APACHE HADOOP

MADIUDIN Radu

Universitatea Tehnică a Moldovei

Abstract: *Articolul dat relevă o scurtă introducere în frameworkul Apache Hadoop - framework predestinat lucrului cu masive de informație, oferind o performanță și flexibilitate deosebită. Este descrisă premisa apariției acestuia. Principiile ce stau la baza funcționării reprezintă nucleul frameworkului, constituit din Hadoop HDFS (Hadoop Distributed File System) si MapReduce. Arhitectura aleasă a fost creată în jurul algoritmului MapReduce. Este arătat domeniul utilizării Hadoop și de ce este unul dintre soluțiile viitorului de stocare și prelucrare a informației.*

Cuvinte cheie: *HADOOP, framework, big data, MapReduce, HDFS, distributed, YARN, Spark, Hive.*

1. Premise

Ne aflăm în era IOT, rata cantității de informație obținute creste la un nivel comparabil celui exponențial. Domeniile din care informația provine sunt rețele sociale, medicina, secvența genomului, meteo, sateliți, comportamentul cumpărătorilor, informația pe vânzări, finanțe, log file etc. Tot mai multe domenii implementează tehnologiile informaționale cu scopul creșterii productivității creând un aflux puternic de date. Toata informația obținută nu are nici o valoare daca nu este analizată. Prin big data înțelegem o colecție de datasets ce nu pot fi procesate utilizând tehnici tradiționale de computare. Aceasta include un volum și o varietate mare ce poate fi împărțită în trei tipuri: informație structurata (informație relațională), informație semi structurata (informație XML), informație nestructurata (Word, PDF, Text, Media Logs).

Metodele tradiționale de soluționare (sistemul centralizat de stocare) ale acestei probleme au devenit inefective. Un sistem centralizat de prelucrare și stocare a informației este efectiv în cazul unui volum mai mic de informație, odată cu creșterea volumului de date apare problema capacității de procesare și network throughput.

2. Apache Hadoop

Google a soluționat problema prin utilizarea algoritmului numit MapReduce (figura 1). Acest algoritm divizează taskurile în parți mici și le atribuie la mai multe calculatoare conectate într-o rețea, colectând rezultatul sub forma de final result dataset.

Hadoop este un open source framework Apache scris în Java care a fost creat pe baza principiului sus menționat. Fiind proiectat sa fie scalabil de la un singur server la mii de mașini, fiecare oferind stocare și computare locală , acesta este capabil sa permită dezvoltarea unei aplicații ce ar putea efectua o statistica completa la cantități mari de informație. Toate modulele în Hadoop sunt proiectate cu presupunerea fundamentală ca eșecurile hardware se întâlnesc des și trebuie sa fie tratate automat de framework.

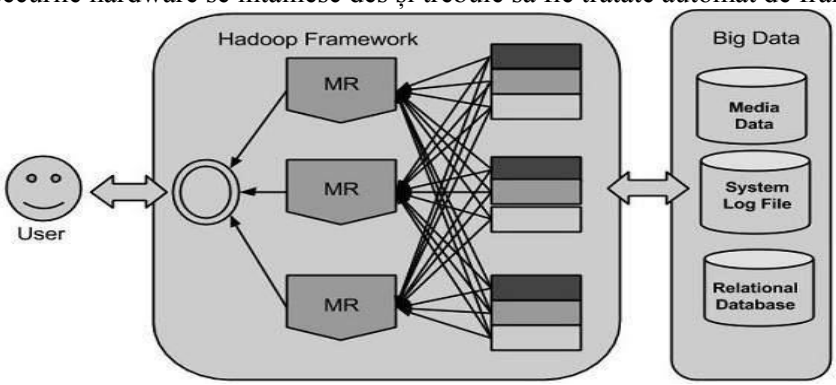

Fig.1. MapReduce

3. Hadoop Architecture

Nucleul Apache Hadoop constă din partea storage, cunoscută ca Hadoop Distributed File System (HDFS), și partea procesuala care este modelul programat al algoritmului MapReduce. Frameworkul de baza este compus din următoarele module:

- Hadoop Common: conține librarii si utilități necesare funcționării altor module Hadoop. Aceste librarii asigură abstracții la nivel filesystem și OS, și conțin java files și scripts necesare pentru startarea Hadoop.
- Hadoop YARN: este un framework pentru job scheduling și cluster resource management.
- Hadoop Distributed File System (HDFS): un sistem distribuit de fișiere care asigura capacitate ridicata de acces la informația aplicației.
- Hadoop MapReduce: este un sistem YARN utilizat în procesarea paralelă a seturilor largi de informație.

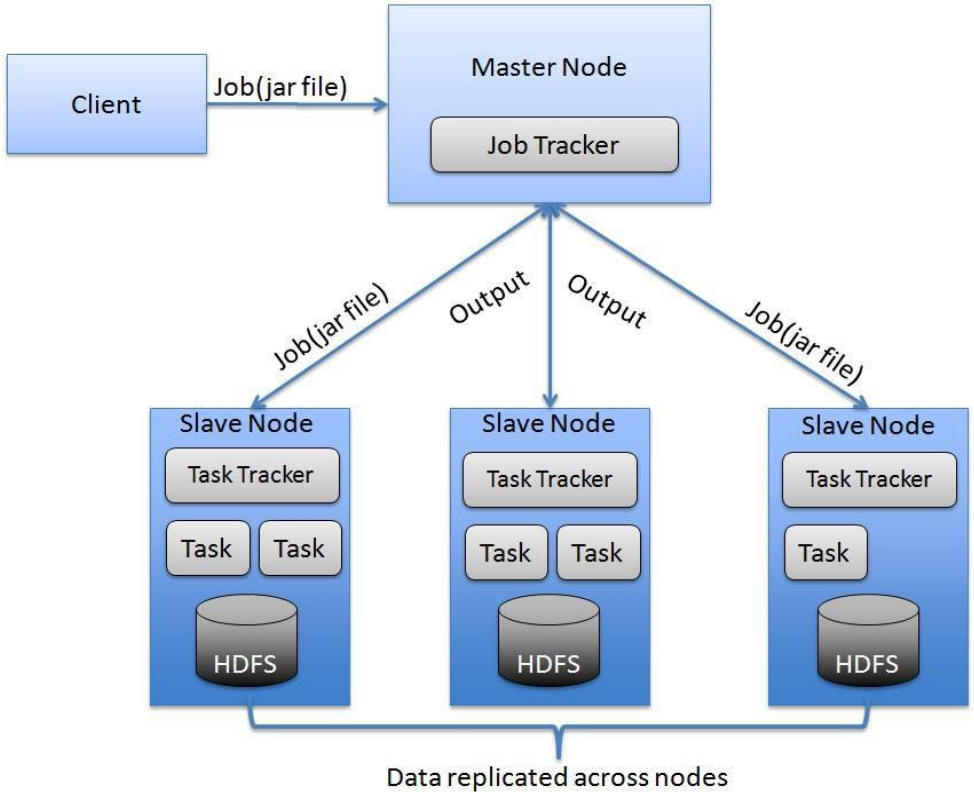

Fig. 2. Procesarea informației

4. Cum funcționează Hadoop?

Stage 1:

Un utilizator/aplicație poate înainta un job către Hadoop (figura 2) pentru procesul necesar prin specificarea următoarelor puncte:

- locația fișierelor input și output în sistemul distribuit.
- clasele Java sub formă de fișiere jar ce conțin implementarea funcțiilor map și reduce.
- configurația job prin setarea diferitor parametri specifici jobului.

Stage 2:

Clientul job Hadoop înaintează jobul (jar/executabil) și configurația către JobTracker care preia responsabilitatea distribuirii softului/configurației la slaves, planificarea taskurilor și monitorizarea lor, furnizând statusul și diagnosticul informației către job-client.

Stage 3:

TaskTrackers pe noduri diferite execută taskurile conform implementării MapReduce și rezultatul funcției reduce este salvat în fișierele output pe sistemul de fișiere (figura 3, 4).

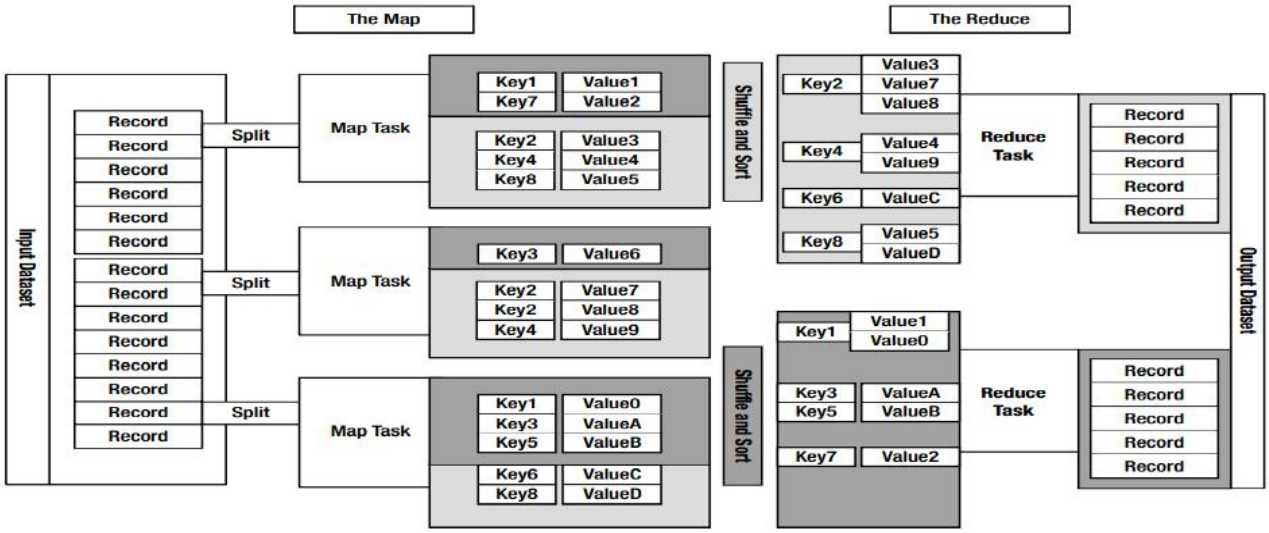

Fig. 3. Principiul de funcţionare a frameworkului

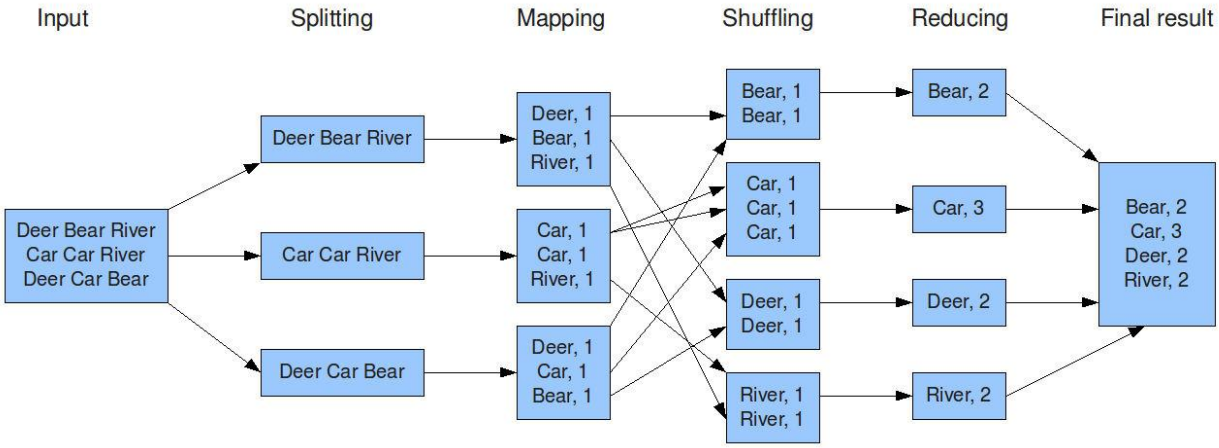

Fig. 4. Exemplu practic a algoritmului MapReduce

5. Hadoop Ecosystem

Ecosistemul Hadoop include atât proiecte Apache open source cât şi o gama largă de instrumente şi soluţii comerciale (figura 5). Unele dintre cele mai cunoscute open source sunt Spark, Hive, Pig, Oozie si Sqoop. Opţiunile comerciale sunt încă mai diverse, incluzând platforme şi distribuţii de la vânzători ca Cloudera, MapR, plus o varietate de instrumente specifice sarcinilor Hadoop de dezvoltate, producere şi mentenanţă

Spark este atât un model de programare cât şi unul computaţional. Acesta oferă o poarta la computarea în memorie pentru Hadoop, ce reprezintă unul din motivele popularităţii lui şi o adopţie larga. Spark oferă o alternativa la MapReduce ce permite executarea sarcinilor în memorie, în schimb pe disk. Utilizând computarea în memorie, volumul de lucru rulează de la 10 la 100 de ori mai repede în comparaţie cu execuţia pe disk.

Hive reprezintă un soft de depozitare a datelor care răspunde de faptul cum informaţia este structurată şi interogată în clusterele distribuite Hadoop. Hive este, de asemenea, un mediu popular utilizat în scrierea interogărilor pentru informaţia în mediul Hadoop. Oferă instrumente ETL (Extract, Transform and Load) ce aduc mediului capabilităţi asemanătoare SQL.

Pig este un limbaj procedural utilizat în dezvoltarea aplicaţiilor procesate în paralel pentru seturi mari de data în mediul Hadoop. Pig este o alternativă programării Java. Pig include Pig Latin, ce este un scripting limbaj. Pig traduce scripturile Pig Latin în MapReduce, ce apoi pot fi rulate pe YARN şi procesarea informaţiei în clusterele HDFS. Este popular din motivul automatizării unei părţi complexe a dezvoltării MapReduce.

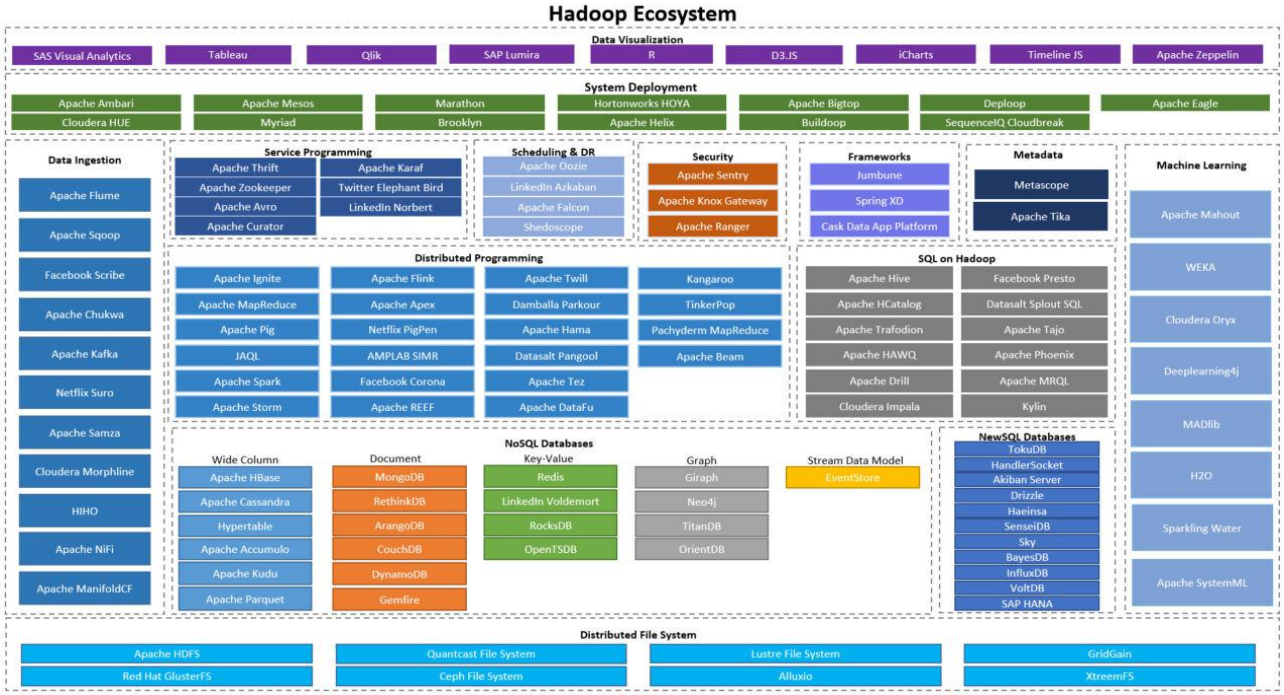

Fig. 5. Familia Hadoop

6. Concluzii

Hadoop este una dintre cele mai performante și larg utilizate sisteme în manipularea volumelor mari de informație rapid și ieftin. Acesta permite o flexibilitate deosebită prin diversitatea modulelor oferite, o posibilă scalabilitate lineară fară pierderea vitezei de analiză și procesare. Ofera o redundanța a datelor datorita stocarii informației în clustere de masini. Datorită acestor calitați Hadoop a fost pe larg implementat de marii giganți software ca Google, Amazon, Facebook etc. Acest framework a devenit o parte a dezvoltării curentului IOT facilitând și iplementând aspectul descentralizat al acestuia, oferind suportul necesar funcționării eficiente.

Bibliografie

1. Alex Homes. *Hadoop în Practice.* Second Edition. Publisher: Manning Publications, September 2014, ISBN: 9781617292224.
2. Jonathan R. Owens, Jon Lentz, Brin Femiano. *Hadoop Real World Solutions CookBook.* [Resursa electronica] – Regim de acces: https://www.amazon.com/Hadoop-Real-World-Solutions-Cookbook-ebook/dp/B00AIVQE3I
3. Chuck Lam. *Hadoop în Action.* Publisher: Manning Publications. December 2010, ISBN: 9781935182191.
4. *Tutorialspoint* [Resursa electronica] – Regim de acces: https://www.tutorialspoint.com/hadoop
5. Apache_Hadoop#Architecture, Wikipedia [Resursa electronica] – Regim de acces: https://en.wikipedia.org/wiki/Apache_Hadoop#Architecture
6. Hadoop Ecosystem and Components [Resursa electronica] – Regim de acces: http://www.bmc.com/guides/hadoop-ecosystem.html