**FILE SANITIZATION UTILITY**

A **file sanitization utility** is a software tool designed to clean files of any malicious or potentially harmful content, ensuring they are safe to open and use.

**METHODS IN FILE SANITIZTION UTILITY:**

**1.Content disarm and reconstruction:**

**It** is a method used to remove potential threats from files by breaking them down, eliminating any harmful elements, and then rebuilding them into a safe, usable format.

**How It Works:**

1. **Disassemble the File**:

The file is broken down into its basic components (e.g., text, images, metadata).

1. **Analyze and Clean**:

Each component is scanned for malicious content, such as malware.

Any harmful elements are removed or neutralized.

1. **Reconstruct the File**:

The cleaned components are put back together to form a new, safe version of the original file.

The new file maintains the original data and usability but without the harmful content.

**2.signature-based detection:**

**It** is a method used to identify malware and other threats by comparing files against a database of known malware signatures.

**How It Works:**

1**. Database of Signatures**:

A collection of unique patterns, identifiers are identified in known malware. These signatures can be binary patterns, file hashes, strings of text, that are unique to specific malware.

2. **Scanning Files**:

When a file is scanned, its contents are analyzed to see if they match any of the signature in the database.

The scanner checks various aspects of the file, such as its code, structure, and behaviour, against the known signatures.

3. **Match Detection**:

If the file contains patterns that match a signature in the database, the scanner flags it as malicious.

The flagged file can then be quarantined, deleted, or otherwise handled to prevent harm to the system.

**3. Heuristic analysis:**

It is a method used to detect new or unknown malware by examining a file.

**How it works:**

1. **Behaviour analysis:**

The file is analyzed for suspicious behaviour.

2. **Pattern Recognition**:

File check for patterns that is found in malicious software.

3. **Rule Application**:

**Using predefined heuristic rules and algorithms you can access file behaviour.**

**4. Scoring and decision:**

**Now file behaviour is being scored high for being suspicious.**

**4.** **Macro stripping:**

Is the process of removing macros from files, particularly from documents such as Microsoft Word (.doc, .docx) or Excel (.xlsx) files. Macros are scripts or programs embedded within these documents that can execute automated tasks.

**How It Works:**

**1. Identifying Macros**:

**You need to scan file if there are any embedded micros.** Macros are typically written in scripting languages like Visual Basic for Applications (VBA).

2. **Disabling or Removing Macros**:

If macros are detected you need to either disable it or remove it.

When you disable file macros cannot execute, while removal removes the macro code entirely.

3. **Preserving Document Content**:

The process ensures that the main content of the document (text, images, formulas) remains intact and unaffected.

Only the macro code, which poses potential security risks, is altered or removed.

**5.** **Binary padding removal:**

It is a process used to eliminate unnecessary data that may be added to a file.

**How it works:**

**1. Identifying Binary Padding:**

Identifying binary padding in a file means looking for extra data inside the file that doesn't actually do anything useful. This extra data is like filler and can sometimes be used to hide harmful software.

**2. Removing Extraneous Data**:

Once identified, the unnecessary data, which could potentially hide malicious code is cleaned from the file.

This ensures that the file contains only the necessary information and removes any potential security risks posed by hidden content.

**3. Preserving File Integrity**:

Throughout the process, care is taken to preserve the integrity.

Only the non-essential or suspicious data is removed, leaving the file in a safe and functional state.

**6. File Type Validation and Filtering:**

File type validation and filtering is the process of checking files to ensure they are of an expected type and blocking or allowing them based on their type. This helps prevent harmful files from being uploaded, shared, or executed.

**How It Works:**

**1.Check the File Type**:

When a file is received, its type is checked to see if it matches what is expected or allowed. This is often done by examining the file's extension (like .jpg, .pdf, .exe) and its internal structure or metadata.

2. **Validation**:

You should check whether .jpg extension is from jpeg file or not

**3.Filtering**:

Based on the validation, the file is either allowed or blocked.

Safe or allowed file types are let through, while dangerous or disallowed types are blocked or quarantined.

**7. Active Content Removal:**

Taking out parts of a file that can run actions or scripts, like macros, embedded code, or active links. This makes the file safer by stopping it from performing any potentially harmful actions.

1. **Identify Active Content**:

The file is scanned to find any active elements like macros, JavaScript, embedded objects, or executable code.

2. **Remove Active Content**:

These active elements are removed or disabled, ensuring the file can no longer execute any scripts or actions.

3. **Preserve File Content**:

The main content of the file, such as text and images, is kept intact. Only the parts that can perform actions are removed.

**8.Embedded Object Extraction and Analysis:**

Extracts embedded objects (e.g., images) from files for separate analysis and sanitization.

**1. Identify Embedded Objects**:

The file is scanned to detect any embedded objects such as images, documents, spreadsheets, or other files that are included within the main file.

2. **Extract Embedded Objects**:

These embedded objects are separated and extracted from the main file for individual examination.

**3. Analyze Extracted Objects**:

Each extracted object is thoroughly analyzed for any signs of malicious content or behaviour.

This analysis can include scanning for viruses, checking for suspicious code, and validating the file type and integrity.

**4. Clean or Remove Malicious Objects**:

If any embedded object is found to be malicious, it is either cleaned (if possible) or removed.

The cleaned or sanitized main file is then reconstructed without the harmful embedded objects.