# EurOpen: Zpracování dat pomocí nástrojů Splunk a Elastic Stack

*“Splunk is the new Excel!”*

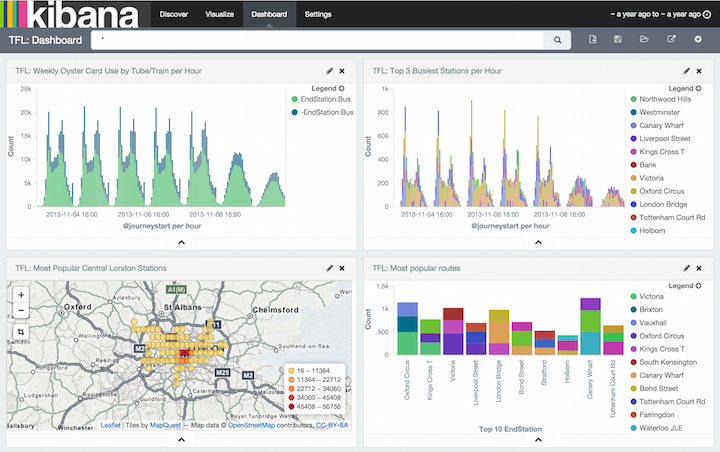
Není tomu tak dávno, co se ke zpracování mnoha a mnoha dat nezřídka používal Microsoft Excel. Ještě několik let nazpět bylo možné potkat bezpečnostní experty převádějící celou řadu forenzních nálezů a logů do CSV souborů. Na internetu se daly najít dobře dokumentované postupy využívající pokročilé vlastnosti Excelu. Jeho schopnost rychlého a interaktivního filtrování dat, jednoduchost ovládání i převodu dat do formátu, který umí zpracovat, v kombinaci s vytvářením nejrůznějších grafů, dlouho nenalézala přemožitele. A ačkoliv pro menší analýzy stále může být nástrojem více než vhodným, některá jeho omezení vedla k hledání dalších řešení.



*Obr. 1: Ukázka vizualizace dat v prostředí Splunk*

V tutoriálu si představíme nástroje Splunk a Elastic Stack, jejichž použití může být stejně snadné nebo snazší než u Excelu, jejich schopnosti zpracování velkých dat ovšem mnohem větší. A nejde jen o data velká — oba nástroje umí, na rozdíl od Excelu, výborně zpracovávat data získávaná v reálném čase z jiných systémů (např. Twitter), obohacovat je o další informace, provádět nad nimi netriviální dotazy (do jisté míry podobné SQL dotazům nad tradičními databázemi), využívat paralelních dotazů pro násobně rychlejší zpracování dat, vytvářet jejich grafové vizualizace a nad nimi i dashboardy a v neposlední řadě nastavovat nad zpracovanými daty notifikace.

Splunk je komerční řešení pro zpracování velkých objemů více či méně strukturovaných dat, Elastic Stack (dříve a častěji známý jako ELK, kombinace řešení Elasticsearch, Logstash a Kibana) je open source varianta téhož.



*Obr. 2: Ukázka vizualizace dat v prostředí Kibana4*

Čemu se tutoriál nebude věnovat je škálování těchto systémů pro opravdu velké datové objemy, neboť s tím nemá ani jeden z autorů žádné zkušenosti. Naopak mají zkušenosti s využitím těchto systémů pro praktické každodenní aplikace v oblasti bezpečnosti či monitoringu dat.

K čemu se tedy takové systémy dají využít právě ve vaší organizaci a pro zlepšení bezpečnosti?

Dle zkušenosti autorů stále existují týmy, které pro zpracování logů používají své ručně psané programy vzniklé před mnoha lety. Ačkoliv dostatečně fungují a reporty z nich stále chodí, jejich udržování, opravy a přizpůsobování novým zařízením i novým typům dat vyžadují nemalé úsilí.

Zkusíme si proto nastavit vlastní Splunk a ukázat si, jakým způsobem je možné do něj dostat požadovaná data, jak vypadá dotazovací jazyk a jaké možnosti vizualizace poskytuje. Totéž se poté pokusíme do nějaké míry replikovat i nad ElasticStackem, ukážeme si rozdíly i společné vlastnosti.

Nejčastěji se obě řešení používají na zpracování dat v časových řadách, tedy právě na logy, záznamy událostí či měření daných veličin v časových intervalech. Zkusíme alespoň v rychlosti nastínit, jakým způsobem je možné posílat logy do jednoho či druhého systému tak, aby byl možné využít centralizovaného zpracování dat jejich pomocí.

Z vizualizací pak bývá velmi oblíbenou časový histogram, který ukazuje, kolik událostí se objevilo v kterém okamžiku. Krom pravidelného přehledu o tom, že všechny systémy fungují jak mají, nám taková informace může usnadnit budoucí plánování infrastruktury (např. u velikosti datových toků, počtu zpracovaných e-mailů, četnosti HTTP dotazů) či zvýraznit věci na první pohled méně viditelné (může se týkat výpadku záloh, příchozích phishingových kampaní či nadměrného provozu v nečekaných částech sítě).

Vývojáři aplikací mohou do obou řešení zasílat aplikační chyby, aplikační statistiky či audity přístupů a následně je přehledně vyhodnocovat a prioritizovat.

Právě rychlá analýza nad vybranou podmnožinou dat, různorodé datové zdroje, schopnost interaktivního filtrování a rychlého nalezení podobností či anomálií se používá nejen v monitoringu infrastruktur, ale i při vyšetřování bezpečnostních incidentů. Moderní bezpečnostní operační centra využívají mnoha datových zdrojů pro korelace dat a rychlé, automatizované přípravy podkladů pro vyšetřování incidentů, často provázané s některým ze známých systémů pro správu incidentů. Z pohledu manažerů pak systémy typu Splunk či Elastic Stack umožňují snadné vizualizace nastavených metrik, což může zrychlit rozhodovací procesy v organizaci. Tomu všemu bychom rádi věnovali alespoň část našeho tutoriálu.

Na základě zájmu a zkušeností publika se také pokusíme předvést několik dalších projektů využívajících ElasticSearch, například *nightHawk Response* [4] pro rychlejší analýzu bezpečnostních incidentů.

Pro potřeby tutoriálu budeme využívat zdarma dostupnou verzi Splunku [1] s omezením na 500 MB dat denně. Vše potřebné pro využití Elastic Stacku bude předpřipravené v obrazu pro některé z virtualizačních řešení (VMWare nebo VirtualBox), což je i největší nezbytná nutná podmínka pro praktickou část tutoriálu. A pokud se vám nechce čekat až na tutoriál nebo trávit čas instalací ať již Splunku nebo Elastic Stacku, můžete směle využít možností řešení 21. století – obě řešení (Splunk Cloud [2] či u Elastic Stacku např. ve variantě Logsene [3]) je možné zdarma za určitých podmínek využívat v cloudu. Jen si řádně rozmyslete, jaká data tam chcete či můžete posílat.

Reference

1. Try Splunk Cloud For Free. *Splunk, Inc.*[online]. [cit. 2016-09-01]. Dostupné z: *https://www.splunk.com/page/sign\_up/cloudtrial*
2. Free Trials and Downloads. *Splunk, Inc.*[online]. [cit. 2016-09-01]. Dostupné z: *https://www.splunk.com/en\_us/download.html*
3. Log Management & Analytics. *Sematext.*[online]. [cit. 2016-09-01]. Dostupné z: *https://sematext.com/logsene/*
4. nightHawk Response. *Asynchronus forensic data presentation.*[online]. [cit. 2016-09-15]. Dostupné z:*https://github.com/biggiesmallsAG/nightHawkResponse*