File fingerprinting of the ZIP format for identifying and tracking provenance

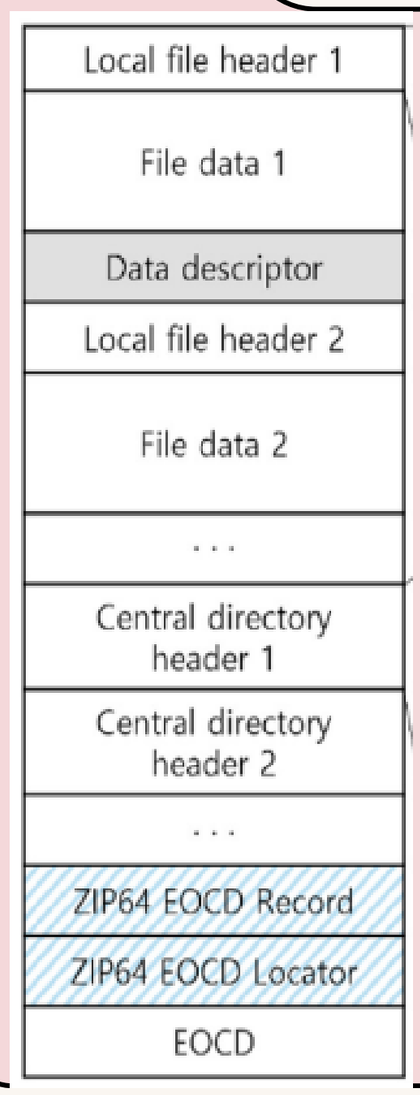
Nhóm 8: Nguyễn Văn Anh Tú, Lâm Hải Đăng, Nguyễn Đình Kha, Trương Long Hưng

Lớp: NT334.P11.ATCL
GVHD: ThS. Lê Đức Thịnh

Giới thiệu

ZIP là định dạng file lưu trữ nén dữ liệu mà không làm mất dữ liệu. File ZIP có thể chứa một hoặc nhiều file, folder được nén lại. File ZIP, dù tuân theo một cấu trúc chung, vẫn tồn tại những chi tiết khác biệt phụ thuộc vào hệ điều hành và ứng dụng tạo tệp. Những khác biệt này tạo ra dấu vân tay tệp (file fingerprint) độc nhất, giúp xác định môi trường tạo tệp, hành vi người dùng và phát hiện khả năng chỉnh sửa trái phép. Bài báo đề xuất có thể tìm ra được môi trường tạo và chỉnh sửa tệp bằng cách phân tích cấu trúc chi tiết của tệp ZIP và so sánh các đặc điểm giải nén với các ứng dụng được cài đặt trên hệ thống

Cấu trúc của file ZIP



File ZIP có chữ ký bắt đầu với (0 x 50 4B 03 04)
Cấu trúc tổng thể gồm:

- Local File Header
- Central Directory Header
- End of Central Directory (EOCD)

Mỗi header chứa một trường bổ sung (Extra field) khác nhau
Cấu trúc của tệp ZIP phản ánh đặc điểm của hệ điều hành và ứng dụng. Do đó, nếu các tệp ZIP được tạo bởi cùng một ứng dụng nhưng trên các hệ điều hành khác nhau, chúng sẽ có dấu vân tay tệp khác nhau.

Nhận dạng file ZIP

1) Phân biệt thời gian tạo tệp ZIP
2) Xác định UID và GID
3) Phân tích mã hóa tên tệp ZIP
4) Bộ phân loại kết hợp các đặc điểm của hệ điều hành và ứng dụng thông qua phân tích cấu trúc file ZIP
5) Kiểm tra các đặc điểm nổi bật như mô tả dữ liệu và các trường bổ sung trong tiêu đề

Thực nghiệm

Sử dụng công cụ như FTK Imager trên Windows và zipdetails trên Linux để phân tích cấu trúc và đặc điểm của file ZIP

Kết luận và hướng phát triển

Kết luận:

- ZIP fingerprinting hiệu quả trong việc xác định nguồn gốc file.
- Hỗ trợ pháp chứng số và điều tra.

Hướng phát triển:

- Phân tích thêm các định dạng nén khác (OOXML, JAR).
- Tăng độ chính xác và mở rộng dataset.