되고 있다. ITWorld/CIO는 이런 변화를 확인하기 위해 국내 IT 전문가를 대상으로 클라우드 및 AI 관련 현황과 전망을 묻는 설문 조사를 실시했다. 클라우드 도입 및 활용, 과제 등 기본 현황과 클라우드와 온프레미스의 균형, 비용 최적화, 그리고 국내 클라우드 기반 AI 도입 및 활용 관련 현황과 과제까지 확인했다.

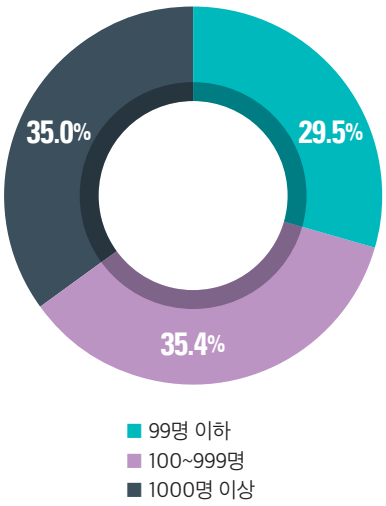
DATADOG IBM GCORE INZENT bizplay MongoDB.

AI 열풍을 뒷받침하는 클라우드
“유연성이 핵심 가치”
2024년 국내 클라우드 컴퓨팅 및 AI 현황과 전망

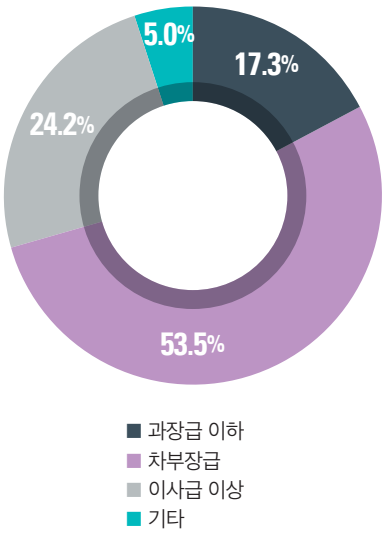
2024년 현재 IT 업계의 핫 이슈는 단연 AI이다. 지난해 초 챗GPT로 촉발된 생성형 AI에 대한 관심이 AI 전반으로 확산됐고, 열풍은 2024년에도 계속될 것으로 보인다. 클라우드 분야 역시 이런 AI 바람을 타고 있다. 주요 클라우드 서비스 업체 모두 클라우드 기반 AI 서비스를 전면에 배치하며 둔화되고 있는 클라우드 시장의 성장률을 되살리는 데 진력하고 있다.

그렇다고 클라우드가 AI 열풍을 이용만 하고 있는 것은 아니다. 클라우드 기반의 AI 서비스가 아니라면, AI 구현을 추진할 만한 자금과 인력을 갖춘 기업은 제한적이었을 것이다. 지금과 같이 기업 규모와 산업군을 가리지 않고 AI가 확산되는 데는 접근성 높은 클라우드 서비스가 한몫하고 있다.

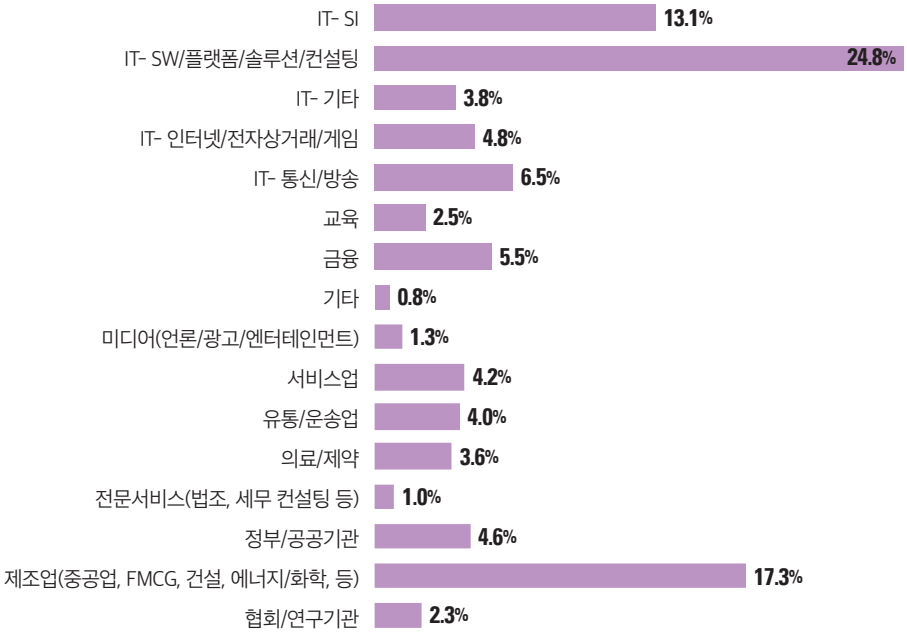
응답자 소속 기업 규모



응답자 직급



응답자 산업군



AI 외에 클라우드가 직면한 과제는 타당성 증명이다. 불확실한 경제 상황으로 기업은 비용 지출에 매우 민감한 상태이다. 이 때문에 예상을 넘는 클라우드 비용은 클라우드 자체에 대한 의문으로, 그리고 온프레미스 인프라에 대한 재검토로 이어지고 있다.

ITWorld/CIO는 이런 맥락에서 국내 클라우드 컴퓨팅 및 AI 관련 현황과 과제를 파악하기 위해 국내 기업 IT 전문가를 대상으로 설문 조사를 실시했다. Tech Survey 플랫폼을 통해 진행한 설문 조사에는 2024년 2월 1일부터 2월 27일까지 총 487명의 유효 응답자가 참여했다.

응답자의 소속 업종은 소프트웨어/플랫폼 등의 IT 솔루션 업종이 24.8%로 가장 많았고, 제조업이 17.3%, IT SI 업종이 13.1%를 차지했다. 이외에 통신/방송, 금융, 공공기간, 유통 등 다양한 업종이 5% 내외로 참여했다.

기업 규모별로는 99명 이하 중소기업과 100~999명의 중견기업, 1,000명 이상의 대기업이 모두 30~35% 정도로 고르게 참여했다. 직급별로는 차부장급이 53.5%로 절반을 넘었으며, 이사급 이상이 29.2%, 과장급 이하가 17.3%였다.

높아지는 클라우드의 중요성…AI/ML 활용 10%p 증가

국내 기업의 클라우드 도입 현황은 1년 만에 큰 폭의 변화를 보일 수 있는 단계는 지난 것으로 보인다. 여전히 미션 크리티컬 업무를 제외한 업무 중 일부만을 클라우드로 구동하고 있다는 응답이 36.6%로 가장 많았고, 어떤 식으로든 클라우드를 활용하고 있다는 기업은 전체의 75.4%로 전년도 조사와 큰 차이가 없었다.

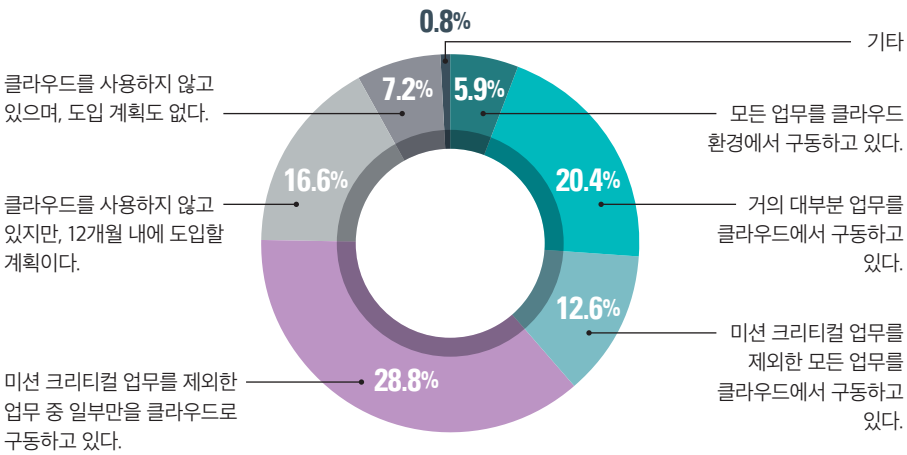
하지만 변화 추이는 긍정적이다. 핵심 업무를 제외하고 일부에서만 클라우드를 활용화고 있다는 응답은 점점 줄어들어 40% 밑으로 떨어졌고, 거의 대부분 업무를 클라우드에서 구동하고 있다는 응답은 조금씩 증가해 20%를 넘었다.

2025년 예상치 역시 같은 추이를 보였다. 현재 수준을 유지할 계획이라는 응답 27.7%를 응답자의 도입 현황에 대입해 보면, 미션 크리티컬 업무를 제외한 업무 중 일부만을 클라우드로 구동한다는 응답은 36.3%로 소폭이지만 줄어든 반면, 다른 응답은 모두 5%p 이상 증가할 것이라고 답했다.

물론 지난 조사 결과에서 알 수 있듯이, 전체적인 응답률은 공격적인 클라우드 확대를 예상하지만, 실제 다음 해의 도입 현황은 예상과는 상당한 격차를 보인다. 하지만 기업 환경에서 클라우드의 양적 질적 비중이 꾸준히 증가하고 있음은 분명하다.

클라우드 활용 목적은 웹/웹 앱/모바일이 작년과 같은 46.0%의 응답률로 1위를 기록했다. 개발 및 테스트, 빅데이터/분석/BI가 30% 중후반 대의 응답률로 뒤를 이었다. 전체적으로 변화폭이 그리 크지 않은 상황이다. 현재 가장 뜨거운 주제인 AI/ML은 지난 해보다 10%p 가까이 증가한 29.2%로, 비즈니스 앱과 같은 수준으

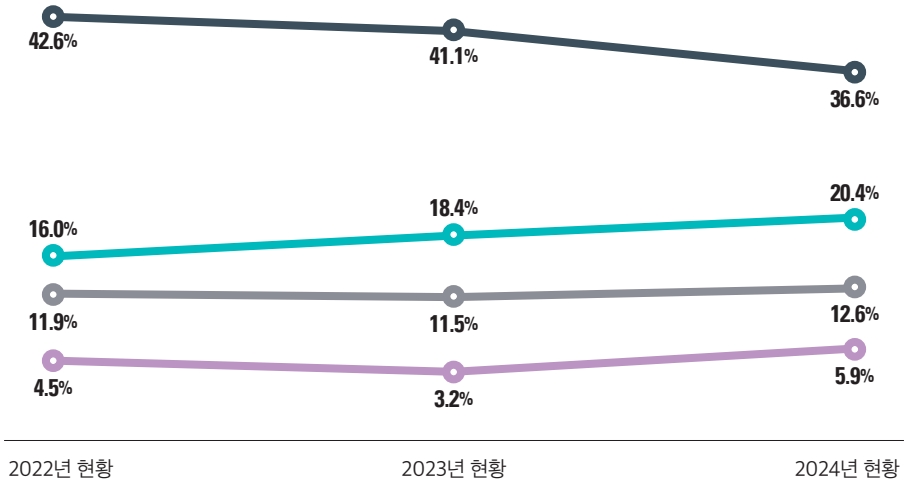
국내 기업의 클라우드 활용 현황



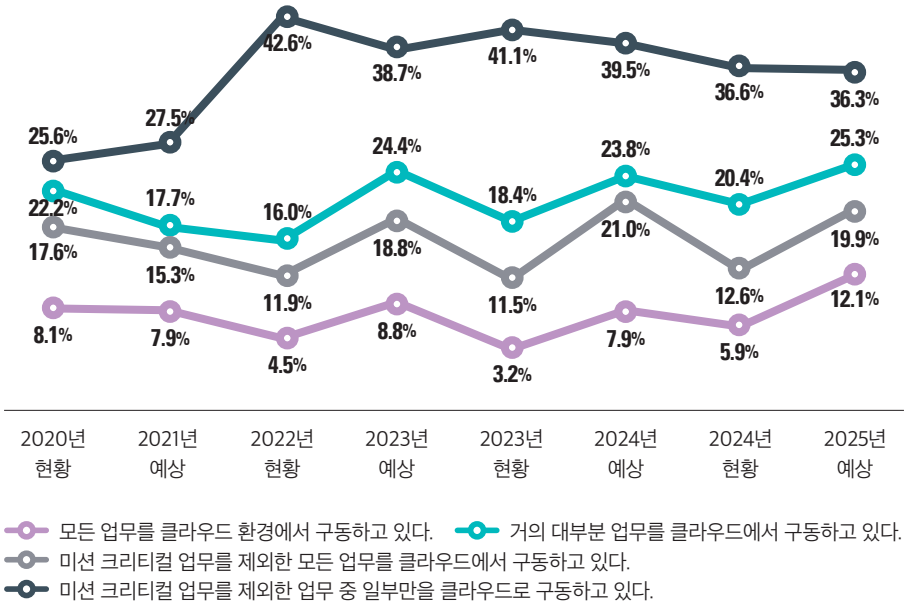
로 올라왔다.

다만, AI/ML은 시류에 따라 변동이 좀 큰 편으로, 그동안 알파고의 등장, 자동화의 확대 등과 같은 시장 이슈에 따라 증가와 감소를 보였다. 이번 조사에서 나타난 증가세는 생성형 AI에 대한 관심이 반영된 것으로 볼 수 있는데, 과연 내년도 조사에

클라우드 활용 현황 추이



클라우드 활용 현황과 전망



서도 성장세를 이어갈지 지켜볼 필요가 있다.

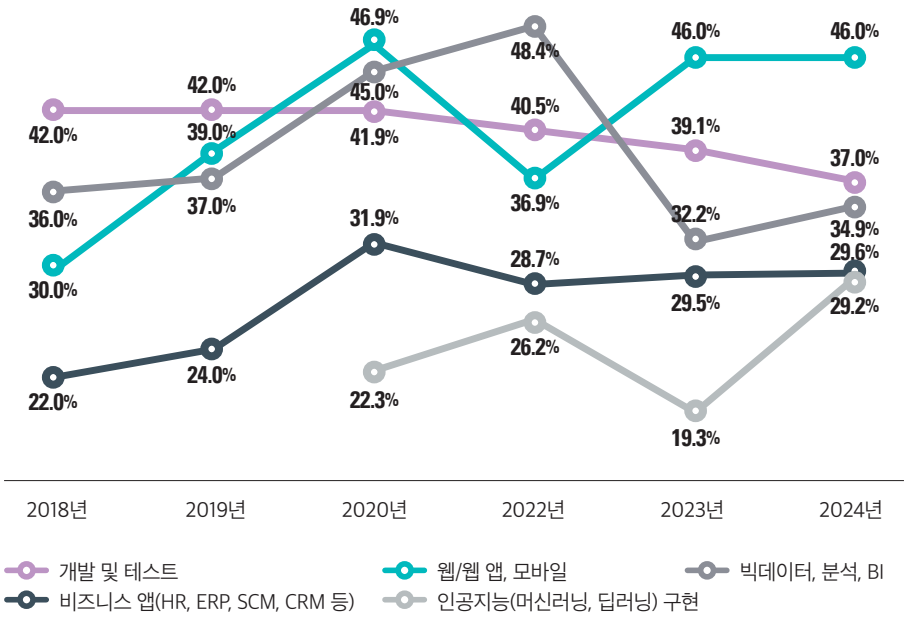
유연하고 탄력적인 IT 자원 기대…비용과 인력은 여전히 과제

기업이 클라우드 도입으로 얻고자 하는 효과는 역시 유연하고 탄력적인 IT 자원 활용이 67.6%로 가장 많았다. 이어서 개발 및 운영 편의성, 비용 절감, 백업 및 안정성 확보가 30~40%대의 응답률을 기록했다. 하지만 클라우드가 보편적인 IT 인프라의 하나로 확산하면서 기업이 기대하는 효과는 산업군과 기업 규모, 그리고 클라우드 도입 현황에 따라 차이를 보였다.

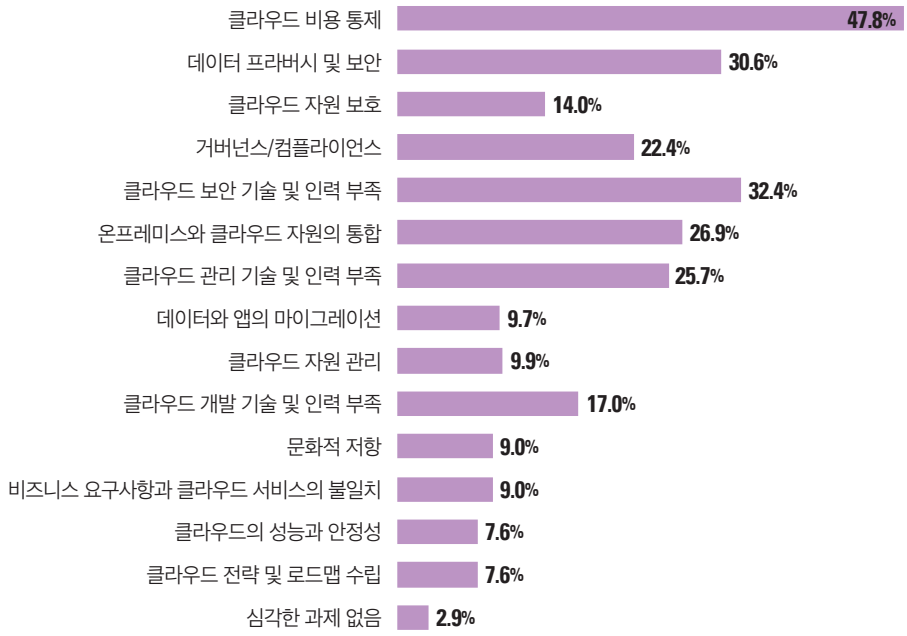
제조업의 경우, 유연성에 대한 기대는 52.4%로 평균보다 낮고, 비용 절감에 대한 기대가 38.1%로 가장 높았다. 반면 IT 통신/방송 업종은 유연성과 개발 및 운영 편의성에 대한 기대가 각각 75.0%, 53.1%로 평균보다 높은 반면, 비용 절감에 대한 기대는 15.6%로 가장 낮았다. 규모별로는 기업 규모가 클수록 유연성에 대한 기대가 높았다.

모든 업무를 클라우드 환경에서 구동하고 있다고 응답한 기업은 유연성에 대한 기대치가 90.3%, 개발 및 운영 편의성에 대한 기대가 48.4%로 상대적으로 높았고, 비용 절감에 대한 기대는 32.3%로 평균보다 낮았다.

주요 클라우드 사용 용도의 변화



클라우드 도입 및 활용 과정의 어려움



기업이 클라우드 도입 및 활용 과정에서 느끼는 가장 큰 어려움은 클라우드 비용 통제가 47.8%로 가장 높은 응답률을 기록했다. 지난 해 조사의 42.0%보다 5%p나 증가하며 클라우드 보안 기술 및 인력 부족을 제치고 1위를 차지했다. 모든 업무를 클라우드 환경에서 구동하고 있는 기업의 경우, 클라우드 비용 통제를 어려움으로 꼽은 응답자가 무려 71.0%인 데서 알 수 있듯이, 클라우드의 비중이 증가하면 할수록 비용을 최적화하고 통제하는 과제는 점점 더 중요해질 것으로 보인다.

그렇다고 인력 부족 문제가 해소된 것으로 보이지는 않는다. 지난 해보다 줄어들긴 했지만, 클라우드 보안, 관리, 개발로 나눠진 인력 부족 응답을 모두 합하면, 무려 75.1%의 응답자가 인력 부족 문제를 겪고 있는 셈이기 때문이다.

데이터 프라이버시 및 보안과 거버넌스/컴플라이언스도 각각 30.6%와 22.4%로 적지 않은 응답률을 기록했는데, 데이터 프라이버시 및 보안은 지난 해보다 5%p나 증가한 수치이다. 하이브리드 클라우드 추세를 반영하는 온프레미스와 클라우드 자원의 통합도 26.9%로 비교적 높은 수치를 기록했는데, 모든 업무를 클라우드 환경에서 구동하고 있는 기업을 제외하고는 대부분 30% 내외의 응답률을 보였다.

여전히 비중이 큰 온프레미스, 무게 중심은 클라우드로 이동 중

클라우드 비용 문제가 대두되면서 전통적인 인프라, 즉 온프레미스 환경에 대한 관

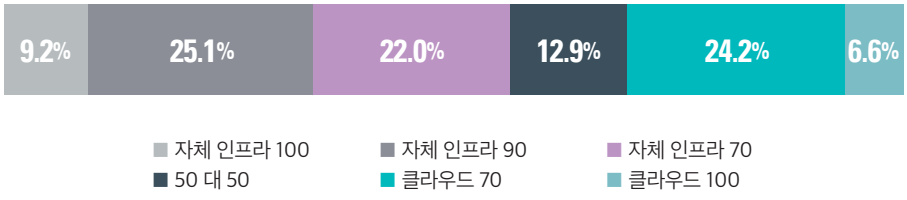
심도 다시 높아지고 있다. 많은 시행 착오를 통해 클라우드에 적합하지 않은 워크로드가 지적되기도 하고, 기존 인프라 솔루션의 발전과 솔루션 업체의 노력이 합쳐져 비용이나 효율성 등에서 경쟁력 있는 해법으로 재등장했기 때문이다.

이번 조사에서는 국내 기업의 인프라에서 온프레미스 환경과 클라우드가 차지하는 비중을 확인하고, 향후 이 비중이 어떻게 변화할 것인지 확인했다. 조사 결과, 대다수 기업은 자체 인프라와 퍼블릭 클라우드를 함께 사용하는 것으로 나타났다. 자체 인프라가 하나도 없거나 퍼블릭 클라우드를 전혀 사용하지 않는다는 응답은 각각 6.6%와 9.2%로 낮았으며, 나머지 84.2%는 비율의 차이는 있어도 자체 인프라와 퍼블릭 클라우드를 함께 사용한다고 답했다.

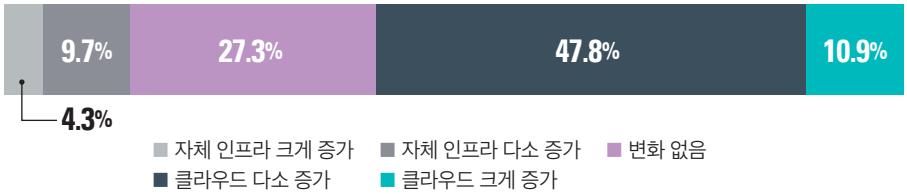
두 환경의 비율과 응답자 수를 기반으로 추산해 보면, 국내 기업의 인프라에서 온프레미스 환경과 클라우드의 비중은 3:2 정도이다. 아직까지 온프레미스의 비중이 20% 정도 많은 셈이다. 하지만 불과 몇 년 뒤에는 이 비율이 뒤바뀔 수 있을 것으로 보인다. 자체 인프라와 퍼블릭 클라우드의 비율이 향후 1~2년 내에 어떻게 변화할 것으로 예상하는지 묻는 설문에서는 자체 인프라가 증가할 것이라는 응답은 12.0%에 불과한 반면, 퍼블릭 클라우드의 비율이 증가할 것이라는 응답은 58.7%로 절반을 넘었기 때문이다.

인프라의 비율이 바뀔 것으로 생각하는 주된 이유는 클라우드 도입으로 얻고자 하는 기대 효과와 마찬가지로 유연하고 탄력적인 IT 활용이 41.1%로 압도적으로 많

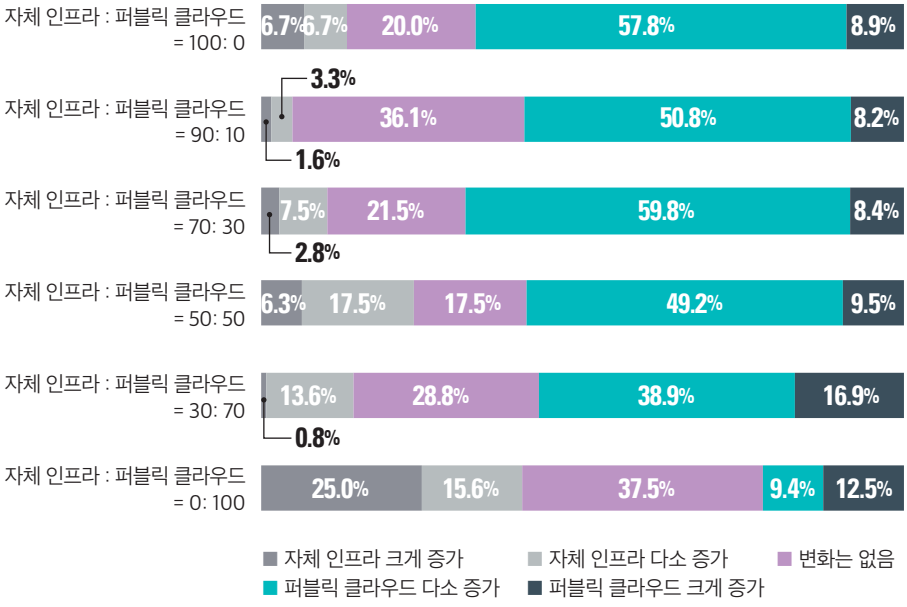
자체 인프라와 퍼블릭 클라우드의 비율



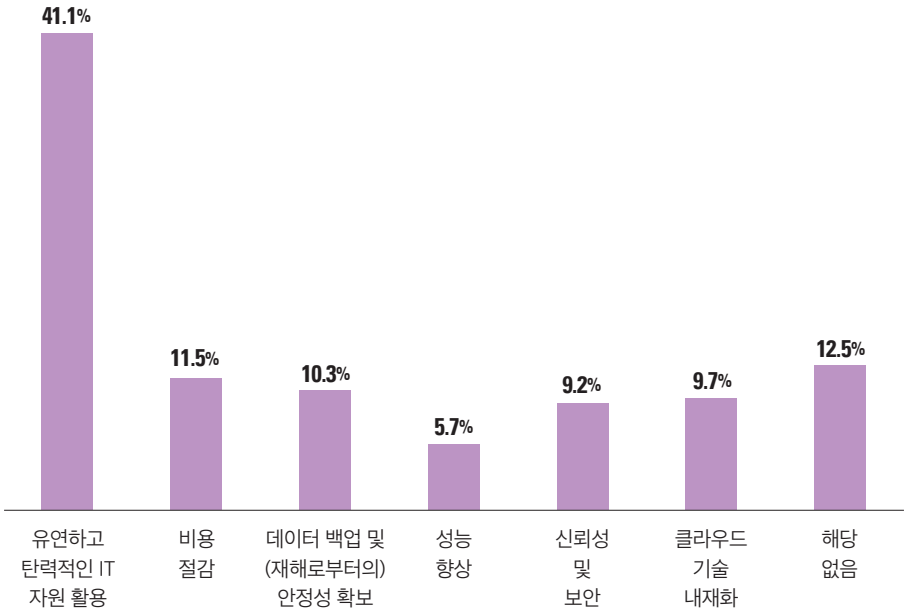
향후 1~2년 내 인프라 비율의 변화



자체 인프라와 클라우드의 비율에 따른 변화



인프라의 비율이 변화하는 이유



았다. 흥미로운 것은 자체 인프라의 비율이 증가할 것이라고 답한 기업 역시 인프라의 비율이 바뀌는 가장 큰 이유로 유연성을 꼽았다는 점이다. 실제로 자체 인프라의 비중이 높은 기업은 퍼블릭 클라우드가 증가할 것으로, 클라우드 비중이 높은 기업은 자체 인프라가 증가할 것이라는 답변이 상대적으로 많았다. 퍼블릭 클라우드가 중심이 되겠지만, 자체 인프라 역시 중요한 역할을 수행하는 하이브리드 클라

우드가 지향점이라는 것을 다시 한번 확인할 수 있다.

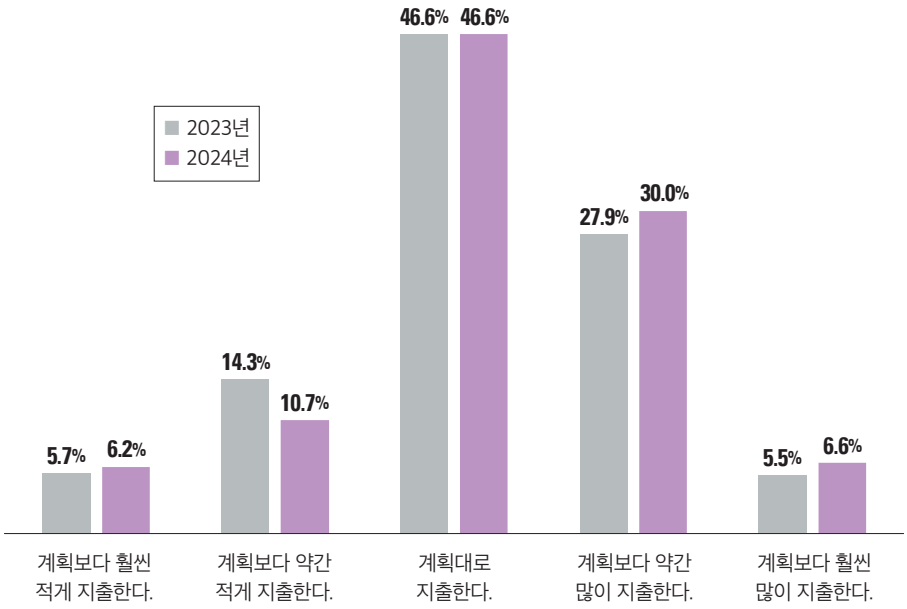
"많이 쓸수록 비용 통제도 어렵다" 관련 정책과 툴 사용 확대 필요

응답자의 절반이 비용 통제를 클라우드 도입 및 활용 과정의 어려움으로 꼽은 만큼, 클라우드 비용을 계획대로 지출한다는 응답은 절반을 넘지 못했다. 전년도 조사 결과와 동일한 46.6%였다. 계획보다 적게 지출한다는 응답은 15% 정도에 불과한 반면, 계획보다 많이 지출한다는 응답은 35%를 넘어 클라우드 비용 통제가 적지 않은 과제임을 확인할 수 있었다.

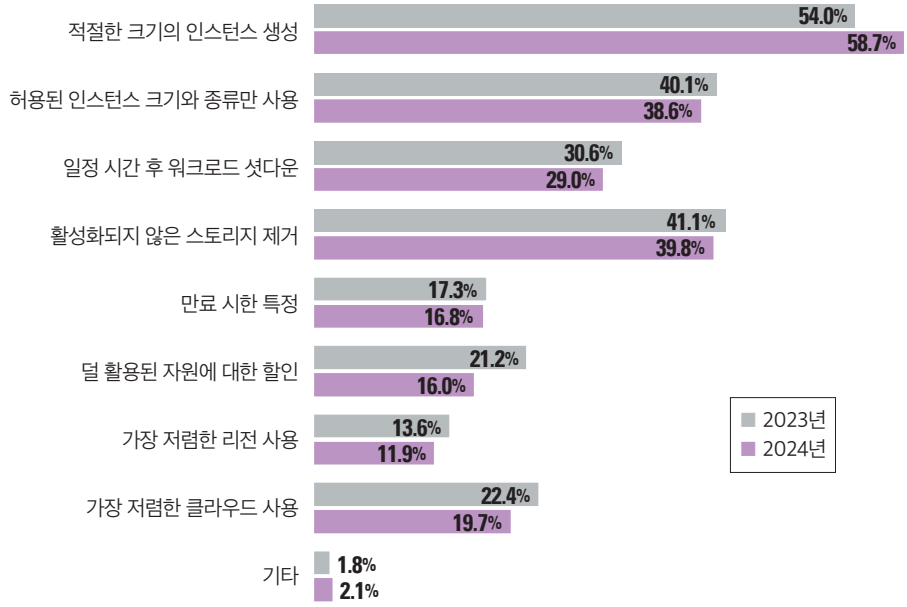
클라우드 비용이 계획대로 지출되는지 여부는 산업군이나 기업 규모와는 관계가 없지만, 클라우드를 얼마나 사용하고 있는지와는 상당한 연관 관계가 있는 것으로 나타났다. 클라우드 활용 비중이 클수록 비용을 적게 지출하거나 많이 지출하는 응답이 많았는데, 계획대로 지출한다는 응답은 모든 업무를 클라우드 환경에서 구동하는 기업이 29.0%, 거의 대부분 업무를 클라우드에서 구동하는 기업이 40.2%로 평균값과 적지 않은 차이를 보였다.

비용 통제가 당면 과제임에도 불구하고, 많은 기업이 클라우드 비용 최적화를 위한 정책을 적용하는 데는 아직 미온적인 것으로 보인다. 가장 기본적인 정책인 적절한 크기의 인스턴스 생성은 58.7%로 절반 이상의 기업이 채용한 것으로 나타났지만, 다소 강제적인 기준이 적용되는 정책의 채택율은 30~40%의 응답률을 보였다.

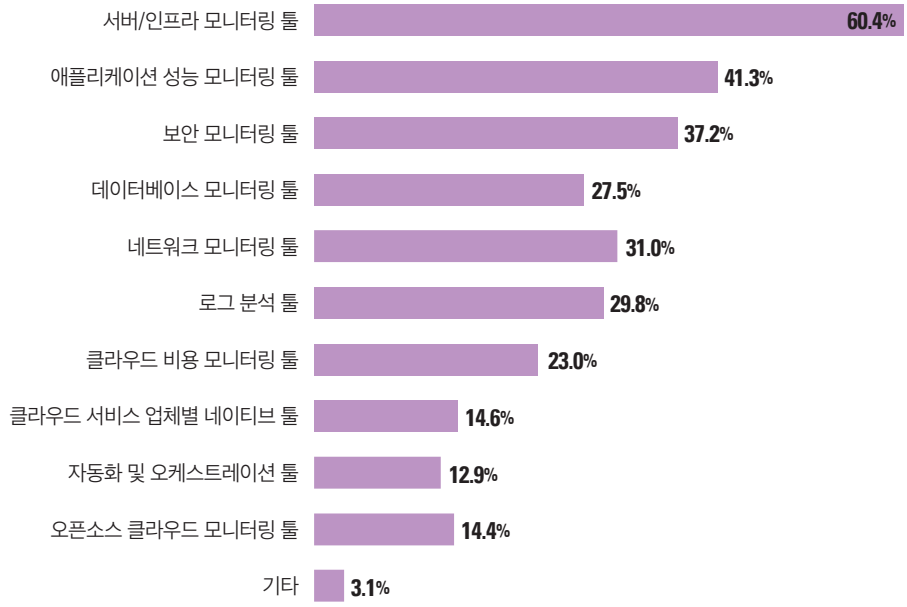
클라우드 비용 지출 계획과 실행



활용 중인 클라우드 비용 최적화 정책



활용 중인 클라우드 모니터링 툴 현황



특히 할인이나 저렴한 리전 선택 등 클라우드 서비스 업체와의 관계를 기반으로 한 정책을 적용한다는 응답은 15% 수준에 그쳤다.

클라우드 모니터링 툴 활용 현황도 같은 맥락에서 볼 수 있다. 클라우드와 온프레미스가 혼재하는 인프라 환경에서 국내 기업이 사용하는 모니터링 툴의 수는 평균

3개 정도인 것으로 나타났다. 서버/인프라 모니터링 툴처럼 전통적인 인프라 환경에서 사용하던 툴의 활용도는 높았지만, 클라우드 비용 모니터링 툴이나 클라우드 서비스 업체별 네이티브 툴, 오케이스트레이션 툴처럼 조금 더 클라우드 환경 지향적인 툴의 활용도는 30%를 넘지 못했다.

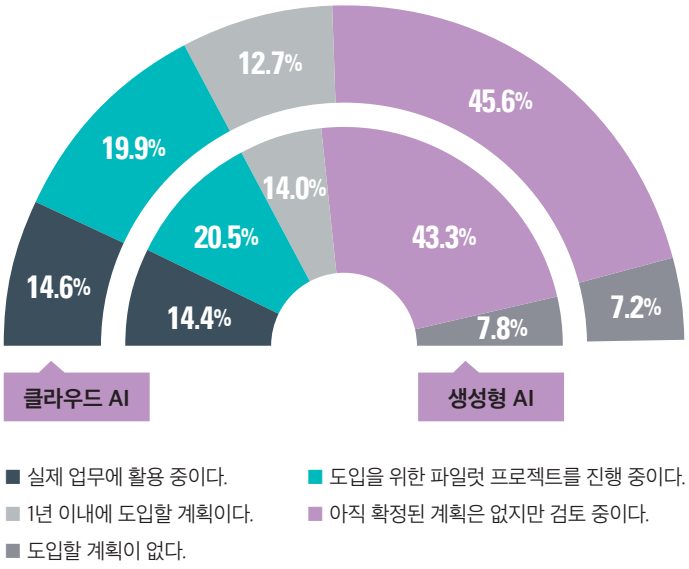
클라우드 기반 AI, 본격적인 활용 위한 준비 단계

국내 기업의 클라우드 기반 AI 도입 현황은 '본격적인 활용을 위한 준비 단계'로 평가할 수 있다. 실제 업무에 활용 중이라는 응답은 14.6%에 그쳤지만, 파일럿 프로젝트를 진행하고 있다는 응답 19.9%와 1년 이내에 도입할 계획이라는 응답 12.7%를 합치면 47.2%로, 절반에 가까운 기업이 AI를 활용하고 있거나 본격적인 활용을 위한 준비 단계를 진행하고 있는 셈이다. 45.6%는 확정된 계획은 없지만 검토 중이라고 답했으며, 도입할 계획이 없다는 응답은 7.2%였다.

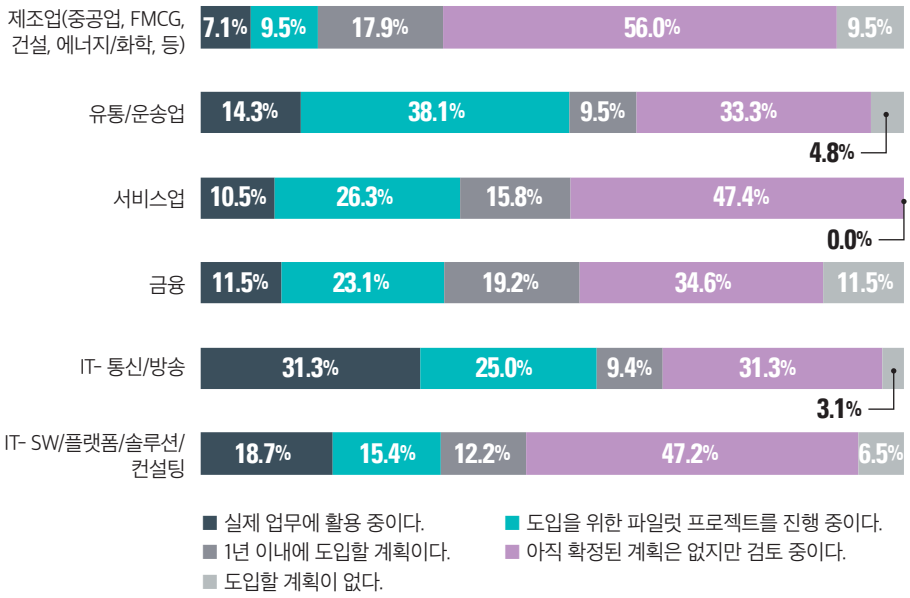
산업군별로는 상당한 차이를 보였다. 실제 업무에 활용 중이거나 파일럿 프로젝트를 진행 중이라는 응답을 기준으로 볼 때, IT 관련 산업군은 대체로 평균을 웃도는 응답을 보였고, 특히 IT-통신/방송은 실제 업무에 활용 중이라는 응답이 31.3%에 달했다. IT 관련 산업군 외에는 금융, 서비스, 유통/운송 산업군의 응답률이 높았다. 하지만 제조 산업군은 각각 7.1%, 9.5%로 낮았다.

기업 규모별로는 1,000명 이상의 대기업은 절반 가까운 기업이 클라우드 기반 AI

클라우드 기반 AI 및 생성형 AI 활용 현황



주요 산업별 클라우드 기반 AI 활용 현황

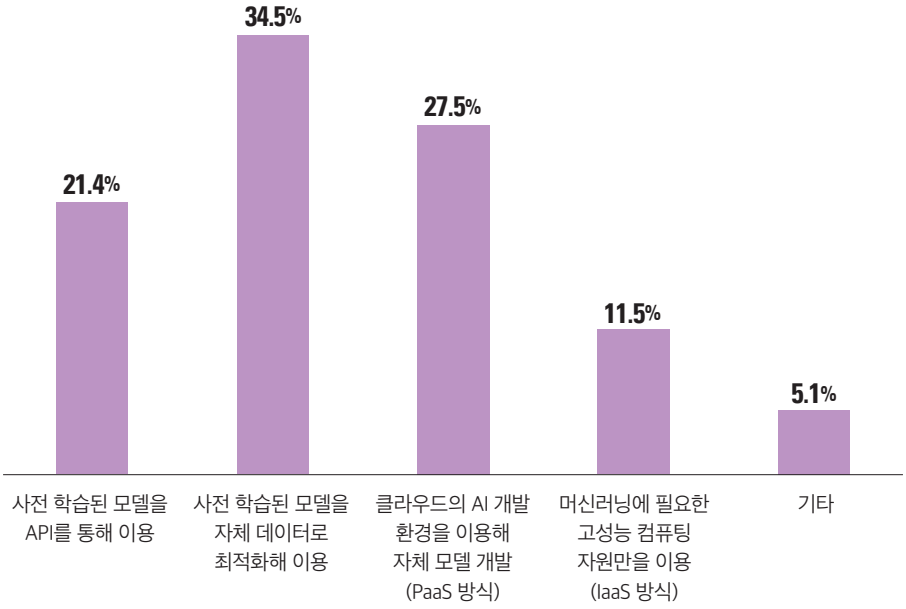


를 실제 업무에 활용 중이거나 파일럿 프로젝트를 진행 중이라고 답한 반면, 중견 중소기업은 이 비율이 30%에 못미쳤다.

챗GPT와 같은 생성형 AI의 도입 및 활용 현황 역시 AI 서비스와 거의 비슷한 것으로 나타났다. 클라우드 기반 AI 서비스가 클라우드 기반 생성형 AI보다 훨씬 넓은 범주임에도 비슷한 결과를 보인 것은 생성형 AI에 대한 높은 관심을 반영하는 것으로 볼 수 있다. 한편으로는 많은 기업이 AI와 생성형 AI를 같은 것으로 간주하거나 AI 프로젝트의 최종 목표를 생성형 AI로 정한 것으로 볼 수도 있다. 산업군별로 보면, 제조 산업군도 생성형 AI를 실제 업무에 활용 중인 비율은 8.3%에 불과하지만 파일럿 프로젝트를 진행 중이라는 응답은 16.7%로 훨씬 더 적극적으로 나서고 있는 것으로 나타났다. 금융 산업군의 경우, 파일럿 프로젝트를 진행 중이라는 응답이 42.3%에 달했다.

클라우드 기반 AI 서비스 이용 방식은 클라우드 서비스 업체가 제공하는 사전 학습된 모델을 자체 데이터로 최적화해 이용하는 방식이 34.5%로 가장 많았다. 하지만 클라우드의 AI 개발 환경을 이용해 자체 모델을 개발하는 방식이나 사전 학습된 모델을 API를 통해 이용하는 방식도 각각 27.5%, 21.4%로 적지 않은 비율을 기록했다. 한 가지 방식이 압도적인 응답을 보이지 않는 것은 기업이 자사의 AI 도입 목적과 환경에 맞춰 다양한 접근 방식을 고려하는 것으로 추정할 수 있다.

클라우드 기반 AI 이용 방식



서비스 이용 방식은 산업군 별로는 눈에 띄는 편차가 없었지만, 기업 규모별로는 차이를 보였다. 1,000명 이상 대기업은 사전 학습된 모델을 API를 통해 이용한다는 응답이 17.0%로 낮은 반면, 자체 모델을 개발한다는 응답은 30.1%로 평균보다 다소 높았다. 중견 중소기업의 경우는 반대의 경향을 보였다. AI 활용 현황과도 관련성이 있는 것으로 나타났다. 현재 AI를 실제 업무에 활용 중인 기업은 사전 학습된 모델을 API를 통해 이용한다는 응답이 33.8%로 평균보다 훨씬 높은 반면, 1년 이내에 도입할 계획인 기업은 자체 모델을 개발한다는 응답이 43.5%로 높았다. 발빠르게 AI 활용에 나선 기업은 구현 속도가 빠른 API 방식을 선택한 것으로 추정할 수 있다.

전문 인력 확보과 데이터 관리가 AI 활용의 최대 과제

클라우드 기반 AI 서비스는 AI 활용의 기반이 되는 요소를 갖추고 기업이 최대한 편리하게 이용할 수 있는 서비스를 제공하지만, 그렇다고 모든 것이 저절로 구현되지는 않는다. 기업은 어떤 업무에 어떤 AI를 적용해 어떤 효과를 얻을지 계획을 세워야 하며, 데이터를 처리하고 비즈니스 프로세스와 통합하는 데도 적지 않은 노력을 기울여야 한다. 최근에는 보안이나 프라이버시, 컴플라이언스 관련 문제도 적지 않게 거론되고 있다.

클라우드 기반 AI 서비스를 활용하는 데 가장 큰 장애물이 무엇인지 묻는 질문에서 절반이 넘는 53.0%응답자가 전문 인력 확보의 어려움을 꼽았다. 데이터 부족과 데

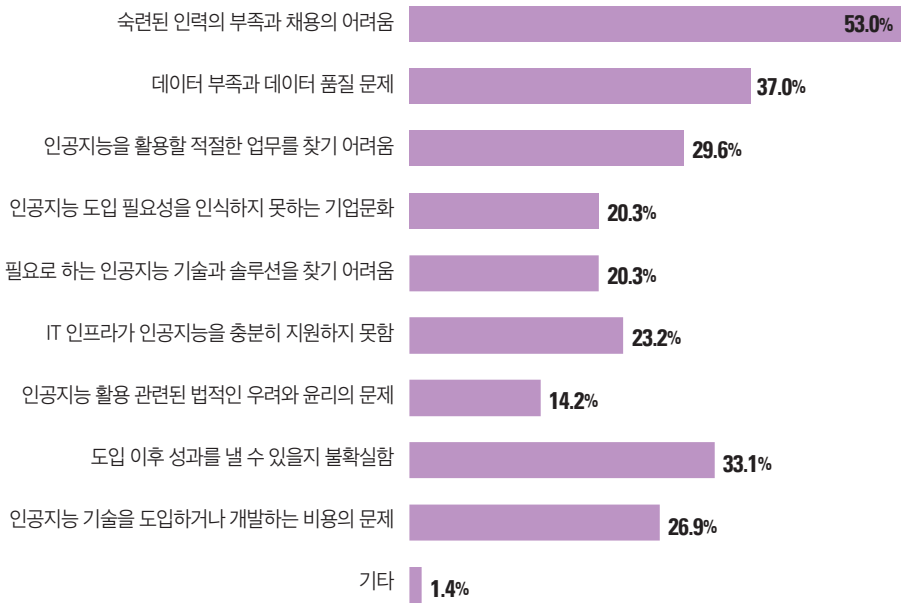
이터 품질 문제가 37.0%로 두 번째로 많았다. 이외에도 AI 도입 및 활용과 관련해 예상할 수 있는 어려움, 즉 적절한 적용 업무, 기업 내의 AI에 대한 인식, 도입 및 개발 과정에 드는 비용 등이 20~30%의 고른 응답률을 보였다.

흥미로운 것은 도입 이후 성과를 낼 수 있을지 불확실함을 어려움으로 꼽은 응답자가 33.1%로 3위를 차지한 것이다. AI가 대세로 떠오르면서 많은 기업이 경쟁에 뒤처지지 않기 위해서라도 AI 도입을 추진하고 있지만, AI가 가져올 성과에 대해 확신하지 못하는 기업이 적지 않다는 것을 보여준다. 한편으로는 실제 업무에 활용 중인 기업은 이런 불안감이 23.9%로 평균보다 낮은 반면, AI 도입을 검토 중인 기업은 40.0%가 실제 성과에 대한 불안감을 어려움으로 토로했다.

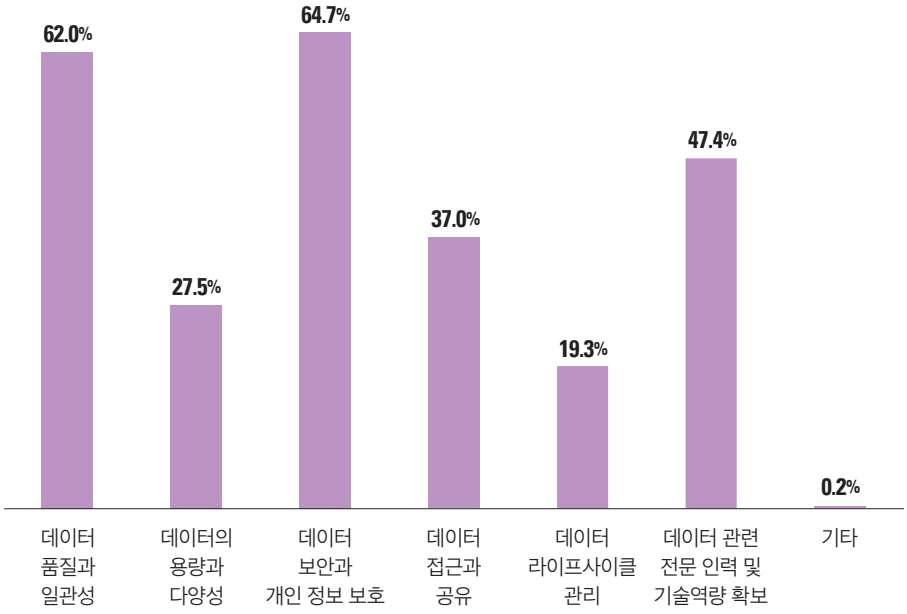
두 번째로 많은 응답자가 어려움으로 꼽은 데이터 관리는 사실 AI 도입이 아니라도 현재의 모든 기업이 고민하는 문제일 것이다. 특히 AI의 성능이 데이터 품질에 좌우된다는 것이 전문가들의 일관된 지적이다 보니, AI 활용을 고려하는 기업으로는 데이터 관리에 신경이 쓰이지 않을 수 없다.

AI 서비스 도입 및 활용과 관련한 데이터 과제 3가지를 묻는 질문에 가장 많은 응답을 기록한 것은 64.7%의 데이터 보안과 개인정보보호로, 62.0%의 데이터 품질과 일관성보다 높았다. 챗GPT 같은 공개 생성형 AI 서비스가 데이터 유출이나 개인정보 노출 등의 문제를 여러 차례 겪은 것이 크게 작용한 것으로 보인다. 데이터

클라우드 기반 AI 도입 및 활용의 어려움



클라우드 기반 AI 도입 및 활용 관련 데이터 과제



관련 전문 인력 및 기술 역량 확보도 47.4%의 응답률을 기록했는데, 응답자의 절반 가까이가 인력 및 기술력 확보를 가장 중요한 과제로 꼽은 셈이다.

이외에 데이터 용량과 다양성, 접근과 공유도 1/3 내외의 응답률을 기록했으며, 가장 적은 응답자가 과제로 지목한 데이터 라이프사이클 관리도 19.3%의 응답률을 기록했다. 데이터 품질과 보안의 중요성이 높지만, 이외에도 데이터와 관련된 모든 요소가 AI를 활용하려는 기업에는 적지 않은 과제가 되고 있음을 알 수 있다.

온프레미스와 공존 속에 가치 증명 위한 최적화 필요

클라우드 컴퓨팅 초기에 클라우드 지지자들은 퍼블릭 클라우드가 기업의 기존 온프레미스 환경을 완전히 대체할 것이라는 야심찬 전망을 내놓기를 주저하지 않았다. 하지만 2024년 현재 국내 기업의 인프라는 여전히 온프레미스의 비중이 클라우드보다 높다.

클라우드 컴퓨팅의 장점은 분명하지만, 그렇다고 기업의 IT 인프라를 하루 아침에 갈아치우지는 못할 뿐만 아니라 앞으로도 완전히 대체하지 못한다는 것이 클라우드 전문가들의 분석이다.

하지만 클라우드가 기업 IT 인프라에서 차지하는 비중은 서서히, 그리고 확실하게

커지고 있다. 이번 조사에서도 많은 응답자가 앞으로 클라우드의 비중이 크게 증가할 것이라고 답했다. 물론, 이런 예상과 현실은 격차가 있기 마련이지만, 피할 수 없는 추세라는 것은 부인할 수 없다.

다만 클라우드의 비중이 커지면서 클라우드 관련 비용이나 관리의 문제 역시 이전과 중요성이 달라졌다. 대표적인 것이 클라우드 비용 통제이다. 클라우드의 비중이 큰 기업일수록 클라우드 비용 통제를 주요 해결과제로 안고 있는 실정이다. 클라우드 비용 최적화에 대한 관심이 필요한 시점이다.

AI 열풍이 전 세계를 휩쓸고 있지만, 국내 기업의 클라우드 기반 AI 및 생성형 AI 도입 및 활용 현황은 아직 "준비 단계"라고 평가할 수 있다. AI를 발 빠르게 도입해 활용하고 있는 기업보다는 파일럿 프로젝트를 진행하고 도입 계획을 세우고 있는 기업이 더 많기 때문이다. 초기 단계인 만큼 전문 인력과 기술력 확보도 쉽지 않고 AI의 연료가 되는 데이터에 대한 우려도 크다. 국내 기업이 산재한 해결 과제를 어떻게 극복하고 AI 관련 계획을 수행해 나가는지, 그리고 클라우드는 이 과정에서 어떤 역할을 하는지 지켜봐야 할 것이다.

Session Brief

“다양한 지표의 통합과 연계 분석이 핵심”

데이터독의 AI 서비스 모니터링 및 AI옵스 활용 전략

AI가 일상에 스며들고 있다. 이미 많은 직장인이 일상 업무에 챗GPT와 같은 생성형 AI 서비스를 활용하고 있으며, 전사적인 AI/ML 도입을 준비하는 기업도 적지 않다. ITWorld의 설문 조사에 따르면, 클라우드 기반 AI 서비스를 이용하고 있거나 검토 중이라고 답한 기업은 무려 93%에 달한다. AI 관련 특허 출원 수와 AI 투자도 매년 꾸준히 증가하는 추세다.

하지만 AI/ML을 도입하는 기업은 다양한 장애물에 맞닥뜨린다. 가장 대표적인 문제가 관리 및 운영의 어려움이다. 비즈니스 요구사항에 맞춰 AI 기술 스택을 기존 서비스와 융합해 제공하는 것은 매우 까다로우며, 그렇다 보니 AI 모델의 운영과 비용을 예측하고 성능 문제를 파악하는 것도 복잡하다.

▲3월 21일 서울 그랜드 인터컨티넨탈 서울 파르나스 국화룸에서 개최된 'Cloud & AI Research Summit 2024'에서 데이터독 조용원 세일즈 엔지니어(SE)가 발표하고 있다.

DATADOG

IT WORLD CIO

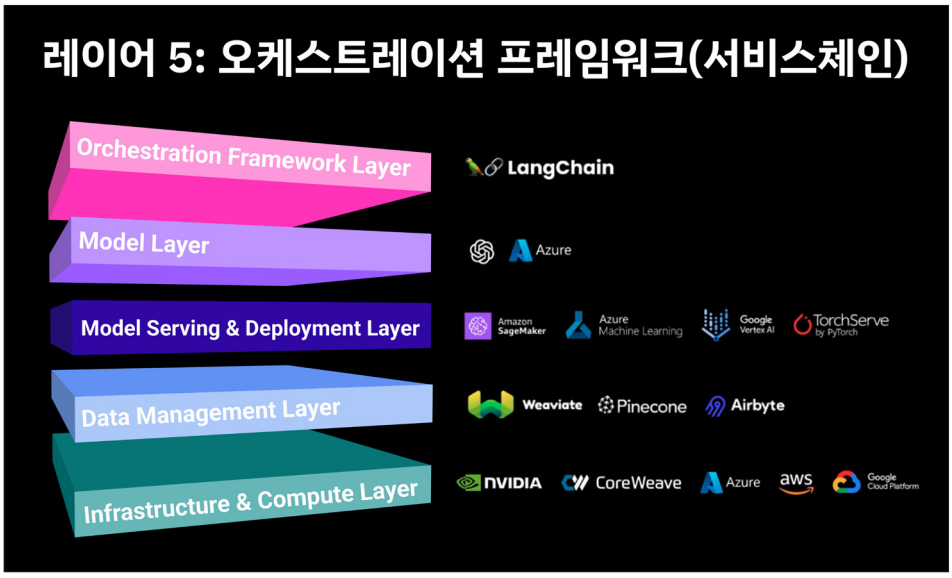
모니터링 플랫폼 업체 데이터독(Datadog Korea)의 조용원 세일즈 엔지니어(SE)는 3월 21일 ITWorld와 CIO가 주최한 'Cloud & AI Research Summit 2024'에서 AI/ML 서비스의 관리 복잡성을 최소화할 수 있는 방법으로 통합된 생성형 AI 모니터링과 AI옵스(AIOps) 도입을 제안했다. AI 서비스를 구성하는 기술 스택별로 필요한 지표를 함께 분석하고 AI옵스를 통해 자동화하면 생성형 AI 서비스를 보다 효율적으로 관리할 수 있다는 설명이다.

기술 스택별 필수 모니터링 지표를 연계 분석하라

조용원 SE는 우선 생성형 AI의 기술 스택을 ▲인프라 및 컴퓨팅 ▲데이터 관리 ▲모델 서빙 및 배포 ▲모델 ▲오케스트레이션 프레임워크 5가지 계층으로 구분하고, 각 계층에서 어떤 지표를 모니터링해야 하는지 설명했다. 먼저 인프라 및 컴퓨팅 계층에서는 빠른 처리 속도와 병렬 처리가 중요하다. 따라서 GPU 상태, 온도 및 팬 속도, 클라우드 사용량과 같은 지표를 확인하면 좋다. 데이터 관리 계층에서는 벡터 데이터베이스의 자료 구조를 반영한 모니터링이 필요하다. 예를 들면 샤드별 활성 LSM 세그먼트 개수, LSM 세그먼트별 엔트리 개수, 샤드 내 블룸 필터 오퍼레이션 응답 시간 등이 있다.

모델 서빙 및 배포 계층은 AI 모델을 비즈니스 요구사항에 맞게 구성하는 단계에서 활용한다. 조용원 SE는 "각 모델을 디자인, 테스트, 배포하는 단계이므로 AI 모델의 학습/추론 작업이 원활하게 수행되는지 품질을 확인할 수 있는 지표가 필요하다"라고 설명했다. 구체적으로 학습 및 결과 출력 시 GPU, 메모리, 네트워크와 같은 리소스 사용량과 예측 건수, 에러, 응답 시간과 같은 모니터링 지표를 확인하면 운영 전 성능 검증에 도움이 된다.

AI 모델 계층에서는 모델 혹은 API가 애플리케이션과 연결된다. AI 모델 및 서비스별 성능과



▲조용원 SE는 생성형 AI의 기술 스택별 모니터링 지표를 연계 분석해야 한다고 강조했다.

사용량, 비용에 대한 가시성 확보가 요구되는 계층이다. 여기서는 서비스/모델/API의 토큰 사용량, 요청 및 응답 시간과 같은 지표 모니터링을 통해 가시성을 확보할 수 있다.

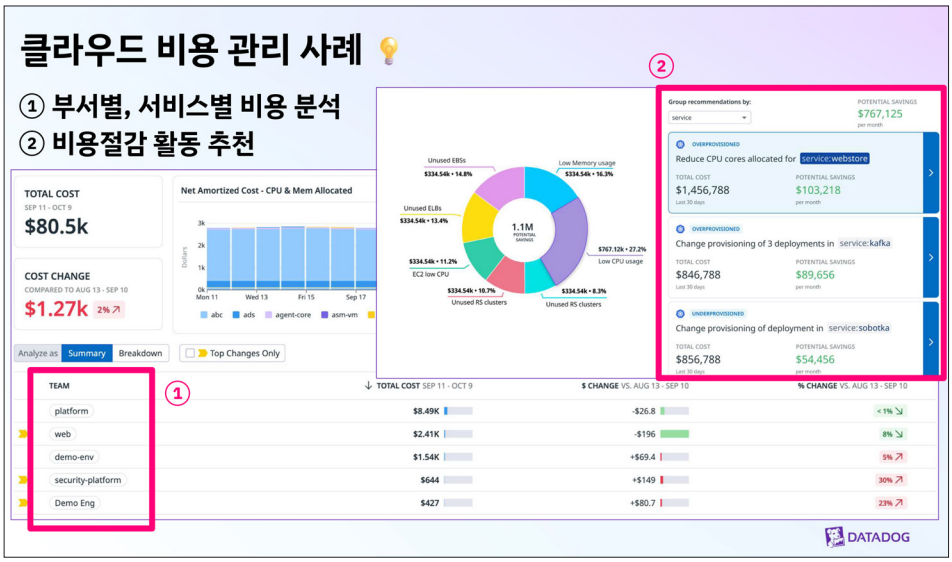
마지막으로 서비스 체인이라고 불리는 오케스트레이션 프레임워크 계층에서는 구성한 AI 모델을 조합해 애플리케이션을 모듈화한다. 서비스 단계에 대한 모니터링이기 때문에 애플리케이션에 사용되는 요청량과 에러, 응답 시간, 토큰 사용량을 모니터링해 성능 및 비용 현황과 문제 발생 여부를 빠르게 파악할 수 있다.

인프라부터 서비스 체인까지 각 계층에서 확인해야 하는 모든 모니터링 지표는 서로 연계 분석할 필요가 있다. 조용원 SE는 "응답 시간이 길어져서 모델 성능이 저하된다고 판단되면 실제 이용되고 있는 토큰량과 리소스 사용량을 검토한 후 성능에 어떤 문제가 있는지 검토할 수 있다. 이것이 AI 서비스를 효과적으로 운영/관리하는 모니터링 전략"이라고 강조했다.

비용 절감을 위한 모니터링 전략을 채택하라

이어 비용을 절감할 수 있는 모니터링 전략이 소개됐다. 조용원 SE는 "클라우드의 사용량 기반 소비 모델은 비용을 유연하게 운용할 수 있다는 이점이 있지만, 과도한 프로비저닝의 위험이 존재한다. 즉, 클라우드 비용을 잘 관리하기 위해서는 실제 리소스 사용량을 모니터링해야 한다"라고 말했다.

기술 스택 모니터링과 마찬가지로 클라우드에서도 성능 및 비용 지표를 종합적으로 검토함으로써 실제 사용량에 기반해 비용을 평가할 수 있다. 조용원 SE에 따르면, 데이터독은 부서별, 서비스별 클라우드 사용 현황을 기간별로 제시하고 어떤 곳에 위치한 서비스의 CPU 코어를



▲데이터독은 부서 및 서비스별 비용을 분석하고 비용 절감에 도움이 되는 방법을 제안한다.

줄여야 하는지, 어떤 배포의 프로비저닝을 변경해야 하는지 등 비용을 절감하는 활동을 추천하는 서비스를 제공한다.

또한 데이터독 같은 모니터링 플랫폼은 매출 등의 비즈니스 지표 대비 비용 분석을 지원하며, 컴퓨팅 자원 사용량 시각화 및 분석을 통해 클라우드 원가율 데이터를 제공한다. 구매 관리 부서 혹은 클라우드 리소스 관리 부서에서는 이런 데이터를 KPI 지표로 삼을 수 있다.

통합 모니터링 플랫폼을 통해 AI옵스를 구현하라

또한 조용원 SE는 "AI 서비스 모니터링 전략을 세울 때는 AI 서비스와 함께 구성되는 다른 서비스도 함께 모니터링해야 한다는 점을 잊지 말아야 한다"라며 통합된 모니터링 환경이 제공하는 이점을 설명했다.

분산된 환경에서는 기술 스택마다 다른 모니터링 채널을 사용하고 데이터 저장소 및 파이프라인이 분산돼 데이터를 활용해 부가가치를 창출하기 어렵다는 단점이 있다. 하지만 SaaS 기반 통합 모니터링 솔루션을 사용하면, 도입하려는 기술에 대한 TTM(Time to market)을 줄이고 TCO(Total cost of ownership) 절감에 기여할 수 있다. 또한 여러 기술 스택에 대해 통합된 모니터링 뷰와 동일한 데이터를 제공하므로 부서별 소통 비용을 줄일 수 있으며, 데이터 간 연계 분석을 지원해 대응 시간 단축에도 기여한다.

특히 조용원 SE는 SaaS 기반 통합 모니터링 솔루션이 AI옵스 구현에 이점이 있다고 역설했다. AI옵스는 플랫폼에서 수집한 데이터를 AI가 학습해 이상 현상 탐지, 원인 분석, 연계 분석, 분류 및 조사, 사후 조치까지 일련의 관리를 자동화하는 관리 방식이다.

데이터독과 같은 통합 모니터링 솔루션은 발견한 장애를 조직 전반에 알리고 여러 팀이 장애 분석 내용과 처리 현황을 확인할 수 있는 타임라인을 제공해 매끄러운 협업을 지원한다. 또한 롤백과 같은 프로세스를 자동화해 피해를 최소화하며, 생성형 AI를 활용해 재발 방지를 위한 개선 대책까지 제안한다.

효과적인 AI 서비스 모니터링 전략은 결국 AI옵스 도입으로 귀결된다. 조용원 SE는 AI옵스를 기반으로 "자원 및 비용 관리를 위해 서비스 성능 지표와 리소스 사용량, 비용 지표를 모두 함께 분석하는 전략을 권장하며, AI 서비스와 기존 서비스의 연관 데이터를 통합해 모니터링하는 전략을 고민할 필요가 있다"라고 설명했다. 또한 AI옵스를 위한 프로세스와 조직 문화도 기업 전반에 정착해야 한다. 조용원 SE는 "데이터 수집과 이슈 탐지, 자동화를 플랫폼에 맡기더라도 운영자는 선제적 대응을 위한 워크플로우를 만들고 AI옵스 플랫폼을 활용한 운영 시나리오를 설계해야 한다"라고 조언했다.

Session Brief

“낭비되는 클라우드 비용과 자원을 찾아라”…
IBM이 전하는 핀옵스 접근법

많은 기업이 클라우드를 도입하면서 새로운 비즈니스 기회를 모색하고 있다. 실제로 기업의 클라우드 지출 규모는 매년 증가하고 있으며, 2024년 클라우드 지출 비용은 6,788억 달러(약 900조 원)로 예상되고 있다.[1] 전년 대비 20% 증가한 수치다. 하지만 클라우드 도입으로 기업은 또 다른 고민을 마주하고 있다. ‘비용 최적화’라는 문제다. 가트너는 기업에서 구매한 클라우드 자원 중 30%는 미사용되고 있다는 분석을 내놓기도 했다. 규모가 크거나 여러 클라우드 서비스를 사용하는 기업일수록 비용 관리 문제로 더욱 어려움을 겪을 확률이 높다.

클라우드 비용 관리가 기업 내 주요 과제로 떠오르면서 ‘핀옵스’라는 개념도 주목받고 있다. 사전적으로 핀옵스는 ‘클라우드 재무 관리 분야이자 문화적 실천 방법’이다. 쉽게 말해 우리 조직에 필요한 클라우드 자원을 최적의 비용으로 쓰기 위해 도입하는 각종 관리법이다. 재무와 관련 있지만 핀옵스를 실행하기 위해선 엔지니어링팀, 업무팀 등 다양한 팀의 협업이 필요하다. 아직 낯선 개념인 핀옵스를 제대로 실행하려면 무엇을 중점적으로 살펴봐야 할까? 마침 3월 21일 IT월드와 CIO코리아가 주최한 클라우드&AI 리서치 컨퍼런스에서 한국 IBM의 조상철 상무가 성공적인 핀옵스 도입을 위한 접근법을 소개했다.

지출 규모에 따라 달라지는 핀옵스 핵심 과제

핀옵스는 크게 3가지 영역을 거쳐 수행할 수 있다. 첫째, 클라우드 리소스 사용량과 관련 비용을 구체적인 수치로 확인해야 한다. 둘째, 그렇게 확보한 데이터를 기반으로 사용량과 비용을 최적화하는 방안을 찾아야 한다. 셋째, 그에 필요한 작업을 지속적으로 수행하며 핀옵스 작업

1 https://www.gartner.com/en/newsroom/press-releases/11-13-2023-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-679-billion-in-20240

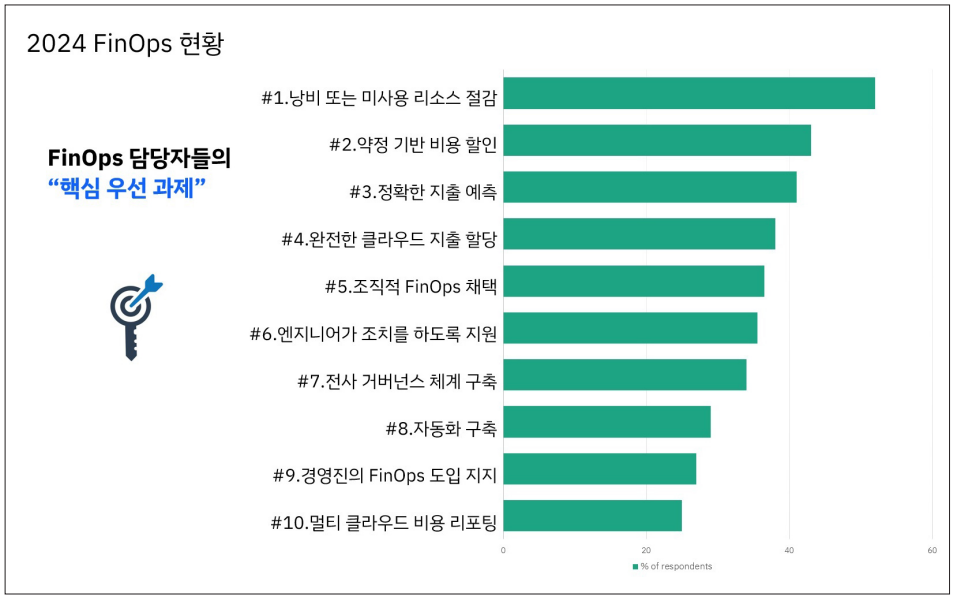
IBM IT WORLD CIO

을 고도화해야 한다.

이러한 세 가지 틀 안에서 핀옵스 담당자는 다시 세부적인 과제를 처리해야 한다. 비영리 단체 핀옵스재단의 설문조사에 따르면, 보통 다음과 같은 과제를 조직에서 고민하고 있었다.

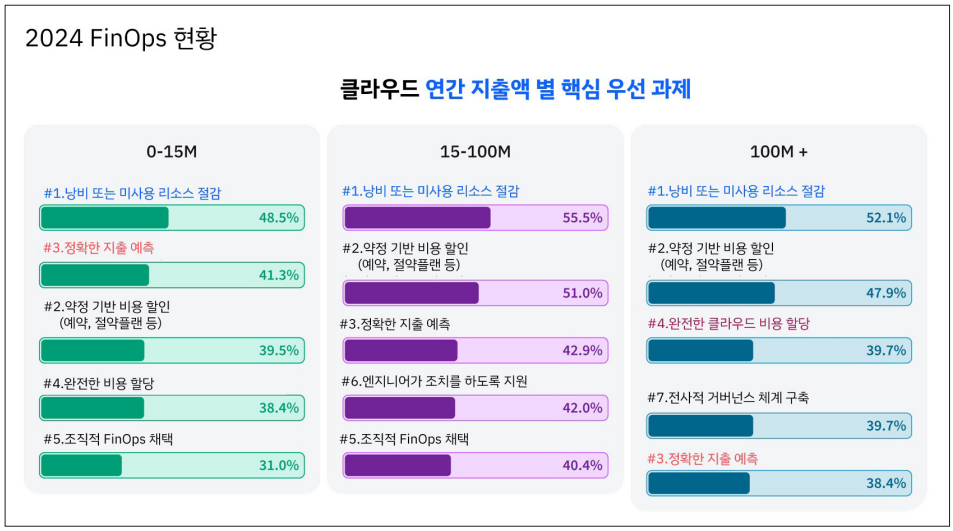
이런 과제의 우선순위는 클라우드 지출 규모에 따라 다시 달라질 수 있다. 가령 클라우드에 연간 1,500만 달러 이하로 쓰는 기업은 정확한 지출 금액을 예측하는 것을 두 번째로 중요한 과제로 꼽았다. 한국IBM 오토메이션 SME 조상철 상무는 "클라우드 지출이 적으면 비용 증가율

그림 1 | 핀옵스 담당자의 주요 과제 상위 10개



(출처:핀옵스재단/IBM)

그림 2 | 클라우드 지출 규모별 주요 과제 상위 5개



(출처:핀옵스재단/IBM)

이 상대적으로 두드러져 보인다. 자연스레 핀옵스 관련 작업에서 예상대로 비용이 나오고 있는지를 더 중점적으로 볼 수 있다"라고 설명했다.

반면 클라우드에 연 1억 달러 이상을 쓰는 기업은 클라우드 비용 할당과 전사적 거버넌스 체계 구축을 보다 중요한 과제로 보고 있었다. 조상철 상무는 "연간 1억 달러 이상 클라우드 비용을 쓰는 곳은 대기업일 확률이 높다. 대기업은 어느 정도 클라우드 비용을 쓰는 것에 익숙한 상태이면서 비용 관리 문화에 더 관심을 쏟고 있어 주요 과제가 달라진 것"이라고 설명했다.

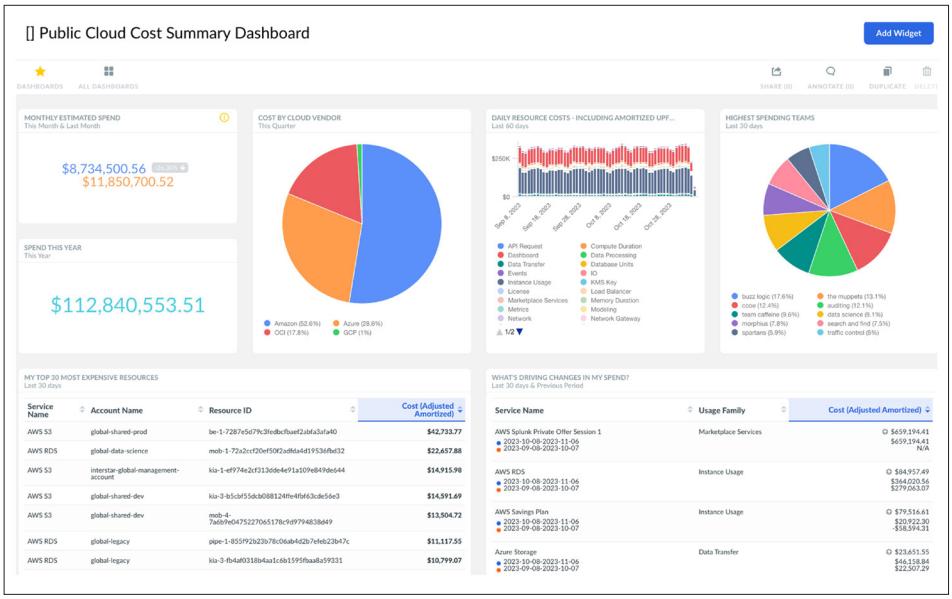
지출 비용과 상관없이 핀옵스 담당자가 공통적으로 관심을 두는 과제도 있었다. '자동화 구축'이라는 영역이다. 조상철 상무는 "비용을 절감하는 작업은 일회성이 아니고 계속 반복이 필요한 과정이다. 그러다 보니 핀옵스의 여러 작업을 자동화하려는 수요가 점점 늘고 있다"라고 분석했다.

'수동' 작업 한계를 극복하기 위해 만들어진 핀옵스 '솔루션'

핀옵스는 문화이자 방법론이기에 기업이 핀옵스를 구체적으로 도입하기 위해서는 담당 인력을 배정하고 필요한 도구도 구비해야 한다. IBM은 핀옵스 영역에 마켓 리더로 애플리케이션 사용 기반 최적화 추천과 클라우드 비용 최적화를 제공하는 터보노믹 솔루션에 이어 핀옵스 재단 창립 멤버 기업인 앱티오를 인수하였고 앱티오 클라우드어빌리티(Apptio Cloudability)와 터보노믹을 연계하여 핀옵스 과제를 적절하게 해결할 수 있는 포트폴리오를 완성했다.

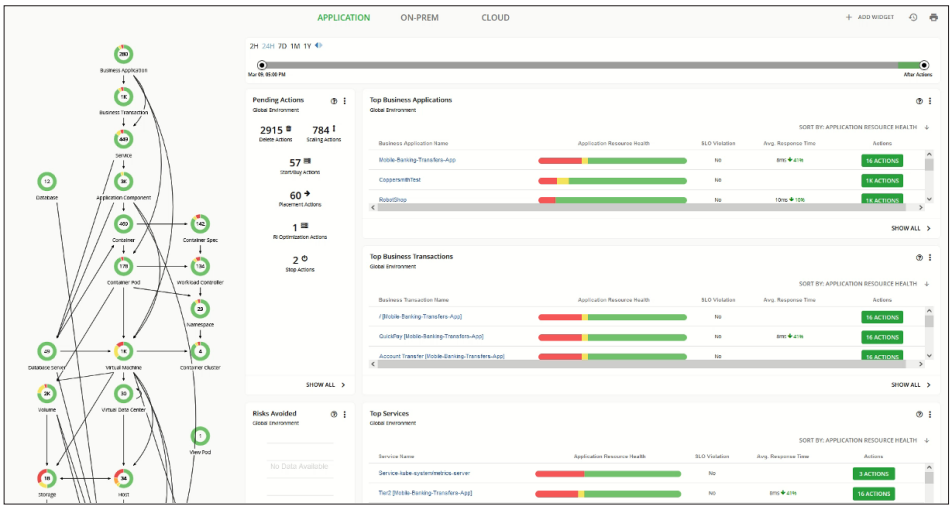
조상철 상무는 "앱티오 클라우드어빌리티 솔루션은 재무 담당자에게 유용하도록 '파이낸스

그림 3 | IBM 앱티오 핀옵스 솔루션 예시



(출처:IBM)

그림 4 | IBM 터보노믹 핀옵스 솔루션 예시



(출처:IBM)

(Finance)' 중심 과제 해결에 최적화된 솔루션이다. 터보노믹은 클라우드 운영 엔지니어에게 도움을 주기 위해 추천, 자동화 같은 '오퍼레이션(Operation)' 업무에 최적화된 솔루션이다. 따라서 두 솔루션으로 앞서 언급한 핀옵스 주요 과제 10가지를 보다 쉽게 해결할 수 있다"라고 설명했다.

구체적으로 살펴보자면, 앱티오 클라우드어빌리티는 멀티 클라우드 비용에 대한 투명성을 높여주며 클라우드 사용 주체인 업무팀에서 비용을 직접 확인하고 관리할 수 있게 도와준다. 특히 업무팀은 다양한 대시보드로 비용을 종합적으로 분석하거나 팀 또는 부서 별로 비용을 쉽게 할당할 수 있다.

터보노믹은 애플리케이션 성능 기반 자원 최적화 관리와 자동화에 초점을 둔 제품이다. 터보노믹에 내장된 다차원 AI 모델이 애플리케이션의 성능 데이터를 수집 및 분석해 성능 최적화를 위해 진행해야 하는 조치를 추천하는 식이다. 조상철 상무는 "터보노믹의 자동화 기능으로 엔지니어는 추천 결과를 손쉽게 반영할 수 있으며 결과적으로 업무 효율성을 높일 수 있다"라며 "여러 최적화 수행 활동으로 애플리케이션 성능을 안정적으로 유지하면서 궁극적으로 클라우드 비용도 최적화할 수 있을 것"이라고 소개했다.

글로벌 기업의 핀옵스 도입 성과

글로벌 기업 상당수는 핀옵스를 활용해 비용 절감 효과를 보고 있다. 가령 코흐(Koch)라는 기업을 보자. 코흐는 석유, 에너지, 금융 서비스 등을 제공하며 60개가 넘는 지사를 보유한 글로벌 대기업이다. 직원 수는 10만 명이 넘는다. 코흐는 클라우드 마이그레이션 전략을 수립하면서 클라우드 비용을 스프레드시트로 관리했었다. 문제는 스프레드시트에 데이터를 입력하는 데만 매달 40시간이 필요했다는 점이다. 그뿐만 아니라 수동으로 비용을 관리하면서 클라우

드 관련 초과 지출이 자주 발생하고, 비용이 아예 다른 팀에 잘못 할당되기도 했다.

이런 문제를 해결하기 위해 코흐는 앱티오 클라우드어빌리티를 도입했다. 덕분에 수백 명의 담당자가 지출 영역을 자동화된 대시보드에서 직접 관리하고 있다. 동시에 경영진에게 투명하고 정확하게 사용 현황과 비용을 보고하는 문화를 구축할 수 있었다. 여기에 AWS의 예약 인스턴스 구매 프로젝트를 위해 500만 달러 규모의 비용을 절약하는 데 성공했다.

미국의 글로벌 은행 A는 IBM의 터보노믹을 도입해 인프라 자원을 보다 효율적으로 관리하고 있다. A 은행은 100개 이상의 쿠버네티스 클러스터와 6만 개 이상의 컨테이너를 운영하면서 인프라 성능 관리에 많은 리소스를 투입했어야 했다. 하지만 터보노믹을 통해 인프라 관리 영역을 상당수 자동화하고 지원 요청 티켓 건수가 이전보다 1만 4,000건 감소하면서 2,700만 달러 규모의 비용을 절감하는 효과를 보았다.

미국의 석유회사인 셰브론(Chevron)은 서비스 인스턴스 1만개와 애저 인스턴스 7,000개를 운영하고 있었다. 이렇게 리소스가 많으면 가상화로 리소스 경합 상황이 발생할 수 있는데, 실제로 셰브론은 성능 하락 문제로 모니터링에 여러 노력을 투자하고 있었다. 터보노믹을 도입한 셰브론에선 성능 문제가 이전보다 95%를 감소했고, 운영 및 관리에 들어갔던 3만 7,000시간이 절약됐다.

조상철 상무는 “많은 기업은 기술을 도입할 때 효과와 더불어 ROI를 생각한다. 클라우드도 예외는 아니다”라며 “IBM의 핀옵스 솔루션은 클라우드의 성능을 높이는 동시에 비용을 줄여주면서 클라우드 ROI를 높이는 데 효과적이다”라고 밝혔다.

Session Brief

"CDN 캐시 서버에 AI 칩을 넣는다면?" 지코어가 제안하는 비용 효율적인 AI 구축 전략

바야흐로 AI 시대다. 챗GPT가 불과 2달 만에 사용자 1억 명을 모으며 대성공을 거둔 이후 다양한 AI 서비스가 우후죽순 등장했다. 상당수는 무료 버전으로도 사용하는 데 무리가 없다. 그렇다면 이들 AI 서비스는 어떻게 수익을 낼까? 수익 없이 얼마나 더 운영될 수 있을까?

가트너의 AI 하이프 사이클 보고서에 따르면, AI는 이미 지난해 '부풀려진 기대의 정점(peak of inflated expectations)'에 도달했다. 상용화 2년 만에 이 단계에 진입한 기술은 AI가 처음이다. 더 놀라운 것은, 2025년이면 '환멸의 계곡(trough of disillusionment)'에 빠질 것이라는 전망이다. 부풀려진 기대에서 깨어나 투자와 관심이 줄어드는 시기다. 과거 야후, 라이코스, 알타비스타 같은 유명 검색엔진이 이 환멸의 계곡 시기를 버티지 못하고 사라졌다.

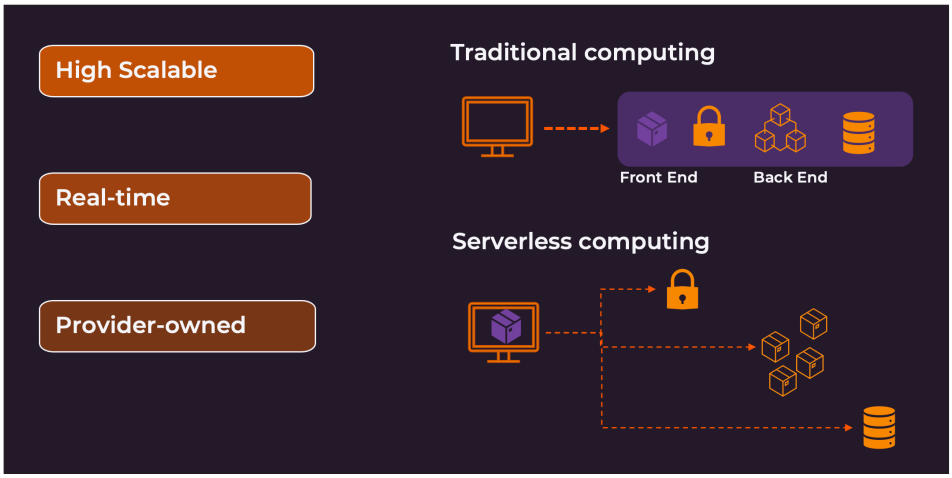
AI 업계 전체가 이 어려운 시기를 견디고 AI 기술이 '생산성의 안정 단계(Plateau of Productivity)'에 접어들 때까지 지속적으로 발전할 수 있는 방법을 찾아야 한다.

"CDN 서버에 AI 칩을 심으면?"

지난 21일 한국IDG가 '지속적인 비즈니스 성장을 이끄는 AI 및 클라우드'를 주제로 진행한, <클라우드 & AI 리서치 서밋 2024> 행사에서, 김진용 지코어 코리아 팀장은 이런 방법을 찾는 출발점으로 AI 서비스를 구성하는 핵심적인 두 요소, 즉 학습과 추론을 구분해 효율성을 검토할 것을 제안했다.

현재 대부분 AI 서비스는 학습과 추론을 구분하지 않고 고가의 GPU 칩을 활용해 작업을 처리한다. 하지만 학습 관련된 작업이 아니라 이미 만들어진 모델로 추론하는 작업이라면 굳이 값비싼 GPU가 필요하지 않다. 따라서 학습과 추론을 분리해 각각에 필요한 최적화된 인프라를

그림 1 | AI 추론을 위한 서버리스 컴퓨팅



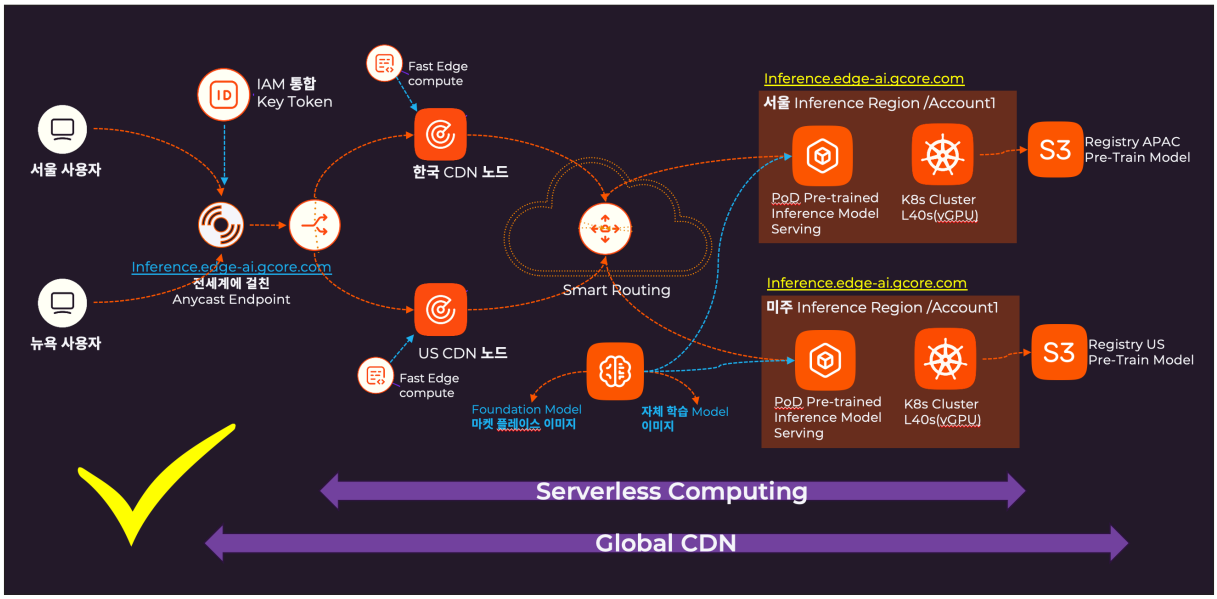
제공하면 전체 AI 서비스의 비용 효율성을 극대화할 수 있다. 스퀴즈비츠 같은 기업이 추론 과정에 필요한 메모리와 연산량을 줄인 경량화된 AI 모델을 만드는 것도 이런 고민의 결과다.

김진용 팀장에 따르면, 지속가능한 비용 효율적인 AI 서비스를 고민할 때 고려해야 할 또 다른 요소가 바로 네트워크다. 김진용 팀장은 "더 정확히 말하면 CDN 네트워크 속 캐시 서버에 AI 처리에 특화된 칩을 심는 아이디어다. 기존 캐시 서버는 이미지나 영상을 빠르게 보내는 역할을 해 왔다. 그런데 이 서버에 엔비디아 L40S 같은 작지만 강력한 AI 칩을 넣으면 어떨까? 전 세계에 펼쳐져 있는 CDN 네트워크를 통해 필요한 사용자에게 가장 가까운 캐시 서버에서 AI 작업을 처리할 수 있다"라고 말했다. 이는 일종의 서버리스 방식이므로 확장성이 뛰어나고 실시간으로 작동한다. CDN 서비스 업체가 컴퓨팅 파워부터 로직까지 담당하므로 AI 서비스 기업은 관리 부담을 덜고, 서비스 사용자는 더 편리하게 원하는 서비스를 쓸 수 있다.

이렇게 되면 AI 추론 작업을 처리하는 과정이 완전히 달라진다. <그림 1>과 같이 기존 컴퓨팅 환경에서는 프론트엔드, 보안, 백엔드, 데이터베이스가 사용자의 반대편, 한쪽으로 몰려 묶여 있는 반면, 서버리스 컴퓨팅 환경에서는 PC나 스마트폰의 컴퓨팅 성능을 이용해 즉, 프론트엔드 기능의 일부를 직접 담당해 효과적으로 AI 작업을 처리할 수 있다. 보안이나 백엔드, 데이터베이스 기능 역시 전문 업체를 통해 외부에서 처리할 수 있다. 결과적으로 AI 서비스 기업은 백엔드 전체를 더 유연하고 비용 효율적으로 운영할 수 있다.

이처럼 CDN 캐시 서버에 AI 칩을 넣어 '엣지 AI'를 구성하면 스마트 라우팅이 가능해진다. 사용자 혹은 AI 서비스 기업의 설정에 따라 빠른 응답이 필요하면 엣지에서 처리하고, 더 많은 컴퓨팅 파워가 필요하면 충분한 처리 역량을 지원하도록 연결한다. 모델 응답을 캐시하는 것도 가능하다. 보통 아침에 일어나 가장 먼저 AI에 물어보는 질문은 오늘의 날씨일 것이다. 이런 질문을 미리 캐시 서버에 설정해 두면 더 빠르게 서비스할 수 있다. 콘텐츠를 즐길 때 사용자가

그림 2 | 지코어 엣지 AI가 작동하는 방식



CDN을 통해 누리던 혜택이 AI 서비스에도 그대로 적용된다.

김진용 팀장은 "지역 특성에 맞춰 AI 서비스를 맞춤화할 수도 있다. 글로벌 서비스를 하다 보면 정치적인 문제 등으로 지역에 따라 답변을 다르게 내놓아야 하는 경우가 있다. 엣지 AI를 활용하면 실제 사용자의 위치에 따라 적절하게 작동하도록 할 수 있다"라고 말했다.

지코어 '엣지 AI'가 특별한 이유

캐시 서버에 컴퓨팅 파워를 할당해 CDN 같은 AI를 구현해 비용 효율적이고 지속가능한 서비스 방식을 만들자는 아이디어 자체는 간단하다. 그러나 이런 서비스를 실제 구현하려면 전 세계 규모의 빠른 네트워크가 필수적이다. 거대 클라우드 업체도 이런 네트워크를 확보한 기업은 거의 없다.

김진용 팀장은 "지코어가 AI 플랫폼과 엣지 AI를 내놓을 수 있었던 것도 전 세계 규모의 초고속 CDN 네트워크를 확보하고 있기 때문이다. 지코어는 IPU(Intelligent Processing Unit), 엔비디아 GPU, 머신러닝 가속 클라우드 인프라 등을 퍼블릭/프라이빗 방식으로 지원한다. 전 세계 160곳 이상의 캐시 서버를 통해 183개국에 초고속 CDN 서비스를 운영 중이다"라고 말했다.

지코어의 엣지 AI 운영 방식은 <그림 2>와 같다. 서울 사용자와 뉴욕 사용자가 있다면, 이들은 보안 기능이 통합된 단일 서비스 UI를 통해 접속한다. 서울 사용자는 한국 CDN 노드로, 뉴욕 사용자는 미국 CDN 노드로 연결된다. 프론트 엔드의 기능 일부를 이들 노드에 배치해 더 빠르게 처리할 수 있다.

이제 두 사용자는 스마트 라우팅을 통해 분류된다. 간단한 로직은 엣지 노드에서 바로 처리한다. 반면 그 이상의 컴퓨팅이 필요할 때가 있다. 예를 들어 청소 로봇의 경우 평상시에는 문제가 없지만 애완견 같은 인식 대상이 추가되면 새로운 연산 작업이 요구된다. 이럴 때는 더 강력한 연산을 지원하는 시스템으로 넘겨 계산하게 된다.

이외에도 AI 서비스를 제공하는 기업이 필요로 하는 모델이 있으면 마켓플레이스에서 즉시 로딩해 사용하거나 자체 학습 모델을 사용해 처리하는 것도 가능하다. 이런 모델을 빠르게 불러오고 사용 이후 할당한 리소스를 해제하려면 빠른 네트워크가 필요하다. 지코어의 글로벌 CDN과 서버리스 컴퓨팅을 결합했을 때 강력한 시너지가 나는 이유다.

마지막으로 남은 문제는 보안이다. 최근 저렴한 애완견 모니터링 기기가 큰 인기를 끌고 있는데, 이는 곧 외부에서 우리집을 볼 수 있다는 의미이기도 하다. AIoT(Artificial Intelligence of Things) 아키텍처가 필요한 이유다. 로보틱스, 자율주행차까지 고려하면 보안의 중요성은 더 커진다.

김진용 팀장은 "AI 보안의 핵심은 사용자의 데이터가 얼마나 안전하게 AI 핵심 시스템까지 전달되느냐다. 해법은 5G 모바일 제로트러스트 네트워크를 구성하는 것이다. 지코어는 인프라 투자 없이 제로 트러스트 5G 프라이빗 네트워크를 지원한다. 다양한 IoT 기기에서 별도의 VPN이나 와이파이 없이도 전 세계에서 보안 네트워크에 연결할 수 있다. 모든 트래픽은 암호화되어 글로벌 네트워크를 통해 라우팅된다"라고 말했다.

AI 시대를 살아갈 미래 세대를 위한 준비

지난해 챗GPT를 서비스하는 오픈AI의 파산 가능성을 제기한 보도가 있었다. 결과적으로는 잘못된 전망으로 정리되고 있지만, 더 비용 효율적인 AI 서비스 방안에 대한 고민이 필요하다는 것은 분명하다.

김진용 팀장은 이런 고민에 추가해야 할 한 가지로 AI의 윤리성을 꼽았다. 김진용 팀장은 "AI가 전 산업으로 급속히 확산하면서 현재는 물론 미래 세대의 삶에도 막대한 영향을 줄 수 있다는 분석이 나오고 있다. GDPR 같은 법률이 이미 시행 중이고 AI를 규제하는 국가는 점점 더 늘어날 것이다. 지속가능한 AI에 대한 논의는 미래를 위한 고민이기도 하다"라고 말했다.

마지막으로 김진용 팀장은 "이번 행사에서는 CDN과 AI를 결합하는 방안을 살펴봤지만 이것만이 비용 효율적이고 지속가능한 AI 서비스를 구현하는 유일한 해법은 아니다. 앞으로 더 다양한 논의가 이뤄지길 기대한다"라고 말했다.