PDF File Format

PDF dosyaları günümüzde iş başvurularında, davetiye dağıtımında, veya alınan eğitimler sonucunda kazanılan dijital sertifikalar paylaşmak istendiğinde ve daha birçok alanda yaygın olarak kullanılmaktadır. Kullanım oranının yüksek olması oluşturduğu tehdit kapsamının da büyüklüğünü gözler önüne sermektedir. Benim bu tehtidi dijital kitap ve döküman okumaya başladığımda fark etmeye başladım. Okumak için açtığım PDF dosyalarının zararlı yazılım içerip içermediği konusunda şüphe duymaya başladım. Bu şüphelerden kurtulabilmek adına birkaç örnek analiz yazısı okuyup videolarını izledikten sonra okumak istediğim PDF dökümanlarını kendim analiz etmeye karar veridm.

PDF dosya formatı 1993 yılında Adobe tarafından oluşturulmuştur. İşletim sisteminden bağımsız belgelerin sunulabilmesini sağlayan metin tabanlı bir yapı olarak oluşturuldu. Metinlerin yanı sıra görseller, videolar ve daha birçok türden içerik PDF dosyaları kullanılarak sunulabilmektedir.

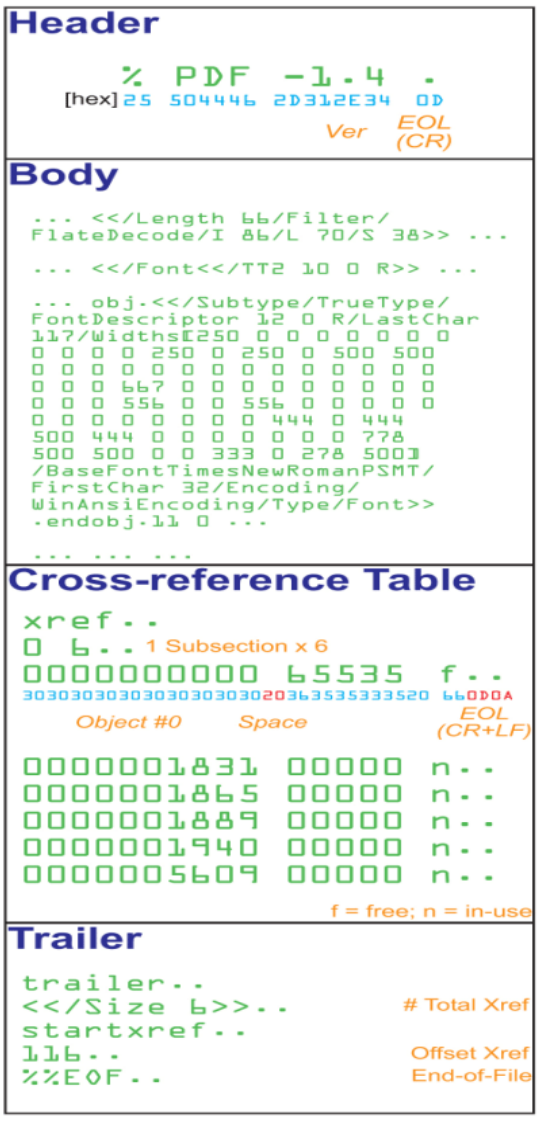
Analiz edebilmek için öncelikle PDF dosyalrının yapısını anlamamız gerekiyor. Bir PDF dosyası 4 ana bölümden oluşuyor. Bu bölümler;

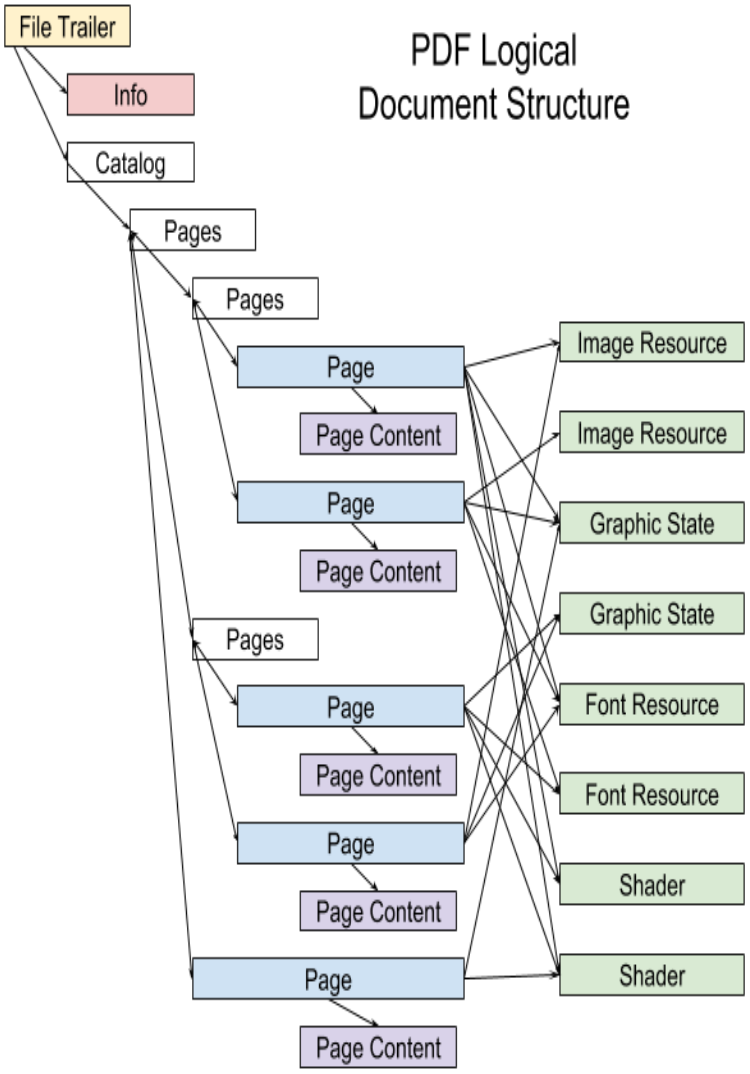
- Header, dosyaın bir PDF dosyası olduğunu belirtir ve sürüm bilgisini bulunur.
- The Body, PDF dosyasında kullanıcıya gösterilen metin, görsel, video ve benzeri içeriklerin tutulduğu kısımdır. Bu kısımda bulunan içerikler aslında bir nesne içerisinde tutulmaktadır. Bu nesneler farklı türkerde olabilir.
    - Names
    - Numbers
    - Boolean
    - String
    - Arrays
    - Dictionary
    - Streams
    - Indirect Objects
- Cross-reference table, the body kısmında bulunan her nesnenin/içeriğin adres (20 bayt uzunluğunda – 3 parçadan oluşan) bilgisinin tutulduğu tablodur (Bir PDF tablosunda birden fazla Cross-reference tablosu bulunabilir). Bu tablo sayesinde bir nesnenin tutulduğu konumu belirlemek için bütün dosyanın okunmasına gerek kalmıyor ve büyük boyutlu dosyaları açarken de zaman kazandırıyor. Cross-reference tablosunda tutulan her nesne için iki satır bulunmaktadır.
    - İlk satır iki sayı içermektedir. Bu sayılardan ilki nesnenin sayısal kimliğidir. İkinci sayı ise bu nesnenin altında bulunan nesnelerin sayısıdır.
    - İkinci satırda ise ilk 10 baytı PDF dosyasının başlangıcından o nesnenin başlangıcına kadar olan uzaklığını tanımlamak için kullanılır. Ardından gelen 5 baytlık parçada nesnenin üretim numarası belirtilir. Son parçada ise “n” ve “f” harfleri kullanılarak nesnenin kullanımda olup olmadığı belirtiliyor.
- Trailer, PDF okuyucular dosyayı sondan okumaya başlar. Trailer kısmıysa dosyanın sonunda bulunur ve PDF okuyucunun nesneleri bulabilmesi için Cross-reference tablosu ve özel nesneler hakkında bilgiler tutar. Trailer kısmı birçok parçadan oluşuyor. Bu parçalar;
    - Size (Integer), Cross-reference tablosundaki girişlerin sayısını tanımlar.
    - Prev (Integer), dosyada birden fazla Cross-reference tablosu kullanıldığı durumda dosyanın başlangıcından önceki Cross-reference tablosuna olan uzaklığı tanımlar.

- Root, (Dictionary), tabloda kök nesne adresini tutan girişin bilgisini tutar. Bu nesne belgenin kataloğudur. Yani dosyanin nasıl sunulacağı hakından bilgi ver belgenin içeriğini açıklayan diğer nesnelere referanslar verir.
- Info, dosya hakkında genel bilgilerin bulunduğu alandır.
- Startxref, Cross-reference tablosunun başlangıcının dosya başlangıcından ne kadar uzak olduğunu temsil eden alandır.
- %%EOF, Trailer bölümünün sonladığını gösterir.

PDF dosyaları birçok farklı yöntem kullanılarak zararlı aktivite gerçekleştirebiliyor. Bunlardan en yaygın kullanılan yöntemler;

- Javascript; PDF dosyaları JavaScript kodları içerebiliyor. Bu sayede istenilen zararlı kodlar kurban cihazlarda çalıştırabiliyor.
- URLs; PDF dosyalarına zararlı bir domain/ip adresi gömülerek kullanıcılar phishing sayfalarına yönlendirilebiliyor veya kurban cihaza zararlı dosyalar indirtilerek çalıştırılması sağlanabiliyor (Genelde bu işlem PDF dosyası açıldığında tetikleniyor).
- Embedded files;PDF dosyalarının içerisine farklı uzantılarda dosyalar gömülebiliyor (Bazı PDF okuyucu yazılımlar blacklist kullanarak belirli dosya uzantılarını açmayı kabul etmese de zararlı yazılım içerebilecek daha birçok dosya uzantısı mevcttur).
- Encrypted PDFs; PDF dosyalarının içerisine şifrelenmiş farklı PDF dosyaları da gömülebiliyor.





PDF dosyasının yapısı hakkında fikir sahibi olduktan sonra (Zararlı PDF dosyalarını incelerken PDF dasya yapısının daha net anlaşılacağına inanıyorum) zararlı yazılımları çalıştırabilmek için PDF dosyalarının nasıl kullanıldığını araştırmaya başlayabiliriz.

Kaynaklar

- https[:]//www.intezer.com/blog/incident-response/analyze-malicious-pdf-files/
- https[:]//resources.infosecinstitute.com/topic/pdf-file-format-basic-structure/
- https[:]//labs.appligent.com/pdfblog/pdf_cross_reference_table/
- https[:]//www.oreilly.com/library/view/pdf-explained/9781449321581/ch04.html
- https[:]//medium.com/aia-sg-techblog/basic-structure-of-portable-document-format-pdf-79db682579c9
- https[:]//www.youtube.com/watch?v=U8xExM3ykYA