# BIG DATA – CONCEPTE, ARHITECTURI ŞI TEHNOLOGII

**Adriana ALEXANDRU** **Dora COARDOŞ**

adriana@ici.ro coardos@ici.ro

Institutul Naţional de Cercetare-Dezvoltare în Informatică - ICI Bucureşti

**Rezumat:** Lucrarea prezintă principalele abordări legate de Big Data. Conceptul Big Data introduce modificări în cadrul a trei dimensiuni: (1) tipuri de date, (2) viteza de acumulare a acestora şi (3) volumul lor. Big Data reprezintă o nouă generaţie de tehnologii şi arhitecturi destinate extragerii de valoare din cadrul volumelor foarte mari de date, cu o mare varietate, permiţând prelucrarea şi analiza acestora în timp real. În articol sunt prezentate seturi de tip Big Data, precum şi arhitecturi şi tehnologii pentru Big Data.

**Cuvinte cheie:** Big Data, seturi Big Data, Hadoop, NoSQL.

**Abstract:** This paper presents the main approaches related to Big Data. Big Data concept introduces changes in three dimensions: (1) data types, (2) speed of data accumulation and (3) data volume. Big Data represents a new generation of technologies and architectures for extracting value from the very large volumes of data that have a wide variety, and allow real time processing and analysis. This article presents Big Data sets, architectures, and technologies.

**Keywords:** Big Data, Hadoop, NoSQL, Big Data sets.

## 1. Introducere

În ultimele decenii, organizaţiile au început să acorde importanţă sporită datelor şi să investească mai mult în colectarea şi gestionarea lor. Dincolo de informaţiile colectate în interiorul organizaţiilor şi de volumul crescând de date pe care le generează calculatoarele în funcţionarea lor, sunt utilizate date obţinute din exteriorul organizaţiei, fie structurate sau nestructurate, care au surse multiple care pot include de la informaţii postate pe reţele de socializare şi produse vizionate în magazine virtuale, la informaţii citite de către senzori, semnale GPS de pe dispozitivele mobile, adrese IP ale compu-terelor, cookie-uri, coduri de bare ş.a.m.d. Unele tipuri de date precum text şi voce, există de mult timp, însă volumul acestora în mediul Internet şi în alte structuri digitale anunţă începutul unei noi ere, precum şi a unor noi tehnologii care permit analizarea acestor tipuri de date.

Multe dintre cele mai importante surse de date sunt însă relativ noi. Se argumentează că explozia volumului de date caracteristic fenomenului Big Data provine din datele de natură nestructurată. În cadrul acestora, spre deosebire de datele generate de către utilizatori, care au la origine informaţii furnizate voluntar în diferite medii de diseminare Web, există şi datele interceptate. Acestea din urmă se referă la informaţii colectate în mod pasiv din comportamentul online al indivizilor, cum sunt, de pildă, termenii de căutare online sau localizarea indivizilor prin aplicaţiile prezente pe dispozitivele mobile.

În comparaţie cu instrumentele analitice tradiţionale, conceptul Big Data introduce modificări în cadrul a patru dimensiuni: (1) tipuri de date, (2) viteza de acumulare a acestora, (3) volumul lor şi (4) calitatea datelor. Odată cu lansarea mediului Web 2.0, o mare parte din datele de valoare pentru întreprinderi sunt generate în exteriorul organizaţiei, de către consumatori şi în general, utilizatori Web.

Ne aflăm astăzi la un punct de inflexiune în care volumul şi varietatea datelor generate în organizaţii creează provocări, dar şi oportunitatea de a atinge noi posibilităţi de afaceri şi valoare adăugată. Cei care vor avea capacitatea de a construi infrastructurile corespunzătoare pentru managementul informaţiei vor putea transforma aceste provocări în avantaj competitiv şi vor putea propulsa afacerea lor spre rezultate mult mai bune. Adăugarea de valoare prin utilizarea potenţialului Big Data este încă în faza emergentă, însă reprezintă o schimbare de paradigmă pe care orice afacere trebuie să o ia în considerare.

Aproape orice companie descoperă că trebuie nu doar să gestioneze volume de date din ce în ce mai mari în sistemele lor în timp real, dar şi să analizeze aceste informaţii astfel încât să poată lua rapid deciziile potrivite pentru a concura eficient pe piaţă.

Cererea crescândă pentru platforme

analitice de generație următoare care oferă clienților răspunsuri aproape în timp real, declanşate de date în timp real cum ar fi istoria accesărilor sau parcursul vizitatorilor (clickstreams), social media, senzori, combinate cu puterea de executare distribuită a seriilor de comenzi, demonstrează faptul că inteligența ar trebui să se afle, implicit, în centrul oricărei aplicații software. Astfel, aplicațiile moderne pun accent pe utilizarea noilor tehnologii Big Data.

Soluția oferită prin utilizarea seturilor Big Data ajută afacerile să îşi administreze mai bine fluxurile de date cu volum mare, varietate mare şi viteză mare şi să transforme aceste date în informații care să genereze profit.

# 2. Conceptul Big Data

## 2.1 Definiții şi caracteristici

În fiecare zi, creăm un număr mare de date. Aceste date sunt de peste tot: senzori care înregistrează vremea, post-uri pe social media, fotografii digitale şi clipuri video, tranzacții online, semnale GPS de pe telefon şi multe altele. Acestea reprezintă ”Big Data”.

*Conceptul de Big Data* este în prim-planul temelor actuale în cele mai multe cercuri de IT. Înțelegerea conceptului de Big Data, la fel ca orice altă tehnologie în curs de dezvoltare, necesită mai întâi ca acesta să fie definit.

Conform [1], există cel puțin 43 definiții ale termenului Big Data. Câteva sunt prezentate în continuare:

- în 2011, un raport al International Data Corporation a definit Big Data ca fiind “o nouă generație de tehnologii şi arhitecturi, proiectate pentru a extrage valoare economică din volume foarte mari de date de o largă varietate, prin asigurarea unei viteze ridicate de captare, descoperire şi/sau analiză” [2];
- Gartner definețte Big Data fiind "date de volum mare, de mare viteză şi de mare varietate care necesită forme inovatoare rentabile de prelucrare a informațiilor pentru a intensifica înțelegerea domeniului şi luarea deciziilor" [3].

Cercetătorii domeniului Big Data sunt în unanimitate de acord că toate sistemele de tipul Big Data au următoarele caracteristici definitorii pentru datele lor: *volumul (volume), varietatea (variety), viteza (velocity/virality), veridicitatea (veracity), validitatea (validity), variabilitatea (variability), volatilitatea (volatility), vâscozitatea (viscosity), vizualizarea (visualization), şi valoarea (value)*. Trebuie să menționăm că numai primii "patru V", dacă au valori “mari”, definesc Big Data, restul celor "şase V" se regăsesc la orice fel de date.

Întrucât cei "patru V" sunt considerați definitorii pentru acest concept, este oportună o detaliere a semnificației acestor caracteristici:

*1. Volum*: creşterea volumelor de date în sisteme de tip întreprindere este cauzată de volumul tranzacțiilor şi a altor tipuri de date tradiționale, precum şi de noi tipuri de date. Un volum prea mare de date reprezintă o problemă de stocare, dar prea multe date au în egală măsură şi un mare impact asupra complexității analizei datelor;

*2. Viteză*: se referă atât la rapiditatea cu care datele sunt produse, cât şi la rapiditatea cu care datele trebuie să fie prelucrate pentru a satisface cererea. Acest lucru implică fluxuri de date, crearea de înregistrări structurate, precum şi disponibilitatea pentru acces şi livrare. Viteza de generare, prelucrare şi analiză a datelor creşte continuu, în principal din următoarele motive: specificul de timp real al proceselor de generare, cererile care rezultă din combinarea fluxurilor de date cu procesele de afaceri, specificul proceselor de luare a deciziilor. Viteza de prelucrare a datelor trebuie să fie ridicată, în timp ce capacitatea de prelucrare depinde preponderent de tipul de prelucrare al fluxurilor de date;

*3. Varietate*: liderii IT au avut întotdeauna o problemă cu transformarea volumelor mari de informații tranzacționale în decizii, deşi tipurile de date generate sau prelucrate erau puțin diversificate, mai simple şi majoritar structurate. În prezent, există mai multe tipuri de informații pentru analiză generate de noile canale şi tehnologii apărute - în principal provenind din *social media*, *Internetul lucrurilor, surse mobile* (sensibile la context) şi publicitatea online – care generează date semistructurate sau nestructurate. Varietatea include date tabelare (baze de date), date ierarhice, documente, XML, e-mailuri, blog-uri, mesaje instant, click stream-uri, fişiere log, date de contorizare, imagini statice, audio,

video, date despre cursul acțiunilor (stoc ticker), tranzacții financiare etc.

*4. Veridicitate*: se referă la cât de încredere sau de îndoielnice sunt datele. Calitatea datelor Big Data este mai puțin controlabilă deoarece provine din diferite surse pentru care nu se poate garanta calitatea conținutului şi forma lui de prezentare. Pentru analistul de date experimentat, este esențială capacitatea de a evalua conformitatea, acuratețea şi sinceritatea datelor supuse analizei. Aici discuția se poartă în jurul responsabilității generatorului inițial al datelor, scopului pentru care datele sunt emise şi reacțiilor receptorilor.

În anticiparea oportunităților Big Data, companiile din toate mediile industriale colectează şi stochează provizoriu un număr imens de date operaționale, publice, comerciale sau sociale. În majoritatea mediilor, în special guvernamentale, producție şi educație, combinarea acestor surse cu "dark data", cum ar fi email-uri, multimedia etc., reprezintă de cele mai multe ori cea mai nouă oportunitate de a transforma afacerile.

## 2.2 Seturi Big Data

Pentru procesarea Big Data, datorită datelor de complexitate şi dimensiune foarte mare, nu pot fi utilizate aplicații standard fiind necesare aplicații capabile să ruleze în mod paralel pe un număr foarte mare de servere. Printre dificultățile întâlnite în procesarea acestor date se numără: capturarea, curățarea, stocarea, căutarea, partajarea, transferul, analiza şi vizualizarea.

Sistemele de Big Data pot furniza informație atât organizațiilor guvernamentale cât şi cetățenilor, provenind din diferite surse care pot fi identificate după cum urmează: document pe hârtie (mediu fizic), documente digitale, puncte de acces la rețeaua de Internet guvernamentală, site-uri localizate pe platformele online de socializare şi sisteme operaționale disponibile.

Informația furnizată de sistemele Big Data nu include informații personale sau informații restricționate de mecanisme de control şi confidențialitate.

Potrivit Garter [4], pentru a gestiona un volum mare de date, informațiile ar putea fi incluse în categorii, în funcție de sursă. Firma de consultanță americană a identificat cinci astfel de tipuri de informații:

1. *Date operaționale*: sunt date despre consumatori, furnizori, parteneri şi angajați deja accesibile pe baza unor procese de tranzacție sau din baze de date;

2. *Date ascunse (Dark Data)*: sunt informațiile adunate de-a lungul vremii în arhive, dar care nu pot fi clar structurate. Ele pot fi utilizate ulterior pentru luarea de decizii, analize de afaceri, etc. În acest caz ar fi incluse mail-urile, contractele, informațiile multimedia;

3. *Date comerciale*: sunt date care pot veni prin intermediul agregatoarelor de date (care citesc RSS-urile) specifice, în funcție de industrie;

4. *Date publice*: sunt datele publice care aparțin instituțiilor statului (informații care vin de la Guvern, de la ministere);

5. *Date din social media*: sunt datele care arată activitatea unui utilizator pe un blog, pe rețelele de socializare. Ele sunt utile pentru a stabili trenduri, atitudini, preferințe.

Big Data reprezintă seturi mari de informații complexe care în urma unei analize pot determina creşterea inteligenței în afaceri prin identificarea trendurilor şi îmbunătățirea operațiunilor de afaceri şi proceselor decizionale, pot contribui la prevenirea bolilor şi chiar combate rata criminalității.

Dintre domeniile în care proiectele Big Data sunt realizabile amintim: Sănătate (analiza statistică a cazurilor, telemedicină etc.), Cultură, eCommerce, Securitate națională.

# 3. Arhitecturi şi tehnologii Big Data

## 3.1 Arhitecturi pentru sistemele Big Data

Din cauza complexității sistemelor Big Data, a fost necesară dezvoltarea *unei arhitecturi specializate*. BDAF (Big Data Arhitecture Framework) are ca scop implementarea unei colecții specifice de elemente de design, asigurarea abordării unui design consistent, reducerea complexității sistemului, maximizarea reutilizării, legăturilor slabe (loose-coupling), reducerea dependențelor şi creşterea productivității.

Pentru Big Data, *cea mai frecventă arhitectură utilizată este Hadoop*. Această inovație a redefinit managementul datelor, deoarece prelucrează cantități mari de date, cu costuri reduse şi în timp util.

### 3.1.1 Ecosistemul Hadoop

Fiind o colecție de seturi de date atât de mare şi de complexă, utilizarea uneltelor manuale de gestionare a bazelor de date devine incomodă în cazul Big Data. Atunci când lucrăm cu volume mari de date avem nevoie de o soluție care să ne permită atât stocarea la un cost cât mai mic, dar şi să asigure o performanță bună la procesare. Un posibil răspuns la această provocare este platforma de aplicații *Apache Hadoop*.

***Hadoop*** [5] este un proiect open-source dezvoltat de Apache care îşi propune realizarea de procesări distribuite a unor seturi de date de dimensiuni mari, rulând pe mai multe clustere, folosind modele de programare simple. Proiectarea acestui framework a fost realizată astfel încât să fie scalabilă chiar şi în situația în care sarcinile sunt rulate pe mii de calculatoare, fiecare dintre acestea punând la dispoziție o anumită capacitate de procesare şi de stocare.

Începând cu anul 2010, Hadoop a fost adoptat pe scară largă de organizații atât în scopul de a stoca volume mari de date, cât şi ca platformă de analiză a acestora. În prezent, Hadoop este folosit de numeroase companii pentru care volumul de date generat zilnic depăşeşte capacitățile de procesare şi stocare specifice sistemelor convenționale: Adobe, AOL, Amazon.com, EBay, Facebook, Google, LinkedIn, Twitter, Yahoo.

Apache Hadoop *este un* ***ecosistem*** *de unelte gândite pentru a funcționa împreună ca o soluție eficientă de stocare şi procesare a datelor*. Aceste unelte sunt dezvoltate de către o comunitate diversificată de dezvoltatori într-un mod colaborativ sub umbrela Apache Software Foundation.

Produsele Hadoop integrate în cele mai multe dintre distribuții sunt HDFS, MapReduce, HBase, Hive, Mahout, Oozie, Pig, Sqoop, Whirr, ZooKeeper, Flume [6].

Nucleul Apache Hadoop este format din **două componente**: *un sistem de fişiere distribuit (HDFS – Hadoop Distributed File System)* şi *un framework pentru procesare distribuită (MapReduce)*. Hadoop a fost gândit să funcționeze într-o arhitectură de tip cluster construită pe echipamente server obişnuite.

- **HDFS** oferă o stocare extrem de fiabilă şi distribuită, prin replicarea datelor pe mai multe noduri. Spre deosebire de un sistem de fişiere obişnuit, atunci când datele sunt trimise la HDFS, acestea se vor împărți în mod automat în mai multe blocuri şi depozitează datele în diferite „DataNodes”. Acest lucru asigură disponibilitate ridicată şi toleranță la erori.
- **MapReduce** oferă un sistem de analiză care poate efectua calcule complexe, pe seturi de date de dimensiuni mari. O procesare de tip MapReduce presupune că problema care trebuie rezolvată poate să fie împărțită în probleme mai mici care pot să fie rezolvate independent (faza de *map*), într-o manieră “divide et impera”, fiecare fiind executată cât mai aproape de datele pe care trebuie să opereze urmând ca apoi rezultatele să fie reunite în funcție de necesități (faza de *reduce*) [7]. Această componentă este responsabilă de efectuarea calculelor şi de împărțirea unui calcul de complexitate ridicată în mai multe task-uri.
- **HBase** este o bază de date distribuită de tip NoSQL, orientată pe coloane având la baza modelul Google BigTable, care foloseşte ca şi mediu de stocare HDFS, fiind utilizată în cazul aplicațiilor Hadoop care necesită operații de citire / scriere aleatoare în seturi de date foarte mari. HBase reprezintă o soluție pentru seturi de informații de dimensiuni foarte mari (de ordinul milioanelor şi miliardelor de înregistrări) sau pentru aplicații ce utilizează date care sunt accesate de foarte mulți clienți (cererile şi răspunsurile generate ca urmare a acestei interacțiuni implică un volum de date foarte mare). Totodată, funcționează optim în cazul unor scheme variabile, unde structura înregistrărilor diferă (datorită unor atribute care pot să existe sau nu). HBase are trei componente principale: biblioteca clientului, un server de tip master, mai multe servere de regiune.
- **Zookeeper** este un serviciu de coordonare pentru aplicațiile distribuite. Zookeeper menține, configurează şi denumeşte

cantități mari de date. De asemenea, furnizează servicii distribuite de sincronizare şi de grup. Singur, Zookeeper conține noduri master şi slave şi stochează informații de configurare.

- **HCatalog** stochează metadate şi generează tabele pentru cantități mari de date. HCatalog simplifică comunicarea utilizator folosind datele HDFS şi este o sursă de partajare a datelor între instrumente şi platformele de execuție.
- **Hive** este o platformă de depozitarea datelor (de tip data warehouse) care permite interogarea şi gestionarea seturilor de date de mari dimensiuni din depozite distribuite, stocate în HDFS. Hive este o sub-platformă în ecosistemul Hadoop şi foloseşte un limbaj de interogare de tipul SQL, care este numit HiveQL. Limbajul, de asemenea, permite programatorilor tradiționali ai MapReduce să se conecteze la mediul lor specific de interogare şi de reducere atunci când este incomod sau ineficient.
- **Pig** este o platformă de nivel înalt folosită pentru analizarea unor seturi de date mari având un limbaj propriu, pentru descrierea programelor de analiză a datelor. Caracteristica principală a Pig este că prin natura programelor Pig, permite paralelizarea lor la momentul rulării. Compilatorul Pig produce joburi MapReduce. Arhitectura Pig generează un limbaj de scripting de nivel înalt (Pig Latin) şi operează pe o platformă în timp real, platformă care permite utilizatorilor să execute MapReduce pe Hadoop. Pig este mai flexibil decât Hive referitor la formatul datelor, furnizând propriul model de date. Pig are propriul tip de date, hartă, care reprezintă datele semistructurate, inclusiv JSON şi XML.
- **Mahout** este o bibliotecă pentru algoritmi de învățare automată (machine-learning) şi data mining, incluzând algoritmi de clasificare şi de clustering. Mulți algoritmi sunt scrişi pentru compatibilitate cu MapReduce, astfel încât ei sunt scalabili la seturi de date mari. Această componentă este împărțită în patru grupe principale: filtrare colectivă, clasificare, clustering şi extragere de modele paralele frecvente (mining of parallel frequent patterns).
- **Oozie** este un instrument pentru managementul workflow-ului /coordonarea joburilor MapReduce. Apache Oozie permite combinarea mai multor elemente într-o unitate logică de lucru. Apache Oozie este o aplicație Java Web, care rulează într-un servlet Java Tomcat şi foloseşte o bază de date pentru a stoca definiții ale fluxului de lucru şi execuții curente ale fluxului de lucru.
- **Avro** serializează datele, conduce apelurile de proceduri la distanță şi transferă datele de la un program sau limbaj la altul. În această arhitectură, datele se auto-descriu şi sunt întotdeauna stocate în funcție de propria lor schemă, deoarece aceste calități sunt potrivite limbajelor de scripting, cum ar fi Pig.
- **Chukwa** este un tool pentru monitorizarea aplicațiilor distribuite, bazându-se pe arhitectura HDFS şi Map Reduce. Chukwa este o arhitectură pentru colectarea şi analiza datelor. Chukwa colectează şi prelucrează datele din sistemele distribuite şi le stochează în Hadoop.
- **Flume** este un serviciu distribuit care permite colectarea, agregarea şi mutarea unor volume mari de date tip log. Are o arhitectură bazată pe fluxuri de date şi care permite construirea de aplicații analitice. Foloseşte două canale, şi anume, surse şi colectoare (sinks). Sursele includ date Avro, fişiere şi fişierele jurnal (log) de sistem, în timp ce colectoarele fac referire la HDFS şi HBase. Prin motorul său personal de prelucrare, interogare, Flume transformă fiecare nou batch de Big Data înainte de a fi transportat în colector.

Deşi Hadoop are diverse componente, fiecare companie utilizează anumite componente ale Hadoop în funcție de necesitățile ei.

### 3.1.2 Integrare Big Data cu Hadoop

Arhitectura de Big Data nu este una fixă, care să se potrivească în toate situațiile. Fiecare strat de procesare în arhitectură are mai multe soluții şi tehnici care pot fi implementate pentru a crea un mediu robust. Fiecare soluție are propriile avantaje şi dezavantaje pentru un anumit volum de muncă [8].

În general orice arhitectură de date se compune din patru componente logice principale: sursele de date, transformarea datelor, prelucrarea datelor sau integrarea datelor, cereri de date. Printre problemele legate de integrarea Big Data se numără varietatea surselor de date, calitatea datelor ce urmează a fi integrate şi vizualizarea datelor [9]. Arhitectura unui ecosistem pentru integrarea Big Data (vezi figura 1) include următoarele componente:

1. *Sursele de date structurate şi nestructurate.* Introducerea bazelor de date stocate în cloud şi a infrastructurii mobile, au dus la o creştere semnificativă a dimensiunii şi complexităţii seturilor de date, acestea devenind componenta principală a ecosistemelor de integrare a datelor. Astfel arhitectura de integrare a datelor trebuie să includă strategii multiple pentru accesarea şi stocarea unei cantităţi foarte mari şi diversificate de date.
2. *Platforma pentru descoperirea datelor.* Platforma pentru descoperirea datelor reprezintă un set de instrumente şi tehnici pentru lucrul cu fişiere pentru Big Data pentru găsirea de modele şi răspunsuri la întrebări de business. În prezent, aceasta este mai mult o activitate adhoc, iar organizaţiile întâmpină dificultăţi în dezvoltarea unor procese în jurul ei. În cadrul activităţii de descoperire a datelor, informaţiile obţinute pot deveni uneori neutilizabile după doar câteva ore. Arhitectura pentru integrarea datelor trebuie să ţină cont de aceste informaţii volatile pentru asigurarea calităţii datelor. Infrastructura pentru integrarea datelor trebuie să fie capabilă să răspundă rapid la cerinţele utilizatorilor.
3. *Depozite de date tradiţionale.* Depozitele de date tradiţionale oferă necesarul de

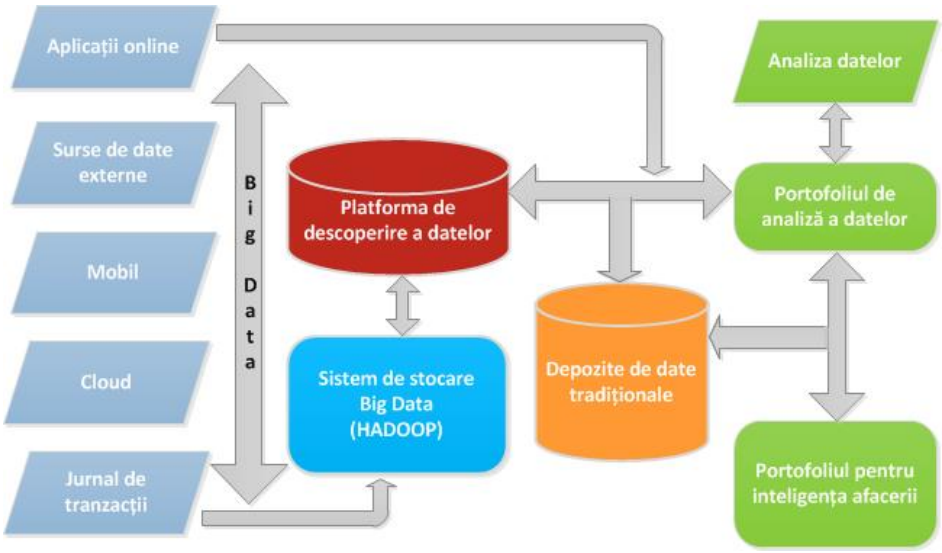


**Figura 1. Arhitectura unui ecosistem pentru integrarea Big Data [9]**

*Sisteme de stocare Big Data.* În timp ce sistemele de stocare a datelor foarte mari precum Hadoop asigură mijloace de stocare şi organizare a unor volume mari de date, procesarea acestora pentru extragerea de informaţii utile rămâne în continuare o activitate dificilă. Arhitectura MapReduce a acestor sisteme dă posibilitatea de stocare rapidă a unor cantităţi foarte mari de date şi oferă suport pentru realizarea de analize pe baza acestor date. Platforma pentru integrarea datelor trebuie să construiască structura pentru stocarea datelor şi să realizeze conexiunile cu celelalte surse de date.

informaţii de bază, dar trebuie să includă noi funcţionalităţi pentru o mai bună integrare a surselor de date nestructurate şi pentru a satisface nivelul de performanţe solicitat de platformele de analiză. Organizaţiile au început să dezvolte noi modalităţi de separare a analizelor operaţionale de analizele în pro-funzime pe baza istoricului pentru deciziile strategice. Platforma pentru integrarea datelor trebuie să fie capabilă să separe informaţiile operaţionale, de sursele de date utilizate în elaborarea strategiilor pe termen lung. Totodată infrastructura de

integrare a datelor trebuie să permită un acces rapid la datele cel mai des accesate.

4. *Portofoliul pentru inteligența afacerii.* Portofoliul pentru inteligența afacerii se concentrează pe rezultatele şi performanțele din trecut, chiar dacă va exista o creştere a cererii pentru rapoarte şi performanțe operaționale. Evoluția necesității de autoservire a inteligenței afaceri şi inteligenței afacerii pe dispozitive mobile va continua să genereze probleme arhitecturale platformelor de integrare a datelor. Un alt aspect foarte important îl reprezintă capacitatea portofoliului de inteligență a afacerii de integrare cu portofoliul de analiză. Aceasta poate conduce la o creştere a cererilor pentru integrarea informațiilor.
5. *Portofoliul de analiză a datelor.* Activitatea de analiză din cadrul acestui portofoliu trebuie să gestioneze atât problemele legate de activitatea companiei, cât şi cele legate de date. Platformele de integrare a datelor joacă două roluri în ceea ce priveşte asigurarea suportului necesar portofoliului de analiză. În primul rând, ecosistemul de integrare a datelor trebuie să asigure accesul la date structurate şi nestructurate pentru activitatea de analiză. În al doilea rând, trebuie să permită reutilizarea analizelor efectuate anterior, reducând astfel situațiile care ar necesita repetarea unor paşi.

Ecosistemul de integrare a datelor trebuie deci să includă posibilitatea de procesare a unor volume foarte mari de date şi să facă față unor solicitări de a lucra cu o varietate mare de surse de date.

Una dintre principalele provocări, în ceea ce priveşte prelucrarea de seturi foarte mari de date, este manipularea fluxurilor de date în timp real. În timp ce ambele tipuri de date, offline şi online, pot fi în mod independent prelucrate, adesea este nevoie să furnizăm răspunsuri la întrebările cu privire la evenimente online bazate pe trecut. ***Arhitectura Lambda*** vine ca un răspuns la aceste provocări [10]. Această arhitectură integrată cu Hadoop este prezentată în figura 2.

Arhitectura, aşa cum este prezentată în figura 2, este alcătuită din următoarele componente:

- *Stratul de loturi* (Batch Layer) - responsabil pentru gestionarea setului de date master şi de precalcularea vizualizărilor batch;
- *Stratul de servire* (Serving Layer) - indexează vizualizările batch pentru interogări ad-hoc;
- *Stratul de viteză* (Speed Layer) - serveşte doar datelor noi, care nu au fost încă procesate de Nivelul batch.

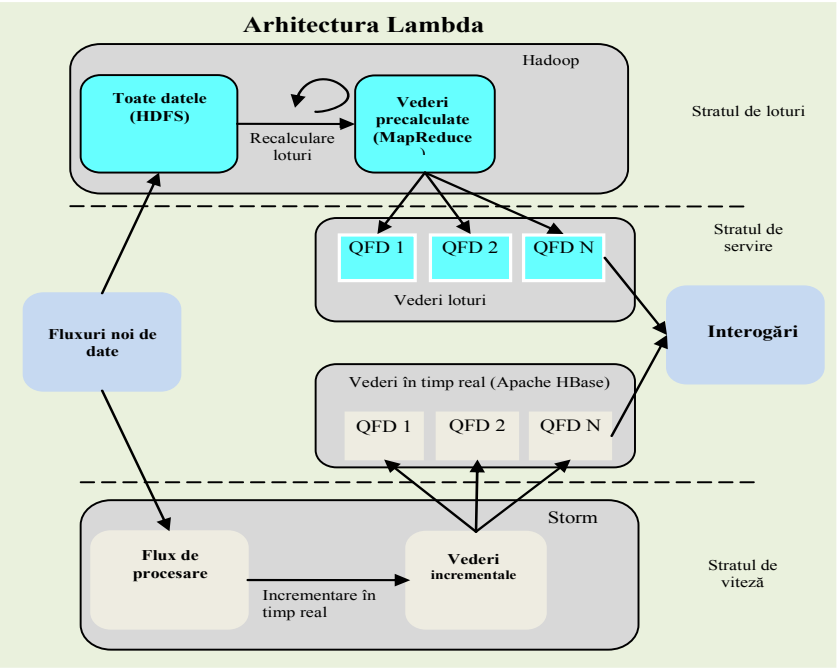


**Figura 2. Arhitectura Lambda integrată cu Hadoop [10]**

## 3.2 Tehnologii pentru Big Data

*Tehnologiile Big Data* reprezintă un domeniu aflat în continuă dezvoltare, ce se ocupă cu analiza şi gestionarea volumelor mari de date. Această definiţie cuprinde atât echipamentele hardware, cât şi sistemele software care integrează, organizează, gestionează, analizează şi prezintă Big Data.

Unele tipuri de date precum text şi voce, există de mult timp, însă volumul acestora în mediul Internet şi în alte structuri digitale anunţă începutul unei noi ere, precum *şi a unor noi tehnologii* care permit analizarea acestor tipuri de date.

Tehnologiile Big Data pot fi clasificate în şase categorii [11]):

**1. Tehnologii suport pentru infrastructură** constau în:

- *platformele Cloud pentru Big Data* – au resurse eficiente de programare şi gestionare,
- *tehnologii de stocare* - legate de compresia datelor şi de virtualizare de stocare,
- *tehnologii de virtualizare* – se referă la procesul de partajare a resurselor şi de izolarea hardware-ului de la bază,
- *tehnologii de reţea* - măresc considerabil frecvenţa şi viteza de transmisie a datelor,
- *tehnologii de monitorizare a resurselor* - gestionează resursele conectate la reţea în scopul identificării erorilor apărute în sistem;

**2. Tehnologii pentru achiziţionarea datelor -** obţinerea datelor neprelucrate (brute) de la senzori sau alte surse dedicate prin intermediul:

- *tehnologiei de achiziţionare a datelor bazată pe senzori* - care permite ca informaţiile oferite de senzori să fie transferate către o bază de date cu ajutorul reţelelor wireless,
- *tehnologiei* de achiziţionare a datelor bazată pe reţele de date;

**3. Tehnologii pentru transferul datelor** - pentru colectarea informaţiilor înainte de procesarea datelor;

**4. Tehnologii pentru memorarea şi arhivarea datelor** - se aplică în mod distribuit în noduri de stocare multiple, pentru care se pun la dispoziţie mecanisme de back-up, securitate, interfeţe de acces şi protocoale şi includ următoarele:

- *sisteme de fişiere distribuite*: pentru prelucrarea datelor trebuie adoptată, de asemenea, o arhitectură şi soluţii distribuite – *HDFS*, *HBase, Cassandra, MongoDB* open source,
- *baze de date relaţionale*: tradiţionale, caracterizate prin lipsă de scalabilitate şi extensibilitate, nu sunt adecvate. Interogările pe bazele de date SQL (*MySQL* şi *Oracle*) de pe disc sunt lente,
- *tehnologii NOSQL*: noi tehnologii pentru baze de date ce nerelaţionale, care nu oferă garanţiile ACID (Atomicitate, Consistenţă, Izolare) şi sunt utilizate în prelucrarea datelor nestructurate şi analiza Big Data;

**5. Tehnologii pentru procesarea datelor** - se referă la aspectele de procesare a datelor şi utilizarea tehnicilor de bază ale tehnologiilor Big Data, pentru analizarea, prelucrarea şi exploatarea datelor, extragerea de informaţii şi cunoştinţe importante şi apoi transformarea în modele utile şi aplicarea acestora la procesele de cercetare şi operare;

**6. Tehnologii pentru afişarea datelor şi interacţiune -** urnizarea de vizualizări interactive, care permit utilizatorilor să navigheze prin seturile de date. Permite utilizatorilor să ia decizii care să sprijine producţia, operarea şi planificarea.

### 3.2.1 Tehnologia NoSQL

Bazele de date relaţionale tradiţionale nu pot face însă faţă provocărilor actuale aduse de către Big Data. În ultima vreme bazele de date de tipul NoSQL sunt din ce în ce mai populare pentru stocarea datelor de mari dimensiuni. Au apărut din necesitatea unor companii precum Google, Facebook sau Twitter de a manipula cantităţi imense de date cărora bazele de date tradiţionale pur şi simplu nu le pot face faţă. Aşa că bazele de date NoSQL au fost proiectate pentru a stoca volume foarte mari de date în general fără o schemă fixă şi partiţionate pe multiple servere. Bazele de date NoSQL oferă moduri flexibile de lucru, suport pentru copierea datelor mult mai simplu şi mai uşor, un API simplu, şi coerenţa eventuală a datelor. Bazele de date NoSQL devin astfel

tehnologia de bază pentru Big Data.

NoSQL (Not Only SQL) sunt baze de date non relaționale [12]. Principalul avantaj al utilizării bazelor de date NoSQL este acela că permit lucrul eficient cu date structurate, precum e-mailul, multimedia, procesoare de text. Bazele de date NoSQL, ca nouă generație de baze de date: nu sunt relaționale, sunt distribuite, sunt Open Source şi se caracterizează prin scalabilitate orizontală. O altă caracteristică importantă a sistemelor NoSQL este arhitectura "shared nothing" prin care fiecare nod-server este independent, nu partajează memorie sau spațiu.

Bazele de date NoSQL au o structură mai simplă şi o tehnologie diferită pentru stocarea şi extragerea datelor decât bazele de date relaționale şi oferă performanțe mai bune pentru analize în timp real sau pe volume mari de date. Într-o bază de date NoSQL nu există o schemă propriu-zisă a datelor, ele fiind stocate ca perechi cheie-valoare (foarte eficient şi flexibil, dar datele nu sunt self-describing), sau de coloane (folosit pentru date împrăştiate), sau document (folosit pentru depozite XML, dar ineficient ca performanță), sau graf (folosit pentru traversări relaționate, dar ineficient la căutări) [13].

Astfel mişcarea NoSQL reprezintă o încercare de a depăşi limitările modelului relațional şi un pas de trecere către NewSQL şi anume relațional plus extra funcționalități NoSQL.

Cele mai populare baze de date NoSQL în acest moment sunt: Cassandra, Mongodb, CouchDB, Redis, Riak, Membase, Neo4j şi HBase.

## 4. Concluzii

Big Data poate adăuga valoare şi oferi o nouă perspectivă prin îmbunătățirea practicilor de analiză şi modelare predictivă. Volumele masive de date provenind din surse diferite au un efect pozitiv în procesul de luare a deciziilor în timp real. Varietatea surselor de date, calitatea datelor care urmează să fie integrate şi vizualizarea lor sunt unele dintre provocările pentru integrarea Big Data.

Noi capacități şi tehnologii trebuie să fie adoptate în scopul de a transforma informațiile prin gestionarea şi de analiza datelor. Principalele provocări sunt acceptarea şi utilizarea noilor tehnologii, precum şi reglementarea lor. Cele mai notabile probleme de depăşit rezidă în dificultatea de a analiza volume mari de date pentru a obține rezultate precise în timp util, necesitatea de standardizare, interoperabilitatea, securitatea, confidențialitatea, precum şi expertiza şi finanțarea pentru dezvoltarea infrastructurii Big Data şi integrarea seturilor de date deja disponibile.

Noile metode, instrumente şi abordări statistice tehnologie, cum ar fi cloud computing şi tehnologii de securitate trebuie să fie explorate. Mai mult decât atât, ar trebui să se investească în instruirea personalului cu privire la utilizarea BD.

Big Data constituie o oportunitate de a utiliza noi tipuri de date în scopul de a crea întreprinderi mai agile, care să rezolve probleme care anterior au fost considerate nesoluționabile, conducând la rezultate mai bune în afaceri. Aceasta duce la schimbări radicale în funcționarea întreprinderilor, care se schimbă de la utilizarea unui model bazat în principal pe experiența decidentul, la un model bazat pe informații, care dă o valoare reală a afacerii şi organizației în sine.

## BIBLIOGRAFIE


1. **DUTCHER, J.:** What Is BD?. Berkeley School of Informatics, Sept. 2014.
2. **GANTZ, J.; REINSEL, D.:** Extracting value from chaos. IDC iView, 2011, pp 1-12.
3. **GARTNER:** "IT glossary: big data" [webpage on the Internet], Stamford, CT; 2012, Available from: http://www.gartner.com/it-glossary/big-data.
4. **GARTNER:** 10 Big Data Software Requirements, http://www.information-management.com/gallery/Big-Data-Required-Software-Applications-10026664-1.html, accesat august 2015.
5. **HADOOP, A.:** Hadoop, 2009, http://hadoop.apache.org/.
6. **HARVEY, C.:** Hadoop and Big Data: 60 Top Open Source Tools, http://www.datamation.com/applications/hadoop-and-big-data-60-top-open-source-tools-1.html, iunie 2015.

7. **KHAN, N.; YAQOOB, I.; HASHEM, I. A. T. et al.:** Big Data: Survey, Technologies, Opportunities, and Challenges, The Scientific World Journal, vol. 2014, Article ID 712826, 18 pagini, 2014. doi:10.1155/2014/712826.
8. **ANUGANTI, V.:** Typical "Big" Data Architecture. 2012. Retrieved from: http://venublog.com/2012/11/30/typical-big-data-architecture/.
9. **BODAPATI, V.:** Data Integration Ecosystem for Big Data and Analytics. 2013. Retrieved from: http://smartdatacollective.com/raju-bodapati/103326/data-integration-ecosystem-big-data-and-analytics.
10. **MARZ, N.; WARREN, J.:** Big data - Principles and best practices of scalable realtime data systems (Chapter 1), 2014.
11. **TIAN, W.; ZHAO, Y.:** Optimized Cloud Resource Management and Scheduling - Theories and Practices, Morgan Kaufmann, Elsevier Inc, 269 p., 2015, ISBN: 978-0-12-801476-9;
12. **MC CREARY, D.; KELLY, D. A.:** Making Sense of NoSQL: A guide for managers and the rest of us, Manning, 2014, ISBN-13: 978-1617291074, ISBN-10: 1617291072.
13. **SADALAGE, P. J.; FOWLER, M.:** NoSQL Distilled: A Brief Guide to the Emerging, World of Polyglot Persistence, resource.mitfiles.com/CSE/.../NoSQL%20Distilled.pd