КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Кафедра інтелектуальних програмних систем

КВАЛІФІКАЦІЙНА РОБОТА
НА ЗДОБУТТЯ СТУПЕНЯ БАКАЛАВРА

на тему:

за спеціальністю 6.050103 Програмна Інженерія

ВЕЛИКІ ТА ВІДКРИТІ ДАНІ. РОЗРОБКА ВЕБ СЕРВІСУ ДЛЯ ПОШУКУ АВТОТРАНСПОРТУ

Виконав студент 4-го курсу
Шанаах Алі Махмуд

Науковий керівник:
кандидат фізико-математичних наук, доцент
Верес Максим Миколайович

Засвідчую, що в цій роботі немає
запозичень з праць інших авторів без
відповідних посилань

Дипломна робота заслухана на засіданні кафедри
Протокол №11 від 13 травня 2019 року
Проватар О. І. _____

Київ — 2019

РЕФЕРАТ

Обсяг роботи 40 сторінок, 8 ілюстрацій, 1 таблиця, 13 джерел посилань.

ВЕЛИКІ ТА ВІДКРИТІ ДАНІ. РОЗРОБКА ВЕБ СЕРВІСУ ДЛЯ ПОШУКУ АВТОТРАНСПОРТУ.

Об'єктом роботи є процес ефективної обробки великих масивів даних, що зберігаються на єдиному порталі даних України.

Предметом роботи є вивчення методів ефективної обробки великих масивів даних на прикладі даних про український транспорт.

Метою роботи є розробка прототипу швидкісного веб-сервісу, що надасть чітку інформацію про український транспорт за державним номерним знаком.

Методи розроблення: комп'ютерне моделювання, розробка програмного продукту на основі ітеративної моделі. Інструменти розроблення: інтегроване середовище розробки Goland IDE, мова програмування Go, реляційна база даних PostgreSQL, єдиний державний портал відкритих даних https://data.gov.ua.

Результати роботи: досліджені сучасні методи оброки великих масивів даних, запропонований алгоритм роботи з даними на єдиному державному порталі відкритих даних, розроблено програмний продукт та приклад його використання у вигляді чат-бота, який дозволяє знайти інформацію про транспорті засоби за державним номерним знаком.

Розроблений програмний продукт має відкритий код, та може використовуватися у комерційних цілях будь-якого підприємця чи компанії. Система була розгорнута, тож будь-який користувач системи Telegram може її використовувати.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАЧЕННЯ

IDE – Integrated Design Environment, інтегроване середовище розробки;

SQL – Structured query language, Мова структурованих запитів;

БД – База даних;

ЄС – Європейський Союз;

ІКТ – Інформаційно-комунікаційні технології;

МВС – Міністерство внутрішніх справ України;

ОБСЄ – Організація з безпеки та співробітництва в Європі;

ПО – Програмне забезпечення;

ВСТУП

Доступність та відкритість даних, що мають суттєве суспільне значення, стали світовим трендом. За цих умов питання забезпечення доступу до публічної інформації набувають особливої ваги для регіональних та місцевих рівнів публічного управління. Адже, незважаючи на напрацьований протягом останніх років досвід, публічні службовці нині мають розв'язувати нові комплексні та складні проблеми за цим напрямом діяльності, у тому числі з використанням сучасних інформаційно-комунікаційних технологій. Постійне удосконалення чинної нормативно-правової бази та організаційної структури органів влади спонукає запроваджувати та використовувати у практиці публічного управління та адміністрування нові технологічні інструменти, спрямовані на поліпшення інформаційних обмінів між органами влади та суспільством.

Актуальність теми дипломної роботи тісно пов'язана із висхідною тенденцією відкриття даних у світі та Україні. У країнах ЄС тема відкритості державних даних вже понад десятиріччя набуває широкого поширення задля боротьби з бюрократією та корупцією.

За результатами дослідження Open Data Barometer Україна зайняла 2 місце у світі за темпами відкриття державних даних. Наша країна розвивається у цьому напрямку тільки останні декілька років, тому питання обробки цих даних є досить нагальним.

Портал відкритих даних — єдиний державний веб-портал відкритих даних, створений з метою зберігання публічної інформації у формі відкритих даних та забезпечення надання доступу до неї широкому колу осіб за принципами, визначеними у Міжнародній хартії відкритих даних, до якої Україна приєдналася у жовтні 2016 року. Портал створено на вимогу Закону України «Про доступ до публічної інформації» та постанови Кабінету Міністрів від 21 жовтня 2015 року No 835 «Про затвердження Положення про набори даних, які підлягають опри-

людненню у формі відкритих даних».

Водночас більшість громадян України зустрічаються з проблемою неможливості знаходження інформації в обробленому для користувача вигляді, саме тому **об'єктом дослідження** цієї роботи було обрано процес ефективної обробки великих масивів даних, що зберігаються на єдиному порталі даних України. А оскільки, не існує єдиного способу обробки різних типів даних, за **предмет дослідження** було обрано вивчення методів ефективної обробки великих масивів даних на прикладі даних про український транспорт.

Очікуваним результатом і **метою** дипломної роботи є розробка прототипу швидкісного веб-сервіса, що надасть чітку інформацію про український транспорт за державним номерним знаком.

Практична значущість дипломної роботи полягає в можливості застосування її результату на практиці з метою надання користувачам різних інструментів для пошуку транспортних засобів.

У ході виконання дипломної роботи на першому етапі буде розроблена архітектура веб-сервісу, наступним кроком буде реалізація та тестування програмного забезпечення. Кінцевим результатом роботи буде чат-бот з використанням розробленого продукту, аби продемонструвати використання сервісу на прикладі.

Серед **методів розроблення** будуть долучені комп'ютерне моделювання та розробка програмного продукту на основі ітеративної моделі. У якості **інструментів розроблення** будуть використовуватися мова програмування Go, реляційні бази даних та єдиний державний портал відкритих даних.

РОЗДІЛ 1
ВІДКРИТІ ДАНІ

1.1. Поняття відкритих даних

Переважна більшість організацій користується визначенням, що було дано всесвітньою організацією «Open Knowledge Foundation» [12]. Це визначення відкритості подає точне значення терміну «відкрите» стосовно знань, сприяючи стійким громадам, в яких кожен може взяти участь і де можливість взаємодії досягає максимуму.

Повне значення відкритості є надто чітким для цієї роботи [11], виділимо з визначення основні риси відкритості:

Доступність. Дані повинні бути доступні у повній мірі. Ресурси на їх отримання повинні бути у розумних межах, переважно шляхом завантаження через Інтернет. Дані також повинні бути доступні у зручній і змінюваній формі.

Повторне використання та розповсюдження. Дані повинні надаватися на умовах, що дозволяють повторне використання та розповсюдження, включаючи використання з іншими наборами даних.

Загальна участь. Кожен повинен мати змогу використовувати та розповсюджувати дані. Будь-які обмеження щодо сфер діяльності, осіб чи груп повинні бути відсутніми. Наприклад, не комерційні обмеження, які запобігають комерційному використанню, або обмежується використання для певних цілей (наприклад, тільки в освіті), не допускаються.

Якщо вам цікаво, чому важливо дати чітке визначення відкритості, є проста відповідь: сумісність. Сумісність означає здатність різних систем та організацій взаємодіяти разом [9]. У даному випадку це можливість взаємодіяти або змішувати різні набори даних. Сумісність є дуже важливою, оскільки дозволяє різним компонентам працювати разом.

Здатність до складання й підключення компонентів має важливе значення для побудови великих, складних систем. Без сумісності це стає неможливим, про що свідчить найвідоміший міф про Вавилонську вежу, де проблема взаємодії призвела до повного краху.

Ми бачимо схожий випадок по відношенню до даних. Сутність сумісності даних полягає у тому, що один фрагмент відкритих даних може бути використаний разом з будь-яким іншим фрагментом відкритих даних.

Сумісність — це ключ до реалізації основних переваг відкритості. Різко зростає здатність об'єднувати набори різноманітних даних і тим самим розробляти більші й кращі продукти й послуги.

Надання чіткого визначення відкритості гарантує, що при отриманні двох відкритих наборів даних з двох різних джерел, ми будемо мати змогу об'єднати їх разом, іншими словами ми не потрапимо у ситуацію де багато наборів даних, проте взагалі неможливо об'єднати їх у великі системи, які у свою чергу мають найбільшу цінність для бізнесу та подальшої обробки.

1.2. Доступ до відкритих даних

Сучасний етап суспільного розвитку позначений динамічним проникненням інформаційних технологій в усі сфери людської діяльності. За таких умов недостатня увага органів публічної влади до характеру, змісту інформаційних обмінів з громадськістю, більше того – їх недооцінка, можуть зумовити дисбаланс у відносинах влади й громадськості [13]. З одного боку, це може загрожувати збільшенням кількості нерезультативних і неефективних рішень, з іншого – зниженням ступеня підтримки органів публічної влади суспільством, а то й відкритим протистоянням. Унаслідок такого стану справ значна кількість суспільно важливих проблем залишаться невирішеними, належним чином не будуть прогнозовані та попереджені нові проблемні ситуації. Для нормального функціонування описаної системи взаємовідносин важливо досягти такого стану речей, коли інформаційні потоки між органами публічної влади та громадськістю циркулюють максималь-

но безперешкодно. Така модель може бути реалізована, коли кожен громадянин матиме реальну можливість на отримання гарантованої законом повної, правдивої та всебічної інформації про функціонування органів влади, про їх плани, можливі напрями дій, стан окремих сфер суспільного життя, використання державних та комунальних ресурсів [1].

За останні роки було ухвалено міжнародно-правові документи, які визнають право на доступ до інформації основоположним правом людини, а представники Організації Об'єднаних Націй, Ради Європи, ОБСЄ у своїх доповідях звертають увагу на те, що право на доступ до інформації є необхідною умовою для участі громадян в ухваленні державних рішень, запорукою розвитку демократії та фундаментом у боротьбі з корупцією [4].

У європейських державах діяльність, що пов'язана з роботою з публічною інформацією, в цілому ґрунтується на принципі максимальної відкритості, який полягає у тому, що державні органи зобов'язані розкривати інформацію, а кожний член суспільства має відповідне право отримувати її. При цьому обов'язок державних органів щодо оприлюднення інформації, що має особливе суспільне значення, здебільшого реалізується не лише як реакція на вимогу надати інформацію у відповідь на запити, але й шляхом оприлюднення та широкого розповсюдження документів, які становлять особливий суспільний інтерес, залежно від наявних ресурсів і можливостей. Це, зокрема, випливає з положень статті 10 Конвенції Ради Європи про доступ до офіційних документів, якою встановлено, що органам державної влади належить із власної ініціативи і в тих випадках, коли це виправдано, вживати всіх необхідних заходів щодо опублікування офіційних документів, які перебувають в їхньому розпорядженні, в інтересах підвищення прозорості та ефективності органів державного управління та сприяння інформованому залученню громадськості у справи, які становлять суспільний інтерес

Аналогічні положення відображені й у національному законодавстві. Так, у Республіці Болгарія законом про доступ до публічної інформації саме з метою максимального полегшення доступу до публічної інформації, передбачено обов'язок щодо розкриття публічної інформації державними органами про свою діяльність

шляхом її регулярної публікації або з використанням інших форм оприлюднення. Також принцип сприяння доступу до інформації реалізується і в такий спосіб, що усі державні органи зобов'язані створити відкриті й доступні внутрішні системи забезпечення права громадян на отримання інформації. Для прикладу, Уряд Словенії зобов'язав усі державні установи створити на своїх веб-сайтах та регулярно оновлювати каталоги публічної інформації. Тобто, в питаннях забезпечення доступу до інформації за кордоном, зокрема в Європі, пріоритетом є проактивна, а не реактивна позиція органів влади.

Одним з яскравих прикладів застосування інноваційних механізмів забезпечення доступу до публічної інформації в умовах розвитку інформаційного суспільства є Фінляндія, де давня традиція відкритості влади, поєдналася з інтенсивним розвитком ІКТ. Як наслідок, політика відкритості та можливість електронного доступу до інформації є основною причиною низької корупції у Фінляндії, факти корупційних діянь у країні трапляються скоріше як винятки, ніж як закономірність. Отже, Фінляндія є прикладом превентивного забезпечення громадськості публічною інформацією.

Як свідчить закордонна практика, з налагодженням діяльності із забезпечення доступу до публічної інформації в органах публічної влади з часом зменшується кількість звернень про доступ до такої інформації. При цьому зацікавленість нею не знижується – просто громадяни, журналісти, окремі організації починають отримувати необхідну інформацію з джерел, що стають дедалі доступнішими. Зокрема, про це свідчить досвід Болгарії, де за останні п'ять років кількість запитів на інформацію до органів влади постійно зменшувалася.

Таким чином, європейський досвід забезпечення доступу до публічної інформації свідчить про перспективну зміну тенденцій і в українській практиці – з розширенням використання інформаційно-комунікаційних технологій споживачі такої інформації зможуть самостійно отримувати доступ до необхідних документів, що зменшить необхідність безпосередніх контактів запитувача і службовця, підвищить ефективність управлінських процедур.

1.3. Класифікація відкритих даних

Для того, щоб зрозуміти, які можуть бути форми відкритих даних, ми також звернемося до відомої класифікації «Five Star Open Data» [6], де якість даних та рівень відкритості визначається кількістю зірок від 1 до 5, чим більше – тим краще. Відкритість даних залежить від способів доступу, форматів та кількості додаткових дій, які потрібні для отримання кінцевої інформації, її обробки та збереження у власному сховищі або базі даних.

Одну зірку отримує будь-яка інформація вільно доступна через Інтернет в будь-якому форматі. Під цю класифікацію підпадає файл в форматі PDF або інша копія документу, на який веде пряме посилання на офіційному сайті державного органу. Якщо цей файл можна відкрити на власному екрані, прочитати, роздрукувати та отримати звідти потрібну інформацію, то це відкриті дані з однією зіркою.

Дві зірки отримує структурована інформація, яку можна обробляти автоматично, наприклад, в форматах для веб-браузерів чи офісних програм (відкриті формати – TXT, HTML, RSS; пропрієтарні формати, Excel – XLS, Word – DOC, RTF). Якщо дані знаходяться в тілі вихідної веб-сторінки, але не мають чіткої структури, містять зайві елементи оформлення, навігації, якщо дані потребують додаткових дій – спеціального розбору, то вони вважаються «з двома зірками».

Три зірки може отримати інформація, представлена у відомих, добре описаних відкритих структурованих форматах (наприклад, CSV, JSON, XML, YAML) і якщо автоматизована її обробка не потребує від користувача особливих ліцензій та додаткових плат. До відкритих форматів також відносяться пов'язані дані (HTML+RDFa) з узгодженою розміткою елементів в атрибутах або текстові файли таблиць, поля яких розділені табуляцією, комами, крапками з комою або іншими символами.

Чотири зірки надаються у випадку, якщо можна отримати первинні необроблені набори відкритих даних у вигляді файлів (довідники, списки, таблиці у відкритому форматі, зліпок бази даних, архів документів тощо) або фільтро-

вані дані у запиті до API за вказаними параметрами. Це дає змогу отримувати тільки потрібну інформацію, актуальну на момент запиту, заощаджує ресурси та час користувача. Безумовно, API має бути описаний так само, як і формати даних, а доступ до нього може бути анонімний без обмежень або з реєстрацією, за вказаним ідентифікатором, лімітами на кількість одночасних запитів.

Останній рівень – **п'ять зірок** – надається інформації, якщо набори відкритих даних пов'язані між собою (мають спільні довідники, класифікатори, ідентифікатори, посилання між документами та іншими елементами) і представляють собою семантичну мережу, що постійно оновлюється й змінюється відповідно до сучасних запитів.

Слід зазначити, що більшість даних представлених на єдиному державному порталі відкритих даних має формат CSV та JSON, тобто оцінка якості даних за цією технологією поки не дуже висока.

1.4. Формати даних

В залежності від специфіки даних, їх розміру та тематики, одні проекти відкритих даних створювались на базі наборів PDF чи DOC файлів, таблиць XLS, що перетворювались на прості текстові таблиці CSV, а інші брали за основу формат розмітки XML, проектували власні схеми XSD і використовували складні структури.

Як свідчить остання статистика використання форматів відкритих даних, найбільш поширений в світі формат (як по кількості, так і по об'єму даних) – PDF. Для українських органів влади, де найбільш розповсюджені операційні системи Microsoft Windows, переважають формати DOC та XLS. Разом з новими версіями офісних програм в Інтернет почали з'являтися документи DOCX та XLSX, рідко ODF (Open Document Format). Після поширення ініціативи відкриття державних даних та створення порталів, кількість наборів в форматі XML та інших відкритих форматах почала суттєво збільшуватись.

Необроблені дані, сформовані державними структурами за багато років, мо-

жуть бути досить неоднорідними, а деякі набори навіть дублюються в різних форматах для зручності користування.

Серед доступних форматів відкритих даних, які можна автоматично обробляти електронними засобами, є:

- CSV – дані, розділені комами або іншими розділовими символами.
- JSON – формат, орієнтований на обробку складних структур.
- XML – універсальний текстовий формат розмітки.

CSV – текстовий відкритий формат, призначений для представлення масивів даних, де кожний рядок – це запис таблиці, а значення окремих полів у рядку розділені спеціальними символами, зазвичай комами. Щоб завантажити записи таблиці за найменуванням полів, додатково потрібно мати опис її структури – назви та формат полів. Більшість програм широко трактують цей формат і допускають використання інших розділових символів, наприклад, табуляції (TSV) чи коми з крапкою.

JSON – текстовий відкритий формат, оснований на Javascript представлені та призначений для обміну даними в мережі Інтернет між сервером та клієнтом або сервером і сервером. Хоча він позиціюється, як незалежний від системи і мови програмування, частіше за все використовується за допомогою програм на Javascript, але як і інші текстові формати, легко читається людиною.

Найбільшу популярність JSON набув після створення інтерактивних веб-сторінок, дані до яких через API передавались під час взаємодії користувача з елементами інтерфейсу. За рахунок своєї лаконічності, на відміну від XML, простоті й швидкості використання саме в програмах на Javascript, широкими можливостями в обробці даних – рекурсивного перетворення в текстовий вигляд складних об'єктів, формат активно використовується для формування «на льоту» та передачі структур даних в Інтернет в різних інформаційних системах і сервісах.

Проте існує велика кількість інших форматів даних, які є менш розповсюдженими чи орієнтованими на певну сферу даних, наприклад відеодані. Дані такого незвичайного виду зазвичай зберігають в спеціалізованих форматах.

XML – найстаріший текстовий відкритий формат, створений в 1994 році та ре-

комендований Консорціумом Всесвітньої павутини, як основний для обміну інформацією в Інтернет. Гіпертекстова розмітка (HTML) – це один з різновидів XML. Разом з таблицями каскадних стилів CSS, які формують зовнішній вигляд документів, вони є тими основними форматами, що обумовлюють розвиток технологій. Перевагами XML є простота та гнучкість розмітки, яка не вимагає формальних, фіксованих назв тегів чи параметрів, і будь-який розробник може доповнювати та змінювати формат, створювати власну схему XSD.

За довгий час існування XML на його базі було розроблено багато форматів і стандартів зі схожим синтаксисом. Зазвичай цю групу форматів називають загальною назвою – XML, тому що вони мають єдині механізми опису схем XSD, перевірки правильності даних, доступ до елементів XPath та трансформації для автоматичного конвертування у інші схеми чи формати (наприклад, альтернативні JSON та YAML) за допомогою мови перетворення XSLT (eXtensible Stylesheet Language Transformations).

Використання формату XML (а саме LegalXML) у якості відкритого стандарту нормативноправового документа – це сучасний спосіб забезпечити обмін інформацією (документами, картками, довідниками тощо) між інформаційними системами або в межах однієї системи при опрацюванні документів.

В одному файлі XML в текстовому вигляді, крім основних даних та тексту електронного нормативного документа, можна розміщати метадані, вкладені файли, необхідні структури чи довідники. Це дозволяє зручно не тільки зберігати, передавати, обробляти документ, отримувати PDF версію для друку, формувати зміст чи робити посилання на конкретну главу, статтю, пункт, підпункт тощо, а й автоматизовано вносити зміни, підготовлені у вигляді, що дозволяє їх програмну обробку.

1.5. Відкриті дані в Україні

В багатьох країнах світу розвиток відкритих даних підтримується на державному рівні вже досить давно. І цей процес включає створення відповідної законо-

давчої бази, виконавчих органів та інформаційних ресурсів [7].

У цілому українське законодавство про доступ до публічної інформації відповідає вимогам міжнародних стандартів про свободу доступу до інформації. В їх основу покладені принципи максимальної відкритості та доступності громадськості на отримання інформації, обов'язку оприлюднення інформації, яка має особливе значення, надання відповіді у стислі терміни, забезпечення безкоштовної допомоги при оформленні запиту. Зокрема, Закон України «Про доступ до публічної інформації» за міжнародним рейтингом забезпечення права на інформацію, розроблений міжнародними організаціями «Access Info Europe» та «Centre for Law and Democracy», посів 8 місце серед 89 країн світу.

Міжнародне визнання. Україна посіла друге місце серед країн, що досягнули найбільшого прогресу за чотири роки за рівнем публікації та використання відкритих даних для підзвітності влади, розвитку інновацій і соціального впливу. Про це свідчить звіт найбільш впливового світового рейтингу Open Data Barometer за вересень 2018 року. Крім того, Україна посіла 17 сходинку серед 30 країн-лідерів, що взяли на себе конкретні зобов'язання щодо розвитку відкритих даних. Це результат нашого уряду щодо забезпечення принципів прозорості та відкритості у державному секторі, які стали першочерговими для України після приєднання у 2016 році до Міжнародної хартії відкритих даних. У листопаді 2018 року Кабмін схвалив план дій з реалізації принципів хартії, тому очікуємо подальшого динамічного розвитку сфери відкритих даних у 2019 році.

Вплив на Економіку. Відкриті дані — це цінний ресурс, який допоможе посилити «цифрову» та «реальну» економіки країни, адже про це свідчать результати глобальних досліджень. Очікується, що економіки країн Великої двадцятки зростуть на 1,1% ВВП завдяки відкриттю пріоритетних наборів даних. «Київська Школа Економіки» та «Open Data Institute» використовуючи авторитетну методологію дослідження Європейської комісії, проаналізували економічний потенціал відкритих даних для України. Ми стали першою країною за межами країн Великої двадцятки, де був проведений такий аналіз. Дослідники виявили, що вже у 2017 році відкриті дані принесли нашій державі понад 700 млн. доларів Якщо рух

за відкриті дані в Україні й надалі набиратиме обертів, ця цифра може зрости до 1,14 млрд доларів або 0,92% ВВП до 2025 року.

Українські дані на Європейському порталі. Україна стає частиною єдиного Європейського інформаційного простору і ще більше наближається до інтеграції з Євросоюзом. З вересня 2018 року українські державні дані, як і дані країн ЄС, публікуються на «European Data Portal». Портал містить понад 860 тис. наборів даних, що охоплюють 35 країн і 78 місцевих та національних порталів. Що це дає Україні? По-перше — імідж. Публікація наборів даних на «European Data Portal» — це потужний сигнал про відкритість діяльності держави на рівні європейських країн. По-друге — підвищення інвестиційної привабливості. Європейські компанії можуть використовувати відкриті дані для отримання більш детальної інформації для аналізу бізнес-потенціалу України. Наприклад, «Open Corporates», найбільша відкрита база даних компаній у світі, що містить інформацію про майже 140 млн компаній, була інтегрована з Єдиним державним реєстром юридичних осіб. Це робить Український ринок більш привабливим для іноземних інвесторів, оскільки можна швидко та зручно знаходити необхідні дані.

Портал відкритих даних. У 2018 році Державне агентство з питань електронного урядування за підтримки проекту «Прозорість та підзвітність у державному управлінні та послугах» запустили модернізований державний портал відкритих даних. Розробила проект команда «Opendatabot». Це найвідоміший стартап у сфері відкритих даних та один з найбільших користувачів порталу. Це унікальна світова практика, коли портал модернізують користувачі відкритих даних. На оновленому порталі публікуються нові дані, які одразу можуть використовуватися бізнесом, громадськими організаціями та журналістами без додаткової обробки. До того ж, тепер розпорядники можуть оновлювати дані автоматично, що дозволяє отримувати інформацію в реальному часі.

Пріоритетні набори даних. Одне з найбільш вагомих досягнень 2018 року — відкриття пріоритетних наборів, починаючи від транспортної сфери й закінчуючи даними місцевих бюджетів.

Найбільш популярним набором даних стали відомості про транспортні засоби

від МВС, що показує реальний стан авторинку України. Лише за перші три місяці опубліковану інформацію завантажили 20 тис. разів. На базі відомостей про транспортні засоби «Texty.org.ua» та «Opendatabot» створили сервіси, якими громадяни скористалися понад 1 млн разів. Окремо слід відзначити відкриття у 2018 році даних щодо ліцензій транспортних засобів на перевезення пасажирів і вантажів. Тепер за кілька секунд кожен охочий за номером транспортного засобу може перевірити, чи він користується послугами одного з 42 794 легальних перевізників. З ініціативи Міністерства Фінансів України на порталі почали публікуватися дані 9 603 місцевих бюджетів. Тепер кожен громадянин може контролювати використання бюджетних коштів на рівні області та села. Також контрольні державні органи відкрили дані про понад 143 400 перевірок бізнесу, запланованих на 2019 рік. Тепер будь-який власник бізнесу може дізнатися, коли перевірка постукає у двері його компанії. У сфері екології найбільший вплив мало відкриття даних державного моніторингу поверхневих вод Державним агентством водних ресурсів. Дані оприлюднені за 16 основними показниками із 445 пунктів збору води на річках за останні п'ять років. Хоча дані про якість води вважаються одним з найважливіших наборів для суспільства, за кордоном вони залишаються доволі закритими. Згідно з останнім випуском «Global Open Data Index», ці дані публікують лише 15 країн світу. Завдяки відкритим даним Міністерства Екології та природних ресурсів щодо дозвільних документів та процедур промислових та інших забруднювачів довкілля активісти з Дніпра створили перший екологічний бот «SaveEcoBot».

Національний конкурс. Відкриті дані мають найбільший вплив, коли їх використовують для створення продуктів та сервісів для економічного розвитку та інновацій.

У 2018 році в Україні вдруге відбувся національний конкурс інноваційних IT-проектів на основі відкритих даних «Open Data Challenge». Заявки на участь у заході подали 190 команд, 15 з них пройшли інкубацію, а шість команд-переможців розділили фінансову допомогу на загальну суму 2,5 млн грн. Переможці конкурсу роблять суттєвий внесок у розвиток економіки країни. Так, проект «Moni-

tor Estate» дозволяє перевіряти ризики купівлі новобудови у Києві або Львові. «NORA» аналізує дані з відкритих джерел і виявляє неочевидні зв'язки між учасниками будівельного ринку, а «LvivCityHelper» інформує львів'ян про ремонти та обслуговування будинків. Як показала практика, найбільший попит на відкриті дані — у сферах інфраструктури, будівництва та екології.

Третій цикл «Open Data Challenge» почнеться у лютому 2019 року, тож кожен може подати заявку та виграти фінансову підтримку проекту на основі відкритих даних. У 2017 році уряд схвалив оновлену постанову №835, що регулює відкриття понад 600 наборів даних. Попит на відкриті дані росте, тому очікується ухвалення третьої редакції постанови та розширення кількості наборів. Публікація даних про транспорт та охорону здоров'я може стати найбільшим проривом, у тому числі завдяки великому антикорупційному потенціалу. Ключовим пріоритетом залишається стимулювання використання даних для розвитку економіки. Одну з провідних ролей відіграватиме третій цикл «Open Data Challenge».

1.6. Публічна інформація у формі відкритих даних

Закон України «Про доступ до публічної інформації» надав публічній інформації статус доступності, але передбачав спочатку лише доступ до публічної інформації за запитами. Зміни до цього закону, пов'язані з виокремленням нового типу публічної інформації, а саме, публічної інформації у формі відкритих даних. Тепер документи, в яких знаходяться відкриті дані, обов'язково повинні бути електронними, оприлюдненими та придатними для повторного використання з можливістю наступного використання з метою проведення наукових досліджень, забезпечення інновації, запровадження бізнес-проектів, забезпечення підзвітності і суспільного контролю за органами публічної влади.

Можна виділити такі основні мотиви відкриття державних даних: запобігання корупції, надання послуг та забезпечення інновацій, посилення громадської участі та контролю, укріплення правопорядку та законності. Концепція відкритих даних активно підтримується та розвивається міжнародними ініціативами та ор-

ганізаціями, зокрема, в рамках ініціативи партнерство «Відкритий Уряд», до якої Україна приєдналась у 2011 році.

Відкриті дані повинні бути придатні до повторного використання без обмежень авторського права, патентів та інших механізмів контролю [2]. Більшість країн світу використовує вільну ліцензію некомерційної організації «Creative Commons», яка розробила вільні та відкриті публічні ліцензії, за допомогою яких автори та правовласники можуть поширювати свою інформаційну продукцію, а користувачі контенту легально їх використовувати.

Відкриті ліцензії «Creative Commons» дозволяють повторне використання інформації державного сектора без необхідності розробляти та оновлювати на замовлення ліцензії на національному рівні. Таку ліцензію використовують Австралія, Бразилія, Великобританія, Франція, Нідерланди.

Хартія відкритих даних формулює шість основних принципів роботи з відкритими даними та визначає пріоритетні сфери, в яких відкриті дані матимуть найбільший ефект.

Відкриті дані виступають як фундамент відкритого публічного управління, сприяють прозорості роботи органів публічної влади, формується база для громадського контролю та створюються нові послуги для громадян та бізнесу.

Одним з ключових елементів концепції відкритих даних є консолідація державних даних, тому єдиний державний портал відкритих даних виступає ключовим елементом та є засобом консолідації цих даних [3]. Завданням порталу є систематизація та зручне представлення даних відповідно до інтересів та запитів користувачів.

РОЗДІЛ 2
ВЕЛИКІ ДАНІ

2.1. Поняття великих даних

Поняття великих даних з'явилося досить недавно. Сервіс аналізу пошукових запитів у мережі Інтернет «Google Trends» демонструє початок активного росту використання словосполучення починаючи з 2011 року. Як ми бачимо на рис. Б.2 піком актуальності теми великих даних був 2015 рік, проте тема все ще залишається досить актуальною. Тут простежується прямий зв'язок с початком розвитку сфери відкритих даних, а тобто можна припустити, що з розвитком теми відкритих даних буде продовжуватися вивчення способів їх ефективної обробки.

Величезна кількість необроблених даних оточує нас у світі [5]. Дані, які не можуть бути безпосередньо розглянуті людьми. Інтернет, держава та бізнес генерують нові дані з неабиякою швидкістю завдяки розробці потужних засобів зберігання та об'єднання даних. Організовані дані чи інформація не можуть бути просто зрозумілі або автоматично оброблені через їх величезну кількість та різноманіття. Ці передумови призвели до розвитку науки про дані та аналіз даних, відомої дисципліни, яка все більше і більше присутня в сучасному інформаційному світі.

Сучасний обсяг даних, що керуються створеними людиною системами, перевершує можливості обробки традиційних систем. Виникнення нових технологій і послуг, а також зниження вартості обладнання призводять до постійно збільшення інформації в мережі «Інтернет». Це явище, безумовно, є великим викликом для спільноти аналітиків даних. Поняття великих даних може бути визначене як великий обсяг різноманітних даних, що вимагають нового підходу, а тобто більш ефективної обробки.

Розподілені обчислення широко використовувалися аналітиками даних до по-

яви терміну великих даних. Багато стандартних та складних алгоритмів були замінені їх паралельними версіями з метою зменшення часу виконання програмного забезпечення, що витрачається на обробку даних. Проте сьогодні, для більшості сучасних проблем, розподілений підхід стає обов'язковим, оскільки жодна архітектура не може розв'язати усі ці проблеми.

Міжнародна компанія McKinsey, що спеціалізується на вирішенні завдань, пов'язаних зі стратегічним управлінням, виділяє 5 методів і технік аналізу, які можна застосувати до великих даних.

Методи класу Data Mining – сукупність методів виявлення в даних раніше невідомих, нетривіальних, п рактично корисних знань, необхідних для прийняття рішень. До таких методів, зокрема, відносяться навчання асоціативним правилами, класифікація (розбиття на категорії), кластерний аналіз, регресійний аналіз, виявлення та аналіз відхилень.

Краудсорсинг – класифікація і збагачення даних силами широкого, невизначеного кола осіб, які виконують цю роботу без вступу в трудові відносини.

Змішування й інтеграція даних – набір технік, що дозволяють інтегрувати різнорідні дані з різноманітних джерел з метою проведення глибинного аналізу. Наприклад, цифрова обробка сигналів, обробка природної мови та ін.

Машинне навчання, включаючи навчання з учителем і без вчителя — використання моделей, побудованих на базі статистичного аналізу або машинного навчання для отримання комплексних прогнозів на основі базових моделей.

Візуалізація аналітичних даних – подання інформації у вигляді малюнків, діаграм, з використанням інтерактивних можливостей та анімації як для отримання результатів, так і для використання як вихідних даних для подальшого аналізу. Дуже важливий етап аналізу великих даних, що дозволяє представити найважливіші результати аналізу в найбільш зручному для сприйняття виді. Прикладом візуалізації даних у сфері українського транспорту є рис. Б.5, що демонструє просту діаграму розподілення кольорів транспорту.

Розглянемо основні принципи роботи з великими даними:

Горизонтальна масштабованість — базовий принцип обробки великих да-

них. Як вже говорилося, великих даних з кожним днем стає все більше. Відповідно, необхідно збільшувати кількість обчислювальних вузлів, за якими розподіляються ці дані, причому обробка повинна відбуватися без погіршення ефективності.

Відмовостійкість. Цей принцип випливає з попереднього. Оскільки обчислювальних вузлів в кластері може бути багато і їх кількість, не виключено, буде збільшуватися, зростає і ймовірність виходу машин з ладу. Методи роботи з великими даними повинні враховувати можливість таких ситуацій і передбачати превентивні заходи.

Локальність даних. Оскільки дані розподілені на великій кількості обчислювальних вузлів, то, якщо вони фізично знаходяться на одному сервері, а обробляються на іншому, витрати на передачу даних можуть стати невиправдано великими. Тому обробку даних бажано проводити на тій же машині, на якій вони зберігаються. Ці принципи відрізняються від тих, які характерні для традиційних, централізованих, вертикальних моделей зберігання добре структурованих даних. Відповідно, для роботи з великими даними розробляють нові підходи й технології. Підсумовуючи проаналізовано інформацію, щодо поняття «Big Data», сформулюємо вичерпне визначення в рамках цієї роботи.

Big Data — набір підходів, інструментів і методів обробки структурованих і неструктурованих даних величезних обсягів і значного різноманіття з метою отримання зрозумілої для людини інформації, ефективної в умовах безперервного приросту.

2.2. Технологія MapReduce

У 2004 вченими корпорації «Google» було розроблено інноваційну технологію для розподілених обчислень [10]. MapReduce – це технологія для розподілених обчислень великих масивів даних. Користувачі визначають так звану функцію Map, що обробляє пари ключ/значення для створення набору проміжного результату, який у свою чергу отримує так звана функція Reduce, головною ціллю якої є об'єднання усіх проміжних значень, пов'язаних з одним й тим самим проміжним

ключем. Величезна кількість справжніх задач можуть бути виражені за допомогою цієї моделі.

Програмне забезпечення, що написане у такому функціональному стилі, може бути легко розбито на велику кількість паралельних операцій, тобто виконуватися одночасно на великій кількості процесорів/машин. Системи виконання програмних продуктів мають змогу займатися тонкощами розбиття вхідних даних, плануванням виконання програми на декількох процесорах, обробкою помилок і керуванням необхідних міжпроцесорних операцій комунікації. Це дозволяє розробникам, що мають невеликий досвід роботи з паралельними й розподіленими системами легко використовувати ресурси великих розподілених систем.

Працівники корпорації «Google» розробили тисячі спеціалізованих інструментів, що обробляють великі обсяги необроблених даних, проте більшість обчислень концептуально прості. Через велику кількість вхідних даних, обчислення повинні бути розподілені на сотні або тисячі паралельних одиниць для виконання роботи у розумний час.

Через складнощі з обробкою даних, кількість яких безперервно росте, вченими було розроблено нові абстракції, що дозволяють нам виконувати прості обчислення, але позбутися складнощів розпаралелювання, відмовостійкості, розподілу даних та балансування навантаження. Створена абстракція надихається функціями Map та Reduce, що присутні в багатьох функціональних мовах програмування.

Вчені зрозуміли, що в більшості обчислень залучали операції Map до кожного логічного рядка у вхідних даних задля обчислення масивів проміжних пар ключ/-значення, та подальшого застосування операції Reduce до всіх спільних значень, щоб відповідним чином поєднати отримані дані.

Використання функціональної моделі з реалізованою користувачем операції Map та Reduce, дозволяють легко розпаралелювати обчислення великих даних та використовувати повторне виконання як основний механізм відмовостійкості [8].

Програмна модель приймає на вхід велику кількість пар ключ/значення, та повертає на вихід пари ключ-значення. Так званий користувач «MapReduce» виражає обчислення як дві функції: Map та Reduce.

Функція Map, реалізована користувачем, приймає вхідну пару і виробляє набір проміжних пар ключ/значення. Алгоритм об'єднає всі проміжні значення, пов'язані з одним й тим самим проміжним ключем, та передає їх далі до функції Reduce.

Функція Reduce, також реалізована користувачем, приймає проміжний ключ й набір значень для цього ключа. Функція повинна об'єднати ці значення, щоб утворити менший набір значень. Зазвичай нульове значення або одне вихідне значення виробляється на кожному виклику Reduce. Проміжні значення подаються до функції Reduce через ітератор. Це дозволяє обробляти величезні масиви даних та ефективно використовувати динамічну пам'ять.

В загальному випадку можливі різні реалізації інтерфейсу MapReduce. Правильний вибір залежить від середовища виконання. Наприклад, одна реалізація може підходити для сервера з невеликою кількістю спільної пам'яті, інша реалізація для кластерів з великою кількістю пам'яті та вузлів. Виклики Map розподілені на декілька шляхом розбиття вхідних даних у масив з M елементів. Вхідні розбиття можуть оброблятися паралельно різними потоками. Виклики Reduce розподілені шляхом розбиття проміжних пар ключ/значення на R фрагментів з використанням функції розподілу (наприклад, F = hash(key) mod R). Кількість фрагментів R і F користувачем. На рис. Б.3 показаний загальний потік дій у MapReduce.

В даному пункті був представлений загальний огляд технології MapReduce та приклад її використання. У будь-якому випадку, перед використанням технології необхідно розуміти для яких саме задач вона призначена та чи доцільно її використовувати.

РОЗДІЛ 3
ПРАКТИЧНА ЧАСТИНА

3.1. Постановка задачі

Реалізація програмного продукту ведеться за вимогами, які розглядаються як технічне завдання на розробку. Побудова архітектури й реалізація програмного продукту повинні вестися відповідно до поставлених завдань. Серверна частина повинна бути реалізована мовою програмування Go. Кінцевий продукт повинен являти собою сервіс пошуку автотранспорту за індивідуальним номерним державним знаком з двома методами взаємодії:

- Веб-додаток, що працює за протоколом «HTTP».
- «Telegram» чат-бот.

Створюваний продукт повинен забезпечувати можливість надання кінцевому користувачеві послуг з пошуку інформації про автотранспорт за індивідуальним державним номером або зображенні, на якій чітко видно номерний державний знак.

База даних, на основі якої будується сервіс, повинна забезпечувати охоплення всієї широти можливої інформації, яка може бути цікава кінцевому користувачеві. Під кінцевим користувачем сервісу розуміється користувач чат-бота чи будь-якої іншої реалізації клієнтської частини.

Передбачається, що сервіс повинен компілюватися на більшості операційних системах, а тобто «Linux», «MacOS», «Windows». Оскільки основною мовою програмування повинна бути Go, ми отримаємо змогу розгорнути проект на наступних операційних системах: «Linux», «Windows», «MacOS», «OpenBSD», «DragonFly BSD», «FreeBSD», «NetBSD», «Native Client», «Solaris». Ця мова програмування є компільованою на багатьох операційних системах задля того, щоб процес розгортання системи був максимально простим, саме тому її називають «Cloud Native».

Веб-сервіс повинен обмінюватися повідомленнями з кінцевими користувачами за «REST» протоколом, також він повинен надавати кілька типів пошуку за параметрами і мати можливість повертати результат як одиничною відповіддю, так і масивом відповідей. Під відповіддю мається на увазі полегшена модель даних автотранспорту.

REST — це архітектурний стиль взаємодії компонентів розподіленого додатка в мережі, а точніше узгоджений набір обмежень, що враховуються при проектуванні розподіленої системи. У певних випадках це призводить до підвищення продуктивності та спрощення архітектури. В мережі Інтернет виклик віддаленої процедури може являти собою звичайний «HTTP-запит», а необхідні дані передаються як параметри запиту. До веб-додатків, побудованих з урахуванням «REST», застосовують термін «RESTful».

Сервіс повинен бути захищений від найбільш розповсюджених уражень. Найбільш відомою організацією, що досліджує ураження в мережі Інтернет є «OWASP», тож будемо спиратися на їх дослідження.

1. Помилки у перевірці.
2. Недоліки в системі аутентифікації.
3. Розкриття чутливої інформації.
4. Порушений контроль доступу.
5. Впровадження зовнішніх XML-сутностей.
6. Помилки в конфігурації.
7. Міжсайтовий скриптинг.
8. Небезпечна обробка JSON-сутностей.
9. Використання компонентів з відомими ураженнями.
10. Недостатній рівень моніторингу.

Принципова схема роботи системи представлена на рис. В.1.

3.2. Побудова архітектури системи

Під архітектурою програмного продукту зазвичай мають на увазі сукупність найважливіших рішень щодо організації структури програмного продукту. У цьому розділі розглядаються такі аспекти моделі продукту:

- Склад вхідних даних.
- Модульність компонентів.
- Схеми взаємодії компонентів.
- Структура таблиці бази даних.

Найбільш сучасний підхід до реалізації програмних продуктів такого типу це мікро-сервіси.

Мікро-сервіси — архітектурний стиль за яким єдиний застосунок будується як сукупність невеличких сервісів кожен з яких працює у своєму власному процесі і комунікує з рештою використовуючи певні протоколи обміну, зазвичай HTTP. Ці сервіси будуються навколо бізнес-потреб та розгортаються незалежно з використанням повністю автоматизованого середовища. Самі по собі сервіси можуть бути написані з використанням різних мов і технологій зберігання даних.

Основні переваги:

- Високий рівень незалежності сервісів.
- Простота заміни однієї реалізації сервісу іншою.
- Простота додавання нового функціоналу в систему.
- Ефективне використання ресурсів.
- Вихід з ладу одного сервісу не призводить до виходу з ладу всієї системи.
- Сервіси організовані відносно бізнес логіки яку вони виконують.
- Кожен сервіс незалежно від інших може бути реалізований.

Отже, розділимо додаток на п'ять логічних складових, кожна з яких відповідає безпосередньо на одну функцію. Це дає змогу розробляти якісніші програмні продукти, оскільки кожен з додатків вирішує конкретне завдання.

Схема роботи системи зображена на рис. В.3. Задача розгортання сховища даних є досить тривіальною. За сховище було обрано реляційну базу даних PostgreSQL,

яка вже багато разів зарекомендувала себе як швидка та надійна.

Чат-бот сервіс.

Інтерфейс комунікації між месенджером та чат-бот сервісом влаштований наступним чином: коли користувач відправляє повідомлення, Telegram відправляє HTTP запит до сервісу, який у свою чергу обробивши запит робить відповідний виклик до Bot API.

Детекція державних номерних знаків на зображенні. Задача розпізнавання тексту на зображеннях є дуже цікавою темою. У відкритому доступі існує величезна кількість цікавих рішень, проте знайти бібліотеку що є безкоштовною та підтримує українські державні номери автотранспорту виявилося досить просто.

«OpenALPR Cloud API» — це веб-сервіс, що аналізує зображення транспортних засобів та розпізнає дані ліцензійних номерів, а також колір, марку, модель і тип кузова. Насправді, цей сервіс досить дорогий, проте він реалізований на основі бібліотеки «OpenALPR», яка у свою чергу знаходиться у вільному доступі в мережі.

Таким чином, розробимо веб сервіс, що приймає на вхід зображення, досліджує її за допомогою бібліотеки та повертає результат у вигляді державного номера.

Оброблювач вхідних даних. Міністерство Внутрішніх Справ України публікує дані про транспортні засоби та їх власників раз у місяць, зазвичай у перших числах календарного місяця. Дані надходять у вигляді величезних CSV документів, які можуть змінені у будь-який момент. Оскільки ніхто не гарантує, що формат цих даних незмінний — автоматизувати їх обробку є досить складним завданням, а тому не будемо робити дослідження цієї проблеми в рамках цієї роботи.

Отже кожного місяця потрібно ефективно та швидко обробляти нову інформацію про транспортні за засоби. Оскільки дані зберігаються у вигляді CSV документів, а час пошуку будь-яких даних у цих документах досить великий, оброблювач повинен читати ці дані та записувати їх у спеціальному вигляді до бази даних, а пошук у проіндексованих таблицях досить швидкий. Дослідивши сучасні

підходи до розробки багато-поточних систем за технологією MapReduce у попередніх розділах, можна реалізувати досить швидкий та надійний до запобігання помилок оброблювач CSV документів.

Алгоритм роботи та деталі реалізації цієї компоненти буде розглянуто у наступних підрозділах.

Головний веб-сервіс. Кінцевий крок — швидкісний сервер пошуку автотранспорту за державним номером у базі даних.

Розділивши логіку на складові, задача пошуку номерів тепер здається на рівень легшою, потрібно лише обрати правильні інструменти, а тобто швидку та надійну мову програмування та базу даних.

Алгоритм роботи та деталі реалізації цієї компоненти буде розглянуто у наступних підрозділах.

3.3. Розробка серверної частини

До серверної частини відносяться сервіси, що мають наступні три мети:

- Обробка вхідних даних.
- Пошук даних.
- Розпізнання державних номерів на зображеннях.

Пропонуємо більш детально оглянути, яким чином це буде реалізовано, у наступних підрозділах.

Обробка вхідних даних. У цьому підрозділі буде детально розглянуто алгоритм обробки даних. Головною метою цього розділу є знаходження та демонстрація чіткого алгоритму, який може бути застосований до будь-яких типів даних, що знаходяться у відкритому доступі на єдиному порталі відкритих даних України.

Отже, одним з найважливіших аспектів цієї дипломної роботи є швидкість оброки великих масивів даних. Підсумовуючи викладений у попередніх розділах матеріал, можна стверджувати, що одним з найкращих технологій оброки даних є MapReduce, тож саме цей метод був узятий за основу розробленого алгоритму.

Слід зазначити дуже гарні результати у пошуку проіндексованих даних у ре-

ляційних базах даних. На рис. Б.1 бачимо близько десяти тис. запитів на секунду, що є достатнім майже для будь-якої системи. У цьому дослідження було використано базу даних PostgreSQL, тому що вона має перевагу над іншими у швидкості пошуку та індексації, а також ця база підтримує складні структури даних, має велику кількість різноманітних типів даних та дуже надійна у збереженні великих масивів даних.

Розглянемо детальніше структуру CSV документів та алгоритм оброки. У табл. А.1 можна побачити список полів, що містить кожний документ. Як бачимо, дані можуть бути оброблені в альтернативний формат, тож побудуємо SQL таблицю для збереження даних про операції над транспортними засобами.

На рис. В.2 роботи обробника CSV документів у SQL можна побачити, що алгоритм має п'ять кроків. На першому кроці програма створює новий об'єкт читання з документа формату CSV, для того щоб читати та групувати дані з файлу у невеличкі масиви рядків. З першого на другий крок дані потрапляють через так званий «channel». Це примітив у мові програмування Go, що є аналогом черги для декількох процесів. На другому кроці програма постійно групує дані у масиви по N елементів, а потім передає їх через до наступної частини програми. Процес безпосередньої обробки та перевірці є досить складним через виділення нової пам'яті та роботи зі строками. Для пришвидшення було створено декілька процесів, що постійно чекають на масиви даних, які потрібні бути оброблені. Після третього кроку, що насправді є прикладом функції Map у MapReduce, оброблені дані потрапляють до 4 кроку. На четвертому кроці програма знову групує дані у масиви по M елементів, а потім передає їх через до фінального кроку. На останньому кроці масив отриманих даних публікується до сховища.

Навіщо потрібно групувати дані на масиви, адже це може сповільнити процес обробки спитає недосвідчений розробник. Операції на масивах даних відбуваються в декілька разів швидше, тому що мінімізується час на відправлення та оброку повідомлень. Гарним прикладом може бути процес читання з файлу. Читання одразу декількох стрічок пришвидшує алгоритм у 5 разів, головне не перебільшити та раціонально використовувати оперативну пам'ять машини.

Пошук даних. Пошук даних реалізовано через веб-інтерфейс, що отримує на вхід державний номер та повертає список операцій, у яких фігурує даний державний номер. Ця компонента має прямий доступ до бази даних, у якій зберігаються дані про транспорті засоби. Такий підхід гарантує мінімальний час на пошуковий SQL запит.

Цей інтерфейс реалізує технологію REST, що була розглянута у теоретичній частині цієї дипломної роботи. Тож протокол який використовується — HTTP, а формат відповіді та запиту сервера — JSON. Такі характеристики дають можливість використовувати цей інтерфейс без будь-яких обмежень.

Таким чином, при отриманні нового HTTP GET запиту, сервер оброблює його та декодує передані параметри, а тобто номерний державний знак та максимальну кількість операцій, що можуть бути повернені. На наступному кроці відбувається пошуковий SQL запит, що повинен отримати усі потрібні дані. Результат отриманий від бази обробляється та повертається до кінцевого користувача у вигляді JSON відповіді. Увесь процес не повинен займати більш ніж декілька мілісекунд, якщо дані проіндексовані та правильно оброблені.

Важливим аспектом пошуку є форматування символів номерного державного знаку. Дані у сховищі зберігаються в вигляді кирилиці, а літери у запиті можуть мати інші формати. Аби запобігти цю проблему, на кожному запиті номер транслітерується у кирилічний аналог, таким чином гарантується запобігання будь-яких інших літер.

Усі функції цієї компоненти були протестовані, також було оцінено швидкість роботи інтерфейсу. На рис. Б.4 бачимо, що сервіс може обробляти від 200 до 500 запитів на секунду. Приблизно такі самі результати можна отримати, якщо використовувати сервіс за допомогою чат-бота. Якщо використовувати більші ресурси, наприклад спеціалізований окремий сервер з базою даних, то можна отримати більш ніж 3 тис. запитів на секунду, тож є величезний запас швидкості.

Сервіс реалізований стандартними засобами мови програмування Go, що були створені кращими світовими розробниками з компанії Google.

Розпізнання державних номерів на зображенні. OpenALPR — це бі-

бліотека автоматичного розпізнавання державних номерних знаків з відкритим вихідним кодом, написана на C++ з можливостями використовуватися у Java, Node.js і Python. Бібліотека аналізує зображення та відео для ідентифікації номерних знаків. Вихідні дані — це текстове представлення державного номерного знака.

Програмне забезпечення можна використовувати різними способами. У нашому випадку потрібно розробити веб-сервіс, вхідними даними якого буде зображення, а вихідними буде номерний знак.

Дана бібліотека створена для американських номерних знаків, проте її можна навчити розпізнавати й українські державні номерні знаки.

Оптичне розпізнавання тексту — це механічне або електронне переведення зображень рукописного, машинописного або друкованого тексту в послідовність кодів, що використовуються для представлення в текстовому редакторі. Розпізнавання широко використовується для конвертації книг і документів в електронний вигляд, для автоматизації систем обліку в бізнесі або для публікації тексту на веб-сторінці. Оптичне розпізнавання тексту дозволяє редагувати текст, здійснювати пошук слова або фрази, зберігати його в компактнішій формі, демонструвати матеріал, не втрачаючи якості, аналізувати інформацію, а також застосовувати до тексту електронний переклад, форматування або перетворення в мовлення. Оптичне розпізнавання тексту є досліджуваною проблемою в галузях розпізнавання образів, штучного інтелекту і комп'ютерного зору.

Системи оптичного розпізнавання тексту вимагають калібрування для роботи з конкретним шрифтом, тому бібліотеку OpenALPR потрібно навчити розпізнавати шрифт українських номерних знаків. Таким чином, за допомогою двох сотень зображень та декількох годин тренування моделі отримано результат у вигляді 80% розпізнання номерних знаків.

3.4. Розробка клієнтської частини

До клієнтської частини належить лише сервіс чат-боту.

Чат-бот — це програмне забезпечення штучного інтелекту, що може імітувати бесіду з користувачем природною мовою через спеціалізовані програми для обміну повідомленнями. Ця технологія часто описується як один з найбільш передових і перспективних методів взаємодії між людьми й машинами. Однак, з технологічної точки зору, чат-бот представляє лише природну еволюцію системи відповіді на запитання. Формулювання відповідей на питання природною мовою є одним з найбільш типових прикладів обробки природної мови, що застосовується в кінцевих додатках різних підприємств.

У контексті системи Telegram, боти — це програми сторонніх розробників, які працюють у системі Telegram. Користувачі можуть взаємодіяти з ботами, надсилаючи їм повідомлення, команди та спеціалізовані запити. Розробники можуть керувати своїми ботами за допомогою HTTP-запитів до Telegram Bot API.

Насправді ж, чат-бот є лише простим сервером, що відповідає на певний запит від Telegram. На початку роботи чат-бота, сервер повідомлює адресу системі. При надходженні повідомлення до бота, система робить запит на сервер з ботом, після оброки запиту сервіс чат-бота відповідний запит з повідомленням, що повинно бути відправлене користувачу.

Таким чином, наш сервер повинен обробити текстове повідомлення або зображення від користувача. Чат-бот зробить запит до сервісу пошуку даних та сконструювати текстову відповідь з отриманого JSON, а потім відправити її до серверів Telegram.

ВИСНОВКИ

На прикладі відкритих даних про українські транспортні засоби, що публікуються на єдиному державному порталі відкритих даних, були вивчені ефективні методи обробки великих масивів даних та детально розібрано процес обробки відкритих даних з цього порталу. Запропонований алгоритм роботи з даними на єдиному державному порталі відкритих даних.

Отже, можна стверджувати, що сфера даних має великий потенціал у нашій країні, проте якість даних, що викладаються на порталі, на жаль, є досить низькою. Стало зрозуміло, що формат документів повинен бути змінений якнайшвидше, оскільки із зростанням кількості даних - зростає складність та час їх обробки.

Таким чином, вища якість даних повинна призвести до більших часових та грошових інвестицій у сфері оброки даних в Україні, а отже можна отримати додаткові 1.5 млрд. доларів у бюджет країни.

Результатом проведеної дипломної роботи є спроектований та створений програмний продукт. Для усіх мешканців України сервіс є простим і зручним інструментом пошуку інформації про автотранспорт.

Під час дослідження ми дізналися, що Україна стає частиною єдиного Європейського інформаційного простору і ще більше наближається до інтеграції з Євросоюзом. Завдяки відкритим даним Україна стає більш привабливою для західних інвестицій. Український уряд намагається розвивати сферу та заохочує нових спеціалістів за допомогою створення відповідних конкурсів та тендерів на розробку додатків з використанням відкритих даних.

В ході написання роботи на першому етапі була розроблена архітектура майбутнього сервісу з урахуванням всіх вимог і побажань. На другому етапі програмний продукт був реалізований в суворій відповідності з встановленими вимогами та створеної архітектурою. Під час реалізації продукту були використані сучасні те-

хнології. Так само було приділено увагу питанню подальшого можливого коригування вихідного продукту. Продукт вийшов досить гнучким на випадок бажання замовника змінити середу або тип бази даних. У ході реалізації було приділено увагу використанню надійних технологій збереження даних. На останньому етапі увесь функціонал був перевірений в ручному режимі.

Розроблений програмний продукт має відкритий код, та може використовуватися у комерційних цілях будь-якого підприємця чи компанії. Система була розгорнута, тож будь-який користувач системи Telegram може її використовувати.

Було вдало оброблено понад 9 млн записів з документів формату CSV, які були викладені Міністерством Внутрішніх Справ України у якості відкритих даних. Загальний час оброки усіх файлів займає менш як 10 хвилин. На жаль, у цих даних відсутня інформація щодо ідентифікаційних номерів транспортних засобів, тож не можливо прослідкувати історію транспорту, а лише побачити поточні дані.

Підбиваючи підсумки виконаної роботи, можна стверджувати, що мета дипломної роботи повністю виконана — прототип швидкісного веб-сервісу, що надає інформацію про український транспорт за державним номерним знаком створений, вимоги враховані і реалізовані.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] А. Газін. Екосистема відкритих даних в Україні: рекомендації щодо впровадження політики. — 2015.

[2] Д. Гурський. Що читати, слухати та куди подаватися Open Data Lovers? — 2019.

[3] Постанова від 21 жовтня 2015 р. № 835: положення про набори даних, які підлягають оприлюдненню у формі відкритих даних. — 2015.

[4] Пігарєв Ю. Дрешпак В. Куспляк І. Доступ до публічної інформації: частина 11. — EGAP, 2017. — ISBN: 9789662214789.

[5] B. Peter. Big Data Fundamentals: Concepts, Drivers Techniques. — 1 вид. — Prentice Hall, 2015. — 12. — Т. 1. — ISBN: 9780134291079.

[6] B. Tim. Five Star Open Data. — 2012. — https://5stardata.info.

[7] G. Brett. Beyond Transparency: Open Data and the Future of Civic Innovation. — 1 вид. — Code for America Press, 2013. — 10. — Т. 1. — ISBN: 9780615889085.

[8] George C. Jean D. Tim K. Distributed Systems: Concepts and Design. — 5 вид. — Pearson, 2011. — 6. — Т. 1. — ISBN: 9780132143011.

[9] Jean-Louis M. Soraya S. Big Data, Open Data and Data Development. — Willey, 2016. — 3. — Т. 3. — ISBN: 9781848218802.

[10] Jeffrey D. Sanjay G. MapReduce: Simplified Data Processing on Large Clusters. — 2004. — https://ai.google/research/pubs/pub62.

[11] Open Data Handbook. — 2015. — http://opendatahandbook.org.

[12] Open Definition. — 2006. — http://opendefinition.org.

[13] Stefaan V. Andrew Y. The Global Impact of Open Data. — O'Reilly, 2016. — 7. — ISBN: 9781492042785.

Додаток А

Назва	**Опис**
oper_code	Код
oper_name	Назва
d_reg	Дата операції
dep	Місце проведення
brand	Марка
model	Модель
make_year	Рік випуску
color	Колір
kind	Вид
body	Тип кузова
purpose	Характеристика
fuel	Вид пального
capacity	Двигун
total_weight	Загальна вага
n_reg_new	Номерний знак

Таблиця А.1

Додаток Б

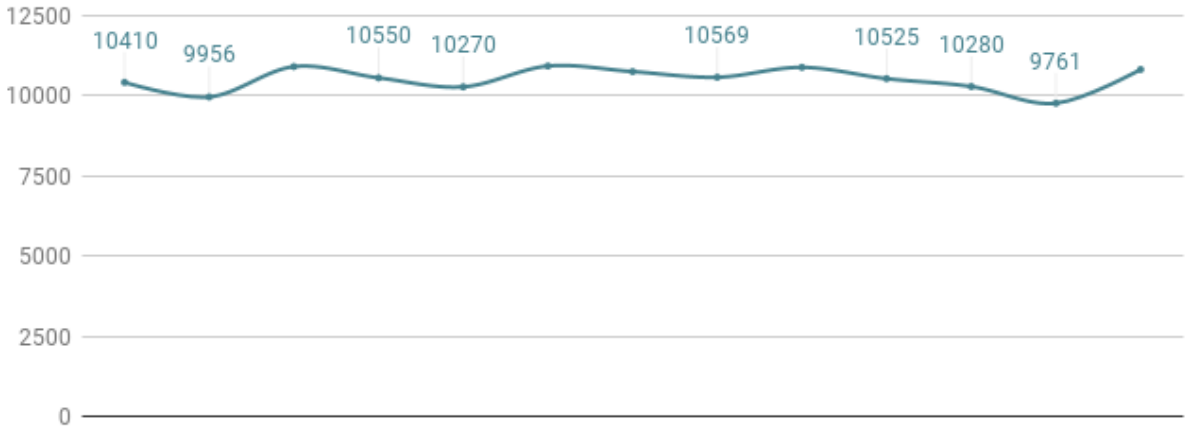

Рис. Б.1. Оцінка роботи обраної бази даних PostgreSQL (запитів на сек.)

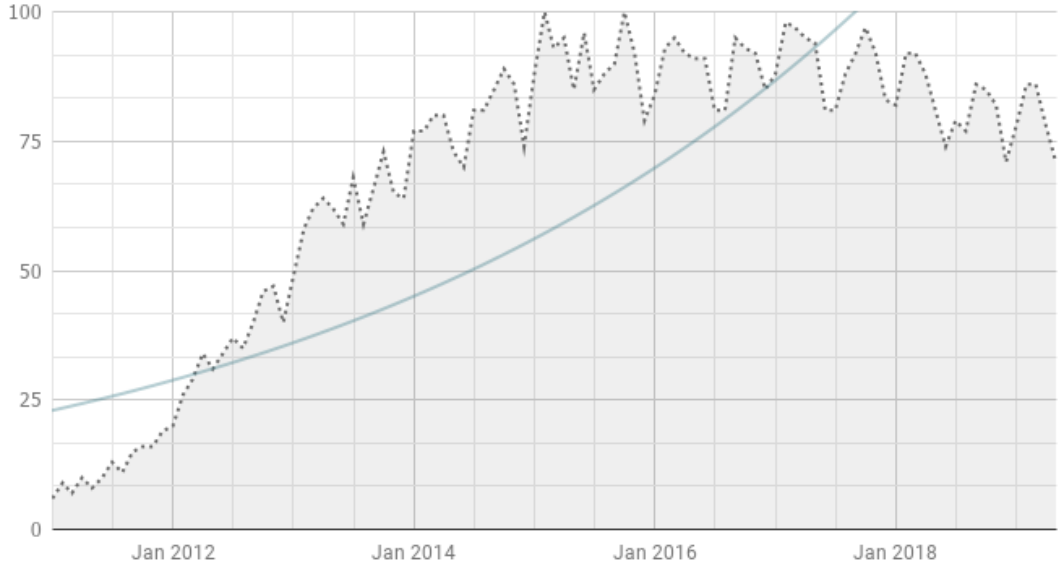

Рис. Б.2. Тренд пошукових запитів із словосполученням «Big Data»

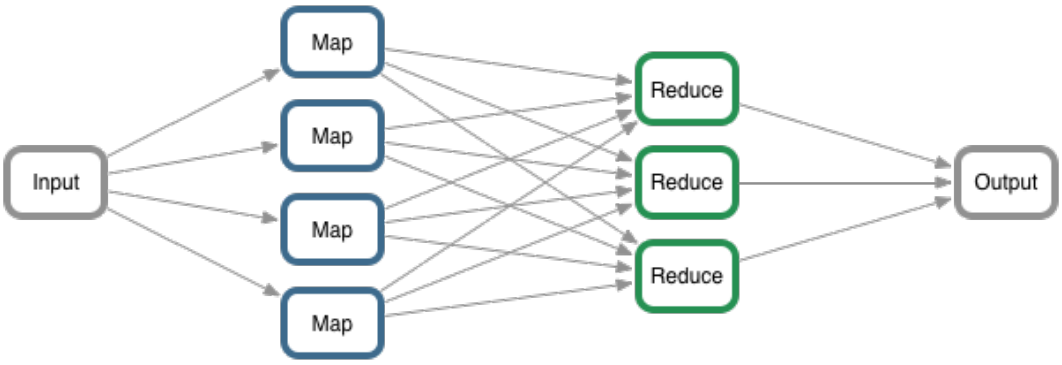

Рис. Б.3. Схема роботи MapReduce

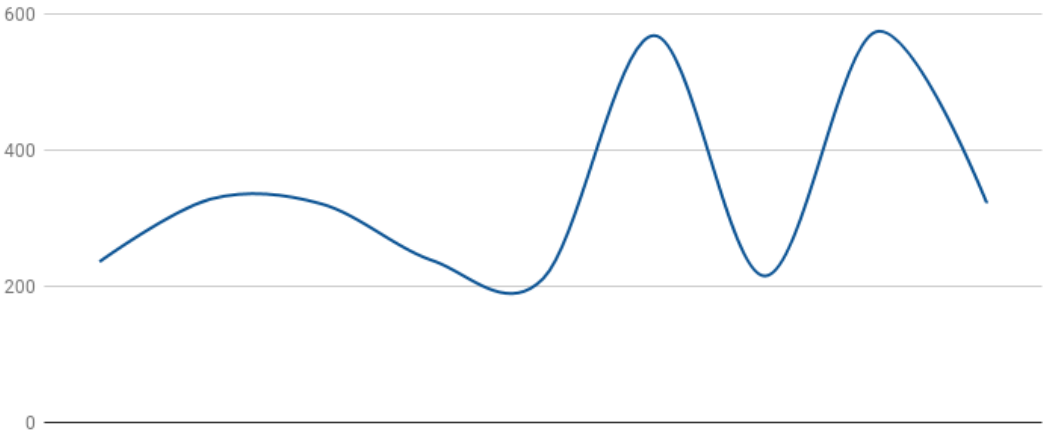

Рис. Б.4. Оцінка роботи розробленого веб-інтерфейсу

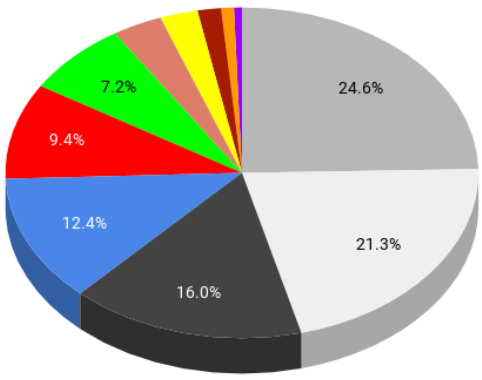

Рис. Б.5. Розбиття кольорів Українського автотранспорту

Додаток В

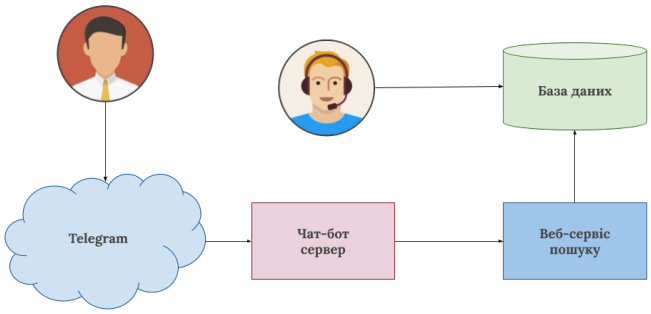

Рис. В.1. Схема роботи користувацької системи

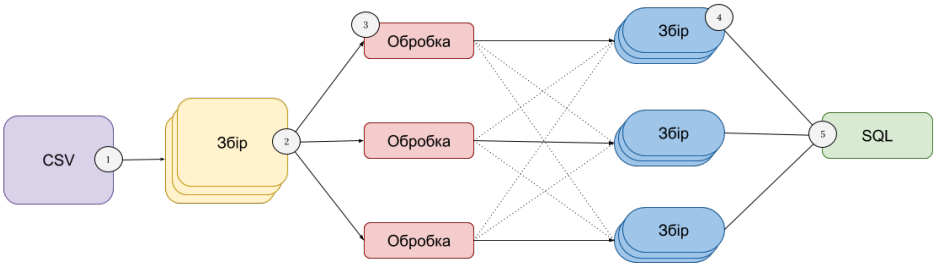

Рис. В.2. Обробка документів формату CSV у SQL

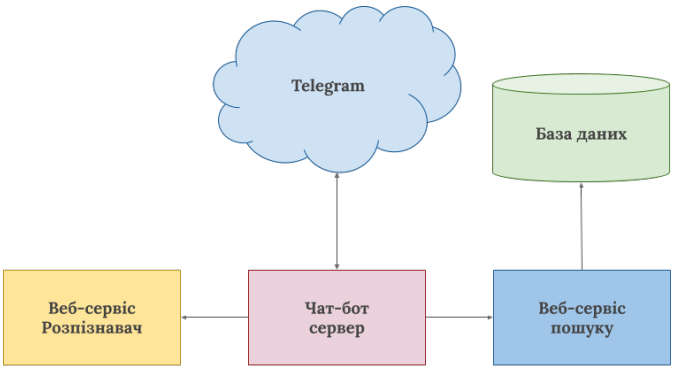

Рис. В.3. Повна схема роботи системи

```
1  func mapper(input chan []string, output chan model.Operation) {
2      for {
3          msg, opened := <- input
4          if !opened {
5              return
6          }
7
8          output <- *model.NewOperation(msg)
9      }
10 }
```

Лістинг В.1. Реалізація функції Map.

```
1  func reducer(input chan []model.Operation, output chan struct{}) {
2      db := database.Must(database.DB())
3      defer db.Close()
4
5      for {
6          operations, open := <-input
7          if !open {
8              output <- struct{}{}
9              return
10         }
11
12         if err := db.Insert(&operations); err != nil {
13             // Handle error.
14         }
15     }
16 }
```

Лістинг В.2. Реалізація функції Reduce.