Techniky spracovanie veľkých dát*

Tomáš Zenka
Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
xzenka@stuba.sk

5. november 2023

Abstrakt

Článok skúma a porovnáva techniky spracovania veľkého množstva dát, čo je kľúčové v dnešnej digitálnej dobe, kde sa generuje obrovské množstvo informácií. Cieľom je poskytnúť prehľad o moderných prístupoch a nástrojoch určených na manipuláciu s masívnymi dátovými súbormi. Tieto nástroje zahŕňajú distribuované systémy na spracovanie dát, algoritmy strojového učenia a metriky na hodnotenie kvality dát. Dôraz sa kladie na potrebu rýchleho spracovania dát v reálnom čase, čo umožňuje rýchlu analýzu a tvorbu hodnotných poznatkov z týchto objemných dátových zdrojov. Článok taktiež uvádza príklady aplikácií v rôznych odvetviach, ako je medicína, finančníctvo a priemysel. Spracovanie veľkého množstva dát sa stáva nevyhnutným nástrojom pre konkurencieschopnosť a inovácie v súčasnom digitálnom prostredí.

*Semestrálny projekt v predmete Metódy inžinierskej práce, ak. rok 2023/24, vedenie: Vladimír Mlynarovič

1 Úvod

Pojem "veľké dáta" (Big Data) odkazuje na súbor dát, ktorých veľkosť, komplexnosť a rýchlosť rastu je rapídna. Preto sú zložité na spracovanie a analýzu. [1] Cieľom článku je poskytnúť prehľad o prístupoch a technikách na manipuláciu s týmito dátami. Tieto nástroje zahŕňajú distrubuované systémy na spracovanie dát 2.1 a metriky na hodnotenie kvality dát 2.2. Článok sa zameriava aj na rýchle spracovanie dát v reálnom čase 3. Na záver sú uvedené aplikácie v rôznych odvetviach 4.

2 Techniky spracovania veľkého množstva dát

Rýchle tempo digitalizácie vytvára obrovské množstvo dát.V dnešnej digitálnej dobe je dôležitá výzva spracovanie veľkého množstva dát. Len za posledné desaťročia sa celkový počet dát na svete zvýšil na 1,8 ZB [3]. Preto boli na tieto účely spracovanie týchto dát vyvinuté rôzne techniky a nástroje. Umožňujú používateľovi efektívne manipulovať a pracovať s masívnymi dátovými súbormi.

2.1 Distribuované systémy na spracovanie dát

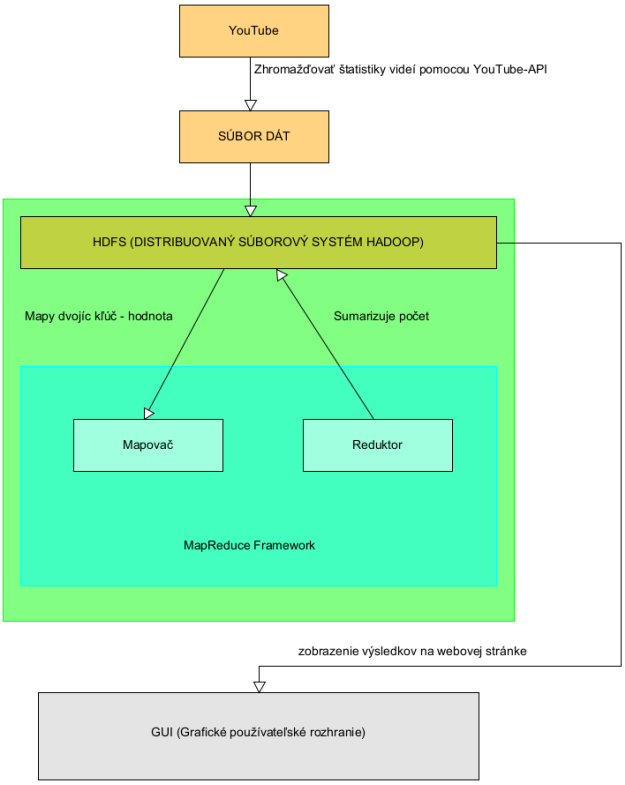
Distribuované systémy na spracovanie dát predstavujú kľúčový prvok v digitálnom svete. Rýchlosť a škálovateľnosť sú najdôležitejšie aspekty. Rozoberieme si základné systémy, ktoré umožňujú rýchlo a efektívne získavať relevantné informácie z masívnych dátových sád.

2.1.1 Hadoop: MapReduce

Hadoop je programovací framework (rámec) na podporu spracovania veľkých dátových súborov. Bol vyvinútý spoločnosťou Google MapReduce. V súčastnosti sa v praxi používa Appache Hadoop, ktorý sa rozdeľuje na rôzne časti:

1. Hadoop Kernel
2. MapReduce
3. HDFS

Hlavnou výhodou frameworku Hadoop je úložný systém odolný voči chýbam s názvom Hadoop Distributed File System (HDFS). HDFS rozdeluje systém súborov do 128 MB blokov [9]. Na obrázku 1 je vykreslená architektúra HDFS. Je schopný uložiť obrovské množstvo informácií, postupne sa škálovať a to najdôležitej je, že dokáže prežiť zlyhanie významných častí infraštruktúry úložsika.



Obr. 1: Architektura HDFS, spracované podľa [6]

Hadoop vytvára zhluky strojov a koordinuje prácu medzi nimi. Ak jeden zlyhá tak Hadoop pokračuje bez straty. Prácu zlyhaného článku prehodí na zvyšné počítače.

Hlavnou zložkou ekosystému Hadoop je MapReducce framework. Umožňuje rozdelenie problému a dát na menšie a spustiť ich paralelne. MapReduce má dve funkcie:

- map - ako vstup hodnota/klúč páru a generuje intemediate set párov klúčov/hodnôt
- reduce - spája intermediate set hodnôt s rovnakým intermediate set klúčov

[1]

2.1.2 Apache Spark

Apache Spark je nová generácia systémov na spracovanie veľkých dát. Systém Apache Spark pozostáva z hlavných systémov:

- Spark core (jadro)
- Upper-level libraries

V porovnaní s Hadoop MapReduce je Apache Spark rýchlejší a všestranejší. Vďaka knižniciam sa dá použiť na strojové učenie (knižnica Spark's MLlib), grafická analýza (knižnica GraphX), prúdové spracovanie (knižnica Spark Streaming) a aj na spracovanie štruktúrovaných dát (knižnica Spark SQL). Kombinuje jadro pre distrubuované výpočty s pokročilým programovacím modelom pre spracovanie v pamäti (in-memory processing) 3.1. Zachováva rovnakú možnosť škálovania a odolnosti voči chybám ako Hadoop MapReduce, avšak poskytuje viacstupňový model programovania. Celkovo je rýchlejší a oveľa jednoduchši na používanie. [7]

2.2 Metriky na hodnotenie kvality dát

Veľké dáta sú novým konceptom a ešte nie je zaužívaný štandart na hodnotenie kvality dát v oblasti veľkých dát. V literatúre nájdeme mnoho rôznych definícii, ale jednu vec majú spoločnú. A to, že kvalita údajov závisí od viacerých vlastností. Predovšetkým od podnikateľského prostredia , ktoré dáta používa.

[2]

3 Rýchle spracovanie dát v reálnom čase

Táto sekcia sa zameriava a kladie dôraz na rýchle a efektívne spracovanie dát v danom čase. V súčastnom digitálnom prostredí sa obzvlásť kladie dôraz na analýzu dát v okamihu kedy sú generované.

3.1 Spracovanie v pamäti

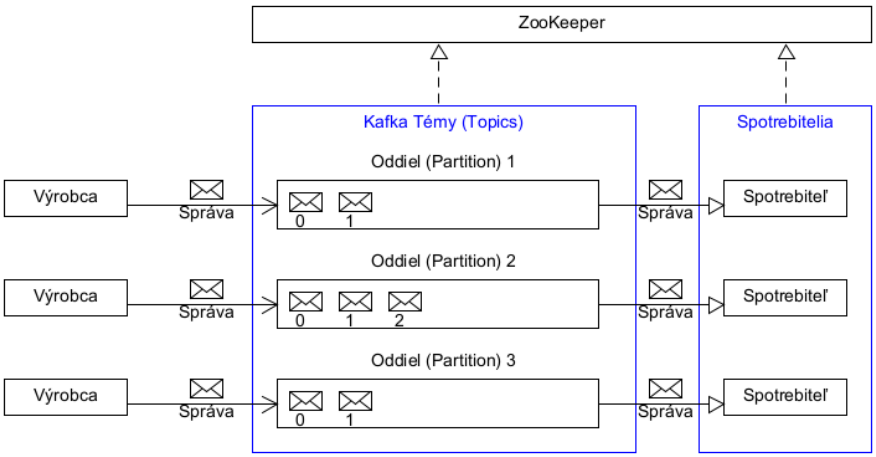
Spracovanie v pamäti je základom pre Apache Spark. Umožňuje mu to ukladať prechodné dáta do pamäte, čo má za následok to, že namiesto aby všetky dáta išli na disk a neskôr sa vyberali z disku, tak sú uložené dočasne v pamäti kde čakujú na spracovanie. Toto urýchluje celý proces spracovania dát. [7]

V databázach, ktoré používajú in-memory processing sa veľké množstvo dát generuje v reálnom čase. Je potrebné ho v reálnom čase aj spracovať. Väčšina dát prichádza priamo do pamäte na spracovanie, len málo dát ide na pevný disk na dlhodobé uloženie. [5]

3.2 Systémy na spracovanie tokov

Jedná sa o klúčové nastroje v prípade ak nám priebežne prúdi veľké množstvo dát. Dokážu ich rýchlo a efektívne spracovať. Takéto dáta môžu byť napríklad dáta zo senzorov alebo zo sociálnych medií.

Momentálne najpoužívanejší sytém je Apache Kafka. Na obrázku 2 je rozpísaná architektúra Apache Kafka. Výhody Apache Kafka sú, že je škálovatelná a spoľahlivá. [4]



Obr. 2: Architektura Apache Kafka, spracované podľa [4]

4 Aplikácie v rôznych odvetviach

V praxi sa využíva hlavne framework Apache Spark v rôznych odvetviach ako napríklad:

- Zdravotníctvo 4.1
- Financie 4.2
- Zábavný priemysel 4.3

4.1 Zdravotníctvo

Využíva sa na analýzu zdravotných záznamov pacienta. Pomáha identifikovať či je pacient náchylný ku zdravotným komplikácia v budúcnosti. Taktiež spracováva dáta z genómového sekvenovania (genomic sequencing).

4.2 Financie

Poskytuje aktuálny prehľad, ktorý pomáha pri robení správnych rozhodnutí. Napríklad v oblasti segmentácie zákazníkov, hodnotení úverového rizika alebo pri cielenej reklame.

4.3 Zábavný priemysel

Hlavne v oblasti videoherného priemyslu pomáha s rozpoznávaním vzorov. Neskôr využité na selektívnu reklamu alebo automatická zmena náročnosti na základe hráčskych schopností.

[8]

5 Záver

Spracovanie veľkého množstva dát sa stalo kľúčovou súčasťou 21. storočia. S narastajúcim objemom dát, ktoré sa podľa predpokladov búdu zdojnásobovať každé dva roky v blízkej budúcnosti [3], sa nástroje na ich spracovanie stávajú povinnosťou pre konkurencieschopnosť a inováciu. Predstavili sme si moderné techniky spracovania veľkých dát, ktoré sú rýchle a efektívne. Taktiež je veľmi dôležité spracovanie v realnom čase. S ohľadom na budúcnosť si myslím, že táto oblasť bude ďalej rásť a rozvíjať sa.

Literatúra

[1] Harshawardhan S Bhosale and Devendra P Gadekar. A review paper on big data and hadoop. *International Journal of Scientific 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.

[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 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.