

UNIVERSITATEA POLITEHNICA DIN BUCUREȘTI
FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE

Tehnologii informatice pentru analiza proceselor de afaceri

COORDONATOR

Prof. Dr. Ing. Valentin Sgârciu

ABSOLVENT

Ing. Bogdan Nedelcu

2018

Cuprins

Introducere	5
I. Stadiul actual în managementul și ingineria proceselor de afaceri	6
I.1. Procesul de afaceri	7
I.1.1. Definirea și clasificarea proceselor de afaceri	7
I.1.2. Managementul proceselor de afaceri	10
I.1.3. Metodologii de management al proceselor de afaceri	11
I.2. Limbaje pentru modelarea proceselor de afaceri	15
I.2.1. Limbajul BPMN	18
I.2.2. Diagrame de activitate UML	19
I.2.3. Analiza comparativă a limbajelor BPMN și UML	19
II. Tehnici informatice de analiză a proceselor de afaceri	26
II.1. Tehnologia volumelor mari de date	27
II.1.1. Noțiuni fundamentale	27
II.1.2. Etapele utilizării volumelor mari de date	32
II.1.3. Provocările utilizării tehnologiei volumelor mari de date	38
II.1.4. Depozite de date	45
II.1.5. Baze de date NoSQL	48
II.1.6. Analiza comparativă a bazelor de date NoSQL	57
II.2. Tehnologia extragerii cunoștințelor din date	62
II.2.1. Modele și algoritmi de extragere a cunoștințelor din date	64
II.2.2. Tehnici de extragere a cunoștințelor din date	66
II.2.3. Metodologii de extragere a cunoștințelor din date	80
II.2.4. Analiza comparativă a metodologiilor de extragere a cunoștințelor din date	83
II.3. Tehnologii informatice pentru analiza datelor	86

II.3.1. Cubul de date și OLAP.....	87
II.3.2. Tehnici utilizate pentru analiza datelor	90
II.3.3. Analiza comparativă a instrumentelor de analiza datelor	93
III. Soluții informatice pentru modelarea proceselor de afaceri	98
III.1. Tehnici și algoritmi de modelare a proceselor de afaceri	99
III.1.1. Integrarea inteligenței afacerii în managementul proceselor de afaceri	99
III.1.2. Integrarea tehnologiei volumelor mari de date și a depozitelor de date	102
III.1.3. Aplicarea tehnicilor de extragere a cunoștințelor din date.....	106
III.1.4. Algoritmi de învățare profundă asistată (<i>deep machine learning</i>).....	108
III.2. Soluții informatice pentru modelarea resurselor umane	111
III.2.1. Volume mari de date în domeniul resurselor umane	111
III.2.2. Procesarea datelor în domeniul resurselor umane	116
III.2.3. Predicții în domeniul resurselor umane	119
IV. Studii de caz privind recrutarea avansată a forței de muncă	125
IV.1. Studii de caz în domeniul resurselor umane	126
IV.1.1. Utilizarea tehnologiei volumelor mari de date în vederea recrutării avansate a forței de muncă	126
IV.1.2. Utilizarea extragerii cunoștințelor din date în vederea recrutării avansate a forței de muncă.....	132
IV.2. Proiectarea unui prototip de recrutare avansată a forței de muncă	145
IV.2.1. Nivelul actual.....	145
IV.2.2. Analiza prototipului informatic	149
IV.2.3. Proiectarea prototipului informatic.....	151
IV.2.4. Implementarea prototipului informatic.....	158
IV.2.4.1. Tehnologii informatice utilizate	158
IV.2.4.2. Realizarea prototipului informatic	164

IV.2.4.3. Prezentarea prototipului informatic	171
IV.2.5. Potențialul impact al prototipului informatic.....	182
Concluzii	184
Concluzii generale	184
Contribuții originale.....	186
Concluzii finale	190
Diseminarea rezultatelor	191
Bibliografie	193

Introducere

Odată cu dezvoltarea tehnologică, toate domeniile au dorit să profite de avansul tehnologic al IT-ului pentru a își re tehnologiza procesele, în scopul de a obține un avans tehnologic în fața competiției.

În cadrul lucrării se identifică soluții tehnice de re tehnologizare a fluxului unui proces de afaceri și se evidențiază influența implementării soluțiilor prezentate prin construirea unui prototip de recrutare avansată a forței de muncă.

Primul capitol al lucrării, intitulat ”Stadiul actual în managementul și ingineria proceselor de afaceri” prezintă concept generale în domeniul proceselor de afaceri și modalitățile de modelare a acestora. A fost realizată o analiză comparativă a limbajelor de modelare a proceselor de afaceri cu scopul de a identifica limbajul adecvat în vederea utilizării viitoare în cadrul lucrării.

Al doilea capitol, ”Tehnici de analiză a proceselor de afaceri”, conține analize comparative precum analiza sistemelor relaționale de baze de date față de NoSQL, a metodologiilor de extragere a cunoștințelor din date sau a sistemelor de analiză a datelor.

Următorul capitol, ”Soluții informatice pentru modelarea proceselor de afaceri” pornește de la analize realizate în cadrul capitolului precedent și cuantifică influența acestora pentru un anumit tip de date.

În capitolul al patrulea, ”Studiu de caz privind recrutarea avansată a forței de muncă” este exemplifică în mod practice utilizarea algoritmilor prezentați anterior și folosesc rezultatele experimentale obținute pentru a evidenția influența noilor tehnologii informatice asupra unui proces de afaceri specific domeniului resurselor umane.

Ultima parte a lucrării conține concluziile finale ale cercetării, diseminarea rezultatelor și bibliografia utilizată.

I. Stadiul actual în managementul și ingineria proceselor de afaceri

Cunoașterea stadiului actual în managementul și ingineria proceselor de afaceri, presupune, în primul rând, cunoașterea noțiunilor fundamentale despre proceselor de afaceri și modelarea acestora.

Ne aflăm într-o perioadă în care fiecare detaliu contează enorm în lupta companiilor pentru supremație, din acest motiv, cunoașterea în detaliu a unui proces de afaceri este o etapă esențială în definirea și înțelegerea procesului respectiv.

Au fost necesari mulți ani, și multe cercetări, pentru a ajunge de la abordarea clasică a managementului proceselor de afaceri, când companiile au făcut primii pași spre îmbunătățirea performanțelor cu impact asupra calității, la dezvoltarea unor metodologii de management al proceselor de afaceri.

În ultimii ani, companiile au afișat un interes sporit pentru îmbunătățirea proceselor de afaceri, cu scopul de a răspunde competiției necruțătoare a economie globale. Un prim pas în această direcție l-a reprezentat folosirea unui limbaj adecvat de modelare al procesor de afaceri.

Necesitatea unei evaluări a limbajelor de modelare a proceselor de afaceri existente ar fi foarte utilă pentru a facilita luarea deciziei corecte. Evaluarea realizată în cadrul lucrării se concentrează pe notațiile grafice utilizate cel mai des pentru reprezentarea proceselor de afaceri: *Business Process Modeling Notation* și *UML Activity Diagram*.

I.1. Procesul de afaceri

I.1.1. Definirea și clasificarea proceselor de afaceri

Un proces de afaceri este o colecție de activități sau de sarcini similare care au un punct de plecare și un punct de final, precum și intrări și ieșiri clar definite. Accentul este pus pe modul în care activitatea se desfășoară în cadrul unei organizații. Un proces de afaceri poate fi descompus în mai multe sub-procese, cu caracteristici specifice care contribuie împreună la obiectivele procesului de bază.

Această afirmație este susținută și prin definițiile oferite de alți autori în lucrările lor. De exemplu, Schedlbauer spune în cartea sa, că procesul de afaceri reprezintă un activ intelectual strategic, care trebuie înțeles și gestionat în mod proactiv [SCH10]¹.

Conform lui J. Pyke, J. O'Connel și R. Whitehead (Mastering Your Organization's Processes: A Plan Guide to BPM) procesele sunt colecții de activități care creează valoare pentru întreprindere sau pentru oamenii care intră în contact cu organizația, cum ar fi acționarii sau clienții. Valoare este creată prin transformarea materiilor prime, producerea de bunuri, oferirea de servicii, cunoștințe sau efort uman. În general, procesele de afaceri pot fi considerate un lanț de activități structurate care transformă intrările în ieșiri comerciale utile. Procesele de afaceri nu sunt un concept nou, dar sunt implicit prezente încă de la începutul afacerii și a comerțului [PYKE]².

În lucrarea "What is BPM", Wurtzel descrie procesul de afaceri ca fiind un set de activități realizate coordonat având un anumit obiectiv de afaceri. Un proces este o traducere de planuri și politici într-un model standard de decizii și acțiuni, planificate corespunzător și având o anumită succesiune [WUR13].³

¹ [SCH10] Schedlbauer, M. (2010) The Art of Business Process Modeling: The Business Analyst's Guide to Process Modeling with UML & BPMN, CreateSpace

² [PYKE] J. Pyke, J. O'Connel și R. Whitehead - Mastering Your Organization's Processes: A Plan Guide to BPM, 2006

³ [WUR13] Wurtzel M.M. - What is BPM, McGraw-Hill Engineering, 2013

Abordarea clasică a managementului proceselor de afaceri a început în anii '80 când companiile au avut mai multe inițiative de a-și îmbunătăți performanțele cu impact asupra calității. Acest lucru a dus la concluzia că toate activitățile de lucru din cadrul unei companii sunt procese de afaceri. Ulterior, calitatea obiectivelor a fost introdusă pentru a îmbunătăți eficiența proceselor de afaceri. În particular, ciclul unui proces de afaceri include următoarele etape: definirea, reproiectarea, implementarea și îmbunătățirea permanentă.

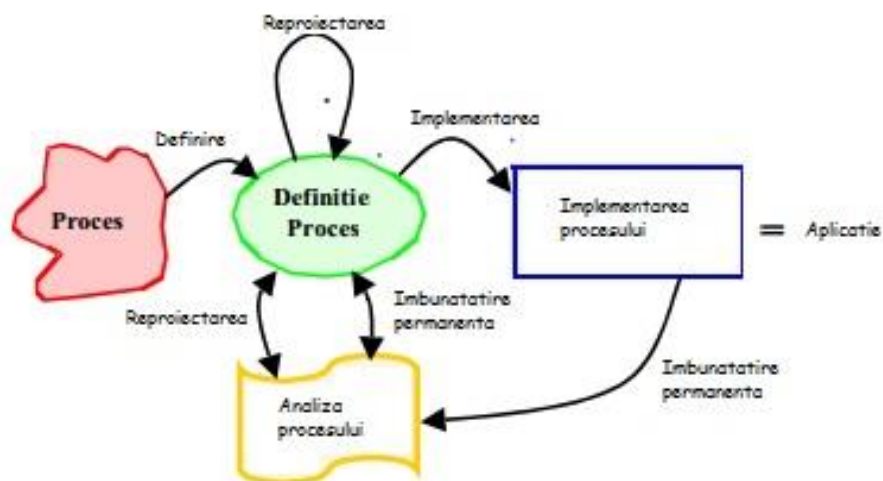


Fig. 1 Ciclul tradițional de viață al unui proces de afaceri [MKER]⁴

Pentru a putea defini un proces trebuie în primul rând să înțelegem procesul respectiv. Acest lucru presupune, în general, interviuarea persoanelor care cunosc foarte bine procesul. Definirea procesului nu poate fi realizată până când nu se obțin suficiente informații despre procesul respectiv.

Definiția procesului reprezintă o abstractizare a acestuia. Nivelul de abstractizare al procesului din cadrul definiției depinde de modul în care aceasta o să fie utilizată ulterior. De exemplu, o definiție poate descrie procesul la cel mai ridicat nivel conceptual, necesar înțelegerii, evaluării și reproiectării procesului. Pe de altă parte, o altă definiție poate descrie același proces la un nivel conceptual mult mai scăzut în vederea implementării acestuia.

⁴ [MKER] Marc Kerremans, Nicholas Kitson - Aligning Business Process Management and Business Intelligence to Achieve Business Process Excellence

Pentru a realiza o definire a procesului este necesar un model al procesului. Un model include în general un set de concepte care sunt folosite pentru descrierea proceselor, activităților, coordonarea activităților și funcțiile necesare care să poată realiza activitățile respective. Aceste concepte sunt înglobate într-un limbaj de definire al procesului. Validarea definiției procesului este necesară pentru a determina dacă definiția procesului corespunde într-adevăr acestuia.

Reproiectarea procesului presupune proiectarea unui proces nou care este intensiv, revoluționar, complet, suportat de soluțiile sistemului și cu rezultate vizibil îmbunătățite. Reproiectarea procesului ar trebui să fie ghidată de obiective de afaceri foarte clare, ca de exemplu creșterea satisfacției clienților, reducerea costurilor afacerii, reducerea timpilor de realizare al unui produs nou sau al unui serviciu.

Implementarea procesului presupune realizarea unui proces utilizând calculatoare, utilitare software și sisteme informatice. Acest lucru nu presupune neapărat ca toate activitățile procesului să fie automatizate, din moment ce unele pot fi realizate de oameni fără suport tehnologic.

Implementarea procesului a fost în general realizată în mod indirect prin introducerea unor părți ale procesului în sisteme software și folosind acțiunile angajaților care să asigure funcționarea întregului proces. În acest caz, definiția procesului servește ca design pentru funcțiile sistemului și comportamentul uman. Personalul IT asigură în general implementarea, care adeseori are loc fără să aibă loc discuții cu echipa managerială în legătură cu procesul.

Implementarea unui proces (nou sau îmbunătățit) include în general o instruire adecvată și instrucțiuni bine gândite care să conducă procesul către performanțele dorite.

Îmbunătățirea unui proces presupune realizarea unor mici corecții ale funcționării mai degrabă decât realizarea unei reproiectari integrale. Măsurătorile execuției procesului reprezintă baza pentru îmbunătățirea procesului (și a definiției sale). Măsurătorile pot arăta cum sunt abordate în general anumite direcții, care sunt timpii necesari rulării unui ciclu, ce costuri au fost necesare, etc. Analizând aceste date se poate ajunge la idei noi privind îmbunătățirea procesului bazându-se pe rezultatele efective ale procesului. În contrast, îmbunătățirile aduse procesului după definire, dar înainte de implementare, sunt bazate pe intuiția umană sau pe eventuale simulări bazate pe date estimate.

În general, măsurătorile sunt realizate prin adăugarea de instrumente în sistemele software și elaborarea unor modalități de măsurare a activității umane. În general, astfel de date trebuie colectate din mai multe surse.

Procesele de afaceri se pot clasifica după mai multe criterii, astfel: în funcție de nivelul organizatoric susținut de proces, în funcție de interacțiunea cu alte organizații, în funcție de nivelul de automatizare al proceselor, în funcție de repetabilitatea procesului și în funcție de nivelul de structurare al procesului.

În funcție de nivelul organizatoric susținut de proces, procesele se pot clasifica în următoarele subcategorii: procese de management – asigură buna funcționare a sistemului, procese operaționale – achiziții, producție, vânzări, marketing și procese organizaționale – contabilitate, resurse umane, suport tehnic.

În funcție de interacțiunea cu alte organizații, procesele se împart în două categorii: procese intraorganizaționale și procese interorganizaționale.

În funcție de nivelul de automatizare al proceselor, acestea se împart astfel tot în două categorii: procese automatizate complet și procese automatizate parțial.

În funcție de repetabilitatea procesului, există două categorii de procese: procese frecvente și procese ocazionale.

În funcție de nivelul de structurare al procesului, procesele pot fi de două tipuri: procese structurate și procese parțial structurate.

I.1.2. Managementul proceselor de afaceri

La începutul anilor '90, un curent numit re tehnologizarea proceselor de afaceri (Process Reengineering - BPR) a apărut din dorința de a face procesele de afaceri mai eficiente. Un termen des utilizat în acea perioadă îl reprezenta reducerea de personal. Retehnologizarea proceselor de afaceri nu a reușit să își atingă obiectivul, reducerea de personal fiind asociată cu disponibilizare și cu un volum de muncă mai ridicat pentru personalul rămas și nicidecum cu obiectivul final de eficiență globală. Retehnologizarea propunea să se renunțe la procesele existente și unele noi să fie create de la zero. Aceasta a fost o abordare curajoasă și foarte ambițioasă, care presupunea o schimbare radicală pentru companie.

La începutul anilor 2000, termenul de managementul proceselor de afaceri (Business Process Management – BPM), a fost inventat pentru a se referi la o nouă abordare de management, un mod holistic de a produce o organizație agilă și adaptivă bazată pe îmbunătățirea proceselor de afaceri.

Managementul proceselor de afaceri (BPM), spre deosebire de re tehnologizarea proceselor de afaceri (BPR), are la bază o abordare etapizată, îmbunătățind procesul etapă cu etapă, cu suficiente realizări pentru a justifica transformarea. Managementul proceselor de afaceri este o veche abordare pentru gestionarea proceselor, care a fost reînnoit prin utilizarea tehnologiei ca instrument de îmbunătățire.

Eficiența nu este singurul obiectiv al managementului proceselor de afaceri. Companiile încearcă să obțină un avantaj în fața competitorilor prin perfecționarea proceselor. Un proces mai bun poate să câștige clienți companiei; de exemplu, în cazul serviciilor cu clienții un proces mai bun presupune și reducerea timpului de lucru și al birocrăției interne.

I.1.3. Metodologii de management al proceselor de afaceri

O metodologie a unui proces de afaceri este o descriere formală a unei proceduri, pe care o echipă o poate urma pentru a restructura sau îmbunătăți un proces de afaceri. Unii autori preferă să descrie o astfel de procedură ca un "cadru", pentru a sugera că acesta oferă o descriere generală a modului de a proceda, dar evită să fie prea prescriptivă. Din moment ce avem tendința de a folosi termenul "cadru" pentru a face referire la un șablon care poate fi folosit pentru a defini un set de procese, am prefera să vorbim de un set de măsuri procedurale ca metodologie, și apoi pur și simplu putem face distincția între metodologiile mai exacte și mai puțin exacte. Un alt mod de a vorbi despre distincții este făcând referire la diferențele dintre metodologii bazate pe o percepție detaliată și cele bazate pe reguli.

Așa cum am folosit termenul în cadrul lucrării, o metodologie de management al proceselor de afaceri se concentrează pe reproiectarea sau îmbunătățirea unui proces de afaceri, și nu pe dezvoltarea unui sistem IT software. Astfel, nu voi lua în considerare o metodă bazată pe UML, ca de exemplu Rational Unified Process, ca o metodologie de management al proceselor de afaceri. RUP (Rational Unified Process) este o metodologie de dezvoltare de software.

Există mai multe metodologii dezvoltate pentru managementul proceselor de afaceri. Probabil, cea mai cunoscută și mai utilizată, este metodologia Rummler Brache, definit în cartea lui Geary Rummler [RUMM]⁵.

Un alt exemplu îl reprezintă metodologia propusă de Six Sigma, DMAIC (Definire, Măsurare, Analiză, Îmbunătățire, și Control) care a fost dezvoltată într-o metodologie în anii optzeci și continuă să fie utilizată pe scară largă de către practicieni Six Sigma chiar și astăzi. În plus, există mai multe alte metodologii acceptate de firme de consultanță, cum ar fi Catalyst CSC și metodologia ARIS IDS Sheer, care au tendința de a combina procesele de afaceri și metodele IT.

În ciuda variațiilor de denumire a pașilor și notațiile utilizate în diagramele de fluxuri de lucru, dacă privim cu atenție la specificul acestor abordări diferite, este posibil să distingem două mari categorii de metodologii de management al proceselor de afaceri: cele care se concentrează strict pe reproiectarea și îmbunătățirea proceselor de afaceri și cele care se concentrează mai mult pe organizarea și stabilirea a unui cadru pentru guvernarea procesului de afaceri.

Analizând metodologiile enumerate am considerat că metodologia de management al proceselor de afaceri propusă de BPTrends Associates (BPTA) are o abordare globală, integrată a managementului proceselor de afaceri și se adresează atât nivelului de întreprindere, cât și nivelului de proces și nivelului de implementare.

Metodologia recunoaște că organizațiile sunt axate pe o varietate de obiective diferite în eforturile lor de dezvoltare a proceselor. Metodologia BPTrends poate fi numită metodologia "cele mai bune practici". Aceasta nu caută să introducă noi tehnologii sau termeni inutili. Ori de câte ori este posibil, metodologia include utilizarea de instrumente existente și tehnici (cum ar fi BPMN, Balanced Scorecard, etc) oferind un cadru comun sau un context în care aceste instrumente și tehnici pot fi integrate în mod corespunzător și coordonat. În multe organizații, diferite grupuri luptă cu o parte din aceste metode diferite, instrumente și tehnici pentru a realiza etape ale misiunii lor generale de proces. Metodologia BPTrends Associates oferă o abordare coerentă în ceea ce privește coordonarea și gestionarea tuturor inițiativelor tehnologice ale unei întreprinderi, oferind cea mai mare performanță posibilă și asigurând randamentul investițiilor.

⁵ [RUMM] Geary Rummler and Alan Brache - Improving Performance: How to Manage the White Space on the Organization Chart, 1990, San Francisco: Jossey-Bass

În figura următoare sunt descrise toate serviciile bazate pe metodologia BPTrends de management al proceselor de afaceri.

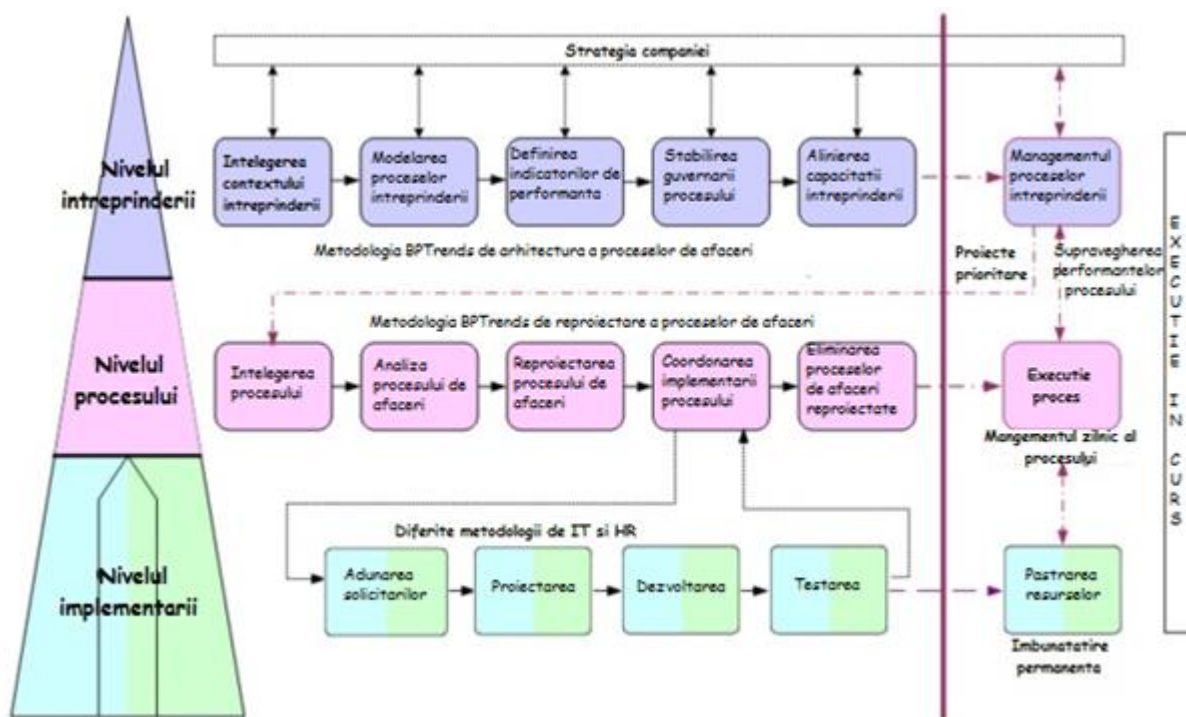


Figura 2. Servicii bazate pe metodologia BPTrends de management al proceselor de afaceri

La nivel de întreprindere, companiile se concentrează pe definirea unei arhitecturi a procesului de afaceri, dezvoltarea de sisteme de măsurare a performanțelor și sisteme de guvernare a proceselor, și crearea de centre de excelență de management al proceselor de afaceri care să coordoneze și să prioritizeze eforturile de proces. La nivel de întreprindere, metodologia BPTrends definește o abordare sistematică pentru a crea resursele și structurile organizatorice necesare pentru a genera și a menține o arhitectură de proces de afaceri, un centru de excelență de management al proceselor de afaceri și un sistem de guvernare care poate planifica, gestiona și monitoriza proiectele de reproiectare a proceselor de afaceri și executarea în curs a proceselor de afaceri în întreaga întreprindere.

La nivelul procesului, echipele de management al proceselor de afaceri sunt axate pe definirea, reproiectarea și îmbunătățirea proceselor existente. La nivelul procesului metodologia BPTrends este concepută pentru a preda practicanților la managementul proceselor de afaceri

cum să definească în mod constant și să modeleze corespunzător procesele lor de afaceri și cum să modeleze, analizeze, proiecteze sau reproiecteze viitoarele procese de afaceri.

La nivel de implementare, echipele de la resurse umane și de la IT se concentrează pe proiectarea sistemelor umane și a sistemelor software pentru a implementa procesele de afaceri. Metodologia BPTrends nu se extinde la dezvoltarea de software sau dezvoltarea resurselor umane. Cu toate acestea, o definește o interfață și folosește notații standard, cum ar fi managementul proceselor de afaceri, pentru a asigura o tranziție lină între reproiectarea procesului și HR și eforturile de implementare IT.

I.2. Limbaje pentru modelarea proceselor de afaceri

În ultimii ani, s-a observat un interes tot mai mare al organizațiilor pentru îmbunătățirea proceselor lor de afaceri, în scopul de a fi mai competitive într-o economie globalizată. Primul pas în realizarea acestui obiectiv este folosirea unui limbaj adecvat de modelare al procesor de afaceri pentru a putea reprezenta procesele de afaceri.

În acest scop, o evaluare a limbajelor de modelare a proceselor de afaceri existente ar fi foarte utilă pentru a facilita luarea deciziei corecte.

Evaluarea realizată în această lucrare se concentrează pe cele două notații grafice utilizate cel mai des pentru procese de afaceri: *Business Process Modeling Notation* (BPMN) și *UML Activity Diagram* (Diagrame de activitate UML). Criteriile de evaluare sunt: capacitatea de a fi ușor înțelese, adecvarea elementelor grafice ale BPMN și UML pentru a reprezenta procesele de afaceri reale ale unei organizații și maparea acestora la limbajele BPL (*Business Process Execution*). Rezultatele evaluării comparative a BPMN și UML sunt deasemenea prezentate în lucrare.

Reprezentarea proceselor de afaceri este o preocupare care datează din secolul trecut. Inițial, procesele care au loc în cadrul organizațiilor au fost reprezentate cu ajutorul diagramelor de fluxuri (Workflow Diagram), care au fost centrate pe activitatea fiecărui departament. Ulterior modele proceselor de afaceri (Business Process Models) au fost dezvoltate, si au reprezentat procesele care acoperă mai multe departamente, reușind captarea întregii organizații. Diagramele de flux de lucru sunt centrate pe procesele efectuate de persoane, pe baza documentelor, în timp ce, modele proceselor de afaceri sunt axate atât pe oameni cat și pe procesele de sistem.

Principalul scop al unei organizații economice este de a genera avantaje financiare pentru părțile implicate. Din acest motiv, mulți autori menționează că "pentru o lungă perioadă de timp, și chiar și astăzi, venitul net continuă să fie considerat principalul indicator de măsurare a performanței economice a unei entități" [JIA11]⁶. În procesul de gestionare a unei afaceri "există mai multe tipuri de decizii, care trebuie luate pentru dezvoltarea în termeni de eficiență a

⁶ [JIA11] Jianu, I., Jianu, I. & Gușatu, I. (2011) „Net income versus comprehensive income for professional investors”, Proceedings of the sixth edition of the International Conference Accounting and Management Information Systems: 966-987

activității economice a companiei" [TAR11]⁷. Modelele proceselor de afaceri sunt create "pentru a înțelege mecanismele cheie ale unei afaceri deja existente; să orienteze crearea unor sisteme informatice adecvate care sprijină activitatea; să pună în aplicare îmbunătățiri în activitatea curentă; pentru a arăta structura unei afaceri inovate; de a experimenta noi concepte de afaceri; și de a identifica elementele de afaceri care nu sunt considerate parte a nucleului, care ar putea fi delegată unui furnizor extern " [ERIK]⁸.

Ținând cont și de aceste mențiuni ale celorlalți autori, putem spune că modele proceselor de afaceri ajută managementul organizației din punct de vedere economic în luarea deciziilor adecvate în probleme importante, cu impact direct asupra venitului net pentru a fi în conformitate cu așteptările părților implicate.

De-a lungul anilor, diferite organizații au elaborat o serie de standarde pentru proiectarea, execuția, administrarea și monitorizarea proceselor de afaceri. Aceste standarde pot fi folosite separat sau în combinație, în funcție de compatibilitățile suportate. În ceea ce privește limbajele de notație, mai mulți autori [KAVI] consideră că două standarde sunt cele mai populare și sunt utilizate astfel pe scară largă în prezent: limbajul de notare al proceselor de afaceri (BPMN - *Business Process Languages Notation* și referit în continuare în cadrul lucrării ca BPMN) și diagramele de activitate UML (*UML Activity Diagrams* referite în continuare în cadrul lucrării ca diagrame UML)⁹.

Ținând cont că în ultima perioadă au fost dezvoltate tot mai multe modele arhitecturale, dedicate sistemelor informatice, modelarea proceselor de afaceri, folosind BPMN sau diagrame UML, are un rol foarte important în dezvoltarea sistemelor informatice, indiferent de arhitectura utilizată. Acest lucru este susținut și în alte lucrări [CZG], unde se spune că modelarea proceselor de afaceri, folosind BPMN sau diagrame UML, se poate utiliza în descrierea

⁷ [TAR11] Țarțavulea, R.I., Belu, M.G. & Dieaconescu, V.C. (2011) "Spatial modeling in logistics decision-making processes. Identifying the optimal location for a single central warehouse", *Annals of the University of Oradea, Economic Science Series*, Tom XX, vol. 1: 137-143

⁸ [ERIK] Eriksson, H. & Penker, M. (2000) *Business Modeling with UML: business patterns at work*, John Wiley & Sons

⁹ [KAVI] Kalnins, A. & Vitolins, V. (2006) "Use of UML and Model Transformations for Workflow Process Definitions", *Databases and Information Systems IV - Selected Papers from the Seventh International Baltic Conference, DB&IS 2006*: 3-15

algoritmilor folosiți în sistemele informatice, inclusiv aplicațiile bazate pe inteligența artificială deoarece "inteligența artificială ar putea deveni principala alternativă pentru rezolvarea problemelor financiare, care necesită calcule matematice sau de optimizare complexe".

În continuare, în cadrul cercetării o să evidențiez care din tehnologiile de modelare a proceselor de afaceri prezentate în lucrare (BPMN, sau diagrame UML), ar trebui să fie aleasă de către organizații pentru a își modela procesele de afaceri. Astfel, mă voi focaliza pe analiza diagramelor UML și BPMN din mai multe perspective: nivelul de înțelegere al utilizatorilor, lejeritatea elementelor grafice a limbajelor de a exprima procesul real de afaceri al companiei și lejeritatea de mapare a acestor limbaje la limbajul de execuție al proceselor de afaceri (BPEL - *Business Process Execution Languages*).

Analiza și evaluarea limbajelor de modelare a proceselor de afaceri a fost abordată în multe lucrări de specialitate, mai ales a diagramelor UML și BPMN care sunt cele mai utilizate pentru reprezentarea proceselor de afaceri. În cadrul lucrării de cercetare, am luat în considerare și aceste analize, am introdus parametrii noi de evaluare și am realizat o comparare diferită a limbajelor bazându-mă pe cât mai mulți parametrii. În general, cercetările menționate analizează modelarea proceselor de afaceri dintr-o singură perspectivă: puterea de exprimare, lizibilitatea sau capacitatea acestora de mapare la limbajul de execuție al proceselor de afaceri. De asemenea, analizele sunt bazate pe versiunile de BPMN și UML care au fost în uz la data la care au fost făcute cercetările, care nu sunt versiunile utilizate în prezent. Deși schimbările nu sunt majore între ultimele versiuni aparute ale limbajelor de modelare, în cadrul lucrării am evaluat și comparat ultimele versiuni apărute ale acestora.

Mai mulți autori au realizat studii despre capacitatea BPMN și a diagramelor UML de mapare la limbajul de execuție al proceselor de afaceri. Un studiu [MAZ] arată cum bazându-ne pe tehnici de programare logică putem construi o transformare bidirecțională între BPMN și

BPEL¹⁰. Un alt studiu [ZHA] propune o modalitate de a transforma diagramele de activitate UML în BPEL¹¹.

1.2.1. Limbajul BPMN

Business Process Modeling Notation (BPMN) este un limbaj de notație grafică, dezvoltat de Business Process Initiative Management (BPIM), foarte utilizat pentru modelarea proceselor de afaceri. Începând cu anul 2005, BPMN este gestionat de OMG (Object Management Group), după fuziunea dintre această organizație și BPIM. În ianuarie 2011, OMG a lansat BPMN versiunea 2.0 care își extinde domeniul de aplicare și capacitățile față de versiunea anterioară, BPMN 1.2, în mai multe domenii (OMG, 2011a) precum: formalizarea execuției semantice a tuturor elementelor BPMN, definirea unui mecanism de extensie pentru extensiile de modelare a proceselor și extensiile grafice și îmbunătățirea și corelarea componentelor evenimentelor. În decembrie 2013 este lansată versiunea BPMN 2.0.2 (ultima versiune disponibilă la momentul efectuării cercetării) care aduce modificări minore versiunii 2.0. În noua versiune 2.0.2 a fost modificată clauza 15 care se referă la formatele de schimb.

Conform documentației asociate BPMN, scopul principal al BPMN este "de a oferi un mod de notație, care este ușor de înțeles de către toți utilizatorii din mediul de afaceri, de la analiștii de afaceri care creează proiectele inițiale ale proceselor, la dezvoltatorii tehnici responsabili de punerea în aplicare a tehnologiei, care vor efectua aceste procese, și în cele din urmă, pentru oamenii de afaceri care vor să administreze și să monitorizeze aceste procese" [OMGa]¹².

BPMN permite crearea de procese de afaceri complete, fiind conceput pentru a acoperi mai multe tipuri de modelare, astfel încât să poată comunica o mare varietate de informații unui

¹⁰ [MAZ] Mazanek, S. & Hanus, M. (2011) "Constructing a bidirectional transformation between BPMN and BPEL with a functional logic programming language", Journal of Visual Languages & Computing, vol. 22, no. 1: 66–89

¹¹ [ZHA] Zhang, M. & Duan, Z. (2008) "From Business Process Models to Web Service Orchestration: The case of UML 2.0 Activity Diagram to BPEL", Lecturer Notes in Computer Science, vol. 5364: 505-510

¹² [OMGa] Object Management Group – Document Associated with BPMN version 2.0.2

public la fel de variat. Un model BPMN complet conține trei tipuri de sub-modele de bază: de procese, de coregrafie și de colaborare. Prin combinarea celor trei tipuri de sub-modele de bază, o poate fi obținută o reprezentare detaliată a proceselor de afaceri, dar este recomandat ca designerul să se concentreze pe un anumit aspect al analizei proceselor, pentru a evita crearea de diagrame prea complexe, care sunt greu de înțeles.

I.2.2. Diagrame de activitate UML

Diagramele de activitate UML au fost elaborate și sunt menținute de OMG. Prima versiune a UML a fost lansată în 1995. Versiunea actuală (la momentul efectuării cercetării), UML 2.5, a fost lansată în martie 2015. Obiectivul principal al UML este "de a oferi arhitecților de sistem, proiectanților și dezvoltatorilor software, instrumente de analiză, proiectare, implementare, precum și pentru modelarea proceselor de afaceri și a altor procese similare" [OMGb]¹³.

Nucleul modelării utilizate în UML pentru modelarea proceselor de afaceri este reprezentat de diagrama de activitate, care face parte din modelele comportamentale. Diagramele de activitate sunt reproiectate în mod semnificativ odată cu trecerea la versiunea 2.0 a UML, atât la nivel de sintaxă, precum și în ceea ce privește semantica, prin trecerea de la semantica stării mașinii (*state machine*) la semantica fluxului jetoanelor (*token flow*). Aceste schimbări au îmbunătățit capacitatea diagramelor de activitate UML de a reprezenta procesele de afaceri.

I.2.3. Analiza comparativă a limbajelor BPMN și UML

Analiza din punct de vedere al puterii de exprimare

Rezultatele modelării proceselor de afaceri sunt de interes pentru mai multe părți implicate: analiștii de afaceri care descriu procesele folosind notații și instrumente specifice,

¹³ [OMGb] Object Management Group – Document Associated with Unified Modeling Language (UML) version 2.5

dezvoltatorii tehnici care implementează tehnologia utilizată pentru a exercita aceste procese și utilizatorii de afaceri care vor gestiona și monitoriza procesele.

Ținând cont că în cadrul cercetării dorim să evaluăm lejeritatea de înțelegere a limbajelor de modelare, trebuie să avem în vedere și faptul că utilizatorii de afaceri nu trebuie să fie experți în limbajele de modelare a proceselor de afaceri, ci trebuie doar să înțeleagă rezultatele modelării, mai precis, trebuie să știe cum să citească diagramele proceselor de afaceri. Prin urmare, ar trebui ca diagramele procesului modelat să fie ușor de utilizat și de înțeles de către toate părțile care sunt, direct sau indirect, implicate în acest proces.

Atât dezvoltatorii standardului BPMN, cât și cei ai UML susțin că principalul țel îl reprezintă furnizarea unui mijloc de notare grafică care este ușor de înțeles de către toți utilizatorii din mediul de afaceri [OMGa, OMGb]¹⁴.

Pentru a evalua aceste afirmații, a fost efectuat un studiu [PEX] pentru a analiza lizibilitatea modelelor proceselor de afaceri, realizate folosind diagramele UML sau BPMN¹⁵. Participanții la experiment au fost persoane din IT nefamiliarizate cu limbajele de modelare și cu domeniul modelat. Concluzia studiului a fost că, pentru modelele de flux de lucru analizate, nivelul de dificultate pentru înțelegerea procesului de afaceri, în ambele limbaje, este același. Un alt studiu similar [BIRK] a făcut o comparație empirică între BPMN și diagramele UML¹⁶. Rezultatele acestui studiu au indicat faptul că diagramele UML sunt cel puțin la fel de utilizabile ca BPMN, deoarece nici BPMN nu diferă în mod semnificativ în eficacitate, eficiență, sau satisfacția utilizatorilor.

¹⁴ [OMGa] Object Management Group – Document Associated with BPMN version 2.0.2

[OMGb] Object Management Group – Document Associated with Unified Modeling Language (UML) version 2.5

¹⁵ [PEX] Peixoto, D.C.C., Batista, V.A., Atayde, A.P., Borges, E.P., Resende, R. F. & Pádua, C.I. (2008) “A Comparison of BPMN and UML 2.0 Activity Diagrams”, VII Simpósio Brasileiro de Qualidade de Software: 1-12

¹⁶ [BIRK] Birkmeier, D., Klöckner, S. & Overhage, S. (2010) “An empirical comparison of the usability of BPMN and UML Activity Diagrams for business users”, 18th European Conference on Information Systems: 1-12

Analiza din punct de vedere al lizibilității

Așa cum este susținut și în alte lucrări de specialitate [AAL, RUSS], puterea de reprezentare a limbajelor de modelare a proceselor de afaceri poate fi evaluată cu ajutorul unui cadru general de evaluare acceptată – cadrul modelelor fluxurilor de lucru (Workflow Models).¹⁷ Cadrul modelelor fluxurilor de lucru oferă un set general de modele al proceselor de afaceri, care pot fi utilizate pentru a evalua în ce măsură limbajul fluxului de lucru analizat sau a limbajului de modelare a proceselor de afaceri este în măsură să reprezinte un anumit model de flux de lucru specificat.

Modelele fluxului de lucru se împart în patru categorii: de control, de date, de resurse și de manipulare a excepțiilor. Modelul fluxului de control poate fi utilizat pentru a analiza aspectele legate de controlul fluxului dependențelor dintre diferite sarcini. Modelul fluxului de date se referă la modul în care datele sunt reprezentate și utilizate în fluxurile de lucru. Modelul fluxului de resurse oferă o imagine cuprinzătoare din perspectiva resurselor, captând aspecte legate de distribuția muncii la resursele asociate cu un proces de afaceri, și modul în care acest lucru este gestionat de aceste resurse. Modelul fluxului de manipulare a excepțiilor surprinde cauzele excepțiilor și acțiunile care trebuie luate atunci când apar excepții.

Cât de adecvată este utilizarea diagramelor UML sau BPMN pentru a reprezenta procesele de afaceri a fost dezbătută de o serie de autori. Unii autori [DUMH] au examinat expresivitatea și caracterul adecvat al diagramelor UML pentru fluxul de lucru și au evaluat capacitatea diagramelor de a capta o colecție de modele similare de flux de lucru¹⁸. Alți autori, ca de exemplu [RUSS], au susținut oportunitatea utilizării diagramelor UML pentru modelarea proceselor de afaceri, folosind modelele de flux de lucru ca un cadru de evaluare, în timp ce alți

¹⁷ [AAL] van der Aalst, W.M.P, ter Hofstede, A.H.M., Kiepuszewski, B. & Barros, A.P. (2003) “Workflow Patterns”, Distributed and Parallel Databases, vol. 14, no. 3: 5-51

[RUSS] Russell, N., van der Aalst, W.M.P, ter Hofstede, A.H.M. & Wohed, P. (2006c) “On the Suitability of UML 2.0 Activity Diagrams for Business Process Modelling”, Proceedings of the 3rd Asia-Pacific conference on Conceptual modelling, vol. 53: 95-104

¹⁸ [DUMH] Dumas, M. & ter Hofstede, A. (2001) “UML activity diagrams as a workflow specification language”, Proceedings of the Fourth International Conference on the Unified Modeling Language (UML 2001): 76–90

autori [SAL] investighează capacitatea diagramelor de activitate UML de a modela perspectiva resurselor proceselor de afaceri și compară diagramele de activitate cu rețelele Petri¹⁹.

Un alt aspect care trebuie luat în considerare atunci când se analizează puterea de reprezentare a BPMN și a diagramelor UML este reprezentat de complexitatea simbolurilor grafice utilizate pentru a reprezenta procesele de afaceri reale ale unei organizații. În multe cazuri, BPMN și diagramele UML folosesc simboluri similare pentru a descrie procesele de afaceri. Cu toate acestea, există aspecte ale proceselor de afaceri care pot fi modelate în BPMN utilizând un singur simbol, dar pentru care reprezentarea în diagrame UML necesită utilizarea unui grup de simboluri. Această ultimă situație vine ca urmare a faptului că BPMN nu folosește întotdeauna un singur simbol pentru reprezentarea fiecărei componente a unui proces de afaceri; se folosesc, de asemenea, simboluri complexe pentru a descrie o serie de informații ca un întreg. Pe de altă parte, diagramele UML folosesc un simbol pentru fiecare componentă a proceselor de afaceri.

Pentru a analiza simbolurile grafice utilizate pentru modelarea procesului de afaceri am elaborat un studiu de caz care constă în modelarea unui proces de afaceri folosind atât BPMN (Figura 3), cât și diagrame UML (Figura 4).

În studiul de caz am ales să descriu procesele implicate de reparațiile efectuate de un service auto asupra vehiculelor avariate aduse de clienții lor. Procesul începe cu solicitarea făcută de un client la serviceul auto pentru repararea vehiculului. Urmează etapa de programare de către serviceul auto a reparației. La momentul de start al reparației (momentul programării), clientul aduce vehiculul și serviceul auto efectuează reparațiile necesare. Când serviceul a terminat toate reparațiile se generează o factură care trebuie plătită de către client pentru a putea ridica ulterior vehiculul reparat.

¹⁹[RUSS] Russell, N., van der Aalst, W.M.P, ter Hofstede, A.H.M. & Wohed, P. (2006c) “On the Suitability of UML 2.0 Activity Diagrams for Business Process Modelling”, Proceedings of the 3rd Asia-Pacific conference on Conceptual modelling, vol. 53: 95-104

[SAL] Sarshar, K. & Loos, V. (2007) “Modeling the Resource Perspective of Business Processes by UML Activity Diagram and Object Petri Net”, Enterprise Modeling and Computing with UML: 204-215

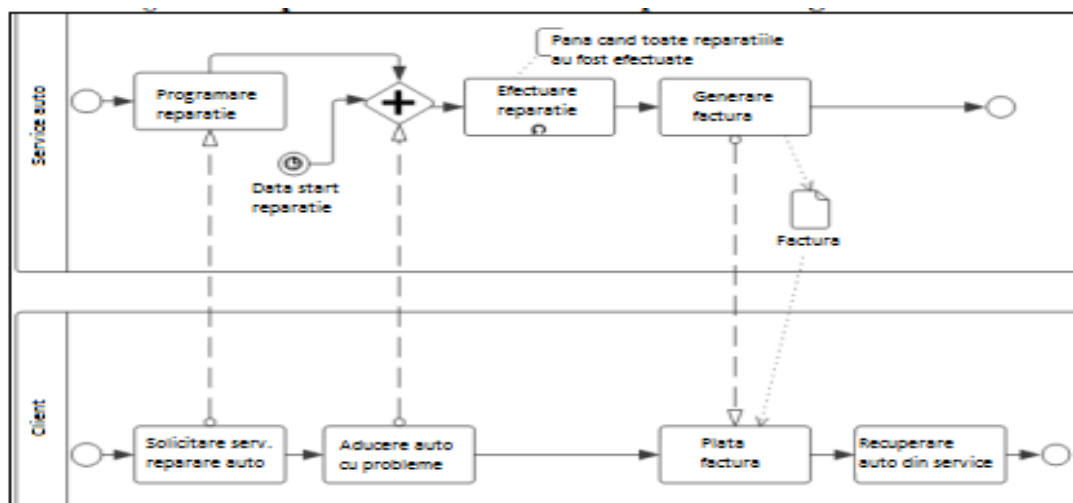


Figura 3. Reprezentarea unui proces de afaceri utilizând BPMN

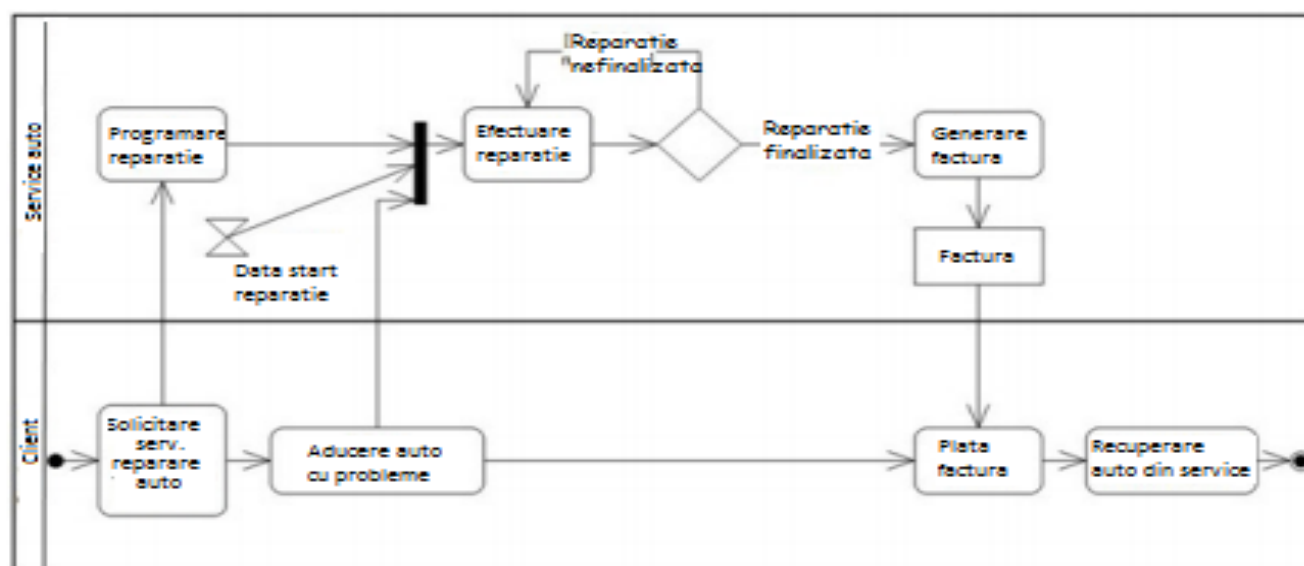


Figura 4. Reprezentarea unui proces de afaceri utilizând diagrame UML

Prin analiza simbolurile grafice utilizate pentru reprezentarea procesului de afaceri descris în figura 3 și figura 4 putem trage următoarele concluzii:

- Simbolurile grafice utilizate pentru reprezentarea majorității etapelor procesului sunt similare în BPMN și diagramele UML.
- Pentru reprezentarea reparațiilor efectuate de serviceul auto, BPMN folosește doar un simbol (un obiect sarcină cu un marker buclă standard), în timp ce diagramele UML folosește un grup de simboluri (un nod acțiune, un nod decizie și două margini de activitate) - Figura 5.

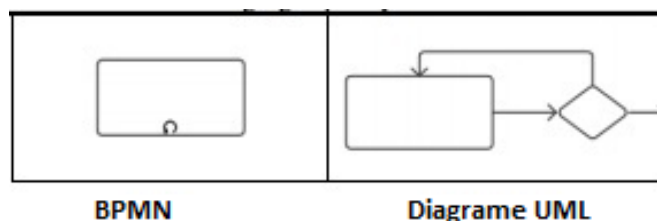


Figura 5. Evidențierea utilizării simbolurilor BPMN – diagrame UML

Analiza din punct de vedere al capacității de mapare

Următorul pas, după crearea unei reprezentări vizuale a proceselor de afaceri (folosind limbajele de modelare a proceselor de afaceri, cum ar fi BPMN și diagramele UML), îl reprezintă execuția. Pentru a atinge acest obiectiv, este necesar să se mapeze reprezentările vizuale ale proceselor de afaceri (modele BPMN și UML) într-un limbaj de execuție a proceselor de afaceri (BPEL).

Cea mai recentă versiune de BPEL este WS-BPEL 2.0 [OASIS], care este un limbaj pentru specificarea comportamentului proceselor de afaceri bazate pe Web Services. WS-BPEL 2.0 a adus îmbunătățiri semnificative la versiunea anterioară - BPEL4WS 1.1. "WS-BPEL definește un model de integrare interoperabilă care ar trebui să faciliteze extinderea integrării proceselor automate, atât în cadrul aceleiași companii cât și în spațiile *business-to-business*" [OASIS]²⁰. Informațiile procesului WSBPEL sunt exportate și importate numai prin utilizarea interfețelor de servicii web.

BPMN 2.0.2 include o cartografiere a unui subset de BPMN la un limbaj de execuție al proceselor de afaceri, respectiv WS-BPEL. "Mapările la alte standarde în curs de dezvoltare sunt considerate a fi eforturi individuale" [OMGa]²¹. Între BPMN și BPEL există unele diferențe importante. De exemplu, în BPMN sarcinile pot fi legate în orice formă, în timp ce fluxurile de legătură BPEL permit doar conexiuni cronologice, nu și bucle. Prin urmare, cartografierea nu este simplă. În secțiunea "Maparea modelelor BPMN la WS-BPEL", a actului normativ BPMN

²⁰ [OASIS] Online Community for the Web Services Business Process Execution Language OASIS Standard - <http://bpel.xml.org/>

²¹ [OMGa] Object Management Group – Document Associated with BPMN version 2.0.2

2.0 [OMGa], specificațiile descriu atât "maparea de bază", cât și "maparea extinsă", care se referă la blocurile BPMN care pot fi mapate folosind multiple modele WS-BPEL.

În ceea ce privește diagramele UML, nici ultimul act normativ [OMGb], nici versiunea anterioară a standardului, nu includ o precizare de mapare a diagramelor UML la un limbaj de execuție al proceselor de afaceri²². Cu toate acestea, în ultimii ani, definirea unei mapări între diagramele UML și BPEL a fost în zona de interes a numeroase cercetări. Au fost autori care au propus un model de transformare a diagramelor UML 2.0 la BPEL prin descompunerea unui model al diagramei în regiuni și identificarea modelelor structurale separat [ZHA]²³. Alții [HLB], propun o transformare pe bază de meta-model de la diagrame de activitate UML la limbajul BPEL4WS²⁴. Deși rezultatele acestor cercetări sunt aplicabile în practică, ele nu oferă soluții pentru o mapare automată completă a diagramelor UML la limbajele de execuție a proceselor de afaceri.

²² [OMGb] Object Management Group – Document Associated with Unified Modeling Language (UML) version 2.5

²³ [ZHA] Zhang, M. & Duan, Z. (2008) "From Business Process Models to Web Service Orchestration: The case of UML 2.0 Activity Diagram to BPEL", *Lecturer Notes in Computer Science*, vol. 5364: 505-510

²⁴ [HLB] Hlaoui, Y.B. & Benayed, L.J. (2011) "A Model Transformation Approach Based on Homomorphic Mappings between UML Activity Diagrams and BPEL4WS Specifications of Grid Service Workflows", *Computer Software and Applications Conference Workshops (COMPSACW) - 2011 IEEE 35th Annual*: 243-248

II. Tehnici informatice de analiză a proceselor de afaceri

Dorința companiilor de a câștiga teren în fața competitorilor a condus la o luptă crâncenă pentru cifra de afaceri, profit, sau diminuarea costurilor.

Companiile au început astfel să descopere puterea oferită de date și au început să stocheze cât mai multe informații. Cu creșterea volumului de date, a apărut și conceptul nou de *big data* (volume mari de date) care se atribuie seturilor mari de date structurate, sau nestructurate, generate într-un interval scurt de timp.

Necesitatea exploatarea informațiilor din aceste volume mari de date, a arătat limitele sistemelor relaționale de gestiune a bazelor de date și a condus la apariția unui concept nou, de NoSQL.

De asemenea, au fost dezvoltate tehnici și algoritmi pentru extragerea cunoștințelor din date, în vederea obținerii informațiilor „ascunse” în date. Cu timpul, procedurile s-au standardizat conducând astfel la apariția metodologiilor de extragere a cunoștințelor din date.

Observând valoarea informațiilor extrase din date, tot mai multe companii, indiferent de dimensiune, sau sfera de activitate, au dorit să profite de influența tehnicilor informatice de analiză a datelor. În acest scop, companiile IT au dezvoltat, atât soluții, cât și unele mai intuitive, pentru a acoperi cererea de instrumente de analiză a datelor.

II.1. Tehnologia volumelor mari de date

II.1.1. Noțiuni fundamentale

Volumele mari de date (*big data*) reprezintă o nouă putere care schimbă toate domeniile cu care interacționează și este considerată de unii a fi energie electrică a secolului XXI [MCK11]²⁵. Acest concept, de care am auzit doar la începutul secolului XXI, a fost primul care a introdus unui volum de date atributele de prea mare sau prea nestructurat.

Primul atribut principal al volumelor mari de date este volumul. Unele volume de date au fost cuantificate prin numărarea înregistrărilor, tranzacțiilor, tabelelor sau fișierelor, pentru altele în schimb, a fost considerat mai util cuantificarea datelor în termeni de timp. De exemplu, în SUA unele norme specifică păstrarea datelor disponibile pentru analiză juridică timp de șapte ani.

Al doilea atribut îl reprezintă varietatea datelor. Acest lucru se întâmplă deoarece datele provin dintr-o varietate de surse cum ar fi loguri, fluxuri, rețele sociale, date de tip text, date semi-structurate din procesele B2B, etc..

Ultimul atribut al volumelor mari de date este viteza, care se referă la viteza în timp real la care analizele trebuie să fie aplicate.

²⁵ [MCK11] McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity – June 2011

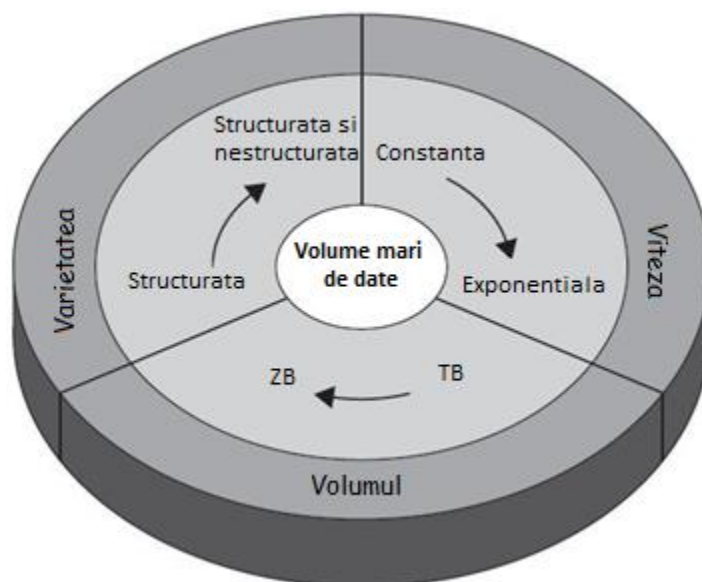


Fig 1. Cele trei componente ale volumelor mari de date [PAN13]²⁶

Printre cele mai populare definiții ale volumelor mari de date se regăsesc și următoarele trei:

Volumele mari de date reprezintă termenul folosit pentru o colecție de seturi de date atât de mare și complexă, care devine dificilă de prelucrat cu ajutorul instrumentelor de gestionare a bazelor de date sau folosind aplicațiile tradiționale de prelucrare a datelor [WIKI]²⁷.

Volumele mari de date reprezintă activele informaționale de mare volum, de mare viteză și de mare diversitate, care necesită o eficiență a costurilor, forme inovatoare de prelucrare a informațiilor, pentru o înțelegere sporită și pentru luarea deciziilor [GART]²⁸.

Volumele mari de date se referă la seturile masive de date care se colectează în timp, care sunt dificil de analizat și manipulat folosind instrumente de gestionare a bazelor de date comune.

²⁶ [PAN13] Andrei Pandre – Datawatch has 3 Vs - 19 Noiembrie 2013

²⁷ [WIKI] Wikipedia – The Free Encyclopedia - http://en.wikipedia.org/wiki/Big_data

²⁸ [GART] Gartner, Inc. – <http://www.gartner.com/it-glossary/big-data>

Volumele mari de date includ tranzacțiile de afaceri, mesaje e-mail, fotografii, clipuri video de supraveghere și log-uri de activitate [PCMG]²⁹.

Pornind de la aceste definiții putem afirma că tehnologia volumelor mari de date reprezintă o tehnologie de nouă generație, cu o nouă arhitectură, menită să extragă informațiile valoroase dintr-un set mare de date alcătuit din diverse surse și cu un flux ridicat de generare.

Dacă privim elementele comune din aceste definiții, începem să vedem unele tendințe și cerințe necesare pentru a măsura ce reprezintă de fapt utilizarea unui volum mare de date. Astfel, orice analiză care se bazează pe volume mari de date trebuie:

- să utilizeze seturi de date care sunt prea mari pentru a fi gestionate utilizând instrumentele standard ale unei baze de date;
- să utilizeze date din mai multe surse;
- să interpreteze datele pentru a ajuta la ușurarea procesului de luare a deciziilor.

Se observă astfel că tehnologia volumelor mari de date nu are ca scop doar stocarea unei cantități cât mai mari de date. Această tehnologie își propune să creeze o infrastructură flexibilă cu o putere ridicată de procesare și analizare a datelor în așa fel încât să poată oferi informații strategice companiei.

Chiar și cu toate această nebulozitate și incertitudine cu privire la ceea ce este și ceea ce nu este un volum mare de date, se va continua să se vorbească despre această tehnologie, deoarece are un potențial extrem de ridicat. Un raport efectuat de [MCK11] afirmă următoarele: “Analiza volumelor mari de date va deveni un element de concurență, care va sta la baza noilor valuri de creștere a productivității, inovației, și surplusului de consum atât timp cât politicile sunt corecte și facilitatorii sunt la locul lor”.³⁰

Tehnologia volumelor mari de date implică mai mult decât simpla abilitate de a stoca volume mari de date. Printre primele companii care au imbratisat încă de la început acest nou

²⁹ [PCMG] PC Mag - <http://www.pcmag.com/encyclopedia/term/62849/big-data>

³⁰ [MCK11] McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity – June 2011

concept, și s-au dezvoltat ulterior în jurul său, putem aminti Google, Twitter, Facebook sau eBay. Aceste companii au generat un volum mare de date într-un format nou și mai puțin structurat (loguri servere web, relațiile rețelei de socializare, etc.) și au fost nevoite astfel să implementeze tehnologii noi și să încerce alte abordări manageriale.

Aceste companii nu au fost singurele care au început să aibă aceste probleme. De exemplu, în industria de turism începeau să apară aceleași dificultăți. Fiecare bilet de avion rezervat, cazare la hotel, închiriere de mașină sau orice altă rezervare, reprezintă date care sunt stocate, iar aceste date s-au adunat în cursul anilor formând volume mari de date nestructurate care pot fi de ordinul sutelor de terabytes sau chiar zetabytes.

Multe echipe de cercetare s-au adunat pentru a aduna informații și pentru a studia valoarea totală a datelor generate, stocate, precum și consumate în lume. Deși au avut de realizat estimări în scopuri diferite și, prin urmare, rezultatul lor variază, toate indică o creștere exponențială a volumelor de date în anii care urmează.

MGI estimează că întreprinderile vor stoca la nivel global mai mult de 7 exabytes de date noi, în următorul an, în timp ce consumatorii vor stoca mai mult de 6 exabytes de date noi pe dispozitive, cum ar fi PC-uri și notebook-uri. Un exabyte de date este echivalentul a mai mult de 4.000 de ori informațiile stocate în Biblioteca Congresului SUA. Într-adevăr, generăm atât de multe date în ziua de azi, încât a devenit fizic imposibil să fie stocate toate.

În zilele noastre fiecare sector din economia mondială se confruntă cu o problemă mare de date. Până în 2009, aproape toate sectoarele din economia Statelor Unite au avut o medie de cel puțin 200 de terabytes de date stocate pentru fiecare companie cu mai mult de 1.000 de angajați [MCK11]³¹.

Potrivit Reuters [NASS], volumele mari de date vor crește de la 3,2 miliarde dolari în 2010 pentru a ajunge la o industrie 25 miliarde dolari până în 2015.³²

³¹ [MCK11] McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity – June 2011

³² [NASS] Nasscom – Crisil GR&A Analysis

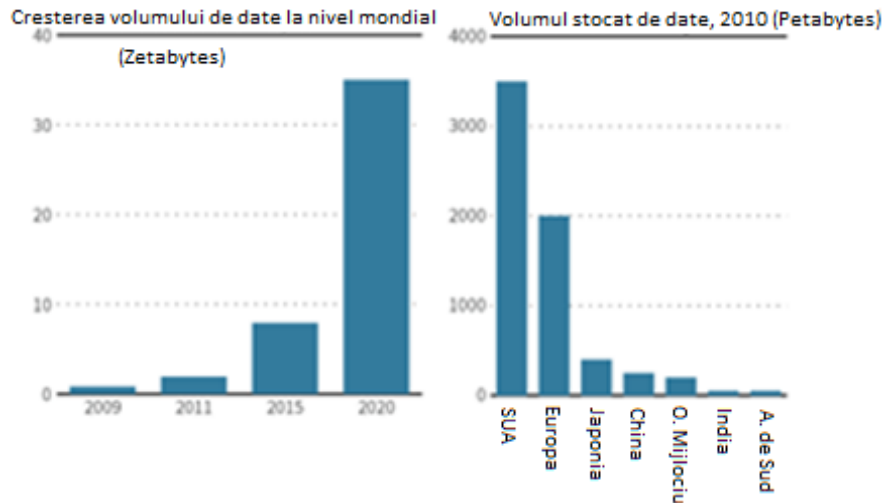


Figura 2. Evoluția volumelor mari de date [NASS]³³

Nu întâmplător, raportul [PCAST10] a identificat tehnologia volumelor mari de date ca fiind „frontiera cercetării care poate accelera progresul pe o gamă largă de priorități”³⁴. Tehnologia volumelor mari de date a devenit, chiar și pentru mass-media, un subiect extrem de popular. Încă de câțiva ani, publicații renumite, precum The Economist [ECO11], The New York Times [NYT12], sau National Public Radio [NPR11a, NPR11b], au încercat să aducă în atenția publicului lor importanța tehnologiei volumelor mari de date.³⁵

³³ [NASS] Nasscom – Crisil GR&A Analysis

³⁴ [PCAST10] Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology. PCAST Report, Dec. 2010

³⁵ [ECO11] Drowning in numbers -- Digital data will flood the planet—and help us understand it better. The Economist, Nov 18, 2011

[NYT12] The Age of Big Data. Steve Lohr. New York Times, Feb 11, 2012

[NPR11a] Following the Breadcrumbs to Big Data Gold. Yuki Noguchi. National Public Radio, 29 Noiembrie 2011

[NPR11b] The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. National Public Radio, 30 Noiembrie 2011

II.1.2. Etapele utilizării volumelor mari de date

Achiziția și stocarea datelor

Volumele mari de date nu apar din neant, ci se formează prin stocarea datelor generate de sursele de date. De exemplu, dacă pornim de la capacitatea noastră de a simți și observa lumea din jurul nostru, ritmul cardiac al unui cetatean, sau prezența toxinelor în aerul pe care îl respirăm și ajungând la telescoapele de ultima generație, vom realiza că datele brute generate într-o singură zi pot depăși 1 milion TB. În mod similar, experimente științifice și simulări pot produce cu ușurință un volum la fel de mare de date.

O mare parte din aceste date nu reprezintă date de interes, și astfel, volumul poate fi filtrat și comprimat cu câteva ordine de mărime. O provocare este de a defini aceste filtre în așa fel încât informațiile utile să nu fie aruncate. De exemplu, să presupunem că datele oferite de un senzor diferă substanțial de restul: este probabil ca acest lucru să fie din cauza unei defecțiuni a senzorului, dar înainte de a filtra aceste date, trebuie să ne asigurăm că ele nu reprezintă defapt o abatere de la modul normal de funcționare. Exact același principiu trebuie avut în considerare și atunci când sortăm cv-urile candidaților. Atfel, trebuie să avem grijă să nu eliminăm un cv care iese din timpar considerându-l nepotrivit, deoarece tocmai acest cv care nu respectă tiparul poate constitui cv-ul candidatului ideal.

Avem nevoie de cercetare în știința comprimării și analizei datelor pentru a putea procesa în mod inteligent aceste date brute obținând o dimensiune adecvată astfel încât utilizatorii să nu piardă informațiile esențiale. Mai mult, sunt necesare tehnici de analiză în timp real (on-line), tehnici care pot procesa aceste date cu o viteză ridicată, deoarece nu este posibilă întâi stocarea și ulterior comprimarea și exploatarea datelor.

A doua mare provocare este de a genera automat metadatele ideale care să descrie ce date sunt înregistrate și modul în care se înregistrează și se măsoară. De exemplu, în experimente științifice, sunt necesare detalii consistente în ceea ce privește condițiile și procedurile experimentale specifice, acestea putând fi necesare pentru a putea interpreta corect rezultatele, și din acest motiv este important ca astfel de metadate să fie înregistrate cu date observaționale.

Sistemele de achiziție de metadate pot minimiza povara umană în procesul de înregistrare al metadatelor.

Un alt aspect important îl reprezintă proveniența datelor. Stocarea datelor în momentul apariției lor nu este utilă, cu excepția cazului în care aceste informații pot fi interpretate ulterior de sistemul de analiză al datelor. De exemplu, o eroare de procesare într-o etapă poate face inutilă analiza ulterioară; cu atenția potrivită, se poate identifica cu ușurință toată prelucrarea ulterioară care depinde de acest pas. Astfel, este nevoie de cercetare, atât în generarea metadatelor adecvate, cât și în sistemele de date care transportă datele și metadatele prin fluxurile de analiză a datelor.

Pregătirea datelor în vederea exploatării

În mod frecvent, datele colectate nu sunt în formatul potrivit pentru analiză. De exemplu, putem lua în considerare datele colectate din cv-urile depuse de candidați pe site-urile de recrutare, care pot cuprinde date provenite din mai multe surse (eventual cu o oarecare incertitudine legată de versiunea finală de curriculum vitae), date structurate de la anumite teste și verificări, date video, cum ar fi interviurile.

Pentru a putea fi analizate cu succes, datele nu pot rămâne în această formă. Mai degrabă avem nevoie de un proces de extragere a informației, care să selecteze informațiile solicitate din sursele care stau la baza volumului de date și să le remodeleze într-o formă structurată adecvată pentru analiză. Realizarea completă a acestui lucru într-un mod corect reprezintă o provocare tehnică continuă. Trebuie luat în considerare că aceste date pot include fișiere text, imagini sau chiar video.

Nu trebuie să ne bazăm pe faptul că tehnologia volumelor mari de date ne va spune întotdeauna adevărul, deoarece acest lucru este departe de realitate. De exemplu, candidatii pot alege să ascundă sau să altereze anumite informații esențiale din cv-ul lor, precum anumite competențe dobândite sau experiența în muncă, ceea ce poate duce la o lipsă de informații sau la o neacuratețe a acestora. Exploatarea datelor folosind tehnologia volumelor mari de date presupune și utilizarea anumitor constrângeri sau modele de eroare. Din păcate însă, pentru

anumite domenii, ca de exemplu recrutarea inteligentă a candidaților, nu există studii de specialitate care să precizeze aceste constrângeri sau modele de eroare, necunoscându-se astfel acuratețea rezultatelor în urma procesării și analizării datelor.

Integrarea, agregarea și reprezentarea datelor

Având în vedere eterogenitatea volumului de date, nu este suficientă doar înregistrarea și stocarea datelor într-un depozit. Trebuie luat în considerare, de exemplu, cazul datelor dintr-o serie de experimente științifice. Dacă în depozit se găsesc doar o grămadă de seturi de date, este puțin probabil ca cineva să poată vreodată să găsească, să reutilizeze singur, oricare dintre aceste date.

Același aspect trebuie avut în vedere și în cazul recrutării inteligente a candidaților. În cazul în care compania are un post disponibil, pentru care a avut deja loc în trecut o recrutare inteligentă, este inutilă reparcurgerea întregului volum de date. Astfel, folosind metadata adecvate, se pot regăsi mai ușor candidații selectați ca potențiali la recrutarea precedentă. Utilizarea acestor metadata, ridică anumite provocări din cauza diferențelor din structura de înregistrare a datelor.

Analiza datelor este mult mai dificilă decât localizarea, identificarea, înțelegerea, și citirea datelor. Pentru o analiză eficientă pe scară largă, aceasta trebuie să fie realizată într-un mod complet automatizat. Acest lucru necesită anumite modificări în structura de date și o semantică care să fie exprimată într-o formă care să fie ușor de înțeles de calculator, iar apoi rezolvabilă automat. Există un volum ridicat de lucrări de cercetare în procesul de integrare a datelor, care pot oferi unele răspunsuri. Cu toate acestea, este nevoie de lucrări suplimentare considerabile pentru a obține o rezoluție automată fără eroare.

Chiar și pentru o analiză simplă, care depinde de un singur set de date, proiectarea unei baze de date adecvate rămâne o chestiune importantă. De obicei, există multe metode alternative în care aceleași informații pot fi păstrate. Anumite modele vor avea avantaje față de altele pentru anumite scopuri, și, eventual, dezavantaje pentru alte scopuri. Proiectarea bazelor de date este astăzi o artă, și este atent executată de profesioniști. Trebuie create modele de baze de date

eficiente, fie prin elaborarea unor instrumente care să ajute în procesul de proiectare sau prin renunțarea completă la procesul de proiectare și dezvoltarea de tehnici, astfel încât bazele de date să poată fi utilizate în mod eficient în lipsa unei proiectari inteligente.

Interogarea, modelarea și analiza datelor

Metodele pentru interogarea și exploatare volumelor mari de date sunt fundamental diferite de analiza statistică tradițională pe eșantioane mici. Tehnologia volumelor mari de date este uneori zgomotoasă, dinamică, eterogenă, inter-conectată și nedemnă de încredere. Cu toate acestea, chiar zgomotul produs de tehnologia volumelor mari de date ar putea fi mai valoros decât probele mici, deoarece statisticile generale obținute din modelele și analizele frecvente, înving, de obicei, fluctuațiile individuale și de multe ori dezvăluie modele ascunse mai fiabile și cunoștințe valoroase. Mai mult, interconectivitatea volumelor mari de date creează rețele de informații eterogene mari, cu informații redundante ce pot fi explorate pentru a compensa lipsa de date, pentru a valida relațiile valoroase, și pentru a descoperi relații ascunse și modele.

Tehnicile de extragere a informațiilor din date necesită date accesibile, curățate, demne de încredere, interfețe eficiente de interogare, algoritmi de extragere scalabili, și medii capabile să proceseze volumele mari de date. În același timp, extragerea cunoștințelor din date poate fi de asemenea folosită pentru a ajuta la îmbunătățirea calității și credibilității datelor, oferind funcții inteligente de interogare. După cum a fost menționat anterior, cv-ul încărcat de un candidat pe un portal de recrutare poate să nu corespundă realității din viața reală, deoarece datele au erori, sunt eterogene, și adesea sunt distribuite pe mai multe sisteme.

Valoarea de analiză a volumelor mari de date în domeniul recrutării inteligente a persoanelor, poate fi realizată numai în cazul în care aceasta poate fi aplicată cu fermitate chiar și în aceste condiții dificile. De exemplu, un candidat care aplică pentru un post de referent de specialitate în domeniul informatic poate nota în cv că deține experiența BMS (Building Management System – Sistem de management al clădirii), în timp ce un alt candidat, care aplică pe același post poate nota în cv că deține experiența BMS (Business Management System – Sisteme de management al afacerii). Cu toate că aceste două exprimări au aceeași abreviație, sensul lor diferă complet. Dacă primul candidat deține experiență în domeniul de activitate

aferent postului pe care încearcă să îl ocupe, cel de al doilea, are o experiență complementară într-un alt domeniu. Astfel, pe baza celorlalte date din curriculum vitae, sistemul inteligent ar trebui să corecteze neglijența candidaților și să determine la care termen se făcea referire.

Tehnologia volumelor mari de date trebuie să permită, de asemenea, implementarea unei noi generații interactive de analiză a datelor cu răspunsuri în timp real. În viitor, interogările volumelor mari de date vor fi generate automat pentru crearea de conținut pe site-uri, pentru a popula anumite liste sau recomandări, precum și pentru a oferi o analiză ad-hoc a unui set de date pentru a decide valoarea și necesitatea de stocare a acestuia.

O problemă curentă a analizei bazată pe tehnica volumelor mari de date o reprezintă lipsa de coordonare între sistemele de baze de date, care găzduiesc datele și care oferă interogările SQL, utilizând pachete de analiză care efectuează diverse forme de prelucrare non-SQL, cum ar fi extragerea cunoștințelor din date și analiza statistică. Analistii sunt împiedicați de procesul lent de export al datelor din baza de date, aplicarea unui proces non-SQL și returnarea datelor. Acesta este un obstacol în implementarea tranziției de prima generație de sisteme OLAP conduse SQL, la tipul de extragere a cunoștințelor din date bazat pe o analiză, care este în continuă creștere.

Interpretarea rezultatelor

În ciuda capacității ridicate de analiză oferită de utilizarea tehnologiei volumelor mari de date, valoarea acesteia este limitată în cazul în care utilizatorii nu pot înțelege rezultatele oferite de analiză. În cele din urmă, un factor de decizie, care să dispună și de rezultatul analizei, trebuie să interpreteze aceste rezultate. În general, luarea deciziei presupune examinarea tuturor ipotezelor făcute și reconstituirea analizei. Mai mult, așa cum am arătat în capitolele precedente, există multe surse posibile de eroare: sistemele informatice pot avea erori (*bug-uri*), modelele au aproape întotdeauna la bază ipoteze (și nu fapte concrete), iar rezultatele pot fi bazate pe date eronate.

Din toate aceste motive, nici un utilizator responsabil nu va ceda factorul decizional sistemului informatic. Mai degrabă va încerca să înțeleagă, și să verifice rezultatele obținute de calculator. Sistemul informatic trebuie să ușureze utilizatorilor îndeplinirea acestor sarcini. Acest

lucru reprezintă o provocare deosebită pentru tehnologia volumelor mari de date, datorită complexității sale. Există de multe ori presupuneri cruciale în spatele datelor înregistrate. Anumite analize pot implica mai multe etape, din nou, fiecare etapă construită pornind de la o nouă presupunere.

Pe scurt, uneori nu este suficient a oferi doar rezultatele. Mai degrabă, trebuie furnizate informații suplimentare care să explice modul în care a fost derivat fiecare rezultat, și care au fost presupunerile inițiale. Astfel de informații suplimentare se referă la proveniența datelor rezultat.

Prin studierea celui mai bun mod de a captura, stoca, și interogara proveniența datelor (metadatele), și utilizând tehnici de a capta metadatele adecvate, se poate crea o infrastructură care să ofere utilizatorilor posibilitatea, atât de a interpreta rezultatele analitice obținute, cât și de a repeta analiza modificând anumite ipoteze, parametri, sau seturi de date.

Sistemele cu o paletă bogată de vizualizări au devenit importante în transmiterea utilizatorilor rezultatele interogărilor, într-un mod care să fie mai ușor de înțeles. Întrucât utilizatorii sistemelor timpurii bazate pe inteligența afacerii s-au mulțumit cu prezentări tabelare, analiștii de astăzi trebuie să împacheteze și să prezinte rezultatele în vizualizări puternice care să faciliteze interpretarea, și colaborarea utilizatorilor.

Mai mult decât atât, cu câteva clicuri, utilizatorii ar trebui să poată detalia fiecare bucată de date, pentru a vedea și înțelege proveniența ei, ceea ce reprezintă un element cheie pentru înțelegerea datelor. Utilizatorii nu trebuie doar să vadă rezultatele, ci trebuie să le și înțeleagă, de aceea, este foarte important să le fie explicat de ce văd acele rezultate. Cu toate acestea, proveniență exactă, în special cu privire la fazele etapelor de analiză, este posibil să fie prea tehnică pentru mulți utilizatori, și imposibil de înțeles complet. O alternativă este de a permite utilizatorilor să interacționeze cu pașii din analiză, prin posibilitatea efectuării unor mici schimbări de parametrii. Utilizatorii ar putea apoi vizualiza rezultatele acestor schimbări incrementale. Prin aceste mijloace, utilizatorii pot dezvolta un sentiment intuitiv pentru analiză și, de asemenea, pot verifica dacă sistemul îndeplinește sarcinile la nivelul așteptat.

II.1.3. Provocările utilizării tehnologiei volumelor mari de date

Având descrise fazele procesului de analiză bazat pe utilizarea tehnologiei volumelor mari de date, se pot detalia provocările comune care stau la baza unora, și, uneori, tuturor, acestor faze. Acestea sunt afișate în figura următoare:

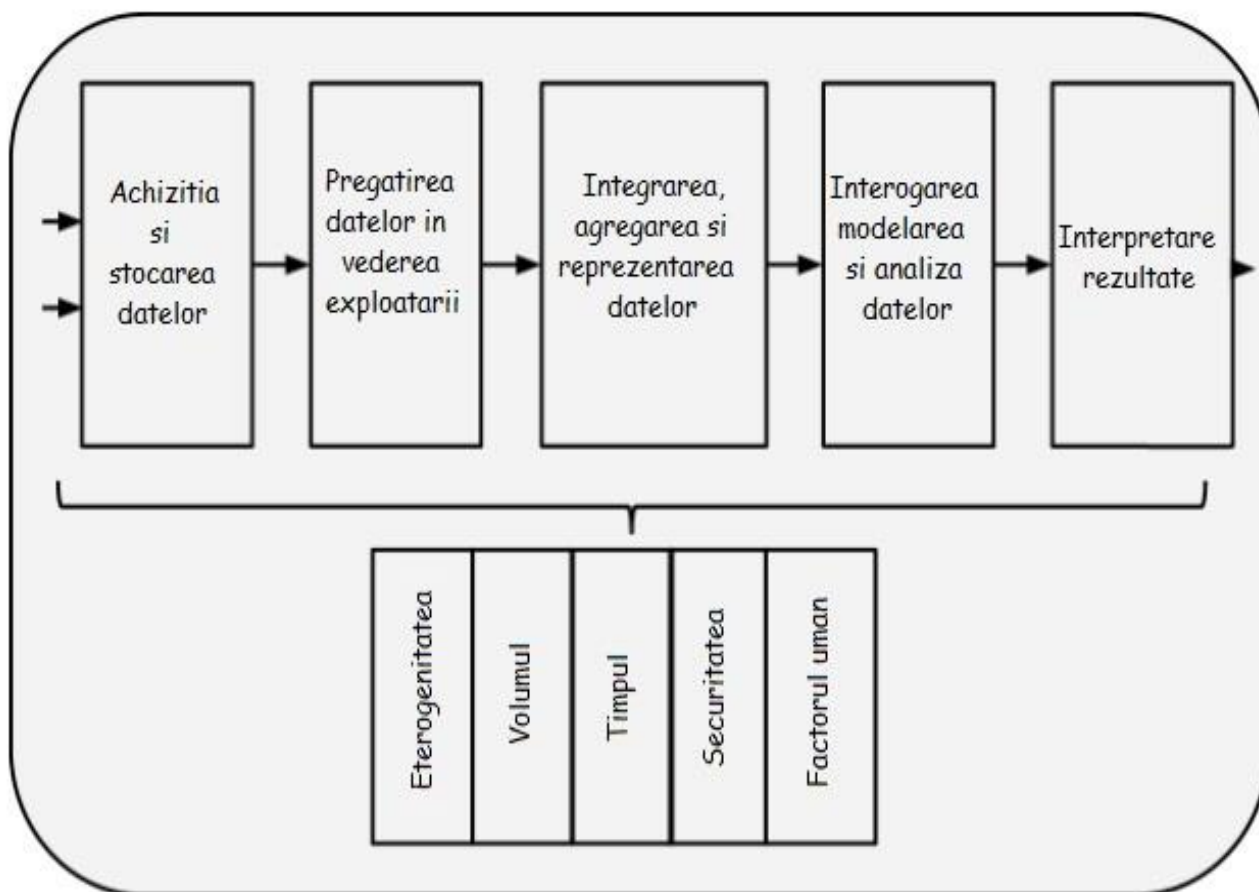


Figura 3. Fazele procesului de analiză și provocările utilizării tehnologiei volumelor mari de date

Când oamenii consumă informații, o mare de eterogenitate este confortabil tolerată. De fapt, nuanța și bogăția limbajului natural poate oferi o adâncime valoroasă. Cu toate acestea, algoritmi de analiză așteaptă date omogene, și nu pot înțelege nuanța. În consecință, datele trebuie să fie atent structurate, acesta fiind primul pas în (sau înainte de) analiza datelor. De exemplu, un candidat care aplică pentru mai multe posturi la o companie. Astfel, se poate crea o înregistrare pentru fiecare aplicare efectuată, o înregistrare pentru toate testele efectuate la o aplicare, sau un record (o clasă) pentru toate interacțiunile candidatului cu compania. În general, numărul de înregistrări asociate unui candidat o să difere în funcție de interacțiunea pe care o are cu compania.

O structură mai mare ar putea fi solicitată de mai multe sisteme de analiză a datelor (sisteme tradiționale). Cu toate acestea, designul mai puțin structurat este probabil să fie mai eficient în mai multe scopuri - de exemplu, aspecte legate de progresia în timp a angajatului la locul de muncă necesită o operație mai costisitoare decât în cazul celorlalte modele, dar acest lucru poate fi evitat. Cu toate acestea, sistemele de analiză funcționează cel mai eficient în cazul în care pot stoca mai multe elemente care să fie toate identice ca mărime și structură. Reprezentarea eficientă, accesul, precum și analiza datelor semi-structurate necesită eforturi suplimentare.

Dacă luăm în considerare informațiile furnizate în general de candidați atunci când aplică pentru un job, ar trebui să ne așteptăm să putem popula și câmpurile pentru data nașterii, ocupație, precum și tipul de experiență pentru fiecare candidat. În acest caz, trebuie luată în considerare și situația în care datele despre candidat sunt insuficiente pentru a popula câmpurile. Evident, că se poate păstra în baza de date înregistrarea candidatului respectiv, dar cu valorile atributelor corespunzătoare stabilite implicit la NULL, dar acest lucru poate influența modul în care se efectuează analiza candidaților. Astfel, în momentul în care se desfășoară o căutare a unor candidați și se aplică mai multe filtre, trebuie avut în vedere și faptul că anumiți candidați ar putea îndeplini criteriile de selecție, dar necompletând în întregime anumite formulare, vor fi excluși în urma analizei.

Chiar și după curățarea datelor și corectarea erorilor, un anumit grad de incompletitudine și unele erori în date sunt susceptibile de a rămâne. Această incompletitudine și aceste erori trebuie să fie gestionate în timpul analizei datelor. Realizarea acestui lucru în mod corect este o provocare. Gestionarea volumelor mari de date și creșterea rapidă a acestora a fost o problemă dificilă în ultimii ani. În trecut, această provocare a fost atenuată de puterea de procesare a noilor procesoare, care furnizează resursele necesare pentru a face față creșterii volumelor de date.

Volumul datelor

Desigur, primul lucru pe care se gândește oricine atunci când vine vorba despre volumele mari de date este dimensiunea. La urma urmei, cuvântul "mare" este acolo chiar în numele tehnologiei. Gestionarea volumelor de date mari și aflate în creștere rapidă a fost o problemă dificilă în ultimii ani. În trecut, această provocare a fost atenuată de îmbunătățirea puterii de procesare care a furnizat resursele necesare pentru a face față creșterii volumului de date. Dar, există o schimbare fundamentală privind această abordare în momentul de față: cantitatea de date generată crește exponențial mai rapid decât se îmbunătățesc resursele de procesare ale sistemului.

În primul rând, în ultimii ani, a avut loc o schimbare în tehnologia procesoare, din cauza constrângerilor de putere, vitezele de ceas au stagnat în mare măsură, iar procesoarele sunt construite cu număr tot mai mare de nuclee. În trecut, sistemele mari de prelucrare a datelor trebuiau să își facă griji cu privire la paralelismul peste noduri, în timp ce acum trebuie să se ocupe de paralelism într-un singur nod. Din păcate, tehnicile paralele de prelucrare a datelor care au fost aplicate în trecut pentru prelucrarea datelor în toate nodurile nu se aplică în mod direct și pentru paralelismul intra-nod, deoarece arhitectura arată foarte diferit; de exemplu, sunt mult mai multe resurse hardware, cum ar fi cache-ul și canalele de memorie ale procesorului care sunt partajate pe nuclee într-un singur nod. În plus, trecerea la sistemele cu putere superioară de procesare (fiecare cu un număr de 10 nuclee) adaugă un alt nivel de complexitate pentru paralelismul intra-nod.

A doua schimbare dramatică, care este în curs de desfășurare o reprezintă trecerea la tehnologia *cloud (cloud computing)*, care înglobează acum mai multe sarcini de lucru cu diferite obiective de performanță.

Acest nivel de partajare al resurselor necesită noi modalități de a determina modul de rulare și execuție al sarcinilor privind procesarea datelor, astfel încât obiectivele de eficiență ale fiecărui volum de lucru să fie îndeplinite, precum și rezolvabilitatea unor erori de sistem, care apar mai frecvent datorită desfășurării activității pe grupuri tot mai mari (care sunt necesare pentru a face față cu creșterea rapidă a volumului de date). Acest lucru presupune o abordare specială pentru dezvoltarea programelor, mai ales a celor care execută sarcini complexe, din moment ce o uniformizare globală a multiplelor platforme folosite de utilizatori este necesară pentru a atinge performanțe cât mai bune.

O a treia schimbare importantă, care este în curs de desfășurare o reprezintă transformarea subsistemului tradițional de intrare / ieșire (I/O). Pentru mai multe decenii, s-au folosit unități HDD pentru a stoca datele persistente. HDD-urile au avut o performanță de intrare / ieșire aleatorie mult mai scăzută decât performanța secvențială de intrare / ieșire, și astfel, motoarele de prelucrare a datelor își formatau datele și proiectau metodele de prelucrare și interogare astfel încât să compenseze această limitare. Dar, la momentul actual, HDD-urile sunt din ce în ce mai des înlocuite cu SSD-uri, și alte tehnologii, cum ar fi etapa de schimb a memoriei (Phase Memory Change) sunt utilizate tot mai des. Aceste tehnologii noi de stocare nu au aceeași răspândire și nici o performanță atât de mare comparativ cu metodele clasice, ceea ce impune o regândire a modului de concepere al subsistemelor de stocare pentru sisteme de prelucrare a datelor. Implicațiile acestui subsistem de stocare au potențialul de a schimba fiecare aspect al prelucrării datelor, inclusiv algoritmi de procesare a interogărilor, programarea interogărilor, proiectarea bazelor de date, metodele de control concurente și metodele de recuperare.

Reversul dimensiunii îl reprezintă viteza. Cu cât setul de date care trebuie să fie prelucrat este mai mare, cu atât mai mult va dura până se va analiza. Proiectarea unui sistem care să rezolve în mod eficient problema dimensiunii volumului de date probabil că o să conducă, de asemenea, și la realizarea unui sistem care să poată procesa o dimensiune specificată mai rapid.

Există multe situații în care rezultatul analizei este necesar imediat. De exemplu, dacă se suspectează că a avut loc o tranzacție frauduloasă cu un card de credit, ar trebui să fie în mod ideal marcată înainte ca tranzacția să fie finalizată – existând astfel posibilitatea de a preveni producerea tranzacției. Evident, o analiză completă a unui utilizator nu este posibilă în timp real. Mai degrabă, trebuie dezvoltate rezultate parțiale în avans, astfel încât o cantitate mică de calcul incremental cu noi date să poată fi folosită pentru a se ajunge la o determinare rapidă.

Având la dispoziție un set mare de date, este adesea necesar pentru a găsi în el elementele care îndeplinesc un anumit criteriu specificat. În cursul procesului de analiză a datelor, acest tip de căutare este probabil să apară în mod repetat. Scanarea întregul set de date pentru a găsi elemente adecvate este, evident, imposibilă. Mai degrabă, sunt create în prealabil structuri de indexare pentru a permite identificarea rapidă a elementelor corespunzătoare. Problema este că fiecare structură index este concepută pentru a sprijini doar unele clase de criterii. Cu fiecare nouă analiză dorită folosind tehnologia volumelor mari de date, există noi tipuri de criterii specificate, precum și necesitatea de a elabora noi structuri index pentru a sprijini astfel de criterii. De exemplu, în cazul în care se recrutează personal pentru posturi diferite, este necesară aplicarea unui set diferit de criterii de selecție. Sistemul are nevoie să precizeze candidatul ideal pentru fiecare post, și să sugereze alternative, iar acest lucru necesită evaluarea mai multor interogări. Astfel, sunt necesare structuri index noi pentru a sprijini astfel de cereri. Proiectarea unor astfel de structuri devine deosebit de dificilă în cazul în care volumul de date este în creștere rapidă și interogările au limite strânse de timp de răspuns.

Confidențialitatea

Confidențialitatea datelor este o altă preocupare foarte mare, și care este în creștere în contextul volumelor mari de date. Pentru dosarelor medicale electronice, există legi stricte care reglementează ceea ce se poate și ce nu se poate face public. Pentru alte date, reglementările, în special în Statele Unite, sunt mai puțin stricte. Cu toate acestea, există o mare frică a publicului cu privire la utilizarea necorespunzătoare a datelor cu caracter personal, în special prin corelarea datelor din mai multe surse. Gestionarea confidențialității este atât o problema tehnică, cât și o problemă sociologică, care trebuie să fie abordată în comun din ambele perspective pentru a atinge întregul potențial promis de utilizare a tehnologiei volumelor mari de date.

Există anumite persoane care nu doresc publicarea anumitor informații referitoare la viața lor privată, precum informațiile legate de sănătate (diagnostice, tratamente efectuate, etc), preferințele religioase (apartenența la o minoritate religioasă), alte preferințe (restaurante frecventate, parcuri, mall-uri, centre de relaxare, etc.). Anumite informații pot fi determinate prin analiza anumitor parametrii. De exemplu, prin analiza datelor referitoare la poziția geografică (*gps tracking*) se pot determina anumite preferințe ale unei persoane. De asemenea, analizând mai mulți parametri ai unei persoane anonime, se poate încerca descoperirea identității acesteia. Barabasi a arătat în lucrarea sa [GON08] că există o stransă legătură între identitățile persoanelor și modul de deplasare al acestora³⁶.

Există multe alte probleme de cercetare care ridică provocări. De exemplu, nu știe încă cum să se realizeze un schimb privat de date, limitând în același timp scurgerea de informații și asigurând utilizatorul de existența suficientor date în datele partajate. Paradigma existentă de confidențialitate diferențială reprezintă un pas foarte important în direcția bună, dar, din păcate, se reduce prea mult conținutul de informații, pentru a mai fi utile în unele cazuri extrem de practice. În plus, datele reale nu sunt statice, ci își măresc volumul și se modifică în timp.

Cu toate acestea, o altă direcție foarte importantă este regândirea securității pentru cazurile care utilizează un schimb de informații de mare volum. Multe servicii online de astăzi

³⁶ [GON08] Understanding individual human mobility patterns. Marta C. González, César A. Hidalgo, and Albert-László Barabási. *Nature* 453, 779-782 - 5 June 2008

obligă utilizatorii să realizeze schimburi de informații private (prin cererile de conectare la Facebook), dar dincolo de controlul accesului la acest nivel record, nu toți utilizatorii înțeleg ce înseamnă a partaja date, modul în care datele partajate pot fi legate, și cum sunt oferite utilizatorilor drepturi de control asupra acestor informații partajate.

Utilizatorii

În ciuda progreselor enorme realizate în analiza asistată de calculator, rămân multe modele pe care oamenii le pot detecta cu ușurință, și pe care algoritmi de calculator le detectează cu dificultate. Într-adevăr, CAPTCHA, reprezintă cel mai bun exemplu de acest tip, și se utilizează tocmai pentru a diferenția utilizatorii web umani de programele software. În mod ideal, analiza volumelor mari de date nu va fi în totalitate realizată de calculator - mai degrabă va fi proiectată în mod explicit pentru a avea un om în buclă. Noul sub-domeniu al analizei vizuale este exact încercarea de a face acest lucru posibil, cel puțin în ceea ce privește faza de modelare și analiză în curs de pregătire.

În lumea tot mai complexă de astăzi, de multe ori este nevoie de mai mulți experți din diferite domenii pentru a înțelege cu adevărat ce se întâmplă. Un sistem de analiză a volumelor mari de date trebuie să sprijine contribuția de la mai mulți experți umani, precum și explorarea în comun a rezultatelor. Acești experți pot fi separați în spațiu și timp, atunci când asamblarea unei întregi echipe, într-o singură cameră, este prea scumpă. Sistemul de date trebuie să accepte această intrare distribuită a expertului, și să sprijine colaborarea lor.

II.1.4. Depozite de date

Un depozit de date este, de obicei modelat printr-o structură de date multidimensională, numită cub de date, în care fiecare dimensiune corespunde unui atribut sau unui set de atribute în schemă, și fiecare set de celule corespunde valorii unor măsuri agregate, cum ar fi numărul de elemente (*count*) sau suma vânzărilor (*sum*). Un cub de date oferă o vedere multidimensională a datelor și permite precompilarea și accesul rapid la date rezumate.

Arhitectura tipică a unui depozit de date este prezentată în figura următoare. Așa cum se poate observa, datele provenite din surse multiple, sunt prelucrate în prealabil pentru a putea fi stocate în depozitul de date, pentru ca ulterior să se poată interoga, astfel încât, beneficiarul final să obțină informații utile.

Sistemele de depozite de date utilizează instrumente și funcții pentru popularea și reîmprospătarea datelor. Aceste instrumente și utilitare includ următoarele funcții:

- Extragerea datelor, reunește în mod obișnuit date din surse multiple, eterogene, și externe.
- Curățarea datelor, presupune detectarea erorilor în date și rectificarea acestora, atunci când este posibil.
- Transformarea datelor, este procesul care transformă datele din formatul moștenit sau gazdă, în formatul depozitului.
- Încărcarea, presupune sortare, rezumare, consolidare, calcularea punctelor de vedere, verificarea integrității, și crearea indicșilor și a partițiilor.
- Împrospătarea, presupune propagarea actualizărilor de la sursele de date la depozit.

În afară de curățare, de încărcare, împrospătare, și instrumentele de definire a metadatelor, sistemele de depozitare a datelor oferă, de obicei, un set bun de instrumente de management al depozitelor de date.

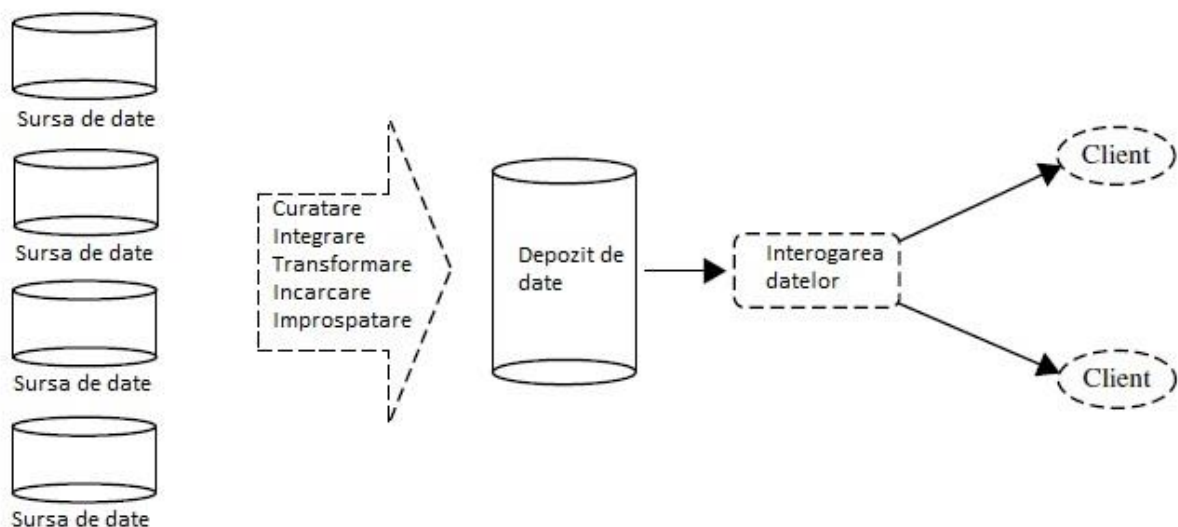


Figura 4. Infrastructura tipică a unui depozit de date

Prin furnizarea de vizualizări multidimensionale de date și precompilarea datelor rezumate, sistemele de depozite de date pot oferi sprijin inerent pentru OLAP. Operațiunile OLAP fac uz de cunoștințe generale cu privire la domeniul datelor care urmează a fi studiate pentru a permite prezentarea datelor la diferite niveluri de abstractizare. Astfel de operațiuni pot oferi utilizatorilor puncte de vedere diferite. Exemple de operațiuni OLAP includ *drill-down* (spre profunzime) și *roll-up* (pornind din profunzime, spre general), care permit utilizatorului vizualizarea datelor de la diferite grade de sumarizare. De exemplu, putem detalia în profunzime, pe datele de vânzări rezumate pe trimestru pentru a vedea datele rezumate pe lună. În mod similar, putem rula pe date de vânzări rezumate după oraș pentru a vedea datele rezumate în funcție de țară.

Deși depozitele de date sprijină analiza datelor, instrumente suplimentare pentru extragerea cunoștințelor din date sunt deseori necesare pentru analiza în profunzime. Extragerea multidimensională (de asemenea, numită explorarea cunoștințelor multidimensionale) efectuează un minerit al datelor (*data mining*) în spațiu multidimensional, într-un stil OLAP. Acest lucru, permite explorarea mai multor combinații de dimensiuni la diferite niveluri de granularitate în mineritul datelor, și are, astfel, un potențial mai mare de a descoperi modele interesante care reprezintă cunoștințe.

Din punct de vedere al arhitecturii, există trei modele de depozitare a datelor: depozit întreprindere (*enterprise*), depozit specific (*data mart*), și depozitul virtual.

Depozit întreprindere (*enterprise*): un depozit întreprindere colectează toate informațiile despre subiectele care acoperă întreaga organizație. Acesta oferă integrare de date la nivelul întregii companii, de obicei, de la unul sau mai multe sisteme de operare sau de la furnizorii de informații externe, și este multifuncțional, fiind aplicabil în multiple domenii. Acesta conține de obicei informații detaliate, precum și date rezumate, și poate varia în mărime de la câțiva MB, TB, sau chiar mai mult. Un depozit de date întreprindere poate fi pus în aplicare pe infrastructuri tradiționale, folosind superservere, sau platforme cu arhitecturi paralele. Acesta necesită o modelare extinsă a afacerii și poate dura ani pentru a putea fi proiectat și construit.

Depozit specific (*data mart*): Un astfel de depozit conține un subset de date la nivelul întregii companii, care este de valoare doar unui anumit grup de utilizatori. Domeniul de aplicare este limitat la subiecte specifice selectate. De exemplu, un depozit specific de marketing poate limita informațiile stocate la nivel de client, tranzacții, sau vânzări. Datele conținute în depozitele specifice tind să fie sumarizate. Depozitele specifice sunt de obicei puse în aplicare pe servere departamentale de buget redus (*low-cost*), care sunt Unix / Linux sau Windows. Ciclul de implementare a unui astfel de depozit este mult mai probabil să fie măsurat în săptămâni, mai degrabă decât luni sau ani. Cu toate acestea, se poate implica integrarea complexă pe termen lung în cazul în care proiectarea și planificarea sa nu au fost la nivel de întreprindere. În funcție de sursa de date, depozitele specifice pot fi clasificate ca independente sau dependente. Depozitele independente sunt obținute din datele capturate de la unul sau mai multe sisteme de operare sau a furnizorilor de informații externe, sau de la datele generate local într-un anumit departament sau zonă geografică, în timp ce datele celor dependente sunt obținute direct de la depozitele de date de întreprindere.

Depozit virtual: Un depozit virtual este un set de vederi pentru bazele de date operaționale. Pentru prelucrarea eficientă a interogărilor, doar o parte din punctele de vedere posibile sumarizate, pot fi materializate. Un depozit virtual este ușor de construit, dar necesită capacități excedentare pe serverele de baze de date operaționale.

Dezvoltarea de sus în jos (*top-down*) a unui depozit întreprindere servește ca o soluție sistematică și reduce problemele de integrare. Cu toate acestea, este scump, necesită o lungă

perioadă de timp pentru dezvoltare, și nu are flexibilitate datorită dificultății în atingerea consistenței și a consensului pentru un model de date comun pentru întreaga organizație.

Abordarea de jos în sus (*bottom-up*) la proiectarea, dezvoltarea și implementarea depozitelor de date independente oferă flexibilitate, costuri reduse, precum și întoarcerea rapidă a investițiilor. Cu toate acestea, acest tip de depozite pot produce probleme atunci când se încearcă integrarea diferitelor depozite independente într-un depozit de date de întreprindere consistent.

II.1.5. Baze de date NoSQL

În trecut, companiile au folosit bazele de date relaționale pentru a stoca datele lor structurate. În momentul de față, în ciuda impactului enorm asupra lumii bazelor de date și a deblocării datelor pentru multe aplicații, bazelor de date relationale le lipsesc caracteristicile necesare pentru a face față tranzacțiilor rapide de date în epoca volumelor mari de mare. Bazele de date NoSQL sunt răspunsul care rezolvă multe din aceste probleme, deoarece ele oferă o nouă perspectivă asupra lumii bazelor de date.

Deși nu au schemă, bazele de date NoSQL sunt rapide și se adaptează cu ușurință la nevoile companiilor. De exemplu, NoSQL poate lucra cu date nerelaționale distribuite și date nestructurate, acestea fiind tipul de date pe care cele mai multe companii îl generează în mod obișnuit [Rij14].³⁷

Riak, Apache HBase, MongoDB și New4J sunt doar câteva exemple de baze de date NoSQL. Este deosebit de important atunci când se creează o strategie de afaceri, să se înțeleagă capacitățile și constrângerile fiecărui tip de baze de date, în scopul de a alege pe cel mai potrivit pentru îndeplinirea obiectivelor.

³⁷ [RIJ14] Mark van Rijmenam - “How the Next Generation of Databases Could Solve Your Problems”, Dataflop, 18 Octombrie 2014 <https://dataflop.com/read/generation-database-solve-problems/139>

Bazele de date cheie / valoare, nu se concentrează pe structura de date, ci pe capacitatea de a stoca și prelua datele. Ca urmare, interogările sunt mai eficiente și mai rapid de implementat, asigurându-și aplicații scalabile și rapide care ar trebui să citească și să scrie o mare varietate de date în această eră a volumelor mari de date.

Această bază de date de tip cheie / valoare permite clienților să citească și să scrie valorile folosind o cheie, după cum urmează [PGLB]³⁸: Get(key) – returnează valoarea asociată respectivei chei, Put(key, value) – asociază valoarea precizată respectivei chei, Multi-get(key1, key2,..., keyN) – returnează lista de valori asociate listei de chei și Delete(key) – elimină înregistrarea cheii respective din dicționarul de date.

Riak este o bază de date cheie / valoare distribuită în care valorile pot fi orice - de la text simplu, JSON sau XML, la imagini sau clipuri video - toate accesibile printr-o interfață simplă HTTP. Astfel, putem afirma că, baza de date Riak poate stoca date indiferent de tipul acestora.

Riak este, de asemenea, tolerant la defecțiuni (erori). Acest lucru înseamnă că serverele pot fluctua în orice moment, fără a da nici o notificare sau vreun motiv pentru care au eșuat. Cluster-ul continuă să funcționeze în timp ce alte servere sunt adăugate, eliminate, sau picate (defecte). Unul dintre marile avantaje este că, folosind Riak nu trebuie avute griji cu privire la grup (cluster), deoarece, chiar dacă, un nod a cedat, aceasta nu este o urgență și nu este nevoie ca problema să fie rezolvată imediat.

Cu toate acestea, această flexibilitate are și dezavantaje. Riak este proiectat astfel încât să nu ofere suport solid pentru interogări ad-hoc, și depozite cheie-valoare, având probleme în ceea ce privește conectarea valorilor (cu alte cuvinte, nu are chei externe).

Riak este o alegere excelentă pentru centrele de date, cum ar fi Amazon, care trebuie să servească multe cereri cu latență scăzută. Dacă fiecare milisecundă petrecută în așteptare poate

³⁸ [PGLB] 3 Pillar Global, Exploring the Different Types of NoSQL Databases Part II <http://www.3pillarglobal.com/insights/exploring-the-different-types-of-nosql-databases>

reprezenta un potențial client pierdut, Riak este greu de învins. Este ușor de gestionat, ușor de configurat, și poate crește cu nevoile clientului. Dacă se utilizează Amazon Web Services, cum ar fi SimpleDB sau S3, se pot observa unele asemănări în formă și funcție. Aceasta nu este o coincidență, Riak fiind inspirat de proiectul Dynamo Amazon [GDC]³⁹.

Riak permite utilizatorului să controleze citirile și scrierile în cluster prin modificarea a trei valori: N, W, și R. N este numărul de noduri peste care scrierea se poate replica în cele din urmă, cu alte cuvinte, numărul de copii din cluster. W este numărul de noduri care trebuie să fie scrise cu succes înainte de a considera un răspuns ca fiind pozitiv. Dacă W este mai mic decât N, o scriere va fi considerată de succes, chiar dacă în acest timp Riak încă copiază valoarea. În cele din urmă, R este numărul de noduri necesare pentru a citi cu succes o valoare. Dacă R este mai mare decât numărul de copii disponibile, cererea va eșua.

Există o mare diferență între bazele de date relaționale și Riak. Absența tranzacțiilor, a limbajului SQL și a schemelor conduce la dificultăți în înțelegerea și utilizarea bazei de date Riak. Deși există chei, procedura de legătură între ele nu este asemănătoare unei joncțiuni. Cu toate acestea, unele probleme pot fi mai bine rezolvate cu ajutorul Riak, deoarece acesta are capacitatea de a se adapta la cerințele în creștere ale serverelor, în termeni de performanță, mai degrabă decât creșterea în dimensiune pentru serverele individuale mari, care, astfel, ajută la rezolvarea problemelor de scalabilitate Web . De asemenea, Riak transmite date bidirecțional atunci când se analizează structura HTTP, ceea ce permite o flexibilitate maximă pentru orice *framework* sau sistem web.

Puncte forte

Unul dintre punctele forte, se referă la faptul că elimină posibilitatea de eșec sprijinind un timp de funcționare (*uptime*) maxim adaptându-se pentru a satisface diferitele cerințe. Nu contează dacă datele sunt complexe sau nu, Riak poate stoca date simple, dar permite și introducerea informațiilor mai sofisticate, dacă este necesar. Există multe biblioteci client pentru

³⁹ [GDC] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Voss, Werner Vogels - Dynamo: Amazon's Highly Available Key-value Store, Amazon.com, 2007

Riak, inclusiv Java, Python, Perl, Erlang, Ruby, PHP, .NET, și multe altele [RDD]⁴⁰. În cazul în care este nevoie de o viteză mai mare decât cea suportată de HTTP, comunicarea se poate realiza prin Protobuf, fiind o soluție mai bună, pentru că este un protocol mai eficient de codificare și transport binar.

Puncte slabe

Există mai multe dezavantaje atunci când vine vorba despre Riak. Există anumite funcții pe care acest tip de bază de date nu le poate suporta, de exemplu interogarea simplă, structuri de date complexe, o schemă rigidă sau posibilitatea de a scala orizontal cu servere proprii. Unul dintre dezavantajele majore cu privire la Riak este faptul că acest cadru al interogării a rămas același - ușor și robust. Deși, MapReduce oferă funcționalități puternice, ar fi fost mult indicat să fie disponibile mai multe acțiuni bazate pe URL sau pe bază de alte interogări. În cele din urmă, în cazul în care Erlang nu este limbajul preferat de programare al utilizatorului, există câteva limitări atunci când se utilizează JavaScript, cum ar fi lipsa de *post-commit* sau execuția lentă MapReduce.

Hbase

HBase este un sistem de management al bazelor de date orientate pe coloană, care rulează pe suport HDFS. Este perfect pentru seturi de date rare, care sunt comune în multe cazuri care implică volume mari de date. HBase nu acceptă SQL, care este un limbaj structurat de interogare. Aplicațiile HBase sunt scrise în Java, mai mult ca aplicații tipice MapReduce. HBase suportă dezvoltarea de aplicații în Avro, REST, și Thrift.

Un sistem HBase conține un set de tabele. Fiecare tabelă conține rânduri și coloane, la fel ca o bază de date tradițională. Fiecare tabelă trebuie să aibă un element definit ca o cheie primară, și toate încercările de acces la tabele HBase trebuie să utilizeze această cheie primară. O coloană HBase reprezintă un atribut al unui obiect. De fapt, HBase permite și gruparea mai multor atribute în ceea ce sunt cunoscute ca familii coloană, astfel încât, elementele unei familii

⁴⁰ [RDD] Basho, Riak Distributed Database <http://basho.com/riak/>

coloană sunt toate stocate împreună. Acest lucru este diferit față de o bază de date relațională, unde toate coloanele unui rând dat sunt stocate împreună. În HBase trebuie stabilită anticipat schema tabelii și specificată familia coloană [IHB]⁴¹.

Principalele caracteristici ale HBase sunt [CDH]⁴²: arhitectură scalabilă - adaugă servere pentru creșterea capacității, replicare automată – replică transparent și eficient datele în toate mașinile din cluster, consistență - previne eșecurile nodurilor sau scrierea simultană la aceeași înregistrare, replicare activă – transferă datele în toate locațiile pentru a putea fi recuperate în caz de dezastru și pentru o mai bună protecție, disponibilitate ridicată - noduri principale (*master*) multiple care asigură accesul continuu la date, căutarea - oferă utilizatorilor non-tehnici și aplicațiilor o experiență interactivă familiară și foarte puternică de căutare, securitatea – tabelă de securitate și coloană de acces la nivel de familie, utilizând Kerberos și acces SQL - interogarea interactivă a datelor folosind Cloudera Impala și Apache Hive pentru prelucrarea loturilor.

Puncte forte

Caracteristicile notabile ale HBase includ o arhitectură robustă la scară largă și încorporarea versiunilor și capabilității de compresie. De exemplu, păstrarea istoriei versiunilor paginilor wiki este o caracteristică esențială pentru întreținere. Dacă se lucrează cu cantități mari de date, măsurate în multi GB sau TB, HBase poate fi o bună soluție.

Puncte slabe

Comunitatea Hbase pare să fie de acord că cinci noduri reprezintă numărul minim de noduri dorit de utilizator să folosească. Deoarece este proiectat pentru a fi uriaș, poate fi, de asemenea, mai greu de gestionat. Rezolvarea problemelor mici este un punct slab al HBase.

HBase nu oferă posibilități de sortare sau de indexare, cu excepția cheilor de rând. Rândurile sunt păstrate în ordine sortată pe baza cheilor de rând, dar nici o astfel de sortare nu se face pe orice alt atribut (coloană). Deci, dacă se dorește găsirea unor rânduri pe baza altor atribute, diferite de cheia rândului, este nevoie de scanarea tabelii sau de menținerea propriului

⁴¹[IHB] IBM, Hadoop <http://www-01.ibm.com/software/data/infosphere/hadoop/hbase/>

⁴²[CDH] Cloudera, CDH <http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/hbase.html>

index. Un alt concept care lipsește îl reprezintă tipul de date. Astfel, nu există nici o distincție între, o valoare întreagă, un șir de caractere, și o dată. Toate acestea sunt considerate biți de HBase, rămânând astfel în sarcina dezvoltatorului să interpreteze biții [RED]⁴³.

MongoDB

Când vorbim despre MongoDB, putem afirma că, abilitatea utilizatorului de a finaliza o sarcină este bazată în mare parte pe componentele alese pentru utilizare. Puterea lui MongoDB constă în versatilitate, putere, ușurință de utilizare, precum și capacitatea de a gestiona atât sarcini mari, cât și mici. Lansat public în 2009, MongoDB încă este considerat o stea în ascensiune în lumea NoSQL.

Acesta a fost conceput ca o bază de date scalabilă, numele de Mongo provenind de la enormitate (engl. „humongous”) – având ca obiective principale de proiectare performanța și accesul ușor la date. Este o bază de date de documente, care permite datelor să persiste într-o stare imbricată, și mai important, se pot interoga datele imbricate într-un mod ad-hoc. Nu se impune nici o schemă (similar cu Riak dar, diferit de Postgres), astfel încât documentele pot conține opțional domenii sau tipuri pe care nici un alt document din colecție nu le conține.

Caracteristicile principale ale lui MongoDB se concentrează pe flexibilitate, putere, viteză, și ușurință de utilizare [MDB]⁴⁴.

Flexibilitatea - MongoDB stochează datele în documentele JSON (pe care le serializează - BSON). JSON oferă un model de date bogat care se potrivește perfect limbajelor native de programare.

⁴³ [RED] Eric Redmond, Jim R. Wilson - Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement, 2012

⁴⁴ [MDB] MongoDB, Introduction to MongoDB <http://www.mongodb.org/about/introduction>

Puterea - MongoDB oferă o mulțime de caracteristici, cum ar fi indici secundari, interogări dinamice, sortare, actualizări bogate, *upserts* (actualizare – *update* dacă există documente, inserare - *insert* dacă nu există), și agregarea ușoară.

Viteza / Scalarea - Prin păstrarea datelor legate împreună în documente, interogările pot fi mult mai rapide decât în cazul bazelor de date relaționale unde datele similare sunt separate în mai multe tabele, trebuind să fie alăturate ulterior. De asemenea, MongoDB, face mai ușor procesul de scalare al bazei de date. Prin *autosharding* se permite scalarea liniară a unui *cluster* adăugând mai multe mașini. Este posibilă creșterea capacității, fără nici o întrerupere (*downtime*) a sistemului, lucru care este foarte important pe web, atunci când sarcina poate crește brusc și oprirea site-ului pentru o întreținere extinsă poate costa afacerea sume mari din venituri.

Ușurința de utilizare - MongoDB lucrează din greu pentru a fi foarte ușor de instalat, configurat, întreținut și utilizat. În acest scop, MongoDB oferă puține opțiuni de configurare, și încearcă în schimb să facă în mod automat "ceea ce trebuie", ori de câte ori este posibil. Acest lucru înseamnă că, utilizând MongoDB, dezvoltatorii se pot arunca direct în dezvoltarea de aplicații, în loc de a petrece o mulțime de timp pentru ajustarea configurației bazei de date.

Puncte forte

Puterea principală a lui MongoDB constă în capacitatea sa de a gestiona volume mari de date (și cantități uriașe de cereri) prin replicare și scalare orizontală. De asemenea, are beneficiul suplimentar al unui model de date foarte flexibil.

În cele din urmă, MongoDB fost construit pentru a fi ușor de utilizat. Acest lucru poate fi observat și în asemănarea dintre comenzile Mongo și conceptele bazelor de date SQL (mai puțin partea de joncțiune a serverelor). Acest lucru nu este întâmplător și este un motiv pentru care Mongo câștigă atât de mult din cota utilizatorilor fostului model de obiecte relaționale (ORM). E destul de diferit și ridică suficiente probleme dezvoltatorilor, dar nu atât de diferit, încât să devină de neînțeles și înfricoșător.

Puncte slabe

Cum Mongo încurajează renunțarea la scheme (de a nu avea nici măcar una), acest lucru ar putea reprezenta un concept mult prea diferit pentru utilizatori. Poate fi periculos să se

introducă orice valoare veche, de orice tip, în orice colecție. Un singur tip de date poate provoca multe neplăceri, dacă utilizatorul nu ia în calcul numele câmpului sau numele colecției ca o posibilă cauză. În general, flexibilitatea lui Mongo nu este importantă, mai ales dacă modelul de date este deja destul de matur și închis.

Deoarece Mongo este axat pe volume de date de mari dimensiuni, acesta funcționează cel mai bine în grupuri mari (*clustere*), care pot necesita un efort de proiectare și gestionare. Spre deosebire de Riak, în cazul în care adăugarea de noi noduri este transparentă și relativ nedureroasă pentru operațiuni, crearea unui grup Mongo necesită un pic mai multă chibzuire [RED]⁴⁵.

New4J

Neo4j este un tip nou de bază de date NoSQL, numit bază de date de tip grafic. După cum sugerează și numele, datele stocate arată ca un grafic (în sens matematic). Acest tip de bază de date se concentrează mai mult pe relațiile dintre valori, mai degrabă decât pe elementele comune dintre seturile de valori (cum ar fi colecții de documente sau tabele de rânduri). Ca o chestiune de fapt, datele pot fi stocate într-un mod natural și simplu. Având dimensiuni reduse, este posibil încorporarea Neo4j în aproximativ orice aplicație. Totuși acesta are capacitatea de a stoca zeci de miliarde de noduri și la fel de multe ramificații și cu sprijinul clusterului cu replicare *master-slave* peste multe servere, se poate ocupa de aproape orice problemă indiferent de dimensiunile acesteia.

⁴⁵ [RED] Eric Redmond, Jim R. Wilson - Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement, 2012

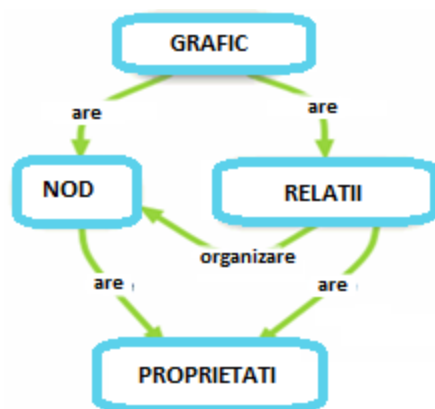


Figura 5. Structura unei baze de date de tip graf

Puncte forte

Neo4j este una dintre cele mai bune exemple de baze de date *open source*, de tip grafic. Bazele de date grafice pot fi considerate o soluție pentru stocarea datelor nestructurate. Chiar dacă Neo4j nu are tipuri de date și nici schemă, constrângerile pe care le pune asupra modului în care datele sunt legate, sunt esențiale. În momentul de față, Neo4j are capacitatea de a sprijini 34,4 miliarde noduri și un număr la fel de mare de relații, ceea ce este suficient pentru cele mai multe utilizări (de exemplu, Neo4j ar putea deține mai mult de 42 de noduri pentru fiecare din cei 800 de milioane de utilizatori ai Facebook, într-un singur grafic [RED]⁴⁶).

Dincolo de ușurința de utilizare, Neo4j este și rapid. Spre deosebire de operațiile din bazele de date relaționale sau operațiunile *map-reduce* din alte baze de date, Neo4j are beneficiul că traversările grafice sunt constante ca timp. Majoritatea bazelor de date realizează, de obicei, joncțiunea valorile în orb și filtrează apoi rezultatele dorite. Bazele de date grafice, spre deosebire de restul, acționează ca și când datele sunt doar la un nod distanță și nu este obligatoriu să se știe cât de mare devine graficul, trecerea de la nodul A la nodul B este întotdeauna văzută ca un singur pas în cazul în care nodurile au o relație.

⁴⁶ [RED] Eric Redmond, Jim R. Wilson - Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement, 2012

Puncte slabe

Există, de asemenea, câteva dezavantaje pentru Neo4j. Unul dintre ele este faptul că, în Neo4j, marginile (ultima frunză) nu pot conduce o ramificație nouă către sine. De asemenea, există o problemă cu nomenclatura pentru că este utilizată noțiunea de nod în loc de frunză, și de relație în loc de arc (legătură), astfel, adăugând o complexitate mai mare în comunicare. Se poate reproduce doar un singur graf complet pe alte servere, chiar dacă Neo4J este excelent la replicare. În cele din urmă, dacă utilizatorul este în căutarea unei licențe gratuite (*open source*), Neo4j nu este alegerea potrivită pentru companie.

II.1.6. Analiza comparativă a bazelor de date NoSQL

În tabelul următor au fost sintetizate diferențele dintre principalele baze de date de tip NoSQL.

Caracteristici	Riak	HBase	MongoDB
Modelul datelor	Stochează perechile cheie/valoare într-un spațiu de nume (<i>namespace</i>) superior numit <i>bucket</i> .	Stochează datele într-un format predefinit al unei familii de coloane. Datele sunt sortate, rarefiate, și grupate fizic pe familii de coloane	Stochează datele în formatul BSON (echivalentul binary al JSON), păstrate ca documente (înregistrări de sine stătătoare, fără legături). Documentele păstrate de MongoDB pot stoca oricare dintre tipurile BSON definite și sunt grupate în colecții.

Modelul de stocare	<p>Riak are un sistem modular, extensibil local, de depozitare care oferă posibilități <i>backend</i> conectabile, concepute pentru a se potrivi unei varietăți de cazuri de utilizare. Magazinul <i>backend</i> implicit al lui Riak este Bitcask.</p>	<p>HDFS (Hadoop Distributed File System) este sistemul de stocare utilizat de HBase. Datele sunt stocate în zone de memorie (<i>MemStores</i>) și în fișiere (<i>StoreFiles</i>), caz în care datele sunt transmise pe disc (implementat prin intermediul HFiles, un format bazat pe <i>SSTable - Bigtable</i>). Implementările folosesc, în general, sistemul JVM nativ de gestiune a fluxului de intrare - ieșire.</p>	<p>Sistemul de stocare implicit utilizat de MongoDB este motorul de stocare al memoriei (<i>Memory - Mapped Storage Engine</i>). Acesta utilizează fișiere de memorie mapate pentru toate operațiile de intrare - ieșire de pe disc. Este responsabilitatea sistemului de operare să gestioneze fluxul datelor către disc și datele de intrare – ieșire.</p>
Interogarea datelor	<p>Există în prezent patru moduri de interogare în Riak:</p> <ul style="list-style-type: none"> • Operații pe baza cheii primare (GET, PUT, DELETE, UPDATE); • MapReduce • Indecși secundari; • Căutarea Riak. 	<p>HBase are două opțiuni de interogare: căutarea valorilor de achiziție / scanare prin cheile ordonate (opțional filtrarea valorilor sau folosirea unui index secundar), sau prin utilizarea Hadoop pentru a efectua MapReduce.</p>	<p>MongoDB are o interfață de interogare care prezintă unele similitudini cu bazele de date relaționale, inclusiv indicii secundari care pot fi derivați din documentele stocate. MongoDB are de asemenea capacitatea de a efectua interogări MapReduce și interogări ad-hoc pe documente. De asemenea, este disponibil și suportul Hadoop.</p>

<p>Concurența</p>	<p>În Riak, orice nod din cluster poate coordona o operație de citire / scriere pentru orice alt nod. Riak subliniază disponibilitatea pentru scriere și citire, și pune povara rezoluției pe client la momentul citirii.</p>	<p>HBase garantează atomicitatea scrierii și focalizează pe rând. De asemenea, HBase a adăugat recent și facilitatea unor tranzacții locale cu acțiuni multiple (<i>multi – action</i>) sau rânduri multiple (<i>multi – row</i>), dar, eliminând posibilitatea de a realiza simultan acțiuni de scriere – citire.</p>	<p>MongoDB prezintă o consistență puternică. Eventuala consistență la citire poate fi realizată prin intermediul secundarelor. Un <i>cluster</i> MongoDB (cu <i>auto-sharding</i> și replicare) are un server master la un moment pentru fiecare <i>shard</i>.</p>
<p>Replicarea</p>	<p>Sistemul de replicare utilizat de Riak este puternic influențat de cartea Dynamo și teorema enunțată de dr. Eric Brewer [GDC].</p> <p>Riak folosește un <i>hashing</i> consistent pentru a reproduce și distribui N copii ale fiecărei valori în jurul unui grup Riak compus din oricâte mașini fizice. De asemenea, Riak folosește noduri virtuale să se ocupe de distribuirea și reechilibrarea dinamică a datelor, astfel decuplând distribuția datelor din active fizice.</p> <p>API-ul Riak oferă consistență și</p>	<p>HBase susține replicarea în <i>cluster</i> și între <i>cluster</i>e. În cluster - replicarea este manipulată de către HDFS și reproduce fișierele de date care stau la bază în funcție de setările Hadoop. Între cluster - reproduce un eventual eveniment consistent master / slave, sau cel mai recent eveniment (experimental) master / master adăugat și replicare ciclică (în cazul în care fiecare nod joacă rolul de stăpân și supus).</p>	<p>MongoDB se bazează pe blocaje pentru consistență. Începând cu versiunea 2.2, MongoDB are un nivel DB de blocare pentru toate operațiunile.</p>

	acordă parametrii disponibili care să permită selecția celui mai bun nivel de configurare pentru cazul utilizat.		
Scalarea	<p>Riak oferă elasticitate în manipularea dimensiunii cluster-ului, acesta putând fi redimensionat în funcție de necesități. Aceași elasticitate o oferă și în cazul rebalansării sarcinilor de lucru între noduri.</p> <p>În Riak niciun nod nu este special, toate nodurile având aceleași particularități. În momentul adăugării unui nou nod, Riak îl descoperă prin interogarea stării membrilor săi. În momentul descoperirii noului membru, acesta primește un procent egal de sarcini de lucru. În momentul eliminării unui nod, sarcinile de lucru ale acestuia sunt redistribuite egal către celelalte noduri rămase funcționale.</p>	HBase divide regiunile pentru a redistribui creșterea volumului datelor. Un proces eșuat pe o regiune, presupune doar restaurarea regiunii respective. În cazul HBase scalarea se realizează prin intervenția dezvoltatorului sau a administratorului bazei de date.	<p>Mongo utilizează partiționarea orizontală a datelor în momentul scalării. Acest lucru implică utilizarea unui server pentru gestiunea anumitor segmente de date, în momentul creșterii volumului de date.</p> <p>Pentru procesul invers celui de scalare prezentat mai sus, MongoDB oferă facilitatea de eliminare a partițiilor orizontale de date din baza de date.</p>
Replicarea multi-datacenter	<p>Riak dispune de două tipuri distincte de replicare. Utilizatorii pot reproduce până la orice număr de noduri într-un cluster (care este conținut, de</p>	HBase utilizează regiunile pentru a se putea replica la nivelul mai multor centre de date.	MongoDB poate fi configurat pentru a rula peste mai multe centre de date, prin intermediul diferitelor opțiuni.

	<p>obicei, într-un singur centru de date). Pentru utilizarea replicării la nivelul mai multor centre de date este necesară o licențiere comercială. Aceasta permite utilizarea activă a clusterelor într-un număr nelimitat de centre de date.</p>		
<p>Monitorizare interactivă / Consola de administrare</p>	<p>Riak se livrează împreună cu Riak control, o consolă grafică open source pentru monitorizarea și gestionarea clusterelor Riak</p>	<p>HBase are câteva unelte grafice sprijinite de comunitate, și o consolă de administrare de tip linie de comandă.</p>	<p>MongoDB nu este livrat cu o consolă grafică de monitorizare / administrare. Cu toate acestea, mai multe proiecte comunitare au elaborat programe grafice de monitorizare / administrare.</p> <p>10Gen oferă un serviciu de monitorizare găzduit.</p>

Tabelul 1. Tabel comparativ al bazelor de date NoSQL

II.2. Tehnologia extragerii cunoștințelor din date

Domeniul interdisciplinar de *Data Mining* (extragerea cunoștințelor din date) provine din confluența statisticii și a capacității de învățare a mașinii (inteligenta artificială). Acesta reprezintă o tehnologie care ajută la analiza și înțelegerea informațiilor conținute în bazele de date, care a fost folosită într-un număr mare de domenii sau aplicații.

Mai exact, conceptul de extragerea cunoștințelor din date derivă din similitudinea între căutarea de informații valoroase în bazele de date și mineritul mineralelor valoroase într-un munte. Ideea este că materia primă este reprezentată de datele care urmează să fie analizate și pentru care se vor folosi un set de algoritmi care acționează în calitate de excavatoare de învățare pentru a căuta bucați valoroase de informații [BIGUS]⁴⁷.

Termenul de “descoperirea cunoștințelor în sisteme de baze de date” a apărut prima dată în 1989. Prin definiție, descoperirea cunoștințelor în sisteme de baze de date este “un proces de extragere a informațiilor, date anterior necunoscute și potențial utile”, dar și ca “știința de a extrage informații utile din baze de date foarte mari” [FAY]⁴⁸.

Unii autori au arătat că extragerea cunoștințelor din date este formată din "analiza (de multe ori vastă) a unor seturi de date de observație pentru a găsi relații nebănuite și pentru a rezuma datele în moduri noi, care sunt atât de ușor de înțeles și utile pentru proprietarul datelor" [HMS]⁴⁹, sau, mai simplu, "căutarea de informații valoroase în volume mari de date" [WSMI]⁵⁰, sau "descoperirea de structuri interesante, neașteptate sau valoroase în baze de date de mari dimensiuni" [HAND]⁵¹. Alți autori au definit extragerea cunoștințelor din date ca "explorarea și

⁴⁷ [BIGUS] Bigus, J.P. (1996). *Data Mining with neural networks: solving business problems from application development to decision support*. New York: McGraw-Hill

⁴⁸ [FAY] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth – *From Data Mining to Knowledge Discovery in Databases*

⁴⁹ [HMS] Hand, D.J., Mannila, H. & Smith, P. (2001). *Principles of Data Mining*. London: The MIT Press

⁵⁰ [WSMI] Weiss, S.M. & Indurkha, N. (1998). *Predictive Data Mining. A Practical Guide*. San Francisco, CA: Morgan Kaufman

⁵¹ [HAND] Hand, D.J. (2007). Principles of Data Mining. *Drug Safety*, 30, 7, 621-622

analiza unor cantități mari de date, în scopul de a descoperi modele semnificative și reguli" [BELI]⁵².

Aceste definiții arată clar că extragerea cunoștințelor din date este un proces adecvat pentru detectarea relațiilor și modelelor în bazele de date de mari dimensiuni.

Extragerea informației din tipuri de date complexe ridică probleme dificile, pentru care există mai multe linii dedicate de cercetare și dezvoltare. Acest capitol prezintă o imagine de ansamblu la nivel înalt de exploatare a tipurilor de date complexe.

Așa cum am prezentat și în capitolul anterior, continua creștere a datelor, la nivel mondial, ne face să afirmăm că perioada aceasta este cu adevărat epoca datelor. Instrumente puternice și versatile sunt extrem de necesare pentru a descoperi în mod automat informații valoroase în cantitățile enorme de date și de a transforma aceste date în cunoștințe organizate. Această necesitate a dus la nașterea data mining-ului. Acest domeniu este încă tânăr, dinamic, și foarte promițător.

Tehnicile de extragere a cunoștințelor din date sunt implementări specifice ale algoritmilor care sunt utilizați pentru a efectua operațiunile de extragere a cunoștințelor din date.

Etapile procesului de extragere a cunoștințelor din date sunt următoarele:

1. Filtrarea datelor – Se elimină datele inconsistente sau perturbatoare;
2. Integrarea datelor – Integrarea datelor din mai multe surse;
3. Selecția datelor – Se selectează datele relevante pentru analiză;
4. Prelucrarea datelor – Datele sunt transformate și consolidate într-o formă acceptabilă în vederea facilitării extragerii cunoștințelor;
5. Extragerea cunoștințelor din date – Aplicarea algoritmilor inteligenți pentru extragerea informațiilor valoroase din date;
6. Evaluarea similarităților – Se realizează și se evaluează șabloanele (*patterns*);
7. Vizualizarea analizei – Informațiile obținute în urma extragerii cunoștințelor din date sunt transformate într-o formă prezentabilă utilizatorului.

⁵² [BELI] Berry, M. & Linoff, G. (2004). *Data Mining Techniques. For marketing, sales, and customer relationship management* (2nd ed.). Indianapolis: Wiley

II.2.1. Modele și algoritmi de extragere a cunoștințelor din date

Obiectivul extragerii cunoștințelor din date este de a identifica perspective noi, cu potențial, și de a realiza modele și corelații ușor de înțeles în datele existente [CHHS]⁵³. Algoritmii extragerii cunoștințelor din date pot fi modelați fie ca predictivi, fie ca descriptivi.

Un model predictiv face o predicție cu privire la valorile datelor utilizând rezultatele găsite și cunoscute din diverse date, în timp ce modelul descriptiv identifică modele sau relații în date. Spre deosebire de modelul predictiv, modelul descriptiv servește ca o modalitate de explorare a proprietăților datelor examinate, nu și pentru a prezice noi proprietăți. Algoritmii modelelor predictive de extragere a cunoștințelor din date includ clasificarea, predicția, regresia și analiza seriilor de timp. Cele descriptive cuprind metode cum ar fi clusterizarea, rezumarea, regulile de asociere, și analiza secvenței (Fig. 6.).

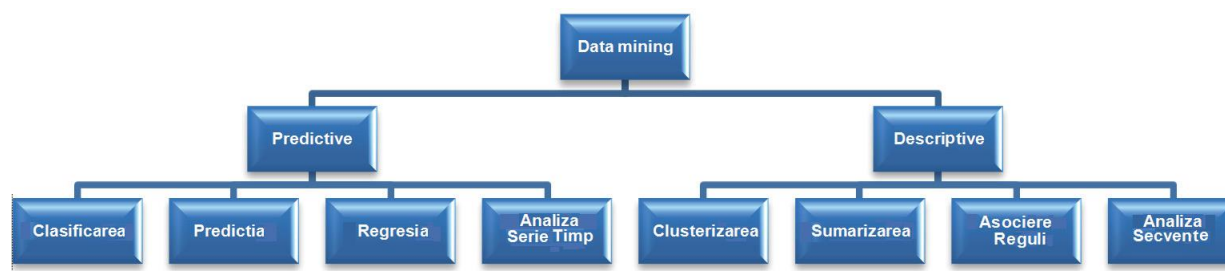


Fig. 6. Modele și algoritmi – data mining

Printre modelele de predicție, clasificarea este, probabil, cea mai bine înțeleasă abordare a extragerii cunoștințelor din date. Trei caracteristici comune ale algoritmilor de clasificare sunt: învățarea supravegheată, variabila dependentă este categorică și modelul construit este capabil de a atribui noi date uneia din clasele unui set de clase bine definit.

De exemplu, având în vedere clasele de pacienți care corespund răspunsului tratamentului medical; este identificată forma de tratament la care un nou pacient este probabil să răspundă cel mai bine. Spre deosebire de un model de clasificare, un model de predicție are scopul de a determina rezultatul viitor, mai degrabă decât comportamentul curent. Predicția poate fi o

⁵³ [CHHS] Chang, H. C. & Hsu, C.C. (2005). Using Topic Keyword Clusters for automatic Document Clustering. *Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)*, Kota Kinabalu, Sabah

valoare categorica sau numerică. De exemplu, având în vedere un model de predicție a tranzacțiilor cu carduri de credit, probabilitatea ca o tranzacție specifică sa fie frauduloasă poate fi prezisă.

Un alt model predictiv, cunoscut sub numele de regresie statistică este o tehnică de învățare supravegheată care implică analiza dependenței unor valori de attribute față de valorile altor attribute în aceeași clasă, precum și dezvoltarea unui model care poate prezice aceste valori de attribute pentru cazurile noi. De exemplu, având în vedere un set de date de tranzacții cu carduri de credit, poate fi construit un model nou care poate să prezică probabilitatea de fraudă pentru tranzacțiile noi. Aplicații de predicție cu unul sau mai multe attribute dependente de timp sunt numite probleme serie de timp.

Analiza seriilor de timp, de obicei, implică estimarea rezultatelor numerice, cum ar fi prețul viitor al acțiunilor, cotațiile bursiere, cotații financiare, etc.

A doua abordare a extragerii cunoștințelor din date este cunoscută sub numele de metoda descriptivă. Această metodă este, în mod normal, utilizată pentru a genera frecvențe, încrucișări și corelări. Metoda descriptivă poate fi definită pentru a descoperi regularități interesante în date, pentru a descoperi și a găsi modele de subgrupuri interesante în cea mai mare parte a datelor.

Clusterizarea, presupune ca un set de valori din date să fie împărțit într-un set de clase, astfel încât produsele cu caracteristici similare să fie grupate împreună. Clusterizarea este utilizată mai mult pentru a găsi grupuri de elemente care sunt similare. De exemplu, având în vedere un set de date ale clienților, subgrupurile de clienți care au un comportament similar de cumpărare pot fi identificate.

Analizele prin asociere sau identificare sunt folosite pentru a descoperi relațiile dintre attribute și elemente, cum ar fi prezența unui model care implică prezența unui alt model (adică în ce măsură un element este legat de un alt element în termeni de cauză-efect). Acest lucru este normal în stabilirea tipurilor de relații statistice între diferitele variabile interdependente ale unui model. Aceste relații pot fi asociații între attributele din cadrul aceluiași articol de date (ex: din cumpărătorii care au cumparat lapte, 64% au optat de asemenea și pentru pâine) sau asociații între diferite elemente de date (ex: de fiecare dată când acțiunile unei companii scad cu 5%, influențează cu aproximativ 13% acțiunile unei alte companii într-un interval de 2 - 6 săptămâni). Regulile de asociere sunt o tehnică populară pentru analiza coșului de piață, deoarece pot fi explorate toate combinațiile posibile de grupuri de produse potențial interesante.

Investigarea relațiilor între elemente pe o perioadă de timp este, de asemenea adesea menționată ca analiza secvențelor [HJKM]⁵⁴. Analiza secvenței este folosită pentru a determina modele secvențiale în date. Modelele din setul de date sunt bazate pe secvența de timp a acțiunilor, iar acestea sunt similare cu datele asociate, însă relația se bazează pe timp.

II.2.2. Tehnici de extragere a cunoștințelor din date

În această secțiune mi-am propus să ofer o viziune integrată a utilizării tehnologiei extragerii cunoștințelor din date, inclusiv metodologia și tehnicile, prin prezentarea procedurilor comune ale acestor tehnici în procesul de obținere a modelelor predictive, și printr-o descriere a particularităților metodologice asociate lor. Mai exact, în continuare, lucrarea se concentrează pe o descriere a tehnicilor care ne permit generarea unui răspuns categoric folosind modele predictive (modele de clasificare).

Tabelul II.2. prezintă o clasificare a unor tehnici de extragere a cunoștințelor din date în funcție de natura datelor analizate. În acest sens, voi prezenta tehnicile disponibile în funcție de natura variabilelor predictor și a variabilelor de ieșire. În cazul în care variabila de ieșire este continuă sau categorică avem de-a face cu modele de învățare supravegheate, în timp ce în cazul în care nu există nici o variabilă de ieșire avem de-a face cu modele de învățare nesupravegheate.

Mai exact, tehnicile abordate în cadrul lucrării vor fi: rețele neuronale, arbori de clasificare, k-NN (cel mai apropiat vecin), Naive Bayes și regresia logistică. Au fost alese aceste tehnici deoarece acestea permit să se analizeze variabilele de ieșire categorice pentru a genera modele de clasificare.

Clasificarea este o sarcină care face parte din categoria de învățare supravegheată și se referă la sarcina de a analiza un set de obiecte pre-clasificate pentru a învăța un model (sau funcție), care poate fi folosită pentru a clasifica datele necunoscute într-una din mai multe clase predefinite.

⁵⁴ [HJKM] Han, J. & Kamber, M. (2001). Data mining: concepts and techniques (Morgan-Kaufman Series of Data Management Systems), Academic Press, San Diego

Din perspectiva învățării supravegheate, tehnica de analiză estimează modelul din cunoștințele pe care le are despre comportamentul fiecăreia dintre variabilele de ieșire selectate, în așa fel încât tehnicile supravegheate, în sine, supraveghează dacă modelul pe care îl construiește se ajustează sau nu la cunoașterea realității. În acest sens, scopul de învățare supravegheată este de a genera modele bazate pe cunoaștere, care vor ajuta la prezicerea comportamentului datelor noi.

	Răspuns continuu	Răspuns categoric	Fără răspuns
Predicție continuă	Regresie liniară Rețele neuronale k-NN cel mai apropiat vecin	Regresie logistică Rețele neuronale Analiză discriminantă k-NN cel mai apropiat vecin	Componente principale Analiza clusterului
Predicție categorică	Regresie liniară Rețele neuronale Arbori de regresie	Rețele neuronale Arbori de clasificare Regresie logistică Naive Bayes	Reguli de asociere

Tabelul 2. Tehnicile de extragere a cunoștințelor din date, grupate în funcție de natura datelor [SHM]⁵⁵

O cerință comună în tehnicile de modelare predictivă o reprezintă utilizarea unui eșantion de date (date de testare), care este independent de cel utilizat în construcția modelului (datele de formare), cu intenția de a evalua capacitatea de generalizare a modelului.

Pe de altă parte, în procesul de învățare nesupravegheată nu există rezultate cunoscute pentru a ghida algoritmul în obținerea modelului, ci mai degrabă acesta explorează proprietățile datelor cu scopul de a identifica modele de comportament, fără o cunoaștere "a priori" a acestora. În acest fel, scopul învățării nesupravegheate este de a genera modele bazate pe cunoaștere având o intenție descriptivă, nu predictivă.

⁵⁵ [SHM] Shmueli, G., Patel, N.R. & Bruce, P.C. (2007). *Data Mining for Business Intelligence. Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. New Jersey: John Wiley & Sons

Rețele neuronale artificiale

Rețelele neuronale artificiale sunt sistemele de prelucrare a datelor a căror structură și mod de funcționare sunt inspirate de rețele neuronale biologice. Aceste rețele au fost elaborate pe baza următoarelor orientări:

- prelucrarea informațiilor are loc în elemente simple, numite neuroni.
- neuronii transmit semnale prin conexiuni stabilite.
- fiecare conexiune (legătură de comunicare) are asociată o pondere.
- fiecare neuron aplică o funcție de activare (de obicei neliniară) la totalul de neuroni conectați primiți ca intrare (suma intrărilor ponderate în funcție de ponderile de conectare), obținându-se astfel o valoare de ieșire, care va acționa ca valoare de intrare și va fi transmisă către restul rețelei.

Caracteristicile fundamentale ale acestor rețele sunt procesarea paralelă, memoria distribuită și adaptabilitatea la împrejurimi.

Unitatea de procesare este neuronul artificial, care primește intrările de la neuronii vecini și calculează o valoare de ieșire, care este trimisă tuturor neuronilor rămași.

În ceea ce privește reprezentarea informațiilor de intrare și de ieșire, putem găsi rețele cu date continue de intrare și de ieșire, rețele cu date discrete sau binare de intrare și ieșire și rețele cu date de intrare continue și date de ieșire discrete.

O astfel de rețea este formată din trei tipuri de noduri de bază sau straturi: noduri de intrare, ieșire și noduri intermediare (strat ascuns). Nodurile de intrare sunt responsabile de primirea valorilor inițiale ale datelor din fiecare caz, în scopul de a le transmite rețelei. Nodurile de ieșire primesc date de intrare și calculează valoarea de ieșire.

Folosirea acestui set de noduri de rețea, împreună cu funcția de activare, face posibilă reprezentarea cu ușurință a relației non-liniare, care este de departe cea mai dificilă dintre tehnicile multivariate.

Funcțiile de activare cele mai utilizate sunt: funcția de pas, funcția de identitate, sigmoid sau funcția logistică și tangenta hiperbolică.

Există o mare varietate de tipuri de rețele neuronale artificiale. O combinație a topologiei (numărul de neuroni și straturi ascunse, și modul în care acestea sunt conectate), paradigma învățării și algoritmul de învățare definesc un model de rețea neuronală artificială [BIGUS]⁵⁶.

Se poate spune că o rețea neuronală artificială are trei avantaje care o fac foarte atractivă în manipularea datelor: învățarea adaptivă prin exemple, robustețea în manipularea informațiilor redundante și inexacte și paralelismul masiv.

Metoda cea mai utilizată în aplicațiile practice ale rețelelor neuronale artificiale este cea a perceptronului multistrat. Acest model începe cu un model de intrare, în care fiecare nod sau neuron corespunde unei variabile predictor. Acești neuroni de intrare sunt conectați cu fiecare dintre neuronii care alcătuiesc stratul ascuns. Nodurile din stratul ascuns, sunt la rândul lor conectate cu neuronii din alt strat ascuns. Stratul de ieșire este format dintr-unul (predicție binară) sau mai mulți neuroni de ieșire. În acest tip de arhitectură, informația este întotdeauna transmisă de la stratul de intrare spre stratul de ieșire.

Atunci când rețeaua este folosită pentru a clasifica în mod normal, ieșirea are la fel de multe noduri ca numărul de clase și nodul de ieșire cu stratul cu cea mai mare valoare oferă estimarea clasei, pe care rețeaua o realizează pentru o anumită intrare. În cazul special a două clase se obișnuiește să se aibă un nod în stratul de ieșire, iar clasificarea între cele două clase se realizează prin aplicarea unui punct de tăiere la valoarea nodului.

Unul din avantajele rețelelor neuronale artificiale îl reprezintă posibilitatea modelării oricărui fel de relație funcțională (liniare sau neliniare) între variabile și, prin urmare, acționează ca funcție universală de aproximare, un alt avantaj remarcabil al acestei tehnici, comparativ cu tehnicile de modelare clasice, este faptul că nu impune nici un fel de restricții cu privire la datele de pornire (tipul relației funcționale între variabile).

Un alt avantaj al tehnicii constă în capacitatea sa de a estima modelele bune, chiar în ciuda existenței perturbațiilor în informațiile analizate, așa cum se întâmplă atunci când există valori omise sau valori aberante în distribuirea variabilelor. Prin urmare, aceasta este o tehnică robustă, atunci când se ocupă de probleme cu perturbații în informațiile prezentate, dar acest lucru nu înseamnă că criteriile de curățare ale matricei de date ar trebui să fie mai lejere.

⁵⁶ [BIGUS] Bigus, J.P. (1996). *Data Mining with neural networks: solving business problems from application development to decision support*. New York: McGraw-Hill

Cu toate acestea, flexibilitatea sa extremă se găsește în necesitatea de a avea suficiente date de formare și în nevoia de a avea mai mult timp pentru execuție decât alte tehnici. Este important de amintit că, în rețelele neuronale artificiale, un set de date de formare este necesar pentru a construi modelul precum și setul de date independent (date de testare), în scopul de a evalua capacitatea de generalizare, un al treilea set de date independente (set de validare) este folosit pentru a evita potrivirea excesivă a modelului (în timpul procesului de învățare), care poate provoca un număr excesiv de parametri sau greutatea în ceea ce privește problema.

În ciuda avantajelor prezentate, unul dintre cele mai importante dezavantaje care a fost ridicat împotriva utilizării rețelelor neuronale artificiale sugerează că o cunoaștere a greutăților în cadrul rețelei nu ajută, în general, modul de interpretare a procesului de bază generat de predicția unei anumite valori de ieșire.

Cu alte cuvinte, reproșurile aduse împotriva folosirii acestei tehnici sunt limitate la dificultatea în înțelegerea naturii reprezentărilor interne generate de rețea, ca răspuns la o anumită problemă. În ciuda acestui fapt, această percepție a rețelelor neuronale artificiale ca o complexă "cutie neagră" nu este complet adevărată. În acest sens, au apărut încercări diferite de interpretare a ponderilor rețelei neuronale, dintre care cel mai des utilizat este așa-numita analiză de sensibilitate [MOPA], pusă în aplicare în programele rețelelor neuronale artificiale, prezentate sub numele de sensibilitatea rețelei neuronale⁵⁷.

Arbori decizionali

Arborii decizionali sunt partiții secvențiale ale unui set de date care maximizează diferențele unei variabile dependente (variabilă de ieșire sau de răspuns). Aceștia oferă un mod concis de a defini grupuri care sunt consecvente în atributele lor, dar care variază în ceea ce privește variabila dependentă.

Arborii decizionali sunt alcătuiți din noduri (variabile de intrare), sucursale (grupuri de intrări în variabilele de intrare) și frunze sau noduri frunze (valori ale variabilei de ieșire). Construcția unui arbore se bazează pe principiul "divide și cucerește": prin intermediul unui

⁵⁷ [MOPA] Montano, J.J. & Palmer, A. (2003). Numeric sensitivity analysis applied to feedforward neural networks. *Neural Computing and Applications*, 12, 2, 119-125

algoritm de învățare supravegheată, diviziuni succesive ale spațiului multivariabil sunt procesate în scopul de a maximiza distanța dintre grupuri și fiecare divizie. Procesul de divizare este finalizat atunci când toate intrările de la o sucursală au aceeași valoare în variabila de ieșire (nodul frunză pur), dând naștere modelului complet (maximă specificată). Cu cât în partea inferioară variabilele de intrare sunt în arbore, cu atât mai puțin importante sunt în clasificarea de ieșire (și cu atât mai puțin permit generalizarea, din cauza scăderii numărului de intrări în ramurile inferioare).

Pentru a evita potrivirea excesivă a modelului, arborele poate fi “tuns” prin eliminarea ramurilor cu câteva intrări sau cu intrări ne semnificative. Ca urmare, dacă pornim de la modelul complet, după tăierea arborelui acest lucru va conduce la îmbunătățirea calității de generalizare (evaluat cu datele de testare), în detrimentul reducerii gradului de puritate al frunzelor sale.

Există diferiți algoritmi de învățare destinați obținerii modelelor de arbori decizionali (vezi Tabelul 3.). Algoritmul de învățare determină următoarele aspecte:

- compatibilitatea specifică cu tipul de variabile: natura variabilelor de intrare și variabilelor de ieșire;
- procedura de evaluare a distanței între grupurile din fiecare divizie: criterii de diviziune;
- restricțiile pot fi plasate pe numărul de sucursale în care fiecare nod poate fi împărțit;
- parametri de tăiere [pre-tăiere / post-tăiere]: Numărul minim de intrări pe nod sau sucursale, valoarea critică a diviziei, diferența de performanță între arborele extins și redus. Pre-tăierea presupune folosirea unor criterii de oprire în timpul construcției arborelui, în timp ce post-tăiere se aplică asupra parametrilor de tăiere ai întregului arbore.

Cei mai utilizați algoritmi (Tabelul 2) sunt CART (clasificare și regresie arbori), CHAID (Chi-Squared detectarea automată interactivă), QUEST (arbore statistic rapid, imparțial, eficient) și C4.5 / C5.0.

Algoritmul CART generează arbori de decizie binari, în cazul în care fiecare nod este împărțit exact în două ramuri. În acest fel, în cazul în care variabila de intrare este nominală și are mai mult de două categorii, ea grupează diferitele categorii într-o singură ramură. În cazul în care variabila de intrare este nominală sau continuă, generează încă două ramuri, și asociază un set de valori limitate în funcție de operatorii dintre ei, "mai mic sau egal cu" sau "mai mare"

decât o anumită valoare. Algoritmul CART face posibilă introducerea datelor de intrare nominale, ordinale și continue în model. Variabila de ieșire a modelului poate fi de asemenea nominală, ordinală sau continuă.

Algoritmul CHAID a fost inițial conceput pentru a gestiona doar variabile categorice. Cu toate acestea, în zilele noastre este posibil să se ocupe și de date categorice nominale și ordinale de ieșire și variabile continue. Procesul de construcție a arborelui se bazează pe calcularea semnificației unui contrast statistic ca un criteriu pentru a decide ierarhia importanței variabilelor predictor, și de a stabili grupuri de valori similare (omogene statistic) cu privire la variabila de ieșire, păstrând toate valorile care se dovedesc a fi eterogene (distincte) nealterate. Valorile similare sunt grupate într-o categorie, care face parte dintr-o ramură a arborelui. Testul statistic utilizat depinde de nivelul de măsurare a variabilei de ieșire. Dacă variabila menționată mai sus este continuă, se utilizează testul F. În cazul în care variabila de ieșire este categorică, este utilizat testul Chi-square.

O caracteristică diferențială între algoritmi CART și CHAID este că acesta din urmă permite împărțirea fiecărui nod în mai multe sucursale. Prin urmare, acesta tinde să creeze arbori mult mai largi decât metodele de dezvoltare binare.

Algoritmul QUEST poate fi utilizat în cazul în care rezultatul este nominal-categoric (permite crearea arborilor de clasificare). Procesul de construcție al arborelui este, de asemenea, bazat pe calcularea semnificației unui contrast statistic. Pentru fiecare variabilă de intrare, în cazul în care acest lucru este o variabilă categorică nominală, se calculează nivelul critic al unui contrast de independență Pearson Chi-pătrat între variabila de intrare și variabila de ieșire. În cazul în care variabila de intrare este ordinală sau continuă, se folosește testul F.

Algoritmul C5.0 admite numai variabilele de ieșire categorice. Variabilele de intrare pot fi categorice sau continue. Acest algoritm este rezultatul evoluției algoritmului C4.5 care are ca nucleu versiunea ID3. Algoritmul ID3 se bazează pe conceptul de a obține informații pentru a selecta cel mai bun atribut.

Algoritm	Variabile intrare	Variabile ieșire	Tip predicție	Ramuri divizare	Criteriu divizare
CHAID	Categorice / numerice	Categorice / numerice	Clasificare / regresie	≥ 2	Chi-square / F
QUEST	Categorice / numerice	Categorice	Clasificare	$= 2$	Chi-square / F
CART	Categorice / numerice	Categorice / numerice	Clasificare / regresie	$= 2$	GINI / Deviație pătratică
C4.5/C5.0	Categorice / numerice	Categorice	Clasificare	≥ 2	Ratia de îmbunătățire

Tabelul 3. Comparație între algoritmi de învățare pentru arborii decizionali [GERV]⁵⁸

Unul dintre cele mai importante avantaje ale arborilor decizionali este natura lor descriptivă, care ne permite să înțelegem cu ușurință și să interpretăm deciziile luate de model, așa cum avem acces la normele care sunt utilizate în activitatea de predicție (un aspect care nu este luat în considerare în alte tehnici de învățare mașină, cum ar fi rețelele neuronale artificiale). Astfel, arborii decizionali permit reprezentarea grafică a unei serii de norme referitoare la decizia care trebuie făcută în atribuirea unei valori de ieșire pentru o anumită intrare, oferind o explicație prietenoasă și intuitivă a rezultatelor.

Pe de altă parte, normele de decizie furnizate de un model de arbore au o valoare predictivă (nu numai descriptivă) din momentul în care exactitatea acestora este evaluată din date independente (date de testare) cu cele utilizate în construcția modelului (date de formare).

O altă caracteristică atractivă a arborilor decizionali este aceea că sunt intrinseci și robuști la valorile lipsă, ocupându-se de acestea fără să atribuie valori sau să elimine observații. Cu toate acestea, arborii decizionali au și unele puncte slabe precum: aceștia sunt sensibili la mici schimbări în setul de date și, spre deosebire de modelele care presupun un anumit raport între răspuns și predicție (de exemplu: o relație liniară, cum ar fi o regresie liniară), arborii decizionali

⁵⁸ [GERV] Gervilla, E., Jiménez, R., Montaña, J.J., Sesé, A., Cajal, B. & Palmer, A. (2009). The methodology of *Data Mining*. An application to alcohol consumption in teenagers. *Adicciones*, 21, 1, 65-80

nu sunt liniari și nu sunt parametrizabili. Acest lucru permite o gamă largă de relații între predictorii și răspuns, dar poate fi o slăbiciune: având în vedere faptul că partițiile sunt efectuate pe predictorii unici, mai degrabă decât de combinații de predictorii, arborii decizionali probabil omit relațiile dintre predictorii, în special structurile liniare, cum ar fi modelele liniare și logistice de regresie.

Un alt dezavantaj în construcția arborilor decizionali este problema potrivirii excesive a modelului, adică, include nu numai modelele reale sau structurile prezente în date, ci și o parte din erori. Pentru a reduce această problemă, cât mai mult posibil, există mai multe strategii:

- strategii care încetinesc creșterea arborelui înainte de a ajunge la clasificarea perfectă a exemplilor din setul de test (de exemplu, algoritmul CHAID);
- strategii care fac posibil ca arborele să crească complet și apoi să efectueze unele tăieri (de exemplu, CART și algoritmi C4.5). Acestea din urmă s-au dovedit a fi mai eficiente decât prima.

Un ultim dezavantaj al arborilor decizionali îl reprezintă nevoia unui set mare de date, în scopul de a construi un clasificator bun. Cu toate acestea, Breiman și Cutler [BRCU] au introdus "seturile aleatorii", care se ocupă cu aceste limitări⁵⁹. Ideea de bază este de a crea mai mulți arbori decizionali din date (și, astfel, a obține un "set") și de a combina rezultatul lor pentru un clasificator mai bun [BRE]⁶⁰.

K-NN

Când ne aflăm în situații noi, suntem ghidați de amintirile situațiilor asemănătoare pe care le-am experimentat în trecut. Aceasta este baza tehnicii celui mai apropiat vecin k-NN. Adică, tehnica k-NN se bazează pe conceptul de similitudine. Mai mult decât atât, această tehnică construiește o metodă de clasificare fără a face presupuneri cu privire la forma funcției la care se referă variabila dependentă cu variabilele independente.

⁵⁹ [BRCU] Breiman, L. & Cutler, A. (2004). *Random Forests*. Retrieved from http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm

⁶⁰ [BRE] Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 1, 5-32

Scopul este de a identifica într-o modalitate dinamică observațiile k în datele de formare care sunt similare cu o nouă observație pe care ne-o dorim să o clasificăm. În acest fel, k observații (începând) similare sunt utilizate pentru a clasifica o observație specifică. Mai precis, K-NN caută observații în datele de formare care sunt similare sau aproape de observația care trebuie să fie clasificată, în funcție de valorile variabilelor independente (atribute). Apoi, în funcție de clasele acestor observații din apropiere, se atribuie o clasă la observația care se dorește a se clasifica, luând votul majorității vecinilor pentru a determina clasa. Cu alte cuvinte, se contorizează numărul de cazuri din fiecare clasă și se atribuie noua cauză celei căreia aparțin cei mai mulți dintre vecinii săi.

Chiar dacă metoda poate părea intuitivă, aceasta este capabilă să concureze cu celelalte metode de clasificare mai sofisticate. Prin urmare, în cazul în care modelul liniar este rigid, k -NN este extrem de flexibil. Performanța acestei tehnici pentru date de aceeași mărime depinde de k , și de măsurările realizate pentru a determina care observații sunt în apropiere.

Ca urmare, în aplicarea tehnicii trebuie să luăm în considerare cât de mulți vecini pot fi luați în considerare (valoarea k), modul de măsurare al distanței, modul de combinare a informațiilor pentru mai mult de o observație și dacă toți vecinii ar trebui să aibă aceeași pondere.

După cum am spus, tehnica k -NN clasifică un exemplu necunoscut în cea mai comună clasă, cu ajutorul vecinilor săi K din apropiere. Se presupune că toate exemplele corespund punctelor într-un spațiu n -dimensional. Un vecin este considerat în apropiere în cazul în care are cea mai mică distanță în spațiul n -dimensional de atribute $[AN]$ ⁶¹. Dacă ne-am stabilit $k = 1$, exemplul necunoscut este clasificat în clasa celui mai apropiat vecin în setul de date.

Deși nu există o formulă pentru a alege numărul k , este de remarcat faptul că, dacă vom alege o valoare k mică, clasificarea se poate, eventual, să fie prea afectată de valori aberante sau observații neobișnuite. Pe de altă parte, alegerea unei valori k nu foarte mici, tinde să amortizeze orice comportament individual învățat de la datele de instruire. Cu toate acestea, dacă vom alege o valoare k , care este prea mare, comportamentul interesant la nivel local va fi trecut cu vederea.

⁶¹ [AN] An, A. (2006). Classification Methods. In J. Wang (Ed.), *Encyclopedia of Data Warehousing and Mining* (pp. 144-149). Hershey, PA: Idea Group Inc

Se recomandă astfel să ne folosim de ajutorul datelor pentru a rezolva această problemă, urmând o procedură de validare încrucișată. Aceasta presupune să încercăm mai multe valori ale lui k cu seturi diferite de formare alese la întâmplare și alegerea unei valori pentru k care minimizează eroarea de clasificare.

În ceea ce privește modul de măsurare a distanței, cea mai comună funcție de distanță este distanța euclidiană (1), unde x și y reprezintă valorile atributelor celor două cazuri.

$$d_{Euclidiană}(x, y) = \sqrt{(x_i - y_i)^2} \quad (1)$$

Deși, în mod alternativ, poate fi utilizată și distanța Manhattan (2).

$$d_{Manhattan}(x, y) = |x_1 - x_2| + |y_1 - y_2| \quad (2)$$

Distanța euclidiană are trei dezavantaje:

- depinde de unitățile alese pentru măsurarea variabilelor;
- nu ia în considerare variabilitatea diferitelor variabile;
- ignoră corelația dintre variabile.

O soluție o reprezintă utilizarea unei măsuri numită distanță statistică (sau distanța Mahalanobis).

Printre avantajele tehnicii k -NN se pot evidenția: în primul rând, aceasta nu simplifică distribuirea obiectelor în spațiu într-un set de caracteristici de înțeles; în schimb, setul de date de test este stocat complet ca o descriere a acestei distribuții. În plus, metoda k -NN este intuitivă, ușor de implementat și eficientă în practică. Se poate construi o aproximare diferită pentru funcția țintă pentru ca fiecare exemplu nou să fie clasificat, ceea ce este avantajos atunci când funcția țintă este foarte complexă, dar poate fi descrisă printr-o colecție de aproximări locale mai puțin complexe. În cele din urmă, această tehnică construiește o metodă de clasificare fără să efectueze ipoteze referitoare la forma funcției la care se referă variabila dependentă (variabila clasificare) cu variabilele independente (atribute).

Pe de altă parte, cel mai important dezavantaj este acela că tehnica k -NN este foarte sensibilă la prezența parametrilor relevanți. Alte dezavantaje:

- timpul necesar, pentru a găsi vecinii din apropiere într-un set de date de test mare, poate fi foarte mare;

- numărul de observații necesare într-un set de date de test crește exponențial cu numărul de dimensiuni (variabile).

Naive Bayes

Metodele bayesiene folosesc regula Bayes sau formula (3) (în funcție de teorema lui Bayes), care exprimă un cadru foarte puternic pentru a combina informațiile relevante din eșantion (probabilitate anterioară), astfel încât să producă o informație relevantă actualizată (probabilitate posterioară) [GIUD]⁶².

Mai exact, tehnica Naïve Bayes (NB) este o tehnică de clasificare foarte puternică, fiind una dintre cele mai utilizate, datorită procesării sale simple. Deoarece se bazează pe teorema lui Bayes, se poate prezice probabilitatea ca un anumit caz să aparțină unei anumite clase. Simplitatea sa de calcul se datorează ipotezei cunoscute sub numele de clasă independentă condiționată (acest lucru presupune ca efectul valorii unui atribut asupra unei anumite clase să fie independent de valorile altor atribute), și, în acest sens, este considerată un clasificator "naiv" [HAND]⁶³.

Acest clasificator prezice dacă un caz A va aparține clasei C_i care are cea mai mare probabilitate X condiționată posterior (set de atribute ale cazului în variabilele predictor).

Teorema Bayes ne permite să definim formula lui Bayes (3), pe care această probabilitate posterioară o oferă; și din moment ce $P(X)$ este constantă pentru toate clasele, avem nevoie doar de a maximiza $P(X | C_i) P(C_i)$, în procesul de clasificare.

$$P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)} \quad (3)$$

Dintr-un set de date de formare $P(C_i)$ se poate estima de câte ori fiecare clasă C_i apare în setul de date. Pentru a reduce costul de calcul al estimării $P(X | C_i)$ pentru toate variantele x_k posibile (variabile predictor), clasificatorul folosește tocmai ipoteza "naivă" conform căreia

⁶² [GIUD] Giudici, P. (2003). *Applied Data Mining. Statistical Methods for Business and Industry*. England: John Wiley & Sons

⁶³ [HAND] Hand, D.J. (2007). Principles of Data Mining. *Drug Safety*, 30, 7, 621-622

atributele folosite pentru a descrie X sunt independent condiționate unul de altul pentru clasa dată C_i . Această independență condiționată poate fi găsită în expresia (4), unde valoarea indică numărul de variabile predictor care participă în clasificare.

$$P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)} \quad (4)$$

Studiile care compară algoritmi de clasificare (ca de exemplu, [MICH]) au arătat că de multe ori tehnica Naive Bayes este comparabilă cu tehnica rețelelor neuronale artificiale și arborii decizionali, și, de fapt, chiar depășește aceste metode clasificatoare sofisticate dacă atributele sunt independent condiționate, având în vedere clasa⁶⁴. Analize teoretice recente au arătat de ce tehnica Naive Bayes este atât de robustă [DOPA, RISH]⁶⁵.

Esența clasificărilor Naive Bayes constă în simplitatea lor, eficiența de calcul și performanța bună în clasificare. Mai mult de atât, Naive Bayes poate manipula cu ușurință valori necunoscute sau valori lipsă. Cu toate acestea, ea are trei dezavantaje importante: în primul rând, este nevoie de un număr mare de cazuri, în scopul de a obține rezultate bune [SHM]⁶⁶; în al doilea rând, în cazul în care o categorie de predicție nu este prezentă în datele de formare, tehnica presupune că noul caz cu această categorie în predictor are probabilitate zero; în cele din urmă, chiar dacă am obține o performanță bună în cazul în care scopul este de a clasifica sau de a ordona cazurile în funcție de probabilitatea lor de apartenență la o anumită clasă, această metodă oferă rezultate foarte părtinitoare atunci când scopul este acela de a estima probabilitatea de apartenență la o clasă.

Dincolo de aceste neajunsuri, tehnica Naive Bayes este simplă de utilizat, aceasta adaptându-se ușor la setul de date și fiind ușor de interpretat. În plus, este nevoie de o singură explorare a datelor.

⁶⁴ [MICH] Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. New York: Ellis Horwood Ltd

⁶⁵ [DOPA] Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29, 103-130

[RISH] Rish, I. (2001). An empirical study of the naive Bayes classifier. *Proceedings of IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*

⁶⁶ [SHM] Shmueli, G., Patel, N.R. & Bruce, P.C. (2007). *Data Mining for Business Intelligence. Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. New Jersey: John Wiley & Sons

Această simplitate, economie și interpretabilitatea a făcut-o să se bucure de popularitate pe scară largă, în special în literatura învățării asistate.

Regresia logistică

Regresie liniară este utilizată pentru a aborda relația dintre o variabilă de răspuns continuă și un set de variabile predictor. Cu toate acestea, atunci când variabila de răspuns este categorică, regresia liniară nu este adecvată.

Regresia logistică este un model liniar generalizat. Acesta este utilizat în principal pentru a prezice variabilelor binare. Astfel, tehnica regresiei logistice poate fi utilizată pentru a clasifica o nouă observație, a cărei grupare este necunoscută, într-una din grupe, bazată pe valorile variabilelor predictor.

Ca și în cazul regresiei liniare, în care încadrarea depinde de combinația liniară a atributelor, funcția logistică (5) transformă combinația liniară într-un interval [0,1] [YEN]⁶⁷. Astfel, în scopul de a utiliza regresia logistică, variabila dependentă este transformată într-o valoare continuă care este o funcție a probabilității de apariție.

$$p = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}} \quad (5)$$

Regresia logistică urmărește două etape: prima etapă constă în estimarea probabilității de apartenență la fiecare grupă în timp ce în a doua etapă se analizează probabilitățile în vederea clasificării fiecărui caz într-una din grupe. Parametrii modelului sunt estimați prin metoda probabilității maxime printr-un proces de iterații succesive.

În cele din urmă, trebuie amintit că regresia logistică poate produce rezultate stabile cu date relativ puține. Pe de altă parte, faptul că regresia tradițională este atât de larg acceptată, ușor de pus în aplicare și, în general, inteligibilă o face și mai atractivă.

⁶⁷ [YEN] Ye, N. (Ed.) (2003). *The handbook of Data Mining*. New Jersey: Lawrence Erlbaum Associates

II.2.3. Metodologii de extragere a cunoștințelor din date

Există mai multe metodologii diferite de extragere a cunoștințelor din date, dar nu există nici o metodologie standard aplicabilă pentru extragerea cunoștințelor din date. Prin urmare, mai mulți vânzători au creat propriile lor metodologii. Acestea au unele dezavantaje. Furnizorii de software au proiectat abordări, care sunt puternic corelate cu proiectarea propriilor lor soluții și a pachetelor software oferite.

O problemă metodologică este aceea că extragerea cunoștințelor din date a fost considerată ca un fel de artă în care fiecare analist ar putea urma propria lui "rețetă". *Cu toate acestea, această afirmație este doar parțial adevărată. În timp ce preferințele individuale și intuiția pot contribui la metodologii ingenioase de extragere a cunoștințelor din date, în același timp, există pași esențiali pe care metodologiile de data mining trebuie să ii includă și moduri elegante, eficiente de a combina elemente metodologice pentru a obține rezultate superioare.*

Două metodologii populare sunt SEMMA și CRISP-DM. Acestea sunt descrise în următoarele secțiuni.

Metodologia SEMMA

SEMMA este o metodologie utilizată pentru extragerea cunoștințelor din date. Aceasta a fost propusă de Institutul SAS - una dintre cele mai importante companii care dezvoltă aplicații software statistice - cu pachetul software Enterprise Miner. În SEMMA, SAS oferă un proces de exploatare a datelor, care constă în cinci etape: eșantionare, explorare, modificare, modelare și evaluare.

Această metodologie începe prin analizarea unei mici părți dintr-un set mare de date. Următorul pas este de a explora datele și informațiile pentru căutarea tendințelor și anomaliilor în date cu scopul de a obține informații despre date. În a treia fază, datele sunt modificate pentru a crea, selecta, și transforma variabilele pentru studiu. Un model valid este apoi creat folosind instrumente software, care caută în mod automat combinații de reguli și modele care să indice

rezultatele observate. În cele din urmă, ultimul pas al metodologiei SEMMA constă în evaluarea utilității și fiabilității rezultatelor.

Deși metodologia SEMMA conține unele dintre elementele esențiale ale oricărui proiect de extragere a cunoștințelor din date, aceasta se referă numai la statistică, modelare, și la părțile de manipulare a datelor din procesul de extragere a cunoștințelor din date. Îi lipsesc o parte din părțile fundamentale ale oricărui proiect de sisteme informatice, inclusiv fazele de analiză, proiectare, și implementare.

Dar, chiar mai important, metodologia SEMMA nu ia în considerare rolul organizației și părților interesate pe parcursul proiectului. *Astfel, putem afirma că metodologia aceasta nu vede extragerea cunoștințelor din date ca un element integral într-o perspectivă a sistemelor.*

SEMMA este special concepută pentru a lucra cu software-ul Enterprise Miner, software-ul de extragere a cunoștințelor din date al institutului SAS. Aceasta nu poate fi aplicată în afara limitelor acestui sistem.

Metodologia CRISP-DM

O altă metodologie de extragere a cunoștințelor din date este CRISP-DM. CRISP-DM a fost concepută inițial la sfârșitul anului 1996, dar nu a fost finalizată până în 1999. Acesta este destinat să fie neutră din punct de vedere al industriei, utilităților, și aplicabilității. Metodologia a fost dezvoltată de un consorțiu de furnizori de soluții de extragere a cunoștințelor din date și companii printr-un efort finanțat de către Comisia Europeană. Cei patru parteneri ai acestui proiect au fost NCR, Daimlerchrysler, OHRA, și Integral Solutions Limited (ISL), care a devenit parte a SPSS în 1998.

Sondajele de opinie efectuate la unul și același site (KDNuggets) în 2002, 2004, 2007 și 2014 [KDM02, KDM04, KDM07, KDM14] arată că metodologia a fost cea mai utilizată metodologie din industria extragerii cunoștințelor din date, de cei care au decis să răspundă la sondaj. Singura altă metodologie menționată în aceste sondaje a fost SEMMA. Cu toate acestea, de 3-4 ori mai multe persoane au raportat utilizarea metodologiei CRISP-DM.

Eforturile de a actualiza metodologia au început în 2006, dar nici până în prezent (iunie 2015), nu au dus la o nouă versiune.

Metodologia CRISP-DM cuprinde o defalcare ierarhică în care procesul de extragere a cunoștințelor din date este împărțit în patru nivele de abstractizare: fazele, sarcinile generice, sarcinile specializate, și instanțele de proces. CRIPS-DM recunoaște, de asemenea, cele patru dimensiuni diferite de extragere a cunoștințelor din context care conduc nivelurile generice și de specialitate ale CRISP-DM. Cele patru dimensiuni sunt: domeniul de aplicare, tipul problemă, aspectul tehnic, și instrumentele și tehnicile.

Tehnicile metodologiei CRISP-DM pot fi aplicate, deoarece acestea sunt încorporate în instrumentele disponibile pentru organizare și nu pentru că sunt cu adevărat necesare. Prin urmare, rezultatele de la această abordare nu pot corespunde în mod adecvat principiilor obiective ale organizației, iar modelele generate în acest fel nu pot reprezenta cu adevărat comportamentul entităților pentru care studiul a fost destinat în primul rând. Acest lucru poate fi valabil pentru anumite aplicații specifice de inginerie industrială, cum ar fi cele legate de controlul calității, monitorizarea procesului, planificare, optimizarea proceselor, și multe altele; dar ele nu sunt limitate doar la inginerie industrială și pot fi aplicate la mai multe proiecte diferite de extragere a datelor din alte domenii.

Cu toate acestea, metodologia CRISP-DM este utilă. Aceasta descrie un proiect de extragere a datelor ca un ciclu format din șase faze, în care secvența de faze nu este rigidă. Fazele luate în considerare de metodologia CRISP-DM sunt înțelegerea afacerii, înțelegerea datelor, pregătirea datelor, modelarea, evaluarea, și implementarea. Această abordare include în prima fază elementele foarte importante precum obiectivele de afaceri, cerințele, constrângerile și resursele disponibile pentru proiect, în scopul de a stabili obiectivele de extragere a cunoștințelor din date. Documentație este, de asemenea, promovată încă de la începutul planului.

Alte elemente esențiale care sunt luate în considerare sunt colectarea datelor în timpul fazei de înțelegere a datelor, dar numai pentru analiza datelor disponibile (nu neapărat datele necesare). Datele sunt, de asemenea, analizate și verificate pentru a se asigura calitatea și că acestea vor permite obținerea rezultatelor scontate ale modulelor.

În cadrul metodologiei CRISP, pregătirea datelor cuprinde selecția, curățarea, construcția, integrarea, și formatarea necesară datelor, în scopul de a crea orice model. Cu toate acestea, de asemenea, presupune că toate informațiile necesare sunt deja disponibile și continuă să fie valabile, astfel încât noile date nu ar trebui să fie colectate.

II.2.4. Analiza comparativă a metodologiilor de extragere a cunoștințelor din date

Metodologia CRISP-DM are aplicații mai largi decât SEMMA, dar unul dintre dezavantajele majore este faptul că aceasta combină instrumentele (pachetele software) și tehnicile în aceeași categorie. Dacă instrumentele și tehnicile sunt combinate și selectate simultan, tehnicile pot fi alese pentru că sunt susținute de instrumente specifice de extragere a cunoștințelor din date, și nu pentru că acestea sunt cele mai relevante pentru scopul studiului, sau pentru că acestea sunt necesare. Acest lucru, de asemenea, poate influența obiectivele și cerințele organizației, existând posibilitatea unei sub-analizări ale studiului. În proiectele de extragere a cunoștințelor din date, este important să se analizeze nevoile organizației, cerințele, obiectivele, și strategiile.

O altă problemă cu abordarea sugerată de metodologia CRISP-DM este că selecția tehnicii este amânată până în faza de modelare; în cazul în care datele solicitate nu sunt disponibile sau sunt în format greșit, modelul trebuie să revină la faza de analiză a datelor.

Metodologia CRISP-DM subliniază importanța evaluării instrumentelor și tehnicilor la începutul procesului, dar, de asemenea, afirmă că selecția instrumentelor poate influența întregul proiect. Acesta este un impediment major pentru extragerea cunoștințelor din date pentru inginerie industrială, deoarece în cazul în care selecția de instrumente influențează complet proiectul, rezultatele, regulile, și modelele și predicțiile obținute cu modele care rezultă, nu garantează că rezolvă problemele pe care organizațiile încearcă să le rezolve.

Metodologiile ar trebui să fie selectate în funcție de obiectivele și cerințele unei organizații și nu ar trebui să depindă numai de datele disponibile. În cazul în care datele disponibile nu sunt suficiente pentru organizație pentru a efectua un proiect de extragere a cunoștințelor din date, noi date și informații ar trebui colectate; în caz contrar, alegerea tehnicii necesare poate fi influențată numai de datele existente, astfel încât modelele rezultate pot fi părtinitoare și nu vor corespunde intențiilor reale ale organizației. O problemă similară este că, deși ipotezele sunt declarate în mod clar, acestea nu pot fi revizuite în mod suficient. Modificările în date, din trecut spre viitor, pot duce la presupuneri cu privire la datele care au fost o dată valide pentru a fi incorecte.

Atât metodologia CRISP-DM cât și altele, subliniază faptul că, în funcție de metodologia selectată, datele trebuie uneori să fie împărțite în seturi de formare și validate. După construcția

modelelor cu datele de formare, un test de validare este apoi utilizat pentru a se asigura că modelul obținut se comportă cu acuratețe adecvată pentru sistemul real.

În cele din urmă, în metodologia CRISP-DM, după crearea a suficient de multe modele, faza de evaluare continuă cu o analiză a rezultatelor, o revizuire a procesului, și în final cu faza de implementare. Faza de desfășurare a metodologiei CRISP-DM constă în crearea unui plan de desfășurare, un plan de monitorizare și de întreținere, un raport final, și evaluarea finală a proiectului. Pe lângă dificultățile pe care metodologia CRISP-DM le prezintă, aceasta este o abordare buna pentru procesul general de exploatare a datelor și pentru ciclul de extragere a cunoștințelor din date.

Conform unui studiu KDnuggets [KDN], CRISP-DM rămâne cea mai populara metodologie de extragere a cunoștințelor din date, analiză și alte proiecte științifice, cu o pondere de 43%, dar un înlocuitor pentru această metodologie neîmbunătățită recent este așteptat de mult timp. Participanții la studiu au fost nevoiți să răspundă la întrebarea: “Care este principala metodologie pe care o utilizați pentru extragerea cunoștințelor din date, analiză și alte proiecte științifice?”. Acest studiu a fost prima data realizat în anul 2007 și reluat apoi în anul 2014⁶⁸.

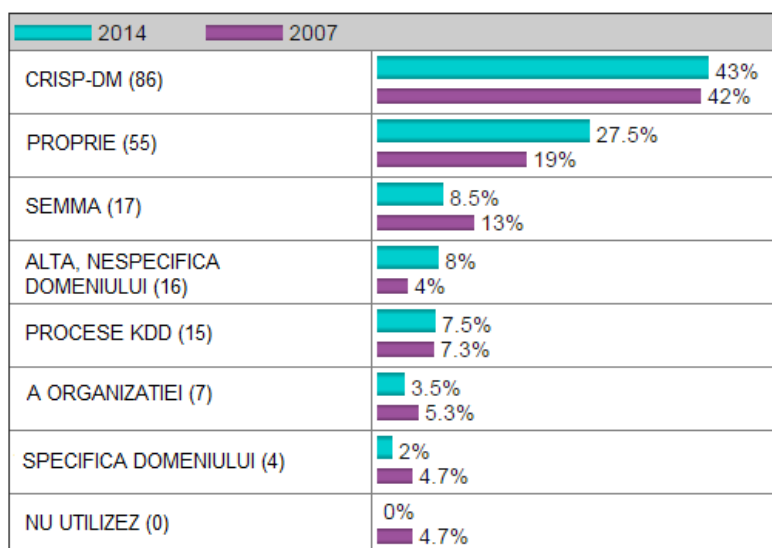


Fig 7. Studiu privind popularitatea metodelor [KDN]

⁶⁸ [KDN] Gregory Piatetsky, CRISP-DM, still the top methodology for analytics, data mining, or data science projects, Kdnuggets

Analizând rezultatul studiului, putem spune că lipsa metodologiilor moderne a condus la creșterea numărului celor care adoptă o metodologie proprie (specifică sau nu domeniului), în timp ce ponderea utilizării metodologiilor consacrate a rămas aproximativ aceeași (doar SEMMA pierzând în popularitate).

Încă o concluzie importantă pe care o putem formula în urma analizei acestui studiu, se referă la importanța utilizării unei metodologii. Dacă în studiul realizat în anul 2007, 4.7% din cei întrebați au răspuns că nu utilizează o metodologie, în anul 2014, procentul acestora a scăzut la 0. Este evident că în această perioadă tot mai multe companii au conștientizat importanța adoptării unei metodologii și au făcut pași importanți în această direcție.

II.3. Tehnologii informatice pentru analiza datelor

Analiza avansată a datelor a continuat să se dezvolte permanent. Această tehnologie oferă un mare impuls favorabil industriei bazelor de date, și permite ca un număr foarte mare de baze de date și arhive de informații să fie disponibile pentru gestionarea tranzacțiilor, regăsirea de informații, și analiza datelor. Datele pot fi acum stocate în mai multe tipuri de baze de date și arhive de informații.

O arhitectură de depozit de date în curs de dezvoltare este: depozitul de date (*data warehouse*). Acesta este un depozit format din mai multe surse de date eterogene organizate în cadrul unei scheme unificate într-o singură locație pentru a facilita luarea deciziilor de management. Tehnologia depozitelor de date include curățarea datelor, integrarea datelor, și OLAP (OLAP), care este, o tehnică de analiză cu funcționalități, precum sumarizarea, consolidarea, și agregarea, precum și cu posibilitatea de a vizualiza informații din diferite perspective. Deși instrumentele OLAP sprijină analiza multidimensională și de luare a deciziilor, instrumente suplimentare de analiză a datelor sunt necesare pentru o analiză în profunzime, de exemplu, instrumente de *data mining* care furnizează clasificarea datelor, *clustering*, *outlier* - detectarea anomaliilor, și caracterizarea modificărilor datelor în timp.

Pentru a facilita procesul decizional, datele dintr-un depozit de date sunt organizate în jurul subiectelor majore (de exemplu, clienți, produse, furnizori, și activități). Datele sunt stocate pentru a furniza informații dintr-o perspectivă istorică, cum ar fi în ultimele 6 până la 12 luni, și sunt de obicei sumariate. De exemplu, în loc să se depoziteze detaliile fiecărei tranzacții de vânzare, depozitul de date poate stoca un rezumat al tranzacțiilor pe produs pentru fiecare magazin sau, rezumate la un nivel superior, pentru fiecare regiune de vânzări.

II.3.1. Cubul de date și OLAP

În 1995 a fost înființat Consiliul OLAP, în vederea standardizării tehnologiilor prin stabilirea unor standarde deschise (OLAP API). Consiliul OLAP a publicat următoarea definiție: “OLAP este o tehnologie software ce permite analiștilor, managerilor și persoanelor cu funcție de conducere să analizeze datele printr-un acces rapid, consistent și interactiv și să le vizualizeze într-un mod cât mai variat.”⁶⁹ [OLAP].

Connolly, definește prelucrarea analitică on-line (OLAP) ca fiind sinteza, analiza și consolidarea dinamică a unor volume vaste de date multidimensionale [CONN]⁷⁰.

În modelul multidimensional, datele sunt organizate în mai multe dimensiuni, și fiecare dimensiune conține mai multe niveluri de abstractizare definite de ierarhii conceptuale. Această organizație oferă utilizatorilor flexibilitatea de a vizualiza datele din perspective diferite. Există un număr de operațiuni OLAP pe cuburile de date pentru a materializa aceste vederi diferite, permițând interogarea interactivă și analiza rapidă a datelor. Prin urmare, OLAP oferă un mediu ușor de utilizat pentru analiza datelor interactive.

Depozitele de date și instrumentele OLAP se bazează pe un model de date multidimensional. Acest model privește datele sub forma unui cub de date.

Un cub de date permite ca datele să fie modelate și vizualizate în mai multe dimensiuni, acesta fiind definit de dimensiunile și atributele sale.

În termeni generali, dimensiunile sunt perspectivele sau entitățile despre care o organizație dorește să țină evidențe. De exemplu, o companie poate crea un depozit de date de vânzări, în scopul de a ține evidența vânzărilor magazinului cu privire la dimensiunile de timp, obiecte, filiale, și locații. Aceste dimensiuni permit magazinului să țină evidența lucrurilor, cum ar fi vânzări lunare de produse și locațiile în care produsele au fost vândute. Fiecare dimensiune poate avea un tabel asociat, numit tabel dimensiune, care descrie în continuare dimensiunea. De exemplu, tabela dimensiune pentru un articol poate conține atributele: numele obiectului, marca

⁶⁹ [OLAP] OLAP Council. OLAP AND OLAP Server

⁷⁰ [CONN] CONNOLLY, T., BEGG, C. și STRACHAN, A. – „Baze de date. Proiectare. Implementare. Gestionare”, Editura Teora, București, 2001, ISBN 973-20-0601-3

și tipul. Tabelele dimensiune pot fi specificate de către utilizatori sau experți, sau pot fi generate automat și ajustate în funcție de distribuțiile de date.

Un model de date multidimensional este, de obicei, organizat în jurul unei teme centrale, cum ar fi vânzările. Această temă este reprezentată de o tabelă de atribute. Atributele sunt măsuri numerice, ca de exemplu, cantitățile pentru care dorim să analizăm relațiile dintre dimensiuni. Exemple de atribute pentru un depozit de date de vânzări includ prețul produselor vândute, unitățile vândute, etc. Tabela de atribute mai conține și cheile de la fiecare dintre tabele de atribute interconectate.

Deși ne gândim, de obicei, la cuburi ca structuri geometrice 3-D, în depozitele de date, cubul de date este n -dimensional. Pentru a obține o mai bună înțelegere a cubului de date și a modelului de date multidimensional, să încercăm să ne imaginăm un simplu cub de date 2-D, care este, de fapt, un tabel sau o foaie de calcul pentru datele de vânzări dintr-o companie. În special, ne vom gândi la datele de vânzări ale companiei pentru produsele vândute pe trimestru într-un anumit oraș. În reprezentarea 2-D, vânzările sunt prezentate în funcție de momentul de timp la care s-a efectuat vânzarea (organizat în sferturi) și obiectul vândut (organizate în funcție de tipul de produs vândut).

Să presupunem că ne dorim să vizualizăm datele de vânzări din figura 2 cu o a patra dimensiune suplimentară, cum ar fi furnizorii. Vizualizarea lucrurilor în 4-D devine complicată. Cu toate acestea, ne putem gândi la un cub 4-D ca fiind o serie de cuburi 3-D, după cum se arată în figura 2. Dacă vom continua în acest fel, este posibil să afișăm toate datele n -dimensionale ca o serie de $(n-1)$ cuburi dimensionale. Cubul de date este o metaforă pentru stocarea de date multidimensionale. Depozitarea fizică reală a acestor date poate diferi de reprezentare logică. Cel mai important lucru de reținut este că aceste cuburi de date sunt n -dimensionale și nu limitează datele pentru modelul 3-D.

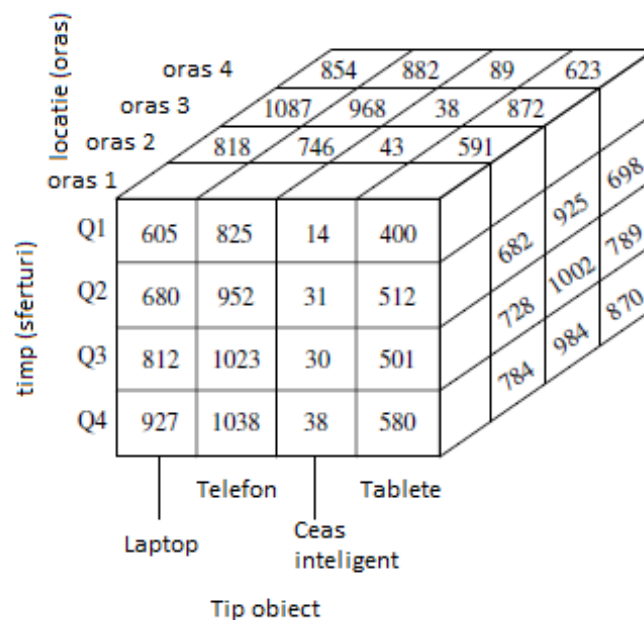


Figura II.8. Exemplu reprezentare date sub forma unui cub 3-D

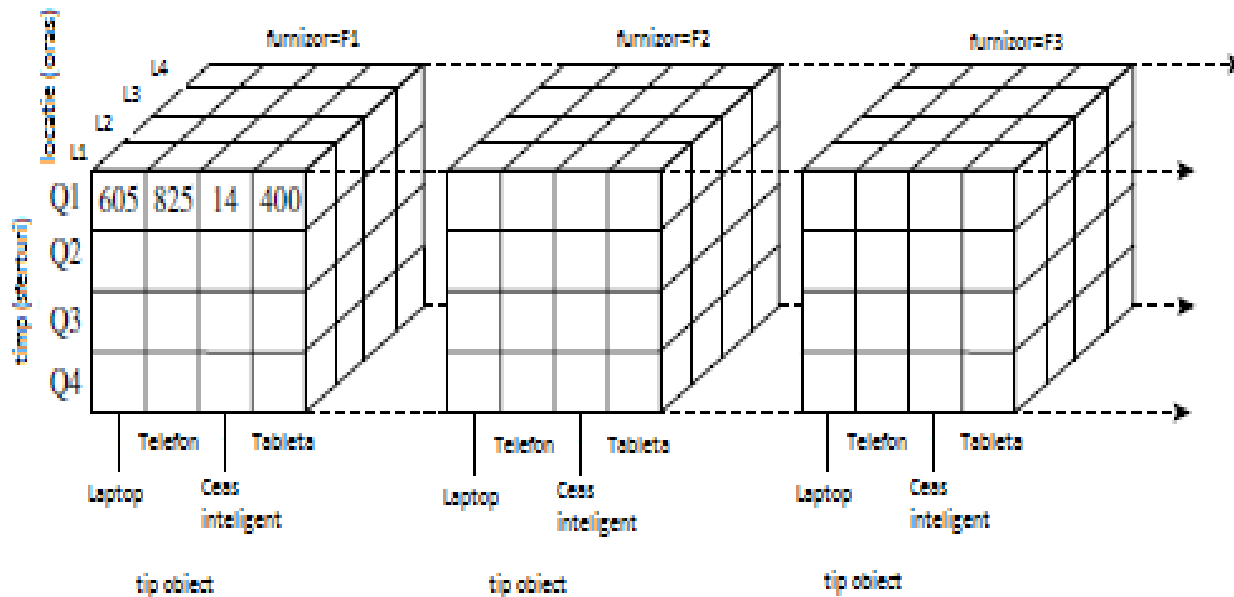


Figura 8. Exemplu reprezentare date sub forma unui cub 4-D

În literatura de specialitate a depozitelor de date, un cub de date ca cele prezentate în Figurile 7 și 8 este adesea menționat ca un cuboid. Având în vedere un set de dimensiuni, putem

genera un cuboid pentru fiecare dintre posibilele subgrupuri de dimensiuni indicate. Rezultatul ar forma o rețea de cuboide, fiecare prezentând datele la un nivel diferit de sumarizare, sau mod de grupare. Rețeaua de cuboide este apoi menționată ca un cub de date. Cuboidul care deține cel mai scăzut nivel de rezumare este numit cuboidului bază.

Caracteristicile multor sisteme OLAP (ca de exemplu, utilizarea unui model de date multidimensional și a conceptului de ierarhii, asociere a măsurilor cu dimensiuni, precum și noțiunile de *roll-up* și *drill-down*), există, de asemenea, în activitatea anterioară a bazelor de date statistice (SBDS). O bază de date statistică este un sistem de baze de date, care este conceput pentru a sprijini aplicații statistice. Similitudinile dintre cele două tipuri de sisteme sunt rareori discutate, în principal din cauza diferențelor de terminologie și al domeniilor de aplicare.

Cu toate acestea, sistemele OLAP și SBDS, au diferențe distinctive. În timp ce SBDS tind să se concentreze pe aplicațiile socio-economice, OLAP a fost orientat către aplicațiile de afaceri. Probleme de confidențialitate cu privire la ierarhiile conceptuale reprezintă o preocupare majoră pentru SBDS. De exemplu, având în vedere datele socio-economice sumarizate, este destul de delicat a permite utilizatorilor vizualizarea corespunzătoare a datelor de nivel scăzut. În cele din urmă, spre deosebire de SBDS, sistemele OLAP sunt proiectate pentru manipularea eficientă a unei cantități imense de date.

II.3.2. Tehnici utilizate pentru analiza datelor

În decursul anilor au fost dezvoltate diferite tehnici care încă sunt utilizate pentru analiza datelor. În continuare în cadrul lucrării, vor fi prezentate cele mai des întâlnite astfel de tehnici: prognoza serie de timp, analiza unităților, analiza cluster, segmentarea, clusterizarea, analiza factorilor și a componentelor principale, analiza corespondențelor și analiza supraviețuirii.

Prognoza serie de timp

Prognoza serie de timp este o formă simplă de tehnică de prognoză, în care unele puncte de date sunt disponibile pe intervale de timp regulate: zile, săptămâni sau luni. Dacă unele

modele pot fi identificate în datele istorice, este posibil să se proiecteze aceste modele în viitor ca o previziune. Prognoza vânzărilor este o utilizare populară a seriilor de timp prognoză. În plus față de tendințe, previziunea seriei de timp poate arăta, de asemenea, caracterul sezonier, care este pur și simplu un șablon repetat, care este observat într-un an sau mai puțin (cum ar fi mai multe vânzări de articole pentru cadouri la ocazii, cum ar fi de Crăciun sau de Ziua Îndrăgostiților).

Analiza unităților

Analiza unităților este o tehnică statistică, utilizată în principal în cercetarea de piață, pentru a determina ce produs (sau serviciu), caracteristici, sau preț ar fi atractiv pentru majoritatea clienților, în scopul de a influența decizia lor de cumpărare în mod pozitiv.

În studiile de unități, persoanelor care participa la studiu le este afișat un produs cu caracteristici și niveluri de preț diferite. Preferințele lor, favorabile cât și nefavorabile, sunt înregistrate pentru profilele alternative de produse. Cercetătorii apoi aplică tehnici statistice pentru a determina contribuția fiecăreia dintre aceste caracteristici ale produsului asupra percepției generale sau asupra unei decizii potențiale de cumpărare. Pe baza acestor studii, se poate realiza un model de marketing, care poate estima rentabilitatea, cota de piață, și veniturile potențiale care pot fi realizate din diferite modele de produse, prețuri, sau combinații ale acestora.

Analiza cluster

Intenția oricărui exercițiu de analiză cluster este de a diviza datele sau observațiile existente în grupuri similare și discrete. Fiecare observație este împărțită înțelept în grupuri pe baza tipului de clasificare al problemei, în timp ce în analiza cluster, scopul este de a determina numărul și componența grupurilor care pot exista într-un anumit set de date sau set de observare.

Segmentarea

Segmentarea este similară cu clasificarea, în cazul în care trebuie găsite criteriile de divizare a observațiilor în grupe distincte. Numărul de grupări poate fi evident chiar la începutul analizei, în timp ce scopul analizei cluster este de a identifica zonele cu concentrații diferite de alte grupuri. Prin urmare, clusterizarea este realizată pentru a descoperi prezența granițelor între diferitele grupuri, în timp ce segmentarea folosește limite sau criterii distincte pentru a forma grupuri.

Clusterizarea

Clusterizarea presupune împărțirea populației în grupuri diferite, bazându-se pe toți factorii disponibili. Segmentarea presupune, de asemenea, împărțirea populației în grupuri diferite, dar pe baza unor criterii predefinite, cum ar fi maximizarea profitului variabil, minimizând defectele, și așa mai departe. Segmentarea este utilizată pe scară largă în marketing pentru a crea campania potrivită pentru segmentul de client care produce avantaje maxime.

Analiza factorilor si a componentelor principale

Aceste metodologii statistice sunt utilizate pentru a reduce numărul de variabile sau dimensiuni într-un exercițiu de construire a unui model. Acestea sunt de obicei variabile independente.

Analiza componentelor principale este o tehnică care combină un număr mare de variabile într-un număr mic de subseturi, în timp ce analiza factorilor este o tehnică utilizată pentru a determina structura sau relația de bază prin calcularea factorilor ascunși care determină relațiile variabile.

Unele studii de analiză pot începe cu un număr mare de variabile, dar din cauza constrângerilor practice, cum ar fi manipularea datelor, timpul de colectare a datelor, bugetele, resursele de calcul disponibile, și așa mai departe, poate fi necesar să se reducă drastic numărul

de variabile care vor apărea pe modelul de date final. Numai acele variabile independente care au cel mai mare sens pentru afacere, trebuie să fie păstrate.

De asemenea, ar putea exista interdependență între unele variabile. De exemplu, nivelul de venit al unor indivizi dintr-o analiză tipică ar putea fi strâns legat de cheltuielile lunare. Cu cât este mai mare venitul, cu atât mai mare o să fie totalul cheltuielilor lunare. Într-un astfel de caz, este mai bine să se păstreze o singură variabilă pentru analiză și îndepărtate cheltuielile lunare din analiza finală.

Analiza corespondențelor

Analiza corespondențelor este similară cu analiza componentelor principale, dar se aplică datelor care nu sunt cantitative sau categorice cum ar fi sexul, statutul de admis sau respins, culoarea obiectelor și domeniul de specializare. Analiza corespondențelor oferă o modalitate de a reprezenta grafic structuri încrucișate de tabele, fiecare rând și coloană fiind reprezentate ca un punct.

Analiza supraviețuirii

Analiza de supraviețuire este de obicei folosită atunci când variabile, cum ar fi timpul de moarte, durata de spitalizare și timpul pentru a finaliza o teză de doctorat, trebuie să fie prezise. Practic, aceasta se ocupă de analiza datelor privind timpul până la producerea evenimentelor.

II.3.3. Analiza comparativă a instrumentelor de analiza datelor

Analiza datelor își propune să modeleze datele în scopul de a descoperi informații pentru a sprijini luarea deciziilor. Este nevoie de diverse operații asupra datelor, cum ar fi revizuire, curățare, transformare și modelare, validare și interpretare pentru a obține perspective de afaceri

utile. Uneori, seturile de date pot avea un milion de înregistrări sau poate chiar mai mult, fiind astfel necesare instrumente de analiză automată pentru manipularea acestor seturi uriașe de date.

Din fericire, tot mai multe astfel de instrumente bune sunt disponibile în prezent. O simplă căutare pe Google pentru instrumente de analiză a datelor oferă o listă cu o serie de astfel de instrumente. Multe dintre ele sunt *open source* și pot fi utilizate gratuit. SAS, SPSS, și R sunt pachetele software cele mai utilizate pe scară largă în prezent, cel puțin pentru aplicațiile de analiză de afaceri.

SAS este un instrument care a fost la îndemana utilizatorilor încă din anii 1970. Există atât de multe realizări construite folosind acest instrument, încât majoritatea companiilor din lumea afacerilor, care sunt în domeniul analizei afacerii, la orice nivel, continuă să îl folosească.

R a fost introdus în 1996. De-a lungul anilor, o mulțime de noi capacități și caracteristici au fost construite în jurul acestui instrument, devenind unul dintre cele mai puternice instrumente de analiză a datelor *open source*, disponibil în prezent. Acest lucru, îl face popular atât în cercetare, cât și în comunitatea academică. Multe companii din lumea afacerilor au început, de asemenea, cu ajutorul R.

SPSS există, de asemenea, de peste 20 de ani și are o bază de utilizatori puternică în științele sociale și în multe alte comunități.

În continuare în cadrul lucrării, se vor detalia unele software-uri de analiză a datelor, utilizate în mod obișnuit, și modurile de alegere al unui astfel de utilitar. SAS, SPSS și R sunt considerate cele mai frecvent utilizate software-uri în industria de profil. Alte utilitare similare sunt: Statistica, KXEN Modeler, GNU Octave, Knime, Minitab.

În tabelul următor sunt prezentate principalele instrumente de analiză a datelor, disponibile pe piață.

Instrument	Descriere
SAS	Cel mai utilizat instrument de analiză avansată Are o mulțime de algoritmi de modelare predictivă
SPSS	Are algoritmi foarte buni de extragere a cunoștințelor din date și din text
R	Cel mai utilizat instrument de analiză open source Are mai multe pachete de analiză a datelor

MATLAB	Utilizat pe scară largă pentru analiză numerică și calcul
RapidMiner	Instrument bun, bazat pe interfață grafică pentru segmentare și clusterizare; poate fi utilizat pentru modelarea convențională Open source
Wekka	Instrument de învățare mașină Open source
SAP	Instrument pentru gestionarea operațiunilor de afaceri și a relației cu clienții Cel mai adesea utilizat pentru vizualizarea informațiilor despre operații
Apache Mahout	Instrument avansat de analiză a volumelor mari de date. Open source

Tabelul 4. Instrumente de analiza a datelor

Alegerea instrumentului potrivit de analiză a datelor se face în funcție de mai multe considerente, printre care: aplicația de afacere, complexitatea acesteia și strategia de analiză a organizației, procesele organizaționale, constrângerile bugetare, mediul proiectului și structura de conducere, structura și dimensiunea datelor care urmează să fie modelate, tehnica de analiză utilizată și frecvența de utilizare a acesteia, integrarea unui depozit de date, așteptările de recuperare a investiției, etc.

Mult mai multe considerente specifice unei organizații sau unui proiect pot fi adăugate la această listă. Ordinea importanței acestor considerații poate varia de la o persoană la alta, de la un proiect la altul, și de la o organizație la alta. Decizia finală, cu toate acestea, nu este una ușoară.

În continuare, în cadrul lucrării se vor analiza câteva caracteristici comparative ale SAS, SPSS, și R, care ar putea ajuta la procesul de luare a deciziilor cu privire la alegerea instrumentului care se potrivește cel mai bine nevoilor de afaceri.

În unele cazuri, s-ar putea concluziona că o simplă aplicație de calcul tabelar, cum ar fi Microsoft Excel, este cea mai convenabilă și eficientă și oferă o perspectivă suficientă pentru a rezolva problema de afaceri. Alteori, un singur proiect de analiză ar putea necesita utilizarea a mai mult de un singur instrument de analiză. Un analist de date va trebui să aplice instrumente software diferite în funcție de problema în discuție.

SAS și SPSS au sute de funcții și proceduri și pot fi în general împărțite în cinci părți: funcții de management al datelor de intrare, care facilitează citirea, transformarea, și organizarea datelor înainte de analiză, procedee de analiză a datelor, care ajută la analiza statistică reală, sistem de livrare de ieșire (ODS), în cazul SAS, și sistemul de management al producției în cazul SPSS (OMS), care ajută la extragerea datelor de ieșire de reprezentare finală sau pentru a fi utilizate de alte proceduri ca intrări, limbaje macro, care pot fi utilizate pentru a da seturi de comenzi în mod repetat pentru a efectua sarcini legate de programare, limbaje matrix (SAS IML și SPSS Matrix), care pot fi folosite pentru a adăuga noi algoritmi.

R are toate aceste cinci domenii integrate într-unul singur. Cele mai multe dintre procedurile R sunt scrise folosind limbajul R, în timp ce SAS și SPSS nu folosesc limbaj matern cu care să își scrie procedurile. Fiind *open source*, procedurile R sunt disponibile utilizatorilor atât pentru a le folosi cât și pentru a le edita.

Așa cum spune și site-ul web SAS, suita de software SAS de analiză în afaceri are peste 35 ani de experiență și 60,000 site-uri de clienți la nivel mondial. De asemenea, are cea mai mare cotă de piață la nivel global în ceea ce privește analiza avansată. Acesta, poate face, practic, tot în legătură cu analiza avansată a informațiilor de afaceri, de gestionare a datelor, precum și analiza predictivă.

Dezvoltarea SAS a început inițial de la Universitatea Carolina de Nord, unde a fost dezvoltat între anii 1966 și 1976. Institutul SAS, fondat în 1976, deține acest software la nivel mondial. Din 1976, noi module și funcționalități au fost adăugate în aplicația de bază. Modulul de analiză de *social media* a fost adăugat în 2010.

Software-ul SAS este, în general, foarte mare și are mai mult de 200 de componente. Unele dintre componente îndeplinesc următoarele facilități: proceduri de bază și de gestionare a datelor, analiză statistică, grafică și prezentare, econometrie și analiza seriilor de timp, limbaj matrice interactiv, extragerea cunoștințelor din date.

Limbajul de programare SAS este un limbaj procedural de programare, de nivel înalt, care oferă o multitudine de funcții predefinite statistice și matematice. Acesta oferă, de asemenea, atât capacități grafice neliniare cu caracteristici avansate de raportare liniară. Este posibilă manipularea și administrarea convenabilă a datelor folosind limbajul de programare SAS, înainte de aplicarea tehnicilor statistice.

Capacitățile de manipulare a datelor oferite de SAS au devenit și mai importante, deoarece până la trei sferturi din timpul petrecut în cele mai multe proiecte de analiză este investit în extragerea de date, transformare, precum și curățare. Această capacitate este inexistentă în alte pachete de analiză a datelor populare, care pot necesita ca datele să fie manipulate sau transformate cu ajutorul mai multor alte programe, înainte de a putea fi supuse la procedurile de analiză statistică efective. Unele tehnici statistice, cum ar fi analiza procedurilor de varianță (ANOVA), sunt deosebit de puternice în mediul SAS.

R este un instrument integrat de gestionare și manipulare a datelor, de analiza și grafică, care dispune, de asemenea, și de capabilități de raportare. Într-un mediu integrat, acesta poate fi utilizat pentru îndeplinirea următoarelor funcții: funcții de management de date, cum ar fi extracție, manipulare, transformare, și depozitare, procedee de analiză statistică și grafică și capabilități avansate de raportare.

Extensibilitatea oferită de R este unul dintre cele mai mari avantaje oferite. Mii de pachete opționale sunt disponibile ca extensii ale software-ului de bază. Dezvoltatorii pot vedea codul din spatele procedurilor și îl pot modifica pentru a scrie propriile pachete.

Limbaje de programare mai populare, cum ar fi C ++, Java, sau Python pot fi conectate la mediul R. SPSS ofera o legătură cu R pentru utilizatorii care, în primul rând, folosesc mediul SPSS pentru analiza datelor. SAS oferă, de asemenea, unele mijloace pentru a muta datele și grafica între cele două pachete.

Numele de SPSS înseamnă Pachetul Statistic pentru Științe Sociale (*Statistical Package for the Social Sciences*). Inițial, acesta a fost un pachet software destinat analizelor statistice, și a fost dezvoltat de SPSS Inc. Ulterior, acesta a fost achiziționat în 2009 de către IBM și redenumit SPSS Statistics.

Conform unui articol [SK15], majoritatea utilizatorilor SPSS, consideră că utilitarul folosit de ei, conține un meniu de comandă cu mai multe opțiuni, comparativ cu R sau SAS⁷¹. De asemenea, în același articol se mai subliniază că majoritatea utilizatorilor au considerat că timpul necesar acomodării cu SPSS este mai mic, în comparație cu R sau SAS. Aceștia, au mai susținut și că SAS și SPSS sunt destul de asemănătoare, și că trecerea de la un utilitar la altul este destul de lejeră, în timp ce R pare destul de diferit după o primă utilizare.

⁷¹ [\[SK15\] Sheilendra Kadre - How to Choose Your Data Analysis Tools, 22 Aprilie 2015](#)

III. Soluții informatice pentru modelarea proceselor de afaceri

În climatul economic actual, companiile trebuie să fie cât se poate de eficace și eficiente și să aleagă cele mai bune decizii posibile.

Un mod prin care se poate realiza acest lucru este prin integrarea inteligenței afacerii și a managementului proceselor de afaceri. Implementarea acestei combinații de tehnologii reprezintă o prerechizită obligatorie a afacerilor inteligente.

Companiile de astăzi deja utilizează, și apreciază valoarea, inteligenței afacerilor (*business intelligence*). Datele afacerilor sunt analizate în mai multe scopuri: o companie poate efectua o analiză a jurnalului de sistem și o analiză *social media* pentru evaluarea riscurilor, păstrarea clienților, managementul companiei (*brand-ul*), și așa mai departe. De obicei, astfel de sarcini variate au fost manipulate de sisteme separate, chiar dacă fiecare sistem includea etape similare de extragere de informații, curățenie date, prelucrări relaționale, modelare statistică și predictivă, precum și explorare și vizualizare utilizând instrumente adecvate.

Utilizarea sistemelor separate bazate pe tehnologia volumelor mari de date, devine prohibitiv de scumpă, având în vedere dimensiunea mare a seturilor de date. Cheltuiala se datorează nu numai costurilor sistemelor în sine, ci, de asemenea, timpului necesar pentru a încărca datele în mai multe sisteme.

În consecință, tehnologia volumelor mari de date a făcut necesară rularea mai multor sarcini de lucru eterogene pe o singură infrastructură care este suficient de flexibilă pentru a se putea ocupa de toate aceste sarcini de lucru. Provocarea nu este de a construi un sistem care să fie ideal pentru toate sarcinile de procesare. În schimb, este nevoie ca arhitectura sistemului de bază să fie suficient de flexibilă astfel încât componentele construite pe infrastructura acesteia, pentru realizarea diferitelor tipuri de sarcini de procesare, să poată fi reglate pentru a rula eficient aceste sarcini de lucru diferite.

III.1. Tehnici și algoritmi de modelare a proceselor de afaceri

III.1.1. Integrarea inteligenței afacerii în managementul proceselor de afaceri

Realizând o integrare a managementului proceselor de afaceri, cu inteligența afacerii și unele aplicații de afaceri, într-o inițiativă a cumparatorului, se va putea atinge avantajul competițional dorit. Este necesară interoperabilitatea structurată între managementul proceselor de afaceri al companiei, inteligența afacerii și aplicațiile afacerii. Această interoperabilitate trebuie să fie axată pe afacere, ceea ce înseamnă că inițiativa trebuie să pornească de la o graniță a proceselor: arhitectura afacerii, modelul afacerii și valoarea lanțului afacerii.

Alinierea managementului procesului de afaceri și a inteligenței afacerii poate fi văzută din două perspective principale: injectarea inteligenței afacerii în procesele de afacere pentru îmbunătățirea luării deciziilor și oferirea unei viziuni din interior spre exterior asupra performanțelor de management al lanțului decizional. Acestea reprezintă două moduri de a privi aceeași problemă, și anume aceea de a lega informațiile și rezultatele procesului, accesibile prin managementul procesului de afaceri cu capacitatea inteligenței afacerii de a lua decizii.

Utilizarea doar a managementului proceselor de afaceri nu este suficientă, pentru că nu se poate vedea contextul dinamic al proceselor. Prin adăugarea inteligenței afacerii se poate obține o buclă închisă pentru managementul performanței, cazul în care valorile sunt comparate cu obiectivele de afaceri, precum și rezultatele sunt refolosite pentru a îmbunătăți procesele și deciziile.

Dinamica contextului poate fi înțeleasă completând managementul proceselor de afaceri cu mecanismul analitic al inteligenței afacerii. În mod egal, pentru a obține potențialul maxim al inteligenței afacerii este necesară atât existența unui model al afacerii, cât și o perspectivă a procesului de afaceri. Doar utilizând împreună managementul proceselor de afaceri și inteligența afacerii se pot atinge performanțe sporite.

Multe procese de afaceri necesită luarea unei decizii de către utilizatori, ca de exemplu aprobarea unei cereri de achiziție. Deseori, decizia nu a fost modelată ca o parte a procesului, lăsând utilizatorul să se bazeze pe *spreadsheet-uri*, convorbiri telefonice, e-mailuri, sau pe experiență pentru a determina cea mai bună decizie în contextul respectiv.

În situații cu o gamă largă și variată de posibilități, această lipsă a suportului în luarea deciziei conduce către o inconsistență dezastruoasă. Introducerea analizei inteligenței afacerii în proces îmbunătățește și accelerează luarea deciziilor, oferind rezultate mai bune (Fig. 1.).

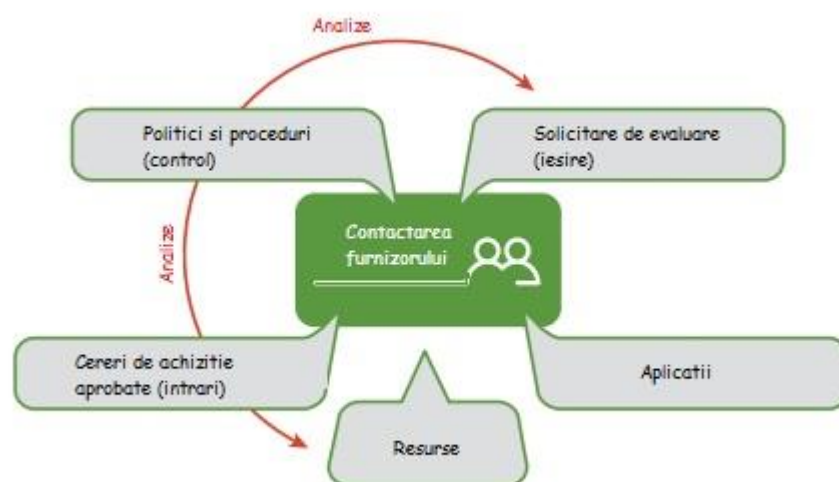


Fig. 1. Analize care se integrează în procesele de afaceri – activate de contactare a furnizorilor [MKER]⁷²

Regulile în afaceri sunt adesea folosite pentru a putea insera inteligența afacerii în managementul proceselor de afaceri și, astfel, să se ofere suportul luării deciziilor. Aceasta este o abordare validă, care face procesul să fie mai flexibil, oferind capacități analitice adiționale și care ascunde mult din complexitatea des întâlnită în inteligența afacerilor.

Definirea regulilor este un proces complex, care presupune echilibrarea priorităților în diferite segmente ale companiei. Aceste reguli pot ajuta oferind posibilități de măsurare și cuantificare, transparentă și eficientă. Din nefericire însă, regulile nu sunt suficiente, acestea

⁷² [MKER] Marc Kerremans, Nicholas Kitson - Aligning Business Process Management and Business Intelligence to Achieve Business Process Excellence

neputând coopera cu scenariul tipic de proliferare al modelului afacerii, produselor, serviciilor, segmentelor de clienți, și profitul estimat așteptat de acționari. Totodată, acestea nu pot oferi abilități vitale de a înțelege schimbările de mediu sau oferirea unui răspuns la aceste schimbări.

Majoritatea inițiativelor de management al proceselor de afaceri efectuează măsurători pentru a evalua eficiența unui proces. Câteva inițiative au arătat că, totuși, multe pot fi realizate prin setarea măsurătorilor unui proces în granița contextuală a unui model de afacere. Devine apoi posibil de înțeles impactul unui proces dat asupra performanțelor generale ale afacerii, fie asupra obiectivelor strategice generale, fie asupra unei campanii specifice de afacere. Această înțelegere poate conduce la o luare de decizii mult mai bună la nivelul organizațional (Fig. 2.).

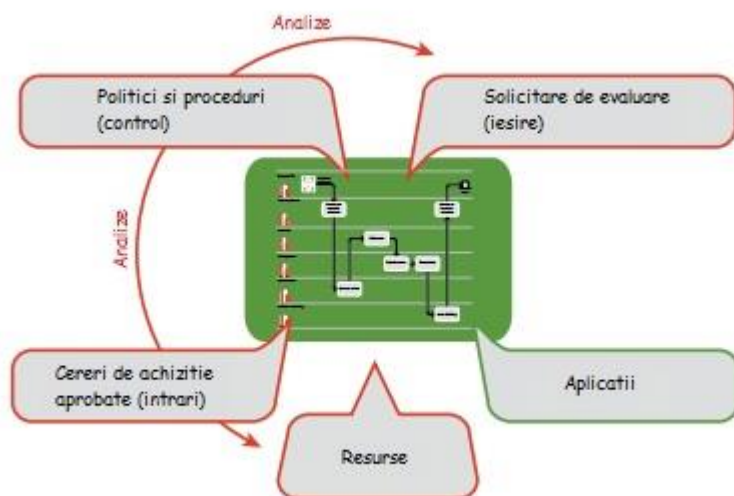


Fig. 2. Analiza despre procesele de afaceri – procesul de achiziții [MKER]⁷³

Rezultatele monitorizării performanțelor pot deveni și mai importante dacă sunt folosite pentru ajustarea proceselor de afaceri și a obiectivelor. Această buclă închisă a managementului performanțelor poate adesea să includă intervenția umană pentru a îmbunătăți modul în care sunt luate deciziile. Feedback-ul câștigat poate de asemenea să conducă afacerea către o automatizare a procesului de afaceri. Dintr-o perspectivă tehnologică, soluțiile de management al performanțelor în buclă închisă necesită cel puțin o combinație a managementului procesului de afaceri și a inteligenței afacerii, și câteodată și a tehnologiei procesării evenimentelor complexe (*Complex Event Processing*).

⁷³ [MKER] Marc Kerremans, Nicholas Kitson - Aligning Business Process Management and Business Intelligence to Achieve Business Process Excellence

III.1.2. Integrarea tehnologiei volumelor mari de date și a depozitelor de date

Dacă utilizatorii doresc să construiască sisteme analitice complexe având la bază tehnologia volumelor mari de date, este esențial ca acestea să aibă primitive de nivel înalt care să precizeze nevoile lor în astfel de sisteme flexibile. Cadrul MapReduce a fost extrem de valoros, dar reprezintă doar un prim pas. Chiar și limbajele declarative care îl exploatează, cum ar fi Pig Latin (limbajul platformei de nivel înalt Pig, utilizată pentru crearea programelor MapReduce folosite cu Hadoop), sunt la un nivel destul de scăzut, atunci când vine vorba de sarcini de analiză complexe. Specificații declarative similare sunt necesare la un nivel mai ridicat pentru a satisface nevoile de programare și compoziția acestor sisteme de analiză. Pe lângă necesitățile tehnice de bază, există, de asemenea, și un impuls de afaceri foarte puternic. Întreprinderile, de obicei, vor externaliza prelucrarea volumelor mari de date, sau multe aspecte ale acesteia.

Specificatiile tehnice sunt necesare nu numai pentru a cunoaște structura sistemului, dar și pentru operațiunile individuale în sine. Fiecare potențială operațiune (curățenie, extracție, modelarea etc.) rulează pe un set foarte mare de date. Mai mult decât atât, fiecare operațiune în sine este suficient de complexă și există multe opțiuni și optimizări posibile în modul în care este pusă în aplicare.

În bazele de date, este necesar un efort de lucru considerabil pentru optimizarea individuală a operațiunilor, cum ar fi joncțiunile. Este bine cunoscut faptul că pot exista mai multe ordine de mărime diferență, între costul a două moduri diferite de a executa aceeași interogare. Din fericire, utilizatorul nu trebuie să facă această alegere – aceasta fiind îndeplinită de sistemul de baze de date.

În cazul tehnologiei volumelor mari de date, aceste optimizări pot fi mai complexe, deoarece nu toate operațiunile de intrare/ieșire vor fi la fel de intensive ca în cazul bazelor de date. Unele operațiuni pot fi, dar altele pot fi intensive pentru procesor, sau pentru mai multe resurse. Așadar, tehnicile standard de optimizare a unei bazei de date nu pot fi utilizate direct. Cu toate acestea, ar trebui să fie posibil să se dezvolte noi tehnici pentru operațiunile cu volum mare de date, inspirate de tehnicile bazelor de date.

Însuși faptulul că analiza volumelor mari de date implică de obicei mai multe etape, evidențiază o provocare care apare în mod curent în practică: sistemele de producție trebuie să ruleze sisteme complexe analitice, sau fluxuri de lucru, la intervale de rutină, de exemplu, oră de

oră, sau zi de zi. Datele noi trebuie să fie contabilizate, luând în considerare rezultatele unei analize prealabile și a datelor pre-existente. Și, bineînțeles, proveniența trebuie să fie păstrată, și trebuie să includă fazele procesului analitic. Sistemele actuale nu oferă nici un pic de sprijin pentru astfel de sisteme bazate pe tehnologia volumelor mari de date, iar acest lucru este, în sine, un obiectiv provocator.

Când structura datelor pare să fie aleatorie (varietatea), când viteza fluxului de date este în continuă creștere (viteza), când cantitatea de informații este în creștere în fiecare secundă (volumul) și când există informații suplimentare ascunse în date (valoarea), doar o singură soluție poate fi aleasă pentru a gestiona acest haos: tehnologia volumelor mari de date. Această sintagmă a fost atât de mult promovată de companiile mari de software, că se pare că nu există o soluție software care să mai fie viabilă dacă nu are capacități de analiză a volumelor mari de date.

Adevărul este că există unele domenii, cum ar fi telecomunicațiile, rețelele sociale, resursele umane, etc., care sunt predispuse celor patru V (varietate, viteză, volum, valoare). Desigur, nu numai aspectele domeniului în care se aplică tehnologia contează. Depinde foarte mult dacă datele sunt istorice sau nu, dacă ar trebui să fie analizate în mod continuu, dacă sunt implicate în procesele decizionale, dacă sunt strategice sau secrete, dacă sunt structurate, semi-structurate sau nestructurate, etc.

Utilizarea depozitelor de date

Depozitele de date oferă arhitecturi și instrumente managerilor de afaceri pentru a organiza în mod sistematic, a înțelege, și a folosi datele lor pentru a lua decizii strategice. Sistemele de depozitare a datelor sunt instrumente valoroase în lumea competitivă de astăzi. În ultimii ani, multe firme au cheltuit milioane de dolari în construirea depozitelor de date la nivel de întreprindere. Mulți oameni simt că, odată cu creșterea concurenței în fiecare industrie, depozitale de date sunt cea mai recentă armă de marketing care trebuie exploatată, pentru a păstra clienții, învățatând cât mai multe despre nevoile lor.

În concluzie, un depozit de date este un magazin semantic consistent de date, care servește ca o implementare fizică a unui model al datelor suport de decizie. Acesta stochează

informațiile unei întreprinderi care trebuie să ia decizii strategice. Un depozit de date este, de asemenea, de multe ori privit ca o arhitectură, construit prin integrarea datelor din mai multe surse eterogene pentru a sprijini interogările structurate și / sau ad-hoc, raportarea analitică, și luarea deciziilor.

Pe baza acestor informații, am putea vedea tehnologia depozitelor de date ca fiind procesul de construire și utilizare a depozite de date. Construirea unui depozit de date necesită o curățare a datelor, integrare a datelor, și o consolidare a datelor. Utilizarea unui depozit de date necesită de multe ori o colecție de tehnologii de asistare a deciziei. Acest lucru permite utilizatorilor specializați (de exemplu, manageri, analiști, și directori) să utilizeze depozitul pentru a obține rapid și comod o prezentare a datelor, precum și pentru a lua decizii pe baza informațiilor din depozit.

Unii autori se folosesc de termenul de depozit de date pentru a se referi doar la procesul de construcție a depozitului de date, în timp ce termenul de SGBDD (sistem de gestiune al bazelor de date depozit) este folosit pentru a se referi la gestionarea și utilizarea depozitelor de date.

Multe organizații folosesc informațiile din depozite pentru a sprijini activitățile de luare a deciziilor de afaceri, inclusiv:

1. creșterea orientării către client, care include analiza modelelor de cumpărare a clientului (cum ar fi preferințele de cumpărare, timpul de cumpărare, ciclurile bugetare, precum și apetitul pentru cheltuieli);
2. repoziționarea produselor și gestionarea portofoliilor de produse prin compararea performanțelor vânzărilor pe trimestru, pe an, și pe regiuni geografice, în scopul de a ajusta strategiile de producție;
3. analiza operațiunilor și căutarea surselor de profit;
4. gestionarea relațiilor cu clienții, efectuarea corecțiilor de mediu, și gestionarea costului activelor corporative.

Stocarea datelor în depozite de date este, de asemenea, foarte utilă din punct de vedere al integrării bazelor de date eterogene. Organizațiile colectează, de obicei, diverse tipuri de date și încearcă să mențină baze de date mari, din surse de informare multiple, eterogene, autonome, și distribuite. Este de dorit, dar provocator, integrarea acestor date și oferirea unui acces ușor și

eficient la aceasta. A fost depus mult efort în industria bazelor de date și cea de cercetare pentru atingerea acestui obiectiv.

Abordarea tradițională a bazelor de date privind integrarea bazelor de date eterogene este de a construi integratori (sau mediatori) peste multiple baze de date, eterogene. Când o interogare este pusă pe un site client, un dicționar de metadate este folosit pentru a traduce interogarea în interogări adecvate pentru site-urile individuale eterogene implicate. Aceste interogări sunt apoi mapate și trimise la procesoare de interogare locale. Rezultatele returnate din diferite site-uri sunt integrate într-un set de răspuns global. Această abordare bazată pe interogări necesită procese de filtrare a informațiilor și procese de integrare complexe, și concurează cu site-uri locale pentru resurse de procesare. Este ineficient și destul de scump, în cazul interogărilor frecvente, mai ales pentru interogările care necesită agregate.

Depozitele de date oferă o alternativă interesantă la această abordare tradițională. Mai degrabă, decât folosirea unei abordări bazate pe interogări, depozitele de date oferă o abordare bazată pe date, în care informațiile din surse multiple, eterogene sunt integrate în prealabil și stocate într-un depozit pentru interogare directă și analiză.

Spre deosebire de bazele de date on-line de procesare a tranzacțiilor, depozitele de date nu conțin cele mai recente informații. Cu toate acestea, un depozit de date aduce o înaltă performanță sistemului de baze de date eterogene integrat deoarece datele sunt copiate, preprocesate, integrate, adnotate, rezumate, și restructurate într-un singur depozit de date semantic.

În plus, prelucrarea interogărilor în depozite de date nu interferează cu procesarea surselor locale. Mai mult decât atât, depozitele de date pot stoca și integra informații istorice și pot sprijini interogări multidimensionale complexe. Ca urmare, stocarea datelor folosind depozitele de date a devenit o metodă foarte populară în industria de specialitate.

III.1.3. Aplicarea tehnicilor de extragere a cunoștințelor din date

Utilizând tehnologia extragerii cunoștințelor din date, o vastă colecție de date se poate transforma într-o importantă sursă de informații valoroase.

Pe lângă datele din bazele de date relaționale, datele din depozitele de date și datele tranzacțiilor, există multe alte tipuri de date, care au forme și structuri diferite. Astfel de tipuri de date pot fi văzute în multe aplicații: date legate de timp sau secvențe (de exemplu, înregistrări istorice, date bursiere, și date de tip biologic), fluxuri de date (de exemplu, date provenind din supravegherea video și senzori, care sunt transmise în mod continuu), date spațiale (de exemplu, hărți), date de proiectare de inginerie (de exemplu, proiectarea clădirilor, componente de sistem, sau circuite integrate), date hipertext și multimedia (inclusiv text, imagine, video și date audio), grafice și date de rețea (de exemplu, sociale și rețele de informare), și date de pe Web (un imens depozit de informații pe scară largă, distribuite, și puse la dispoziție pe internet).

Aceste aplicații aduc noi provocări, cum ar fi modul în care se transferă datele care transportă structuri speciale (de exemplu, secvențe, arbori, grafice, și rețele) și semantică specifice (cum ar fi conținutul de comandă, de imagine, audio și video, și conectivitate), și cum să se exploateze modele care conțin astfel de structuri valoroase.

Diferite tipuri de cunoaștere pot fi exploatate de la aceste tipuri de date. În ceea ce privește datele temporale, de exemplu, putem extrage cunoștințe din datele bancare pentru a observa schimbarea tendințelor, care pot ajuta la programarea de observatori bancari în funcție de volumul de trafic al clientului. Datele bursiere pot fi exploatate pentru a descoperi tendințele care ar putea ajuta planificarea unor strategii de investiții (de exemplu, cel mai bun timp pentru a achiziționa un pachet de acțiuni la o companie). Am putea exploata fluxurile de date în rețelele de calculatoare pentru a detecta intruziunile pe baza anomaliilor fluxului de date, care pot fi descoperite prin clustering, construcții dinamice de modele de flux sau prin compararea benzilor frecvente curente cu cele de la un moment anterior. În ceea ce privește datele spațiale, am putea căuta modele care descriu modificările ratelor sărăciei metropolitane bazate pe distanța orașului față de autostrăzile majore. Relațiile dintre un set de obiecte spațiale pot fi examinate în scopul de a descoperi care subseturi de obiecte sunt autocorelate sau asociate spațial. Exploatând datele de tip text, cum ar fi literatura în domeniul *data mining* din ultimii zece ani, putem identifica evoluția subiectelor de interes din domeniu. Prin exploatarea comentariile utilizatorilor pe

produse (care sunt adesea prezentate ca mesaje text scurte), putem evalua sentimentele clienților și putem înțelege cât de bine un produs este îmbrățișat de o piață. Din datele multimedia, putem extrage imagini pentru a identifica obiecte și le putem clasifica prin atribuirea de etichete semantice sau tag-uri. Din datele video a unui joc de hochei, putem detecta secvențe video corespunzătoare golurilor. Exploatarea datelor web ne poate ajuta să învățăm despre distribuirea de informații cu privire la WWW, în general, caracterizarea și clasifica paginilor web, și descoperirea dinamică a web-ului și alte relații între diferite pagini web, utilizatori, comunități, și activități bazate pe web.

Este important să ținem cont de faptul că, în multe aplicații, sunt prezente mai multe tipuri de date. De exemplu, în extragerea informațiilor din web, există de multe ori date text și date multimedia (de exemplu, imagini și clipuri video), cum ar fi date grafice web, sau hărți de date. În bioinformatică, secvențe genomice, rețele biologice, și structuri spațiale 3-D ale genomilor pot coexista pentru anumite obiecte biologice.

Exploatarea mai multor surse de date complexe conduce de multe ori la descoperiri importante din cauza consolidării reciproce. Pe de altă parte, este de asemenea o provocare din cauza dificultăților în curățarea și integrarea datelor, precum și interacțiunile complexe între multiplele surse ale acestor date. În timp ce aceste date necesită facilități sofisticate de depozitare eficiente, regăsire, și actualizare, ele oferă, de asemenea un teren fertil și ridică probleme dificile de cercetare și de punere în aplicare pentru extragerea cunoștințelor din date.

Caracterizarea datelor este o centralizare a caracteristicilor generale sau a funcționalităților unei clase țintă de date. Datele corespunzătoare clasei specificate de utilizator sunt de obicei colectate de o interogare. De exemplu, pentru a studia caracteristicile produselor software cu vânzări care au crescut cu 10% în anul precedent, datele referitoare la astfel de produse pot fi colectate prin executarea unei interogare SQL în baza de date de vânzări.

Există mai multe metode pentru centralizarea datelor efective și caracterizarea ca rezumate de date, bazată pe măsurători statistice și grafice. Operațiunea OLAP roll-up bazat pe cubul de date poate fi utilizată pentru a efectua rezumarea datelor controlate de utilizator de-a lungul unei dimensiuni specificate. O tehnică de inducție orientată pe atribut poate fi utilizată pentru a efectua generalizarea datelor și caracterizarea pas-cu-pas fără a interacționa cu utilizatorul.

Discriminarea datelor este o comparare a caracteristicilor generale ale obiectelor de date de clasă țintă, cu caracteristicile generale ale obiectelor de la una sau mai multe clase contrastante. Clasele contrastante și clasa referință pot fi specificate de către un utilizator, iar obiectele de date corespunzătoare pot fi recuperate prin interogări de baze de date. De exemplu, un utilizator poate dori să compare caracteristicile generale ale produselor software cu vânzări care au crescut cu 10% anul trecut, cu caracteristicile celor cu vânzări care au scăzut cu cel puțin 30% în aceeași perioadă. Metodele utilizate pentru discriminarea datelor sunt similare cu cele utilizate pentru caracterizarea lor.

Extragerea cunoștințelor din date transformă o colecție mare de date în cunoștințe. Un motor de căutare (de exemplu, Google) primește sute de milioane de interogări în fiecare zi. Fiecare interogare poate fi privită ca o tranzacție în care utilizatorul descrie informațiile sale de interes. Ce cunoștințe noi și utile poate un motor de căutare învăța de la o astfel de colecție mare de interogări colectate de la utilizatori în timp? Interesant, unele modele găsite în interogările de căutare ale utilizatorilor pot dezvălui cunoștințe de neprețuit, care nu pot fi obținute prin citirea individuală a elementelor datelor. De exemplu, pentru a evalua tendințele unei gripe, Google utilizează termeni de căutare specifici ca indicatori ai activității de gripă. Acesta a constatat o relație strânsă între numărul persoanelor care caută informații legate de gripă și numărul de persoane care au de fapt simptome de gripă. Un model apare atunci când toate interogările de căutare referitoare la gripă sunt agregate. Utilizând și analizând datele de căutare Google, tendințele gripei pot estima activitatea de gripă de cu până la două săptămâni mai repede decât sistemele tradiționale. Acest exemplu arată cum extragerea cunoștințelor din date poate transforma o mare colecție de date în cunoștințe care pot ajuta la satisfacerea unei provocări globale.

III.1.4. Algoritmi de învățare profundă asistată (*deep machine learning*)

Scopul principal al învățării asistate este reprezentarea datelor de intrare și generalizarea modelelor învățate pentru a fi utilizate pe viitoarele date necunoscute. Calitatea reprezentării datelor are un mare impact asupra performanței de învățare asistate a datelor: o slabă reprezentare a datelor poate să reducă performanța chiar și a unui sistem avansat de învățare asistată, în timp ce o bună reprezentare a datelor poate duce la o înaltă performanță, chiar și

pentru un sistem relativ simplu de învățare asistată. Astfel, această tehnică, care se concentrează pe construirea de caracteristici și reprezentarea datelor utile din datele brute [DOMP], este un element important al învățării mașinii⁷⁴.

Tehnica consumă o mare parte din efort, într-o sarcină de învățare mașină, și este în mod cert un domeniu destul de specific și implică o intervenție umană considerabilă. De exemplu, histograma gradientilor orientați [DNTB]⁷⁵ și scala transformării elementului invariant [LWDG]⁷⁶ sunt algoritmi populari caracterizați de o metodă special dezvoltată pentru domeniul vizualizării pe calculator. Utilizarea tehnicilor asistate de calculator într-un mod mai automatizat și general, ar fi un progres major în procesul de învățare mașină deoarece acest lucru ar permite practicanților să extragă în mod automat astfel de caracteristici fără un efort uman direct.

Algoritmii de învățare profundă (*deep learning*) sunt o promițătoare cale de cercetare în extragerea automată a reprezentărilor de date complexe (caracteristicilor) la un nivel ridicat de abstractizare. Astfel de algoritmi dezvoltă o arhitectură stratificată ierarhic, de reprezentare și învățare a datelor astfel încât, caracteristicile unui nivel superior (mai abstract) sunt definite în funcție de caracteristicile unui nivel inferior (mai puțin abstract). Arhitectura de învățare ierarhică a algoritmilor de învățare profundă este motivată de inteligența artificială care emulează procesul de învățare profundă și extrage automat caracteristici și abstracțiuni din datele care stau la bază [ARK]⁷⁷.

Algoritmii de învățare profundă sunt destul de benefici atunci când se ocupă cu învățarea din cantități mari de date nesupravegheate, și învață, de obicei, reprezentarea datelor într-un mod înțelept și lacom. Studiile empirice au demonstrat că reprezentările de date obținute din prelucrarea datelor cu caracteristici neliniare (la fel ca în învățare profundă), de multe ori, oferă

⁷⁴ [DOMP] Domingos P, *A few useful things to know about machine learning*, Commun ACM 55(10), 2012

⁷⁵ [DNTB] Dalal N, Triggs B, *Histograms of oriented gradients for human detection*. In: *Computer Vision and Pattern Recognition*, 2005. CVPR 2005. IEEE Computer Society Conference On. IEEE Vol. 1. pp 886–893

⁷⁶ [LWDG] Lowe DG, *Object recognition from local scale-invariant features*. In: *Computer Vision*, 1999. The Proceedings of the Seventh IEEE International Conference On. IEEE Computer Society Vol. 2. pp 1150–1157

⁷⁷ [ARK] Arel I, Rose DC, Karnowski TP, *Deep machine learning-a new frontier in artificial intelligence research*, Research Frontier, 2010, IEEE Comput Intell 5:13–18

rezultate mai bune de învățare asistată, ca de exemplu, îmbunătățirea modelării clasificate, o calitate mai bună a probelor generate de modelele probabilistice generative, și proprietatea invariantă a datelor reprezentări [LBLL].⁷⁸

Soluțiile de învățare profundă au avut rezultate remarcabile în diferite aplicații de învățare automată, precum recunoașterea vorbirii și procesarea limbajului natural.

Volumele mari de date reprezintă domeniul general al problemelor și a tehnicilor utilizate pentru domenii de aplicare, care colectează și mențin volume masive de date brute pentru tehnologii mari, consumatoare de analize de date. Tehnologiile moderne de date, specifice unui domeniu, precum și creșterea resurselor de stocare, de calcul și de date au contribuit în mare măsură la dezvoltarea domeniului volumelor mari de date [NRC]⁷⁹.

Mineritul și extragerea modelelor semnificative din date de intrare masive pentru luarea deciziilor, predicție, și alte scopuri constituie nucleul analizei volumelor mari de date. Pe lângă analiza volume mari de date, domeniul ridică și alte provocări unice pentru învățarea mașinii și analiza datelor, inclusiv variația formatului datelor brute, deplasarea rapidă a fluxurilor de date, credibilitatea analizei datelor, sursele de intrare extrem de distribuite, zgomotoase și sărace în date de calitate, dimensionalitate ridicată, scalabilitatea algoritmilor, date de intrare neechilibrate, date nesupravegheate și ne-clasificate, date supravegheate / etichetate limitate, etc. Stocarea datelor adecvate, indexarea datelor / etichetarea și regăsire rapidă a informațiilor sunt alte probleme-cheie în domeniul analizei volumelor mari de date . În consecință, analiza de date și soluțiile inovatoare de gestionare a datelor sunt justificate atunci când se lucrează cu volume mari de date.

⁷⁸ [LBLL] Larochelle H, Bengio Y, Louradour J, Lamblin P (2009) Exploring strategies for training deep neural networks. J Mach Learn Res 10:1–40

⁷⁹ [NRC] National Research Council, *Frontiers in Massive Data Analysis*. The National Academies Press, Washington, DC, 2013

III.2. Solutii informatice pentru modelarea resurselor umane

III.2.1. Volume mari de date în domeniul resurselor umane

Odată cu apariția tehnologiei volumelor mari de date, resursele umane (HR) se confruntă cu o oportunitate fără precedent de a deveni mai axate pe date, analitice și strategice în felul în care se dobândește talent. Într-adevăr, HR a petrecut o mare parte din ultimul deceniu încercând să găsească modalități din ce în ce mai bune de a aplica valori și de analiză a deciziilor de capital uman.

Acum, tehnologii noi și puternice fac posibil ca HR-ul să amestece datele sale interne, cu o cantitate fără precedent de date din surse externe pentru a lua decizii de management al talentelor bazat pe dovezi și pentru a ridica profilul departamentului ca partener strategic pentru conducerea superioară. Acest lucru este valabil mai ales atunci când vine vorba de portaluri de locuri de muncă, care continuă să fie o sursă de top pentru angajatori în încercarea de a găsi candidați pentru locurile de muncă oferite.

Portalurile de locuri de muncă mai evolute reprezintă o mină de aur de informații, dar și o provocare tot mai mare pentru HR în încercarea de a găsi cadidatul ideal. Datorită numărului tot mai mare de furnizori de software care au colectat cu succes și analizează volume mari de date, HR poate apela acum la acești experți pentru asistența atât de necesară în lansarea propriilor inițiative de căutare a candidatului ideal în volumele mari de date.

Pornind de la definițiile volumelor mari de date prezentate la începutul lucrării, am prezentat în următoarea figură (Fig. 3.) ce reprezintă volumele mari de date în domeniul resurselor umane, exemplificând fiecare caracteristică fundamentală.

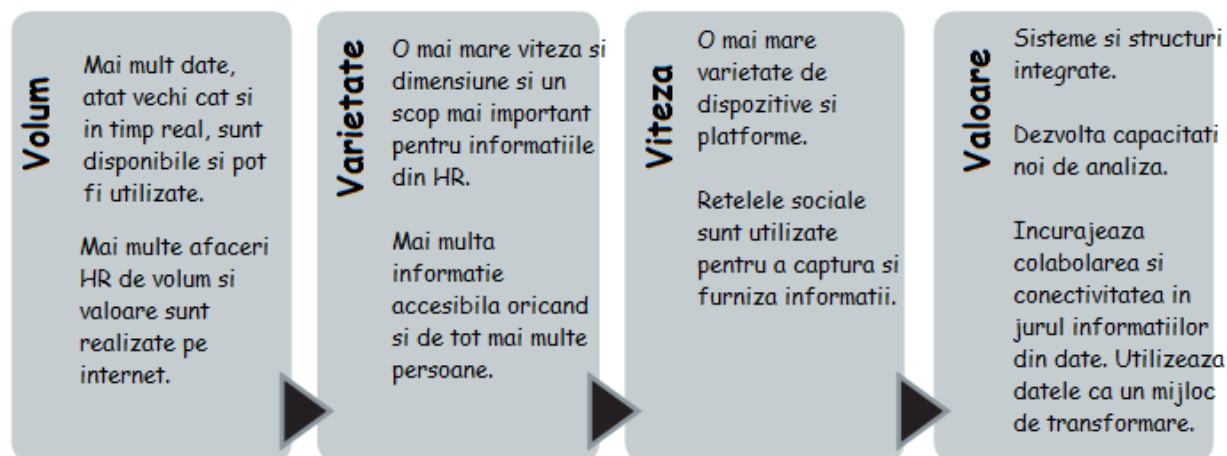


Figura 3. Cei 4 V ai volumelor mari de date în HR

Există o dezbatere serioasă cu privire la potențialul utilizării volumelor mari de date în domeniul resurselor umane. De exemplu, Thomas Davenport, un analist de renume de la Harvard, sugerează că organizațiile care se concentrează pe importanța datelor sunt cu 6% mai productive și cu 5% mai profitabile [DAVE13]⁸⁰. Cu toate acestea, există și scepticism în rândurile analiștilor din HR cu privire la utilizarea volumelor mari de date, unii sceptici susținând că volumele mari de date reprezintă doar un cuvânt pe care unii oameni îl folosesc pentru că vor să se simtă mai inteligenți.

Pentru domeniul resurselor umane, în special, volumele mari de date oferă o oportunitate istorică: posibilitatea de a lua cele mai riguroase decizii cu privire la capitalul uman. Volumele mari de date pot ajuta solidificarea reputației HR ca un partener de afaceri strategic care ia decizii pe bază de analiză, bazate pe o evidență, în special atunci când vine vorba de achiziționarea talentului uman. Utilizarea mijloacelor de angajare cu posibilitatea introducerii tehnologiei volumelor mari de date facilitează obținerea oamenilor potriviți în companie la momentul oportun, de prima dată. Aceasta înseamnă îmbunătățirea metodologiei de selecție a

⁸⁰ [DAVE13] Davenport, T. H. (2013). Analytics 3.0. Harvard Business Review, 91(12), 64-72

candidaților, accelerarea procesului de angajare și reducerea costurilor, toate acestea echivalând cu obținerea de avantaje competitive semnificative.

Departamentele de resurse umane au investit mult timp în aplicații software și în sisteme dedicate în întregime la captarea, raportarea și stocarea în siguranță a datelor sale despre persoane. Tot în cadrul resurselor umane s-au raportat, de asemenea, cele mai mari investiții din ultimul deceniu în soluții pentru gestionarea mai eficientă a capitalului uman. Potrivit Gartner, cheltuielile globale pe software pentru recrutarea inteligentă a persoanelor au crescut la 3,8 miliarde dolari în 2011, o creștere cu 15 la sută față de 2010 [GART]⁸¹.

În continuare sunt prezentate etapele evoluției resurselor umane față de volumele mari de date:

- Etapa 1: Utilizarea sistemelor de gestiune – Înainte de apariția sistemelor de gestiune a informațiilor HR și uneltelor similare, și disponibile pe scara largă, departamentele de resurse umane erau obligate să facă evaluarea candidaților și deciziile de angajare bazându-se pe experiențele trecute, opinii și intuiție.
- Etapa 2: Utilizarea datelor interne - Cum instrumentele eficiente de procesare și de colectare a datelor au devenit disponibile, HR a început să le utilizeze pentru a își îmbunătăți deciziile.
- Etapa 3: Utilizarea evaluării interne – În continuare, HR a îmbunătățit calitatea deciziilor sale prin examinarea datelor operaționale ale companiei. Aceasta a reprezentat începutul schimbării HR către adevărata recrutare inteligentă.
- Etapa 4: Aplicarea analizei descriptive – Următoarea schimbare HR a fost către analiza propriu-zisă a datelor - utilitatea informațiilor (cum ar fi ratele de uzură) și analizarea evenimentelor din trecut pentru a obține informații utile cu privire la modul de a face evaluările viitoare și de a lua deciziile de angajare.
- Etapa 5: Aplicarea analizei predictive - În cele din urmă, HR folosește tehnologia volumelor mari de date, pentru a determina probabilele rezultate viitoare ale deciziilor privind capital uman, ceea ce reprezintă un salt semnificativ în ceea ce privește extragerea valorii de date.

⁸¹ [GART] Gartner, Inc. – <http://www.gartner.com/it-glossary/big-data>

Tehnologia volumele mari de date oferă resurselor umane posibilitatea de a efectua procesul de decizie bazându-se pe dovezi, la un nivel cu totul nou de factori, folosind o cantitate fără precedent de date, de la o gamă largă de surse, dintre care unele nu au fost luate niciodată în considerare anterior. Tehnologia volumelor mari de date va permite HR să testeze teoriile, să rezolve probleme în mod proactiv și să efectueze analize predictive mai complexe legate de strategiile de eficientizare și de angajare.

Valoarea reală a volumelor mari de date este că oferă HR posibilitatea de a valorifica tehnologii mai puternice și o cantitate exponențial mai mare de date pentru a lua decizii bazându-se pe mai multe dovezi precise, și apoi să ia măsuri rapide cu privire la aceste decizii. Acest lucru creează avantaje competitive semnificative, nu numai în războiul pentru angajarea candidatului ideal, dar și în capacitatea organizației de a executa mai eficient strategiile sale de afaceri.

Angajatorii recrutează pentru un motiv: ei au o nevoie imediată de forță de muncă, care este critică pentru executarea obiectivelor lor de afaceri. Practicile clasice de selecție și de angajare duc inevitabil la angajarea cu întârzieri și la o selecție nu tocmai ideală de candidați, iar acest lucru duce la întârzieri în realizarea obiectivelor de afaceri.

Potențialul uriaș al tehnologiei volumelor mari de date în domeniul recrutării inteligente a forței de muncă a fost ilustrat și de Joseph Walker în articolul său publicat în *The Wall Street Journal* [WALK12]⁸². Articolul subliniază faptul că Xerox a redus rata de uzură a angajaților din call center aproximativ la jumătate după ce a început să folosească tehnologia volumelor mari de date pentru a evalua cererile de angajare depuse pentru obținerea unui nou loc de muncă în cadrul companiei. În timp ce, în trecut, compania angaja oameni pe baza experienței lor de muncă, dovezi noi au arătat Xerox că experiența nu contează în angajarea unui lucrător bun în call center (în acest caz, cel care va rămâne angajat suficient de mult, astfel încât Xerox să poată recupera investiția în instruirea noului angajat). În schimb, dovada a subliniat că personalitatea era mai relevantă decât experiența. După o perioadă de probă, informațiile furnizate de datele procesate, s-au dovedit a fi corecte, și Xerox a început angajarea tuturor angajaților săi din call center bazându-se pe această recomandare oferită de informațiile extrase din date.

⁸² [WALK12] Joseph Wlaker – Meet the New Boss: Big Data – *The Wall Street Journal*, 20 Sept. 2012

Desigur, găsirea surselor potrivite pentru recrutarea personalului este la fel de importantă ca și găsirea tipului potrivit de persoană, și utilizarea tehnologiei volumelor mari de date poate fi de neprețuit în această calitate. Tehnologia volumelor mari de date poate demonstra angajatorilor cu precizie care sunt sursele care au cea mai mare probabilitate de generare a tipurilor ideale de candidați.

Angajatorii trebuie să gestioneze un amestec de strategii pentru a construi adevărate modele de recrutare. Aceștia trebuie să fie atenți să aloce resurse precum timp, bani și forță de muncă activităților care produc rezultatele optime. Tehnologia volumelor mari de date poate fi utilizată pentru a identifica atât sursele cele mai eficiente, dar și care va fi volumul preconizat al fluxului de candidați. O prognoză anticipată pentru sursele optime și fluxul de candidați poate fi făcută acum cu o mai mare acuratețe și precizie. Avantajul competitiv, de a ști în avans care sunt sursele care vor produce cele mai bune rezultate, este clar.

În ciuda apariției relativ recentă a tehnologiei volumelor mari de date, multe companii pierd deja din vedere ceva important, elementul uman. Acest lucru este de înțeles, deoarece aspectele tehnice și tehnologice pot deveni copleșitoare destul de ușor. Dar adevărul este că, tehnologia volumelor mari de date este inutilă fără oamenii potriviți.

În esența sa, valoarea reală a tehnologiei volumelor mari de date constă în răspunsurile pe care le poate oferi, de perspectivele care pot fi derivate și de capacitatea acesteia de a permite utilizatorilor din mediul de afaceri să enunțe noi tipuri de întrebări la care nu ar fi putut afla răspuns în prealabil. Tehnologia își atinge pe deplin potențialul doar atunci când aceste informații sunt analizate, interpretate, raportate și puse în folosință de angajații unei organizații. În acest sens, oportunitatea oferită de această tehnologie ar putea fi ușor irosită dacă HR nu implementează informațiile oferite, în special în ceea ce privește recrutarea inteligentă a angajaților.

Nu este o exagerare să spunem că HR va avea un aport important în a își ajuta organizațiile să găsească oamenii potriviți și cu aptitudinile potrivite pentru a putea exploata beneficiile oferite de tehnologia volumelor mari de date. Conform unui raport publicat anii trecuți de McKinsey, acest lucru nu este foarte ușor de realizat [MCK11]⁸³. Acest raport afirmă

⁸³ [MCK11] McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity – June 2011

că: „Principalul motiv pentru care tehnologia volumelor mari de date nu va putea fi exploatată la maxim îl reprezintă lipsa persoanelor competente, cu cunoștințe avansate în statistică și sisteme inteligente, care să o exploateze și a managerilor și analiștilor care să știe cum conducă o organizație profitând de informațiile valoroase oferite de tehnologia volumelor mari de date.” Analistii McKinsey preconizează că sunt necesari încă 1,5 milioane de manageri, doar în Statele Unite ale Americii, pentru a analiza, interpreta și lua decizii pe baza informațiilor oferite de tehnologia volumelor mari de date. Evident, tot HR o să trebuiască să găsească aceste persoane și să asiste crearea echipelor de lucru folosind mixajul perfect de analize și expertize tehnice pentru a dezlănțui ulterior puterea tehnologiei volumelor mari de date.

A venit timpul pentru HR să se ridice la nivelul unui partener de afaceri strategic prin utilizarea tehnologiei volumelor mari de date și să îmbunătățească strategiile de recrutare a angajaților la un nou nivel de succes.

III.2.2. Procesarea datelor în domeniul resurselor umane

Imaginați-vă o lume în care volumele mari de date permit o analiză predictivă care poate identifica viitoarele tendințe de afaceri, determină calitatea cercetării, prevenirea bolilor, prezicerea și prevenirea crimelor, și determinarea condițiilor de trafic în timp real. Sau, în contextul resurselor umane (HR), arată de ce un agent de vânzări își surclasează colegii, sau prezice dacă un candidat se va integra bine într-o organizație, etc.

Această oportunitate a fost explicată și de specialistul Oracle în strategii *big-data*, Paul Sonderegger: “Tehnologia volumelor mari de date este pretabila oricărei companii care poate observa setul de date generat de interacțiunea dintre comerciant și cumpărător și care își propune să experimenteze rafinarea, analiza, și injecția datelor înapoi în procesele de afaceri pentru a face și a decide lucrurile mai bine.” [SOND13]⁸⁴

Din cauza puterii deținute de volumele mari de date, Oracle investește în dezvoltarea de soluții care extind gama de date analizată la prezentarea de informații de afaceri pentru a include

⁸⁴ [SOND13] Paul Sonderegger – Forbes – Pokerbots Bet On Big Data Strategy, 11/21/2013

surse non-tradiționale de informații - inclusiv date structurate și nestructurate; și informații din surse externe de încredere, inclusiv arhive Oracle Cloud și terțe părți interne.

În prezent, informațiile cu privire la cererea și oferta de pe piața forței de muncă sunt stocate electronic ca CV-uri sub formă de baze de date de text. Aceste date, de obicei semi-structurate, provin de la portaluri și site-uri de recrutare. Dar există o cantitate mare de informații, în formă nestructurată, pe rețelele de socializare, platformele colaborative ale universităților și forumuri de specialitate. Aceste date sunt nestructurate. În scopul de a utiliza atât datele semi-structurate și nestructurate, este necesar să se utilizeze metodele și tehnicile de prelucrare în paralel, extracție, curățare, transformare și integrare într-o bază de date NoSQL. Dificultatea problemei în acest caz este de a analiza și de a identifica soluții și tehnologii pentru volumele mari de date, care pot fi aplicate pentru organizare și prelucrare.

Din cauza complexității tehnologiile care vor fi utilizate, și datorită schimbărilor rapide de pe piața forței de muncă, crearea unei arhitecturi care să permită introducerea de noi surse de date, cu capacitatea de a integra surse multiple și eterogene, care include un nivel complex de analiză, modelare și determinare a profilelor, a condus la crearea unui management bazat pe cunoaștere resurselor umane. Din acest punct de vedere, dificultatea constă în alegerea elementelor și construirea unei platforme care să permită procesarea paralelă eficientă, extragerea de informații în timp util, analiza de date interactive și îndeplinirea cerințelor de performanță impuse de paradigma analizei volumelor mari de date (Big Data Analytics).

Datorită volumului tot mai mare de date personale colectate, stocate și disponibile cu ajutorul internetului, resursele umane au depășit capacitatea umană de înțelegere fără ajutorul puternicelor tehnologii inteligente. Astfel, în loc să fie bazat pe informații relevante, deciziile importante se fac în ceea ce privește recrutarea, în mod intuitiv, subiectiv sau pe baza unor criterii stabilite, fără însă a lua în considerare complexitatea naturii și comportamentului uman. Pentru a obține informații relevante, metode cum ar fi analiza multivariată ar trebui să fie utilizată pentru prelucrarea datelor, extragerea cunoștințelor, metode statistice și metode matematice care pot fi aplicate la volume mari de date. Pentru aceste aplicații, datele trebuie să fie bine organizate și indexate astfel încât să asigure ușurința de utilizare și regăsirea ușoară a informațiilor. Studii recente orientate spre organizarea și prelucrarea datelor de la portaluri de

recrutare [NERM14, EQU14], se referă la importanța acestei analize pentru procesul de selecție și impactul pe care aceste tehnici le au asupra performanței afacerii⁸⁵.

În ceea ce privește determinarea profilelor candidaților, există studii publicate [SADA13, JANT11] privind aplicarea algoritmilor de extragere a cunoștințelor din date (arbori de decizie, reguli de asociere, grupare) pentru selecția candidaților și pentru a determina metodele de pregătire pentru personalul recrutat⁸⁶. Cu toate acestea, aceste studii nu iau în considerare datele de la rețelele sociale și platformele de colaborare, sau din alte surse cum ar fi universitățile sau forumurile.

Prelucrarea informațiilor text și aplicabilitatea tehnicilor de extragere a cunoștințelor pe date din aceste surse sunt luate în considerare din ce în ce mai mult. Au fost dezvoltate numeroase metode de *text mining*, dar, de obicei, ele sunt orientate pe documentele de selecție (în cazul în care interogarea este considerată un furnizor de constrângeri) sau pe documentele de evaluare (în cazul în care interogarea este folosită pentru a clasifica documentele în ordinea relevanței) [GYOR10]⁸⁷.

Scopul este de a prelua cuvinte cheie dintr-o interogare a documentelor text și de a evalua fiecare document, în funcție de cât de mult satisface interogarea. În acest fel este evaluată relevanța unui document pentru interogarea efectuată. O altă metodă de clasificare a documentelor este modelul spațiului vectorial [KAO07, SRIV]⁸⁸. Aceasta implică reprezentarea

⁸⁵ [NERM14] C.Nermey - How HR analytics can transform the workplace

[EQU14] eQuest Headquarters - Big Data: HR's Golden Opportunity Arrives

⁸⁶ [SADA13] L.Sadath - Data Mining: A Tool for Knowledge Management in Human Resource, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-2, Issue-6, April 2013

[JANT11] H. Jantan, A. Hamdan, Z. Ali Othman - Data Mining Classification Techniques for Human Talent Forecasting, Knowledge-Oriented Applications in Data Mining, InTech Open, 2011, ISBN 978-953-307-154-1

⁸⁷ [GYOR10] C.Györödi, R.Györödi, G.Pecherle, G. M. Cornea - Full-Text Search Engine Using MySQL, Journal of Computers, Communications & Control (IJCCC), Vol. 5, Issue 5, December 2010, pag. 731-740

⁸⁸ [KAO07] A.Kao, S. Poteet - Natural Language Processing and Text Mining, Springer-Verlag London Limited 2007, ISBN 1-84628-175-X

unui document și vectori de interogare și utilizarea unei măsuri ca un etalon adecvat pentru a determina compatibilitatea vectorului de interogare și a vectorului de documente. Clasificare automată este un punct important în *text mining*, pentru că atunci când există un număr mare de documente on-line, posibilitatea de organizare automată a acestora în clase pentru a facilita regăsirea documentelor și analiza lor este esențială.

Pentru dezvoltarea de software, există în prezent tehnologii de inteligența afacerii (*business intelligence*) care pot fi utilizate. De asemenea, evoluțiile actuale din domeniul tehnologiei informațiilor au condus la apariția unor concepte și modalități noi de organizare și de prelucrare, în vederea îmbunătățirii accesului la date. Arhitectura *cloud* pentru aplicații cu putere de calcul, baze de date, stocare și software coexistă într-o rețea complexă și completă de servere care oferă utilizatorilor informații ca un serviciu, accesibile prin internet folosind dispozitive mobile. O astfel de arhitectură flexibilă care permite conectarea mai multor tipuri de subsisteme poate fi utilizată pentru a crea o platformă pentru recrutare. Există, de asemenea platforme de date mari disponibile în arhitectura *cloud computing* care pot fi utilizate și adaptate la necesitățile de realizare a prototipului.

III.2.3. Predicții în domeniul resurselor umane

Rețelele sociale prin volumul mare de date generat pot oferi perspective valoroase despre comportamentele oamenilor, cum ar fi probabilitatea lor de a se angaja în comportamente de risc sau contractarea unei boli. Deși se află încă la început, studiile în acest domeniu oferă predicții promițătoare.

Tehnologiile media sociale au apărut rapid și au devenit o necesitate în viața de zi cu zi a tot mai multor persoane. În cel mai scurt timp de la apariția rețelelor sociale, cercetătorii și corporațiile și-au dat seama de valoarea studierii datelor generate de acestea. Acest lucru s-a dovedit a fi valabil în special în sănătate și medicină, în cazul în care datele furnizate de rețelele

[SRIV] A. Srivastava, M.Sahami - Text Mining: Classification, Clustering, and Applications. Boca Raton, FL: CRC Press. ISBN 978-1-4200-5940-3

sociale oferă promisiuni pentru monitorizarea și supravegherea comportamentelor de risc și a focarelor de boli la distanță [BRO13, CHEW10, YOUN14]⁸⁹.

Această lucrare oferă o privire de ansamblu asupra abordărilor actuale ale utilizării rețelelor sociale pentru a monitoriza și anticipa comportamentul persoanelor și oferă recomandări cu privire la instrumentele și abordările necesare pentru a promova acest domeniu.

Tehnologia volumelor mari de date poate conține nu numai date relaționale și structurate (așa cum se găsesc în mai multe seturi de date medicale și genetice care utilizează valorile obținute din măsurători cantitative), dar, de asemenea nestructurate (de exemplu, text liber) datele evaluărilor calitative și conversații pe rețelele de socializare [MUR13]⁹⁰.

De exemplu, site-urile de socializare și motoarele de căutare pot fi folosite pentru a colecta posturi nestructurate, mesaje, căutări, și actualizări care furnizează informații despre utilizatori. Rețelele sociale, precum Facebook și Twitter, permit utilizatorilor să comunice cu ușurință și în mod liber, trimițându-și imagini, mesaje scurte, link-uri, și altele. Astfel, site-urile de socializare au devenit o parte integrantă a cercetării volumelor mari de date și ca urmare a numărului mare de utilizatori (de exemplu, peste 500 de milioane de tweet-uri pe zi pe Twitter [TWI13]⁹¹, peste 2.7 miliarde de like-uri pe zi pe Facebook [DON12])⁹². Aceste date pot fi modelate alături de alte seturi de date și utilizate pentru a anticipa comportamentul persoanelor.

⁸⁹ [BRO13] Broniatowski, D.A. et al. (2013) National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. PLoS ONE 8, e83672

[CHEW10] Chew, C. and Eysenbach, G. (2010) Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS ONE 5, e14118

[YOUN14] Young, S.D. et al. (2014) Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev. Med. 63, 112–115

⁹⁰ [MUR13] Murdoch, T.B. and Detsky, A.S. (2013) The inevitable application of big data to health care. JAMA 309, 1351–1352

⁹¹ [TWI13] Twitter, Inc. – United States Securities and Exchange Commission, October 3, 2013

⁹² [DON12] Donna Tam, Facebook processes more than 500 TB of data daily, CNET, August 22, 2012

Cercetări recente au aratat că interacțiunile oamenilor pe rețelele de socializare pot fi analizate pentru a furniza informații psihologice despre atitudinile și comportamentele acestora, inclusiv comportamentele legate de sănătate [MYS13, YOUN09]⁹³.

Utilizatorii rețelelor sociale devin tot mai confortabili oferind public din ce în ce mai multe tipuri de informații, inclusiv povești personale și informații de sănătate, furnizând astfel date care pot fi extrase, clasificate ca date psihologice și comportamentale și pot fi utilizate pentru analiză.

Deși metodele bazate pe volumele mari de date, generate de rețelele de socializare sunt fezabile și pot conduce la implementarea unor instrumente cu potențial uriaș, utilizarea acestora presupune disponibilitatea volumului mare de date și o frecvență ridicată de actualizare a acestuia.

Datele rețelelor sociale au influențat rapid cercetarea volumelor mari de date și au devenit unul din instrumentele principale utilizate în acest domeniu în curs de dezvoltare. Deoarece aceste tehnologii oferă bogate date psihologice și oamenii sunt dispuși în mod liber să împărtășească informații personale de sanatate pe rețelele de socializare, aceste tehnologii vor continua să fie monitorizate și explorate pentru potențialul lor în predicția stărilor de sănătate și comportamentale.

Înțelegând limitările cercetării bazate pe volumele mari de date, generate de rețelele sociale (de exemplu, validitatea datelor, lipsa de date, reprezentativitatea eșantionului) și metodele necesare abordării acestor limitări va îmbunătăți valoarea acestei cercetări. Stabilirea și documentarea unor metode de utilizare a rețelelor de socializare în cercetarea volumelor mari de date este importantă, deoarece datele sociale pot avea un impact mai larg și în alte domenii, în care aceste abordări nu au fost încă studiate sistematic.

Rețelele sociale nu sunt singurul furnizor de date utile tehnologiei volumelor mari de date. Atunci când se caută o persoană în vederea angajării, în general, se caută o persoană care are deja experiență în acel domeniu și astfel, atenția se focalizează pe analiza cv-urilor depuse de candidați pe platformele de recrutare.

⁹³ [MYS13] Mysli'n, M. et al. (2013) Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. J. Med. Internet Res. 15, e174

[YOUN09] Young, S. et al. (2009) Extrapolating psychological insights from Facebook profiles: a study of religion and relationship status. Cyberpsychol. Behav. 12, 347–350

Javid Muhammedali [JAV14] ne prezintă cazul unei companii multinaționale care este prezentă în topul Fortune 500 și care încearcă să recruteze persoane potrivite pentru a ocupa cele 400 de locuri disponibile în cadrul companiei.⁹⁴ În general, pentru un post scos la concurs, aplică în medie 250 de persoane, ceea ce înseamnă că o astfel de companie are de filtrat aproximativ 100.000 de cv-uri depuse. Dar această nu reprezintă întreaga bază de selecție a candidaților. Pe lângă aceasta, se mai adaugă toate cv-urile depuse de candidați cu ocazia recrutărilor precedente organizate de companie și milioanele de cv-uri depuse de candidați și disponibile pe site-urile de recrutare online. Din acest motiv, companiile au nevoie de soluții noi, bazate pe tehnologia volumelor mari de date, care să le poată permite să analizeze și, ulterior, să recruteze cel mai bun candidat disponibil pentru postul respectiv. Analizând volumul de cv-uri folosind tehnologia volumelor mari de date, companiile ar realiza o recrutare inteligentă, într-un timp mai scurt care le-ar oferi și un uriaș avantaj competițional.

Din păcate, analizarea CV-urilor candidaților depuse pe platformele de recrutare, nu reprezintă întotdeauna cel mai bun criteriu de selecție. Din fericire însă, tehnologia volumelor mari de date, oferă posibilitatea de a pătrunde în viața personală a candidatului și de a îi cunoaște mai bine capacitățile și personalitatea, depășind astfel, granița obișnuită reprezentată de curriculum vitae.

În cazul companiei Xerox [WALK12], recrutarea angajaților din call-center pe baza analizei cv-urilor (punând accent pe experiența similară deținută de candidați) s-a dovedit un eșec.⁹⁵ O analiză efectuată pentru a elucida cauza a arătat că experiența similară precedentă a candidaților nu este atât de relevantă în cazul unui angajat din call-center. Studiul a arătat că Xerox are nevoie să realizeze recrutările viitoare bazându-se mai mult pe personalitatea angajatului decât pe experiența sa, deoarece personalitatea poate spune cu o precizie mai mare dacă investiția companiei în pregătirea candidatului o să fie rentabilă sau dacă aceasta o să fie irosită, candidatul părăsind locul de muncă înainte de amortizarea ei. Datele au arătat că angajații, cu un tip de personalitate mai creativă, au tendința de a își păstra locul de muncă

⁹⁴ [JAV14] Javid Muhammedali – How Big Data is Impacting the Job-Hunting and Hiring Experience Today – Official Monster Blog, 30 Sept. 2014

⁹⁵ [WALK12] Joseph Wlaker – Meet the New Boss: Big Data – The Wall Street Journal, 20 Sept. 2012

suficient de mult încât investiția în pregătirea post-angajare să fie recuperată, în timp ce, cei cu un tip de personalitate mai curioasă, de cele mai multe ori schimbă mai repede locul de muncă.

Prin supunerea candidaților la o serie de teste și apoi urmărirea performanței lor la diferite locuri de muncă, Xerox a fost dezvoltat un model ideal al lucrătorului din call-center. Datele spun că persoanele care locuiesc în apropierea locului de muncă, nu au probleme legate de transport și utilizează una sau mai multe platforme de socializare, dar nu mai mult de patru. Deasemenea, candidatul ideal tinde să nu fie prea curios sau empatic, dar este mai creativ.

Utilizarea tehnologiei volumelor mari de date pentru o recrutare inteligentă a personalului s-a dovedit a fi de mare folos nu doar companiei Xerox, pentru recrutarea angajaților săi din call-center. În cazul companiei Richfield Management LLC (companie care are ca domeniu principal de activitate salubritatea), s-a apelat la utilizarea tehnologia volumelor mari de date pentru a evita angajarea unor persoane predispuse accidentelor de muncă și utilizării excesive a fondului destinat plății compensatorii. Astfel, pentru noii candidați înscriși pentru ocuparea unui loc de muncă, compania a pus mai mult accent pe evaluarea stării emoționale, a eticii în muncă și a predispoziției la utilizarea de droguri și alcool. În urma angajărilor efectuate pe baza noilor criterii de selecție a candidaților, compania a raportat o scădere cu 68% a plângerilor efectuate de angajații săi [WALK12]⁹⁶.

Utilizarea unor criterii prea specifice de selecție a candidaților poate expune compania riscului de încălcare a dreptului egalității de șanse. Chiar dacă, și neintenționat, criteriile de selecție exclud persoanele mai în vârstă sau minoritățile, acesta este un motiv suficient care poate expune compania unui litigiu. În cazul în care o practică de angajare este contestată în instanța de judecată ca fiind discriminatorie, compania trebuie să arate că criteriile pe care le utilizează asigură succesul în muncă. De exemplu, Matthew Camardella, un partener în cadrul firmei de avocatură Jackson Lewis LLP, specializat în a determina dacă dreptul egalității de șanse este respectat de angajatori, a afirmat că un număr tot mai mare de companii îi solicită ajutorul pentru a evalua dacă aplicațiile inteligente de recrutare pe care le folosesc respectă, sau nu, principiul egalității de șanse [WALK12]. Astfel, putem considera că utilizarea unui volum mai mare de date și a unor condiții mai restrictive de evaluare poate reprezenta o problemă juridică pentru companie. Din acest motiv, deși Xerox a descoperit că un candidat care locuiește mai departe de

⁹⁶ [WALK12] Joseph Wlaker – Meet the New Boss: Big Data – The Wall Street Journal, 20 Sept. 2012

locul de muncă, are o șansă mai mare de a demisiona, decât un candidat care locuiește mai aproape, compania nu utilizează această informație atunci când recrutează persoane noi, deoarece, cel puțin în Statele Unite ale Americii, această informație poate fi considerată o discriminare.

Testele de personalitate au o istorie lungă în angajare, dar importanța lor este tot mai mare. Aplicații software tot mai sofisticate au făcut posibilă evaluarea mai multor candidați, cumulara mai multor date din diverse surse și astfel s-a reușit o pătrundere tot mai adâncă în viața și în interesele applicantului.

Unele companii consideră că printre variabilele care ar trebui monitorizate se află și atitudinea față de consumul de alcool sau distanța de la domiciliul candidatului față de locul de muncă. Acest proces ar putea crea companiilor anumite probleme juridice în cazul în care procesul se finalizează cu respingerea minorităților sau cu a persoanelor cu handicap.

Noile instrumente de angajare sunt parte a unui efort mai larg de a colecta și analiza datele angajaților. La nivel global, cheltuielile pentru așa-numitul software de recrutare inteligentă a persoanelor, au crescut la 3,8 miliarde dolari în 2011, cu 15% față de 2010, potrivit firmei de cercetare Gartner [GART] .

IV. Studii de caz privind recrutarea avansată a forței de muncă

În procesul de management al cunoștințelor, tehnici de exploatare a datelor pot fi folosite pentru a extrage și a descoperi cunoștințe valoroase și semnificative într-o cantitate mare de date. În zilele noastre, extragerea de date a dat o mare îngrijorare și atenție în industria informațională și în societate, în ansamblul ei.

Printre cele mai importante sarcini în exploatarea datelor sunt clasificarea și predicția, descrierea conceptului, excluderea asociațiilor, analiza grupului, analiza similitudinilor, analiza tendințelor și evaluarea, analiza statistică și altele. Clasificarea și predicția sunt printre cele mai populare sarcini în exploatarea datelor, fiind utilizate pe scară largă în multe domenii, în special pentru analiza tendințelor și planificarea viitoare.

Există mai multe domenii care au adaptat această abordare pentru a rezolva problemele lor, cum ar fi în domeniul financiar, medical, marketing, telecomunicațiile, producția, sănătatea, relația cu clienții, etc. Cu toate acestea, cererea de exploatare a datelor nu a atras prea mult atenția în cadrul resurselor umane. Mai mult, cele mai multe cereri de predicție sunt folosite pentru a prezice cererea de angajare, rata, riscul, și altele; dar există studii destul de limitate asupra predicției umane. În plus, aplicațiile de predicție sunt dezvoltate în principal în domeniile de afaceri și industriale; și studiile sunt destul de restrânse în ceea ce privește implicarea talentului uman într-o organizație.

Datele din resurse umane pot reprezenta o resursă bogată pentru descoperirea de cunoștințe și pentru dezvoltarea sistemului de asistare a deciziilor.

IV.1. Studii de caz în domeniul resurselor umane

IV.1.1. Utilizarea tehnologiei volumelor mari de date în vederea recrutării avansate a forței de muncă

Extragerea cunoștințelor din date este una dintre tehnologiile emergente în domeniul instrumentelor de analiză a datelor, și are un impact semnificativ în domeniul resurselor umane în descoperirea modelelor și informațiilor ascunse în bazele de date, în scopul de a utiliza aceste informații pentru a dobândi cunoștințe care ajută factorii de decizie să dezvolte strategii și planurile de viitor ale pieței forței de muncă. Acest lucru este sugerat și de alți autori care spun despre extragerea cunoștințelor din date că este o abordare care primește în prezent o mare atenție și este recunoscută ca un instrument nou de analiză în curs de dezvoltare [HAND]⁹⁷.

În continuare vom prezenta o parte din problemele ridicate de procesul recrutării inteligente, rezolvabile prin utilizarea tehnologiei extragerii cunoștințelor din date.

Aplicațiile din domeniul resurselor umane care integrează tehnici de inteligența artificială pot fi utilizate pentru a ușura procesul decizional. Odată cu progresul tehnologiilor inteligente, au apărut tot mai multe tehnici care ar putea fi folosite și de aplicațiile din domeniul resurselor umane, ca de exemplu extragerea cunoștințelor din date. Această tehnică a fost dezvoltată pentru a facilita explorarea și analiza cantităților mari de date cu scopul de a descoperi reguli și modele similare. În realitate, astfel de date descoperite în datele resurselor umane pot reprezenta o resursă bogată pentru descoperirea de cunoștințe și un instrument de sprijin în luarea deciziilor. Până în prezent, tehnicile și aplicarea extragerii cunoștințelor din date nu a atras multă atenție în domeniul resurselor umane.

Citind tot mai multe articole care încurajează extragerea cunoștințelor din datele motoarelor de cautare, am decis să realizez un studiu de caz pentru a evalua relevanța datelor extrase. Astfel, am decis să evaluez frecvența de căutare a cuvântului “Paris” în decursul anului 2015. Pentru realizarea acestei evaluări am fost folosit Google Trends, care este un utilitar web public al Google Inc., bazat pe Google Search, care arată cât de des se introduce un anumit

⁹⁷ [HAND] Hand, D.J. (2007). Principles of Data Mining. *Drug Safety*, 30, 7, 621-622

termen în căutare, în raport cu numărul total de căutări, în diferite perioade timp, în diferite regiuni ale lumii, și în diferite limbi [WIKIa].⁹⁸

Graficul următor arată cât de mult a fost căutat cuvântul “Paris” folosind motorul de căutare Google. Am ales în mod special Parisul, datorită evenimentelor cu impact mondial major care au avut loc în capitala franceză în ultimul an.

Se poate observa cum au existat 2 momente care nu se încadrează în tiparul de căutare. Primul, înregistrat în luna ianuarie 2015, se datorează atentatului împotriva revistei Charlie Hebdo din 7 ianuarie 2015 [WIKIb]⁹⁹, în timp ce, al doilea se datorează atentatelor cu bombă din 13 noiembrie 2015 [WIKIc]¹⁰⁰.

Nivelul de interes



Figura IV.4. Nivelul de căutare al cuvântului “Paris” în motorul de cautare Google în 2015

Analizând în profunzime datele furnizate de Google Trends din perioada noiembrie 2015, vom găsi cuvântul “Paris” utilizat cel mai mult în următoarele întrebări: “Ce s-a întâmplat noaptea trecută în Paris?”, “Sinucigașul din Paris a fost un refugiat?”, “Ești în siguranță dacă mergi azi la Paris?”.

În continuare am selecționat termenii “oracle dba” (administrator baze de date oracle) pentru a analiza interesul pentru un astfel de job la nivel mondial. Deși nivelul de interes manifesta variații în decursul anului, acestea oscilează în jurul valorii de 80.

Nivelul de interes



Figura IV.5. Nivelul de căutare al termenilor “oracle dba” în motorul de căutare Google în 2015

În schimb, mai multe informații ne oferă analiza interesului regional asupra termenilor “oracle dba”. Analizând interesul la nivel mondial, se observă că această combinație de termeni este căutată în mod excesiv în India.

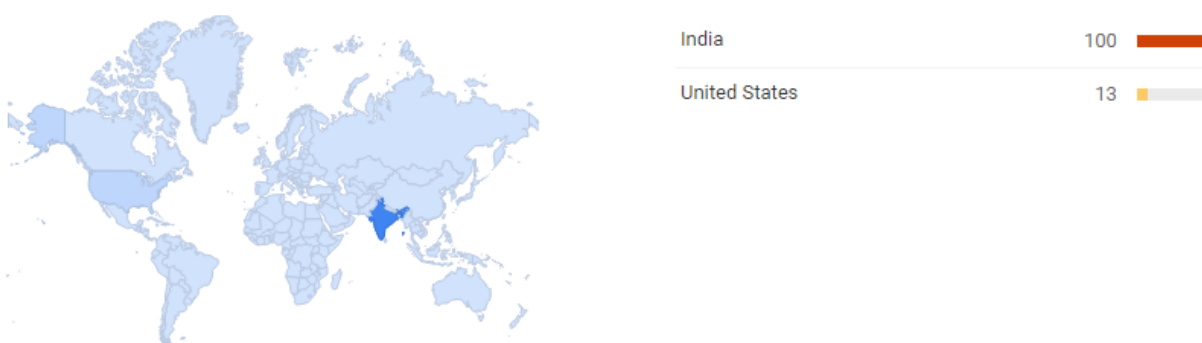


Figura IV.6. Nivelul de căutare al termenilor “oracle dba” în motorul de căutare Google în 2015

Analizând și mai în profunzime India, la nivel de oraș, vom observa că cel mai des căutarea acestor termeni are loc în orașele: Chinchwad, Hyderabad și Bangalore. Fapt deloc surprinzător ținând cont că India a devenit încă de mulți ani o alegere a companiilor care investesc în cercetare și dezvoltare [HNWS]¹⁰¹.

¹⁰¹ [HNWS] Hot News – India, magnetul marilor companii care investesc în cercetare și dezvoltare

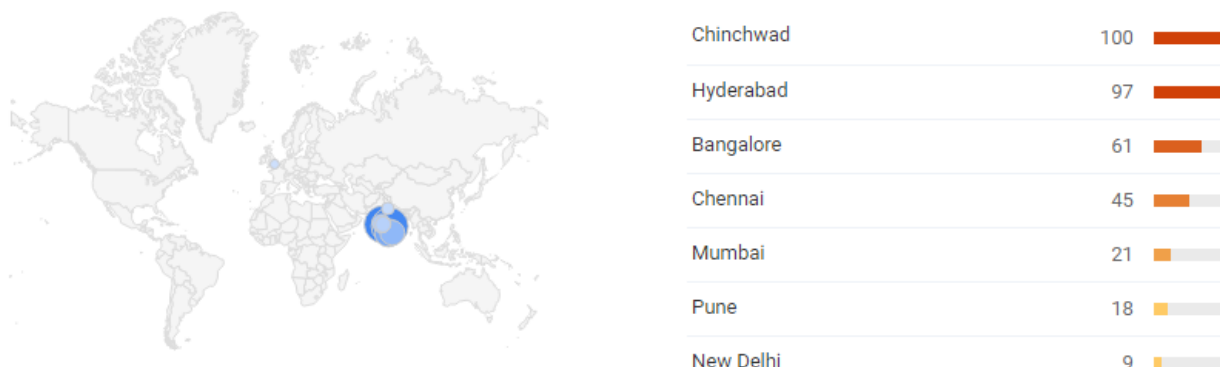


Figura IV.7. Nivelul de căutare al termenilor “oracle dba” în motorul de căutare Google în 2015

Mai mult, dacă se analizează și combinația de termeni în care regăsim “oracle dba” vom descoperi că aceste cuvinte au predominat în următoarele tipuri de interogări: “oracle dba jobs”, “oracle dba salary”, “oracle dba interview”. Acest lucru ne indică existență unui nivel ridicat de interes în ceea ce privește administrarea bazelor de date, foarte multe persoane încercând să descopere informații suplimentare privind domeniul, în special datorită numărului destul de redus al specialiștilor și a salariilor atractive.

Cunoașterea interesului cetățenilor pentru un anumit job, poate oferi avantaje suplimentare managerilor în luarea deciziilor corecte. De exemplu, graficul din figura IV.8. arată nivelul de interes la nivelul României privind termenii “call center”.

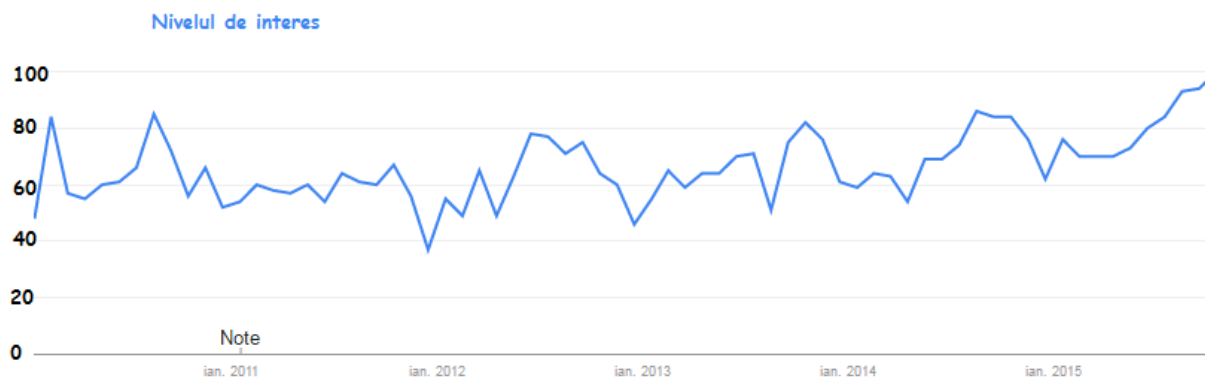


Figura IV.8. Nivelul de căutare al termenilor “call center” în motorul de căutare Google în perioada 2010 - 2015

După cum se poate observa, nivelul de interes a fost într-o permanentă creștere în anul 2015. La nivel de regiune, distribuția acestuia a fost următoarea:

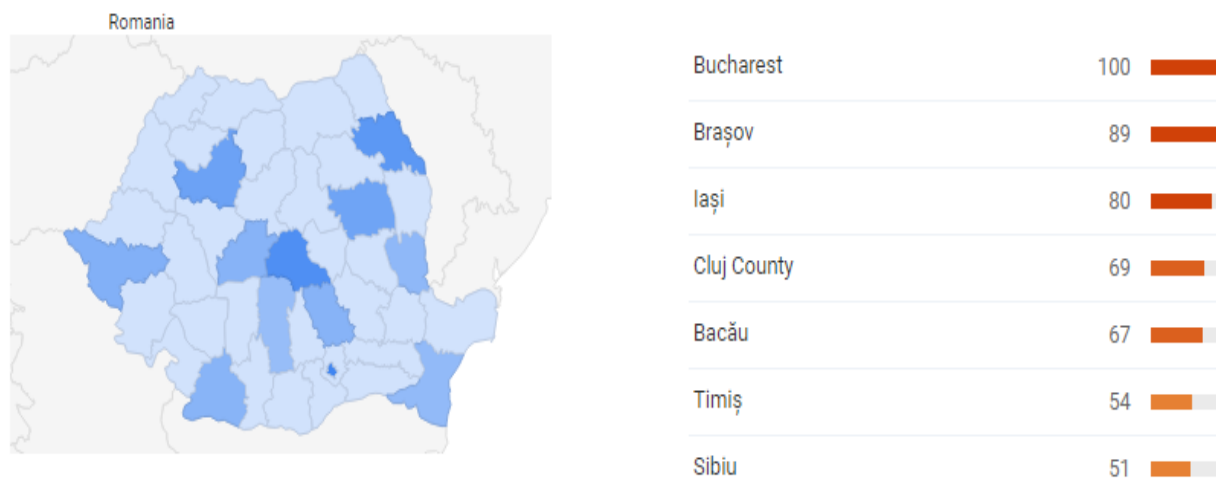


Figura IV.9. Distribuția nivelului de interes a cuvintelor “call center” în motorul de căutare Google în perioada 2010 - 2015

Concluziile studiului de caz

Cunoscând astfel de informații, managerul unei companii care activează în domeniul serviciilor de tip “call center”, poate să opteze sau renunțe la dezvoltarea companiei într-o nouă regiune. Astfel, în cazul în care compania dorește să își extindă activitatea în afara Bucureștiului

(pentru a beneficia de avantaje financiare precum: salarii mai mici, chirii mai mici, probabilitatea mai ridicată de rămânere a salariatului pe termen mai lung, etc), aceste statistici se pot dovedi extrem de utile pentru alegerea celei mai potrivite zone. Astfel, se poate opta pentru județul Brașov, în defavoarea județului Covasna, sau se poate alege județul Iași, în locul județului Suceava. Luând o astfel de decizie, compania ar putea să beneficieze de angajați cu experiență în domeniul “call center”, ceea ce ar reduce costurile de training al personalului. Deasemenea, în cazul în care compania dorește să angajeze un număr mai mare de salariați, o astfel de statistică o poate ajuta să estimeze dacă se va putea găsi personalul căutat în timp util.

Dacă rezultatul căutării este unul nefavorabil, o companie își poate schimba decizia de a recruta personal nou, și poate opta pentru achiziția unor servicii similare furnizate de alte companii prestatoare de servicii. O astfel de decizie, ar ajuta compania, nu doar să nu irosească timp prețios, ci și să câștige timp în fața competitorilor direcți.

Din nefericire, Google nu oferă suport API pentru Google Trends, ceea ce împiedică dezvoltarea unui produs care să includă astfel de facilități. O altă soluție ar fi interogarea manuală a datelor de căutare și interpretarea rezultatelor, dar nici această soluție nu este viabilă deoarece nici Google, nici un alt motor de căutare, nu oferă publicului un volum de date de căutare. Astfel, am decis să nu aprofundez această direcție în cadrul cercetării și să nu integrez informațiile extrase din date de căutare. Aprofundarea acestor date se poate realiza și integra în momentul în care Google va furniza suport API și pentru Google Trends.

IV.1.2. Utilizarea extragerii cunoștințelor din date în vederea recrutării avansate a forței de muncă

Sarcinile de exploatare a cunoștințelor din date sunt, în general, clasificate astfel: gruparea, asocierea, clasificarea și predicția [CHN, RANJ]¹⁰². De-a lungul anilor, extragerea cunoștințelor din date a evoluat și au apărut diverse tehnici noi pentru a îndeplini sarcinile care includ tehnici de baze de date orientate, statistice, învățare mașină (*machine learning*), de recunoaștere a tiparului, rețele neuronale, etc.

Bazele de date sau depozitele de date sunt bogate în informații ascunse, care pot fi utilizate pentru a asigura luarea de decizii inteligente. Decizia inteligentă se referă la capacitatea de a lua o decizie automat, care este destul de similară cu decizia umană. Clasificarea și predicție în învățarea mașină se numără printre tehnicile care pot produce decizii inteligente. În acest moment, multe tehnici de clasificare și de predicție au fost propuse de către cercetători în procesul de învățare mașină, precum recunoașterea modelelor și statisticile.

Clasificarea și predicția, în exploatarea datelor, sunt două forme de analiză a datelor, care pot fi utilizate pentru a extrage modele, pentru a descrie clase importante de date sau pentru a anticipa tendințele viitoare ale datelor [HJKM]¹⁰³. Procesul de clasificare are două faze: prima fază este procesul de învățare, în care, datele de instruire vor fi analizate prin algoritmul de clasificare. Astfel, modelul învățat sau clasificator este reprezentat sub forma unor norme de clasificare. În continuare, a doua fază este reprezentată de procesul de clasificare, în cazul în care datele de testare sunt utilizate pentru a estima acuratețea modelului de clasificare sau clasificator. În cazul în care precizia este considerată acceptabilă, regulile pot fi aplicate la clasificarea noilor date (Fig. III.3).

¹⁰² [CHN] Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems and Applications*, 34(1), 380-290

[RANJ] Ranjan, J. (2008). Data Mining Techniques for better decisions in Human Resource Management Systems. *International Journal of Business Information Systems*, 3(5), 464-481

¹⁰³ [HJKM] Han, J. & Kamber, M. (2001). Data mining: concepts and techniques (Morgan-Kaufman Series of Data Management Systems), Academic Press, San Diego

Mai mult, tehnicile care sunt utilizate pentru clasificarea datelor sunt arborii de decizie, metodele bayesiene, rețeaua Bayesian, algoritmi bazați pe reguli, rețelele neuronale, vectorii de suport mașină, regulile de asociere minieră, cel mai apropiat vecin k-NN, raționamentul bazat pe caz, algoritmi genetici, seturile brute și logica fuzzy.

În continuare în cadrul lucrării, voi folosi trei tehnici principale de clasificare și anume arborii de decizie, rețelele neuronale și cel mai apropiat vecin k-NN. Am ales acești algoritmi deoarece, arborele de decizie și rețele neuronale sunt găsite utile în dezvoltarea de modele predictive în multe domenii. Această afirmație este întărită și de alte studii similare [TSO]¹⁰⁴.

Avantajul tehnicii arborelui de decizie este acela că nu are nevoie de nici o setare de cunoștințe de domeniu sau de parametrii, și este adecvată pentru descoperirea de cunoștințe de explorare.

A doua metodă este cea a rețelelor neuronale, care au toleranță mare la zgomotul datelor, precum și capacitatea de a clasifica un model pe care acestea nu au fost instruite. Această tehnică poate fi utilizată atunci când avem puține cunoștințe despre relația dintre atribute și clase.

În continuare, tehnica celui mai apropiat vecin k-NN este o învățare bazată pe exemple, folosită pentru a măsura distanța metrică a similitudinii instanțelor.

Toate aceste trei tehnici de clasificare au propriile lor avantaje și dezavantaje, din aceste motive, lucrarea explorează aceste tehnici de clasificare a datelor în domeniul resurselor umane. Mai mult de atât, tehnica de extragere a cunoștințelor din date a fost aplicată în multe domenii, dar, așa cum susțin și alți autori, aplicarea ei în domeniul resurselor umane este destul rară [CHN].

¹⁰⁴ [TSO] Tso, G. K. F., & Yau, K. K. W. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32, 1761-1768

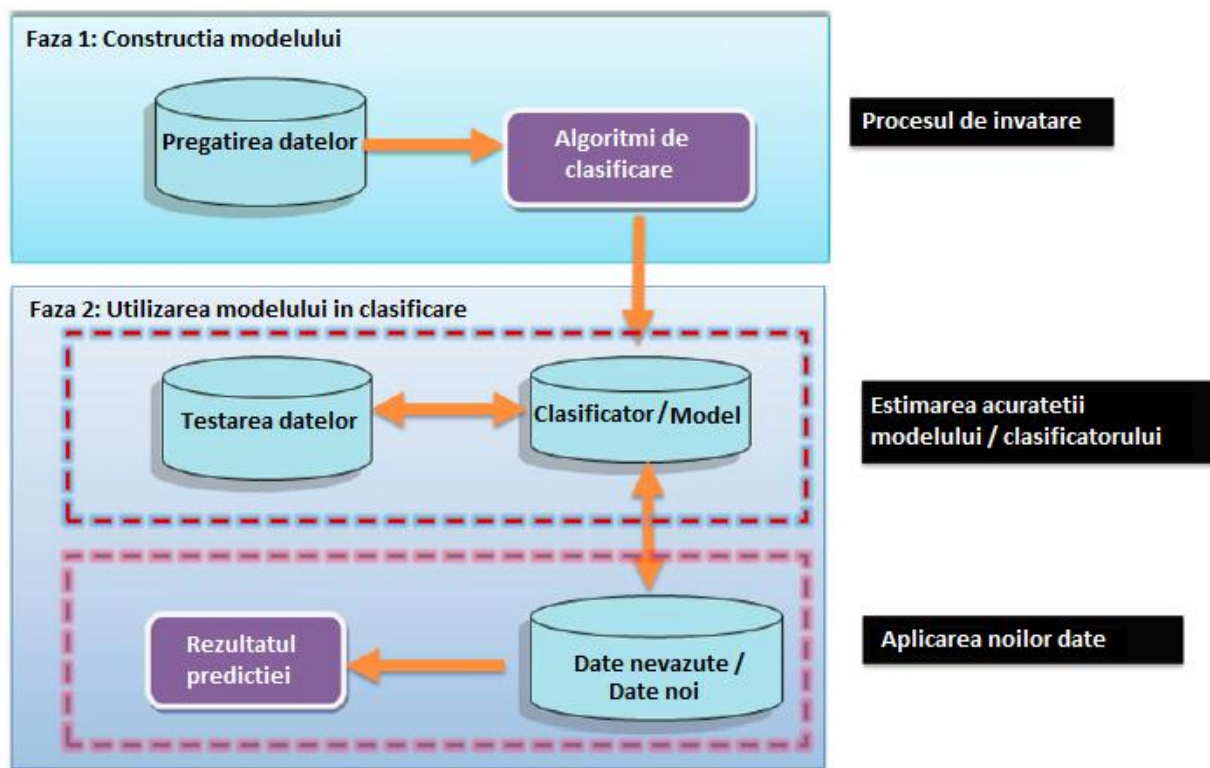


Fig III.3. Clasificarea și predicția în extragerea cunoștințelor din date

Există unele studii, care arată un mare interes pentru rezolvarea problemelor de HR folosind extragerea cunoștințelor din date [RANJ]¹⁰⁵. Tabelul III.1. arată unele dintre sarcinile din domeniul resurselor umane care utilizează tehnica de extragere a datelor. În plus, până în prezent există discuții destul de limitate cu privire la managementul talentelor, cum ar fi prognoza talentului, planificarea carierei și recrutarea talentelor printr-o abordare care să utilizeze extragerea cunoștințelor din date.

În domeniul resurselor umane, tehnicile de extragere a cunoștințelor din date utilizate se concentrează asupra selecției personalului, în special pentru a alege candidații potriviți pentru un loc de muncă. Clasificarea și predicția în exploatarea datelor pentru probleme de resurse umane

¹⁰⁵ [RANJ] Ranjan, J. (2008). Data Mining Techniques for better decisions in Human Resource Management Systems. *International Journal of Business Information Systems*, 3(5), 464-481

sunt rare și există câteva exemple, cum ar fi pentru a prezice durata unui serviciu, primele de vânzări, persistența indicilor agenților de asigurare și analiza comportamentului de lucru al operatorilor [CHN]¹⁰⁶. Din aceste motive, această lucrare încearcă să folosească tehnici de clasificare de extragere a cunoștințelor din date pentru a realiza o prognoză a potențialilor angajați, care să arate cât de potriviți sunt aceștia pentru sarcinile de lucru care îi așteaptă, folosind experiența din trecut.

	Tehnica de extragere a cunoștințelor din date
Selecția personalului	Arbori de decizie [CHN] Logica fuzzy și extragerea cunoștințelor din date [TAHS] ¹⁰⁷ Teoria seturilor brute [CCWL] ¹⁰⁸
Pregătirea angajatului (training)	Reguli de asociere minieră [CCWL]
Dezvoltarea angajatului	Logica fuzzy și rețelele neuronale artificiale [HTL] ¹⁰⁹ Arbori de decizie [THCS] ¹¹⁰
Evaluarea performanțelor	Arbori de decizie [ZHAO] ¹¹¹

Tabelul III.1. Tehnici de extragere a cunoștințelor din date

¹⁰⁶ [CHN] Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems and Applications*, 34(1), 380-290

¹⁰⁷ [TAHS] Tai, W. S., & Hsu, C. C. (2005). A Realistic Personnel Selection Tool Based on Fuzzy Data Mining Method

¹⁰⁸ [CCWL] Chen, K. K., Chen, M. Y., Wu, H. J., & Lee, Y. L. (2007). *Constructing a Web-based Employee Training Expert System with Data Mining Approach*. Paper presented at the Paper in The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and Eservices (CEC-EEE 2007)

¹⁰⁹ [HTL] Huang, M. J., Tsou, Y. L., & Lee, S. C. (2006). Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge. *Knowledge-Based Systems*, 19(6), 396-403

¹¹⁰ [THCS] Tung, K. Y., Huang, I. C., Chen, S. L., & Shih, C. T. (2005). Mining the Generation Xer's job attitudes by artificial neural network and decision tree - empirical evidence in Taiwan. *Expert Systems and Applications*, 29(4), 783-79

¹¹¹ [ZHAO] Zhao, X. (2008). *An Empirical Study of Data Mining in Performance Evaluation of HRM*. Paper presented at the International Symposium on Intelligent Information Technology Application Workshops, Hangzhou, China

În orice organizație, managementul talentelor a devenit o abordare din ce în ce mai importantă în funcțiile resurselor umane. Talentul este considerat ca fiind capacitatea unei persoane de a face o diferență semnificativă a performanței actuale și viitoare ale organizației [LYN]¹¹². De fapt, gestionarea talentelor implică planificarea resurselor umane, care pune accentul pe procesele de gestionare a persoanelor din organizație. În afară de faptul că, managementul talentelor poate fi definit ca procesul de a asigura continuitatea conducerii în poziții cheie și de a încuraja progresul individual, acesta mai poate fi considerat și decizia de a gestiona oferta, cererea și fluxul de talent prin intermediul motorului de capital uman [CNGM].¹¹³

Managementul talentelor este foarte important și are nevoie de puțină atenție din partea profesioniștilor resurselor umane. De exemplu, un raport de cercetare [TPTR]¹¹⁴ a constatat că printre provocările actuale și viitoare de top ale managementului talentelor se găsesc: dezvoltarea talentelor existente, prognozarea nevoii de talente, atragerea și menținerea talentului potrivit de conducere, identificarea talentelor existente, atragerea și reținerea conducerii potrivite și a persoanelor cheie, implementarea talentelor existente, lipsa capacității de conducere la nivel superior și asigurarea unui bazin de talente divers.

Procesul de management al talentelor constă în recunoașterea domeniilor-cheie de talent în organizație, identificarea persoanelor din organizație care constituie talent cheie, precum și desfășurarea de activități de dezvoltare a bazei de selecție a talentului pentru a păstra și a angaja și, de asemenea, pentru a fi pregătiți să transfere persoane cheie în alte roluri importante [CNGM] (Fig. III.4.). Aceste procese implică activități de resurse umane, care trebuie să fie integrate într-un sistem eficient [CHNU]¹¹⁵ (Fig. III.4.).

¹¹² [LYN] Lynne, M. (2005). *Talent Management Value Imperatives: Strategies for Execution*: The Conference Board

¹¹³ [CNGM] Cubbingham, I. (2007). Talent Management: Making it real. *Development and Learning in Organizations*, 21(2), 4-6

¹¹⁴ [TPTR] TP Track Research Report - *TalentManagement: A State of the Art*: Tower PerrinHR Services, 2005

¹¹⁵ [CHNU] CHINA UPDATE. (2007). HR News for Your Organization: The Tower Perrin Asia Talent Management Study. Retrieved from www.towersperrin.com. 7/1/2008

În aceasta lucrare, mă voi concentra pe una dintre provocările de management al talentelor, mai precis pe aceea de a identifica talentul existent în ceea ce privește talentul-cheie într-o organizație prin anticiparea performanțelor folosind înregistrările anterioare de performanță ale angajaților din bazele de date. În acest caz, voi folosi datele aferente angajaților din trecut în ceea ce privește talentul lor prin utilizarea tehnicii de clasificare în extragerea

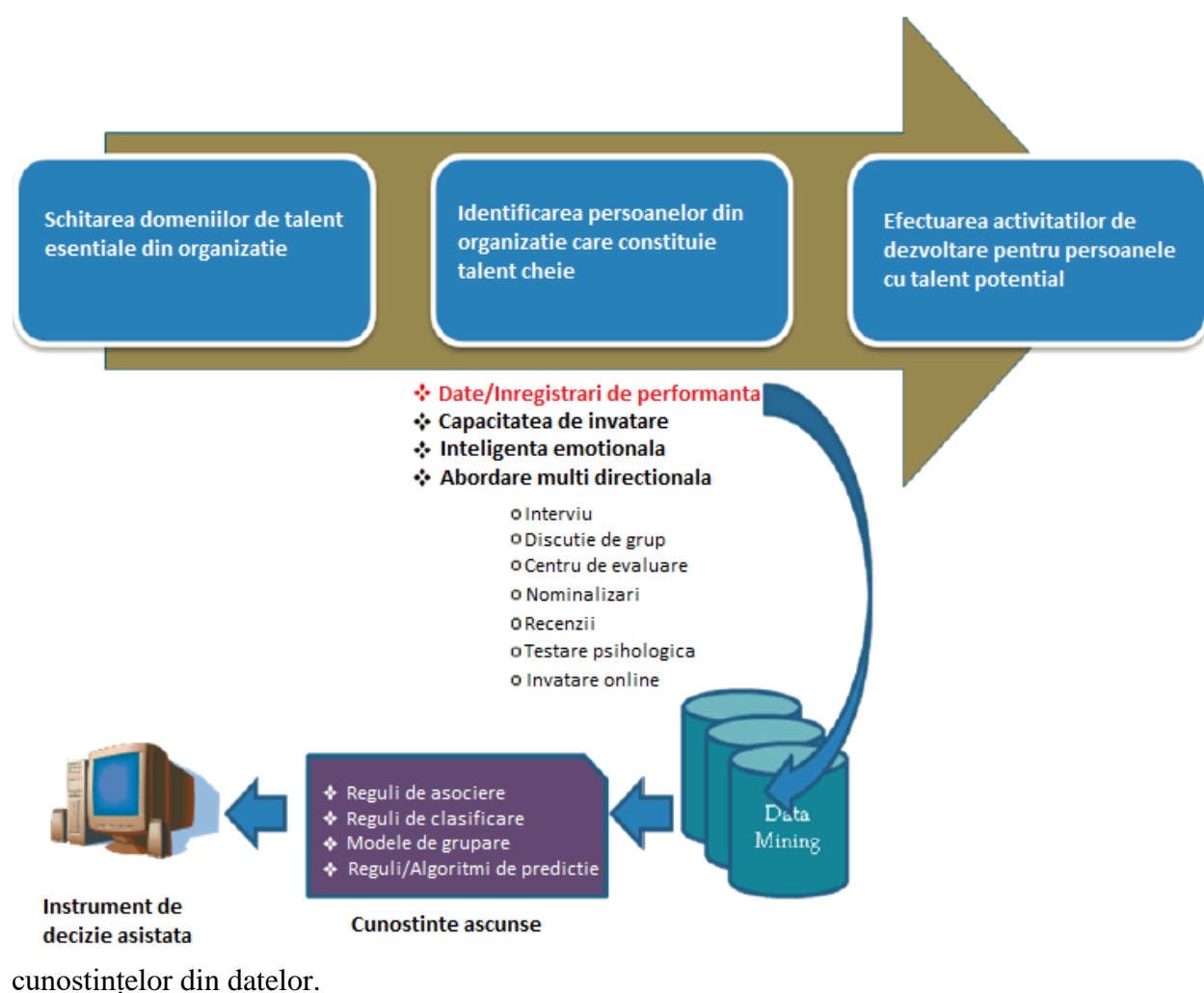


Fig III.4. Managementul talentului și extragerea cunoștințelor din date

Recent, odată cu noua cerere și sporirea vizibilității, resursele umane caută un rol mai strategic prin valorificarea metodelor de exploatare a datelor [RANJ]¹¹⁶. Acest lucru poate fi realizat prin descoperirea modelelor generate în cunoștințe utile din datele existente în bazele de date HR. Astfel, studiul realizat se concentrează pe identificarea modelelor care se referă la talentul uman.

Modelele pot fi generate prin utilizarea unora dintre cele mai importante tehnici de exploatare a datelor, cum ar fi gruparea, pentru a lista angajații cu caracteristici similare, pentru a grupa performanțele și etc. Din tehnica de asociere, modelele care sunt descoperite pot fi folosite pentru a asocia profilul angajatului pentru cel mai adecvat program / loc de muncă, asociat cu atitudinea angajatului față de performanță și etc. În etapa de predicție și de clasificare, modelul descoperit poate fi utilizat pentru a prezice acuratețea procentuală a performanței, comportamentul și atitudinile angajaților, a prezice evoluția performanței pe tot parcursul perioadei de performanță și identifică, de asemenea, cel mai bun profil pentru diferiți angajați (fig. III.5.).

Legăturile dintre problemele de exploatare a datelor și nevoile managementului talentelor sunt foarte importante. Prin urmare, este foarte important să se determine tehnicile adecvate de exploatare a datelor pentru problemele de management al talentelor.

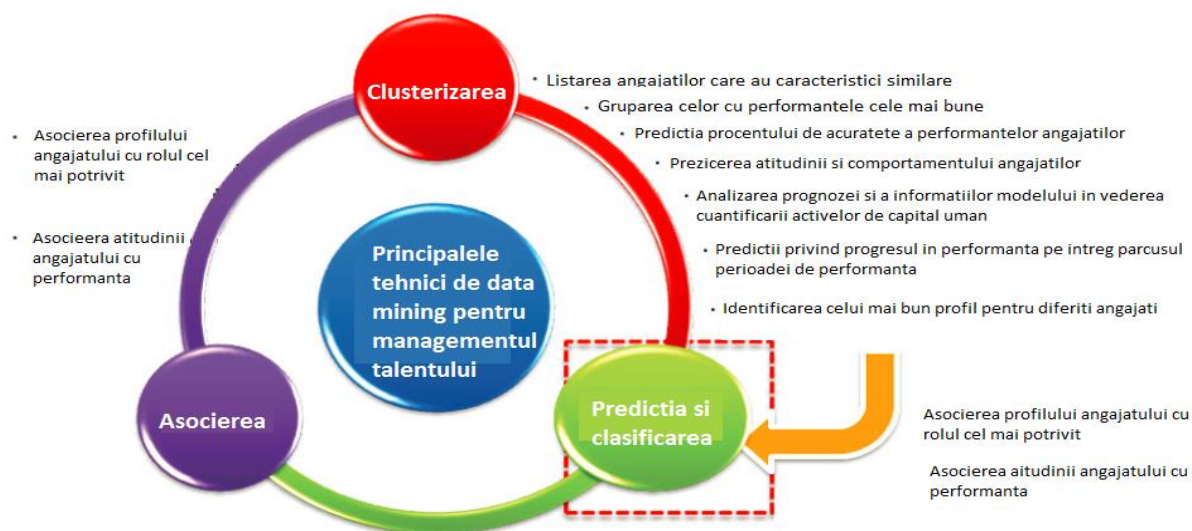


Fig. III.5. Tehnici de extragere a cunoștințelor din date pentru managementul talentului

¹¹⁶ [RANJ] Ranjan, J. (2008). Data Mining Techniques for better decisions in Human Resource Management Systems. *International Journal of Business Information Systems*, 3(5), 464-481

Configurarea experimentului în acest studiu are mai multe sarcini, cum ar fi simularea construcției de date, plasarea valorilor excepționale, reducerea atributelor și precizarea determinării modelului așa cum este prezentat în Fig. III.6. Cu toate acestea, din cauza dificultăților de a obține date reale de la un departament de resurse umane, din cauza problemelor de confidențialitate și de securitate, în scopul de explorare, acest studiu simulează seturile de date de talente umane folosind un generator de date.

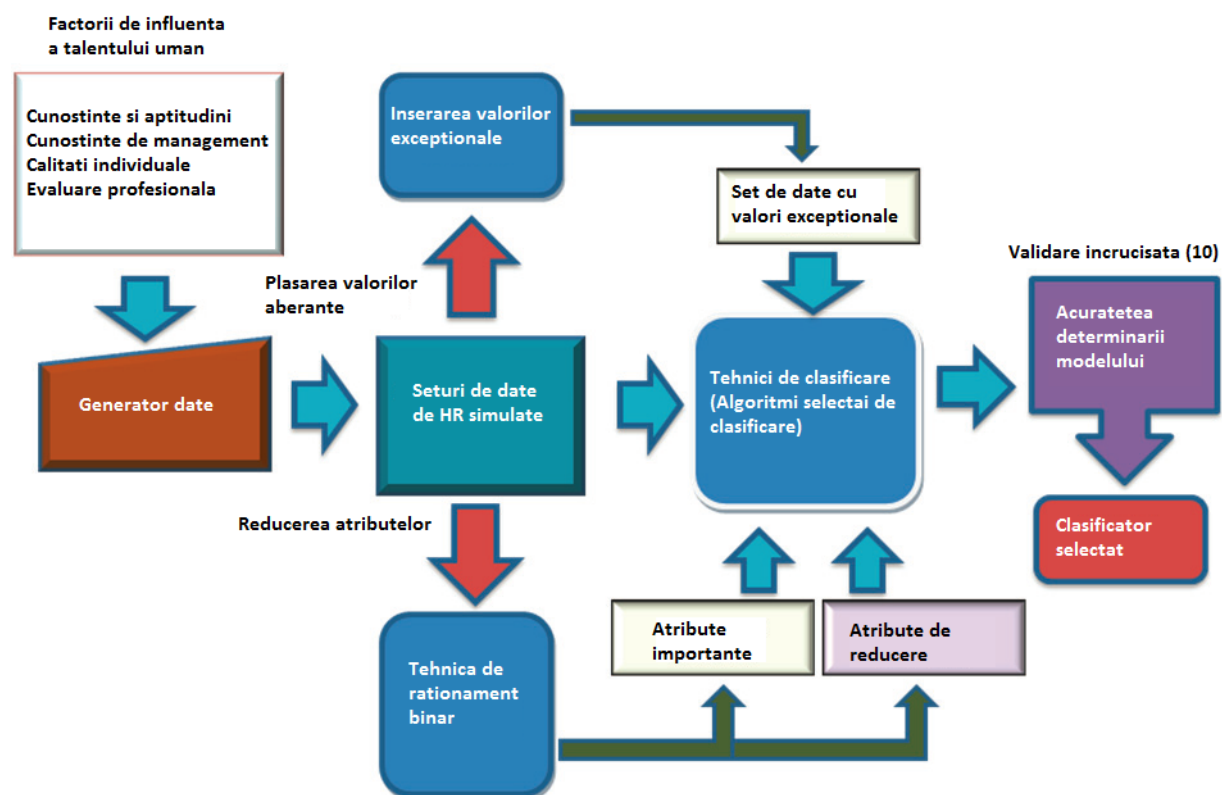


Fig III.6. Pregătirea experimentului

Primul set de date conține o sută de înregistrări (*setdate1*), în timp ce al doilea set de date conține o mie de înregistrări (*setdate2*). În multe cazuri, datele simulate pot fi foarte apropiate de datele ideale și pot produce astfel, un bun model de exploatare a datelor. Din acest motiv, în acest studiu se inserează valori excepționale în primul set de date (*setdate1*) tocmai pentru a se evita această problemă și se formează astfel un nou set de date cunoscut sub numele de *setdate3*.

În această lucrare, tehnicile de clasificare selectate și utilizate se bazează pe tehnicile obișnuite utilizate pentru clasificare și predicție în exploatare a datelor. Așa cum am menționat deja, tehnicile de clasificare alese sunt rețelele neuronale, care sunt destul de populare în comunitatea minieră de date și utilizate ca model de clasificare. Arborele de decizie cunoscut sub

numele de abordarea "divide-și-cucerește" este format dintr-un set de instanțe independente pentru clasificare, în timp ce, cel mai apropiat vecin este folosit pentru clasificarea care se bazează pe distanțe metrice.

Tehnici de extragere a cunoștințelor din date	Algoritm de clasificare
Arbore de decizie	<ul style="list-style-type: none"> • C4.5 – Arbore decizional de inducție - obiectivul este nominal, iar intrările pot fi nominale sau interval. Uneori dimensiunea arborilor induși este redusă în mod semnificativ atunci când se adoptă o strategie diferită de tăiere. • Padure aleatoare - Alege un test bazat pe un anumit număr de caracteristici aleatoare la fiecare nod, care nu efectuează tăieri. Pădurea aleatoare construiește o mulțime de ansambluri împachetând copaci aleatorii.
Rețele neuronale	<ul style="list-style-type: none"> • Percepția multi-strat - Un predictor precis pentru problemele de clasificare care stau la bază. Având în vedere o structură de rețea fixă, trebuie să se stabilească ponderi adecvate pentru conexiunile în rețea • Rețea radială de funcții de bază (<i>radial basis function network</i>) - Un alt tip popular de rețea de alimentare, care are două straturi, fără să numărăm stratul de intrare, și diferă față de un perceptron multistrat în modul în care unitățile ascunse efectuează calcule.
Cel mai apropiat vecin	<ul style="list-style-type: none"> • K* - O învățare bazată pe exemplu, folosind distanța metrică pentru a măsura similitudinea de instanțe și funcției generalizată a distanței bazată pe o transformare.

Tabelul III.2. Tehnici de clasificare

Tabelul III.2. rezumă tehnicile de clasificare selectate în exploatarea datelor, cum ar fi arborele de decizie, rețelele neuronale și cel mai apropiat vecin. În acest studiu, voi încerca să folosesc C4.5 și pădurile aleatorii pentru categoria arborilor de decizie, percepția multistrat și

rețeaua radială de funcții de bază pentru categoria rețelelor neuronale și K-Star pentru cea mai apropiată categorie de vecin.

Seturile de date conțin informații privind cunoștințele și aptitudinile unei persoane, cunoștințele de management, calitățile individuale și evaluare profesională. Pentru cunoștințe și aptitudini au fost generate valori pentru 16 variabile, în timp ce pentru cunoștințele de management, calitățile individuale și evaluarea profesională, câte 8 variabile. Aceste seturi de date au fost prelucrate folosind Weka.

În urma analizării seturilor de date au fost obținute următoarele rezultate, rotunjite la 2 zecimale:

Algoritm de clasificare	Setdate1	Setdate2	Setdate3
Padure aleatoare (<i>random forest</i>)	66,81	89,46	59,72
K*Star	79,30	86,04	62,29
Percepția multi-strat	77,12	92,63	74,94
Rețea radială de funcții de bază (<i>radial basis function network</i>)	70,56	94,28	80,72
C4.5	86,35	96,18	82,77

Tabelul III.3. Rezultate obținute în cazul utilizării tuturor atributelor

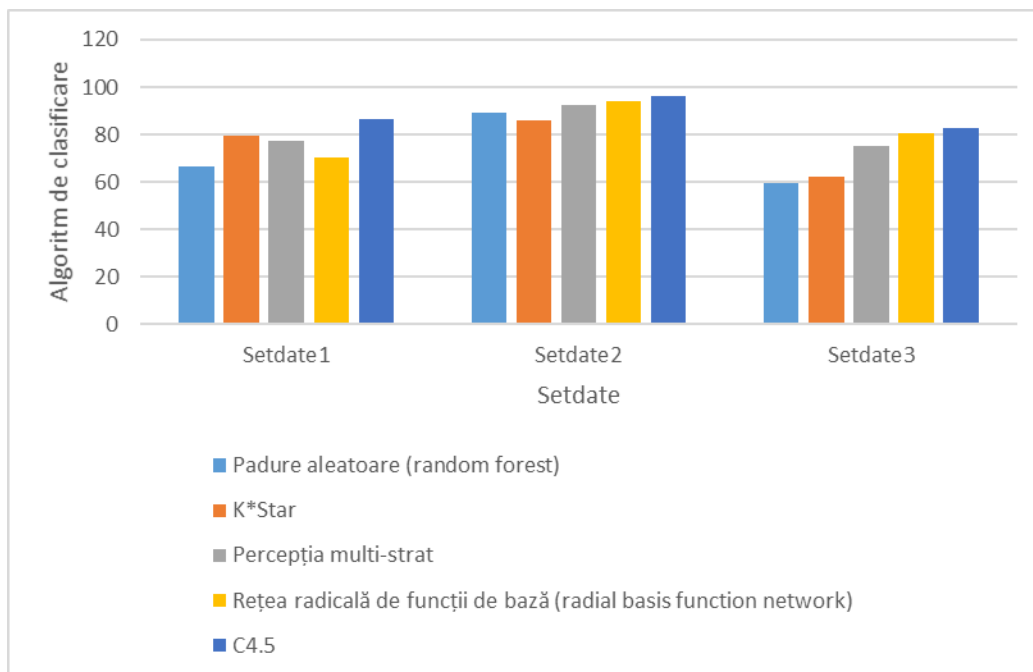


Fig III.7. Acuratețea modelului în cazul utilizării tuturor atributelor

Rezultatul pentru attributele complete ne arată că atunci când folosim mai multe date (Setdate2) în procesul de instruire, poate fi dezvoltat un model cu o precizie mai mare. În afară de faptul că, precizia pentru Setdate3 care conține valori aberante, este mai mică pentru toți clasificatorii, ne demonstrează efectul valorilor aberante în setul de date pentru precizia modelului.

Cea de a doua fază a experimentului constă în realizarea unei analize relevante, în scopul de a determina acuratețea tehnicii de clasificare selectate folosind seturi de date cu reducerea atributului. În acest experiment, m-am concentrat pe seturile de date 1 și 2. Scopul acestui proces de reducere a atributului este de a selecta atributul cel mai relevant în setul de date. Procesul de reducere este pus în aplicare cu ajutorul tehnicii de raționament boolean. Prin reducerea atributului, putem reduce timpul și spațiul de preprocesare și de prelucrare. Tabelul următor prezintă rezultatele analizei relevante pentru reducerea atribut, cinci attribute fiind selectate. Prin utilizarea acestor variabile de reducere a atributelor, a doua etapă a experimentului este pusă în aplicare. Scopul acestui experiment este de a afla precizia tehnicilor de clasificare cu o reducere de atribut folosind cele mai scurte attribute de lungime și prin combinarea importanței atributelor după procesul de reducere.

Tabelul următor arată acuratețea algoritmului de clasificare cu reducerea atributului pentru cele mai scurte metode de lungime (cu reducerea atributelor seturilor de date). Clasificatorul C4.5 are cel mai mare procent de acuratețe în prima etapă a celei de a doua faze, dar precizia a scăzut în acest stadiu.

Algoritm de clasificare	Setdate1	Setdate2
Padure aleatoare (<i>random forest</i>)	60,17	64,87
K*Star	62,39	65,34
Percepția multi-strat	57,83	62,19
Rețea radială de funcții de bază (<i>radial basis function network</i>)	61,24	66,55
C4.5	63,17	65,43

Tabelul III.4. Rezultate obținute în cazul reducerii atributelor

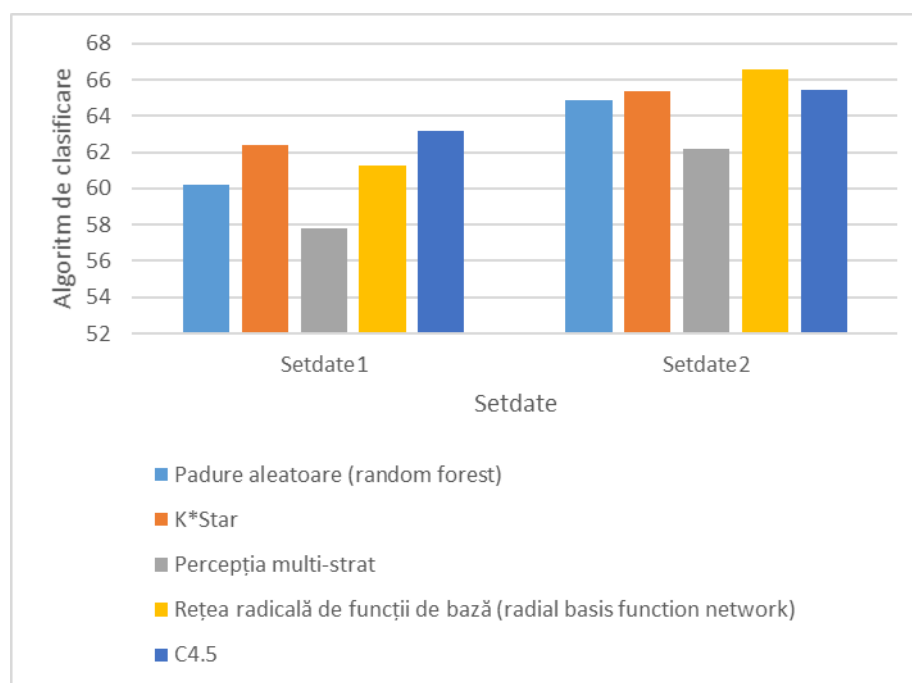


Fig III.8. Acuratețea modelului în cazul reducerii atributelor

În cazul acestui experiment, rezultatul ne arată că folosirea mai multor atribute în setul de date va afecta precizia clasificatorului. Prin urmare, acest rezultat ilustrează că cele mai multe dintre atributele din setul de date sunt importante și ar trebui să fie luate în considerare. Cu toate

acestea, cu o combinație de attribute din procesul de reducere din cea de a doua etapă a experimentului, precizia clasificatorului este mai mare comparativ cu cele mai scurte attribute de lungime (AR seturilor de date). Tabelul III.5 arată acuratețea clasificatorului pentru attributele importante pentru seturile de date 1 și 2. Clasificatorul C4.5 are cea mai mare precizie pentru ambele seturi de date în această etapă a experimentului. Figura următoare arată acuratețea modelului pentru seturile de date AR și seturi de date de AI în a doua faza a experimentului.

Algoritm de clasificare	Setdate1	Setdate2
Padure aleatoare (<i>random forest</i>)	83,50	94,86
K*Star	75,38	97,44
Percepția multi-strat	77,12	96,63
Rețea radială de funcții de bază (<i>radial basis function network</i>)	74,39	98,06
C4.5	92,61	99,13

Tabelul III.5. Rezultate obținute în cazul importanței atributelor

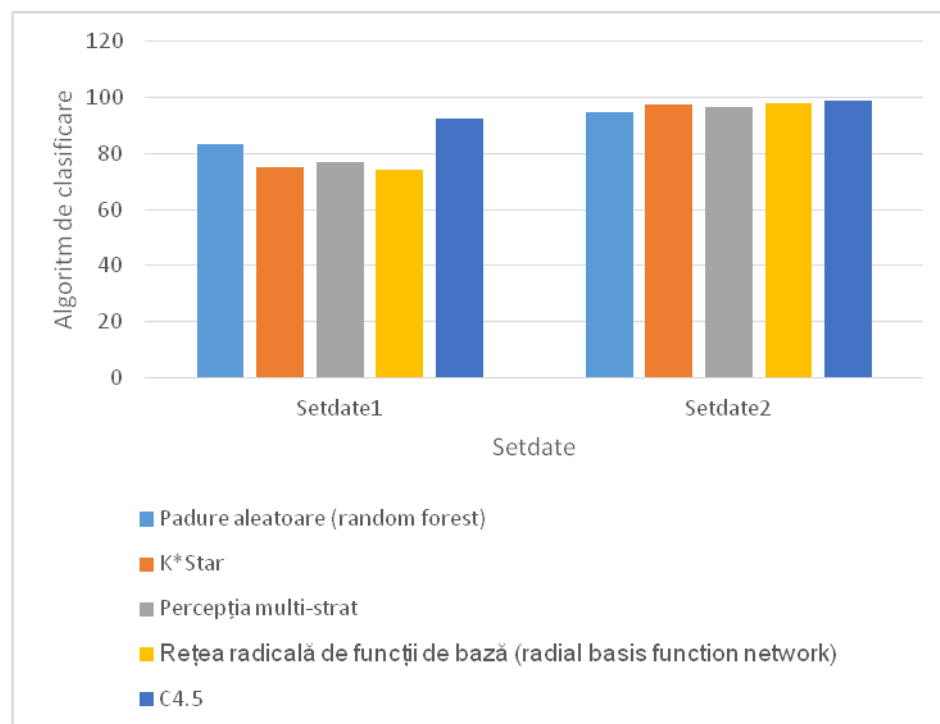


Fig III.9. Acuratetea modelului în cazul importanței atributelor

IV.2. Proiectarea unui prototip de recrutare avansată a forței de muncă

IV.2.1. Nivelul actual

Încă de la lansarea lucrării „Razboiul pentru talent” [MCKa]¹¹⁷, previziunile sale au început să devină realitate. Cererea pentru talent a crescut în timp ce oferta a scăzut, iar aceste tendințe îngrijorătoare par să nu arate nici un semn de schimbare în viitorul apropiat. Potrivit Bloomberg Businessweek, SUA, Canada, Marea Britanie, și Japonia (printre multe altele) se vor confrunta cu diferite grade de deficit de talente în aproape fiecare industrie în anii 2020 – 2030 [BLM]¹¹⁸.

Predicțiile realizate de directorii resurselor umane ne arată o anumită presiune pe piața forței de muncă în mai multe profesii, generată de lipsa personalului, în perioada 2020 – 2030. Această presiune este evidențiată în graficul următor:

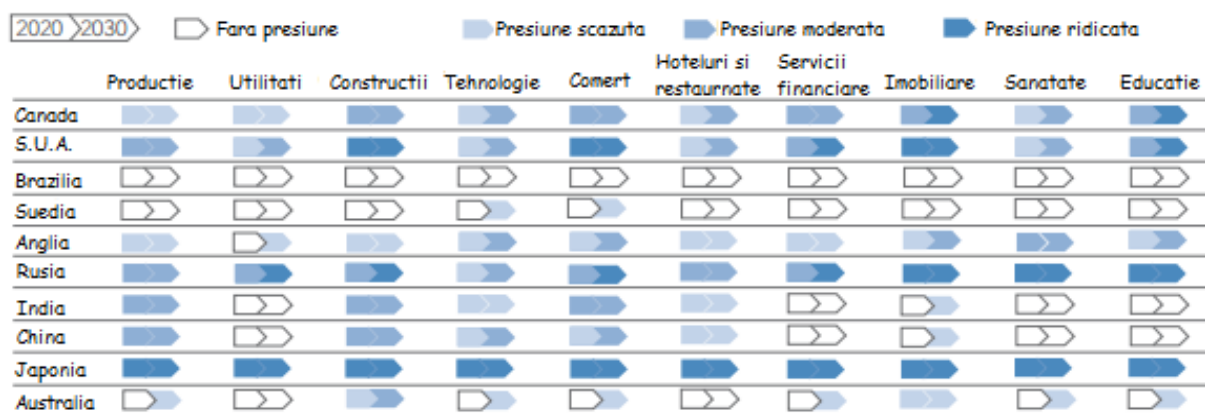


Figura IV.1. Predicția mondială a lipsei personalului calificat în perioada 2020-2030 [BLM]¹¹⁹

¹¹⁷ [MCKa] Elizabeth G. Chambers, Mark Foulton, Helen Handfield-Jones, Steven M. Hankin, Edward G. Michaels III 1998, The War for Talent,

¹¹⁸ [BLM] Bloomberg Businessweek, September 13 – September 19, 2010 issue, 54

¹¹⁹ [BLM] Bloomberg Businessweek, September 13 – September 19, 2010 issue, 54

Deși recesiunea globală a forțat multe organizații să recurgă la disponibilizări și să se concentreze mai puțin asupra deficitului de talente pe termen scurt, incertitudinile de pe piața mondială au dat directorilor mai multe motive să își facă griji cu privire la talentul pe care îl au și la talentul de care au nevoie. Acest lucru s-a concretizat într-o intensificare a interesului privind capitalul uman, ceea ce a dat naștere termenului de managementul integrat al resurselor umane (Integrated Talent Management). Astfel, domeniul HR, ar fi integrat într-un ciclu continuu de acțiuni (de exemplu, previzionarea, recrutarea, implementarea, dezvoltarea, păstrarea talentului uman) care se alimentează reciproc și formează o vedere de ansamblu a capitalului uman al unei organizații, care apoi formează strategia de afaceri și de capital uman. Deși termenul managementul integrat al resurselor umane a început să prindă contur, se pare că practica actuală încă nu a făcut acest pas. Cele mai multe firme au încă o funcție de HR destul de izolată cu utilitare software care încearcă cu disperare să fie un sprijin, ceea ce face dificilă analiza datelor, dacă nu chiar imposibilă.

Un studiu recent [ADP15]¹²⁰, evidențiat și în figura IV.2, realizat pe un eșantion de 725 de companii multinaționale, cu peste 5000 de angajați la nivel mondial, a arătat că, în medie, companiile încearcă să utilizeze 33 de sisteme de salarizare și 31 de sisteme pentru managementul resurselor umane. Mai mult, aceste numere au crescut cu 40% în ultimul an. Astfel, putem afirma că se impune standardizarea la nivel de companie a sistemelor de management, prin implementarea unui sistem inteligent care să faciliteze îndeplinirea funcțiilor manageriale în cadrul organizației. Un sistem standard ar ușura implementarea unui sistem de mobilitate internă a angajaților, exploatând și mai mult calitățile unui angajat prin relocarea sa pe un alt post pentru a oferi un randament mai bun. Deasemenea, ar oferi posibilitatea angajaților să considere o nouă carieră, în cadrul unui alt departament al companiei, înainte de a opta pentru o ofertă de muncă din exterior. Acest lucru ajută la păstrarea angajaților în companie, ceea ce conduce la o diminuare a investițiilor în specializarea noilor angajați.

¹²⁰ [ADP15] ADP Research Institute: Harnessing Big Data: The Human Capital Management Journey to Achieving Business Growth – ADP Global Human Capital Management Decision Makers Survey, 2015

Nr total participanti:
 2013 n = 461
 2014 n = 566
 2015 n = 724

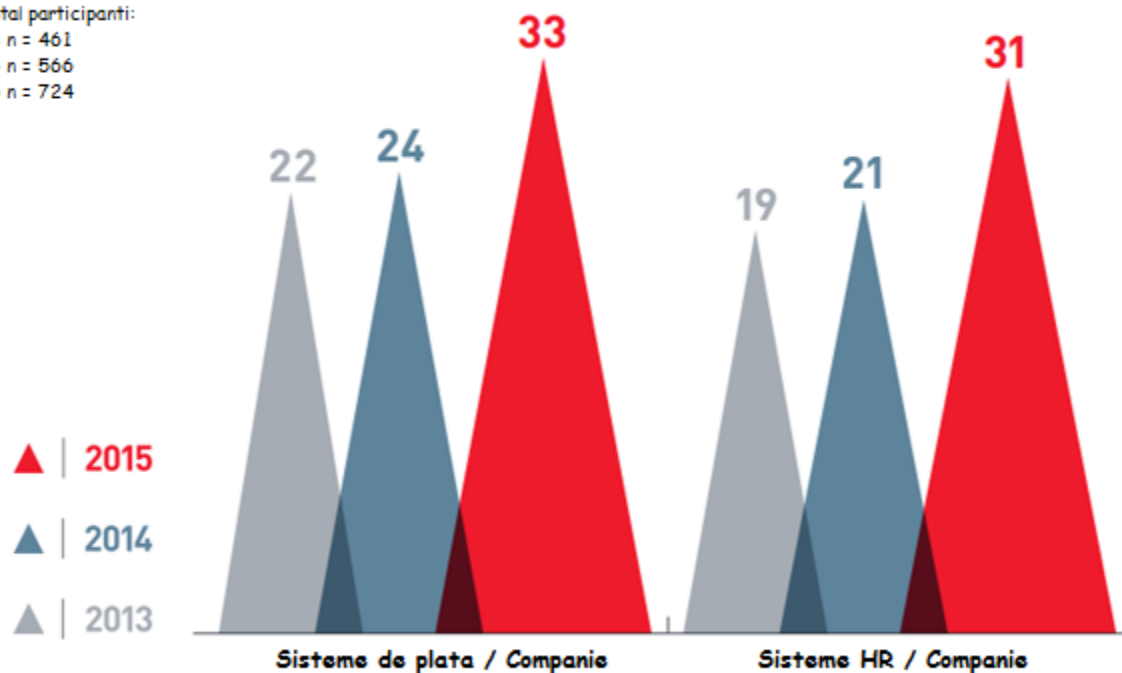


Figura IV.2. Graficul sistemelor de salarizare și de management
 al resurselor umane la nivel mondial [ADP15]¹²¹

Caracteristica cea mai evidentă a volumelor mari de date este volumul. Tot mai mulți oameni folosesc dispozitive inteligente conectate la o rețea, sau la Internet, și produc date la fiecare secundă. Știința are acum o bază solidă pentru a face tot felul de presupuneri pe baza datelor primite de la pacienți, clienți, sportivi, etc. Este un model care implică întregul nostru univers în colectarea, procesarea și distribuirea datelor. Este important să se beneficieze de acest flux de date, prin stocarea în mod corespunzător, utilizând soluții bazate pe volumele mari de date.

Importanța folosirii volumelor mari de date în vederea luării unor decizii informate a devenit, recent, extrem de discutată în majoritatea organizațiilor. În timp ce finanțele, marketingul și alte departamente din cadrul unei companii primesc sisteme de date și de analiză

¹²¹ [ADP15] ADP Research Institute: Harnessing Big Data: The Human Capital Management Journey to Achieving Business Growth – ADP Global Human Capital Management Decision Makers Survey, 2015

personalizate, resursele umane sunt încă neasistate de sisteme specializate în procesarea volumelor mari de date [KRUS15]¹²².

Nu există o infrastructură bazată pe tehnologia volumelor mari de date (big data) care să se potrivească oricărei infrastructuri. Companiile investesc masiv pentru a personaliza tehnologiile bazate pe volumele mari de date și instrumentele analitice, în general pentru o utilizare la nivel de departament, în vederea asigurării nevoilor specifice și a obiectivelor propuse. Din nefericire, resursele umane, încă nu beneficiază de o infrastructură care să satisfacă complet necesitățile departamentului [KRUS15].

Așa cum afirmă și O'Reilly [OREI11]¹²³, cele mai utilizate soluții privind tehnologia volumelor mari de date sunt Cassandra și HBase. Cassandra este lider în realizarea celui mai mare randament pentru numărul maxim de noduri, așa cum arată și experimentele realizate în cadrul lucrării *“Solving Big Data Challenges for Enterprise Application Performance Management”* [RSJV12]¹²⁴. Desigur, există și un aspect negativ: operațiunile de intrare / ieșire necesită un timp mai ridicat. Cassandra a fost dezvoltat de Facebook, în timp ce HBase este o parte a proiectului Apache Hadoop și are sprijinul Google, fiind utilizat pe seturi de date extrem de mari.

Tehnologia actuală permite stocarea eficientă și interogarea seturilor de date mari, punându-se accentul pe utilizarea întregului set de date și nu doar a unor probe (eșantioane) [SIGU14]. Tehnologia volumelor mari de date este ideală procesului de analiză, deoarece scopul final al colectării unui volum atât de mare de date îl reprezintă procesarea și analizarea în vederea obținerii de informații valoroase. Analiza nu funcționează direct pe date, fiind necesară extragerea prealabilă a datelor din baza de date, folosind un limbaj specific și doar apoi putându-se utiliza instrumentele analitice.

¹²² [KRUS15] JoAnne Kruse – With Big Data, HR Departments Too Often Get Short Shrift – The Wall Street Journal, 22 Feb. 2015

¹²³ [OREI11] O'Reilly Media - Big Data Now, September 2011, ISBN: 978-1-449-31518-4

¹²⁴ [RSJV12] Rabl, Sadoghi, Jacobsen, Villamor, Mulero, Mankovskii - Solving Big Data Challenges for Enterprise Application Performance Management, 2012-08-27, VLDB, Vol. 5, ISSN 2150-8097

Până la utilizarea tehnologiei volumelor mari de date, cel mai bun mod de interogare a datelor dintr-o bază de date se realiza prin utilizarea limbajului SQL, care este specific pentru tabele relaționale structurate. Atunci când datele au început să fie stocate în baze de date NoSQL, limbajul SQL a devenit utilizat doar complementar în cadrul interogărilor. De exemplu, joncțiunile nu sunt disponibile în interogările NoSQL. Mai presus de toate, a fost declarat recent (septembrie 2014), că limbajul SQL este mult mai important decât se credea pentru tehnologia volumelor mari de date, Oracle lansând limbajul Big Data SQL, care reprezintă o extindere a limbajului SQL pentru Hadoop și NoSQL.

IV.2.2. Analiza prototipului informatic

Analiza sistemului existent

În prima figură (fig. IV.9.) am prezentat un model standard de recrutare al candidaților pe perioadă de probă (3 luni). Figura IV.10., prezintă fluxul care are loc în cadrul unui departament, din momentul angajării unei noi persoane pe perioadă de probă și până la angajarea definitivă a acesteia pe perioadă nedeterminată sau, până la încheierea colaborării.

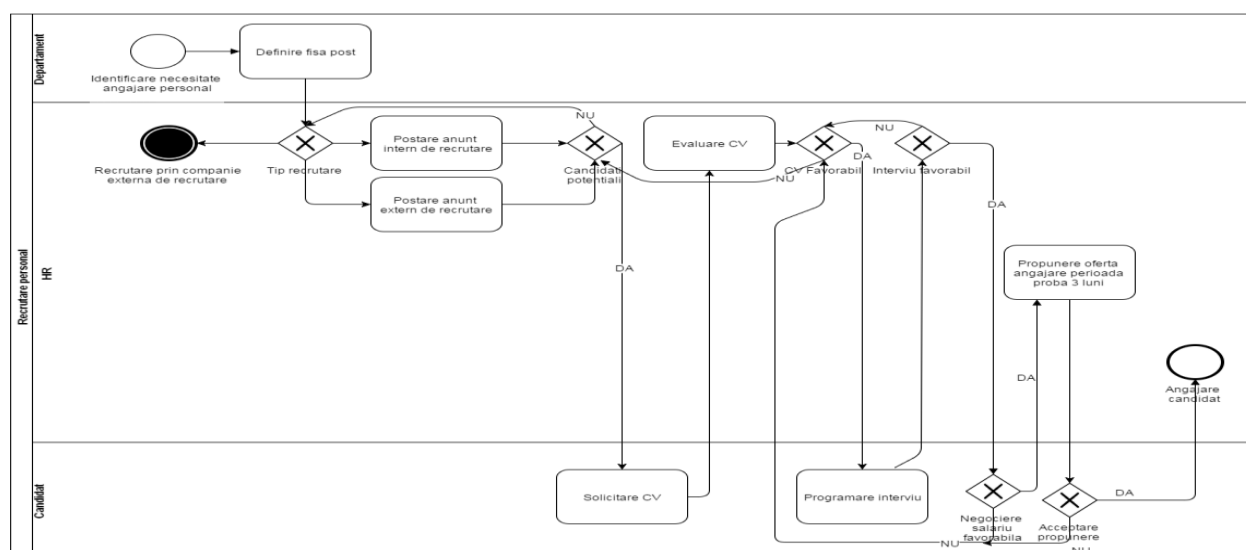


Fig. IV.9. – Diagrama BPMN a procesului de recrutare pe perioadă determinată

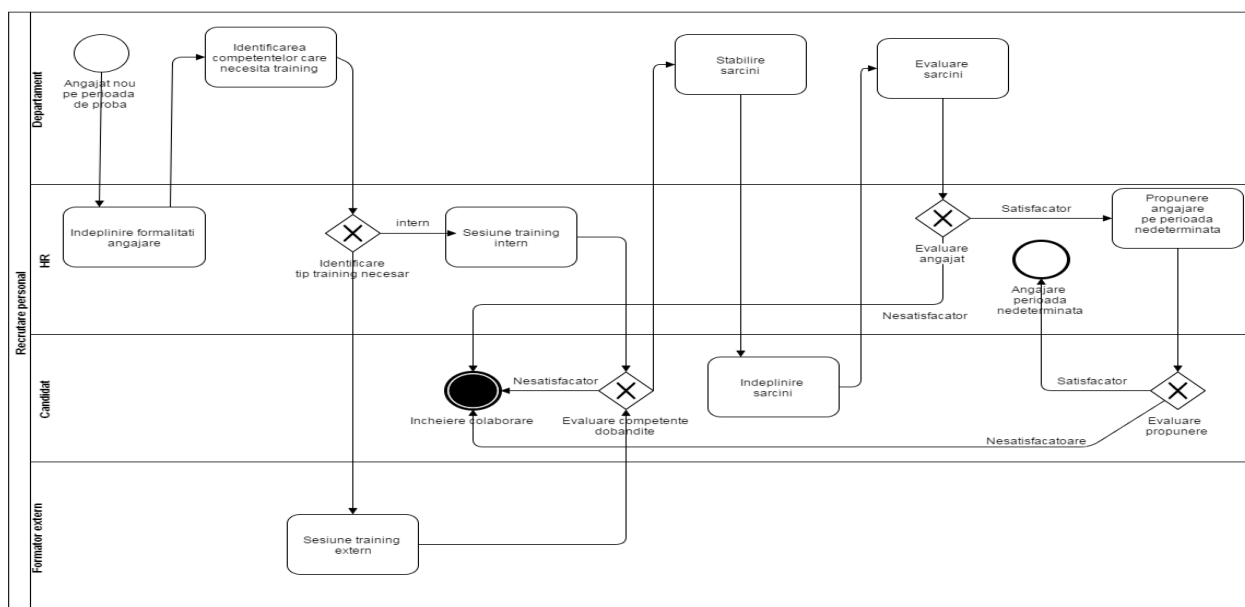


Fig. IV.10. Diagrama BPMN a procesului de recrutare pe perioadă nedeterminată

Specificarea cerințelor prototipului informatic

Datorită noilor tehnologii care își pun aparența în mod favorabil asupra modului de lucru în cadrul departamentului de resurse umane, influențând radical întreg procesul de recrutare, soluția prezentată în figura IV.11. reprezintă o opțiune cu mare potențial pentru sectorul resurselor umane.

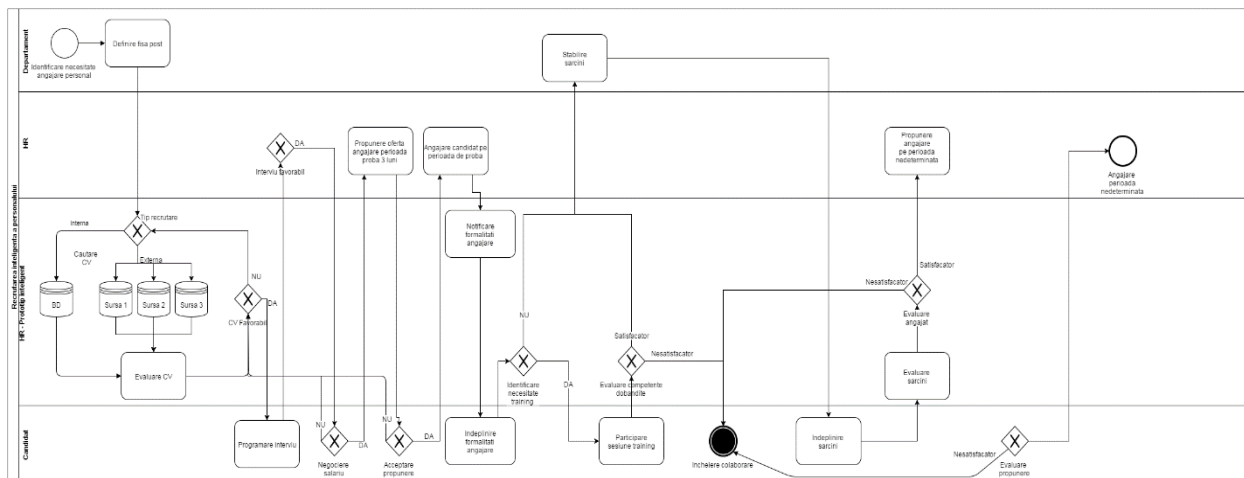


Fig IV.11. Diagrama BPMN a sistemului inteligent de recrutare folosind prototipul proiectat

IV.2.3. Proiectarea prototipului informatic

Figura următoare (fig. IV.12.) prezintă fluxul general de utilizare al aplicației. Astfel, în diagramă se poate identifica actorul principal, angajatul departamentului de resurse umane. Rolul diagramei este de a arăta modul de utilizare al prototipului.

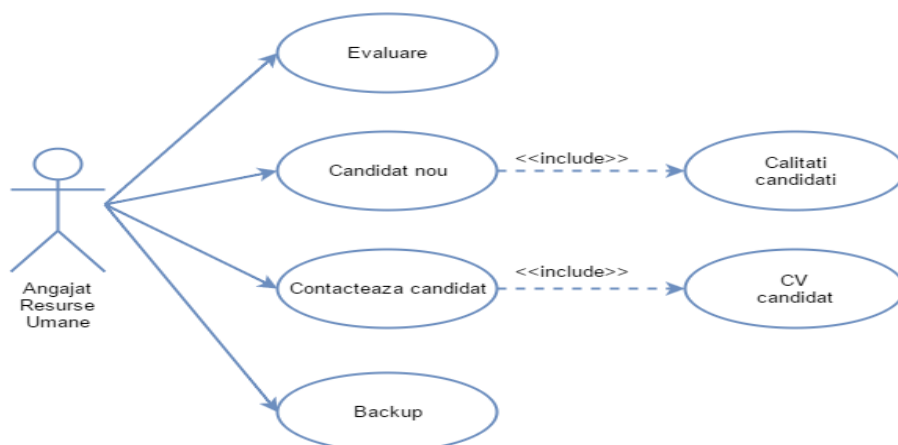


Fig. IV.12. Diagrama generală a cazurilor de utilizare

Diagramele de activitate sunt realizate pentru a oferi o imagine clară a pașilor de execuție a unei secvențe de acțiuni, realizată pentru obținerea unui anumit rezultat. Acest flux este descris detaliat, de la primul pas (momentul începerii acțiunii) și până la final (terminarea acțiunii). În cadrul lucrării voi prezenta diagramele tuturor cazurilor de utilizare.

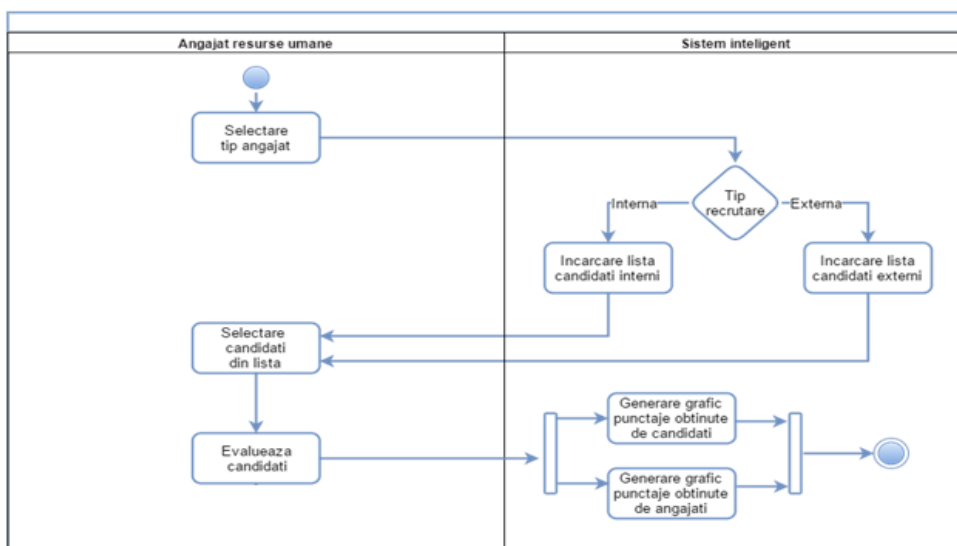


Fig. IV.13. Diagrama de activitate evaluare candidat

În imaginea precedentă (fig. IV.13.) este prezentată diagrama de activitate pentru evaluarea unui candidat. Se poate observa cum angajatul departamentului de resurse umane are posibilitatea de selecție a tipului de angajat, a candidaților din listă și de evaluare a acestora. Fiecare selecție efectuată de angajatul departamentului de resurse umane este interpretată adecvat de prototipul inteligent. În final, prototipul inteligent returnează angajatului graficele cu punctajele obținute de candidați în urma evaluării.

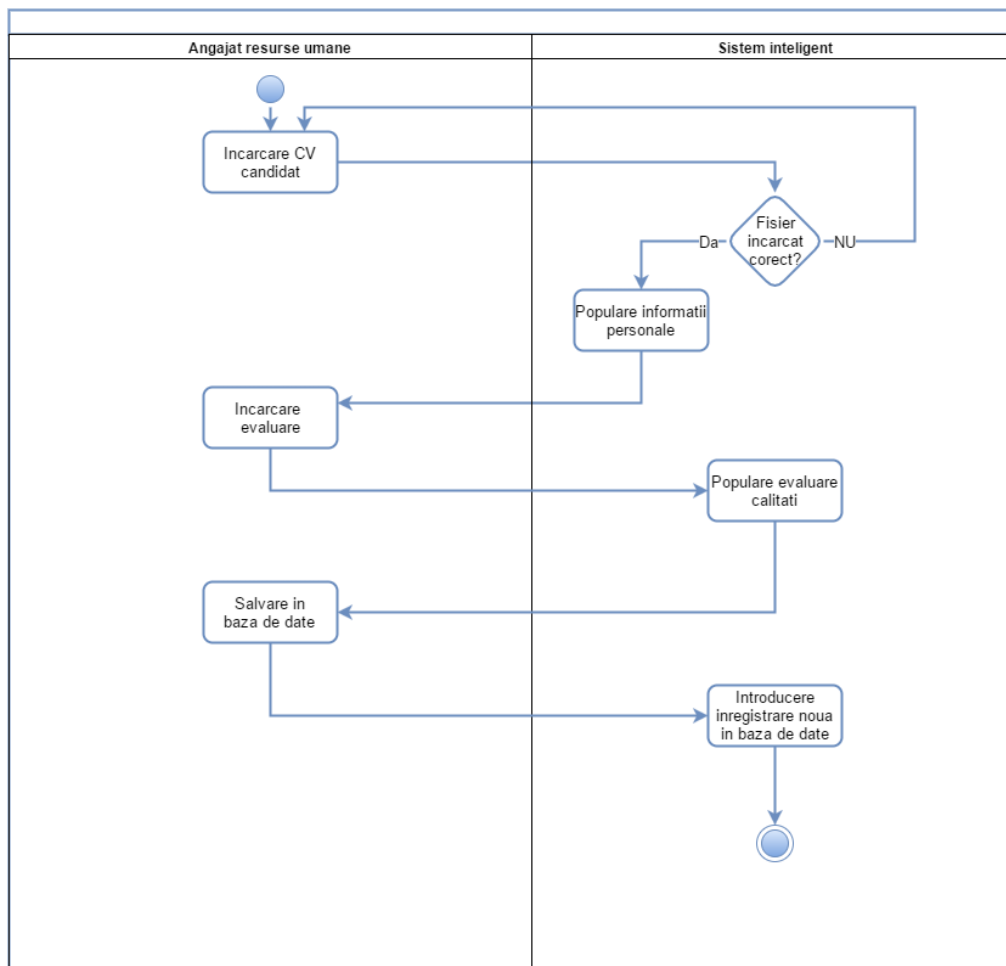


Fig IV.14. Diagrama de activitate candidat nou

În figura IV.14. este ilustrată diagrama de activitate pentru un candidat nou. Se poate observa și în diagramă că angajatul departamentului de resurse umane are posibilitatea de a încărca CV-ul unui angajat, iar, ulterior, posibilitatea de a încărca evaluările obținute de acesta la testele efectuate și de a salva datele în baza de date. În funcție de alegerile angajatului

departamentului de resurse umane, prototipul inteligent afișează pe ecran mesaje corespunzătoare activității desfășurate.

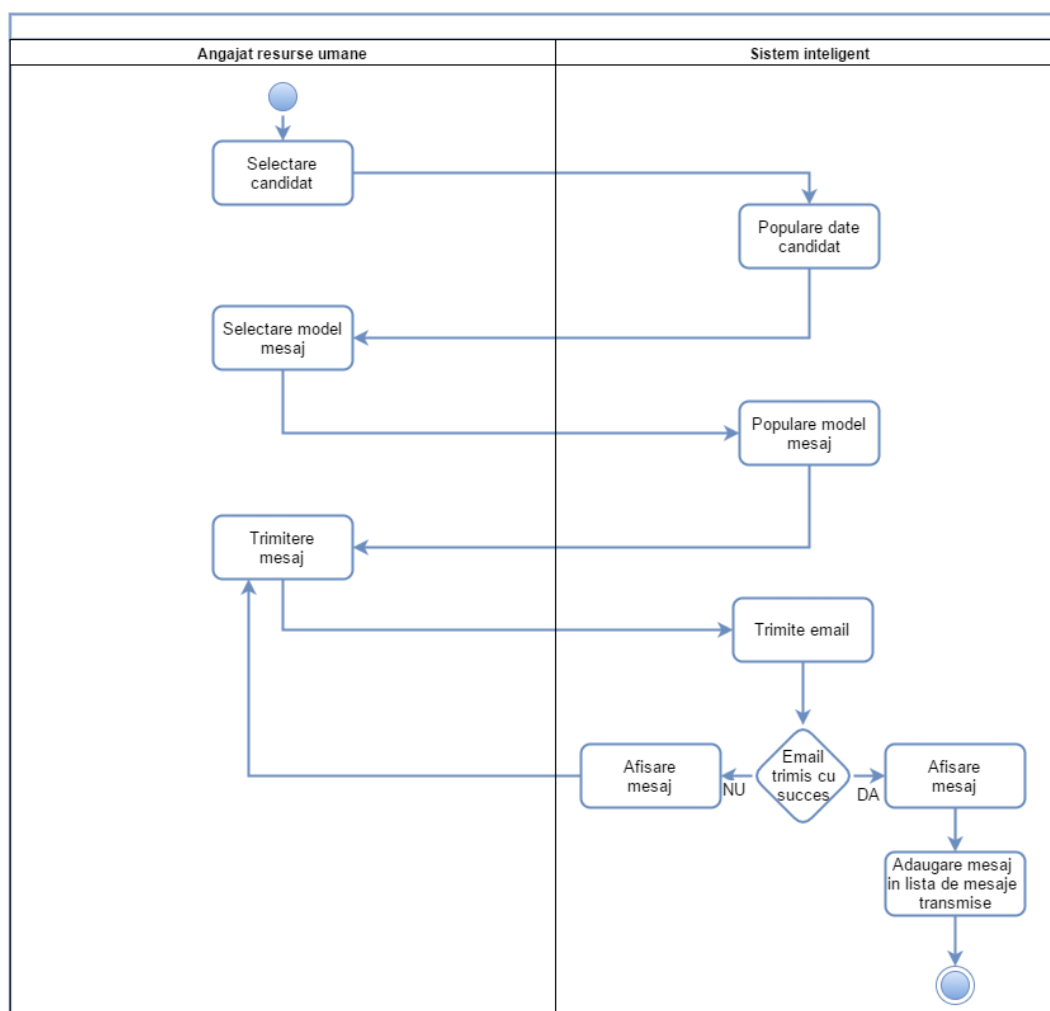


Fig IV.15. Diagrama de activitate contactează candidat

Diagrama de activitate pentru contactarea unui candidat este prezentată în figura IV.15. După cum se poate observa și în diagramă, angajatul departamentului de resurse umane are posibilitatea de a selecta candidatul pe care dorește să îl contacteze, de a selecta un model standard de mesaj și de a trimite mesajul către candidatul selectat. De asemenea, se poate observa modul în care prototipul informatic răspunde solicitărilor angajatului.

În figura următoare, figura IV.16., arată toate stările prin care trece un candidat în funcție de acțiuni, de la candidat afișat, la candidat evaluat și în final la candidat salvat.

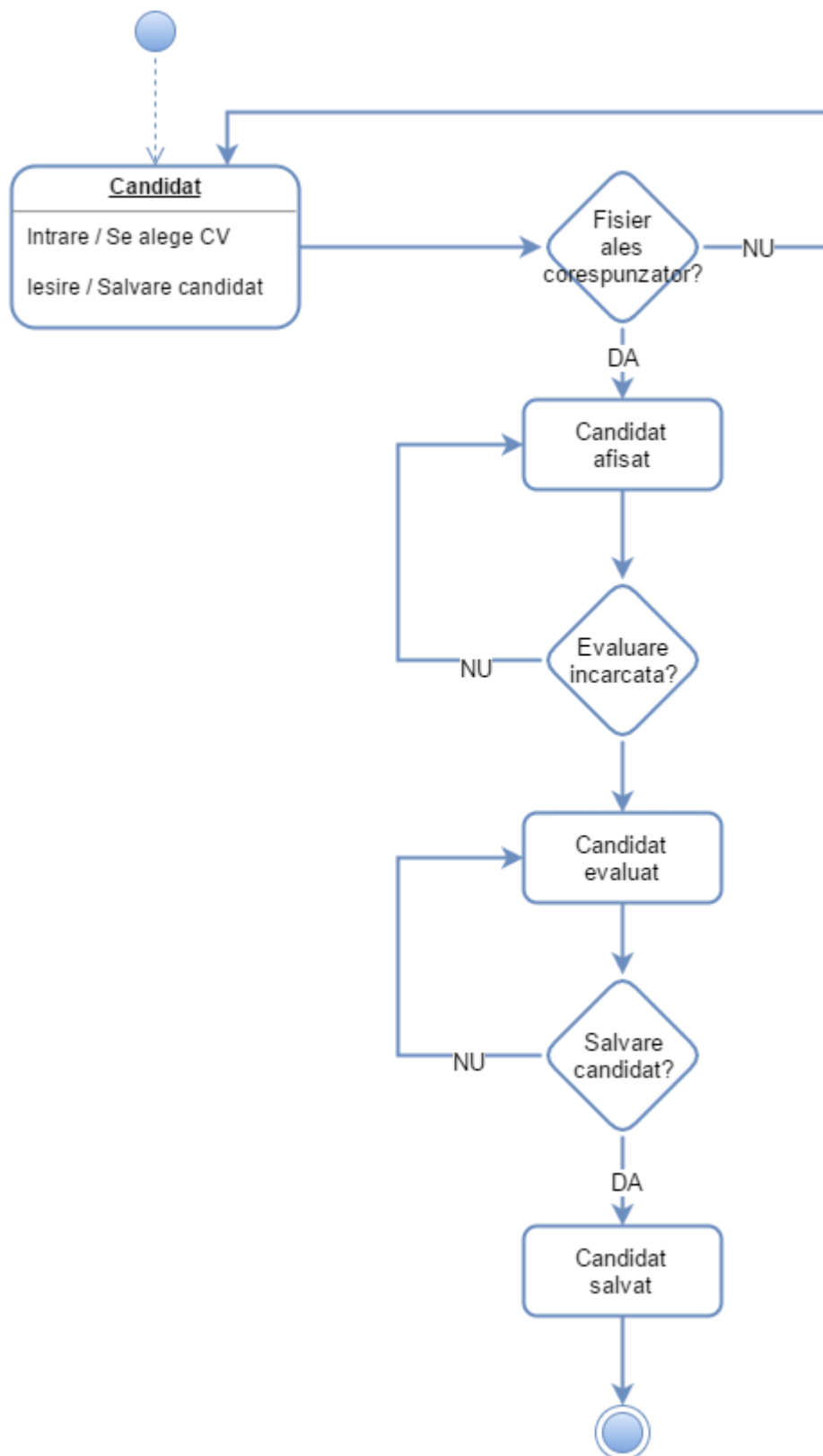


Fig IV.16. Diagrama stare candidat

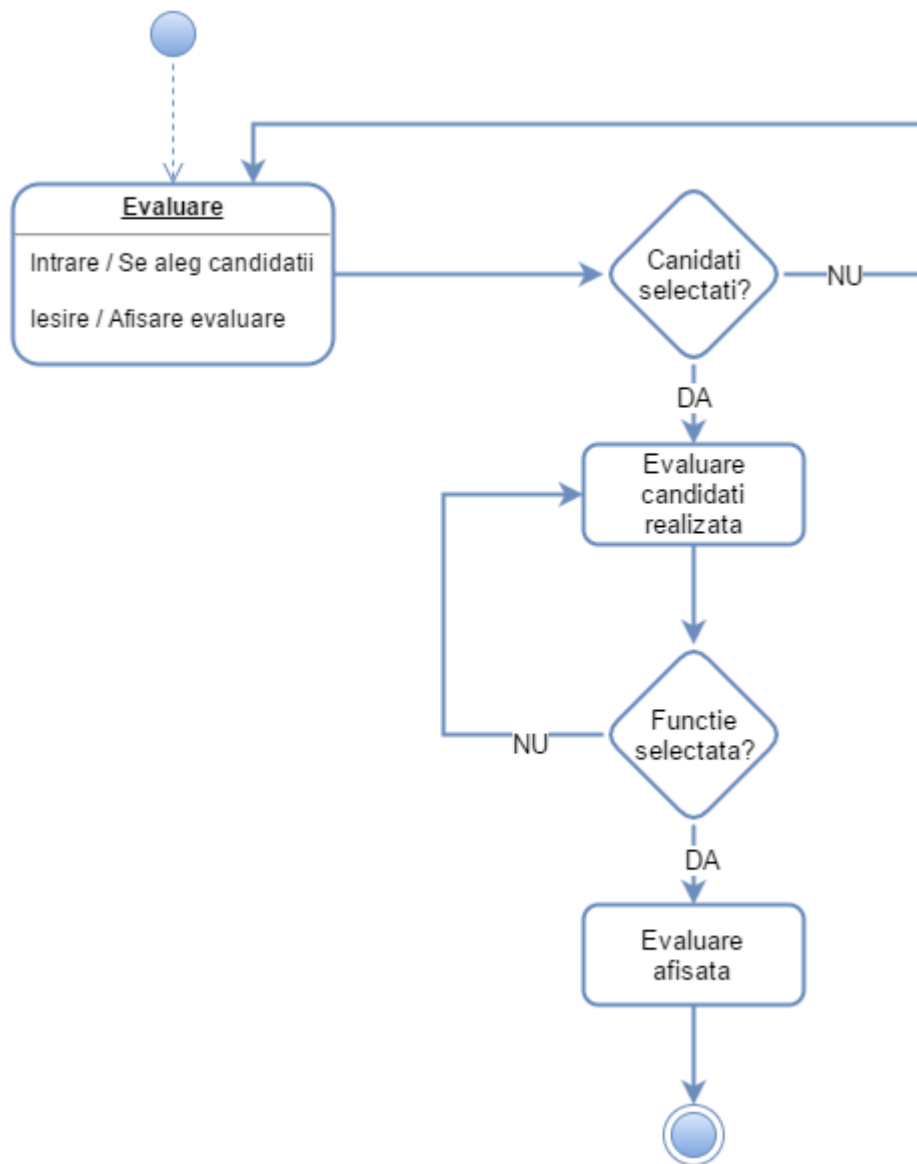


Fig. IV.17. Diagrama de stare evaluare

Diagrama precedentă, ilustrată în figura IV.17., prezintă toate stările posibile ale unei evaluări în funcție de acțiuni, pornind de la evaluarea realizată și ajungând la evaluare afișată.

Diagrama de interacțiune evidențiază obiectele, relațiile dintre ele și mesajele pe care acestea le trimit.

În figura IV.18. este prezentată diagrama de interacțiune evaluare candidat, care evidențiază obiectele implicate în efectuarea unei evaluări.

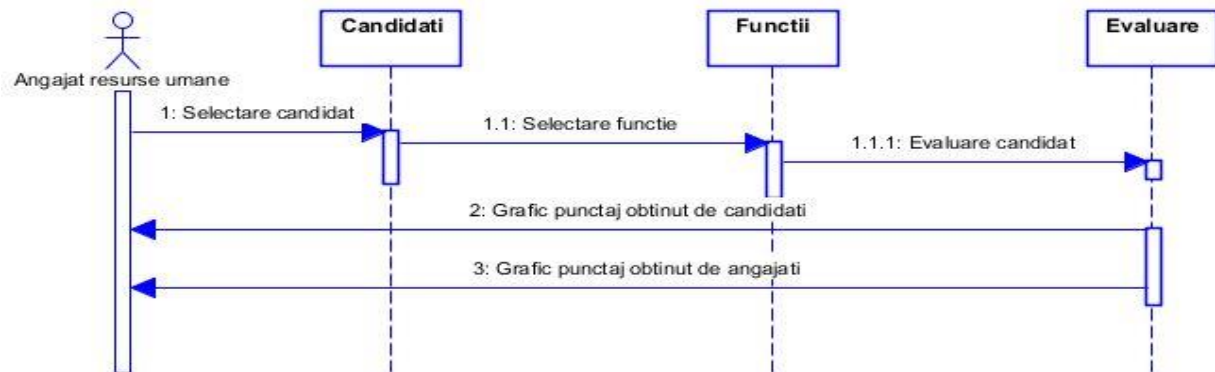


Fig IV.18. Diagrama de interacțiune evaluare candidat

În mod similar, figura IV.19., prezintă diagrama de interacțiune pentru un candidat nou, în timp ce în figura IV.20. este prezentată diagrama de interacțiune pentru contactarea unui candidat.

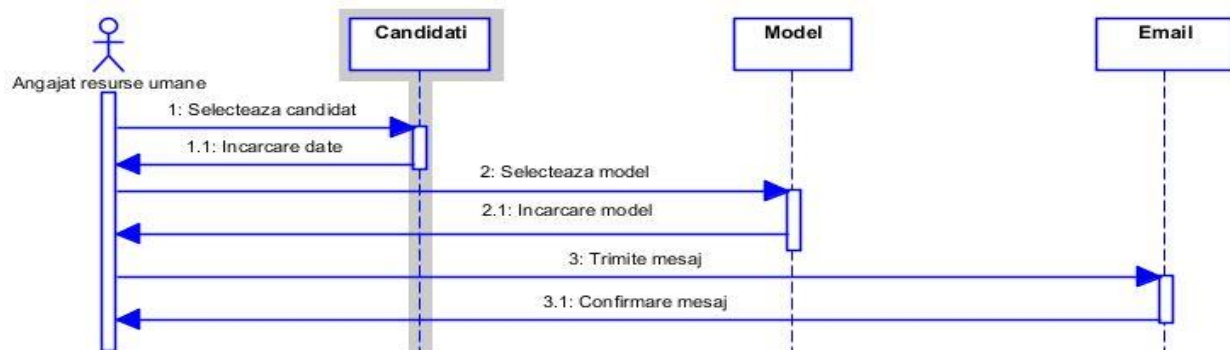


Fig. IV.19. Diagrama de interacțiune candidat nou

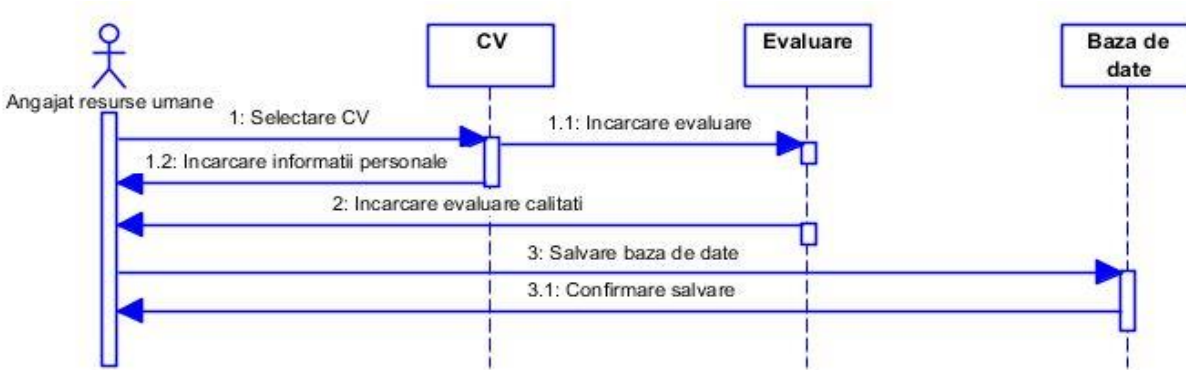


Fig IV.20. Diagrama de interacțiune contactează candidat

Ultima etapă de proiectare, o reprezintă proiectarea interfeței utilizatorului. Această etapă este utilizată pentru vizualizarea schemei interfeței care urmează a fi implementată cu ajutorul platformei Embarcadero RAD Studio.

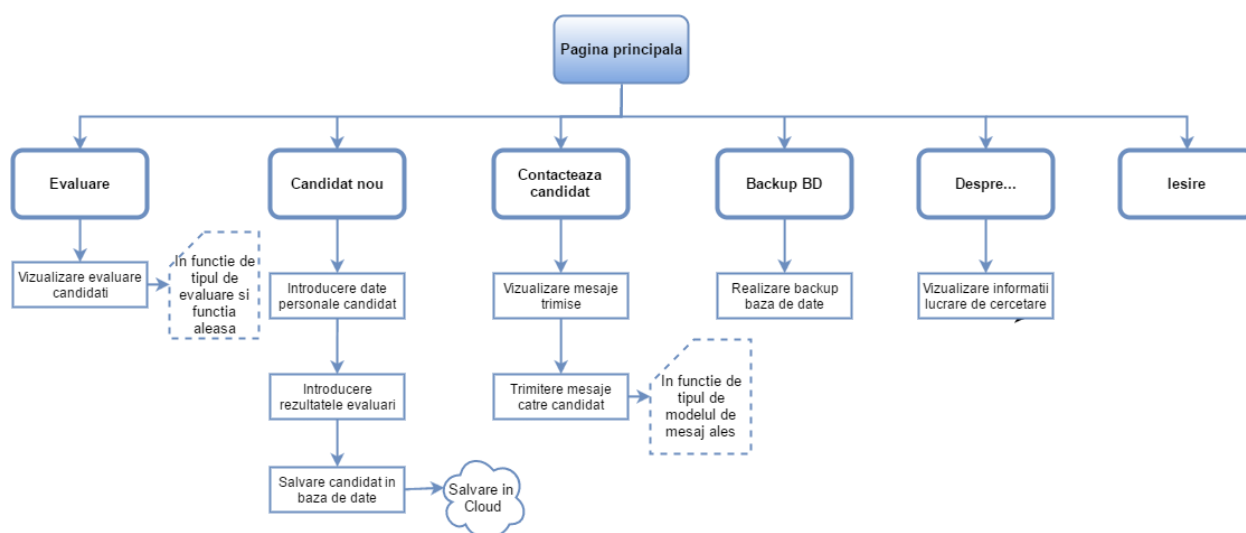


Fig. IV.21. Diagrama interfețelor utilizator

IV.2.4. Implementarea prototipului informatic

IV.2.4.1. Tehnologii informatice utilizate

Tehnologiile informatice utilizate în cadrul lucrării sunt: Embarcadero RAD Studio 10 Seattle, Oracle Database 12c, SQLDeveloper, Putty, PuttyGen, WinSCP, Draw.IO și GIMP. Am folosit Embarcadero RAD Studio 10 deoarece am dorit să realizez un prototip multi-platformă și Delphi reprezintă prima mea opțiune de programare.

Cu toate că Delphi dispune de suport nativ pentru utilizarea bazelor de date Interbase și SQLite, am decis să implementez tehnologia Oracle Database 12c. Am ales în mod special această versiune (la momentul realizării aplicației, 12.1.0.2) deoarece am dorit ca prototipul să se adreseze ultimelor cerințe în materie de baze de date și să suporte integrarea tehnologiei cloud.

Am folosit SQLDeveloper pentru o proiectare mai ușoară a bazei de date, pentru realizarea diagramei bazei de date și pentru administrarea bazei de date.

Cu ajutorul PuttyGen am generat cheile de criptare, publică și privată, necesare pentru conexiunea la o bază de date cloud. Ulterior, am folosit utilitarul Putty pentru a mă conecta la serverul găzduit în cloud și WinSCP pentru transferul de fișiere.

Pentru realizarea diagramelor BPMN am apelat la soluția oferită de Google, și anume Draw.IO. Am ales această soluție deoarece este gratuită și oferă un mod rapid de acces (nu necesită instalare) și a posibilității de a lucra direct în cloud.

Pentru editarea imaginilor folosite în aplicație, am folosit GIMP deoarece oferă gratuit toate facilitățile pe care le căutam (editare imagine, eliminare background), cu un consum minim de resurse.

Toate tehnologiile informatice utilizate au fost gratuite (*open-source*) sau cu licență de probă (*trial* 30 zile).

În continuare în cadrul lucrării am oferit detalii despre tehnologiile informatice utilizate pentru realizarea prototipului.

Embarcadero® RAD Studio™ 10 Seattle

Embarcadero® RAD Studio™ 10 Seattle reprezintă cea mai eficientă soluție pentru construirea și actualizarea aplicațiilor bogate în date, hiper-conectate, captivante vizual atât pentru Windows 10, Mac, Mobile, folosind *Object Pascal* și *C++*. Embarcadero aduce rapid și ușor aplicațiile și clienții la Windows 10 cu o gamă largă de caracteristici speciale pentru Windows, cum ar fi controale noi Windows VCL, VCL și stiluri FMX UI și servicii UWP (Universal Windows Platform), sau notificările.

Pot fi sprijinite proiecte mai mari, cu mai multe platforme, dublând memoria de proiect și capacitatea IDE disponibilă. Această versiune de Embarcadero te ajută să fi mai productiv ca oricând, cu suport multi-monitor și zeci de noi caracteristici IDE concepute pentru a ajuta utilizatorii să programeze mai rapid. RAD Studio 10 permite dezvoltatorilor să livreze aplicații de până la 5 ori mai rapid decât alte instrumente și de a construi chiar mai rapid pe mai multe platforme desktop, mobile, cloud și baze de date, inclusiv pe 32 de biți și pe 64 de biți pentru Windows 10, Mac OSX, iOS și Android.

Windows 10 este cea mai importantă versiune de windows din ultimii ani și reprezintă o mare oportunitate pentru dezvoltatorii de software. Acesta se confruntă cu o adoptare rapidă, și este de așteptat să fie pe 1 miliard de dispozitive în următorii câțiva ani. Acum este momentul pentru a muta aplicațiile și utilizatorii la Windows 10. RAD Studio 10 Seattle, aduce aceste noi caracteristici pentru platforma Windows 10 în aplicații, rapid și ușor.

„Windows 10: O nouă oportunitate pentru dezvoltatori” [WIN10] este o evaluare în profunzime a motivului pentru care dezvoltatorii RAD Studio sunt plasați în mod ideal pentru a profita de Windows 10 și de ce acum este momentul să se pregătească pentru noi modalități de dezvoltare și implementare pentru aplicațiile Windows 10¹²⁵.

RAD Studio 10 Seattle oferă facilitățile *FireUI Multi-Device Designer* și un framework cross-platform UI, care oferă singura și unica adevărată soluție pentru aplicațiile compilate nativ. Cei mai mulți alți furnizori, care sprijină dezvoltarea nativă multi-platformă necesită ca interfețe separate de utilizator să fie scrise pe fiecare platformă.

¹²⁵ [WIN10] Embarcadero - *Windows 10: The Big New Opportunity for Developers*, July 2016

Dezvoltatorii de aplicații mobile pot construi în cele din urmă un aspect comun, nativ și pot simți interfața utilizatorului, care funcționează pe factori de formă pe multiple telefoane mobile, tablete și sisteme desktop - toate o dată! Aplicațiile pot avea acces la API-uri platformă, senzori de dispozitive și servicii și pot oferi cea mai bună performanță a aplicației cu GPU nativ și suport CPU toate dintr-o bază de cod comun și partajate. Suportul modern, multi platformă include Windows 10, iOS 8.4, Android 5.1.1 și OS X Yosemite.

Kike Perez, manager de sistem pentru HabitatSoft afirmă că experiența de dezvoltare oferită de Delphi devine din ce în ce mai bună. Personal, menționez că și pentru mine Delphi reprezintă prima opțiune când vine vorba despre alegerea platformei de dezvoltare.

RAD Studio este o soluție de dezvoltare de software premiată, folosită de milioane de dezvoltatori din întreaga lume și sprijinită de o comunitate activă de dezvoltatori de software, parteneri de tehnologie și furnizori de componente.

Oracle Database 12c

Pe măsură ce organizațiile îmbrățișează tehnologia *cloud*, aceștia caută tehnologii care să transforme afacerile și să îmbunătățească agilitatea lor generală operațională și eficiență. Oracle Database 12c este o bază de date de ultimă generație concepută pentru a satisface aceste nevoi, oferind o nouă arhitectură *multitenant* pe infrastructura unei platforme rapide, scalabile, fiabile și sigure de bază de date. Prin conectarea în cloud cu Oracle Database 12c, clienții pot îmbunătăți calitatea și performanța aplicațiilor, pot economisi timp cu arhitectura de disponibilitate maximă și managementul de stocare și pot simplifica consolidarea unei bazei de date, prin gestionarea a sute de baze de date ca una singură.

Oracle Database 12c abordează provocările cheie ale clienților, care consolidează bazele de date într-un model de cloud privat, oferind eficiență mult îmbunătățită și costuri de gestionare mai mici, și păstrând în același timp autonomia de baze de date separate.

Oracle *multitenant* este o nouă caracteristică a Oracle Database 12c, și permite ca fiecare bază de date conectată la noua arhitectură *multitenant* să arate și să se simtă ca un standard Oracle Database pentru aplicații; astfel încât aplicațiile existente pot rula neschimbate. Oracle *multitenant* gestionează mai multe baze de date ca fiind una și poate crește utilizarea resurselor

de server și de a reduce timpul și efortul necesar pentru upgrade-uri de baze de date, backup, recuperare, și multe altele.

Oracle *multitenant* funcționează cu toate caracteristicile de baze de date Oracle, inclusiv Real Application Clusters, partiționare, data guard, compresie, management de stocare automat, testare aplicații reale, criptare date transparente, și multe altele.

Pentru a ajuta clienții să gestioneze în mod eficient mai multe date, cu costuri mai mici de depozitare și de a îmbunătăți performanța bazei de date, Oracle Database 12c introduce noi caracteristici automate de optimizare a datelor. Astfel, o hartă de căldură monitorizează activitatea de citire / scriere a bazei de date și permite administratorilor bazei de date să identifice cu ușurință datele care sunt foarte active, numai citite sau rar citite, stocate în tabele și partiții.

Folosind compresia și stocarea inteligentă, administratorii bazelor de date pot defini cu ușurință politicile pentru a comprima în mod automat și organiza nivelul OLTP, depozitul de date și arhivarea datelor bazate pe activitatea și vârsta datelor.

SQLDeveloper

Oracle SQL Developer este un mediu liber de dezvoltare integrată, care simplifică dezvoltarea și gestionarea bazelor de date Oracle în ambele implementări, tradițională și Cloud. SQL Developer oferă posibilități de dezvoltare complete end-to-end a aplicațiilor PL / SQL, o foaie de lucru pentru rularea interogărilor și script-urilor, o consolă DBA pentru gestionarea bazei de date, o interfață de rapoarte, o soluție completă de modelare a datelor, precum și o platformă de migrare pentru mutarea de baze de date de la alți furnizori la Oracle.

Putty

PuTTY este un emulator de terminal gratuit și open-source, consolă serială și aplicație de transfer de fișiere pe rețea. Acesta suportă mai multe protocoale de rețea, inclusiv SCP, SSH, sau Telnet. Se poate conecta, de asemenea, la un port serial. PuTTY a fost scris inițial pentru Microsoft Windows, dar a fost adaptat pentru diverse alte sisteme de operare. Porturile oficiale sunt disponibile pentru unele platforme Unix, cu porturi de lucru în curs de execuție la Classic

Mac OS și Mac OS X, iar porturile neoficiale au contribuit la platforme, cum ar fi Symbian, Windows Mobile și Windows Phone [PWIKI]¹²⁶. Este scris și menținut în primul rând de Simon Tatham [PUTTY]¹²⁷.

PuTTY este alcătuit din mai multe componente [PUTTY]:

- PuTTY: Telnet, rlogin, și clientul SSH în sine, care se poate conecta, de asemenea, la un port serial
- PSCP: un client SCP, adică linia de comandă de copiere fișiere securizată
- PSFTP: un client SFTP, adică sesiuni generale de transfer de fișiere, cum ar fi FTP
- PuTTYtel: doar ca Telnet-client
- PLINK: o interfață de linie de comandă pentru consola PuTTY
- Pageant: un agent de autentificare SSH pentru chituri, PSCP și PLINK
- PuTTYgen: un utilitar de generare chei RSA și DSA
- pterm: un emulator terminal de sine stătător

WinSCP

WinSCP (Windows Secure Copy) este un program gratuit și open-source SFTP, FTP, WebDAV și client SCP pentru Microsoft Windows. Funcția sa principală este de transfer de fișiere securizat între un calculator local și un calculator la distanță. În plus, WinSCP oferă management de fișiere de bază și funcționalitatea de sincronizare fișier. Pentru transferuri sigure, folosește Secure Shell (SSH) și acceptă protocolul SCP, în plus față de SFTP [WSCPW]¹²⁸.

Conform Wikipedia [WSCPW], dezvoltarea WinSCP a început, aproximativ, în martie 2000 și continuă. Inițial a fost găzduit de Universitatea de Economie din Praga, pentru care autorul lucra la momentul respectiv. Din 16 iulie 2003 este licențiat sub GNU GPL și găzduit pe SourceForge.net.

¹²⁶ [PWIKI] Putty, Wikipedia - <https://en.wikipedia.org/wiki/PuTTY>

¹²⁷ [PUTTY] Putty: A Free SSH and Telnet Client

¹²⁸ [WSCPW] WinSCP, Wikipedia - <https://en.wikipedia.org/wiki/WinSCP>

Draw.IO

Conform descrierii Google [GDRAW], draw.io pro este o aplicație complet gratuită pentru diagrame produsă de Google Drive(TM), care oferă posibilitatea de a desena: diagrame de proces (Flowchart), diagrame structurale (UML), diagrame relaționale (ERD), diagrame de rețea, modele de proces de afaceri, grafice organizaționale, circuite electronice, schematizare și machetare.

Printre facilitățile draw.io se numără și: client nativ HTML 5 cu suport deplin pentru IE 6-8, bibliotecă mare de șabloane inclusă, interfață intuitivă de tip drag and drop, funcționalitate pentru căutare și adăugare de imagini, exportare în format PNG/JPG/XML/SVG/PDF, suport pentru echipamente cu touch-screen, colaborare în timp real, includerea diagramei în bloguri și site-uri de tip wiki.

GIMP

Conform descrierii oficiale [GIMP], GIMP este un editor de imagine multi-platformă disponibil pentru GNU / Linux, OS X, Windows și mai multe sisteme de operare. Acesta este un software gratuit, care permite schimbarea codului sursă și distribuirea modificărilor¹²⁹.

Indiferent dacă utilizatorul este un designer grafic, fotograf, grafician, sau om de știință, GIMP oferă instrumente sofisticate pentru ca oricine să își poată îndeplini obiectivele.

WEKA

Conform prezentării oficiale [WEKA], WEKA este o colecție de algoritmi de învățare pentru sarcini de exploatare a datelor¹³⁰. Acești algoritmi pot fi aplicați, fie direct pe un set de date, sau direct din propriul cod dezvoltat în Java. WEKA conține instrumente pentru date pre-procesare, clasificare, regresie, clusterizare, reguli de asociere și vizualizare. De asemenea, este

¹²⁹ [GIMP] GIMP – *About GIMP*, <https://www.gimp.org/about/>

¹³⁰ [WEKA] The University of Waikato - Weka, <http://www.cs.waikato.ac.nz/ml/weka/>

foarte potrivit pentru dezvoltarea unor noi scheme de învățare automată. WEKA este o aplicație gratuită (*open source*) emisă sub licența GNU General Public.

IV.2.4.2. Realizarea prototipului informatic

În primă fază am configurat mediul pentru ca ulterior să pot proiecta o bază de date. Astfel, urmând etapele descrise în ANEXA A – Instalare BD am instalat o binare Oracle 12c și am creat o nouă instanță de bază de date. Am optat pentru varianta de a crea noua instanță la instalarea binării, pentru a nu folosi ulterior utilitarul *DBCA* (*Database Configuration Assistant*), deoarece în cadrul proiectului voi folosi o singură instanță de bază de date.

Următoarea etapă a constat în configurarea SQLDeveloper, realizând astfel conexiunea pentru userul SYS. O dată realizată conexiunea, am creat o nouă schemă (c##dw) conform scriptului:

```
CREATE USER C##DW IDENTIFIED BY <parola>;  
ALTER USER "C##DW"  
DEFAULT TABLESPACE "USERS"  
TEMPORARY TABLESPACE "TEMP"  
ACCOUNT UNLOCK;  
ALTER USER "C##DW" DEFAULT ROLE "CONNECT", "RESOURCE";
```

Odată creată această schemă, am realizat o conexiune către aceasta în SQLDeveloper, pe care am folosit-o pentru proiectarea bazei de date. Scripturile folosite pentru crearea tabelor și a constrângerilor au fost detaliate în ANEXA – BD.

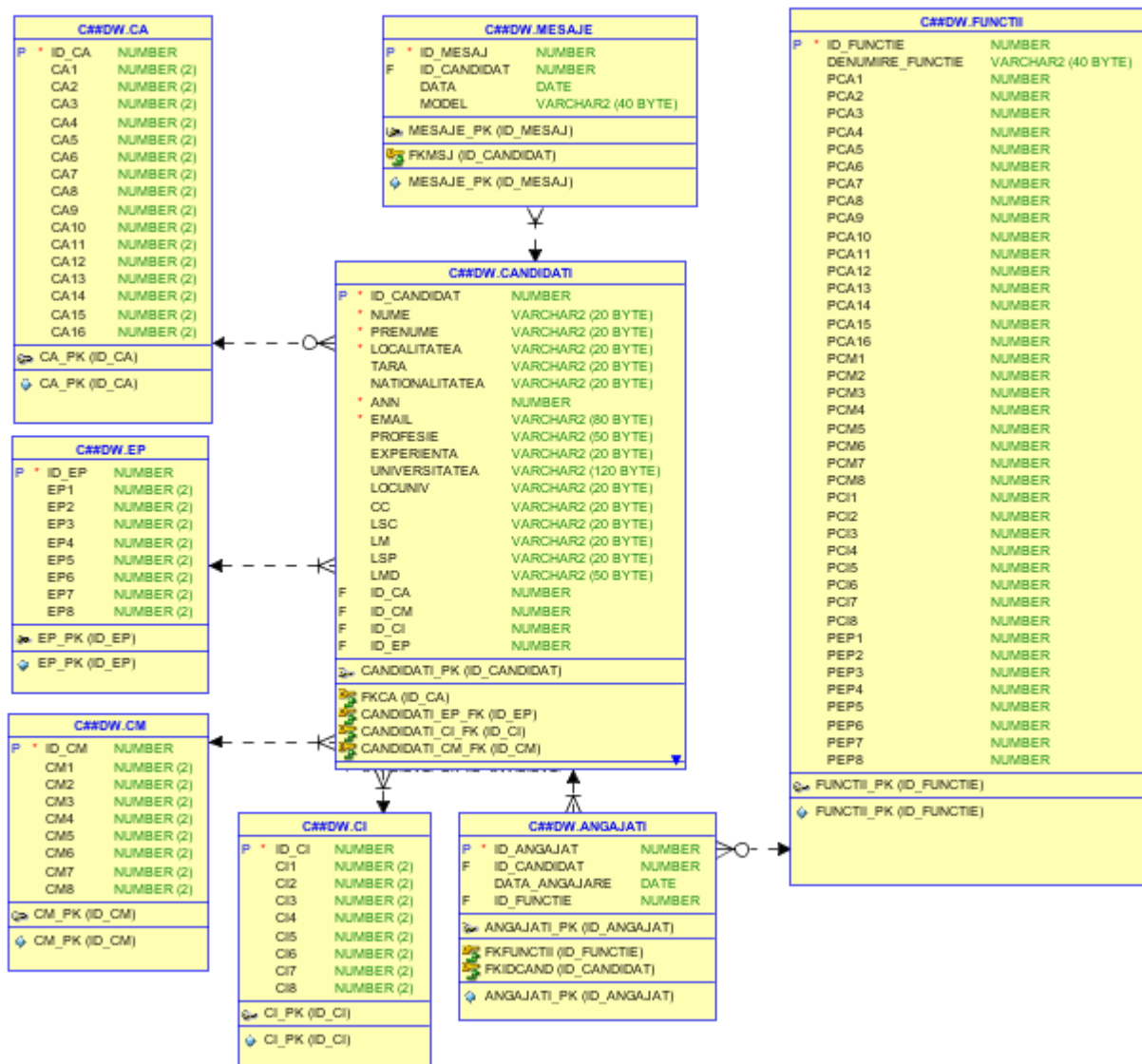


Fig. IV.22. Diagrama bazei de date

Pentru a asigura unicitatea tuturor cheilor primare, am decis să creez câte o secvență pentru cheia primară a fiecărei tabele. Au fost astfel create secvențele SEQ_IDCA, SEQ_IDCANDIDAT, SEQ_IDCI, SEQ_IDCM, SEQ_IDEP, SEQ_IDMSJ. Mai jos se poate observa scriptul de creare a secvenței SEQ_IDCA, celelalte fiind detaliate în ANEXA – BD.

```
CREATE SEQUENCE SEQ_IDCA START WITH 100 INCREMENT BY 1 MAXVALUE 1000 NOCYCLE;
```

Am considerat că ar fi util și un backup al bazei de date, și am decis să implementez varianta cu data-pump. Am optat pentru un export doar al schemei de lucru și nu al întregii bazei de date. În prealabil, am creat un director temporar, denumit *temp*, la nivelul bazei de date, în care să salvez fișiere de export și logurile.

```
CREATE DIRECTORY temp AS 'C:\APP\TEMP';
```

Am efectuat apoi un export de test, conform scriptului de mai jos, pentru a valida procedura.

```
expdp 'c##dw/dw123' schemas=c##dw directory=temp dumpfile=export_dw.dmp  
logfile=export_dw.log reuse_dumpfiles=y
```

Ulterior, pentru a putea apela exportul din aplicație, a fost necesară crearea unui fișier *batch* pe baza scriptului de export prezentat.

Aplicația a fost testată cu baza de date proiectată local și a funcționat corespunzător. Am decis că este important ca aplicația să beneficieze de tehnologie de ultimă generație și am pregătit o infrastructură similară de baza de date în cloud, în vederea migrării datelor de pe platforma locală către cloud.

Urmând pașii prezentați în ANEXA – Cloud am configurat o bază de date Oracle 12c Standard Edition în Oracle Cloud. Am optat pentru versiunea Standard Edition, deoarece în urma dezvoltării efectuate local, am remarcat că nu sunt necesare facilitățile versiunii Enterprise Edition. Astfel, pentru a avea și un cost de licențiere cât mai redus, aplicația poate rula corespunzător și pe versiunea Standard Edition.

După confiurarea mediului Oracle Cloud, am configurat la nivelul instanței nou create un director temporar, denumit *temp*, pentru a putea realiza migrarea datelor de pe mediul local și viitoarele backup-uri ale bazei de date, păstrând tot varianta cu *data-pump*.

```
CREATE DIRECTORY temp AS '/HOME/ORACLE/TMP';
```

Am folosit scriptul de backup pentru a realiza un export al bazei de date locale în vederea migrării acesteia în cloud. Am copiat utilizând WinSCP fișierul *dump* exportat de pe mediul local, în directorul */home/oracle/tmp* din cloud. Apoi, utilizând importul cu *data-pump* am finalizat procesul de migrare al datelor în cloud.

```
impdp \"/ as sysdba \" directory=temp dumpfile=export_dw.dmp  
logfile=import_dw.log
```

Scriptul a rulat cu succes și datele au fost migrate în cloud. În continuare, a fost necesară modificarea aplicației și actualizarea datelor de conectare la baza de date, astfel încât, conectarea să nu se mai facă la baza de date locală, ci la baza de date din Oracle Cloud. De asemenea, a fost necesară modificarea fișierului *batch* pe baza căruia aplicația realizează backup-ul.

```
expdp 'c##dw/dw123@cloud' schemas=c##dw directory=temp dumpfile=export_dw.dmp
logfile=export_dw.log reuse_dumpfiles=y
```

Noul script de backup se conectează la baza de date din cloud folosind datele de conectare către Oracle Cloud, adăugate în TNSNAMES.ORA. De asemenea, am utilizat și clauza reuse_dumpfiles=y pentru a optimiza spațiul consumat de backup-uri, printr-o suprascriere a vechiului fișier exportat.

În a doua fază am folosit Weka pentru a calcula ponderile pentru toate atributele în funcție de funcțiile din companie. Astfel, au fost calculate ponderi pentru următoarele funcții: Dezvoltator Java, Administrator BD, Administrator rețea și Șef serviciu IT.

Pentru calcularea ponderilor, a fost necesară formatarea datelor conform cerințelor Weka. Astfel, au fost realizate mai multe fișiere .arff care să conțină datele pe baza cărora au fost determinate ponderile atributelor fiecărei funcții.

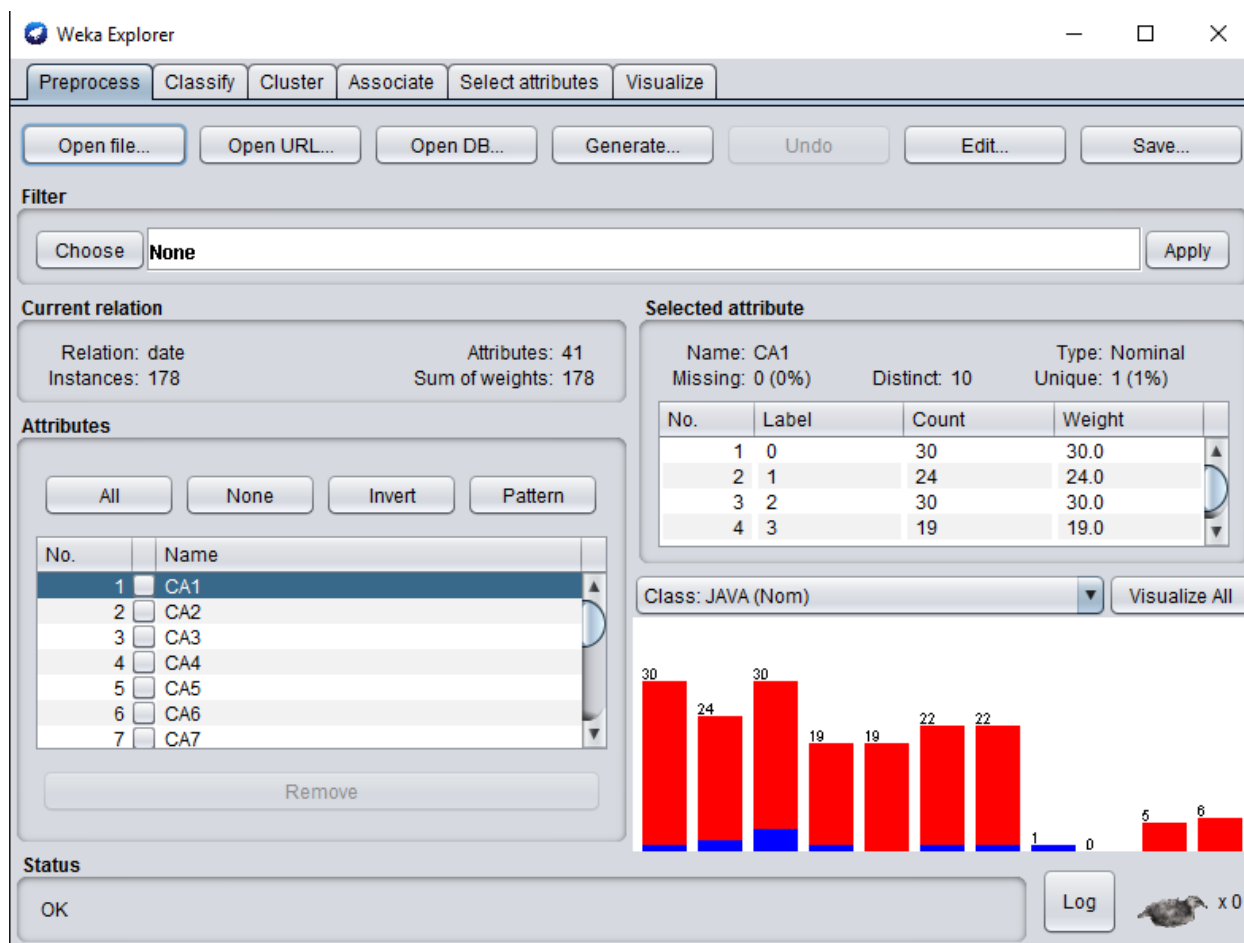


Fig. IV.23. Vizualizarea datelor în Weka

Datele introduse în program au fost prelucrate folosind componenta de evaluare a atributelor prin funcția *InfoGainAttributeEval*. În acest mod, am calculat câștigul de informație (*Information Gain*), pentru fiecare funcție în parte (Dezvoltator Java – Fig ZZ, Administrator rețea, Administrator BD și Șef serviciu IT). Acest indicator este util atunci când se dorește să se stabilească importanța (influența) atributelor. Se poate aplica atât pentru un atribut, cât și pentru mai multe, sau chiar pentru toate attributele, așa cum am aplicat în cadrul studiului.

După cum se poate observa și în imaginea următoare, programul ne-a furnizat în ordine descrescătoare a importanței, attributele și ponderea acestora.

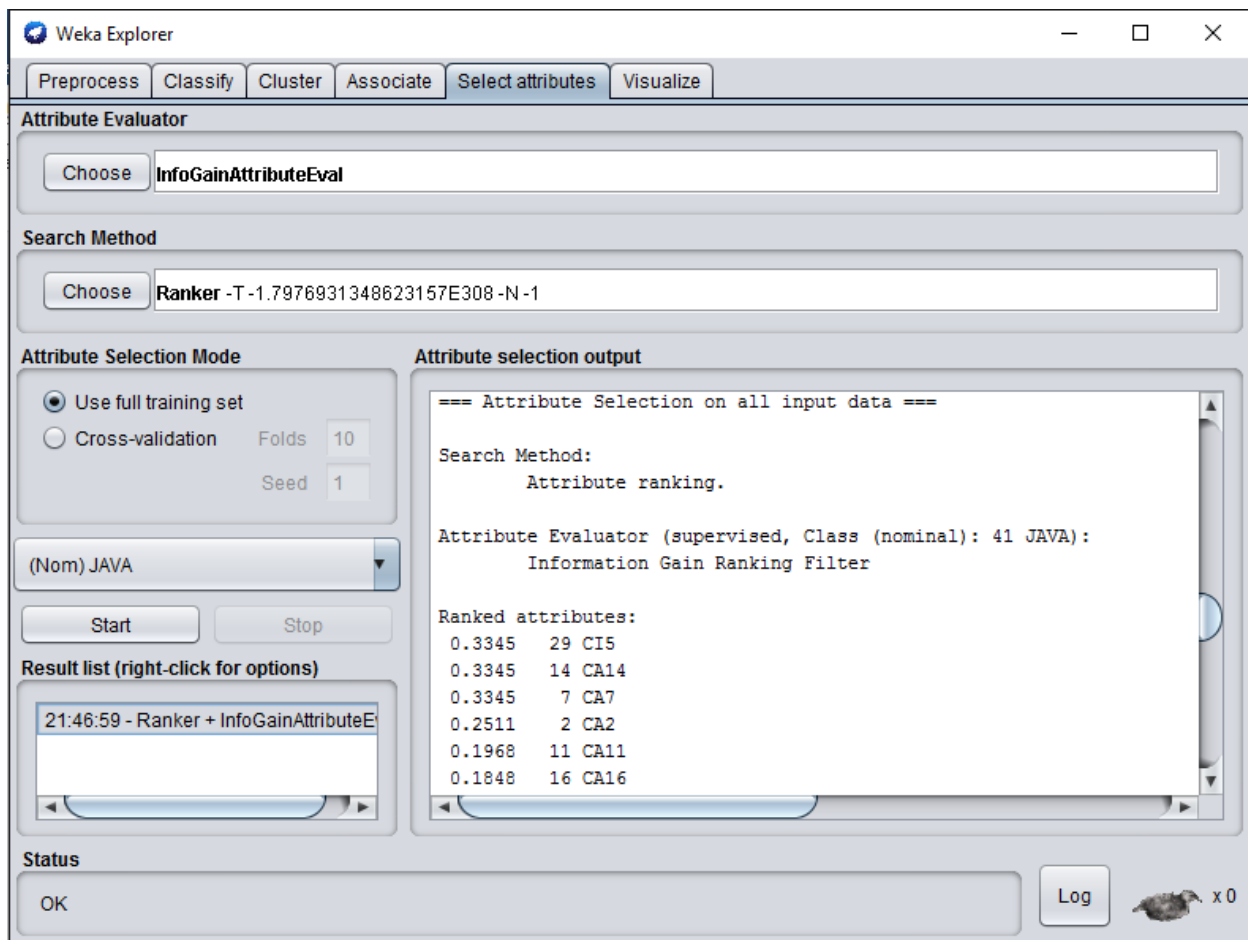


Fig. IV.24. Selecția atributelor în Weka

Atribut	Dezvoltator Java	Administrator BD	Administrator rețea	Șef serviciu IT
CA1	0.05	0.33454	0.04819	0.02495
CA2	0.2511	0.04969	0.02753	0.24213
CA3	0.1556	0.13337	0.16054	0.15103
CA4	0.1078	0.31906	0.0422	0.02348
CA5	0.0977	0.01616	0.37714	0.02271
CA6	0.0513	0.03681	0.34345	0.05129
CA7	0.3345	0.03572	0.05036	0.04597
CA8	0.0513	0.02628	0.07083	0.04412
CA9	0.059	0.29957	0.03049	0.03386
CA10	0.0445	0.02774	0.37714	0.03197
CA11	0.1968	0.1543	0.03967	0.19231
CA12	0.1433	0.02638	0.0257	0.01524
CA13	0.0575	0.30051	0.02471	0.03609
CA14	0.3345	0.06063	0.05681	0.06348
CA15	0.1005	0.03594	0.35687	0.03435
CA16	0.1848	0.06249	0.19783	0.20364
CM1	0.148	0.13462	0.16027	0.17507
CM2	0.1328	0.14017	0.15571	0.16255
CM3	0.0251	0.02057	0.02474	0.01489
CM4	0.049	0.06775	0.24394	0.1804
CM5	0.0311	0.01754	0.04084	0.02124
CM6	0.0373	0.0515	0.00982	0.01436
CM7	0.0498	0.02486	0.0429	0.03174
CM8	0.0285	0.04091	0.02249	0.2844
CI1	0.0385	0.02941	0.04209	0.03111
CI2	0.0648	0.02212	0.02055	0.30236
CI3	0.0454	0.02454	0.01062	0.30051
CI4	0.1337	0.13368	0.15899	0.13788
CI5	0.3345	0.0315	0.07589	0.04428

CI6	0.0231	0.00722	0.03746	0.04899
CI7	0.0442	0.33454	0.02625	0.00991
CI8	0.0166	0.04549	0.02095	0.07653
EP1	0.1494	0.14735	0.16098	0.141
EP2	0.0448	0.19743	0.20493	0.13657
EP3	0.0308	0.24121	0.04029	0.18689
EP4	0.0135	0.02576	0.03753	0.26826
EP5	0.0224	0.03116	0.02082	0.26725
EP6	0.0286	0.01747	0.0555	0.02245
EP7	0.0423	0.01705	0.04271	0.04145
EP8	0.0249	0.00699	0.00853	0.03034

Tabel IV.X. Ponderi atribuite determinate folosind Weka

Aceste ponderi au fost introduse manual în baza de date creată, în tabela funcții, pe baza interogărilor prezentate în ANEXA – BD. Pe baza acestora se va calcula punctajul fiecărui candidat.

IV.2.4.3. Prezentarea prototipului informatic

Dupa cum se poate observa și în imaginea următoare, în care este prezentat ecranul principal al aplicației, meniul aplicației oferă următoarele opțiuni: Evaluare, Candidat nou, Contactează candidat, Backup BD, Despre... și Ieșire.

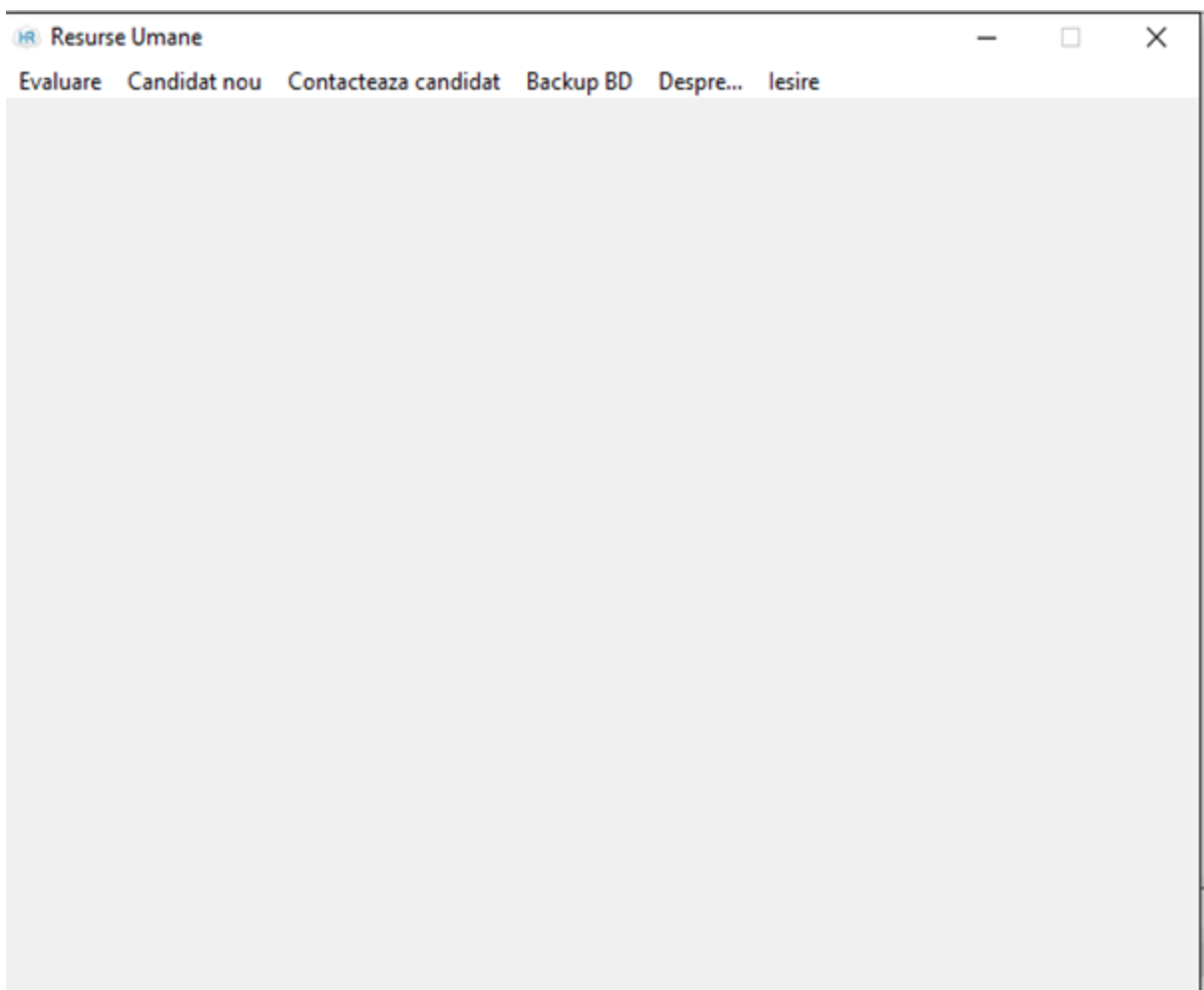


Fig IV.25. Ecranul principal al aplicației

Modulul Evaluare, conține o listă de persoane, din care, se pot alege candidații pentru care se dorește să se realizeze evaluarea. Implicit, la pornirea modulului, este activată varianta externă de recrutare a angajaților. În acest caz, popularea listei de persoane se realizează pe baza următoarei interogări:

```
select Nume||' '||Prenume from candidati where candidati.ID_CANDIDAT NOT IN (SELECT  
ID_CANDIDAT FROM Angajati) order by Nume,Prenume
```

Dacă se dorește evaluarea unor persoane, care sunt deja angajate în cadrul companiei, se poate alege opțiunea internă de recrutare. Comutarea între cele două variante s-a realizat prin utilizarea a doua butoane RadioButton. În momentul comutării pe varianta internă de recrutare, popularea listei de persoane se realizează pe baza următoarei interogări:

```
select Nume||' ' ||Prenume from candidati where candidati.ID_CANDIDAT IN (SELECT ID_CANDIDAT FROM Angajati) order by Nume,Prenume
```

De asemenea, în cadrul listei se pot selecta unul sau mai mulți candidați pentru evaluare.

Fig IV.26. Modulul evaluare

Modulul *Evaluare* oferă posibilitatea de selecție a funcției din cadrul companiei pentru care se evaluează candidații. Astfel, se pot realiza evaluări pentru Administrator BD, Administrator retea, Dezvoltator Java și Șef serviciu IT.

Pentru a putea rula evaluarea candidaților (butonul *Evalueaza*), este obligatoriu să avem o selecție în lista de funcții și cel puțin o selecție în lista de candidați.

Evaluarea candidaților se realizează pe baza valorilor introduse în baza de date pentru atributele lor și ponderile acestora. Ponderile au fost determinate folosind Weka iar formula de calcul pentru punctajul final este:

$$Punctaj = \frac{\sum_{i=1}^{16} CA_i * PCA_i + \sum_{i=1}^8 CM_i * PCM_i + \sum_{i=1}^8 CI_i * PCI_i + \sum_{i=1}^8 EP_i * PEP_i}{\sum_{i=1}^{16} CA_i + \sum_{i=1}^8 CM_i + \sum_{i=1}^8 CI_i + \sum_{i=1}^8 EP_i} \quad (1)$$

Astfel, folosind formula precedentă (1), se calculează punctajele candidaților și se realizează un grafic pentru a oferi utilizatorului posibilitatea de a interpreta ușor rezultatul calculului și diferențele dintre candidați. De asemenea, în funcție de funcția aleasă se realizează un grafic similar cu persoanele angajate în companie, pentru a se evidenția nivelul angajaților și diferențele dintre ei și candidați.

Butonul *Inapoi* oferă posibilitatea utilizatorului de a părăsi modulul *Evaluare*, prin întoarcerea la pagina principală a aplicației.

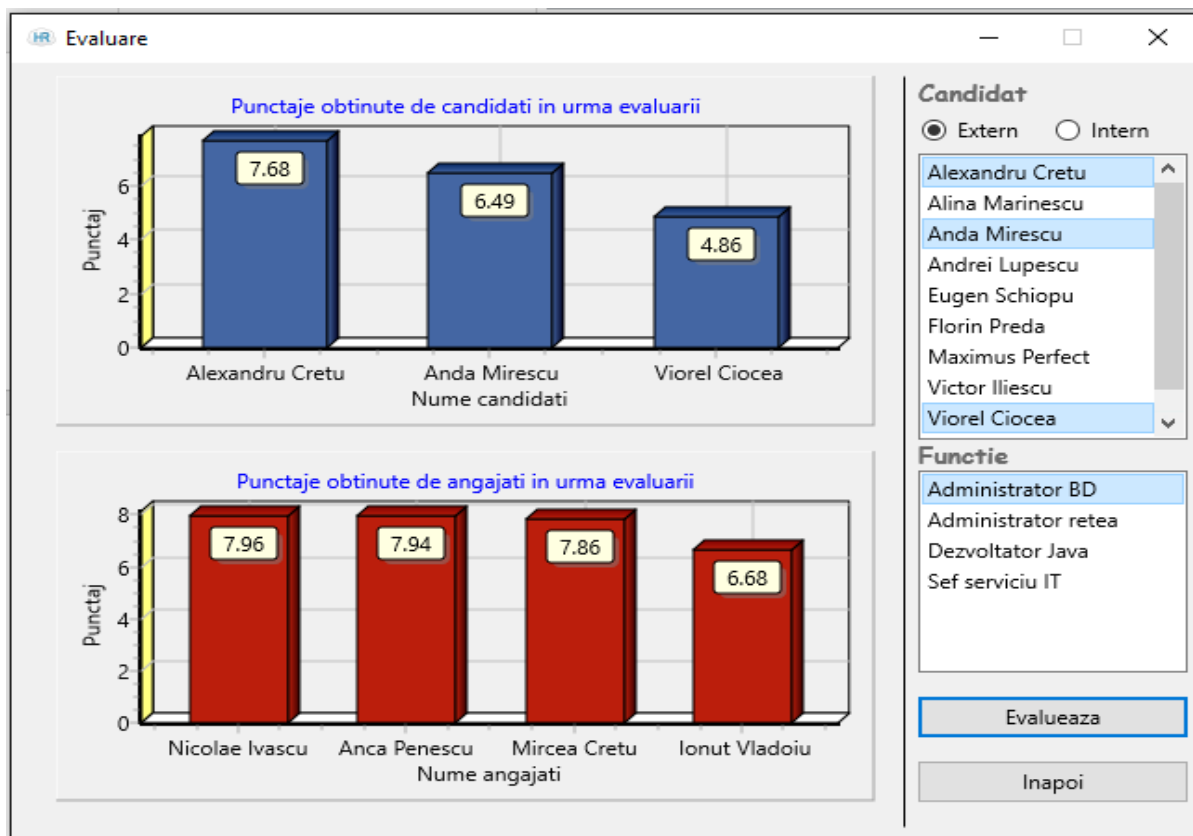


Fig IV.27. Modulul evaluare – Exemplu de evaluare a candidaților

Modulul *Candidat* nou oferă posibilitatea utilizatorului de a introduce o nouă persoană în baza de date. În zona de *Informații personale* a modulului, sunt afișate datele personale ale candidatului. Acestea se vor popula automat prin încărcarea CV-ului candidatului. De asemenea, programul oferă și posibilitatea de editare manuală a câmpurilor după încărcarea din fișier. În zona de *Evaluare calități* a modulului, sunt afișate valorile determinate în urma testelor (cunoștințe și aptitudini, cunoștințe de management, calități individuale și evaluarea

Informatii personale

Nume Prenume

Localitatea Tara Nationalitate

Anul nasterii Email

Locul de munca dorit

Profesie Experienta profesionala

Universitate Localitatea

Certificate de competenta Limbi straine cunoscute

Limba materna Limba straina principala

Evaluare calitati

CA1 ☐ CA2 ☐ CA3 ☐ CA4 ☐ CA5 ☐ CA6 ☐ CA7 ☐ CA8 ☐

CA9 ☐ CA10 ☐ CA11 ☐ CA12 ☐ CA13 ☐ CA14 ☐ CA15 ☐ CA16 ☐

CM1 ☐ CM2 ☐ CM3 ☐ CM4 ☐ CM5 ☐ CM6 ☐ CM7 ☐ CM8 ☐

CI1 ☐ CI2 ☐ CI3 ☐ CI4 ☐ CI5 ☐ CI6 ☐ CI7 ☐ CI8 ☐

EP1 ☐ EP2 ☐ EP3 ☐ EP4 ☐ EP5 ☐ EP6 ☐ EP7 ☐ EP8 ☐

CA - Cunoștințe și aptitudini CI - Calități individuale
CM - Cunoștințe de management EP - Evaluare profesională

Incarcare XML

Incarcare evaluare

Salvare in BD

Inapoi

profesională).

Fig IV.28. Modulul Candidat nou

Pentru a putea salva noul candidat în baza de date, este obligatoriu ca în prima fază să se încarce CV-ul acestuia și ulterior rezultatele obținute la evaluări, lucru observabil și în starea butoanelor. În primă fază, doar butonul *Încărcare XML*, care permite încărcarea CV-ului, este

activ, urmând ca apoi să devină activ și butonul *Încărcare evaluare*, și doar în ultima fază butonul *Salvare în BD*.

Încărcarea CV-ului candidatului se face automat prin apăsarea în prima fază a butonului *Încărcare XML*, și selectarea fișierului care urmează să fie încărcat.

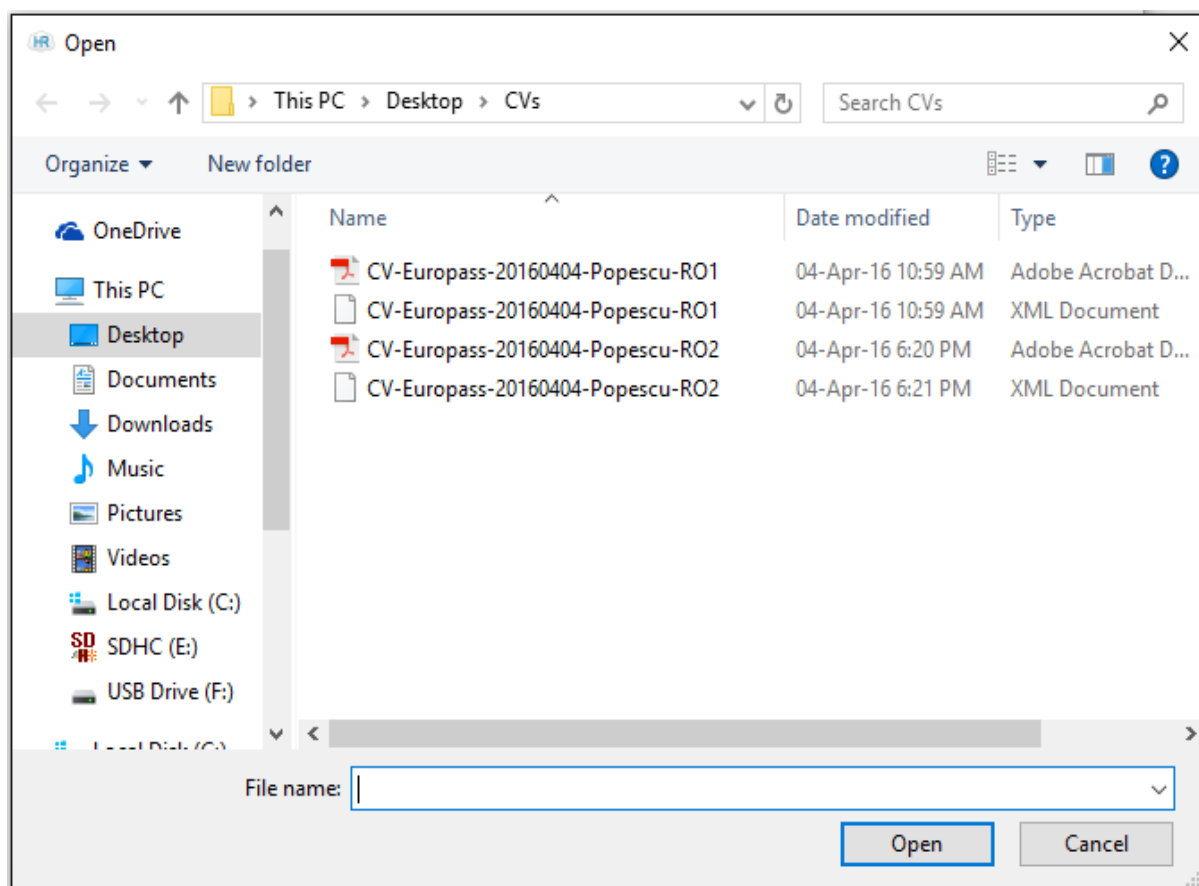


Fig IV.29. Modulul Candidat nou – Încărcare XML

Fișierul selectat pentru încărcare trebuie să fie obligatoriu în format XML și să corespundă normelor de formatare a CV-urilor Europass. Programul returnează un mesaj de eroare, în cazul în care fișierul selectat nu are extensia XML, sau un mesaj de confirmare, în cazul în care fișierul selectat are extensia potrivită. Dacă fișierul este XML, dar nu este formatat corespunzător, există riscul, ca la parcurgerea fișierului să nu se găsească toate informațiile. Din acest motiv, programul permite și introducerea și editarea manuală a datelor personale.

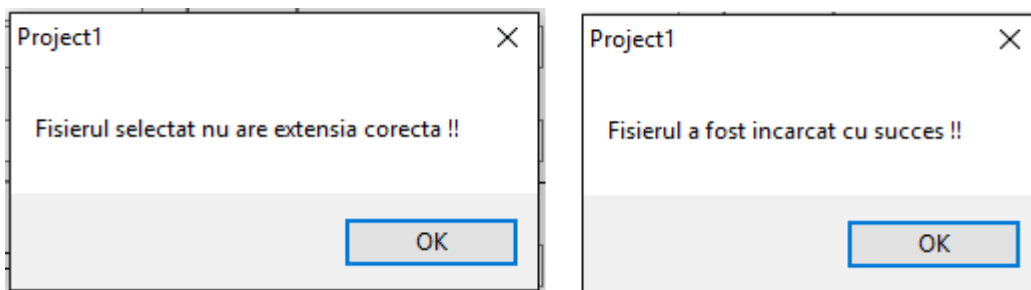


Fig IV.30. Modulul Candidat nou – Mesaje notificare

Pentru a afișa informația din fișierul XML, a fost necesară utilizarea unei funcții care caută un nod în fișierul XML și returnează informația stocată în acesta. Astfel, am dezvoltat o funcție recursivă pe care am apelat-o în mod repetat pentru a încărca toate datele personale necesare, din fișierul XML.

```
function RecursiveFindNode(ANode: IXMLNode; const SearchNodeName:
string): String;
var
    I: Integer;
begin
    if CompareText(ANode.NodeName, SearchNodeName) = 0 then
        begin
            if Assigned(ANode) then result := ANode.Text;
        end
    else if not Assigned(ANode.ChildNodes) then Result := ''
    else begin
        for I := 0 to ANode.ChildNodes.Count - 1 do
            begin
                Result := RecursiveFindNode(ANode.ChildNodes[I], SearchNodeName);
                if Result <> '' then Exit;
            end;
        end;
    end;
end;
```

Funcție căutare nod xml și returnare text din nod

După încărcarea datelor personale ale candidatului, se pot introduce rezultatele obținute de acesta la evaluări. Rezultatele se pot introduce fie manual, fie automat prin apăsarea butonului *Încărcare evaluare*. În cadrul lucrării, nu am mai dezvoltat un modul de testare al candidaților considerând că pentru a sublinia rezultatul cercetării sunt suficiente datele generate aleatoriu. Astfel, prin apăsarea butonului menționat programul generează aleatoriu valori între 0 și 10 pentru toate atributele candidatului. Programul poate fi dezvoltat ulterior prin adăugarea unui modul de testare, care să prelucreze răspunsurile candidatului și să genereze un fișier XML care să conțină valorile atributelor acestuia. Ulterior, aceste rezultate stocate în fișiere XML se pot încărca automat în program pe baza aceluiași metode folosite la încărcarea datelor personale.

Informatii personale

Nume Prenume

Localitatea Tara Nationalitate

Anul nasterii Email

Locul de munca dorit

Profesie Experienta profesionala **DA**

Universitate Localitatea

Certificate de competenta **1** Limbi straine cunoscute **2**

Limba materna Limba straina principala

Evaluare calitati

CA1	8	CA2	0	CA3	7	CA4	3	CA5	0	CA6	8	CA7	0	CA8	4
CA9	9	CA10	2	CA11	10	CA12	7	CA13	1	CA14	1	CA15	2	CA16	5
CM1	1	CM2	4	CM3	7	CM4	0	CM5	3	CM6	2	CM7	9	CM8	7
CI1	5	CI2	7	CI3	9	CI4	0	CI5	8	CI6	4	CI7	7	CI8	10
EP1	7	EP2	9	EP3	3	EP4	8	EP5	0	EP6	1	EP7	10	EP8	8

CA - Cunostinte si aptitudini CI - Calitati individuale
CM - Cunostinte de management EP - Evaluare profesionala

Incarcare XML

Incarcare evaluare

Salvare in BD

Inapoi

Fig. IV.31. Modulul Candidat nou – Informații completate

După completarea tuturor informațiilor (fig. IV.31), utilizatorul are posibilitatea de încărcare a acestora în baza de date, prin apăsarea butonului *Salvare în BD*. În funcție de rezultatul încărcării, utilizatorului îi este afișat un mesaj corespunzător.

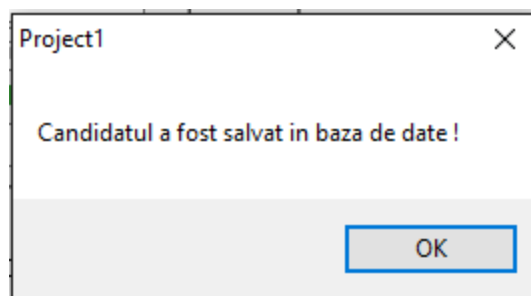


Fig. IV.32. Modul Candidat nou – Mesaj încărcare cu succes a datelor în baza de date

Modulul *Contactează candidat*, prezentat și în imaginea următoare, oferă utilizatorului posibilitatea de a contacta rapid un candidat în diferitele etape ale recrutării.

Transmise			
Anda Mirescu	Invitatie interviu	16-May-16 5:20:10 AM	
Eugen Schiopu	Invitatie interviu	06-May-16 5:22:37 AM	
Alexandru Cretu	Acceptare candidat	06-May-16 5:19:34 AM	
Alexandru Cretu	Invitatie interviu	20-Apr-16 5:20:46 AM	

Fig IV.33. Modul Contactează candidat

Lista de candidati, este populată automat la încărcarea formulii, pe baza următoarelor interogări: `select Nume||'' '||Prenume, email,ID_Candidat from candidati where`

```
candidati.ID_CANDIDAT NOT IN (SELECT ID_CANDIDAT FROM Angajati) order by
Nume,Prenume.
```

Lista de modele conține tipurile de modele de mesaje utilizate cu preponderență de angajații departamentului de resurse umane în dialogul lor cu candidații. Această listă a fost populată static, iar mesajele sunt stocate local în fișiere text. La selectarea unui tip de model, căsuța de text de tip *Memo* se populează cu textul conținut în fișierul corespunzător modelului. După încărcare, textul respectiv, poate fi modificat în funcție de necesități.

În partea de jos a modulului se poate observa o listă cu mesajele transmise ordonate în funcție de data la care au fost trimise. Această listă este populată utilizând următoarea interogare: `Select A.Nume ||' ' ||A.Prenume, B.Model, B.Data From Candidati A, Mesaje B where A.ID_Candidat=B.ID_Candidat order by data desc, nume, prenume asc.`

Dacă mesajul este transmis cu succes, se primește un mesaj care confirmă acest lucru (fig. IV.34) și se inserează o nouă înregistrare în tabela *Mesaje* pe baza următoarei instrucțiuni SQL: `INSERT INTO MESAJE(ID_MESAJ, ID_CANDIDAT, DATA, MODEL) VALUES (Seq_IDMSJ.NextVal, ' + Listbox4.ItemByIndex(ListBox1.ItemIndex).Text + ', SYSDATE, ' +'''+ Listbox2.Selected.Text +'''+ ').`

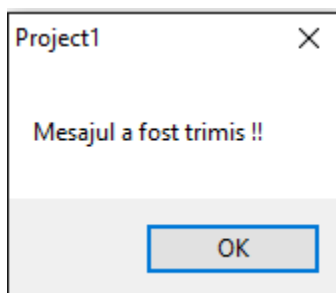


Fig. IV.34. Modul Contactează candida – Confirmare trimitere mesaj

Pentru a putea trimite mesajul, este necesară configurarea unui obiect de timp `TIdSMTP` și a unui mesaj de tip `TIdMessage`. Acestea au fost configurate folosind o adresa de email de test a unui domeniu privat, iar celelalte setări folosite se pot observa mai jos:

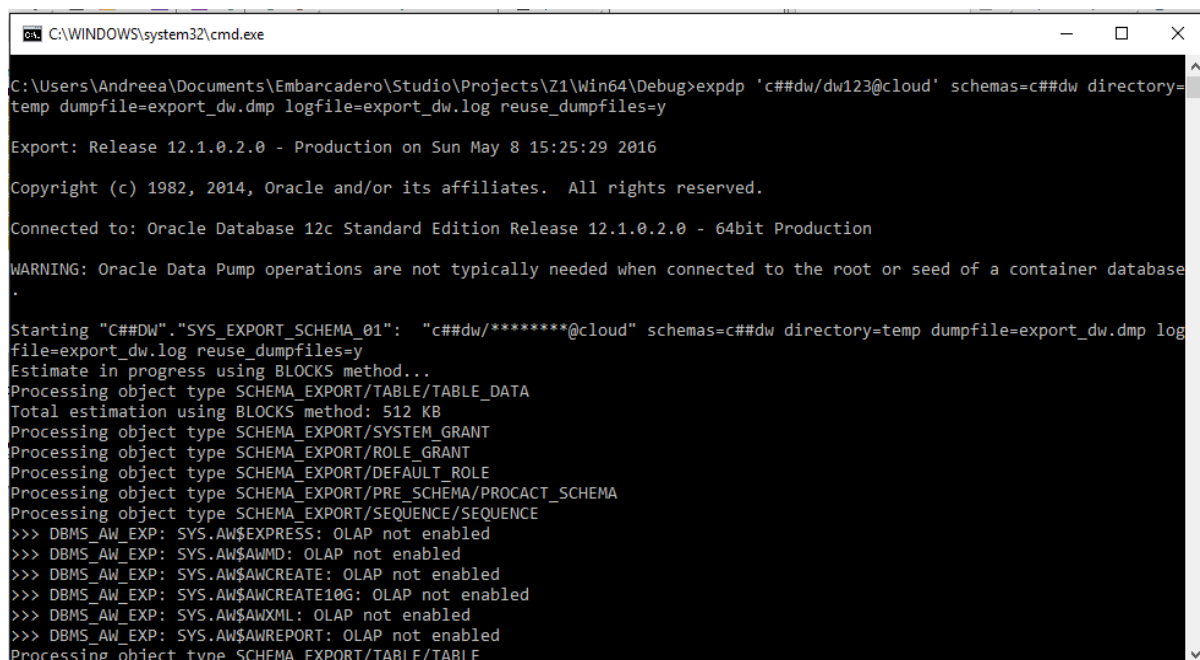
```

Msg := TIdMessage.Create(nil);
try
    Msg.From.Address := 'test@dbc93.ro';
    Msg.Recipients.EmailAddresses := Edit2.Text;
    Msg.Body.Text := Memo1.Text;
    Msg.Subject := 'Contact HR';
    SMTP := TIdSMTP.Create(nil);
    try
        SMTP.Host := 'mail.dbc93.ro';
        SMTP.Port := 26;
        SMTP.AuthType := satDefault;
        SMTP.Username := 'test@dbc93.ro';
        SMTP.Password := '####';
        SMTP.Connect;
        SMTP.Send(Msg);
    finally
        SMTP.Free;
    end;
finally
    Msg.Free;
end

```

Bloc de comenzi utilizat pentru trimiterea unui email

Modulul *Backup BD* realizează un backup al bazei de date, cu *data pump*, așa cum am prezentat la începutul capitolului. Aplicația rulează un fișier *batch* salvat local (fig IV.35.), *backupdb.bat*, utilizând comanda: `ShellExecute(0, nil, 'backupdb.bat', nil, nil, 1)`.



```
C:\WINDOWS\system32\cmd.exe

C:\Users\Andreea\Documents\Embarcadero\Studio\Projects\Z1\Win64\Debug>expdp 'c##dw/dw123@cloud' schemas=c##dw directory=
temp dumpfile=export_dw.dmp logfile=export_dw.log reuse_dumpfiles=y

Export: Release 12.1.0.2.0 - Production on Sun May 8 15:25:29 2016

Copyright (c) 1982, 2014, Oracle and/or its affiliates. All rights reserved.

Connected to: Oracle Database 12c Standard Edition Release 12.1.0.2.0 - 64bit Production

WARNING: Oracle Data Pump operations are not typically needed when connected to the root or seed of a container database
.

Starting "C##DW"."SYS_EXPORT_SCHEMA_01": "c##dw/*****@cloud" schemas=c##dw directory=temp dumpfile=export_dw.dmp log
file=export_dw.log reuse_dumpfiles=y
Estimate in progress using BLOCKS method...
Processing object type SCHEMA_EXPORT/TABLE/TABLE_DATA
Total estimation using BLOCKS method: 512 KB
Processing object type SCHEMA_EXPORT/SYSTEM_GRANT
Processing object type SCHEMA_EXPORT/ROLE_GRANT
Processing object type SCHEMA_EXPORT/DEFAULT_ROLE
Processing object type SCHEMA_EXPORT/PRE_SCHEMA/PROCACT_SCHEMA
Processing object type SCHEMA_EXPORT/SEQUENCE/SEQUENCE
>>> DBMS_AW_EXP: SYS.AW$EXPRESS: OLAP not enabled
>>> DBMS_AW_EXP: SYS.AW$AWMD: OLAP not enabled
>>> DBMS_AW_EXP: SYS.AW$AWCREATE: OLAP not enabled
>>> DBMS_AW_EXP: SYS.AW$AWCREATE10G: OLAP not enabled
>>> DBMS_AW_EXP: SYS.AW$AWXML: OLAP not enabled
>>> DBMS_AW_EXP: SYS.AW$AWREPORT: OLAP not enabled
Processing object type SCHEMA_EXPORT/TABLE/TABLE
```

Fig. IV.35. – Rulare backup

În modulul *Despre...* sunt prezentate pe scurt informații privind lucrarea curentă de cercetare. Astfel, după cum se pot observa și în figura IV.36, în partea superioară numele universității și titlul lucrării de cercetare, central sunt afișate obiectivele lucrării, iar în partea inferioară a modulului este afișat numele profesorului coordonator și cel al studentului doctorand. Modulul nu oferă alte facilități, singurul buton disponibil fiind folosit pentru întoarcerea la pagina principală a aplicației.

Ultima opțiune a meniului, *Ieșire*, oferă utilizatorului posibilitatea de părăsire a aplicației.

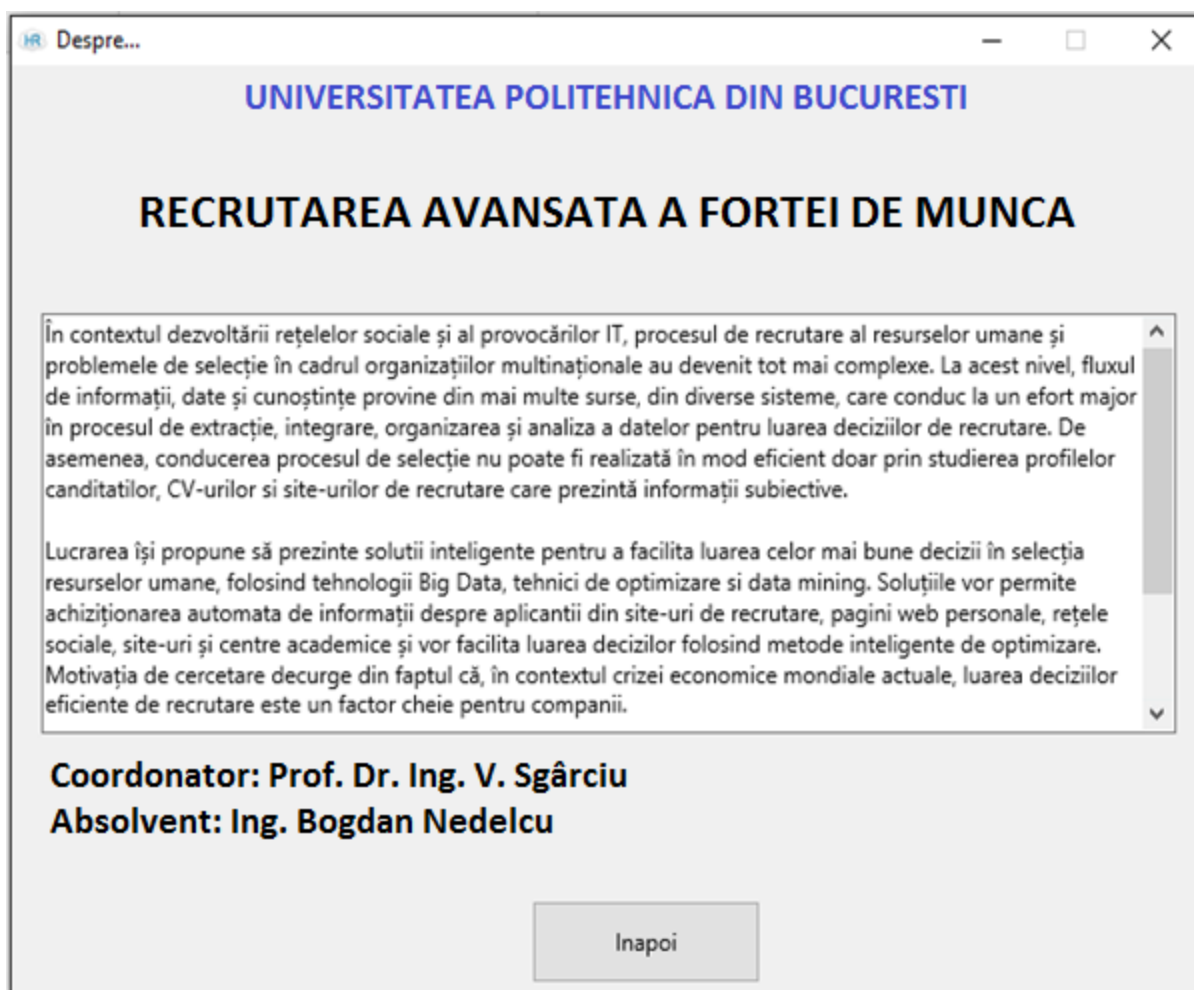


Figura IV.36. Modulul Despre...

IV.2.5. Potențialul impact al prototipului informatic

Implementarea prototipului va avea un impact important și va influența facilitarea accesului la informații relevante susținând deciziile managerilor de recrutarea a personalului, prin minimizarea timpului necesar procesului de selecție prin oferirea rapidă a unor informații cheie despre candidați și îmbunătățirea relevanței informațiilor pe baza cărora managerii iau deciziile de recrutare.

Punerea în funcțiune a unui astfel de sistem oferă un avantaj competitiv în ceea ce privește selecția personalului, ceea ce aduce un plus de valoare companiei.

Acest proiect va avea un impact major și din următoarele puncte de vedere:

- Economic – Dezvoltarea unei platforme online pe arhitectura Cloud Computing poate duce la o activitate mult mai ușoară de recrutare a resurselor umane în cadrul companiei. Prin utilizarea prototipului se facilitează accesul la date, se reduce cantitatea de informație care ajunge la factorii de decizie, minimizând astfel timpul pentru deciziile de recrutare facilitând accesul la informații și profile prin utilizarea graficelor.
- Social - Principalii beneficiari ai prototipului sunt managerii și candidații. Prin implementarea unei platforme scalabile on-line, managerii companiei pot selecta direct candidații și pot crește eficiența procesului de recrutare, astfel încât angajările viitoare să adăuge valoare companiei.
- Mediu - Folosind o arhitectura scalabilă, cum ar fi Cloud Computing, companiile nu vor mai investi în propriul lor hardware, reducând costurile de achiziție, consumul de energie și climatizare a centrului de date, minimizând impactul asupra mediului.

Concluzii generale

Această lucrare analizează impactul tehnologiilor informatice asupra proceselor de afaceri și asupra managementului resurselor umane. Analizele efectuate arată că utilizarea corespunzătoare a infrastructurii IT îmbunătățește managementul talentelor, care, la rândul său, permite executarea strategiilor optime de recrutare, crescând astfel performanța companiei.

Managementul talentelor reprezintă capacitatea dinamică a firmei de a recruta, dezvolta, și păstra talentul pentru a atinge obiectivele propuse și pentru a executa strategii de afaceri. S-a constatat că tehnologiile informatice îmbunătățesc semnificativ managementul talentelor.

În ultimii ani, motoarele NoSQL au devenit tot mai atrăgătoare pentru dezvoltatori, mulți alegând soluții NoSQL precum MongoDB, Cassandra, Redis, sau Hadoop, ca prima lor alegere pentru construirea de aplicații, considerându-le o singură familie de produse, superioară motoarelor SQL vechi.

Cele mai multe sisteme IT nu sunt nevoite să proceseze un volum foarte mare de date. Așa că, principala mea motivație pentru explorarea soluțiilor NoSQL a fost de a determina dacă ar putea exista un motiv pentru a alege una dintre ele atunci când se construiește un sistem regulat, cu un volum rezonabil de date.

Recrutarea are o importanță ridicată influențând direct fiecare parte a organizației, nimic altceva neavând un impact la fel de important asupra capacității organizației de a-și atinge obiectivele de afaceri. Astfel, HR trebuie să se ridice la nivelul unui partener de afaceri strategic prin utilizarea tehnologiei volumelor mari de date și să îmbunătățească strategiile de recrutare a angajaților.

Cu ajutor din partea partenerilor externi adecvați, resursele umane pot demara eforturile de implementare a tehnologiei volumelor mari de date în cadrul organizațiilor, reușind astfel să ia decizii care sunt exponențial mai valoroase, fiind bazate pe dovezi indiscutabile și oferind avantaje strategice.

Asistarea deciziei pentru recrutarea persoanelor în domeniul resurselor umane se bazează foarte mult pe accesul la date fiabile, analiza datelor, precum și vizualizare de informații. Există speranța de a îmbunătăți acuratețea inferenței pe care un om o poate realiza fără ajutor, capacitate uneori menționată ca "extensie cognitivă". Aceasta poate lua diverse forme, de la calcule relativ simple la modele extrem de complexe și complete de simulare la scară largă.

Ca și în cazul altor aplicații de suport decizional, cu cât apropierea de utilizator și de fluxul de lucru este mai mare, și cu cât efortul necesar pentru a invoca și utiliza răspunsul este mai mic, cu atât mai mari sunt așteptările în urma utilizării și beneficiile.

Pentru tot mai multe companii, responsabilitatea angajării revine unui algoritm din ce în ce mai complex. Factorii care sunt luați în considerare diferă tot mai mult astfel încât candidații au ajuns să nu știe la ce să se aștepte. Locuri de muncă, care au fost odată ocupate pe baza experienței precedente și în urma interviurilor, sunt acum ocupate în urma unor teste de personalitate și de analiză a datelor, deoarece angajatorii doresc să fie cât mai siguri că viitorul angajat este cel potrivit pentru post. Sub presiunea de a reduce costurile și a crește productivitatea, angajatorii încearcă să prezică rezultate specifice.

Deși angajarea este o funcție de afaceri importantă, metodele convenționale nu sunt suficient de exigente. În funcție de cine decide, factorii care transformă un candidat într-un angajat pot varia extrem de mult (realizări academice, experiență anterioară, aspectul fizic, personalitatea).

Totuși, utilizarea unor criterii prea specifice de selecție a candidaților poate expune compania riscului de încălcare a dreptului egalității de șanse. Chiar dacă, și neintenționat, criteriile de selecție exclud persoanele mai în vârstă sau minoritățile, acesta este un motiv suficient care poate expune compania unui litigiu.

Managerii care se bazează pe instinctul lor ar putea, uneori, să obțină candidatul dorit, dar bănuielile lor au, în general, o valoare mică în estimarea modului în care candidatul se va descurca la locul de muncă. Companiile care recomandă o abordare statistică pentru efectuarea unei angajări spun că pot îmbunătăți rezultatele prin reducerea influenței distorsiunilor unui manager.

Încrederea în proces și acceptarea recomandărilor de sistem sunt componente cheie ale succesului pe termen lung a suportului decizional pe baza volumelor mari de date generat de rețelele sociale.

Contribuții originale

În primul capitol al lucrării a fost realizată o *analiză comparativă BPMN – UML* în vederea identificării limbajului potrivit pentru realizarea diagramelor studiului de caz. Întrucât majoritatea studiilor comparative realizau o analiză dintr-o singură perspectivă, în cadrul lucrării analiza limbajelor a fost realizată din mai multe perspective precum capacitatea de înțelegere, adecvarea elementelor grafice și capacitatea de mapare.

În ceea ce privește capacitatea de a fi ușor de înțeles, putem spune că atât BPMN, cât și diagramele UML sunt la fel de ușor de înțeles de către părțile implicate în modelarea proceselor de afaceri (analști de afaceri, dezvoltatori tehnici și utilizatorii de afaceri).

Adecvarea elementelor grafice ale BPMN și a diagramelor UML pentru a reprezenta procesele de afaceri reale ale unei organizații a fost analizată în această lucrare din două perspective: capacitatea limbajelor de modelare a proceselor de afaceri pentru a capta modele de flux de lucru și complexitatea simbolurilor grafice utilizate pentru reprezentarea proceselor de afaceri reale ale unei organizații.

Evaluarea BPMN și a diagramelor UML utilizând cadrul modelelor fluxurilor de lucru a relevat faptul că atât a limbajele de modelare a proceselor de afaceri oferă soluții similare pentru cele mai multe dintre modele. Așa cum am arătat în cadrul analizei, rezultatele cercetărilor efectuate de o serie de autori privind capacitatea versiunile anterioare ale BPMN și a diagramelor UML de a capta modele de flux de lucru a arătat că ambele notații oferă suport complet pentru perspectivele cu flux de control și de date, dar oferă un număr limitat de soluții pentru modele de resurse de flux de lucru și modele de manipulare a excepțiilor. Aceste rezultate sunt confirmate și de analiza pe care am efectuat-o în cadrul lucrării cu privire la specificațiile actuale ale BPMN și a diagramelor UML.

Complexitatea simbolurilor grafice utilizate pentru a reprezenta procesele de afaceri reale ale unei organizații este evidențiată în această cercetare în primul rând prin analiza documentelor normative BPMN și UML și în al doilea rând prin intermediul studiului de caz realizat.

Rezultatele analizei indică faptul că, în cele mai multe cazuri, BPMN și diagramele UML utilizează simboluri similare pentru a descrie procesele de afaceri, dar că există cazuri în care componente ale proceselor de afaceri sunt modelate folosind doar un simbol în BPMN, în timp ce în diagramele UML este necesară utilizarea unui grup de simboluri.

În ceea ce privește maparea limbajelor de modelare a proceselor de afaceri în limbaje de execuție a proceselor de afaceri, documentul normativ curent al BPMN include o mapare a unui subset de BPMN la WSBPEL, în timp ce documentul normativ UML nu definește maparea la un BPEL. Soluții pentru maparea între diagramele UML și BPEL au fost descrise într-o serie de cercetări, dar aceste soluții nu oferă o mapare complet automatizată.

În cel de *al doilea capitol* al lucrării a fost realizată evaluarea mai multor soluții NoSQL pentru a determina dacă acestea sunt adecvate implementării propuse pentru studiul de caz. În urma analizei pot spune că acestea aduc, fără îndoială, o mai mare complexitate decât simplitate. În primul rând, este necesară alegerea unui furnizor și, pentru că nu există nici un standard, această alegere ar putea fi determinantă pentru succesul proiectului.

Creșterea volumului de date, la nivel global, va continua, cu siguranță, și va alimenta piața NoSQL. Cu toate acestea, o mulțime vastă de aplicații nu vor trebui să se ocupe de volume gigantice de date și pentru aceștia, vechiul SQL va rămâne cu siguranță prima opțiune.

În afară de scalabilitate, se pare că, într-adevăr, bazele de date NoSQL nu au o altă caracteristică extraordinară. Acest lucru nu mă va împiedica să îmi mențin aprins interesul pentru ele, dar, neavând un volum de date extrem de mare la dispoziție pentru studiul de caz și nici o încărcare semnificativă pentru aplicație, am decis să nu construiesc o aplicație din studiul de caz folosind o soluție NoSQL.

Cu toate acestea, analiza realizată în cadrul lucrării reprezintă o noutate pentru literatura de specialitate întrucât analizează comparativ mai multe tipuri de baze de date NoSQL și nu doar două tipuri cum se găsesc în majoritatea studiilor de specialitate. De asemenea, tratează tipuri și structuri NoSQL diferite pentru a determina structura potrivită pentru un set de date similar celui din studiul de caz.

O contribuție importantă a acestei lucrări, este aceea de a oferi o mai bună înțelegere a conceptelor de procese de afaceri, managementul proceselor de afaceri, analiza proceselor de afaceri, tehnologiile informatice utilizate în analiza proceselor de afaceri, și de a prezenta un exemplu concret de aplicare a tehnologiilor informatice în domeniul resurselor umane.

De asemenea, analizele realizate, arată că tehnologiile informatice cresc performanța firmei printr-un management inteligent al talentelor.

Nu în ultimul rând, lucrarea analizează, în cel de *al treilea capitol*, domeniul resurselor umane și avansează ideea de a introduce într-un prototip inteligent, analize pe baza datelor de căutare ale motoarelor. Experimentul realizat contribuie la susținerea teoriei conform căreia există o corelație între datele de căutare ale motoarelor și interesul manifestat de populație către anumite funcții (job-uri).

Cel mai mare impediment al acestui studiu l-a reprezentat lipsa volumelor de date. Astfel, pentru efectuarea experimentelor a fost necesară construirea seturilor de date. În ciuda efortului depus pentru realizarea unor seturi de date adecvate, diferențele dintre seturile create și un set autentic pot fi semnificative. În viitor, aceste experimente pot fi reluate pe baza unor seturi reale de date, iar rezultatele pot fi comparate cu cele obținute în lucrare.

În ceea ce privește infrastructura resurselor umane, aceasta este în mod clar, o componentă critică. Există prea multe sisteme de resurse umane și de salarizare care împiedică aplicarea eficientă a tranzacțiilor de date a angajaților. Abilitatea de a manipula aceste date în avantajul organizației, și de a le și înțelege, este și mai complicată - presupunând că există resurse pentru a crea legătura dintre informații după curățarea provocărilor ridicate de tranzacțiile în curs de desfășurare. În plus, pentru o companie multinațională, achiziția unei alte companii, o fuziune, sau intrarea pe o nouă piață, introduce, de obicei, un nou sistem divergent.

O altă contribuție importantă a lucrării, evidențiată în cel de *al patrulea capitol* al lucrării, o reprezintă analiza posibilității de utilizare a datelor din motoarele de căutare în domeniul resurselor umane. Astfel, a fost evaluată acuratețea datelor oferite de Google Trends, stabilindu-se că acestea sunt relevante și pot influența pozitiv luare deciziilor.

În urma acestei analize am considerat că extragerea de cunoștințe din datele motoarelor de căutare ar aduce un mare aport și în domeniul resurselor umane. Pentru o companie multinațională, această facilitate se poate dovedi extrem de utilă pentru alegerea punctelor în care

se vor realiza investiții ulterioare. Astfel, se poate ține cont de interesul local pe piața locurilor de muncă și se pot extinde anumite departamente, sau se poate lua decizia creării unui departament într-o altă locație.

Pe baza acestor informații, o companie își poate orienta intenția de recrutare în zona respectivă, având astfel șanse mai mari de a găsi rapid un candidat potrivit. De asemenea, analizând nivelul ridicat de interes, o companie poate dezvolta o politică de recrutare a unor persoane care nu au cunoștințe solide în domeniul respectiv de activitate. Astfel, compania poate opta fie pentru un program de internship, fie pe o recrutare cu o perioadă mai lungă de training.

În funcție de domeniul de activitate al companiei, sau de profilul căutat al unui candidat se pot stabili termeni de interes pe baza cărora să se efectueze căutările în datele de căutare ale motorului Google.

Contribuția esențială a acestei lucrări, evidențiată tot în cel de *al patrulea capitol* al lucrării, este reprezentată de evidențierea potențialului clasificatorului de exploatare a datelor pentru datele de resurse umane. Clasificatorul propus poate fi utilizat pentru a genera modele de clasificare a performanței, utilizând bazele de date care conțin datele de performanță ale angajaților. Ulterior, modelele de clasificare generate pot fi folosite ca instrument de suport decizional pentru predicția talentului uman.

În cadrul acestei secțiuni a lucrării am evidențiat importanța studiului folosind extragerea cunoștințelor din datele pentru managementul talentelor, în special pentru clasificare și predicție. Cu toate acestea, la finalul analizei pot afirma că ar trebui să existe mai multe tehnici de exploatare a datelor aplicate diferitelor domenii problematice în domeniul cercetării resurselor umane, în scopul de a lărgi orizontul de muncă academică și practică privind extragerea cunoștințelor din date în HR.

În plus, algoritmul clasificator C4.5 este potențialul clasificator în acest experiment. Astfel, această tehnică poate fi folosită pentru date reale, similare, în următoarea fază de predicție. Aceste reguli de clasificare generate pot fi folosite pentru a prezice talentul potențial pentru o sarcină specifică într-o organizație. În managementul resurselor umane, există mai multe sarcini care pot fi rezolvate folosind această abordare, ca de exemplu: selectare angajați noi, potrivirea persoanelor pentru locuri de muncă, trasee de planificare a carierei, planificarea nevoilor de formare pentru angajați noi și seniori, predicția performanței angajaților.

În concluzie, capacitatea de a schimba în mod continuu și de a obține o nouă viziune cu privire la clasificarea și predicția în domeniul resurselor umane, poate, astfel, să devină cea mai mare contribuție la extragerea cunoștințelor din date de HR.

Concluzii finale

Se așteaptă ca resursele umane (HR) să susțină creșterea economică și extinderea la nivel global. În companiile mari, resursele umane trebuie să construiască și să mențină agilitatea de afaceri. În toate acestea, pătrunderea sistemelor analitice HR va fi un element cheie pentru succes.

Este incontestabil că s-a intrat într-o eră a volumelor mari de date. Prin îmbunătățirea analizei volumelor mari de date care devin disponibile, există potențialul de a face progrese mai rapide în mai multe discipline științifice și de a crește rentabilitatea și succesul multor întreprinderi. Cu toate acestea, multe provocări tehnice descrise în această lucrare trebuie să fie aprofundate înainte ca acest potențial să poată fi realizat pe deplin.

Provocările includ nu doar aspectele evidente de dimensiune, dar și eterogenitatea, lipsa de structură, de tratare a erorilor, confidențialitatea, actualitatea, proveniența și vizualizarea, în toate etapele procesului de analiză, de la achiziția datelor și până la interpretarea rezultatelor. Aceste provocări tehnice sunt comune într-o mare varietate de domenii de aplicare, și, prin urmare, nu este rentabil să se abordeze doar în cadrul unui singur domeniu.

Mai mult decât atât, aceste provocări vor necesita soluții de transformare, și nu vor fi abordate în mod natural de către următoarea generație a produselor industriale. Trebuie să se sprijine și să se încurajeze cercetarea fundamentală față de abordarea acestor provocări tehnice, dacă se dorește obținerea beneficiilor promise de tehnologia volumelor mari de date.

Diseminarea rezultatelor

➤ Carti:

- Ioan RUSU, Vlad Al. GROSU, Bogdan NEDELCU, Georgiana MUSCALU (2016) – Baze de date: Indrumar de laborator: mediul Oracle Ed. II, Bucuresti, Editura Matrix Rom, ISBN 978-606-25-0230-0
- Vlad Al. GROSU, Bogdan NEDELCU, Andreea JIGAU (2016) – Baze de date: Indrumar de laborator: mediul Oracle Ed. III, Bucuresti, Editura Matrix Rom

➤ Articole:

- Bogdan NEDELCU, Madalina-Elenea STEFANET, Ioan-Florentin TAMASESCU, Smaranda-Elena TINTOIU, Alin VEZEANU (2015) - CLOUD COMPUTING AND ITS CHALLENGES AND BENEFITS IN THE BANK SYSTEM, Database Systems Journal, vol VI, nr. 1, pg 44-58, ISSN 2069-3230, <http://www.dbjournal.ro>
- Bogdan NEDELCU, Andreea Maria IONESCU, Ana Maria IONESCU, Alexandru George VASILE (2015) - RESHAPING SMART BUSINESSES WITH CLOUD DATABASE SOLUTIONS, Database Systems Journal, vol. V, nr. 4, pg. 21-38, ISSN 2069-3230 <http://www.dbjournal.ro>
- Nedelcu Bogdan (2014) - BUSINESS INTELLIGENCE SYSTEMS, Database Systems Journal, vol. IV, nr. 4, pg. 12-20, ISSN 2069-3230, <http://www.dbjournal.ro/14.html>
- Nedelcu Bogdan (2014) - ABOUT BIG DATA AND ITS CHALLENGES AND BENEFITS IN MANUFACTURING, Database System Journal, vol. IV, nr. 3, pg. 10-19, ISSN 2069-3230, <http://www.dbjournal.ro/13.html>

➤ Conferinte:

- Adela BĂRA, Iuliana ȘIMONCA (BOTH), Anda BELCIU, Bogdan NEDELCU (2015) - BIG DATA CHALLENGES FOR HUMAN RESOURCES MANAGEMENT, The 14th International Conference on Informatics in Economy, Education, Research & Business Technologies, IE 2015, 4/30/2015, Bucuresti, România, Proceedings of the IE 2015 International Conference, pg. 1-6, WOS:000362796900059, ISSN: 2284-7472, <http://www.conferenceie.ase.ro>
- Bogdan NEDELCU – Human Talent Forecasting, The 11th International Conference on Business Excellence Strategies, Complexity and Energy in

Changing Times, ICBE 2017, Thomson Reuters ISI Web of Science (WOS) Conference Proceedings Citation Index 2017, WOS:000431004400047, ISSN: 2502-0226, eISSN: 2558-9652, <http://www.bizexcellence.ro/icbe/11th-edition>

- Bogdan NEDELCU – Cybersecurity awareness in NBR, BSMCySec 2018 Edition, Constanta Maritime University, Blacksea Maritime Cybersecurity Conference 7-8 June 2018
- Bogdan NEDELCU, Andreea JIGAU, Valentin Sgarciu – Mining Medical Data, ECAI 2018 International Conference 10th Edition, Electronics, Computers and Artificial Intelligence, 28 June-30 June 2018

Bibliografie

- [1] [SCH10] Schedlbauer, M. (2010) The Art of Business Process Modeling: The Business Analyst's Guide to Process Modeling with UML & BPMN, CreateSpace
- [2] [PYKE] J. Pyke, J. O'Connell și R. Whitehead - Mastering Your Organization's Processes: A Plan Guide to BPM, 2006
- [3] [WUR13] Wurtzel M.M. - What is BPM, McGraw-Hill Engineering, 2013
- [4] [MKER] Marc Kerremans, Nicholas Kitson - Aligning Business Process Management and Business Intelligence to Achieve Business Process Excellence
- [5] [RUMM] Geary Rummler and Alan Brache - Improving Performance: How to Manage the White Space on the Organization Chart, 1990, San Francisco: Jossey-Bass
- [6] [JIA11] Jianu, I., Jianu, I. & Gușatu, I. (2011) „Net income versus comprehensive income for professional investors”, Proceedings of the sixth edition of the International Conference Accounting and Management Information Systems: 966-987
- [7] [TAR11] Țarțavulea, R.I., Belu, M.G. & Dieaconescu, V.C. (2011) “Spatial modeling in logistics decision-making processes. Identifying the optimal location for a single central warehouse”, Annals of the University of Oradea, Economic Science Series, Tom XX, vol. 1: 137-143
- [8] [ERIK] Eriksson, H. & Penker, M. (2000) Business Modeling with UML: business patterns at work, John Wiley & Sons
- [9] [KAVI] Kalnins, A. & Vitolins, V. (2006) “Use of UML and Model Transformations for Workflow Process Definitions”, Databases and Information Systems IV - Selected Papers from the Seventh International Baltic Conference, DB&IS 2006: 3-15

- [10] [MAZ] Mazanek, S. & Hanus, M. (2011) “Constructing a bidirectional transformation between BPMN and BPEL with a functional logic programming language”, *Journal of Visual Languages & Computing*, vol. 22, no. 1: 66–89
- [11] [ZHA] Zhang, M. & Duan, Z. (2008) “From Business Process Models to Web Service Orchestration: The case of UML 2.0 Activity Diagram to BPEL”, *Lecturer Notes in Computer Science*, vol. 5364: 505-510
- [12] [OMGa] Object Management Group – Document Associated with BPMN version 2.0.2
- [13] [OMGb] Object Management Group – Document Associated with Unified Modeling Language (UML) version 2.5
- [14] [OMGa] Object Management Group – Document Associated with BPMN version 2.0.2
- [15] [PEX] Peixoto, D.C.C., Batista, V.A., Atayde, A.P., Borges, E.P., Resende, R. F. & Pádua, C.I. (2008) “A Comparison of BPMN and UML 2.0 Activity Diagrams”, *VII Simpósio Brasileiro de Qualidade de Software*: 1-12
- [16] [BIRK] Birkmeier, D., Klöckner, S. & Overhage, S. (2010) “An empirical comparison of the usability of BPMN and UML Activity Diagrams for business users”, *18th European Conference on Information Systems*: 1-12
- [17] [AAL] van der Aalst, W.M.P, ter Hofstede, A.H.M., Kiepuszewski, B. & Barros, A.P. (2003) “Workflow Patterns”, *Distributed and Parallel Databases*, vol. 14, no. 3: 5-51
- [18] [DUMH] Dumas, M. & ter Hofstede, A. (2001) “UML activity diagrams as a workflow specification language”, *Proceedings of the Fourth International Conference on the Unified Modeling Language (UML 2001)*: 76–90
- [19] [RUSS] Russell, N., van der Aalst, W.M.P, ter Hofstede, A.H.M. & Wohed, P. (2006c) “On the Suitability of UML 2.0 Activity Diagrams for Business Process Modelling”, *Proceedings of the 3rd Asia-Pacific conference on Conceptual modelling*, vol. 53: 95-104

- [20] [OASIS] Online Community for the Web Services Business Process Execution Language OASIS Standard - <http://bpel.xml.org/>
- [21] [OMGa] Object Management Group – Document Associated with BPMN version 2.0.2
- [22] [OMGb] Object Management Group – Document Associated with Unified Modeling Language (UML) version 2.5
- [23] [ZHA] Zhang, M. & Duan, Z. (2008) “From Business Process Models to Web Service Orchestration: The case of UML 2.0 Activity Diagram to BPEL”, Lecturer Notes in Computer Science, vol. 5364: 505-510
- [24] [HLB] Hlaoui, Y.B. & Benayed, L.J. (2011) “A Model Transformation Approach Based on Homomorphic Mappings between UML Activity Diagrams and BPEL4WS Specifications of Grid Service Workflows”, Computer Software and Applications Conference Workshops (COMPSACW) - 2011 IEEE 35th Annual: 243-248
- [25] [MCK11] McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity – June 2011
- [26] [PAN13] Andrei Pandre – Datawatch has 3 Vs - 19 Noiembrie 2013
- [27] [WIKI] Wikipedia – The Free Encyclopedia - http://en.wikipedia.org/wiki/Big_data
- [28] [GART] Gartner, Inc. – <http://www.gartner.com/it-glossary/big-data>
- [29] [PCMG] PC Mag - <http://www.pcmag.com/encyclopedia/term/62849/big-data>
- [30] [MCK11] McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity – June 2011
- [31] [MCK11] McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity – June 2011
- [32] [NASS] Nasscom – Crisil GR&A Analysis

- [33] [NASS] Nasscom – Crisil GR&A Analysis
- [34] [PCAST10] Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology. PCAST Report, Dec. 2010
- [35] [ECO11] Drowning in numbers -- Digital data will flood the planet—and help us understand it better. *The Economist*, Nov 18, 2011
- [36] [GON08] Understanding individual human mobility patterns. Marta C. González, César A. Hidalgo, and Albert-László Barabási. *Nature* 453, 779-782 - 5 June 2008
- [37] [RIJ14] Mark van Rijmenam - “How the Next Generation of Databases Could Solve Your Problems”, Datafloq, 18 Octombrie 2014 <https://datafloq.com/read/generation-database-solve-problems/139>
- [38] [PGLB] 3 Pillar Global, Exploring the Different Types of NoSQL Databases Part II <http://www.3pillarglobal.com/insights/exploring-the-different-types-of-nosql-databases>
- [39] [GDC] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, Werner Vogels - Dynamo: Amazon’s Highly Available Key-value Store, Amazon.com, 2007
- [40] [RDD] Basho, Riak Distributed Database <http://basho.com/riak/>
- [41] [IHB] IBM, Hadoop <http://www-01.ibm.com/software/data/infosphere/hadoop/hbase/>
- [42] [CDH] Cloudera, CDH <http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/hbase.html>
- [43] [RED] Eric Redmond, Jim R. Wilson - Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement, 2012
- [44] [MDB] MongoDB, Introduction to MongoDB <http://www.mongodb.org/about/introduction>

- [45] [RED] Eric Redmond, Jim R. Wilson - Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement, 2012
- [46] [RED] Eric Redmond, Jim R. Wilson - Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement, 2012
- [47] [BIGUS] Bigus, J.P. (1996). *Data Mining with neural networks: solving business problems from application development to decision support*. New York: McGraw-Hill
- [48] [FAY] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth – From Data Mining to Knowledge Discovery in Databases
- [49] [HMS] Hand, D.J., Mannila, H. & Smith, P. (2001). *Principles of Data Mining*. London: The MIT Press
- [50] [WSMI] Weiss, S.M. & Indurkha, N. (1998). *Predictive Data Mining. A Practical Guide*. San Francisco, CA: Morgan Kauffman
- [51] [HAND] Hand, D.J. (2007). Principles of Data Mining. *Drug Safety*, 30, 7, 621-622
- [52] [BELI] Berry, M. & Linoff, G. (2004). *Data Mining Techniques. For marketing, sales, and customer relationship management* (2nd ed.). Indianapolis: Wiley
- [53] [CHHS] Chang, H. C. & Hsu, C.C. (2005). Using Topic Keyword Clusters for automatic Document Clustering. *Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)*, Kota Kinabalu, Sabah
- [54] [HJKM] Han, J. & Kamber, M. (2001). *Data mining: concepts and techniques* (Morgan-Kaufman Series of Data Management Systems), Academic Press, San Diego
- [55] [SHM] Shmueli, G., Patel, N.R. & Bruce, P.C. (2007). *Data Mining for Business Intelligence. Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. New Jersey: John Wiley & Sons

- [56] [BIGUS] Bigus, J.P. (1996). *Data Mining with neural networks: solving business problems from application development to decision support*. New York: McGraw-Hill
- [57] [MOPA] Montaña, J.J. & Palmer, A. (2003). Numeric sensitivity analysis applied to feedforward neural networks. *Neural Computing and Applications*, 12, 2, 119-125
- [58] [GERV] Gervilla, E., Jiménez, R., Montaña, J.J., Sesé, A., Cajal, B. & Palmer, A. (2009). The methodology of *Data Mining*. An application to alcohol consumption in teenagers. *Adicciones*, 21, 1, 65-80
- [59] [BRCU] Breiman, L. & Cutler, A. (2004). *Random Forests*. Retrieved from http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm
- [60] [BRE] Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 1, 5-32
- [61] [AN] An, A. (2006). Classification Methods. In J. Wang (Ed.), *Encyclopedia of Data Warehousing and Mining* (pp. 144-149). Hershey, PA: Idea Group Inc
- [62] [GIUD] Giudici, P. (2003). *Applied Data Mining. Statistical Methods for Business and Industry*. England: John Wiley & Sons
- [63] [HAND] Hand, D.J. (2007). Principles of Data Mining. *Drug Safety*, 30, 7, 621-622
- [64] [MICH] Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. New York: Ellis Horwood Ltd
- [65] [DOPA] Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29, 103-130
- [66] [SHM] Shmueli, G., Patel, N.R. & Bruce, P.C. (2007). *Data Mining for Business Intelligence. Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. New Jersey: John Wiley & Sons
- [67] [YEN] Ye, N. (Ed.) (2003). *The handbook of Data Mining*. New Jersey: Lawrence Erlbaum Associates

- [68] [KDN] Gregory Piatetsky, CRISP-DM, still the top methodology for analytics, data mining, or data science projects, Kdnuggets
- [69] [OLAP] OLAP Council. OLAP AND OLAP Server
- [70] [CONN] CONNOLLY, T., BEGG, C. și STRACHAN, A. – „Baze de date. Proiectare. Implementare. Gestionare”, Editura Teora, București, 2001, ISBN 973-20-0601-3
- [71] [SK15] Sheilendra Kadre - How to Choose Your Data Analysis Tools, 22 Aprilie 2015
- [72] [MKER] Marc Kerremans, Nicholas Kitson - Aligning Business Process Management and Business Intelligence to Achieve Business Process Excellence
- [73] [MKER] Marc Kerremans, Nicholas Kitson - Aligning Business Process Management and Business Intelligence to Achieve Business Process Excellence
- [74] [DOMP] Domingos P, *A few useful things to know about machine learning*, Commun ACM 55(10), 2012
- [75] [DNTB] Dalal N, Triggs B, *Histograms of oriented gradients for human detection*. In: *Computer Vision and Pattern Recognition*, 2005. CVPR 2005. IEEE Computer Society Conference On. IEEE Vol. 1. pp 886–893
- [76] [LWDG] Lowe DG, *Object recognition from local scale-invariant features*. In: *Computer Vision*, 1999. The Proceedings of the Seventh IEEE International Conference On. IEEE Computer Society Vol. 2. pp 1150–1157
- [77] [ARK] Arel I, Rose DC, Karnowski TP, *Deep machine learning-a new frontier in artificial intelligence research*, Research Frontier, 2010, IEEE Comput Intell 5:13–18
- [78] [LBLL] Larochelle H, Bengio Y, Louradour J, Lamblin P (2009) Exploring strategies for training deep neural networks. J Mach Learn Res 10:1–40
- [79] [NRC] National Research Council, *Frontiers in Massive Data Analysis*. The National Academies Press, Washington, DC, 2013

- [80] [DAVE13] Davenport, T. H. (2013). Analytics 3.0. Harvard Business Review, 91(12), 64-72
- [81] [GART] Gartner, Inc. – <http://www.gartner.com/it-glossary/big-data>
- [82] [WALK12] Joseph Wlaker – Meet the New Boss: Big Data – The Wall Street Journal, 20 Sept. 2012
- [83] [MCK11] McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity – June 2011
- [84] [SOND13] Paul Sonderegger – Forbes – Pokerbots Bet On Big Data Strategy, 11/21/2013
- [85] [NERM14] C. Nermey - How HR analytics can transform the workplace
- [86] [SADA13] L. Sadath - Data Mining: A Tool for Knowledge Management in Human Resource, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-2, Issue-6, April 2013
- [87] [GYOR10] C.Györödi, R.Györödi, G.Pecherle, G. M. Cornea - Full-Text Search Engine Using MySQL, Journal of Computers, Communications & Control (IJCCC), Vol. 5, Issue 5, December 2010, pag. 731-740
- [88] [KAO07] A.Kao, S. Poteet - Natural Language Processing and Text Mining, Springer-Verlag London Limited 2007, ISBN 1-84628-175-X
- [89] [BRO13] Broniatowski, D.A. et al. (2013) National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. PLoS ONE 8, e83672
- [90] [MUR13] Murdoch, T.B. and Detsky, A.S. (2013) The inevitable application of big data to health care. JAMA 309, 1351–1352
- [91] [TWI13] Twitter, Inc. – United States Securities and Exchange Commission, October 3, 2013

- [92] [DON12] Donna Tam, Facebook processes more than 500 TB of data daily, CNET, August 22, 2012
- [93] [MYS13] Mysli'n, M. et al. (2013) Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. *J. Med. Internet Res.* 15, e174
- [94] [JAV14] Javid Muhammedali – How Big Data is Impacting the Job-Hunting and Hiring Experience Today – Official Monster Blog, 30 Sept. 2014
- [95] [WALK12] Joseph Wlaker – Meet the New Boss: Big Data – The Wall Street Journal, 20 Sept. 2012
- [96] [WALK12] Joseph Wlaker – Meet the New Boss: Big Data – The Wall Street Journal, 20 Sept. 2012
- [97] [HAND] Hand, D.J. (2007). Principles of Data Mining. *Drug Safety*, 30, 7, 621-622
- [98] [WIKIa] Wikipedia – Google Trends - https://en.wikipedia.org/wiki/Google_Trends
- [99] [WIKIb] Wikipedia – Atentatul impotriva revistei Charlie Hebdo
- [100] [WIKIc] Wikipedia – November 2015 Paris attacks
- [101] [HNWS] Hot News – India, magnetul marilor companii care investesc in cercetare si dezvoltare
- [102] [CHN] Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems and Applications*, 34(1), 380-290
- [103] [HJKM] Han, J. & Kamber, M. (2001). Data mining: concepts and techniques (Morgan-Kaufman Series of Data Management Systems), Academic Press, San Diego
- [104] [TSO] Tso, G. K. F., & Yau, K. K. W. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32, 1761-1768
- [105] [RANJ] Ranjan, J. (2008). Data Mining Techniques for better decisions in Human Resource Management Systems. *International Journal of Business Information Systems*, 3(5), 464-481
- [106] [CHN] Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems and Applications*, 34(1), 380-290

- [107] [TAHS] Tai, W. S., & Hsu, C. C. (2005). A Realistic Personnel Selection Tool Based on Fuzzy Data Mining Method
- [108] [CCWL] Chen, K. K., Chen, M. Y., Wu, H. J., & Lee, Y. L. (2007). *Constructing a Web-based Employee Training Expert System with Data Mining Approach*. Paper presented at the Paper in The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and Eservices (CEC-EEE 2007)
- [109] [HTL] Huang, M. J., Tsou, Y. L., & Lee, S. C. (2006). Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge. *Knowledge-Based Systems*, 19(6), 396-403
- [110] [THCS] Tung, K. Y., Huang, I. C., Chen, S. L., & Shih, C. T. (2005). Mining the Generation Xer's job attitudes by artificial neural network and decision tree - empirical evidence in Taiwan. *Expert Systems and Applications*, 29(4), 783-79
- [111] [ZHAO] Zhao, X. (2008). *An Empirical Study of Data Mining in Performance Evaluation of HRM*. Paper presented at the International Symposium on Intelligent Information Technology Application Workshops, Hangzhou, China
- [112] [LYN] Lynne, M. (2005). *Talent Management Value Imperatives: Strategies for Execution*: The Conference Board
- [113] [CNGM] Cubbingham, I. (2007). Talent Management: Making it real. *Development and Learning in Organizations*, 21(2), 4-6
- [114] [TPTR] TP Track Research Report - *TalentManagement: A State of the Art*: Tower PerrinHR Services, 2005
- [115] [CHNU] CHINA UPDATE. (2007). HR News for Your Organization: The Tower Perrin Asia Talent Management Study. Retrieved from www.towersperrin.com. 7/1/2008
- [116] [RANJ] Ranjan, J. (2008). Data Mining Techniques for better decisions in Human Resource Management Systems. *International Journal of Business Information Systems*, 3(5), 464-481
- [117] [MCKa] Elizabeth G. Chambers, Mark Fouldon, Helen Handfield-Jones, Steven M. Hankin, Edward G. Michaels III 1998, *The War for Talent*,
- [118] [BLM] Bloomberg Businessweek, September 13 – September 19, 2010 issue, 54
- [119] [BLM] Bloomberg Businessweek, September 13 – September 19, 2010 issue, 54
- [120] [ADP15] ADP Research Institute: *Harnessing Big Data: The Human Capital Management Journey to Achieving Business Growth – ADP Global Human Capital Management Decision Makers Survey*, 2015

- [121] [ADP15] ADP Research Institute: Harnessing Big Data: The Human Capital Management Journey to Achieving Business Growth – ADP Global Human Capital Management Decision Makers Survey, 2015
- [122] [KRUS15] JoAnne Kruse – With Big Data, HR Departments Too Often Get Short Shrift – The Wall Street Journal, 22 Feb. 2015
- [123] [OREI11] O'Reilly Media - Big Data Now, September 2011, ISBN: 978-1-449-31518-4
- [124] [RSJV12] Rabl, Sadoghi, Jacobsen, Villamor, Mulero, Mankovskii - Solving Big Data Challenges for Enterprise Application Performance Management, 2012-08-27, VLDB, Vol. 5, ISSN 2150-8097
- [125] [WIN10] Embarcadero - *Windows 10: The Big New Opportunity for Developers*, July 2016
- [126] [PWIKI] Putty, Wikipedia - <https://en.wikipedia.org/wiki/PuTTY>
- [127] [PUTTY] Putty: A Free SSH and Telnet Client
- [128] [WSCPW] WinSCP, Wikipedia - <https://en.wikipedia.org/wiki/WinSCP>
- [129] [GIMP] GIMP – *About GIMP*, <https://www.gimp.org/about/>
- [130] [WEKA] The University of Waikato - Weka, <http://www.cs.waikato.ac.nz/ml/weka/>