Un grupo de académicos demostró nuevos ataques que aprovechan los modelos Text-to-SQL para producir código malicioso que podría permitir a los atacantes obtener información confidencial y realizar ataques de denegación de servicio (DoS).

> *«Para interactuar mejor con los usuarios, una amplia gama de aplicaciones de bases de datos emplean técnicas de IA que pueden traducir preguntas humanas en consultas SQL (a saber, Text-to-SQL)»*, dijo Xutan Peng, investigador de la Universidad de Sheffield.

> *«Descubrimos que al hacer algunas preguntas especialmente diseñadas, los crackers pueden engañar a los modelos Text-to-SQL para producir código malicioso. Como dicho código se ejecuta automáticamente en la base de datos, la consecuencia puede ser bastante grave (por ejemplo, violaciones de datos y ataques DoS)».*

Los hallazgos, que se validaron con dos soluciones comerciales BAIDU-UNIT y AI12sql, marcan la primera instancia empírica en la que los modelos de procesamiento de lenguaje natural (NLP) se han explotado como un vector de ataque en la naturaleza.

Los ataques de caja negra son análogos a las vulnerabilidades de inyección SQL en las que la incrustación de una carga maliciosa en la pregunta de entrada se copia en la consulta SQL construida, lo que genera resultados inesperados.

Las cargas útiles especialmente diseñadas, según el estudio, podrían armarse para ejecutar consultas SQL maliciosas que podrían permitir que un atacante modifique las bases de datos back-end y realizar ataques DoS contra el servidor.

Además, una segunda categoría de ataques exploró la posibilidad de corromper varios modelos de lenguaje preentrenados (PLM), modelos que han sido entrenados con un gran

conjunto de datos mientras se mantienen agnósticos a los casos de uso en los que se aplican, para desencadenar la generación de comandos maliciosos basados en ciertos factores desencadenantes.

> *«Hay muchas formas de plantar puertas traseras en marcos basados en PLM al envenenar las muestras de capacitación, como hacer sustituciones de palabras, diseñar indicaciones especiales y alterar los estilos de las oraciones»,* explicaron los investigadores.

Los ataques de puerta trasera en cuatro modelos de código abierto diferentes (BART-BASE, BART-LARGE, T5-BASE y T5-3B) utilizando un corpus envenenado con muestras maliciosas lograron una tasa de éxito del 100% con poco impacto perceptible en el rendimiento, lo que hace que los problemas sean difíciles de detectar en el mundo real.

Como medidas de mitigación, los investigadores sugieren incorporar clasificadores para verificar cadenas sospechosas en las entradas, evaluar modelos listos para usar para prevenir amenazas en la cadena de suministro y adherirse a buenas prácticas de ingeniería de software.