

RAPEX Searcher – User setup

Quick start setup guide to work with RAPEX Searcher system

27/11/2019

TABLE OF CONTENTS

1.	INS	TALL	ING REQUIRED SOFTWARE4	ŀ
	1.1.	Opei	NVPN4	r
	1.2.	Мов	AXTERM6	j
	1.3.	WINE	DOWS REMOTE DESKTOP9)
	1.4.	PGA	DMIN11	
2.	RAF	PEX S	EARCHER EXECUTION AND DATA FLOW14	ŀ
	2.1.	ETL		;
	2.2.	RAP	EX SEARCHER	ò
	2.2.	1.	Searcher	7
	2.2.2	2.	Scraper	7
	2.2.3	3.	Text mining	}
	2.3.	ELAS	STICSEARCH INDEXING	\$
3.	ACC	ESS	ING KIBANA21	

Document characteristics

Property	Value
Release date	27/11/2019
Status:	Initial version
Version:	1.0
Authors:	Everis
Reviewed by:	
Approved by:	

Document history

Version	Description	Date
1.0	Initial version	27/11/2019

1. Installing required software

This guide is made on a Windows 10 environment. In other versions some steps may change.

The necessary software to run the RAPEX Searcher are:

- OpenVPN
- MobaXterm
- Windows Remote Desktop
- pgAdmin

These technologies and their purpose will be explained in the following parts of this document.

1.1. OpenVPN

RAPEX Searcher is deployed in a private network in the AWS Amazon using a Bastion as a Jump machine, so the first requirement to connect to this network is to use a VPN.

There are multiple software to make VPN connections, in this manual we will use OpenVPN software.

To install it, go to <u>https://openvpn.net/community-downloads/</u> and download OpenVPN software version fits better with your computer. Download the .ovpn file that is needed to connect to the Rapex Searcher network and put it in an easy accessible location. This file is the following:



Follow the steps as any common installation. Once it is ready, it should appear in task bar. If not, search it and launch it.



Now, check again if it appears in your task bar with the 🗐 icon. Right-click on it and select "Import file...":



Now you will see an open dialog. Search here the file where you saved it before and open it. The program should show you the following info message:

OpenVPN GUI 🛛 🗙
File imported successfully.
Aceptar

Now right-click again in OpenVPN icon and select the entry that has been added named "rapex" (or like your .ovpn filename):



A window showing OpenVPN log connection like the following appears.

OpenVPN Connection (rap						
Current State: Connecting						
$\label{eq:constraints} \begin{array}{c} \end{tabular}{lllllllllllllllllllllllllllllllllll$	TIONS IMPORT: TIONS IMPORT: TIONS IMPORT: TIONS IMPORT: A Channel: using J oging Data Chann ming Data Chann ming Data Chann active service ms JTE_GATEWAY n_tun -tun -Windows Driver fied TAP-Window cessful ARP Flue VAGEMENT: >S1	route options modified peerid set adjusting link, mitu to 1622 data channel capto option negotiated cipher 'AES-22 et'. Opter 'AES-285-65CM et'. Opter 'AES-285-65CM et'. Opter 'AES-285-65CM g, channel-580 10.99.234.1/255.255.254 Ethermet 5] opened: \\\\C Version 9.21 vs driver to set a DHCP IP vs driver to set a DHCP II h on interface [15] (8C4D FATE-1574787680.ASSIG	rs modified i6-GCM' initialized wit i0 I=20 HWA i0bal \8C4D7 /netmask of 1 731B-15AE-4 N_IP.,192.16	h 256 bit h 256 bit DDR=28 731B-15A 92.168.6 86A-B6F6 8.61.6	key key E-486A-B 1.6/255.: -4D522C	*
<					>	
, Bytes in: 0 Bout: 88 B			OpenVPN	GUI 11.1	2.0.0/2.4	1.7
Disconnect	Reconnect				<u>H</u> ide	

If everything went right, the OpenVPN icon on the task bar will be green Now, check again if it appears in your task bar with the aparience.

1.2. MobaXterm

MobaXterm is a software used to stabilise SSH connections to the RAPEX Ubuntu server that contains the searcher and scrapper components and the Logstash component.

This server is deployed in the AWS EC2 service with the name: RPXPREUBBUBU01

To install it, go to <u>https://mobaxterm.mobatek.net/download-home-edition.html</u> and download the version you want (portable or installable). If it is not compatible with your system you can use other SSH clients as PuTTY. This configuration can be used as reference anyway.

Once it is installed in your computer, open it.



You should see a screen like the following:

🔯 MobaXterm		-	o ×
Terminal Sessions View X server Tools Games	Settings Macros Help		
Quick connect	*		Ø
K User sessions			*
Sessions			
*			
Tools	Koba>	Kterm	
Macros	Start local terminal	C Recover previous sessions	
1	Find existing session or serve	r name	
	Recent session	IS	
	🔊 ubuntu PRE 📉 dev	📉 ubuntu	
	INT-1.3_ckan-back I Empleabilidad SEP	INT-1.3_ckan-front	
	PPI Node 1 Victory Datanode 1	INT-1.3_elda-solr	
	Enable advanced features and enhance security wit	h MobaXterm Professional Edition!	
UNREGISTERED VERSION - Please support MobaXterm by sub	scribing to the professional edition here: https://mobaxterm.mobatek.net		

Right-click on "User sessions" and then "New session". Select SSH option in the top left part of the window. Now we are going to configure the login settings for the Ubuntu machine that launches the RAPEX Searcher process. The information you need is the following:

- Remote Host: 10.2.1.91
- Port: 22
- Username: cliente
- Private key: You have to put here the path to the RPX.pem key in order to authenticate in the system. You can download it here:



Take a look to the next picture to know where to put this parameters.

ession settings
SSH Telnet Rsh Xdmcp RDP VNC FTP SFTP Serial File Shell Browser Mosh Aws S3 WSL
Basic SSH settings
Remote host * 10.2.1.91 Specify username cliente 😒 Port 22
Advanced SSH settings Terminal settings Network settings Bookmark settings X11-Forwarding Compression Remote environment: Interactive shell ~ Execute command: Do not exit after command ends Image: Comparison of the setting
SSH-browser type: SFTP protocol V Group Follow SSH path (experimental)
Use private key C:\PATH_TO\RPX.pem
Execute macro at session start: <pre></pre>
OK Cancel

The setup will ask you for a name for this new session. Enter any valid text ("Ubuntu PRE" for example) and click ok to save. Now you can enter the into the Ubuntu terminal double-clicking session's name. If everything is ok, you should have available the shell:



If you click in the "Sftp" tab at the left you will see the system directories where you can see easily what files and directories are available in the machine. This easies file manipulation tasks (upload/download, open, removal, etc.).

Every interaction you need to do with RAPEX Searcher module should be managed from here.

1.3. Windows Remote Desktop

The ETL component is deployed in the AWS EC2 Windows machine called: RPXPREWSBETL01

You do not need any additional software in Windows 10 to connect to the Windows server (ETL machine). Press "Windows" key and search "Remote desktop".



Now you need to introduce the host, user and password for the remote connection. These are:

- Computer: 10.2.1.69
- Username: cliente
- Password: TFWRjzE98EGcxXwU

The password shall be supplied after clicking on "Connect" button:

₩ Remote Desktop Connection - □ ×									
	Remote Conn	e Desk ectio	top n						
General [Display Local I	Resources	Experience	Advanced					
Log-on se	ettings								
	Enter the nam	e of the ren	note computer						
	<u>C</u> omputer:	10.2.1.69)		~				
	Usemame:	cliente							
	You will be as	ked for crea	dentials when	you connect.					
	Allow me to	o save c <u>r</u> ec	lentials						
Connectio	on settings								
	Save the current connection settings to an RDP file or open a saved connection.								
	Save Sa <u>v</u> e As Op <u>e</u> n								
Hide Options Connect Help									

Accept the certificate prompt that appears after introducing the password. You know can interact with the Windows machine. You can logout easily clicking on the close button at the top (by default).



1.4. pgAdmin

In order to access to the PostgreSQL database you need an administration tool or development environment like Oracle SQL Developer or Microsoft SQL Server. In this case, for PostgreSQL we recommend pgAdmin.

Download pgAdmin 4 from its official site: <u>https://www.pgadmin.org/download/</u>. Choose your version (Windows for this guide <u>https://www.pgadmin.org/download/pgadmin-4-windows/</u>) and install it as usual.

Once it is installed, you should have available the executable:



Launch it. Now the elephant from PostgreSQL should appear in the task bar. Double-click on it to open a new tab in your browser with pgAdmin service.

🗙 rapex.sql - Visua	🗸 🖷 🕼 🤹	📴 🖅 🌈 🗘 अ	:57 📮
			o ×
≣ Untitled-11	≣ Untitled-9	🛢 rapex.sql ×	□ …

In this new tab, double-click on "Servers", then "Create" and "Server…" last. You need to enter the credentials for the database. In this window, enter the name of the connection as you want ("rapex-pre" for example). Then, for "Connection", you need the following information:

- Host: rpxprerdspsq01.c9vfzi8b257o.eu-west-1.rds.amazonaws.com
- Port: 5432
- Database: RPX
- User: cliente
- Password: Z69yycK6M8mMWReg

Remember checking "Save password?" option. At last, click "Save".

Pg pgAdmin 4	× +				-		×
\leftarrow \rightarrow (b) (c) (c	:51193/browser/		⊌ ↔ ☆	lii\ O	۵	•	Θ Ξ
PgAdmin File V Object V	Teolow Holpw Create - Server		¢				
> Servers	General Connec	tion SSL SSH Tunnel Advanced					
	Host name/address	rpxprerdspsq01.c9vfzi8b257o.eu-west-1.rds.amazona	oted object.]	
	Port	5432					
	Maintenance database	RPX					
	Username	cliente					
	Password	•••••					
	Save password?						
	Role						
	Service						
	i ?	🗶 Cancel 🗳 Reset 🖺 Save					

You can see the tables opening the recently created connection. Go to RPX database, then "Schemas", "public" and "Tables".

2. RAPEX Searcher execution and data flow

The following diagram shows the flow of the program and how it interacts with the database and its tables.



Next there is an explanation by process.

2.1. ETL

The weekly file is obtained through the following URL: <u>https://ec.europa.eu/consumers/consumers_safety/safety_products/rapex/alerts/?event=main.w</u> <u>eeklyReports.XML</u>

To execute this component, go to Windows ETL machine (explained in Windows Remote Desktop chapter). There are two ways to refresh data, automatically or manual:

1. The automatic way is a schedule task, it's happen every night at 12:00 p.m.

The manual way; when we are in windows remote desktop, we go to RAPEX folder -> 2.ETL -> 2. Executables -> master and we found a batch file "master run", if we click twice it will be start the data refresh.

← → · ↑ 📘	> Thi	s PC > Local Disk (C:) > RAPEX	> 2. ETL > 2. Executables > maste	r	
		Name	Date modified	Туре	Size
📌 Quick access		itemas	0/22/2010 1-54 PM	Cile de Islan	
📃 Desktop	*	items	9/23/2019 1:54 PM	File folder	
Downloads	*	rapex	9/23/2019 1:54 PM	Filefolder	
	<u> </u>	src	9/23/2019 1:54 PM	File folder	
Documents	×.	xmlMappings	9/23/2019 1:54 PM	File folder	
Pictures	*	🕌 alerts_0_1	11/7/2019 3:30 PM	Executable Jar File	361 KB
🏪 Local Disk (C:)		🕋 log4j	11/7/2019 3:30 PM	XML Document	2 KB
RAPEX		🕌 master_0_1	11/7/2019 3:30 PM	Executable Jar File	36 KB
_		💿 master_run	11/7/2019 3:30 PM	Windows Batch File	1 KB
This PC		📓 master_run	11/7/2019 3:30 PM	Windows PowerS	1 KB
E Desktop		master_run.sh	11/7/2019 3:30 PM	SH File	2 KB
Documents		실 weekly_reports_0_1	11/7/2019 3:30 PM	Executable Jar File	45 KB
👆 Downloads					
👌 Music					
Pictures					
😽 Videos					
🏪 Local Disk (C:)					
Network					

Other way to run by is by console. Launch "cmd" console open Windows dialog and type "cmd":

Best match										
6°N	Command Prompt Desktop app									
T	۲	ß		2	đ	ła				
cmd										

And then write this sentence



The ETL takes RAPEX XML and processes the data present there storing it in alert, alert_i18n, picture and weekly_report tables.

2.2. RAPEX Searcher

This component (divided between three "minor" components) shall be executed from the Ubuntu machine (connection explained in MobaXterm chapter). Despite being three different parts with three different inputs and outputs, the whole component is launched as a single one.

Launch it with:

sudo systemctl start rapex

Replace start by stop to stop the service and replace it by restart to restart it.

You can see the output of the service executing the following command:

sudo journalctl -u rapex -f

With this command the shell will print every log line as it is working. You cannot write any command in this mode. To exit, press Ctrl+C.

Every new registry added in "out" tables contains a "timestamp" for each registry processed.



2.2.1. Searcher



The component that makes the query to Google API takes the information from the alerts imported by the ETL. Every entry in alert table is an alert from RAPEX. Every alert will have as many entries in alert_i18n table as languages configured. This table contains alert information that has multi-language values (as brand, name or description for example).

2.2.2. Scraper



The scraper component takes the results from Google (URLs mainly). It tries to access these URLs to extract their HTML code. This scraping could not be done by some reasons that appear in scraping_error table.

- In case the scraping can be done, the process adds a row in scraped_product table by URL. Each URL should have a scraped_source row and a scraped_screenshot associated.
- In case the scraping cannot be done, the process adds a row in scraped_product showing that the scraping was not done and why (associating an error from scraping_error table).

2.2.3. Text mining



This process takes from every scraped_product that could be scraped its HTML code (scraped_source table). Based on the analysis done, it will add or not a new row in found_product table. This means that every row in found_product is a positively identified URL.

2.3. Elasticsearch indexing

The indexing is managed by Logstash. This program is launched by a command with RAPEX config file location as parameter.

Execute the following command to launch Logstash and index in Elasticsearch the data from the output of RAPEX Searcher.

```
sudo /usr/share/logstash/bin/logstash -f /etc/logstash/conf.d/rapex.conf
```

It will execute two queries to the PostgreSQL database to make the indices used later in Kibana.

This SQL query for rapex_searched_products_last index:

```
SELECT sp.id AS searched_product_id,
    sp.alert_id,
    sp.searched_date,
    sp.url,
    sp.domain,
    wr.publication_date,
    al.value AS alert_level,
    at.value AS alert_type,
    c.value AS alert_type,
    c.value AS category,
    rt.value AS risk_type,
    ct.iso_value
FROM searched_product sp
        JOIN alert a ON sp.alert_id = a.id
        JOIN weekly_report wr ON a.weekly_report_id = wr.id
        JOIN alert level al ON a.alert level id = al.id
```

```
JOIN alert_type at ON a.alert_type_id = at.id
JOIN category c ON a.category_id = c.id
JOIN alert_risk_type ar ON ar.alert_id = a.id
JOIN risk_type rt ON ar.risk_type_id = rt.id
JOIN country ct ON ct.id = sp.country id
```

And this one for rapex_found_products_last index:

During the execution, the data will follow the next flow.



As said before, Kibana needs Elasticsearch indices for its visualisations. To do so, the indexing process takes the information resultant from found_product and searched_product. These

tables are the necessary to build the dashboard showed before. Some other information are taken too (like alert ID from alert table) but these two are the most relevant.

3. Accessing Kibana

Kibana URL is: https://prerpxesaut.everincloud.com/ plugin/kibana/

You need a user and password to access Kibana. This user will be created based on the username and email provided to the team. There could be as many users as desired. These credentials are used in the following window:

	Signin			× +			—			X
←	→ @	G	Ū	https://rpxpreauth.au	th.eu-west-1. amazoncognito.com /login?dient_id=50aqiiboc5c	cusss67c 🛛 🕶 🛇 🏠	lii\ 🕻)	Θ	≡
					Sign in with your username and password					
					Username					
					Username					
					Password					
					Password					
					Forgot your password?					

Once you are in, click on the dashboard button on the left panel to see the dashboard configured.

	📕 Rapex Searcher Analysis - Kiban 🗙	+					-				
← -	→ A C' 0 A mo https://pre	rpxesaut.everincloud.com/_plugin/kiba	na/app/kibana#	/dashboard/d64d1ee	20-d945-11e9-83ff-1ffde8cb3	3a40? ••• 🕑 🟠	III\ 🙂 🖥	0 E			
	Dashboard / Rapex Searcher A	nalysis						٥			
G	Full screen Share Clone Edit										
Ø	Filters Search		KQL	〕 ✓ This year Show date			s C Refresh				
ش	 										
5		Alext True				Alest Diele					
A	Select	Select		Select		Select					
e e											
<u></u>	Apply changes Cancel changes Clear form										
100											
	Total Web pages found	Total analyse	i products		Years since	ce first product reviewed					
	0			0		-					
	Web pages found by category	/		Web pages four	nd by risk type						
\rightarrow											