DĚLAT
DOBRÝ SOFTWARE
NÁS BAVÍ

PROFINIT

A4M33BDT
Technologie pro velká data

MapReduce cvičeni

Sergii Stamenov
Kadlec Tomáš

12`.4.17

WordCount

- Zkopírovat data z `/user/pascepet/map_reduce/imdb_data`
- Upravit kod mapperu aby spracovaval vstupní formát

  `review_id \t review_text`
- Spustit program, prozkoumat vystup
- **(!)** Seřadit vystup podle četností

PROFINIT

Inverted Index

- Napsat MapReduce job který pro každé slovo vypusuje seznam dokumentů ve kterých slovo se vyskytuje
- (!) Odfiltrovat stop-slova. Seznam slov načíst ze souboru. (tip na googlovaní `DistributedCache`.)

PROFINIT

Hadoop Streaming

PROFINIT

Domácí úkol

- Vyzkoušet map-side / reduce-side join

PROFINIT

Další zdroje informace

- Hadoop tutorial https://developer.yahoo.com/hadoop/tutorial/
- MOOC kurz na Udacity https://classroom.udacity.com/courses/ud617
- knížka Hadoop: The Definitive Guide
- knížka MapReduce Design Patterns

Děkujeme
za pozornost

PROFINIT

Profinit, s.r.o.
Tychonova 2, 160 00 Praha 6

Telefon
+ 420 224 316 016

Web
www.profinit.eu

LinkedIn
linkedin.com/company/profinit

Twitter
twitter.com/Profinit_EU