Binary Diff 및 Dataset 수집 도구 구현

캡스톤디자인 2022 팀36

[최종발표]

발표자: 윤형준

2022-05-27

목차

프로젝트 소개

(1) 프로젝트 소개

팀 정보

- 산학분반 자유주제 36팀
- 산학기관: 한국과학기술원 사이버보안연구센터
- 국가연구과제 “바이너리 분석 플랫폼 B2R2 구현”에 참여하여 Bianry Diff 파트 단독 담당

팀 멤버

- 윤형준 학부생 (팀장)
  - 설계 및 구현, 연구, 논문 조사 및 발표, 프로젝트 일정 및 산출물 관리
- 차상길 교수님 (산학 멘토)
  - 연구 및 개발 방향성 제시, 코드 리뷰, 논문 추천, 매주 온라인 미팅 및 피드백

(1) 프로젝트 소개

수행 배경

- Binary Diff
  - 동일한 바이너리의 서로 다른 두 버전 간의 차이점을 식별하는 기술
  - 바이너리 분석은 소프트웨어 보안을 실현하기 위해 필수
- 오픈 소스 바이너리 분석 도구
  - 현존하는 잘 알려진 바이너리 분석 도구인 IDA Pro는 라이센스 비용이 수천만 원
  - 뛰어난 오픈 소스 바이너리 분석 도구가 필요
- 실무적인 binary diff 도구 부재
  - 기존 도구들은 연구 수준의 아이디어를 실현하는 데에 초점
- Binary Diff 연구용 공통의 Dataset 부재 문제
  - 서로 다른 binary diff 연구를 평가할 만한 공통적이고 양질의 Dataset 부재

(1) 프로젝트 소개

주제: Binary Diff 및 Dataset 수집 도구 구현

Binary Diff 도구 (B2R2-BinDiff)

- 오픈소스 바이너리 분석 플랫폼인 B2R2에 Binary Diff 도구 구현
- 두 프로그램을 바이너리 수준에서 분석하여 diff
- binary-pair로 구성된 dataset을 특정 metric으로 측정

Dataset 수집 도구 (Binary-Gleaner 구현)

- B2R2와 별개로 구현
- binary diff에 사용할 dataest을 수집
- 소스코드 수준과 바이너리 수준의 데이터 제공

- 두 도구의 모든 구현 과정 및 결과는 오픈 소스하여 공개

(1) 프로젝트 소개

산출물 및 기대효과

Binary Diff 도구 (B2R2-BinDiff)

- 오픈 소스 소프트웨어 및 보안 생태계에 기여
- 보안 전문가, 해커, 해킹대회 참가자 등 실무의 분석가들이 무료로 사용 및 기여 가능
- 연구자가 학술적 연구를 위해서도 사용 가능

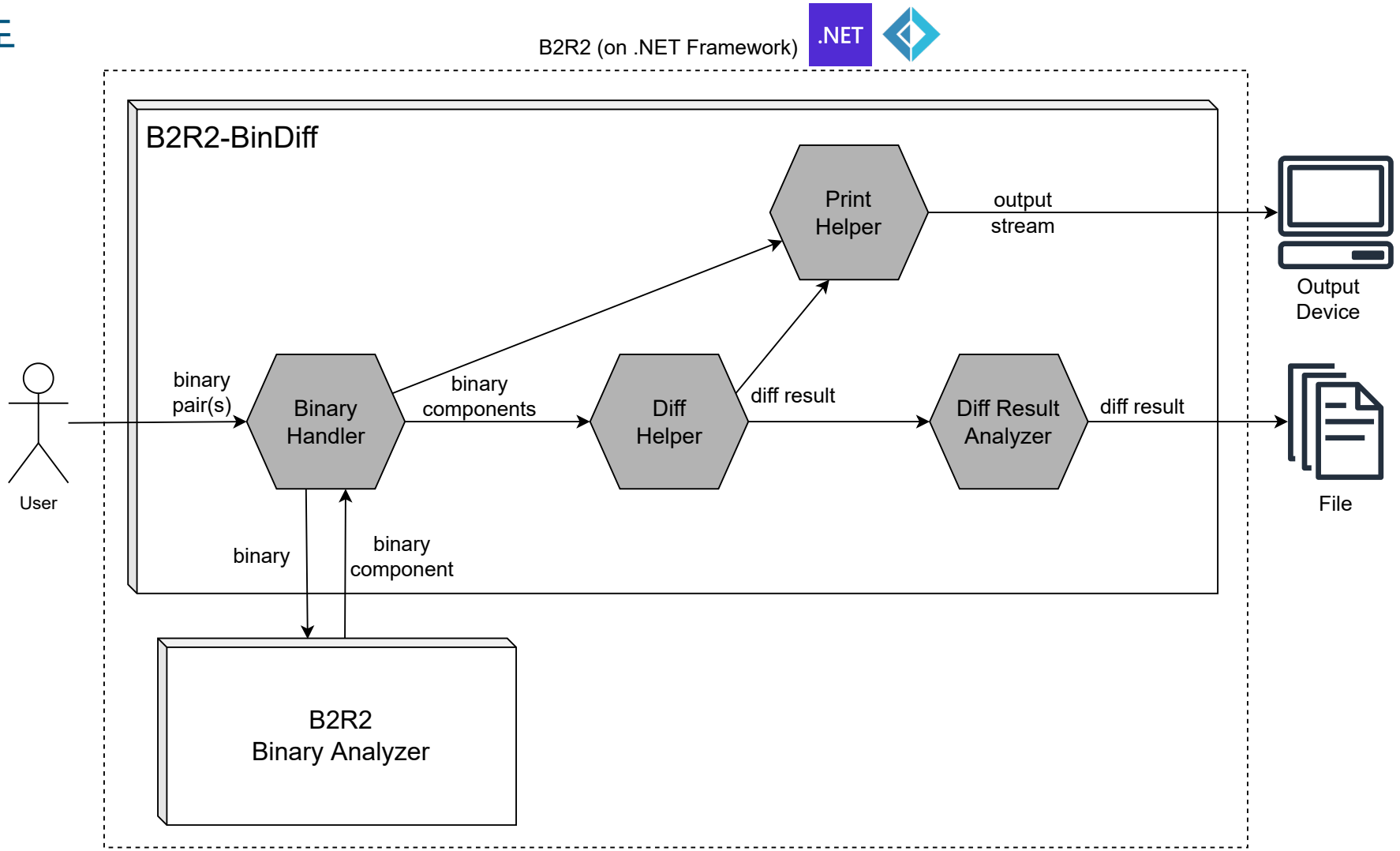
Dataset 수집 도구 (Binary-Gleaner)

- 학계의 공통적인 binary diff dataset 부재 문제 해결
- 서로 다른 binary diff 연구들을 체계적이고 정확하게 평가 가능

Binary Diff 도구

(2) Binary Diff 도구

시스템 구조도



(2) Binary Diff 도구

Diff 결과 예시

```
<foo>                                                    <foo>
0000000000001169: endbr64                                0000000000001169: endbr64
000000000000116d: push RBP                               000000000000116d: push RBP
000000000000116e: mov RBP, RSP                           000000000000116e: mov RBP, RSP
0000000000001171: mov dword ptr [RBP-0xc], 0xa           0000000000001171: mov dword ptr [RBP-0x8], 0xa
0000000000001178: mov dword ptr [RBP-0x8], 0x14          0000000000001178: mov dword ptr [RBP-0x4], 0x14
000000000000117f: mov EDX, dword ptr [RBP-0xc]

0000000000001182: mov EAX, dword ptr [RBP-0x8]           000000000000117f: mov EAX, dword ptr [RBP-0x8]
0000000000001185: add EAX, EDX                           0000000000001182: sub EAX, dword ptr [RBP-0x4]
0000000000001187: add EAX, 0x1e                          0000000000001185: add EAX, 0x1e
000000000000118a: mov dword ptr [RBP-0x4], EAX
000000000000118d: mov EAX, dword ptr [RBP-0x4]
0000000000001190: pop RBP                                0000000000001188: pop RBP
0000000000001191: ret                                    0000000000001189: ret

<main>                                                   <main>
0000000000001192: endbr64                                000000000000118a: endbr64
0000000000001196: push RBP                               000000000000118e: push RBP
0000000000001197: mov RBP, RSP                           000000000000118f: mov RBP, RSP
000000000000119a: lea RDI, qword ptr [0x2004]            0000000000001192: lea RDI, qword ptr [0x2004]
00000000000011a1: call -0x141                            0000000000001199: call -0x139
00000000000011a6: mov EAX, 0x0
00000000000011ab: call -0x42
00000000000011b0: mov ESI, EAX
00000000000011b2: lea RDI, qword ptr [0x200c]
00000000000011b9: mov EAX, 0x0                           000000000000119e: mov EAX, 0x0
00000000000011be: call -0x14e                            00000000000011a3: call -0x15a
                                                         00000000000011a8: mov ESI, EAX
                                                         00000000000011aa: lea RDI, qword ptr [0x2009]
                                                         00000000000011b1: mov EAX, 0x0
                                                         00000000000011b6: call -0x146
00000000000011c3: mov EAX, 0x0                           00000000000011bb: mov EAX, 0x0
00000000000011c8: pop RBP                                00000000000011c0: pop RBP
00000000000011c9: ret                                    00000000000011c1: ret
00000000000011ca: nop word ptr [RAX+RAX+0x0]             00000000000011c2: nop word ptr [CS:RAX+RAX+0x0]
                                                         00000000000011cc: nop dword ptr [RAX+0x0]

<__libc_csu_init>                                        <__libc_csu_init>
00000000000011d0: endbr64                                00000000000011d0: endbr64
00000000000011d4: push R15                               00000000000011d4: push R15
00000000000011d6: lea R15, qword ptr [0x3db0]            00000000000011d6: lea R15, qword ptr [0x3db0]
00000000000011dd: push R14                               00000000000011dd: push R14
00000000000011df: mov R14, RDX                           00000000000011df: mov R14, RDX
```

[CODE 영역 diff 예시 사진]

```
(.rodata)                                                (.rodata)
01 00 02 00 41 41 41 41  42 42 42 00 25 64 0A 00 | *.*.AAAABBB.%d_.   01 00 02 00 42 42 42 42  00 25 64 0A 00         | *.*.BBBB.%d_.

(.eh_frame_hdr)                                          (.eh_frame_hdr)
01 1B 03 3B 4C 00 00 00  08 00 00 00 10 F0 FF FF | ***;L...*...*...   01 1B 03 3B 4C 00 00 00  08 00 00 00 10 F0 FF FF | ***;L...*...*...
80 00 00 00 40 F0 FF FF  A8 00 00 00 50 F0 FF FF | ....@.......P...   80 00 00 00 40 F0 FF FF  A8 00 00 00 50 F0 FF FF | ....@.......P...
C0 00 00 00 70 F0 FF FF  68 00 00 00 59 F1 FF FF | ....p...h...Y...   C0 00 00 00 70 F0 FF FF  68 00 00 00 59 F1 FF FF | ....p...h...Y...
D8 00 00 00 82 F1 FF FF  F8 00 00 00 C0 F1 FF FF | ....?...........   D8 00 00 00 7A F1 FF FF  F8 00 00 00 C0 F1 FF FF | ....z...........
18 01 00 00 30 F2 FF FF  60 01 00 00             | **..0...`*..       18 01 00 00 30 F2 FF FF  60 01 00 00             | **..0...`*..

(.eh_frame)                                              (.eh_frame)
14 00 00 00 00 00 00 00  01 7A 52 00 01 78 10 01 | *.......*zR.*x**   14 00 00 00 00 00 00 00  01 7A 52 00 01 78 10 01 | *.......*zR.*x**
1B 0C 07 08 90 01 00 00  14 00 00 00 1C 00 00 00 | *_**.*..*...*...   1B 0C 07 08 90 01 00 00  14 00 00 00 1C 00 00 00 | *_**.*..*...*...
00 F0 FF FF 2F 00 00 00  00 44 07 10 00 00 00 00 | ..../....D**....   00 F0 FF FF 2F 00 00 00  00 44 07 10 00 00 00 00 | ..../....D**....
24 00 00 00 34 00 00 00  88 EF FF FF 30 00 00 00 | $...4.......0...   24 00 00 00 34 00 00 00  88 EF FF FF 30 00 00 00 | $...4.......0...
00 0E 10 46 0E 18 4A 0F  0B 77 08 80 00 3F 1A 3A | .**F**J*_w*..?*:   00 0E 10 46 0E 18 4A 0F  0B 77 08 80 00 3F 1A 3A | .**F**J*_w*..?*:
2A 33 24 22 00 00 00 00  14 00 00 00 5C 00 00 00 | *3$"....*...\...   2A 33 24 22 00 00 00 00  14 00 00 00 5C 00 00 00 | *3$"....*...\...
90 EF FF FF 10 00 00 00  00 00 00 00 00 00 00 00 | ....*...........   90 EF FF FF 10 00 00 00  00 00 00 00 00 00 00 00 | ....*...........
14 00 00 00 74 00 00 00  88 EF FF FF 20 00 00 00 | *...t....._...     14 00 00 00 74 00 00 00  88 EF FF FF 20 00 00 00 | *...t....._...
00 00 00 00 00 00 00 00  1C 00 00 00 8C 00 00 00 | ........*.......   00 00 00 00 00 00 00 00  1C 00 00 00 8C 00 00 00 | ........*.......
79 F0 FF FF 29 00 00 00  00 45 0E 10 86 02 43 0D | y...).....E**.*C_   79 F0 FF FF 21 00 00 00  00 45 0E 10 86 02 43 0D | y...!....E**.*C_
06 60 0C 07 08 00 00 00  1C 00 00 00 AC 00 00 00 | *`_**...*.......   06 58 0C 07 08 00 00 00  1C 00 00 00 AC 00 00 00 | *X_**...*.......
82 F0 FF FF 38 00 00 00  00 45 0E 10 86 02 43 0D | ....8....E**.*C_   7A F0 FF FF 38 00 00 00  00 45 0E 10 86 02 43 0D | z...8....E**.*C_
06 6F 0C 07 08 00 00 00  44 00 00 00 CC 00 00 00 | *o_**...D.......   06 6F 0C 07 08 00 00 00  44 00 00 00 CC 00 00 00 | *o_**...D.......
A0 F0 FF FF 65 00 00 00  00 46 0E 10 8F 02 49 0E | ....e....F**.*I*   A0 F0 FF FF 65 00 00 00  00 46 0E 10 8F 02 49 0E | ....e....F**.*I*
18 8E 03 45 0E 20 8D 04  45 0E 28 8C 05 44 0E 30 | *.*E*_.*E*(.*D*0   18 8E 03 45 0E 20 8D 04  45 0E 28 8C 05 44 0E 30 | *.*E*_.*E*(.*D*0
86 06 48 0E 38 83 07 47  0E 40 6E 0E 38 41 0E 30 | .*H*8.*G*@n*8A*0   86 06 48 0E 38 83 07 47  0E 40 6E 0E 38 41 0E 30 | .*H*8.*G*@n*8A*0
41 0E 28 42 0E 20 42 0E  18 42 0E 10 42 0E 08 00 | A*(B*_B**B**B**.   41 0E 28 42 0E 20 42 0E  18 42 0E 10 42 0E 08 00 | A*(B*_B**B**B**.
10 00 00 00 14 01 00 00  C8 F0 FF FF 05 00 00 00 | *...**.....*...    10 00 00 00 14 01 00 00  C8 F0 FF FF 05 00 00 00 | *...**.....*...
00 00 00 00 00 00 00 00                          | ........           00 00 00 00 00 00 00 00                          | ........

(.init_array)                                            (.init_array)
60 11 00 00 00 00 00 00                          | `*......           60 11 00 00 00 00 00 00                          | `*......

(.fini_array)                                            (.fini_array)
20 11 00 00 00 00 00 00                          | _*......           20 11 00 00 00 00 00 00                          | _*......

(.dynamic)                                               (.dynamic)
01 00 00 00 00 00 00 00  01 00 00 00 00 00 00 00 | *.......*.......   01 00 00 00 00 00 00 00  01 00 00 00 00 00 00 00 | *.......*.......
0C 00 00 00 00 00 00 00  00 10 00 00 00 00 00 00 | _.......*.......   0C 00 00 00 00 00 00 00  00 10 00 00 00 00 00 00 | _.......*.......
0D 00 00 00 00 00 00 00  48 12 00 00 00 00 00 00 | _.......H*......   0D 00 00 00 00 00 00 00  48 12 00 00 00 00 00 00 | _.......H*......
19 00 00 00 00 00 00 00  B0 3D 00 00 00 00 00 00 | *.......=......    19 00 00 00 00 00 00 00  B0 3D 00 00 00 00 00 00 | *.......=......
1B 00 00 00 00 00 00 00  08 00 00 00 00 00 00 00 | *.......*.......   1B 00 00 00 00 00 00 00  08 00 00 00 00 00 00 00 | *.......*.......
1A 00 00 00 00 00 00 00  B8 3D 00 00 00 00 00 00 | *.......=......    1A 00 00 00 00 00 00 00  B8 3D 00 00 00 00 00 00 | *.......=......
```

[DATA 영역 diff 예시 사진]

(2) Binary Diff 도구

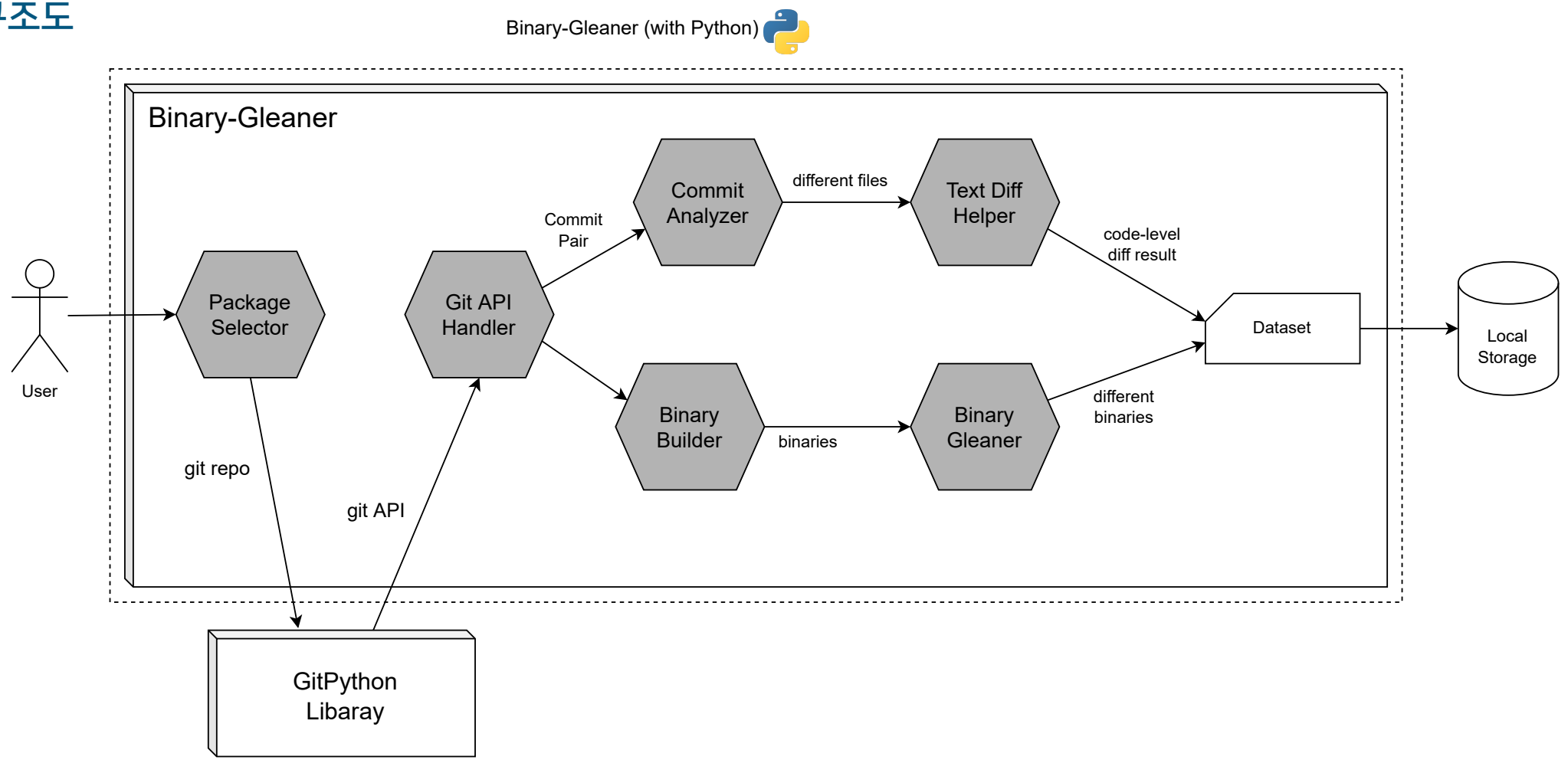
중요 기술 소개

- Diff 알고리즘
  - Git Diff의 네 가지 알고리즘 분석
  - 특히 문서화가 되어있지 않은 Histogram Diff는 소스코드를 분석
- Binary Diff에 Binary Comopnent를 이용
  - 예를 들어, 특정 어셈블리 코드의 주소가 달라져도 matching line으로 인지
- Binary Diff 관련 연구 Survey
  - Binary Diff 연구자들이 흥미롭게 생각하는 metric 도출

Dataset 수집 도구

(3) Dataset 수집 도구

시스템 구조도



(3) Dataset 수집 도구

주요기능 소개

- 특정 git repository의 모든 commit별 데이터를 수집
- High-level Data
  - commit 정보 (예: commit hash, parent commit 정보 등)
  - 소스코드 수준의 정보 (예: 달라진 파일의 수, 각 파일의 변화된 라인의 수)
- Low-level Data
  - binary-pair 정보

➔ high와 low 두 수준간의 diff 양상 비교 연구 가능

시연

(4) 시연

시연 영상

발표 들어주셔서 감사합니다.

Reference

[1] [논문 리뷰] "WYSINWYX: What you see is not what you execute." - CPUU

[2] 바이너리 분석에 대한 오해와 진실 - CSRC

[3] A Survey of Binary Code Similarity - Arxiv

[4] How different are different diff algorithms in Git? - Springer

[5] An O(ND) difference algorithm and its variations - Springer

[6] Myers Diff Algorithm source code – Github Repository of Git

[7] Patience Diff Advantages - Bram Cohen

[8] Histogram Diff Algorithm source code – Github Repository of Git

Appendix A.