

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a

Indexing

Hive

Hadoop

Hosting

Google Cloud

Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta

Využití distribuovaných databázových systémů pro správu vektorových dat v GIS

Diplomová práce

Matěj Krejčí

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Katedra geomatiky

23. 6. 2016

Vývoj počtu vektorových prvků v OSM History datasetu

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

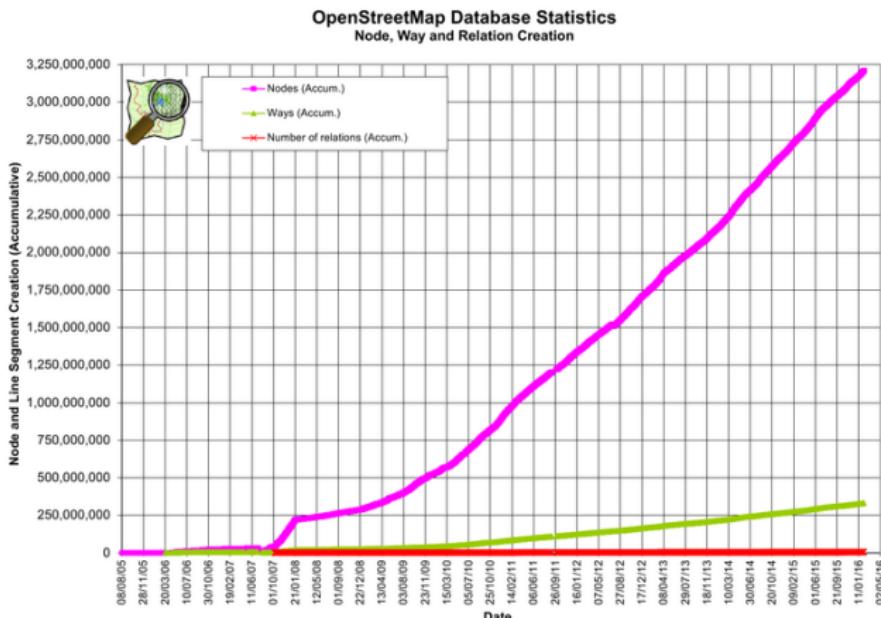
Závěr

Výsledky práce

Navazující práce

Otázky

oponenta



Zdroj: OpenStreetMap Statistics

Vizualizace OSM History dataset pro Evropu

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a

Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

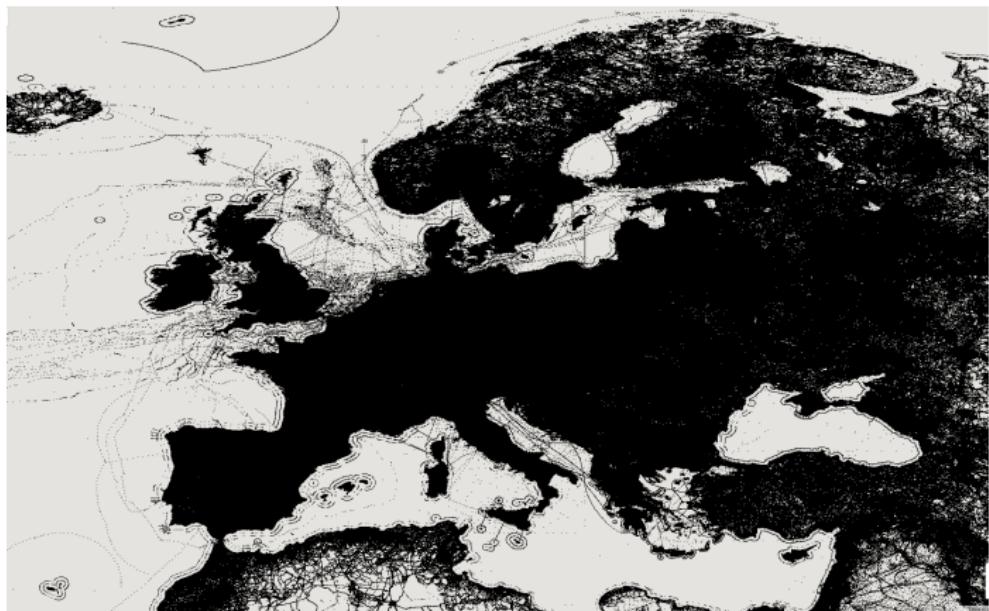
Závěr

Výsledky práce

Navazující práce

Otázky

ponenta



Zdroj: SpatialHadoop

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta



Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

Výsledky práce

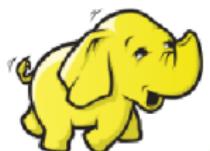
Navazující práce

Otázky

oponenta

Hadoop-GIS

Spatial Big Data Solutions



Počet publikací v databázi ResearchGate s kličovým slovem BigData

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

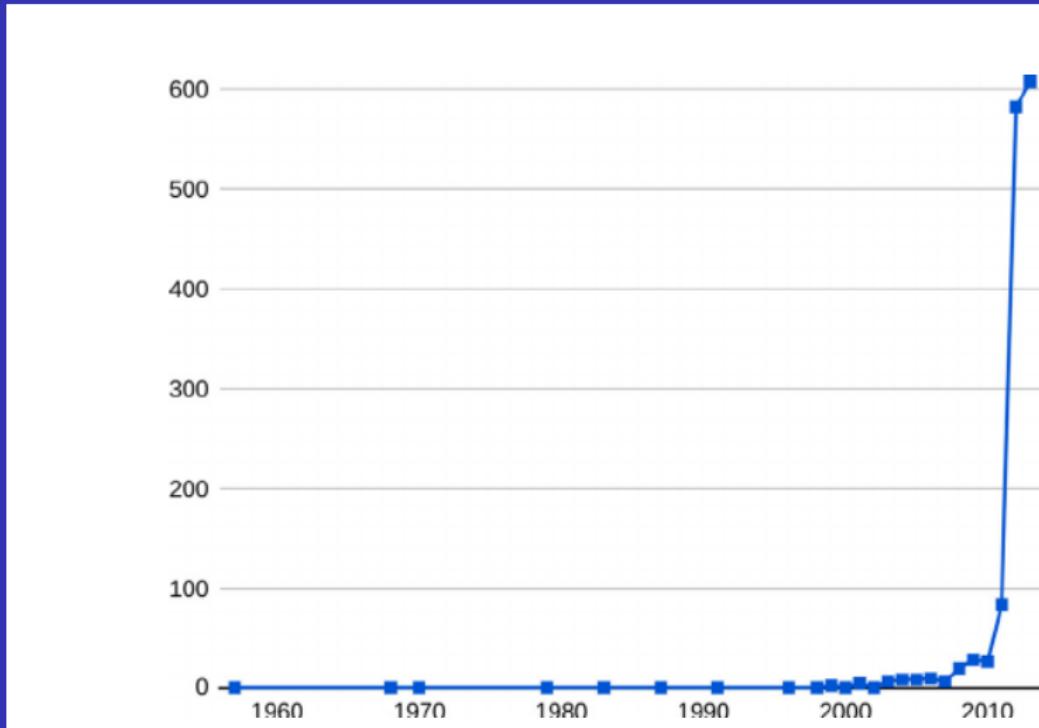
Závěr

Výsledky práce

Navazující práce

Otázky

oponenta



Zdroj: ResearchGate

[Motivace](#)[Úvod](#)[Hadoop](#)[HDFS](#)[MapReduce](#)[GIS Hadoop](#)[GIS Extenze](#)[MapReduce](#)[Partitioning a Indexing](#)[Hive](#)[Hadoop Hosting](#)[Google Cloud Platform](#)[GHF](#)[Funkcionality](#)[Využití](#)[Případová studie](#)[Závěr](#)[Výsledky práce](#)[Navazující práce](#)[Otázky](#)[oponenta](#)

Diplomová práce se věnuje analýze technologií pro zpracování velkého množství dat (tzv. bigdata). Konkrétně je zaměřena na uložení a správu časoprostorových vektorových dat v prostředí distribuovaných databázových systémů jako je např. Hadoop. Cílem praktické části práce je jejich zpřístupnění a umožnění analýz v prostředí desktopového open source GIS nástroje GRASS GIS. Testování navrženého řešení bude prováděno s využitím virtualizované sítě uzlů tvořících cluster.

[Motivace](#)[Úvod](#)[Hadoop](#)[HDFS](#)[MapReduce](#)[GIS Hadoop](#)[GIS Extenze](#)[MapReduce](#)[Partitioning a](#)[Indexing](#)[Hive](#)[Hadoop
Hosting](#)[Google Cloud
Platform](#)[GHF](#)[Funkcionality](#)[Využití](#)[Případová studie](#)[Závěr](#)[Výsledky práce](#)[Navazující práce](#)[Otázky](#)[oponenta](#)

Návrh postupu pro zpracování objemných vektorových datasetů z prostředí GIS

- Teoretický základ Hadoop a jeho GIS extenzí
- Konfigurace Hadoop clusteru hostovaného na Google Cloud Platform
- Vývoj GRASS Hadoop Framework umožňující interakci desktop GIS s Hadoop/Hive a jeho GIS exetenzemi.
- Otestovaní navrženého workflow



Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky
oponenta

- Úvod do Hadoop
- GIS extenze pro Hadoop
- Hadoop konfigurací a cloudové služby
- Seznámení s využitým software
- Případová studie
- Závěr

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a

Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

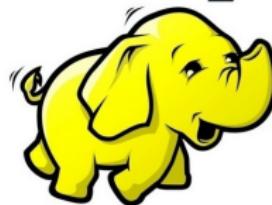
Navazující práce

Otázky

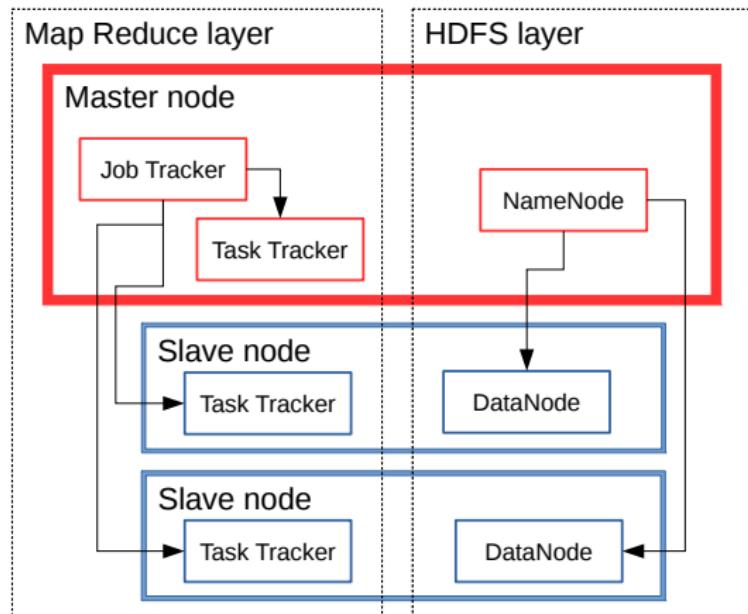
oponenta

- Sada **open-source** komponent pro zpracování velkého množství dat(v řádech petabytů a exabitů)
- Uložení a zpracování objemných dat na velkém množství běžných počítačů.
- Dvě hlavní komponenty:
 - ① Distribuovaný file systém, nativně **HDFS**.
 - ② Framework **MapReduce** pro paralelní výpočty.

hadoop



Dvě hlavní komponenty - HDFS a MapReduce



Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a

Indexing

Hive

Hadoop

Hosting

Google Cloud

Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

Výsledky práce

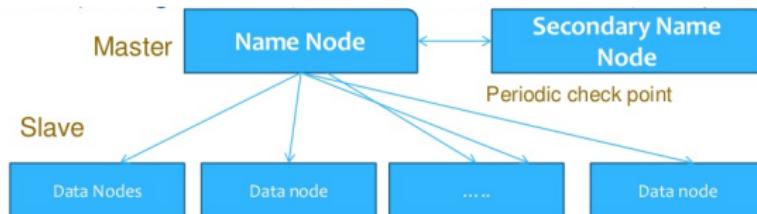
Navazující práce

Otázky

oponenta

Serverové komponenty HDFS

- **NameNode - master** - Udržuje metadata filesystému.
- **DataNode - slave** - Jsou instalovány na každém uzlu a disponují data bloky. Zajišťují čtení a zápis dat od klienta.
- **Secondary NameNode** - periodicky zajišťuje zálohu metadata logů. V případě selhaní NameNode, umožňuje obnovu z poslední zálohy.



Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta

Hlavní vlastnosti

- abstraktní vrstva - umožňuje využití jiné file systémy
 - Nativní distribuovaný filesystem - **Hadoop Distributed File System(HDFS)**
- škálovatelnost
- replikace ->fault tolerant
- vysoká propustnost dat

[Motivace](#)[Úvod](#)[Hadoop](#)[HDFS](#)[MapReduce](#)[GIS Hadoop](#)[GIS Extenze](#)[MapReduce](#)[Partitioning a](#)[Indexing](#)[Hive](#)[Hadoop
Hosting](#)[Google Cloud
Platform](#)[GHF](#)[Funkcionality](#)[Využití](#)[Případová studie](#)[Závěr](#)[Výsledky práce](#)[Navazující práce](#)[Otázky](#)[oponenta](#)

Programovací model pro pralelní zpracování velkého objemu dat s využitím počítačevého clusteru.

Princip MapReduce

- ① **Map** fáze - Každý **Slave** počítač zavolá funkci *Map()* nad lokálními daty a dočasně zapíše mezivýsledky. Master počítač zajišťuje, že je zpracována pouze jedna replika.
- ② **Reduce** fáze - Slave počítače paralelně zpracují jednotlivé skupiny key-value pomocí uživatelem definované funkce *Reduce()*

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
HostingGoogle Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

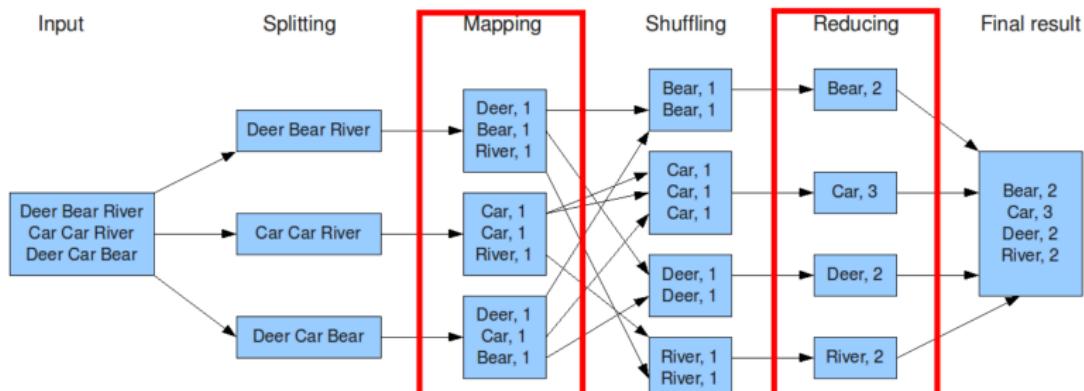
Výsledky práce

Navazující práce

Otázky

oponenta

Četnost slov



Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta

GIS vektorové analýzy paralelně

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a

Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta

Pro MapReduce framework jsou dostupné extenze, které podporují základní GIS operace nad prostrovými daty. Analogii v běžných relačních databázích je PostGIS pro Postgres.

- ① Esri Spatial processing framework for Hadoop
- ② HadoopGIS
- ③ SpatialHadoop

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

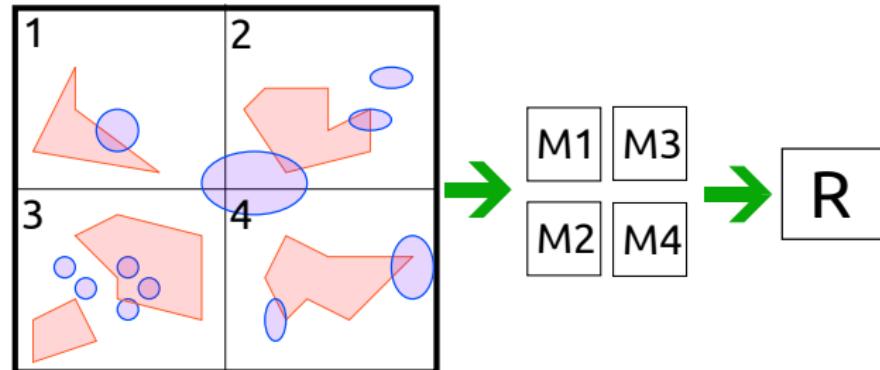
Závěr

Výsledky práce

Navazující práce

Otázky

oponenta



Analýza prostorového vztahu dvou datasetů

- ① Filtering
- ② Partitioning
- ③ Indexing

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

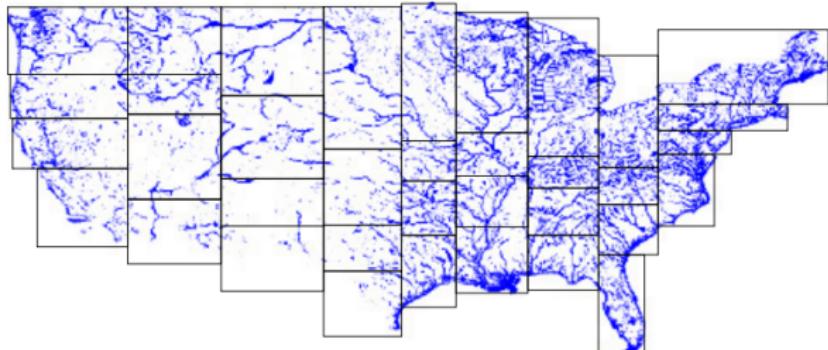
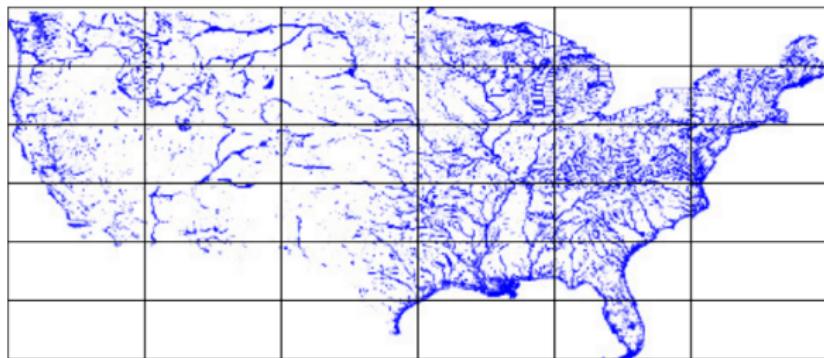
Výsledky práce

Navazující práce

Otázky

oponenta

Grid partitioning a R-tree partitioning



[Motivace](#)[Úvod](#)[Hadoop](#)[HDFS](#)[MapReduce](#)[GIS Hadoop](#)[GIS Extenze](#)[MapReduce](#)[Partitioning a
Indexing](#)[Hive](#)[Hadoop
Hosting](#)[Google Cloud
Platform](#)[GHF](#)[Funkcionality](#)[Využití](#)[Případová studie](#)[Závěr](#)[Výsledky práce](#)[Navazující práce](#)[Otázky](#)[oponenta](#)

Co je Hive? Hadoop nadstavba, která umožňuje optimální správu dat s expresivním dotazovaním podobným jako SQL.



- ① Podpora indexů
- ② Podpora formátů pro uložení dat - text, RCFile, HBase, ORC a další.
- ③ Úložiště pro metadata je v externí RDBMS (default Derby databáze)
- ④ UDFs- funkce definované uživatelem např. BinEnvelope, Intersects
- ⑤ SQL-like query, které jsou implicitně konvertovány do MapReduce funkcí

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta

Hadoop - konfigurace a cloud služby

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta

Konfigurace Hadoop serveru, Google Cloud a localhost

- ① Konfigurace Google Cloud Platform prostředí
- ② Konfigurace Hadoop serveru
- ③ Konfigurace počítačové sítě a firewall
- ④ Konfigurace lokalního počítače

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta

Vývoj software - GRASS Hadoop Framework

GRASS Hadoop Framework - Úvod

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

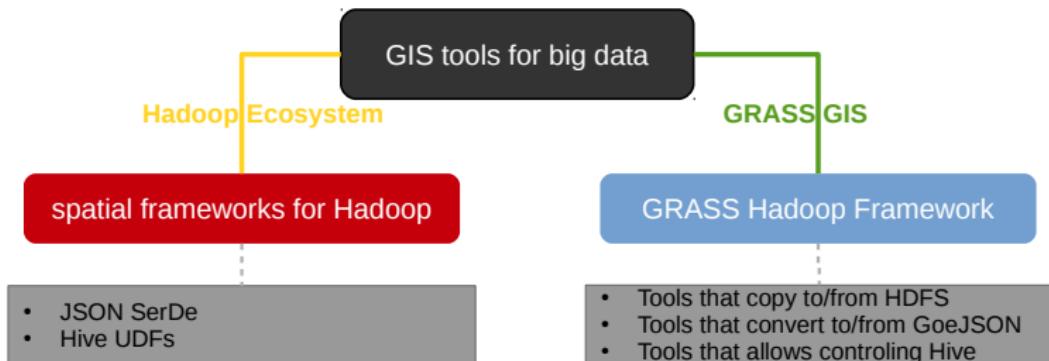
Výsledky práce

Navazující práce

Otázky

oponenta

GRASS Hadoop Framework(GHF) je sada nástrojů umožňující interakci GRASSu s Hadoop/Hive a správu uživatelských připojení.



GRASS Hadoop Framework - Funkce

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

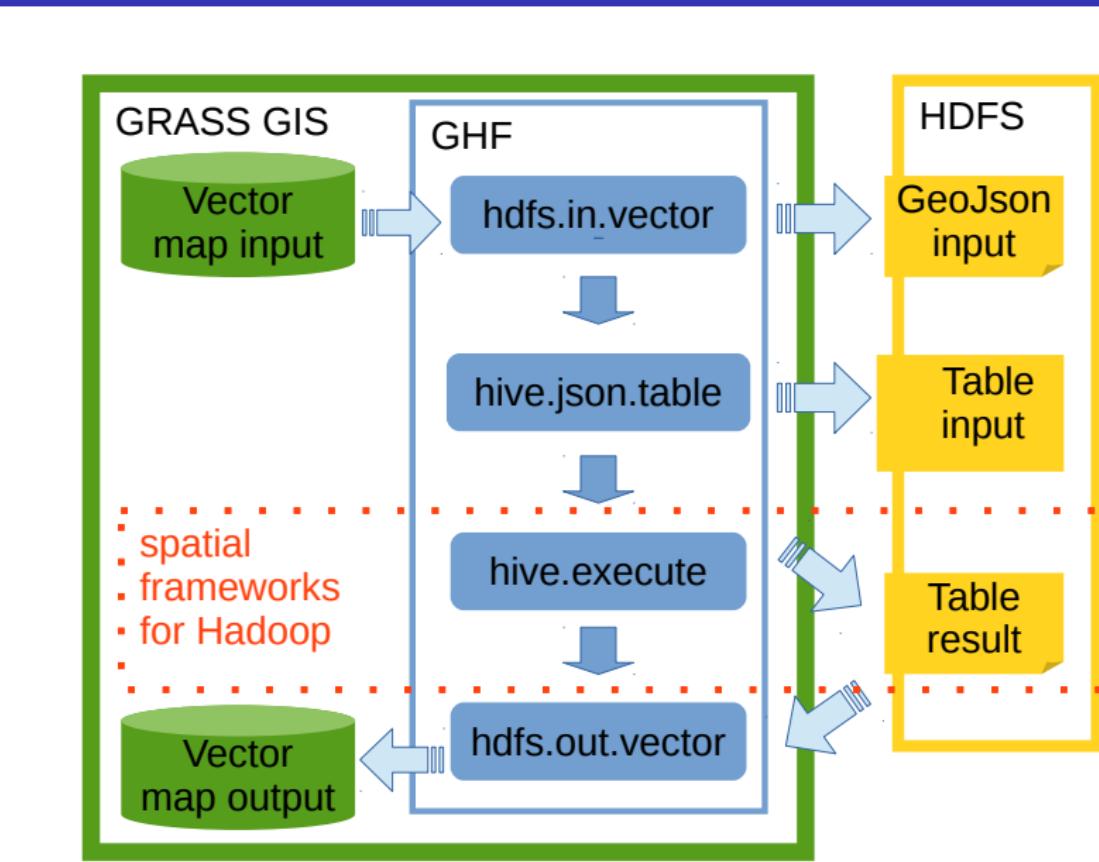
Závěr

Výsledky práce

Navazující práce

Otázky

ponenta



GRASS Hadoop Framework - Implementace

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
HostingGoogle Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

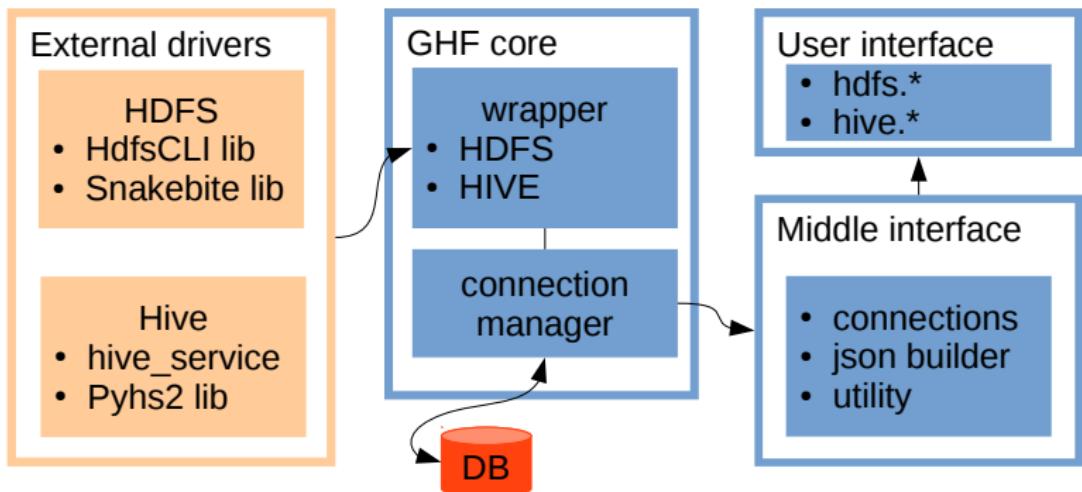
Závěr

Výsledky práce

Navazující práce

Otázky

oponenta



Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky
oponenta

Ovladání modulů:

- z příkazové řádky GRASSu = parametry a přepínače
- z vygenerovaného GUI GRASSu

[Motivace](#)[Úvod](#)[Hadoop](#)[HDFS](#)[MapReduce](#)[GIS Hadoop](#)[GIS Extenze](#)[MapReduce](#)[Partitioning a
Indexing](#)[Hive](#)[Hadoop
Hosting](#)[Google Cloud
Platform](#)[GHF](#)[Funkcionalita](#)[Využití](#)[Případová studie](#)[Závěr](#)[Výsledky práce](#)[Navazující práce](#)[Otázky](#)[oponenta](#)

Informace o datasetu:

- Obsahuje veškeré změny prostorových informací od založení projektu
- Extrakce Evropy obsahuje cca 1.3 miliardy bodů tj. 116 Gb

S využitím *binning* funkce byly body z OSM agregovány do bin(obdelníků) o velikosti 0.1 stupně tj. cca 10km

Vizualizace 1.3 miliardy bodů

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

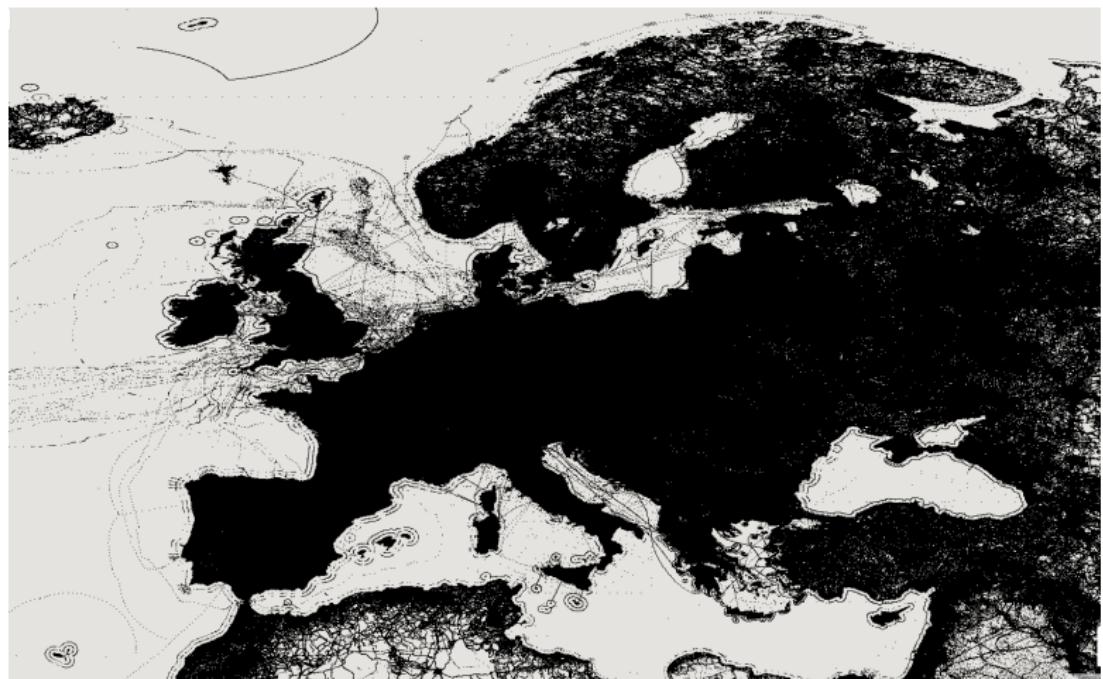
Závěr

Výsledky práce

Navazující práce

Otázky

oponenta



Vizualizace 1.3 miliardy bodů po prostorové agregaci

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

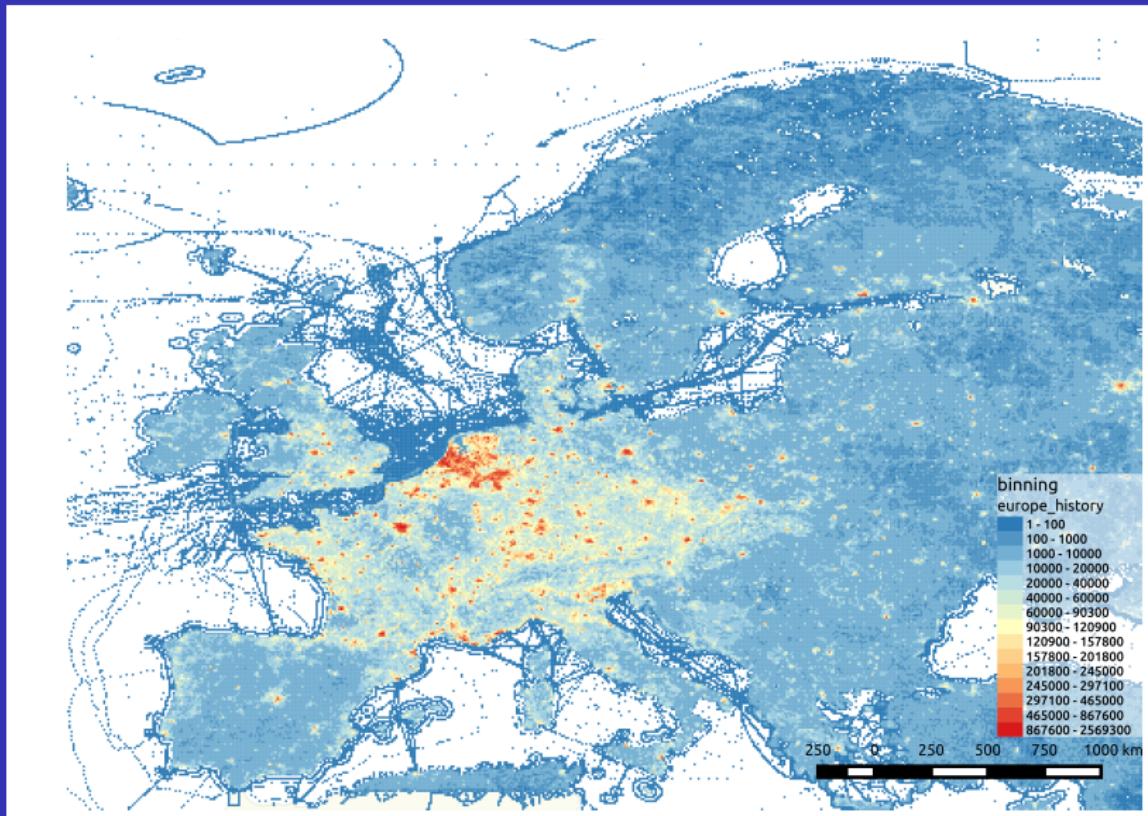
Závěr

Výsledky práce

Navazující práce

Otázky

ponenta



Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
HostingGoogle Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta

Výsledky práce:

- ① Vysvětlení teoretických principů distribuované databáze Hadoop
 - Hadoop Distributed File System
 - MapReduce
- ② Úvod do problematiky GIS extenzí pro Hadoop
 - Analýza teoretických aspektů
 - Porovnání dostupných GIS extenzí pro Hadoop
- ③ Využití cloud služby Google Cloud Platform
 - Konfigurace projektu, sítě, firewall
 - Konfigurace Hadoop clusteru
- ④ Implementace vlastních nástrojů pro GRASS GIS
 - Umožnění využití Hadoop GIS extenzí z desktop GIS

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky

oponenta

Navazující práce:

- Návrh knihoven systematicky implementuje core API nezávisle na GRASSu = možnost implementace Hadoop frameworku pro QGIS bez zásahů do těchto knihoven. Je třeba dopsat pouze konverzi map, která je v QGIS také založena na OGR driveru + implementace GUI pro QGIS.
- Implementace podpory Oozie workflow pro plánování Hadoop úloh

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionality

Využití

Případová studie

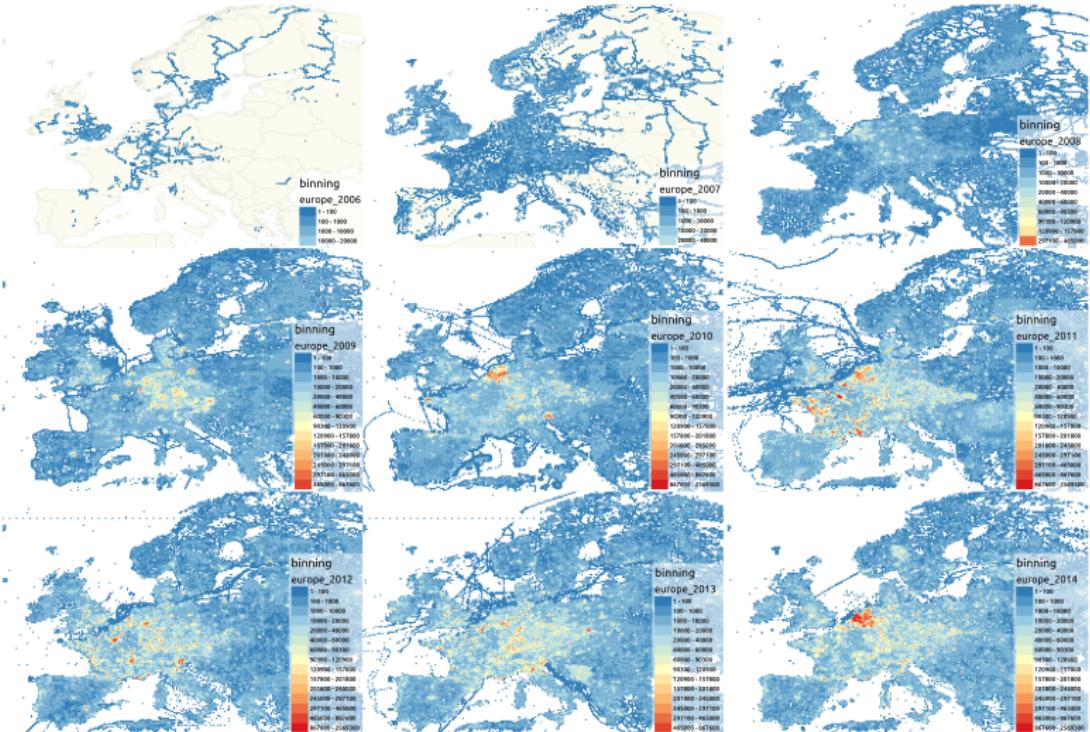
Závěr

Výsledky práce

Navazující práce

Otázky

oponenta



Reakce na poznámky oponenta

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky
oponenta

1

2

3

Reakce na poznámky oponenta

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky
oponenta

1

2

3

Reakce na poznámky oponenta

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky
oponenta

1

2

3

Reakce na poznámky oponenta

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky
oponenta

4

5

6

Reakce na poznámky oponenta

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky
oponenta

4

5

6

Reakce na poznámky oponenta

Matěj Krejčí

Motivace

Úvod

Hadoop

HDFS

MapReduce

GIS Hadoop

GIS Extenze

MapReduce

Partitioning a
Indexing

Hive

Hadoop
Hosting

Google Cloud
Platform

GHF

Funkcionalita

Využití

Případová studie

Závěr

Výsledky práce

Navazující práce

Otázky
oponenta

4

5

6