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.

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, algoritmy strojového učenia **??** a metriky na hodnotenie kvality dát **??**. Článok sa zameriava aj na rýchle spracovanie dát v reálnom čase **??**. Na záver sú uvedené aplikácie v rôznych odvetviach **??**.

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 [2]. Preto boli na tieto účely spracovanie týchto dát vyvinuté rôzne techniky a nástroje. Umožňujú používateľovi efektívne manipulovať a pracovať s masívnymi dátovými súbormi.

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

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). Je schopný uložiť obrovské množstvo informácií, postupne sa škálovať a to najdôležitej je, že dokáže prežiť zlyhanie významných častí infraštruktúry úložsika.

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). Zachováva rovnakú možnosť škálovania a odolnosti voči chybám ako Hadoop MapReduce, avšak poskytuje viacstupňový model programovania. Celkovo je rýchlejší a oveľa jednoduchši na používanie.

3 Záver

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] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *MOBILE NETWORKS & APPLICATIONS*, 19(2):171–209, APR 2014.