DETEKCE ANOMÁLNÍHO CHOVÁNÍ UŽIVATELŮ KATASTRÁLNÍCH MAPOVÝCH SLUŽEB

VYSOKÁ ŠKOLA BÁŇSKÁ - TECHNICKÁ UNIVERZITA OSTRAVA
Hornicko-geologická fakulta
Institut geoinformatiky
Ostrava 2014

Autorka:
Bc. Radka **MATOLÁKOVÁ**

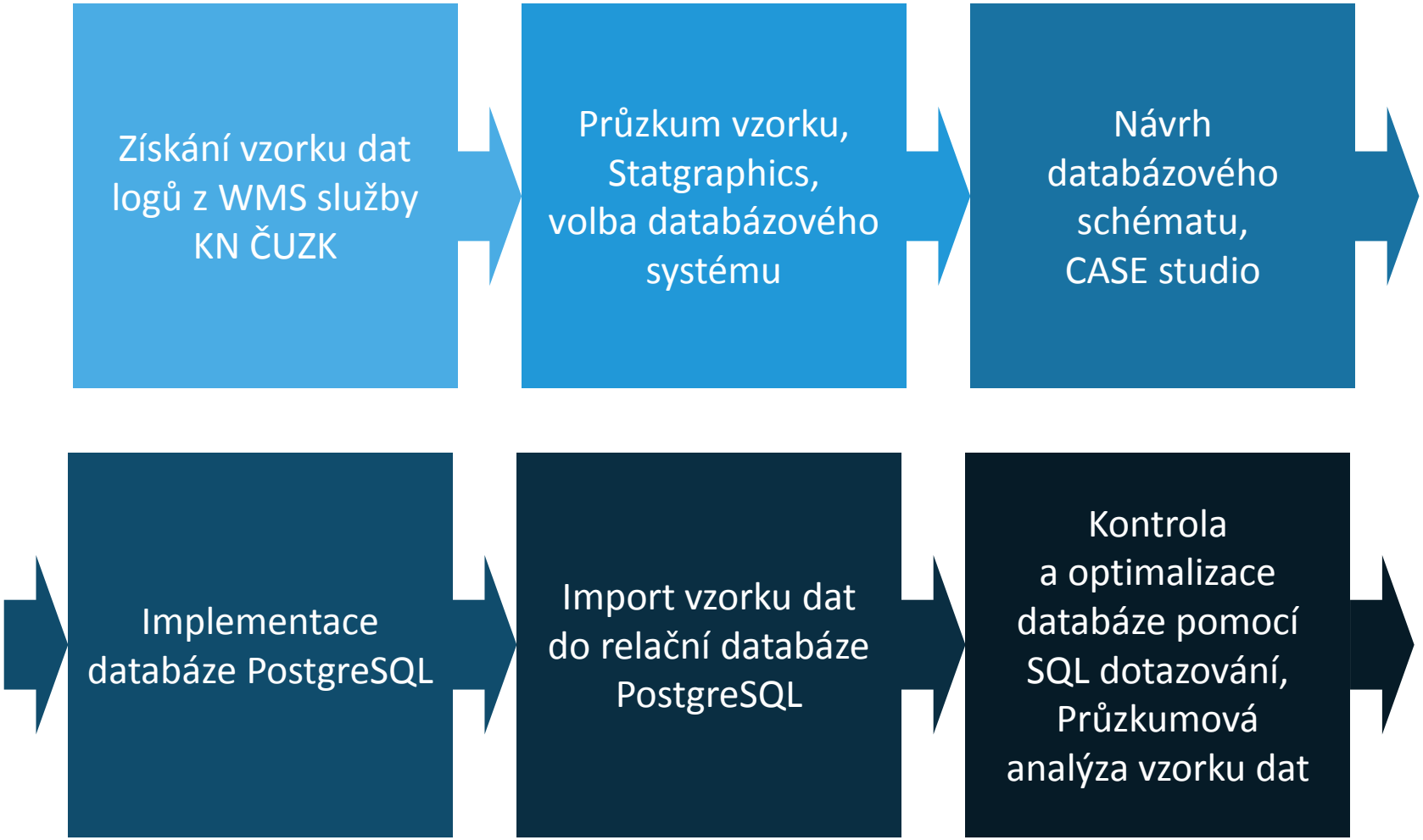
Vedoucí diplomové práce:
Ing. Jan **RŮŽIČKA**, Ph.D.

OSNOVA

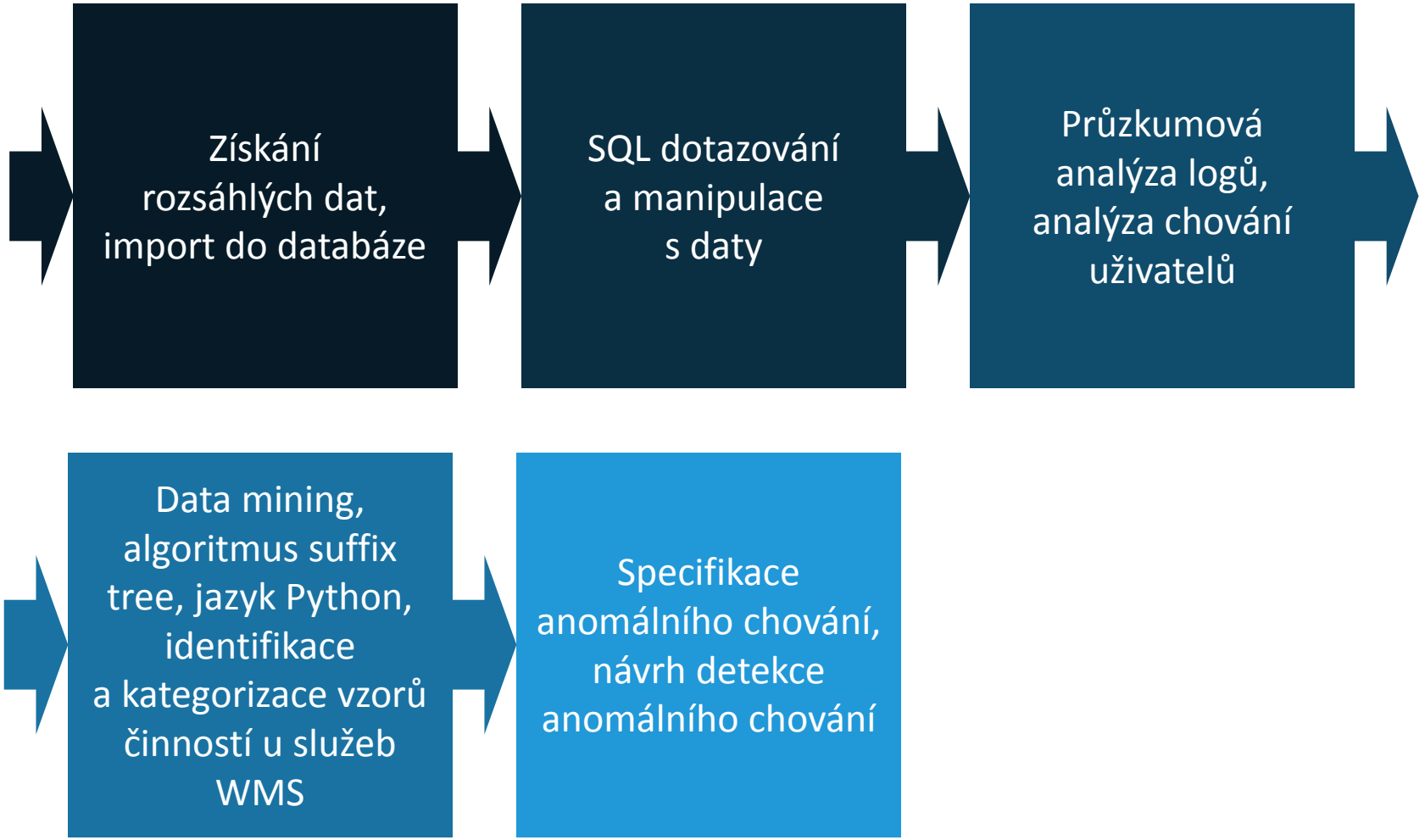
CÍLE

- Návrh datové struktury pro ukládání a zpracování logů webových služeb WMS
- Implementace databáze
- Průzkumová analýza dat logů
- Analýza chování uživatelů WMS
- Identifikace a kategorizace vzorů činností u služeb WMS
- Návrh detekce anomálního chování uživatelů

POSTUP ZPRACOVÁNÍ VZORKU DAT



POSTUP ZPRACOVÁNÍ ROZSÁHLÝCH DAT



LOG SOUBORY

- Rozsáhlé logy webových služeb
- Období od 3. března 2014 do neděle 9. března 2014
- 2,5 mil záznamů za den

Datum akce	Čas akce	Cíl akce	Požadavek	Otisk IP adresy
HTTP status kód	Počet byte, které server odeslal	Počet byte, které server přijal	Délka trvání akce [ms]	

- Anonymizovaná IP adresa - zakódování md5 funkcí
- Otisk IP adresy Cb0447e9c75448bbaee0ecd300e47ea8

UKÁZKA POŽADAVKU „GetMap“

- &REQUEST = GetMap (nebo GetCapabilities)
- &SERVICE=WMS
- &VERSION=1.1.1
- &LAYERS=RST_KN_I,RST_KMD_I,parcelni_cisla_i,hranice_parcel_i,omp
- &STYLES=&FORMAT=image/png
- &BGCOLOR=0xFFFFFF
- &TRANSPARENT=TRUE
- &SRS=EPSG:4326
- &BBOX=14.3701171875,50.10648772767332,14.39208984375,50.12057809796007
- &WIDTH=256&HEIGHT=256

IMPLEMENTACE DATABÁZE

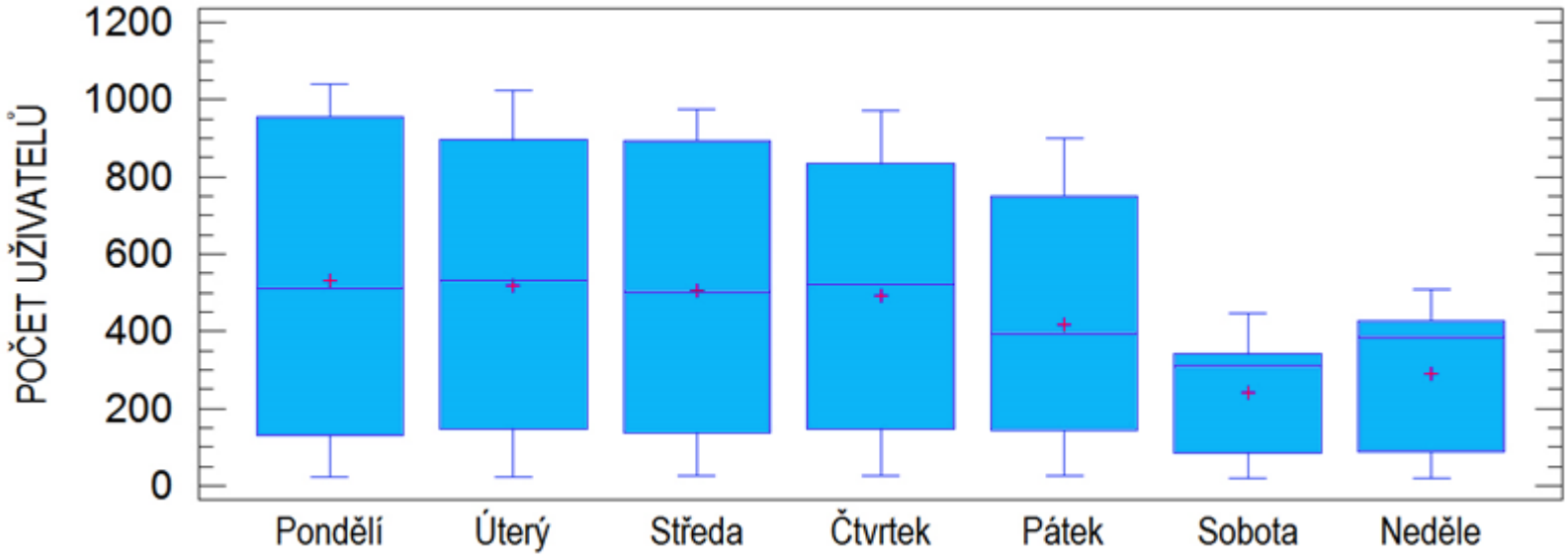
- Návrh schématu - software CASE STUDIO
- Export schématu do jazyka SQL
  Kontrola, korekce a začištění skriptu
- Vytvoření databáze
- Import dat pomocí skriptu v jazyce Python
  ošetření výjimek
  a rozdělení dle mezer a znaku “&”
- Vytvoření zálohy schématu
  i dat funkcí pg_dump

PRŮZKUMOVÁ ANALÝZA DAT 1/2

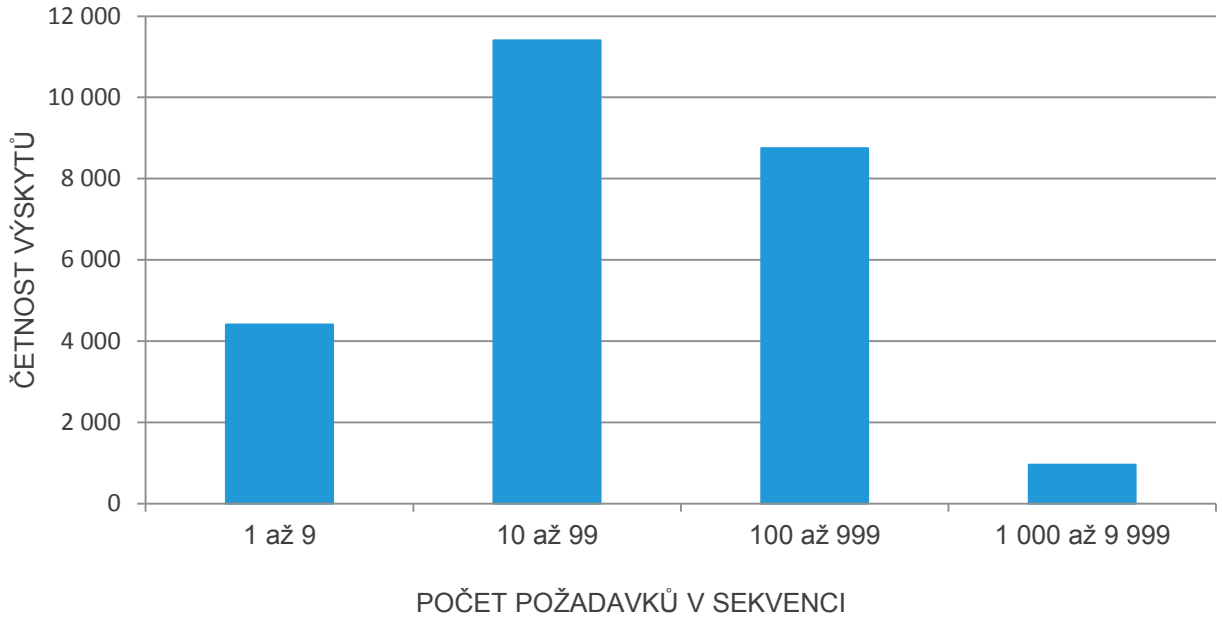
- Denně v průměru:
  - 5 975 uživatelů
  - 2,4 mil. požadavků na mapu, tj. 108,6 tisíc požadavků za hodinu
- Průměrný počet požadavků z 1 IP adresy je 229
- Průměrná délka vyřízení požadavku 661 ms
- Maximální serverová zátěž mezi 10. a 11. hodinou SEČ
  64,2 dotazů za sekundu, tj. 231 tis. dotazů za hodinu
- Minimální zátěž mezi 2 a 6 hodinou SEČ
  méně než 15 tis. dotazů za hodinu

PRŮZKUMOVÁ ANALÝZA DAT 2/2

- Pomocí metody neparametrické analýzy rozptylu prokázána závislost mezi:
  - průměrným počtem dotazů (uživatelů) za hodinu a denní dobou
  - průměrným počtem dotazů (uživatelů) za hodinu a délkou vyřízení dotazu
- Neprokázána závislost mezi průměrným počtem dotazů (uživatelů) za hodinu a dnem v týdnu
- Pomocí Spearmanova korelačního koeficientu prokázána pozitivní korelace počtu dotazů a délkou vyřízení požadavku



ANALÝZA CHOVÁNÍ UŽIVATELŮ



Pohyb na mapě

- Pohyb pouze směrem na východ 9,5 % požadavků
- pohyb pouze ve směru na západ 6,4 %
- kombinace pohybu severozápadně a oddálení 5,9 %
- jen přiblížení, nebo jen oddálení (bez posunu) << 0,1 %

IDENTIFIKACE A KATEGORIZACE VZORŮ ČINNOSTÍ

- Každý pohyb po mapě nahrazen alfanumerickým kódem
- Kódy jednotlivých uživatelů zřetězeny v sekvenci
- Pomocí algoritmu „suffix tree“ nalezeny vzory činností

Méně než 5 kroků	Více než 10 kroků	Testováno 5-9 kroků	Vhodná délka posloupnosti je 6 kroků

- Nejčastější vzory činností (nad 200 výskytů) :
  - Posun v ose X oběma směry
  - Opětovné zobrazení výsledku předchozího požadavku

ANOMÁLNÍ CHOVÁNÍ UŽIVATELŮ

Jakékoli záměrné i neúmyslné chování vymykající se očekávanému chování

Nejčastěji nalezené anomální chování

- Dotaz mimo rozsah ČR 203 tis. ze 20mil.
- Extrémní délka vyřízení dotazu

Příklady nežádoucího chování záměrného

- Snaha o narušení funkce serveru
  např. vykrádání dat pomocí robotů

NÁVRH DETEKCE ANOMÁLNÍHO CHOVÁNÍ

Detekce na základě odlišnosti od nalezených vzorů činností
není možná vzory vykazují přílišnou variabilitu
a nízkou četnost výskytů

Detekce na základě obecných poznatků:
například algoritmem spuštěném na PROXY serveru
testující zda je požadována oblast uvnitř ČR
a zda netrvá nepřiměřeně dlouhou dobu

PRAKTICKÉ VYUŽITÍ

Možnost využít vzory chování pro **simulaci** běžných uživatelů při testování služeb

Možnost **posílit logiku aplikace** předvídáním požadavků uživatele

Možnost **detekce** práce robotů a vykrádání dat

LITERATURA A ZDROJE

- J. SRIVASTAVA, *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data* University of Minnesota. 2000
- RŮŽIČKA, J., CIBULKA, et al. : Web Map Service Performance Testing based on Extents Generated Randomly or by Algorithm Simulating General User Behaviour SOMAP 2012 Vídeň
- HORÁK, J. a ARDIELLI, J.: Dostupnost a výkonové parametry nového WMS serveru čuzk z pohledu klienta GIS Ostrava 2011

DĚKUJI ZA POZORNOST