AI를 Google Cloud에 안전하게 배포하기 위한 모범 사례

Anton Chuvakin, Luis Urena

Office of the CISO, Google Cloud

Google Cloud

소개 **3**
시나리오 **3**
AI 위협 **5**
AI 보안 및 책임 공유 **7**
AI를 위한 Google의 안전한 집 **8**
고객 모범 사례 **8**
모델 보안 8
애플리케이션 보안 8
인라인 프롬프트 및 응답 보호 8
인프라 보안 10
소프트웨어 공급망 10
서빙 인프라 10
데이터 보안 12
저장 데이터 암호화 12
민감한 데이터 보호 12
데이터 준비단계에서의 보호 13
로깅, 탐지 및 대응 13
로깅 및 모니터링 13
Vertex AI를 위한 탐지 13
Google Kubernetes Engine의 AI 워크로드 탐지 14
탐지 강화 14
AI와 함께 미래를 건설하다 **15**
부록 A: 모범 사례 체크리스트 **16**

소개

Google Cloud를 포함한 클라우드 플랫폼의 채택 증가는 인공 지능(AI)의 빠른 혁신을 촉진합니다. 모든 클라우드 앱이 보호되어야 하는 일반적인 위협 외에도 AI 시스템은 직면하는 고유한 위협으로부터 보호되어야 합니다.

본 문서는 Google Cloud에서 AI 워크로드를 보호하는 측면에서 공유 책임 모델을 살펴보고 Google이 제공하는 사항과 고객이 수행해야 하는 중요한 단계를 강조합니다. 필수 보안 영역을 검토하고 프롬프트 삽입과 같은 모델별 문제를 해결하며 현재 및 미래의 위험에 대해 AI 시스템을 더욱 탄력적으로 만드는 사전 예방적 조치를 간략하게 설명합니다. 이 실용적인 접근 방식은 AI 시스템 보안을 위한 개념적 프레임워크인 Google의 보안 AI 프레임워크(SAIF)와 보안 AI와 기존 시스템 간의 차이점 및 유사점에 대한 이해에 기반합니다.

Google Cloud 또는 기타 최신 클라우드 환경에 AI 워크로드를 배포할 때 특정 보안 모범 사례는 기존 클라우드 보안에서 시작됩니다. 예를 들어, 다른 환경에 배포하는 것과 같이 ID 관리, 네트워크 제어 및 액세스 관리에 고객이 집중하도록 권장합니다.

그러나 데이터 및 네트워크 보안, **훈련 데이터 보안**, **입력 및 출력 필터링**(안전, 개인정보 보호 및 허용 가능한 사용과도 관련됨)을 포함한 다른 제어 기능도 있으며 VPC 서비스 제어(VPC Service Controls, VPC-SC) 및 민감한 데이터 보호(Sensitive Data Protection, SDP)와 같은 클라우드 네이티브 기술을 사용합니다. 이러한 도구를 사용하면 AI 워크로드를 클라우드에 더욱 안전하게 배포할 수 있습니다.

시나리오

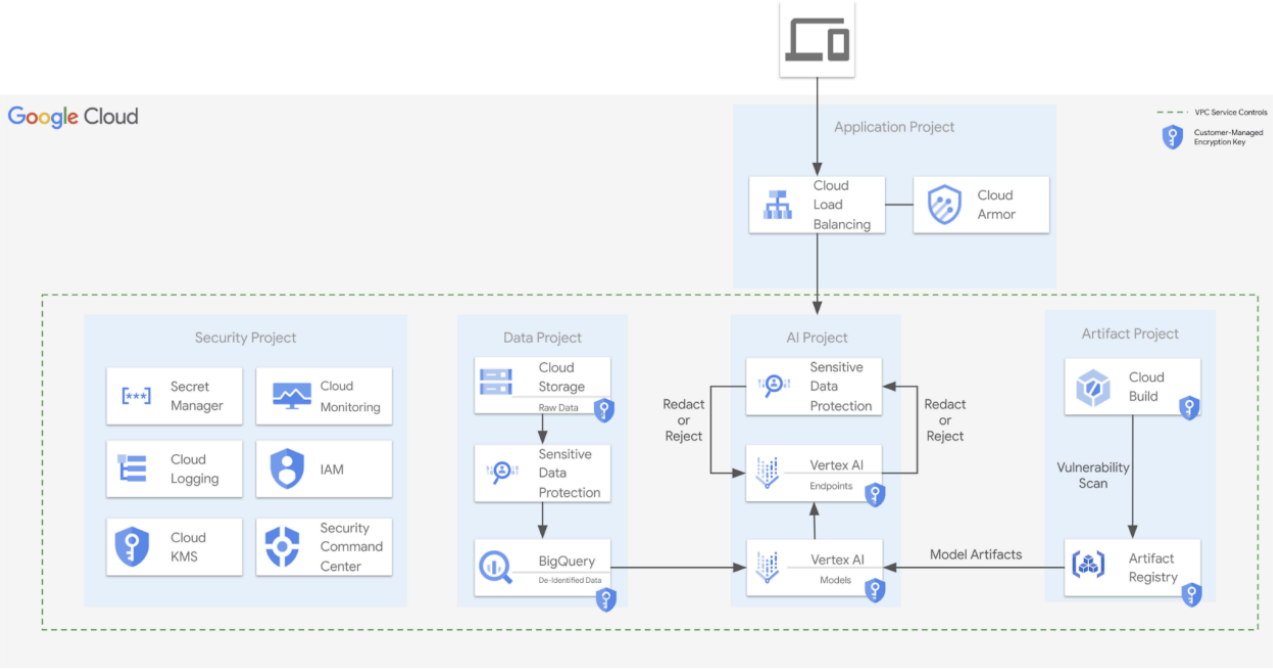
모든 조직에서 AI의 힘을 활용하기로 한 결정은 다음을 포함한 수많은 질문에 달려 있습니다.

- AI를 사용하여 비즈니스를 개선, 확장, 가속화하려면 어떻게 해야 합니까?
- 데이터 과학팀에 가장 효과적인 도구와 프로세스는 무엇입니까?
- 팀이 솔루션을 제공하는 데 필요한 데이터 양(및 유형)은 얼마나 됩니까? 데이터는 어디로부터 가져오나요?
- AI는 우리 조직에 어떤 종류의 위험이나 위협을 가져오는가?
- **어떻게 하면 조직에서 이러한 위험을 최소화하면서 AI의 기능을 활용하도록 도울 수 있나요?**

이 백서는 비즈니스 운영에 잠재적으로 영향을 미칠 수 있는 보안 위험 및 관련 위협을 최소화하면서 조직이 AI의 기능을 활용하도록 돕는 데 중점을 두고 최종 질문을 해결하려고 합니다.

본 문서의 목적을 위해 귀하의 조직에 AI/ML 전략과 Google Cloud의 통합 AI 플랫폼인 Vertex AI에서 모델을 구축하고 훈련할 데이터 과학자 팀이 있다고 가정합니다. 내장된 보안 기능을 강화하기 위해 고객은 조직 정책과 함께 작동하여 정책 및 AI 리소스 구성 변경 사항을 거의 실시간으로 감지하는 Google Cloud의 멀티 클라우드 보안 및 위험 관리 솔루션인 Security Command Center와 같은 다른 서비스를 배포할 수 있습니다.

본론에 들어가면서 샘플 애플리케이션을 나타내는 다음의 **아키텍처**를 제안합니다. 여기에는 데이터 과학자가 Vertex AI에서 즉각적인 이점을 얻을 수 있도록 하는 데 필요한 최소한의 보안 제어 집합이 포함되어 있으며 보안 팀이 강력한 보안 태세를 유지하는 데 자신감을 가질 수 있도록 합니다.



특히 이 아키텍처는 다음을 수행하는 데 도움이 됩니다.

- Cloud Armor로 **외부 트래픽을 관리**하고 제어합니다.
- SDP를 활용하여 **민감한 데이터를 식별**하고 입력 및 출력을 허용하거나 차단할지 결정합니다.
- Cloud KMS(Key Management Service)를 사용하여 민감한 데이터 및 컨테이너 이미지를 **암호화**합니다.
- VPC 서비스 제어(VPC-SC)를 사용하여 **백엔드 인프라를 보호**합니다.
- Cloud IAM 및 조직 정책으로 강력한 **ID 및 액세스 관리**를 유지합니다.
- 로깅, 모니터링 및 경고를 통해 강력한 **사고 대응**을 지원합니다.

이러한 각 보안 제어는 AI 워크로드와 관련된 위협을 해결하는 데 도움이 되는 다양한 방어 계층 역할을 합니다.

AI 위협

AI 시스템은 기존의 사이버 보안 위험을 물려받는 동시에 **고유한 위협 스펙트럼에 직면**합니다. 구체적인 영향과 완화 방법은 비즈니스에서 AI를 사용하는 방식과 관련된 데이터의 민감도에 따라 다릅니다. 중요한 것은 AI 애플리케이션을 개발하는 주체(본인 또는 제3자)에 따라 위험 관리가 달라질 수 있다는 것입니다. 일반적으로 AI 위협은 **모델**, **애플리케이션**, **인프라** 및 **데이터**의 네 가지 범주를 중심으로 발생합니다. 이러한 범주에서 대부분의 AI 배포에 대한 몇 가지 일반적인 보안 위험과 관련 위협을 주요 영향별로 그룹화하여 요약했습니다.

모델 조작 및 회피

- **프롬프트 인젝션**: 모델의 취약점을 악용하거나 유해한 콘텐츠 또는 부정확한 응답을 생성하도록 유도하는 언어 모델에 대한 악의적인 입력 제작.
- **모델 회피**: 잘못된 분류 또는 예상치 못한 동작을 유발하도록 설계된 입력 데이터의 미묘한 변형(예: 이미지의 픽셀 변경)과 관련된 적대적 공격.
- **안전하지 않은 모델 출력**: 출력 필터링 또는 모니터링이 부적절한 경우 AI 모델은 오해의 소지가 있거나 불쾌하거나 안전하지 않은 콘텐츠를 생성할 수 있습니다.

애플리케이션 손상

- **ML 서비스 거부**: AI 시스템에 과부하를 주거나 리소스를 조작하는 공격은 서비스를 사용할 수 없게 만들어 AI 출력을 사용하는 운영을 중단시킬 수 있습니다.
- **안전하지 않은 통합 시스템**: AI 모델과 상호 작용하는 소프트웨어의 취약점은 공격자가 액세스 권한을 얻거나 악성 코드를 삽입하거나 시스템 운영을 손상시키는 데 악용될 수 있습니다.
- **모델 리버스엔지니어링**: 모델의 입력, 출력 또는 동작을 분석하여 지적 재산 또는 독점 알고리즘을 추출하려는 시도.
- **무단 모델 작업**: 의도된 용도 이상의 작업을 수행하기 위해 AI 모델의 제어권을 확보하여 잠재적으로 유해한 결과를 초래합니다.

AI를 위한 인프라 공격

- **모델 백도어 삽입**: 나중에 특정 입력으로 공격자가 모델의 동작을 조작할 수 있도록 모델 훈련 과정 중에 숨겨진 트리거를 삽입합니다.
- **모델 유출**: 기능을 복제하거나 지적 재산을 추출하기 위해 AI 모델을 무단으로 도용합니다.
- **모델 서빙 손상**: 무단 모델을 삽입하거나 응답을 손상시키기 위해 서빙 인프라(모델이 예측하는 곳)를 조작하는 공격.

데이터 노출

- **민감한 데이터 유출**: 기밀 훈련 데이터 또는 입력을 교묘하게 사용하고 응답을 관찰하여 모델에서 추출한 데이터의 노출.
- **추론된 민감한 데이터**: AI 모델은 관련성 있고 정확한 정보를 제공하기 위해 LLM 프롬프트에 새로운 또는 독점 데이터를 풍부하게 하는 검색 증강 생성(RAG)과 같은 프레임워크를 통해 모델이 발견한 패턴 또는 상관 관계를 통해 파생되어 훈련 데이터에 명시적으로 포함되지 않은 민감한 정보를 의도치 않게 공개할 수 있습니다.
- **데이터 및 모델 포이즈닝**: 모델 성능을 저하시키거나 특정 결과로 결과를 왜곡하거나 숨겨진 백도어를 생성하기 위해 훈련 또는 재훈련 중에 악성 데이터를 삽입합니다(Vertex AI에서 사용하는 것과 같은 frozen 모델에는 적용되지 않음).
- **과도한 데이터 보존**: AI 모델을 훈련하는 데 사용되는 데이터를 불필요하게 저장하여 윤리적 및 법적 위험을 모두 초래합니다.

AI 영역에서 위협 환경은 빠르게 진화합니다. 강력한 AI 보안을 위해서는 새로운 위협에 대해 경계를 늦추지 않고 정보를 얻는 것이 중요합니다. Mandiant 및 기타 출처에서 최신 사이버 위협에 대해 자세히 알아볼 수 있습니다.

Google Cloud

AI 보안 및 책임 공유

Google Cloud에서는 기업이 생성형 AI(gen AI)의 잠재력을 최대한 활용할 수 있도록 효과적인 AI 위험 관리 전략을 개발하도록 지원하기 위해 노력하고 있습니다. 우리는 책임공유가 이러한 노력의 핵심 구성 요소이며 Google Cloud에서 AI 워크로드를 보호하는 데 중요한 개념이라고 믿습니다. 즉, Google과 고객 모두 AI 시스템을 보호하는 데 필수적인 역할을 수행합니다. 조직이 타사에서 개발한 AI 애플리케이션을 사용하는 것이 아니라 자체 애플리케이션을 개발하는 사용 사례에서 Google과 고객 간의 핵심 책임은 다음과 같습니다.

Google의 책임은 다음과 같습니다.

- **인프라 보안**: Google은 데이터 센터, 네트워크 및 운영 체제를 포함한 기본 Google Cloud 인프라의 물리적 및 논리적 보안을 책임집니다.
- **플랫폼 보안**: Google Cloud는 암호화, 취약점 관리, ID 및 액세스 관리(IAM) 제어와 같은 기능을 제공하여 보안을 염두에 두고 구축되었습니다.
- **플랫폼 규정 준수**: Google은 데이터 개인 정보 보호 및 보안을 보장하기 위해 FedRAMP, HIPAA 및 PCI DSS와 같은 엄격한 규정 준수 표준을 준수합니다.

고객의 책임은 다음과 같습니다.

- **Identity and Access Management (IAM)**: 고객은 AI 리소스 및 데이터에 대한 액세스를 제어하기 위해 IAM 모범 사례를 구현할 책임이 있습니다. 여기에는 최소 권한, 서비스 계정 사용, IAM 정책의 정기적인 검토 및 업데이트가 포함됩니다.
- **네트워크 보안**: 고객은 VPC 분할, 방화벽 규칙, 웹 애플리케이션 방화벽 보호 및 API 관리와 같은 AI 워크로드를 보호하기 위해 네트워크 보안 제어를 구성해야 합니다.
- **애플리케이션 보안**: AI 모델과 상호 작용하는 애플리케이션을 보호하는 것이 중요합니다. 여기에는 입력의 유효성 검사, 안전한 코딩 실시 및 정기적인 보안 테스트가 포함됩니다.
- **데이터 보안 및 거버넌스**: 고객은 AI 교육 및 운영에 사용되는 민감한 데이터를 보호하기 위해 강력한 데이터 보안 조치를 구현해야 합니다. 여기에는 암호화, 데이터 손실 방지(DLP) 및 적절한 데이터 수명 주기 관리가 포함됩니다.
- **로깅 및 모니터링**: 보안 사고를 감지하고 대응하려면 AI 워크로드를 지속적으로 모니터링하는 것이 필수적입니다. 고객은 Google Cloud 로깅, 모니터링 및 SIEM 도구를 활용하여 AI 시스템에 대한 가시성을 확보해야 합니다.
- **사고 대응**: 보안 침해의 영향을 최소화하려면 잘 정의된 사고 대응 계획을 세우는 것이 중요합니다. 이 계획은 보안 사고를 식별, 격리 및 수정하는 단계를 간략하게 설명해야 합니다.

각자의 책임을 이해하고 이행함으로써 Google과 고객은 협력하여 Google Cloud에서 안전하고 신뢰할 수 있는 AI 워크로드를 구축할 수 있습니다.

Google Cloud

AI를 위한 Google의 안전한 집

수년에 걸쳐 Google은 사이버 보안에 대한 개방적이고 협력적인 접근 방식을 채택해 왔습니다. SAIF는 데이터 포이즈닝, 프롬프트 삽입, 모델 도용 및 민감한 데이터 유출과 같은 특정 AI 시스템 위험을 완화하도록 설계되었습니다. Gen AI는 Google Cloud 고객이 혁신적인 기술을 구축하는 방식과 내용 모두에서 변화를 나타냅니다. Google은 Gen AI 보안을 보장하기 위해 보안 및 데이터 보호 분야에서 독보적인 강점을 보유하고 있습니다.

우리의 접근 방식에는 최전선 인텔리전스, 전문 지식 및 혁신을 다른 사람들과 위협 정보를 공유하려는 약속과 결합하여 사이버 공격에 대응하고 예방하는 것이 포함됩니다. 수십 년간의 투자로 인해 Google은 플랫폼 및 컨테이너 보안과 자세 관리, 콘텐츠 악용, 보안 인텔리전스 및 운영, 개발자, AI 및 인간 전문가 지원 등 보안 및 데이터 보호 분야에서 독보적인 강점을 보유하고 있습니다.

고객 모범 사례

Enterprise Foundations Blueprint에 요약된 Google Cloud 보안 모범 사례를 검토하고 구현하는 것이 좋습니다. Enterprise Foundations Blueprint는 Google Cloud 환경의 모든 워크로드에서 일관된 거버넌스, 보안 제어, 확장, 가시성 및 공유 서비스에 대한 액세스를 가능하게 하는 잘 설계된 기반입니다. 해당 문서에 설명된 제어 및 거버넌스를 배포한 후 아래에 요약된 AI 보안 모범 사례 집합을 적용해야 합니다. 이 모범 사례 집합은 AI 시스템에 특별히 적용되는 위협을 완화하는 데 중점을 둡니다.

위에서 설명한 네 가지 광범위한 범주와 일관되게 모델, 애플리케이션, 인프라 및 데이터 보안에 중점을 둡니다.

모델 보안

애플리케이션 및 인프라 보안은 AI 보안의 중요한 구성 요소이지만 AI 모델 자체의 보안을 우선시해야 합니다. AI 모델은 시스템의 핵심 decision-making 엔진이므로 정교한 공격의 주요 대상이 됩니다. 본 문서에서 설명하는 시나리오에서 고객은 모델을 훈련하거나 개선하는 과정(시나리오 섹션 참조) 보다는 모델 보안은 해결해야 할 비즈니스 문제에 모델을 맞추고 무단 액세스, 수정 및 공개로부터 모델을 보호하는 것으로 집중하겠습니다.

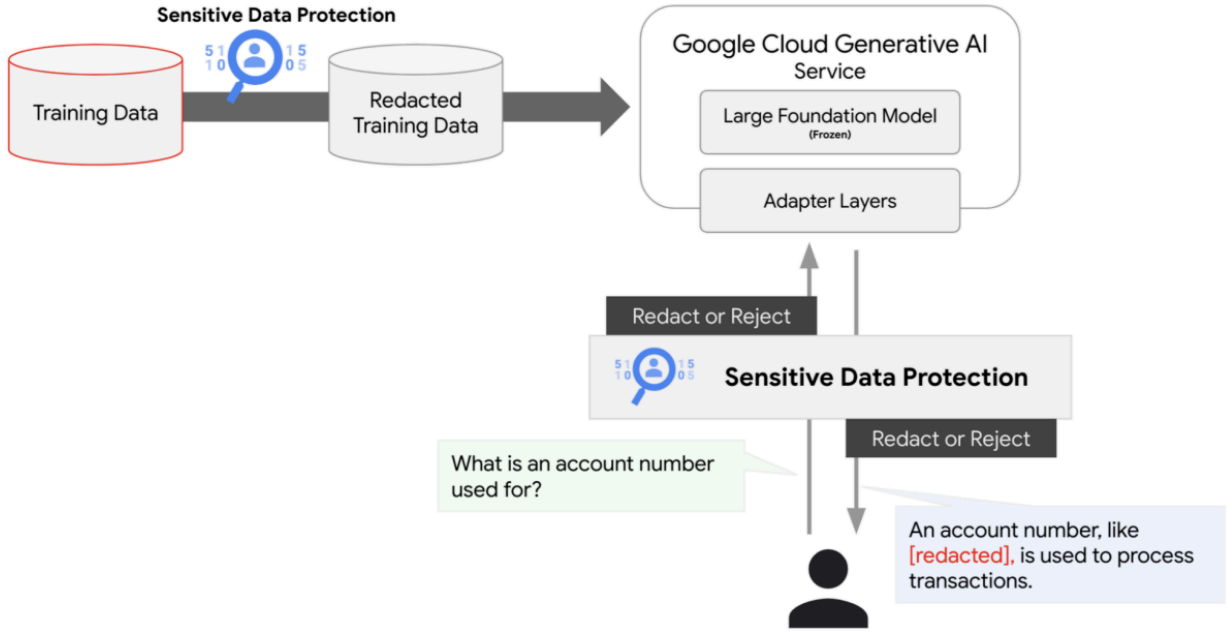
애플리케이션 보안

인라인 프롬프트 및 응답 보호

인라인 프롬프트 및 응답 보호 훈련 파이프라인을 보호하는 것은 중요하지만 SDP에서 제공하는 방어의 일부일 뿐입니다. Gen AI 모델은 사용자로부터 구조화되지 않은 프롬프트를 가져와서 새롭고 보이지 않는 응답을 생성하므로 인라인으로 민감한 데이터를 보호할 수도 있습니다. 많은 알려진 프롬프트 삽입 공격이 실제로 목격되었습니다. 이러한 공격의 주요 목표는 모델을 조작하여 의도하지 않은 정보를 공유하도록

하는 것입니다.

프롬프트 삽입으로부터 보호하는 방법에는 여러 가지가 있지만 SDP는 입력 프롬프트와 생성된 응답을 스캔하여 민감한 요소가 식별되거나 제거되었는지 확인하여 Gen AI 기반 모델로 이동하고 이동하는 데이터에 대한 데이터 중심 보안 제어를 제공할 수 있습니다.



많은 공격자가 민감한 데이터를 유출하기 위해 프롬프트 삽입 기술을 시도할 수 있지만, 또 다른 공격자는 모델을 조작하여 불쾌하거나 오해의 소지가 있거나 위험한 콘텐츠를 생성하려고 할 수 있습니다. AI를 통해 처리된 콘텐츠는 "유해한 범주" 및 민감하다고 간주될 수 있는 주제를 포함하는 안전 속성 목록에 대해 평가됩니다. 기본적으로 이러한 API는 안전 속성 목록과 구성된 차단 임계값을 기반으로 안전하지 않은 콘텐츠를 차단합니다. 각 안전 속성에 대해 다른 임계값을 적용하여 AI 애플리케이션이 수락하거나 생성하는 콘텐츠 유형을 제어할 수 있습니다.

모델 조작 및 회피로부터 보호하기 위해 텍스트 임베딩을 세 번째 방어 계층으로 사용하는 것도 고려해야 합니다. 텍스트 임베딩은 텍스트의 벡터 표현이며 유사한 항목을 찾는 데 여러 가지 방법으로 사용됩니다. 텍스트 임베딩을 생성하면 자연어 텍스트의 벡터 표현을 부동 소수점 숫자 배열로 얻게 되며 모든 입력 텍스트에 숫자 표현이 할당됩니다. 두 텍스트 조각의 벡터 표현 사이의 숫자 거리를 비교하여 애플리케이션은 텍스트 또는 텍스트로 표현된 개체 간의 유사성을 확인할 수 있습니다.

이는 보안 컨텍스트에서 매우 유용합니다. 팀이 공격자가 모델을 조작하기 위해 프롬프트를 작성할 수 있는 방법을 고려하는 데 많은 노력과 주의를 기울였지만 가능한 모든 시도를 식별할 수는 없습니다. 따라서 텍스트 임베딩은 새 프롬프트가 알려진 악성 프롬프트와 유사한지 확인할 수 있으며 방어자는 이 정보를 사용하여 보안 보호 장치를 적용할 수 있습니다.

인프라 보안

AI 보안의 기본 핵심은 이러한 시스템이 작동하는 기본 인프라를 보호하는 데 있습니다. 시스템의 인프라(하드웨어, 네트워크 및 소프트웨어)를 손상시키면 민감한 AI 모델, 훈련 데이터 및 전체 시스템이 유해한 공격 및 조작에 노출될 수 있습니다. 따라서 AI 인프라는 권한이 없는 사용자가 모델에 대한 무단 액세스를 설정할 수 없도록 하고 모델의 무단 도용을 금지하며 무단 모델을 삽입하거나 응답을 손상시킬 수 없도록 해야 합니다.

소프트웨어 공급망

예시 아키텍처에서는 Cloud Build 및 Artifact Registry를 사용하여 소프트웨어 공급망을 보호하여 사용자 지정 컨테이너에서 취약점을 생성하고 찾습니다. 사용자 지정 컨테이너를 사용하면 Vertex AI에서 지원되지 않는 ML 프레임워크, 비 ML 관련 종속성, 라이브러리 및 바이너리를 사용하여 훈련 작업을 실행할 수 있습니다.

Cloud Build로 이미지를 빌드하면 이미지의 빌드 출처가 자동으로 기록됩니다. 빌드 출처는 검증 가능한 데이터 모음이며 빌드된 이미지의 다이제스트, 입력 소스 위치, 빌드 인수 및 빌드 기간과 같은 세부 정보를 포함합니다. 이 빌드 출처를 활용하여 빌드 아티팩트가 신뢰할 수 있는 소스 및 빌더에서 생성되고 있는지 확인하고 빌드 프로세스를 설명하는 출처 메타데이터가 완전하고 신뢰할 수 있는지 확인할 수 있습니다.

실제로 Cloud Build가 이 출처 메타데이터를 생성하지 않으면 빌드가 실패하도록 빌드 구성 파일을 수정할 수 있습니다. 각 빌드에 대해 생성되는 고유한 임시 고객 관리 암호화 키(CMEK)로 build-time persistent disk를 암호화하도록 선택할 수도 있습니다. 빌드가 시작되면 키는 최대 24시간 동안 필요한 빌드 프로세스만 액세스할 수 있습니다. 그런 다음 키는 메모리에서 지워지고 삭제됩니다.

커스텀 이미지가 빌드를 통과하면 통합된 Google Cloud 환경의 일부로 아티팩트를 중앙에 저장하고 종속 항목을 빌드할 수 있는 서비스인 Artifact Registry로 푸시됩니다. Artifact Registry에 이미지를 푸시하기 전에 Container Analysis API를 사용 설정하는 것이 좋습니다. Container Analysis API는 이미지가 Artifact Registry에 푸시되면 자동 취약점 스캔을 시작합니다. 새로운 취약점이 발견되면 취약점 정보가 지속적으로 업데이트되고 Security Command Center에서 제공됩니다.

서빙 인프라

Google Cloud의 웹 애플리케이션 방화벽 서비스인 Cloud Armor는 WAF 및 DDoS 방지 기능을 제공하여 레이어 3, 4 및 7 공격, OWASP(Open Web Application Security Project) Top 10 및 정교한 애플리케이션 악용으로부터 애플리케이션을 보호합니다.

Google Cloud Application Load Balancer는 애플리케이션 API에 도달하려는 트래픽의 첫 번째 진입점입니다. 트래픽이 External Application Load Balancer에 도달하면 Cloud Armor는 즉시 트래픽을 평가하여 네트워크 공격을 탐지하고 완화합니다. 이 평가는 트래픽이 API 및 백엔드 서비스에 도달하기 전에 요청을 허용, 거부, 속도 제한 또는 리디렉션하는 데 적용할 수 있는 사전 구성된 사용자 지정 보안 정책을 기반으로 합니다. Cloud Armor가 트래픽이 합법적이라고 판단하면 사용자는

애플리케이션에 액세스할 수 있습니다.

다중 테넌트 서비스를 격리하여 데이터 유출 위험을 완화할 수 있도록 백엔드 인프라를 VPC Service Controls Perimeter 내에 포함하는 것이 좋습니다. VPC Service Controls Perimeter는 콘솔, 개발자 워크스테이션 및 리소스를 배포하는 데 사용되는 기반 파이프라인을 포함하여 경계 외부에서 시작되는 트래픽에서 제한된 Google Cloud 서비스에 대한 액세스를 거부합니다. 경계가 생성되기 전에 의도한 액세스 경로를 허용하는 경계에 대한 ingress 및 egress 규칙 예외를 설계해야 합니다. 또한 다음을 수행해야 합니다.

- 경계를 적용하기 전에 애플리케이션 중단 없이 API 액세스 위반을 식별하기 위해 드라이 런 모드를 사용합니다.
- 새 프로젝트를 경계에 일관되게 추가하는 프로세스를 설계합니다.
- 개발자가 현재 경계 구성에서 거부된 새로운 사용 사례(use case)를 가질 때 예외를 설계하는 프로세스를 설계합니다.

VPC Service Controls로 **백엔드 인프라를 보호**하고 보안 소프트웨어 **공급망과 위협 탐지 및 예방**을 결합하는 이 **계층화된 접근 방식**을 통해 모델 백도어 삽입, 모델 유출 또는 모델 제공 손상 공격으로 이어질 수 있는 취약점을 탐지하고 수정할 수 있습니다.

Google Cloud

데이터 보안

AI/ML 시스템을 구축하려면 모델을 적절하게 훈련하기 위해 대규모 데이터 코퍼스(corpus)가 필요하며 데이터가 민감하다고 간주되는 경우가 많습니다. 해당 데이터를 적절하게 보호하는 것이 가장 중요하며 **암호화** 및 **익명화 기술**로 데이터 유출 위험을 방지할 수 있습니다.

저장 데이터 암호화

Google Cloud에 저장된 모든 데이터는 Google이 자체 암호화된 데이터에 사용하는 것과 동일한 강화된 키 관리 시스템을 사용하여 저장 시 암호화됩니다. 이러한 키 관리 시스템은 엄격한 키 액세스 제어 및 감사를 제공하고 AES-256 암호화 표준을 사용하여 저장 시 사용자 데이터를 암호화합니다. 설정, 구성 또는 관리가 필요하지 않습니다. 기본 암호화는 조직에 암호화 자료의 규정 준수 또는 지역성과 관련된 특정 요구 사항이 없는 경우 가장 적합한 선택입니다.

그러나 Cloud KMS(Cloud Key Management Service)에서 CMEK를 사용할 수도 있습니다. Cloud KMS는 암호화 키를 관리할 수 있는 서비스입니다. AES-256, RSA 2048, RSA 3072, RSA 4096, EC P256 및 EC P384 암호화 키를 생성, 사용, 순환 및 삭제할 수 있습니다. 고객 관리 암호화 키를 사용하면 키 생성, 키 순환 빈도 및 키 위치를 보다 효과적으로 제어할 수 있습니다. 제어 기능이 추가됨에 따라 책임도 추가됩니다. 따라서 다음을 통해 키를 적절하게 관리하고 제어하는 것이 좋습니다.

- Cloud KMS Admin, CryptoKey Decrypter and Decrypter, KMS Viewer와 같은 사전 정의된 역할을 사용하여 세분화된 액세스 제어를 적용합니다.
- 조직 정책을 설정을 통해 다음을 보장합니다.
    - 폴더, 프로젝트 또는 조직 수준의 모든 리소스에 CMEK가 필요합니다.
    - 특정 키만 AI 워크로드를 암호화하는 데 사용할 수 있습니다.
    - 키 버전은 삭제 예정 전에 비활성화됩니다.
    - 최소 "삭제 예정" 기간을 설정하여 키가 실수로 삭제되지 않도록 보호합니다.
- 암호화 키를 중앙 집중식 Google Cloud 프로젝트에 호스팅합니다.

이러한 추가된 책임으로 인해 기본 암호화로 충분한지 또는 키를 직접 관리하기 위해 Cloud KMS를 사용해야 하는 규정 준수 요구 사항이 있는지 평가하는 것이 좋습니다. 자세한 내용은 저장 시 암호화에 대한 규정 준수 요구 사항을 충족하는 방법을 참조하세요.

민감한 데이터 보호

Sensitive Data Protection 서비스에는 이름, 개인 식별자, 금융 데이터, 의료 컨텍스트 또는 인구 통계 데이터와 같은 민감한 데이터 요소를 신속하게 식별하는 데 도움이 되는 150개 이상의 기본 infoType이 포함되어 있습니다. 이러한 요소를 식별하여 파이프라인에서 제거할 레코드를 선택하거나 인라인 변환을 활용하여 주변 컨텍스트를 유지하면서 민감한 요소만 가릴 수 있습니다. 이렇게 하면 데이터의 유용성을 유지하면서 위험을 줄일 수 있습니다. 인라인 변환은 AI 모델에 대한 훈련 또는 조정 데이터를 준비할 때 사용할 수 있으며 실시간으로 AI 생성 프롬프트 및 응답을 보호할 수도 있습니다.

데이터 준비단계에서의 보호

고객은 prediction 엔드포인트에 AI 모델을 배포하는 경우와 같이 사용자 지정 AI 모델을 훈련하기 위해 데이터 세트를 생성하기 위해 자체 데이터를 자주 사용합니다. 또 다른 일반적인 예로 고객은 자체 데이터를 사용하여 LLM을 파인튜닝하여 모델 출력을 향상시키고 관련 비즈니스 우선 순위를 더 잘 발전시킵니다.

위에서 설명한 튜닝 프로세스는 고객별 데이터 세트를 사용하고 추론시에 사용되는 매개 변수를 생성합니다. 매개 변수는 "frozen" 기반 모델 앞단에 사용자 프로젝트 내부의 있습니다. 이러한 데이터 세트에 민감한 데이터가 포함되지 않도록 조직에서 민감한 데이터 보호 서비스를 사용하여 데이터 세트를 만드는 데 사용된 데이터를 스캔하는 것이 좋습니다. 마찬가지로 Vertex AI 검색에도 이 방법을 사용하여 업로드된 데이터에 민감한 정보가 포함되지 않도록 할 수 있습니다.

로깅, 탐지 및 대응

로깅 및 모니터링

Google Cloud의 로깅은 스토리지, 검색, 분석 및 모니터링 지원을 갖춘 실시간 로그 관리 시스템인 Cloud Logging으로 수행됩니다. 기본적으로 Cloud Logging은 관리 활동 로그, 시스템 이벤트 및 정책 거부 로그를 캡쳐하여 리소스 및 메타데이터 구성 수정과 리소스 액세스 실패 시도에 대한 통찰력을 제공합니다. 이러한 기본 로그와 함께 액세스 투명성을 활성화하여 Google 직원이 지침을 수행하는 동안 오류를 범하지 않았는지 확인하고 법적 또는 규제 의무 준수를 추적하는 로그를 캡처할 수 있도록 하는 것이 좋습니다.

Google Cloud는 AI 애플리케이션과 **최종 사용자와의 상호 작용을 기록하지 않는다는 점**에 유의해야 합니다. 이러한 상호 작용을 로깅하는 것은 애플리케이션 계층에서 수행해야 합니다. Cloud Logging 클라이언트 라이브러리를 사용하여 최종 사용자의 프롬프트와 AI 애플리케이션의 응답을 로깅하는 것이 좋습니다.

보안 및 감사와 관련된 로그를 장기 보존, 분석 및 외부 시스템으로 내보내기 위해 중앙 집중식 프로젝트로 집계하고 저장해야 합니다. 이렇게 하려면 로그 버킷을 생성해야 하며 로그 버킷은 중앙 집중식 프로젝트에 있어야합니다. 이러한 로그 버킷은 보존 기간을 늘리지 않는 한 30일 동안만 로그를 저장한다는 점에 유의해야 합니다. 내부 및 외부 규정 준수 요구 사항에 맞게 보존 기간을 조정하는 것이 좋습니다.

Vertex AI를 위한 탐지

Security Command Center의 보안 상태(security posture) 기능에는 Vertex AI 워크로드를 보호하는 데 도움이 되는 사전 정의된 상태가 포함되어 있습니다. 이러한 정책에는 공개 IP 주소 사용제한, 파일 다운로드, 루트 액세스 및 Vertex AI Workbench 터미널 비활성화가 포함되지만 이에 국한되지는

않습니다. 이러한 상태에는 예를 들어 Vertex AI 모델, 데이터 세트, 엔드포인트, 훈련 파이프라인, 데이터 레이블링 및 사용자 지정 작업에서 CMEK가 비활성화된 경우를 식별하는 Security Health Analytics를 사용하는 탐지 제어가 포함됩니다.

Google Kubernetes Engine의 AI 워크로드 탐지

Google Kubernetes Engine(GKE)에서 AI 워크로드를 실행하는 경우 IaaS(Infrastructure as a Service) 서비스 실행과 관련된 보안 책임으로 인해 더 넓은 공격 표면을 가질 수 있습니다. 그럼에도 불구하고 Security Command Center 기능인 컨테이너 위협 탐지는 컨테이너 최적화 OS 노드 이미지의 상태를 지속적으로 모니터링합니다. 이 서비스는 변경 사항과 원격 액세스 시도를 평가하여 거의 실시간으로 런타임 공격을 탐지합니다. 이러한 공격에는 악성 바이너리, 라이브러리, URL 및 스크립트 실행, 역방향 셸 및 예기치 않은 하위 셸이 포함되지만 이에 국한되지는 않습니다.

탐지 강화

여러 보안 및 개인 정보 보호 제어를 통해 조직은 위협을 방지, 탐지 및 대응할 수 있는 중앙 집중식 위치가 필요합니다. Security Command Center는 AI 워크로드의 보안 상태를 정의, 평가 및 관리할 수 있는 보안 상태 제어 및 위협 경고를 제공합니다.

Cloud Sensitive Data Protection을 사용하여 데이터를 준비하고 인라인 프롬프트 및 응답을 보호하므로 처리하는 데이터 유형을 이해해야 할 수 있습니다. Cloud Sensitive Data Protection 결과를 Security Command Center로 보내 결과를 풍부하게 하는 것이 좋습니다. 이렇게 Security Command Center에 제공하면 보안팀은 더 큰 보안 및 규정 준수 위험을 초래하는 결과의 우선 순위를 더 쉽게 지정하고 특정 문제를 더 잘 해결하기 위해 정보에 입각한 결정을 내릴 수 있습니다.

마찬가지로 데이터 유출에 대한 예방 조치로서의 중요성을 고려하여 VPC 서비스 제어에 대한 사전 정의된 상태를 활용하는 것이 좋습니다. 이러한 사전 정의된 상태는 VM 중첩 가상화, 비활성화된 로그 및 네트워크 라우팅 변경과 같은 네트워크 기반 잘못된 구성 및 위협을 사전에 해결하는 데 도움이 될 수 있습니다.

컨테이너 위협 탐지와 함께 이러한 사전 정의된 사용자 지정 보안 상태는 정의된 벤치마크에서 벗어난 드리프트를 탐지하고 완화하는 데 도움이 될 수 있습니다. Security Command Center는 검토, 필터링 및 대응할 수 있는 결과로 드리프트 인스턴스를 보고합니다.

Google Cloud

AI와 함께 미래를 건설하다

전 세계가 AI의 잠재력에 주목하고 정부와 산업계가 AI의 안전과 보안을 보장하기 위한 규제 접근 방식을 모색하는 가운데, 우리는 AI가 디지털 보안의 변곡점을 나타낸다고 믿습니다. 이 잠재력을 활용하기 위해 Google Cloud에서 AI 기능을 안전하게 배포하기 위해 이 백서에 요약된 모범 사례를 구현하도록 조직에 권장합니다. 추가 정보는 다음 리소스를 검토하는 것이 좋습니다.

- Google의 보안 AI 프레임워크 소개
- AI 보안: 유사점과 차이점
- Gen AI 거버넌스: AI 프로그램 수준을 높이는 10가지 팁
- 조직이 AI 시스템을 보호하는데 레드팀이 중심적인 역할을 하는 이유
- 민감한 데이터 보호가 생성형 AI 워크로드 보안에 어떻게 도움이 되는가

우리는 새로운 기능을 개발하고 배포하면서 안전한 방식으로 AI를 활용하는 데 도움이 되는 연구를 계속 공유하고 방법을 모색할 것입니다. 이는 Google을 고객의 보안팀의 일부로 만들고, 이는 매우 유용한 기술이 모든 사람에게 작동하도록 보장하려는 노력의 중요한 부분입니다.

Google Cloud

부록 A: 모범 사례 체크리스트

안전한 AI 기술 구축은 다른 안전한 소프트웨어 제품 구축과 유사합니다. 둘 다 개발자가 목표를 더 빠르고 효율적으로 달성하고 개발 체인에서 위험으로 변할 수 있는 실수를 줄이는 데 도움이 되도록 설계된 모범 사례를 따라야 합니다.

각 주제 범주에 대한 연결된 모범 사례를 참조하십시오. 여기에는 모델 개발, 애플리케이션 보안, 인프라 및 데이터 관리에 대한 모범 사례가 포함됩니다.

Model

- [ ] AI가 해결할 특정 비즈니스 문제를 정의하고 해당 문제와 관련된 위험을 평가합니다.
- [ ] 무단 액세스, 수정 및 공개로부터 모델을 보호합니다.
- [ ] 보안 소프트웨어 개발 및 소프트웨어 공급망 보안을 기반으로 안전한 모델 개발 사례를 활용합니다.

Application

- [ ] 적대적인 사례를 탐지하고 악의적인 조작을 방지하도록 설계된 입력 및 출력 필터링 및 검증을 구현합니다.
- [ ] AI로 생성된 유해 콘텐츠의 위험을 완화하기 위해 콘텐츠 안전 정책을 시행합니다.
- [ ] 보안 및 AI를 위해 대한 레드팀 테스트를 실행합니다.
- [ ] 잠재적인 실패나 악의적인 사용을 포함하여 AI 관련 문제에 대한 사고 대응 계획을 수립합니다.

Infrastructure

- [ ] AI 워크로드 및 훈련 환경을 포함하여 조직 전체에서 일관된 보안을 보장하기 위해 플랫폼 수준 제어를 조화시킵니다.
- [ ] 최소 권한 원칙을 포함하여 강력한 ID 및 액세스 관리(IAM) 모범사례에 맞춥니다.
- [ ] 위 구성요소로부터 로그를 수집하고 보관합니다.
- [ ] AI에 특정한 공격 및 남용에 대한 탐지 콘텐츠를 개발합니다.

Data

- [ ] 데이터 거버넌스 및 관리 프로세스에 AI를 포함합니다.
- [ ] 저장된 데이터 및 전송 중인 데이터에 강력한 암호화를 사용합니다.
- [ ] 데이터 레이블링 및 출처 추적을 포함하여 AI 데이터 세트에 대한 데이터 거버넌스 프로세스를 확장합니다.
- [ ] 훈련, 튜닝 및 기타 시스템을 둘러싼 컨텍스트에 사용되는 데이터에 대한 비즈니스 위험을 평가합니다.