O abordare metodologică privind structurarea ofertei CDI TIC pentru mediul de afaceri. Studiu de caz

Gabriel NEAGU, Mădălina ZAMFIR, Vladimir FLORIAN,

Alexandru STANCIU, Mihnea Horia VREJOIU

Institutul Naţional de Cercetare-Dezvoltare în Informatică, ICI București

gabriel.neagu@ici.ro; madalina.zamfir@ici.ro; vladimir.florian@ici.ro;

alexandru.stanciu@ici.ro; mihnea.vrejoiu@ici.ro

Rezumat: Statutul Tehnologiei Informației și Comunicațiilor (TIC) de sector caracterizat printr-o dinamică accentuată de dezvoltare este confirmat în documentele strategice la nivel național în domeniul economic. Din punct de vedere științific, TIC este inclus în sectoarele de specializare inteligentă pentru cercetare-dezvoltare și inovare (CDI), care asigură suport pentru competitivitate și performanță economică. Pentru valorificarea acestui potențial este important ca definirea și structurarea ofertei de servicii CDI TIC pentru beneficiarii din economie să aibă în vedere evidențierea specificului și complementarității acestei oferte în raport cu produsele și serviciile industriei de profil. Lucrarea propune o abordare metodologică în trei etape pentru structurarea unei asemenea oferte: focalizarea tematică și analiza state-of-the-art a tematicii selectate; identificarea tipologiei de soluții generice (caracterizate printr-o arie largă de aplicabilitate) pentru tematica analizată și constituirea unui portofoliu de profil; stabilirea categoriilor de servicii ce pot fi configurate prin agregarea soluțiilor generice și prototipizarea acestora. Metodologia prezentată este exemplificată în contextul unui proiect dedicat furnizării de produse și servicii CDI TIC pentru sectoarele de specializare inteligentă din economie.

Cuvinte cheie: TIC, CDI, date masive, analiza avansată a datelor, soluții generice, servicii CDI TIC.

A methodological approach to structuring the ICT RDI offer for business environment. Case Study

Abstract: The status of Information and Communication Technology (ICT) as a sector characterized by a strong dynamic of development is confirmed in the strategic documents at national level in the economic field. From the scientific point of view, ICT is included in the list of research, development and innovation (RDI) sectors of smart specialization, which provide support for competitiveness and economic performance. In order to capitalize on this potential, it is important that the definition and structuring of the ICT RDI service offer to the beneficiaries of the economy to highlight the specificity and complementarity of this offer in relation to the products and services of the IT industry. The paper proposes a three-stage methodological approach for structuring such a offer: thematic focus and state-of-the-art analysis of selected topic; identifying the typology of generic solutions (characterized by a wide applicability) for the analyzed topic and creating a dedicated portfolio; establishing the categories of services that can be configured by aggregating generic solutions and prototyping them. The methodology is d exemplified in the context of a project devoted to deliver ICT RDI products and services to economical sectors of smart specialization.

Keywords: ICT, RDI, Big Data, data analytics, generic solutions, ICT RDI services.

1. Introducere

Documentele strategice actuale privind dezvoltarea economică și tehnico-stiințifică confirmă caracterul dinamic de dezvoltare a domeniului tehnologiei informației și comunicațiilor (TIC), potențialul său de a susține dezvoltarea și competitivitatea celorlalte domenii și, în consecință, prioritatea pe care o reprezintă pentru activitatea de cercetare-dezvoltare-inovare. Astfel, Strategia Naţională pentru Competitivitate 2014-2020 include TIC în grupa de sectoare de specializare inteligentă la nivel național, caracterizate printr-o dinamică accentuată de dezvoltare. Una din cele două axe prioritare ale Programului Operaţional Competitivitate 2014-2020, care implementează

strategia de profil, este TIC pentru o economie digitală competitivă. Strategia Națională privind Agenda Digitală pentru România 2020 evidențiază suportul CDI TIC pentru dezvoltarea economică și socială și definește patru domenii de acţiune: (1) eGuvernare, interoperabilitate, securitate cibernetică, cloud computing, open data, big data şi media sociale; (2) TIC în Educaţie, Sănătate, Cultură şi eInclusion; (3) eCommerce, cercetare-dezvoltare şi inovare în TIC; (4) broadband şi infrastructura de servicii digitale. În sfârșit, Strategia Naţională de Cercetare-Inovare 2014-2020 include TIC între cele patru domenii prioritare de specializare inteligentă CDI, dar evidențiază în același timp suportul pe care îl asigură cercetărilor în celelalte domenii, implicit rolul său în abordările cu caracter interdisciplinar, promovate prin această stategie.

În acest context, în conformitate cu cerințele Programului "Agenda Digitală pentru România" (2015-2017), a fost lansat proiectul "*Cercetare-Dezvoltare şi Inovare în TIC: Dezvoltarea de produse şi servicii inovative care să deservească cele 10 sectoare identificate în domeniul Smart Specialization*".

Din punct de vedere managerial, principala provocare a acestui proiect a constituit-o punerea de acord a tematicii sale generoase cu nivelul resurselor alocate. Pentru a răspunde acestei provocări, metodologia elaborată a avut ca principale obiective fundamentarea delimitării ariei tematice a proiectului și tratarea specificității ofertei CDI pentru beneficiarii potențiali din economie. Lucrarea prezintă această abordare metodologică și rezultatele implementării sale pentru una din temele relevante selectate.

În continuare, conținutul lucrării este structurat după cum urmează: capitolul 2 prezintă în sinteză metodologia proiectului, care structurează derularea proiectului în trei etape principale. Capitolele 3 și 4 sunt dedicate studiului de caz. Capitolul 3 detaliază trei soluții generice din portofoliul proiectului pentru tematicile date massive (*Big Data*) și analiza avansată a acestora (*Big Data Analytics-BDA*). Capitolul 4 prezintă un serviciu de tip transfer de cunoștințe și expertiză de specialitate pentru dezvoltarea de soluții BDA, dezvoltat pe baza soluțiilor generice prezentate în capitolul 3, precum și recomandări de utilizare a acestui serviciu. Contribuția lucrării este sintetizată în capitolul de concluzii.

2. Metodologia proiectului

Din punct de vedere metodologic, derularea proiectului a fost structurată în 3 etape, prezentate în sinteză în continuare.

2.1. Selectarea unor tematici relevante CDI TIC

Etapa a avut la bază studierea tematicilor TIC din Strategia națională CDI, din Programul European Orizont 2020, precum și din analizele privind tendințe și priorități în informatizarea mediului de afaceri, ca de exemplu (Schlack, 2015).

Pentru selectarea tematicilor abordate au fost utilizate 3 criterii principale:

- reprezentativitatea tematicii respective pentru elaborarea unei soluții inovative de informatizare;
- complementaritatea funcțională, importantă pentru realizarea de soluții care beneficiază de avantajele mai multor tematici;
- expertiza și interesul profesional la nivelul echipei de cercetare.

Tematicile selectate în final au fost următoarele:

- managementul, guvernanţa şi analiza datelor masive;
- suport decizional bazat pe soluţii de inteligenţa afacerilor;
- timp real şi conectivitate extinsă;
- aplicaţii pentru dispozitive mobile inteligente.

Raportat la *criteriul de reprezentativitate*, aceste tematici susțin tendințele actuale de evoluție a soluțiilor informatice (sisteme, aplicații, servicii), care vizează:

- "*Datele*" – activ informatic prioritar la nivelul unei întreprinderi, a cărui importanță se amplifică în contexul Big Data;
- "*Suportul decizional*" – funcționalitate cu impact asupra performanţei manageriale, amplificat de posibilitatea valorificării Big Data în beneficiul calităţii deciziilor;
- "*Timpul real*" – regim de funcționare performant, specific pentru o tipologie din ce în ce mai largă de aplicații și servicii, în contextul dezvoltării Internetului lucrurilor (IoT);
- "*Mobilitatea*" – modalitate de interfațare cu utilizatorul care devine dominantă pentru aplicațiile și serviciile informatice moderne.

Complementaritatea funcţională poate fi ilustrată prin câteva exemple: reţelele de senzori reprezintă o sursă importantă de fluxuri de date masive; reciproc, implementările de tip IoT au nevoie de soluţii eficiente de prelucrare şi analiză a fluxurilor de date colectate, specifice Big Data; valorificarea resurselor de date masive necesită metode performante de analiză; Big Data şi BDA reprezintă o contribuţie importantă pentru inteligenţa afacerilor, cu referire la valorificarea datelor nestructurate sau semistructurate; valorificarea pentru suportul decizional a fluxurilor de date culese în timp real necesită eficienţă în analiza acestora; categorii reprezentative de aplicaţii mobile se bazează pe utilizarea senzorilor şi a analizei de tip predictiv a datelor.

Din punct de vedere al *expertizei echipei de realizare*, nucleul de bază al echipei a inclus cercetători cu experiență pentru cele patru tematici selectate.

2.2. Identificarea unei tipologii de soluții generice

Această etapă a avut la bază analiza fiecărei tematici selectate în etapa 1, care s-a axat pe conceptele de bază, potenţialul de impact pentru competitivitate, soluţii de referinţă specifice tematicii respective, reprezentative ca vizibilitate pe piață și experiență de implementare. Soluțiile generice reprezintă un subset al acestor soluții de referință, caracterizate printr-o arie largă de aplicabilitate. Au fost identificate următoarele grupe tipologice:

– analiza evoluţiei domeniilor tematice şi priorităţi curente de dezvoltare (AEP);
– îndrumare, instruire şi suport tehnic (IST);
– asistenţă pentru investigarea şi adoptarea de soluţii de informatizare (AIS).

Portofoliul de soluții generice al proiectului cuprinde 10 poziții:

- 2 soluţii de tip AEP :
  - Arhitecturi, platforme şi soluţii software pentru prelucrarea fluxurilor de date
  - Integrarea IoT cu Cloud computing
- 4 soluţii de tip IST :
  - Metodologie de dezvoltare a unei soluții de tip analiză avansată a datelor pentru mediul de afaceri
  - Expert în date (*Data Scientist*) - profil profesional şi intergrarea în structura companiei
  - Metode de baza în analiza avansată a datelor
  - Lucru în Cloud - platforma ICIPRO
- 4 soluţii de tip AIS :
  - Ecosistemul Hadoop
  - Ghid de evaluare soluții *Business Intelligence* / *Business Analytics* și studii de caz
  - Platformă pilot IoT pentru captare şi memorare date
  - Evaluare soluții de referinţă pentru platforme de dezvoltare a aplicaţiilor mobile.

2.3. Structurarea ofertei de servicii CDI TIC

Etapa a demarat cu analiza specificului grupelor tipologice ale soluțiilor de portofoliu identificate în etapa 2, din punct de vedere al relevanței pentru diversele forme de colaborare între

comunitatea de cercetare și colectivele implicate în utilizarea TIC din diverse domenii aplicative. Au fost identificate următoarele categorii de servicii:

a) consiliere pentru orientarea deciziilor de informatizare;
b) transfer de cunoştinţe și expertiză de specialitate;
c) suport în adoptarea unor soluții inovative de informatizare.

Pentru exemplificarea acestor categorii au fost elaborate exemple concrete de servicii specifice ofertei CDI, susținute de soluțiile de protofoliu.

- pentru categoria (a):
  - *Îndrumar pentru dezvoltarea de aplicații de prelucrare a fluxurilor de date,* bazată pe următoarele soluții de portofoliu: *Arhitecturi, platforme şi soluţii software pentru prelucrarea fluxurilor de date (AES)* și *Ecosistemul Hadoop (AIS).*
  - *Soluții de inteligența afacerilor,* bazată pe soluția de protofoliu *Ghid de evaluare soluţii Business Intelligence / Business Analytic (AIS).*
- pentru categoria (b):
  - *Suport pentru implementarea soluțiilor de analiză avansată a datelor masive,* bazată pe soluțiile de portofoliu *Ecosistemul Hadoop (AIS), Metodologie de dezvoltare a unei soluţii de tip Data Analytics pentru mediul de afaceri* și *Metode de bază în Data Analytics (IST).*
- pentru categoria (c):
  - *Serviciul de sensing (SNaaS),* bazat pe soluția de portofoliu *Integrarea IoT cu Cloud computing (AEP).*

3. Exemple de soluții generice

3.1. Soluția generică de tip AIS - Ecosistemul Hadoop

Soluția generică vizează componentele proiectului Hadoop și principalele componente ale ecosistemului dezvoltat în jurul acestui proiect. De asemenea, include expertiză privind evaluarea comparativă a motoarelor de procesare specifice acestui ecosistem, analizează perspectiva de evoluție a utilizării Hadoop, prezintă rezultatele experimentării procedurii de instalare a Hadoop 2.7.0 şi a unor componente din ecosistemul acestuia.

3.1.1. Proiectul Hadoop

Proiectul a fost lansat ca o implementare a motorului de procesare MapReduce împreună cu un sistem distribuit de fișiere (http://hadoop.apache.org/). Există o bogată literatură referitoare la acest proiect, o sinteză relevantă fiind furnizată în lucrarea (Landset ș.a., 2015). Cele patru module ale proiectului Hadoop sunt menționate în continuare.

Sistemul distribuit de fișiere *HDFS* (https://wiki.apache.org/hadoop/HDFS) include nodurile de date (stocare și regăsire date, raportare către nodul master a referințelor la metadate și locațiile fișierelor) și nodul master (memorare referințe și direcționare trafic către nodurile de date, ca răspuns la cererile clienților). Sistemul este sacabil și tolerant la erori, păstrând min. trei copii ale blocurilor de date.

Motorul de procesare *MapReduce* (http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html) îşi structurează funcţionarea în faza "*map*" (organizarea datelor brute în perechi {cheie, valoare}) şi faza "*reduce*" (agregarea valorilor care corespund aceleiaşi chei).

Administratorul de resurse *YARN* (http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html) permite rularea pe clusterul Hadoop a unuia sau mai multor motoare de procesare.

Setul de instrumente *Common* (http://hadoop.apache.org/docs/current/) include utilitare necesare altor module Hadoop: compresia datelor, operații de I/O și detectarea de erori, interfețe și instrumente pentru autorizarea utilizatorilor de tip proxy, autentificare, confidențialitate date, administrare chei criptografice de acces.

3.1.2. Structura ecosistemului Hadoop

Este organizată pe trei niveluri:

a. Nivelul de stocare - conține soluţii diverse de administrare a datelor: HDFS, baze de date nonrelaționale, de tip cheie-valoare, de tip document sau orientate pe coloană, modele bazate pe grafuri.

Bazele de date nonrelaționale (*NoSQL - Not only SQL*) folosesc unul din cele patru tipuri de bază de modele de date:

- *cheie-valoare*: fiecărui obiect de date îi este atașată o cheie unică. Exemple de sisteme construite pe acest model sunt *Voldemort* (http://www.project-voldemort.com/voldemort/) sau *Redis* (http://redis.io/);
- *document*: un model de tip cheie-valoare, în care cheia trimite către o colecție de înregistrări cheie-valoare. Exemple: *CouchDB* (http://couchdb.apache.org/) și *MongoDB* (https://www.mongodb.org/);
- *orientat pe coloană*: datele sunt stocate în coloane, care sunt grupate în familii de coloane. Exemple: *HBase* (http://hbase.apache.org/) şi *Cassandra* (http://cassandra.apache.org/);
- *bazat pe grafuri*: pentru datele care pot fi reprezentate sub formă de grafuri. Exemple: *Titan* (http://thinkaurelius.github.io/titan/) şi *OrientDB* (http://orientdb.com/orientdb/).

b. Nivelul de procesare: pe lângă YARN, acest nivel include:

- *motoare de procesare*: care pot fi departajate după modelul de procesare (pe loturi-*batch* sau flux de date-*streaming*), latență, productivitate, toleranța la erori, uzabilitatea, consumul de resurse, scalabilitatea. *MapReduce* funcţionează în regim batch, se caracterizează prin lipsa eficienței în ceea ce privește viteza și resursele de calcul consumate. Mecanismul de toleranță la erori este bazat pe replicarea datelor, iar creșterea dimensiunii datelor afectează scalabilitatea sistemului. Nu asigură implementarea facilă a procesării iterative, care se regăsește în multe proiecte de analiză a datelor. *Spark* (https://spark.apache.org/) se bazează pe MapReduce. Suportă calculul iterativ, îmbunătățește performanțele de viteză și utilizare resurse folosind calculul în memorie (seturile de date distribuite reziliente). Pentru activitățile de analiză a datelor lucrează cu bibliotecile MLib și GraphX. Oferă Spark Streaming, a cărui funcţionare bazată pe micro-loturi poate fi considerată o simulare de procesare în timp real a datelor. *Storm* (https://storm.apache.org/) este folosit pentru procesarea datelor în timp real și a fost inițial conceput să depășească deficiențele altor procesoare referitoare la colectarea și analizarea fluxurilor din rețelele de socializare. Arhitectura Lambda reprezintă o modalitate de a executa simultan job-uri pe MapReduce și Storm și de a combina rezultatele, unificând în acest fel prelucrările datelor de timp real și celor istorice. *Flink* (https://flink.apache.org/) oferă facilităţi de procesare de tip batch și streaming, permițând astfel implementarea arhitecturii Lambda. El este scalabil, are opțiune de calcul în memorie, poate fi integrat cu HDFS și YARN sau poate rula în mod independent de ecosistemul Hadoop. Similar Spark oferă procesare pe loturi de tip iterativ, precum şi opțiunea de streaming, bazată pe evenimente, similară cu regimul de timp real de la Storm. *H_2O* (http://www.h2o.ai/) este un framework open source care oferă un motor de procesare paralelă, funcţionalitate de tip data analytics, instrumente de preprocesare și evaluare a datelor. Integrarea cu Storm permite prelucrarea de fluxuri de date în timp real.
- instrumente diverse: *Flume* (https://flume.apache.org/) - pentru colectarea, agregarea şi migrarea datelor de tip jurnal în HDFS, *Kafka* (http://kafka.apache.org/) - sistem distribuit de mesagerie de tip „*publish-subscribe*”, *Sqoop* (http://sqoop.apache.org/) - pentru transferul pachetelor de date între bazele de date relaţionale şi HDFS. Pentru interacţiune sunt folosite motoarele de interogare *Hive* (http://hive.apache.org/) şi *Drill*

(http://drill.apache.org/). *Pig* (http://pig.apache.org/) oferă un framework de execuție și limbajul de flux de date Pig Latin. *Cascading* (http://www.cascading.org/) facilitează integrarea a unui număr mare de surse de date diferite, oferă limbajul de marcare PMML (*Predictive Model Markup Language*).

c. Nivelul de administrare: include instrumente de nivel înalt pentru interacțiunea cu utilizatorul. Planificatorul fluxurilor de lucru *Oozie* (http://oozie.apache.org/) administrează activitățile pentru instrumentele de pe nivelul de procesare. *Zookeeper* (https://zookeeper.apache.org/) este un serviciu pentru coordonarea și sincronizarea sistemelor distribuite. *Hue* (http://gethue.com/) este o interfață web pentru proiectele Hadoop, care include navigatoare de fișiere pentru HDFS și HBase, un navigator pentru activități MapReduce/YARN, instrumente pentru vizualizarea datelor.

3.1.3. Utilizarea Hadoop în mediul de afaceri

În anul 2016, în condiţiile în care organizațiile deveneau tot mai conştiente de avantajul pe care Big Data îl poate aduce afacerii lor, principalele orientări privind utilizarea Hadoop se refereau la (Garcia, 2016):

- creşterea numărului de implementări Apache Spark rulând pe Hadoop;
- utilizarea fluxurilor de date şi a surselor de date în timp real pentru probleme privind detectarea fraudelor, analiza securității datelor, validarea cererilor de asigurare, IoT;
- renunțarea la platforme scumpe în favoarea Hadoop, inclusiv migrarea în Hadoop a unor procese din depozitele de date ale întreprinderii;
- reducerea investițiilor în infrastructură, prin opţiunea pentru un singur mediu software necesar accesării tuturor datelor (loturi sau flux de date) și prin opţiunea cloud.

La nivelul anului 2018, aceste orientări sunt confirmate, Hadoop devenind un instrument de date care contribuie, alături de alte abordări tehnologice (Spark, Business Intelligence, magazii de date), la implementarea unor soluții balansate de analiză a datelor pentru afaceri (Burst, 2018).

3.2. Metodologie de dezvoltare a unei soluţii de tip analiza avansată a datelor pentru mediul de afaceri

Această soluție generică propune o abordare metodologică specifică, adoptată de *EMC Education Services* (Dietrich, Helle, Yang, 2015), precum și expertiză de instruire pentru asimilarea acestei abordări. Specificul acestei abordări derivă din caracterul exploratoriu al proiectelor BDA, spre deosebire de proiectele tradiţionale de inteligenţa afacerilor sau a celor de analiză a datelor. Din acest motiv este esenţial ca un asemenea proiect să se bazeze pe un proces riguros structurat, care să asigure completitudine şi repetitivitate în desfăşurarea analizei, utilizarea corectă a timpului şi efortului pentru atingerea cât mai devreme a consensului celor implicaţi privind înţelegerea clară a problemei de business care urmează a fi rezolvată.

Metodologia se bazează pe 7 roluri cheie și 6 etape.

Principalele competenţe şi reponsabilităţi specifice celor *7 roluri cheie* sunt următoarele:

(a) *Utilizatorul din mediul de afaceri* – înţelege specificul domeniului şi, de obicei, beneficiază de pe urma rezultatelor soluției dezvoltate; poate oferi consultanţă membrilor echipei despre contextul proiectului, valoarea rezultatelor şi modul în care rezultatele proiectului pot fi operaţionalizate; de obicei acest rol este jucat de un analist de business sau un expert în domeniul specific al proiectului.

(b) *Sponsor al proiectului* – este responsabil de iniţierea proiectului; formulează necesitatea şi cerinţele pentru proiect, defineşte problema principală de business; este cel care asigură finanţarea şi estimează valoarea rezultatelor finale; de asemenea, stabileşte prioritățile pentru proiect şi precizează rezultatele dorite.

(c) *Managerul de proiect* – asigură ca bornele cheie şi obiectivele să fie atinse la timp şi la calitatea aşteptată.

(d) *Analistul BI* – aduce expertiza domeniului specific pe baza unei înţelegeri aprofundate a datelor, indicatorilor de performanţă şi metricilor cheie, a inteligenţei afacerii din perspectiva raportărilor; generează tablouri de bord şi rapoarte, are cunoştinţă despre sursele și fluxurile de date.

(e) *Administratorul de baze de date* – încarcă şi configurează mediul bazei de date pentru a sprijini cerințele echipei de lucru privind rularea modelelor de analiză avansată a datelor.

(f) *Inginerul de date* – dispune de abilităţi tehnice avansate pentru formularea interogărilor SQL şi extragerea datelor necesare managementului; oferă suport pentru asimilarea de date în mediul de analiză; în timp ce administratorul bazei de date stabileşte şi configurează bazele de date care urmează să fie utilizate, inginerul de date asigură extragerea şi transformarea substanţială a datelor pentru a facilita procesul de analiză; lucrează în strânsă colaborare cu expertul în date.

(g) *Expertul în date* – oferă expertiză în modelarea datelor şi aplicarea unor tehnici adecvate de analiză avansată la probleme de business date; asigură îndeplinirea obiectivelor globale ale activităţii de analiză avansată a datelor; proiectează şi execută metodele de analiză pentru datele disponibile proiectului.

Conținutul celor *6 etape ale metodologiei* este, în sinteză următorul:

(1) *Descoperirea*: înţelegerea domeniului de afaceri, încadrarea tipologică a problemei, identificarea cerinţelor, dezvoltarea ipotezelor iniţiale, identificarea posibilelor surse de date;

(2) *Pregătirea datelor*: pregătirea mediului de lucru pentru analiza avansată a datelor, efectuarea ETLT (extract-transform-load-transform), familiarizarea cu datele, condiţionarea datelor (curăţarea, normalizarea seturilor de date şi efectuarea transformărilor asupra datelor), studierea şi vizualizarea datelor;

(3) *Planificarea modelului*: determinarea metodelor, tehnicilor şi fluxului de lucru pentru etapa de construire a modelului; explorarea datelor, studierea relațiilor între acestea şi selectarea variabilelor cheie, selectarea modelului.

(4) *Construirea modelului*: dezvoltarea de seturi de date cu scopul testării, antrenării, şi producţiei; construirea şi executarea de modele bazate pe rezultatele etapei de planificare a modelului; evaluarea instrumentelor existente din punct de vedere al cerințelor de rulare a modelelor; identificarea cerințelor de performanță și robustețe a mediului pentru execuţia modelelor şi a fluxurilor de activități.

(5) *Comunicarea rezultatelor*: informarea acţionarilor cu privire la concluziile principale privind valoarea de business a rezultatelor, pentru a conveni de comun acord asupra utilității proiectului pe baza criteriilor elaborate în etapa 1.

(6) *Operaţionalizarea proiectului*: furnizarea de rapoarte finale, informări, cod şi documente tehnice; execuția unui proiect pilot pentru validarea modelelor într-un mediu de real de utilizare.

3.3. Metode de bază în analiza avansată a datelor

Metodele de tip „inteligenţă în afaceri” furnizează rapoarte, tablouri de bord, interogări pe probleme de afaceri, vizualizări pentru perioada curentă sau trecută, adică răspund unor întrebări referitoare la ”*când*” şi ”*unde*” apar evenimente. Prin comparaţie, analiza avansată a datelor tinde să folosească date disparate într-un mod explorator, concentrându-se pe analizarea prezentului şi creând posibilitatea luării unor decizii competente raportate la viitor. Această abordare vizează analize mai sofisticate, furnizează o viziune profundă despre activitatea curentă şi prevede evenimente viitoare, în general concentrându-se pe probleme legate de ”*cum*” şi ”*de ce*” apar evenimentele.

Lucrarea (Hu ș.a., 2014) identifică 6 tipuri de aplicaţii intensive ca date, de tip analiză avansată a datelor, funcţie de tipul de date cu care operează: date structurate, text, multimedia, pagini web, rețele sociale, mobile. O structurare similară este propusă și în lucrarea (Dietrich, Helle, Yang, 2015). În plus, soluția generică include expertiză de îndrumare privind selectarea metodelor funcție de specificul problemei, precum și suport tehnic pentru implementarea acestora.

Principalele *clase de metode utilizate în analiza avansată a datelor* sunt următoarele:

(a) *clusterizarea* – gruparea obiectelor după caracteristici similare, cu exemple de utilizare în segmentarea clienţilor pe profiluri comportamentale, gruparea pacienţilor după anumite simptome, prelucrarea de imagini;

(b) *reguli de asociere* – identificarea unor relaţii între obiecte pe baza frecvenţei asocierii aparent întâmplătoare a acestora (produse achiziţionate împreună de clienţi diferiţi, link-uri accesate în sesiuni de lucru derulate independent), cu exemple de aplicare în comercializarea încrucişată de produse sau în structurarea conţinutului unui site;

(c) *regresia* – identificarea variabilelor de intrare cu cea mai mare influenţă statistică asupra ieşirilor, cu exemple de utilizare privind estimarea preţurilor imobilelor, prognozarea cererii pentru un anumit produs, prognozarea efectului unui tratament medical pentru diverşi pacienţi (cazul regresiei liniare) sau probabilitatea ca un solicitant de credit să-l poată plăti, probabilitatea unei reacţii pozitive la un anumit tratament (cazul regresiei logistice);

(d) *clasificarea* – atribuirea de etichete de clasă unor noi observaţii pe baza unor exemple de asociere învăţate (arbori de decizie, clasificator bayesian naiv, regresie logistică), cu utilizări multiple, de ex. în diagnostic medical sau tehnic, în marketing sau în filtrarea mesajelor spam;

(e) *analiza seriilor de timp* – identificarea şi modelarea structurii observaţiilor derulate în timp, precum şi prognozarea pe această bază a unor valori viitoare a seriilor de timp (metodologia Box-Jenkins), având ca exemple de utilizare prognozarea vânzărilor cu amănuntul, planificarea pieselor de schimb sau fundamentarea deciziilor de tranzacţionare la bursă;

(f) *analiza de text* – extragerea de informaţii utile din prelucrare şi modelarea datelor de tip text.

4. Suport pentru implementarea soluțiilor de analiză avansată a datelor masive

4.1. Justificarea serviciului

Acest serviciu de tip ”transfer de cunoştinţe şi expertiză de specialitate” valorifică soluțiile generice de protofoliu prezentate în capitolul precedent:

- AIS - Ecosistemul Hadoop
- IST - Metodologie de dezvoltare a unei soluţii de analiză avansată pentru mediul de afaceri
- IST - Metode de bază în analiza avansată a datelor.

Justificarea oportunității acestui serviciu are la bază conceptul de transformare a afacerii (*business transformation*), care are ca obiectiv îmbunătățirea competitivității pe baza valorificării oportunităților de inovare oferite de noile tehnologii (Cray, 2014). Conform Harvard Business Review (2014), analiza datelor masive în conexiune cu noile tehnologii informatice (cloud, mobile și social networking) va avea impact asupra transformării afacerii, cu efecte asupra îmbunătățirii serviciilor către clienți, creșterii productivității, dezvoltarea de noi servicii, modele de afaceri și produse. Valorificarea acestor oportunități este îngreunată de rutină, organizare deficitară, lipsa gândirii orientate spre inovare, rezistența la schimbare, cultură informatică deficitară (Whitehurst, 2015). În contrast cu aceste bariere, companiile orientate spre punerea în valoare a datelor masive se disting prin abordări moderne ca de exemplu atenția acordată profesiei de tip expert în date, migrarea activității de analiză a datelor masive de la departamentul IT la departamentele de business și operaționale, atenția acordată fluxurilor de date (Davenport, Barth, Bean, 2012). În acest context, implementarea conceptului de analiză avansată a datelor masive deschide calea de migrare către companii de tip reactiv, bazate pe date, capabile să reacționeze în timp util la stimulii contextului de afaceri (Florian și Neagu, 2016). O analiză cuprinzătoare a ofertei actuale pentru platformele de inteligență în afaceri și de analiză avansată a datelor este furnizată de Gartner în lucrarea (Sallam et al., 2017).

4.2. Stivă de instrumente pentru dezvoltarea de soluții BDA

4.2.1. Structura stivei

Termenul ”stivă” desemnează o ierarhizare a tipologiei de instrumente după etapele fluxului de execuție a unei soluții BDA. Aceste etape au fost selectate în conformitate cu metodologia adoptată în cadrul proiectului, dar și pe baza altor puncte de vedere prezentate în literatura de

specialitate, în principal lucrarea (Bahga și Madisetti, 2014). Sursa principală de configurare a stivei o reprezintă componentele ecosistemului Hadoop-Spark.

Principalele etape avute în vedere au fost:

- *colectarea datelor*: preluarea datelor din surse multiple, utilizând conectori specifici;
- *pregătirea datelor*: rezolvarea înainte de procesare a diverselor probleme ale datelor colectate, ca de exemplu: înregistrări corupte, valori lipsă sau duplicate, abrevieri nepotrivite, formatări incorecte. Implică activități specifice, ca de exemplu curățarea datelor (*data cleaning*), parsarea și transformarea datelor colectate dintr-un format în altul (*data munging*), deduplicare, normalizare (uniformizarea scalelor, unităților de măsură, abrevierilor), eșantionare și filtrare (pentru procesarea selectivă a datelor care îndeplinesc anumite reguli);
- *selectarea tipului de analiză avansată* pentru o aplicație, care la nivel generic, pot fi caracterizate ca fiind descriptive, de diagnosticare, predictive, prescriptive;
- *determinarea regimului de lucru*, în funcție de cerințele aplicației: pe loturi, în timp real sau interactiv;
- *stabilirea modelului de procesare a datelor*, în funcție de precedentele două decizii, ca de exemplu: MapReduce pentru analiză statistică și regim pe loturi, Stream processing pentru analize tip regresie și regim timp real;
- *vizualizări ale rezultatelor*: *statice* (afișarea rezultatelor analizei stocate într-o bază de date), *dinamice* (actualizarea rezultatelor în mod repetat) sau *interactive* (afișarea rezultatelor pe baza intrărilor furnizate de utilizatori).

4.2.2. Componentele stivei

Sunt exemplificate principalele componente ale stivei BDA, după destinația lor:

(a) Surse de date brute pentru o aplicație BDA: fișiere de tip jurnal, date tranzacționale, date din rețelele de socializare, baze de date, date provenite de la senzori, date de tip clickstream, date de la sistemele de supraveghere, date generate de înregistrările medicale și alte aplicații pentru îngrijirea sănătății.

(b) Conectori de acces la date pentru colectarea și ingerarea acestora. Diferă funcție de tipul sursei de date, ca de exemplu:

- conectorii de tip *publish-subscribe messaging*: sursa de date trimite datele corespunzător temelor (topics) gestionate de broker. *Apache Kafka* și *Amazon Kinesis* (https://aws.amazon.com/kinesis/) sunt framework-uri de acest tip;
- conectorii de tip *source-sink* : permit colectarea, agregarea și transformarea datelor din diferite surse într-un depozit de date centralizat (cum ar fi un sistem de fișiere distribuit). *Apache Flume*, a cărui arhitectură se bazează pe fluxuri de date, include următoarele componente: sursa, care poate să primească date, de exemplu, de la o rețea de socializare (folosind API pentru streaming); canalul, în care sunt scrise datele de către sursă; colector (sink) care filtrează datele dintr-un canal într-un sistem distribuit de fisiere; agent-o colecție de surse, canale și colectoare de date; eveniment-generat de o sursă externă de date și stocat de canalul la care aceasta este conectată;
- conectorii de tip bază de date, folosiți la importarea datelor din baze de date. *Apache Sqoop* importă datele din baze de date relaționale în tabele HDFS, Hive sau HBase;
- conectorii de tip ”*cozi de mesaje*”: operează cu mesajele de tip *push-pull*, în care producătorii împing datele în cozi, iar consumatorii le extrag. Producătorii și consumatorii acționează independent. Câteva exemple: *RabbitMQ* (https://www.rabbitmq.com/), *ZeroMQ* (http://zeromq.org/), *RestMQ* (http://restmq.com/), *Amazon SQS* (https://aws.amazon.com/sqs/);
- conectori personalizați: sunt dedicați unor cerințe specifice de colectare din rețelele de socializare, din bazele de date NoSQL, din rețele IoT (*AWS IoT* și *Azure IoT Hub*).

(c) Soluții de stocare a datelor colectate din sursele de date primare prin conectorii de acces la date. Cu ajutorul HDFS, datele stocate pot fi analizate cu diferite framework-uri BDA construite peste acest sistem. Pentru unele aplicații BDA este de preferat stocarea datelor într-o bază de date NoSQL, cum ar fi HBase. Adiacent soluțiilor de stocare sunt instrumente dedicate diverselor

operații de pregătire a datelor, ca de exemplu: curățarea și reformatarea datelor - *Open Refine* (http://openrefine.org/); deduplicarea datelor - Open Refine, Hive, Pig, Spark SQL; normalizare, eșantionare, filtrare - MapReduce, Hive, Pig, Spark SQL.

(d) Analiza datelor:

- *analiza pe loturi* - framework-uri care permit acest tip de analiză: Hadoop MapReduce, Pig, Oozie, Spark, Solr (http://lucene.apache.org/solr/), Machine Learning (http://mahout.apache.org/, https://spark.apache.org/mllib/).
- *analiza în timp real* – este exemplificată de framework-urile Apache Storm și Spark Streaming (https://spark.apache.org/streaming/).
- *interogarea interactivă* – se bazează pe utilizarea limbajelor de tipul SQL: *Spark SQL* (componentă a lui Spark pentru interogări SQL pe date structurate și semistructurate), *Hive* (oferă un shell pentru crearea tabelelor și interogarea datelor), *Amazon Redshift* (specializat în gestionarea interogărilor pe seturi de date de dimensiuni de până la 1 PB sau mai mult), *Google BigQuery* (serviciu pentru interogarea seturilor masive de date folosind interogări SQL).

(e) Vizualizarea rezultatelor: rezultatele obținute în urma procesării și analizei datelor sunt stocate în baze de date SQL sau NoSQL pentru operații ulterioare de interogare și vizualizare în aplicații web. Exemple de instrumente:

- baze de date: MySQL, Cassandra, MongoDB, *Amazon DynamoDB* (https://aws.amazon.com/dynamodb/).
- framework-uri de vizualizare: biblioteca *Pyton PyGal* (http://pygal.org/en/) pentru construirea hărților în diferite formate (SVG, PNG); *Pyton Seaborn* (https://seaborn.pydata.org/) - bibliotecă de vizualizare Python pentru reprezentarea graficelor statistice; *Lightning* (http://lightning-viz.org/) - pentru interactivitate și fluxuri de date actualizate continuu;
- framework-uri pentru dezvoltare aplicații web: *Django Python* (https://www.djangoproject.com/), *Flask* (http://flask.pocoo.org/).

4.3. Recomandări de instrumente pe tipuri de analiză

Recomandările vizează cele 6 clase de metode de analiză prezentate în cap. 3.3 și sunt structurate pe modele de analiză și regimuri de prelucrare (Bahga și Madisetti, 2014):

- Analize statistice:
  - counts, Max, Min, Mean, Top-N, Distinct
    - Batch: Hadoop-MapReduce, Pig, Spark
    - Timp real: Storm, Spark Streaming
    - Interactiv: Spark SQL
- Clusterizare
  - K-Means:
    - Batch: Hadoop-MapReduce, H2O
    - Batch & timp real: Spark MLib
  - Gaussian mixture, LDA (Latent Dirichlet allocation):
    - Batch: Spark MLlib
- Reguli de asociere:
    - Batch: Spark MLlib
- Regresie:
  - Cele mai mici pătrate, regresie izotonică:
    - Batch, timp real: Spark MLlib
  - Model liniar generalizat
    - Batch: H_2O

- Clasificare:
  - KNN, Arbori de decizie, Random Forest, SVM (Support Vector Machine):
    - Batch, timp real: Spark MLlib
  - Random Forest, Clasificatorul bayesian naiv, Deep Learning:
    - Batch: H_2O
- Analiza seriilor de timp:
  - Filtrarea Kalman, modele timp-frecvență, detectare:
    - timp real: Spark, Storm
- Analiza textului:
  - Clasificare, analiza sentimentelor, explorarea textului:
    - Batch, timp real: Spark
    - Timp real: Storm
  - Sumarizare:
    - Batch: Spark.

5. Concluzii

Lucrarea prezintă o abordare metodologică de structurare a ofertei de servicii CDI TIC pentru beneficiarii din economie. Abordarea are la bază necesitatea tratării specificului activității CDI în scopul evidențierii complementarității ofertei sale de servicii în raport cu produsele și serviciile industriei de profil. Abordarea este exemplificată prin experiența de utilizare în cadrul unui proiect dedicat contribuției CDI TIC privind dezvoltarea de produse şi servicii inovative care să deservească sectoarele de specializare inteligentă din economie. Dintre concluziile acestei implementări menționăm:

- importanța criteriilor de selecție și delimitare a tematicii abordate, pentru asigurarea nivelului corespunzător de inovare pentru oferta respectivă;
- nivelul de exigență a analizei tematicii selectate pentru asigurarea relevanței realizărilor de referință identificate (metodologii, arhitecturi, medii de dezvoltare, studii de caz);
- justificarea riguroasă a configurației portofoliului de soluții generice prin criterii referitoare la concordanța cu tipologia de soluții propusă în metodologie, complementaritatea funcțională a soluțiilor ca premisă a posibilității de configurare a unor servicii, expertiza existentă la nivelul echipei proiectului;
- acoperirea echilibrată a categoriilor de servicii specifice CDI TIC identificate în metodologie, în concordanță și cu evaluarea cerințelor pieței țintă;
- importanța prototipizării acestor servicii în forme specifice, pentru facilitarea asimilării lor de către potențialii beneficiari.

Recunoaștere. Prezenta lucrare a beneficiat de suportul proiectului CS 360 (contract 144/2015) din cadrul Planului sectorial 2015-2017 al Ministerul Comunicaţiilor şi Societății Informaţionale, Programul *Agenda Digitală pentru România.*

BIBLIOGRAFIE

1. Bahga, A., Madisetti, V. (2014). Internet of Things: A Hands-On Approach, *Published by Bahga & Madisetti*, ISBN: 978-099605515;
2. Burst, A. (05.03.2018). Seven Data and Analytics Trends for 2018, *Datameer Blog post*. Internet: https://www.datameer.com/blog/seven-data-analytics-trends-2018/;
3. Cray, P. (2014). The digital transformation of business. *Harvard Business School Publishing*, Internet: https://hbr.org/sponsored/2014/09/the-digital-transformation-of-business;

4. Davenport, T.H., Barth, P., Bean, R. (2012). How 'Big Data' is different, *MIT Sloan Management Review* 54(1), 2012. Internet: http://sloanreview.mit.edu/article/how-big-data-is-different/;

5. Dietrich, D., Helle, B., Yang, B. (2015). *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. John Wiley & Sons, ISBN: 978-1-118-87613-8;

6. Florian, V., Neagu, G. (2016). Abordări şi soluţii specific în managementul, guvernanţa şi analiza datelor de mari dimensiuni (Big Data). *Revista Română de Informatică şi Automatică*, 26(1), 5-22, ISSN: 1220-1758;

7. Garcia, J. (18.12.2015). BI and Analytics Predictions for 2016. Interent: https://www3.technologyevaluation.com/research/article/6-BI-and-Analytics-Predictions-for-2016.html;

8. Harvard Business Review (2014). The Leadership Edge in Digital Transformation, *A Report by Harvard Business Review Analytic Services*. Internet: http://www.oracle.com/us/c-central/oracle-leadership-edge-digital-2276804.pdf;

9. Hu, H., Wen, Y., Chua, T.S., Li, X. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*, 652-687;

10. Landset, S., Khoshgoftaar, T.M., Aaron, N. Richter, A.N., Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem, *Journal of Big Data*, 2-24, DOI 10.1186/s40537-015-0032-1.

11. Sallam, R., Howson, C. Idoine, C., Oestreich, T., Richardson, J. Tapadinhas, J. (2017). Magic Quadrant for Business Intelligence and Analytics Platforms, *Garner Report*, 16.02.2017, ID: G00 301340. Interent: https://www.gartner.com/doc/reprints?id=1-3RTAT4N&ct= 170124& st=sb.

12. Schlack, M. (2015). IT Priorities - Editorial Global. TechTarget. Interent: http://docs.media. bitpipe.com/io_10x/io_102267/item_465972/2015%20IT%20Priorities%20Global.pdf.

13. Whitehurst, J. (2015). Driving digital transformation: new skills for leaders, new role for the CIO, *Harvard Business School Publishing*. Internet: https://hbr.org/resources/pdfs/comm/ RedHat/RedHatReportMay2015.pdf.

Gabriel NEAGU este cercetător gradul I și director științific în ICI București. A obținut titlul de doctor în Informatică Aplicată la Universitatea *Politehnica* București, în anul 1998. Principalele domenii de interes pentru activitatea de cercetare includ: arhitecturi de sisteme distribuite, analiza avansată a datelor masive, servicii IoT-Cloud, tehnologia blockchain.

Gabriel NEAGU is senior researcher 1st degree and Scientific Director at ICI Bucharest. He received a PhD in Applied Informatics at the *Politehnica* University of Bucharest, in 1998. His main topics of interest for research activity include: distributed system architectures, data analytics, Cloud- IoT services, blockchain technology.

Mădălina ZAMFIR este cercetător științific în Departamentul ”Sisteme inteligente distribuite, intensive ca date” din ICI București. Este doctorand în domeniul suportului IoT pentru sisteme de afaceri, la Universitatea *Politehnica* București. Subiectele de interes în activitatea de cercetare acoperă infrastructurile de tip Cloud, suportul IoT pentru soluțiile de tip Big Data, data analytics, modele de afaceri pentru IoT și Big Data.

Mădălina ZAMFIR is a scientific researcher in the ”Distributed and Data Intensive Intelligent Systems” Department at ICI Bucharest. She is PhD student in IoT support for Business Systems at the *Politehnica* University of Bucharest. Topics of interest in the research activity include Cloud infrastructures, IoT support for Big Data solutions, data analytics, business models for IoT and Big Data.

Vladimir FLORIAN este inginer dezvoltare tehnologică gradul I în Departamentul ”Sisteme inteligente distribuite, intensive ca date” din ICI București. A obținut titlul de doctor în Știința Calculatoarelor la Universitatea *Politehnica* București, în anul 2006. Domeniile sale de interes includ: arhitecturi și sisteme distribuite, BigData, Big Data Analytics, IoT.

Vladimir FLORIAN is technology development engineer 1st degree in the ”Distributed and Data Intensive Intelligent Systems” Department at ICI Bucharest. He received a PhD in Computer Science at the *Politehnica* University of Bucharest, in 2006. His topics of interest include: distributed architectures and systems, Big Data, Big Data Analytics, IOT.

Alexandru STANCIU este cercetător științific gradul III la ICI București. A obținut titlul de doctor în domeniul Ingineria Sistemelor la Universitatea *Politehnica* București, în anul 2013. Este interesat și a avut contribuții în domenii precum arhitecturi cloud pentru sistemele de control distribuite, gestionarea și analiza datelor de mari dimensiuni, IoT și tehnologia blockchain.

Alexandru STANCIU is scientific researcher 3rd degree in ICI Bucharest. He received a PhD in System Engineering at the *Politehnica* University of Bucharest in 2013. He is interested in, and has contributed to such domains as: cloud architectures for distributed control systems, Big Data administration and analysis, IoT, blockchain technology.

Mihnea Horia VREJOIU este cercetător ştiinţific gradul III în Departamentul "Sisteme inteligente distribuite, intensive ca date" din ICI București. Domeniile şi subiectele sale de expertiză şi interes cuprind: Vedere Artificială (prelucrare şi analiză de imagini, recunoaşterea formelor, recunoaşterea optică de caractere – OCR, recunoaşterea numerelor de înmatriculare – LPR) şi Învăţare Automată (clasificatoare, memorii asociative).

Mihnea Horia VREJOIU is scientific researcher 3rd degree in the "Distributed and Data Intensive Intelligent Systems" Department at ICI Bucharest. His main areas and topics of expertise and interest cover: Artificial Vision (Image Processing and Analysis, Pattern Recognition, Optical Character Recognition – OCR, License Plate Recognition – LPR) and Machine Learning (classifiers, associative memories).