Techniky spracovanie veľkých dát

Metódy inžinierskej práce 2023/2024

Tomáš Zenka

Ústav informatiky, informačných systémov a softvérového inžinierstva
Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave

26. november 2023

O čom to je

V súčasnej ére, kedy sa množstvo dát neustále zväčšuje, stáva sa kľúčovým porozumenie a efektívne spracovanie veľkých objemov informácií. V tejto prezentácii sa ponúka pohľad na nové technológie v oblasti spracovania veľkých dát. Je nevyhnutné pochopiť, aký potenciál majú tieto dáta pre rôzne odvetvia a aké výzvy a príležitosti prinášajú.

Prehľad

Úvod do sveta veľkých dát

- Súbor dát, ktorých veľkosť, komplexnosť a rýchlosť rastu je rapídna
- Zložité na spracovanie a analýzu
- Rýchle tempo digitalizácie
- Za posledné desaťročie sa celkový objem dát zvýšil na 1,8 ZB

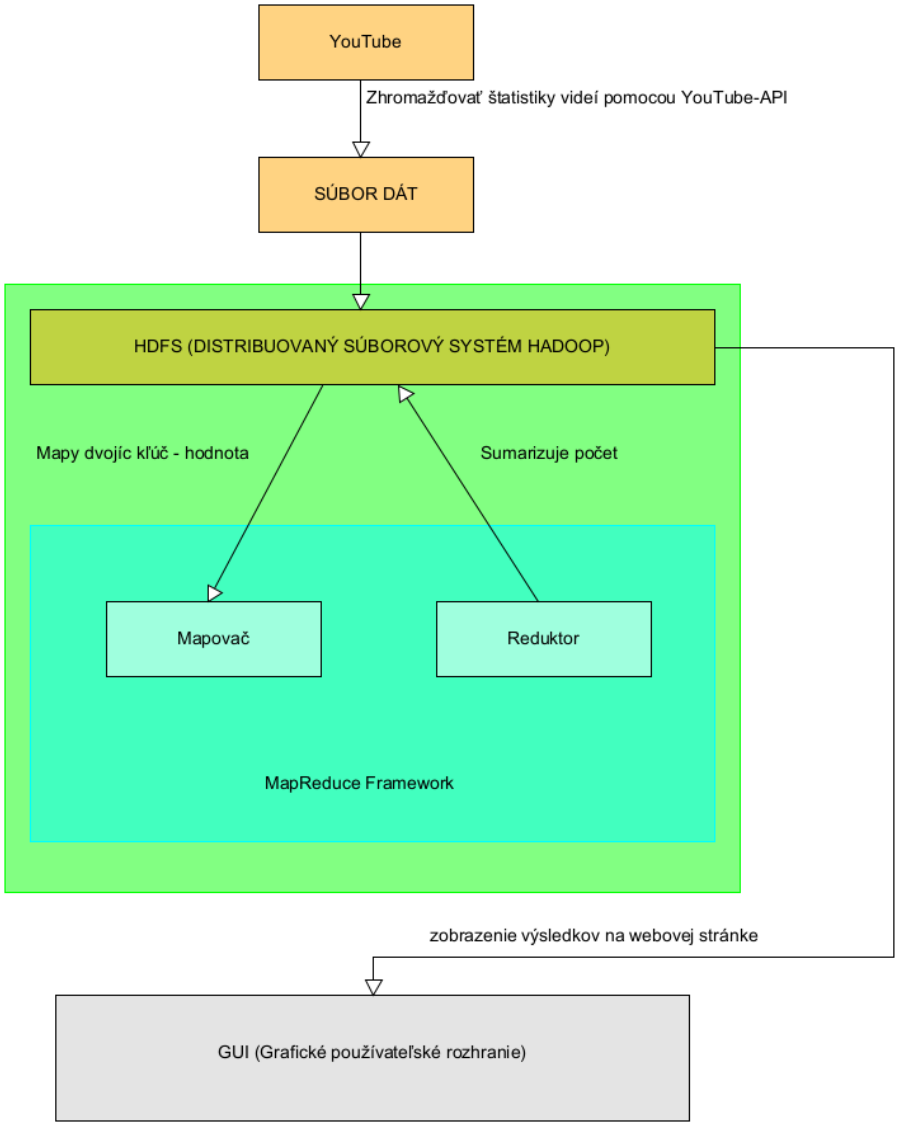
Distribuované systémy na spracovanie dát

- Kľúčový prvok v digitálnom svete
- Distribuovanie výpočtových úloh na viaceré počítače alebo uzly v sieti
- Rýchlosť a škálovateľnosť sú najdôležitejšie aspekty
- Najpoužívanejšie:
  - Hadoop: MapReduce
  - Apache Spark

Hadoop: MapReduce

- Spoločnosť Google MapReduce
- V súčastnosti Appache Hadoop:
  - Hadoop Kernel
  - MapReduce
  - HDFS (Hadoop Distributed File System)
- Uložiť obrovské množstvo dát
- Škálovateľnosť
- Dokáže prežiť zlyhanie významných častí infraštruktúry úložiska

Architektúra Apache Hadoop



Spracované podľa: [6]

Apache Spark

- Nová generáciasystémov na spracovanie veľkých dát
- Hlavné systémy:
  - Spark jadro (core)
  - Upper-level knižnice
- Rýchlejší a všestranejší
- Vďaka knižniciam:
  - Strojové učenie - Spark´s MLlib
  - grafická analýza - GraphX
  - prúdové spracovanie - Spark Streaming
  - spracovanie štruktúrovaných dát - Spark SQL

Hadoop MapReduce verzus Apache Spark

Pri porovnaní by *Apache Spark* bol v mnohých bodoch lepší a to z nasledujúcich dôvodou:

- Zachováva rovnakú možnosť škálovania a odolnosti
- Poskytuje viacstupňový model programovania
- Rýchlejší a jednoduchší na používanie

Zvýraznenie syntaxe

- Na zvýraznenie syntaxe stačí použiť balík listings so správne nastaveným programovacím jazykom

```
int na_druhu(int i) {
    return i * i;
}

int main() {
    printf("%d", na_druhu(118));
    return 0;
}
```

- Jazyk C++ je ešte zaujímavejší: je multiparadigmový[1]

[1] J. O. Coplien. Multi-Paradigm Design for C++. Addison-Wesley, 1998.

Rámiky

Text možno uviesť v rámiku

- Program

```
void main() {
   printf("%d", na_druhu(118));
}

void na_druhu(int i) {
   return i * i;
}
```

- Výstup

```
13924
```

Zhodnotenie a ďalšia práca

- Každá prezentácia musí byť nejako uzavretá
- Ale vždy je čo robiť ďalej...

Zdroje I

[1] Harshawardhan S Bhosale and Devendra P Gadekar. A review paper on bigdata and hadoop. International Journal of Scienti c and Research Publications, 4(10):1-7, 2014.
[2] Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. Data science journal, 14:2-2, 2015.
[3] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. MOBILE NETWORKS & APPLICATIONS, 19(2):171-209, APR 2014.
[4] Bhole Rahul Hiraman, Chapte Viresh M., and Karve Abhijeet C. A study of apache kafka in big data stream processing. In 2018 International Conference on Information , Communication, Engineering and Technology (ICICET), pages 1-3, 2018.

Zdroje II

[5] Wu Jun and Huang Zhixiong. Research on in-memory computing model and data analysis. In 2015 8th International Conference on Intelligent Computation Technology and Automation (ICICTA), pages 726-729, 2015.
[6] PrathyushaRani Merla and Yiheng Liang. Data analysis using hadoop mapreduce environment. In 2017 IEEE International Conference on Big Data (Big Data), pages 4783-4785, 2017.
[7] Salman Salloum, Ruslan Dautov, Xiaojun Chen, Patrick Xiaogang Peng, and Joshua Zhexue Huang. Big data analytics on apache spark. International Journal of Data Science and Analytics, 1:145-164, 2016.
[8] Eman Shaikh, Iman Mohiuddin, Yasmeen Alufaisan, and Irum Nahvi. Apache spark: A big data processing engine. In 2019 2nd

Zdroje III

IEEE Middle East and North Africa COMMunications Conference (MENACOMM), pages 1-6, 2019.
[9] Chitresh Verma and Rajiv Pandey. Comparative analysis of gfs and hdfs: Technology and architectural landscape. In 2018 10th International Conference on Computational Intelligence and Communication Networks (CICN), pages 54-58, 2018.