

# Иновативни подходи за обработка на уеб данни в епохата на дигитализацията

Явор Табов\*

**Резюме:** Нарастващият обем и значимост на уеб данните в съвременната дигитална ера предоставят както възможности, така и предизвикателства за бизнеса и организациите. Уеб данните се отличават със своята обширност, разнообразие и динамична природа, както и с тенденцията си за бързо разрастване, което налага използването на сложни подходи за тяхната обработка и управление. Статията разглежда основните етапи в обработката на уеб данни – от събирането, подготовката и трансформирането до тяхното съхранение, анализ и представяне. Обсъдени са ключови предизвикателства, включително проблеми, свързани с качеството и структурата на данните, трудности в процесите на обработка и въпроси, касаещи управлението на данните. На този фон са представени популярни решения, които използват съвременни технологии и стратегии за мащабируема интеграция и ефективен анализ на данни. В заключението се подчертава

важността за бизнеса и организациите да инвестират в разработването и внедряването на иновативни решения за обработка на уеб данни, които са от решаващо значение за извличането на стойностна информация и поддържането на конкурентоспособност в съвременната дигитална епоха.

**Ключови думи:** обработка на уеб данни, анализ на данни, NoSQL, обработка в реално време.

**JEL:** C88, L86.

## Увод

С развитието на цифровите технологии и бързото разширяване на уеб пространството, уеб средата се утвърждава като един от най-значимите източници на данни. Данните, генерирани от уеб платформи, варират от текст и изображения до видео и аудио файлове, което разкрива огромен потенциал за анализ и използване в различни области. Затова интересът на бизнеса и организациите към обработката на уеб данни нараства експоненциално, тъй като те разчитат на

\* Явор Табов е доктор, главен асистент в катедра „Информационни технологии и комуникации“ на УНСС.

тези ресурси за формулиране на стратегически и управленски решения.

Настоящата статия разглежда особеностите на данните в уеб пространството, които включват неструктурирани, полуструктурирани и структурирани формати. Специален акцент е поставен върху предизвикателствата при обработката им, свързани с обема, разнообразието и качеството на информацията, както и с динамиката на нейното генериране и съхранение.

Освен анализ на предизвикателствата, статията представя обзор на популярни решения в областта на обработката на уеб данни, включително инструменти и технологии за събиране, почистване, анализ и визуализация. Основната цел е да се илюстрира значимостта на разработването на ефективни методи и решения за справяне с предизвикателствата на информационната среда в уеб, среда, която постоянно нараства и се развива с все по-голяма скорост.

Тази статия не само подчертава нарастващата роля на уеб данните, но и насърчава търсенето на иновации в обработката им, които да подпомогнат организациите в техните усилия за вземане на информирани решения и адаптиране към предизвикателствата на съвременната цифрова ера.

### **1. Особенности на данните в уеб. Обработка на уеб данни**

Обработка на данни включва преобразуването на сурови данни в смислена информация чрез тяхното организиране, индексирание и манипулиране с цел

разкриване на ценни връзки и модели, полезни за решаване на проблеми. Технологичният напредък значително подобри възможностите за обработка на данни, като замени ръчните, трудоемки методи с автоматизирани процеси, изпълнявани от машини и компютри. Специфичните техники, използвани в обработката на данни, зависят от типа данни (Huang, 2022).

Обработката на уеб данни се концентрира върху обработването на информация от уеб базирани източници като уебсайтове, социални мрежи и онлайн бази данни. Това често включва техники като уеб скрапинг, достъп до програмни интерфейси за приложения, парсинг на HTML или JSON данни. Данните придобиват стойност едва когато бъдат трансформирани в полезна информация чрез анализ, обработка и интерпретация (Vaughan и др., 2021). Обработката на уеб данни може да включва структурирани, полуструктурирани и неструктурирани данни от различни уеб източници.

Структурираните данни представляват високо организирана информация, в която може лесно да се търси и която се вписва в предварително дефинирани формати или схеми, като бази данни и електронни таблици. Този тип данни е организиран в редове и колони и е лесен за управление чрез структуриран език за заявки. Релационният модел на базите данни поддържа структурирани данни, като ги организира в редове и таблици, което улеснява тяхното управление и обработка (Praveen и др., 2017).

Неструктурираните данни, създавани както от машини, така и от хора, могат да бъдат текстови и нетекстови. Нетекстовите включват мултимедия (изображения, видео и аудио файлове), докато текстовите обхващат имейли, съобщения, документи и презентации. Стандарти като XML, SMTP, SMS и CSV са разработени за управление на неструктурирани данни (Kanimozhi и др., 2015). Неструктурираните данни обикновено липсват в стандартизиран формат, което ги прави сложни и времеемки за обработка. Организацията днес притежава огромни количества неструктурирани данни, но особеностите на тяхната форма затрудняват извличането на стойност (Chowdhury, 2020).

Полуструктурираните данни, които технически са подмножество на структурирани данни, нямат фиксиран модел. Езици за маркиране като XML и JSON обикновено се използват за управление на полуструктурирани данни (Kanimozhi и др., 2015). Този тип данни комбинират елементи на структурирани и неструктурирани данни, предлагайки гъвкавост и мащабируемост, като избягват строгите модели на базите данни, както се вижда във формати като CSV файлове (Jiang и др., 2022).

Всеки ден в световен мащаб се създават безброй структурирани, полуструктурирани и неструктурирани документи (Zaman и др., 2020). Дигиталните данни, генерирани от различни цифрови платформи и устройства, нарастват с поразителна скорост (Nti и др., 2022). За илюстрация на този ръст

могат да бъдат използвани анализи от германската платформа *statista.com*, специализирана в събиране и визуализация на данни. Тези анализи показват, че количеството произведена информация е нараснало десетократно в периода от 2014 до 2023 г. Очаква се до 2028 г. този дневен обем да се увеличи до 394 зетабайта (Statista Research Department, 2024). Тази тенденция на растеж на данните е илюстрирана на фигура 1.

Обработката на уеб данни е процес, който обхваща етапите на събиране, подготовка, трансформиране, съхранение, анализ и визуализиране на данни, генерирани от различни уеб източници. С нарастващото значение на цифровите платформи и бързо увеличаващото се количество данни в интернет, обработката на уеб данни се е утвърдила като основен инструмент за бизнеса, науката и технологиите. Дигитализацията на тези процеси изисква внимателно проектиране на бизнес логика и функционалности, които отговарят на спецификата на предметната област. Както отбелязва Милев (2022), успешното изграждане на информационни системи за управление на данни предполага дефиниране на специфични модели на потребителски случаи и бизнес процеси, които интегрират нормативните изисквания с технологичните решения. Пример за това са системите за отчитане и рецензиране на научни проекти, които предлагат структурирани подходи за въвеждане, обработка и анализ на данни. Етапите в обработката



Фигура 1. Тенденция на нарастване на данните в уеб



Фигура 2. Етапи в обработката на уеб данни

на данни уеб данни са представени на фигура 2.

Събирането на уеб данни е процесът на извличане на данни от различни онлайн източници като уебсайтове, социални мрежи и програмни интерфейси за приложения. Тези данни обикновено се събират от уеб роботи и намират приложение в областта на индексирането, извличането на информация и знание (Liu, 2013).

Предварителната подготовка включва преобразуването на суровите данни в по-подходящ формат, който улеснява и ускорява тяхната последваща обработка според нуждите на потребителя. Основната цел е извличането на стандартизирани данни от оригиналния формат, които ще бъдат подготвени за последващи алгоритми за анализ (Losarwar и gr., 2012).

Процесът на трансформация включва множество стъпки, всяка от които променя данните по различен начин. Това може да включва коригиране на структурата на данните (трансформации, свързани със схемата) или промяна на действителните стойности на данните (трансформации на инстанции). В контекста на хранилищата на данни трансформацията означава преобразуване на данни от формата на източника във формат, съвместим с изискванията на дестинацията, за да се осигури съответствие и използваемост (Ali и gr., 2013).

Съхранението на данни е процесът на запазване и съхранение на цифрова информация на устройство или носител с цел бъдещ достъп. С бързото

развитие на технологиите, съхранението на данни е ключов аспект в съвременните компютърни системи (Хи, 2023).

Анализът на данни има за цел извличането на смислена информация от данните, за да се подкрепи процесът на вземане на решения. Тази практика изисква организиране и структуриране на данните, което е от съществено значение за разбирането на съдържащата се информация (Islam, 2020).

Визуализацията на данни се свързва не само с графично представяне на данни, но и с разкриване и комуникация на прозренията в тях. Ефективната визуализация помага на зрителите да разберат структурата и значението на информацията (Chen и gr., 2007).

## 2. Прегизвикателства в обработката на уеб данни

Обработката на уеб данни е от съществено значение за бизнеса, научните изследвания и технологиите, тъй като предоставя ценни възможности за оптимизация на процеси, подобряване на вземането на решения и генериране на иновации. Въпреки това, тя поставя редица предизвикателства, които трябва да бъдат преодолени, за да се осигури по-висока ефективност и точност. Обработката на уеб данни изисква стратегии за управление на техния голям обем, различните формати, надеждността и скоростта на създаване. Необходимо е използването на ефективни методи за извличане на прозрения, като същевременно се осигурява сигурност на данните

и се уважава личното пространство на потребителите.

Предизвикателствата при обработката на уеб данни могат да бъдат групирани в три основни категории, като първата обхваща аспекти, свързани с данните, включително техния обем, формат и качество, които често затрудняват тяхната обработка. Втората категория се отнася до процесите, където ефективността, времето за обработка и точността на извлечената информация играят ключова роля за успешното управление на данните. Третата категория включва управленски предизвикателства, свързани със сигурността, защитата

на поверителността и спазването на законови изисквания, което налага използването на устойчиви и надеждни подходи. Категориите на предизвикателствата са илюстрирани на фигура 3.

Предизвикателствата, свързани с данните, се отнасят до препятствията, произтичащи от вътрешните характеристики на данните (Sinha, 2020). Основните предизвикателства от този тип, идентифицирани в научната литература, включват следното:

- **Обем.** Обемът описва количеството генерирани и събирани данни. Огромният мащаб и растежът на данните надхвърлят възможностите на тра-



Фигура 3. Предизвикателства при обработката на данни от уеб

диционните методи за съхранение и анализ (Shukla и гр., 2015). Въпреки че напредъкът в технологиите за съхранение и намаляването на разходите са смекчили проблемите, свързани с капацитета за съхранение, обработката остава сериозно предизвикателство. Бързото увеличение на обема на данните основно влияе върху обработката, управлението и вземането на решения, тъй като растежът на данните често изпреварва наличната изчислителна мощност. Това несъответствие създава затруднения за организациите, които се стремят да анализират и използват големи количества данни ефективно (Khan и гр., 2018). С разширяването на капацитетите за съхранение обемите, които преди са били неуправляеми, стават възможни за обработка в бъдеще (Bhadani и гр., 2016).

- **Разнообразие.** Разнообразието на данните се отнася до богатството и многообразието на техните представления, включително текст, изображения, видео и аудио. От аналитична гледна точка това е едно от най-големите предизвикателства при ефективно използване на големи обеми данни. Проблеми като несъвместими формати, некоординирани структури и непоследователна семантика създават значителни препятствия, които могат да доведат до раздробеност в анализа (Khan и гр., 2018).
- **Скорост.** Скоростта на данните се отнася до темпа, с който те се генерират. Има два основни метода за обработка на данни: пакетна обработка и обработка на потоци в реално време.

При пакетната обработка данните се събират и съхраняват, преди да бъдат обработени, докато при обработката на потоци в реално време процесът се извършва непрекъснато (Rawat и гр., 2021). Обработката на потоци е от съществено значение за избора на аналитично решение, тъй като обикновено изисква навременни и бързи аналитични резултати (Khan и гр., 2018). Тя изисква висока скорост на приложение и аналитични възможности, които позволяват на организациите да вземат решения въз основа на най-актуалната информация, което представлява значително предизвикателство.

- **Достоверност.** Достоверността на данните се отнася до тяхното качество и точност, определящи степента на доверие в събраните данни при вземането на критични решения. Данните могат да бъдат класифицирани като добри, лоши или неопределени. При несъответствия в източниците, липса на важна информация, наличие на заблуждаващи данни или неясноти, често възникват проблеми, които затрудняват анализа и намаляват ефективността на процесите. Наличието на тези проблеми значително влияе върху способността на организациите да разчитат на данните за вземане на решения. Неточните или подвеждащи данни могат да доведат до грешни стратегически избори, неефективност и в крайна сметка до финансови загуби. За да се преодолеят тези предизвикателства, организациите е необходимо да прилагат ефективни стратегии за защита на чувствител-

ността на данните и да осигурят съответствие с регулаторните изисквания (Khan и гр., 2018).

Предизвикателствата, свързани с процеса на обработка на уеб данни, включват различни трудности, които възникват на всеки етап от жизнения цикъл на данните. Те могат да се проявят при събирането и интегрирането на данни от множество източници, при анализа и интерпретацията на сложни модели и тенденции, както и при представянето на резултатите в удобна и разбираема форма за крайния потребител. Основните предизвикателства от този тип включват:

- Събиране и съхранение на данни. Събирането на данни е процес на придобиване и съхраняване на информация за бъдещо използване с цел извличане на ценни знания. Данните се събират от различни източници като сензори, социални мрежи, блогове и други. Това води до наличието на данни в различни форми — структурирани, полуструктурирани и неструктурирани, което създава значителни предизвикателства при тяхната обработка. Съхранението на тези данни е следващото предизвикателство. Генерирането на данни чрез различни устройства не гарантира, че всяка част информация е полезна, затова е необходимо да се прилагат умни филтри, които да създават релевантни набори от данни. Съхранението на масивни обеми данни може да доведе до високи разходи за мащабируеми системи, които да се

справят с тяхната обработка (Rawat и гр., 2021).

- Почистване и подготовка на данни. Почистването и подготовката на мащабни, неструктурирани данни е трудно поради тяхната сложност, променливост и взаимосвързани характеристики (Sinha, 2020). При работата с големи данни, ниското им качество често води до значително време, отделено за тяхното почистване (Staegemann и гр., 2021). Откриването и коригирането на непълни или грешни фрагменти е едно от основните предизвикателства в анализа на данни, а неуспехът в този процес може да доведе до неточни аналитични резултати и ненадежни решения. През последните години се наблюдава значителен интерес както от индустрията, така и от академичните среди към проблемите на почистването на данни, включително нови абстракции, интерфейси, подходи за мащабируемост и статистически техники (Chi и гр., 2016).
- Интеграция. Интеграцията на разнообразни източници на данни е от ключово значение, тъй като комбинирането на данни от различни платформи и формати може да бъде сложно и да изисква усъвършенствани техники за осигуряване на съгласуваност и използваемост. Интеграцията на данни се отнася до процеса на обединяване на данни от множество източници за създаване на единен изглед. Тя поражда значителни предизвикателства, тъй като различните среди често включват широк набор от устройства, всяко от които може да използва различни протоколи,

формати и стандарти. Целта е да се преодолеят тези предизвикателства, като се осигури съгласуваност на данните от различните източници и тяхната ефективна употреба, което да доведе до увеличаване на извлечената стойност (Dave и гр., 2024).

- Анализ и моделиране на данни. Традиционните аналитични методи са недостатъчни за справяне с бързия растеж и неструктурирания характер на съвременните уеб данни, което налага използването на усъвършенствани инструменти и методологии. Анализът на данни е процес, който разкрива скритата информация в тях и помага на организациите да вземат по-добри решения. За ефективно извличане на знания от големи набори от данни са необходими специализирани подходи, включително използването на статистически методи и техники за машинно обучение за разкриване на скрити закономерности в обширни масиви от данни (Rawat и гр., 2021).
- Визуализация. Визуализацията на данни включва смесица от технически ограничения, човешко възприятие и културни норми. Различни фактори, като възприятие, визуални конвенции, обработка на данни, социални влияния и етични съображения, могат да доведат до грешки и недоразумения при представянето на крайните резултати. Визуализацията на големи и сложни набори от данни става все по-трудна. Независимо дали става въпрос за прости или сложни проекти, проблеми като грешни, липсващи или дублирани записи трябва да бъдат адресирани преди об-

работката. Освен това трябва да се вземат предвид фактори, като размер на данните, скорост, мащабируемост в реално време, интерактивност и разнообразие. Управлението на нарастващия обем от статични и динамични данни остава основно предизвикателство (Siddiqui, 2021). Инструментите за визуализация на данни трябва да могат да обработват полуструктурирани и неструктурирани данни, тъй като това е типичният формат на големите данни (Ali и гр., 2016). Създаването на персонализирани визуализации е сложна задача, изискваща умения в анализа на данни, гизайна и разработката. Дизайнерите и разработчиците първо трябва да проучат набора от данни, за да разберат неговите характеристики, а след това да създадат подходящи визуализации, графични гизайни и интерактивни функции (Walny и гр., 2019).

Предизвикателствата, свързани с управлението, се отнасят до трудностите, свързани с достъпа, наблюдението и ефективното управление на данни. Основните аспекти на тези предизвикателства са:

- Поверителност. Поверителността е ключов аспект от управлението на данни, който подчертава връзката между събирането на чувствителна информация и нейното разпространение. Предизвикателствата за защита на личните данни, като здравни записи, финансова информация и биологични характеристики, стават все по-съществени, тъй като хората са

Все по-наясно как се обработват техните лични данни. За справяне с тези проблеми са въведени закони за поверителност, които регулират събирането и обработката на данни, а компаниите са задължени да предоставят прозрачни политики за поверителност и да назначават длъжностни лица за защита на данните, които да наблюдават практиките за тяхното управление (Bonello и др., 2015).

- Сигурност. Защитата на данните от нарушения е друг критичен въпрос, тъй като много от предизвикателствата за сигурност, свързани с традиционните бази данни, се усилват от разпределения и свързан характер на системите за големи данни. Освен това, управлението на сигурността на мрежовите журнали и разпределените среди добавя допълнителна сложност (Gandhi и др., 2017). Рисковете за сигурността се увеличават поради големия мащаб на данните и разнообразието от устройства, които участват. За подобряване на сигурността са от съществено значение техники като автентикация, авторизация и криптиране. Приложенията за големи данни се сблъскват с уникални предизвикателства, свързани със сигурността, включително мащаб на мрежата, мониторинг в реално време и липса на системи за предотвратяване на атаки. Поради това се наблюдава нарастваща необходимост от разработване на многослойни модели за сигурност и системи за превенция, които да отговарят ефективно на тези рискове (Acharjya и др., 2016). Информационна-

та сигурност обхваща различни технически аспекти, насочени към защита на системи, мрежи и бази данни от кибератаки. Тези атаки могат да включват вируси, фишинг, киберизнудване и разпределени атаки за отказ на услуга (Sapinski, 2023).

### 3. Решения и подходи за обработката на уеб данни в съвременната дигитална среда

В съвременната среда, ориентирана към данните, ефективната обработка на уеб данни е ключова за организациите, които се стремят да се възползват от потенциала на големите обеми информация. С нарастващата зависимост на бизнеса и научните изследвания от огромното количество данни, генерирани от различни източници, търсенето на надеждни решения за управление, анализ и визуализация на тези данни никога не е било по-голямо. Тези решения включват широка гама от технологии, всяка създадена за справяне с конкретни предизвикателства, свързани със съхранението, обработката и анализа на данни. По-долу ще бъдат разгледани някои от най-популярните сред тях.

Apache Storm е мощна изчислителна рамка, специално създадена за обработка на големи обеми данни, които постъпват с висока скорост в реално време. За разлика от традиционните системи за обработка на партиди, които могат да предизвикат забавяне на процеса, Storm се отличава в ситуации, когато незабавният анализ и реакция са от съществено значение.

Рамката използва архитектурата YARN (Yet Another Resource Negotiator), която осигурява висока производителност и цели да улесни ефективно групиране и управление на различни механизми за обработка на данни, позволявайки по-надеждна интеграция и изпълнение на разнообразни изчислителни задачи. Това прави тази рамка идеално решение за приложения, които изискват анализ в реално време, като например мониторинг на финансови транзакции, машинно обучение и непрекъснат надзор върху оперативните процеси. Благодарение на възможността да анализира данни в потока им, Apache Storm дава възможност на организациите да вземат навременни решения и да извличат ценни прозрения от потока от данни (Bonello и др., 2015).

NoSQL се отнася до група бази данни, които не следват традиционните системи за управление на релационни бази данни (RDBMS). Тези бази данни често се използват за работа с големи набори от данни в уеб мащаб. NoSQL включва различни продукти и идеи, свързани със съхранението и обработката на данни. Основната идея е, че докато релационните бази данни са полезни в определени ситуации, NoSQL предлага алтернативи, когато релационните не са най-доброто решение. MongoDB е пример за NoSQL база данни, която използва разпределено съхранение на файлове. Тя заема място между релационните и нерелационните бази данни. MongoDB е особено подходяща за съхранение на документи и се фокусира върху подобряване на ефективността

на съхранението и достъпа до големи обеми данни (Tan, 2018).

Инструментите за визуализация на данни могат да бъдат категоризирани в три основни типа: електронни таблици, специализиран софтуер и библиотеки за програмиране. Електронни таблици, като Microsoft Excel и Google Sheets, често се използват за основни визуализации, като стълбовидни диаграми, линейни графики и точкови диаграми. За разлика от тях, специализираните софтуери за визуализация на данни, като Tableau, QlikView и Power BI, предлагат разширени функции за създаване на интерактивни визуализации, включително табла, топлинни карти и мрежови диаграми. За специалистите с умения за програмиране, библиотеки като Matplotlib, ggplot2 и D3.js предлагат гъвкава възможност за създаване на персонализирани визуализации, въпреки че изискват по-високо ниво на техническа експертиза. Всички тези инструменти позволяват на потребителите ефективно да анализират и представят данни в различни области (Srivastava, 2023).

OpenRefine е настолен инструмент, създаден за почистване на неструктурирани данни. Той функционира като база данни в уеб браузър, предоставяйки достъпен и удобен вариант за обработка на данни, сравним с Microsoft Excel (Miller и др., 2022).

Докато прости Python скриптове могат да обработват малки набори от данни, по-големите системи изискват стабилни рамки за ефективна обработка на големи обеми данни. Един от

Водещите примери е Apache Hadoop — мащабируема рамка за партидна обработка, която може да се разширява от един компютър до големи кластери сървъри, позволявайки ефективни паралелни изчисления. Друга важна рамка е Apache Spark, която съчетава възможностите за партидна и потокова обработка. Spark обработва данни по-бързо от Hadoop, като оптимизира обработката в паметта, и е гъвкава при внедряване. Тази рамка опростява писането на програми чрез своята екосистема от библиотеки, въпреки че може да изисква повече оперативна памет, което води до увеличаване на разходите. Като цяло, и двете рамки играят ключова роля в управлението и обработката на големи обеми данни (Huang, 2022).

MapReduce е програмен модел, който използва подхода „разделяй и владей“ за обработка на големи набори от данни, разпределени в две основни фази. В първата фаза входните данни се разделят на множество по-малки задачи, които се обработват паралелно от различни възли в системата. Във втората фаза резултатите от тези паралелни обработки се събират и обобщават от основния възел, за да се получи крайният резултат. Hadoop и MapReduce са особено ефективни при съхранение на данни с устойчивост на грешки и при обработка с висока пропускателна способност, което ги прави ефективни за решаване на проблеми с големи данни (Acharjya и др., 2016).

Облачните решения за съхранение на данни играят критична роля в

контекста на обработката на уеб данни, като предоставят мащабируеми и гъвкави платформи за съхранение и достъп до огромни обеми информация. Тези решения улесняват интеграцията на различни системи и предоставят възможности за автоматизиране на обработката и анализа на данни, като по този начин подобряват ефективността и намаляват разходите. Най-популярните облачни платформи за съхранение включват решения като Amazon Web Services (AWS), Google Cloud Storage и Microsoft Azure, които предлагат надеждност и мащабируемост на глобално ниво (Gupta и др., 2021).

Apache Kafka е мащабируема софтуерна рамка, създадена за управление на големи обеми потокови данни. Основните ѝ характеристики включват разпределена обработка, висока пропускателна способност, постоянни съобщения и ефективни дискови структури. Kafka поддържа обработка на данни в реално време (онлайн) и е широко използвана за изграждане на системи за стрийминг и събитийни потоци. Със способността си да наблюдава потоци от данни и да събира статистическа информация от отдалечени приложения, Kafka е особено подходяща за обработка и анализ на данни в реално време (Kekevi и др., 2022).

TensorFlow е популярна библиотека за дълбочинно обучение, разработена от изследователи в Google, известна със своята гъвкавост и мащабируемост при анализ и моделиране на данни. Тя поддържа различни модели на невронни мрежи и значително улеснява

имплементирането на тези модели и свързаните с тях алгоритми за оптимизация, които иначе биха били трудоемки и склонни към грешки. Основните ѝ функции включват инструменти за изграждане на изчислителни графики, изпълнение и визуализация (Pang и др., 2020).

В съвременната дигитална среда обработката на уеб данни е основен фактор за оптимизацията на процеси, вземането на информирани решения и разработването на иновативни решения в бизнеса и науката. Въпреки че предизвикателствата, свързани със събирането, анализирането и защитата на данни, остават значителни, наличието на модерни технологии и инструменти предоставя ефективни подходи за тяхното управление. Изборът на подходящи методи за обработка зависи от конкретния контекст, но все по-често организациите интегрират облачни платформи, NoSQL бази данни и технологии за анализ в реално време, за да повишат ефективността си. Тези решения не само подобряват възможностите за управление на големи обеми данни, но и предопределят бъдещите насоки в развитието на сферата. Именно тази еволюция подчертава необходимостта от непрекъснато адаптиране към новите технологични тенденции и разработване на стратегии за по-ефективна обработка на уеб данни.

## Заклучение

С нарастващото значение на уеб пространството като източник на данни, обработката на информация, генерирана от онлайн източници, се превръща в критична дейност за бизнеса и организациите. Тази тенденция подчертава необходимостта от справяне с редица предизвикателства, които могат да бъдат разделени в три основни категории: свързани с характеристиките на данните, сложността на процесите и управлението на сигурността и поверителността. За да се преодолеят тези предизвикателства, се наблюдава ускорено развитие на нови подходи в областта на обработката на уеб данни. Сред ключовите бъдещи тенденции се открояват:

- Интеграция на изкуствен интелект и машинно обучение – автоматизираните методи за анализ и обработка на данни ще подобрят ефективността и точността на резултатите.
- Облачни и разпределени изчисления – компаниите все по-често се насочват към облачни платформи за съхранение и анализ на големи масиви от данни.
- Повишена сигурност и регулации – с нарастващите опасения за киберсигурността ще се засилят механизмите за защита и спазване на правните норми.
- Анализ в реално време – организации и платформи ще разчитат все повече на технологии, които позволяват бързо извличане на информация от динамични потоци от данни.

За бизнеса и организациите е от съществено значение да разработват и внедряват решения, които обхващат всички етапи от жизнения цикъл на данните – от обработка на потоци в реално време и почистване на информация до съхранение и визуализация. Тези усилия не само улесняват вземането на стратегически решения, но и осигуряват конкурентно предимство в условията на динамична и информационно

наситена среда. Ефективната обработка на уеб данни вече не е само техническо предизвикателство, а стратегическа необходимост за организациите, които искат да се адаптират и успяват в цифровата епоха.

*Публикацията съдържа резултати от изследване, финансирано със средства от целева субсидия за НИД на УНСС по договор № НИД НИ-22/2024/А.*

### Цитирани източници (References):

1. Милев, П. (2022). Технологични особености на информационна система за отчитане и рецензиране на научни проекти. *Икономически и социални алтернативи*, (1), 58-68. (Milev, P. (2022). Tehnologichni osobenosti na informatsionna sistema za otchitane i retsenziranje na nauchni proekti. *Ikonomicheski i sotsialni alternativi*, (1), 58-68)
2. Acharjya, D.P., & Ahmed, K. (2016). A survey on big data analytics: challenges, open research issues and tools. *International Journal of Advanced Computer Science and Applications*, 7(2), 511-518. ISSN: 2156-5570
3. Ali, A.A., Abdelrahman, T.A., & Mohamed, W.M. (2013). USING SCHEMA MATCHING IN DATA TRANSFORMATION FOR WAREHOUSING WEB DATA. *International Journal Information Technologies & Knowledge*, Vol.7, Number 3, 2013. ISSN: 1313-048X
4. Ali, S.M., Gupta, N., Nayak, G.K., & Lenka, R.K. (2016, December). Big data visualization: Tools and challenges. In 2016 2nd International conference on contemporary computing and informatics (IC3I) (pp. 656-660). IEEE.
5. Bhadani, A.K., & Jothimani, D. (2016). Big data: challenges, opportunities, and realities. Effective big data management and opportunities for implementation, 1-24. ISBN: 9781522501824
6. Bonello, J., & Cachia, E. (2015). Data processing: challenges and tools. 7th Workshop in Information and Communication Technology (WICT 2015), Msida. 1-4.
7. Chen, C.H., Härdle, W.K., & Unwin, A. (Eds.). (2007). Handbook of data visualization. Springer Science & Business Media. ISBN: 978-3-540-33036-3
8. Chowdhury, A. (2020). Computer & Information Systems into Next Frontier: The Emergence. New Delhi Publishers, 2020.
9. Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and emerging challenges. In Proceedings of the 2016 international conference on management of data (pp. 2201-2206).

10. Dave, D.M.K., & Mittapally, B.K. (2024). Data integration and interoperability in IoT: challenges, strategies and future direction. *International journal of computer engineering and technology (IJCET)*, 15(1), 45-60. ISSN: 0976-6375
11. Gandhi, R. V., Kumar, C. R., & Krishna, P. V. (2017). BIG DATA: ISSUES AND CHALLENGES. *International Journal of Software & Hardware Research in Engineering*, Volume 5, issue 7, July 2017. ISSN: 2347-4890
12. Gupta, B., Mittal, P., & Mufti, T. (2021, March). A review on amazon web service (aws), microsoft azure & google cloud platform (gcp) services. In Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020, 27-28 February 2020, Jamia Hamdard, New Delhi, India.
13. Huang, F. (2022). Data processing. In Encyclopedia of big data (pp. 312-316). Cham: Springer International Publishing.
14. Islam, M. (2020). Data analysis: types, process, methods, techniques and tools. *International Journal on Data Science and Technology*, 6(1), 10-15. ISSN: 2472-2235
15. Jiang, S., Nocera, A., Tatar, C., Yoder, M., Chao, J., Wiedemann, K., Finzer, W., Rosé, C. (2022). An empirical analysis of high school students' practices of modelling with unstructured data. *British Journal of Educational Technology*, Volume 53, issue 5, 2022, ISSN: 1467-8535
16. Kanimozhi, K. and Venkatesan, M. (2015). Unstructured Data Analysis-A Survey. *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 4, Issue 3, 2015.
17. Kekevi, U., & Aydın, A.A. (2022). Real-time big data processing and analytics: Concepts, technologies, and domains. *Computer Science*, 7(2), 111-123. ISSN: 2548-1304
18. Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A. A., & Salehian, S. (2018, March). The 10 Vs, issues and challenges of big data. In Proceedings of the 2018 international conference on big data and education (pp. 52-56). ISBN: 978-1-4503-6358-7
19. Liu, W. (2013, March). Web Page Data Collection Based on Multithread. In Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013) (pp. 2023-2026). Atlantis Press. ISSN: 1951-6851
20. Losarwar, V., & Joshi, D.M. (2012, July). Data preprocessing in web usage mining. In International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July (pp. 15-16).
21. Miller, M., & Vielfaure, N. (2022). OpenRefine: An Approachable Open Tool to Clean Research Data. Bulletin-Association of Canadian Map Libraries and Archives (ACMLA), (170). ISSN: 2561-2263
22. Nti, I., Quarcoo, J., Aning, J., Fosu, G. (2022). A Mini-Review of Machine Learning in Big Data Analytics: Applications, Challenges, and Prospects. *Big Data Mining and Analytics*, Volume 5, Number 2, June 2022, ISSN: 2096-0654
23. Pang, B., Nijkamp, E., & Wu, Y.N. (2020). Deep learning with tensorflow: A review. *Journal of Educational and Behavioral Statistics*, 45(2), 227-248. ISSN: 1076-9986

24. Praveen, S., & Chandra, U. (2017). Influence of structured, semi-structured, unstructured data on various data models. *International Journal of Scientific & Engineering Research*, 8(12), 67-69. ISSN: 2229-5518
25. Rawat, R., & Yadav, R. (2021). Big data: Big data analysis, issues and challenges and technologies. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012014). IOP Publishing.
26. Sapinski, A. (2023). The importance and challenges of information security in the digital age: analysis of the current situation and prospects for development. *The Scientific Journal of Bielsko-Biala School of Finance and Law*, Vol 27, No. 1 (2023). ISSN: 2543-9103
27. Shukla, R.K., Pandey, P., & Kumar, V. (2015). Big data frameworks: at a glance. *Int. J. Innovations Adv. Comput. Sci. IJACS*, 4(1), 169-175. ISSN: 2347-8616
28. Siddiqui, A.T. (2021). Data visualization: A study of tools and challenges. *Asian Journal of Technology & Management Research (AJTMR)*, ISSN: 2249-0892
29. Sinha, D.S. (2020). Big Data Analysis: Concepts, Challenges and Opportunities. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, ISSN: 2347-5552
30. Srivastava, D. (2023). An Introduction to Data Visualization Tools and Techniques in Various Domains. *International Journal of Computer Trends and Technology*, 71(4), 125-30. ISSN: 2231-2803
31. Staegemann, D., Volk, M., Saxena, A., Pohl, M., Nahhas, A., Häusler, R., ... & Turowski, K. (2021). Challenges in Data Acquisition and Management in Big Data Environments. In *IoTDBS* (pp. 193-204).
32. Statista Research Department (2024). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028 [Infographic]. Statista. <https://www.statista.com/statistics/871513/worldwide-data-created/>
33. Tan, Q. (2018, October). Application of MongoDB technology in NoSQL database in video intelligent big data analysis. In 8th International Conference on Management and Computer Science (ICMCS 2018) (pp. 104-108). Atlantis Press. ISSN: 2352-538X
34. Vaughan, J.P., C. Victora, A. Chowdhury (2021). *Data Processing and Analysis, Practical Epidemiology: Using Epidemiology to Support Primary Health Care*, Oxford Academic, 1 Oct. 2021, ISBN: 9780191944000
35. Walny, J., Frisson, C., West, M., Kosminsky, D., Knudsen, S., Carpendale, S., & Willett, W. (2019). Data changes everything: Challenges and opportunities in data visualization design handoff. *IEEE transactions on visualization and computer graphics*, 26(1), 12-22.
36. Xu, C. (2023). Exploring the World of Data Storage. *International Journal of Sensor Networks and Data Communications*, Volume 12:2, 2023, ISSN: 2090-4886
37. Zaman, G., H. Mahdin, K. Hussain, A. Rahman (2020). Information extraction from semi and unstructured data sources: a systematic literature review. *ICIC Express Letters*, Volume 14, Number 6, June 2020, ISSN: 1881-803X

## Inovativni podhodi za obrabotka na ueb danni v epohata na digitalizatsiyata

Yavor Tabov

## Innovative Approaches for Web Data Processing in the Digital Age

Yavor Tabov

**Abstract:** The growing volume and importance of web data in the modern digital era present both opportunities and challenges for businesses and organizations. Web data is characterized by its vastness, diversity, and dynamic nature, as well as its tendency for rapid growth, necessitating complex approaches to its processing and management. This article examines the primary stages of web data processing—ranging from collection, preparation, and transformation to storage, analysis, and presentation. Key challenges are discussed, including issues related to data quality and structure, difficulties in processing workflows, and concerns regarding data governance. Against this backdrop, the article presents popular solutions leveraging modern technologies and strategies for scalable integration and efficient data analysis. The conclusion highlights the importance for businesses and organizations to invest in developing and implementing innovative solutions for web data processing, which are critical for extracting valuable insights and maintaining competitiveness in the contemporary digital age.

**Key words:** web data processing, data analysis, NoSQL, real-time processing.

**JEL:** C88, L86.