Soluţii de prelucrare specifice Big Data

Dragoş Cătălin BARBU

Academia de Studii Economice București, Școala Doctorală de Informatică Economică

bcatalin_ro@yahoo.com

Rezumat: Lucrarea de faţă îşi propune să descrie contextul ştiinţific actual pentru domeniul Big Data Processing. Subiectele abordate au fost structurate în două mari direcţii: domeniul tehnologiei informaţiei - Big Data Processing, Big Data Analytics, Machine Learning şi IoT, pe de o parte, şi implementarea principiilor teoretice în zona de business, respectiv reglementări şi politici. În contextul Big Data, tehnicile şi platformele tradiţionale de date sunt mai puţin eficiente, existând o reacţie lentă şi o lipsă de scalabilitate, performanţă şi precizie. În ultimul deceniu provocările generate de Big Data au fost evidenţiate atât de oamenii de ştiinţă, programatori şi matematicieni cât şi specialişti în domeniu, experţi, consultanţi chiar politicieni, rezultatele lor fiind publicate atât în lucrări de cercetare, articole, promovând dezvoltarea de soluţii şi tehnologii avansate, dar şi creând un cadru legislativ necesar implementării de strategii prin regulamente sau acte normative ale unor instituţii abilitate sau organizaţii, ca de exemplu Parlamentul European.

Cuvinte cheie: Big Data analytics, IoT, Data mining, Hadoop, Data Storage, Cloud computing, Big Data processing.

Big Data Processing Solutions

Abstract: This paper aims to describe the current scientific context for Big Data Processing. The topics discussed were structured in two major directions: information technology - Big Data Processing, Big Data Analytics, Machine Learning and IoT, and the implementation of the theoretical principles in the business area, respectively regulations and policies. In the context of Big Data, traditional data processing techniques and platforms are less efficient, with a slow response and lack of scalability, performance and precision. Over the last decade, the challenges that Big Data has generated have been highlighted by both scientists, programmers and mathematicians as well as specialists in the field, experts, consultants and even politicians, their results being published both in research papers, articles, promoting the development of solutions and advanced technologies, but also creating a legal framework necessary for the implementation of strategies through regulations or normative acts of competent institutions or organizations, such as the European Parliament.

Keywords: Big Data analytics, Internet of Thing, Data mining, Hadoop, Cloud computing, Data storage, Big Data processing.

1. Introducere

Sfârșitul ultimului deceniu ne găsește într-un moment în care evoluția tehnologică, internetul, rețelele interconectate au devenit parte integrantă din viața noastră. Impactul digitalizării și importanța spațiului cibernetic a crescut exponențial în ultimii ani. Un procent foarte mare al populației globului, companiile și țările de pe întreg mapamondul s-au conectat în acest spațiu și depind din ce în ce mai mult de sistemele de tehnologia informației și a comunicațiilor din ce în ce mai complexe. Fluxul de informații nu are frontiere, iar volumul de date și informații a crescut foarte rapid.

Mediul de afaceri, instituțiile statului, transporturile, siguranța publică, sănătatea, comunicațiile, sistemul financiar bancar, serviciile de urgență, utilitățile, apărarea națională, componente ale societății moderne, depind de funcționarea corectă a tehnologiilor informaționale, de operarea structurilor critice de informații, de disponibilitatea, integritatea și confidențialitatea informațiilor.

Obiectivul principal al lucrării îl reprezintă procesarea și analiza datelor în contextul de Internet of Things și Big Data. Big Data a devenit vârful de lance în ceea ce privește inovarea și competitivitatea, furnizând provocări dar și oportunități în cadrul peisajului global IT. Colectarea, prelucrarea și analiza de volume mari de date poate ajuta semnificativ procesul decizional în afaceri și poate crea avantaje competitive pentru cei ce aleg soluțiile adecvate de Big Data. Tehnologiile folosite în Big Data au cunoscut o dezvoltare accelerată datorită creșterii cererii aplicațiilor care generează și procesează volume mari de date. De asemenea Cloud Computing și IoT sunt principalele motoare pentru dezvoltarea de soluții pentru organizațiile economice care sprijină Business Intelligence.

În contextul Big Data, tehnicile și platformele tradiționale de date sunt mai puțin eficiente, existând o reacție lentă și o lipsă de scalabilitate, performanță și precizie.

În ultimul deceniu provocările generate de Big Data au fost evidenţiate atât de oameni de ştiinţă, programatori şi matematicieni cât şi de specialişti în domeniu, experţi, consultanţi chiar politicieni, rezultatele lor fiind publicate atât în lucrări de cercetare, articole, promovând dezvoltarea de soluţii şi tehnologii avansate, dar şi creând un cadrul legislativ necesar implementârii de strategii prin regulamente sau acte normative ale unor instituţii abilitate sau organizaţii, ca de exemplu Parlamentul European.

În aprilie 2019, în ultima lor sesiune de dinaintea alegerilor, o majoritate uluitoare de membri ai Parlamentului European au aprobat pachetul legislativ pentru următorul program al Uniunii Europene pentru cercetare și inovare, Horizon Europe.

Strategia Uniunii Europene în domeniul computerelor de înaltă performanță a fost dezbătută în cadrul Euro HPC - Întreprinderea Comună EuroHPC, inițiativă comună între Uniunea Europeană și țările europene în valoare de 1 miliard pentru dezvoltarea unui ecosistem de supercomputing de clasă mondială în Europa. Astfel, EuroHPC va permite țărilor participante și Uniunii Europene să-și coordoneze eforturile și să împartă resursele având ca obiectiv implementarea în Europa a unei infrastructuri de clasă mondială și un ecosistem de inovare competitiv în tehnologiile, aplicațiile și abilitățile supercomputing.

În lucrarea (Oussous A., 2018) se prezintă unul din cele mai recente studii de dezvoltare pentru Big Data, cu scopul de a ajuta la selectarea și adoptarea combinației potrivite a diferitelor tehnologii Big Data în conformitate cu nevoile tehnologice ale acestora și cu cerințele aplicațiilor specifice. Aceştia oferă nu numai o viziune globală asupra principalele tehnologii Big Data, dar și comparații în funcție de diferitele nivele ale sistemului, cum ar fi nivelul de stocare al datelor, nivelul de procesare a datelor, nivelul de interogare al datelor, nivelul de acces al datelor şi cel de management, unde sunt prezentate clasificarea și caracteristicile principalelor tehnologii, avantaje, limite și utilizări.

Qiu et. all (Qiu, 2016) au publicat un studiu cu privire la cele mai recente progrese în cercetarea privind machine learning pentru prelucrarea Big Data, analizând tehnicile de învățare a mașinilor, evidențiind câteva metode de învățare în studiile recente, cum ar fi învățarea prin reprezentare, învățarea profundă, învățarea distribuită și paralelă, învățarea transferului, învățarea activă și învățarea bazată pe kernel. Autorii se concentrează pe analiza și discuțiile despre provocările și posibilele soluții machine learning pentru big data, investigând conexiunile strânse ale procesului de învățare a mașinilor cu tehnici de procesare a semnalelor pentru prelucrarea Big Data.

În urma unor investigații preliminare, conform studiului din lucrarea *Prelucrarea în timp real a Big Data pentru detectarea anomaliilor* (Habeeb, 2019), abordările existente pentru detectarea anomaliilor în rețea nu sunt suficient de eficiente pentru a putea fi detectate în timp real, motivul datorându-se, în principal, acumulării de volume masive de date prin intermediul dispozitivelor conectate. Prin această lucrare autorii dezvoltă problema detectării anomaliilor în timp real, evidenţiind ca esențială propunerea unui cadru care să se ocupe efectiv de prelucrarea mare a datelor în timp real și să detecteze anomalii în rețele. Studiul a analizat tehnologiile avansate de procesare a datelor în timp real legate de detectarea anomaliilor și caracteristicile vitale ale algoritmilor de învățare a mașinilor asociate, explicând contextele lor esențiale și a taxonomiei

proceselor de procesare a datelor în timp real, a algoritmilor de detectare anormală și de învățare a mașinilor, urmată de revizuirea tehnologiilor de prelucrare a datelor.

Vladimir Schreiner şi Marko Topolnik în *Ghidul de referinţă pentru procesarea fluxurilor de date* (Schreiner, 2018) explică aspectele cheie procesării fluxurilor de date (stream), când şi cum se folosec, transformările pentru procesarea și interogarea fluxului de date prin filtrarea, conversia, gruparea, agregarea și conectarea acestuia, ferestrele de selecţie a sub-fluxurilor finite din fluxurile de date infinite şi modul de rulare a unei aplicații de streaming într-un motor de procesare a fluxului.

Dintre lucrările ce tratează modalitatea de punere în practică a tehnicilor Big Data cât şi ultimelor reglementări fiscale amintesc Programul Fiscal 2020 (E.C., SWD(2019) 151 final, 2019) şi studiul OECD din 2017 (OECD, 2017) . În *Raportul Fiscal pe 2017 al Comisiei Europene, Direcţia Generală de Impozitare şi Uniune Vamală,* sunt prezentate o serie de activităţi şi indicatori din cadrul programului ce urmăresc îndeplinirea obiectivelor cu rol important în facilitarea funcționării corespunzătoare a sistemelor de impozitare pe piața internă prin finanțarea sistemelor informatice europene, a acțiunilor comune și a activităților comune de formare. Studiul *Instrumente tehnologice pentru combaterea evaziunii fiscale și a fraudei fiscale* (OECD, 2017) prezintă Tehnologii pentru combaterea a fraudei fiscale în contextul transformării digitale.

2. Contextul global al conceptului Big Data

În ultimul deceniu, s-a înregistrat o schimbare palpabilă în amploarea influenței Europei asupra guvernării și direcției cercetării globale. Iar ambiția acesteia nu se oprește aici: UE dorește, de asemenea, să conducă abordarea mondială la o serie de agende politice informate de știință, inclusiv schimbările climaterice, reglementarea substanțelor chimice și protecția datelor.

De la introducerea lor la începutul anilor 1980, programele-cadru europene pentru cercetare și inovare au crescut în mod constant în buget și în complexitate. Scopul lor a evoluat, de asemenea devenind o formă de sprijinire a cercetării și dezvoltării legată de o serie de sectoare industriale, promovarea coordonării și coeziunii cercetării și consolidarea capacității, mobilității și infrastructurii în statele membre UE (Reillon, 2017). Astăzi, trăsătura cea mai izbitoare a programelor este măsura în care acestea proiectează și încorporează principiile de funcționare pentru cercetare în Europa și, implicit, în lumea întreagă. Aceste principii variază de la știința deschisă și datele deschise la alinierea cercetării și dezvoltării la prioritățile societății și la obiectivele globale. Pentru a realiza acest lucru, cu un buget care se ridică la doar 10% din totalul investițiilor publice în cercetare și dezvoltare din statele membre ale UE este chiar mai remarcabil.

Cerințele de date și de putere de calcul ale oamenilor de știință și industriei europene nu corespund în prezent capacităților computaționale disponibile în Uniunea Europeană. Niciun supercomputer din UE nu se află în top 10 la nivel mondial, iar cele existente depind de tehnologia non-europeană. Acest lucru aduce un risc tot mai mare Uniunii Europene de a fi lipsită de know-how-ul strategic sau tehnologic pentru inovare și competitivitate. În plus, Europa consumă în prezent aproximativ 29% din resursele HPC la nivel mondial, însă industria UE oferă doar aproximativ 5% din aceste resurse (European Parliament, 2018).

Răspunsul UE a fost de a investi împreună într-o strategie de infrastructură supercomputing ambițioasă, ambiția Uniunii Europene fiind aceea de a deveni unul din liderii mondiali în domeniul supercomputing-ului (EuroHPC).

În ceea ce privește infrastructura, EuroHPC dorește să cumpere și să instaleze cel puțin două mașini pre-exascale până în 2020 și cel puțin alte 3 mașini de tip petascale. Aceste mașini doresc să fie interconectate cu supercomputerele naționale existente și să fie puse la dispoziție în toată Europa, utilizatorilor publici și privați pentru dezvoltarea unor aplicații științifice și industriale de vârf.

Facând o scurtă trecere în revistă a literaturii de specialitate apărută în ultimul deceniu despre Big Data, vom găsi răspunsul la una din cele mai vehiculate întrebări din domeniu „*Cum ne va schimba Big Data viețile noastre în viitor?*” (ITPro, 2019)

În lumina recentelor îngrijorări cu privire la utilizarea necorespunzătoare a datelor cu caracter personal, big data şi potenţialul său uriaş sunt considerate de unii destul de riscante. Pe de altă parte volumul din ce în ce mai mare de date disponibile despre toate aspectele vieții noastre este una dintre minunile legate de epoca digitalizării. Zilnic sunt stocaţi 2.5 quintilioni bytes de date și această cifră crește în fiecare zi. În ciuda preocupărilor reale privind utilizarea abuzivă a datelor cu caracter personal, toate aceste informații au potențialul de a revoluționa fiecare zonă a vieții noastre într-un mod benefic.

Ingredientul cheie este modul în care se interpretează și utilizează toate aceste date și acest lucru nu reprezintă decât un infim procent a ceea ce se poate obține cu toate aceste informații. Acest lucru se datorează faptului că Big Data sunt prin definiție dincolo de ceea ce software-ul tradițional de prelucrare a datelor este proiectat să facă față.

Într-un raport de cercetare al META Group (acum Gartner) (Gartner, 2001), analistul Doug Laney, a inventat modelul "3 Vs" pentru a defini în 2001 Big Data, în funcție de volum, viteză și varietate. Recent s-a ajuns la cei „10 Vs", celor 3 V adăugându-se: variabilitatea, veridicitatea, validitatea, vulnerabilitatea, volatilitatea, vizualizarea şi valoarea.

Big Data sunt adesea văzute ca parte integrantă a strategiei de date a unei companii, ce au caracteristici și proprietăți specifice care ne pot ajuta să înțelegem atât provocările cât și avantajele inițiativelor legate de conceptul Big Data. George Firican, de la Universitatea British Columbia, descrie pe scurt cele 10 principii, caracteristici ale Big Data (Firican, 2017):

- **Volumul.** Probabil cea mai cunoscută caracteristică a Big Data, având în vedere că peste 90% din toate datele de astăzi au fost create în ultimii ani. Valoarea actuală a datelor poate fi destul de uluitoare şi în acest sens avem câteva statistici:
  - 300 de ore de videoclipuri sunt încărcate pe YouTube în fiecare minut,
  - aproximativ 1,2 trilioane de fotografii preluate în 2017. Întrucât aceeași fotografie are în mod obișnuit stocuri multiple pe diferite dispozitive, servicii de partajare a fotografiilor sau documentelor, precum și servicii media sociale, numărul total de fotografii stocate în 2017 este de 4,7 trilioane,
  - traficul mobil global în 2016 a fost de 6,2 exabyte pe lună (6,2 miliarde de gigabytes).
- **Viteza**. Se referă la viteza cu care se generează, se produce, se creează sau se actualizează datele. De exemplu depozitul de date Facebook stochează peste 300 de petabytes de date, dar trebuie luată în considerare viteza la care se creează date noi. Facebook solicită 600 terabiți de date pe zi. Google singur procesează în medie mai mult de 40.000 de interogări de căutare în fiecare secundă, ceea ce înseamnă aproximativ peste 4,5 miliarde de căutări pe zi.
- **Varietate.** Avem diverse tipuri de date: structurate, semistructurate și nestructurate, cele din urmă fiind cele mai des întâlnite (fișiere audio, imagini, fișiere video, actualizări media sociale cât și alte formate de text cum ar fi fișiere de jurnal, date de interacțiune, date despre mașini și senzori).
- **Variabilitatea.** Se referă la:
  - numărul de inconsecvențe din date ce trebuie găsite prin metode de detectare a anomaliilor și a unor dereglări extraordinare pentru a se produce o analiză relevantă,
  - Big data sunt, de asemenea, variabile datorită multitudinii de dimensiuni de date care rezultă din mai multe tipuri și surse diferite de date,
  - viteza inconsistentă la care Big Data sunt încărcate în baza de date.
- **Veridicitate sau veracitatea.** Este una dintre caracteristicile nefericite ale Big Data. Pe măsură ce toate proprietățile de mai sus cresc, veracitatea (încrederea în date) scade. Aceasta este similară, dar nu aceeași, cu valabilitate sau volatilitate. Veracitatea se referă mai mult la proveniența sau fiabilitatea sursei de date, la contextul acesteia și la cât de semnificativă este pentru analiza bazată pe ea.

- **Validitate.** Similar veridicităţii, validitatea se referă la cât de precise și corecte sunt datele pentru utilizarea dorită. Potrivit Forbes, un procent estimat de 60% din timpul cercetătorilor de date este cheltuit pentru curățarea datelor înainte de a putea face orice analiză. Beneficiile analizei Big Data sunt la fel de bune ca și datele de bază, şi trebuie adoptate practici bune de guvernanță a datelor pentru a asigura o consistenţă a calităţii datelor, definiții comune și metadate.
- **Vulnerabilitatea.** Big Data aduc noi probleme de securitate, o încălcare a securităţii lor reprezintând o încălcare majoră. Un astfel de caz îl reprezintă cel din iulie 2015 când un grup denumit "Echipa de Impact" a furat datele de utilizator ale Ashley Madison, un site comercial căutat pentru relaţii extraconjugale. Grupul a copiat informații personale despre baza de date a utilizatorilor site-ului și a amenințat că va face publice numele utilizatorilor și informațiile personale de identificare dacă Ashley Madison nu se închide imediat.
- **Volatilitatea.** Cât de vechi trebuie să fie datele înainte de a fi considerate irelevante, istorice sau inutile? Cât timp trebuie păstrate datele? Înainte să apară conceptul de Big Data, organizațiile au avut tendința să stocheze datele pe o perioadă nedeterminată - câteva terabyte de date ar putea să nu creeze cheltuieli aşa mari de depozitare; ar putea fi chiar păstrate în baza de date live, fără a provoca probleme de performanță. Într-un context clasic de date, nu ar putea exista nici măcar politici de arhivare a datelor în vigoare. Datorită vitezei și volumului Big Data, totuși, volatilitatea sa trebuie să fie luată în considerare cu atenție. Trebuie stabilite reguli privind moneda de schimb și disponibilitatea datelor, precum și să fie asigurată recuperarea rapidă a informațiilor atunci când este necesar. Trebuie ştiut care sunt nevoile și procesele de business - cu Big Data, costurile și complexitatea unui proces de stocare și recuperare fiind amplificate.
- **Vizualizarea.** O altă caracteristică a Big Data este cât de dificil este să le vizualizăm. Instrumentele actuale de vizualizare a datelor se confruntă cu provocări tehnice datorită limitărilor tehnologiei în memorie și scalabilitate redusă, funcționalitate și timpul de răspuns redus. Nu ne putem baza pe graficele tradiționale atunci când încercăm să parcurgem un miliard de data points, având nevoie de modalități diferite de reprezentare a datelor, cum ar fi gruparea de date (data clustering) sau folosind tree maps, coordonate paralele, diagrame de rețea circulară sau conuri. Combinând acest lucru cu multitudinea de variabile care rezultă din varietatea și viteza mare a datelor și relațiile complexe dintre ele se poate vedea că dezvoltarea unei vizualizări semnificative nu este ușoară.
- **Valoarea.** Este cea mai importantă carateristică, fără ea celelalte caracteristici ale Big Data neavând sens dacă nu există valoare dedusă din date. Valoarea substanțială poate fi găsită în Big Data, inclusiv înțelegerea mai bună a clienților, direcționarea acestora în mod corespunzător, optimizarea proceselor și îmbunătățirea performanței mașinilor sau afacerilor. Trebuie înţeles potențialul, împreună cu caracteristicile mai dificile, înainte de a începe o strategie de amploare pentru Big Data.

Toţi acești parametrii ilustrează că nu numai cantitatea de informații care definește Big Data este importantă ci şi viteza la care se ajunge precum și numeroasele categorii diferite de date implicate. De exemplu, articole de îmbrăcăminte cum ar fi ceasurile sport colectează o mulțime de informații despre obiceiurile de exercițiu ale oamenilor și acest lucru include adesea detalii cum ar fi batăile inimii, localizarea pe tot parcursul unei rutine, cadenţa bătăilor pentru ciclism și alergare și chiar nivelul de oxigen din sânge.

Companiile se bazează pe datele pe care le colectează despre clienții lor, deci modul în care acest lucru este utilizat în mod eficient de către angajați este de o importanță capitală. Epoca modernă a companiilor "digitale", cum ar fi Google, Facebook, Uber și Airbnb, se referă mai mult la modul în care utilizează datele pe care le colectează decât pe cele pe care le comercializează sau le produc. Există o dezbatere uriașă despre relația dintre aceste tipuri de companii și utilizatorii acestora. În cazul companiilor cu date pure precum Facebook, există un schimb de valori mai complex decât cel al comerțului tradițional.

La o companie cum ar fi Facebook sau Google, utilizatorul final nu plăteşte nimic pentru serviciul pe care îl primeşte - cum ar fi partajarea de rețele sociale sau rezultatele căutării pe internet, aplicațiile prin e-mail și cloud. În schimb, ceea ce schimbă sunt informațiile personale ale acestora. Datele sunt moneda de schimb pe care utilizatorii o cheltuiesc pentru a primi serviciile furnizate. Cadrele de reglementare, cum ar fi GDPR, au apărut ca o recunoaștere a valorii părții utilizatorilor de date atunci când accesează aceste servicii. Cu toate acestea, mulți utilizatori nu își dau seama cât de multe date cu caracter personal le oferă.

Majoritatea companiilor colectează date despre clienții lor și dacă utilizatorii se simt inconfortabili în legătură cu acest lucru depind de modul în care sunt utilizate aceste informații, precum și de ceea ce primesc în schimb. Transmiterea de detalii personale în scopuri de marketing către terți nu este de obicei apreciată. Cu toate acestea, capacitatea de a utiliza un sistem precum Apple Pay pentru a comanda o livrare de produse alimentare folosind doar o amprentă pentru a verifica identitatea și a transfera fondurile necesare este mult mai convenabilă decât căutările în buzunar pentru un card de credit. Predarea detaliilor cărții de credit către Apple este necesară pentru această folosire. Atunci când acest proces transmite automat detaliile adresei la compania de livrare, este o experiență fără probleme.

Acesta este doar cel mai mic vârf al aisbergului:

Figura 1. Big Data Iceberg, adaptată după Timo Elliot's Blog

Big Data promite să facă serviciile emergente, cum ar fi partajarea mașinilor, să răspundă mai bine nevoilor utilizatorilor finali. Punând deoparte pe cei care dețin mașini, cea mai mare barieră care împiedică oamenii să treacă la schimbul de mașini de la proprietatea personală la cea de tip sharing este teama de a nu avea autovehiculul lor disponibil exact când au nevoie de el. Dar o analiză predictivă precisă a comportamentului, care aduce factori cum ar fi vremea, evenimentele curente și chiar obiceiurile personale, ar putea însemna că există întotdeauna o mașină în apropiere atunci când este necesar, deoarece analiza datelor a calculat-o. Poate fi un pic înfricoşător, dar fără îndoială convenabil. Abilitatea serviciilor precum Uber și Airbnb de a se potrivi cu nevoile provizorii arată deja potențialul unei analize bine fundamentate a datelor comportamentale. În mod similar, înțelegerea Amazon a fluxului lanțului de aprovizionare îi permite să livreze multe produse a doua zi sau chiar în aceeași zi.

În următorii ani, cantitatea de informații pe care o împărtășim și care este acumulată despre lumea din jurul nostru va crește exponențial. Practic, toate companiile pot beneficia de colectarea datelor corecte și analizarea corespunzătoare a acestora. Dispozitivele de IoT (internetul obiectelor), cum ar fi termostatele din cameră, monitorizarea consumului de energie electrică, patch-urile de monitorizare a sănătății și mașinile conectate, cu urmărire în timp real, se preconizează să se prolifereze. Acestea vor oferi volume imense de date și noi posibilități de analiză. Relația dintre sănătate și stilul de viață, de exemplu, poate fi explorată în mod continuu pentru a găsi îmbunătățiri.

Lățimea de bandă disponibilă pentru dispozitivele wireless va fi cu un ordin de mărime mai mare, iar tehnologia 5G care se află într-un stadiu incipient al implementărilor în orașele europene, fiind deja testată și în Europa (European 5G Observatory, 2019), va fi lansată în cel puțin un oraș până în 2020, conform obiectivelor europene Horizon 2020. Atunci când 5G va fi introdus, va permite ca viteza wireless să fie de până la 1000 de ori mai rapidă decât 4G și promite o latență mult mai mică. În combinație cu revoluția IoT, 5G va permite în continuare creșterea exponențială a acumulării de date, în special furnizarea în timp real de la surse intense precum supravegherea video.

Big Data au, de asemenea, puterea de a uşura munca lucrătorilor din domenii critice, precum serviciile de urgență. Poliția din Regatul Unit, de exemplu, utilizează deja "cartografia predictivă a crimelor", în care sunt procesate cantități uriașe de date privind tipurile de crime, locații și timpuri pentru a genera hărți hotspot care să arate ofițerilor în cazul în care criminalitatea este mai mare. De asemenea, NHS are o bogată colecție de date despre pacienţi pe care să le prelucreze. Acest lucru poate ajuta medicii, de la recunoașterea semnelor de avertizare a diabetului la gestionarea eficientă a fluxului de pacienți.

Posibilitățile pentru IoT sunt nesfârșite, dar va trece ceva timp până când vom înțelege cu adevărat efectul asupra vieții noastre și a economiei. Pentru ca aceasta să se facă cu adevărat, este nevoie însă ca restul industriei și tehnologiei să o ajungă din urmă. Instrumentele de management, sistemele de operare IoT personalizate și standardele de comunicare trebuie dezvoltate în mod corespunzător, în momentul de față, ne aflăm încă în stadiul de a afla cum să folosim cel mai bine tehnologia.

3. Principalele Tehnologii Big Data

Trăim într-o lume condusă de date, iar aceste date cresc exponențial, atât de mult încât schimbarea rapidă a vieții noastre și a organizațiilor din întreaga lume trebuie să se adapteze și să se alinieze la această cantitate vastă de informații.

De la tehnologiile inovatoare de stocare la implementarea IoT și noua legislație GDPR a UE, Big Data sunt la conducerea schimbării economiei. Big Data reprezintă o provocare chiar și pentru cea mai mare organizație, care nu-şi mai permite să ignore potențialul imens pe care îl are pentru a îmbunătăți deciziile de afaceri, pentru a ajunge la clienți cu o precizie mai mare și pentru a raționaliza procesele din cadrul organizaţiei.Pentru a valorifica întregul potențial, companiile au nevoie de instrumentele potrivite pentru a procesa, analiza și stoca informațiile vitale pe care le produc și le colectează zilnic pentru rezultate în timp real.

Cele patru elemente principale ale oricărui proiect de Big Data sunt stocarea datelor (big data storage), de extragere a datelor (data mining), de analiză și de vizualizare. Fiecare element are un număr de instrumente inovatoare și de înaltă tehnologie oferite întreprinderilor.

Dintre cele mai importante și cei mai de succes furnizori de instrumente pentru proiectele Big Data enumerăm următoarele:

A. Stocarea datelor

Pentru proiectele Big Data, instrumentele de stocare a datelor în cloud sunt vitale pentru a maximiza cantitatea de informații care se poate stoca. Opțiunile de stocare în cloud permit stocarea de date într-o manieră sigură și accesibilă, pentru a fi ușor de utilizat.

a) Hbase/Hadoop

Hadoop este o platformă open-source, special concepută pentru a stoca seturi de date foarte mari folosind clustere. Suportă atât date structurate, cât și nestructate, astfel încât este excelent pentru organizațiile care au nevoie de o capacitate suplimentară fără prea multă atenție. De asemenea, se poate ocupa de un număr mare de sarcini fără nici o latență. Aceasta este o opțiune excelentă pentru organizațiile care au resursa dezvoltatorului de a implementa Java, dar necesită un efort pentru a intra în funcțiune.

b) MongoDB

MongoDB este foarte util pentru organizațiile care utilizează o combinație de date semistructurate și nestructurate. Acestea ar putea fi, de exemplu, organizații care dezvoltă aplicații mobile, cele care au nevoie să stocheze date referitoare la cataloage de produse sau date utilizate pentru personalizarea în timp real.

B. Data mining

După ce au fost stocate datele, va trebui să se adauge niște instrumente pentru a găsi informațiile pe care dorim să le analizăm sau să le vizualizăm. Cele trei instrumente de top ne vor ajuta să extragem datele de care avem nevoie fără a avea nevoie de trasarea manuală a acestora (o sarcină imposibilă pentru oameni, oricum, dacă deținem mii de înregistrări).

a) IBM SPSS Modeler

SPSS Modeler al IBM poate fi folosit pentru a construi modele predictive folosind interfața vizuală, mai degrabă decât prin programare. Acesta acoperă analiza textului, analiza entității, gestionarea deciziei și optimizarea și permite extragerea atât a datelor structurate, cât și a datelor nestructurate într-un întreg set de date.

b) KNIME

KNIME este o soluție scalabilă open source cu mai mult de 1.000 de module pentru a ajuta oamenii de știință să obțină informații noi, să facă previziuni și să descopere puncte-cheie din date. Fișierele text, bazele de date, documentele, imaginile, rețelele și chiar datele bazate pe Hadoop pot fi citite, făcându-le o soluție perfectă dacă tipurile de date sunt amestecate. Acesta oferă o gamă imensă de algoritmi și contribuții comunitare pentru a oferi o suită completă de instrumente de extragere a datelor și de analiză.

c) RapidMiner

RapidMiner este un instrument de data mining (extragere de date) open source care permite clienților să utilizeze șabloane, mai degrabă decât să scrie coduri de programare. Acest lucru îl face o opțiune atractivă pentru organizațiile care nu au o resursă specifică sau dacă se caută doar un instrument pentru a începe minig data. O versiune gratuită este, de asemenea, disponibilă, deși este limitată la un procesor logic și 10 000 de rânduri de date. Instrumentul oferă, de asemenea, machine learning, data mining, analize predictive și analize de afaceri pentru a ajuta întregul proces.

C. Analiza datelor

Deținem date, dar chiar avem nevoie de toate aceste date? Pentru acest moment trebuie găsite cele mai puternice instrumente care să ne ajute să analizăm în scopul de a obține informații esențiale despre afacere, clienți sau despre lumea întreagă. Instrumentele preferate de analiză a datelor sunt:

a) Apache Spark

Apache Spark este probabil unul dintre cele mai bine cunoscute instrumente de analiză Big Data, construit cu date importante în prim-planul publicului. Este o sursă deschisă, rapidă, eficientă și funcționează cu toate limbajele majore de Big Data, inclusiv Java, Scala, Python, R și SQL.

Este, de asemenea, unul dintre cele mai folosite instrumente de analiză a datelor și este utilizat de companiile de toate dimensiunile, de la cele mai mici până la organizațiile din sectorul public și giganții de tehnologie precum Apple, Facebook, IBM și Microsoft.

Apache Spark face o analiză cu un pas mai departe, permițând dezvoltatorilor să utilizeze pe scară largă SQL, prelucrarea lotului, procesarea fluxurilor și machine learning într-un singur loc, alături de procesarea graficelor.

Apache Spark este, de asemenea, super-flexibil, şi rulează pe Hadoop (pentru care a fost inițial dezvoltat), Apache Mesos, Kubernetes, ca platformă autonomă sau în cloud, făcându-l potrivit pentru întreprinderi de toate dimensiunile și în toate sectoarele.

b) Presto

Ca și Apache Spark, Presto este un instrument open source, care utilizează interogări SQL distribuite, conceput pentru a rula interogări a datelor ca un puternic motor de analiză interactiv. Acesta suportă atât surse non-relaționale, cum ar fi Hadoop Distributed File System (HDFS), Amazon S3, Cassandra, MongoDB și HBase, precum și surse de date relaționale, cum ar fi MySQL, PostgreSQL, Amazon Redshift, Microsoft SQL Server și Teradata, făcându-l un instrument util pentru întreprinderile care exploatează ambele tipuri de baze de date. De asemenea, este folosit de corporații uriașe precum Facebook contribuind în mod semnificativ la dezvoltarea sa alături de Netflix, Airbnb și Groupon, devenind astfel unul dintre cele mai puternice instrumente de analiză a datelor.

c) SAP HANA

Analiza datelor este doar un aspect al platformei SAP HANA, dar este o caracteristică extrem de bună. SAP HANA se integrează cu Hadoop, R si SAS pentru a ajuta companiile sa ia decizii rapide bazându-se pe date incalculabile.

d) Tableau

Tableau combină instrumentele de analiză și vizualizare a datelor și poate fi utilizat pe un desktop, printr-un server sau online. Versiunea online are un accent deosebit pe colaborare, ceea ce înseamnă că se pot împărtăși cu ușurință descoperirile cu oricine din organizație. Viziunile interactive ne permit să înțelegem cu ușurință toate informațiile și cu opțiunea complet găzduită de Tableau Cloud, nu vom avea nevoie de nicio resursă pentru a configura serverele, pentru a gestiona actualizările de software sau pentru a scala capacitatea hardware.

e) Splunk Hunk

Conceput pentru a funcționa pe baza sistemului Apaches Hadoop, Splunk's Hunk este un instrument de analiză a datelor complet echipat, care poate genera grafice și reprezentări vizuale ale datelor care sunt alimentate, toate gestionabile printr-un tablou de bord. Interogările pot fi făcute pe datele brute prin interfața Hunk, în timp ce grafice, diagrame și tablouri de bord pot fi create rapid și partajate prin interfața Hunk. De asemenea, funcționează și pe alte baze de date și magazine, inclusiv Amazon EMR, Cloudera CDH și Platforma de date Hotronworks, printre altele.

D. Vizualizarea datelor

Nu toată lumea este adepta de a lua informații esențiale dintr-o listă de puncte de date sau de a înțelege ce înseamnă ele. Cea mai bună modalitate de a prezenta datele este transformarea acestora în vizualizări de date, astfel încât toată lumea să poată înțelege ce înseamnă. Instrumentele de vizualizare de top sunt:

a) Plotly

Plotly sprijină crearea de diagrame, prezentări și tablouri de bord din datele analizate folosind JavaScript, Python, R, Matlab, Jupyter sau Excel. O bibliotecă uriașă de vizualizare și un instrument de creare a diagramelor online o face extrem de simplu să creeze o grafică excelentă, folosind un GUI de import și analiză extrem de eficient.

b) DataHero

DataHero este un instrument de vizualizare simplu de utilizat, care poate extrage datele dintr-o varietate de servicii cloud și le poate introduce în diagrame și tablouri de bord care fac mai ușor înțelegerea intregii afaceri. Deoarece nu este necesară codificarea, este adecvată pentru utilizarea de către organizațiile fără cercetători de date în reședința lor.

c) QlikView

Cu o suită de capabilități oferite, QlikView permite utilizatorilor săi să creeze vizualizări de date din toate sursele de date cu instrumente de auto-service care elimină necesitatea existenței unor modele complexe de date. Vizualizarea directă este servită de QlikView care rulează pe

platforma proprie de analiză a companiei, care poate fi împărtășită cu ceilalți, astfel încât deciziile luate în urma tendințelor, datele dezvăluite putând fi colaborative. Capabilitățile mai avansate permit ca analizele vizuale ale QilkView să fie încorporate în aplicații, în timp ce tablourile de bord pot ghida oamenii prin producerea rapoartelor de analiză fără a fi nevoie să aibă o înțelegere a științei datelor.

Conform Cuadrantului Magic pentru platformele de Business Inteligence și Analytics oferit de către Gartner pentru anul 2019, liderii acestor platforme sunt Tableau, Microsoft și Qlik.

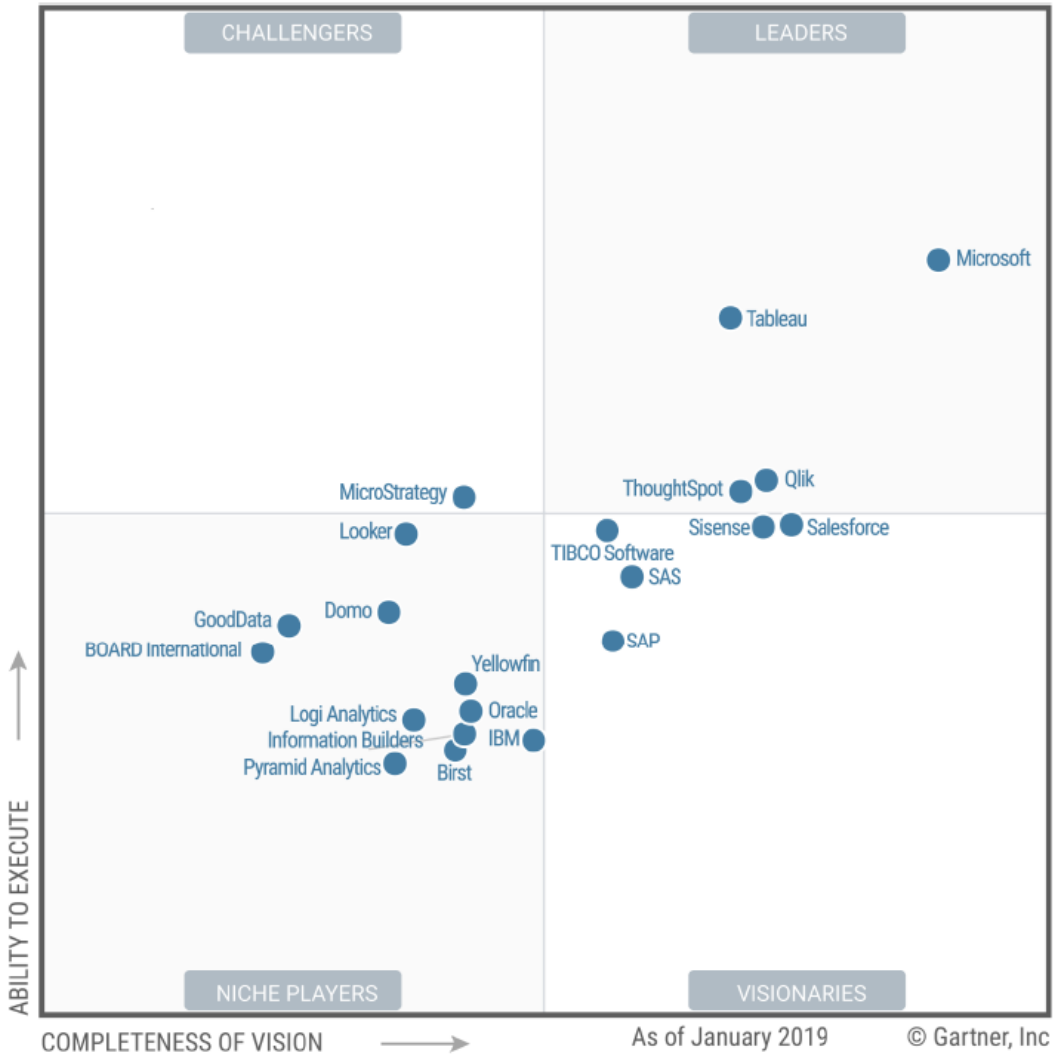


Figura 2. Platforme de Business Inteligence și Analytics – 2018 (Sursă Gartner Inc.)

4. Analiza datelor

A. Explicarea metodelor descriptive, predictive și prescriptive de examinare a datelor

In implementarea strategiilor de business la nivel de companie sau afacere, folosirea puterii de analiză a datelor este esențială. Având în vedere cantitatea impresionantă de date zilnice rezultată din operațiunile de afaceri, atât de la clienți cât şi de la furnizori, trebuie exploatat la maxim acest avantaj (ITPro, 2019).

Există o serie variată de instrumente de analiză a datelor, precum și diferite metode de realizare a acestora privind satisfacerea nevoile specifice. De asemenea, este important de reţinut faptul că concurenții vor dori şi ei să utilizeze acest volum mare de date în avantajul lor. Odată cu cantitatea de date disponibile astăzi, există un impuls şi mai mare de a o analiza cu o viziune orientată spre îmbunătățirea modului în care se desfășoară activităţile din cadrul companiei.

Termenul de Big Data este utilizat pentru a descrie cantitatea vastă de date colectate din surse cum ar fi furnizorii, clienţii sau dispozitivele IoT, ce au fost colectate în mod convențional și stocate în cloud sau pe servere fizice, în formă digitală.

Big Data analytics ajută la sortarea și analizarea datelor structurate şi nestructurate.

- în cazul **datelor structurate** informațiile sunt stocate în "linii" pentru a le face ușor accesibile și căutate,
- spre deosebire de **datele nestructurate**, unde informațiile sunt mult mai puțin organizate.

Avantajul analizei datelor îl reprezintă faptul că eliberează mai mult timp și energie pentru alte sarcini cum ar fi cele creative şi semnificative care folosesc interpretarea modelelor de date în procesul decizional strategic.

Chiar dacă datele pot fi puse într-un format SQL structurat și analizate într-un depozit de date, tot mai multe organizații caută tehnologii precum bazele de date Spark, NoSQL, Hadoop și MapReduce. Aceste sisteme suportă procesarea seturilor de Big Data și variate în cadrul sistemelor grupate. Acest lucru este extrem de util pentru datele nestructurate, cum ar fi datele extrase din dispozitivele IoT și datele media sociale.

Analizele de Big Data utilizează algoritmi statistici și matematici pentru a găsi sens prin expunerea tendințelor, preferințelor consumatorilor, modelelor îngropate și corelațiilor nespecificate. Organizațiile pot utiliza aceste constatări pentru a lua decizii de afaceri mai bune și mai bine informate.

B. Tipuri de Analize Big Data

Nevoile de business trebuie să fie conștiente de cele trei tipuri de analize care pot fi implementate cu Big Data.

- Primul este **descriptiv**. De exemplu, notificări, alerte și tablouri de bord. Acestea ne spun ce s-a întâmplat anterior, dar nu trebuie lăsate motive de ce s-a întâmplat sau ce se poate schimba.
- Următorul este **predictiv**, care este o formă mai utilă de analiză. Aceasta folosește datele anterioare pentru a modela ceea ce s-ar putea întâmpla în viitor. De exemplu, modul în care vânzările ar putea fi afectate de condițiile de marketing sau de modul în care un client ar putea răspunde la o campanie de marketing.
- În sfârșit, există analize **prescriptive**. Aceasta utilizează tehnici precum testare A / B sau testare de optimizare pentru a sfătui managerii și angajații cu privire la cele mai bune modalități de a-și îndeplini rolurile în cadrul unei organizații. De exemplu, ar putea ajuta un ofițer de poliție să prezică activitate criminală, să informeze un agent de vânzări cu privire la tipurile de reduceri care le oferă clienților sau să spună unui dezvoltator de web ce anunț va funcționa cel mai bine pe o pagină web.

5. Concluzii și tendințe în analiza Big Data

Instrumentele de analiză a datelor, fie într-un date lake care stochează date în format nativ, fie într-un data warehouse, sunt încă în curs de dezvoltare. Vor exista o serie de tendințe care vor determina modul în care vor funcționa în viitor datele mari și analizele asociate (ITPro, 2019).

Prima dintre aceste analize este **analiza în cloud.** Hbase/Hadoop poate procesa acum seturi de Big Data în cloud, chiar dacă inițial a fost proiectat pentru a face acest lucru pe clusterele de mașini fizice, deși popularitatea acestuia a scăzut. Printre companiile care oferă servicii bazate pe Hbase/Hadoop în cloud se numără platforma cloud Bluemix de la IBM, Redshift de la Amazon găzduită de BI data warehouse, serviciul de analiză a datelor BigQuery al Google și serviciul de prelucrare a datelor Kinesis.

Utilizarea de **analize predictive** de asemenea crește. Pe măsură ce tehnologiile devin mai puternice, pot fi analizate seturi de date mai mari și, la rândul lor, vor crește predictibilitatea.

Deep learning reprezintă un set de tehnici machine learning care utilizează rețele neuronale pentru a găsi modele interesante în cantități masive de date binare și nestructurate și a deduce relații fără a avea nevoie de programe sau modele explicite. Un algoritm de învățare profundă a fost folosit pentru a analiza datele Wikipedia pentru a afla că California și Texas sunt state din SUA.

Combinația dintre Big Data și analiză este o parte importantă a menținerii organizațiilor cu un pas înaintea competiției. Dar aceste companii trebuie, de asemenea, să creeze condițiile necesare pentru a permite cercetătorilor de date și analiștilor să testeze teorii pe baza datelor pe care le au.

Soluţii existente: Implementări în domeniul fiscal bazate pe Big Data

Ungaria: În contextul unor probleme extinse legate de suprimarea vânzărilor electronice și de pierderile fiscale substanțiale, guvernul maghiar a implementat legislația privind sistemul de înregistrare online a numerarului în 2014 - sectoarele comerțului cu amănuntul și al ospitalității și în 2016 de extindere în sectorul serviciilor. Impactul estimat pentru primul an a fost o creștere cu 15% a veniturilor din TVA (cota de TVA 27%) și o creștere netă a veniturilor de €210.000.000. (RetailInnovation, 2018).

Suedia: Trezoreria suedeză de stat a avut un deficit de venituri fiscale $2.000.000.000 în fiecare an din cauza utilizării industriale pe scară largă a tehnicilor de suprimare a vânzărilor, care a facilitat frauda fiscală și evaziunea. Această provocare a determinat guvernul să lanseze un act de înregistrare în numerar în 2009-2010. Legislația impune oricui să vândă bunuri și servicii în schimbul plății în numerar sau a cardului pentru a utiliza un registru de numerar conectat la o unitate de control certificată, denumită în mod obișnuit "cutie neagră“, pentru sectorul comerțului cu amănuntul, ospitalitate și servicii cu câteva excepții, de exemplu, microîntreprinderi.

Belgia: Guvernul belgian a introdus în 2014 o legislație similară cu cea suedeză, "sistemul de evidență a numerarului certificat: un instrument de combatere a fraudei în domeniul TVA".

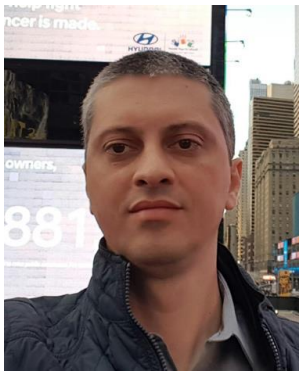
Regulamentele mandatează firmele de ospitalitate -hoteluri și restaurante- pentru a utiliza un registru de numerar integrat cu o unitate de control certificat de falsificată-așa-numitele modulul de date fiscale (FDM). Aproximativ 30 000 de registre de numerar sunt acum echipate cu un FDM.

Austria: În fiecare an, evaziunea fiscală prin manipularea registrelor de numerar costă cetățenii austrieci și Trezoreria austriacă €1.000.000.000. Noua ordonanță privind registrul de numerar "privind detaliile tehnice pentru mecanismele de securitate în registrele de numerar și alte măsuri de protecție a datelor" (Registrierkassensicherheitsverordnung RKSV) urmărește să ajungă la rădăcina problemei. De la jumătatea 2016, utilizarea registrelor de numerar certificate este obligatorie pentru întreprinderile din sectoarele comerțului cu amănuntul, al ospitalității și al serviciilor pentru întreprinderile care ating o cifră de afaceri anuală de €15 000, cu condiția ca tranzacțiile sale în numerar (inclusiv plățile prin card bancar, cardurile de credit) să depășească €7 500 net pe an.

BIBLIOGRAFIE

1. E. C., SWD (2019) 151 final. (2019). Fiscalis 2020 Programme - *Progress Report* 2017.
2. EuroHPC, Euro High Performance Computing Joint Undertaking. https://eurohpc-ju.europa.eu/index.html.
3. European 5G Observatory. (2019). 5G Scoreboard. https://5gobservatory.eu/observatory-overview/5g-scoreboards/.
4. European Parliament (2018). European high-performance computing joint undertaking - briefing. COM(2018) 8, 11.1.2018, 2018/0003(NLE).

5. Firican, G. (2017). The 10 Vs of Big Data. https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx.
6. Gartner. (2001). http://www.gartner.com/resId=2057415.
7. Habeeb, R. N. (2019). Real-time big data processing for anomaly detection: A Survey. *International Journal of Information Management*, Volume 45, 289-307.
8. Hbase/Hadoop. Apache Hadoop. https://hbase.apache.org/
9. ITPro (2019). What is big data analytics? https://www.itpro.co.uk/business-strategy/28163/what-is-big-data-analytics.
10. ITPro (2019). How big data will change our lives. https://www.itpro.co.uk/data-insights/32868/how-big-data-will-change-our-lives.
11. Mongo, D. B., https://www.mongodb.com
12. OECD (2017). Technology Tools to Tackle Tax Evasion and Tax Fraud. Organisation for Economic Co-operation and Development.
13. Oussous, A., B. F.-Z. (2018). Big Data technologies: A survey. Journal of King Saud University. *Computer and Information Sciences*, Volume 30, Issue 4 , 431-448.
14. Qiu, F. W. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016: 67.
15. Reillon, V. (2017). EU Framework Programmes for Research and Innovation: Evolution and Key Data from FP1 To Horizon 2020 in View of FP9. *European Parliamentary Research Service.*
16. Retail Innovation (2018). The Online Cash Register System. https://www.retailinnovation.se/Hungary.
17. Schreiner, V. T. (2018). A Reference Guide to Stream Processing. White Paper. www.hazelcast.com.



Dragoş Cătălin BARBU este doctorand la Universitatea de Studii Economice din București și deține un masterat la Universitatea din București, Facultatea de Matematică și Informatică. În prezent este cercetător științific superior III și șef al departamentului Cloud Computing la Institutul Național de Cercetare-Dezvoltare în Informatică - ICI București. De asemenea, este lector asociat la Facultatea de Matematică și Informatică - Universitatea din București.

Expertiza sa profesională include limbaje de programare C / C++ și Java, baze de date relaționale, medii integrate de dezvoltare (Visual Studio, NetBeans, Eclipse). Din 2004 a fost implicat în următoarele proiecte finanțate de UE: Similar și ARiSE (FP6), UsiXML (Eureka), SPOCS (CIP-ICT PSP), TOOP (H2020), IdealIST2018. Este de asemenea membru al Consiliului Științific al ICI București din 2017 și președintele Comitetului Tehnic de Specialiști (TCS) - parte a Comitetului Tehnico-Economic (TCE) - entitate guvernamentală a Ministerului Comunicațiilor și Societății Informaționale în procesul de elaborare, monitorizare și implementare a politicii guvernamentale. Este autor a 20 de lucrări de jurnal și 14 lucrări de conferință, un capitol de carte și o carte în domeniul securității cibernetice.

Dragoş Cătălin BARBU is a PhD candidate at the Bucharest University of Economic Studies and he holds a M.Sc. from the University of Bucharest, Faculty of Mathematics and Computer Science. He is currently working as a Senior Scientific Researcher III and Head of the Cloud Computing Department at the National Institute for Research and Development in Informatics - ICI Bucharest. He is also an associate lecturer at the Faculty of Mathematics and Computer Science - University of Bucharest.

His professional expertise includes C/C++ and Java programming, relational databases, integrated development environments (Visual Studio, NetBeans, Eclipse). Since 2004 he has been involved in the following EU funded projects: Similar and ARiSE (FP6), UsiXML (Eureka), SPOCS (CIP-ICT PSP), TOOP (H2020), IdealIST2018. Mr. Barbu is also a member of the Scientific Council of ICI Bucharest since 2017 and the President of the Technical Committee of Specialists (TCS) - part of the Technical - Economical Committee (TCE) - governmental entity that assists the Ministry for Communications and Information Society in the process of government policy elaboration, monitoring and implementation. He is author/ co-author of 20 journal papers and 14 conference papers, one book chapter and one book in the cyber-security domain.