

Využití JupyterLab pro interaktivní analýzu nejen kyberbezpečnostních dat

Seminář o bezpečnosti sítí a služeb 2024

Milan Čermák

CSIRT-MU, Masarykova univerzita, Brno



MUNI
CSIRT-MU

Whoami

LinkedIn – <https://www.linkedin.com/in/cermmik/>

Google Scholar – <https://muni.cz/go/cermak-scholar>



- Analytik a výzkumník v oblasti analýzy kyberbezpečnostních dat
- Vedoucí analytického oddělení kyberbezpečnostního týmu MU – [CSIRT-MU](#)
- Řešitel národních a mezinárodních výzkumných projektů spolupracující s analytiky a vyšetřovateli trestné činnosti, NÚKIB a mnoha dalšími výzkumnými týmy a společnostmi
- Ph.D. na Fakultě informatiky Masarykovy univerzity (obhájeno v roce 2020)
- Vyučující předmětu [Network Forensics](#) a kurzu Network Detection and Response (již brzy)
- Odborné zájmy – forenzní analýza digitálních a síťových dat, tvorba analytických systémů využívající moderní technologie (viz [Granef](#)) a web hacking

Pokud vás cokoliv z toho zajímá, tak si s vámi rád popovídám u kávy 😊

Každý větší incident by měl být spojený s analýzou dostupných dat, tak abychom zjistili všechny dopady

- Typicky provádíme stejné kroky, jen s menší obměnou na základě podstaty incidentu
- Velmi často nemáme dostatek času vše detailně prozkoumat a ověřit všechny hypotézy

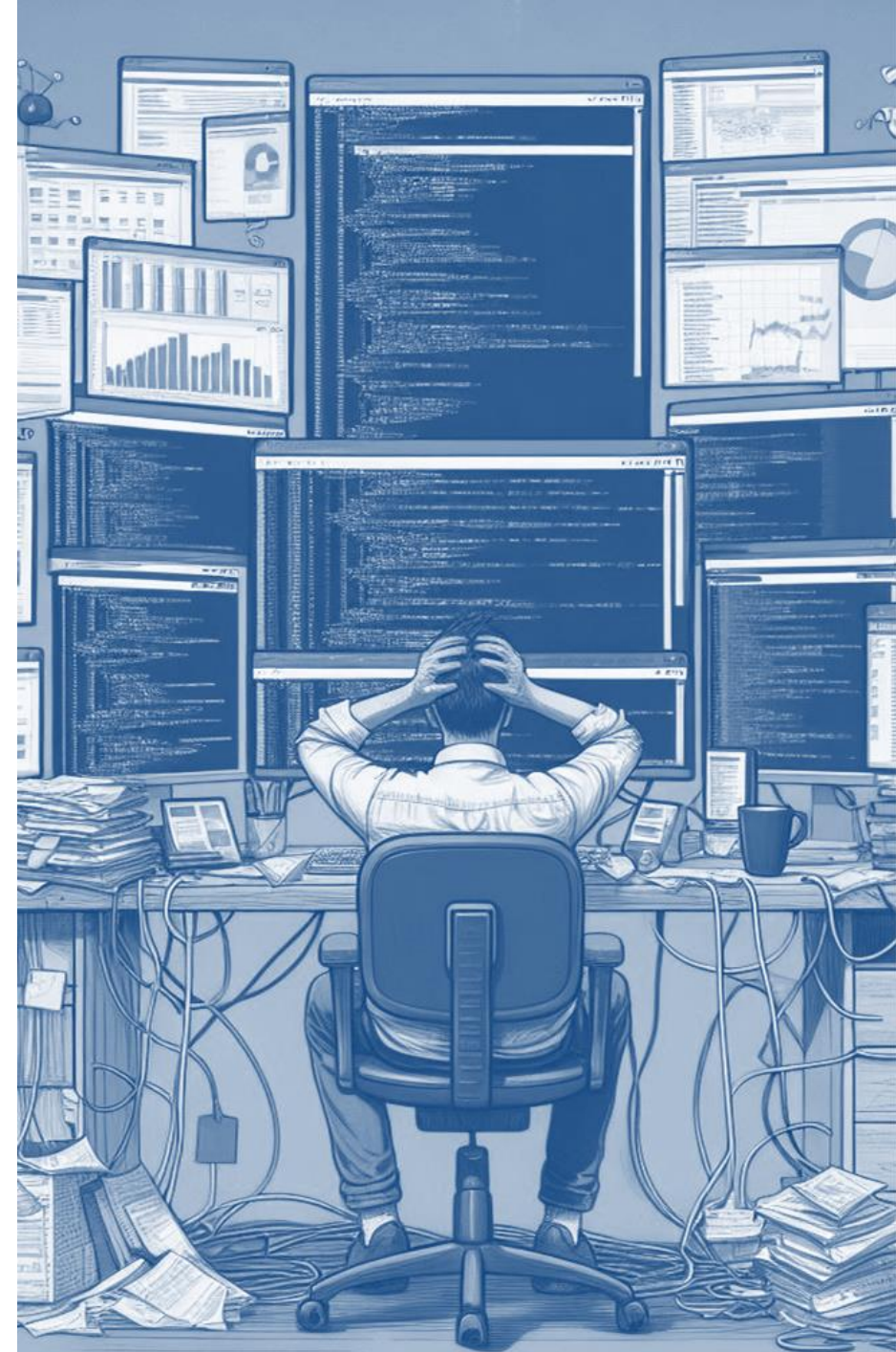
Typický průběh analýzy dat

Scénář: Server v naší síti skenuje své okolí a hádá hesla na RDP



Těžký život analytika

- Často se opakující rutinní analytické postupy
- Dlouhé čekání na výsledky (obzvláště při dotazech na delší časové úseky)
- Používání různých analytických nástrojů s rozdílnou strukturou dat a odlišným uživatelským rozhraním
- Složitá korelace dat z různých zdrojů (jiné formáty dat, nesynchronizované časové známky, ...)
- Velké množství analytických úkolů s požadavkem na co nejrychlejší řešení
- Obtížné sdílení know-how s ostatními analytiky
- Potřeba zpracování výsledků a poznatků do jednoduše pochopitelné formy



V datové analytice se typicky používá nástroj **Jupyter Notebook**, tak proč jej nevyužít i pro naše potřeby?

- Snadné programování v jazyce Python s využití knihoven pro zpracování a vizualizaci dat
- Univerzální řešení umožňující jednoduché sdílení prováděných analýz

Jupyter Notebook a JupyterLab

Jednoduché prostředí umožňující psát analytický kód v jazyce Python, dokumentovat v Markdown a zobrazit výsledky v textové i grafické formě. Vše v rámci jednoho souboru, který je možné jednoduše sdílet.



Odkaz na projekt: <https://jupyter.org/>

- [JupyterLab](#) je webové grafické prostředí umožňující spouštět a zobrazit Jupyter Notebooky (víceuživatelské prostředí je realizováno pomocí [JupyterHub](#))
- Pro jednoduchou instalaci na serveru je možné využít projekt [The Littlest JupyterHub](#), pro nasazení v Kubernetes projekt [Zero to JupyterHub](#)
- Typicky využívané knihovny: [Pandas](#), [Numpy](#), [Plotly](#), [Bokeh](#)
- Pěkný úvod do Jupyter Notebooků nabízí série článků od Pavla Tišnovského na Root.cz: <https://www.root.cz/clanky/jupyter-notebook-nastroj-pro-programatory-vyzkumniky-i-lektory/>

Výhody Jupyter Notebooků

- Python ve spojení s vhodnými knihovnami umožňuje **rychlý vývoj metod** pro zpracování heterogenních dat a jejich korelaci
- Možnost jednoduše změnit proměnnou nebo blok kódu a **spustit jej znovu** při zachování předcházejících výsledků výpočtu
- Lze vytvořit **analytický playbook** se všemi kroky pro danou analýzu s možností dokumentování jednotlivých kroků
- Většina databází, služeb a nástrojů má **dostupný Python modul** nebo umožňuje dotazování skrze API
- Při správném naprogramování je možné **provést celou analýzu na jedno kliknutí** a nechat běžet i několik dní
- Jupyter Notebook je možné **jednoduše sdílet** včetně uložených výsledků analýz

Number of observations that no IP not communicated at all

```
In [5]: noncommunicative_obs = df[df.isnull().all(axis=1)].index
noncommunicative_obs = noncommunicative_obs.shape[0]
print("Noncommunicative Observations: ", noncommunicative_obs)
print("Percentage of total Observations: ", noncommunicative_obs/df
```

```
Noncommunicative Observations: 1329
Percentage of total Observations: 15.17123287671233
```

Number of IPs that have not communicated at all

```
In [6]: noncommunicative_ips = df.columns[df.isnull().all()]
noncommunicative_ips = noncommunicative_ips.shape[0]
print("Noncommunicative IPs: ", noncommunicative_ips)
print("Percentage of total IPs: ", noncommunicative_ips/df.shape[1]*
```

```
Noncommunicative IPs: 703
Percentage of total IPs: 1.07269287109375
```

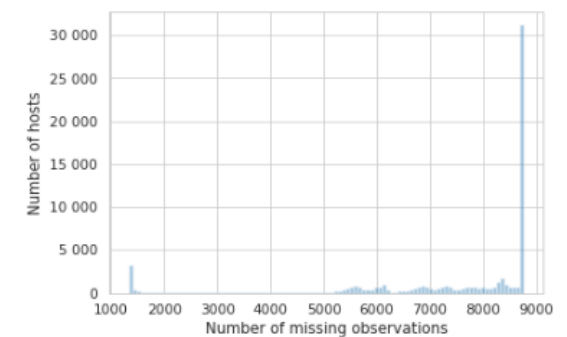
Distribution of the number of missing observation over the whole network

```
In [7]: sns.set_style("whitegrid")
sns.set_context("paper", font_scale=1.2)

# Plot the data
sns_plot = sns.distplot(df.isnull().sum(), bins=100, hist_kws={'cumulat

# Adjust plot properties
sns_plot.set(ylabel = "Number of hosts", xlabel="Number of missing ob
ylabls = ['{:,.0f}'.format(y).replace(', ', ' ') for y in sns_plot.ge
sns_plot.set_yticklabels(ylabls)

#sns_plot.figure.savefig('./fig/NA_distr_big.svg')
plt.show()
```



Teoreticky máme nástroj řešící většinu problémů současných analýz... jen **uživatelská přívětivost** pokulhává

- V rámci analýzy je potřeba vždy nějakým způsobem měnit kód (byť jen hodnotu proměnné)
- Neznalý člověk se v celém prostředí nevyzná a může být zmatený z některých funkcí

Jupyter Widgets a interaktivní prostředí

- Široká nabídka [interaktivních widgetů](#), které nám nabízí možnost interakce v Jupyter Notebooku bez potřeby zasahovat do kódu
- Spojení [output a button widgetů](#) nám umožňuje jednoduše měnit již zobrazený obsah
- Pro zobrazení tabulek doporučuji využít navíc knihovnu [itables](#)

Pick a Time: 06.02.2024 12:30

OS: Linux, Windows, macOS

String: Hello World

Click me

Pizza topping:
 pepperoni
 pineapple
 anchovies

Loading: [Progress bar]

Show 2 entries

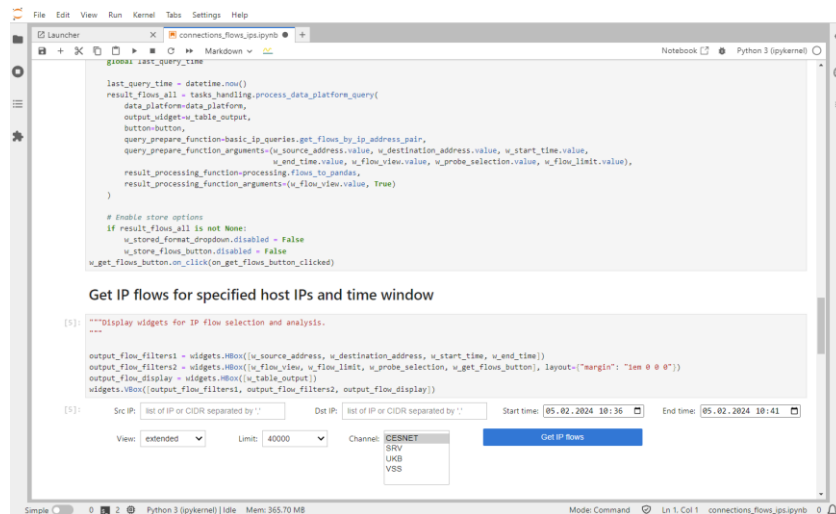
code	region	country	capital	longitude	latitude	flag
CZ	Europe & Central Asia	Czech Republic	Prague	14.4205	50.0878	

Showing 1 to 1 of 1 entries (filtered from 208 total entries)

Previous 1 Next

Voilà aneb zbavme se zobrazení kódu

- Rozšíření [Voilà](#) umožňuje spustit Jupyter Notebook a **zobrazit pouze jeho výstupy** (dokumentace v Markdown, widgety a jakékoliv další výstupy provedených funkcí)
- V kombinaci s Jupyter Widgets nám umožňuje vytvářet **interaktivní analytické dashboardy a playbooky**, které jsou jednoduše použitelné i pro méně znalé uživatele
- **Tip:** Prohlížeč souborů v JupyterLab je možné nastavit tak aby defaultně zobrazoval Jupyter Notebook pomocí [Voilà preview](#) namísto zobrazení Notebooku s kódem



```
from ast import Query

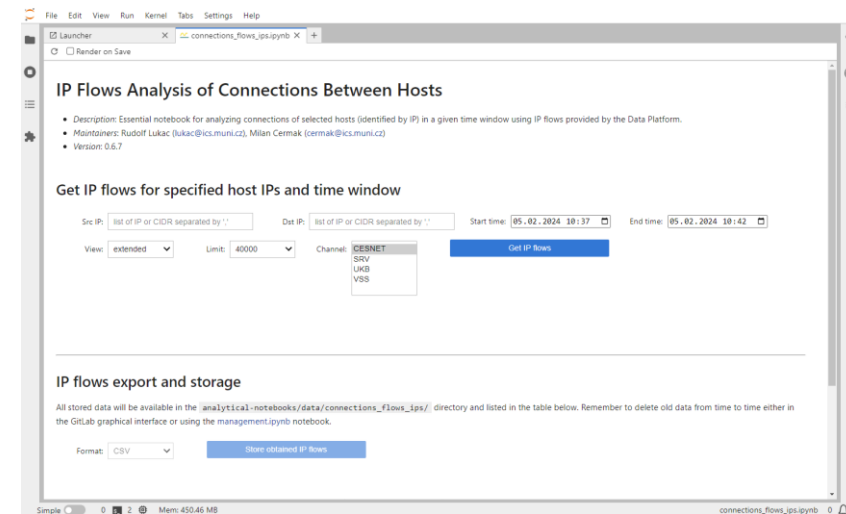
last_query_time = datetime.now()
result_flows_all = tasks.handling_process_data_platform_query(
    data_platform_data_platform,
    output_widget_table_output,
    button_get_ip_flows,
    query_prepare_function_basic_ip_queries.get_flows_by_ip_address_pair,
    query_prepare_function_arguments=(w_source_address.value, w_destination_address.value, w_start_time.value,
    w_end_time.value, w_flow_limit.value, w_probe_selection.value, w_flow_limit.value),
    result_processing_function_processing_flows_to_pandas,
    result_processing_function_arguments=(w_flow_view.value, True)
)

# Enable store options
if result_flows_all is not None:
    w_store_format_dropdown.disabled = False
    w_store_flows_button.disabled = False
w_get_ip_flows_button.on_click(on_get_ip_flows_clicked)

Get IP flows for specified host IPs and time window

[5]: """Display widgets for IP flow selection and analysis.
"""
output_flow_filters1 = widgets.HBox([w_source_address, w_destination_address, w_start_time, w_end_time])
output_flow_filters2 = widgets.HBox([w_flow_view, w_flow_limit, w_probe_selection, w_get_ip_flows_button], layout={"margin": "1em 0 0"})
output_flow_display = widgets.HBox([w_table_output])
widgets.VBox([output_flow_filters1, output_flow_filters2, output_flow_display])

[5]: Src IP: list of IP or CIDR separated by ',' Dst IP: list of IP or CIDR separated by ',' Start time: 05.02.2024 10:36 End time: 05.02.2024 10:41
View: extended Limit: 40000 Channel: CESNET
SRV
UKB
VSS
[Get IP Flows]
```



IP Flows Analysis of Connections Between Hosts

- Description: Essential notebook for analyzing connections of selected hosts (identified by IP) in a given time window using IP flows provided by the Data Platform.
- Maintainers: Rudolf Lukac (lukac@ics.muni.cz), Milan Cermak (cermak@ics.muni.cz)
- Version: 0.6.7

Get IP flows for specified host IPs and time window

Src IP: Dst IP: Start time: End time:

View: Limit: Channel:

IP flows export and storage

All stored data will be available in the `analytical-notebooks/data/connections_flows_ips/` directory and listed in the table below. Remember to delete old data from time to time either in the GitLab graphical interface or using the `management.ipynb` notebook.

Format:

Interaktivní Jupyter Notebooky nám skvěle pomáhají při analýze incidentů, jen jsme museli **vyřešit pár komplikací**

- Vše je dostupné v rámci jednoho GitLab repozitáře, který si uživatel v JupyterLab stáhne
- Pro námi používané služby máme vytvořeny konektory, přičemž uživatel musí pouze vyplnit autentizační údaje v konfiguračního souboru

Správa pomocí Management Notebooku

Problémy v použitelnosti

- Aktualizace pomocí git pull je zbytečně složitá
- Jupyter Notebook může běžet jen v jedné instanci
→ lze obejít pomocí vytváření kopií (duplikování)
- Potřeba průběžného mazání exportovaných dat



Management Jupyter Notebook

- Uživatel jen jednou stáhne celé prostředí a dál už využívá poskytnuté funkce
- Automatické řešení nekonzistence repositáře pokud uživatel udělá nějaké změny

File Edit View Run Kernel Tabs Settings Help

Launcher x management.ipynb x +

Render on Save

Analytical Notebooks Management

- Description: Easy-to-use functions for managing analytical Jupyter notebooks in this repository.
- Maintainers: Milan Cermak (cermak@ics.muni.cz)
- Version: 0.3.4

Repository update

Update all the files in the repository to the latest version. It automatically reverts all changes and resets the laptops. After completion, **open notebooks must be reloaded for the changes to take effect** (if you are asked which file version to use, select the "Revert" option).

Check for repository updates Update the repository

```
2024-02-05 12:33:56 Analytical_Notebooks_Management [INFO]: Checking for repository updates...
2024-02-05 12:33:56 Analytical_Notebooks_Management [INFO]: The repository is up to date
```

Delete duplicated files

Removal of files that were created as copies of the original ones to allow simultaneous view and run in the JupyterLab.

Check for duplicated files Delete selected duplicated files

Select duplicated files to delete:

incident_handling/connections_flows_ips-Copy1.ipynb

Delete stored data files

Clear "./data" directory with exported data files.

Check for data files Delete selected data files

Simple 0 2 Mem: 748.19 MB management.ipynb 0

Typický analytický playbook

- Jupyter Notebook **obsahuje vše** co potřebuje analytik při řešení incidentu
- Vše řízeno přes **jednoduché vstupy** (text, časové rozmezí, selektor, ...) a tlačítka spouštějící danou akci
- Každý krok lze **vysvětlit a zdokumentovat**
- V rámci jednoho úkonu se mohou **korelovat data z více zdrojů**
- Možnost **propojení externích služeb** (např. VirusTotal) bez nutnosti přepínání mezi okny
- Velmi **rychlý vývoj a prototypování**

File Edit View Run Kernel Tabs Settings Help

Launcher x visited_domains.ipynb x visited_domains.ipynb x +

Render on Save

Visited Domains Analysis

- *Description:* Notebook for analysing domains visited by selected hosts using lists of most common domains and VirusTotal.
- *Maintainers:* Rudolf Lukac (lukac@ics.muni.cz)
- *Version:* 0.2.3

Get visited domains and evaluate them using the top domain lists

IP: Start time: End time:

DOMAIN	SUBDOMAIN	DOMAIN IN TOP 10K LIST	DOMAIN IN TOP 1M LIST	SUBDOMAIN IN TOP LIST
gvt2.com	beacons.gcp.gvt2.com	false	true	true
muni.cz	mattermost.csirt.muni.cz	true	true	false
muni.cz	analytics.csirt.muni.cz	true	true	false
gvt2.com	e2c1.gcp.gvt2.com	false	true	true
mediavine.com	scripts.mediavine.com	false	true	true
live.com	roaming-eu.officeapps.live.com	true	true	true
muni.cz	kb.csirt.muni.cz	true	true	false
microsoft.com	eu-office.events.data.microsoft.com	true	true	true
microsoft.com	self.events.data.microsoft.com	true	true	true
live.com	roaming.officeapps.live.com	true	true	true
gvt2.com	beacons4.gvt2.com	false	true	true

Analyse unknown (sub)domains with VirusTotal

Analyse individual (sub)domains with VirusTotal

(Sub)domain:

Simple 0 4 Mem: 4.64 GB visited_domains.ipynb 0

Shrnutí a pár poznámek na závěr

Shrnutí

- Spojení nástrojů JupyterLab, Jupyter Widgets a Voilà umožňuje vytvářet **interaktivní analytické playbooky** jednoduše použitelné i pro běžné uživatele
- **Obecné řešení** pro různé typy analýz nad všemi dostupnými daty (nemusíme se omezovat jenom na kyberbezpečnostní doménu)
- Pomocí knihoven a konektorů je možné **napojit libovolnou službu nebo databázi**
- Analytici kteří neumí programovat GUI aplikace mohou **jednoduše tvořit interaktivní analýzy** (stačí když mají dostupné ukázky již hotových Jupyter Notebooků)
- Kód z Jupyter Notebooků můžeme **jednoduše přepsat do API** a využít v jiných systémech

Pokud vás téma zaujalo a máte nějaké otázky nebo další podněty, tak budu rád, když se mi ozvete na cermak@ics.muni.cz