**BÁO CÁO ĐỊNH KỲ TUẦN 1**

**Chủ đề: Thám mã tệp tin ZIP**

**Chương 1: Tổng quan về tệp tin Zip**

1. **Định nghĩa và cấu tạo cấu trúc tệp tin**

**ZIP** là một tên gọi của một định dạng nén được phát triển bởi **PKWARE** vào năm 1989, với mục tiêu chính là giảm kích thước tệp và lưu trữ nhiều tệp trong một tệp được duy nhất.

Về mặt lợi ích, người dùng có thể gửi dữ liệu tệp nén qua email hoặc tải dữ liệu lên các dịch vụ tiện ích đám mây thuận lợi, tiết kiệm được băng thông và thời gian truyền tải bởi tệp đã được giảm dung lượng. Một ưu điểm của tệp ZIP nữa đáng nói tới đó là định dạng ZIP còn hỗ trợ tính năng mã hóa, giúp người dùng có thể bảo vệ tệp tin của mình thông qua mật khẩu được tạo, theo cách này chỉ những người biết mật khẩu mới có thể unzip được file, đảm bảo sự bảo mật của thông tin qua quá trình chia sẻ hoặc lưu trữ tệp.

Ban đầu, việc nén thông tin nhằm hướng tới mục đích giảm thiểu giảm thiểu kích thước dữ liệu, từ đó giúp trao đổi chúng được hiệu quả. Nhưng trong quá trình đó, việc bảo mật thông tin là không thể thiếu nên những công cụ nén phổ biến như WinZip hay WinRar thường tích hợp thêm khả năng mã hóa thông tin nén (thường là sử dụng những hệ mã đối xứng phổ biến như DES hay AES).

Để thuận tiện cho người dùng, khóa bí mật cho các hệ mã này được sinh ra từ một mật khẩu do người gửi nhập vào thông qua một hàm băm, sau đó được sử dụng để mã hóa tài liệu. Sau đó mật khẩu được người dùng chuyển đến cho người nhận thông qua một kênh an toàn và được người nhận sử dụng mã này để sinh khóa cho quá trình giải mã.

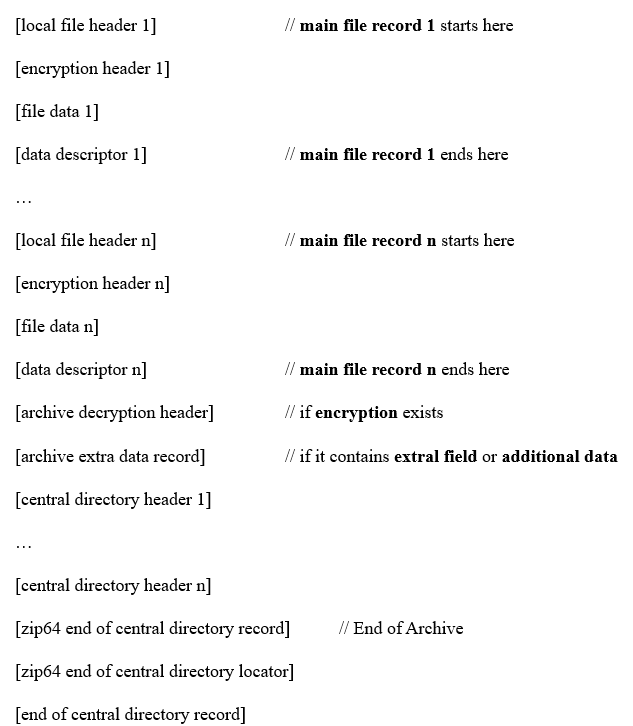
Quá trình mã hóa và giải mã hóa tệp nén ZIP được mô tả như sau:

A diagram of a computer system

Description automatically generated with medium confidence

Hình 1: Mã hóa, giải mã hóa tệp nén ZIP

Cấu trúc tổng quan của một tệp ZIP tiêu chuẩn



Hình 2: Cấu trúc tổng quan của một tệp ZIP

Các công cụ ZIP sẽ tạo ra hai bản ghi cho mỗi file: **main file** và **central directory**, main file được nối cùng nhau, đánh dấu sự kết thúc của main file bằng **Extended Metadata** (không bắt buộc), nối tiếp sau đó là các central directory sắp xếp theo thứ tự từ điển, và cuối cùng là trường **End-of-Archive** đánh dấu kết thúc file ZIP.

**Local file header**

Đây là phần thông tin miêu tả dữ liệu cục bộ đi kèm với mỗi tệp ở trong tệp ZIP nhằm cho biết thông tin về các thông tin mô tả, định nghĩa, sắp xếp, tổ chức và định dạng dữ liệu. Đây là phần thông tin được đặt ở trong phần đầu mỗi file.

Cấu trúc chi tiết:

|  |  |  |
| --- | --- | --- |
| **Thành phần** | **Kích thước (bytes)** | **Mô tả** |
| Local File Header Signature | 4 | Giá trị cố định: 0x04034b50 |
| Version Needed to Extract | 2 | Phiên bản ZIP yêu cầu để giải nén |
| General Purpose Bit Flag | 2 | Cờ chỉ định trạng thái mã hóa/nén |
| Compression Method | 2 | Thuật toán nén |
| Last Modification Time/Date | 4 | Dùng mã MS-DOS để lưu thời gian sửa đổi cuối |
| CRC-32 | 4 | Mã CRC-32 để xác minh tính toàn vẹn của dữ liệu |
| Compressed Size | 4 | Kích thước dữ liệu đã nén |
| Uncompressed Size | 4 | Kích thước dữ liệu trước khi nén |
| File Name Length | 2 | Độ dài tên tệp |
| Extra Field Length | 2 | Độ dài trường mở rộng |
| File Name | Variable | Tên tệp thực tế |
| Extra Field | Variable | Dữ liệu bổ sung tùy chọn |

**File Data**

Đây là phần chứa dữ liệu nội dung của từng tệp, được lữu trữ dưới dạng nguyên trạng hoặc nén. Trên thực tế, nếu tệp đã được mã hóa, phần nãy sẽ bao gồm cả tiêu đề mã hóa (encryption header) và phần dữ liệu nguyên trạng sẽ được thay thế bằng phần dữ liệu đã mã hóa.

**Data Description (Tùy chọn)**

Đây là phần tùy chọn trong cấu trúc tệp ZIP, chỉ xuất hiện trong trường hợp không thể xác định giá trị CRC-32 của dữ liệu trong trường Local File Header. Thường thì là khi tệp Zip được tạo dưới dạng streaming (truyền tải hoặc xử lý dữ liệu theo dòng liên tục mà không cần phải lưu trư toàn bộ dữ liệu đó vào một nơi cố định trước khi sử dụng, có thể truy cập ngay khi xuất hiện) hoặc khi General Purpose Bit Flag có 3 bit được bật.

Cấu trúc chi tiết:

|  |  |  |
| --- | --- | --- |
| Thành phần | Kích thước (bytes) | Mô tả |
| CRC-32 | 4 | Mã CRC-32 để xác minh tính toàn vẹn của dữ liệu |
| Compressed Size | 4 | Kích thước dữ liệu đã nén |
| Uncompressed Size | 4 | Kích thước dữ liệu trước khi nén |

**Central Directory**

Đây là phần đóng vai trò là bảng chỉ mục tổng hợp của tất cả các tệp bên trong. Nó chứa thông tin chi tiết về tên, vị trí, kích thước, trạng thái nén/mã hóa, các thuộc tính… của từng tệp, điều này cho phép các phần mềm đọc tệp ZIP có thể nhanh chóng truy xuất hoặc chỉnh sửa mà không cần đọc toàn bộ. **Central Directory** bao gồm từng **Header** cho mỗi tệp. Trong đó, cấu trúc của từng Central Directory Header được miêu tả như sau:

|  |  |  |
| --- | --- | --- |
| **Thành phần** | **Kích thước (bytes)** | **Mô tả** |
| Central File Header Signature | 4 | Giá trị cố định: 0x02014b50 |
| Version Made By | 2 | Phiên bản ZIP tạo ra tệp này |
| Version Needed to Extract | 2 | Phiên bản ZIP tạo ra tệp này |
| General Purpose Bit Flag | 2 | Trạng thái cờ chỉ định mã hóa hoặc nén |
| Compression Method | 2 | Thuật toán nén |
| Last Modification Time/Date | 4 | Thời gian và ngày chỉnh sửa cuối |
| CRC-32 | 4 | Giá trị checksum CRC-32 của tệp |
| Compressed Size | 4 | Kích thước dữ liệu đã nén |
| Uncompressed Size | 4 | Kích thước gốc của tệp |
| File Name Length | 2 | Độ dài tên tệp |
| Extra Field Length | 2 | |  | | --- | |  |   Độ dài trường mở rộng |
| File Comment Length | 2 | Độ dài của bình luận cho tệp |
| Disk Number Start | 2 | Số đĩa nơi tệp bắt đầu (dành cho việc phân đoạn) |
| Internal File Attributes | 2 | Các thuộc tính nội bộ |
| External File Attributes | 4 | Thuộc tính file theo hệ điều hành |
| Relative Offset of Local Header | 4 | Vị trí của **Local File Header** trong tệp ZIP |
| File Name | Variable | Tên tệp |
| Extra Field | Variable | Dữ liệu mở rộng (nếu có) |
| File Comment | Variable | Bình luận cho tệp (nếu có) |

**End of Central Directory**

Phần đánh dấu kết thúc của tệp ZIP, giúp xác định nhanh Central Directory của tệp

|  |  |  |
| --- | --- | --- |
| **Thành phần** | **Kích thước (bytes)** | **Mô tả** |
| End of Central Directory Signature | 4 | Giá trị cố định: 0x06054b50 |
| Number of This Disk | 2 | Số thứ tự hiện tại (cho ZIP chia nhỏ) |
| Total Number of Central Directory Records | 2 | Tổng số entry trong Central Directory |
| Size of Central Directory | 4 | Tổng kích thước Central Directory |
| Offset of Central Directory | 4 | Vị trí bắt đầu của Central Directory |
| ZIP File Comment Length | 2 | Độ dài của phần bình luận (nếu có) |
| ZIP File Comment | Variable | Nội dung phần bình luận (nếu có) |

Thông qua cấu trúc của một tệp tin ZIP, có thể phân tích được khả năng tấn công vào từng chi tiết theo thành phần định dạng để tìm ra điểm yếu, từ đó xác định được tính khả thi của các phương pháp thám mã. Dưới đây là bảng tổng hợp và so sánh một số khả năng và đánh giá tính khả thi của chúng trong việc phân tích nội dung nhằm mục đích thám mã tệp ZIP.

|  |  |  |
| --- | --- | --- |
| **Phần tấn công** | **Mã hóa truyền thống** | **Mã hóa AES** |
| Central Directory | Dễ bị khai thác | Thông tin hạn chế |
| CRC - 32 | Dễ bị khai thác | Được bảo vệ |
| Local File Header | Không mã hóa | Một phần có thể bị mã hóa |
| Data Descriptor | Có thể khai thác | Phụ thuộc vào chế độ mã hóa |
| Compressed Data | Rất dễ khai thác | Khó khai thác |
| Traditional Encrytion | Nhiều lỗ hổng | Không áp dụng |
| AES Encryption | Không lỗ hổng | Bảo mật rất mạnh |

1. **Phương pháp nén vã mã hóa tệp tin**

**Nén** là quá trình giảm kích thước tệp bằng cách loại bỏ dư thừa hoặc tái cấu trúc dữ liệu một cách hiệu quả. Trong tệp ZIP, các thuật toán nén được áp dụng riêng biệt cho từng tệp, điều này cho phép lưu trữ cả tệp đã nén và tệp không nén trong một tệp ZIP.

Dưới đây là các thuật toán nén phổ biến thường được sử dụng

**Stored (Không nén)**

Với thuật toán này, dữ liệu được lưu dưới dạng nguyên trạng, không có bất kỳ xử lý nén nào được diễn ra. Trong quá trình này, mã CRC-32 vẫn được tính toán để đảm bảo tính toàn vẹn của dữ liệu. Việc không nén giúp truy cập một số tệp trong ZIP trở nên dễ dàng hơn mà không cần tốn đến tài nguyên tính toán, phù hợp với những tệp đã được nén trước đó như video, hình ảnh .jpeg…

**DEFLATE (Nén chuẩn)**

Đây là thuật toán kết hợp giữa thuật toán LZ77 (lặp chuỗi) và mã hóa Huffman, việc sử dụng kết hợp giữa hai thuật toán nhằm hướng tới mục đích giảm kích thước của các dữ liệu mà vẫn cân bằng được tốc độ nén hiệu quả. Tuy nhiên, DEFLATE có độ phức tạp cao hơn, yêu cầu nhiều tài nguyên tính toán hơn và hiệu quả sẽ giảm đáng kể nếu dữ liệu đầu vào đã được nén trước đó.

**BZIP2**

Đây là thuật toán sử dụng **nén khối (block sorting)** để nén dữ liệu. Thuật toán này đem lại hiệu quả cao hơn DEFLATE nhưng bù lại vấn đề yêu cầu nhiều tài nguyên hơn. Từ đó, có thể hiểu rằng, BZIP2 làm giảm nhiều kích thước của tệp hơn DEFLATE nhưng tốc độ chậm lại chậm hơn, đây là phương thức nén thường được dùng với dữ liệu văn bản lớn hoặc cho các hệ thống kỹ thuật nén hiệu quả.

**Mã hóa** là một lớp bảo mật bổ sung nhằm bảo vệ dữ liệu trong tệp ZIP khỏi việc truy cập trái phép. Trên thực tế, việc mã hóa chỉ thay đổi **giá trị** của dữ liệu đầu vào (plaintext) thành một dạng khác (ciphertext) bằng cách kết hợp với một chuỗi khóa giả ngẫu nhiên mà không thêm bất kỳ byte mới nào vào dữ liệu. Đặc điểm này được thiết kế nhằm mục đích bảo toàn kích thước dữ liệu để đạt hiệu quả trong việc xử lý và truyền tải.

Tệp ZIP hỗ trợ **hai** loại mã hóa chính, đó là: mã hóa truyền thống **(Traditional Encryption)** và mã hóa **AES** hiện đại.

**Mã hóa truyền thống (Traditonal PKZIP Encryption)**

**Mã hóa AES (Advanced Encryption Standard)**

**Bảng so sánh Mã hóa truyền thống và AES**

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | **Mã hóa truyền thống** | **Mã hóa AES** |
| Mức độ bảo mật | Thấp | Cao |
| Tính tương thích | Cao (Phù hợp với tệp ZIP cũ) | Trung bình (chỉ với ZIPX) |
| Lỗ hổng bảo mật | Dễ bị tấn công (CRC-32, plaintext) | Không có lỗ hổng lớn |
| Hiệu suất | Nhanh | Chậm hơn nhưng đáng tin cậy |

1. **Phương pháp giải nén và giải mã hóa tệp tin**
2. **Các công cụ hỗ trợ phổ biến**

**Nguồn tài liệu tham khảo**[**https://support.pkware.com/**](https://support.pkware.com/)

[**https://pkware.cachefly.net/webdocs/casestudies/APPNOTE.TXT**](https://pkware.cachefly.net/webdocs/casestudies/APPNOTE.TXT)