Projekt ADAM

Systematické zpracování provozních dat

Ladislav Lhotka • ladislav.lhotka@nic.cz • 12. listopadu 2020

CZ.NIC | SPRÁVCE DOMÉNY CZ

Osnova

- cíle projektu, vize
- architektura systému
- sběr a zpracování dat o DNS provozu
- DNS crawler
- servisní databáze a REST API
- nové grafy a statistiky na webu
- další výsledky

Takový průměrný den ...

- 1,2 miliardy DNS dotazů na autoritivní servery
- 300 milionů dotazů na resolvery ODVR
- 5300 operací v registru domény
- 8000 testů pomocí Netmetru
- 2000 přihlášení pomocí mojeID

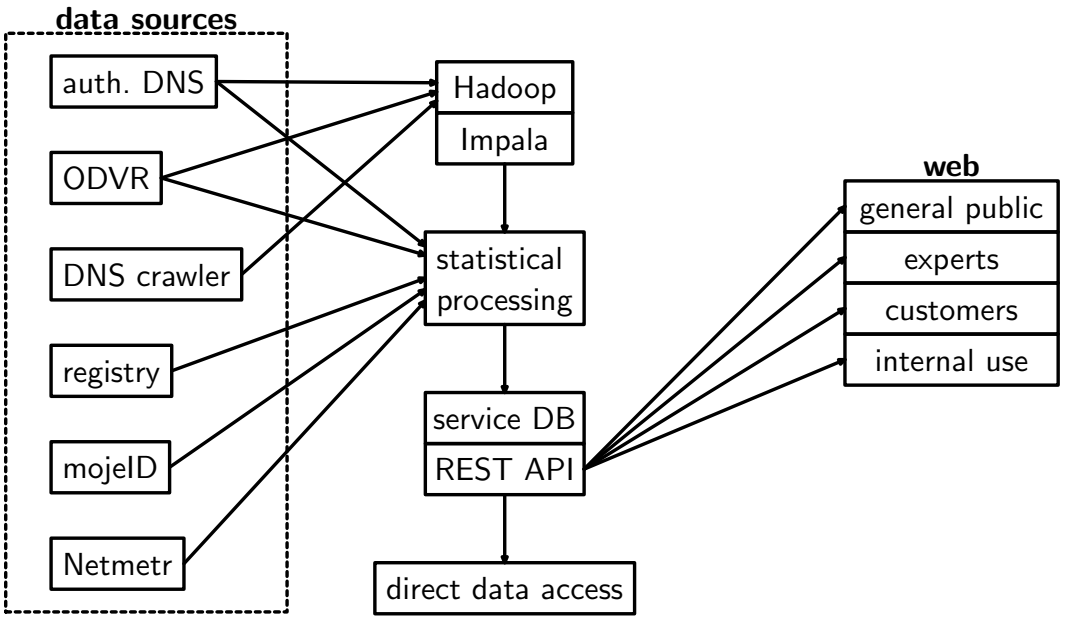
Cíle projektu, vize

Cílem projektu ADAM (*Advanced DNS Analysis and Monitoring*) je vytvoření robustního a rozšiřitelného systému pro sběr, ukládání, zpracování, analýzu a vizualizaci dat z hlavních služeb provozovaných sdružením CZ.NIC.

Očekává se využití mj. v těchto oblastech:

- monitorování provozu DNS serverů a dalších služeb,
- sledování krátko- i dlouhodobých trendů,
- detekce a analýza anomálií a bezpečnostních incidentů,
- plánování rozvoje infrastruktury,
- grafy a statistiky pro interní použití, členy, zákazníky i odbornou veřejnost.

Architektura systému



Provoz na DNS serverech

Na DNS serverech, které spravuje CZ.NIC, je veškerý DNS provoz ukládán do souborů PCAP. Ty jsou pak pravidelně komprimovány a přenášeny do centrálního úložiště, kde jsou průběžně zpracovávány a údaje o všech DNS transakcích ukládány do Hadoopu.

Nevýhody:

- velké objemy dat (přenášené a uchovávané)
- výkyvy zátěže procesoru ovlivňují časové značky paketů
- část transakcí se nezachytí, zejména TCP

DNS sonda

https://gitlab.nic.cz/adam/dns-probe

Software pro monitorování DNS vyvinutý ve spolupráci s FIT VUT Brno. Umožňuje extrahovat DNS dotazy a odpovědi ze živého provozu nebo PCAP souborů (v UDP i TCP), párovat je a exportovat konsolidované záznamy o transakcích.

- záchyt paketů přes raw socket (`AF_PACKET`) nebo DPDK
- exportní formáty: Apache Parquet nebo a C-DNS [RFC 8616] – až o 70 % menší objem dat
- lepší párování dotazů a odpovědí: o 2–3 % více, u TCP o 50–100 %.

DNS crawler

https://gitlab.nic.cz/adam/dns-crawler

Nástroj pro procházení zadaného seznamu domén a získávání dostupných údajů z DNS, ale také z komunikace s webovými a mailovými servery.

Ve spolupráci s CSIRT.CZ procházíme pravidelně všechny domény 2. úrovně pod *.cz* (https://www.csirt.cz/cs/dns-crawler).

Kromě podkladů pro statistiky slouží i ke klasifikaci webových stránek, kontrole dat v DNS a detekci bezpečnostních problémů (webový framework *Nette*).

Servisní databáze

V databázi Apache Hadoop aktuálně udržujeme půl roku provozu, pracujeme na upgradu clusteru s cílem prodloužit uchovávanou historii na jeden rok.

Agregované hodnoty a statistiky ukládáme *servisní databázi* PostgreSQL bez omezení délky historie.

Problém: když se ukáže potřeba nového zpracování originálních dat.

REST API

Nástroj PostgREST vytváří REST API automaticky z tabulek servisní databáze.

Doporučená metoda pro rutinní strojové zpracování dat, integraci do webových stránek apod.

- root endpoint: https://stats.adam.nic.cz
- webové rozhraní OpenAPI/Swagger: https://stats.adam.nic.cz/swagger (dokumentace, praktické vyzkoušení)

Statistické zpracování, vizualizace

Po předchozích zkušenostech se snažíme využívat „klasické“ nástroje se širokou základnou uživatelů a celý proces co nejvíce zpřehlednit, automatizovat a zdokumentovat.

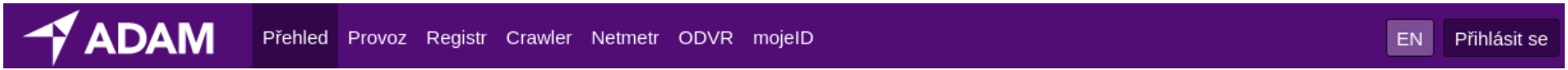
Každý výstup (graf, tabulka, statistika) musí být:

1. dostupný ve dvou jazykových verzích (česky a anglicky)
2. kompletně reprodukovatelný
3. snadno modifikovatelný

Zvolené nástroje

- statistický systém *R*
- osvědčené rozšiřující balíky: *tidyverse*, *ggplot2*, *Plotly*, *kableExtra*
- *R Markdown* – literate programming pro R
- webové stránky: *flexdashboard*
- internacionalizace R kódu: *gettext*
- automatizace: GitLab CI

Nové statistiky na webu



- https://stats.adam.nic.cz/dashboard/cs
- Projekt: https://gitlab.nic.cz/adam/adam-dashboard
- Chyby, přání: https://gitlab.nic.cz/adam/adam-dashboard/-/issues

Zatím jde o provoz v testovacím režimu, chystáme další rozšiřování a úpravy.

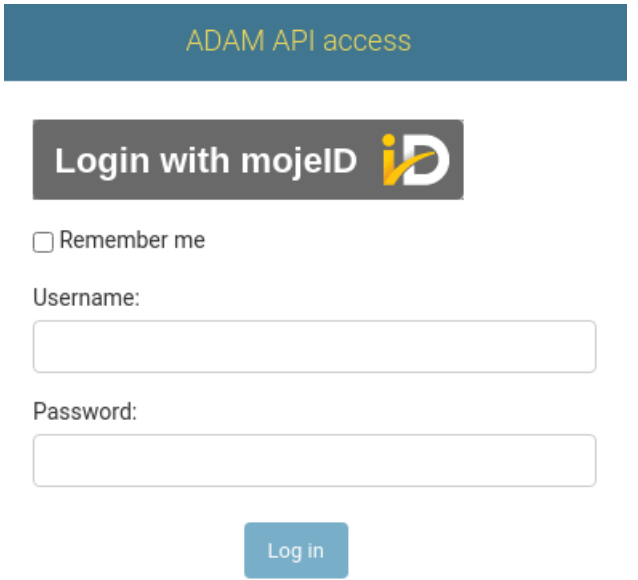
Autentizovaný přístup

Část grafů a statistik je určena jen pro specifické uživatele (např. držitele domény nebo registrátora) či jejich skupiny.

Dva způsoby přihlášení uživatele:

1. přes mojeID
2. pomocí speciálního jména a hesla

Autorizace funguje zároveň pro webové stránky i pro REST API. Uživatel si může také vytvořit tokeny pro strojový přístup.



Další výsledky

- Výroční domain report: https://stats.nic.cz/reports/2019
- ADAM reports: https://adam.pages.nic.cz/reports/adam
  1. COVID-19 v doméně .cz
  2. Základní klasifikace domén druhé úrovně pod .cz

Odkazy

1. Apache Hadoop Ecosystem: https://www.cloudera.com/products/open-source/apache-hadoop.html
2. flexdashboard: https://rmarkdown.rstudio.com/flexdashboard
3. gettext: https://www.gnu.org/software/gettext
4. kableExtra: https://github.com/haozhu233/kableExtra
5. OpenAPI/Swagger: https://swagger.io

6. PostgREST: https://postgrest.org
7. R: https://www.r-project.org
8. tidyverse, ggplot2: https://www.tidyverse.org
9. Plotly R: https://plotly.com/r/

Děkuji za pozornost.

Ladislav Lhotka • ladislav.lhotka@nic.cz • https://adam.nic.cz

CZ.NIC | SPRÁVCE DOMÉNY CZ