# 악성코드 탐지, 암호화 및 패키징 프로젝트

데이터 수집 보고서



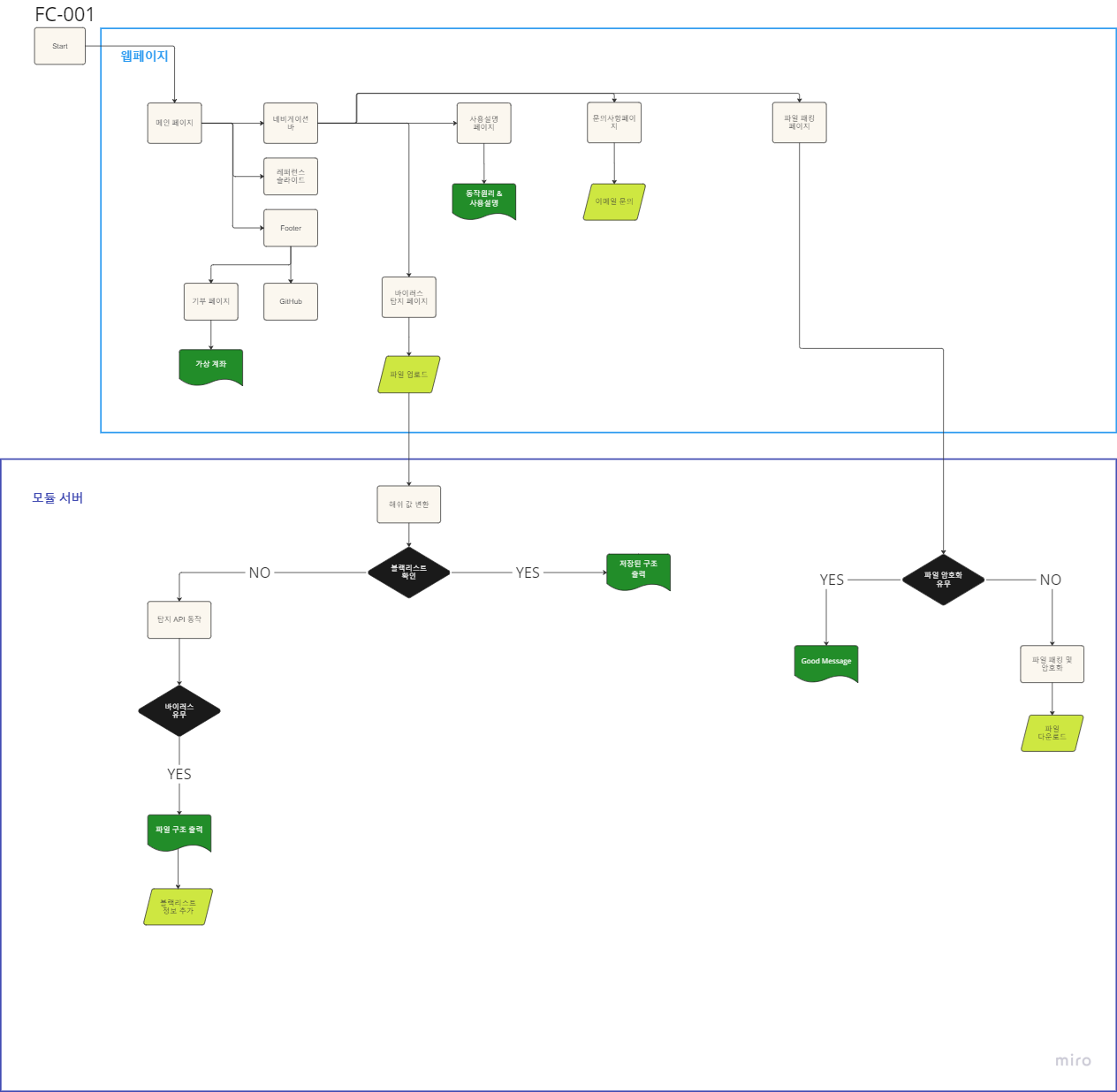
**목**

**차**

1. **데이터 수집 정책**
2. **선정된 데이터**
3. **선정되지 않은 데이터**
4. **결론**

# 데이터 수집 정책

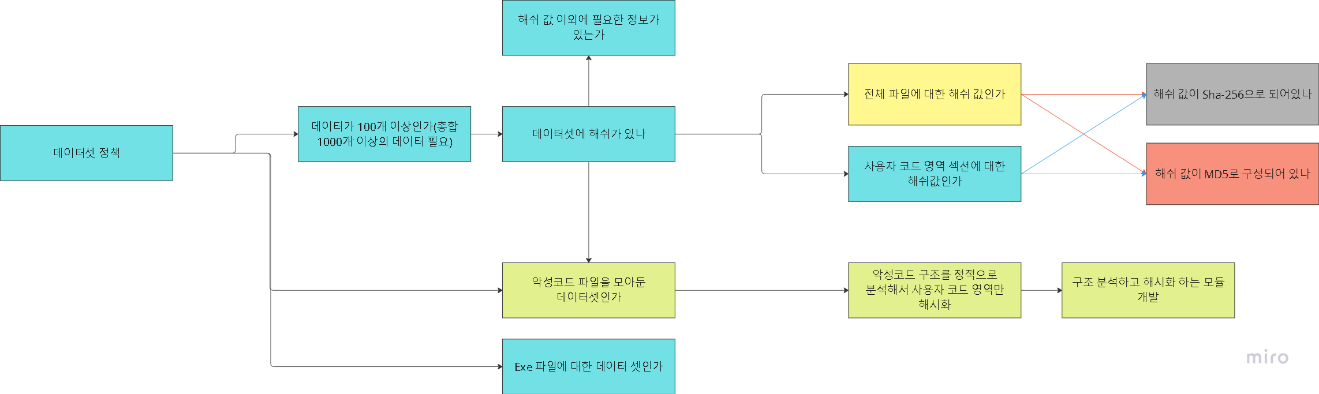
먼저 플로우 차트를 살펴 보면 저희가 만들어야 하는 데이터셋은 블랙리스트에 해당 합니다.



블랙리스트는 사전에 정의된 바이러스 정보가 들어가야 합니다. 바이러스 정보에는 파일의 해시, 파일 타입 등의 정보, 시그니처 정보, pe 정보 등이 포함되어야 합니다. 또한 Virus Total 사이트에서 파일을 업로드 하거나 해시를 통해 검색하면 나오는 정보로 Behavior 정보도 있습니다. 이 또한 API를 통해 받아올 수 있는 정보입니다. 그래서 블랙리스트에는 최종적으로 Detail, Behavior 정보가 포함됩니다.

블랙리스트를 만들기 위해선 사전 정의된 바이러스 정보가 있는 데이터셋을 사용해야 합니다. 가장 좋은 방안으로는 Virus Total에 정의되어 있는 정보를 가져오는 것이지만 서비스에서 따로 제공해주는 데이터가 없었습니다. 그래서 저희는 다른 사이트를 서치하여 데이터셋을 수집하기로 했습니다.

데이터셋 정책



데이터셋을 수집할 때 최종적으로 고려해야 할 사항입니다.

1. 데이터가 100개 이상인가?
2. 데이터셋에 해시 값이 포함되어 있나?
3. 전체 파일에 대한 해시 값인가?
4. MD5 해시가 있는가?

저희가 만들 블랙 리스트는 서비스의 시간 단축이 목적인 데이터 셋입니다. 따라서 사전에 정의된 데이터들이 많으면 좋습니다. 따라서 위에 4개의 정책을 중점으로 데이터셋을 수집했습니다.

1. 선정된 데이터

최종적으로 선정된 데이터셋은 Malware bazzar abuse에서 가져온 csv file입니다. 이 파일은 2020년 2월 13일부터 2024년 8월 29일까지 이 사이트에 보고된 바이러스 파일들에 대한 정보가 포함되어 있습니다.

**포함된 정보**

|  |  |
| --- | --- |
| **이름** | **설명** |
| first\_seen\_ufc | 바이러스가 처음 보고된 시간 |
| sha256\_hash | 파일의 SHA-256 해시 값 |
| md5\_hash | 파일의 md5 해시 값 |
| sha1\_hash | 파일의 SHA-1 해시 값 |
| reporter | 바이러스를 보고한 사람 |
| file\_name | 바이러스 파일의 이름 |
| file\_type\_guess | 바이러스 파일의 형식(추정값) |
| mime\_type | 바이러스 파일의 MIME 유형 |
| signature | 바이러스의 정의된 이름(ex. Loki) |
| clamav | Clamav 동작 후 나온 값(ex. RedLineStealer) |
| vtpercent | Virustotal 검사 점수 |
| imphash | PE 파일의 import table 해시 값 |
| ssdeep | 파일 간 유사성을 비교하기 위한 해시 값 |
| tlsh | TLSH 해시 값. 파일의 유사성을 비교하기 위한 해시 |

이 데이터셋에는 이러한 정보가 담긴 데이터가 약 80만개가량 있습니다. 해당 데이터셋을 가공하여 블랙리스트의 정보를 만드는데 적합하다고 판단했습니다. 최종적으로 블랙리스트에 적용되어야 할 정보는 Virustotal 사이트를 통해서 나온 detail과 behavior 섹션의 값이므로 Virustotal API를 적용시켜 가공할 수 있는 데이터셋으로 MD5 hash를 선정했습니다. 이 데이터셋을 API module을 통해 가공하여 블랙리스트를 정의할 것입니다.

`3. 선정되지 않은 데이터

결론

블랙리스트와 Virus total API를 돌려서 나온 정보가 불균형을 이루면 안되기 때문에 Virustotal API로 데이터를 가공해야 합니다. 따라서 수집해야 할 데이터셋에는 해시 값이 필수로 포함되어 있어야합니다. 해당 해시 값으로 Virustotal API module을 작동시켜 나온 데이터를 데이터베이스에 저장하여 블랙리스트 데이터베이스를 구성해야 합니다.

저희는 최종적으로 Malware Bazzar abuse 의 악성코드 데이터셋을 선정했고 해당 데이터는 약 80만개의 데이터를 포함하고 있습니다. 데이터셋에는 md5, sha-256 등의 중요한 해시정보가 포함되어 있고 저희는 md5 hash 값을 이용해 데이터를 재가공 할 예정입니다.

악성코드의 분석 데이터는 정형화 되어있지 않는 경우가 많기 때문에 데이터베이스 선정에 있어서 Non-SQL 방식의 데이터베이스를 선정해야 했습니다. 저희는 BSON 형식으로 데이터를 저장하는 MongoDB를 선정했고 해당 DB에 블랙리스트를 구축할 것입니다.