Wojewódzki Urząd Pracy w Lublinie

Ekspertyza: Możliwości wykorzystania big data w badaniach regionalnego rynku pracy, ze szczególnym uwzględnieniem ofert pracy

Czerwiec 2023 r.

Zamawiający:

Wojewódzki Urząd Pracy w Lublinie

ul. Obywatelska 4

20-092 Lublin

Wykonawca:

Warsaw Enterprise Institute

Marek Lachowicz
Katarzyna Rumiancew

Publikacja jest dystrybuowana bezpłatnie

Publikacja elektroniczna: WUP w Lublinie

Lublin 2023 r.

Spis treści

1. Wprowadzenie

Rynek pracy jest jednym z kluczowych obszarów badawczych, którego badanie ma istotne znaczenie dla podejmowania decyzji politycznych, gospodarczych i społecznych. Zrozumienie procesów, zachodzących na tym rynku jest niezbędne dla odpowiedniego kształtowania polityk publicznych w zakresie zatrudnienia, dostosowywania systemu edukacji do zmieniających się oczekiwań pracodawców czy planowania rekrutacji przez firmy.

Analiza rynku pracy umożliwia identyfikację trendów, wzorców i zjawisk, które mają wpływ na sytuację zawodową jednostek oraz ogólną kondycję gospodarki. Dostarcza informacji na temat popytu i podaży siły roboczej, struktury zatrudnienia, umiejętności i kompetencji poszukiwanych przez pracodawców, płac, warunków pracy, bezrobocia i innych wskaźników. Pozwala na ocenę potrzeb rynku w zakresie wykwalifikowanej siły roboczej, identyfikację luk kompetencyjnych oraz prognozowanie przyszłych zmian. Dzięki temu możliwe jest podejmowanie świadomych decyzji politycznych, takich jak dostosowywanie programów szkoleniowych i edukacyjnych do aktualnych potrzeb, wspieranie sektorów gospodarki charakteryzujących się wysokim popytem na pracowników czy rozwijanie strategii inwestycyjnych. Odpowiednia ocena stanu rynku pracy jest kluczowa dla formułowania polityk zatrudnienia, zarządzania bezrobociem, tworzenia nowych miejsc pracy oraz promowania inkluzji społecznej. Informacje na temat struktury i dynamiki rynku pracy pozwalają na lepsze zrozumienie sytuacji grup społecznych, takich jak młodzi ludzie, kobiety, osoby starsze czy osoby niepełnosprawne, i podejmowanie odpowiednich działań mających na celu zwiększenie ich aktywizacji zawodowej. Sytuacja na rynku pracy stanowi również informację dla przedsiębiorców i inwestorów, podejmujących decyzje dotyczące lokalizacji biznesu, planowania zasobów ludzkich, prognozowania potrzeb kadrowych czy oceny konkurencyjności rynku. Wiedza na temat struktury rynku pracy i trendów zawodowych pozwala firmom dostosowywać swoje strategie, rekrutować odpowiednich pracowników oraz tworzyć dla nich warunki i ścieżki rozwoju zawodowego.

Dzięki wykorzystaniu nowoczesnych technologii big data możliwe jest pogłębienie dotychczasowych analiz, pozyskanie nowych danych, a tym samym lepsze zrozumienie procesów zachodzących na rynku pracy. Przetwarzanie ogromnych zbiorów danych stwarza zupełnie nowe możliwości, pozwala na identyfikację trendów, wzorców i zjawisk, które wcześniej mogły pozostawać niewidoczne. Dzięki zaawansowanym algorytmom i narzędziom analitycznym można przetwarzać i analizować dane w czasie rzeczywistym, co umożliwia szybkie reagowanie na zmieniające się warunki rynkowe. Dzięki analityce big data można tworzyć precyzyjniejsze prognozy przyszłych zmian i trendów na rynku pracy, które stanowią istotną podstawę dla planowania strategicznego zarówno na poziomie krajowym, regionalnym, jak i lokalnym. Nowe rozwiązania stanowią zarówno szansę, jak

i wyzwanie. Nie są one wolne od ryzyk czy problemów, jednak ich umiejętne wykorzystanie może przynieść szereg korzyści.

Celem niniejszej ekspertyzy jest przedstawienie możliwości wykorzystania dużych zbiorów danych w badaniach regionalnego rynku pracy, ze szczególnym uwzględnieniem analizy ofert pracy. W tym celu zawiera ona omówienie koncepcji big data oraz zastosowania jej do działań administracji publicznej. Następnie zarysowana została kwestia analiz regionalnych rynków pracy z wykorzystaniem nowych narzędzi oraz szczegółowo przedstawiono możliwości związane z analizą ofert pracy. Dopełnieniem dokumentu jest wskazanie na aspekty prawne istotne z punktu widzenia big data.

2. Big data

Żyjemy w dobie rewolucji cyfrowej, zwanej również czwartą rewolucją przemysłową lub Industrie 4.0. To pojęcie po raz pierwszy pojawiło się w 2011 roku w niemieckiej strategii rozwoju wysokich technologii (high-tech) i zostało spopularyzowane przez Klausa Schwaba, założyciela i przewodniczącego Światowego Forum Ekonomicznego. Nawiązuje ono do rewolucji przemysłowych, które dzięki pojawiającym się innowacjom i wynalazkom, stanowiły proces zmian technologicznych, gospodarczych, społecznych i kulturowych. Czwarta rewolucja przemysłowa buduje na osiągnięciach poprzednich epok, ale wprowadza zupełnie nowy poziom zintegrowanych, inteligentnych technologii, a tym samym stwarza ogromne możliwości.

Pierwsza rewolucja zapoczątkowana w Anglii pod koniec XVIII wieku, wprowadziła świat w wiek pary. Jej głównym motorem napędowym było wprowadzenie mechanicznego krosna tkackiego, które całkowicie zmieniło przemysł włókienniczy. Wykorzystanie siły pary i wody jako źródła energii, zastępującego siłę mięśni ludzkich, otworzyło drogę do wynalezienia maszyny parowej na początku XIX wieku. To przełomowe odkrycie zapoczątkowało nowy etap rozwoju przemysłu.

Druga rewolucja przemysłowa, przypadająca na przełom XIX i XX wieku, wiązała się z wynalezieniem elektryczności i wprowadzenie linii montażowej przyczyniły się do dalszego dynamicznego rozwoju przemysłu. Dzięki temu procesy produkcyjne zostały usprawnione oraz pojawiło się szereg wynalazków, takich jak żarówka, telefon czy samochód, będących obecnie standardowymi produktami codziennego użytku.

Trzecia rewolucja przemysłowa miała miejsce w latach 70. XX wieku. Rozwój komputerów i wprowadzenie pierwszych programowalnych układów logicznych (PLC) odegrały kluczową rolę w automatyzacji procesów produkcyjnych. To właśnie wtedy zaczęto wykorzystywać komputery do sterowania maszynami i całymi fabrykami. Dzięki tym osiągnięciom zautomatyzowano produkcję, która stała się efektywniejsza, a szereg procesów może być wykonywanych bez udziału człowieka, jedynie pod jego nadzorem.

Obecnie jesteśmy świadkami czwartej rewolucji przemysłowej, która buduje na fundamentach poprzedniej i wykorzystuje pełnię potencjału nowoczesnych technologii. Jest ona oparta na rozwinięciu automatyzacji i cyfryzacji, ale idzie znacznie dalej niż poprzednia. Wprowadza ona pojęcie "inteligentnych" technologii, które stopniowo zacierają granicę między człowiekiem a maszyną, co doskonale obrazują chociażby możliwości narzędzi generatywnej sztucznej inteligencji.

Czwarta rewolucja przemysłowa przynosi za sobą wiele innowacji, takich jak internet rzeczy (IoT), sztuczna inteligencja (AI), robotyka, big data, łączność 5G i wiele innych. Te technologie rewolucjonizują różne sektory, w tym przemysł, usługi,

transport, ochronę zdrowie i wiele innych. Przemysł 4.0 dąży do stworzenia inteligentnych fabryk, w których maszyny, urządzenia i systemy są w stanie komunikować się ze sobą, analizować dane w czasie rzeczywistym i podejmować autonomiczne decyzje.

Rozwój technologii pozwala na lepszą analizę świata, zarówno w zakresie pozyskiwania danych, jak i ich przetwarzania, dzięki czemu możliwa jest automatyzacja procesów, przyspieszenie pracy i zwiększenie wydajności. Kluczową wartością, oprócz zaawansowanych algorytmów, które stanowią jedynie narzędzie, są informacje, zarówno jako zmienna wejściowa, jak i wytworzony materiał. Precyzyjność i wiarygodność wyników jest pochodną posiadanych zbiorów danych. Dlatego też tak istotną kwestią jest ich odpowiednie gromadzenie i porządkowanie.

Celem niniejszego rozdziału jest omówienie koncepcji big data, uwzględniając nie tylko kwestię samych danych, ale również ich wykorzystanie. Wskazane zostaną narzędzia służące do ich przetwarzania oraz związane z nimi możliwości. Zwrócona zostanie również uwaga na nowe źródła danych oraz zastosowanie big data w administracji publicznej.

2.1. Definicja big data

Głównym walorem obecnie dostępnych technologii jest możliwość przetwarzania dużych zbiorów danych (ang. big data). Termin ten, we współczesnym rozumieniu, pojawił się w latach 90. XX wieku i zazwyczaj definiowany jest jako duże, różnorodne i zmienne zbiory danych powstające za sprawą nowoczesnych urządzeń telekomunikacyjnych, które są przechowywane, przetwarzane i analizowane za pomocą zaawansowanej technologii informatycznej. Pierwotnie wskazywano na trzy cechy big data (model 3V), obejmujące volume (objętość danych), velocity (prędkość napływania nowych danych i ich analizy), variety (różnorodność danych). Następnie koncepcja ta była uzupełniana przez dodawanie kolejnych charakterystyk wskazanych w tabeli 1.

Tabela 1. Charakterystyki big data.

Charakterystyka	Opis
Volume (objętość)	Volume odnosi się do ogromnej ilości danych generowanych i gromadzonych przez różne źródła. Tradycyjne systemy informatyczne często mają trudności z przechowywaniem i przetwarzaniem tak dużych ilości danych, dlatego big data wymaga nowych technologii i narzędzi.

Charakterystyka	Opis
Velocity (prędkość)	Velocity odnosi się do tempa, z jakim dane są generowane i przesyłane. W erze big data dane często napływają w czasie rzeczywistym lub w bardzo krótkich odstępach, na przykład z czujników internetu rzeczy (IoT), mediów społecznościowych czy transakcji online. Systemy big data muszą być w stanie szybko przetwarzać i analizować dane w takim tempie.
Variety (różnorodność)	Variety odnosi się do różnych formatów danych, które są dostępne w środowisku big data. Mogą to być strukturalne dane (np. relacyjne bazy danych), półstrukturalne dane (np. pliki XML) lub nieustrukturalne dane (np. multimedia, treści z mediów społecznościowych). Systemy big data muszą radzić sobie z różnorodnością tych danych i umożliwiać integrację i analizę danych o różnych formatach.
Veracity (prawdziwość)	Veracity odnosi się do jakości danych, w tym do niepewności, niekompletności i błędów. W przypadku big data, ze względu na różnorodne źródła danych, istnieje ryzyko obecności niedokładnych, niekompletnych lub niezgodnych danych. Ważne jest, aby systemy big data uwzględniały te czynniki i miały mechanizmy do oceny i zarządzania jakością danych.
Value (wartość)	Value odnosi się do zdolności do wyciągania wartościowych informacji i wiedzy z dużych zbiorów danych. Głównym celem big data jest wykorzystanie analizy danych, eksploracji i odkrywania wzorców, aby generować wartość biznesową, zrozumienie klientów, optymalizację procesów, podejmowanie lepszych decyzji itp.
Viability (wykonalność)	Viability odnosi się do oceny technologicznej, ekonomicznej i organizacyjnej, czyli określenia, czy dany projekt big data jest wykonalny i opłacalny. Ważne jest uwzględnienie czynników takich jak koszty infrastruktury, zasoby ludzkie, technologia i możliwości organizacyjne w kontekście big data.
Validity (poprawność)	Validity odnosi się do sprawdzenia i potwierdzenia, czy dane są dokładne, prawdziwe i odpowiednie do danego celu. Dane, które są niepoprawne lub nieodpowiednie, mogą prowadzić do błędnych wniosków i podejmowania złych decyzji. Dlatego ważne jest, aby przeprowadzać procesy walidacji danych w celu zapewnienia ich jakości i poprawności.

Charakterystyka	Opis
Volatility (ulotność)	Volatility odnosi się do szybkości zmian w danych i ich krótkotrwałej użyteczności. W niektórych dziedzinach, takich jak analiza rynku finansowego, dane mogą mieć krótki okres ważności. Konieczne jest monitorowanie i przetwarzanie tych danych w czasie rzeczywistym, aby wykorzystać ich pełny potencjał przed utratą ich wartości.
Vagueness (nieprecyzyjność)	Vagueness odnosi się do braku precyzji lub jednoznaczności w danych. Często w big data można napotkać dane o niepełnych informacjach, niewielkiej jakości lub o niejasnym znaczeniu.
Virility (wiralność)	Virility odnosi się do zdolności rozprzestrzeniania się informacji w sieci i wpływania na inne dane. Big data niejako tworzy się same – przetwarzanie danych wpływa na pojawienie się kolejnych.
Vendible (sprzedawalność)	Vendible odnosi się do możliwości sprzedaży danych lub uzyskania z nich wartości. Wiele danych (również tradycyjnie uznawanych za „nierynkowe”) może mieć wartość rynkową i być sprzedawane innym podmiotom.
Voracity (żarłoczność)	Voracity odnosi się do intensywności przetwarzania danych i zasobów potrzebnych do efektywnego przetwarzania dużych zbiorów danych. Big data wymaga wydajnych systemów i infrastruktury, aby sprostać wymaganiom dotyczącym przetwarzania.
Vanity (próżność)	Vanity odnosi się do tendencji do gromadzenia dużych ilości danych bez konkretnego celu lub wartości. Warto zwrócić uwagę, aby zbierane dane miały praktyczne zastosowanie i przyczyniały się do osiągania określonych celów.
Variability (zmiennność)	Variability odnosi się do zmienności i nieregularności danych w czasie. W przypadku niektórych zbiorów danych, wartości mogą zmieniać się w sposób nieprzewidywalny lub mogą występować nagłe skoki, co może wpływać na analizę i przetwarzanie danych. Zrozumienie zmienności danych jest istotne w kontekście big data.

Charakterystyka	Opis
Visualization (wizualizacja)	Visualization odnosi się do prezentacji danych w sposób łatwy do zrozumienia za pomocą różnych technik graficznych. Wielkie zbiory danych mogą być trudne do zrozumienia w formie surowych liczb i faktów. Wizualizacja danych pozwala na łatwiejsze zidentyfikowanie wzorców, trendów i zależności, co ułatwia podejmowanie decyzji.
Venue (miejsce)	Venue odnosi się do lokalizacji danych, ich pochodzenia i przechowywania. Informacje mogą pochodzić z różnych źródeł, takich jak bazy danych, systemy internetowe, urządzenia IoT itp.
Vocabulary (słownictwo)	Vocabulary odnosi się do konieczności ustalenia jednolitego i spójnego sposobu opisywania danych oraz zrozumienia znaczenia terminów i terminologii używanej w kontekście big data. Wielkie zbiory danych często zawierają różne typy danych z różnych źródeł, co może prowadzić do problemów związanych z interpretacją i jednoznacznością terminologii.
Visibility (widoczność)	Visibility odnosi się do dostępności danych dla poszczególnych odbiorców. W szczególności w ramach organizacji, dostęp do danych powinny mieć wszystkie osoby, którym są one niezbędne w pracy niezależnie od stanowiska czy działu, w którym pracują. Jest to zgodne z często podnoszonym postulatem „uwolnienia” danych, czyli dzielenia się informacjami posiadanymi przez poszczególne podmioty.
Vitality (witalność)	Vitality odnosi się do wartościowania danych i umiejętności ich prawidłowego prioretyzowania. Informacje istotniejsze, wartościowsze z punktu widzenia projektu czy organizacji powinny mieć wyższy priorytet i na nich należy się skupić w pierwszej kolejności.
Vincularity (łączność)	Vincularity dotyczy zależności i powiązań między różnymi danymi w kontekście big data. Informacje stanowią połączoną sieć i w wielu przypadkach występują między nimi zależności.

Charakterystyka	Opis
Vulnerability (podatność)	Vulnerability odnosi się do zagrożeń i ryzyka związanego z bezpieczeństwem danych w środowisku big data. Duże zbiory danych mogą być podatne na różne rodzaje ataków, takie jak kradzież danych, naruszenia poufności czy manipulacja informacjami. W kontekście big data istnieje potrzeba zabezpieczania danych i infrastruktury przed zagrożeniami oraz zapewnienia prywatności i integralności informacji.
Verification (weryfikacja)	Verification odnosi się do procesu sprawdzenia i standaryzacji danych, w celu potwierdzania ich wiarygodności.
Virality (wirusowość)	Virality odnosi się do szybkości rozprzestrzeniania się danych, które znajdują się w przestrzeni publicznej.
Valor (waleczność)	Valor odnosi się do koniecznej postawy użytkowników. Big data to nie tylko ogromne możliwości, ale również szereg wyzwań, którym trzeba stawić czoła.
Verbosity (gadatliwość)	Verbosity odnosi się do nadmiarowych treści, znajdujących się w danych. Zwłaszcza w informacjach pierwotnych przed przetworzeniem większość ich zawartości może być nieprzydatna dla użytkownika. Dla zapewnienia wydajności i jakości analiz big data konieczne jest odpowiednie czyszczenie danych.
Versality (wszechstronność)	Versality odnosi się do możliwości wykorzystania danych w różnych sytuacjach i przez różne grupy interesariuszy, mimo że zostały one stworzone w określonym, innym celu. Niezbędne są do tego wiedza o pochodzeniu, znaczeniu, jakości i kontekście informacji.
Viscosity (lepkość)	Lepkość odnosi się do trudności w przepływie danych między różnymi systemami, platformami lub procesami. Często w środowisku big data występują trudności z integracją danych z różnych źródeł, co utrudnia płynność przepływu danych i wymianę informacji między nimi.

Źródło: opracowanie własne na podstawie S. Dhamodharavadhani, R. Gowri, R. Rathipriya, Unlock different V's of big data for analytics, International Journal of Computer Sciences and Engineering, 6.4, 2018, s. 183-190.

Należy przy tym zaznaczyć, że kolejne cechy odnoszą się nie tylko do opisu atrybutów big data, ale również do zagadnień związanych z ich przetwarzaniem. Kluczowe bowiem nie jest samo występowanie dużych zbiorów danych, ale możliwość ich wykorzystania tych danych w celu pozyskania wartościowych

informacji. Nieodzowną częścią konceptu są nowoczesne technologie, które pozwoliły na ich techniczne gromadzenie i przetwarzanie. Istotą czwartej rewolucji przemysłowej było stworzenie narzędzi – nowością nie są same duże zbiory danych, ale zdolność do ich wykorzystania.

Dlatego też pod pojęciem big data należy rozumieć nie tylko duże wolumeny danych, ale również metody ich analizy oraz wykorzystywane do tego technologie i narzędzia. Podstawowe z nich to uczenie maszynowe (machine learning) oraz eksploracja danych (data mining). W analizach wykorzystywane są algorytmy bazujące m.in. na drzewach decyzyjnych, modelu Markowa, modelach regresji logistycznej, czy modelowaniu bayesowskim. Warto również wspomnieć o takich narzędziach jak Hadoop, Spark, Hive czy Pig, które umożliwiają równoległe przetwarzanie danych na klastrach komputerowych. Ponadto, rozwój technologii baz danych, takich jak NoSQL czy NewSQL, umożliwił skuteczne zarządzanie i przetwarzanie danych w czasie rzeczywistym.

Jedną z głównych korzyści wynikających z wykorzystania big data jest możliwość wyciągania cennych wniosków i informacji z ogromnych ilości danych, które wcześniej były niedostępne lub trudne do zauważenia. Analiza dużych zbiorów danych pozwala na wykrywanie ukrytych wzorców, trendów rynkowych, zachowań klientów czy predykcję przyszłych zdarzeń. Dzięki temu możliwe jest podejmowanie lepszych decyzji, optymalizowanie procesów, zwiększanie efektywności czy poprawa konkurencyjności.

Niemniej jednak, przetwarzanie big data wiąże się również z pewnymi wyzwaniami i problemami. Jednym z nich jest odpowiednie zarządzanie i przechowywanie ogromnych ilości danych. Konieczna jest odpowiednia infrastruktura i procedury, aby zapewnić skalowalność, niezawodność czy bezpieczeństwo danych, co może również wiązać się z istotnymi kosztami.

2.2. Analityka big data

Duże zbiory danych generują nowe wyzwania wynikające z ich charakterystyki. Ogromna ilość, złożoność, różnorodność oraz wielość źródeł informacji powoduje, że tradycyjne formy oprogramowania mogą nie być wystarczające do zbierania i przetwarzania big data. Dane są często generowane z dużą prędkością i posiadają zróżnicowaną formę – od ustrukturyzowanych (tabele, arkusze), przez częściowo ustrukturyzowane (pliki XML, strony internetowe), po nieustrukturyzowane (obrazy, pliki audio i video). W związku z tym konieczne jest zastosowanie nowych technologii dedykowanych dużym zbiorom. Kluczową kwestią są narzędzia do zbierania i przechowywania danych oraz ich przetwarzania. Bowiem posiadając już dane w ustrukturyzowanej formie (a do tego prowadzą wskazane procesy) można zastosować tradycyjne pakiety statystyczne i ekonometryczne pozwalające na przeprowadzanie analiz i wnioskowania.

Kluczową koncepcją w zakresie gromadzenia danych jest data lake. Pojęcie odnosi się do repozytoriów, gdzie informacje są przechowywane w swojej pierwotnej, również nieustrukturyzowanej formie w klastrach wielu serwerów. W przeciwieństwie do tradycyjnych magazynów danych, magazyny data lake nie wymagają przetworzenia danych do określonego formatu, dzięki czemu doskonale sprawdzają się w gromadzeniu big data. W tym kontekście warto wskazać na zalety tego rozwiązania, obejmujące[1]:

- znacznie mniejszy koszt procesowania i przechowywania danych;
- możliwość procesowania danych na wielu komputerach jednocześnie;
- możliwość przechowywania całego zbioru danych we wszystkich typach (bez konieczności dostosowywania do schematu baz danych);
- łatwą przyswajalność nowych typów danych;
- możliwość szybkiego wglądu do całości danych bez konieczności wyszukiwania;
- podzbiorów danych w poszczególnych izolowanych silosach.

Kolejnym istotnym rozwiązaniem jest przechowywanie i przetwarzanie danych w chmurze obliczeniowej – cloud computing. Jest to model oparty na użytkowaniu usług dostarczanych przez zewnętrznych dostawców i odciążenie komputerów odbiorcy. Ciężar świadczenia usług informatycznych, obejmujących przechowywanie danych, posiadanego oprogramowanie czy mocy obliczeniowej, przenoszony jest na serwer dostawcy, a odbiorca korzysta z nich za pośrednictwem internetu. Dostęp jest możliwy z wielu różnych komputerów, zatem zapewnia możliwość jednoczesnej pracy przez kilku użytkowników. Dzięki temu nie trzeba posiadać w swoich zasobach serwerów czy oprogramowania, co jest szczególnie istotne przy ograniczeniu kosztów. Większość dostępnych rozwiązań w zakresie big data wykorzystuje ten model działania.

Korzystając z nowoczesnych technologii stworzono szereg infrastruktur informatycznych z myślą o big data. Wśród najważniejszych należy wskazać[2]:

- NoSQL;
- Map Reduce;
- Hadoop.

[1] R. Ładysz, M. Kiedrowicz, M Bliźniuk G, Przetwarzanie i analiza zasobów typu Big Data w przeciwdziałaniu przestępstwom finansowym [w:] M. Kiedrowicz, Metody i narzędzia informatyczne wykorzystywane w zwalczaniu przestępstw finansowych, Redakcja Wydawnictw WAT, 2019.

[2] K. Racka, Big Data – znaczenie, zastosowania i rozwiązania technologiczne, Nauki Ekonomiczne, t. XXIII, 2016, s. 311-323.

1. NoSQL

Bazy NoSQL powstały w opozycji do baz SQL, stąd też ich nazwa oznaczająca „nonSQL” lub „not only SQL”. Tradycyjne bazy są relacyjne (tabelaryczne), natomiast bazy NoSQL nie wymagają, aby przechowywane w nich typy danych były zgodne ze stałym schematem lub strukturą. Pozwala to na obsługę różnych modeli i formatów, dzięki czemu ma zastosowanie do danych pierwotnych i nieustrukturyzowanych. Istotną cechą baz NoSQL jest ich elastyczność, która pozwoliła na zwiększenie szybkości i lepszą skalowalność, co oznacza, że mogą one efektywnie obsługiwać duże ilości danych i dynamicznie rosnące zasoby. Ten rodzaj baz danych oferuje również wysoką dostępność i odporność na awarie. Często działają w środowiskach rozproszonych, w których dane są replikowane i partycjonowane, aby zapewnić wyższą wydajność i niezawodność. Należy jednak zaznaczyć, że duża dostępność danych jest uzyskiwana kosztem spójności danych, dlatego bazy NoSQL nie powinny być stosowane, jeżeli wymagana jest wysoka dokładność danych. Dodatkowo, problemem mogą być ograniczenia związane z brakiem jasno sformalizowanego języka zapytań czy intuicyjnych, standardowych interfejsów.

Podstawowe typy baz danych NoSQL zostały przedstawione w tabeli 2.

Tabela 2. Rodzaje baz danych NoSQL.

Rodzaj bazy	Charakterystyka
Bazy dokumentowe (document store)	Bazy dokumentowe są oparte na modelu danych dokumentowych, gdzie dane są przechowywane w postaci dokumentów. Każdy dokument jest zazwyczaj reprezentowany w formacie JSON lub BSON i może mieć zróżnicowaną strukturę. Bazy dokumentowe są elastyczne, ponieważ nie wymagają ściśle określonego schematu, co pozwala na łatwe dodawanie, modyfikowanie i usuwanie pól wewnątrz dokumentów. Są one przydatne w przypadku, gdy dane są heterogeniczne lub ich struktura może się zmieniać.
Bazy klucz-wartość (key-value store)	Bazy klucz-wartość są najprostszym modelem baz danych, gdzie dane są przechowywane jako pary klucz-wartość. Każdy rekord w bazie jest identyfikowany unikalnym kluczem, a wartość może być dowolnym obiektem lub danymi binarnymi. Bazy klucz-wartość są bardzo szybkie i skalowalne, ponieważ dostęp do danych odbywa się bezpośrednio za pomocą klucza. Są one często

Rodzaj bazy	Charakterystyka
	wykorzystywane do przechowywania informacji o pamięci podręcznej, sesji użytkownika lub prostych operacji odczytu/zapisu.
Bazy kolumnowe (columnoriented stores)	Bazy kolumnowe są zoptymalizowane pod kątem efektywnego przechowywania i przetwarzania dużych zbiorów danych. Dane są przechowywane w postaci kolumn, a nie w tradycyjnym wierszowym układzie. Każda kolumna zawiera wartości dla jednej konkretnej cechy lub atrybutu, co umożliwia efektywne wykonywanie operacji agregacji, analizy i filtrowania danych. Bazy kolumnowe są szczególnie przydatne w przypadku zapytań, które wymagają analizy szerokiego zakresu danych.
Bazy grafowe (graphdatabase)	Bazy grafowe skupiają się na przechowywaniu i przetwarzaniu relacji między danymi. Dane są reprezentowane jako węzły (node) i krawędzie (edge) grafu, gdzie węzły reprezentują obiekty, a krawędzie reprezentują relacje między nimi. Bazy grafowe są skoncentrowane na efektywnym wyszukiwaniu i analizie tych relacji, co czyni je idealnym narzędziem do analizy sieci społecznych, rekomendacji, trasowania w sieciach drogowych i innych zastosowań, gdzie relacje między danymi są kluczowe.
Bazy obiektowe (object database)	Bazy obiektowe są zaprojektowane do przechowywania i manipulowania obiektami, które są bardziej zaawansowaną strukturą danych niż tradycyjne wiersze i kolumny. Bazy obiektowe umożliwiają bezpośrednie przechowywanie obiektów programowych w bazie danych, co ułatwia mapowanie obiektowo-relacyjne. Pozwalają na złożone zapytania i dziedziczenie obiektowe, co jest przydatne w przypadku aplikacji, które silnie opierają się na obiektach i wymagają kompleksowych relacji między nimi.

Źródło: opracowanie własne na podstawie K. Racka, Big Data – znaczenie, zastosowania i rozwiązania technologiczne, Nauki Ekonomiczne, t. XXIII, 2016, s. 311-323.

2. MapReduce

MapReduce to model programowania, opracowany przez Google w celu efektywnego procesowania dużych zbiorów danych równolegle na rozproszonych systemach. Pozwala to na przetwarzanie danych w miejscu ich przechowywania, a tym samym nie wymaga przesyłania informacji z komputerów magazynujących dane do serwerów. Proces składa się z dwóch głównych kroków: mapowania (mapping) i redukcji (reducing). Krok mapowania polega na przekształceniu wejściowych danych na pary klucz-wartość. Funkcja mapowania jest stosowana do każdego rekordu wejściowego niezależnie, a wynikiem jest zestaw par klucz-wartość. Klucz identyfikuje grupę rekordów, a wartość reprezentuje rekordy w tej grupie. Krok redukcji następuje po kroku mapowania i polega na agregacji rekordów na podstawie ich kluczy. W tym kroku, rekordy o tych samych kluczach są grupowane razem, a funkcja redukcji jest stosowana do każdej grupy, generując wynik końcowy.

MapReduce wykorzystuje mechanizm automatycznego podziału pracy na wiele węzłów przetwarzających. Dane wejściowe są podzielone na fragmenty, które są przetwarzane niezależnie przez wiele węzłów. Każdy węzeł wykonuje funkcje mapowania na swoim fragmencie danych, generując lokalne pary klucz-wartość. Następnie, wyniki z wszystkich węzłów są sortowane i przekazywane do funkcji redukcji. Ostateczny wynik jest generowany przez funkcję redukcji i zwracany jako wynik końcowy.

Model jest szczególnie przydatny do przetwarzania dużych zbiorów danych, które nie mieszczą się w pamięci jednego komputera. Dzięki rozproszeniu obliczeń na wiele węzłów, zapewnia skalowalność i wydajność przetwarzania. Ponadto, MapReduce automatycznie obsługuje odporność na awarie, ponieważ w przypadku awarii jednego węzła, zadanie jest automatycznie przydzielane innemu węzłowi.

Model MapReduce jest szeroko stosowany w systemach przetwarzania danych takich jak Apache Hadoop i jest wykorzystywany do analizy big data, generowania raportów, indeksowania danych czy przetwarzania logów systemowych.

3. Hadoop

Apache Hadoop jest jedną z pierwszych platform programistycznych stworzoną na potrzeby analizy big data przez Douga Cuttinga oraz Mike’a Cafarellę. Jest to ekosystem typu open source przeznaczony do przechowywania i przetwarza danych w rozproszonym środowisku obliczeniowym przy pomocy klastra komputerów. Platforma jest wysoce

skalowalna, co zapewnia elastyczność i ekonomiczność zarządzania informacjami.

Hadoop składa się z szeregu modułów (Common, HDFS, YARN, MapReduce) oraz projektów powiązanych (m.in. Ozone, Hive, Pig, Spark, HBase) zapewniających różne funkcjonalności. Wśród kluczowych możliwości należy wskazać klastrowanie danych i ich niezależne przetwarzanie na różnych węzłach klastra w czasie rzeczywistym, zapewnienie automatycznego podziału pracy, monitorowanie jej postępów, generowania raportów czy wykorzystania narzędzi uczenia maszynowego.

Hadoop jest wykorzystywany w wielu dziedzinach, takich jak analiza big data, eksploracja danych, przetwarzanie logów, indeksowanie i wyszukiwanie, rekomendacje, uczenie maszynowe i wiele innych. Jego elastyczność, skalowalność i zdolność do obsługi dużych zbiorów danych sprawiają, że Hadoop jest popularnym narzędziem dla organizacji, które muszą przetwarzać i analizować dane na dużą skalę.

2.3. Nowe źródła danych

Rozwój nowoczesnych narzędzi i technik analitycznych oraz sukcesywnie zwiększające się możliwości technologiczne pozwalają na wykorzystanie dużych zbiorów danych. Nie byłoby to możliwe gdyby nie zachodziły zmiany takie jak chociażby miniaturyzacja sprzętu i wzrost mocy obliczeniowej, opracowywanie bardziej wydajnych algorytmów i oprogramowania, rozwój usług (w tym np. sieciowych) czy wzrost zastosowań do nowych obszarów.

Co istotne, prowadzi to nie tylko do umożliwienia analizy już istniejących zbiorów danych, zgromadzonych w bazach czy rejestrach, ale tworzy również nowe możliwości w pozyskiwaniu i tworzeniu takich zbiorów. Powstają nowe, alternatywne źródła informacji, niedostępne w przeszłości, wśród których można wskazać dane pochodzące z sieci, w tym stron internetowych i serwisów społecznościowych czy dane pochodzące z różnego rodzaju czujników, kamer czy sieci komórkowych.

Global Working Group on Big Data for Official Statistics działająca w ramach Organizacji Narodów Zjednoczonych stworzyła klasyfikację typów big data[3], wyróżniając:

- sieci społecznościowe (informacje generowane przez ludzi) – informacje będące zapisem ludzkich doświadczeń, wcześniej rejestrowanymi w książkach i dziełach sztuki, a później w fotografiach, nagraniach audio i wideo. Są one

[3] Classification of Types of Big Data, Extract from UNECE website, United Nations Department Of Economic And Social Affairs Statistics Division, Meeting of the Expert Group on International Statistical Classifications New York, 19-22 May 2015.

obecnie prawie całkowicie zdigitalizowane i przechowywane w różnych miejscach, od komputerów osobistych po sieci społecznościowe. Dane te mają luźną strukturę, często nie są zarządzane i nie podlegają konkretnym regulacjom. Obejmują one:
    - sieci społecznościowe (np. Facebook, Twitter, Tumblr);
    - blogi i komentarze;
    - dokumenty osobiste;
    - zdjęcia (np. Instagram, Flickr, Picasa);
    - filmy (np. Youtube);
    - wyszukiwanie w internecie;
    - zawartość danych mobilnych – wiadomości tekstowe;
    - mapy generowane przez użytkowników;
    - poczta elektroniczna – wiadomości e-mail.
- tradycyjne systemy biznesowe (dane pośredniczone przez procesy) – informacje pochodzące z procesów, rejestrujących i monitorujących określone zdarzenia biznesowe, takie jak rejestracja klienta, wytwarzanie produktu, przyjmowanie zamówienia itp. Gromadzone w ten sposób dane są wysoce ustrukturyzowane i obejmują transakcje, tabele referencyjne i relacje, a także metadane określające ich kontekst. Tradycyjne dane biznesowe stanowią zdecydowaną większość zasobów, które są zarządzane i które przetwarzane komputerowo, zarówno w systemach operacyjnych, jak i systemach Business Intelligance. Dane te zazwyczaj są ustrukturyzowane i przechowywane w systemach relacyjnych baz danych. Niektóre źródła należące do tej klasy stanowią dane administracyjne. Obejmują one:
    - dane opracowane przez agencje publiczne:
        - dokumentację medyczną;
    - dane wytwarzane przez przedsiębiorstwa:
        - transakcje handlowe;
        - ewidencja bankowa lub inwentaryzacyjna;
        - handel elektroniczny;
        - karty kredytowe.
- internet rzeczy (dane generowane maszynowo) – informacje pochodzące z czujników i maszyn używanych do mierzenia i rejestrowania zdarzeń i sytuacji w świecie fizycznym. Dane wyjściowe tych czujników to dane generowane maszynowo, które są dobrze ustrukturyzowane, od prostych zapisów czujników po złożone dzienniki komputerowe. W miarę rozprzestrzeniania się czujników i wzrostu ilości danych staje się to coraz ważniejszym elementem informacji przechowywanych i przetwarzanych przez wiele firm. Dzięki dobrze ustrukturyzowanemu charakterowi, dane te nadają się do przetwarzania komputerowego, ale ich rozmiar i szybkość pozyskiwania wykraczają poza tradycyjne podejście. Obejmują one:
    - dane z czujników:
        - czujniki stacjonarne:

                - automatyka domowa;
                - czujniki pogody i zanieczyszczeń;
                - czujniki ruchu, kamery internetowe;
                - czujniki naukowe;
                - filmy i obrazy z monitoringu;
            - czujniki mobilne (śledzące):
                - lokalizacja telefonu komórkowego;
                - samochody;
                - zdjęcia satelitarne;
    - dane z systemów komputerowych:
        - dzienniki (logs);
        - dzienniki sieciowe (web logs).

2.4. Wykorzystanie big data w administracji publicznej

Rozwój analizy dużych zbiorów danych miał przede wszystkim charakter komercyjny. Pierwotnie big data było wykorzystywane głównie przez duże korporacje do optymalizowania działalności biznesowej. Jednak obecnie, w związku z dużo większą dostępnością, znajduje zastosowanie zarówno w sektorze prywatnym, jak i publicznym. Chociaż należy zaznaczyć, że nadal w wyniku ograniczeń finansowych czy organizacyjnych zdecydowanie większe stosowanie nowoczesnych narzędzi i technologii obserwowane jest w działalności prywatnych podmiotów. Wykorzystanie big data w sektorze publicznym jest zjawiskiem stosunkowo nowym. Generuje to szereg problemów, takich jak brak przygotowania, zasad czy wytycznych. W związku ze swoim charakterem urzędy czy instytucje publiczne działają na podstawie ustalonych procedur, a proces zmiany jest dużo trudniejszy i bardziej sformalizowany niż w prywatnej firmie. Administracja publiczna zbiera duże ilości danych w celu zarządzania miastami, dostarczania usług publicznych i podejmowania decyzji politycznych. Dane te mogą obejmować informacje o mieszkańcach, infrastrukturze miejskiej, transporcie zbiorowym, zdrowiu publicznym, edukacji, bezpieczeństwie państwowym oraz wielu innych obszarach. Przykładowe dane mają związek z ewidencją ludności, opodatkowaniem osób fizycznych i firm, monitoringiem miejskim czy chociażby zmianami klimatycznymi.

Należy zaznaczyć, że sektor publiczny ma ogromny potencjał w postaci posiadanych zasobów. Funkcjonowanie państwa i jego struktur prowadzi do masowego zbierania danych o obywatelach i ich aktywności. Zapewnienie podstawowych potrzeb związanych z bezpieczeństwem, edukacją, ochroną zdrowia, zabezpieczeniem społecznym czy prawidłowością obrotu gospodarczego wiąże się chociażby z tworzeniem rejestrów czy gromadzeniem informacji w sposób umożliwiający skuteczną realizację tych zadań. Istnieje szereg agend państwa odpowiedzialnych za monitorowanie istotnych procesów i zjawisk zarówno społeczno-gospodarczych, ale również przyrodniczych czy kulturalnych. Podstawowym problemem w sektorze publicznym nie jest brak informacji – można nawet stwierdzić, że skala gromadzenia

danych okazuje się znacznie większa niż w przypadku podmiotów prywatnych. Sedno sprawy dotyczy tego, że posiadane zasoby nie są efektywnie wykorzystywane. Wynika to nie tylko z braku odpowiednich narzędzi, ale również innych czynników. Dane gromadzone przez poszczególne jednostki często są wykorzystywane jedynie w ich działalności, brakuje więc wymiany informacji między współpracującymi instytucjami oraz holistycznego podejścia. Nawet w ramach poszczególnych obszarów, np. ochrony zdrowia, występują braki powiązań między różnymi rejestrami i bazami, które zazwyczaj tworzone są na potrzeby konkretnych działań.

Oprócz danych zbieranych w ramach codziennego funkcjonowania państwa, administracja posiada wyspecjalizowane instytucje w zakresie statystyki publicznej. Obszar ten reguluje ustawa z 29 czerwca 1995 r. o statystyce publicznej[4], która określa zasady i tworzy podstawy rzetelnego, obiektywnego, profesjonalnego i niezależnego prowadzenia badań statystycznych, których wyniki mają charakter oficjalnych danych statystycznych, oraz ustala organizację i tryb prowadzenia tych badań i zakres związanych z nimi obowiązków. Służby statystyki publicznej obejmują nie tylko Główny Urząd Statystyczny (GUS) i jeden urząd statystyczny w każdym województwie, ale również: Centrum Informatyki Statystycznej (CIS), Zakład Wydawnictw Statystycznych (ZWS), Centrum Badań i Edukacji Statystycznej (CBiES GUS) oraz Centralną Bibliotekę Statystyczną im. Stefana Szulca.

Główny Urząd Statystyczny udostępnia narzędzia do statystyki publicznej. W ramach Platformy Analitycznej SWAiD (System Wspomagania Analiz i Decyzji) uruchomiono 23 Dziedzinowe Bazy Wiedzy, które zostały podzielone na cztery obszary: Statystykę wielodziedzinową (w której znalazły się bazy o charakterze przekrojowym: Rachunki Narodowe, Nauka, Technika i Społeczeństwo Informacyjne; Przekroje Terytorialne Bank Danych Lokalnych wraz z Atlasem Regionów) oraz bazy tematyczne zaliczone do obszarów: Gospodarka, Społeczeństwo i Środowisko. Wymienione zbiory umożliwiają korzystanie z różnych form wizualizacji danych, takich jak tabele, wykresy i mapy. Dodatkowo użytkownicy mają możliwość pobierania wygenerowanych zestawień i wykorzystywania ich do własnych opracowań i analiz. Dziedzinowe Bazy Wiedzy są również źródłem bogatej wiedzy dodatkowej, która obejmuje metodyczne wyjaśnienia, zbiory linków do rekomendowanych publikacji, opracowań i konferencji oraz interesujących faktów związanych z daną dziedziną.

Inne z wymienionych zasobów, które opracował i udostępnił GUS, to m.in. TERYT (Krajowy Rejestr Urzędowy Podziału Terytorialnego Kraju). Dane z tego rejestru są bezpłatnie przekazywane podmiotom administracji publicznej oraz podmiotom komercyjnym. Dostępne jest pięć katalogów: identyfikatory i nazwy jednostek podziału terytorialnego kraju (TERC), identyfikatory i nazwy miejscowości (SIMC), Centralny Katalog Ulic (ULIC), zbiór określeń rodzajów miejscowości WMRODZ oraz

[4] Dz.U. 2023 poz. 773.

Nomenklatura Jednostek Terytorialnych do Celów Statystycznych (NTS). Można w nich znaleźć: listę jednostek podziału terytorialnego, miejscowości i ulic (pliki pełne), zmiany w rejestrze powstałe w danym okresie czasu (pliki aktualizacyjne), określone zakresy danych o jednostkach podziału terytorialnego (listy obiektów) oraz dane adresowe do poziomu ulicy. Ponadto uprawnione organy administracji rządowej i jednostek samo-rządu terytorialnego, inne instytucje rządowe oraz podmioty komercyjne mogą wykorzystywać dane rejestrowe REGON (Krajowy Rejestr Urzędowy Podmiotów Gospodarki Narodowej), które nie ograniczają się jedynie do identyfikatorów REGON, ale obejmują również bazę NIP (Numer Identyfikacji Podatkowej) i KRS (Krajowy Rejestr Sądowy).

Analityka danych wymaga zasobów i umiejętności, zatem ograniczeniem są również koszty, które trzeba ponieść w ramach przetwarzania i wykorzystywania informacji. Jednocześnie łączenie materiałów z różnych źródeł bywa problematyczne z uwagi na charakterystyki poszczególnych zbiorów i ich potencjalną niekompatybilność – inny format czy strukturyzacja danych stanowią istotny problem dla wspólnych analiz. Po-nadto pułapką może być koszt przechowywania, przetwarzania i archiwizowania ogromnego strumienia danych. Ich zakres i zbiór powinien być zawsze dostosowany do potrzeb i możliwości podmiotu.

Wśród możliwych obszarów zastosowania rozwiązań big data w sektorze publicznym wskazuje się[5]:

- statystykę publiczną;
- proces legislacyjny i tworzenie polityk publicznych;
- usługi publiczne dla obywateli i biznesu;
- bezpieczeństwo państwa i walka z przestępstwami.

1. Statystyka publiczna

   Tradycyjne gromadzenie danych w ramach statystyki publicznej obejmuje badania ankietowe ludności (np. spisy powszechne czy badania prowadzone w ramach określonego obszaru takie jak Badanie Aktywności Ekonomicznej Ludności), sprawozdania statystyczne wypełniane przez przedsiębiorstwa oraz pozyskiwanie informacji ze źródeł administracyjnych (np. w zakresie przestępczości od organów policji). Ten sposób pozyskiwania informacji powoduje przede wszystkim publikowanie danych z dużym opóźnieniem. Przy czym należy zauważyć, że z uwagi na wiarygodność źródeł mogą być one wyższej jakości. Jednocześnie koszty badań ankietowych powodują, że objęte są nimi jedynie najbardziej istotne zjawiska i na dość dużym poziomie ogólności. Stosowanie statystyki publicznej w podejmowaniu decyzji jest

[5] K. Kosior, Big data w sektorze publicznym–szanse, ograniczenia, perspektywy, Kultura i polityka, 20, 2016, s. 20–33.

istotne, ale w związku ze wskazanymi ograniczeniami, informacje uzyskane w ten sposób mogą nie pokazywać pełnego obrazu zjawisk oraz najaktualniejszej sytuacji.

W kontekście danych dotyczących rynku pracy, gromadzonych w ramach statystyki publicznej, nie można nie wspomnieć o istotnej roli urzędów pracy w pozyskiwaniu informacji. Zgodnie ze wspomnianą ustawą z 29 czerwca 1995 r. o statystyce publicznej, służby statystyki publicznej przetwarzają dane pochodzące z systemów i rejestrów m.in. prowadzonych przez odpowiednich ministrów czy organy jednostek samorządu terytorialnego. Zadania w zakresie aktywizacji zawodowej i polityk rynku pracy na szczeblu lokalnym są właśnie urzędy pracy (wojewódzkie i powiatowe), będąc źródłem danych pierwotnych. Szczegółowy zakres przekazywanych danych ustalany jest corocznie w rozporządzeniu obejmującym program badań statystycznych statystyki publicznej na dany rok. Obecnie urzędy pracy są zobowiązane do comiesięcznego sprawozdawania danych w zakresie[6]:

- aktywnych form przeciwdziałania bezrobociu;
- bezrobotnych zarejestrowanych w urzędach pracy;
- innych klientów urzędów pracy;
- poszukiwania pracy;
- wolnych miejsc pracy zgłoszonych do urzędów pracy;
- zgłoszeń zwolnień grupowych, zwolnień grupowych i zwolnień monitorowanych zgłoszonych do urzędów pracy.

Nie należy całkowicie rezygnować z ugruntowanych rozwiązań, ale można je uzupełnić o dodatkowe metody pozyskiwania i badania danych. Istotną kwestią jest kreatywne podejście do posiadanych zasobów i ustalenie w jaki sposób je efektywnie wykorzystać. Przede wszystkim technologie big data umożliwiają pozyskiwanie danych w czasie rzeczywistym, skracając tym samym czas opublikowania statystyk i eliminując opóźnienie. Również pozyskiwanie danych może być mniej kosztowne, dzięki czemu mona rozszerzyć zakres analizowanych zjawisk i procesów. Warto również zastanowić się nad alternatywnymi miernikami, które można uzyskać z dużych zbiorów danych. Niektóre zagraniczne instytucje publiczne wykorzystują big data do szacowania aktualnej stopy bezrobocia (np. w oparciu o treści zamieszczane w mediach społecznościowych) lub do oceny aktualnego wskaźnika wzrostu PKB (np. w oparciu o bazy danych zawierające informacje o płatnościach elektronicznych). Istotną kwestią jest zatem świadomość narzędzi i możliwości, które za sobą niosą oraz podjęcie próby wyjścia poza

[6] Rozporządzenie Rady Ministrów z 7 października 2022 r. w sprawie programu badań statystycznych statystyki publicznej na rok 2023 (Dz.U. 2022 poz. 2453).

ustalone standardy. Poniższa tabela obrazuje szanse, wyzwania i zagrożenia w wykorzystaniu big data w statystyce publicznej.

Tabela 3. Szanse, wyzwania i zagrożenia w wykorzystaniu big data w statystyce publicznej.

Szanse	Zmniejszenie kosztów wybranych badań (np. stosowanie automatycznego pobierania danych z internetu)
	Zmniejszenie obciążenia respondentów poprzez wykorzystanie już dostępnych danych.
	Możliwość wykorzystania danych z big data jako zmiennych pomocniczych w podejściu modelowym (na przykład w statystyce małych obszarów).
	Uzyskanie nowych informacji niedostępnych w statystyce publicznej (np. pochodzących z internetu rzeczy).
	Możliwość zastąpienia, uzupełnienia czy poprawy istniejących zbiorów danych (np. skrócenie czasu od zebrania do publikacji danych).
Wyzwania	Konieczność nawiązania współpracy z „dostawcami” i „producentami” dużych zbiorów danych (na przykład z operatorami telefonii komórkowej).
	Uzyskanie dostępu do danych (na przykład z Twittera, Facebooka czy z sieci komórkowej) oraz ich integracja z istniejącym systemem statystyki publicznej.
	Zagadnienia prawne i regulacyjne (m.in. poufność danych oraz ich bezpieczeństwo).
	Ocena jakości danych z punktu widzenia ich dokładności, przydatności, porównywalności, spójności, terminowości i punktualności.
	Ocena reprezentatywności danych oraz możliwość porównania z istniejącymi źródłami statystyki publicznej – konieczność przeprowadzenia badań.
	Problemy związane z dopasowaniem istniejącej struktury oraz metod stosowanych w statystyce publicznej do dużych danych (np. czyszczenia i edycji, imputacja, kalibracji).
Zagrożenia	Dostępność danych, które są najczęściej w rękach prywatnych, brak chęci współpracy z urzędami statystycznymi.

	Zagadnienia prawne (m.in. ochrona prywatności, konieczność zachowania tajemnicy statystycznej, „permanentna inwigilacja”).
	Niewystarczające pokrycie informacyjne badanej zbiorowości (m.in. ograniczenia w dostępie do internetu, problemy z publikacją danych w różnych subpopulacjach, na niskim poziomie agregacji przestrzennej).
	Obciążenie będące konsekwencją selektywności oraz braku reprezentatywności danych.
	Na ogół dane nie spełniają wymagań metodologicznych statystyki oficjalnej (m.in. w kontekście definicji stosowanych przez statystykę publiczną).
	Jakość danych (m.in. błędy nielosowe na poziomie jednostki oraz źródła, pomiar w różnych odstępach czasu, dane nieustrukturyzowane).
	Integracja z istniejącymi źródłami danych statystycznych – brak wspólnych identyfikatorów.

Źródło: M. Beręsewicz, M. Szymkowiak, Big data w statystyce publicznej – nadzieje, osiągnięcia, wyzwania i zagrożenia, Econometrics. Ekonometria. Advances in Applied Data Analytics, 2, 2015, s. 9-22.

2. Proces legislacyjny i tworzenie polityk publicznych

Wprawdzie decydenci przy kształtowaniu polityk publicznych głównie kierują się względami ideologicznymi, jednak niezależnie od wyboru konkretnego kierunku istotny jest również sposób jego wdrażania. Zastosowane rozwiązania powinny pozwalać na osiągnięcie zamierzonego skutku oraz być efektywne kosztowo. Wybór konkretnych metod należy poprzedzić odpowiednią analizą, która pozwoli na określenie, która z nich będzie najlepsza. W tym zakresie przydatna może być analityka predykcyjna oraz nowe źródła danych. Kluczowe jest posiadanie aktualnych informacji dotyczących określonych problemów czy zjawisk, których dana polityka ma dotyczyć. Jednocześnie prognozy umożliwiają ocenę efektów planowanych decyzji w czasie rzeczywistym, dzięki czemu możliwe jest wcześniejsze niwelowanie negatywnych skutków, które mogą się pojawić. Dostarczenie wiarygodnych, kompletnych informacji stanowi podstawę do podejmowania lepszych, trafniejszych decyzji. Analizy big data mogą być przydatne na różnych etapach podejmowania decyzji dostarczając informacje o preferencjach i potrzebach społecznych, prognozując konsekwencje różnych wariantów czy wskazując zależności, których nie da się ustalić tradycyjnymi metodami analitycznymi.

3. Usługi publiczne dla obywateli i biznesu

   W tym kontekście warto wskazać na koncepcję nowego zarządzania publicznego, którego podstawowym założeniem było wykorzystanie doświadczeń i metod zarządzania stosowanych w sektorze prywatnym oraz posłużenie się mechanizmami rynkowymi, uznawanymi za efektywniejsze. Zgodnie ze wskazaną teorią odbiorca świadczeń administracji publicznej powinien być traktowany jako klient, który płacąc podatki lub w inny sposób opłacając usługi administracji, ma prawo oczekiwać świadczeń o wysokiej jakości udzielanych na warunkach (w miarę możliwości) konkurencyjnych. Skojarzenie jest o tyle oczywiste, że firmy prywatne oferując towary czy usługi już od dawna korzystają z zaawansowanych metod analitycznych, żeby jak najlepiej dostosować się do potrzeb konsumentów i je zaspokajać. Analogicznie państwo mogłoby poprawić jakość i skuteczność oferowanych usług publicznych, zindywidualizować je poprzez dopasowanie do potrzeb konkretnych odbiorców i nagłych sytuacji, jak również określać nowe obszary, w których realizacja usług publicznych byłaby pożądana. Wykorzystanie metod charakterystycznych dla usług komercyjnych pozwoliłoby na zwiększenie skuteczności działań państwa, poprawę poziomu zadowolenie odbiorców usług, ale również na obniżenie kosztów, dzięki ich efektywniejszemu wydatkowaniu. Analizy big data można zastosować do projektowania usług rynku pracy, opieki społecznej, opieki medycznej, oświaty, infrastruktury publicznej czy zarządzanie przestrzenią miejską.

4. Bezpieczeństwo państwa i walka z przestępstwami

   Zastosowanie rozwiązań big data może również służyć jako dodatkowe narzędzia kontroli i monitorowania aktywności jednostek, grup, przedsiębiorstw oraz innych organizacji. Pozwala to chociażby na wcześniejsze wykrywanie nieprawidłowości lub niepokojących zjawisk czy lepsze dostosowanie działań prewencyjnych. W tym przypadku można wykorzystać analizy stron internetowych czy mediów społecznościowych. Stosowane są również specjalne algorytmy predykcyjne, które umożliwiają przewidywanie potencjalnych zagrożeń. Są to narzędzia przydatne szczególnie w walce z przestępczością zorganizowaną czy terroryzmem, ale również z przestępstwami pospolitymi. Wnioskowanie na podstawie statystyk policyjnych może wskazywać miejsca, które wymagają dodatkowych patroli. Narzędzia big data mogą być także pomocne do wykrywania oszustw i przestępstw gospodarczych, dzięki odkrywaniu niestandardowych zachowań i wzorców lub jednoczesnemu porównywaniu danych pochodzących z różnych źródeł.

Nowe technologie stwarzają szereg możliwości, jednak wiążą się również z ryzykami. Administracja publiczna posiada ogromne ilości danych o obywatelach, również dane sensytywne. Należy dbać o to, by nie wykorzystywać aparatu państwa do gromadzenia nadmiarowych informacji, ingerując jednocześnie w prywatne, wrażliwe sfery mieszkańców. Pozyskiwanie tak istotnych danych musi zawsze iść w parze z ogromną odpowiedzialnością, ostrożnością oraz skrupulatną dbałością o obowiązujące procedury bezpieczeństwa – analiza i interpretacja dużych zbiorów może prowadzić do błędów i nieprawidłowych wniosków, szczególnie jeśli nie jest przeprowadzana przez wykwalifikowanych specjalistów. Co istotne, im większe zbiory danych, tym większe ryzyko naruszenia bezpieczeństwa. Ataki hakerskie, kradzież danych czy wycieki informacji mogą skutkować poważnymi konsekwencjami, takimi jak kradzież tożsamości czy oszustwa finansowe. Ponadto należy zwrócić uwagę na fakt, że zaawansowane rozwiązania big data są w stanie wydobywać wzorce i zależności, które mogą posłużyć do wdrażania dyskryminacyjnych rozwiązań, umacniania nierówności czy szerzenia uprzedzeń. Analiza danych oparta na niepełnych lub błędnych informacjach może prowadzić do nieuczciwego traktowania grup społecznych.

Zakres wykorzystania big data w administracji publicznej może być bardzo szeroki i mieć zastosowanie do wielu różnych obszarów. Poniżej przedstawione zostało studium przypadku koncepcji Smart City, który pokazuje jak, w szerszym wymiarze, nowoczesne technologie mogą zmienić funkcjonowanie państwa.

Case Study – koncepcja Smart City

Wykorzystanie big data nie musi wiązać się jedynie z raportowaniem czy optymalizacją, w wąskim zakresie zastosowań, w administracji publicznej. Przykładem kompleksowego wykorzystania dostępnych narzędzi koncepcji Smart City (inteligentne miasto). Za sprawą rozwoju technologicznego ośrodki miejskie w niedalekiej perspektywie będą w stanie rozpocząć proces zbierania, a później przetwarzania danych z czujników urządzeń podłączonych do systemu internetu rzeczy (Internet of Things, IoT), aby w ten sposób rozpoznawać wzorce i potrzeby mieszkańców. Analiza może pomóc zmniejszyć liczbę wypadków drogowych i korków, a także umożliwić kierowcom szybkie znalezienie miejsca parkingowego. Ponadto, dzięki pozyskanym w ten sposób informacjom, można wnioskować na temat obszarów wzmożonej przestępczości bądź obniżać koszty infrastruktury miejskiej. Dane mogą również poprawić efektywność wydatkowania budżetu, np. poprzez skuteczniejsze wykorzystywanie systemów wodnych i energetycznych czy wprowadzenie inteligentnego oświetlenia miejskiego. Osiągnięcie podobnych celów wiąże się ze znaczną oszczędnością zasobów.

Na gruncie polskim pokazał to przykład Wrocławia. Po wdrożeniu projektu SmartFlow miasto odnotowało 9-procentową redukcję zużycia wody (500 milionów litrów). System analizował parametry pracy urządzeń wodociągowych, jak również przyczyniał się do szybszej reakcji na zakłócenia w sieci. Dzięki temu innowacyjnemu rozwiązaniu procesy lokalizacji i naprawy ukrytych wycieków wody mogły zostać skrócone nawet do 72 godzin, co znacznie usprawniło pracę diagnostów sieci i brygad wodociągowych działających na terenie miasta.

Aplikacja SmartFlow umożliwiła pełną kontrolę nad siecią wodociągową, gromadząc wszystkie dane dotyczące jej stanu w jednym systemie. Przedsięwzięcie zostało zrealizowane dzięki współpracy inżynierów gliwickiej spółki Future Processing oraz ekspertów Miejskiego Przedsiębiorstwa Wodociągów i Kanalizacji S.A. (MPWiK SA) we Wrocławiu, co świadczy o pomyślnej kooperacji administracji publicznej z sektorem prywatnym w sferze modernizacji państwa. Opisane zmiany umożliwiają rozwój miast oraz wyciąganie wniosków w kwestii poprawy jakości życia obywateli, np. efektywnego wykorzystywania środków finansowych.

Nie należy jednak abstrahować od potencjalnych zagrożeń związanych z rozwojem koncepcji Smart City. W inteligentnych miastach zbierane są ogromne ilości danych osobowych, takich jak dane lokalizacyjne, preferencje użytkowników czy informacje o nawykach. Istnieje ryzyko naruszenia prywatności mieszkańców, jeśli takie bazy byłyby nieodpowiednio chronione lub wykorzystywane w sposób niezgodny z intencjami obywateli, np. do monitorowania ich aktywności. Co więcej, wraz z wprowadzeniem zaawansowanych technologii informacyjno-komunikacyjnych (Information and Communications Technology, ICT) do infrastruktury miejskiej, zwiększa się prawdopodobieństwo ataków cybernetycznych. Hakerzy mogą dążyć do przechwycenia danych, zakłócenia systemów zarządzania miastem lub wywołania chaosu w infrastrukturze krytycznej. Inny aspekt wdrażania rozwiązań spod znaku Smart City dotyczy uzależnienia ośrodka miejskiego od technologii. W przypadku awarii systemów lub braku odpowiedniej infrastruktury komunikacyjnej wzrasta szansa zakłócenia podstawowych usług miejskich, a w rezultacie – pogorszenia jakości życia mieszkańców.

3. Analiza regionalnego rynku pracy przy użyciu big data

Analiza rynku pracy jest kluczowym narzędziem, które pozwala zrozumieć dynamikę zatrudnienia, trendy zawodowe oraz wyzwania związane z zaspokajaniem potrzeb pracodawców i pracowników w danym regionie. Tradycyjne metody zbierania danych dotyczących rynku pracy, takie jak badania ankietowe, są często czasochłonne, kosztowne i ograniczone pod względem skali. Jednak wraz z rozwojem technologii big data, możliwe stało się wykorzystanie ogromnych zbiorów danych do przeprowadzania dogłębnej analizy rynku pracy, również na poziomie regionalnym.

Współczesne narzędzia informatyczne umożliwiają łatwe gromadzenie ogromnych ilości informacji z nowych źródeł, w szczególności platform rekrutacyjnych czy portali społecznościowych. Jednocześnie stanowią szansę na uzupełnienie dotychczasowych metod badawczych, pozwalając na szerszą i dokładniejszą analizę. Tym samym dostarczają nowego, wartościowego materiału do badań, dostępnego w czasie rzeczywistym Zastosowanie narzędzi big data do rynku pracy pozwala na lepszą identyfikację trendów zatrudnienia, prognozowanie popytu na konkretne umiejętności, analizę rozwoju zawodowego oraz ocenę efektywności programów szkoleniowych i polityk rynkowych. Przykładowe zastosowania obejmują identyfikację niedopasowania popytu i podaży, identyfikację luk w umiejętnościach, analizę płacową czy wykrywanie nowych trendów i nisz rynkowych.

Korzystając z technologii big data, analitycy i decydenci są w stanie uzyskać głębsze i bardziej precyzyjne spojrzenie na regionalne rynki pracy. Zrozumienie aktualnych potrzeb i tendencji pozwala na podejmowanie lepiej ugruntowanych decyzji, w tym opracowywanie strategii rozwoju gospodarczego i polityk publicznych, planowania programów szkoleniowych i edukacyjnych czy restrukturyzacji zasobów ludzkich.

Celem niniejszego rozdziału jest rozpoznanie wykorzystania narzędzi big data do analizy regionalnego rynku. Omówione zostaną badania rynku pracy, przedmiot ich zainteresowania oraz źródła danych. Szczególna uwaga zostanie poświęcona analizie internetowych ofert pracy jako jednemu z zastosowań narzędzi big data. Wskazane zostaną również metody analityczne, które mogą zostać wykorzystane w dalszych badaniach oraz kwestie techniczne dotyczące tworzenia i prowadzenia bazy danych.

3.1. Charakterystyka badań rynku pracy

Podstawową kwestią jest zdefiniowanie rynku pracy jako kategorii ekonomicznej. W ekonomii przez pojęcie rynek rozumie się miejsce wymiany określonych dóbr lub usług, na którym spotykają się kupujący oraz sprzedający. Kupujący generują popyt, czyli wielkość, którą są w stanie nabyć po określonej cenie. Natomiast podaż oznacza zasób dóbr czy usług oferowanych przez sprzedających w zależności od ceny, którą mogą uzyskać. W obu przypadkach kluczowa jest cena, bowiem do transakcji dojdzie jedynie w przypadku, gdy będzie ona akceptowalna dla obu stron.

Zgodnie z klasycznym ujęciem, wraz ze wzrostem ceny popyt maleje, a rośnie podaż, natomiast spadkowi ceny towarzyszy wzrost wielkości popytu i spadek podaży.

W tym kontekście rynek pracy jako kategoria ekonomiczna obrazuje kształtowanie się relacji pomiędzy podażą zasobów siły roboczej a popytem na nie, a więc całokształt stosunków wymiennych zachodzących pomiędzy pracownikami a pracodawcami. Stronę podażową (zasoby pracy) stanowią osoby zdolne i chętne do pracy, czyli co do zasady osoby aktywne zawodowo[7]. Natomiast strona popytowa identyfikowana jest z zapotrzebowaniem na pracę, mającym odzwierciedlenie w oferowanych miejscach pracy. Ceną na tym rynku jest wynagrodzenie, które pracownicy otrzymują od pracodawców. Jeżeli jest ono satysfakcjonujące dla obu stron dochodzi do transakcji, czyli zatrudnienia.

Należy również zwrócić uwagę na kwestię nierównowagi na rynku pracy, która objawia się niezaspokojonym popytem lub podażą. W przypadku, gdy podaż przewyższa popyt pojawia się zjawisko bezrobocia, gdzie część osób zdolnych i gotowych do podjęcia pracy jej nie znajduje. Natomiast w odwrotnej sytuacji, gdy popyt przewyższa podaż, mamy do czynienia z deficytem siły roboczej (niedoborem kadr lub niedoborem talentów).

Celem badań rynku pracy jest analiza i zrozumienie zjawisk, które na nim zachodzą, identyfikacja trendów i prognozowanie zmian. Kluczową kwestią jest analiza niedopasowania popytu i podaży, ale również monitorowanie warunków pracy i zabezpieczenia socjalnego. Informacje pozyskiwane w ten sposób pozwalają na ocenę i kształtowanie polityk publicznych, ale również są istotną wskazówką dla pracodawców. Jednocześnie mogą być użyteczne w kontekście tworzenia programów kształcenia, dostosowując je do oczekiwań i realiów rynku pracy.

W praktycznym wymiarze, rynek pracy nie jest jednolity, zarówno w kontekście geograficznym, jak i przedmiotu wymiany. Prowadzi to do sytuacji, gdzie można mówić o lokalnych rynkach o różnych poziomach zatrudnienia i płac. W zależności od obszaru mogą być poszukiwane różne kwalifikacje, które jednocześnie mogą być inaczej wynagradzane. Tym samym analizy na poziomie regionalny mogą być precyzyjniejsze i lepiej obrazować lokalną, często specyficzną, sytuację. Jest to również zbieżne ze strukturą instytucjonalną służb zatrudnienia – urzędy pracy odpowiedzialne są za obszary województw i powiatów.

[7] Tradycyjnie pojęcie to obejmuje osoby w wieku produkcyjnym, nieuwzględniające osób, którym przysługuje emerytura. Przy czym również w tej grupie mogą znaleźć się osoby, które poszukują pracy np. w niepełnym wymiarze godzin, a ich aktywność jest przedmiotem zainteresowania.

3.1.1. Rodzaje analiz rynku pracy

Na rynku pracy istnieje szereg zjawisk, które warto monitorować i analizować. Wśród najważniejszych obszarów badawczych, które mogą być realizowane przez urzędy pracy wskazuje się[8]:

- analizy stanu zatrudnienia;
- analizy bezrobocia;
- analizy niedopasowania podaży do popytu;
- analizy stanu bierności;
- analizy przepływów na rynku pracy;
- analizy skuteczności instytucji i polityk rynku pracy.

1. Analizy stanu zatrudnienia

   Podstawową kwestią jest analiza stanu zatrudnienia, jej dynamika i struktura w czasie. Pozwala to na określenie zachodzących zmian i ustalenie ich przyczyn. Oprócz ogólnej liczby pracujących na danym obszarze zwraca się uwagę również na ich cechy, obejmujące kwestie demograficzne, kwalifikacje, rodzaj wykonywanej pracy czy wynagrodzenie. Istotne jest również miejsce pracy i charakterystyka pracodawcy. Dodatkową kwestią jest zachowanie pracujących na rynku pracy i ich postawy, w szczególności udział w kształceniu ustawicznym czy sposoby poszukiwania zatrudnienia. Dzięki temu można określić jakie czynniki wpływają na znalezienie i utrzymanie pracy, co może być wykorzystywane we wdrażanych politykach.

2. Analizy bezrobocia

   Kolejnym kluczowym zagadnieniem jest analiza bezrobocia. Analogicznie, istotne jest nie tylko mierzenie tego zjawiska, ale również poszukiwanie jego przyczyn i sposobów na przeciwdziałanie. Jednakże kwestia faktycznego oszacowania bezrobocia sprawia większe trudności niż w przypadku osób pracujących. Bezrobocie rejestrowe, widoczne w statystykach urzędów pracy, może nie odzwierciedlać realnego stanu na rynku. Z jednej strony część osób, zwłaszcza o wyższych kwalifikacjach, nie dokonuje rejestracji, z drugiej zaś niektórzy zarejestrowani mogą nie być faktycznie zainteresowani podjęciem pracy, a zgłoszenie do urzędu wiąże się np. z uzyskaniem zabezpieczenia społecznego. Analiza bezrobocia wymaga również określenia charakterystyk związanych z tym stanem. Oprócz podstawowych cech demograficznych, istotną kwestią jest również bezpośrednia przyczyna tego stanu. Może ona wynikać w szczególności z wejścia na rynek po zakończeniu edukacji,

[8] M. Góra, U. Sztanderska, Wprowadzenie do analizy lokalnego rynku pracy: Przewodnik, Ministerstwo Pracy i Polityki Społecznej, 2006.

rezygnacji z zatrudnienia czy zwolnienia wynikającego z likwidacji miejsca pracy lub decyzji pracodawcy. Dodatkowo mogą wystąpić osoby, dotychczas nieaktywne zawodowo, które zdecydowały się ponownie podjąć pracę z przyczyn finansowych lub rozwojowych. We wszystkich tych przypadkach należy zwrócić również uwagę na czas pozostawania bez pracy (zarówno celowy, jak i wynikający z braku znalezienia odpowiedniego zatrudnienia).

Badania bezrobocia powinny prowadzić do ustalenia jego charakteru. Może być ono frykcyjne, krótkotrwałe, wynikające z poszukiwania nowej pracy i będące rezultatem ruchu zatrudnionych na rynku pracy. Występowanie jedynie tego typu bezrobocia oznacza w praktyce równowagę na rynku i nie stanowi problemu, który wymaga rozwiązania. Jednakże częściej brak pracy wynika z niedostosowań strukturalnych lub przyczyn instytucjonalnych. Należy zatem ustalić w szczególności czy osoby bezrobotne posiadają deficyty kwalifikacyjne, czy próbują je uzupełnić, czy poszukują pracy dostępnej na danym terenie i czy ich oczekiwania dotyczące warunków zatrudnienia są zgodne z sytuacją rynkową. Problematyczne w tej kwestii jest dokonanie obiektywnej oceny, bowiem chociażby w kontaktach z urzędem pracy (będącym jedynym z podstawowych źródeł wiedzy o tej grupie) bezrobotni mogą nie prezentować prawdziwych postaw, np. w kontekście oczekiwań finansowych czy innych źródeł dochodu, które pozwalają na dłuższe poszukiwania pracy.

3. Analizy niedopasowania podaży do popytu

   Istotnym przedmiotem badań jest również niedopasowanie podaży do popytu. Korzystając z tradycyjnych narzędzi analitycznych, brakowało dobrych miar do określenia faktycznego zapotrzebowania na pracę. Można było posługiwać się porównaniem struktury pracujących, przybliżającej strukturę popytu, i struktury bezrobotnych, przybliżającą niezrealizowaną podaż. Pewnym rozwiązaniem jest wykorzystanie big data i przeprowadzenie analizy ofert internetowych, które obrazują popyt pracodawców. Do oceny niedopasowań istotne są cechy obrazujące wyposażenie w kapitał ludzki – wykształcenie, kwalifikacje, rodzaj wykonywanej pracy, staż pracy etc.

4. Analizy stanu bierności

   Analogicznie jak w przypadku, pracujących i bezrobotnych, ważne jest analizowanie dynamiki i struktury trzeciej grupy, czyli osób, które decydują się na niepodjęcie pracy. Oprócz cech demograficznych czy kwalifikacyjnych, kluczowe są okoliczności pozostawania biernym zawodowo. Wśród podstawowych motywów dezaktywizacji zawodowej wskazuje się proces kształcenia lub otrzymywanie świadczeń, które zostałyby wstrzymane lub ograniczone w przypadku podjęcia pracy. Inną przyczyną może być również

konieczność opieki nad dziećmi czy chorymi lub niepełnosprawnymi członkami rodziny, przy braku odpowiednich placówek opiekuńczych lub innego wsparcia. Poza ustaleniem okoliczności, ważne jest również określenie czy i w jakiej perspektywie bierni zawodowo zamierzają wrócić na rynek pracy oraz jakie czynniki mogłyby sprzyjać takiej decyzji.

5. Analizy przepływów na rynku

   Badanie przepływów pozwala na stwierdzenie jaka jest dynamika każdego ze stanów, jak również określenie charakterystyk osób, które go zmieniają poprzez aktywizację zawodową lub znalezienie pracę. Istotne jest również ustalenie czasu pozostawania w określonych stanach, zarówno w kontekście określenia przyczyn trudności w znalezieniu pracy, jak i cech gwarantujących stabilność wynagrodzenia. Na podstawie tych danych można prześledzić procesy restrukturyzacyjne w zatrudnieniu i zidentyfikować ich konsekwencje.

6. Analizy skuteczności polityk rynku pracy

   Ewaluacja działań podejmowanych w celu przeciwdziałania bezrobociu czy promocji zatrudnienia pozwala na ich ocenę oraz ulepszenie. Przede wszystkim obejmuje ona beneficjentów programów wsparcia oraz ich funkcjonowanie na rynku pracy. Tego typu badania przeprowadza się poprzez porównanie z grupą kontrolną o podobnych charakterystykach, którzy nie korzystali z podobnej pomocy. W celu zaobserwowania faktycznych, długoterminowych efektów, nie powinny być realizowane bezpośrednio po udzieleniu wsparcia.

3.1.2. Wykorzystanie big data w badaniach rynku pracy

Nowoczesne technologie stwarzają nowe możliwości badań, poszerzając dotychczasowe analizy. Pozwalają zarówno na pozyskanie danych z niedostępnych dotąd źródeł, jak i przetwarzanie ich dużych zbiorów. W szczególności dostępne stały się informacje o pracodawcach oraz osobach aktywnych zawodowo udostępniane przez nich samych w internecie. Zamiast przeprowadzania kosztownych i czasochłonnych badań ankietowych można wykorzystać dane z portali rekrutacyjnych, społecznościowych czy edukacyjnych. Dlatego też obecnie jednym z najpopularniejszych obszarów wykorzystania big data jest analiza stron internetowych.

Strony internetowe z ogłoszeniami o pracę pozwalają na uzyskanie danych o preferencjach i oczekiwaniach pracodawców, ale również oferowanych warunkach. Pozwala to zarówno ustalić popyt na konkretne zawody i umiejętności w czasie rzeczywistym, jak również posiadając dane historyczne można te kategorie prognozować. Informacje z ofert pracy wskazują również jakie dodatkowe umiejętności wymagane są na danych stanowiskach oprócz formalnego

wykształcenia (np. znajomość języków, określonych programów komputerowych). Mogą być one wykorzystane do dostosowania form kształcenia, ale również pozwalają zaplanować kursy podnoszące kwalifikacje lub pozwalające na przekwalifikowanie.

Analiza ofert pracy pozwala również na szacowanie liczby wakatów na rynku oraz częściowo przepływów. Czas aktywności oferty wskazuje ile zajmuje poszukiwanie pracownika i proces rekrutacyjny, a ponowne pojawienie się tożsamych ofert po dłuższym czasie (np. kilku miesięcy) może sugerować niedopasowanie pozyskanej osoby do posady lub rozwój firmy i pojawienie się nowych stanowisk. Przykład ten pokazuje, że tego typu dane powinny stanowić jedynie jedno z pozyskiwanych źródeł, a łączenie różnych zasobów pozwala na kompleksową analizę i identyfikację zjawisk czy wzorców. Co istotne, analityka big data dostarcza również narzędzi pozwalających na takie działania.

Dodatkowo ogłoszenia dostarczają danych o zmianach i trendach na rynku pracy np. w zakresie elastycznych form pracy – pracy zdalnej czy pracy w niepełnym wymiarze godzin. Jak również pozwalają na analizę i prognozowanie wynagrodzeń oraz kształtowania się benefitów pozapłacowych. Często pokazują również informację o aplikacjach na dane stanowisko co pokazuje zainteresowanie poszczególnymi ofertami.

Portale społecznościowe, w szczególności biznesowe, dostarczają informacji o osobach aktywnych zawodowo, ich wykształceniu, kwalifikacjach czy doświadczeniu zawodowym. Stanowią praktycznie interaktywne CV, które pozwalają określić charakterystyki zasobów siły roboczej. Możliwa jest analiza okresów zatrudnienia u danego pracodawcy, gotowość do zmiany pracy czy częstotliwość awansów.

W określeniu i prognozowaniu cech potencjalnych pracowników mogą również pomóc informacje dotyczące edukacji. Zarówno serwisy rekrutacyjne szkół wyższych wskazujące na liczbę miejsc na danym kierunku, ale również popularne portale oferujące kursy online np. Coursera czy edX. Kursy te są zazwyczaj tworzone we współpracy z uczelniami wyższymi i umożliwiają uzyskanie certyfikatu potwierdzającego nabycie określonych umiejętności, a czasem nawet pozwalają na uzyskanie formalnego wykształcenia. Jest to popularna forma dokształcania, również w kontekście rozwoju zawodowego. Dzięki temu można uzyskać informację w jakich obszarach użytkownicy chcą poszerzać swoją wiedzę i jakie kwalifikacje uznają za pożądane.

Oprócz tego możliwe jest badanie aktywności w sieci. Może to dotyczyć zarówno analizy odwiedzanych stron, jak i wyszukiwanych fraz związanych z zatrudnieniem. Z jednej strony może to pokazywać aktualne trendy w zakresie poszukiwanych treści, ofert czy kursów oraz wskazywać najpopularniejsze strony np. wśród osób

poszukujących pracy. Z drugiej zaś analiza długoterminowa może uwidocznić zachodzące zmiany, ale również wskazywać na cykliczność pewnych zjawisk.

Wykorzystanie danych internetowych w badaniach ma oczywiście swoje ograniczenia. W szczególności nie wszystkie informacje mogą być rzetelne czy pełne, co należy uwzględniać w ich analizie. Nowe zbiory danych powinny stanowić uzupełnienie i rozszerzenie dotychczasowych zasobów, z których nie należy całkowicie rezygnować. Co istotne, potencjał big data pozwala nie tylko na gromadzenie danych, ale również ich łączenie oraz dostarcza narzędzi analitycznych, chociażby bazujących na sztucznej inteligencji, rozpoznawaniu obrazów czy uczeniu maszynowym.

3.1.3. Źródła danych o rynku pracy

Dane do analiz mogą być pozyskiwane z różnych źródeł. Warto korzystać zarówno z danych gromadzonych w ramach statystyki publicznej, będących w posiadaniu instytucji publicznych, w tym zasobów własnych, dostarczanych przez podmioty prywatne czy pozyskiwanych samodzielnie ze źródeł ogólnodostępnych np. stron internetowych. Zestawienie wybranych źródeł zostało zawarte w tabeli 4.

Tabela 4. Źródła danych przydatne do badań rynku pracy.

Rodzaj źródła	Zakres posiadanych danych
Urzędy pracy	Dane dotyczące bezrobotnych i pracodawców
Główny Urząd Statystyczny oraz urzędy statystyczne	Dane badań przedsiębiorstw, (zatrudnienie, płace, popyt na pracę, koszty pracy), badań ankietowych ludności (zatrudnienie, bezrobocie, bierność zawodowa, płace)
Ministerstwo Finansów oraz urzędy skarbowe	Dane podatkowe
Ministerstwo Rozwoju i Technologii	Dane dotyczące działalności gospodarczej
Ministerstwo Rodziny i Polityki Społecznej oraz ośrodki pomocy społecznej	Dane dotyczące pomocy społecznej
Zakład Ubezpieczeń Społecznych oraz jego oddziały	Dane dotyczące ubezpieczonych, składek i świadczeń społecznych
Kasa Rolniczego Ubezpieczenia Społecznego oraz jej oddziały	Dane dotyczące ubezpieczonych, składek i świadczeń społecznych rolników
Podmioty prywatne, organizacje pozarządowe, ośrodki badawcze	Dane gromadzone w celu analizy rynku pracy

Rodzaj źródła	Zakres posiadanych danych
Placówki edukacyjne (szkoły, uniwersytety)	Dane dotyczące absolwentów, realizowanych programów kształcenia
Strony internetowe	Ogólnodostępne dane np. dotyczące ofert pracy

Źródło: opracowanie własne.

Identyfikując potencjalne źródła, umożliwiające pozyskiwanie danych, warto rozpocząć od zasobów własnych urzędu pracy. Jednymi z jego podstawowych zadań jest rejestracja osób bezrobotnych oraz pośrednictwo pracy. Dzięki temu urząd pracy posiada dane o bezrobociu rejestrowanym oraz aktualnych ofertach pracy zgłoszonych przez pracodawców, chcących skorzystać ze wsparcia urzędu. Główną zaletą tego źródła jest możliwość dostosowania zakresu oraz rodzajów danych w momencie ich pozyskiwania. Odpowiednie przygotowanie formularzy czy stosowanie ustandaryzowanych pojęć jest istotnym ułatwieniem. Jednakże wiele osób poszukujących pracy nie rejestruje się w urzędzie, jak również nie wszyscy pracodawcy korzystają z pośrednictwa. W szczególności może to dotyczyć pracowników i zawodów wysokospecjalizowanych, co istotnie może zaburzyć wiarygodność analiz czy prognoz. Zatem zasoby własne należy uznać za punkt wyjścia przy konstruowaniu pełnej bazy.

Istotnym elementem mogą być dane pochodzące z instytucji i organów publicznych, jednak posiadają one istotne ograniczenia. Co do zasady urzędy i ministerstwa podają do publicznej wiadomości jedynie dane wtórne, przetworzone i zagregowane w postaci sprawozdań, których nie można zakwalifikować jako big data. Uzyskanie mikrodanych zawartych w bazach czy rejestrach wymagałoby odpowiedniej zgody oraz przygotowania, w szczególności anonimizacji. O ile udostępnianie przez polskie instytucje danych pierwotnych nie jest standardową praktyką, możliwe jest ubieganie się dostęp do nich i zgodnie z najlepszą wiedzą autorów takie sytuacje miały miejsce dla celów badawczych.

Uzyskanie danych od podmiotów prywatnych, organizacji pozarządowych czy ośrodków badawczych nie powinno być problematyczne, jednak najczęściej odbywa się na zasadach komercyjnych i może generować wysokie koszty, zwłaszcza w przypadku zbierania danych na zlecenie. Istotnym czynnikiem kosztotwórczym jest wolumen oczekiwanych informacji, zwłaszcza w kontekście dużych zbiorów danych.

Dane mogą być również pozyskiwane ze szkół ponadpodstawowych, policealnych oraz uczelni wyższych. Informacja o liczbie uczniów i studentów wybierających określone profile czy kierunki pozwala prognozować kształtowania się podaży pracy o określonych kwalifikacjach.

W kontekście planowanego projektu kluczowym źródłem danych są treści oferty pracy publikowane w internecie. Podstawową kwestią w tym zakresie jest identyfikacja stron, które posiadają największy potencjał informacyjny. Przede wszystkim należy zidentyfikować najpopularniejsze portale rekrutacyjne z uwzględnieniem liczby ofert i odsłon. Wśród najbardziej rozpoznawalnych komercyjnych serwisów ogłoszeniowych należy wskazać:

- pracuj.pl;
- pl.indeed.com;
- praca.pl;
- infopraca.pl;
- olx.pl;
- jobs.pl;
- lento.pl;
- pracatobie.pl;
- gowork.pl;
- www.careerjet.pl.

Jednocześnie wiele ofert pojawia się w mediach społecznościowych, zarówno profesjonalnych, związanych z działalnością zawodową (LinkedIn, GoldenLine), jak i ogólnych (Facebook). Należy przy tym zaznaczyć, że dostęp do tych ofert może być utrudniony w związku z ograniczoną widocznością postów dla osób niezalogowanych.

Warto również ustalić jakie dane (kategorie) mają być pozyskiwane z tego źródła. Mając na względzie wykorzystanie informacji zawartych w ofertach do dalszej analizy zapotrzebowania na zawody, kwalifikacje i kompetencje, wśród istotnych danych należy wskazać (przy czym jest to jedynie przykładowa lista):

- numer ewidencyjny;
- nazwę stanowiska;
- wykształcenie;
- kwalifikacje;
- umiejętności;
- doświadczenie zawodowe;
- wynagrodzenie;
- dodatkowe benefity;
- rodzaj umowy;
- wymiar czasu pracy;
- dane pracodawcy;
- branżę;
- lokalizację miejsca pracy;
- datę publikacji oferty;
- nazwę portalu.

3.2. Analiza ofert pracy

Jednym z podstawowych zadań „Lubelskiego Obserwatorium Rynku Pracy 2023-2028” ma być analiza ofert pracy, dzięki której można określić zapotrzebowanie na zawody, kwalifikacje i kompetencje. Pozwala to ustalić orientacyjne oczekiwania pracodawców w poszczególnych branżach. W efekcie możliwe jest lepsze dopasowanie ofert do kandydatów, przygotowanie ich do procesu rekrutacji i pracy na stanowisku oraz ogółem poprawienia jakości pośrednictwa pracy. Trudne, ale wyjątkowo ważne jest prognozowanie kierunków rozwoju rynku pracy i określenie przynajmniej orientacyjnego zapotrzebowania na pracowników w minimum kilkuletnim horyzoncie. Pozwala to zidentyfikować przyszłe przesunięcia sektorowe i wyjść z ofertą do pracowników branż ograniczających produkcję tak, by mogli oni się przekwalifikować w terminie. Z perspektywy urzędu pracy wyniki analiz stanu obecnego oraz prognoz umożliwią chociażby lepsze dostosowanie szkoleń, kursów oraz innych form wspierających podnoszenie kwalifikacji do aktualnego i prognozowanego zapotrzebowania. Prognozy rozwoju rynku pracy są także niezmiernie ważne dla uczelni oraz szkół zawodowych, by mogły dopasować program kształcenia do zmieniających się potrzeb. Nieadekwatne dopasowanie oferty szkolnictwa do potrzeb rynku pracy jest od dawna wymieniane jako słabość Polski w różnorakich rankingach konkurencyjności.

3.2.1. Ograniczenia analizy

Analiza ofert pozwala na oszacowanie popytu na pracę w podziale na sektory gospodarki, zawody, określone umiejętności i kwalifikacje. Jednak wykorzystanie dużych zbiorów danych jako źródła informacji do dalszych badań statystycznych czy ekonometrycznych również obarczone jest pewnymi ograniczeniami, o których należy pamiętać.

- Ogłoszenia internetowe odnoszą się tylko do części rzeczywistych ofert pracy i nie wszyscy pracodawcy korzystają z tego kanału. Niekiedy stosowane mogą być bardziej tradycyjne metody, takie jak ogłoszenia w gazetach, czy nawet ‘word of mouth’ polegające na pytaniu po znajomych. Z przyczyn oczywistych tego rodzaju ogłoszeń nie da się uwzględnić w analizie. Postępująca informatyzacja kraju oznacza, że ogłoszeń nie publikowanych w internecie będzie coraz mniej, aż do ich całkowitego zaniku (oczywiście ogłoszenie może być np. opublikowane jednocześnie w gazecie oraz w internecie). Popularność internetu jako platformy dla ogłoszeń o pracę zależy także od rodzaju stanowiska. Gdzie indziej pracy poszukiwał będzie programista, a gdzie indziej budowlaniec. W efekcie, wystąpić może nadreprezentacja ofert dużych przedsiębiorstw czy zawodów częściej ogłaszających się w internecie, wymagających wyższego lub specjalistycznego wykształcenia. W szczególności chociażby poszukiwania pracowników do gastronomii czy na stanowiska ekspedientów odbywa się często poprzez ogłoszenia wywieszane w lokalach czy sklepach. Jednocześnie rekrutacja na wyższe stanowiska menedżerskie zazwyczaj prowadzone są przez rekruterów

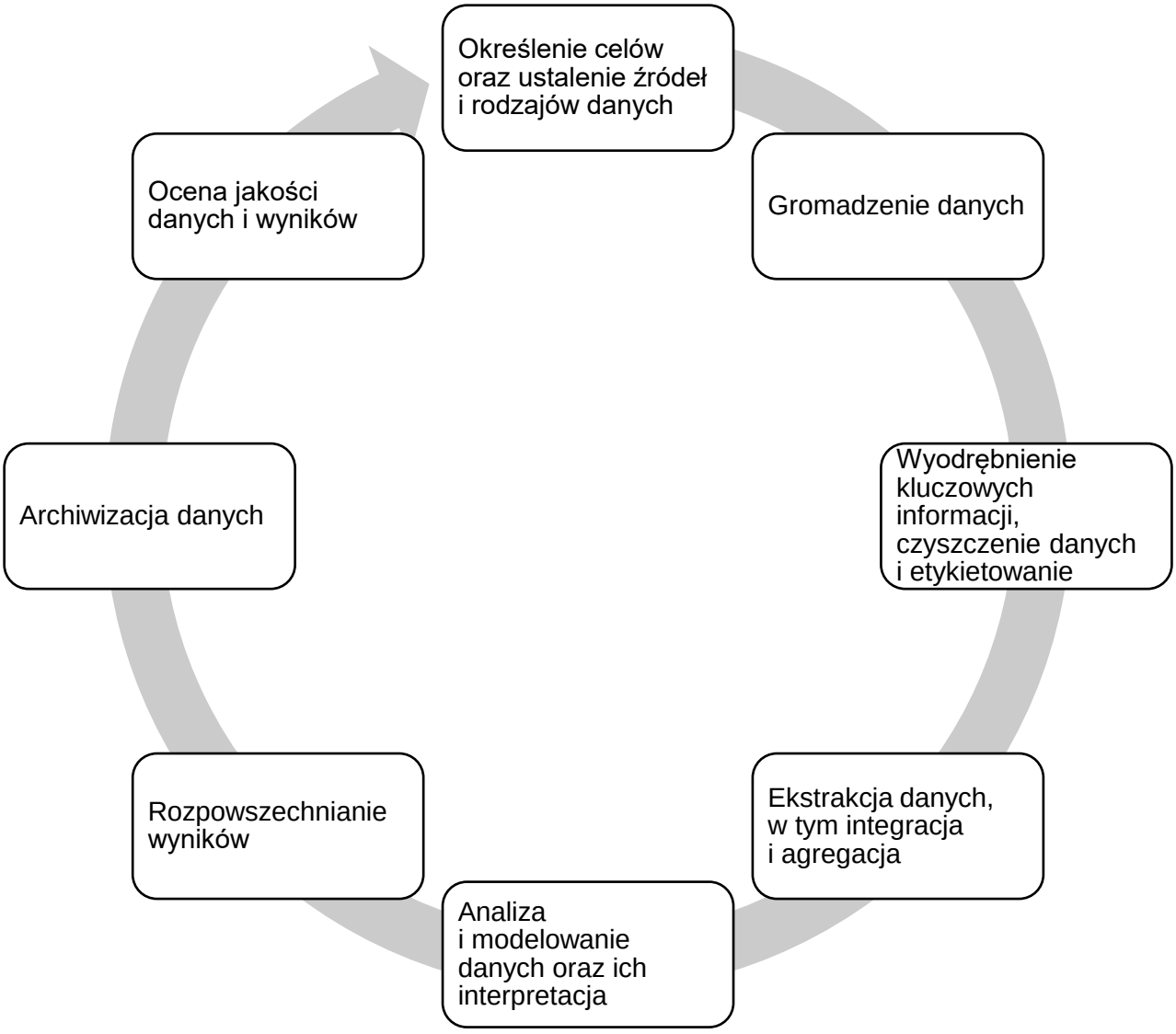
kontaktujących się bezpośrednio z kandydatami wyselekcjonowanymi w internetowych sieciach biznesowych (np. LinkedIn czy GoldenLine). Warto zatem korzystać z różnych źródeł (np. analiza ofert pracy wraz z analizą zasobów własnych), mając jednak świadomość, że nie stanowią one całości dostępnych ofert i uwzględniać ten fakt w dalszych analizach statystycznych czy ekonometrycznych.

- Nie ma jednego określonego źródła informacji o ofertach pracy. Mogą być one publikowane jednocześnie w różnych kanałach internetowych np. na kilku serwisach ogłoszeniowych, stronach agencji rekrutacyjnych czy konkretnych pracodawców. Dodatkowo część ofert może być promowana, a tym samym pojawiać się kilkukrotnie w ramach jednego portalu. Pozyskanie danych z kilku źródeł pozwala na uzyskanie większego zakresu informacji, jednak może również wiązać się z powielaniem tych samych ofert. Eliminowanie powtarzających się obserwacji jest jednym z podstawowych problemów wykorzystania big data do monitorowania rynku pracy. Jednocześnie oferty nie zawsze zawierają nazwę pracodawcy, zwłaszcza w zakresie rekrutacji przeprowadzanych przez agencje pracy, które często używają określeń rodzajowych. Nie można również wykluczyć zapotrzebowania na więcej niż jeden etat określonego rodzaju w danym miejscu. Eliminowanie powtarzających się ogłoszeń jest zadaniem skomplikowanym, z uwagi na brak jednolitego identyfikatora. Taki mógłby być nadawany przez urząd, po zgłoszeniu zapotrzebowaniu na pracownika przez firmę, wakatowi nadawany jest numer ID, który jest następnie umieszczany we wszystkich ogłoszeniach na ów wakat. Rozwiązanie tworzy jednak znaczną komplikację dla przedsiębiorstw i rekruterów, zatem należy traktować je raczej w kategoriach koncepcji. Pewnym rozwiązaniem są algorytmy przetwarzania języka naturalnego, które pozwolą na identyfikację tożsamych ofert w związku z ich konstrukcją i sposobem sformułowania. Pozwoli to przykładowo na badanie różnic w wynagrodzeniu na równorzędnych stanowiskach między firmami, a nawet pozwoli wykryć ewentualne różnice w ogłoszeniach na to samo stanowisko w tej samej firmie, umieszczane w różnych miejscach w internecie.
- Oferty pracy mogą mieć różny stopień szczegółowości i zawierać różne zakresy informacji. W szczególności może brakować wspomnianej wcześniej nazwy pracodawcy czy poziomu wynagrodzenia. Mogą również występować różnice w opisach stanowisk, wymagań, lokalizacji czy benefitów pozapłacowych. W tym kontekście istotna jest standaryzacja gromadzonych informacji (np. w zakresie ujednolicenia nazewnictwa stanowisk) czy odpowiedni wybór istotnych z perspektywy urzędu pracy elementów oferty, które będą zapisywane w bazie. Należy również rozważyć cykliczną ewaluację ustalonych kategorii, chociażby w związku z pojawianiem się nowych zawodów czy ewolucją warunków na rynku pracy. Jednocześnie pewne braki danych w utworzonej bazie mogą być nieuniknione lub może wystąpić nieprawidłowa klasyfikacja, zwłaszcza w kontekście danych nieustrukturyzowanych.

3.2.2. Proces analizy ofert pracy

Mając świadomość występujących ograniczeń, warto ustalić podstawowe działania w ramach procesu analizy ofert pracy. Jest to swego rodzaju powtarzający się cykl, którego etapy zostały przedstawione na grafie.

Graf 1. Proces analizy ofert pracy.



Źródło: opracowanie własne.

1. Określenie celów oraz ustalenie źródeł i rodzajów danych

   Podstawowym celem analizy jest ustalenie zapotrzebowanie na zawody, kwalifikacje i kompetencje, czyli oczekiwań strony popytowej. Dostępnymi źródłami będą portale, na których publikowane są ogłoszenia rekrutacyjne. Z przyczyn oczywistych ogłoszenia umieszczane w mediach innych niż internet nie będą uwzględniane w tworzeniu bazy. Choć teoretycznie możliwe

jest np. skanowanie gazet i ekstrahowanie tekstu ogłoszenia ze skanu, w praktyce jest to zbyt czaso- i pracochłonne. Wzbogaceniem tworzonej bazy mogą być inne, wskazywane wcześniej źródła: zasoby własne, informacje od pracodawców czy agencji rekrutacyjnych. Zaznaczyć należy, że tego rodzaju dane przypuszczalnie nie będą stanowić jednej tabeli danych z informacjami o ogłoszeniach pobieranymi z internetu, gdyż tych drugich będzie najpewniej znacznie więcej. Baza danych jest jednak zwykle tworzona z wielu tabel (udostępnianych pracownikom w formie widoków), zatem ewentualne rozbieżności w strukturze pozyskanych dodatkowych danych nie przekreślają ich użyteczności w prowadzeniu analiz.

Kluczowe w tworzeniu bazy są informacje o treści ofert, które następnie będą przetwarzane. Należy również określić kategorie danych, które znajdą się w bazie (rodzaj stanowiska, wymagane wykształcenie i kwalifikacje, oczekiwany staż pracy, wynagrodzenie etc.). W kontekście danych pochodzących ze stron internetowych istotne może być również zachowanie informacji dodatkowych takich jak data publikacji czy adres strony. W zależności od dostępnych zasobów, tworzyć można kopie całych ofert.

By rozszerzyć analitykę na inne aspekty rynku pracy np. w zakresie kształtowania się bezrobocia, zmian w wynagrodzeniach, najlepiej będzie pozyskać dodatkowe dane, oprócz analiz ofert internetowych. Analitykę w ograniczonym zakresie będzie można jednak wykonać za pomocą bazy ogłoszeń. Przykładowo, częstotliwość publikacji ofert pracy na konkretne stanowisko w danej firmie pozwala ustalić skalę rotacji, co pozwala wyciągnąć istotne wnioski na temat kondycji branży, w której przedsiębiorstwo działa. Idealnie, dane z ogłoszeń zostaną uzupełnione mikrodanymi takimi jak zeznania podatkowe firm, czy informacje z opłacanych składek. Szersza analityka pozwala wówczas oszacować skalę problemów takich jak praca na czarno.

2. Gromadzenie danych

Ten etap obejmuje pobieranie surowych danych z ustalonych źródeł. Pozyskiwanie informacji może być wykonywane ręcznie jedynie w bardzo małych bazach. W analityce big data wykorzystuje się w tym celu wyspecjalizowane boty, automatyzujące tę czynność i istotnie zwiększające pozyskiwane wolumeny danych w jednostce czasu. Uzyskane informacje są zapisywane w celu dalszego przetwarzania.

Gromadzenie danych nie ogranicza się jednak wyłącznie do web scrapingu. Ważnych informacji dla analityków pracujących na bazie dostarczą także dane posiadane i przekazane w drodze współpracy przez zaprzyjaźnione instytucje sektora publicznego czy firmy prywatne. Gromadzenie danych może również

obejmować inne czynności np. przekazywanie ich przez inne podmioty (np. instytucje publiczne) z posiadanych przez nie baz, tworzenie formularzy i narzędzi do zbierania danych czy wykorzystywanie informacji statystycznych.

3. Wyodrębnienie kluczowych informacji, czyszczenie danych i etykietowanie

   Surowe dane pozyskiwane zwykle w formie ciągu znaków nie nadają się do analizy. Najpierw należy je przetworzyć i dostosować, żeby nadawały się do dalszej analizy. Ważny jest wybór typu zmiennej. Oczywistym jest rozróżnienie między ciągiem znaków a liczbą. Mniej oczywisty jest np. wybór między zapisaniem kolumny liczb w postaci int a zmiennoprzecinkowej. O ile te drugie oferują dużo większy zakres akceptowanych liczb, to zajmują zdecydowanie więcej miejsca. Dane następnie należy przefiltrować oraz posortować. Pozwoli to na ustalenie braków lub błędów. Konieczne jest również wyczyszczenie i korekta danych, usunięcie duplikatów czy błędnych rekordów. Warto zwrócić uwagę na format, który powinien umożliwiać wykorzystanie informacji do dalszych analiz. Przydatne może być również etykietowanie danych poprzez przypisanie im określonych kategorii.

   Osobną kwestią, wartą poruszenia, jest traktowanie braków danych. Niektóre metody np. uczenia maszynowego nie akceptują w kodzie braków danych, istotne jest zatem zastąpienie tychże jakąś wartością. Jedną z opcji jest wzięcie średniej z pozostałych wartości w zbiorze (lub mediany w przypadku zmiennych typu char). Innym rozwiązaniem jest tzw. hot deck imputation, gdzie braki wypełniane są na podstawie analizy podobnych przypadków, w których wartość występuje. Podczas tworzenia modelu można też zdecydować pominąć się obserwacje, w których występuje brak danych. Może to jednak prowadzić do problemów, kiedy braki są liczne.

   Anonimizacja często bywa niezbędna przy tworzeniu baz danych, w których zawarte są informacje pozwalające jednoznacznie zidentyfikować dany podmiot, w tym przypadku firmę. Wprawdzie NIP czy REGON nie są podawane w ogłoszeniach, jednak nazwa w nich widnieje. Ta powinna zostać zhashowana, tj. nazwom rzeczywistym powinny zostać przyporządkowane losowe kombinacje znaków w relacji jeden do jednego. Podobnie potraktowane lub wyeliminowane powinny zostać wszelkiego rodzaju informacje pozwalające jednoznacznie zidentyfikować dany podmiot. Nie oznacza to jednak, że identyfikacja nie jest możliwa w żadnym przypadku. Jeżeli jakiemuś podmiotowi przypisać można kombinację pewnych rzadko spotykanych wartości zmiennych, wówczas rozpoznanie przedsiębiorstwa kryjącego się pod danym numerem ID jest możliwe. Zapobieżenie temu jest praktycznie niemożliwe, dlatego osoby pracujące przy danych należy na to uczulić.

4. Ekstrakcja danych, w tym integracja i agregacja

   Ekstrakcja polega na wydobyciu z surowych danych istotnych elementów. Jest to proces przekształcenia niestrukturalnych informacji osadzonych w tekstach w dane strukturalne. W przypadku treści pochodzących z różnych źródeł, należy je zintegrować, aby stworzyć spójny zbiór. Dane warto sklasyfikować i poddać kodowaniu, czyli przetworzyć na wartość liczbową, dzięki czemu można skrócić zapis, który będzie czytelny dla programów statystycznych. W przypadku tworzenia bazy danych istotne jest nadawanie numerów porządkowych (ID), na które zakładany jest tzw. klucz. Pozwala to znacznie przyśpieszyć pracę na zbiorze. Na tabelach właściwych warto stworzyć tzw. widoki, co pozwala ochronić dane właściwe przed przypadkowym błędem ze strony analityków pracujących na bazie.

5. Analiza i modelowanie danych oraz ich interpretacja

   Odpowiednio przygotowane dane mogą zostać poddane analizom statystycznym i ekonometrycznym, które pozwolą na określenie wzorców i zależności. Badania mogą być zarówno ilościowe, jak i jakościowe, pozwalając na określenie trendów na rynku pracy czy tworzenie predykcji. Oprócz samego uzyskania wyników warto również zastanowić się nad oczekiwaną formą wizualizacji, która pozwoli na przystępne zapoznanie się z ich treścią. Wytworzone informacje stanowią podstawę do oceny aktualnej i prognozowanej sytuacji, dlatego warto zwrócić również uwagę na ich wiarygodność, chociażby w kontekście statystycznej istotności wyników.

6. Rozpowszechnianie wyników

   Uzyskane wyniki mogą stanowić źródło informacji dla urzędu pracy, ale również stanowić istotne materiały dla innych podmiotów czy instytucji. Jest to podstawa do tworzenia raportów, prezentacji, informatorów czy innych publikacji, które mogą być udostępniane publicznie lub wybranym interesariuszom na rynku pracy, w szczególności pracodawcom, agencjom pracy czy placówkom edukacyjnym, zarówno w kontekście dostosowywania programów nauczania, jak i jako materiał badawczy do dalszych analiz rynku pracy.

7. Archiwizacja danych

   Istotnym elementem jest odpowiednie przechowywania danych w tym surowych danych wejściowych, metadanych czy wyników przetwarzania. Należy zapewnić ich integralność oraz dostępność w przyszłości, pozwalając na ich dalsze wykorzystanie. Możliwe jest również tworzenie kopii zapasowych. Danym należy zapewnić bezpieczeństwo. Osoby niepowołane nie powinny mieć do nich dostępu. Pracownicy nie powinni mieć możliwości

wykonania kopii danych na zewnętrzny nośnik, a ewentualne połączenia internetowe muszą być odpowiednio zabezpieczone.

8. Ocena jakości danych i wyników

   Cały proces powinien podlegać ewaluacji, dzięki czemu może być doskonalony, jeżeli efekty nie są zadawalające. Dotyczy to zarówno jakości pozyskanych danych, poprawności metodologii, efektywności wykorzystywanych technologii czy prawidłowego wyboru metod analizy. Ewaluacja jest szczególnie istotna po wdrożeniu nowych procedur czy działań, ale warto rozważyć również wprowadzenie cyklicznej oceny.

3.2.3. Szczegółowy opis procesu gromadzenia danych

Gromadzenie rozpoczyna się od wyboru źródeł, z których następnie pozyskiwane są dane. Pozyskiwanie polega na identyfikowaniu ofert pracy dostępnych w danym źródle i pobieranie ich treści. Dostęp do danych można uzyskać z frontendów oraz backendów serwisu. W przypadku stron internetowych określa się część „widoczną" dla użytkowników i obejmuje technologie uruchamiane w przeglądarce, natomiast backend to skrypty uruchamiane po stronie serwera (systemy i bazy danych zasilające stronę).

Wśród sposobów pozwalających na dostęp do treści można wskazać:

- web scraping;
- web crawling;
- bezpośredni dostęp za pośrednictwem interfejsu programowania aplikacji (API).

1. Web scraping

   Umożliwia wyodrębnianie danych strukturalnych ze stron internetowych. Zastosowanie tej metody wymaga, żeby dane były ustrukturyzowane na stronie i znana jest pozycja każdego pola, zawierającego określoną informację. Narzędzie jest programowane dla konkretnego serwisu, więc jest to najlepsze rozwiązanie dla portali zawierających wiele ofert pracy.

2. Web crawling

   Jest to wykorzystanie zaprogramowanego bota indeksującego, który systematyczne przeglądaniu i pobieraniu stron. Jest to technika operująca na wyższym poziomie ogólności niż web scraping oraz łatwiejsze do opracowania. Jednak wiąże się to z pobieraniem szerszego zakresu treści, a tym samym większej ilości nieistotnych informacji, które wymagają

usunięcia. Zatem proces czyszczenia danych jest bardziej rozbudowany niż w przypadku web scrapingu.

3. Bezpośredni dostęp za pośrednictwem interfejsu programowania aplikacji (API)

   Umożliwia pobieranie treści bezpośrednio z baz danych strony internetowej (niedostępnej dla użytkownika). Dane uzyskane w ten sposób mają najwyższą jakość, bowiem są pobierane u źródła, jednak dostęp wymaga formalnej zgody i umowy z administratorem strony, co może wiązać się z dodatkowymi kosztami finansowymi i administracyjnymi.

Web scraping i web crawling bazują na wykorzystaniu botów pobierających treści ze stron. Ich zaletą jest brak konieczności czynienia ustaleń z właścicielem czy administratorem strony, jednak mogą oni podejmować środki utrudniające takie działania. Z jednej strony może to być podyktowane ochroną treści – przeciwdziałaniu kopiowania i wykorzystania np. przez konkurencyjne portale. Z drugiej zaś może służyć zapobieganiu przeciążeniom serwerów w wyniku zbyt dużej liczby i częstotliwości żądań, co może wpływać na wydajność i dostępność witryny dla faktycznych użytkowników. Przed przystąpieniem do zbierania danych warto sprawdzić regulamin portalu, w którym mogą być określone kwestie z tym związane. Jednocześnie wiele stron posiada plik robots.txt wskazujący zasady indeksowania przez boty (np. dozwolony zakres, częstotliwość), którego należy przestrzegać.

Istnieją różne metody ograniczające web scraping i web crawling. Obejmują one m.in.

- limitowanie częstotliwości żądań (rate limiting);
- zmienianie kodu HTML lub korzystanie z dynamicznego generowania zawartości;
- test CAPTCHA;
- pułapki „honey pot”;
- identyfikowanie automatycznego ruchu.

1. Limitowanie częstotliwości żądań (rate limiting)

   Pozwala ograniczyć liczbę żądań, które można przeprowadzić z jednego adresu IP w określonym czasie (np. sekundy lub minuty). Po przekroczeniu limitu generowany jest błąd, dzięki czemu działanie bota jest spowalniane. Wykorzystanie zmiennego IP (np. poprzez VPN) może pomóc w uniknięciu tego ograniczenia.

2. Zmienianie kodu HTML lub korzystanie z dynamicznego generowania zawartości

   Boty bazują na wzorcach, uczą się struktury strony i postępują według ustalonego schematu. Częsta zmiana kodu lub wykorzystanie dynamicznego generowania zawartości (np. z użyciem JavaScript) może utrudnić analizę i znalezienie interesującej informacji, powodując braki danych. Dlatego narzędzie powinno być dostosowywane do struktury strony, którą należy monitorować w celu wykrycia ewentualnych zmian.

3. Test CAPTCHA

   Jest to technika wykorzystującą testy Turinga, które mają na celu odróżnienie ludzi od robotów. Zazwyczaj polega na przepisaniu określonego tekstu lub wybraniu obrazów o wskazanych charakterystykach w celu przejścia do kolejnej strony lub podstrony. Dostępne są narzędzia umożliwiające automatyczne rozwiązanie testu, wykorzystujące np. rozpoznawanie obrazów lub uczenie maszynowe, które należy dostosować do występującego ograniczenia.

4. Pułapki „honey pot”

   Pułapki to ukryte elementy na stronie, które nie są widoczne dla przeciętnego użytkownika przeglądającego stronę w przeglądarce, ale mogą znaleźć się na drodze bota. Może to być link, który, jeżeli zostanie kliknięty, wskazuje na wykorzystanie narzędzi indeksujących. Bot może zostać zidentyfikowany poprzez adres IP oraz możliwe jest monitorowanie jego aktywności, skutkujące np. wprowadzeniem wskazanych powyżej ograniczeń.

5. Identyfikowanie automatycznego ruchu

   Automatyczny ruch może być zidentyfikowany poprzez analizę „nieludzkich” wzorców w ruchu sieciowym, takich jak duży wolumen żądań z jednego adresu IP w krótkim czasie lub regularne żądania w stałych odstępach czasowych. Wykorzystuje się do tego np. techniki analizy statystycznej lub uczenia maszynowego. Analogicznie do pułapek, pozwala na ustalenie indeksowania i stanowi informację do podejmowania dalszych działań.

Należy jednak zaznaczyć, że nie istnieją metody, które całkowicie uniemożliwiłyby wykorzystanie web scrapingu czy web crawlingu. Jednakże utrudnienia w pozyskiwaniu informacji mogą powodować konieczność wykorzystania bardziej zaawansowanych narzędzi lub częstsze ich dostosowywanie i udoskonalanie. Istotne jest cykliczne monitorowanie stron, przeprowadzanie testowych indeksowań, których zadaniem byłoby identyfikowanie możliwych ograniczeń, a następnie uwzględnianie ich w projektowanych rozwiązaniach. Czynniki te mogą wpłynąć na zwiększenie

kosztu działań, jednak są niezbędne w celu osiągnięcia zamierzonych efektów i uzyskania miarodajnych danych. Wydaje się jednak, że techniki przeciwdziałające web scrapingowi i web crawlingowi nie są standardową praktyką w przypadku stron z ogłoszeniami o pracę, jednak należy poddać tę kwestię dalszej analizie. W szczególności przydatny może być kontakt z portalami, bowiem takie działania mogą nie być wymierzone w badania rynku pracy, a konkurentów lub mieć na celu zapewnienie odpowiednie wydajności strony dla jej odbiorców.

Dane pozyskiwane z różnych stron internetowych mogą się różnić jakością oraz zakresem pobranych informacji. Konieczne jest ich oczyszczenie z elementów, które nie są użyteczne w dalszej analizie, takich jak reklamy, prezentacje profilu pracodawcy czy inne niecelowo pobrane treści. Istotną kwestią jest również usuwanie tych samych ofert opublikowanych na różnych portalach. Na marginesie należy zaznaczyć, że informacja o preferencjach pracodawców w zakresie wyboru różnych stron może również być ciekawą podstawą do analiz. Identyfikowanie tożsamych ofert odbywa się przede wszystkim na podstawie treści, ale przydatne mogą być również metadane np. numery referencyjne, adresy URL.

Ważnym elementem procesu jest klasyfikowanie danych przy wykorzystaniu algorytmów uczenia maszynowego. Do tego celu konieczne jest przygotowanie ontologii czyli listy pojęć i kategorii wykorzystywanych w klasyfikacji. Celem algorytmu jest przypisanie ustalonych, wystandaryzowanych terminów do danego ogłoszenia na podstawie dopasowanie tekstu lub jego podobieństwa. Wśród podstawowych kategorii można wskazać zawód, poziom wykształcenia, określone umiejętności, branże. W przypadku zawodów adekwatne wydaje się zastosowanie systematyki zawartej w rozporządzeniu Ministra Pracy i Polityki Społecznej z 7 sierpnia 2014 r. w sprawie klasyfikacji zawodów i specjalności na potrzeby rynku pracy oraz zakresu jej stosowania[9]. Przy czym warto rozważyć dodatkowe wykorzystanie europejskiej klasyfikacji umiejętności/kompetencji, kwalifikacji i zawodów (ESCO) lub międzynarodowego standardu klasyfikacji zawodów (ISCO), zwłaszcza w kontekście badań porównawczych z regionami innych krajów. Proces uczenia maszynowego powinien być stosowany oddzielnie dla każdej kategorii. Model powinien zostać uprzednio przeszkolony na odpowiedniej liczbie ofert w ramach trzech faz: szkolenia, testowania wydajności oraz oceny dokładności. Kluczowe jest ponadto odpowiednie dopasowanie zbiorów treningowych oraz testowych, tak, by model odpowiednio dobrze radził sobie później z pracą na realnych danych. Zbiór testowy pozwala upewnić się, że trenowany model nie został nadmiernie dobrze dopasowany do zbioru treningowego. Okazać się bowiem może, że model świetnie sprawdzający się w prognozowaniu na zbiorze treningowym działa zupełnie przeciętnie na danych spoza zbioru treningowego.

[9] Dz.U. 2018 poz. 227 ze zm.

3.2.4. Zapewnienie reprezentatywności i stabilności źródeł danych

Reprezentatywność odnosi się do stopnia, w jakim dana próba, zbiór lub grupa jest adekwatnym odzwierciedleniem większej populacji, do którego należy. Jest to istotna kategoria w badaniach ekonomicznych i społecznych, bowiem wyniki uzyskane na podstawie próby reprezentatywnej można uogólniać na całą populację bez większego ryzyka wystąpienia błędów statystycznych. Aby próba była reprezentatywna powinna posiadać takie same lub zbliżone cechy, właściwości lub charakterystyki jak populacja, która ma być reprezentowana.

Ważne jest również rozróżnienie między reprezentatywnością próby a reprezentatywnością wyników. O ile próba może być reprezentatywna, to wyniki mogą nadal zawierać błędy pomiarowe lub inne czynniki, które wpływają na dokładność uogólnień na populację. Dlatego konieczne jest również stosowanie odpowiednich technik analizy danych w celu interpretacji wyników i oceny stopnia reprezentatywności.

Jak już zostało wskazane w ograniczeniach analizy ofert pracy, portale ogłoszeniowe mogą nie obejmować wszystkich ofert. W związku z tym może wystąpić problem związany z nadreprezentacją określonych rodzajów ofert. W celu ograniczenia ryzyka należy uwzględnić dywersyfikację źródeł danych. Najlepiej korzystać z kilku lub kilkunastu stron internetowych, zamiast opierać się na jednym źródle oraz uwzględnić alternatywne źródła np. zasób ofert służb zatrudnienia. Jednocześnie przy niewielkiej liczbie źródeł warto unikać portali specjalistycznych np. dedykowanych branży IT. Zwiększanie wielkości próby oraz różnicowanie źródeł jest podstawowym i najlepszym sposobem zapewnienia reprezentatywności.

Warto również przeprowadzać analizę jakości danych, obejmującej ocenę czy zawierają one informacje o różnych charakterystykach. Jednocześnie można je porównywać ze znanymi trendami na rynku pracy, dostępnymi prognozami czy badaniami dotyczącymi regionalnych rynków pracy o podobnych cechach.

Jeżeli w zbiorze danych występuje nadreprezentacja ofert o określonych charakterystykach, której nie można uzasadnić specyfiką rynku pracy lub innymi czynnikami i zachodzi podejrzenie, że zebrane informacje mogą nie być reprezentatywne można zastosować korektę wag. Wymaga ona jednak wiedzy o charakterystyce badanego rynku pracy, a wagi powinny zostać określone zgodnie z panującymi na nim warunkami.

Stabilność źródeł danych odnosi się do ich trwałości i niezmienności w czasie. W tym kontekście problematyczne mogą być zmiany interfejsu stron internetowych, ich struktury lub sposobu wyświetlania ofert pracy. Zmiany te mogą wpływać na sposób pobierania danych i prowadzić do niestabilności w analizie, np. poprzez pominięcie części istotnych informacji skutkującej brakami danych.

Podstawową kwestią jest odpowiedni wybór narzędzia. Wprawdzie istnieje wiele gotowych programów służących do web scrapingu czy web crawlingu, ale uznaje się, że najtrwalszą i najstabilniejszym sposobem jest ręczne pisanie kodu dostosowanego do danej strony internetowej i uwzględniającego jej specyfikę[10]. Dzięki temu można uzyskać narzędzie bardziej uodpornione na zmiany w budowie portalu, np. poprzez zastosowanie komend niekończących działania skryptu, gdy nie uda się pobrać określonych danych.

Niezależnie od odpowiedniego przygotowania, należy mieć świadomość, że w przypadku istotnego zmodyfikowania struktury strony internetowej konieczne będzie aktualizowanie kodu. Należy tę kwestię uwzględnić również w planowaniu budżetu, poprzez zabezpieczenie odpowiednich środków na działania dostosowawcze. Rezygnacja z aktualizacji i bazowanie na pierwotnie stworzonym narzędziu może prowadzić do pozyskiwania niekompletnych, niedokładnych lub przestarzałych danych, wpływając na wiarygodność i rzetelność analiz przeprowadzanych na ich podstawie.

W ramach procesu gromadzenia informacji należy uwzględnić monitorowanie zmian na stronach internetowych. Narzędzie do web scrapingu i web crawlingu może mieć wbudowaną funkcję, które ułatwią to zadanie np. rejestrowanie zdarzeń lub wysyłanie powiadomień w przypadku wystąpienia problemów. Warto sprawdzać cyklicznie plik robots.txt, mapę witryny lub kanał RSS w celu zidentyfikowania ewentualnych zmian w regułach indeksowania lub strukturze stron.

Monitorowanie stron może odbywać się ręcznie poprzez sprawdzanie układów wizualnych danych, przy pomocy zautomatyzowanych testów, dostępnych w ramach modułów standardowych bibliotek języków programowania wbudowanych w narzędzie do web scrapingu lub web crawlingu, jak również z wykorzystaniem specjalistycznych rozwiązań wyspecjalizowanych do takich działań dostarczanych przez zewnętrzne podmioty.

Jednym z podstawowych sposobów, przeprowadzanych w ramach web scrapingu lub web crawlingu, jest weryfikacja i walidacja pobieranych danych. Polega na stworzeniu zestawu testowego, który spełnia oczekiwania dotyczące jakości i zakresu pobieranych informacji, oraz porównywaniu go z wynikami uzyskiwanymi w kolejnych etapach gromadzenia. Dzięki temu można weryfikować różnice między nowo zebranymi danymi a danymi referencyjnymi.

Wykrycie zmian wpływających na pozyskiwanie informacji wiąże się z koniecznością aktualizacji i adaptacji stosowanego narzędzia. Trudno ocenić zakres dostosowań nie analizując konkretnego przypadku – będzie on zależał od zakresu i złożoności wprowadzonych na danej stronie modyfikacji. Może być wymagana np. zmiana

[10] Por. A. Juszczak, Zastosowanie danych scrapowanych w pomiarze dynamiki cen. Acta Universitatis Lodziensis. Folia Oeconomica 1.352, 2021, s. 25-37.

selektorów, atrybutów lub parametrów używanych do wyszukiwania i ekstrakcji danych, dodanie lub usunięcie nagłówków, plików cookie lub proxy używanych do wysyłania zapytań na stronę czy dostosowanie wzorców URL, paginacji lub logiki nawigacji stosowanej podczas indeksowania strony. W przypadku wdrożenia, wspomnianych wcześniej, środków przeciwdziałających automatycznemu pobieraniu treści trzeba będzie wprowadzić rozwiązywanie testów CAPTCHA czy ograniczenie tempa pobierania danych.

Ręczne modyfikacje narzędzi mogą być uciążliwe i czasochłonne, zwłaszcza gdy występują częste lub nieprzewidywalne zmiany na stronach. Dlatego tak istotne jest odpowiednie przygotowanie pierwotnego kodu, uwzględniając funkcje do obsługi takich zdarzeń, stosować budowę modułową (ograniczając zakres kodu wymagający modyfikacji), korzystać z dobrych praktyk programistycznych, odpowiednio dokumentować proces tworzenia narzędzia oraz zaplanować zadania związane z monitorowaniem.

3.2.5. Wykorzystanie danych w dalszych analizach

Stworzenie zbioru danych zawierającego informacje z ofert pracy stanowi punkt wyjścia do przeprowadzenia dalszych analiz. Wnioskowanie na ich podstawie wymaga zastosowania metod analitycznych, wśród których można wskazać:

- analizę eksploracyjną;
- modele ekonometryczne;
- modele uczenia maszynowego;
- analizę lokalizacyjną;
- text mining.

1. Analiza eksploracyjna

   Dostarcza podstawowych informacji o analizowanej bazie danych i poszczególnych zmiennych. Wykonuje się ją zwykle przed właściwym modelowaniem. Do popularnych charakterystyk eksploracyjnych różnego rodzaju zmiennych należą:

   - średnia;
   - mediana;
   - zakres;
   - grupy decylowe;
   - rozkład gęstości oraz jego charakterystyki, np. ewentualna skośność czy kurtoza;
   - liczba kategorii (w przypadku zmiennych poziomowych);
   - liczba braków danych;
   - korelacja między zmiennymi.

Analiza eksploracyjna, poza przedstawieniem podstawowych informacji o analizowanych danych, ułatwia podjęcie decyzji co do dalszego postępowania ze zbiorem przy budowie konkretnego modelu. Przykładowo, kwestia braków danych często wymaga od analityka decyzji, postąpić bowiem można na kilka sposobów. Jeżeli braków jest mało, można nie robić z nimi nic. Braki można też wypełnić np. średnią lub medianą pozostałych wartości. Innym powszechnym problemem jest postępowanie z outlierami. Outliery bywają powodowane np. błędami w konstrukcji zbioru danych (niekoniecznie z winy analityka) i mogą wypaczyć wyniki. Przykładowo, jeżeli 98 obserwacji mieści się w przedziale 0-100 a dwie wynoszą po 10 tysięcy, wówczas te dwie stanowią outliery, które może być warto usunąć z dalszej pracy.

W przypadku analizy ofert pracy, analiza eksploracyjna dostarcza informacji takich jak:

- liczba ofert danego typu;
- zakres widełek zarobków;
- poziom wymaganego wykształcenia;
- liczba ofert w danej lokalizacji.

Warto pamiętać, że wiele opracowań eksperckich prezentujących dane ekonomiczne poprzestaje na analizie eksploracyjnej, przedstawiając jej wyniki w formie graficznej. Prostota analizy eksploracyjnej powoduje, że łatwo trafia ona do odbiorcy posiadającego nawet minimalne przygotowanie merytoryczne.

2. Modele ekonometryczne

Ekonometria to podstawowa metoda badania związków przyczynowo-skutkowych między analizowanymi zmiennymi oraz prognozowania ich przyszłych wartości.

Modele ekonometryczne można rozdzielić na dwie kategorie. Pierwsza to analizy szeregów czasowych, które będzie można wykonać dopiero po zebraniu danych z odpowiedniej liczby okresów. Wyklucza to zatem zastosowanie tego rodzaju narzędzi zaraz po rozpoczęciu analizy danych przez urząd. W zależności od częstotliwości zbierania danych stosować można różne rodzaje modeli. Większość modeli ekonomicznych wykorzystuje dane kwartalne, ostatnio pojawiły się również makromodele konstruowane na danych miesięcznych (np. w zespole Prof. Welfe w Łodzi). Dane o wyższej częstotliwości są rzadko wykorzystywane w analizach ekonomicznych. Wyjątkiem są finanse, gdzie dane o częstotliwości dziennej, czy godzinowej są normą. Do opisywania zmienności takich szeregów wykorzystuje się modele klasy GARCH oraz zmienności stochastycznej.

Drugim rodzajem modeli są modele danych punktowych, tzw. mikrodanych (zgodnie z definicją Gruszczyńskiego). Dobór modelu zależy od typu zmiennej objaśnianej. Jednym z najpopularniejszych jest logit. W oryginalnej postaci służy on do modelowania zmiennej binarnej, natomiast szybko pojawił się tzw. logit wielomianowy, umożliwiający analizę zmiennych o kilku poziomach (factor). Model logitowy służy sprawdzeniu w jakim stopniu zmienne objaśniające wpływają na prawdopodobieństwo pojawienia się danej kategorii zmiennej objaśnianej. W odróżnieniu od klasyfikatorów w uczeniu maszynowym, konstrukcja modelu jest znana, można również zbadać wpływ poszczególnych zmiennych na zmienną objaśnianą. Ma to znaczenie np. w przypadku modeli credit scoringowych, które muszą być zatwierdzone przez Komisję Nadzoru Finansowego, a co za tym idzie, nie mogą być tzw. black boxem, gdzie znany jest tylko input oraz output, nie wiadomo jednak co dokładnie model wykonał z danymi.

Przykładowy model logitowy wykonywany na danych rynku pracy pozwoli określić, jakie czynniki i w jakim stopniu wpływają na znalezienie pracy. Z kolei klasyfikator zbudowany przy pomocy metod uczenia maszynowego może pomóc przewidzieć szanse kandydata na uzyskanie zatrudnienia, czy szanse na obsadzenie wakatu w określonym przedziale czasu. Więcej o modelach uczenia maszynowego poniżej.

3. Modele uczenia maszynowego

Mimo nowocześnie brzmiącej nazwy wiele modeli uczenia maszynowego znanych jest od lat 60. Obecny boom na IT i coraz większa moc obliczeniowa komputerów umożliwia coraz częstsze wykorzystanie uczenia maszynowego. Podkreślić należy, że modele uczenia maszynowego pracują na zasadzie czarnej skrzynki. Dane wsadowe są wprowadzane do modelu, który następnie przygotowuje wynik. Pracą narzędzia można sterować przy pomocy parametrów, różnych dla każdego modelu, które określają jak model się uczy[11].

[11] Modele uczenia maszynowego to najczęściej klasyfikatory. Bardzo popularne są zbiory drzew decyzyjnych tzw. lasy losowe (random forests). Technika ta, choć leciwa, jest poprawiana do dziś, np. w pracy Chena i Guestrina z 2016 gdzie zaprezentowali oni metodę extreme gradient boost, która poprawiła efektywność lasów losowych. W pakiecie caret do języka R zaprogramowano 238 różnych modeli, które można wykorzystać zarówno do klasyfikacji, jak i regresji. Modele regresyjne są rzadziej spotykane od klasyfikatorów, gdyż do badania zależności między szeregami czasowymi nadal dobrze sprawdza się ekonometria. Regresory uczenia maszynowego można jednak wykorzystać w sytuacji, w której nie można wyspecyfikować poprawnej postaci modelu ekonometrycznego, tj. tej, w której jest to, co niezbędne. W ekonometrii zmiennych objaśniających jest co najwyżej kilka, gdyż specyfikację najczęściej ustala się w oparciu o literaturę i hipotezy badawcze. Modele ekonometryczne nie skorzystają z informacji, jakie dać może rozszerzenie zbioru zmiennych objaśniających. Sztucznej zwiększenie liczby zmiennych w równaniu może wręcz utrudnić estymowanie parametrów i całą pracę analityka. Modele regresyjne uczenia maszynowego tymczasem służą do korzystania z dużej liczby zmiennych objaśnianych, bez podawania konkretnej specyfikacji.

Metody uczenia maszynowego pozwalają wykonać szereg prac przy analizie ofert pracy. Można na przykład wytrenować klasyfikator na podstawie tych bezrobotnych, którzy znaleźli już pracę i na tej podstawie profilować nowych kandydatów do konkretnego typu stanowisk albo informować ich, kiedy, mniej więcej, mogą spodziewać się zatrudnienia. Regresje uczenia maszynowego pozwolą wykonać prognozy w sytuacji, w której modelu ekonometrycznego nie można wyspecyfikować, co przy pobieraniu dużej ilości danych różnego typu może okazać się normą.

4. Analiza lokalizacyjna

   Umieszczenie w zbiorze danych informacji geograficznych otwiera przed analitykami cały obszar analiz przestrzennych. Podstawowe statystyki przestrzenne pozwalają np. sprawdzać grupowanie się poszczególnych rejonów w zależności od intensywności występowania danego zjawiska (tzw. hot and cold spots). Za pomocą innych statystyk mierzyć można tzw. efekty lokalne czyli ten fragment całościowego wzrostu w danej branży, w danym rejonie, który pozostaje po wprowadzeniu poprawki na ogólny wzrost gospodarczy w rejonie oraz na ogólny wzrost w branży. Niewielkie rozszerzenie tego rodzaju analiz umożliwia sprawdzenie specjalizacji i konkurencyjności.

   Ekonometrię przestrzenną wykorzystuje się do analizy danych panelowych. W tradycyjnym podejściu, różnice geograficzne „umieszczane” są albo w stałej równania albo uznaje się, że jest ona zawarta w składniku losowym (fixed and random effects). Ekonometria przestrzenna wykorzystuje macierze odległości, które następnie mnoży się z macierzami wartości zmiennych i w ten sposób buduje model.

   Informacje lokalizacyjne o danych rejonach można pozyskać z GIODO, a następnie połączyć z danymi zebranymi w trakcie pozyskiwania informacji o ofertach pracy. Następnie można je analizować i przedstawiać z uwzględnieniem informacji geograficznych.

5. Text mining

   Text mining, znany również jako analiza tekstu, to zbiór metod służących do ekstrakcji i analizy informacji zawartych w tekście. Proste metody text miningu umożliwiają badanie częstotliwości występowania poszczególnych wyrazów, natomiast bardziej zaawansowane techniki umożliwiają ocenę nacechowania emocjonalnego tekstu oraz analizę struktury zdaniowej.

   Warto zaznaczyć, że większość bibliotek i narzędzi text miningowych została opracowana głównie do analizy tekstów pisanych w języku angielskim, który cechuje się stosunkowo prostą strukturą. Natomiast konstrukcja języka

polskiego, szczególnie odmiana przez przypadki, wprowadza znaczne trudności w przeprowadzaniu analiz text miningowych.

Mimo tych trudności, text mining znajduje zastosowanie w wielu obszarach. Analiza tekstów ogłoszeń o pracę może dostarczyć cennych informacji dotyczących bonusów i wymagań stawianych kandydatom. Przeszukiwanie ogłoszeń danego typu (np. w ramach konkretnej branży i stanowiska) pod kątem powtarzających się wymagań pozwala urzędowi pracy tworzyć wzorcowe profile pożądanego pracownika. Dzięki temu możliwe staje się bardziej precyzyjne dopasowywanie ofert pracy do indywidualnych potrzeb poszukujących pracy, co przyczynia się do bardziej efektywnego procesu rekrutacji oraz podnoszenia kwalifikacji bezrobotnych zarejestrowanych w urzędzie.

Text mining może obejmować następujące etapy:

- przetwarzanie tekstu;
- analizę częstotliwości
- analizę sentymentu;
- klasyfikację i kategoryzację;
- wykrywanie związków i wzorców;
- modelowanie języka i uczenie maszynowe.

a) Przetwarzanie tekstu

   Pierwszym krokiem w text miningu jest przetworzenie tekstu w celu przygotowania go do analizy. Może to obejmować usuwanie znaków interpunkcyjnych, przekształcanie tekstu na małe litery, tokenizację (podział tekstu na pojedyncze wyrazy) oraz usuwanie słów stopu (często występujących, ale nieprzydatnych słów, takich jak "a", "i", "jest", itp.).

b) Analiza częstotliwości

   Prosta analiza częstotliwości polega na obliczaniu liczby wystąpień poszczególnych wyrazów w tekście. Może to pomóc zidentyfikować najważniejsze wyrazy lub tematy poruszane w tekście. Analiza częstotliwości może być również stosowana do analizy n-gramów, czyli sekwencji kolejnych słów.

c) Analiza sentymentu

   Zaawansowane metody text miningu umożliwiają ocenę nacechowania emocjonalnego tekstu. Dzięki temu można automatycznie określić, czy tekst jest pozytywny, negatywny czy neutralny. Analiza sentymentu jest

szczególnie przydatna w przypadku analizy opinii klientów, recenzji produktów lub nastrojów społecznych.

d) Klasyfikacja i kategoryzacja

Text mining umożliwia automatyczną klasyfikację tekstów na podstawie określonych kategorii. Może to być przydatne, na przykład, w analizie tematów poruszanych w wiadomościach lub klasyfikacji dokumentów według ich zawartości.

e) Wykrywanie związków i wzorców

Text mining pozwala również na odkrywanie ukrytych związków i wzorców w tekście. Można stosować metody analizy asocjacyjnej do identyfikowania często współwystępujących wyrazów lub fraz w tekście oraz analizy sekwencji do identyfikowanie sekwencji słów lub zdarzeń, które występują w określonym porządku.

f) Modelowanie języka i uczenie maszynowe

Zaawansowane techniki text miningu, takie jak modelowanie języka i uczenie maszynowe, umożliwiają budowanie modeli predykcyjnych na podstawie danych tekstowych. Dzięki nim można automatycznie generować tekst, przewidywać zachowania na podstawie analizy treści lub tworzyć systemy rekomendacyjne.

3.2.6. Kwestie techniczne związane z tworzeniem i prowadzeniem bazy danych

W kontekście tworzenia i prowadzenia bazy danych, istnieje wiele kwestii technicznych, które mają kluczowe znaczenie dla skutecznego zarządzania bazą danych i zapewnienia optymalnego funkcjonowania systemu. Wśród nich można wskazać:

- analizę potrzeb zamawiającego i technicznych;
- wybór odpowiedniego systemu zarządzania bazą danych (DBMS);
- projektowanie bazy danych;
- utworzenie fizycznej instancji bazy danych;
- konfigurację bazy danych;
- optymalizację bazy danych;
- zapewnienie bezpieczeństwa bazy danych;
- systematyczne monitorowanie bazy danych;
- skalowanie bazy danych;
- definiowanie hierarchii uprawnień;
- ustalanie uprawnień dostępu;
- prowadzenie audytu i monitorowanie zmian w bazie danych;

- zastosowanie środków technicznych i procedur wewnętrznych w celu ograniczenia możliwości nieuprawnionych zmian w bazie danych;
- szkolenie pracowników.

1. Analiza potrzeb zamawiającego i technicznych

   Przed rozpoczęciem procesu tworzenia, prowadzenia, utrzymywania i dostępu do bazy danych, konieczne jest przeprowadzenie dokładnej analizy potrzeb zamawiającego i technicznych związanych z big data. Obejmuje ona identyfikację celów, rodzajów danych do przechowywania, wymaganej skalowalności, wydajności i elastyczności schematu danych.

2. Wybór odpowiedniego systemu zarządzania bazą danych (DBMS)

   Po dokładnej analizie potrzeb należy dokonać wyboru odpowiedniego systemu zarządzania bazą danych (DBMS). W przypadku big data, konieczne jest uwzględnienie DBMS dostosowanych do obsługi dużych zbiorów danych, takich jak Apache Hadoop, Apache Cassandra czy MongoDB. Wybór powinien być oparty na kryteriach takich jak skalowalność, wydajność, elastyczność schematu i zdolność do obsługi dużej ilości danych.

3. Projektowanie bazy danych

   Projektowanie bazy danych w kontekście big data wymaga uwzględnienia specyficznych wymagań dotyczących struktury, modelu danych i sposobu przechowywania informacji. Elastyczne schematy danych, takie jak NoSQL lub rozproszony model bazy danych, często są bardziej odpowiednie dla big data. Projektowanie powinno uwzględniać analizę wymagań zamawiającego, aby stworzyć efektywną i dostosowaną do potrzeb strukturę bazy danych.

4. Utworzenie fizycznej instancji bazy danych

   Po zaprojektowaniu bazy danych należy utworzyć fizyczną instancję bazy danych w wybranym DBMS. W przypadku big data, może to obejmować wykorzystanie rozproszonych systemów plików, takich jak Hadoop Distributed File System (HDFS), lub platform przetwarzania danych, takich jak Apache Spark. Utworzenie bazy danych wymaga wdrożenia odpowiednich narzędzi i technologii zgodnie z zaprojektowaną strukturą.

5. Konfiguracja bazy danych

   Konfiguracja bazy danych dotyczy zarówno samych serwerów bazodanowych, jak i innych komponentów infrastruktury, takich jak klastry Hadoop czy systemy przetwarzania równoległego. W przypadku big data, konieczne jest uwzględnienie skalowalności, wydajności i dostępności danych. Konfiguracja

obejmuje ustawienie odpowiednich parametrów, narzędzi i technologii, aby zapewnić optymalną wydajność i dostępność danych.

6. Optymalizacja bazy danych

   Optymalizacja bazy danych ma na celu zoptymalizowanie przepustowości, wydajności i czasu odpowiedzi. W przypadku big data, optymalizacja może obejmować skalowanie poziome, partycjonowanie danych, indeksowanie i optymalizację zapytań. Wprowadzanie optymalizacji wymaga dogłębnego zrozumienia charakterystyki danych oraz wykorzystanie odpowiednich technik i narzędzi dostępnych w wybranym DBMS.

7. Zapewnienie bezpieczeństwa bazy danych

   Bezpieczeństwo bazy danych jest kluczowe dla ochrony poufności, integralności i dostępności danych. Obejmuje to mechanizmy autoryzacji i uwierzytelniania, zarządzanie uprawnieniami użytkowników, szyfrowanie danych, zarządzanie kluczami, audyt i zabezpieczenia przed zagrożeniami cybernetycznymi. W przypadku big data, zastosowanie odpowiednich środków bezpieczeństwa jest szczególnie istotne, biorąc pod uwagę duże zbiory danych i potencjalne ryzyka związane z ich przetwarzaniem i przechowywaniem.

8. Systematyczne monitorowanie bazy danych

   Systematyczne monitorowanie bazy danych jest niezbędne do identyfikacji problemów, optymalizacji wydajności i zapewnienia dostępności danych. Narzędzia monitorujące, takie jak Apache Ambari czy Cloudera Manager, mogą dostarczać metryki dotyczące wykorzystania zasobów, przepustowości, czasu odpowiedzi oraz stanu systemu. Regularne zadania konserwacyjne, takie jak tworzenie kopii zapasowych, zarządzanie wersjami oprogramowania oraz monitorowanie wydajności, są istotne dla utrzymania stabilnej i efektywnej bazy danych. Tworzenie kopii zapasowych bazy danych jest niezbędne dla zapewnienia ochrony danych przed utratą w przypadku awarii lub innych nieprzewidzianych zdarzeń. Ważne jest, aby stosować odpowiednie strategie tworzenia kopii zapasowych, uwzględniając częstotliwość, lokalizację i przechowywanie. Dodatkowo, konieczne jest przeprowadzanie regularnych testów przywracania danych, aby upewnić się, że proces przywracania działa poprawnie.

9. Skalowanie bazy danych

   Skalowanie bazy danych w przypadku big data jest kluczowym aspektem. Może obejmować zarówno skalowanie poziome (dodawanie nowych węzłów) jak i skalowanie pionowe (uaktualnienie sprzętu). Dodatkowo, technologie

takie jak Hadoop czy Spark zapewniają elastyczność skalowania poprzez możliwość dodawania nowych węzłów do klastra w celu obsługi większych zbiorów danych i zwiększenia przepustowości.

10. Definiowanie hierarchii uprawnień

    Aby zarządzać dostępem i ograniczyć możliwość zmiany bazy danych przez osoby mające do niej dostęp, konieczne jest definiowanie hierarchii uprawnień. W tym celu można zastosować role i uprawnienia, które określają, które osoby mają dostęp do poszczególnych tabel, pól i funkcji bazy danych. Hierarchia uprawnień pozwala kontrolować i ograniczać możliwość zmiany danych w bazie.

11. Ustalanie uprawnień dostępu

    Po zdefiniowaniu hierarchii uprawnień, konieczne jest ustalenie precyzyjnych uprawnień dostępu dla poszczególnych użytkowników lub grup użytkowników. Obejmuje to określenie, czy użytkownicy mają prawo do odczytu, zapisu, aktualizacji czy usuwania danych. Uprawnienia dostępu powinny być ściśle kontrolowane i przyznawane tylko tym osobom, które faktycznie ich potrzebują.

12. Prowadzenie audytu i monitorowanie zmian

    Audyt i monitorowanie zmian w bazie danych są istotne dla śledzenia operacji zmiany danych oraz identyfikacji potencjalnych zagrożeń i nieprawidłowości. Rejestrowanie operacji zmiany danych, analiza logów i monitorowanie aktywności użytkowników pomagają w wykrywaniu nieautoryzowanych zmian, błędów czy prób naruszenia bezpieczeństwa bazy danych.

13. Zastosowanie środków technicznych i procedur wewnętrznych

    Aby ograniczyć możliwość nieuprawnionych zmian w bazie danych, konieczne jest zastosowanie odpowiednich środków technicznych i procedur wewnętrznych. Przykłady obejmują dwuskładnikową autoryzację, uwierzytelnianie wielopoziomowe, stosowanie zasad najmniejszych uprawnień, regularne aktualizacje oprogramowania oraz audyty bezpieczeństwa.

14. Szkolenie pracowników

    Szkolenie pracowników oraz podnoszenie świadomości na temat zasad bezpieczeństwa bazy danych i procedur zarządzania jest kluczowe dla utrzymania kontroli nad dostępem i ograniczenia możliwości zmiany bazy. Pracownicy powinni być świadomi swoich obowiązków w zakresie

bezpieczeństwa danych, procedur i polityk dotyczących zarządzania bazą danych.

Wszystkie powyższe kwestie są istotne w kontekście tworzenia, prowadzenia, utrzymania i dostępu do bazy danych na poziomie organizacji, zwłaszcza w przypadku big data. Ich uwzględnienie i implementacja są niezbędne dla efektywnego zarządzania dużymi zbiorami danych i zapewnienia bezpieczeństwa oraz integralności informacji.

3.3. Propozycja badań jakościowych

W celu efektywnego przeprowadzenia zamierzonego projektu analiz ofert pracy, jak również wdrażania narzędzi big data w działaniach urzędu pracy warto przeprowadzić dodatkowe, pogłębione badania, które mogłyby obejmować:

- wybór optymalnego narzędzia do web scrapingu i web crawlingu;
- stosunek administratorów portali ogłoszeniowych do udostępniania danych;
- informacje istotne dla pracodawców, które mogłyby być pozyskiwane przy pomocy narzędzi big data;
- narzędzia big data wykorzystywane przez podmioty zajmujące się analizami rynku pracy;
- potencjalne implikacje etyczne i społeczne związane z wykorzystaniem narzędzi big data.

1. Wybór optymalnego narzędzia do web scrapingu i web crawlingu

   Pytania badawcze: Jakie narzędzie będzie optymalne z punktu widzenia zamierzonych celów, czy lepiej wykorzystać istniejące narzędzia, czy stworzyć nowe w ramach projektu dostosowane do zamierzonych działań, jakie języki programowania zapewnią najlepsze efekty, jakie konkretne rozwiązania warto wdrożyć w celu zapewnienia reprezentatywności i stabilności źródeł danych, jakie zasoby ludzkie i rzeczowe są niezbędne, żeby efektywnie gromadzić i analizować dane, jakie są różnice w kosztach w zależności od wybranego modelu.

   Kategorie uczestników: firmy programistyczne, niezależni programiści, ośrodki naukowe specjalizujące się w informatyce.

2. Stosunek administratorów portali ogłoszeniowych do udostępniania danych

   Pytania badawcze: czy portale są zainteresowane współpracą i udostępnianiem danych do celów badawczych, czy w ramach takich badań byłby możliwy bezpośredni dostęp za pośrednictwem interfejsu programowania aplikacji (API), z jakimi kosztami wiązałby się taki dostęp, jaki jest stosunek do pozyskiwania danych za pomocą web scrapingu i web

crawlingu, czy stosowane są jakieś ograniczenia związane z takim sposobem pozyskiwania danych.

Kategorie uczestników: administratorzy portali z ogłoszeniami o pracę.

3. Informacje istotne dla pracodawców, które mogłyby być pozyskiwane przy pomocy narzędzi big data

   Pytania badawcze: jakie dane, rodzaje analiz ułatwiłyby rekrutację pracowników, planowanie zatrudnienia oraz zarządzanie zasobami ludzkimi, czy możliwa byłaby współpraca obejmująca przekazywanie dodatkowych danych wewnętrznych do analiz urzędu pracy.

   Kategorie uczestników: pracodawcy z obszaru województwa lubelskiego, zarówno małe, średnie, jak i duże podmioty, prywatne firmy i instytucje publiczne.

4. Narzędzia big data wykorzystywane przez podmioty zajmujące się analizami rynku pracy

   Pytania badawcze: jakie narzędzia big data wykorzystywane są w bieżącej pracy, jak kształtują się trendy i jakie kategorie danych są najbardziej pożądane, jakie ograniczenia występują w ramach dotychczasowego stosowania narzędzi big data.

   Kategorie uczestników: ośrodki naukowe i prywatne podmioty zajmujące się analizami rynku pracy.

5. Potencjalne implikacje etyczne i społeczne związane z wykorzystaniem narzędzi big data

   Pytania badawcze: czy istnieją ryzyka wynikające z wykorzystania narzędzi big data do analiz rynku pracy, czy mogą się one przyczynić do powstawania nierówności społecznych lub dyskryminacji określonych grup osób, jak konstruować badania, żeby uniknąć ewentualnych ryzyk.

   Kategorie uczestników: ośrodki naukowe specjalizujące się w badaniach rynku pracy, kapitału społecznego, socjologii, etyce, filozofii.

4. Aspekty prawne wykorzystywania big data

Pojęcie big data czy dużych zbiorów danych nie jest bezpośrednim przedmiotem regulacji. Istnieje jednak wiele aktów, które swoim przedmiotem obejmują określone kategorie danych. Należy zatem rozróżnić dane chronione i dane neutralne z prawnego punktu widzenia. Ochrona danych może wynikać z aktu prawnego lub umowy. Jeżeli dane nie są objęte ochroną, można je uznać za neutralne, a ich wykorzystanie i przetwarzanie nie powinno wiązać się z wystąpieniem ryzyk prawnych.

Kluczową kategorią danych chronionych przez prawo są dane osobowe, stanowiące wszelkie informacje dotyczące zidentyfikowanej lub możliwej do zidentyfikowania żyjącej osoby fizycznej. Występują również dane objęte tajemnicą prawnie chronioną np. tajemnicą przedsiębiorstwa, bankową, skarbową, telekomunikacyjną czy zawodową. Dodatkowo zbiory danych mogą być objęte ochroną prawno-autorską lub prawem wyłącznym do baz danych.

Na uwagę zasługują również inicjatywy podejmowane w ramach Unii Europejskiej. Przy czym są one zgodne z ideą uwalniania danych i co do zasady mają się przyczynić do ich większej dostępności, aniżeli ograniczać ich wykorzystanie. Więc w kontekście wykorzystania big data w analizach nie powinny wiązać się z ryzykami, a dodatkowymi szansami.

W 2020 r. została opublikowana Europejska strategia w zakresie danych. Jej celem jest stworzenie w Europie jednolitej przestrzeni danych i prawdziwie jednolitego rynku otwartego na dane. W takim środowisku zarówno dane osobowe, jak i dane nieosobowe pozostaną bezpieczne, a przedsiębiorstwa będą miały łatwy dostęp do wysokiej jakości danych przemysłowych w niemal nieograniczonej ilości.

Przedstawiona koncepcja Komisji Europejskiej ma przyczynić się do wzrostu gospodarczego oraz tworzenia wspólnych wartości przy jednoczesnym minimalizowaniu śladu węglowego i środowiskowego człowieka. Zaplanowane w ramach strategii działania opierają się na czterech filarach, które wraz z kluczowymi działaniami zostały przedstawione w tabeli 5.

Tabela 5. Filary Europejskiej strategii w zakresie danych.

Filar	Kluczowe działania
Międzysektorowe ramy zarządzania w zakresie dostępu do danych i ich wykorzystywania	• zaproponowanie ram prawnych dotyczących zarządzania wspólnymi europejskimi przestrzeniami danych • przyjęcie aktu wykonawczego w sprawie zbiorów danych o wysokiej wartości • zaproponowanie aktu w sprawie danych • analiza znaczenia danych w gospodarce cyfrowej (np. za pośrednictwem Obserwatorium Gospodarki Platform Internetowych) oraz przegląd istniejących ram politycznych w kontekście pakietu związanego z aktem o usługach cyfrowych
Czynniki sprzyjające rozwojowi: inwestycje w dane oraz wzmocnienie zdolności i rozbudowa infrastruktury Europy na potrzeby hostingu, przetwarzania i wykorzystywania danych, interoperacyjność	• inwestycje w projekt o dużym oddziaływaniu dotyczący europejskich przestrzeni danych, obejmujący struktury wymiany danych (w tym standardy w zakresie wymiany danych, najlepsze praktyki, narzędzia) i mechanizmy zarządzania, a także w europejską federację energooszczędnej i godnej zaufania infrastruktury chmurowej i powiązanych usług • podpisanie protokołów ustaleń z państwami członkowskimi w sprawie federacji chmur obliczeniowych • uruchomienie europejskiego rynku usług w chmurze, obejmującego kompleksową ofertę usług w chmurze • stworzenie unijnego zbioru przepisów (samo)regulacyjnych dotyczących chmury obliczeniowej
Kompetencje: wzmocnienie pozycji osób fizycznych, inwestowanie w umiejętności i w MŚP	• zbadanie możliwości zwiększenia prawa do przenoszenia danych przez osoby fizyczne, zapewniając im większą kontrolę nad tym, kto może uzyskać dostęp do danych generowanych maszynowo i je wykorzystywać

Filar	Kluczowe działania
Wspólne europejskie przestrzenie danych w strategicznych sektorach i dziedzinach interesu publicznego	Utworzenie dziewięciu wspólnych europejskich przestrzeni danych: • przemysłowych (produkcyjnych) • dotyczących Zielonego Ładu • dotyczących mobilności • dotyczących zdrowia • finansowych • dotyczących energii • dotyczących rolnictwa • dla administracji publicznej • dotyczących umiejętności

Źródło: opracowanie własne na podstawie Europejskiej strategii w zakresie danych.

Jednocześnie na poziomie Unii Europejskiej przyjęto lub przygotowano kilka aktów związanych z tematyką danych, w szczególności danych nieosobowych. Wśród nich warto wskazać:

- Rozporządzenie w sprawie ram swobodnego przepływu danych nieosobowych w Unii Europejskiej[12];
- Akt w sprawie zarządzania danymi[13];
- Akt w sprawie danych[14].

1. Rozporządzenie w sprawie ram swobodnego przepływu danych nieosobowych w Unii Europejskiej

   Rozporządzenie weszło w życie w 2018 r. i miało na celu ograniczenie przeszkód w rozwoju opartej na danych obejmujących kwestie przepływu danych. W ramach aktu zakazano wprowadzania przez państwa członkowskie wymogów nakazujących lokalizację danych nieosobowych na terytorium danego państwa. Jednocześnie zadeklarowano wsparcie w opracowywaniu samoregulacyjnych kodeksów postępowania na poziomie Unii w zakresie przenoszenia danych nieosobowych, w szczególności między różnymi dostawcami usług przetwarzania danych.

[12] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2018/1807 z 14 listopada 2018 r. w sprawie ram swobodnego przepływu danych nieosobowych w Unii Europejskiej.

[13] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2022/868 z 30 maja 2022 r. w sprawie europejskiego zarządzania danymi i zmieniające rozporządzenie (UE) 2018/1724 (akt w sprawie zarządzania danymi).

[14] Wniosek dotyczący rozporządzenia Parlamentu Europejskiego i Rady w sprawie zharmonizowanych przepisów dotyczących sprawiedliwego dostępu do danych i ich wykorzystywania (akt w sprawie danych).

2. Akt w sprawie zarządzania danymi

Akt wszedł w życie w 2022 r., ale zacznie obowiązywać od września 2023 r. Jego celem jest zwiększenie zaufania do wymiany danych, wzmocnienie mechanizmów zwiększających dostępność danych i przezwyciężenie przeszkód technicznych utrudniających ponowne wykorzystywanie danych.

Regulacja wprowadza mechanizm umożliwiający ponowne wykorzystanie niektórych kategorii danych z sektora publicznego, które są chronione szczególnie ze względu na ochronę danych osobowych, lecz także ochronę praw własności intelektualnej i tajemnicy handlowej. Zobowiązuje ona podmioty sektora publicznego do:

- zakazu (z określonymi wyjątkami) uzgodnień dotyczących wyłączności udostępniania danych;
- zapewnienie niedyskryminujących, przejrzystych, proporcjonalnych i obiektywnie uzasadnionych warunków dostępu do danych;
- zapewnienie zachowania chronionego charakteru danych np. poprzez zanonimizowanie lub poddawanie modyfikacji, agregowaniu lub przekształcaniu przy pomocy innej niż zanonimizowanie metody, zapobiegające ujawnieniu danych;
- zapewnienie bezpiecznego środowiska przetwarzania danych;
- zapewnienie zachowania poufności i nieujawniania danych, które stwarzają zagrożenie dla praw i interesów osób trzecich, a które pomimo wprowadzonych zabezpieczeń mogłyby się znaleźć w posiadaniu ponownego użytkownika.

Dodatkowo akt obejmuje przepisy regulujące warunki i zasady świadczenia usług pośrednictwa danych, zmierzające do zapewnienia większego bezpieczeństwa podmiotom udostępniającym dane. Podmioty świadczące takie usługi mają działać jako neutralni dostawcy, którzy jedynie pośredniczą w transakcjach, w związku z czym nie mogą wykorzystywać wymienianych danych do innych celów. Dodatkowo działalność pośrednictwa danych podlega zgłoszeniu do rejestru oraz jest monitorowana i nadzorowana przez właściwe organy.

Akt wprowadza również rejestr uznanych organizacji o altruistycznym podejściu do danych, który ma zwiększyć zaufanie do tych podmiotów. Altruistyczne podejście do danych opiera się na dobrowolnym udostępnianiu danych przez osoby fizyczne lub przedsiębiorstwa dla wspólnego dobra. Podmioty prowadzące taką działalność mogą się zarejestrować się, aby zwiększyć zaufanie do swojej działalności.

3. Akt w sprawie danych

   Regulacja jest obecnie w procesie legislacyjnym i obejmuje ona zasady udostępniania danych. Wśród proponowanych rozwiązań znalazły się:

   - umożliwienie użytkownikom urządzeń podłączonych do internetu uzyskania dostępu do generowanych przez nie danych, udostępniania tych danych osobom trzecim w celu świadczenia usług posprzedażowych lub innych innowacyjnych (podmioty trzecie będą uprawnione do użytkowania danych tylko w celu, który został uzgodniony z użytkownikiem, później musza je usunąć, nie będzie można też przekazać tych danych dalszym osobom);
   - zrównoważenie pozycji MŚP przez zapobieganie nadużyciom z nierównowagi kontraktowej w umowach, których przedmiotem jest udostępnianie danych (posiadacz danych będzie mógł żądać wynagrodzenia za udostępnienie danych, jednakże w przypadku mikro, małych lub średnich przedsiębiorstw, opłata ta nie może wynosić więcej niż koszty bezpośrednio związane z udostępnieniem danych);
   - umożliwienie organom sektora publicznego dostępu do danych, będących w posiadaniu podmiotów prywatnych, które są niezbędne w wyjątkowych okolicznościach, zwłaszcza w przypadku sytuacji nadzwyczajnych (jak powodzie czy pożary lasów), albo do wykonania uprawnień, jeżeli pozyskanie danych nie jest inaczej możliwe;
   - umożliwienie klientom skutecznej zmiany dostawców usług przetwarzania danych w chmurze oraz wprowadzenie zabezpieczeń przed niezgodnym z prawem przekazywaniem danych.

4.1. Aspekty wykorzystania danych prawnie chronionych

Dane są obecnie niezwykle cennym zasobem, jednak pomimo dążenia do ich jak najszerszego udostępnienia, część z nich jest chroniona, ograniczając możliwość ich gromadzenia czy przetwarzania. Prawne nadanie ochrony może wynikać z wielu przyczyn, przede wszystkim ze względu na znaczenie czy wrażliwość danych, a tym samym z uwagi na szkody, które mogą powstać w przypadku ich ujawnienia, nieuprawnionego dostępu, kradzieży, manipulacji czy utraty. Wśród najważniejszych przyczyn ochrony danych należy wskazać:

- prywatność;
- bezpieczeństwo;
- ochronę interesów;
- prawa autorskie;
- własność intelektualną.

1. Prywatność

   Prawo do prywatności wiąże się z ochroną danych osobowych, Jest to jedno z podstawowych praw człowieka respektowane w większości jurysdykcji na świecie. Tradycyjnie obejmuje ono poszanowanie życia prywatnego i rodzinnego, domu i komunikowania się. Jednak obecnie rozumie się przez nie również autonomię informacyjną, czyli prawo do decydowania o ujawnieniu informacji o sobie i do kontroli nad tymi informacjami

2. Bezpieczeństwo

   Ochrona danych może się również wiązać z zapewnieniem bezpieczeństwa. Chociażby tajemnice państwowe czy służbowe urzędników służą zachowaniu poufności danych istotnych z punktu widzenia racji stanu. Jednocześnie tajemnice zawodowe np. bankowa lub odnosząca się do transakcji na giełdach towarowych, mogą być istotne przy zapewnieniu bezpieczeństwa obrotu i stabilności rynków.

3. Ochrona interesów

   Dane w szczególności takie jak tajemnice handlowe, know-how czy plany rozwoju są istotną wartością niematerialną, która może mieć jednak przełożenie na wyniki osiągane przez danego przedsiębiorcę czy jego pozycję na rynku. Celem tajemnicy przedsiębiorstwa jest ochrona przed nieuczciwą konkurencją, wykorzystaniem informacji w szpiegostwie gospodarczym czy wyciekom informacji, które mogą zaszkodzić interesom danego podmiotu.

4. Prawa autorskie

   Ochrona prawno-autorska wyraża przekonanie, iż autor powinien mieć zapewnioną kontrolę nad swoim dziełami oraz ich wykorzystaniem, a także możliwość czerpania korzyści ze swojej twórczości. Należy przy tym zaznaczyć, że o ile autor może zbyć prawa majątkowe, o tyle prawa osobiste są niezbywalne.

5. Własność intelektualna

   Ochrona własności intelektualnej ma na celu zachęcenie do innowacji, badań naukowych i rozwijania nowych rozwiązań. Działania te wymagają czasu, nakładów finansowych i pracy, dlatego brak ochrony mógłby zniechęcać do ich podejmowania, ograniczając możliwość czerpania korzyści z wytworzonych efektów.

Wykorzystanie danych chronionych wiąże się zazwyczaj z dodatkowymi obowiązkami, w szczególności takimi jak uzyskanie zgody lub zapewnienie

odpowiedniego poziomu bezpieczeństwa. Jednocześnie nieuprawnione przetwarzanie danych może wiązać się z konsekwencjami prawnymi np. karą finansową, ale również odpowiedzialnością karną. Dlatego też niezwykle istotna jest świadomość, które informacje objęte są ochroną, jaki jest jej zakres i jakie obowiązki się z nimi wiążą. Jednocześnie tworząc bazę danych należy mieć świadomość o charakterze danych jakie są w niej gromadzone.

4.1.1. Dane osobowe

Ochrona danych osobowych jest w ostatnim czasie tematem bardzo głośnym i popularnym. Główne regulacje w tym zakresie to unijne ogólne rozporządzenie o ochronie danych osobowych, tzw. RODO[15], oraz krajowa ustawa z 10 maja 2018 r. o ochronie danych osobowych[16]. Akty mają charakter komplementarny – ustawa doprecyzowuje i dostosowuje przepisy rozporządzenia do polskiego porządku prawnego. Zgodnie z RODO przez dane osobowe rozumie się informacje o osobie zidentyfikowanej lub osobie, którą można bezpośrednio lub pośrednio zidentyfikować, w szczególności na podstawie identyfikatora takiego jak imię i nazwisko, numer identyfikacyjny, dane o lokalizacji, identyfikator internetowy lub jeden bądź kilka szczególnych czynników określających fizyczną, fizjologiczną, genetyczną, psychiczną, ekonomiczną, kulturową lub społeczną tożsamość osoby fizycznej. Reasumując, dane osobowe stanowią informację, która pozwala na identyfikację tożsamości danej osoby.

W kontekście analiz big data istotne jest również pojęcie przetwarzania danych, które jest ujęte bardzo szeroko. Obejmuje ono operację lub zestaw operacji wykonywanych na danych osobowych lub zestawach danych osobowych w sposób zautomatyzowany lub niezautomatyzowany, taką jak zbieranie, utrwalanie, organizowanie, porządkowanie, przechowywanie, adaptowanie lub modyfikowanie, pobieranie, przeglądanie, wykorzystywanie, ujawnianie poprzez przesłanie, rozpowszechnianie lub innego rodzaju udostępnianie, dopasowywanie lub łączenie, ograniczanie, usuwanie lub niszczenie. Zatem, de facto dokonanie praktycznie dowolnej czynności związanej z posiadaniem lub analizą danych osobowych stanowi ich przetwarzanie.

Określenia zakresu obowiązków wymaga ustalenia roli, która jest pełniona. Najszerszą odpowiedzialność ma administrator, czyli podmiot, który samodzielnie lub wspólnie z innymi ustala cele i sposoby przetwarzania danych. To on odpowiada za zapewnienie odpowiedniego przetwarzania danych i spełnienia większości wymogów prawnych. Może również wystąpić podmiot przetwarzający, który wykonuje te

[15] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych).

[16] Dz.U. 2019 poz. 1781.

czynności na zlecenie administratora na podstawie zawartej umowy. Jest to zazwyczaj zewnętrzna firma świadczące usługi analityki big data. Podmiot przetwarzający jest związany zakresem umowy oraz poleceniami administratora, musi również stosować wymogi w zakresie bezpieczeństwa.

RODO określa sześć zasad ochrony danych osobowych, które muszą zostać spełnione przy przetwarzaniu danych. Zostały one przedstawione w tabeli 6.

Tabela 6. Zasady ochrony danych osobowych.

Zasada	Opis
Zgodność z prawem, rzetelność i przejrzystość	Dane muszą być przetwarzane zgodnie z prawem, rzetelnie i w sposób przejrzysty dla osoby, której dane dotyczą.
Ograniczenie celu	Dane muszą być zbierane w konkretnych, wyraźnych i prawnie uzasadnionych celach i nieprzetwarzane dalej w sposób niezgodny z tymi celami.
Minimalizacja danych	Dane muszą być adekwatne, stosowne oraz ograniczone do tego, co niezbędne do celów, w których są przetwarzane.
Prawidłowość	Dane muszą być prawidłowe i w razie potrzeby uaktualniane; należy podjąć wszelkie rozsądne działania, aby dane osobowe, które są nieprawidłowe w świetle celów ich przetwarzania, zostały niezwłocznie usunięte lub sprostowane.
Ograniczenie przechowywania	Dane muszą być przechowywane w formie umożliwiającej identyfikację osoby, której dane dotyczą, przez okres nie dłuższy, niż jest to niezbędne do celów, w których dane te są przetwarzane (dane można przechowywać przez okres dłuższy, o ile będą one przetwarzane wyłącznie do celów archiwalnych w interesie publicznym, do celów badań naukowych lub historycznych, do celów statystycznych, z zastrzeżeniem że wdrożone zostaną odpowiednie środki techniczne i organizacyjne w celu ochrony praw i wolności osób, których dane dotyczą).

Zasada	Opis
Integralność i poufność	Dane muszą być przetwarzane w sposób zapewniający odpowiednie bezpieczeństwo danych osobowych, w tym ochronę przed niedozwolonym lub niezgodnym z prawem przetwarzaniem oraz przypadkową utratą, zniszczeniem lub uszkodzeniem, za pomocą odpowiednich środków technicznych lub organizacyjnych.

Źródło: opracowanie własne na podstawie RODO.

Jak już wskazano, osobą odpowiedzialną za zapewnienie spełnienia powyższych zasad jest administrator.

Z zasady zgodności z prawem, rzetelności i przejrzystości wynika konieczność każdorazowego posiadania podstawy prawnej do przetwarzania danych. RODO wymienia następujące:

- zgodę wyrażoną dobrowolnie, konkretnie, świadomie i jednoznacznie;
- umowę, jeżeli jej stroną jest osoba, której dane dotyczą, a przetwarzanie jest niezbędne do jej wykonania;
- obowiązek prawny ciążący na administratorze, jeżeli do jego wykonania jest niezbędne przetwarzanie;
- żywotne interes osoby, której dane dotyczą, lub innej osoby fizycznej, jeżeli do ich ochrony jest niezbędne przetwarzanie;

- zadanie realizowane w interesie publicznym lub w ramach sprawowania władzy publicznej powierzonej administratorowi, jeżeli do jego wykonania jest niezbędne przetwarzanie;
- prawnie uzasadnione interesy realizowane przez administratora lub przez stronę trzecią, z wyjątkiem sytuacji, w których nadrzędny charakter wobec tych interesów mają interesy lub podstawowe prawa i wolności osoby, której dane dotyczą, wymagające ochrony danych osobowych, w szczególności gdy osoba, której dane dotyczą, jest dzieckiem (nie ma zastosowania do organów publicznych w ramach realizacji ich działań).

Odnosząc wskazane kategorie do potencjalnych badań rynku pracy realizowanych przez urząd pracy wykorzystujących dane osobowe, wydaje się, że w większości przypadków wymagana będzie zgoda osób, których dane miałyby być przetwarzane. Wprawdzie urząd może realizować zadania w interesie publicznym, jednak wątpliwości budzi niezbędność danych osobowych do tego celu. Zdecydowana większość nie wymaga informacji pozwalających na identyfikację konkretnych jednostek. Jednak można wyobrazić sobie szerokie badania, obejmujące różne dane (np. podatkowe, dotyczące zabezpieczenia społecznego), gdzie na wstępnym etapie

może pojawić się jakiś zakres danych osobowych, pozwalających na połączenie informacji z różnych źródeł. Jednakże w każdej sytuacji, kiedy miałoby dojść do przetwarzania danych osobowych bez uzyskania zgody, należy zasięgnąć porady prawnej oceniającej ryzyka takiego działania.

Jednocześnie w kontekście planowanego projektu analizy internetowych ofert pracy gromadzone informacje nie powinny co do zasady zawierać danych osobowych. Istnieje jednak ryzyko, zwłaszcza w odniesieniu do ofert zamieszczanych na platformach społecznościowych, że takie dane zostaną przypadkowo pobrane. Należy zatem odpowiednio dostosować narzędzia, żeby zakres zbieranych danych był adekwatny do oczekiwanego celu i ograniczał się jedynie do treści ofert.

W przypadku innych badań warto zwrócić uwagę na kwestię anonimizacji danych. Jest to proces, powodujący brak możliwości identyfikacji osób, których dane dotyczą. Odpowiednie przeprowadzenie anonimizacji powoduje, że administrator nie ma technicznych możliwości odwrócenia tego procesu i przywrócenia pełnych danych. Co istotne, dane zanonimizowane nie podlegają przepisom RODO. Należy jednak odróżnić anonimizację od pseudoanonimizacji. Drugi z procesów jest bowiem odwracalny, dzięki oddzielnie przechowywanym kluczom. W takiej sytuacji dane dalej pozwalają (za pomocą klucza) na identyfikację osoby, której dotyczą, tym samym mają do nich zastosowanie regulacje dotyczące ochrony danych osobowych.

4.1.2. Dane objęte tajemnicą

W przypadku analiz big data dokonywanych przez urząd pracy większość tajemnic prawnie chronionych nie będzie miało zastosowania, bowiem są one adresowane do konkretnych podmiotów, zazwyczaj określanych przez zawód albo miejsce pracy (np. bank). Chociażby w przypadku tajemnicy skarbowej to na pracownikach izby skarbowej czy innych wskazanych osobach będzie spoczywał obowiązek jej przestrzegania. Zatem w sytuacji, w której doszłoby do przekazania danych to te podmioty będą odpowiedzialne za ich odpowiednie przygotowanie (np. zanonimizowanie), aby nie zawierały informacji objętych tajemnicą.

Jednak w celu rozwiania ewentualnych wątpliwości, warto wskazać na tajemnicę przedsiębiorstwa, zdefiniowaną w ustawie z 16 kwietnia 1993 r. o zwalczaniu nieuczciwej konkurencji[17]. Obejmuje ona informacje techniczne, technologiczne, organizacyjne przedsiębiorstwa lub inne informacje posiadające wartość gospodarczą, które jako całość lub w szczególnym zestawieniu i zbiorze ich elementów nie są powszechnie znane osobom zwykle zajmującym się tym rodzajem informacji albo nie są łatwo dostępne dla takich osób, o ile uprawniony do korzystania z informacji lub rozporządzania nimi podjął, przy zachowaniu należytej staranności, działania w celu utrzymania ich w poufności.

[17] Dz.U. 2022 poz. 1233.

Jedną z istotnych cech tajemnicy przedsiębiorstwa jest podjęcie działań w celu utrzymania ich w poufności. Zatem pobieranie różnych informacji z publicznie dostępnych stron internetowych nie powinno generować ryzyk, bowiem skoro zostały upublicznione przeczy to przesłance utrzymania poufności.

4.1.3. Ochrona prawno-autorska i ochrona sui generis baz danych

Przedmiotem ochrony mogą być nie tylko określone informacje, ale również całe bazy danych. Zgodnie z ustawą z 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych[18], baza danych może być przedmiotem praw autorskich, jeżeli spełnia cechy utworu. Przy czym może ona zawierać niechronione materiały, jeżeli przyjęty w nich dobór, układ lub zestawienie ma twórczy charakter. Zatem baza danych, żeby podlegała ochronie, musi być przejawem działalności twórczej o indywidualnym charakterze, ustalony w jakiejkolwiek postaci, niezależnie od wartości, przeznaczenia i sposobu wyrażenia. Należy przy tym zaznaczyć, że ochrona ta obejmuje strukturę bazy, a nie jej zawartość.

W kontekście baz danych tworzonych na potrzeby big data, gromadzenie danych ma co do zasady charakter automatyczny. Trudno zatem mówić o indywidualnym charakterze czy przejawie działalności twórczej. Jednakże nie można a priori przesądzać, że wszystkie bazy obejmujące duże zbiory danych automatycznie nie podlegają prawu autorskiemu. Utworzenie niestandardowej bazy, według nowej koncepcji, ze specyficzną strukturą dostosowaną na potrzeby danego projektu może spełniać przesłanki utworu. Zatem ocen należy dokonywać na gruncie konkretnych przypadków, uwzględniając charakterystykę danej bazy.

Twórcy bazy danych przysługują autorskie prawa osobiste (niezbywalne) i autorskie prawa majątkowe. Prawa osobiste obejmują:

- autorstwo utworu;
- oznaczenia utworu nazwiskiem lub pseudonimem twórcy albo do udostępniania go anonimowo;
- nienaruszalność treści i formy utworu oraz jego rzetelnego wykorzystania;
- decydowanie o pierwszym udostępnieniu utworu publiczności;
- nadzoru nad sposobem korzystania z utworu.

Natomiast prawa majątkowe stanowią wyłączne prawo do korzystania z utworu i rozporządzania nim na wszystkich polach eksploatacji oraz do wynagrodzenia za korzystanie z utworu. Te prawa są zbywalne i mogą zostać przeniesione na inny podmiot, jednak niezależnie od dokonania takiej transakcji prawa osobiste zawsze będą przysługiwać twórcy.

[18] Dz.U. 2022 poz. 2509.

Jeżeli utwór zostanie rozpowszechniony, można z niego nieodpłatnie korzystać bez zgody twórcy w zakresie własnego użytku osobistego. Jednak nie uprawnia to do korzystania z elektronicznych baz danych (spełniających cechy utworu), chyba że dotyczy to własnego użytku naukowego niezwiązanego z celem zarobkowym.

Należy również zaznaczyć, że ochronie nie podlega oprogramowanie używane do sporządzenia lub obsługi baz danych dostępnych przy pomocy środków elektronicznych.

Ochrona zawartości bazy danych jest możliwa na podstawie ustawy z 27 lipca 2001 r. o ochronie baz danych[19] i jest niezależna od uprawnień wynikających z prawa autorskiego. W rozumieniu ustawy baza danych oznacza zbiór danych lub jakichkolwiek innych materiałów i elementów zgromadzonych według określonej systematyki lub metody, indywidualnie dostępnych w jakikolwiek sposób, w tym środkami elektronicznymi, wymagający istotnego, co do jakości lub ilości, nakładu inwestycyjnego w celu sporządzenia, weryfikacji lub prezentacji jego zawartości.

Również w tym przypadku pojawiają się wątpliwości, czy ochrona obejmie bazy danych przeznaczone dla big data. Pierwszą kwestią jest wymóg określonej systematyki lub metody. Gromadzenie danych nieustrukturyzowanych, mimo że podlega pewnym zasadom, jednak może nie mieć sprecyzowanej metodyki. Informacje nie muszą być uporządkowane, mogą mieć różne formaty.

Jednocześnie występuje konieczność poniesienia istotnych nakładów w celu sporządzenia, weryfikacji lub prezentacji. Bazy danych na potrzeby big data zazwyczaj nie obejmują weryfikacji czy prezentacji, a ich głównym celem jest gromadzenie informacji. Wizualizacja danych odbywa się na kolejnych etapach przetwarzania, często przy zastosowaniu dodatkowych programów czy pakietów. Istotne jest zatem ustalenie znaczenia pojęcia „sporządzanie”. Zgodnie z wyrokiem Trybunału Sprawiedliwości Unii Europejskiej[20] ochrona sui generis przysługuje bazom danych, które zawierają dane już istniejące. Bowiem koszty poszukiwania i gromadzenia informacji stanowią nakłady inwestycyjne. Natomiast nakłady poniesione na stworzenie zawartości bazy nie można uznać za inwestycję.

Mając na uwadze wskazane wątpliwości, ustalenie czy dana baza danych jest objęta ochroną sui generis wymaga uwzględnienia jej specyfiki. Wydaje się jednak, że ten typ ochrony może mieć szersze zastosowanie do baz danych tworzonych na potrzeby big data niż ochrona prawno-autorska.

[19] Dz.U. 2021 poz. 386.

[20] Wyrok TSUE z 9 listopada 2005 r. Fixtures Marketing Ltd przeciwko Organismos prognostikon agonon podosfairou AE (OPAP), sygn. akt C-444/02.

Analogicznie jak w przypadku prawa autorskiego, ochrona przyznana bazom danych nie obejmuje programów komputerowych użytych do sporządzenia baz danych lub korzystania z nich.

W ramach ochrony sui generis producentowi bazy danych, czyli podmiotowi ponoszącemu ryzyko nakładu inwestycyjnego przy tworzeniu bazy, przysługuje wyłączne i zbywalne prawo pobierania danych i wtórnego ich wykorzystania w całości lub w istotnej części, co do jakości lub ilości. Jeżeli jednak baza została udostępniona publicznie, nie można zabronić użytkownikowi, korzystającemu z niej zgodnie z prawem, pobierania lub wtórnego wykorzystania w jakimkolwiek celu nieistotnej, co do jakości lub ilości, części jej zawartości. Natomiast korzystanie z istotnej części zawartości jest możliwe na osobisty użytek osobistego, ale tylko z zawartości nieelektronicznej bazy danych, ewentualnie w określonych celach dydaktycznych lub badawczych. Czas ochrony bazy danych wynosi 15 lat.

5. Zakończenie

Rozwój cyfrowego świata zapewnia dostęp do najnowocześniejszych technologii, które tworzą się na naszych oczach. Możliwości analityczne stają się coraz większe, pozwalając jeszcze lepiej poznawać i opisywać świat. Należy korzystać z potencjału big data, pamiętając jednak o jego ograniczeniach. Wykorzystanie nowoczesnych technologii i narzędzi w analizie regionalnego rynku pracy stanowi idealne uzupełnienie dotychczasowych metod badawczych. Pozyskanie dodatkowych informacji może przyczynić się do lepszego zrozumienia dynamiki rynku, identyfikacji trendów oraz podejmowania trafnych decyzji.

Wprowadzenie analiz big data do badań rynku pracy umożliwia gromadzenie, przetwarzanie i interpretację ogromnych ilości danych z różnych źródeł. Dzięki temu można uzyskać bardziej szczegółowe i aktualne informacje dotyczące struktury zatrudnienia, preferencji zawodowych, umiejętności poszukiwanych przez pracodawców czy nawet prognozować przyszłe trendy rynkowe. To nieocenione narzędzie, które umożliwia podejmowanie lepiej ugruntowanych decyzji strategicznych, zarówno dla instytucji publicznych, jak i prywatnych przedsiębiorstw.

6. Streszczenie

Jeszcze kilkadziesiąt lat temu sztuczna inteligencja mogła się wydawać futurystyczną wizją. Obecnie na wyciągnięcie ręki jest szereg nowoczesnych rozwiązań, które pozwalają lepiej zrozumieć i analizować świat, zachodzące procesy i zjawiska. Kluczową wartością są dane, które dostarczają cennych informacji, zarówno dla biznesu, jak i instytucji publicznych. W ostatnim czasie coraz bardziej eksplorowany jest potencjał big data, stwarzający ogromne możliwości analityczne. Pod tym pojęciem należy rozumieć zarówno duże zbiory danych, pozyskiwane z wielu źródeł i w różnych formatach (liczbowych, tekstowych, głosowych, obrazowych), ale również zaawansowane technologicznie narzędzia, które umożliwiają ich gromadzenie, przetwarzanie czy wizualizację.

Termin big data jest trudny do precyzyjnego zdefiniowania i wraz z rozwojem technologii podlega ewolucjom. Opisując duże zbiory danych wskazuje się na ich cechy, co zostało zapoczątkowane przez model 3V, na który składają się volume (objętość danych), velocity (prędkość napływania nowych danych i ich analizy) oraz variety (różnorodność danych). Obecnie wskazuje się na szereg cech, których katalog jest nieustannie poszerzany. Należy przy tym zaznaczyć, że obejmują one zarówno pozytywne, jak i negatywne aspekty, bowiem korzystanie z dużych zbiorów danych wiąże się nie tylko z możliwościami, ale również generuje problemy i ryzyka np. związane z jakością danych, ich prawdziwością czy celowością pozyskiwania.

Powstanie i rozwój narzędzi, rozwiązań, algorytmów, w tym sztucznej inteligencji, uczenia maszynowego, analizy głosu czy obrazów, pozwoliły na wykorzystywanie nowych rodzajów danych. Informacje mogą być masowo pobierane ze stron internetowych, sieci społecznościowych, różnego rodzaju czujników. Możliwe jest przetwarzanie zdjęć czy nagrań, ale również precyzyjniejsze wyłuskiwanie określonych danych z plików testowych. Dzięki temu można precyzyjniej identyfikować i analizować wzorce, zachowania czy trendy.

Zakres zastosowania big data jest bardzo szeroki. Pierwotnie duże zbiory danych były wykorzystywane przede wszystkim przez prywatne podmioty, w szczególności duże korporacje. Jednak upowszechnienie narzędzi i zwiększenie ich dostępności doprowadziło do stosowania nowoczesnych rozwiązań również przez instytucje publiczne. Wykorzystanie dużych zbiorów danych przez państwo może istotnie wesprzeć procesy decyzyjne, poprawić jakość statystyki publicznej, prowadzić do lepszej realizacji jego podstawowych funkcji. Pozyskane dane mogą służyć w procesie legislacyjnym, przy tworzeniu polityk publicznych, dostarczaniu usług dla obywateli i biznesu, ale również można je wykorzystać dla lepszego zapewniania bezpieczeństwa i w walce z przestępczością.

Administracja publiczna ma ogromny potencjał w postaci posiadanych zasobów. Funkcjonowanie państwa i jego struktur prowadzi do masowego zbierania danych

o obywatelach i ich aktywności. Zapewnienie podstawowych potrzeb związanych z bezpieczeństwem, edukacją, ochroną zdrowia, zabezpieczeniem społecznym czy prawidłowością obrotu gospodarczego wiąże się chociażby z tworzeniem rejestrów czy gromadzeniem informacji w sposób umożliwiający skuteczną realizację tych zadań. Istnieje szereg agend państwa odpowiedzialnych za monitorowanie istotnych procesów i zjawisk zarówno społeczno-gospodarczych, ale również przyrodniczych czy kulturalnych. Podstawowym problemem w sektorze publicznym nie jest brak informacji – można nawet stwierdzić, że skala gromadzenia danych okazuje się znacznie większa niż w przypadku podmiotów prywatnych. Kwestią jest jednak jak je efektywnie wykorzystywać.

Oprócz informacji zbieranych w ramach codziennego funkcjonowania państwa, administracja posiada wyspecjalizowane instytucje w zakresie statystyki publicznej. Tradycyjnymi źródłami danych w tym zakresie są badania ankietowe ludności, sprawozdania statystyczne wypełniane przez przedsiębiorstwa oraz pozyskiwanie informacji ze źródeł administracyjnych. Generuje to zarówno opóźnienia związane z przekazywaniem danych, jak i koszty wynikające z zaangażowania ankieterów, co ogranicza obszary obejmowane tego typu badaniami, jedynie do najistotniejszych zjawisk i na dość dużym poziomie ogólności. Big data może stanowić pewne rozwiązanie wskazanych problemów, rozszerzając zakres danych i dostarczając je w czasie rzeczywistym, dzięki czemu możliwe jest podejmowanie lepsze podejmowanie decyzji i planowanie działań.

Jednym z obszarów, w którym można zastosować nowoczesne narzędzia analityczne, jest rynek pracy. Identyfikowanie trendów, wzorców i zjawisk związanych z aktywnością zawodową obywateli jest niezwykle istotne z punktu widzenia państwa. Pozwala na podejmowanie świadomych decyzji politycznych w zakresie polityk publicznych wspierających zatrudnienie, przeciwdziałających bezrobociu czy dostosowujących programy kształcenia do aktualnych i przyszłych potrzeb rynkowych.

Dzięki wykorzystaniu narzędzi big data możliwe jest pogłębienie dotychczasowych analiz, pozyskanie nowych informacji, a tym samym lepsze zrozumienie procesów zachodzących na rynku pracy. Przetwarzanie ogromnych zbiorów danych stwarza zupełnie nowe możliwości, pozwala na identyfikację trendów, wzorców i zjawisk, które wcześniej mogły pozostawać niewidoczne. Zaawansowane algorytmy i narzędzia analityczne przetwarzają i analizują dane w czasie rzeczywistym, co umożliwia szybkie reagowanie na zmieniające się warunki rynkowe. Analityka big data prowadzi do tworzenia precyzyjniejszych prognoz przyszłych zmian i trendów na rynku pracy, które stanowią istotną podstawę dla planowania strategicznego zarówno na poziomie krajowym, regionalnym, jak i lokalnym.

Przykładem zastosowania nowoczesnych technologii w badaniach rynku pracy jest gromadzenie i przetwarzanie danych ze stron internetowych m.in. publikujących

oferty pracy. Dostęp do treści można w szczególności uzyskać dzięki technikom web scrapingu lub web crawlingu, bazującym na wykorzystaniu botów pobierających określone informacje ze stron.

Strony internetowe z ogłoszeniami o pracę pozwalają na uzyskanie danych o preferencjach i oczekiwaniach pracodawców, ale również oferowanych warunkach. Pozwala to zarówno ustalić popyt na konkretne zawody i umiejętności w czasie rzeczywistym, jak również posiadając dane historyczne można te kategorie prognozować. Informacje z ofert pracy wskazują również jakie dodatkowe umiejętności wymagane są na danych stanowiskach oprócz formalnego wykształcenia (np. znajomość języków, określonych programów komputerowych). Mogą być one wykorzystane do dostosowania form kształcenia, ale również pozwalają zaplanować kursy podnoszące kwalifikacje lub pozwalające na przekwalifikowanie.

Analiza ofert pracy pozwala również na szacowanie liczby wakatów na rynku oraz częściowo przepływów. Czas aktywności oferty wskazuje ile zajmuje poszukiwanie pracownika i proces rekrutacyjny, a ponowne pojawienie się tożsamych ofert po dłuższym czasie (np. kilku miesięcy) może sugerować niedopasowanie pozyskanej osoby do posady lub rozwój firmy i pojawienie się nowych stanowisk. Przykład ten pokazuje, że tego typu dane powinny stanowić jedynie jedno z pozyskiwanych źródeł, a łączenie różnych zasobów pozwala na kompleksową analizę i identyfikację zjawisk czy wzorców. Co istotne, analityka big data dostarcza również narzędzi pozwalających na takie działania.

Dodatkowo ogłoszenia dostarczają danych o zmianach i trendach na rynku pracy np. w zakresie elastycznych form pracy – pracy zdalnej czy pracy w niepełnym wymiarze godzin. Jak również pozwalają na analizę i prognozowanie wynagrodzeń oraz kształtowania się benefitów pozapłacowych. Często pokazują również informację o aplikacjach na dane stanowisko co pokazuje zainteresowanie poszczególnymi ofertami.

Portale społecznościowe, w szczególności biznesowe, dostarczają informacji o osobach aktywnych zawodowo, ich wykształceniu, kwalifikacjach czy doświadczeniu zawodowym. Stanowią praktycznie interaktywne CV, które pozwalają określić charakterystyki zasobów siły roboczej. Możliwa jest analiza okresów zatrudnienia u danego pracodawcy, gotowości do zmiany pracy czy częstotliwość awansów.

W określeniu i prognozowaniu cech potencjalnych pracowników mogą również pomóc informacje dotyczące edukacji. Zarówno serwisy rekrutacyjne szkół wyższych wskazujące na liczbę miejsc na danym kierunku, ale również popularne portale oferujące kursy online. Kursy te są zazwyczaj tworzone we współpracy z uczelniami wyższymi i umożliwiają uzyskanie certyfikatu potwierdzającego nabycie określonych

umiejętności, a czasem nawet pozwalają na uzyskanie formalnego wykształcenia. Jest to popularna forma dokształcania, również w kontekście rozwoju zawodowego. Dzięki temu można uzyskać informację w jakich obszarach użytkownicy chcą poszerzać swoją wiedzę i jakie kwalifikacje uznają za pożądane.

Oprócz tego możliwe jest badanie aktywności w sieci. Może to dotyczyć zarówno analizy odwiedzanych stron, jak i wyszukiwanych fraz związanych z zatrudnieniem. Z jednej strony może to pokazywać aktualne trendy w zakresie poszukiwanych treści, ofert czy kursów oraz wskazywać najpopularniejsze strony np. wśród osób poszukujących pracy. Z drugiej zaś analiza długoterminowa może uwidocznić zachodzące zmiany, ale również wskazywać na cykliczność pewnych zjawisk.

Wykorzystanie danych internetowych w badaniach ma oczywiście swoje ograniczenia. W szczególności nie wszystkie informacje mogą być rzetelne czy pełne, jak również konieczne jest poniesienie kosztów na odpowiednią infrastrukturę, prace informatyczne i analityczne. W związku z tym niezwykle istotne jest odpowiednie zaplanowanie procesu, określenie oczekiwanych efektów, przygotowanie narzędzi, zaprojektowanie bazy danych, ale również jej właściwe prowadzenie – monitorowanie danych, zapewnienie ich bezpieczeństwa czy odpowiedniej archiwizacji. Jednocześnie należy dbać o reprezentatywność informacji, żeby stanowiły wartościowy wkład do analiz oraz stabilność źródeł, gwarantujące dostęp do potrzebnych danych.

Pomimo ograniczeń big data stanowi przede wszystkim ogromną szansę na nową jakość analityki rynku pracy. Poszerzenie zakresu dostępnych informacji, zaawansowane narzędzia analityczne, możliwość analizowania danych w czasie rzeczywistym to niewątpliwie przyczynek do udoskonalania tradycyjnych badań, precyzyjniejszego wnioskowania i prognozowania, a zatem również lepszego zarządzania obszarem funkcjonowania państwa, przedsiębiorców i obywateli.