Обработка и анализ на неструктурирани данни с помощта на изкуствен интелект

И. Велкова

Unstructured Data Processing and Analysis Using Artificial Intelligence

I. Velkova

Department of Information Technology and Communications, University of National and World Economy, Student District, 19, December 8th St., 1700 Sofia, Bulgaria, ivonavelkova@unwe.bg

Key Words: *IDOL; Hadoop; artificial intelligence; big data; analysis, processing; unstructured data.*

Abstract. *This paper presents at the challenges and prospects of handling and analyzing unstructured big data with artificial intelligence. A huge amount of data like video, text, images, and social content from various applications is created, collected, and accumulated every day in all areas of human activities. This unstructured data cannot be effectively stored, processed, and analyzed using traditional tools and databases. This paper addresses this issue and proposes an approach to help process and subsequently extract knowledge from unstructured data collected from various sources using artificial intelligence. Some of the few existing technological tools for processing unstructured data in a Hadoop environment are presented. As a result of the study, an architecture for such processing and analysis is proposed. It includes a set of technologies that provide a possible solution to the problem.*

1.Въведение

През изминалото десетилетие данните се превърнаха в един от определящите успеха ресурс за всяка област на човешката дейност. Броят на потребителите, използващи интернет и заявките за търсене в онлайн пространството, постоянно нараства. През 2021 г. общото количество генерирани данни е 79 зетабайта (ZB) или това са приблизително 79 трилиона гигабайта [1]. Този процес на увеличаващо се количество на информацията ще продължи. В проучване на Statista се прогнозира, че общото количество данни, генерирани в интернет пространството в световен мащаб, ще нарасне до над 180 ZB до 2025 г. [2]. Forbes от своя страна прогнозират, че повече от 150 ZB данни в реално време ще се нуждаят от анализ до 2025 г. [3]. Предизвикателството, пред което е изправен светът, е, че по-голямата част от тези данни са неструктурирани. Те идват под формата на текст, видео, аудио и уеб съдържание, които са трудни за обработка в сравнение с данните, които се съхраняват в таблици. Неструктурираните данни са от голямо значение за бизнеса, тъй като при тяхната обработка могат да бъдат извлечени тенденции, динамични модели или релации, които да подобрят вземането на решения [4]. Затова с увеличаването на размера на данните на пазара започват да се създават и използват различни облачни и софтуерни продукти, които да спомогнат за съхранението, обработката и анализа на големи данни в реално време. Все повече от предлаганите решения включват в себе си изкуствен интелект (ИИ), който позволява ефективна обработка на големи количества данни, включително и неструктурирани данни [5]. Използването на изкуствения интелект в приложения за големи данни спомага за увеличаване на ефективността от обработката на данни, тяхната визуализация и изготвянето на прогнози. Той развива разбирането на данните чрез непрекъснато учене както от данните, така и от поведението на крайните потребители [5]. Целта на настоящата работа е да представи възможен подход за анализ на големи данни чрез използването на ИИ.

2. Големи данни

По своята същност големите данни са сложни множества от данни, които се характеризират с огромен размер, генерирани от различни източници и достигащи обем до няколко зетабайта, както и с бърза скорост, с която се обработват от специализирани системи. Често големите данни се представят като трите V: V^3 (volume, variety, velocity) [6]. Големите данни от гледна точка на структура могат да бъдат структурирани, полуструктурирани и неструктурирани [7]. За да могат да бъдат ефективно обработвани, те трябва да бъдат подходящо организирани, съхранявани и анализирани.

- **Структурираните данни** съответстват на предварително дефиниран модел на данни и тяхното съхранение е в табличен вид – например SQL релационни бази данни [8].

- **Полуструктурираните данни** се състоят от документи, съхранявани във формат на JSON или XML. Те нямат точно съответствие с дефинирана структура на релационна база данни. При анализ и обработка на този тип данни се използват класифициращи характеристики като етикети, метаданни или маркировки, които разделят и диференцират различни елементи от данни, поставяйки ги в двойки и йерархии [4].
- **Неструктурираните данни** са множества, които нямат предварително дефиниран модел. За тяхното съхранение и анализ се използват сложни системи като Hadoop, Spark, Cassandra, MongoDB и други. Тези софтуерни продукти използват динамичните модели, тенденции и връзки, съществуващи в данните за вземане на интелигентни решения. От всички генерирани и събирани данни в световен мащаб близо 80% са неструктурирани – публикации в социалните медии, изображения, аудио, видео и други [4], [10].

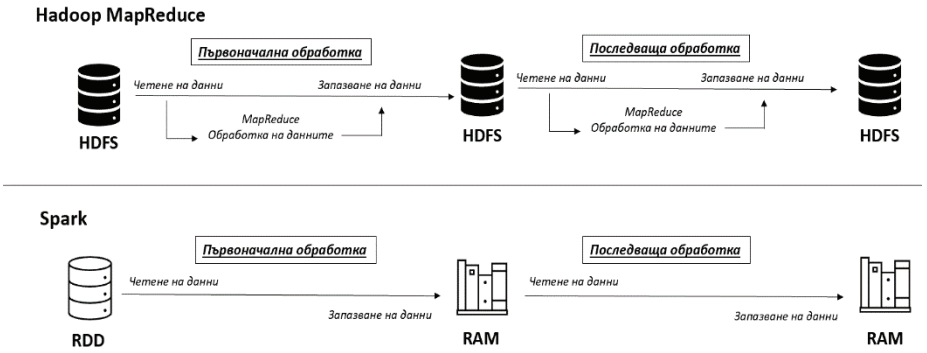
3. Средства за обработка на неструктурирани данни

Изследване на IDC сочи, че 90% от неструктурираните данни не се анализират [11]. Основното предизвикателство, свързано с неструктурираните данни, е липсата на инструменти за извличане на подходяща информация, която да бъде обработена и анализирана, след което да се превърне в знание. За да може да се вникне и да се разбере смисълът на данни от неструктуриран вид, все по-масово започва използването на модели на машинно обучение. Те спомагат на процеса по обработка за извличане на знания от събраните данни и идентифициране на нови възможности за решаване на конкретни проблеми.

Най-популярните решения за обработка на големи данни са Apache Hadoop и Apache Spark [12], [13]. Двата инструмента са широко използвани и съдържат технологии, които подготвят, обработват, управляват и анализират големи множества от данни [14].

- **Apache Hadoop**

Hadoop е система за разпределено съхранение и обработка на големи данни. Най-характерното при Hadoop системата е разделянето на данните на по-малки блокове, които да бъдат прехвърлени и използвани от взаимносвързани машини, съставящи Hadoop клъстери [14]. Ядрото на Hadoop се състои от част за съхранение Hadoop Distributed File System (HDFS) и част за обработка на данните MapReduce [14]. Това е алгоритъм, чиято задача е да обработва блоково файлове, събрани за определен период време. Има две основни фази на обработка на данните: първата е свързана с картографиране на данните, включваща дейности като филтриране, сортиране или разделяне на данните, а втората фаза е свързана с редуцирането или агрегирането на получените резултати [15]. Този процес е представен на *фиг. 1*.



Фиг. 1. Обработка на данните от Hadoop и Spark

- **Apache Spark**

Spark е софтуерно решение с отворен код, предназначено за обработка на данни и приложения с ИИ. Основната характеристика на Spark е възможността да съхранява междинни данни в RAM паметта, като по този начин осигурява по-бърза скорост на обработка – *фиг. 1* [15]. Apache Spark има два режима на обработка на данните: пакетен и поточен. Разликата между двата режима е, че при първия се обработва голямо количество исторически събрани данни. При втория режим непрекъснато се четат входящите данни от различни източници в реално време с включени алгоритми за машинно обучение. Spark използва възможностите на различни приложения като MLlib, съсредоточено върху използването на машинно обучение. Данните се групират на малки парчета, след което се подават на Resilient Distributed Datasets (RDD) – основна структура от данни на Spark, която представлява колекция от обекти [15], [16]. Обработката на данни в двете системи е представена на *фиг. 1*.

При Apache Hadoop данните се четат и записват в HDFS, докато при Apache Spark данните първоначално се четат от RDD и се запазват в паметта, като при всяка следваща обработка данните се четат от паметта, където последно са били запазени.

Засега няма известен подход за обработката на неструктурирани данни с разгледаните дотук продукти. Затова по-долу се предлага подход за такава обработка, използващ изкуствен интелект, с което се цели получаване на знания. За тази цел ще се използва Micro Focus IDOL – Intelligent Data Operating Layer.

- **Micro Focus IDOL – слой за работа с интелигентни данни**

По своята същност IDOL е система, извършваща анализи върху данни от различни източници и от различен тип – структурирани, полу- и неструктурирани. Тези анализи се актуализират непрекъснато и извличат информация от данните в реално време от компоненти, наречени конектори. Те са основен компонент на IDOL, като тяхната цел е да осигурят гъвкавост и прецизност при управлението на събираните данни [17]. Конекторите се избират на базата на вида на източника данни. Такива могат да бъдат различните социални медии – Twitter, Facebook, YouTube и други, облачно базирани хранилища, локални и мрежови файлови системи, интернет сървъри, уебсайтове, Hadoop и други. Предимството на IDOL пред другите софтуерни решения е използването на изкуствен интелект с включено машинно обучение и невронни мрежи, които заедно се саморазвиват и непрекъснато учат за подобряване на анализа, както от получаваните данни, така и събират информация за поведението на крайните потребители. Едни от свойствата на IDOL са откриване на тенденции и разкриване на модели на действие за интелигентно вземане на решения [10], [18].

4. Обработка на неструктурирани данни с изкуствен интелект

Продуктът IDOL предоставя възможности за обработка на данните с изкуствен интелект в реално време, като идентифицира ключови обекти в изображения – лицево разпознаване или разпознаване на цифри (например регистрационни номера на автомобили), а също и анализ на текст чрез система за разпознаване на знаци – OCR (Optical Character Recognition). Може да се извършва търсене по име или може да бъде по зададени шаблони, които идентифицират данни от определен тип – например телефонен номер. Предлаганият подход и архитектура за обработка на данните предоставят възможност да се съвмести продуктът IDOL със скалируемите възможности на Hadoop. Последният има възможност да работи съвместно с други проекти на Apache [19]. Важно е да се отбележи, че към момента Hadoop и Spark нямат възможност да събират данни директно от социални медии в реално време. Затова архитектурата, която се предлага тук, включва IDOL, който предоставя конектори за директно събиране на информация от множество хранилища, включително от социални медии в реално време. Този подход предлага данните от тези социални мрежи да се свържат с данните в Hadoop. По този начин данните от социални медии ще се обработват в реално време с помощта на средствата с ИИ на IDOL. Освен това Hadoop средата може да се използва да съхранява данни от IDOL, където те са индексирани. Това индексиране помага за по-бързо търсене и обработка на неструктурирани данни, съхранявайки метаданните им в базата данни на IDOL. В същото време Hadoop може да бъде източник на данните, които да се обработват от IDOL. Представеният процес за обработка на неструктурирани данни е представен на *фиг. 2*.

Фиг. 2. Процес на обработване на неструктурирани данни

Към текущия момент е инсталирана виртуална машина, върху която е инсталиран IDOL сървърът и NiFi система, която обработва и разпределя данните към IDOL сървъра. Конфигурирани са IDOL Admin и IDOL Find. С инсталирането им се предоставя възможност за създаване на база данни за съхранение на извлечените данни, дефиниране на роли за достъп до IDOL сървъра, както и се създава възможност за използване на вградения ИИ за обработка на извлечените неструктурирани данни. Следва стъпка за вземането на решение какви данни са необходими, както и какъв да е техният източник *(фиг. 2)*. В предложената архитектура източник на данните са социалните мрежи. Затова е направено инсталиране на конекторите за тях – избрани са Twitter и Facebook. Освен това е инсталиран и конекторът на Hadoop към IDOL сървъра. За да могат данните да бъдат извлечени през конекторите, първо се създават приложения в средата за разработчици на приложения в двете социални медии. В системата NiFi се описват конекторите, като те се свързват към API на съответните приложения,

за да се осъществи извличането на желаното съдържание. Тези конектори предлагат възможност за филтриране на данните по ключови думи или фрази, които се съдържат в социалните медии. След като данните започнат да се събират, на следващата стъпка те се запазват под формата на файл, който да се импортира в IDOL и паралелно с това данните могат да се импортират и през конектора в Hadoop средата. Преобразуването на извлечените данни в подходящ формат се осъществява от компонента IDOL KeyView, който представлява решение за откриване на файлови формати, декриптиране на съдържание и извличане на текст. Позволените формати на импортираните данните са .idx и .xml. Данните се индексират в хранилището на IDOL с помощта на NiFi компонента PutIDOL. На следваща стъпка се очаква да бъдат извършени обработки и анализ на данните по допълнителни заявки за търсене, филтриране или сортиране на данните в IDOL Find, който използва изкуствен интелект при тази обработка.

Системата на Micro Focus IDOL съдържа в себе си разширена интелигентност. Това е подход, който използва ИИ и съкращава времето за получаване на прозрения и знания, които да служат за извеждане на нови открития и моделиране на зависимости, помагащи за вземане на интелигентни решения.

5. Заключение

В статията е разгледано предизвикателството пред обработката на неструктурирани данни и е предложен нов метод за решаване на този проблем. В процеса на решаването му може да се използват продуктите на Apache – Hadoop и Spark, като освен тях се предлага нов подход за обработка на данните с продукта IDOL. Той обработва и анализира данните с помощта на ИИ. Това се използва при извличането на неструктурирани данни от социални мрежи и такива в реално време. Предложената архитектура работи със системата NiFi и може да се използва съвместно със средата за големи данни Hadoop. Предложената архитектура вече се прилага, но няма резултати въпреки започналия процес по проверка на работата ѝ. Очаква се следващият етап да бъде верификация на предложената методология с конкретни данни.

Литература

1. Total Data Volume Worldwide 2010-2025. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/ (Accessed Oct. 09, 2022).

2. Statista Research Department. Topic: Big Data. Statista. https://www.statista.com/topics/1464/big-data/ (Accessed Oct. 08, 2022).

3. Coughlin, T. 175 Zettabytes by 2025. Forbes. https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/ (Accessed Oct. 09, 2022).

4. Eberendu, A. Unstructured Data: An Overview of the Data of Big Data. – *Int. J. Comput. Trends Technol*., 38, Aug. 2016, 46-50, doi: 10.14445/22312803/IJCTT-V38P109.

5. Duan, Y., J. S. Edwards, and Y. K. Dwivedi. Artificial Intelligence for Decision Making in the Era of Big Data – Evolution, Challenges and Research Agenda. – *Int. J. Inf. Manag.*, 48, Oct. 2019, 63-71. doi: 10.1016/j.ijinfomgt.2019.01.021.

6. Naeem, M. et al. Trends and Future Perspective Challenges in Big Data. Advances in Intelligent Data Analysis and Applications. Singapore, 2022, 309-325. doi: 10.1007/978-981-16-5036-9_30.

7. Боянов, Л. Дигиталният свят – промяната. ISBN: 978-619-239-637-4, София, Авангард Прима, 2021, 188.

8. B. D. Framework. Data Types: Structured vs. Unstructured Data – Enterprise Big Data. Enterprise Big Data Framework©, Jan. 09, 2019. https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/(Accessed Oct. 08, 2022).

9. Hanna, K. and C. Stedman. What is Unstructured Data? Search Business Analytics. https://www.techtarget.com/searchbusinessanalytics/definition/unstructured-data (Accessed Oct. 08, 2022).

10. MicroFocus. Saba Cloud Pro: IDOL01WBT – IDOL Essentials 12.6 Digital Learning for Administrators with Specialist Exam.

11. Kandukuri, G. Unstructured Data Analytics and the Hidden Value – Saxon AI. Saxon, Apr. 04, 2022. https://saxon.ai/blogs/how-to-tap-the-power-of-unstructured-data-in-2022-and-beyond/ (Accessed Oct. 08, 2022).

12. Techreviewer. The Most Popular Big Data Frameworks in 2022|Techreviewer Blog. https://techreviewer.co/blog/the-most-popular-big-data-frameworks-in-2022 (Accessed Oct. 09, 2022).

13. Inoubli, W., S. Aridhi, H. Mezni, M. Maddouri, and E. Mephu Nguifo. An Experimental Survey on Big Data Frameworks. – *Future Gener. Comput. Syst.*, 86, Sep. 2018, 546-564. doi: 10.1016/j.future.2018.04.032.

14. IBM. Hadoop vs. Spark: what's the Difference? Jul. 15, 2021. https://www.ibm.com/cloud/blog/hadoop-vs-spark (Accessed Oct. 09, 2022).

15. Jevtic, G. Hadoop vs Spark: Detailed Comparison of Big Data Frameworks. Knowledge Base by phoenixNAP, Jun. 04, 2020. https://phoenixnap.com/kb/hadoop-vs-spark (Accessed Oct. 09, 2022).

16. Regarding Unstructured Data Handling in Hadoop. Edureka Community. https://www.edureka.co/community/51491/regarding-unstructured-data-handling-in-hadoop (Accessed Oct. 09, 2022).

17. IDOL Connectors.pdf. Accessed: Oct. 09, 2022. [Online]. Available: https://www.microfocus.com/media/data-sheet/idol_connectors_ds.pdf.

18. Clarke, S. Micro Focus is Using Content Management and AI Data Analytics to Help Enterprises Manage New Work Practices. Micro Focus. https://content.microfocus.com/idol-analytics-21/idol-analyst-report (Accessed Oct. 09, 2022).

19. Lalitha, Y. S. A Spark Implementation on Hadoop System for Big Data Analytics on Acquantic dataset. 07, 2016, No. 02, 13.

За контакти:
Ивона Велкова
Редовен докторант към катедра
"Информационни технологии и комуникации"
Университет за национално и световно стопанство
ivonavelkova@unwe.bg