Investigadores en ciberseguridad han identificado dos fallas de seguridad en la plataforma de aprendizaje automático (ML) Vertex de Google que podrían ser explotadas por atacantes para escalar privilegios y extraer modelos almacenados en la nube.

> *«Al aprovechar los permisos de trabajos personalizados, logramos aumentar nuestros privilegios y acceder de manera no autorizada a todos los servicios de datos del proyecto»*, señalaron Ofir Balassiano y Ofir Shaty, miembros del equipo Unit 42 de Palo Alto Networks, en un informe publicado recientemente.
>
> *«El despliegue de un modelo manipulado en Vertex AI resultó en la extracción de todos los demás modelos ajustados, lo que representa un importante riesgo de exfiltración de datos confidenciales y exclusivos».*

Vertex AI es la plataforma de Google diseñada para entrenar y desplegar modelos personalizados de aprendizaje automático, así como aplicaciones de inteligencia artificial (IA) a gran escala. Fue lanzada en mayo de 2021.

Una característica clave para explotar la vulnerabilidad de escalamiento de privilegios es Vertex AI Pipelines, una herramienta que permite a los usuarios automatizar y supervisar flujos de trabajo de MLOps, utilizados para entrenar y ajustar modelos de aprendizaje automático mediante trabajos personalizados.

El equipo de Unit 42 descubrió que manipulando los pipelines de trabajos personalizados, es posible escalar privilegios y obtener acceso a recursos normalmente restringidos. Esto se logra al crear un trabajo personalizado que ejecuta una imagen especialmente diseñada para iniciar un shell inverso, permitiendo el acceso remoto no autorizado al entorno.

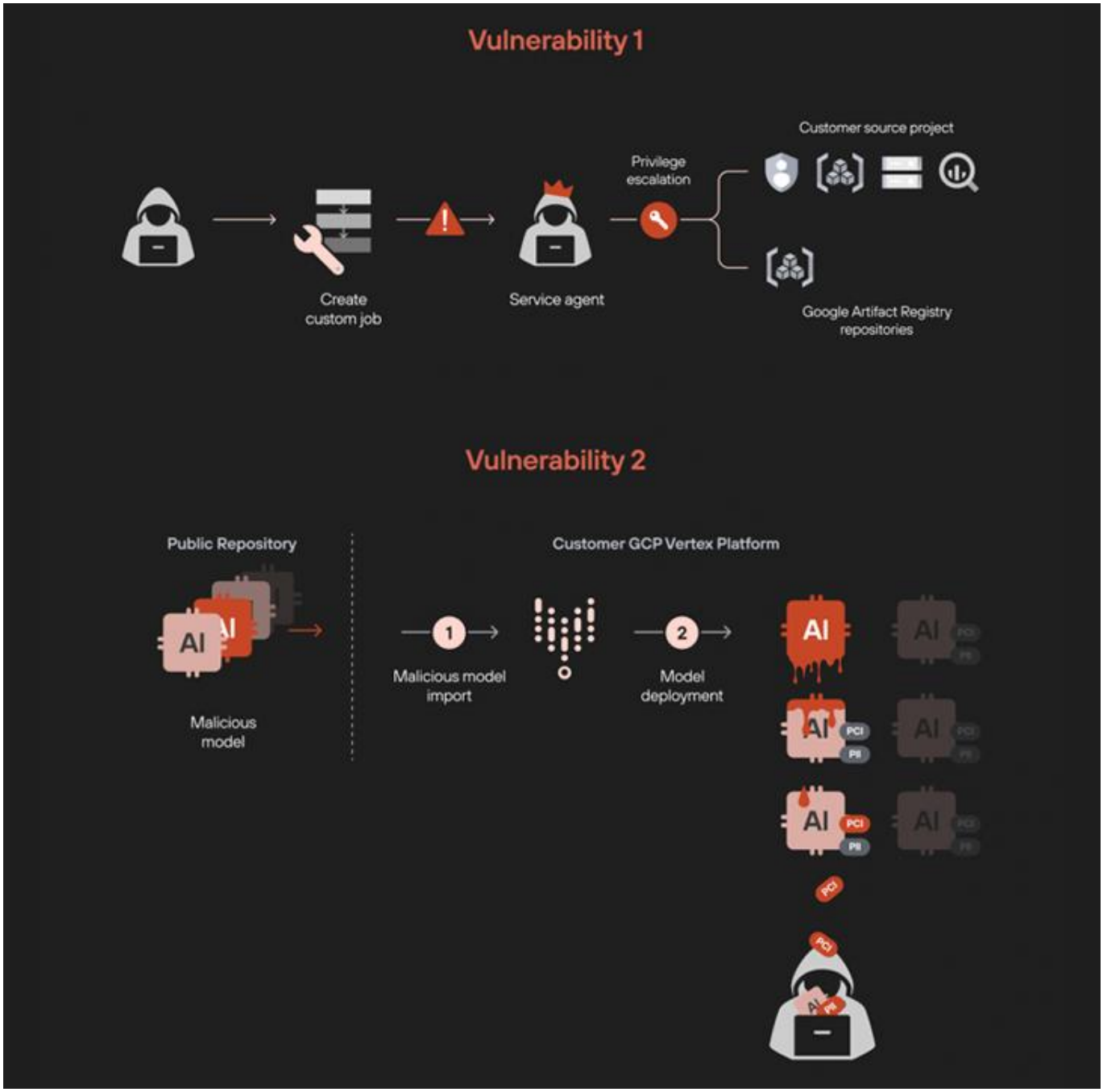
Según los investigadores, estos trabajos personalizados se ejecutan en un proyecto del inquilino utilizando una cuenta de agente de servicio con permisos extensivos, como listar todas las cuentas de servicio, gestionar buckets de almacenamiento y acceder a tablas de

BigQuery. Dichos privilegios podrían ser explotados para ingresar a repositorios internos de Google Cloud y descargar imágenes.

La segunda vulnerabilidad identificada consiste en desplegar un modelo malicioso en un proyecto del inquilino, lo que genera un shell inverso cuando se ejecuta en un endpoint. Esto aprovecha los permisos de solo lectura de la cuenta de servicio «custom-online-prediction» para explorar clústeres de Kubernetes y recuperar sus credenciales, permitiendo ejecutar comandos arbitrarios con kubectl.

> *«Este proceso nos permitió movernos del entorno de Google Cloud Platform (GCP) al de Kubernetes. El movimiento lateral fue posible debido a la vinculación de permisos entre GCP y Google Kubernetes Engine (GKE) a través de IAM Workload Identity Federation«,* explicaron los investigadores.

El análisis también reveló que, mediante este acceso, se puede visualizar la imagen creada en el clúster de Kubernetes y obtener su digest único, lo que permite extraer las imágenes del contenedor utilizando crictl y el token de autenticación asociado a la cuenta de servicio *«custom-online-prediction»*.



Además, un modelo malicioso podría ser explotado para acceder y exportar todos los modelos de lenguaje grande (LLMs) y sus adaptadores ajustados de manera similar.

Esto podría tener un impacto significativo si un desarrollador, sin darse cuenta, implementa un modelo comprometido en un repositorio público, lo que permitiría al atacante sustraer todos los modelos de aprendizaje automático y LLM afinados. Tras la notificación responsable de estas vulnerabilidades, Google tomó medidas para solucionarlas.

> *«Esta investigación subraya cómo la implementación de un único modelo malicioso podría poner en riesgo todo un ecosistema de inteligencia artificial. Un atacante podría aprovechar un modelo no verificado en un sistema en producción para obtener datos sensibles, lo que resultaría en ataques graves de extracción de modelos»*, señalaron los expertos.

Se recomienda a las organizaciones establecer controles rigurosos sobre la implementación de modelos y supervisar los permisos necesarios para desplegarlos en proyectos compartidos.

Este descubrimiento coincide con el anuncio de la Red de Investigación 0Day de Mozilla (0Din), que demostró que es posible interactuar con el entorno de sandbox interno de OpenAI ChatGPT («/home/sandbox/.openai_internal/») mediante comandos. Esto permite subir y ejecutar scripts en Python, mover archivos e incluso acceder al plan de acción del LLM.

Dicho esto, OpenAI considera que estas interacciones son comportamientos previstos, ya que la ejecución del código está confinada dentro del entorno del sandbox y no debería extenderse más allá de él.

> *«Para aquellos interesados en explorar el entorno sandbox de OpenAI ChatGPT, es fundamental comprender que la mayoría de las acciones posibles dentro de este entorno están diseñadas como características intencionales, no como fallas de seguridad»,* explicó Marco Figueroa, investigador en seguridad.

> *«Actividades como extraer información, cargar archivos, ejecutar comandos bash o scripts de Python en el sandbox son válidas, siempre y cuando no superen los límites establecidos por el contenedor».*