

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA ELEKTROTECHNICKÁ

KATEDRA ŘÍDICÍ TECHNIKY



Diplomová práce

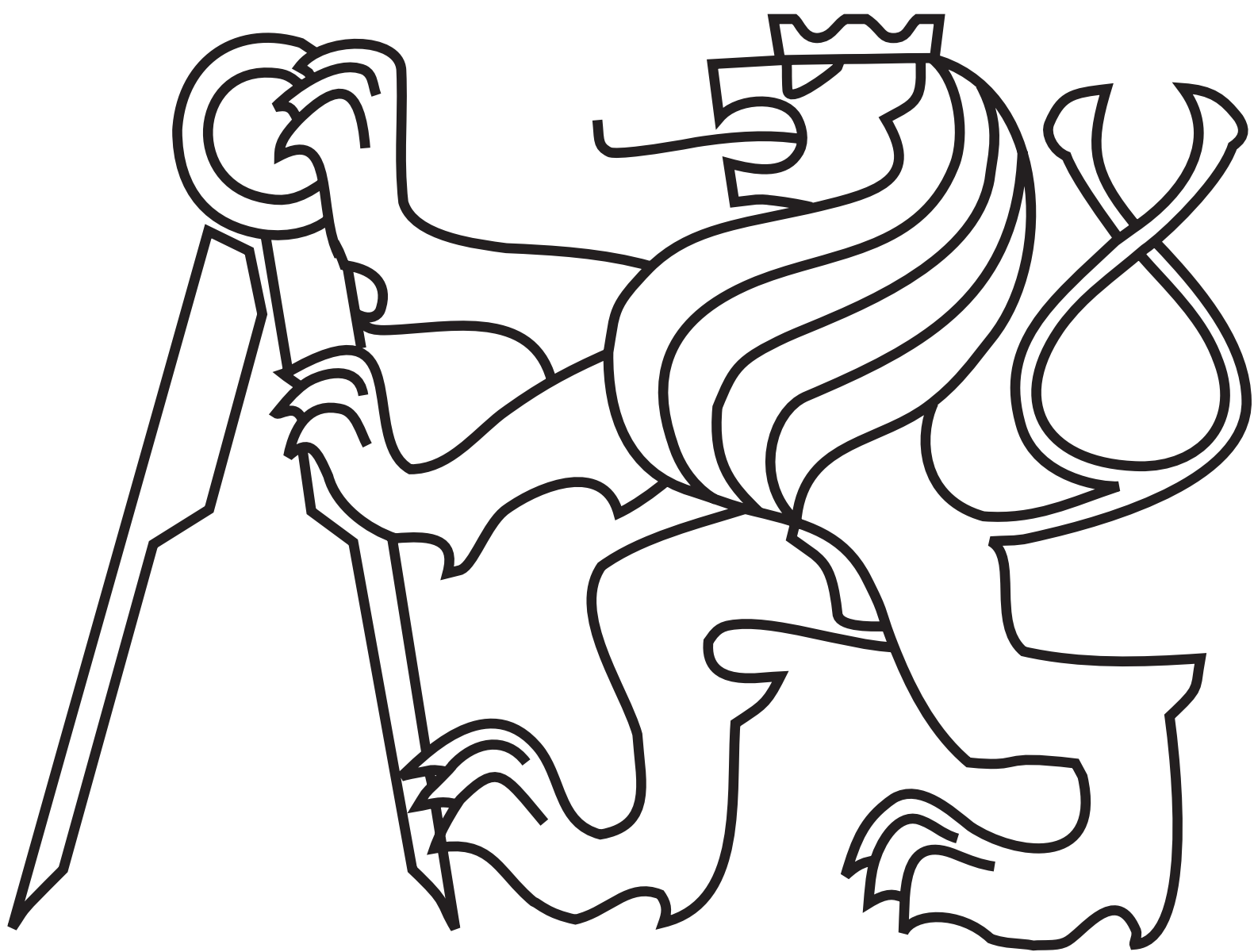
# **Automatizovaný systém stahování webového obsahu potřebného k doplňování cleansetu**

*Michal Staněk*

Vedoucí práce: Ing. Jan Kubr, Ph.D.

12. února 2020







---

## Poděkování

Chtěl bych poděkovat panu Ing. Janu Kubrovi, Ph.D. za odborné vedení mé práce, za pomoc a věcné rady při zpracování této práce.



---

## Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

V Praze dne 12. února 2020

.....

České vysoké učení technické v Praze

Fakulta elektrotechnická

© 2020 Michal Staněk. Všechna práva vyhrazena.

*Tato práce vznikla jako školní dílo na Českém vysokém učení technickém v Praze, Fakultě elektrotechnické. Práce je chráněna právními předpisy a mezinárodními úmluvami o právu autorském a právech souvisejících s právem autorským. K jejímu užití, s výjimkou bezúplatných zákonných licencí, je nezbytný souhlas autora.*

## **Odkaz na tuto práci**

Staněk, Michal. *Automatizovaný systém stahování webového obsahu potřebného k doplňování cleansetu*. Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta elektrotechnická, 2020.



---

# Abstrakt

TODO

**Klíčová slova** Python, Virtualbox, Fiddler, html, js

---

# Abstract

TODO

**Keywords** Python, Virtualbox, Fiddler, html, js



---

# Obsah

<b>1</b>	<b>Úvod</b>	<b>1</b>
1.1	Motivace . . . . .	1
1.2	Cíle práce . . . . .	2
<b>2</b>	<b>Analýza a řešení problému</b>	<b>3</b>
2.1	Výběr programovacího jazyka . . . . .	3
2.2	Předzpracování zadaných adres . . . . .	3
2.3	Získání čistých souborů z webových stránek . . . . .	4
2.3.1	Emulace webového prohlížeče . . . . .	5
2.3.2	Zachytávání komunikace Fiddlerem . . . . .	5
2.4	Zabezpečení stahovacího procesu . . . . .	7
2.5	Nahrání získaného obsahu do databáze cleansetu . . . . .	8
2.6	Začlenění do stávající infrastruktury . . . . .	8
<b>3</b>	<b>Použité technologie</b>	<b>11</b>
3.1	Python 3.7 . . . . .	11
3.2	Fiddler . . . . .	11
3.3	Selenium . . . . .	12
3.4	Jenkins . . . . .	12
3.5	Kafka . . . . .	12
3.6	Virtualbox . . . . .	13
<b>4</b>	<b>Implementace</b>	<b>15</b>
<b>5</b>	<b>Otestování a zhodnocení přínosu</b>	<b>17</b>
	<b>Závěr</b>	<b>19</b>
	<b>Literatura</b>	<b>21</b>



# Úvod

Dnešní doba je plná rizik, která představují hrozbu pro každodenního uživatele internetu. Ať už se jedná o phishing (zisk citlivých údajů pomocí podvodné internetové komunikace) či různé druhy malwaru (nežádoucí programy mající za úkol poškodit uživatele). V boji s těmito riziky je důležité chránit sebe a svoje data pomocí antivirových programů. Jedním z nejrozšířenějších je Avast, který má přes 435 milionů aktivních uživatelů a měsíčně zabránil okolo 2 miliardám útoků[1].

## 1.1 Motivace

Avast, stejně jako většina antivirových programů, uchovává informace o všech známých škodlivých entitách. Tato databáze se denně rozšiřuje o spousty nových záznamů, které obsahují nejen informace o celých souborech, ale i kusy kódu webového obsahu (tzv. string detekce), které jsou považovány za příznak podvodných úmyslů. Může se však stát, že je tento kus kódu moc obecný a dochází tak i k blokování čistého obsahu (tzv. fals-positive detekcím). Aby se těmito situacím předcházelo, je zapotřebí udržovat i databázi s čistými záznamy (tzv. cleanset). Tyto záznamy jsou převážně HTML a js soubory.

Dříve, než se nová string detekce začlení do jádra antiviru, je její obsah porovnán se všemi záznamy na cleansetu a pokud dojde ke shodě (tj. detekční string je součástí nějakého souboru na cleansetu), je tato detekce považována za nevalidní. Tímto dochází k zabránění fals-positive detekcím.

Ideálním stavem je tedy mít záznam o veškerém čistém obsahu internetu, což je samozřejmě nemožné. Avšak čím více záznamů cleanset obsahuje, tím kvalitnější je běh antivirového programu. V současné době dochází k doplňování cleansetu pouze občasné a to převážně manuálně za pomoci jednoduchých scriptů.

### 1.2 Cíle práce

Hlavním cílem této práce je vytvořit plně automatizovaný systém, který bude databázi s čistými záznamy periodicky doplňovat o nový obsah, čímž by mělo dojít ke zlepšení funkčnosti antivirového programu. Dále bude potřeba systém začlenit do již stávající infrastruktury. Primárním úkolem je tedy vytvořit projekt, který by modernizoval doplňování cleansetu, avšak současně je možné systém zobecnit k využití i v jiných aplikacích. Systému bylo dáno kódové označení *Magpie* (česky Straka), protože systém, stejně jako straky, bude shromažďovat data z různých míst a ukládat je na jedno místo. Celý proces bude řádně otestovaný a bude zhodnocen jeho přínos.

# Analýza a řešení problému

Jednotlivé body práce by se daly rozdělit na vícero dílčích podproblémů:

- Výběr programovacího jazyka
- Předzpracování zadaných adres webových stránek
- Získ čistých souborů z webových stránek
- Zabezpečení stahovacího procesu
- Nahrání získaného obsahu do databáze cleansetu
- Začlenění do stávající infrastruktury

## 2.1 Výběr programovacího jazyka

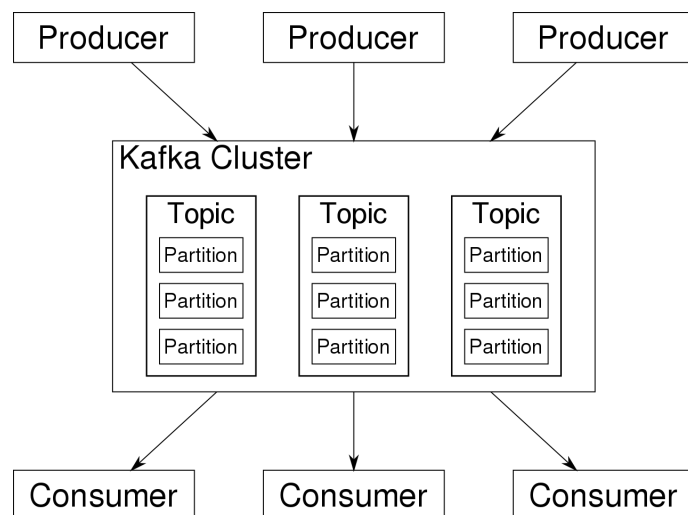
Podle prvních odhadů a požadavků bude práce obsahovat více odlišných částí, které by měly být jednoduše ovladatelné a propojitelné. K tomu by šlo využít programovacího jazyka *Python*(3.1), který obsahuje velké množství dostupných knihoven. Tato volba je i v souladu s firemní politikou.

## 2.2 Předzpracování zadaných adres

Předzpracování zadaných adres se bude řešit pomocí frontového systému *Kafka*(3.5).

Systém by měl reagovat na dva typy vstupů. Prvním vstupem budou adresy s vysokou prevalencí (případem takové adresy může být třeba internetový obchod [www.amazon.com](http://www.amazon.com)), které budou systému periodicky dodávány z externích modulů. Vstupem ale může být i ručně zadaná adresa či seznam adres v případě, kdy by operátor systému potřeboval na cleanset dodat soubory z webových stránek s menší prevalencí, či, ve více obecném řešení, by potřeboval stáhnout zdrojové soubory cílených webových stránek pro odlišné účely.

V obou případech se vložené adresy nahrají do frontového systému, odkud se budou postupně odebírat (Obr.2.1). Využití *Kafky* má výhodu v již naimplementovaném řešení fronty. Jednou z funkcionalit *Kafky*, které by zde šlo využít, je potvrzování zpracované zprávy po přijetí. Tím dojde k zaručení, že se vždy zpracují všechny zprávy ve frontě.



Obrázek 2.1: Využití frontového systému Kafka

### 2.3 Zisk čistých souborů z webových stránek

Jak bylo zmíněno v předchozí kapitole, lze k zisku čistých souborů přistoupit dvěma způsoby. Jednou z možností je otevírat stránky ve webovém prohlížeči a k zisku souborů použít nástroj *Fiddler*, druhým způsobem je emulace webového prohlížeče v *Pythonu* a stahování zdrojových kódů stránek.

Další problematikou je ošetření přesměrovávání (tzv. redirecty), které se na spoustě webových stránek používá. Jedna z možností je redirecty neřešit a zabývat se pouze obsahem dané url. To by proces zisku zdrojových souborů usnadnilo. Není to ale příliš robustní řešení. Bylo by tedy rozumné s přesměrováváním webových stránek počítat.

Také by zde měla být implementována logika selekce pouze HTML a js souborů, ostatní soubory pro funkčnost cleansetu nejsou důležité. To je možné provádět přímo při stahování souborů, nebo vždy pro zdanou adresu stáhnout všechny soubory a selekci provést následně.

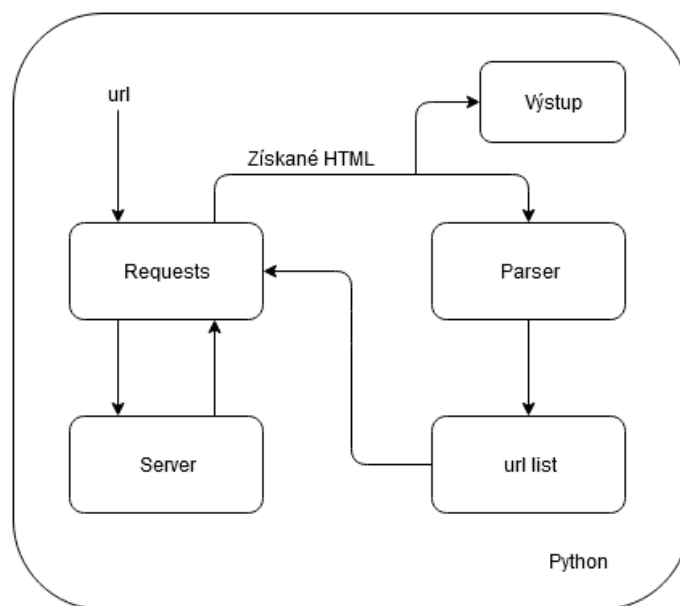
Jednotlivé rozebírané metody zisku čistých souborů z webových stránek jsou tedy následující:

- Emulace webového prohlížeče a následné parsování webových stránek
- Spouštění stránek v prohlížeči a zachytávání komunikace Fiddlerem



### 2.3.1 Emulace webového prohlížeče

Oproti druhé metodě s využitím nástroje *Fiddler* by byla emulace webového prohlížeče bezesporu rychlejší. Pro práci s webovými stránkami v *Pythonu* existuje více knihoven, avšak nejčastěji se používá knihovna *Requests*. Její interface je daleko snazší na použití než u knihovny *Urllib*, jejíž výhodou je pouze fakt, že je již obsažena v základní instalaci *Pythonu* a není nutno ji doinstalovávat. Nevýhoda knihovny *Requests* je v horší práci se stránkami s



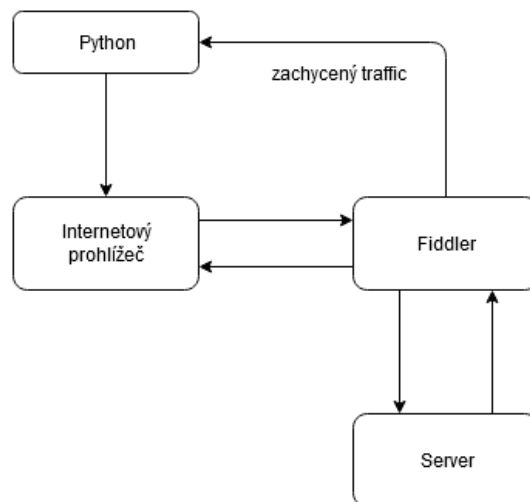
Obrázek 2.2: Diagram zachytávání komunikace pomocí web scrappu

přesměrováváním pomocí javascriptu nebo obecně s načítáním obsahu pomocí javascriptu. Vytvořit komplexní web scrapper (tj. nástroj, který prochází obsah webových stránek), který by dokázal reagovat i na javascriptem řízený obsah, je netriviální úkol (Obr.2.2).

TODO: možná trochu popsat netriviální úkol requests libky, parsování html a hledání js scriptu, zmínit že by nebylo nutné používat virtual

### 2.3.2 Zachytávání komunikace Fiddlerem

Z tohoto důvodu by bylo snazší použít nástroj *Selenium*(3.3), čímž už dojde k určitému zpomalení z nutnosti spouštění internetového prohlížeče, avšak odpadne nutnost implementace sledování redirectů a postupného načítání stránek pomocí javascriptu. Stále je ovšem potřeba načtené stránky nějak zpracovat, což by bylo možné pomocí *Fiddleru*(3.2). Tím sice opět vzniká další zpomalení kvůli spouštění dalšího programu, řešení ale přináší téměř kompletní implementaci rozebraného problému s možnou variací úprav pomocí konfiguračního souboru. Protože je *Fiddler* původně vyvíjen pro testovací účely a



Obrázek 2.3: Diagram zachytávání komunikace Fiddlerem

sledování internetové komunikace (tzv. traffic), zachytává veškerý traffic, který skrze nástroj proudí (Obr.2.3). Bylo by tedy potřeba implementovat logiku pro třídění a následnou selekci HTML a js souborů.

### 2.3.2.1 Selektce HTML a js souborů

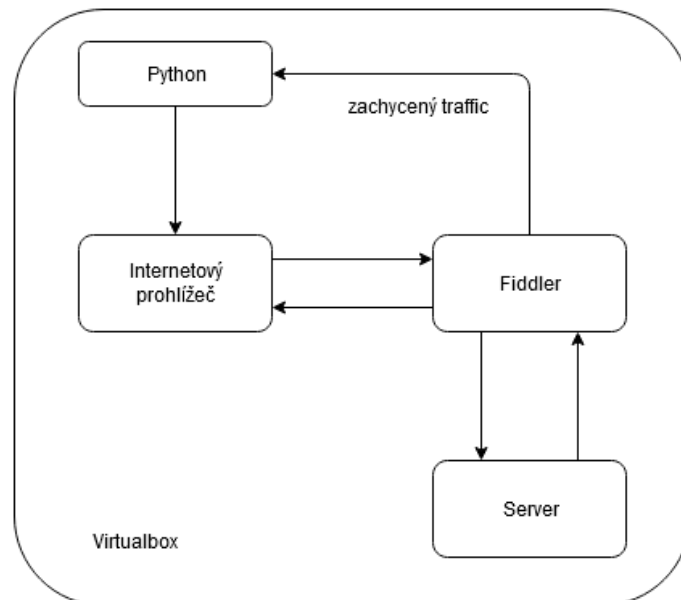
První přístup, emulace webového prohlížeče, přináší výhodu v tom, že se rovnou při parsování HTML stránek přistupuje pouze k js a HTML souborům, čímž odpadá nutnost následně nějakou selekci provádět. Metoda s použitím *Fiddleru* je v tomto složitější. *Fiddler* je komplexní nástroj a automaticky zachytává veškerou komunikaci - nejen soubory potřebné k vykreslení webové stránky, ale i režijní komunikaci mezi prohlížečem a serverem. Tento přístup je možné změnit pomocí již zmíněného inicializačního scriptu.

Avšak problémy této metody přináší i webový prohlížeč, kterého se zde využívá. Téměř vždy při spuštění prohlížeče probíhá nevyžádaná komunikace prohlížeče se servery, která je potřeba vytrždit. I zde je možné využít inicializačního scriptu *Fiddleru*, implementovat třídění již při zachytávání komunikace a zachytávat pouze HTML a js soubory ze serverů, které odpovídají zpracovávané url. To ovšem přináší problémy s přesměrováním, které může seznam chtěných serverů libovolně navyšovat, což už by pro inicializační script představovalo nelehký úkol a takové řešení patrně nebude optimální. Druhý způsob je zachytávat HTML a js komunikaci ze všech serverů a selekci řešit až později při zpracování dat v *Pythonu*. Nedostatkem této možnosti je zvýšená náročnost na paměť, kterou představuje zpracovávání více souborů. Předpokládá se ale, že tato nevyžádaná komunikace bude v poměru s chtěnými soubory minimální, tudíž by k výrazné zvýšení náročnosti na paměť dojít nemělo.

## 2.4 Zabezpečení stahovacího procesu

Z bezpečnostních důvodů je kladen důraz na to, aby byl veškerý proces stahování webového obsahu zabezpečen. Předpokládá se sice, že všechny stažené soubory budou nezavirované, avšak spoléhat se na to není moc bezpečné řešení. Jedním z možných způsobů jak docílit základní bezpečnosti je vytvoření a obsluha virtuálního stroje, na kterém by docházelo ke stahování souborů.

Celý koncept zabezpečení stahování pomocí virtuálního systému je ovlivněn použitím nástroje *Fiddler*(3.2) pro stahování souborů z webových stránek. Tato metoda by využívala *Fiddler*, který by běžel na pozadí, spolu s prohlížečem, v kterém by byly stránky otevírány a pomocí *Fiddleru* zachytávána komunikace (Obr.2.4). Právě otevírání stránek v prohlížeči, během



Obrázek 2.4: Zaobalení stahovacího procesu virtuálním strojem

čehož dochází ke spouštění javascriptových souborů, je z pohledu bezpečnosti potenciálně nebezpečné. Toho by se dalo vyvarovat emulací prohlížeče přímo v *Pythonu*, čímž by se stahovaly rovnou zdrojové kódy webových stránek bez nutnosti jejich spouštění. Tato metoda však přináší problémy, které již byly popsány v subsekcí 2.3.1.

Pokud se jedná o nástroje umožňující virtualizaci systému, lze použít *VMware workstation* od firmy VMware nebo *Virtualbox* vyvíjený firmou Oracle. Firma VMware nabízí pro práci se svými virtuálními stroji infrastrukturu nazývanou *vSphere*. Tato infrastruktura obsahuje vlastní SDK nástroje pro implementaci do programovacího jazyka *Python*[7], avšak celé toto řešení se nenabízí s freeware licencí. Z tohoto důvodu by bylo lepší použít virtualizační nástroj *Virtualbox*(3.6), který je zdarma. Nevýhoda *Virtualboxu* spo-

čívá v horší podpoře *Pythonu*. Existuje ale knihovna *pyvbox*[8], která základní funkčnost podporuje.

Z hlediska bezpečnosti by bylo rozumné spouštět virtuální stroj (neboli VM) pro každou webovou stránku zvlášť, to by ale celý proces výrazně zpomalovalo. Předpokládá se, že nejdelší dobu zabere právě startování VM. Jiné řešení by nabízelo restartovat virtuální stroj vždy po určitém čase nebo po daném množství zpracovaných url adres. Tím by se běh systému výrazně zrychlil.

### 2.5 Nahrání získaného obsahu do databáze cleansetu

Dalším krokem bude přesun získaného obsahu do samotné databáze cleansetu. Ta je součástí většího systému aplikací, který nese interní označení Scavenger. Zjednodušeně lze říci, že tento systém obsahuje záznamy o všech antiviru známých souborech a url adresách (nakažených i čistých). Soubory se v systému Scavenger ukládají v podobě hashe vzniklé pomocí hashovacího algoritmu sha-256. Tímto lze docílit jednoduché kontroly duplicity (stejně soubory mohou mít rozdílné názvy, ale hash souboru je pro identické soubory stejná). K hashi souboru se přikládají metadata mimojiné s informací o původu souboru, času výskytu a prevalenci.

K nahrání souborů do Scavengeru lze využít síťový souborový systém Sambu, který implementuje přenos souborů po síti pomocí síťového protokolu SMB a to převážně v systémech Windows. Tato metoda je však z hlediska firemní infrastruktury zastaralá. Novější způsob představuje využití datové platformy HDP (Hortonworks data platform), která je založená na technologiích od společnosti Apache - Hadoop, Hive a Spark. Tato platforma se specializuje na přesun a zpracování velkého množství dat z různých zdrojů. Ke komunikaci se zmíněným systémem HDP lze využít interně vyvinutý python klient, který přesun souborů usnadní.

### 2.6 Začlenění do stávající infrastruktury

Systém bude spouštěn periodicky, ale měl by být také spustitelný na vyžádání uživatelem. Pro takové požadavky lze použít systém Jenkins(3.4), který je firmou Avast používaný k periodickému spouštění procesů. Jednotlivé části systému Magpie je možné oddělit do samostatných procesů (v terminologii systému Jenkins tzv. jobů), které se dají sekvenčně pouštět v závislosti na úspěšném ukončení předcházejícího jobu. Tento přístup přináší přehledné rozhraní, v kterém je možné jednotlivé části samostatně monitorovat, spolu s jednoduchým přístupem k výstupům jobů. Tímto způsobem by bylo možné přistoupit k získaným datům přímo, bez nutnosti data nahrávat do databáze v systému Scavenger, v případě, kdy by byl systém spuštěný manuálně.

Jiným přístupem je využít firemní mutaci nástroje Kubernetes interně nazývanou Luft, která by umožňovala mít systém sputněný bez přestávky. Velkou výhodou Kubernetes je jednoduché škálování, kdy lze při velkém vytížení jednoduše navýšit výpočetní prostředky danému procesu a tím urychlit jeho běh.

Je také možné pro jednotlivé části systému Magpie využít rozdílné technologie, a to kombinací obou výše zmíněných.



## Použité technologie

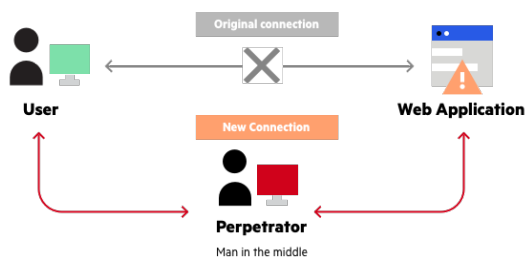
V této kapitole jsou stručně popsány všechny technologie využité při zpracování této práce.

### 3.1 Python 3.7

Python je skriptovací programovací jazyk, jehož syntaxe je lehce odlišná od konvenčních programovacích jazyků (Java, C) v tom, že nepoužívá středníky ani složené závorky. Jedná se o hybridní programovací jazyk, což znamená, že program nemusí být nutně objektově orientovaný, ale části mohou mít více procedurální charakter. Tím dochází k lepší čitelnosti kódu a celkovému zjednodušení. Síla Pythonu je i ve velkém množství balíků s knihovnami, které podporují jeho všestrannost. Kvůli těmto vlastnostem byl vybrán pro tuto diplomovou práci.

### 3.2 Fiddler

Fiddler je nástroj vyvíjen firmou Telerik, sloužící k zachytávání internetové komunikace. Funguje na principu MitM (Man-in-the-middle) útoku, kdy se útočník vtěsná mezi dva účastníky internetového provozu a nechá je komunikovat skrz sebe. Zde je však tento útok chtěný (Obr.3.1). Jeho automatizace



Obrázek 3.1: MitM útok [3]

lze docílit inicializačním souborem, který obsahuje různá pravidla a je psaný v javascriptu. Při správném nastavení je fiddler schopný zachytávat i šifrovanou komunikaci, kvůli čemuž byl použit v této práci.

## 3.3 Selenium

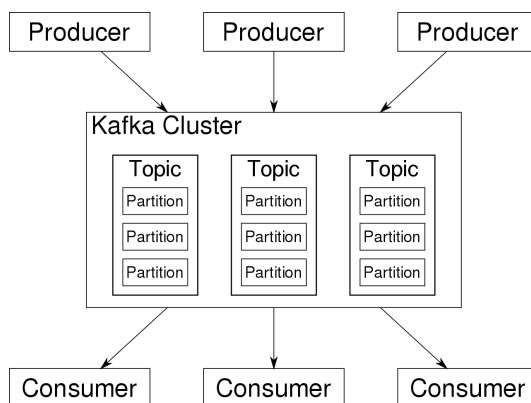
Selenium je opensource nástroj používaný k automatizovanému přístupu k webovým aplikacím. Těchto vlastností se často využívá při testování, avšak v této práci je použit pouze k obsluze webového prohlížeče. Selenium obsahuje vlastní vývojové prostředí, které lze využít bez velké znalosti programování, existují však i jeho implementace do většiny populárních programovacích jazyků.

## 3.4 Jenkins

Jenkins je opensource CI/CD systém (continuous integration and delivery) umožňující vykonávání automatických či periodických operací (tzv. jobů). Používá se převážně k spouštění buildů a udržování testů. Joby lze spouštět automaticky po vzniku nové verze vyvíjeného programu, lze je ale i spouštět pravidelně v předem určený čas, čehož bylo v této práci využito.

## 3.5 Kafka

Kafka by se dala zařadit mezi frontové systémy. Jedná se o opensource platformu sloužící ke streamování dat v reálném čase. Její zaměření je převážně na procesing velkého množství zpráv bez narůstající latence. Zprávy se však neukládají do fronty, jako je tomu třeba u RabbitMQ (tj. jiný frontový systém), avšak do tzv. topiců. Jeden Topic může obsahovat více částí (tzv. partition), do kterých se zprávy distribuují (Obr.3.3). Zprávy do topicu posílá proces,



Obrázek 3.2: Znázornění Kafka systému [4]

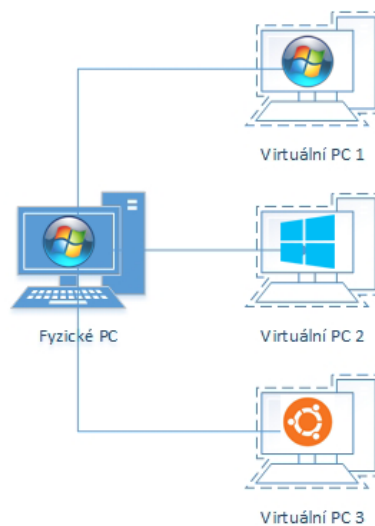


který se nazývá Producer. Obdobně proces, který data čte, je nazýván Consumer. Ten dostává zprávy z topicu v závislosti na indexaci a časové známky. K jednomu topicu může být přihlášeno více nezávislých konzumerů, kteří jsou rozděleni do rozdílných skupin a každá skupina dostává identické zprávy, čím se zabrání vzájemnému čtení stejných zpráv.

Zprávy zůstávají v topicu po určitou dobu, což určuje hodnota retence. Po tuto dobu jsou zprávy přístupny pro každou skupinu konzumerů. Síla kafky je v politice přečtených zpráv. Na rozdíl od zmiňovaného RabbitMQ, který přečtením zprávy zprávu odstraní ze své fronty, funguje v kafce tzv. potvrzovací systém. Konzumer dostane zprávu, a po jejím zpracování vyšle tzv. commit, kterým oznámí úspěšné zpracování. Pokud by v průběhu nastala chyba, tento commit se nepošle a zpráva se vrátí zpátky, kde může být přečtena jiným konzumerem ve skupině.

### 3.6 Virtualbox

Virtualbox je opensource virtualizační nástroj vyvíjen firmou Oracle. Slouží k instalaci virtuálních operačních systémů na jednom fyzickém stroji. Jeho výhodou je multiplatformnost, což znamená, že je možné ho nainstalovat na MS Windows i operační systémy s unixovým jádrem (Linux, Mac OS). Tímto lze docílit například spuštění linuxového systému pod operačním systémem Windows.[5]



Obrázek 3.3: PC s virtuálními stroji [6]

Další důležitou funkcionalitou Virtualboxu jsou tzv. snapshoty. Snapshot zachycuje virtuální stroj a veškeré jeho nastavení v daném čase, v kterém je vytvořen. Tímto lze jednoduše vrátit provedené změny zpět do bodu vytvo-

### 3. POUŽITÉ TECHNOLOGIE

---

ření snapshotu. Toho je možné využít při testování aplikací, využití je avšak možné i v oblasti bezpečnosti. Pokud je vytvořen snapshot čistého systému, lze se do něj vrátit v případě potenciálního nakažení virtuálního stroje. Těto funkcionality se využívá i v této práci.

# Implementace



## **Otestování a zhodnocení přínosu**



---

## **Závěr**





---

## Literatura

- [1] *Avast corporate factsheet*, Dostupné z:  
[https://cdn2.hubspot.net/hubfs/2706737/media-materials/corporate-factsheet/Avast\\_corporate\\_factsheet\\_A4\\_en.pdf](https://cdn2.hubspot.net/hubfs/2706737/media-materials/corporate-factsheet/Avast_corporate_factsheet_A4_en.pdf)
  
- [2] Moravec Jan. *Distribučované řízení kolon vozidel na autodráze*. ©2014, České vysoké učení technické v Praze, vedoucí práce Ing. Ivo Herman, Dostupné z:  
<https://dspace.cvut.cz/bitstream/handle/10467/24299/F3-BP-2014-Moravec-Jan-prace.pdf>
  
- [3] <https://www.imperva.com/learn/application-security/man-in-the-middle-attack-mitm/>
  
- [4] [https://en.wikipedia.org/wiki/Apache\\_Kafka#/media/File:Overview\\_of\\_Apache\\_Kafka.svg](https://en.wikipedia.org/wiki/Apache_Kafka#/media/File:Overview_of_Apache_Kafka.svg)
  
- [5] <https://www.virtualbox.org/manual/ch01.html>
  
- [6] <https://www.virtualnipc.cz/wp-content/gallery/140701-1-uvod-do-virtualizace-na-desktopu/cache/140701-uvod-do-virtualizace-na-desktopu-img-1.png-nggid013-ngg0dyn-640x480x100-00f0w010c010r110f110r010t010.png>
  
- [7] [<https://code.vmware.com/web/sdk/6.7/vsphere-automation-python>]
  
- [8] Dorman Michael. *pyvbox Documentation*. ©2017 Dostupné z:  
<https://buildmedia.readthedocs.org/media/pdf/pyvbox/latest/pyvbox.pdf>



## Obsah přiloženého CD

slotcar-sw.....	adresář s Java projektem
SimulinkControllers	
├─ SISO.....	jednovstupový regulátor
├─ TwoInputSingleOutput .....	dvouvstupový regulátor
├─ MISO .....	vícevstupový regulátor
└─ transferscript.bat .....	skript pro přenos souborů
text	
├─ thesis.pdf .....	text práce ve formátu PDF
├─ thesis.tex .....	text práce ve formátu $\text{\LaTeX}$
└─ pictures .....	zdrojové obrázky pro formát $\text{\LaTeX}$
video	
└─ tutorial.mp4.....	instruktační video