DĚLAT
DOBRÝ SOFTWARE
NÁS BAVÍ

PROFINIT

A4M33BDT
Technologie pro velká data

Spark cvičeni

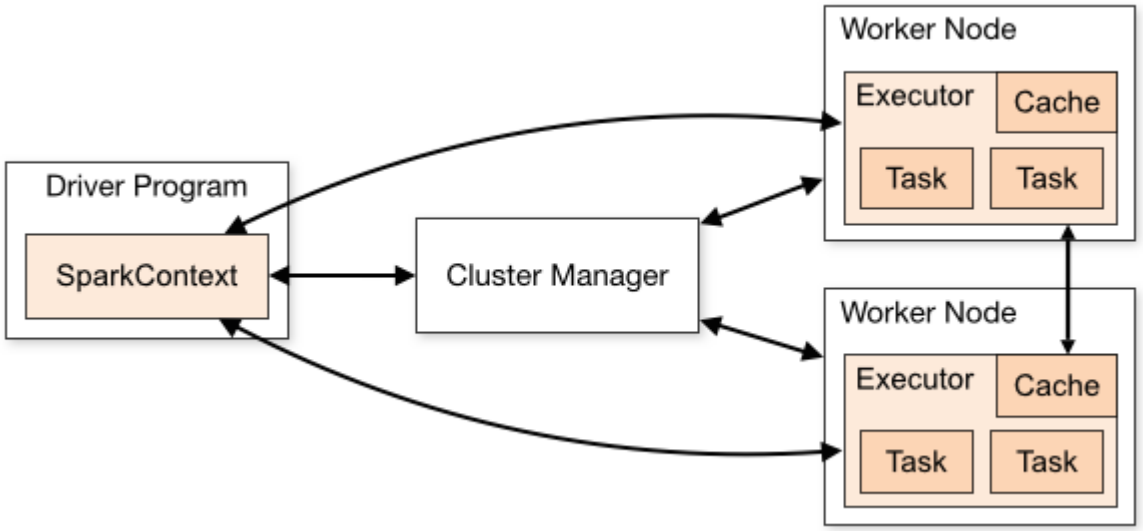
Sergii Stamenov
Hučín Jan

26.4.17

PROFINIT

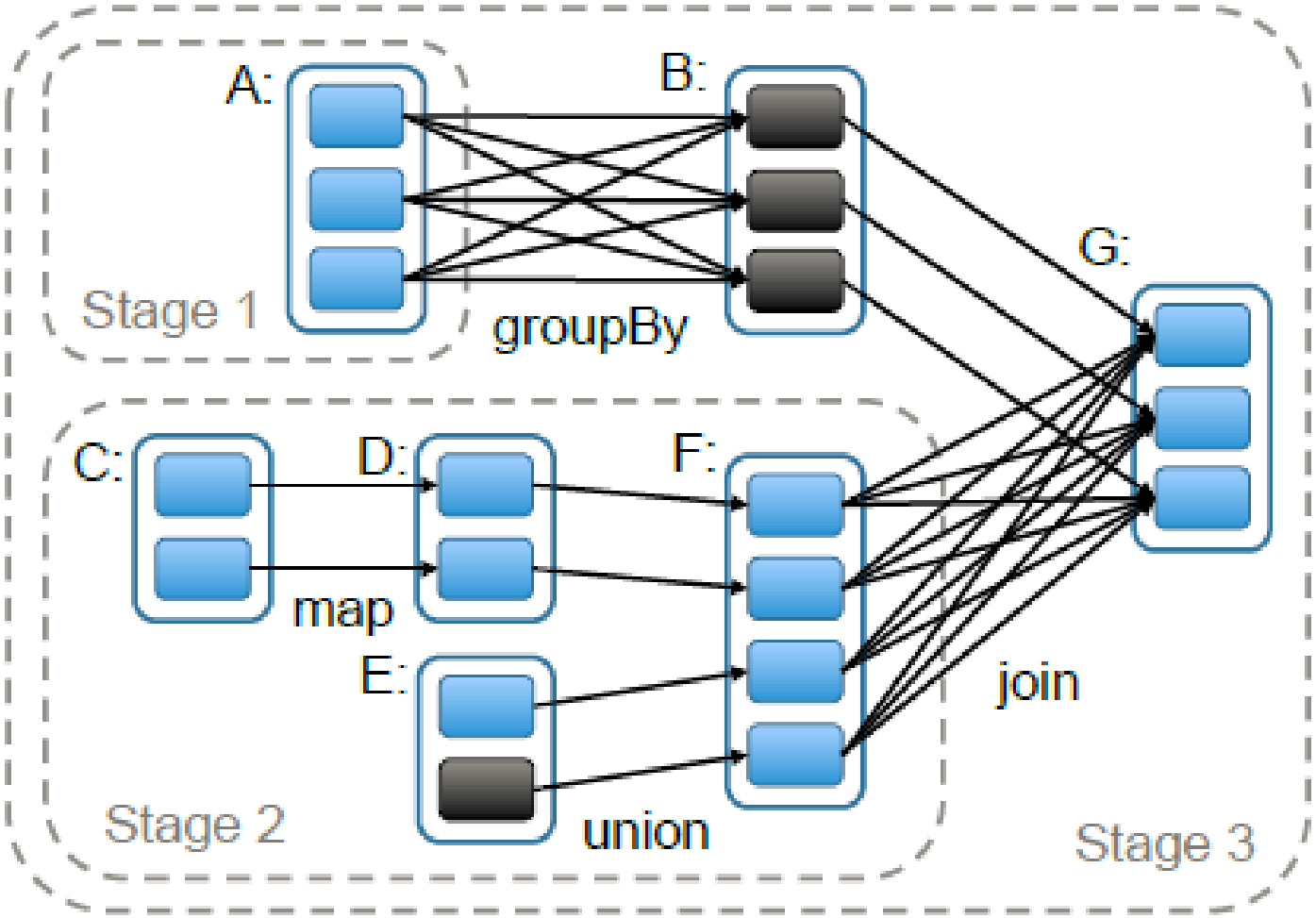
Spark opakováni

Prvky



PROFINIT

Prvky



PROFINIT

PROFINIT

WordCount ukázka

Úloha #1

Data

- 380 000 textu písniček z MetroLyrics
- Formát: csv
- Polička: `index, song, year, artist, genre, lyrics`
- `cesta hdfs:///user/pascepet/lyrics_data/`

Zadáni

- Pro každou písničku spočítat počet slov
- Vypsat 10 nejdelších

Bonus

- Pomoci akumulátoru spočítat počet chyb (viz Spark **Accumulators**)

PROFINIT

Úloha #2

Zadáni

- Pro každé slovo spočítat počet písniček ve kterých se to slovo vyskytuje (aka Inverted Index)
- Vypsat 50 nejčastějších slov

Bonus

- Odfiltrovat stop-slova (viz Spark broadcast variable). Seznam stop-slov `/storage/brno2/home/pascepet/SmartStoplist.txt`
- Napočítat počty per genre

PROFINIT

Další zdroje informace

- Youtube kanal Apache Spark
  https://www.youtube.com/user/TheApacheSpark
- Blogy Cloudery, Databricks

PROFINIT

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