UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Effective code maintenance with continuous data collection

by

David Chenaux

Thesis for the Master of Science in Computer Science
Supervised by Prof. Dr. Philippe Cudré-Mauroux

eXascale Infolab
Department of Informatics - Faculty of Science - University of Fribourg

January 23, 2017

UNIVERSITY OF FRIBOURG
Faculty of Science
Department of Informatics
eXascale Infolab

Thesis for the Master of Science in Computer Science
Supervised by Prof. Dr. Philippe Cudré-Mauroux

by David Chenaux

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Contents

Abstract **2**

Acknowledgements **3**

List of Figures **6**

List of Tables **7**

1 Introduction **8**
1.1 Problem definition . . . 8
1.2 Objectives . . . 9
1.3 Organization . . . 9

2 Related work **10**
2.1 What is Program Analysis ? . . . 10
2.2 Program Analysis approaches . . . 12
2.2.1 Static methods . . . 12
2.2.2 Dynamic methods . . . 13
2.3 Program Analysis tools . . . 14
2.3.1 Static Analysis tools . . . 14
2.3.2 Dynamic Analysis tools . . . 15
2.4 Dynamic Analysis limitations . . . 16
2.5 Concluding remarks . . . 17

3 Development **18**
3.1 Proposed solution . . . 18
3.2 Environment . . . 19
3.3 Data Capture Model . . . 20
3.3.1 Setting up the trace . . . 21
3.3.2 Initialization . . . 21
3.3.3 Event Catching . . . 22
3.3.4 Event Handling . . . 22
3.3.5 Trace Ending . . . 24
3.4 Data model . . . 24
3.4.1 Definition . . . 24

3.4.2 Writing to the database . . . 26
3.4.3 Reading from the database . . . 28
3.5 User interface . . . 28
3.5.1 Index page . . . 28
3.5.2 Viewing a file . . . 29
3.5.3 Comparing files . . . 29
3.6 Concluding remarks . . . 29

4 Installation guide **30**
4.1 Setting up the environment . . . 30
4.2 Use the packaged version . . . 30
4.3 From source code . . . 30

5 Experiments **31**
5.1 Test script and machine . . . 31
5.2 Data extraction analysis . . . 31
5.2.1 Memory usage . . . 32
5.2.2 Run-time overheads . . . 32
5.3 Database performances . . . 32
5.3.1 Some numbers . . . 32
5.4 Concluding remarks . . . 32

6 Conclusion **33**
6.1 Conclusion . . . 33
6.2 Future work . . . 33

A Glossary **34**

B License of the software **35**

Bibliography **36**

List of Figures

3.1 The Jenkins home page 20

List of Tables

2.1 Comparison of Dynamic analysis with Static Analysis 11
2.2 Dynamic Analysis Tools 15
2.3 Dynamic Analysis Techniques comparison 16

Chapter 1

Introduction

Integrated development environments have been around for a few decades already, yet none of the modern IDEs was able to successfully integrate their source code editors with the actual data stream flowing through the code. Ability to display the actual data running through the system promises many potential benefits, including easier debugging and code recall, which results in significantly lower code maintenance costs.

1.1 Problem definition

Every developer is more or less feared about the debugging and code reviewing phase of their software. Obviously, this process can sometimes take several painfully hours and each programmer knows how frustrating it can be to search for a hidden bug in thousands lines of codes. In order to support the programmers in this hated task, debuggers are the most useful existing tools which are part of the so called *static program analysis*.

With the apparition of object-oriented programming language, searching for syntactic errors in the code is not anymore sufficient. Therefore, a new research field was pushed forward which is called the *dynamic program analysis* and consists in analyzing the software during its execution. This procedure allows to take in account some possible inputs which were not probed with the SPA. Yet none of the modern IDEs was able to successfully integrate their source code editors with the actual data stream flowing through the code. This is why the present project, which objectives are defined in the next section, is aiming to contribute to the subject.

1.2 Objectives

The goal of this project is to design a proof-of-concept system in one programming language that allows full code instrumentation. This system should be able to seamlessly capture all values for all variables in source code and store them somewhere, with further possibility to easily retrieve saved values. The system should also provide an API to the storage in order to make the data accessible for navigation and display in third-party applications. Also, a basic visualizing interface will also be included in order to allow an easy review of the results. Finally, an evaluation of system's performances will be established through different experiments.

1.3 Organization

The thesis is divided in four main sections :

1. **Related work**: In this first chapter of the thesis, an insight of the existing work on the field *program analysis* will be presented and in particular the DPA. This is including a definition of the field and its particularities, an overview of some available solutions side by side with the current restrictions.
2. **Development**: This part is focusing on the development of the proof-to-concept system with a presentation of the proposed solution and detailed information about its structure.
3. **Installation guide**: Simply an installation guide of the software which describes the needed environment, the package installation and the compilation of the system.
4. **Experiments**: Finally in this section a few experiments will be conducted in order to test and check the performance and results of the software.

The thesis concludes with some outputs and is proposing some future improvements which could be relevant.

Chapter 2

Related work

"Sharing is good, and with digital technology, sharing is easy."
Richard Stallman

The intention of this thesis, as brought up in the introduction, would be to implement a dynamical program analysis system. In order to meet this ambition, it is unavoidable to build a theoretical understanding of "Program Analysis" and therefore the present chapter will endeavor to do a presentation of the subject. The first part propose a definition of the field, then the second suggest some technical approaches. Following, the third section introduce some trendy analysis tools, to finally discuss the actual limitations of dynamic analysis in the fourth part.

2.1 What is Program Analysis ?

Programming environments are an essential key for the acceptance and success of a programming language. After Ducassé and Noyé [1994], without the appropriate developments and maintenance tools, programmers are likely to have a bad software understanding and therefore produce low-quality code. They will be therefore reluctant to use a language without appropriate programming environments, however powerful the programming language is.

As already introduced in the previous chapter, program analysis is an automated process which aims to analyze the behavior of a software regarding a property such as correctness, robustness, safety and liveness. Program analysis can be separated in two methods : the SPA which is performed without running the software and the DPA which is obviously fulfilled during runtime. [Wikipedia, 2016]

The SPA is a really simple solution because it does not require running the program for analyzing its dynamic behavior. The analysis consists in going through the source code and highlight coding errors or ensure conformance to coding guidelines. A classic example of static analysis would be a compiler which is capable of finding lexical, syntactic and even semantic mistakes. The main advantage of this method is that it allows to reason about all possible executions of a program and gives assurance about any execution, prior to deployment.

Nevertheless, according to Gosain and Sharma [2015], since the widespread use of object oriented languages, SPA is found to be ineffective. This can be explained because of the usage of run-time features like dynamic binding, polymorphism, threads etc. To remedy this situation, developers call on DPA which can, after Marek et al. [2015], gain insight into the dynamics and runtime behavior of those systems during execution. Moreover, because the run-time behavior depends now on many other factors, such as program inputs, concurrency, scheduling decisions, or availability of resources, static analysis does not allow full understanding of the code. The following table, proposed by Gosain and Sharma [2015], is resuming the main differences between static and dynamic analysis.

Dynamic Analysis	Static Analysis
Requires program to be executed	Does not require program to be executed
More precise	Less precise
Holds for a particular execution	Holds for all the executions
Best suited to handle run-time programming lan- guage features like polymorphism, dynamic bind ing, threads etc.	Lacks in handling run-time programming lan- guage features.
Incurs large run-time overheads	Incurs less overheads

Table 2.1: Comparison of Dynamic analysis with Static Analysis

In the light of this comparison, it is well worth noting that Dynamic Program Analysis does not substitute to the Static Analysis. Quite the reverse, both are interdependent tools and even if Static Program Analysis is not sufficient anymore, it still gives relevant information about the code to the programmer. The DPA should come in a second phase when the source code has been validated through SPA. As it can be surmised, the ability to examine the actual and exact run-time behavior of the program might be the DPA main advantage, whereas SPA prime edge could be the independence of input stimuli and the generalization for all executions. To illustrate these characteristics, some program analysis solutions are presented further in this chapter.

2.2 Program Analysis approaches

Now that a definition of Program Analysis has been established, some different approaches have to be exposed in order to fully understand the subject. Since the field is really vast, it is not the aim to cover the entire subject, but the reading of this section should give a good overview to the reader. First, the main static analysis methods will be steered following logically with the dynamic analysis methods.

2.2.1 Static methods

The static methods are regrouped in four different categories proposed by Nielson et al. [2004] and briefly presented here, some information was also gathered from the Wikipedia [2016] page which is proposing a grouping based on the same criteria.

Data Flow Analysis: is a technique which consist in gathering information about the values and their evolution at each point of the program. In the Data Flow Analysis the program is considered as a graph in which the nodes are the elementary blocks and the edges describe how control might pass from one elementary block to another.

Constrained Based Analysis: or Control Flow Analysis, aims to know which functions can be called at various points during the execution ; what "elementary blocks" may lead to what other "elementary blocks".

Abstract Interpretation: consists in proving that the program semantics satisfies its specification according to Cousot [2008]. What the program executions actually do should satisfying ,what the program executions are supposed to do. It can be explained as a partial execution of a program which gather information about its semantics without performing all the calculations.

Type and Effect Systems: are two similar techniques. The first one is using types, which are a concise, formal description of the behavior of a program fragment. Rémy [2017] explains that programs must behave as prescribed by their types. Hence, types must be checked and ill-typed programs must be rejected. Effect systems can be described, after Nielson and Nielson [1999] as an extension of annotated type system where the typing judgments take the form of a combination of a type and an effect. This combination is associated with a program relative to a type environment.

2.2.2 Dynamic methods

Now that the main static analysis methods have been defined in the preceding section, the dynamic methods will be exposed here. As it was already stated, dynamic analysis is a quite recent research field which status could be still defined as academical. Therefore the different techniques are not as well established as for the static analysis and can vary a lot in accordance with the author of the different papers. For this work, the following different method were selected which are proposed by Gosain and Sharma [2015] in their survey of Dynamic Program Analysis Techniques and Tools.

Instrumentation based approach: needs a code instrumenter used as a pre-processor in order to inject instrumentation code into the target program. This can be done at three different stages : source code, binary code and bytecode. The first stage adds instrumentation code before the program is compiled, the second one adds it by modifying or re-writing compiled code and the last one performs tracing within the compiled code.

VM Profiling based technique: uses the profiling and debugging mechanism provided by the particular virtual machine, for example the JPDA for Java SDK or the PDB for Python. These profilers give an insight into the inner operations of a program, especially the memory and heap usage. To capture these profiling information plug-ins are available and can access the profiling services of the VM. Benchmarks are then used for actual run-time analysis which acts like a block-box test for a program. This process involves executing or simulating the behavior of the program while collecting data which is reflecting the performance. Unfortunately this technique has the drawback of generating high run-time overheads.

Aspect Oriented Programming: aims to increase modularity by allowing the separation of cross-cutting concerns. Because there is no need to add instrumentation code as the instrumentation facility is integrated within the programming language, the additional behavior is added to existing code without modifying the code itself. AOP adds the following constructs to a program : aspects, join-point, point-cuts and advices. These constructs can be considered like classes. Most popular languages have their aspect oriented extensions like AspectC++ and AspectJ. In python, there are some libraries who aims to reproduce AOP behavior but there isn't any canonical one. Actually there is a debate to what extent aspect oriented practices are useful or applicable to Python's dynamic nature.

2.3 Program Analysis tools

This section is dedicated to the available solutions in terms of program analysis. As it will be explained in the next chapter, the proof-to-concept system will be coded in *Python* and therefore an additional information will be given for solutions available in this language. As already exposed in this order, first, some Static Analysis solutions will be presented following with the dynamic method ones.

2.3.1 Static Analysis tools

Following, some of the most popular tools (commercial or free) for SPA are described, selected in widespread languages : Java, C/C++ and Python. The description are based on the official website of the tools and also on the Gomes et al. [2009] paper.

Starting with C/C++, **Splint** is a very well known tool, allowing to check for security vulnerabilities and coding mystakes. Splint is based on Lint and tries to minimize the efforts needed for its deployment. Additionally, with some annotation, Splint can extend its performances over Lint. Splint can among others detect : Dereferencing a possibly null pointer, Memory management errors including uses of dangling references and memory leaks, Problematic control flow such as likely infinite loops. **Astrée**, where as it is based on abstract interpretation, is analyzing safety-critical applications written or generated in C. It proves the absence of run-time errors and invalid concurrent behavior for embedded applications as found in aeronautics, earth transportation, medical instrumentation, nuclear energy, and space flight. Another worth mentioning tool is the **PolySpace Verifer** tool developped by MathWorks who also created the famour Matlab software.

Concerning Java, one recognized tool is Findbugs. With the advantage of being a Libre software, the tools uses a series of ad-hoc techniques designed to balance precision, efficiency and usability. FindBugs operates on Java bytecode, rather than source code. Another Libre software is **Checkstyle** which, as his names indicates it, allows to report any breach of standards in the source code. Finally a commercial tool, **Jtest** which is an integrated Development Testing solution, can perform Data-flow analysis Unit test-case generation and execution, static analysis, regression testing, runtime error detection, code review, and design by contract.

In the Python world, **Pylint** is a coding standard checker which follows the style recommended by the PEP 8 specification. It is also capable of detecting coding errors and is integrable in IDEs. Speaking of IDEs, **PyCharm** includes also static analysis functions like PEP8 checks, testing assistance, smart refactorings, and a host of inspections.

2.3.2 Dynamic Analysis tools

As for the static tools, the most popular DPA tools are presented here. Following, a table proposed by Gosain and Sharma [2015] with an summary of some available DPA tools regrouped by technique. The table indicates the concerned language and also which type of dynamic Analysis is done by the tool.

Technique	Tool	Language	Type of Dynamic Analysis done								
			Cache Modelling	Heap Allocation	Buffer Overflow	Memory Leak	Deadlock Detection	Race Detection	Object LifeTime	Metric Computation	Invariant Detection
Instr.Based	Daikon	C,C++									✓
	Valgrind	C,C++				✓		✓			
	Rational Purify	C, C++, Java				✓					
	Parasoft Insure++	C,C++		✓		✓					
	Pin	C	✓								
	Javana	Java	✓						✓		
	DIDUCE	Java									✓
AOP Based	DJProf	Java		✓					✓		
	Racer	Java						✓			
VM Profiling Based	Caffeine	Java							✓		
	DynaMetrics	Java								✓	
	*J	Java								✓	
	JInsight	Java				✓	✓		✓		

TABLE 2.2: Dynamic Analysis Tools

Valgrind, Purify and Insure++ are instrumentation based and can automatically detect memory management and threading bugs among with profiling a program in details. While Valgrind is a instrumentation framework for building dynamic analysis tools, the two others are fully-fledged analysis software. **Javana** comes with an easy-to-use instrumentation framework so that only a few lines of instrumentation code need to be programmed for building powerful profiling tools. **Daikon** and **Diduce** are the most known tools for invariant detection and are respectively an offline and online tool. Last but not least, **Pin** is a dynamic binary instrumentation framework developed by Intel. It enables the creation of dynamic program analysis tools and can be used to observe low level events like memory references, instruction execution, and control flow as well as higher level abstractions such as procedure invocations, shared library loading, thread creation and system call execution.

For AOP based tool, the two selected programs are **DjProf** and **Racer**. The first one is a profiler used for the analysis of heap usage and object life-time analysis and the second one is a data race detector tool for concurrent programs.

***J** and **DynaMetrics** are two academical research projects about Virtual Machine profiling and are proposing solution for computing dynamic metrics for Java. The first one, proposed by Dufour et al. [2003], relies on JVMPI, while the second solution, from Singh [2013], relies on the new JVMTI. **JInsight** is for exploring run-time behaviour of Java programs visually and **Caffeine** helps to check conjectures about Java programs.

In addition to this table, some Python tools are also available even if the field seems to not to be really well developed for this programming language. That could be explainable because of the dynamic nature of the language. This might be why the following tools are developed *in* Python but not *for* it. The first tool is **Angr** which is a python framework for analyzing binaries. It focuses on both static and dynamic instrumentation analysis, making it applicable to a variety of tasks. **Triton** is another binaries analyzer framework and proposes python bindings. Its main components are Dynamic Symbolic Execution engine, a Taint Engine, AST representations of the x86 and the x86-64 instructions set semantics, SMT simplification passes, an SMT Solver Interface

2.4 Dynamic Analysis limitations

As the DPA is a quite new research field, it induce ineluctably some drawbacks and limitations. The following table created by Gosain and Sharma [2015] gives a good overview of the different techniques and some drawbacks.

	Instrumentation		VM Profiling	AOP
	Static	Dynamic		
Level of Abstraction	Instruction/Bytecode	Instruction/Bytecode	Bytecode	Programming Language
Overhead	Runtime	Runtime	Runtime	Design and deployment
Implementation Complexity	Comparatively low	High	High	Low
User Expertise	Low	High	Low	High
Re-compilation	Required	Not Required	Not Required	Required

TABLE 2.3: Dynamic Analysis Techniques comparison

This summary shows straightforwardly some limitations of the different Dynamic Analysis techniques. Instrumentation and VM Profiling based techniques induces high Run-time overheads whereas AOP needs heavy design and deployment efforts. While the implementation complexity is rather high for Dynamic Instrumentation and VM Profiling, a strong user expertise is also needed for the first one. Finally recompilation is needed for two on four techniques.

Additionally, the programmer must be aware that the automated tools cannot guarantee the full test coverage of the source code. More over, however how powerful the tools can be, they might yet produce false positives and false negatives. This is why a human code understanding and reviewing is still an absolute necessity.

2.5 Concluding remarks

In this chapter, we tried to summaries some related work about program analysis. After defining what program analysis is, we briefly presented some of the static and dynamic approaches with their respective techniques. However, this is by no means an exhaustive presentation of all the approaches and the reader must be aware that the field is far more complex than that.

To complete this theoretical explanations, we presented some popular tools for both approaches and spoke about some general dynamic analysis limitations. During the redaction of the chapter, it appeared clearly that the DPA field is quite recent and therefore only a few researches were conducted on the subject.

In the next chapter, we will introduce our own contribution with development of the proof-to-concept system.

Chapter 3

Development

"For me, open source is a moral thing."
Matt Mullenweg

In this chapter, we introduce our contribution to the dynamic program analysis. As explained in the introduction the aim is to develop a proof-to-concept system and all the steps to achieve it will be presented in details including the Setup, Data capture model, Data model and its user interface.

3.1 Proposed solution

While working on a growing project, there is always a point where it becomes difficult to keep an eye on all the variables. In order to give the programmer an overview of the variables evolution, this work is aiming to propose a proof-to-concept system which will not only monitor the data evolution, but also give the possibility to compare the gathered data between different runs.

To achieve such a system, the project is going to be separated in three different parts which will constitute the system. First, a data capture model will monitor all the needed variables and their evolution during the execution of the reviewed program. Then a data model will be created and backup procedure will be implemented to store the data in this model. Finally, a web-application will process the extracted data and show them for reviewing the results. Each mentioned part is exposed in the development section.

3.2 Environment

For the project 4 main technologies were chosen in order to develop the required features.

First the data is captured in *Python* with the help of the integrated Debugger Framework. Python is a widely used high level programming language which has seen these last year an increasing enthusiasm around it, especially for web based applications. Thanks to the dynamic nature of Python, which includes a dynamic type system, the real-time collection of object is a pretty straightforward process and therefore it made plenty sense to use it in our project. More over Python offers good compatibility with other programming language since there are a lot of bindings available. For this project, the newest version 3 of the programming language was chosen because of the better handling of encoding.

Secondly, the extracted data is stored in a *MongoDB* Database. MongoDB is a document-oriented database which enters in the new categorie of No-SQL database systems. MongoDB has the advantage to use JSON like documents with schemas which was a clever choice to store the heterogeneous extracted data.

Finally, the used interface was built with the help of *Python*, *Html/CSS* and *Javascript*. As already said, Python is now an interesting language to develop web applications and was used here, with the help of the Flask framework, for the server side process. HTML/CSS and Javascript were used for the presentation of the results.

Additionally, the chosen IDE was PyCharm academic edition version 2015 and then 2016. PyCharm is a very complete IDE which supports among others Python web frameworks, database support, code inspection. In order to optimize the development management, the GitHub online tool was chosen as version control repository. The deployment during the development of the solution was tested on virtual machine server under Ubuntu Server 14.04. The server is provided by the Department of Informatics at the University of Fribourg and is accessible internally at `http://diufpc115.unifr.ch/`. This server will also be used for the experiments in the chapter 5.

In order to deploy regularly the newest version of the ongoing work an automation server named Jenkins was configured. Jenkins was charged to fetch every day the latest prototype on the GitHub repository, create a package of it and install it on the server. If during this process a bug occurred, an e-mail to the interested persons was sent.

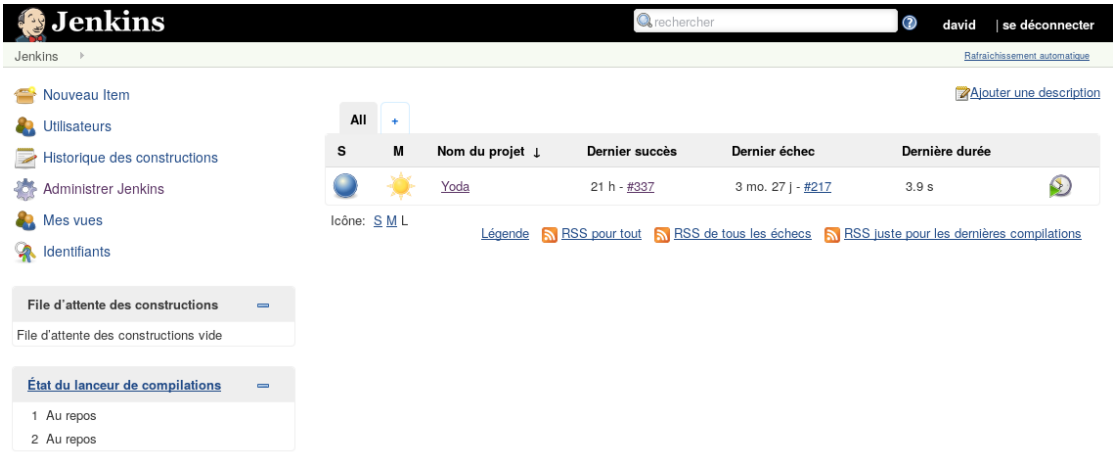

FIGURE 3.1: The Jenkins home page

In the next section, each module of the proposed system will be exposed in details regarding their functionality and their implementations.

3.3 Data Capture Model

This section is presenting the development phase of the data capture model. The data capture model, or *analyser* as it was called during the development, is the core of the system and is based on the Python Debugger Framework (BDB). BDB handles basic debugger functions, like setting breakpoints or managing execution. Thanks to the object-oriented programming, the classes and the function inheritance, it is a straightforward job to rewrite the different functionality as needed for this project.

The development began with study of a script provided by Roman Prokofyev which is implementing some basic data capture functionality derived from the Python debugger framework. The understanding of the developed concepts was the first step to the creation of the data capture model. The analyser consists in 240 lines of code and some of the most important functions are explained here.

3.3.1 Setting up the trace

In order to use the analyzer, some code has to be added at the beginning of the aimed file. The code is necessary to import the module and to set the start of the tracing phase.

```
1 import yoda.analyser
2 yoda.analyser.db.set_trace()
```

The `set_trace()` function is inherited by the BDB and is needed to start debugging with a Bdb instance from caller's frame. It is also absolutely necessary to stop the trace at the end of the aimed code with the `yoda.analyser.db.set_quit()` function which set the quitting attribute to `True`. This raises BdbQuit in the next call to one of the `dispatch_*()` methods.

For further information about the operating of the Python Debugger Framework, we advise the reader to refer to the official documentation [Python-Foundation, 2017].

3.3.2 Initialization

Once the analyser module called, the first step is to setup the `Yoda` class along with some global variables needed during the tracing process. The first variable `json_results` (line 2) will be explained further but is basically where the extracted data will be stored. Then, the `instrumented_types` list (line 3) limits the instrumented objects to this list, it is possible to add further objects if needed. The next 6 variables (line 4-9) are needed for gathering and computing line numbers, frames and files name. Finally, the `next_backup` variable (line 10) defines a limit of how many lines can be analyzed before flushing the information in the database.

```
1  class Yoda(bdb.Bdb):
2      json_results = None
3      instrumented_types = (int, float, str, list, dict)
4      prev_lineno = defaultdict(int)
5      prev_lineno['<module>'] = 0
6      cur_framename = '<module>'
7      file_name = None
8      file_id = None
9      total_linenb = 0
10     next_backup = 1000
```

As the needed variables are now set up, the script continues with the initialization of the `Yoda` class. Within the class, the connection of the database is also created if case of production mode (line 4).

```
1 def __init__(self):
2     bdb.Bdb.__init__(self)
3     if settings.DEBUG is False:
4         mongoengine.connect(settings.MONGODB)
```

3.3.3 Event Catching

BDB can react to various events during the code execution which are handled by 4 functions : `user_call, user_line, user_return, user_exception`. Each function has been rewritten in order to redirect the event to a self-written handling function called `interaction`.

```
1 def user_call(self, frame, args):
2     self.interaction(frame, 'call', None)
3 def user_line(self, frame):
4     self.interaction(frame, 'line', None)
5 def user_return(self, frame, value):
6     self.interaction(frame, 'return', None)
7 def user_exception(self, frame, exception):
8     self.interaction(frame, 'exception', exception)
```

Once the `interaction` function has been called, the first thing to do is to check whenever the `file_name` variable is blank or not. If `file_name` is `None` then a new one is taken and applied from the source code file otherwise the script will continue with the handling of the events.

```
1 if self.file_name is None:
2     self.file_name = inspect.getfile(frame)
```

3.3.4 Event Handling

The first handled event type is the `call` type. This kind of event is normally happening when the frame of the code is changing and thus is really short. Indeed, it just need to capture the frame name (line 2) and catch the line number (line 3). Nothing else special is handled there.

```
if event == 'call':
    self.cur_framename = str(frame.f_code.co_name)
    self.prev_lineno[self.cur_framename] = frame.f_lineno
    self.set_step() # continue
```

Following, the second event type is the the `line` type which occurs at each line-break. This event is the most important for the data collection and its operating has to be explained in separated steps. First, the interaction function checks the type of the event and then proceed to extract the line number which is a key information for the user interface. Then for each line, the interpreted objects have to be caught. This is handled by a external function called `_filter_locals` and called with the frame locals in option.

```
locals = self._filter_locals(frame.f_locals)
```

The function itself create first an empty dictionary which will store the name and the value of each local (line 2). The locals starting with a double underscore are ignored and only the specified object are fetched (line 4 to 6). The function returns the `new_locals` dictionary to the main `interaction` function (line 9).

```
1 def _filter_locals(self, local_vars):
2     new_locals = {}
3     for name, value in list(local_vars.items()):
4         if name.startswith('__'):
5             continue
6         if not isinstance(value, self.instrumented_types):
7             continue
8         new_locals[name] = [copy.deepcopy(value)]
9     return new_locals
```

Then, the locals are stored in a JSON defaultdict object along with the file name, the frame and the line number. At the end, the JSON dictionary is periodically stored in the database in order to flush the data from the memory and enhance the run-time performances. The population of the database is detailed in the next point.

```
if self.total_linenb > self.next_backup:
    self._populate_db()
    self.next_backup += self.next_backup
```

The handling of the `line` event is now finished and interaction function continues with the two last types. The `return` event only occurs at the begging of a file for which we just set the main frame name (line 2) and the `exception` event happens when there is an error in the code which is printed out in the console (line 6).

```
1 if event == 'return':
2     self.cur_framename = '<module>'
3     self.set_step()  # continue
4 if event == 'exception':
5     name = frame.f_code.co_name or "<unknown>"
6     print("exception in", name, exception)
7     self.set_continue()  # continue
```

3.3.5 Trace Ending

Finally, the data capture model is ended by the `set_quit()` BDB function which was remodeled for writing the last traced lines (line 6).

```
1  def set_quit(self):
2      self.stopframe = self.botframe
3      self.returnframe = None
4      self.quitting = True
5      sys.settrace(None)
6
7      if self.json_results:
8          if settings.DEBUG:
9              print(self.json_results)
10         else:
11             self._populate_db()
```

3.4 Data model

The data model is an in-between layer used for the Data capture model and the user interface. Both modules parts will be explained in this section along with the presentation of the data model itself.

3.4.1 Definition

The data model itself evolved a lot during the development and lead to the finale state which will be presented here. This can be observed in the chosen nomenclature, which sometimes does not exactly correspond to the reality. The best example is the use of the substantive ”file” in the code which actually describes more an analysis instance or a run

than the file itself. Thanks to the MongoDB database engine, it is easy to modify the document structure without any database manipulation in opposition with the tabled nature of SQL engines. This was a great asset which allowed tremendous saving time in the development of the data model since it changed a significant number of times.

To understand the data model it is a good reminder to enumerate what the data capture model is actually capturing. First the data capture model is searching for objects, i.e. integer, string, float variables, and their values. These objects are linked with line number, which are them-self linked with frames. Finally, each frame is owned by a file (or more specifically a run as it was pointed out previously).

Keeping that in mind the different data structures can be considered as documents and defined the following way in Python. This notation is used further for reading from the database. First, the *line* which has a number and some data (objects):

```
1 class Line(EmbeddedDocument):
2     lineno = IntField()
3     data = DictField()
```

Secondly, the *frame* which has a name and contains one or many lines :

```
1 class Frame(EmbeddedDocument):
2     name = StringField()
3     lines = ListField(EmbeddedDocumentField(Line))
```

Finally, the *file* which as a name, a time-stamp, the content itself (source code) and additionally a revision number gathered from the git repository when available and also the user name of the person who started the analysis. The file contains logically the different frames and the whole is defined this way :

```
1 class File(Document):
2     user = StringField()
3     revision = StringField()
4     filename = StringField()
5     timestamp = DateTimeField()
6     content = StringField()
7     frames = ListField(EmbeddedDocumentField(Frame))
```

3.4.2 Writing to the database

This part of the database handling is directly implemented in the data capture model along with the capture functionality. The process can be called in two different state of the analyzing phase :

- The program reached the limit of lines and need to flush the gathered data into the database. This occurs inside of the `interaction()` function which has already been described in the previous section.
- The system reached the end of the targeted software and the function `set_quit()` has been called.

Both states induce the call of the `_populate_db()` which is constituted of an **`if...else`** condition. This condition checks whenever it is the first time the system tries to backup the data or not and calls respectively the `_create_new_file()` or the `_update_file()` functions (line 3 and 6).

```
1 def _populate_db(self):
2     if self.file_id is None:
3         self._create_new_file()
4         self._clear_cache()
5     else:
6         self._update_file()
7         self._clear_cache()
```

If the system need to create a new document in the database, as already stated it will call the `_create_new_file()` which is explained here in a simplified and step-by-step version. The complete version of the function includes also a compatibility layer for Python 2 but has been removed here for readability reasons. The first step of the creation of a new entry is to fetch each row of the JSON type dictionary (line 2) where the data has been stored until now and store the data in two different variables (`module_file` and `frames`).

```
1 def _create_new_file(self):
2   for module_file, frames in self.json_results.items():
```

Then, in order to display also the source code in the user interface, the content of the file retrieved (line 1-3) and as all the needed information are already there, the file document type can be created (line=4). The `user` and the `revision` variables are gathered from two function which retrieve the git repository information, but will not be explained in this report.

```
1 file = open(module_file, 'r')
2 file_content = file.read()
3 file.close()
4 item = File(user=self._get_git_username(), revision=self.
      _get_git_revision_short_hash(), filename=module_file,
      timestamp=datetime.now(), content=file_content)
```

The next step is to create the `frame` document (line 2) along side with each `line` document belonging to this frame (line 4). Finally, the frame is linked to the file (line 6) and the file can be saved into the database (line 7). Additionally the variable `file_id`, which was previously defined, is set.

```
1 for name, lines in sorted(frames.items()):
2     frame = Frame(name=name)
3     for lineno, data in sorted(lines.items()):
4         line = Line(lineno = lineno, data = data)
5         frame.lines.append(line)
6     item.frames.append(frame)
7     item.save()
8 self.file_id = item.id
```

Now that a first backup has been created in the database for our run, the system will probably have to update the database with the following analyzed lines . With this end in mind, the `_update_file()` has been implemented and is presented the same way as the foregoing function. First, as in the previous function the JSON dictionary is looped in order to gather the needed data.

```
1 def _update_file(self):
2   for module_file, frames in self.json_results.items():
```

Then for each frame, the system first checks if it is a new frame or not (line 2) and hence create it in the database (line 3-4). Finally the new analyzed lines are created and saved in the database (line 5-7).

```
1 for name, lines in sorted(frames.items()):
2     if not File.objects(id=self.file_id, frames__name=name):
3         frame = Frame(name=name)
4         File.objects(id=self.file_id).update(push__frames=frame)
5     for lineno, data in sorted(lines.items()):
6         line = Line(lineno = lineno, data = data)
7         File.objects(id=self.file_id, frames__name=name).update(
      push__frames__S__lines=line)
```

3.4.3 Reading from the database

Reading from the database exclusively arises in the user interface module. In order to cut down the procedure, the *MongoEngine* library has been chosen. The MongoEngine is a Document-Object Mapper for working with MongoDB from Python. Hence the use of this library gathering the data of a run in the database stands in one line of code.

```
1 file_object = File.objects(id=file_id)
```

This line of code gather the complete data of a run, but thanks to the API of MongoEngine it is also possible to retrieve specifically the needed data.

3.5 User interface

The user interface is a web application which helps the programmer to review the result of the data capture model. It is based on the Python web framework *Flask* intended to be as lightweight as possible. To initialize and start a new web application with the Flask framework only a few lines are needed.

```
1  from flask import Flask, render_template, redirect, url_for,
       flash, jsonify
2  from flask_mongoengine import MongoEngine
3  from flask_debugtoolbar import DebugToolbarExtension
4
5  app = Flask(__name__)
6  app.config['MONGODB_DB'] = settings.MONGODB
7  app.config['SECRET_KEY'] = "xxx"
8  app.debug = True
9  if __name__ == "__main__":
10     app.run()
```

3.5.1 Index page

```
1 @app.route("/")
2 def index():
3     return render_template('index.html', files=File.objects.
      exclude("frames", "content").order_by('-timestamp'))
```

3.5.2 Viewing a file

3.5.3 Comparing files

3.6 Concluding remarks

Chapter 4

Installation guide

"If Microsoft ever does applications for Linux it means I've won."
Linus Torvalds

In this chapter, the complete installation process of the developed script will be presented.

4.1 Setting up the environment

In order to use the developed tool, it is highly recommended to install it on a system providing a GNU/Linux distribution. The tool might work under Windows or MacOS as the used libraries should all be cross-platform, but the software has never been tested under these platforms.

As the main used language is Python you should install it with the following command :

4.2 Use the packaged version

In order to simplify the installation process, a packaged version has been built and is ready to download on the project's GitHub page.

The installation process is really straightforward and since it's a pip package.

4.3 From source code

Chapter 5

Experiments

In this section different type of experients will be conducted in order to test and check the performance of the developped software. In order to test the following variables the same script was used for all experiments.

5.1 Test script and machine

In order to conduct the different experiments, a test script has been chosen.

Lenovo Thinkpad T460p CPU : Intel Core i7-6700HQ @ 2.60GHz x 8 OS : Fedora 25 64bits GPU : Intel HD Graphics 530 RAM : 15.1Gio

5.2 Data extraction analysis

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5.2.1 Memory usage

5.2.2 Run-time overheads

5.3 Database performances

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5.3.1 Some numbers

Speak about db size, memory usage, etc.

5.4 Concluding remarks

Chapter 6

Conclusion

6.1 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

6.2 Future work

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Appendix A

Glossary

AOP Aspect Oriented Programming

AST Abstract Syntax Tree

DPA Dynamic Program Analysis

IDE Integrated development environments

JPDA Java Platform Debugger Architecture

JSON JavaScript Object Notation

JVMPI Java Virtual Machine Profiling Interface

JVMTI Java Virtual Machine Tools Interface

Libre or Free software, is distributed under terms that allow users to run the software for any purpose as well as to study, change, and distribute the software and any adapted versions.

PDB The Python Debugger

pip Pip Installs Packages is a package management system used to install and manage software packages written in Python

SDK Software Development Kit

SMT Satisfiability Modulo Theories

SPA Static Program Analysis

VM Virtual Machine

Appendix B

License of the software

Bibliography

Patrick Cousot. Abstract interpretation, August 2008. URL https://www.di.ens.fr/~cousot/AI/. [Online; accessed 9-January-2017].

Mireille Ducassé and Jacques Noyé. Logic programming environments: Dynamic program analysis and debugging. *The Journal of Logic Programming*, 19:351–384, 1994.

Bruno Dufour, Laurie Hendren, and Clark Verbrugge. * j: a tool for dynamic analysis of java programs. In *Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications*, pages 306–307. ACM, 2003.

Ivo Gomes, Pedro Morgado, Tiago Gomes, and Rodrigo Moreira. An overview on the static code analysis approach in software development. *Faculdade de Engenharia da Universidade do Porto, Portugal*, 2009.

Anjana Gosain and Ganga Sharma. *A Survey of Dynamic Program Analysis Techniques and Tools*, pages 113–122. Springer International Publishing, Cham, 2015. ISBN 978-3-319-11933-5. doi: 10.1007/978-3-319-11933-5_13. URL http://dx.doi.org/10.1007/978-3-319-11933-5_13.

Lukáš Marek, Yudi Zheng, Danilo Ansaloni, Lubomír Bulej, Aibek Sarimbekov, Walter Binder, and Petr Tůma. Introduction to dynamic program analysis with disl. *Science of Computer Programming*, 98, Part 1:100 – 115, 2015. ISSN 0167-6423. doi: http://dx.doi.org/10.1016/j.scico.2014.01.003. URL http://www.sciencedirect.com/science/article/pii/S0167642314000070. Fifth issue of Experimental Software and Toolkits (EST): A special issue on Academics Modelling with Eclipse (ACME2012).

F. Nielson, H.R. Nielson, and C. Hankin. *Principles of Program Analysis*. Springer Berlin Heidelberg, 2004. ISBN 9783540654100.

Flemming Nielson and Hanne Riis Nielson. *Correct System Design: Recent Insight and Advances*, chapter Type and Effect Systems, page 114–136. Springer International Publishing, 1999.

Python-Foundation. Python documentation, pdb — the python debugger¶, 2017. URL https://docs.python.org/3.6/library/pdb.html. [Online; accessed 17-January-2017].

Didier Rémy. Type systems for programming languages, January 2017.

Paramvir Singh. Design and validation of dynamic metrics for object-oriented software systems. 2013.

Wikipedia. Program analysis — wikipedia, the free encyclopedia, 2016. URL https://en.wikipedia.org/w/index.php?title=Program_analysis&oldid=732080552. [Online; accessed 7-January-2017].