Investigadores de ciberseguridad han identificado casi dos docenas de vulnerabilidades de seguridad en 15 proyectos de código abierto relacionados con el aprendizaje automático (ML).

Estos problemas incluyen vulnerabilidades detectadas tanto en el servidor como en el cliente, según un análisis publicado la semana pasada por la empresa JFrog, especializada en seguridad de la cadena de suministro de software.

Las debilidades en el servidor *«permiten a los atacantes tomar control de servidores importantes en la organización, como registros de modelos de ML, bases de datos de ML y pipelines de ML»*, afirmó la compañía.

Las vulnerabilidades, encontradas en Weave, ZenML, Deep Lake, Vanna.AI y Mage AI, se han clasificado en categorías generales que permiten el secuestro remoto de registros de modelos, frameworks de bases de datos de ML y la toma de control de pipelines de ML.

A continuación, una descripción breve de cada una de las fallas encontradas:

- CVE-2024-7340 (Puntuación CVSS: 8.8): Vulnerabilidad de recorrido de directorios en la herramienta Weave ML, que permite leer archivos en todo el sistema, lo que habilita a un usuario autenticado con pocos privilegios a escalar sus permisos al rol de administrador al leer un archivo llamado «api_keys.ibd» (solucionado en la versión 0.50.8).
- Vulnerabilidad de control de acceso incorrecto en el framework de MLOps ZenML que permite a un usuario con acceso a un servidor ZenML administrado aumentar sus permisos de «visor» a permisos completos de administrador, obteniendo la capacidad de modificar o leer el Secret Store (sin identificador CVE).
- CVE-2024-6507 (Puntuación CVSS: 8.1): Vulnerabilidad de inyección de comandos en la base de datos Deep Lake orientada a IA, que permite a los atacantes inyectar comandos del sistema al cargar un conjunto de datos remoto de Kaggle debido a una sanitización inadecuada de entradas (solucionado en la versión 3.9.11).
- CVE-2024-5565 (Puntuación CVSS: 8.1): Vulnerabilidad de inyección de prompt en la

biblioteca Vanna.AI, que puede ser explotada para ejecutar código remoto en el host subyacente.

- CVE-2024-45187 (Puntuación CVSS: 7.1): Vulnerabilidad de asignación de privilegios incorrecta que permite a usuarios invitados en el framework Mage AI ejecutar código arbitrario de forma remota a través del terminal de Mage AI, ya que se les otorgan privilegios elevados y permanecen activos durante 30 días de manera predeterminada, incluso después de ser eliminados.
- CVE-2024-45188, CVE-2024-45189 y CVE-2024-45190 (Puntuación CVSS: 6.5): Varias vulnerabilidades de recorrido de rutas en Mage AI que permiten a usuarios remotos con el rol de «visor» leer archivos de texto arbitrarios del servidor Mage a través de las solicitudes de «File Content», «Git Content» y «Pipeline Interaction», respectivamente.

> *«Dado que los pipelines de MLOps pueden acceder a los conjuntos de datos de ML, a la capacitación de modelos de ML y a la publicación de modelos de ML, explotar un pipeline de ML puede llevar a una violación de seguridad extremadamente grave»,* indicó JFrog.
>
> *«Cualquiera de los ataques mencionados en este artículo (como la inclusión de puertas traseras en modelos de ML, la alteración de datos de ML, etc.) puede ser llevado a cabo por un atacante, dependiendo del acceso del pipeline de MLOps a estos recursos».*

La divulgación llega más de dos meses después de que la empresa identificara más de 20 vulnerabilidades que pueden ser aprovechadas para atacar plataformas de MLOps.

Además, coincide con la presentación de un marco de defensa llamado Mantis, que utiliza la inyección de prompt como una manera de contrarrestar ataques cibernéticos en modelos de lenguaje de gran tamaño (LLMs), logrando más del 95% de efectividad.

> *«Al detectar un ataque cibernético automatizado, Mantis inserta entradas cuidadosamente diseñadas en las respuestas del sistema, provocando que el modelo de lenguaje del atacante interrumpa sus propias operaciones (defensa pasiva) o incluso comprometa la máquina del atacante (defensa activa)»*, comentó un grupo de académicos de la Universidad George Mason.
>
> *«Al desplegar servicios de señuelo vulnerables intencionalmente para atraer al atacante y utilizando inyecciones de prompt dinámicas para el modelo de lenguaje del atacante, Mantis puede realizar un contraataque de forma autónoma».*