Utilizarea tehnicilor de analiză a datelor în procesul de identificare a evaziunii fiscale

Dragoș-Cătălin BARBU

Academia de Studii Economice București, Școala Doctorală de Informatică Economică

bcatalin_ro@yahoo.com

Rezumat: La nivel macroeconomic puterea de analiză a datelor și a informațiilor digitale de natură fiscală joacă un rol esențial în implementarea strategiilor economice. Volumul mare de date zilnice rezultate în timp real la nivel național prin transmiterea bonurilor fiscale de la operatorii economici către Agenția Națională de Administrație Fiscală (ANAF) cu ajutorul aparatelor de marcat electronice fiscale (AMEF) facilitează posibilitatea unei analize predictive ce permite recunoașterea tiparelor și semnalarea celor atipice în vederea identificării evaziunii fiscale. Identificarea parametrilor relevanți din bonurile fiscale, colectarea datelor și integrarea lor în platformă, prelucrarea datelor fiscale cu ajutorul unor instrumente și tehnici specifice Big Data, analiza informațiilor folosind un model de extragere a datelor, precum și organizarea și structurarea acestor date prin diverse tehnici de agregare și îmbunătățire a datelor, vor conduce la validarea rezultatelor pe baza clusterelor și interpretarea lor. Avantajul analizei datelor îl reprezintă faptul că eliberează mai mult timp și energie pentru alte sarcini cum ar fi cele creative și semnificative, care folosesc interpretarea modelelor de date în procesul decizional strategic. Un alt aspect relevant îl reprezintă modul de implementare a principiilor teoretice în zona de business și cum sunt folosite politicile de funcționare a caselor de marcat pentru combaterea infracționalității și a fraudelor fiscale.

Cuvinte cheie: Tehnici Big Data, Analytics, Machine Learning, Data Mining, Cloud Computing, Data Aggregation, Data Enrichment, date digitale fiscale, aparate de marcat electronice fiscale.

Use of data analysis techniques in the process of identifying tax evasion

Abstract: Macroeconomically speaking, the power to analyze digital data and digital information of a fiscal nature, plays a key role in implementing economic strategies. The large volume of daily data resulting in real time, at a national level by sending tax receipts from economic operators to the National Agency for Fiscal Administration (ANAF) using electronic fiscal cash registers (AMEF), facilitates the possibility of a predictive analysis that allows pattern recognition and signaling atypical ones in order to identify tax evasion. Identifying the relevant parameters in tax receipts, collecting data and integrating it into the platform, processing tax data using specific Big Data tools and techniques, analyzing information using a data extraction model, organizing and structuring this data through various aggregation techniques and improved data will lead to the validation of results based on clusters and their interpretation. The advantage of data analysis is that it frees up more time and energy for other tasks, such as creative and meaningful ones, that use the interpretation of data models in the process of strategic decision making. Another relevant aspect is the aspect of implementing the theoretical principles in the business area and how the cash registers operating policies are used to fight crime and tax fraud.

Keywords: Big Data Technics, Analytics, Machine Learning, Data Mining, Cloud Computing, Data Aggregation, Data Enrichment, Fiscal Digital Data, Electronic Fiscal Devices.

1. Big Data și analiza datelor fiscale

Revoluția digitală a determinat avansarea și evoluția tehnologiei de astăzi plecând de la dispozitivele electronice și mecanice analogice la forme ale tehnologiei precum inteligența artificială și Machine Learning, dar a dus și la disfuncționalități în modul în care se iau deciziile de afaceri în fiecare industrie, fie că este vorba de asistență medicală, finanțe, asigurări, educație, divertisment, retail sau energie.

Big Data Analytics și Big Data Processing fac referire la cantitățile uriașe de date generate și procesate, acestea nu doar determinând luarea deciziilor pentru firme, ci și impactul asupra modului în care folosim serviciile în viața noastră de zi cu zi.

https://doi.org/10.33436/v30i4y202003

Cantitatea de date pe care organizațiile le generează astăzi este exponențial mai mare decât ceea ce generam colectiv chiar și acum câțiva ani. Creșterea interesului pentru Big Data este evidentă și se vede peste tot. O serie de proiecte tehnologice dezvoltate se concentrează pe soluții Big Data și o serie de firme au intrat în afaceri care se concentrează numai pe furnizarea de astfel de soluții organizațiilor. Pentru a valorifica întregul potențial, companiile au nevoie de instrumentele potrivite pentru a procesa, analiza și stoca informațiile vitale pe care le produc și le colectează zilnic pentru rezultate în timp real.

Membru al grupului de cercetare Gartner, Doug Laney a inventat modelul "3 Vs" pentru a defini în 2001 Big Data, în funcție de **volum, viteză** și **varietate**. Recent s-a ajuns la cei „10 Vs”, celor 3 V adăugându-se: **variabilitatea, veridicitatea, validitatea, vulnerabilitatea, volatilitatea, vizualizarea** și **valoarea.** Astfel se observă că nu numai cantitatea de informații care definește Big Data este importantă ci și viteza la care se ajunge, precum și numeroasele categorii diferite de date implicate.

Tehnologia Big Data a evoluat pentru a deveni una dintre cele mai căutate domenii tehnologice de către organizații și instituții guvernamentale. Cele patru elemente principale ale oricărui proiect de Big Data sunt: de stocare a datelor (big data storage), de extragere a datelor (data mining), de analiză și de vizualizare.

Eficacitatea abordărilor existente pentru detectarea anomaliilor în rețea nu este suficientă pentru o detectare în timp real, datorită acumulării unui volum mare de date colectate prin intermediul dispozitivelor conectate, conform [1], evidențiindu-se că este esențial să se propună un cadru care să se ocupe efectiv de prelucrarea mare a datelor în timp real și să detecteze anomalii în rețele. Studiul a analizat tehnologiile avansate de procesare a datelor în timp real legate de detectarea anomaliilor și caracteristicile vitale ale algoritmilor asociați Machine Learning, a taxonomiei proceselor de prelucrare a datelor în timp real, a algoritmilor de detectare anormală și de tip Machine Learning.

Combinația dintre Big Data și analiză reprezintă o parte importantă a menținerii instituțiilor de combatere a evaziunii fiscale cu un pas înaintea agenților economici care folosesc diverse modalități de fraudă. Astfel s-au creat condițiile cadru, tehnice și legislative necesare pentru a permite specialiștilor și analiștilor să testeze teorii pe baza datelor pe care le vor colecta.

Contextul legislativ actual din România privind măsurile de reducere a fraudei și evaziunii fiscale presupune obligativitatea agenților economici de a utiliza dispozitive de marcat electronice respectiv casele de marcat cu jurnal electronic și imprimantele fiscale, lucru ce permite transmiterea în timp real a informațiilor fiscale către ANAF.

Comercianții din retail generează o cantitate mare de date într-o varietate de formate din diverse surse, cum ar fi tranzacții POS, detalii de facturare, programe de fidelizare și sisteme CRM (*Customer Relationship Management*). Aceste date trebuie organizate și analizate într-o manieră sistematică pentru a obține perspective semnificative. Clienții pot fi segmentați în funcție de modelele lor de cumpărare și de sumele cheltuite la fiecare tranzacție. De asemenea, companiile pot fi segmentate în funcție de cota de TVA, încasări sau tipuri de produse vândute. Fiscul poate utiliza aceste informații pentru analize în vederea reducerii fraudelor și evaziunii fiscale.

2. Procesul de colectare a datelor

Odată cu apariția Internet of Things (IoT), potențialul de a colecta date din magazin a crescut. Walmart a început să folosească tehnologia de identificare a frecvențelor radio (RFID) în urmă cu aproximativ un deceniu. Etichetele RFID sunt mult mai ușor de citit decât codurile de bare, deoarece nu necesită scanare directă a liniei vizuale. Această ușurință de urmărire permite utilizarea etichetelor pentru a colecta date privind mișcarea produselor prin magazin.

Chip-urile de comunicare NFC (Near Field Communication) sunt utilizate de retaileri pentru a simplifica experiența cumpărătorilor. Cea mai mare parte a utilizării actuale NFC este orientată către plăți. Cu toate acestea, mai mulți comercianți utilizează scanarea NFC ca mijloc de a oferi clienților informații suplimentare despre produs. Acest lucru ajută la colectarea informațiilor despre

produsele pe care le are în vedere un client. Deoarece cititorii NFC nu sunt prezenți în toate telefoanele inteligente, unii comercianți utilizează coduri de răspuns rapid (QR) pentru produsele lor pe care clienții le pot scana folosind o aplicație pentru funcționalități similare.

O altă metodă nouă de colectare a datelor clienților este prin intermediul jaloanelor Bluetooth (*Bluetooth Beacons*). Aceste jaloane utilizează Bluetooth Low Energy, o tehnologie încorporată în smartphone-urile recente. Jaloanele sunt plasate în magazin și pot detecta semnalul Bluetooth de la un smartphone al clientului care se află în apropiere. Aceste dispozitive pot trimite informații către smartphone prin intermediul aplicațiilor specializate. Acest lucru poate fi utilizat pentru a transmite clientului notificări despre produse și oferirea de cupoane în timp real. Mai mult, deoarece clientul interacționează cu aplicația pentru a utiliza aceste informații, efectul trimiterii informațiilor către client poate fi, de asemenea, urmărit imediat.

Este important să avem date corecte și clare în formularea problemelor despre datele pe care le colectăm. Atunci când colectăm date întâmpinăm o serie de provocări și anume:

- ce date trebuie să colectăm? Acest lucru necesită adesea o anumită familiaritate cu/ sau cunoașterea domeniului problematic;
- identificarea sursele de date;
- procesul real de extracție a datelor brute;
- evaluarea calității datelor ce presupune diferite operații de curățare, imputare și alte lucrări de „consiliere" a datelor, care consumă o mare parte din timpul tipic al proiectului de știință a datelor;
- cercetătorii în date *(Data scientist*) trebuie să judece relevanța datelor pentru problema existentă.

Colectarea datelor digitale reprezintă primul pas către analiza datelor în încercarea de a înțelege și a rezolva problemele existente la nivelul sistemului de colectare a TVA-ului. Oamenii de știință trebuie să cunoască caracteristicile datelor în cauză, iar analiștii trebuie să răspundă la întrebări ca de exemplu cum colectăm datele, ce tipuri de date există și de unde provin.

În vara acestui an a fost demarat proiectul pilot de testare a procesului de conectare a caselor de marcat electronice fiscale la serverele ANAF, obiectivul final fiind înregistrarea în timp real pe serverele ANAF a tuturor bonurilor fiscale emise de agenții comerciali. Procesul de colectare a informațiilor digitale a presupus trei etape, iar procedura tehnică oferă toate informațiile necesare pentru a permite ca în acest proiect pilot de testare al platformei să poată intra toți furnizorii de autorizați aparate de marcat electronice fiscale.

Procesul de colectare a datelor pentru cercetare și analiză presupune împărțirea în două tipuri majore: **date primare** (colectate „la sursă") și **date secundare** (colectate anterior pentru un scop care nu este specific cercetării în cauză).

Cele două **metode de colectare a datelor** sunt:

- generarea de date printr-un experiment proiectat atunci când putem controla diferiți factori în vederea studierii efectului unei variabile importante asupra rezultatului. Pentru experimente bine concepute, determinarea efectelor cauzale este ușoară;
- colectarea datelor existente deja (de exemplu cu ajutorul AMEF).

Datele furnizate pot fi considerate atât **structurate** (format XML), cât și **nestructurate** (fișiere text și alte date bazate pe locație). Datele tradiționale de vânzare cu amănuntul au fost structurate și derivate în mare parte din dispozitivele de vânzare (POS) și din datele furnizate de terți. Datele POS captează de obicei informații despre vânzări, numărul de articole vândute, prețuri și time stamps-uri ale tranzacțiilor. Combinate cu păstrarea evidenței stocurilor, aceste date oferă o mulțime de informații despre produsele vândute și, în special, coșurile de produse (colecția de articole din coș) vândute. Comercianții cu amănuntul tind să folosească programe de fidelizare

pentru a atașa datele clienților la aceste informații, astfel încât datele de vânzări la nivel de client să poată fi analizate. Datele provenite de la terți constau de obicei în informații despre prețurile, sortimentele de produse, nivelul TVA. Tendința recentă este de a capta din ce în ce mai mult date nestructurate.

Actul normativ [2] a instituit obligativitatea transmiterii prin conexiunea stabilită de către AMEF, conform structurilor fișierelor de tip XML descrise în [3], a următoarelor informații:

- datele din bonul fiscal sunt structurate astfel: valoarea totală a bonului, inclusiv TVA; numărul de ordine al bonului fiscal, pentru nivelul zilei de lucru; timestamp-ul emiterii bonului fiscal (data, ora, minutul); seria fiscală AMEF; valoarea totală TVA pe cote (pe fiecare nivel al TVA);
- datele din raportul fiscal de închidere zilnică prevăzute la art. 64;
- informații asupra pornirii și opririi AMEF, încetarea funcționării acestuia în urma unor evenimente de tipul: lipsa curentului electric sau întreruperea conexiunii la internet, precum și despre repunerea în funcțiune sau stabilirea comunicației aparatului.

2.1. Tipuri de informații fiscale

În procesul de analiză a datelor digitale colectate de la AMEF privind identificarea unei posibile fraude fiscale am identificat patru tipuri de date asociate cu patru scale primare: *nominal*, *ordinal*, *interval* și *raport*.

Scala nominală se utilizează pentru a descrie categoriile în care nu există nicio ordine specifică, în timp ce scala ordinală se folosește la descrierea categoriile în care există o ordine inerentă. Pentru a transmite informații despre mărimea relativă se utilizează scala de tip interval. Motivul pentru care contează ca scară primară utilizată la colectarea datelor este faptul că analiza din aval este restricționată în funcție de tipul de date.

În tabelul 1 sunt descrise câteva exemple de tipuri de date împărțite pe categorii.

Tabel 1. Descrierea datelor și tipurilor de date

Categorie de date	*Exemple*	*Tipuri de date*
Date Interne		
Date tranzacționate	Tranzacții de vânzare (POS / online), comenzi și tranzacții bursiere, IP-ul clienților și date de geolocalizare	Numeric, text
Date Externe		
Date ale sondajului	Recensământ, sondaj de eșantion național, anchetă anuală a industriilor, sondaj geografic, registru funciar	Text, numeric, imagine, biometrice
Agenții guvernamentale	Guvern, autoritățile de reglementare - Telecom, Energie, Banca Mondială, rapoarte de credit, rapoarte climatice, meteorologice, producție agricolă, indicatori de referință - PIB	Toate tipurile de date
Date despre site-urile social media, date generate de utilizatori	Twitter, Facebook, YouTube, Instagram, Pinterest Wikipedia, videoclipuri YouTube, bloguri, articole, recenzii, comentarii	Toate tipurile de date

Cu **date nominale** putem calcula modulul, unele frecvențe și procente. Nimic dincolo de acest lucru nu este posibil datorită naturii datelor.

Cu **date ordinale** se pot calcula mediile și unele statistici de ordine de rang, în plus față de datele nominale. Acest lucru se datorează faptului că datele ordinale păstrează toate proprietățile tipului de date nominale.

Când mergem mai departe la **datele de interval** și apoi la raportarea datelor, întâlnim un salt calitativ față de ceea ce era posibil înainte, iar, media aritmetică și variația devin semnificative. Prin urmare, cele mai multe analize statistice și teste statistice parametrice (și procedurile de inferență asociate) devin toate disponibile. Pentru ca procesul de analiză a datelor să fie corespunzător sunt necesare informații de calitate, iar monitorizarea în timp real va conduce la o evaluare corectă a datelor.

La momentul actual, operatorii economici folosesc schema de proces din figura 1 pentru depunerea declarațiilor la nivelul ANAF.

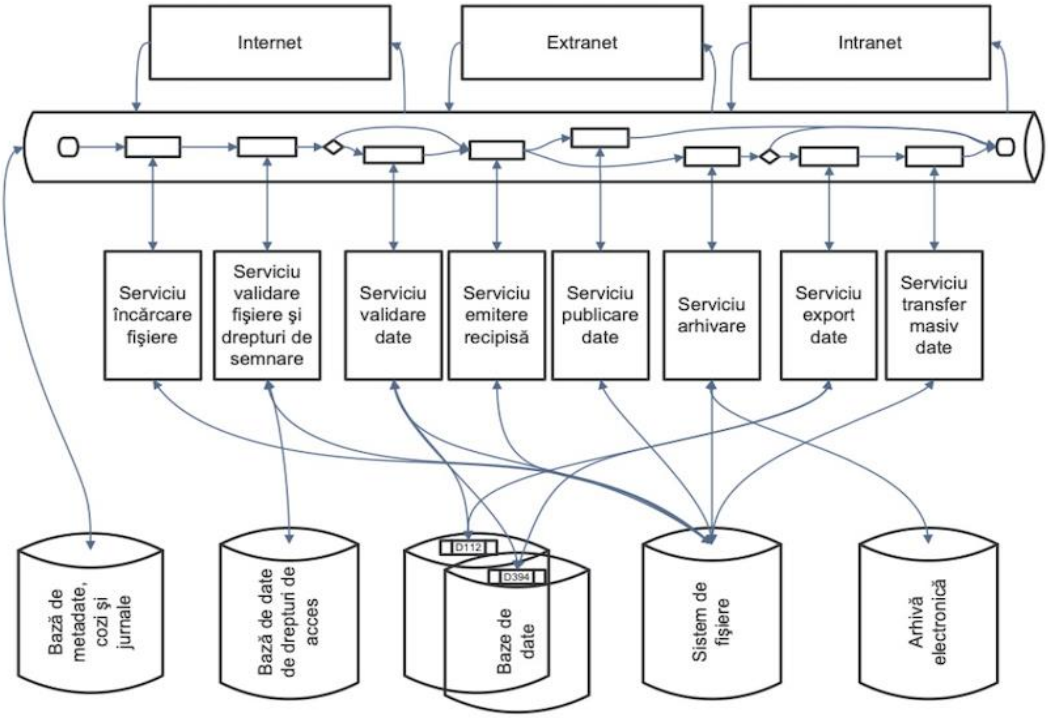


Figura 1. Schemă proces depunere declarații ANAF

Hotărârea de Guvern nr. 479/2003 [2] privind obligativitatea operatorilor economici de a folosi AMEF, conform art. 22, un AMEF trebuie să îndeplinească mai multe condiții, printre acestea aflându-se faptul că trebuie să conțină un modul fiscal propriu ce are un set minim de funcții accesibile disponibile prin comenzi de interfațare cu restul aplicației AMEF, prin intermediul căruia controlează: un dispozitiv de comunicație externă; un modul criptografic certificat; importul certificatului digital folosit de ANAF.

Formatul datelor se regăsește sub formă structurată – fișiere de tip XML și sub formă de date nestructurate (fișiere jurnal electronic, date comerciale de tip adresă, cod CAEN).

Datele conținute în Registrul național de evidență a aparatelor de marcat electronice fiscale conform [3] oferă informații cu privire la AMEF-urile instalate în România, precum și despre metodologia și procedura de înregistrare.

În tabelul 2 sunt prezentate Informațiile din acest Registru:

Tabel 2. Informațiile din Registrul național de evidență AMEF – OP ANAF 4156/2017

	Datele de identificare a producătorului, importatorului				
1	cod de identificare fiscală	denumire	calitate (producător, importator, distribuitor autorizat, unitate acreditata)	număr și dată de emitere a autorizației de distribuție	date de contact (email; nr tel)
	Datele referitoare la persoana desemnată să transmită informații și/sau să efectueze înregistrări privind AMEF				
2	nume, prenume, CNP/CUI	date contact (email, nr telefon)	elemente de identificare a certificatului digital calificat utilizat pentru transmiterea informațiilor (nr serie, autoritatea emitenta)	calitate (tehnician service)	drepturile acordate persoanelor prevăzute la subpct. anterior
	Datele de identificare a aparatului de marcat electronic fiscal				
3	numărul unic de identificare AMEF	elemente de identificare a certificatului digital instalat pe aparat	seria de fabricație	certificat digital inițial/înlocuit	
	Datele privind localizarea/amplasarea aparatului de marcat electronic fiscal				
4	adresa completă a locului de depozitare/ instalare/ intervenție, respectiv județ, localitate, stradă, număr, bloc, scara, etaj, apartament	coordonatele GPS ale locului de depozitare/ instalare/ intervenție, respectiv longitudinea și latitudinea exprimate în grade, minute și secunde (DMS	numărul de înmatriculare al autoturismului utilizat în activitatea de transport în regim de taxi	utilizat în activitățile de comerț/servicii prestate în regim ambulant sau în mijloacele de transport	
	Datele referitoare la starea aparatului de marcat electronic fiscal				
5	furat/ dispărut/ distrus/ dezmembrat pentru piese de schimb	exportat/livrat intracomunitar/exportat temporar/reexportat	adaptat pentru a corespunde prevederilor art. 3 alin. (2) din OUG 28/1999	fiscalizat/ nefiscalizat	alte indicii despre starea de funcționare a aparatului
	Datele cu privire la starea societății ce are calitatea de producător, importator				
6	declarată inactivă potrivit legii și data de la care a fost înscrisă inactivitatea	radiată de către Oficiul Național al Registrului Comerțului			
	Datele cu privire la operatorul economic beneficiar al transferului sau livrării unuia sau mai multor aparate de marcat electronice fiscale				
7	cod de identificare fiscală	denumire	seria, numărul și data avizului de însoțire a mărfii/facturii	seria de fabricație a aparatului de marcat electronic fiscal înscrisă în avizul de însoțire a mărfii/factură	codul/codurile CAEN al/ale activității/activităților pentru care este utilizat aparatul de marcat electronic fiscal
	Datele privind instalarea și intervențiile efectuate asupra aparatului de marcat electronic fiscal				
8	modalitatea agreată pentru notificare defecțiune (posta, fax)	codul de identificare fiscală a distribuitorului autorizat	numărul de identificare a tehnicianului de service	tipul intervenției	data și ora instalării/efectuării intervenției

Un aspect foarte important al activității de colectare în timp real al bonurilor fiscale este monitorizarea acestei activități, indiferent dacă este vorba despre un singur operator economic sau de o anumită perioadă de monitorizare.

Informațiile fiscale ale agenților economici sunt analizate cu ajutorul colectării rapoartelor de închidere zilnice, ce includ informații relevante legate de prețuri, cote tva, în special totalizatoarele la nivelul unei zi de lucru cât și evenimente neprevăzute cum ar fi căderi de tensiune. Aceste date sunt îmbunătățite cu valori referitoare la adresa comerciantului, necesare unor analize geografice, date referitoare la obiectul de activitate al comerciantului (coduri CAEN) cât și extrasele din jurnalul electronic (fișiere text) ce conțin toate informațiile fiscale înregistrate la nivelul unei zi de lucru în format brut.

2.2. Agregarea și transmiterea datelor

Colectarea datelor din mai multe surse nu va avea ca rezultat informații bogate, cu excepția în care datele sunt colectate pentru a-și păstra integritatea [8]. Valabilitatea datelor poate fi compromisă dacă nu este îngrijită în mod corespunzător în timpul colectării. În timpul colectării și agregării datelor pot apărea o serie de probleme cum ar fi:

- fără identificatori comuni: în timp ce se colectează date din mai multe surse apare o

problemă din cauza absenței identificatorilor comuni pe diferite surse. Analistul poate căuta un al treilea identificator care să poată servi drept legătură între două surse de date;

- date lipsă, eroare de introducere a datelor: datele lipsă pot fi ignorate, șterse sau imputate cu statistici relevante;
- diferite niveluri de granularitate: datele ar putea fi agregate la diferite niveluri. Datele primare sunt colectate la nivel individual, în timp ce datele secundare sunt de obicei disponibile la nivel agregat. Se pot agrega datele pentru a aduce toate observațiile la același nivel de granularitate sau se pot împărți datele folosind logica de afaceri;
- schimbarea tipului de date în cursul perioadei de colectare sau în cadrul tuturor eșantioanelor: date financiare/fiscale, de multe ori se modifică perioada de bază sau multiplicatorii, ceea ce ar trebuie contabilizat pentru a obține coerența datelor. Se poate solicita reaprovizionarea tipurilor de date vechi sau noi pentru a aduce datele pentru analiză la același nivel;
- validare și fiabilitate: deoarece datele secundare sunt colectate de o altă persoană, cercetătorul poate dori să le valideze pentru a verifica corectitudinea și fiabilitatea acestora pentru a răspunde la o anumită întrebare de cercetare.

Metoda de prezentare a datelor este foarte importantă pentru a înțelege problemele acestora. Prezentarea de bază poate include diagrame relevante, cum ar fi: parcele de dispersie, histograme și grafice sau statistici sumare, numărul de observații, medii, variație, minim și maxim.

Pentru validarea datelor transmise de către aparatele de marcat electronice fiscale, dar și pentru colectarea informațiilor brute stocate în fișierul jurnal electronic din modulul fiscal se vor folosi fișierele text ale jurnalului electronic.

În cadrul procesului de analiză a datelor digitale colectate de la AMEF privind identificarea unei posibile fraude fiscale se va utiliza o platformă software integrată cu o mostră de date din bazele de date a fișierelor colectate până în prezent de către ANAF, din formularul A4200. Acest aspect are ca principal avantaj faptul că volumul de date transmise pe cale offline și datele stocate începând cu anul 2018, cât și datele noi raportate zilnic vor fi folosite în **analize predictive**.

Astfel se observă că modalitatea de transmitere a datelor poate fi făcută:

- **offline** prin transmiterea formularului A4200 în format PDF inteligent;
- **online** prin transmiterea de fișiere în format XML de către agenții economici. Această colectare a datelor solicitate în timp real, online, se va face prin canalul de comunicare *https* - autentificare pe bază de certificat digital necalificat.

Specificațiile tehnice de nivel minim necesare transmiterii datelor de la AMEF către sistemul informatic al ANAF în cazul conectării la distanță sunt următoarele:

- conectarea la distanță între AMEF și sistemul informatic se face prin orice tip de conexiune la internet, utilizând un canal de comunicație *https* criptat cu protocol TLS/SSL. Dacă AMEF-ul nu trebuie să transmită date către sistemul informatic, într-un interval de timp stabilit prin utilizarea unui indicator din cadrul profilului, acesta încearcă să restabilească automat comunicația;
- certificatul digital instalat pe AMEF asigură autentificarea acestuia, și criptarea și semnarea datelor;
- la cererea ANAF, dispozitivele de marcat electronice fiscale trebuie să permită **înlocuirea certificatului digital instalat** de către producător.

În cazul funcționării AMEF în modul de lucru offline cerințele sunt următoarele:

Modalitatea de transmitere a datelor offline se bazează pe extragerea datelor folosind un mediu de stocare extern, în următoarele cazuri, în funcție de caz:

- conexiunea la internet este temporar indisponibilă, din orice motiv;
- lipsa oricărui mijloc de conectare la internet la adresa unde este instalată/ utilizată AMEF;
- datele nu au putut fi transmise pe canalul de comunicare online.

Pentru transmitere, datele sunt exportate pe un suport de stocare extern într-un fișier semnat utilizând certificatul digital al AMEF. Fișierul PDF cu XML atașat obținut folosind aplicația pusă la dispoziție pe portalul ANAF este transmis sistemului informatic prin mijloace electronice de transmisie de la distanță sau prin prezentare către autoritatea fiscală.

Comunicarea dintre AMEF și sistemul informatic național de supraveghere și monitorizare a datelor fiscale se face prin trimiterea/primirea fișierelor prezentate în secțiunile II.1-II.12 din Ordinul Președintelui ANAF nr.146/2018 [4]. Fișierele XML trebuie să conțină elemente/atribute marcate ca obligatorii, iar succesiunea elementelor este obligatorie pentru fiecare secțiune. Atributele corespunzătoare pentru „Câmp obligatoriu“ ce nu au prevăzută mențiunea „DA“ se vor transmite numai în cazul în care există informația respectivă. Fișierele XML care se transmit offline sunt semnate, și se aplică o semnătură inclusă conformă cu standardul PKCS#7. Structura certificatului digital va avea minim următoarele caracteristici:

Tabel 3. Structura certificatului digital conform cu RFC 5280

ISSUER	NUMELE DISTINCTIV AL AUTORITĂŢII DE CERTIFICARE EMITENTE (DN)
SerialNumber	Numărul de serie al certificatului. Acest număr este unic și va conține minimum 4 octeți cu date generate aleatoriu.
Subject	Numele distinctiv al entității pentru care se emite certificatul. Acesta va include informații de identificare a aparatului de marcat electronic fiscal (numărul de serie al aparatului).
SignatureAlgorithm	Tipul de semnătură cu care autoritatea de certificare a emis certificatul
NotBefore	Data începând cu care certificatul este valid
NotAfter	Data la care expiră certificatul digital
Subject Key Identifier	Identificatorul cheii subiectului
Authority Info Access	[1]Authority Info Access Access Method=On-line Certificate Status Protocol (1.3.6.1.5.5.7.48.1) Alternative Name: URL=url-ul la care autoritatea de certificare pune la dispoziție mecanismul de validare al certificatelor digitale OCSP conform RFC 6960
Certificate Policies	[1]Certificate Policy: Policy Identifier=OID-ul politicii sub care sunt emise certificatele de către autoritatea de certificare [2,1]Policy Qualifier Info: Policy Qualifier Id=CPS Qualifier: url-ul unde se găsește documentul descriptiv privind emiterea certificatelor digitale de către autoritatea de certificare
CRL Distribution Points	url-ul unde sunt publicate CRL-urile emise de autoritatea de certificare

AMEF este presetat cu funcționarea profilului offline (0). Profilul online (1) este importat printr-un fișier XML care conține următoarele categorii de informații cum sunt:

- tipul profil: 1;
- data de la care se va face trecerea în profilul 1;
- numărul de zile după care activitatea AMEF este blocată;

- intervalul în minute la care AMEF încearcă să se conecteze la sistemul informatic național de supraveghere și monitorizare a datelor fiscale;
- intervalul în minute la care AMEF încearcă să se conecteze la sistemul informatic național de supraveghere și monitorizare a datelor fiscale pentru a transmite informațiile prevăzute la art. 3 alin. (1) sau (2) din anexa nr. 11 la normele de aplicare;
- URL-urile ce sunt utilizate în comunicarea AMEF cu sistemul informatic național de supraveghere și monitorizare a datelor fiscale;
- certificatul digital al ANAF în format PKCS#7 (doar certificatul codificat în baza 64).

Trecerea de la profilul 0 la profilul 1 se efectuează exclusiv de către tehnicianul de service al distribuitorului autorizat sau al unității de service acreditate.

Trecerea de la profilul 1 la profilul 0 se face prin importarea unui fișier XML semnat folosind certificatul digital al ANAF prezentat în secțiunea II.11 OP ANAF 146/2018 - Profil AMEF și care trebuie să conțină obligatoriu următoarele categorii de informații: tipul profil – 0 și data până la care AMEF funcționează în profilul 0.

Trecerea de la profilul 1 la profilul 0 se efectuează exclusiv de către tehnicianul de service al distribuitorului autorizat sau al unității de service acreditate.

AMEF-ul care lucrează în profil 1 își blochează activitatea de emitere și tipărire de bonuri fiscale dacă se atinge sau depășește termenul stabilit potrivit art. 8 din anexa nr. 11 la normele metodologice. Deblocarea AMEF se face după transmiterea datelor prevăzute la art. 3 alin. (1) sau alin. (2) din anexa nr. 11 la normele metodologice prin importul unui fișier de tip XML .

Fișierul trebuie să conțină obligatoriu următoarele categorii de informații: tip de profil și cod pentru deblocarea activității AMEF, reprezentat de valoarea id-ului ultimului mesaj registru transmis.

Pentru AMEF-urile care funcționează în modul de lucru offline, mesajul din secțiunea II.10 care conține atât codul de deblocare, cât și continuarea activității, și/sau valorile modificate ale indicatorilor de profil vor fi recepționate după transmiterea datelor.

3. Modelul de analiză a datelor

Instrumentele de **stocare a datelor** în cloud sunt folosite pentru a maximiza cantitatea de informații stocată, într-o manieră sigură și accesibilă, pentru a fi ușor de utilizat. Printre aceste instrumente de stocare regăsim: Hadoop, Mondo DB, RainStor [7].

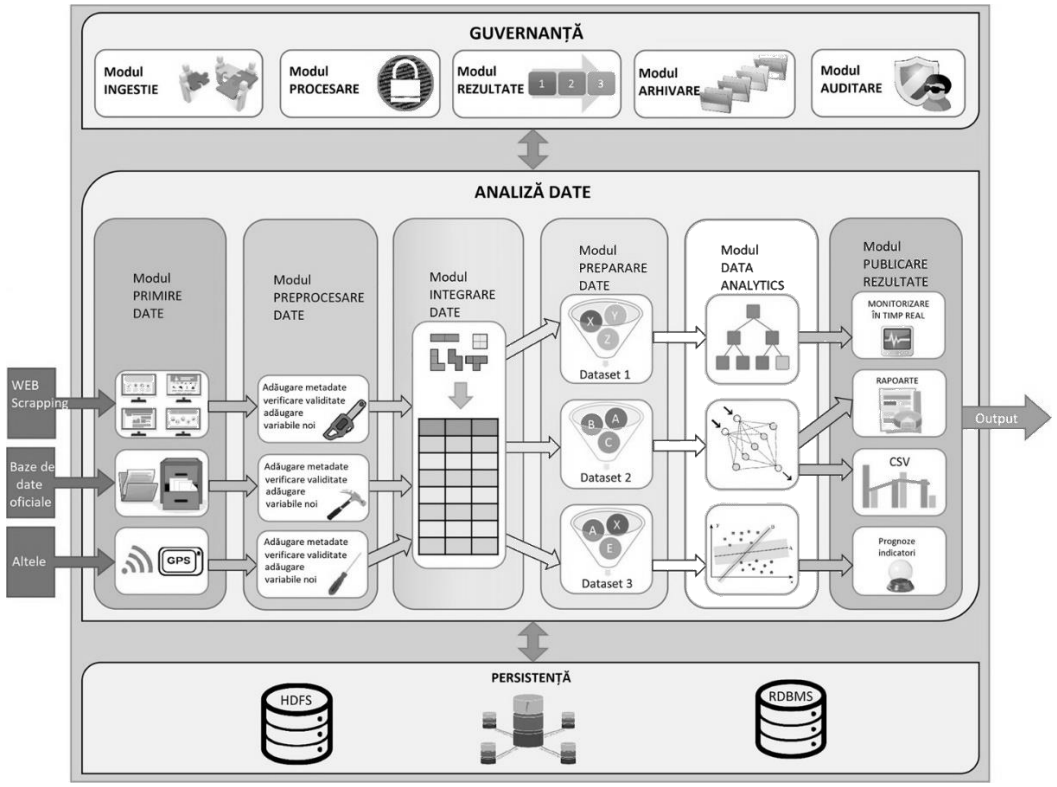
Dacă platformă open-source **Hadoop** este folosită în vederea stocării unor seturi de date foarte mari folosind clustere ce suportă atât date structurate, cât și nestructurate, **MongoDB** este folosită pentru utilizarea combinaților de date semi structurate și nestructurate, iar **RainStor**, în loc să stocheze pur și simplu Big Data, comprimă și dezabonează date, oferind economii de stocare de până la 40:1 fără să piardă vreun set de date în acest proces.

Pentru definirea modelului de analiză a datelor având stocarea și procesarea distribuite se utilizează platforma **Hadoop**, ca și componente de bază, stocarea distribuită este de tip HDFS (*Hadoop Distributed File System*).

Arhitectura este structurată pe 3 nivele – Guvernanță, Analiză Date și Persistență. Pentru nivelul de persistență se folosește stocarea pe HDFS și pe RDBMS (*Relational DataBase Management System*). Principiile de design și caracteristicile HDFS (arhitectură master-slave, High-Availability, redundanță, scalabilitate) oferă împreună cu ecosistemul Hadoop garanția unei arhitecturi de top în materie de Big Data dar în același timp flexibilă și scalabilă datorită diferitelor module integrate cu acesta.

Sistemul de management al bazelor de date relaționale a fost ales și ca o constrângere a implementărilor anterioare la nivelul Agenției Naționale de Administrare Fiscală (bazată pe tehnologii Oracle) ce oferă o funcționalitate SQL superioară.

În figura 2 se poate vizualiza arhitectura modelului de analiză a datelor.



Figură 2. Arhitectura Big Data pentru previziuni și analize economice (cercetare proprie)

Nivelul de analiză a datelor este structurat pe șase secțiuni:

- **Modulul primire date** - agregarea datelor din diferite formate:
  - Nestructurat - fișier jurnal electronic al aparatelor de marcat electronice fiscal,
  - Semi-structurate - fișierele XML cu date sintetice la nivelul zilei de lucru transmise de către aparatele de marcat electronice fiscal (online sau offline),
  - Structurate - informații deținute în bazele de date ale ANAF;
- **Modulul preprocesare date** - adăugarea de metadate, verificarea validității și adăugarea de noi variabile (cod CAEN, date statistice);
- **Modulul Integrare Date** - modul pentru gestionarea tuturor acestor date prin mapare, transformare și *data cleaning*;
- **Modulul Preparare date** - prepararea și selecția seturilor de date pentru Big Data Analytics;
- **Modulul Data Analytics** va folosi datele selectate în modulul anterior pentru antrenare, clusterizare și analiza detecțiilor de anomalii;
- **Modulul publicare rezultate** oferă monitorizare în timp real, generarea de rapoarte și de prognoză a indicatorilor.

Nivelul de guvernanță face separația între diferitele entități implicate în tot acest proces, de la ingestia și procesarea datelor de către ANAF la analizele și rezultatul acestora pentru diferitele direcții ale sistemului fiscal (Antifraudă fiscală, Control Venituri, Planificare, Monitorizare și Sinteză). Datorită cerințelor legale cu privire la arhivarea și păstrarea datelor fiscale apar și cerințe de replicare, arhivare și auditare a acestor date.

Modelul de analiză a datelor oferă o flexibilitate și o scalabilitate pentru a satisface toate cerințele de business cât și utilizarea diversității ecosistemului Hadoop la toate cele 3 niveluri:

- Stocare distribuită – persistență HDFS;

- Managementul resurselor – facil prin modulul YARN;
- Procesare distribuită – framework-urile MapReduce, Hive, Spark.

Pentru a avea rezultate cât mai bune și mai precise, se poate implementa un proces de îmbunătățire a datelor cu noi informații, denumit și **data enrichment**.

Astfel, îmbunătățirea și adăugarea de noi informații despre agenții economici altele decât cele care se regăsesc pe bonurile fiscale, din surse deschise (listă firme, date firme) aduc un plus de valoare analizei datelor prin noi informații financiare precum cifra de afaceri, profit pe ultimii ani cât și informații despre datele de autentificare ale firmei, cod CAEN, localitate, județ.

4. Concluzii

Analiza unui set de profiluri de agenți economici și informațiile de referință din bonurile fiscale colectate în timp real, folosind tehnici adecvate de extragere a datelor și construirea de modele predictive din aceste date ar putea defini strategii utile pentru a identifica comportamente atipice și a îmbunătăți calitatea procesului de depistare timpurie a unor posibile modalități de evaziune fiscală.

Tehnicile de extragere a datelor au acces rapid și extrag informații din bonurile fiscale colectate pentru a produce cu ușurință modele interpretabile, folosind tehnici de partiționare, clustering și pre-procesare. Sarcina de a găsi tiparele în datele digitale a bonurilor fiscale capătă o relevanță mai mare, deoarece ANAF colectează și produce cantități imense de date, inclusiv informații contextuale masive, luând astfel în considerare un număr mai mare de variabile. Folosirea datelor pentru a înțelege și îmbunătăți procesul de analiză și predicție din cadrul ANAF, eficiența și utilitatea este o oportunitate excelentă de a evita evaziunea fiscală.

Modelul de analiză a datelor permite recunoașterea unor tipare și se va putea semnaliza pe cele disfuncționale sau atipice. Interfața sistemului va permite filtrarea datelor prin intermediul mai multor tipuri de date: CUI, sediu, tip de produs/serviciu, stradă, cod CAEN și alți parametri de identificare.

Sistemul informatic pentru colectarea automată a datelor conduce la îmbunătățirea relației dintre autoritatea fiscală și operatorii economici prin simplificarea interacțiunii, utilizând datele colectate pentru a identifica cu o acuratețe foarte mare a acelor operatori economici care se conformează voluntar și pe cei care încearcă să ocolească prevederile legale.

Beneficiile reale pentru firmele din piața locală care se aliniază reglementărilor fiscale constau în aceea că Autoritatea Fiscală nu va mai declanșa controale inopinate decât acolo unde analizele de risc indică existența unor posibile nereguli.

Un pas important în transformarea digitală a sistemului fiscal din România îl reprezintă faptul că acest sistem devine operațional.

Mențiuni

Prezenta lucrare are la bază parte din activitățile și rezultatele temei de cercetare PN 19370201/2020 – „*Creșterea performanțelor serviciilor de cloud prin analiza și dezvoltarea unui sistem de Billing*”, proiect ce a fost finanțat în cadrul Programului Național Nucleu 2019-2022.

BIBLIOGRAFIE

1. Habeeba, R. A. A., Nasaruddina, F., Ganib, A., Hashemb, I. A. T., Ahmedc, E., Imran, M., (2018). *Real-time big data processing for anomaly detection: A Survey*. ELSEVIER.
2. Hotărârea de Guvern nr.479/2003 privind aprobarea Normelor metodologice pentru aplicarea OUG 28/1999 privind obligația agenților economici de a utiliza aparate de marcat electronice fiscale, republicată în Monitorul Oficial al României, Partea I, nr. 348/25.04.2005, cu modificările și completările ulterioare.

3. Ordinul Președintelui ANAF nr. 4156/2017, modificat prin Ordinul ANAF nr. 857/2019.
4. Ordinul Președintelui ANAF nr. 146/2018 - Profil AMEF.
5. Mohammed J. Z., Wagner M., (2014). *Data Mining and analysis - Fundamental concepts and Algorithms*. Cambridge University Press, pp. 25-30.
6. Barbu, D. C. (2019). *Soluţii de prelucrare specifice Big Data*. Revista Română de Informatică și Automatică (Romanian Journal of Information Technology and Automatic Control), ISSN 1220-1758, vol. 29(2), pp. 35-48.
7. Pochiraju, B, Seshadri, S. (2019). *Essentials of Business Analytics - An Introduction to the Methodology and its Applications*. International Series in Operations Research & Management Science, SPRINGER.

Dragoș-Cătălin BARBU este doctorand în cadrul Academiei de Studii Economice din București, în domeniul „Informatica Economică”, a absolvit Facultatea de Matematică și Informatică din cadrul Universității din București și deține diplomă de master în domeniul „Informaticii Teoretice” din cadrul Departamentului de Informatică, Facultatea de Matematică și Informatică, Universitatea din București. În prezent deține funcția de Șef Serviciu „Cloud Computing” și este Cercetător Științific gradul III în cadrul Institutului Național de Cercetare-Dezvoltare în Informatică – ICI București, desfășurând activitate de cercetare în domeniul TIC de peste 15 ani. Este reprezentant supleant al României în Consiliul de conducere al Întreprinderii comune europene de calcul de înaltă performanță (EuroHPC) și, de asemenea, delegatul român în Consiliul de guvernanță al European Open Science Cloud (EOSC). A coordonat proiecte naționale în domeniul „Cloud Computing”, securitate informatică, servicii electronice, librării digitale, inteligență artificială și realitate îmbogățită, a participat la realizarea a peste 25 de proiecte naționale, 8 proiecte internaționale, și a publicat peste 30 de articole la nivel național și 4 articole la nivel internațional.

Dragoș-Cătălin BARBU is a PhD candidate at the University of Economic Studies in Bucharest, he graduated from the Faculty of Mathematics and Computer Science at the University of Bucharest and holds a master's degree in the field of Theoretical Informatics from the Department of Computer Science, the Faculty of Mathematics and Computer Science, the University of Bucharest. He is the Head of the “Cloud Computing” Department and a Senior Researcher III within the National Institute for Research and Development in Informatics – ICI Bucharest. Dragoș-Cătălin Barbu has been carrying out research activity in the ICT field for over 15 years, coordinating national projects in the field of “Cloud Computing”, computer security, electronic services, digital libraries, artificial intelligence and augmented reality. He is the Romanian Substitute Representative in the Governing Board of European High Performance Computing Joint Undertaking (EuroHPC) and also the Romanian delegate in the European Open Science Cloud (EOSC) Governance Board. Moreover, he participated in the implementation of more than 25 national projects, 8 international projects and he has published over 30 articles at a national level and 4 articles at an international level.