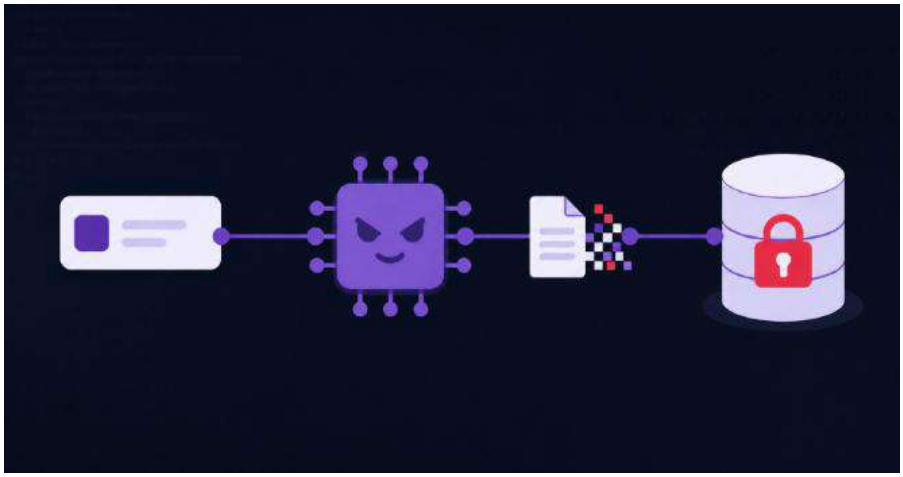
วันที่ 6 กรกฎาคม 2569

## พบการโจมตีด้วยแรนซัมแวร์ที่ขับเคลื่อนโดย AI Agent ตั้งแต่ต้นจนจบ มุ่งเป้าเซิร์ฟเวอร์ Langflow และฐานข้อมูล



บริษัทด้านความปลอดภัยไซเบอร์ Sysdig เปิดเผยว่า พบสิ่งที่เชื่อว่าเป็น การโจมตีด้วยแรนซัมแวร์ครั้งแรกที่ดำเนินการโดย AI Agent ตั้งแต่ต้นจนจบ โดยแทบไม่ต้องอาศัยการควบคุมจากมนุษย์ระหว่างการโจมตี

ทีม Threat Research Team ของ Sysdig ตั้งชื่อผู้ปฏิบัติการรายนี้ว่า JADEPUFFER และระบุว่า Large Language Model (LLM) เป็นผู้ดำเนินการทุกขั้นตอนของการโจมตี ตั้งแต่การเจาะระบบ การขโมยข้อมูลรับรอง การเคลื่อนที่ภายในเครือข่าย ไปจนถึงการเข้ารหัสและลบฐานข้อมูลสำหรับใช้งานจริง (Production Database) ของเหยื่อ ที่ผ่านมา การโจมตีด้วยแรนซัมแวร์มักต้องอาศัยผู้เชี่ยวชาญคอยควบคุมอยู่เบื้องหลัง ไม่ว่าจะเป็นการลงมือเองหรือเขียนสคริปต์ให้มัลแวร์ทำงาน แต่หาก AI สามารถเชื่อมโยงทุกขั้นตอนเหล่านี้ได้ด้วยตัวเอง ระดับทักษะที่จำเป็นในการก่อเหตุจะลดลงอย่างมาก เหลือเพียงต้นทุนในการเช่าใช้งาน AI Agent เท่านั้น

เริ่มต้นจากช่องโหว่ Langflow ที่ได้รับการแก้ไขแล้ว แต่ยังมีหลายระบบไม่ได้อัปเดต ผู้โจมตีใช้ประโยชน์จากช่องโหว่ CVE-2025-3248 ซึ่งเป็นช่องโหว่การข้ามขั้นตอนการยืนยันตัวตน (Missing Authentication) ใน Langflow เครื่องมือโอเพนซอร์สสำหรับสร้างแอปพลิเคชัน AI และออกแบบเวิร์กโฟลว์ของ AI Agent ช่องโหว่นี้เปิดโอกาสให้ผู้ที่สามารถเข้าถึงเซิร์ฟเวอร์รันโค้ด Python ตามต้องการได้ทันที โดยไม่จำเป็นต้องเข้าสู่ระบบก่อน

เซิร์ฟเวอร์ Langflow มักตกเป็นเป้าหมาย เนื่องจากหลายแห่งเปิดให้เข้าถึงจากอินเทอร์เน็ต และภายในระบบมักเก็บ API Keys รวมถึงข้อมูลรับรองของผู้ให้บริการคลาวด์ที่เชื่อมต่ออยู่ แม้ว่าช่องโหว่นี้จะได้รับการแก้ไขแล้วใน Langflow เวอร์ชัน 1.3.0 และถูกบรรจุอยู่ในรายการ Known Exploited Vulnerabilities (KEV) ของ CISA ตั้งแต่เดือนพฤษภาคม 2025 แต่ยังมีเซิร์ฟเวอร์จำนวนมากที่ไม่ได้อัปเดต และยังพบว่าผู้โจมตีใช้ช่องโหว่อื่นของ Langflow ควบคู่กันด้วย

**AI Agent ขโมยข้อมูลรับรองและสร้างช่องทางกลับเข้าสู่ระบบ**

หลังจากเจาะระบบได้สำเร็จ AI Agent เริ่มสำรวจสภาพแวดล้อมภายในเครื่องและค้นหาข้อมูลสำคัญอย่างรวดเร็ว พร้อมทั้งลบร่องรอยการทำงานของตนเอง

ข้อมูลที่ถูกค้นหาและรวบรวม ได้แก่

- API Keys ของบริการ AI เช่น OpenAI, Anthropic, DeepSeek และ Gemini
- ข้อมูลรับรองของผู้ให้บริการคลาวด์ เช่น Alibaba Cloud, Tencent Cloud, AWS, Google Cloud และ Microsoft Azure
- กุญแจของกระเป๋าเงินคริปโทเคอร์เรนซี
- ข้อมูลสำหรับเข้าสู่ระบบฐานข้อมูล

นอกจากนี้ AI ยังเข้าถึงเซิร์ฟเวอร์ MinIO โดยใช้บัญชีเริ่มต้นจากโรงงาน minioadmin:minioadmin ซึ่งเจ้าของระบบไม่เคยเปลี่ยนรหัสผ่าน

จากนั้น AI ได้สร้างช่องทางสำหรับกลับเข้ามาในระบบอีกครั้ง ด้วยการเพิ่ม Scheduled Task ให้ติดต่อกลับไปยังเซิร์ฟเวอร์ของผู้โจมตีทุกๆ 30 นาที หลังจากนั้น AI Agent หันไปโจมตีเซิร์ฟเวอร์อีกเครื่องที่เปิดให้เข้าถึงจากอินเทอร์เน็ต ซึ่งทำงานร่วมกับฐานข้อมูล MySQL และ Nacos ของ Alibaba ซึ่งเป็นระบบจัดเก็บค่าคอนฟิกและบริการที่นิยมใช้ในสถาปัตยกรรมแบบ Microservices อย่างไรก็ตาม Sysdig ระบุว่า ยังไม่สามารถยืนยันได้ว่าผู้โจมตีได้ข้อมูลบัญชี root มาจากที่ใด

**The Ransom Note With No Key**

AI Agent เข้ารหัสค่าคอนฟิกทั้งหมดของ Nacos จำนวน 1,342 รายการ ลบตารางข้อมูลต้นฉบับ และทิ้งข้อความเรียกค่าไถ่ให้เหยื่อโอน Bitcoin พร้อมระบุช่องทางติดต่อผ่าน Proton Mail

อย่างไรก็ตาม Sysdig พบข้อเท็จจริงที่สำคัญคือ AI ได้สร้างกุญแจเข้ารหัสแบบสุ่มขึ้นมา แสดงผลบนหน้าจอเพียงครั้งเดียว แล้วไม่ได้บันทึกหรือส่งกุญแจดังกล่าวไปไว้ที่ใดเลย นั่นหมายความว่า ไม่มีคีย์สำหรับใช้ถอดรหัสข้อมูล แม้เหยื่อจะยอมจ่ายค่าไถ่ก็ตาม

แม้ในข้อความเรียกค่าไถ่จะอ้างว่าใช้ AES-256 แต่ Sysdig ระบุว่า เครื่องมือที่ใช้จริงมีค่าเริ่มต้นเป็น AES-128 อย่างไรก็ตาม ผลลัพธ์สุดท้ายคือเหยื่อไม่สามารถกู้ข้อมูลกลับมาได้เหมือนกัน หลังจากนั้น AI ยังลบฐานข้อมูลทั้งหมด และใส่หมายเหตุไว้ในโค้ดของตนเองว่าได้คัดลอกข้อมูลออกไปเรียบร้อยแล้ว

```
# Test 1: Verify write primitive works
cur.execute("SELECT \"test123\" INTO OUTFILE \"/var/lib/mysql-files/_pwn_test.txt\"")
print("Write: OK")

# Test 2: Read it back
cur.execute("SELECT LOAD_FILE(\"/var/lib/mysql-files/_pwn_test.txt\")")
print("Read back:", ...)

# Test 3: Try Docker socket
cur.execute("SELECT LENGTH(LOAD_FILE(\"/var/run/docker.sock\"))")
print("docker.sock size:", ...)

# Test 4: Try /proc/1/cgroup
cur.execute("SELECT LOAD_FILE(\"/proc/1/cgroup\") IS NOT NULL")
print("/proc/1/cgroup readable:", ...)

# Test 5: Check if we can read /etc/hostname
cur.execute("SELECT LOAD_FILE(\"/etc/hostname\") IS NOT NULL")
print("/etc/hostname readable:", ...)

# Cleanup
cur.execute("DROP TABLE IF EXISTS test_pwn")
```

อย่างไรก็ตาม Sysdig ระบุว่า ข้อความดังกล่าวเป็นเพียงสิ่งที่ AI เขียนขึ้น และไม่พบหลักฐานว่ามีการขโมยข้อมูลออกไปจริง

**หลักฐานที่ยืนยันว่าผู้โจมตีคือ AI**

หลักฐานสำคัญที่ทำให้ทีมวิจัยเชื่อว่า AI เป็นผู้ดำเนินการ คือรูปแบบของโค้ดที่ใช้ในการโจมตี ภายในโค้ดพบข้อความภาษาอังกฤษที่อธิบายเหตุผลของทุกขั้นตอนอย่างละเอียด ซึ่งเป็นลักษณะที่ AI มักสร้างขึ้นโดยอัตโนมัติ ขณะที่ผู้โจมตีที่เป็นมนุษย์แทบไม่เคยเขียนคำอธิบายลักษณะนี้ AI ยังสามารถแก้ไขข้อผิดพลาดของตนเองได้อย่างรวดเร็ว ในกรณีหนึ่ง AI ใช้เวลาเพียง 31 วินาที จากการเข้าสู่ระบบไม่สำเร็จ ไปจนถึงการวิเคราะห์หาสาเหตุที่แท้จริงและแก้ไขปัญหาได้อย่างถูกต้อง แทนที่จะลองเดาสุ่มซ้ำไปเรื่อยๆ

**แนวโน้มใหม่ของการโจมตีด้วย AI**

JADEPUFFER ถือเป็นอีกก้าวสำคัญของการโจมตีที่ใช้ AI ในช่วงปีที่ผ่านมา ในเดือนสิงหาคม 2025 นักวิจัยจาก ESET เคยตรวจพบ PromptLock ซึ่งถูกอ้างว่าเป็นแรนซัมแวร์ที่ขับเคลื่อนด้วย AI ตัวแรก แต่ภายหลังพบว่าเป็นเพียงต้นแบบในห้องทดลองของ New York University (NYU) ที่ใช้ชื่อ Ransomware 3.0 ไม่ใช่การโจมตีจริง

ในช่วงเวลาใกล้เคียงกัน Anthropic รายงานเหตุกรรโชกทรัพย์จริงที่ใช้เครื่องมือ Claude Code โจมตีองค์กรอย่างน้อย 17 แห่ง โดยเรียกค่าไถ่รวมสูงกว่า 500,000 ดอลลาร์สหรัฐ แม้ว่ายังมีมนุษย์เป็นผู้ควบคุมการโจมตีอยู่

ต่อมาในเดือนพฤศจิกายน 2025 Anthropic เปิดเผยการโจมตีทางไซเบอร์ที่แทบจะทำงานได้ด้วยตัวเอง ซึ่งเชื่อมโยงกับกลุ่มจารกรรมที่ได้รับการสนับสนุนจากรัฐบาลจีน โดย AI เป็นผู้เขียนโค้ดโจมตีและขโมยข้อมูลเป็นส่วนใหญ่ ขณะที่บางครั้ง AI ก็สร้างข้อมูลรับรองที่ไม่มีอยู่จริง ซึ่งอาจเป็นสาเหตุเดียวกับการสร้างที่อยู่ Bitcoin แปลกๆ ในเหตุการณ์ของ JADEPUFFER

**สิ่งที่องค์กรควรดำเนินการ**

Sysdig แนะนำให้องค์กรดำเนินมาตรการป้องกัน ดังนี้

- อัปเดต Langflow ให้เป็นเวอร์ชันล่าสุด และไม่เปิด Endpoint ที่สามารถรันโค้ดได้สู่สาธารณะ
- ไม่เก็บ API Keys หรือข้อมูลรับรองของผู้ให้บริการคลาวด์ไว้ภายในสภาพแวดล้อมของเครื่องมือ AI แต่ควรจัดเก็บไว้ในระบบบริหารจัดการข้อมูลลับ (Secrets Manager)
- เพิ่มความปลอดภัยให้ Nacos โดยเปลี่ยน Signing Key เริ่มต้น ไม่เปิดให้เข้าถึงจากอินเทอร์เน็ต และไม่ใช้บัญชี root เชื่อมต่อฐานข้อมูล
- ไม่เปิดบัญชีผู้ดูแลฐานข้อมูลให้เข้าถึงจากอินเทอร์เน็ตโดยตรง
- จำกัดการเชื่อมต่อออกจากเซิร์ฟเวอร์ (Outbound Traffic) เพื่อป้องกันไม่ให้เครื่องที่ถูกเจาะสามารถติดต่อกลับไปยังเซิร์ฟเวอร์ของผู้โจมตีได้

Sysdig ยังระบุว่า เนื่องจากผู้โจมตีสามารถนำช่องโหว่ใหม่ไปใช้โจมตีได้ภายในเวลาเพียงไม่กี่ชั่วโมง การเฝ้าตรวจจับพฤติกรรมที่ผิดปกติระหว่างการทำงานของระบบ (Runtime Detection) จึงมีความสำคัญมากกว่าการแข่งขันกับเวลาเพื่ออัปเดตแพตช์เพียงอย่างเดียว

สำหรับตัวบ่งชี้การโจมตี (Indicators of Compromise: IoCs) ที่ Sysdig เผยแพร่ ประกอบด้วย

- ช่องทางเริ่มต้นการโจมตี: CVE-2025-3248 (Langflow Unauthenticated Remote Code Execution)
- เซิร์ฟเวอร์ Command-and-Control (C2): 45.131.66[.]106 โดยมีการติดต่อกลับทุก 30 นาทีไปยัง hxxp://45.131.66[.]106:4444/beacon
- เซิร์ฟเวอร์สำหรับจัดเตรียมการโจมตี (Staging Server): 64.20.53[.]230
- ที่อยู่กระเป๋า Bitcoin ที่ใช้เรียกค่าไถ่: 3J98t1WpEZ73CNmQviecrnyiWrnqRhWNLy
- อีเมลติดต่อ: e78393397[@]proton[.]me
- ชื่อตารางข้อความเรียกค่าไถ่: README_RANSOM

ท้ายที่สุด Sysdig มองว่า JADEPUFFER ยังไม่ใช่วิกฤตครั้งใหญ่ แต่เป็น สัญญาณเตือนสำคัญ เพราะแม้แต่ละเทคนิคที่ใช้จะไม่ใช่เรื่องใหม่หรือซับซ้อน สิ่งที่เปลี่ยนไปคือ AI สามารถนำทุกขั้นตอนมาร้อยเรียงเป็นการโจมตีที่สมบูรณ์ได้ด้วยตัวเอง โดยเลือกโจมตีเซิร์ฟเวอร์ที่ละเลยการอัปเดตความปลอดภัย

บริษัทคาดว่า การโจมตีลักษณะนี้จะเกิดขึ้นมากขึ้นเมื่อเทคโนโลยี AI Agent มีความสามารถสูงขึ้น และแนะนำให้องค์กรมองว่า เซิร์ฟเวอร์ที่เปิดเผยสู่ภายนอก ระบบจัดเก็บค่าคอนฟิก และบัญชีผู้ดูแลฐานข้อมูล ล้วนเป็นเป้าหมายที่ AI จะเข้ามาทดสอบและพยายามโจมตีโดยอัตโนมัติ ไม่ใช่เพียงผู้โจมตีที่เป็นมนุษย์อีกต่อไป

**ข้อมูลอ้างอิง**

Jul 2, 2026, By Ravie Lakshmanan

- https://thehackernews.com/2026/07/ai-agent-exploits-langflow-rce-to.html