MUNI
ÚVT

PV177 – DataScience seminář
(Semestrální projekty)

Tomáš Rebok & **Martin Macák**
Ústav výpočetní techniky MU
Fakulta informatiky MU

Cíle projektů

Praktické projekty situující vás do různých rolí datové analýzy

- správci systémů pro datovou analýzu
- vývojáři nástrojů pro datovou analýzu
- datoví analytici

Nejde o věci „do šuplíku“

- většinou projekty, které se nám hodí vyřešit
- nebo která prozkoumávají určitou cestu
- případně jejichž výstup je jinak užitečný

Omezená velikost datových sad

- ať neřešíte další složitosti související s velkými objemy či výpočetní náročností
- zpracování však bude probíhat pohledem BigData analytika

Některá témata bez specifikace konkrétního nástroje

- volnost ve výběru
  - během konzultací doporučíme podle nás vhodný nástroj (typ nástroje)
  - Vy však můžete přijít s novým pohledem, který by nás nenapadl
    - třeba se něco naučíme i my ☺

Pravidla projektů

Základní pravidla

- 4 týmy po 3 studentech
  - ustavte si sami nebo vás rozdělíme ☺
  - témata vypíšeme do ISu
- výběr libovolného tématu (i návrh vlastního – např. BP/DP, spolupráce, …)
  - stejné téma mohou realizovat max. 2 skupiny (preference je však max. 1)
- dostupnost konzultující osoby ke každému z témat
  - detaily k tématu budou doplněny konzultantem tématu
    - dnes jen základní představení
  - a zejména pak dostanete data
- pro realizaci témat využijete výpočetní infrastrukturu (grid, cloud)
  - více k infrastrukturám za chvíli
- závěrem semestru odprezentujete dosažené výsledky
  - různá náročnost témat
  - hodnocení zohledňuje i snahu před konkrétním výsledkem

Představení projektů

Analýza dat jednotného přihlášení MU

- jednotné přihlášení MU generuje data o přístupu ke službám
  - kdo (UČO/ID), kdy, kam (služba), odkud (IP)
- cíle projektu:
  - indexace dat (logů) ve vhodném nástroji
  - příprava přehledových pohledů (*dashboardů*) dle zadání či vlastního uvážení
    - časové průběhy přihlášených, geolokace, trendy ke službám, využití služeb uživateli, atp.
  - detekce anomálií, např.:
    - nové, dosud nepoužívané zařízení
    - nestandardní čas přihlášení
    - nová lokace
    - dvě přihlášení s podezřelým prostorovým a časovým rozdílem
  - *bonus:* anomálie v chování uživatele
    - využití metod strojového učení

Představení projektů

Analýza dat meteorologických měření

- náš výzkumný partner (CzechGlobe) zpracovává zachycená data různých meteorologických měření
  - teplota (v různých výškách), tlak, vlhkost, sluneční radiace, síla větru, atp.
    - zkrátka nějaká data typu „typ – hodnota“
    - několik desítek parametrů sbíraných v 15minutových intervalech
  - tyto analyzuje různými grafy (Excel)
- cíle projektu:
  - indexace dat v pokročilém nástroji datové analýzy
  - příprava přehledových pohledů (*dashboardů*) dle zadání či vlastního uvážení
    - součástí zadání bude náhled na aktuální Excel dokument
  - demonstrace možností při analýze dlouhodobějších období
    - včetně interaktivních při práci s nimi
  - *bonus:* analýza anomálií (špatných hodnot) metodami strojového učení
    - integrace s DP prací M. Moravčíka
    - případně doplnění o vlastní metody

Představení projektů

Analýza dat síťových záchytů

- indexace dat zachyceného síťového provozu (formát PCAP) pro potřeby kriminalistického a forenzního zkoumání
- cíle projektu:
  - indexace dat síťového provozu ve vhodném nástroji
    - s důrazem na zachování časových značek, parametrů komunikace, extrakce dat, …
  - příprava přehledových pohledů (*dashboardů*) dle zadání či vlastního uvážení
    - rychlá analýza konkrétních typů útoků, přehledy komunikace (včetně geoinformací), atp.
  - analýza předaných PCAP souborů a poskytnutí odpovědí na předané dotazy
  - *bonus (lze i jako samostatný projekt):*
    - indexace dat ve dvou různých datových modelech (např. *key-value* a *graph db.*)
    - porovnání analytických možností obou přístupů

Představení projektů

Analýza souborových systémů diskových ISO obrazů

- indexace souborových dat předaného úložiště pro potřeby kriminalistického a forenzního zkoumání
- cíle projektu:
  - indexace dat souborového systému ve vhodném nástroji
    - s důrazem na zachování časových značek
  - příprava přehledových pohledů (*dashboardů*) dle zadání či vlastního uvážení
    - typy souborů (skutečné), časy modifikací souborů, atp.
  - analýza předaných souborových systémů a poskytnutí odpovědí na předané dotazy
  - *bonus:*
    - integrace funkcionality do nástroje CopAS (viz příště)
    - podrobnější analýza vybraných typů souborů

      např. podobnostní analýza obrázků (integrace nástroje laboratoře DISA)

Představení projektů

Analýza CSV souborů v ElasticSearch

- průzkum možností analýzy CSV datových souborů v ElasticSearch
- cíle projektu:
  - příprava co nejobecnějšího nástroje pro analýzu CSV souborů
    - a jeho integrace do nástroje CopAS (příště)
  - indexace různých typů CSV souborů a demonstrace funkcionality
    - příprava přehledových pohledů (*dashboardů*) dle zadání či vlastního uvážení
  - analýza předaných CSV souborů a poskytnutí odpovědí na předané dotazy

Představení projektů

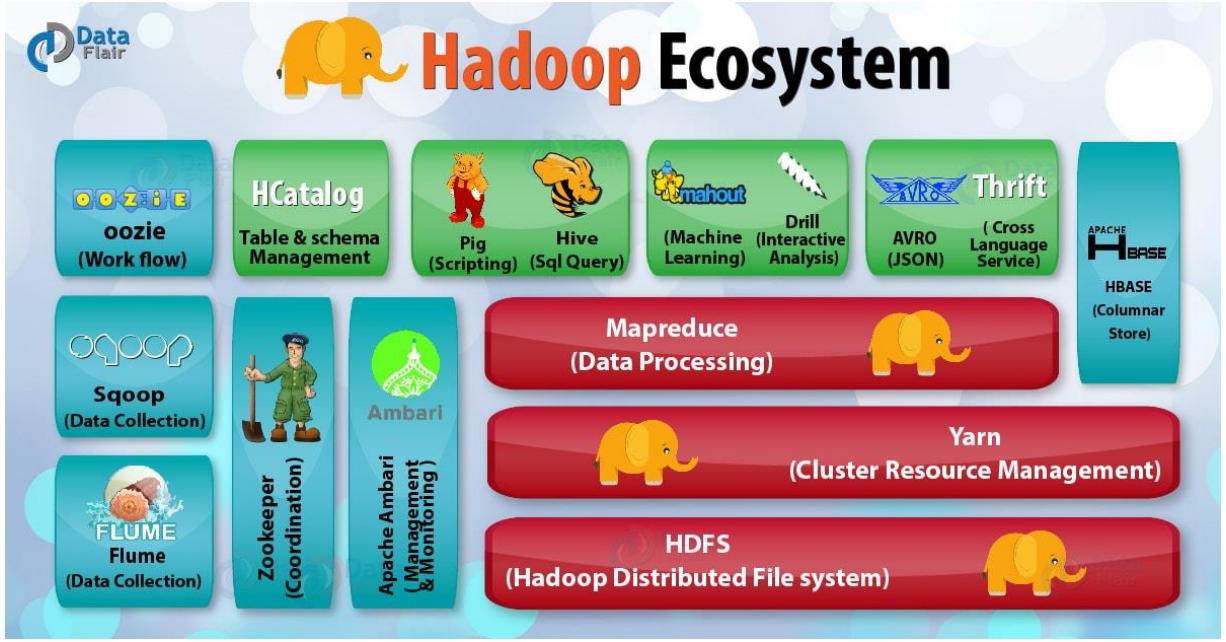
Průzkum a dokumentace (pokročilých) vlastností analytického nástroje Kibana (ElasticSearch)

- s primárním zaměřením na síťové toky
- příprava školících materiálů
  - dokumentace
  - demonstrační příklady
- *bonus:* spolupráce na realizaci školení pro PČR

Představení projektů

Příprava a dokumentace realizace Hadoop ekosystému

- identifikace nejvhodnější cesty pro individuální nasazení
- zmapování dostupných rozšíření
  - z praktického analytického pohledu
- příprava školících materiálů
  - dokumentace
  - demonstrační příklady



MUNI ÚVT

Sumarizace projektů

Analýza dat jednotného přihlášení MU

Analýza dat meteorologických měření

Analýza dat síťových záchytů

- možný samostatný projekt na grafový přístup

Analýza souborových systémů diskových ISO obrazů

Analýza CSV souborů v ElasticSearch

Průzkum a dokumentace (pokročilých) vlastností analytického nástroje Kibana (ELK)

Příprava a dokumentace realizace Hadoop ekosystému

Další, zatím nejisté projekty:

- analýza biomedicínských dat evropského projektu BBMRI (nutnost zajištění anonymizace)
- analýza dat „chytrého města“ (spolupráce s MMB, zatím bez zadání)


thank you
danke
謝謝
merci
gracias
obrigado
bedankt
dziękuję
terima kasih
kop khun krap
go raibh maith agat
teşekkür ederim
ngiyabonga
спасибо
рахмат
Баярлалаа
dank je
tapadh leat
mochchakkeram
sukriya
grazie
arigatô
dakujem
мерси
tack
welalin
barka
mersi
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
djiere dieuf
tau
дякую
akun
dankon
aciu
sulpáy
gràcies
gratias ago
chnorakaloutioun
taiku
takk
trugarez
shukriya
merci
dhanyavadagalu
diolch
rahmet
tanemirt
ありがとう
najis tuke
kam sah hamnida
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
didi madloba
rahmat
sagolun
mèsi
dēkuji
sobodi
enkosi
bayarlalaa
grâcie
hvala
mauruuru
köszönöm
nandri
nanni
dhanyavad
kiitos
dankie
faafetai lava
спасибі
vinaka
blagodaram
kia ora
spas

MUNI ÚVT

MUNI

ÚVT