НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Навчально-науковий інститут прикладного системного аналізу
Кафедра системного проектування

«На правах рукопису»
УДК _____

До захисту допущено:
Завідувач кафедри
_____ Вадим МУХІН
«___»_____2021 р.

Магістерська дисертація

на здобуття ступеня магістра

за освітньо-професійною програмою

“Інтелектуальні сервіс-орієнтовані розподілені обчислювання”

зі спеціальності 122 "Комп'ютерні науки"

на тему: «Методи ефективного зберігання великих обсягів медичних даних в умовах їх неперервного зростання»

Виконала:
студентка ІІ курсу, групи ДА-01мп
Материнська Софія Василівна _____

Науковий керівник:
проф., д.т.н., Рогоза Валерій Станіславович
_____

Консультант з основного розділу:
ас. Яременко Вадим Сергійович _____

Рецензент:
доц., к.т.н., Смаковський Денис Сергійович _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студентка _____,

Київ – 2021

Національний технічний університет України
«Київський політехнічний інститут
імені Ігоря Сікорського»

Інститут/факультет Прикладного системного аналізу
(повна назва)

Кафедра Системного проектування
(повна назва)

Рівень вищої освіти – другий (магістерський) за освітньо-професійною програмою Інтелектуальні сервіс-орієнтовані розподілені обчислення

Спеціальність (спеціалізація) 122 Комп`ютерні науки
(код і назва)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____В.Є.Мухін.___
(підпис) (ініціали, прізвище)
«___»_____2021__р.

ЗАВДАННЯ
на магістерську дисертацію студенту

Материнській Софії Василівні
(прізвище, ім'я, по батькові)

1. Тема дисертації "Методи ефективного зберігання великих обсягів медичних даних в умовах їх неперервного зростання"

науковий керівник дисертації Рогоза Валерій Станіславович, д.т.н. професор,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «_02_»___11____ 2021_ р. №___3651-с__

2. Строк подання студентом дисертації 20 грудня 2021

3. Об'єкт дослідження: методи та підходи, що сприяють оптимізації ефективності обробки та використання великих даних за умов неперервного зростання їх обсягів.

4. Предмет дослідження (Вихідні дані – для магістерської дисертації за освітньо-професійною програмою)

Технології та підходи, що дозволяють покращити ефективність обробки даних та оптимізувати їх зберігання і використання - DataOps, edge computing, in-memory computation, schema optimization.

5. Перелік завдань, які потрібно розробити

1. Дослідження предметної області, огляд останніх досліджень та публікацій за обраним напрямом.
2. Аналіз основних проблем покращення ефективності систем обробки великих даних та варіантів їх вирішення.
3. Детальне дослідження таких напрямків як in-memory computations, edge computing, DataOps, schema optimization та проектування плану реалізації ефективного зберігання великих обсягів медичних даних в умовах їх неперервного зростання.
4. Програмна реалізація запропонованого рішення, аналіз отриманих результатів.
5. Формулювання висновків на основі отриманих результатів.

6. Перелік графічного (ілюстративного) матеріалу

Графічні порівняння результатів, схематичне зображення моделі системи.
Презентація до захисту роботи.

7. Орієнтовний перелік публікацій

1) Julian Ereth, DataOps – Towards a Definition, 2018. (http://ceur-ws.org/Vol-2191/paper13.pdf)
2) Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, Lanyu Xu, Edge Computing: Vision and Challenges, 2016. (https://ieeexplore.ieee.org/abstract/document/7488250)
3) Non Sanprasita, Katechan Jampachaisrib, Taravichet Titijaroonrojc, Kraisak Kesorn, Intelligent approach to automated star-schema construction using a knowledge base, 2021 (https://www.sciencedirect.com/science/article/abs/pii/S0957417421006588)

8. Консультанти розділів дисертації[1*]

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Основний	ас. Яременко В.С.		

[1*] Консультантом не може бути зазначено наукового керівника

9. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Строк виконання етапів магістерської дисертації	Примітка
1	Отримання завдання	01.09.2021	
2	Збір інформації	10.09.2021	
3	Ознайомлення з літературою і підготовка теоретичної частини роботи	20.09.2021	
4	Аналіз вимог завдання, вибір методів і засобів розв’язання поставленої задачі	01.10.2021	
5	Розробка плану дослідження	05.10.2021	
6	Практичне виконання дослідження	20.10.2021	
7	Звіт з проходження практики	25.10.2021	
8	Аналіз результатів	15.11.2021	
9	Формулювання висновків	01.12.2021	
10	Оформлення дипломної роботи	10.12.2021	
11	Отримання допуску до захисту та подача роботи в ДЕК	20.12.2021	

Студент _____ __С.В. Материнська___
(підпис) (ініціали, прізвище)

Науковий керівник дисертації _____ __В.С. Рогоза_____
(підпис) (ініціали, прізвище)

РЕФЕРАТ НА МАГІСТЕРСЬКУ ДИСЕРТАЦІЮ

виконану на тему: «Методи ефективного зберігання великих обсягів медичних даних в умовах їх неперервного зростання»

студенткою Материнською Софією Василівною

Актуальність. Широкий набір методів і підходів, що надає напрямок великих даних, дозволяє ефективно зберігати, обробляти та аналізувати дуже великі обсяги інформації за максимально короткий термін, що і є ключовою перевагою застосування великих даних. Поєднання цих методів зі сферою охорони здоров'я є перспективним напрямком, що активно розвивається. Медична сфера щороку продукує великі обсяги даних, які потребують певної обробки з метою отримання цінної інформації на їх основі. На даний момент є досить багато успішних випадків застосування аналітики великих даних до сфери охорони здоров'я. Подальший розвиток і покращення даної області може принести переваги для пацієнтів, медичного персоналу та інших залучених сторін. Різні дослідження аналітики великих даних для медицини зазначають різні методи зберігання, що обмежуються в основному прикладами конкретних засобів.

Мета і задачі дослідження. Метою цієї роботи є виявлення проблем, що пов'язані зі зберігання великих об'ємів медичних даних та можливих шляхів вирішення однієї з них. Задачі дослідження полягають в аналізі актуальних публікацій, статей та інших джерел інформації, що стосуються застосування аналітики великих даних у сфері медицини, з метою визначення конкретних напрямків їх використання, проблем та переваг, що вони надають, методів зберігання великих даних та покращення їх ефективності. Також важливим завданням є приведення способу для вирішення деяких проблем, що стосуються ефективності зберігання даних.

Об’єкт дослідження. Методи та підходи, що сприяють оптимізації ефективності обробки та використання великих даних за умов неперервного зростання їх обсягів.

Предмет дослідження. Технології та підходи, що дозволяють покращити ефективність та оптимізувати зберігання і використання великих обсягів медичних даних.

Наукова новизна одержаних результатів. Полягає в розробці прототипу дерева прийняття рішень для визначення найоптимальнішого сховища даних.

Практичне значення одержаних результатів. Програмна імплементація отриманого дерева рішень може бути розроблена як окремий засіб допомоги у визначенні найкращого методу зберігання медичних даних для відповідного проекту чи дослідження на моменті його планування та практичного виконання.

На основі виконаної роботи подано статтю на публікацію в періодичне фахове видання категорії Б.

Загальний обсяг роботи – 105 сторінок, 24 таблиці, 9 рисунків, 29 посилань.

Ключові слова: великі дані, охорона здоров'я, медицина, NoSQL, бази даних, колонко-орієнтовані, документо-орієнтовані, ключ-значення, HDFS, Hadoop, хмарні сервіси.

ABSTRACT FOR MASTER'S DISSERTATION

performed on the topic: "Methods of effective storage of large medical data amounts in conditions of their continuous growth"

by student Sofiia Materynska

Actuality of theme. A wide range of methods and approaches, which provide the big data area, allows to efficiently store, process and analyze very large amounts of information in the shortest possible time, which is a key advantage of using big data. The combination of these methods with the field of health care is a promising area that is actively developing. The medical field annually produces large amounts of data that require some processing in order to obtain valuable information based on them. At present, there are many successful applications of big data analytics to healthcare. Further development and improvement of this area can bring benefits to patients, medical staff and other stakeholders. Various studies of big data analytics for medicine point to different storage methods, which are limited mainly to examples of specific tools.

The purpose and objectives of the study. The purpose of this work is to identify problems associated with the storage of large amounts of medical data and possible solutions to one of them. The objectives of the study are to analyze current publications, articles and other sources of information related to the application of big data analytics in medicine, to identify specific areas of their use, problems and benefits, methods of storing big data and improvement of their efficiency. Another important task is to provide a way to solve some problems related to the efficiency of data storage.

Object of study. Methods and approaches that help to optimize the efficiency of processing and use of big data in the conditions of continuous growth of their volumes.

Subject of study. Technologies and approaches to improve the efficiency and optimize the storage and use of large amounts of medical data.

Scientific novelty of the obtained results is to develop a prototype decision tree to determine the best data warehouse.

The practical significance of the results obtained. The software implementation of the resulting decision tree can be developed as a separate tool to help determine the best method of storing medical data for the project or research at the time of its planning and implementation.

Based on the work done, an article was submitted for publication in a periodical professional publication of category B.

The total volume of the work is 105 pages, 24 tables, 9 figures, 29 references.

Keywords: big data, healthcare, medicine, NoSQL, databases, column-oriented, document-oriented, key-value, HDFS, Hadoop, cloud services.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ, ТЕРМІНІВ

BI – Business Intelligence, інтелектуальний аналіз даних

Big data - великі дані

EHR – electronic health record (електронна картка здоров’я)

EMR – electronic medical record (електронний медичний запис)

GWAS – genome-wide association studies (повногеномний пошук асоціацій)

NGS – next-generation sequencing (секвенування «нового» покоління)

NoSQL – not only SQL

SaaS – Software as a Service - програма як послуга

ШІ – штучний інтелект

ВСТУП

Сьогодні індустрія охорони здоров’я стикається з проблемою роботи з великими медичними даними, які швидко розвиваються. Сфера аналітики великих даних розвивається і вже представляє значні результати у різних областях її застосування, проте вона може також надати корисну інформацію для системи охорони здоров’я. Великі дані можуть покращити надання медичної допомоги та знизити її вартість, підтримуючи розширений догляд за пацієнтами, покращуючи результати пацієнтів та уникаючи непотрібних витрат.

Аналітика великих даних зараз використовується для прогнозування результатів рішень, прийнятих лікарями, результатів операції на серці на основі віку пацієнта, поточного стану та стану здоров’я[11]. По суті, ми можемо сказати, що роль великих даних у секторі охорони здоров’я полягає в управлінні наборами даних, пов’язаних з охороною здоров’я, які є складними для керування за допомогою сучасних апаратних засобів, програмного забезпечення та інструментів управління.

Попри значний потенціал, що відкривається за умови інтеграції обробки великих даних з інформацією, що продукує область медицини, є також і низка проблем, які необхідно вирішити, щоб забезпечити їх подальшу ефективну взаємодію. Звичайно, багато досліджень здійснюється з використанням обробки великих медичних даних, але вони пропонують конкретні рішення для конкретних проблем.

Існує багато платформ та підходів для зберігання даних і ще більше засобів, що їх реалізують і кожен з них підходить краще для певної ситуації, кожен володіє набором характеристик, що можуть виступати як перевагою, так і недоліком у різних випадках.

Метою цієї роботи є огляд наявних результатів, їх обробка, подальший аналіз та створення універсальнішого методу, що буде

допомагати розв'язувати проблеми ефективності зберігання великих обсягів медичних даних.

Особливість медичних даних полягає в широкій різноманітності представлених форматів, що необхідно інтегрувати чи уніфікувати. Це пов’язано з нестандартно великою кількістю джерел інформації, яка може бути як структурованою, так і неструктурованою. Розумне й ефективне управління цими даними може стати ключем на шляху до покращення медицини.

Оброблені медичні дані можуть бути використані як основа для діагностування методами машинного навчання та штучного інтелекту, для з’ясування закономірностей та прихованих зв’язків. Правильне застосування отриманих знань може допомогти змінити погляд на сучасну систему охорони здоров’я.

1 ОГЛЯД ДОСЛІДЖУВАНОЇ ОБЛАСТІ

У світі ми спостерігаємо значне зростання об’ємів даних, що можуть нести в собі важливу інформацію. Прийнято вважати, що значні об’єми даних в основному генеруються в Інтернеті, соцмережах та мають “штучне” походження. Проте електронізація даних з інших сфер життя також може принести користь для суспільства. Наприклад, у сфері охорони здоров’я обробка великих масивів даних може принести цікаві та важливі результати, що в перспективі будуть застосовуватись у медицині. Для роботи з великими масивами даних, які є об’єктом цього дослідження, заведено використовувати набір технологій та підходів, що об’єднаний під загальною назвою Big Data (з англійської - Великі Дані).

1.1 Актуальність сфери великих даних

Ми є свідками прогресивної «датафікації» суспільного життя. Діяльність людини та взаємодія з навколишнім середовищем контролюються та реєструються з підвищенням ефективності, створюючи величезний цифровий слід. Отримані «великі дані» є скарбницею для досліджень, і все більш складні обчислювальні інструменти розробляються для вилучення знань з таких даних.

Ми можемо використовувати ці дані для прогнозування поточних тенденцій певних параметрів і майбутніх подій. Оскільки ми все більше і більше усвідомлюємо це, ми почали виробляти та збирати більше даних майже про все, впровадивши технологічні досягнення у цьому напрямку. Сьогодні ми стикаємося з ситуацією, коли ми переповнені безліччю даних з усіх аспектів нашого життя, таких як громадська діяльність, наука, робота, охорона здоров'я тощо. Певним чином ми можемо порівняти нинішню ситуацію з затопленням даними.

Щодня люди, які працюють з різними організаціями по всьому світу, генерують величезну кількість даних. Термін «цифровий Всесвіт» кількісно визначає такий величезний обсяг даних, створених, відтворених та спожитих за один рік. Міжнародна корпорація даних (IDC) оцінила приблизний розмір цифрового Всесвіту в 2005 році до 130 ексабайт (EB). Цифровий Всесвіт у 2017 році розширився приблизно до 16 000 ЕБ або 16 зетабайт (ZB). IDC передбачав, що цифровий всесвіт розшириться до 40 000 EB до 2020 року. Щоб уявити такий розмір, нам доведеться призначити близько 5200 гігабайтів (ГБ) даних для кожної людини зі всіх людей. Це є прикладом феноменальної швидкості, з якою розширюється цифровий всесвіт. Інтернет-гіганти, такі як Google та Facebook, збирали та зберігали величезну кількість даних. Наприклад, залежно від наших уподобань, Google може зберігати різноманітну інформацію, включаючи місцезнаходження користувача, налаштування реклами, список використаних програм, історію перегляду Інтернету, контакти, закладки, електронні листи та іншу необхідну інформацію, пов’язану з користувачем. Так само Facebook зберігає та аналізує понад 30 петабайт (ПБ) даних, створених користувачами. Такі великі обсяги даних становлять «великі дані». За останнє десятиліття великі дані успішно використовуються IT-індустрією для отримання критичної інформації, яка може приносити значний дохід[4].

1.2 Визначення “Big Data”

Як випливає з назви, "Big Data" (“Великі Дані”) являє собою великі обсяги даних, якими не можна керувати за допомогою традиційного програмного забезпечення або інтернет-платформ. Вони перевершують традиційно використовуваний обсяг пам’яті, обробки та аналітичної потужності. Сукупність цифрових даних, створених за допомогою досліджень, зростає шаленими темпами та способами, які неможливо

зрозуміти когнітивною системою людини а, отже, потребують певної форми автоматизованого аналізу.

Можливо, найпростішою характеристикою “Big Data” є великі набори даних, які створюються в цифровій формі та можуть бути проаналізовані за допомогою обчислювальних засобів. Недавні уявлення про великі дані як «все, що не можна легко зафіксувати в електронній таблиці Excel» неминуче швидко змінюється у міру створення нового аналітичного програмного забезпечення, а сама ідея використання електронних таблиць для збору даних відходить у минуле. Крім того, розмір та швидкість даних не враховують різноманітності типів даних, які використовуються дослідниками, які можуть включати дані, які не генеруються у цифрових форматах або їх формат не піддається обчисленню. І акцент на фізичних характеристиках даних приховує постійну залежність інтерпретації даних від обставин їх використання, включаючи конкретні запити, значення, навички та дослідницькі ситуації.

Бойд і Кроуфорд ототожнюють Big Data зі «здатністю шукати, об’єднувати та перехресно посилатися на великі набори даних», тоді як О'Меллі та Сойер зосереджуються на здатності опитувати та взаємопов’язувати різні типи даних, щоб мати можливість ознайомитися з ними як єдиним доказом.

Незважаючи на те, що існує низка визначень великих даних, найпопулярніше та прийняте визначення дав Дуглас Лейні. Лейні помітив, що (великі) дані зростають у трьох різних вимірах, а саме: об’ємі(volume), швидкості(velocity) та різноманітності(variety) (відомі як 3 Vs) [4].

«Велика» частина великих даних свідчить про їх значний обсяг. Окрім обсягу, опис великих даних також містить швидкість та різноманітність. Швидкість вказує на швидкість або частоту збору даних і робить їх доступними для подальшого аналізу; в той час, як різноманітні зауваження щодо різних типів структурованих і неструктурованих даних,

які може збирати будь-яка фірма чи система, такі як дані на рівні транзакцій, відео, аудіо, текстові або файли записів забезпечуються різноманітність. Обсяг і швидкість(Volume and velocity) є найбільш обговорюваними рисами великих даних. Те, що може сприйматися як "великий обсяг" або "висока швидкість", залежить від того, як швидко розвиваються технологій генерування, зберігання, передачі та візуалізації даних. Ці три V стали стандартним визначенням великих даних. Хоча інші люди додали кілька інших V до цього визначення.

У цьому контексті мають сенс інші характерні "v-слова", які пов'язані з великими даними, включаючи:

1. Veracity - достовірність, що розуміється як ступінь, у якому можна гарантувати якість та надійність великих даних. Дані з великим обсягом, швидкістю та різноманітністю мають значний ризик містити неточності, помилки та невраховані упередження. За відсутності відповідної валідації та перевірки якості це може призвести до оманливої або відверто неправильної бази доказів на твердження про знання;
2. Validity - дійсність, яка вказує на вибір відповідних даних щодо цільового використання. Вибір конкретного набору даних як доказової бази вимагає адекватного та чіткого обґрунтування, включаючи звернення до відповідних базових знань для обґрунтування визначення того, що вважається даними у цьому контексті;
3. Volatility - нестабільність, тобто ступінь, в якому можна покладатися на дані, щоб вони залишалися доступними та повторно інтерпретованими, незважаючи на зміни в архівних технологіях. Це важливо з огляду на тенденцію застарівання форматів та інструментів, що використовуються для генерування та аналізу

даних, а також зусилля, необхідні для оновлення інфраструктури даних, щоб гарантувати доступ до них у довгостроковій перспективі;

4. Value - цінність, тобто багатогранні форми значущості, що приписуються великим даним різними верствами суспільства, які залежать як від передбачуваного використання даних, так і від історичних, соціальних та географічних обставин. Поряд з науковою цінністю, дослідники можуть вважати дані фінансово, етично, репутаційно та навіть емоційно значущими, залежно від їх цільового призначення, а також історичних, соціальних та географічних обставин їх використання. Інститути, які беруть участь в управлінні та фінансуванні досліджень, також мають способи оцінки даних, які не завжди можуть збігатися з пріоритетами дослідників.

Рисунок 1 – Список рис, характерних для великих даних[20]

Цей список рис, хоча і не вичерпний (Рисунок 1), підкреслює, що Big Data — це не просто «багато даних». Справжня сила великих даних полягає в їх здатності створювати ланки об'єднання між різними дослідницькими спільнотами, методологічними підходами та теоретичними рамками, які важко пов'язати через концептуальну фрагментацію, соціальні бар'єри та технічні труднощі. І дійсно, звернення до великих даних часто виникає при запитах, які одночасно є технічно, концептуально та соціально складними, і коли існуючі методи та ресурси виявилися недостатніми або невідповідними.

В останні роки термін "великі дані" став надзвичайно популярним у всьому світі. Майже кожен сектор досліджень, незалежно від того, чи стосується він промисловості чи науковців, генерує та аналізує великі дані для різних цілей. Найскладнішим завданням щодо цієї величезної купи даних, які можна організувати та реорганізувати, є управління ними. Враховуючи той факт, що великими даними неможливо керувати за допомогою традиційного програмного забезпечення, нам потрібні технічно просунуті програми та програмне забезпечення, які можуть використовувати швидкі та економічно ефективні обчислювальні потужності високого класу для таких завдань. Реалізація алгоритмів штучного інтелекту (ШІ) та нових алгоритмів злиття була б необхідною, щоб отримати цінність (сутність) з такої великої кількості даних. Справді, було б чудовим подвигом досягти автоматизованого прийняття рішень шляхом впровадження методів машинного навчання (ML), таких як нейронні мережі та інші методи ШІ. Однак за відсутності відповідного програмного та апаратного забезпечення великі дані можуть бути досить туманними. Нам потрібно розробити кращі методи для роботи з цим «нескінченним морем» даних та розумних веб-додатків для ефективного аналізу, щоб отримати справжнє розуміння.

З належним зберіганням та аналітичними інструментами під рукою інформація та ідеї, отримані з великих даних, можуть зробити важливі компоненти та послуги соціальної інфраструктури (наприклад, охорону здоров’я, безпеку чи транспорт) більш обізнаними, інтерактивними та ефективними. Крім того, візуалізація великих даних у зручній для користувача формі стане критичним чинником суспільного розвитку[4].

1.3 Великі дані в сфері охорони здоров’я

Технологічний прогрес допоміг нам генерувати все більше і більше даних, навіть до такого рівня, коли це стало неможливо контролювати за допомогою наявних на даний момент технологій. Це призвело до створення терміну "великі дані" для опису даних, які є об’ємними та некерованими. Для того, щоб задовольнити наші нинішні та майбутні суспільні потреби, нам потрібно розробити нові стратегії організації цих даних та отримання значущої інформації. Однією з таких особливих суспільних потреб є охорона здоров'я. Як і будь-яка інша галузь, організації охорони здоров'я виробляють дані з величезною швидкістю, що представляє багато переваг і проблем одночасно. Медичні працівники, як і підприємці, здатні збирати величезні обсяги даних і шукати найкращі стратегії використання цих цифр.

Великі дані в охороні здоров’я – це термін, який використовується для опису величезних обсягів інформації, створених у результаті впровадження цифрових технологій, які збирають записи пацієнтів і допомагають керувати роботою лікарень, які в іншому випадку занадто великі та складні для традиційних технологій.

Застосування аналітики великих даних у сфері охорони здоров’я має багато позитивних результатів[29]. По суті, великі дані – це величезна кількість інформації, що створюється в результаті оцифрування всього, що консолідується та аналізується за допомогою конкретних технологій.

Застосовані до охорони здоров’я, вони використовуватимуть конкретні дані про здоров’я населення (або окремої особи) і потенційно допоможуть запобігти епідеміям, вилікувати хвороби, скоротити витрати тощо.

Тепер, коли ми живемо довше, моделі лікування змінилися, і багато з цих змін керуються даними. Лікарі хочуть якомога більше розуміти пацієнта, щоб помічати попереджувальні ознаки серйозного захворювання в міру їх появи – лікування будь-якої хвороби на ранній стадії набагато простіше і дешевше. Використовуючи ключові показники ефективності в охороні здоров’я та аналітиці медичних даних, попередження краще, ніж лікування, а вміння скласти вичерпну картину пацієнта дозволить страховці надати індивідуальний пакет. Це спроба індустрії впоратися з проблемами, які виникають у даних пацієнтів: всюди збираються їх уривки й архівуються в лікарнях, клініках, хірургічних установах тощо, з неможливістю належної комунікації.

Дійсно, роками збір величезної кількості даних для медичних потреб був дорогим і тривалим. Завдяки сучасним технологіям, які постійно вдосконалюються, стає легше не тільки збирати такі дані, але й створювати вичерпні звіти про медичне обслуговування та перетворювати їх у відповідні критичні ідеї, які потім можна використовувати для надання кращої допомоги. Це мета аналітики медичних даних: використання отриманих даних, щоб передбачити та вирішити проблему, поки не стало надто пізно, а також швидше оцінити методи та лікування, краще відстежувати інвентаризацію, більше залучати пацієнтів до власного здоров’я та розширювати можливості з інструментами для цього.

1.4 Висновки до розділу 1

Визначення великих даних, як обсягу інформації характеризується відповідними критеріями, які включають в себе швидкість надходження, обсяг, різноманітність, правдивість, цінність, стабільність та інші, тоді як

Big Data - як напрямок розвитку інформаційних технологій забезпечує платформи, інструменти, підходи та стратегії, що дозволяють ефективно справлятись з такими великими даними.

Область великих даних не лише активно розвивається та допомагає вирішувати проблеми, які було неможливо вирішити раніше звичайними прикладними засобами, а й розширює сфери застосування. Таким чином медичні записи, що переносяться в електронний формат, дозволяють отримати важливі результати шляхом обробки та застосування методів машинного навчання та штучного інтелекту.

Потенціал великих даних полягає в тому, що вони можуть змінити результати щодо найбільш відповідного або точного діагнозу пацієнта та точності інформації, яка використовується в системі медичної інформатики. Таким чином, дослідження величезної кількості інформації матиме потужний вплив на систему медичних послуг[11].

2 АНАЛІЗ ПРОБЛЕМИ ЕФЕКТИВНОСТІ СИСТЕМ ОБРОБКИ ВЕЛИКИХ НАБОРІВ МЕДИЧНИХ ДАНИХ

Застосування великих даних в медицині може бути не очевидним. Для того, що зрозуміти природу даних та оцінити перспективи роботи з ними, потрібно мати чітке уявлення про джерела цих даних, їх формат, спосіб збору та застосування, об’єм, доступність. Також нам необхідно точно розуміти, яку вигоду можна отримати з обробки тих чи інших даних, як їх може бути опрацьовано, щоб одержати цінність при цьому максимально збільшуючи ефективність.

2.1 Приклади застосування Big Data в медичній сфері

З метою краще зрозуміти сфери застосування великих даних в галузі медицини, було детально досліджено та проаналізовано 8 досліджень за даним напрямком. Окрім того, було проведено узагальнення та створено таблицю отриманих результатів (Таблиця 1).

Таблиця 1. Сфери застосування великих даних у медицині

Застосування	[16]	[4]	[15]	[8]	[14]	[3]	[5]	[11]
Електронні медичні картки, EMR	*	*	*	*	*	*	*	*
Раннє діагностування	*		*	*	*	*	*	*
Носимі пристрої	*	*	*	*	*	*	*	*
Біомедичні дослідження		*		*	*			*
Виявлення та запобігання шахрайству	*		*	*	*	*		*
Системи підтримки прийняття клінічних рішень	*	*	*	*	*	*	*	*
Персоналізація		*	*	*	*	*	*	*

Відкриття лікарських засобів та клінічні випробування			*	*	*	*		
Іміджева інформатика та теледіагностика	*		*					
Система медичних знань	*		*	*	*		*	*
Громадське здоров'я	*		*	*	*		*	*
Моніторинг клінічних показників	*		*	*	*	*	*	*

2.1.1 Електронні медичні картки (EHRs)

Це найбільш поширене застосування великих даних у медицині. Кожен пацієнт має власну цифрову карту. Записи передаються через захищені інформаційні системи та доступні для постачальників як із державного, так і з приватного секторів. Кожен запис складається з одного файлу, який можна змінювати, а це означає, що лікарі можуть вносити зміни з часом без паперової роботи і небезпеки тиражування даних.

EHR надають багато переваг для роботи з сучасними даними, пов’язаними з охороною здоров’я. Першою перевагою EHR є те, що медичні працівники мають покращений доступ до всієї історії хвороби пацієнта. Інформація включає медичні діагнози, рецепти, дані про відомі алергії, демографічні дані, клінічні описи та результати різних лабораторних досліджень. Таким чином, розпізнавання та лікування захворювань є ефективним за часом завдяки скороченню часу затримки попередніх результатів тесту. З часом ми помітили значне зменшення кількості зайвих та додаткових обстежень, втрачених замовлень і неясностей, спричинених нерозбірливим почерком, а також покращення координації надання медичної допомоги між кількома медичними працівниками. Подолання таких логістичних помилок призвело до зменшення кількості лікарських алергій за рахунок зменшення похибок у дозі та частоті прийому ліків. EHR також можуть ініціювати попередження

та нагадування, коли пацієнт повинен пройти новий лабораторний тест або відстежувати рецепти, щоб перевірити, чи виконує пацієнт вказівки лікарів.

Медичні працівники також знайшли доступ через веб та електронні платформи, щоб значно покращити свою медичну практику, використовуючи автоматичні нагадування та підказки щодо щеплень, аномальних лабораторних результатів, скринінгу на рак та інших періодичних оглядів. Буде більша безперервність надання медичної допомоги та своєчасних втручань за рахунок полегшення спілкування між кількома постачальниками медичних послуг і пацієнтами. Вони можуть бути пов’язані з електронною авторизацією та негайним схваленням страхування через меншу кількість документів. EHR забезпечують швидший пошук даних і полегшують звітування організацій щодо ключових показників якості медичної допомоги, а також покращують нагляд за станом здоров’я шляхом негайного повідомлення про спалахи захворювань. EHR також надають відповідні дані щодо якості догляду для бенефіціарів програм медичного страхування працівників і можуть допомогти контролювати зростання витрат на виплати медичного страхування. Нарешті, EHR можуть зменшити або повністю усунути затримки та плутанину у сфері виставлення рахунків та управління претензіями. EHR та Інтернет разом допомагають забезпечити доступ до мільйонів медичної інформації, пов’язаної зі здоров’ям, важливої для життя пацієнтів[4].

Подібно до EHR, електронна медична карта (EMR) зберігає стандартні медичні та клінічні дані, зібрані від пацієнтів. EHR, EMR, особиста медична картка (PHR), програмне забезпечення для управління медичною практикою (MPM) та багато інших компонентів медичних даних разом мають потенціал для покращення якості, ефективності обслуговування та витрат на медичне обслуговування разом із зменшенням

медичних помилок. Великі дані в галузі охорони здоров'я включають дані платників і постачальників медичних послуг (наприклад, EMR, аптечні рецепти та страхові записи), а також експерименти, керовані геномом (наприклад, генотипування, дані про експресію генів) та інші дані, отримані з розумної мережі Інтернету речей (IoT). Хоча EHR є чудовою ідеєю, багато країн все ще намагаються втілити її повністю. Впровадження EHR було повільним на початку 21 століття, проте після 2009 року воно значно зросло. Згідно з дослідженням HITECH, США зробили великий ривок: 94% лікарень впровадили EHR.

Кайзер Перманенте повністю впровадили систему під назвою HealthConnect, яка обмінюється даними між усіма їхніми закладами та полегшує використання EHR. У звіті McKinsey про охорону здоров'я з великими даними стверджується, що «інтегрована система покращила результати при серцево-судинних захворюваннях і заощадила приблизно 1 мільярд доларів за рахунок скорочення відвідувань лікаря та лабораторних тестів»[16].

Управління та використання таких даних охорони здоров'я все більше залежало від інформаційних технологій. Розробка та використання пристроїв для моніторингу здоров'я та відповідного програмного забезпечення, які можуть генерувати сповіщення та обмінюватися даними, пов'язаними зі здоров'ям пацієнта, з відповідними постачальниками медичних послуг, набрали обертів, особливо у створенні системи біомедичного моніторингу та моніторингу стану здоров'я в реальному часі. Ці пристрої генерують величезну кількість даних, які можна аналізувати для надання клінічної або медичної допомоги в режимі реального часу. Використання великих даних із охорони здоров'я обіцяє покращити результати здоров'я та контролювати витрати.

2.1.2 Надання персоналізованого медичного обслуговування

Збільшення використання технологій повільно змінює напрямок сектору охорони здоров'я від допомоги, орієнтованої на захворювання, до допомоги, орієнтованої на пацієнта. Значну роль у цій трансформації відіграватимуть великі дані. Це дозволить надавати інформацію безпосередньо пацієнтам і дасть їм можливість відігравати активну участь у догляді за ними. Коли пацієнтам надається відповідна інформація, це вплине на прийняття ними рішень і дозволить приймати обґрунтовані рішення. На обґрунтовані рішення також впливатиме посилення комунікації між пацієнтами, провайдерами, а також їхніми спільнотами[5].

Таким чином великі дані відіграють важливу роль у сфері охорони здоров'я для розробки персоналізованої рекомендаційної системи для надання точних та відповідних медичних рекомендацій (порад) особі (пацієнту) на основі їх поточного стану здоров'я та історії хвороби. Також запропонували систему рекомендацій щодо здоров'я на основі дослідження, яке використовує аналітику великих даних для вивчення та дослідження медичних карт пацієнтів, оцінки ризику та тяжкості різних захворювань, а потім надання рекомендацій на основі результатів прогнозування. Окрім того, запропонували систему клінічних рекомендацій, яка є вигідною для пацієнтів, щоб отримати точні рекомендації на основі їх власного стану здоров'я.

Великі дані в охороні здоров'я можуть змінити медичну сферу шляхом виявлення захворювань на ранній стадії та знизити медичні витрати для пацієнтів, використовуючи відповідні аналітичні інструменти комплексним чином. Це допомагає розробити персоналізовану систему охорони здоров'я для зацікавлених сторін у сфері охорони здоров'я[15].

З використанням великих даних цілі персоналізованої медицини можна втілити в клінічну практику. Доступ до великих обсягів даних та їх

обробка мають уможливити персоналізовану реєстрацію ризиків захворювання для конкретного пацієнта. Додатки для великих даних мають на меті зробити цей процес більш ефективним.

2.1.3 Раннє виявлення захворювань

Великі дані забезпечують раннє виявлення захворювань, що допомагає досягти клінічних цілей, пов'язаних із досягненням покращеного лікування та вищих результатів для пацієнтів. Виявлення захворювань на ранніх стадіях дозволяє лікувати легше та ефективніше Саме в цій області знайшли великі перспективи в лікуванні вікових хвороб і захворювань. Поряд із раннім виявленням, аналітика великих даних також може допомогти у запобіганні широкого спектру смертельних захворювань та персоналізованому управлінні та моніторингу захворювань. Це дозволяє постачальникам відстежувати здорову поведінку та допомагає пацієнтам контролювати їхні відповідні стани. Ця здатність має великий потенціал, якщо стикатися або з віковими захворюваннями, або з проблемами здоров'я у всьому світі, такими як кардіологія[5].

Ми вже визнали прогнозну аналітику як одну з найбільших тенденцій бізнес-аналітики, але потенційні програми виходять далеко за межі бізнесу та набагато далі в майбутньому. Optum Labs, американська дослідницька організація, зібрала EHR понад 30 мільйонів пацієнтів, щоб створити базу даних для інструментів прогнозної аналітики, які покращать надання допомоги.

Мета бізнес-аналітики в галузі охорони здоров'я — допомогти лікарям приймати рішення на основі даних протягом секунд і покращити лікування пацієнтів. Це особливо корисно у випадку пацієнтів зі складною історією хвороби, які страждають від багатьох захворювань. Нові рішення та інструменти BI також зможуть передбачити, наприклад, хто знаходиться в групі ризику захворіти на цукровий діабет, і, таким чином, буде

рекомендовано використовувати додаткові скринінги або контроль ваги[16].

2.1.4 Носимі пристрої та інтернет речей

Одним із таких джерел клінічних даних у сфері охорони здоров'я є «Інтернет речей» (IoT). Насправді, IoT став зростаючим рухом у сфері охорони здоров'я. Пристрої IoT створюють безперервний потік даних під час моніторингу здоров'я людей (або пацієнтів), що робить ці пристрої основним внеском у великі дані в охороні здоров'я. Такі ресурси можуть взаємодіяти між різними пристроями для надання надійної, ефективної та розумної медичної допомоги літнім людям та пацієнтам із хронічними захворюваннями.

Використовуючи мережу IoT-пристроїв, лікар може вимірювати та контролювати різні параметри своїх клієнтів у відповідних місцях, наприклад, вдома чи в офісі. Таким чином, завдяки ранньому втручанню та лікуванню пацієнт може не потребувати госпіталізації чи навіть відвідування лікаря, що призведе до значного зниження витрат на охорону здоров'я. Деякі приклади пристроїв Інтернету речей, які використовуються в охороні здоров'я, включають носимі пристрої для фітнесу або відстеження здоров'я, біосенсори, клінічні пристрої для моніторингу життєво важливих показників та інші типи пристроїв або клінічних інструментів. Такі пристрої Інтернету речей генерують велику кількість даних, пов'язаних зі здоров'ям. Якщо ми зможемо інтегрувати ці дані з іншими наявними даними охорони здоров'я, такими як EMR або PHR, ми зможемо передбачити стан здоров'я пацієнта та його прогресування від субклінічного стану до патологічного.

Насправді великі дані, створені за допомогою Інтернету речей, були використовувані в кількох областях, пропонуючи кращі дослідження та прогнозування. У більшому масштабі дані з таких пристроїв можуть

допомогти в моніторингу здоров’я персоналу, моделюванні поширення захворювання та пошуку способів стримування конкретного спалаху захворювання.

Аналіз даних з Інтернету речей вимагатиме оновленого операційного програмного забезпечення через його специфічну природу разом із передовими апаратними та програмними додатками. Нам потрібно буде керувати надходженням даних з інструментів IoT в режимі реального часу та аналізувати їх щохвилини. Співробітники системи охорони здоров’я намагаються скоротити витрати та покращити якість медичної допомоги, застосовуючи передову аналітику як до даних, що генеруються всередині, так і ззовні.

У сучасному цифровому світі кожна людина, здається, одержима відстежувати статистику своєї фізичної активності та здоров’я за допомогою вбудованого крокоміра своїх портативних і носимих пристроїв, таких як смартфони, розумні годинники, фітнес-панелі або планшети. У зв’язку з дедалі більш мобільним суспільством майже в усіх аспектах життя інфраструктура охорони здоров’я потребує реконструкції для розміщення мобільних пристроїв.

Практика медицини та охорони здоров’я за допомогою мобільних пристроїв, відомих як mHealth або мобільне здоров’я, пронизує різні ступені медичної допомоги, особливо для хронічних захворювань, таких як діабет та рак. Організації охорони здоров’я все частіше використовують мобільні послуги здоров’я та оздоровлення для впровадження нових та інноваційних способів надання допомоги та координації здоров’я. Мобільні платформи можуть покращити охорону здоров’я, прискорюючи інтерактивне спілкування між пацієнтами та постачальниками медичних послуг.

Фактично, Apple і Google розробили спеціальні платформи, такі як Apple ResearchKit і Google Fit, для розробки дослідницьких програм для

статистики фітнесу та здоров’я. Ці програми підтримують безперебійну взаємодію з різними споживчими пристроями та вбудованими датчиками для інтеграції даних. Вони допомагають лікарям мати прямий доступ до ваших загальних даних про стан здоров’я. І користувач, і його лікарі дізнаються про стан вашого тіла в режимі реального часу. Ці програми та розумні пристрої також допомагають покращувати наше планування здоров’я та заохочуючи до здорового способу життя. Таким чином користувачі або пацієнти можуть стати захисниками власного здоров'я[4].

Крім того носимі пристрої мають одну важливу функціональність – оповіщення в режимі реального часу. У лікарнях програмне забезпечення Clinical Decision Support (CDS) аналізує медичні дані на місці, надаючи медичним працівникам пораду, коли вони ухвалюють рішення.

Однак лікарі хочуть, щоб пацієнти утримувались від стаціонарного лікування, щоб знизити його вартість. Аналітика може стати частиною нової стратегії. Носимі пристрої постійно збиратимуть дані про здоров’я пацієнтів та надсилатиме їх у хмару.

Крім того, ця інформація буде доступна до бази даних про стан здоров’я населення, що дозволить лікарям порівнювати ці дані в соціально-економічному контексті та відповідно змінювати стратегії доставки. Установи та менеджери по догляду будуть використовувати складні інструменти для моніторингу цього масивного потоку даних і реагувати щоразу, коли результати будуть тривожними.

Наприклад, якщо артеріальний тиск у пацієнта тривожно підвищується, система в режимі реального часу надішле попередження лікарю, який потім вживе заходів, щоб зв’язатися з пацієнтом і вжити заходів для зниження тиску.

Іншим прикладом є компанія Asthmapolis, яка почала використовувати інгалятори з GPS-трекерами, щоб ідентифікувати тенденції астми як на індивідуальному рівні, так і на більшій групі

населення. Ці дані використовуються разом з даними CDC для розробки кращих планів лікування астматиків[16].

2.1.5 Біомедичні дослідження

Біологічна система, така як клітина людини, демонструє молекулярні та фізичні події складної взаємодії. Щоб зрозуміти взаємозалежність різних компонентів і подій такої складної системи, біомедичний або біологічний експеримент зазвичай збирає дані про менший і/або простіший компонент. Отже, для створення широкої карти даного біологічного явища, що цікавить, потрібно провести кілька спрощених експериментів. Це свідчить про те, що більше даних ми маємо, тим краще ми розуміємо біологічні процеси. Завдяки цій ідеї сучасні техніки розвивалися великими темпами. Наприклад, можна уявити, скільки даних було згенеровано після інтеграції ефективних технологій, таких як секвенування наступного покоління (NGS) і дослідження асоціації генома (GWAS) для декодування генетики людини.

Дані на основі NGS надають інформацію на глибинах, які раніше були недоступні, і переносять експериментальний сценарій в абсолютно новий вимір. Це підвищило роздільну здатність, з якою ми спостерігаємо або реєструємо біологічні події, пов’язані з конкретними захворюваннями, у режимі реального часу. Ідея про те, що великі обсяги даних можуть надати нам значну кількість інформації, яка часто залишається неідентифікованою або прихованою в менших експериментальних методах, започаткувала еру «-omics». Дисципліна «геноміка» стала свідком значного прогресу, оскільки замість вивчення одного «гена» вчені тепер можуть вивчати весь «геном» організму в дослідженнях «геноміки» протягом певного проміжку часу. Аналогічно, замість того, щоб вивчати експресію чи «транскрипцію» окремого гена, тепер ми можемо вивчати

експресію всіх генів або весь «транскриптом» організму під час досліджень «транскриптоміки».

Кожен із цих індивідуальних експериментів генерує велику кількість даних із більш глибокою інформацією, ніж будь-коли раніше. Проте цієї глибини та роздільної здатності може бути недостатньо, щоб надати всі деталі, необхідні для пояснення конкретного механізму чи події. Тому зазвичай виявляється, що аналізує велику кількість даних, отриманих в результаті кількох експериментів, щоб отримати нові ідеї. Цей факт підтверджується безперервним зростанням кількості публікацій щодо великих даних у сфері охорони здоров'я. Аналіз таких великих даних із медичних систем та систем охорони здоров'я може бути надзвичайно корисним у створенні нових стратегій охорони здоров'я. Останні технологічні розробки в області генерації, збору та аналізу даних можуть викликати у найближчому майбутньому революцію в галузі персоналізованої медицини[10].

NGS значно спростив секвенування та зменшив витрати на генерацію даних про послідовність всього геному. Вартість повного секвенування геному впала з мільйонів до кількох тисяч доларів. Технологія NGS призвела до збільшення обсягу біомедичних даних, отриманих з геномних і транскриптомічних досліджень. За оцінками, кількість людських геномів, секвенованих до 2025 року, може становити від 100 мільйонів до 2 мільярдів.

Це може значно розширити наші знання про індивідуальний профіль пацієнта — підхід, який часто називають «індивідуальним, персоналізованим або точним медичним обслуговуванням». Систематичний та інтегративний аналіз даних omics у поєднанні з аналітикою охорони здоров'я може допомогти розробити кращі стратегії лікування, спрямовані на точну та персоналізовану медицину. Експерименти, керовані геномою, наприклад, генотипування, експресія

генів і дослідження на основі NGS, є основним джерелом великих даних у біомедичній охороні здоров'я разом з EMR, інформацією про рецепти в аптеках та страховими записами. Охорона здоров'я вимагає міцної інтеграції таких біомедичних даних з різних джерел, щоб забезпечити краще лікування та догляд за пацієнтами. Ці перспективи настільки захоплюючі, що, незважаючи на те, що геномні дані пацієнтів мають багато змінних, які потрібно врахувати, комерційні організації вже використовують дані геному людини, щоб допомогти постачальникам приймати персоналізовані медичні рішення. Це може змінити гру в майбутній медицині та здоров'ї [4].

2.1.6 Виявлення та запобігання шахрайству

Одним із основних і важливих застосувань аналізу даних у сфері охорони здоров'я є виявлення та запобігання шахрайству. Методи штучного інтелекту даних і машинного навчання в основному використовуються для виявлення шахрайства в охороні здоров'я[15].

Шахрайство може статися у разі відвідувань лікарні та запитів від різних служб охорони здоров'я. Більше того, медичні вердикти, аналізи та їх результати, медичне лікування та інші предмети можуть бути проаналізовані і таким чином вказати на можливе зловживання ліками.

На початку аналітики великих даних медичні страхові групи використовують кілька шляхів для виявлення шахрайства та встановлення методів запобігання медичним шахрайствам. Компанії використовують програми, засновані на моделі прогнозування, щоб ідентифікувати тих, хто вчиняє шахрайство, за допомогою даних щодо їхніх попередніх заяв про здоров'я, голосових записів, заробітної плати та демографічних даних. Також можливе запобігання шахрайству, пов'язаному з медичними заявами на ранній стадії, за допомогою використання медичних програм в режимі реального часу, справжніх рахунків за медичні претензії, даних прогнозу

погоди, записів голосових даних та інших джерел даних[11]. Аналітика допомагає спростити обробку страхових відшкодувань, дозволяючи пацієнтам отримувати кращу віддачу від своїх претензій, а особам, які доглядають, отримують швидше. Наприклад, Центри послуг Medicare і Medicaid заявили, що лише за рік вони заощадили понад 210,7 мільйонів доларів на шахрайстві[16].

2.1.7 Системи підтримки прийняття клінічних рішень

Великі дані можна використовувати для підтримки прийняття клінічних рішень. Великі дані дають змогу належним чином використовувати доказову медицину та допомагають постачальникам медичних послуг приймати більш обґрунтовані рішення. Це, у свою чергу, покращує якість надання медичної допомоги пацієнтам. Віддалений моніторинг, аналітика профілю пацієнта та геномна аналітика є прикладами інших застосувань, які впливають на процес прийняття рішень. Процес прийняття рішень можна значно оптимізувати завдяки наявності точної та актуальної інформації, оскільки на прийняття рішень впливає створення нових методів та рекомендацій щодо лікування в рамках клінічних досліджень. Дозволяючи великим даним впливати на прийняття рішень, процес буде швидшим і простішим. Це робиться шляхом підтримки або заміни прийняття людських рішень[5]. Різні компоненти дозволяють лікарям отримувати додаткову інформацію для допомоги в процесі прийняття рішень щодо діагностики захворювань на основі стану здоров’я пацієнта[8].

2.1.8 Відкриття лікарських засобів та клінічні випробування

Аналітика великих даних в галузі охорони здоров’я широко використовується фармацевтичною промисловістю для відкриття ліків, щоб допомогти лікарям, розробникам фармацевтичних засобів та іншим

медичним працівникам отримати потрібний препарат потрібному пацієнту в потрібний час[15].

Також це включає прогнозне моделювання, щоб зменшити виснаження та створити більш економний, швидший, більш цілеспрямований конвеєр досліджень і розробок у ліках та приладах; статистичні інструменти та алгоритми для покращення дизайну клінічних випробувань та рекрутингу пацієнтів, щоб краще відповідати лікування окремим пацієнтам, таким чином зменшуючи кількість невдач у випробуваннях та прискорюючи вихід нових методів лікування на ринок і аналіз клінічних випробувань і записів пацієнтів для визначення подальших показань і виявлення побічних ефектів до того, як продукти потраплять на ринок[14].

2.1.9 Іміджева інформатика та теледіагностика

Інформатика зображень — це вивчення методів створення, керування та представлення інформації зображення в різних біомедичних програмах. Це стосується того, як обмінюються та аналізуються медичні зображення в складних системах охорони здоров'я. Деякі дослідження впроваджують нову телемамографічну систему для раннього виявлення раку молочної залози за допомогою методів обробки зображень і машинного навчання. Комп'ютерна діагностика відіграє значну роль у медичній візуалізації[15].

Медична візуалізація є життєво важливою, і щороку в США виконується близько 600 мільйонів процедур візуалізації. Аналіз і зберігання цих зображень вручну є дорогим як з точки зору часу, так і грошей, оскільки рентгенологам потрібно досліджувати кожне зображення окремо, а лікарням потрібно зберігати їх кілька років.

Постачальник медичних зображень Carestream пояснює, як аналіз великих даних для охорони здоров'я може змінити спосіб зчитування зображень: розроблені алгоритми для аналізу сотень тисяч зображень

можуть ідентифікувати конкретні закономірності в пікселях і перетворити їх у число, щоб допомогти лікарю поставити діагноз. Вони навіть йдуть далі, кажучи, що можливо, що радіологам більше не потрібно буде дивитися на зображення, а замість цього аналізуватиме результати алгоритмів, які неминуче вивчатимуть і запам'ятовуватимуть більше зображень, ніж вони могли б за все життя. Це, безсумнівно, вплине на роль радіологів, їхню освіту та необхідну кваліфікацію[10].

Телемедицина присутня на ринку понад 40 років, але тільки сьогодні, з появою онлайн-відеоконференцій, смартфонів, бездротових пристроїв і носіїв, вона змогла розквітнути. Термін відноситься до надання дистанційних клінічних послуг з використанням технології.

Він використовується для первинних консультацій та первинної діагностики, дистанційного моніторингу пацієнтів та медичної освіти для медичних працівників. Деякі більш конкретні застосування включають телехірургію – лікарі можуть виконувати операції з використанням роботів і високошвидкісної доставки даних у реальному часі, не перебуваючи фізично в одному місці з пацієнтом.

Клініцисти використовують телемедицину, щоб надати персоналізовані плани лікування та запобігти госпіталізації чи повторному госпіталізації. Таке використання аналітики медичних даних може бути пов'язане з використанням прогнозної аналітики, як було показано раніше. Це дозволяє клініцистам заздалегідь передбачити гострі медичні події та запобігти погіршенню стану пацієнта.

Утримуючи пацієнтів від відвідування лікарень, телемедицина допомагає знизити витрати та покращити якість обслуговування. Пацієнти можуть уникнути очікування в чергах, а лікарі не витрачають час на непотрібні консультації та оформлення документів. Телемедицина також покращує доступність медичної допомоги, оскільки стан пацієнтів можна відстежувати та консультуватися в будь-якому місці та в будь-який час[16].

2.1.10 Система медичних знань та доказова медицина

Система управління знаннями розроблена на основі великих даних охорони здоров'я для підтримки прийняття клінічних рішень і діагностики захворювань. Система знань у сфері охорони здоров'я об'єднує різноманітні бази даних, такі як електронна медична карта (EHR), дані медичних зображень, а також неструктуровані клінічні нотатки та генетичні дані, щоб узгодити лікування з результатами, прогнозувати пацієнтів із ризиком захворювання чи повторної госпіталізації та надавати більш ефективну допомогу[15].

Література припускає, що великі дані дозволяють швидко отримувати дані та перетворювати первинні, необроблені та неструктуровані дані у значущу інформацію. Нові знання можуть бути отримані з великих обсягів ефективних даних, що дозволить повторно використовувати дані. Технологія з відкритим кодом підвищує доступність і прозорість даних. Нарешті, якість даних можна підтримувати за допомогою аналітики, щоб позбутися непотрібної інформації.

Великі дані активно допомагатимуть у поширенні знань, отриманих із зібраних даних та відіграють активну роль у використанні практики і знань не лише на регіональному, а й на глобальному рівні. Завдяки глобалізації, дані стають більш доступними, а постачальники можуть отримати доступ до нової інформації з усіх регіонів[5].

Цікавим прикладом використання великих даних в охороні здоров'я є програма Cancer Moonshot. Перед закінченням свого другого терміну президент Обама розробив цю програму, яка мала на меті досягти 10-річного прогресу в лікуванні раку за половину цього часу.

Медичні дослідники можуть використовувати великі обсяги даних про плани лікування та показники одужання хворих на рак, щоб знайти тенденції та методи лікування, які мають найвищі показники успіху в

реальному світі. Наприклад, дослідники можуть досліджувати зразки пухлин у біобанках, які пов’язані з записами лікування пацієнтів. Використовуючи ці дані, дослідники можуть побачити, як певні мутації та білки раку взаємодіють з різними методами лікування, і знайти тенденції, які призведуть до кращих результатів для пацієнтів.

Ці дані також можуть призвести до несподіваних переваг, наприклад, виявити, що дезипрамін, який є антидепресантом, має здатність допомогти вилікувати певні типи раку легенів.

Однак для того, щоб зробити таку інформацію більш доступною, необхідно об’єднати бази даних пацієнтів із різних установ, таких як лікарні, університети та некомерційні організації. Тоді, наприклад, дослідники могли отримати доступ до звітів біопсії пацієнтів з інших установ. Одним із потенційних випадків використання великих даних у сфері охорони здоров’я було б генетичне секвенування зразків ракової тканини від пацієнтів із клінічних випробувань та надання цих даних у ширшу базу даних раку.

Нарешті, рішення лікарів стають все більше і більше заснованими на доказах, а це означає, що вони покладаються на велику кількість досліджень і клінічних даних, а не лише на їхню академічну та професійну думку. Як і в багатьох інших галузях, збір даних і керування ними стають все більшими, і професіоналам потрібна допомога у цьому питанні. Це нове ставлення до лікування означає, що в медичних установах існує більший попит на аналітику великих даних, ніж будь-коли раніше, і зростання інструментів SaaS BI також відповідає на цю потребу[16].

2.1.11 Інформація про громадське здоров'я

Аналітику великих даних в охороні здоров’я також можна використовувати для відстеження та моніторингу стану громадського здоров’я для прийняття рішень та розробки політики, оцінки ризиків, а

також аналізу тенденції захворювань для покращення нагляду за суспільним здоров’ям[15].

Тема управління здоров’ям населення зосереджена на особливих групах населення, а не на громадському здоров’ї. Аналітика великих даних визначає популяції на більш тонкому рівні деталізації, ніж це було досягнуто раніше. Це може допомогти в управлінні загальним здоров’ям населення, а також конкретним індивідуальним здоров’ям. Великі дані можуть забезпечити управління здоров’ям населення з локальної або глобальної точки зору. Ця можливість стає більш помітною з глобальної точки зору, якщо врахувати старіння населення та проблеми, пов’язані зі здоров’ям, пов’язані з багатьма групами населення та підгрупами, багато з яких недостатньо обслуговуються[5].

Аналітика великих даних може також забезпечити аналіз моделей захворювань та відстеження спалахів і передачі захворювань для покращення нагляду за станом здоров’я та швидкого реагування; швидшу розробку точніших вакцин, наприклад, вибір щорічних штамів грипу; і перетворення великих обсягів даних на інформацію, яка може бути використана для визначення потреб, надання послуг, а також прогнозування та запобігання криз, особливо на користь населення[14].

2.1.12 Моніторинг клінічних показників

Існує великий ентузіазм щодо оцінки клінічної ефективності, щоб перевірити та підвищити якість медичних послуг. Реформа лікарні є головною проблемою в стратегічному плані галузі охорони здоров’я. Цього можна досягти шляхом моніторингу та встановлення лікарні відповідно до стандартів медичної ради[15]. Інформація про охорону здоров'я допомагає визначити методи діагностики та лікування пацієнтів, які є більш клінічно важливими та економічно ефективними[14].

Розглянемо одну класичну проблему, з якою стикається будь-який керівник лікарні: скільки людей персоналу необхідно в певний період часу? Якщо ви наймете занадто багато працівників, ви ризикуєте отримати непотрібні витрати на оплату праці. Занадто мало працівників, ви можете мати погані результати обслуговування клієнтів, що може бути фатальним для пацієнтів у цій галузі.

Великі дані допомагають вирішити цю проблему, принаймні в кількох лікарнях Парижа. У дослідженні Intel описано, як чотири лікарні, які входять до Assistance Publique-Hôpitaux de Paris, використовують дані з різних джерел для щоденного та щогодинного прогнозування кількості пацієнтів, які очікується в кожній лікарні.

Одним з ключових наборів даних є записи про госпіталізацію за 10 років, які науковці аналізували за допомогою методів «аналізу часових рядів». Цей аналіз дозволив дослідникам побачити відповідні закономірності у ставках прийому. Потім вони можуть використовувати машинне навчання, щоб знайти найточніші алгоритми, які передбачають майбутні тенденції вступу.

Підсумовуючи результат усієї цієї роботи, команда наукових досліджень розробила вебінтерфейс користувача, який прогнозує завантаженість лікарні й допомагає планувати розподіл ресурсів за допомогою візуалізації даних в режимі онлайн, що забезпечує покращення загального догляду за пацієнтами[16].

2.2 Переваги використання аналітики великих даних в сфері охорони здоров’я

Незважаючи на проблеми, які необхідно подолати з великими даними, передова аналітика, яку обіцяють великі дані, пропонує величезні можливості для більшості зацікавлених сторін у галузі охорони здоров’я

(пацієнтів, постачальників і платників). Єдиною метою великих даних є покращення медичних послуг, даючи лікарям можливість робити точні прогнози щодо стану пацієнта на основі його добре записаної попередньої історії хвороби та алгоритмів, які аналізують його спосіб життя. Маємо потенціал значних переваг шляхом оцифрування, інтеграції та ефективного використання інструментів і методів аналізу великих даних у сфері охорони здоров'я. Якщо навіть деякі з можливостей будуть реалізовані, вони можуть радикально змінити результати пацієнтів і спосіб прийняття рішень постачальниками послуг, а також допомогти вирішити деякі проблеми на макрорівні, пов'язані з охороною здоров'я, а саме:

2.2.1 Для пацієнтів

Інформація про медичне обслуговування може допомогти пацієнтам прийняти правильне рішення в потрібний час і покращити здоров'я пацієнтів.

Підвищення якості догляду - великі дані мають потенціал і здатність покращувати якість та ефективність допомоги, дають можливість передбачати результати, використовуючи доступні первинні або історичні дані, і надавати докази переваг, які можуть змінити встановлені галузеві стандарти надання допомоги. Використання технології на стороні пацієнта також може допомогти у дотриманні ліків. Це, безсумнівно, відіграє важливу роль у покращенні результатів та покращення якості життя, пов'язаної зі здоров'ям. Якість медичної допомоги також буде покращено за рахунок зменшення втрати інформації, що зменшить неефективність. Це також допоможе в аналізі продуктивності використання ресурсів у реальному часі.

Зниження витрат - виявлено потенціал для покращення якості медичної допомоги, знижуючи медичні витрати пацієнтів шляхом вивчення асоціації та розуміння природи медичних даних.

2.2.2 Постачальники медичних послуг

Дані, отримані від медичних організацій, допомагають зацікавленим сторонам розробити нові стратегії охорони здоров’я для пацієнтів, щоб звести до мінімуму непотрібні госпіталізації

Зменшення витрат і часу очікування. Таким чином, використання великих даних полегшує підвищення продуктивності в галузі охорони здоров’я за рахунок розумного використання наявних людських і фінансових ресурсів, а також створення кращих прогнозів на основі прогнозування кількості пацієнтів. Літературні дані свідчать про те, що зниження вартості таких елементів обчислювальної техніки, як зберігання та обробка, призводить до зниження вартості завдань із інтенсивним використанням даних. Таке перенесення заощаджень буде спостерігатися у всьому спектрі медицини та персоналу охорони здоров’я. Економія буде досягнута за рахунок більш економічно ефективних методів лікування та моніторингу для покращення прихильності до лікування. Вважається, що впровадження аналітики великих даних організаціями охорони здоров’я може призвести до економії понад 25% щорічних витрат у найближчі роки. За оцінками McKinsey, аналітика великих даних може дозволити заощадити понад 300 мільярдів доларів на рік на охороні здоров’я США.

Можливість покращити якість, структуру та доступність даних.

2.2.3 Дослідження та розробки

Дані про охорону здоров’я допомагають дослідникам і вченим покращувати медичні послуги за допомогою більш точних і відповідних методів лікування.

2.3 Перешкоди на шляху до широкого поширення великих даних в системи охорони здоров'я

Індустрія охорони здоров’я повинна використовувати весь потенціал багатих потоків інформації для покращення досвіду пацієнтів. У секторі охорони здоров’я це могло б матеріалізуватися з точки зору кращого управління, догляду та недорогого лікування. Ми за багато кілометрів від того, щоб усвідомити переваги великих даних у значущий спосіб і використати знання, які випливають із них. Щоб досягти цих цілей, нам потрібно систематично керувати великими даними та аналізувати їх. 4

Аналітика великих даних у режимі реального часу є ключовою вимогою в охороні здоров’я. Необхідно усунути затримку між збором та обробкою даних. Для широкомасштабного впровадження також необхідна динамічна доступність численних аналітичних алгоритмів, моделей і методів у спадному меню. Необхідно розглянути важливі питання власності, управління та стандартів. І через ці проблеми пов’язані безперервний збір і очищення даних. Дані охорони здоров’я рідко стандартизуються, часто фрагментуються або генеруються в застарілих IT-системах з несумісними форматами. Цей великий виклик також потребує вирішення[14].

2.3.1 Зберігання та передача

Зберігання великого об’єму даних є однією з основних проблем, але багатьом організаціям зручно зберігати дані у власних приміщеннях. Він має кілька переваг, як-от контроль безпеки, доступу та часу роботи. Однак мережа серверів на місці може бути дорогою в масштабуванні та важкою в обслуговуванні. Схоже, що зі зниженням витрат і підвищенням надійності хмарне сховище з використанням IT-інфраструктури є кращим варіантом, який обрали більшість медичних організацій. Організації повинні вибирати хмарних партнерів, які розуміють важливість відповідності

медичним стандартам і питанням безпеки. Крім того, хмарне сховище пропонує менші початкові витрати, швидке аварійне відновлення та легше розширення. Організації також можуть мати гібридний підхід до своїх програм зберігання даних, що може бути найбільш гнучким і працездатним підходом для постачальників із різними потребами в доступі та зберіганні даних.

Генерація даних є недорогою в порівнянні зі зберіганням і передачею даних. Після створення даних витрати, пов’язані з їх захистом і зберіганням, залишаються високими. Витрати також несуть передачу даних з одного місця в інше, а також їх аналіз. Деякі дослідники змогли поєднати теми структури даних, зберігання та передачі, коли вони ілюструють, як структуровані дані можна легко зберігати, запитувати, аналізувати тощо, але неструктурованими даними не так легко маніпулювати. Хмарна інформаційна технологія охорони здоров’я має додатковий рівень безпеки, пов’язаний із вилученням, перетворенням та завантаженням даних, пов’язаних із пацієнтами. Використання великих даних має вирішувати питання, пов’язані зі збільшенням витрат, а також з передачею безпечної або незахищеної інформації.

2.3.2 Обробка

Величезні обсяги різноманітних даних, які генеруються високою швидкістю, зібрані з різних джерел, в основному потрібні для оптимізації споживчих послуг, а не самого споживання. Основна проблема з великими даними полягає в тому, як обробляти цей великий обсяг інформації. Щоб зробити їх доступними для наукового співтовариства, дані повинні зберігатися у форматі файлу, який легко доступний і читабельний для ефективного аналізу.

У контексті даних про охорону здоров’я ще однією серйозною проблемою є впровадження високоякісних обчислювальних інструментів, протоколів і високоякісного обладнання в клінічних умовах. Щоб досягти

цієї мети, необхідно працювати разом експертам з різними знаннями, включаючи біологію, інформаційні технології, статистику та математику. Дані, зібрані за допомогою датчиків, можуть бути доступні в хмарі зберігання з попередньо встановленими програмними засобами, розробленими розробниками аналітичних інструментів. Ці інструменти будуть мати функції аналізу даних і машинного навчання, розроблені експертами з ШІ для перетворення інформації, що зберігається як дані, у знання. Після впровадження це підвищить ефективність отримання, зберігання, аналізу та візуалізації великих даних із охорони здоров'я.

Головне завдання — анотувати, інтегрувати та представити ці складні дані у відповідний спосіб для кращого розуміння. За відсутності такої відповідної інформації дані (медичної допомоги) залишаються досить туманними і можуть не вести біомедичних дослідників далі. Нарешті, інструменти візуалізації, розроблені дизайнерами комп'ютерної графіки, можуть ефективно відобразити ці нещодавно отримані знання.

2.3.3 Очищення

Дані потрібно очистити, щоб забезпечити точність, правильність, послідовність, релевантність та чистоту після отримання. Цей процес очищення може бути ручним або автоматизованим за допомогою логічних правил, щоб забезпечити високий рівень точності та цілісності. Більш складні та точні інструменти використовують методи машинного навчання, щоб скоротити час і витрати, а також запобігти зриву непотрібних даних з рейок проектів великих даних.

2.3.4 Структура даних

Одна з основних причин складності інтегрування аналітики великих даних в сферу охорони здоров'я - це власне природа самих даних. EHR можуть забезпечити розширену аналітику та допомогти у прийнятті клінічних рішень, надаючи величезні дані. Однак значна частина цих даних наразі не структурована.

Неструктуровані дані – це інформація, яка не відповідає попередньо визначеній моделі чи організаційній структурі. Причина такого вибору може бути просто в тому, що ми можемо записати його в безлічі форматів. Іншою причиною для вибору неструктурованого формату є те, що часто параметри структурованого введення (випадні меню, перемикач і прапорці) можуть не впоратися з записом даних складного характеру. Наприклад, ми не можемо записувати нестандартні дані про клінічні підозри пацієнта, соціально-економічні дані, уподобання пацієнтів, ключові фактори способу життя та іншу пов’язану інформацію в будь-який інший спосіб, крім неструктурованого формату. Важко згрупувати такі різноманітні, але критичні джерела інформації в інтуїтивно зрозумілий або уніфікований формат даних для подальшого аналізу за допомогою алгоритмів для розуміння та використання допомоги пацієнтам.

Більшість даних у сфері охорони здоров’я є неструктурованими, наприклад, з обробки природної мови. Вони часто фрагментовані, розпорошені і рідко стандартизовані. Не секрет, що EHR погано розподіляються між організаційними лініями, але з неструктурованими даними, навіть у межах однієї організації, неструктуровані дані важко агрегувати та аналізувати. Для вирішення цієї великої проблеми знадобиться аналітика великих даних. Дані досліджень у секторі охорони здоров’я є більш неоднорідними, ніж дані досліджень, отримані в інших галузях досліджень. Дані як наукових досліджень, так і громадського здоров’я часто отримують у великих обсягах. Інша проблема, пов'язана зі структурою, є результатом зміни моделі оплати послуг охорони здоров'я. Нарешті, великі дані потребують вирішення проблем з прозорістю метаданих.

2.3.5 Метадані

Щоб мати успішний план управління даними, було б обов’язково мати повні, точні й актуальні метадані щодо всіх збережених даних.

Метадані складатимуться з такої інформації, як час створення, мета та особа, відповідальна за дані, попереднє використання (ким, чому, як і коли) для дослідників та аналітиків даних. Це дозволить аналітикам повторити попередні запити та допомогти пізнішим науковим дослідженням і точному порівняльному аналізу. Це збільшує корисність даних і запобігає створенню «смітників даних», які мало корисні або взагалі не використовуються.

2.3.6 Точність

Деякі дослідження помітили, що звітність даних пацієнтів у EMR або EHR ще не зовсім точна. Це може сприяти проблемам якості великих даних протягом усього їхнього життєвого циклу. EHR мають намір покращити якість та передачу даних у клінічних робочих процесах, хоча звіти вказують на розбіжності в цих контекстах. Якість документації може покращитися, якщо використовувати анкети для самозвіту пацієнтів щодо їхніх симптомів.

2.3.7 Неоднорідність

Неоднорідність даних є ще однією проблемою в аналізі великих даних. Величезний розмір і дуже неоднорідна природа великих даних у сфері охорони здоров’я робить їх відносно менш інформативними з використанням звичайних технологій.

Методи управління та аналізу великих даних постійно розробляються, особливо для потокової передачі даних, захоплення, агрегації, аналітики (з використанням машинного навчання та прогнозування) та рішень візуалізації, які можуть допомогти краще інтегрувати EMR з охороною здоров’я. Наприклад, рівень впровадження EHR перевірених на федеральному рівні та сертифікованих програм EHR у секторі охорони здоров’я в США майже завершений. Однак наявність сотень продуктів EHR, сертифікованих урядом, кожен із яких має різну клінічну термінологію, технічні характеристики та функціональні

можливості, призвело до труднощів у взаємодії та обміні даними. Тепер головна мета полягає в тому, щоб отримати корисну інформацію з цих величезних обсягів даних, зібраних у вигляді EMR.

2.3.8 Попередня обробка зображення

Дослідження виявили різні фізичні фактори, які можуть призвести до зміни якості даних та невірної інтерпретації наявних медичних записів. Медичні зображення часто стикаються з технічними бар'єрами, які включають кілька типів шуму та артефактів. Неправильне поводження з медичними зображеннями також може спричинити підробку зображень. Зменшення шуму, очищення артефактів, налаштування контрасту отриманих зображень і налаштування якості зображення після неправильного використання — це деякі з заходів, які можна застосувати для досягнення мети.

2.3.9 Безпека

Існують значні занепокоєння щодо конфіденційності щодо використання аналітики великих даних, особливо в охороні здоров'я. Дані з вільним доступом є дуже вразливими. Крім того, через чутливість даних охорони здоров'я існують значні побоювання, пов'язані з конфіденційністю. Більше того, ця інформація централізована, і як така, вона дуже вразлива до атак. З цих причин забезпечення конфіденційності та безпеки є дуже важливими.

Було багато порушень безпеки, злому, фішингових атак і епізодів програм-вимагачів, тому безпека даних є пріоритетом для медичних організацій. Після виявлення низки вразливостей було розроблено список технічних заходів безпеки для захищеної медичної інформації (PHI). Ці правила, які називаються правилами безпеки HIPAA, допомагають організаціям використовувати протоколи зберігання, передачі, аутентифікації та контролю доступу, цілісності та аудиту. Загальні заходи безпеки, як-от використання сучасного антивірусного програмного

забезпечення, брандмауерів, шифрування конфіденційних даних та багатофакторна аутентифікація, можуть позбавити від багатьох проблем.

2.3.10 Запит

Метадані полегшать організаціям можливість запитувати свої дані та отримувати відповіді. Однак за відсутності належної взаємодії між наборами даних інструменти запитів можуть не отримати доступ до всього сховища даних. Крім того, різні компоненти набору даних повинні бути добре пов’язані між собою та легко доступними, інакше повний портрет здоров’я окремого пацієнта може не бути створений. Медичні системи кодування, такі як ICD-10, SNOMED-CT або LOINC, повинні бути реалізовані, щоб звести концепції вільної форми в спільну онтологію. Якщо точність, повнота та стандартизація даних не піддається сумніву, то для запитів великих наборів даних і реляційних баз даних можна використовувати мову структурованих запитів (SQL).

2.3.11 Візуалізація

Чиста і приваблива візуалізація даних за допомогою діаграм, теплові карти та гістограми для ілюстрації контрастних фігур і правильне маркування інформації для зменшення потенційної плутанини може значно полегшити нам засвоювання інформації та належне її використання. Інші приклади включають стовпчасті діаграми, кругові діаграми та діаграми розсіювання зі своїми специфічними способами передачі даних.

2.3.12 Стандартизація

Хоча EHR обмінюються даними в межах однієї організації, внутрішньоорганізаційні платформи EHR, у кращому випадку, фрагментовані. Дані зберігаються у форматах, які не сумісні з усіма додатками та технологіями. Ця відсутність стандартизації даних також викликає проблеми при передачі цих даних, ускладнює отримання та очищення даних. Обмежена сумісність створює велику проблему для

великих даних, оскільки дані рідко стандартизуються. Це змушує великі дані стикатися з проблемами, пов’язаними із отриманням та очищенням даних у стандартизований формат, щоб уможливити аналіз та глобальний обмін. З глобалізацією даних великі дані будуть мати справу з різноманітними стандартами, мовними бар’єрами та різною термінологією.

Те, як медичні дані поширюються в багатьох джерелах, якими керують різні штати, лікарні та адміністративні департаменти, також є проблемою. Інтеграція цих джерел даних потребує розробки нової інфраструктури, де всі постачальники даних співпрацюють один з одним.

Пацієнти можуть отримувати або не отримувати допомогу в кількох місцях. У першому випадку обмін даними з іншими організаціями охорони здоров’я був би дуже важливим. Під час такого обміну, якщо дані несумісні, переміщення даних між різними організаціями може бути серйозно обмежено. Це могло бути пов’язано з технічними та організаційними бар’єрами. Це може залишити лікарів без ключової інформації для прийняття рішень щодо подальших заходів та стратегій лікування пацієнтів. Такі рішення, як Fast Healthcare Interoperability Resource (FHIR) і загальнодоступні API, CommonWell (некомерційна торгова асоціація) і Carequality (побудована консенсусом, загальна структура взаємодії), роблять взаємодію та обмін даними простими та безпечними. Найбільша перешкода для обміну даними — це ставлення до даних як до товару, який може забезпечити конкурентну перевагу. Тому іноді постачальники навмисно втручаються в потік інформації, щоб блокувати його між різними системами EHR.

2.3.13 Управління

Управління даними потрібно буде підняти в пріоритетному списку організацій, і його слід розглядати як основний актив, а не як побічний продукт бізнесу. Володіння даними та управління даними повинні

створити нові ролі в бізнесі, які враховують аналітику великих даних, а під час обміну даними потрібно буде залучати нові партнерства.

Постачальникам медичних послуг потрібно буде подолати всі труднощі з цього списку та багато іншого, щоб розробити екосистему обміну великими даними, яка надає надійну, своєчасну та значущу інформацію, об'єднуючи всіх учасників континууму допомоги. Щоб подолати ці проблеми, знадобиться час, зобов'язання, фінансування та комунікація[4].

2.4 Висновки до розділу 2

Ці приклади аналізу даних у сфері охорони здоров'я доводять, що медичні програми можуть рятувати життя і мають бути головним пріоритетом експертів у всій галузі. Навіть зараз аналітика, керована даними, полегшує раннє виявлення, а також втручання при хворобах, упорядковуючи установи для швидшого, безпечнішого та точнішого догляду за пацієнтами. У міру розвитку технологій ці безцінні функції можуть тільки посилюватися – майбутнє охорони здоров'я вже тут, і воно полягає в даних.

Області, у яких розширені дані та аналітика дають найкращі результати, включають: точне визначення пацієнтів, які є найбільшими споживачами медичних ресурсів або мають найбільший ризик несприятливих наслідків; надання людям інформації, необхідної для прийняття обґрунтованих рішень і ефективнішого управління власним здоров'ям, а також легше приймати та відстежувати поведінку; визначення методів лікування, програм і процесів, які не приносять очевидних переваг або коштують занадто дорого; скорочення повторних госпіталізацій шляхом виявлення факторів навколишнього середовища або способу життя, які підвищують ризик або викликають небажані явища, і відповідно коригуючи плани лікування; покращення результатів шляхом дослідження

життєво важливих показників за допомогою домашніх моніторів здоров'я; управління здоров'ям населення шляхом виявлення вразливості серед популяції пацієнтів під час спалахів захворювань або катастроф; і об’єднання клінічних, фінансових та операційних даних для продуктивного та в режимі реального часу аналізу використання ресурсів[14].

Щоб розробити систему охорони здоров’я на основі великих даних, яка може обмінюватися великими даними та надавати нам достовірну, своєчасну та значущу інформацію, нам потрібно подолати також всі вказані проблеми. Для їх подолання знадобляться інвестиції з точки зору часу, фінансування та зобов’язань. Однак, як і інші технологічні досягнення, успіх цих амбітних кроків, очевидно, полегшить нинішнє навантаження на охорону здоров’я, особливо з точки зору витрат.

Покращена діагностика та прогнози захворювань за допомогою аналітики великих даних можуть дозволити знизити витрати за рахунок зниження частоти повторних госпіталізацій. Аналітика великих даних також може допомогти в оптимізації персоналу, прогнозуванні потреб операційних, оптимізації догляду за пацієнтами та покращенні ланцюга поставок фармацевтичних препаратів. Усі ці фактори призведуть до остаточного зниження витрат на охорону здоров’я з боку організацій[4].

Загалом ми помітили три ключові тенденції що забезпечує аналітика охорони здоров’я: досвід пацієнтів значно покращиться, включаючи якість лікування та рівень задоволеності; загальне здоров’я населення також може бути зміцнено на стійкій основі, а експлуатаційні витрати можна значно зменшити.

Галузь змінюється, і, як і будь-яку іншу, великі дані починають змінювати її, але попереду ще багато роботи. Сектор повільно впроваджує нові технології, які просувають його у майбутнє, допомагаючи приймати більш обґрунтовані рішення, покращувати роботу тощо.

«Більшість людей у світі прийматиме рішення, здогадуючись або використовуючи свою інтуїцію. Їм або пощастить, або помиляться». – Сухайл Доші, головний виконавчий директор Mixpanel[16].

Завдяки оцифровці, об’єднанню та ефективному використанню великих даних, організації охорони здоров’я отримають значні переваги.

3 АНАЛІЗ МЕТОДІВ ЕФЕКТИВНОГО ЗБЕРІГАННЯ ВЕЛИКИХ ОБСЯГІВ МЕДИЧНИХ ДАНИХ В УМОВАХ ЇХ НЕПЕРЕРВНОГО ЗРОСТАННЯ

Однією з головних проблем застосування обробки великих даних у галузі медицини, як було визначено раніше, є зберігання, передача та управління даними.

Обсяги даних, що продукує система охорони здоров'я є неймовірно великими і продовжують зростати. У 2011 році організації, що працюють у сфері охорони здоров'я, створили понад 150 ексабайт даних, усі з яких необхідно ефективно аналізувати, щоб бути корисними для системи охорони здоров'я. Згідно з дослідженням IDC, очікується, що обсяг даних, які генеруються системами медичної інформації та зображень, зросте зі 153 ексабайт у 2013 році до 2300 ексабайт у 2020 році[11].

3.1 Вимоги до систем зберігання великих даних

Зберігання даних, пов'язаних з охороною здоров'я, відбувається в різних формах. Існують різноманітні аналітичні методики для інтерпретації медицини, які потім можна використовувати для догляду за пацієнтами. Різноманітне походження та форми великих даних ставлять виклик інформаційній спільноті охорони здоров'я у розробці методів обробки даних. Існує великий попит на техніку, яка поєднує різні джерела даних.

З розповсюдженням великих даних управління самими даними викликає більше зацікавлення, ніж управління обчисленнями. Технології значно розвинулись у забезпечення ефективного зберігання та обробки великих даних, та все ж для побудови платформи аналізу великих даних у

сфері охорони здоров’я, вона повинна підтримувати ключові функції, необхідні для обробки даних. Критерії оцінки платформи можуть включати доступність, безперервність, простоту використання, масштабованість, здатність маніпулювати на різних рівнях деталізації, забезпечення конфіденційності та безпеки, а також забезпечення якості[14].

Отже, постачальники медичних послуг повинні розглянути, як вони будуть зберігати всі ці дані, враховуючи вимоги:

1) Масштабованість — сегмент медичних даних, що найшвидше розвивається, — це неструктуровані дані, такі як МРТ, КТ, рентгенівські та ПЕТ-сканування. Оскільки ці дані зростають до петабайтів, медичним організаціям потрібне масштабоване та недороге рішення для зберігання даних.
2) Відповідність — зберігання має відповідати правилам. HIPAA та інші нормативно-правові акти галузі охорони здоров’я вимагають, щоб дані були захищені засобами контролю доступу на основі ролей (RBAC), журналюванням журналу аудиту та шифруванням даних у стані спокою та SSL для даних, що передаються, із належним чином перевіреними процедурами керування ключами.
3) Незалежний від постачальника архів (VNA) — VNA забезпечує один інтерфейс для кількох інформаційних платформ охорони здоров’я. Це полегшує консолідацію багатьох типів медичної інформації в центральне сховище, яке забезпечує центральний перегляд записів пацієнтів. Платформи зберігання повинні забезпечувати інтеграцію VNA.
4) Стійкість і захист даних — дані про стан здоров’я є частою мішенню для кібератак і представляють великий ризик для постачальників у разі їх випадкової втрати або видалення.

Системи зберігання повинні забезпечувати резервування, реплікацію, резервне копіювання даних і кодування стирання, що може розподіляти фрагменти даних між кількома вузлами[7].

3.2 Концептуальна архітектура аналітики великих даних

Для глибокого розуміння, які саме вимоги можуть ставитись для зберігання даних, та з яким форматами доводиться працювати, варто розглянути весь процес обробки даних. Концептуальна основа проекту аналітики великих даних у сфері охорони здоров’я подібна до аналітики для будь-якої іншої сфери, але відрізняється різноманітністю форматів даних, їх структурованістю. Також аналітика великих даних відрізняється від звичайної аналітики охорони здоров’я. Хоча алгоритми та моделі схожі, користувацькі інтерфейси традиційних інструментів аналітики та тих, що використовуються для великих даних, абсолютно різні; традиційні інструменти аналітики здоров’я стали дуже зручними та прозорими. Інструменти аналізу великих даних, з іншого боку, надзвичайно складні, інтенсивні для програмування та вимагають застосування різноманітних навичок.

Отже, концептуальна модель аналітики великих даних у сфері охорони здоров’я, схематично зображена на рисунку 2, охоплює наступні частини.

3.2.1 Джерела “сирих даних”

Вони можуть бути внутрішніми(наприклад, електронні медичні картки, системи підтримки клінічних рішень) або зовнішніми(державні джерела, лабораторії, аптеки, страхові компанії тощо), охоплювати різні формати даних(файли, .csv, реляційні таблиці, ASCII/текст і т. д.) та програми(додатки для обробки транзакцій, бази даних тощо).

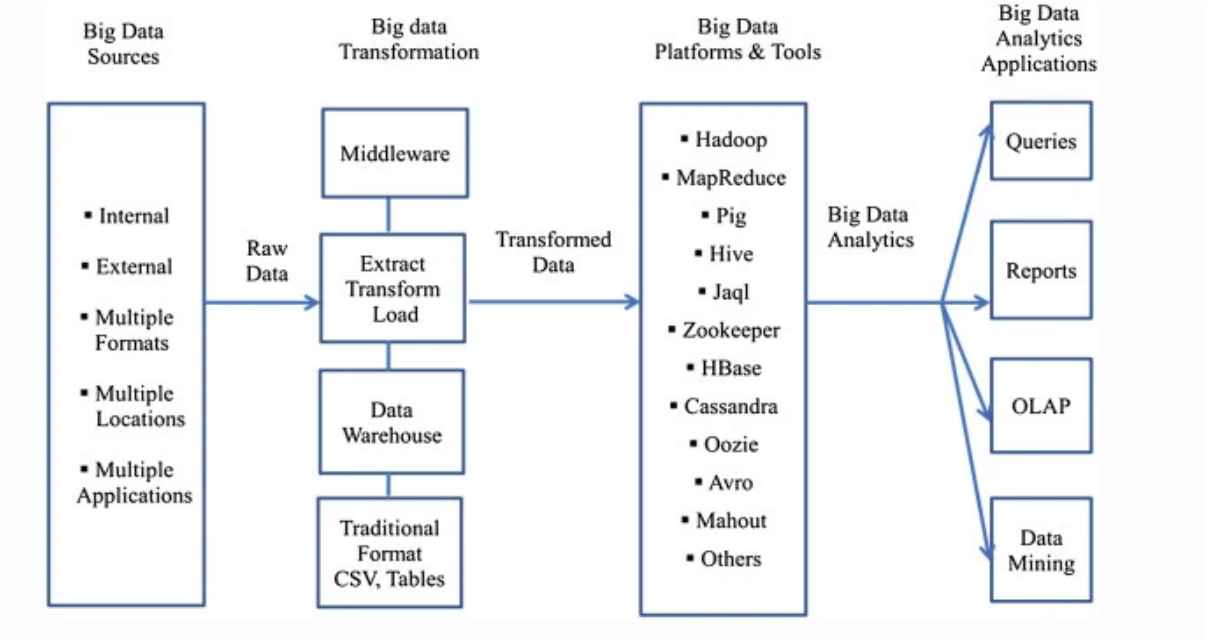

Рисунок 2 – Концептуальна архітектура аналітики великих даних у сфері охорони здоров’я [14]

Джерела та типи даних включають дані Інтернету та соціальних мереж, дані від машини до машини(показання від віддалених датчиків та інших життєво важливих пристроїв), дані про великі транзакції(претензії на медичне обслуговування та інші записи рахунків все більш доступні в напівструктурованих і неструктурованих форматах), біометричні дані(відбитки пальців, генетика, почерк, сканування сітківки ока, рентгенівські та інші медичні зображення, артеріальний тиск, показники пульсу та пульсоксиметрії та інші подібні типи даних, дані, створені людиною(неструктуровані та напівструктуровані дані, такі як EMR, нотатки лікарів, електронна пошта та паперові документи).

3.2.2 Трансформація даних - попередня обробка

У другому компоненті дані знаходяться в необробленому “сирому” стані та потребують обробки або перетворення, після чого доступні кілька варіантів.

- Сервісно-орієнтований архітектурний підхід у поєднанні з вебсервісами (проміжне програмне забезпечення) є однією з можливостей. Дані залишаються необробленими, а служби використовуються для виклику, отримання та обробки даних.
- Іншим підходом є сховище даних, при якому дані з різних джерел агрегуються та готуються до обробки, хоча дані недоступні в режимі реального часу.
- За допомогою етапів вилучення, перетворення та завантаження (ETL) дані з різних джерел очищаються та готуються.
- Залежно від того, структуровані чи неструктуровані дані, в платформу аналізу великих даних можна вводити кілька форматів даних.

Оскільки великі дані за визначенням великі, обробка розбивається і виконується на кількох вузлах. Концепція розподіленої обробки існує десятиліттями. Відносно новим є її використання для аналізу дуже великих наборів даних, оскільки постачальники медичних послуг починають використовувати свої великі сховища даних, щоб отримати уявлення для прийняття більш обґрунтованих рішень, пов’язаних зі здоров’ям. Крім того, платформи з відкритим кодом, такі як Hadoop/MapReduce, доступні в хмарі.

3.2.3 Кінцева обробка даних

У цьому наступному компоненті концептуальної основи приймається кілька рішень щодо підходу до введення даних, розподіленого проектування, вибору інструментів та аналітичних моделей. Обробка здійснюється за допомогою різноманітних засобів та платформ.

3.2.4 Аналітика

Нарешті, переходимо до застосування аналітики великих даних у сфері охорони здоров’я. До них належать запити, звіти, OLAP та

інтелектуальний аналіз даних. Візуалізація є загальною темою в чотирьох програмах. Виходячи з таких галузей, як статистика, інформатика, прикладна математика та економіка, було розроблено та адаптовано широкий спектр методів і технологій для укрупнення, маніпулювання, аналізу та візуалізації великих даних у сфері охорони здоров'я.

3.3 Інструменти для зберігання великих даних

В контексті роботи було опрацьовано 6 досліджень інструментів, що використовується для зберігання та управління великими даних у сфері охорони здоров'я та проаналізовано їх ефективність та особливості використання. Консолідовані дані результатів проведеного аналізу представлені в Таблиці 2.

Таблиця 2. Інструменти для зберігання великих даних

	[4]	[14]	[11]	[2]	[15]	[12]	[18]
Hadoop	*	*	*	*	*	*	
HDFS	*	*	*	*	*		
MapReduce	*	*	*	*	*	*	
Hive		*	*	*	*		
HBase		*	*		*		*
PIG and PIG Latin		*	*		*		
Avro		*	*				
MongoDB		*		*			*
CouchDB		*					*
Cassandra		*					*
Cloud services		*				*	*
Non relational DB						*	
Google Big Query						*	

Як результат, розглянемо найбільш поширені інструменти для роботи з великими даними.

3.3.1 Hadoop

Завантаження великої кількості (великих) даних у пам’ять навіть найпотужніших обчислювальних кластерів не є ефективним способом роботи з великими даними. Тому найкращий логічний підхід для аналізу величезних обсягів складних великих даних — розподілити та обробити їх паралельно на кількох вузлах. Однак розмір даних зазвичай настільки великий, що для розподілу й завершення обробки за розумний проміжок часу потрібні тисячі обчислювальних машин. Працюючи з сотнями чи тисячами вузлів, доводиться вирішувати такі питання, як паралелізація обчислення, розподілення даних та обробка збоїв.

Одним із найпопулярніших поширених програм з відкритим кодом для цієї мети є Hadoop. Hadoop має потенціал для обробки надзвичайно великих обсягів даних, головним чином, розподіляючи розділені набори даних численним серверам (вузлам), кожен з яких вирішує різні частини більшої проблеми, а потім інтегрує їх для отримання кінцевого результату. Hadoop може виконувати подвійну роль організатора даних і інструмента аналітики. Зокрема, Hadoop дає змогу обробляти надзвичайно великі обсяги даних з різними структурами або взагалі без них. Hadoop має різні інструменти, які покращують компоненти зберігання та обробки, тому багато великих компаній, як-от Yahoo, Facebook та інші, швидко застосували його[14].

3.3.2 Розподілена файлова система Hadoop (HDFS)

Це компонент файлової системи, який забезпечує масштабоване, ефективне та засноване на репліках зберігання даних на різних вузлах, які є частиною кластера. Тут розподілена файлова система (DFS) дозволяє користувачам ділитися даними та ресурсами зберігання, використовуючи

загальну файлову систему. DFS надає користувачам високопродуктивне масштабоване рішення. Проте DFS є надійним у порівнянні з кластером баз даних[13].

HDFS відноситься до розподіленої файлової системи Hadoop, яку можна використовувати для обробки неструктурованих даних переважно на стандартному обладнанні. HDFS забезпечує швидке виявлення несправностей і автоматичне відновлення, оскільки містить велику кількість компонентів. Однак існує ймовірність відмови блоку та його непрацездатності. Блочна реплікація пропонується, щоб уникнути відмови вузла, недоступності чи втрати даних. Реплікація забезпечує не тільки доступність, але й надійність системи, і вона автоматично обробляється HDFS NameNode. Замість того, щоб бути просто рівнем зберігання Hadoop, HDFS є окремою розподіленою файловою системою, яка допомагає покращити пропускну здатність системи. HDFS є основним сховищем даних, де кожен файл поділений на блоки фіксованого розміру і розподілений між численними серверами (вузлами). HDFS використовує архітектуру master/slave, використовуючи NameNode (головний вузол) і DataNode (підпорядкований вузол).

Важливо також пам’ятати й про обмеження, що накладає система та проблеми, які можуть виникнути в процесі її використання. Перш за все варто звернути увагу на спосіб зберігання файлів у HDFS. Блочна система зберігання даних (файли розбиваються на блоки по 64 Мб або 128 Мб) може стати як перевагою, наприклад для файлів великого розміру, так і проблемою, оскільки навіть для файла об’ємом менше 64 Мб в пам’яті головного вузла виділяється місце для зберігання даних про цей файл. Тобто велика кількість маленьких файлів буде займати значно більше місця в головному вузлі, ніж великі файли такого ж загального обсягу.

Крім того, зчитування великої кількості маленьких файлів може привести до затримок, оскільки призводить до “перескакування” з одного

вузла даних на інший. Варто також пам’ятати і про типи даних, оскільки в кожного з них є свої переваги. Для прикладу пропоную розглянути 5 найчастіше використовуваних форматів: CSV, JSON, AVRO, ORC, Parquet.

Їх показники ефективності представлені в таблиці 3 та на рисунку 3.

Таблиця 3. Порівняння форматів даних

	avro	csv	json	orc	parquet
Час запису, s	25	14	19	32	15
Випадковий вибір, s	6,275	5,163	17,989	4,955	4,991
Aggregation (count) query, s	43,099	42,834	55,536	25,958	30,123
Filtering query, s	0,982	0,696	13,753	0,785	1,008
Group by query, s	4,698	5,231	17,124	4,144	4,015
Distinct value query, s	5,069	5,816	19,096	3,995	4,475
Space utiliaztion, Mb	308,1480455	651,8611956	1892,795997	202,4573822	203,2187119

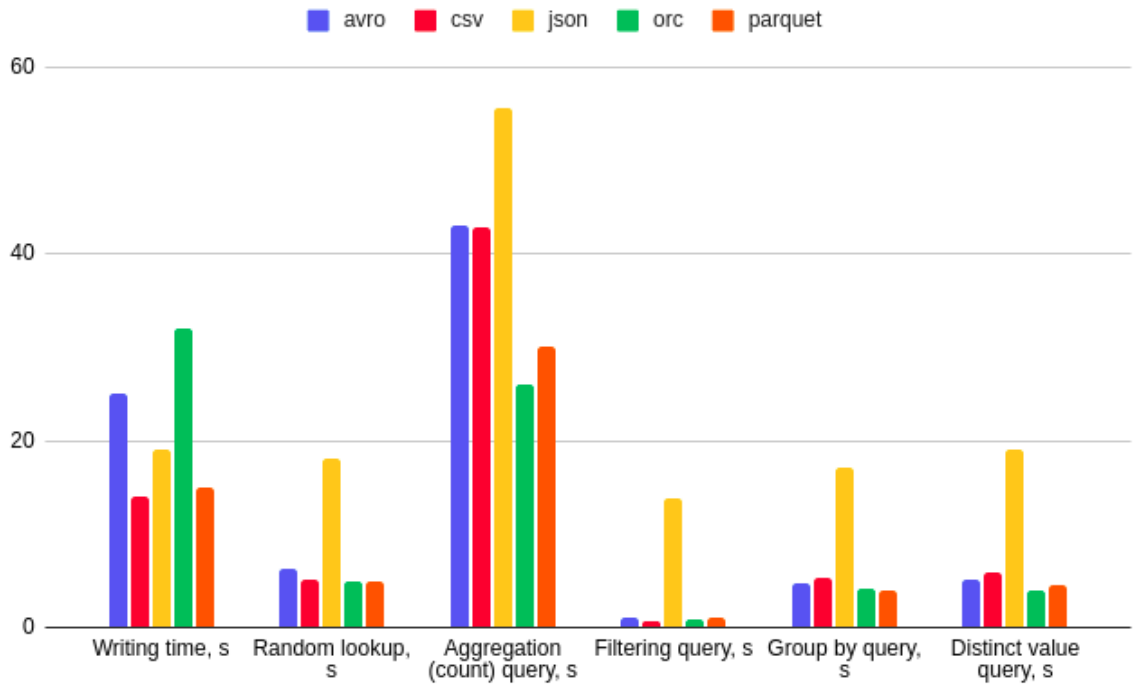

Рисунок 3 – Порівняння ефективності використання різних форматів даних.

3.3.3 Apache Hive

Hive — це рівень сховища даних поверх Hadoop, в якому можна виконувати аналіз і запити за допомогою процедурної мови, схожої на SQL. Apache Hive можна використовувати для виконання спеціальних запитів, узагальнення та аналізу даних. Hive вважається де-факто стандартом для запитів на основі SQL над петабайтами даних за допомогою Hadoop і пропонує функції легкого вилучення даних, перетворення та доступу до HDFS, що містить файли даних або іншу систему зберігання HBase чи сумісні файлові системи, такі як файлова система Amazon S3. Перекладає ці запити в завдання MapReduce і надсилає їх до Hadoop для виконання.

3.3.4 Нереляційні бази даних (NoSQL)

Реляційні бази даних були дуже ефективними для інтенсивних обсягів даних з точки зору процесів зберігання та пошуку протягом багатьох десятиліть. Однак поява та доступністю Інтернет-технології для громадськості перетворили структуру даних на безсхемну, взаємопов’язану та швидко зростаючу. Крім того, складність даних, які генеруються вебресурс, не дозволяє використовувати технології реляційних баз даних. Експоненційне зростання, відсутність структури та різноманітність типів створюють проблеми для зберігання та аналізу даних для традиційних систем управління даними. Перетворення структур великих даних у реляційні моделі даних, суворо визначені реляційні схеми та складність процедур для простих завдань – це жорсткі особливості реляційних баз даних, які неприйнятні для великих даних.

Технології NoSQL пропонують гнучкі моделі даних, горизонтальну масштабованість і моделі даних без схем[24]. Бази даних NoSQL пропонують певний рівень обробки транзакцій, щоб вони були достатніми для соціальних мереж, електронної пошти та інших вебдодатків. Щоб покращити доступність даних для користувачів, дані поширюються та реплікуються на кількох вузлах. Реплікація на одному вузлі не тільки підтримує відновлення даних у разі будь-яких пошкоджень, але також сприяє високій доступності, якщо репліки створюються в різних географічних місцях. Узгодженість — це ще один аспект розподілених систем зберігання даних, коли дані мають кілька копій, підтримувати їх в актуальному стані на кожному сайті стає складніше. Брюер зазначив, що перевага доступності або узгодженості є загальною метою проектування для розподілених баз даних, тоді як мережеві розділи зустрічаються рідко[17].

NoSQL забезпечує кращу продуктивність у надійному зберіганні та управлінні великомасштабними даними незалежно від їх формату. Ця властивість робить його справжнім претендентом на дослідницьку роботу з аналізу великих даних, оскільки це також робить його більш гнучким і може бути корисним в аварійному відновленні, безпеці та масштабованості.

3.3.5 Хмарні сховища

Найпоширенішими платформами для роботи з програмними фреймворками, які допомагають аналізувати великі дані, є високопотужні обчислювальні кластери, доступ до яких здійснюється через інфраструктуру мережевих обчислень. Хмарні обчислення – це така система, яка має віртуалізовані технології зберігання даних і надає надійні послуги. Вони пропонують:

- Високу надійність

- Масштабованість
- Автономність
- Повсюдний доступ
- Динамічне виявлення ресурсів і їх компонування.
- Зменшення витрати на управління ресурсами, надаючи їх через мережу на вимогу.

Такі платформи можуть діяти як приймач даних від всюдисущих датчиків, як комп’ютер для аналізу та інтерпретації даних, а також надавати користувачеві легку для розуміння вебвізуалізацію. В Інтернеті речей обробку та аналітику великих даних можна виконувати ближче до джерела даних за допомогою послуг мобільних периферійних обчислювальних хмар і туманних обчислень. Для реалізації підходів ML та AI для аналізу великих даних в обчислювальних кластерах потрібні розширені алгоритми.

Крім того, найбільш вдалі NoSQL рішення насамперед націлені на вирішення специфічних завдань і створюються гігантами IT-індустрії для власних потреб – це продукти від Google, Amazon, Microsoft та Apache, які обслуговують конкретні проекти[1]. І більшість цих рішень доступні лише в контексті використання хмарних платформ. Тому вважаю важливим розглянути найбільш популярні продукти у сфері хмарних NoSQL баз даних.

3.4 NoSQL бази даних

Оскільки нереляційні бази даних є широко застосовуваними як сховища для великих даних, то в контексті роботи актуально розглянути їх типи, особливості застосування, переваги та недоліки, щоб отримати краще розуміння умов, коли та чи інша база найкраще підходить для вирішення конкретної задачі зберігання та доступу до великих обсягів даних.

Обробка даних у базах даних NoSQL зазвичай швидше, ніж у реляційних базах даних. У цих базах даних ефективно використовуються дані різної структури та розмірності [13]. Крім цього всі наступні розглянуті бази даних мають відкрити код. Вони є масштабованими, відмовостійкими, орієнтованими на обробку великої кількості запитів та зберігання великих об’ємів даних, що постійно зростають.

Основні типи таких баз даних:

3.4.1.Ключ-значення

Бази даних «ключ-значення» розроблені для зберігання великих даних і:

- Зберігають не тільки структуровані, але й особливо неструктуровані дані у вигляді ключа та відповідного значення кожного запису даних[27].
- Мають відповідну структуру зберігання для постійно зростаючих, суперечливих значень великих даних, для яких потрібна швидша відповідь на запити.
- Забезпечують підтримку зберігання великих обсягів даних і виконання одночасних запитів.
- Зберігають дані в невеликих об’єктах замість блоків, і їх легко налаштувати.
- Значення в записах можуть відрізнятися або мати різне представлення. Таким чином, ця структура забезпечує менше споживання пам’яті та гнучкість додавати більше записів[28].
- Основний сценарій - це коли потрібен пошук кількох функцій із записів.

До таких баз даних належать:

1) Scalaris — це масштабоване, розподілене й високо доступне сховище ключів і значень, яке розроблено для виконання інтенсивних вимог

до читання/запису. Для одночасного доступу та підтримки інтенсивних транзакцій звичним є виникнення збої вузлів. Однак Scalaris вдається досягти узгодженості для критичних операцій запису. Структурована мережа накладання запитів реалізована як перший рівень архітектури Scalaris і забезпечує підтримку ряду запитів. Архітектура Scalaris також застосовує реплікацію, щоб зробити систему високо доступною та стійкою до відмов.

a) Строга узгодженість
b) Масштабованість і висока доступність з балансуванням навантаження та відмовостійкістю
c) Мала кількість накладних витрат на технічне обслуговування
d) Самоуправління
e) Мета застосування - досягти узгодженості для транзакцій з інтенсивним читанням/записом; для створення масштабованих онлайн-сервісів

2) Aerospike — це перше сховище даних у режимі реального часу з відкритим вихідним кодом, оптимізоване для флешпам'яті, яке забезпечує масштабованість та надійність з дуже низькою вартістю. Aerospike має архітектуру «нічого спільного», яка підтримує об’єм даних у петабайтовому масштабі з надійністю та лінійною масштабованістю. Aerospike розроблено для забезпечення масштабованості, реплікації, автоматичного відновлення розділу вузла та високої доступності. Ефективний з точки зору робочого навантаження читання та запису, узгодженості та вартості операцій читання та запису.

a) Висока масштабованість, узгодженість та надійність
b) Мета застосування - розробити масштабовану та гнучку платформу для веб-додатків; підтримувати надійність і послідовність як традиційні бази даних

3) Redis — це кеш і «ключ-значення» сховище для великих даних із відкритим вихідним кодом, яке забезпечує ефективну структуру даних для індексації для прискорення виконання запитів і відповідей. Redis має значну підтримку реплікації в середовищі master-slave. Призначена для ситуацій частого доступу до даних. У разі більших вимог до масштабованості, коли в системах не вистачає оперативної пам’яті, сховище в пам’яті Aerospike є кращим, ніж Redis. Всі дані знаходяться в пам’яті. Однак Redis має деякі інші потужні функції, такі як вбудована стійкість і підтримка більшої кількості типів даних. Redis є постійним як справжня база даних, і дані не будуть втрачені при перезапуску. Крім того, підтримка більшої кількості типів даних є унікальною властивістю Redis.
    a) Висока масштабованість, неоднорідність платформи, серверів і додатків, а також дані в пам’яті.
    b) Автоматичне розділення
    c) Ефективний доступ до читання/запису даних
    d) Відмовостійкість
    e) Для ефективної підтримки операцій запитів і реплікації в середовищі master-slave з підвищеною продуктивністю оновлення
4) Voldemort — це універсальне рішення для розподіленого зберігання даних для великомасштабних даних. виконання кластеризації та реплікації даних. Voldemort забезпечує остаточну узгодженість, оскільки операції читання/запису можуть виконуватися на будь-якому вузлі та протягом дуже короткого періоду, як це трапилося, коли представлення даних є суперечливим.
    a) Автоматичне розбиття та реплікація даних
    b) Прозоре усунення несправностей

c) Висока доступність читання/запису та горизонтальна масштабованість
d) Мета застосування - для забезпечення розподіленого та узгодженого зберігання великомасштабних даних лише для читання; для підтримки прозорості розповсюдження, прозорості збоїв і версій для забезпечення цілісності

5) KAI
   a) Висока відмовостійкість з низькою затримкою
   b) Масштабованість з конфігурованими вузлами
   c) Остаточна узгодженість
   d) Мета застосування - забезпечити масштабоване та надійне рішення шляхом впровадження Amazon Dynamo
6) Riak — забезпечує високу доступність і менш коштовну масштабованість.
   a) Надає простіші моделі даних і вирішення конфліктів даних.
   b) Відмовостійкість і висока доступність
   c) Розв'язання конфлікту даних
   d) Підтримка налаштування стандартного обладнання
   e) Мета застосування - для забезпечення високої доступності додатків і платформ
7) MemcacheDB — це сховище з відкритим кодом, розроблене для швидкого та надійного зберігання та доступу до об’єктів і даних. Реалізує такі функції як транзакції та реплікація.
   a) Постійна база даних для динамічних вебдодатків, керованих базою даних.
   b) Можна отримати доступ через будь-який API, оскільки він використовує протокол Memcached.
   c) Швидкий доступу з пам’яті - стає легко забезпечити стабільність даних.

d) Стійкість даних, коли запит на читання даних може бути виконаний будь-яким вузлом, але операція запису виконується лише на головному вузлі.
e) Працює з одним головним і кількома реплікаційними вузлами. Однак він гнучкий, щоб дозволити будь-якому вузлу стати реплікаційним вузлом, але це виявляє проблему безпеки.
f) Мета застосування - для швидкого та надійного зберігання та пошуку

3.4.2 Орієнтовані на стовпці

Ця категорія баз даних NoSQL підходить для вертикально розділених, безперервно збережених і стиснутих систем зберігання[19]. Бази даних, орієнтовані на стовпці:

- Зберігають стовпці даних окремо, на відміну від традиційного сховища, де дані зберігаються у вигляді повних записів.
- Зчитування даних та отримання атрибутів у таких системах є досить швидкими та порівняно менш дорогими, оскільки доступ здійснюється тільки до відповідного стовпця і для кожного стовпця виконується одночасне виконання процесу.
- Мають високу масштабованість і, зрештою, узгоджені[26].
- Забезпечують підтримку додатків для надійного та високо доступного зберігання.
- Зручно додавати нові функції в усі рядки.

До таких баз належать:

1) HBase — забезпечує масштабований, розподілений, відмовостійкий і випадковий доступ до великих даних, орієнтований на читання-запис, поверх Hadoop і HDFS. HBase використовує базову HDFS від Hadoop для зберігання даних таблиці.

a) Що стосується паралельності, має чудову підтримку транзакцій з інтенсивним читанням.
b) Піклується про розподіл і використовує кластерний підхід для управління даними; ось чому це потенційне рішення для надзвичайно великої кількості рядків даних.
c) Підтримка версіонування.
d) Мета застосування - забезпечення послідовного, випадкового та доступу в режимі реального часу до масштабованих таблиць із запитами на читання/запис, зберігання даних секвенування генів[10].

2) Hypertable — це розподілена база даних, яка забезпечує дуже хорошу підтримку узгодженості збережених даних.
   a) Для роботи і сумісна з багатьма розподіленими файловими системами, такими як GFS, HDFS і CloudStore.
   b) Зберігає дані у вигляді таблиць і розбиває таблиці для досягнення розподілу та масштабованості.
   c) Забезпечує високу доступність і узгодженість завдяки розподіленій конфігурації реплік за допомогою розподіленого консенсусного протоколу.
   d) Високо доступне та швидке випадкове читання/запис
   e) Мета застосування - для забезпечення паралельних, високопродуктивних, масштабованих баз даних для даних великого розміру; для підтримки кращої продуктивності запитів для великого розміру даних
3) Cassandra — це децентралізована і високо доступна колонко-орієнтована система зберігання структурованих даних «ключ-значення», яка включає велику кількість базових центрів обробки даних[9].

a) Забезпечує масштабованість, зберігання екземплярів і покращує продуктивність для частих запитів на операції читання/запису.
b) Забезпечує значну підтримку транзакцій з інтенсивним записом.
c) Постійно бореться з відмовами компонентів у масштабних системах і досягає загальної надійності та масштабованості.
d) Кластеризація, розділення та реплікація з дотриманням відмовостійкості, зниження затримки та масштабованості є визначними рисами Cassandra[25].
e) Висока пропускна здатність запису, відсутність компромісів при читанні (блокування)

3.4.3 Орієнтовані на документи

Модель даних, орієнтована на документ, подібна до структури ключ-значення і зберігає дані у формі ключа та значення як посилання на документ[21]. Однак бази даних документів:

- Підтримують складніші запити та ієрархічні зв’язки.
- Зазвичай реалізують формат JSON і пропонують дуже гнучку схему. Хоча архітектура сховища не має схеми для структурованих даних, індекси добре визначені в документно-орієнтованих базах даних.
- Витягують метадані для подальшої оптимізації та зберігають їх як документи.
- Продуктивність евристичної схеми залежить від введених даних користувача та природи запитів.

1) MongoDB — це високо доступне, масштабоване та відмовостійке документо-орієнтоване рішення. MongoDB отримує характеристики MySQl за допомогою моделі даних JSON.

a) Має таку ж горизонтальну масштабованість, легкість розробки з підтримкою динамічних схем для всіх видів даних документів та ефективність керування, як і MySQL.
b) Індексація, динамічні та тимчасові запити, агрегація та динамічні оновлення є одними з великих засобів MySQL, які використовуються MongoDB з незначними змінами[6].
c) MongoDB зберігає документи у вигляді даних у двійковому представленні під назвою BSON, що дозволяє легко відображати дані з додатків у бази даних.
d) Надійність і доступність даних досягаються за допомогою набору реплік, а кілька реплік доступні, коли основний сервер не відповідає.
e) Надає підтримку для зберігання та об'єднання кількох структурних даних, тоді як індексація все ще можлива.
f) Мета застосування - для забезпечення засобів реляційної моделі даних для динамічних схем на основі документів; для підтримки швидкого й узгодженого доступу до даних із різних програм через декілька інтерфейсів, для зберігання EHR [10].

2) Terrastore — це ще одна система розподіленого на основі документів,
    a) Пропонує підтримку з високою масштабованістю, а також узгодженістю та динамічною кластеризацією під час виконання.
    b) Підтримка кількох кластерів.
    c) Автоматичне балансування навантаження для зберігання даних виконується, коли вузли підключаються або відключаються від кластерів.
    d) Оптимізовано для узгодженості, оскільки виконує послідовне хешування для розділення і не створює репліки.

e) Має більше накладних витрат, ніж Redis, що знижує його продуктивність.

f) Мета застосування - щоб досягти узгодженості даних документів через розповсюдження

3) CouchDB — це розподілена, масштабована база даних. Використовує складний механізм паралельності у разі інтенсивних запитів доступу, який не дає системі відмови. Однак таке інтенсивне навантаження спричиняє затримку загальних відповідей.

a) Підтримує динамічну структуру даних, що дає змогу за потреби визначити схему.

b) Реалізує модель даних JSON, яка допомагає CouchDB підтримувати напівструктуровані дані.

c) Дозволяє зберігати будь-які дані у вигляді документів.

d) Легке використання

e) Відмовостійкість

f) Паралельність робочого навантаження запиту

g) Мета застосування - для забезпечення динамічної та автономної схеми для вебдокументів, зберігання даних у вигляді зображень[10]

4) OrientDB є першою багатомодельною базою даних з відкритим вихідним кодом і високою масштабованістю для даних документів з розширеним, прозоро керованим рівнем графів, що забезпечує зв’язки між записами. Рівень вбудовування графів робить OrientDB не тільки стійкою і компактною, але й швидкою в обході даних та управлінні зв’язками даних без збільшення вартості.

a) Багатомодельна платформа дозволяє класифікувати її як документно-орієнтовану базу даних графів.

b) Незалежно від обсягу даних, має чітку швидкість зберігання даних, а також транзакцій читання та запису.

c) Для забезпечення продуктивності та масштабованості кластеризація та реплікація на гетерогенних серверах є значною перевагою
d) Має модель без схем, підтримку вторинних індексів, шардінг для розділення даних і модель реплікації "головний-головний".
e) Має додаткові функції, такі як мультимодельний підхід для підтримки всіх операційних систем.
f) Мета застосування - забезпечити багатомодельне масштабоване сховище для досягнення функцій моделі, орієнтованої на графіки та документи

3.4.4 Графові бази даних

Графові бази даних є найкращим вибором для зберігання даних разом із відношеннями. Вони пропонують:

- Постійне зберігання об’єктів і зв’язків і підтримують прості й зрозумілі запити з власним синтаксисом.
- Забезпечують легкий обхід даних.
- Досягнення низької затримки. Наприклад, для отримання рекомендацій із даних відгуків клієнтів на комерційному вебсайт потрібні самостійні багаторівневі запити до традиційних баз даних, що стає дуже складною операцією. Навпаки, для графової бази даних ці маніпуляції з даними є досить простими: два рядки коду без впливу на структуру даних.

1) Лідером у галузі є Neo4j.
a) Масштабованість, високе навантаження транзакцій, паралельність та продуктивність для робочих навантажень із запитами на читання.

b) Підтримка інтенсивних транзакцій без блокування досягається за допомогою буферизації.
c) Незважаючи на свої переваги, Neo4j забирає багато часу на створення готової до виробництва та надійної бази даних.
d) Мета застосування - для забезпечення баз даних реляційних графів для інтенсивного зв'язування даних; для підтримки маніпулювання відносинами з даними та прийняття рішень

2) HyperGraphDB — це реалізація моделі гіперграфа для розробки графової бази даних із відкритим кодом для штучного інтелекту та вебсемантичних проектів.
a) Забезпечує важливу настроювану функцію індексування, що забезпечує ефективний пошук даних і обхід графіка.
b) Використовує механізм «ключ-значення» для зберігання інформації графа, такої як вузли та ребра, як ключ.
c) Реалізує одноранговий механізм розподілу даних. Кожен одноранговий пристрій працює незалежно, а оновлення виконуються асинхронно і в кінцевому підсумку.
d) Мета застосування - розробка моделі постійної пам'яті для проектів штучного інтелекту та семантичних вебпроєкт; для забезпечення як реляційного, так і об'єктноорієнтованого управління даними.

3.5 Хмарні NoSQL бази даних

1) DynamoDB — це широко використовуваний інструмент для баз даних ключ-значення без схем від Amazon. DynamoDB переважно використовується для зберігання неструктурованих, змінних і масштабованих даних[23].

a) Пропонує нескінченні можливості масштабування для зберігання даних і швидкості доступу.
b) Застосовується, коли потрібні ефективне індексування та адаптивна масштабованість.
c) Доступність і довговічність можна досягти за допомогою автоматичної реплікації.
d) Незалежно від розміру запиту, забезпечує стабільну продуктивність і видиме використання ресурсів.
e) Ефективно керована база даних; тому він підходить для масштабованих додатків з експоненційно зростаючими даними.
f) Мета застосування - для підтримки розподіленого зберігання даних масштабованого розміру; щоб підвищити ефективність пошукових запитів

2) BigTable — це колонко-орієнтований продукт, розроблений компанією Google Inc. для забезпечення гнучкого та високопродуктивного сховища для великомасштабних структурованих даних, розповсюджених на великій кількості стандартних серверів.
    a) Адаптивна, надійна система зберігання для керування даними в петабайтовому масштабі на тисячах машин.
    b) Високопродуктивне та доступне сховище даних.
    c) Високі вимоги до пропускної здатності або затримки.
    d) Забезпечує динамічний контроль над розміщенням даних, представленням, індексуванням і кластеризацією.
    e) Мета застосування - для забезпечення поширення високомасштабованих структурованих даних
3) SimpleDB — це документно-орієнтована база даних з відкритим вихідним кодом, яка доступна як сервіс Amazon. Без навантаження

на адміністрування бази даних, SimpleDB забезпечує високу доступність і довговічність даних за допомогою автоматичної географічної реплікації. Крім того, модель даних є гнучкою, і для даних виконується автоматичне індексування. Таким чином, автоматичне надання адміністрування баз даних робить розробку додатків простою за допомогою SimpleDB. Незважаючи на простоту керування даними, яку забезпечує SimpleDB, масштабованість обмежена 10 ГБ.

3.7 Висновки до розділу 3

В цьому розділі було детально розглянуто, на яких етапах необхідно оптимізувати зберігання даних, та в якому вигляді та форматі ці дані можуть бути збережені. Отже, на етапі збору даних найбільш оптимальним рішенням для малоструктурованих даних є розподілені файлові системи - звичайні або хмарні - або ж набір із кількох розглянутих баз даних. Після попередньої обробки дані найкраще зберігати в одній із баз даних для легкості здійснення наступних кроків. Таких як використання методів машинного навчання та штучного інтелекту, проведення аналітики та візуалізації, побудови складних запитів та звітів.

4 РЕАЛІЗАЦІЯ РІШЕННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

Для того, щоб правильно підібрати базу даних (або їх набір), необхідно розуміти весь шлях використання даних від моменту збору і до кінцевого результату, що був метою обробки цих даних. Яка структура даних? Як часто вони будуть зчитуватись, як часто змінюватись? Які зв’язки між ними існують? Які показники та характеристики є ключовими для майбутньої бази, а якими можна пожертвувати на користь важливіших? Який формат повинні мати дані?

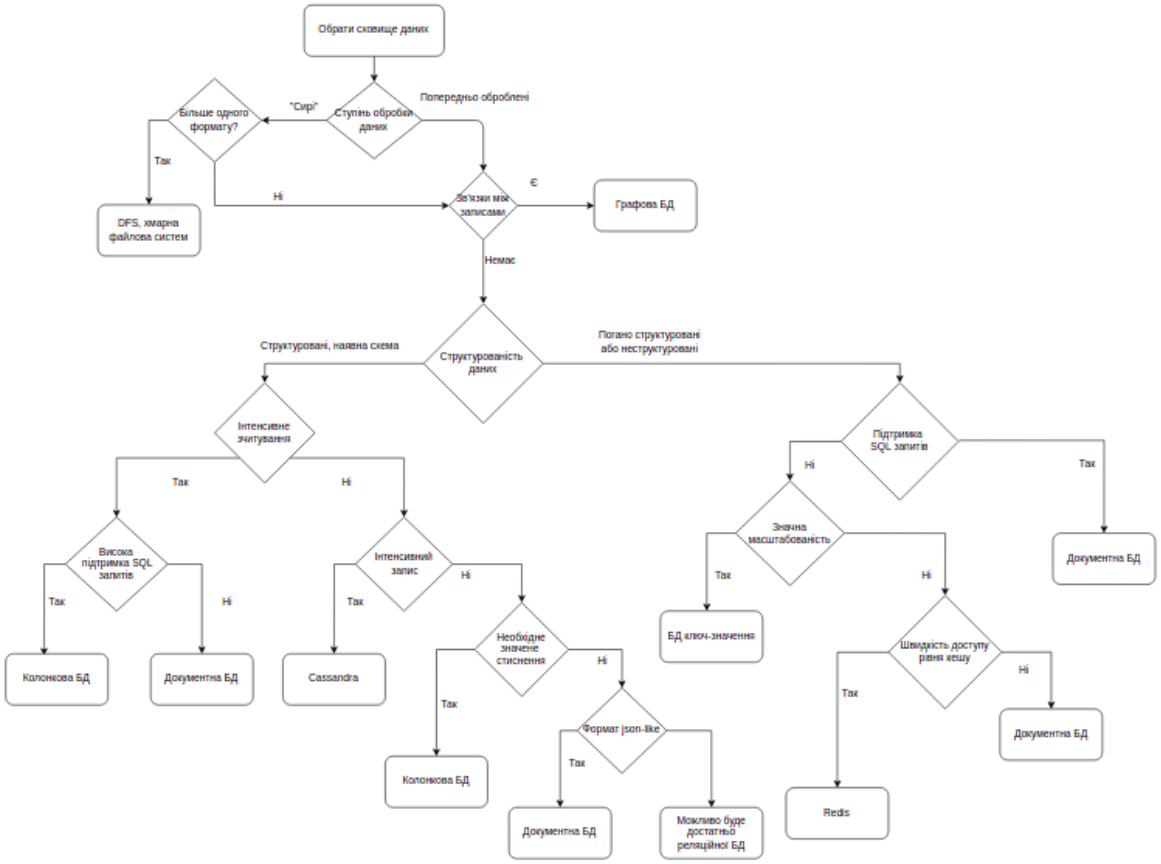

Рисунок 4 – Дерево ухвалення рішення про вибір слушного сховища даних

В процесі аналізу було розроблено прототип дерева ухвалення рішень - Рисунок 4, що призначене для спрощення прийняття рішення, яке саме сховище даних найкраще підходить для вирішення тієї чи іншої проблеми збереження даних. Можливі варіанти охоплюють розподілені файлові системи та хмарні сховища, нереляційні бази даних: графові, документні, колонкові та типу ключ-значення. Кожна з баз забезпечують швидкість доступу, масштабованість, відмовостійкість.

Детальніше розглянемо окремі частини дерева, щоб глибше розуміти, чому саме було враховано ті чи інші атрибути, та які логічні висновки можна зробити на їх основі.

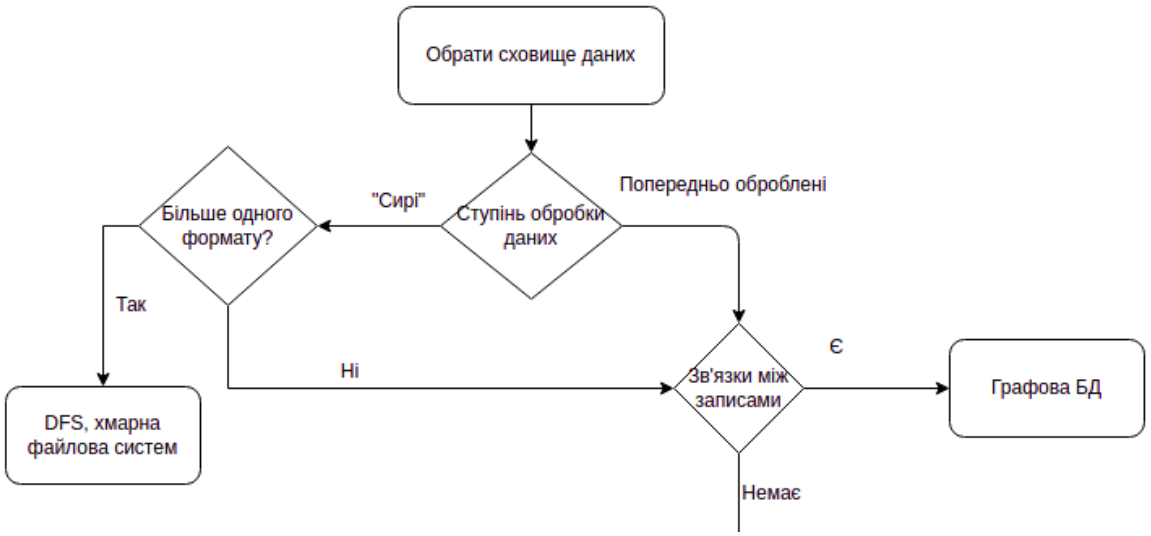

Рисунок 5 – Частина перша дерева рішень.

В корені дерева стоїть одне із найбільш важливих запитань (Рисунок 5). Спершу необхідно подивитись на картину глобально: який ступінь обробки даних, які ми маємо намір зберігати. Якщо це необроблені, “сирі” дані, то є сенс розглянути як сховище даних розподілену файлову систему, а в найкращому випадку - хмарну файлову систему, оскільки остання дозволяє зберігати дуже великі обсяги даних і забезпечує автоматичне масштабування, до того ж оплата здійснюється лише за використовуване

місце. Цей варіант особливо актуальний, коли дані надходять з різного роду джерел, мають різну структурованість, різний формат. До прикладу, дані, що збирає лікарня, які включають в себе EHR, моніторинг показників приладів, дані про кількість пацієнтів та інші. Якщо ж навіть необроблені дані отримані в конкретному структурованому форматі, можливо є сенс використовувати один з інших варіантів, які розглянемо далі.

Якщо дані, що мають бути збереженими, вже зазнали попередньої обробки, то є сенс розглядати як сховище одну із нереляційних баз даних. На цьому етапі переходимо до наступного вузла, де потрібно визначити наявність та важливість зв'язків між записами. Якщо вони присутні й будуть часто використовуватись, то варто розглянути графову базу даних, яка пропонує швидкий обхід графу. Часто такі бази використовують, для виявлення шахрайства, що актуально у сфері охорони здоров'я[22].

Якщо ж дані не передбачають зв'язків між записами, тоді переходимо до наступної частини - визначення структурованості даних (Рисунок 6). Коли дані є структурованими та будуть мати чітко задану схему, то необхідно розглянути бази даних, що дозволяють наявність схем, тобто такі як колонкові та документні, крім того, структурованість даних дозволяє зразу ж відкинути варіант бази типу ключ-значення, оскільки вона не підтримує схеми. Для структурованих даних визначаємо вимоги інтенсивності зчитування даних. Якщо це необхідно, тоді потрібно зрозуміти, чи будемо будувати складні SQL-запити, як такі, що використовуються для аналітики. Якщо це правда, тоді найкраще підійде колонкова база даних. Цей варіант підходить для аналітики завантаженості лікарні, визначення закономірностей, що стосуються певних груп населення або громадського здоров'я в цілому, витрати та інші дані, що можуть потребувати аналітики, створення звітів або побудови досить складних запитів. В протилежному випадку можливо хорошим варіантом буде розглянути документно-орієнтовану базу даних. Як варіант

застосування можна зазначити використання EMR для визначення тенденцій в контексті деяких досліджень та оновлення даних пацієнтів в EHR, оскільки документні БД добре справляються зі зміною вмісту записів.

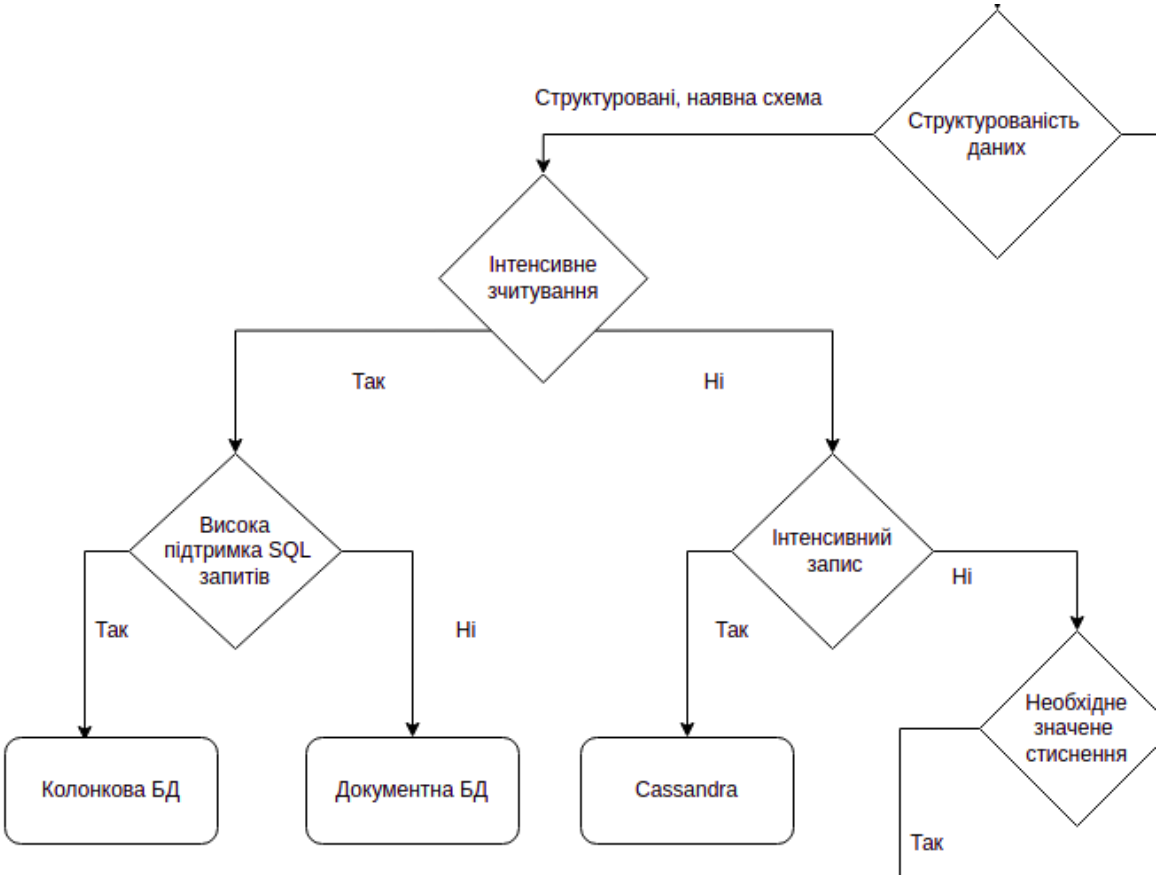

Рисунок 6 – Частина друга дерева рішень.

У випадках, коли вимог до інтенсивності зчитування немає, перевіряємо вимоги, що стосуються інтенсивності запису. Коли це необхідно, хорошою ідеєю було б використати базу даних Apache Cassandra, що поєднує в собі переваги колонкової та ключ-значення БД. Наприклад, для запису показників медичних вимірювальних пристроїв та моніторингу стану пацієнтів у реальному часі.

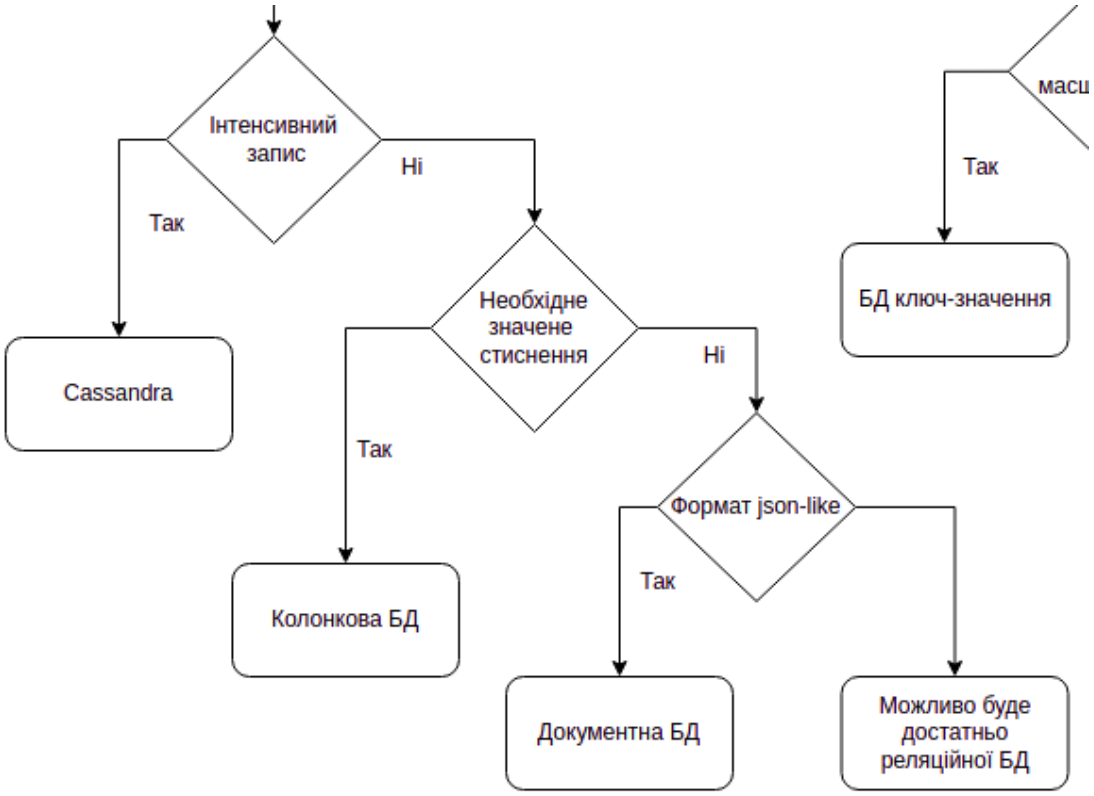

Рисунок 7 – Частина третя дерева рішень.

За відсутності умов інтенсивності запису переходимо на частини 3 дерева рішень (Рисунок 7). Розглядаємо необхідність продуктивного стиснення, тобто об’єм даних буде значним. В такому випадку варто розглянути колонкові бази даних, що пропонують хороше стиснення, оскільки записи зберігають стовпці, а не рядки даних. Прикладом для застосування таких даних можуть стати біомедичні дослідження, що продукують величезні обсяги даних.

Якщо ж стиснення не є настільки важливим, то задаємо питання про те, чи мають дані json-подібний формат. Якщо це так, то варто використовувати документну БД. Як приклад, моніторинг клінічних показників, або баз даних пацієнтів. Якщо ж дані не мають такого формату, то ймовірно, що для задоволення потреб буде достатньо скористатись звичайною реляційною базою даних, зважаючи також на відповіді на всі попередні запитання.

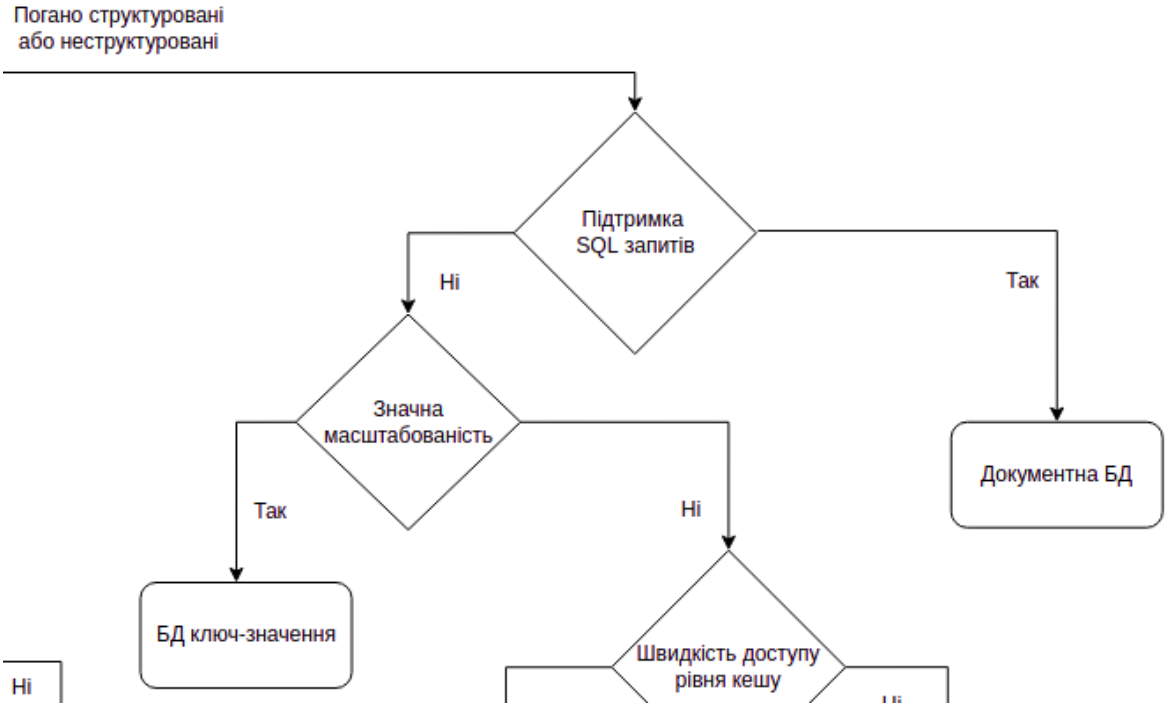

Рисунок 8 – Частина четверта дерева рішення.

Повертаючись, до частини 2, де на етапі структурованості інформації був обраний шлях структурованих даних(рисунок 8). Якщо ж дані мають неструктурований або напівструктурований вигляд, то розглядаємо важливість підтримки SQL-запитів, якщо така потреба є, то може бути використовувана документна БД, оскільки в основному вони надають підтримку запитів до вмісту документів. Тут можна розглянути приклад зберігання малоструктурованих медичних записів. Якщо ж такої необхідності немає, то варто визначити, наскільки значно може знадобитись масштабованість. Якщо це значення є важливим, то варто розглянути базу даних типу ключ-значення, оскільки такі бази мають здатність легко масштабуватись.

Якщо ж це не ключовий показник, розглянемо вимоги до швидкості доступу(рисунок 9). Коли необхідно отримувати якнайшвидше дані, то мова йде про швидкість доступу до оперативної пам’яті. В такому випадку можна використати Redis - базу даних типу ключ значення, що

розміщується в кеші. В іншому ж випадку можна застосовувати або БД типу ключ-значення або документну. Приклад використання таких баз - збереження даних про наявні лікарські засоби та пошук нових.

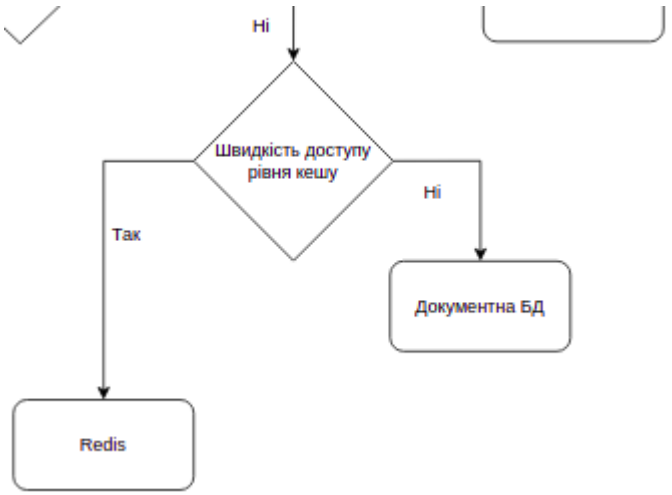

Рисунок 9 – Частина п’ята дерева рішень.

Для детальнішого аналізу було виділено набори переваг та недоліків для кожного типу сховища.

1) Бази даних ключ-значення
    - зберігання в форматі ключ-значення
    - немає підтримки sql запитів
    - широке горизонтальне масштабування
    - швидкість доступу (пошуку) через хеш-таблицю
    - неструктуровані дані
    - ключ і значення можуть мати різного роду формати
    - стандартні типи значень String, JSON, BLOB
    - погано працюють на оновлення
    - проблеми підтримки унікальних ключів
    - відсутність зв'язків між значеннями
2) Колонкові бази даних
    - орієнтовні на швидке зчитування

- добре підходять для аналітичних систем
- працюють із заданою схема в основному
- підтримують високе стиснення даних
- повільніше працюють на запис, ніж на читання
- високий рівень масштабованості
- простота модифікації схеми

3) Документна база даних
   - ієрархічна/деревоподібна система, що побудована на основі зв’язків ключ-значення
   - значення - документи, можуть мати певну структуру
   - XML, JSON або BSON - типові структури
   - можливість робити запити на основі вмісту документу
   - підтримують просте добавлення й оновлення
   - призначені для впорядкованого зберігання, але без зв'язків
   - відсутність схеми - неструктуровані або напівструктуровані
   - забезпечується індексація
   - непогана масштабованість
   - добре підтримують запити на читання великих об'ємів даних
4) Графові бази даних
   - орієнтовані на зв'язки між вузлами
   - зберігають властивості зв'язків і вузлів
   - використовуються коли необхідне швидке виявлення зв'язків
   - актуальний приклад - системи виявлення шахрайства, система рекомендацій
5) Розподілені файлові системи

- зберігають дані різних джерел, різного формату та природи
- добре підходять для зберігання “сирих даних” (до попередньої обробки)
- складність здійснення запитів, що можливі лише за підтримки певних додаткових інструментів
- легка масштабованість

6) Хмарні сервіси
   - забезпечують легкість у використанні, відмовостійкість
   - легкість розширення, постійний супровід
   - налаштування надійного захисту - сховище може бути максимально безпечним
   - можуть бути відносно дорогими

Висновки до розділу 4

Створена діаграма є лише рекомендацією і може не вказати точний тип необхідної бази даних, оскільки часто потрібно враховувати всі конкретні обставини, щоб прийняти найкраще рішення. Та все ж проаналізувавши запропоновану інформацію, що достатньо вичерпно описує переваги та недоліки, можна з високою ймовірністю прийняти правильне рішення, беручи до уваги всі деталі.

Вибір ефективного методу зберігання впливає не лише на кількість спожитих ресурсів та стабільність системи, а я на час доступу та можливістю керування даними. Покращена продуктивність доступу до даних сприяє кращій якості аналізу даних. А отже вибір оптимального сховища даних може стати ключовим в процесі досягнення поставленої мети.

5 РОЗРОБКА СТАРТАП ПРОЕКТУ

5.1 Ідея проекту

Ідея проекту полягає у розробці системи підтримки прийняття рішень про вибір найбільш ефективного методу зберігання великих обсягів даних на основі дерева ухвалення рішень. Сьогодні крім текстового порівняння різних систем зберігання даних, не знайдено ніяких програмних чи алгоритмічних реалізацій даної ідеї. Послідовність вичерпних ключових запитань приводить до правильного рішення без необхідності глибоко розбиратись в деталях роботи кожної системи. Таким чином підтримка рішень може бути реалізована як невеличкий хмарний застосунок у вигляді опитування.

У цьому розділі описано послідовність розробки стартап проекту на основі проведеного дослідження. Сутність ідеї проекту викладена в таблиці 4.

Таблиця 4 - Суть ідеї стартапу

Зміст ідеї	**Напрямки застосування**	**Переваги для користувача**
Система прийняття рішення про вибір найбільш ефективного методу зберігання великих даних	Вибір найефективнішого методу зберігання на основі відповідей на питання стосовно вимог до даних	Користувач може отримати рекомендації та порівняти їх із власними припущеннями
	Надання вичерпної інформації про недоліки, переваги та особливості певного методу зберігання та конкретного сховища даних	Крім отриманої рекомендації, можливість прийняти власне рішення на основі додаткових характеристик

Розглянемо уже наявні системи прийняття рішень стосовно методу зберігання даних. Порівняння техніко-економічних характеристик власного рішення з конкурентними наведено в таблиці 5.

Таблиця 5 - Нейтральні, слабкі та сильні характеристики ідеї додатку

№ п\п	Техніко-економічні характеристики	Потенційні конкуренти			W слабка сторона	N нейтральна сторона	S сильна сторона
		Мій проект	TDS	Medium			
1.	Інтерфейс користувача	+	-	-			+
2.	Інтерактивний підхід	+	-	-			+
3.	Безкоштовне розгортання в хмарі	+	-	-			+
4.	Додатковий набір рекомендацій з врахуванням особливостей	+	+	-			+

Як результат проведення аналізу слабких, нейтральних та сильних сторін системи, можна зробити висновок про її конкурентоспроможність та наявність всього необхідного функціоналу. Тому вона може протистояти конкурентам у даній галузі.

5.2 Технологічний аудит ідеї проекту

Обрані технології конкретного функціонала для реалізації проекту представлені у таблиці 6 в рамках технологічного аудиту ідеї проекту.

Таблиця 5.3 - Технологічний аудит проекту

№ п\п	Ідея проекту	Технології	Наявність технології	Доступність технології
1	Розгортання проекту в хмарі	Heroku	Наявна	Доступна безкоштовно
2	Зручний інтерфейс користувача	Python, HTML, CSS	Наявна	Доступна

Представлений набір технологій є наявним і доступним.

5.3 Аналіз ринкових можливостей

Для запуску стартап-проекту необхідно провести аналіз його ринкових можливостей, щоб визначити конкурентоспроможність і зацікавленість потенційних клієнтів. Щоб провести аналізу попиту, потрібно оцінити основних гравців, що є на ринку. Результати аналізу зображені в таблиці 7.

Таблиця 7 - Характеристика ринку стартап проекту

№ п\п	Показники ринку	Характеристика
1	Кількість головних гравців, од	< 10
2	Динаміка ринку (якісна оцінка)	На етапі стрімкого зростання
3	Наявність обмежень для входу (характер)	Потенційний клієнт - людина, що проектує систему зберігання великих даних
4	Специфічні вимоги до стандартизації та сертифікації	Відсутні
5	Середня норма рентабельності в галузі (або по ринку), %	Невідомо

Потенційні групи клієнтів та їх характеристики представлено у таблиці 8.

Таблиця 8 - Характеристика потенційних клієнтів

№ п\п	Формуюча ринок потреба	Цільова аудиторія	Відмінності у поведінці різних цільових груп	Вимоги споживачів до продукту
1	Різноманітність методів зберігання великих даних і потреба підбору для конкретного вибору	Інженери великих даних	Відсутні	Зручність використання, розширена пропозиція

Варто провести також аналіз середовища ринку та визначити фактори загроз. Виявлення можливих джерел та план реакції на загрози

дозволяє ефективно з ними боротись. Можливі загрози та шляхи їх вирішення описано в Таблиці 9.

Таблиця 9 - Аналіз загроз

№ п\п	Фактор	Зміст загрози	Можлива реакція компанії
1	Низький попит	Обмежена кількість користувачів	Добавлення нового функціоналу, що стосується різних сфер обробки даних
2	Недостатня функціональність	Невелика кількість функцій	Добавлення детальних характеристик

В таблиці 10 представлені фактори можливостей, що матимуть потенційно позитивний вплив на розвиток стартап проекту.

Таблиця 10 - Фактори можливостей

№ п\п	Фактор	Зміст можливості	Можлива реакція компанії
1	Розширення клієнтської бази	Після додавання додаткових функцій, додатком може зацікавитись більше людей	Розширення функціоналу
2	Розширення компетенції	Створення аналогічних систем з підбором необхідних користувачу функцій	Аналіз і розробка нових рішень на основі попереднього
3	Система навчання	Розробка тестування для інженерів, що хочуть перевірити свої навички на конкретних прикладах	Створення нового напрямку розвитку продукту

Проаналізувавши таблицю можна зробити висновок, що основною проблемою стартап-проекту є вузька функціональність і обмежена клієнтська база. Саме тому, спочатку варто розширювати функціональність та відкривати нові напрямки розвитку продукту.

Проведемо аналіз наявних пропозицій на ринку та виокремимо загальні характеристики, що наведено в таблиці 11.

Таблиця 11 - Аналіз конкуренції на ринку

Особливість конкурентного середовища	**В чому проявляється характеристика**	**Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)**
1. Вказати тип конкуренції	Монополія	Надання конкурентно спроможних послуг
2. За рівнем конкурентної боротьби	Національний	Співпраця з компаніями та дослідними центрами, що працюють в напрямку великих даних
3. За галузевою ознакою	Внутрішньогалузева	Безкоштовне користування
4. Конкуренція за видами товарів	Товарно-видова	Пропозиція нових функцій за пропозиціями клієнта
5. За характером конкурентних переваг	Нецінова	Підвищення лояльності до користувачів
6. За інтенсивністю	Марочна	Отримання імені на створених для користувачів проектах

Згідно з проведеним ступеневим аналізом конкуренції на ринку в ролі початкової стратегії можна обрати розширення функціоналу з метою отримання нових клієнтів.

У таблиці 12 показано аналіз умов конкуренції у галузі.

Таблиця 12 – Аналіз конкуренції в галузі за М. Портером

Складові аналізу	**Конкуренти**		**Постачальники**	**Клієнти**	**Товари-замінники**
	Прямі	**Потенційні**			
	Medium, TDS	Невідомі	-	Інженери систем аналітики великих даних	Відсутні
Висновки	Існують конкуренти, що надають відмінні від нашого рішення	Дані відсутні	Умови не диктуються постачальниками послуг	Розширення функціоналу призведе до збільшення клієнтської бази	Відсутні обмеження

Фактори конкурентоспроможності наведені в таблиці 13

Таблиця 13 - Обґрунтування факторів конкурентоспроможності

№ п\п	Фактор конкурентоспроможності	Обґрунтування
1	Масштабованість	Логіка побудови системи дозволяє легко її масштабувати залежно від потреби
2	Адаптивність	Зумовлено гнучкістю системи та забезпечує підлаштування під вимоги ринку

За отриманими факторами конкурентоспроможності було проведено аналіз сильних та слабких сторін стартап проекту. Вони наведені в таблиці 14.

Таблиця 14 - Порівняльний аналіз сильних та слабких сторін

№ п\п	Фактор конкурентоспроможності	Бали 1-20	Рейтинг товарів конкурентів у порівнянні з нашим						
			-3	-2	-1	0	+1	+2	+3
1	Масштабованість	10	+						
2	Адаптивність	17		+					

На основі виділених ринкових загроз та можливостей, сильних та слабких сторін, було підготовано SWOT-аналіз стартап проекту, і наведено в таблиці 15.

Таблиця 15 - SWOT-аналіз стартап проекту

Сильні сторони: Адаптивність, масштабованість, зручний інтерфейс, простота використання, хмарне розгортання проекту	**Слабкі сторони:** Низький попит, недостатня функціональність
Можливості: Розширення функціональності з подальшим розширенням клієнтської бази	**Загрози:** Низька зацікавленість

На основі сильних сторін та можливостей проекту, наведених у матриці SWOT-аналізу, можемо зробити висновок, що даний стартап-проект є конкурентоспроможним і має право на життя.

Таблиця 16 - Альтернативи ринкового впровадження стартап-проекту

№ п\п	**Альтернатива ринкової поведінки**	**Ймовірність отримання ресурсів**	**Строки реалізації**
1	Розробка системи прийняття рішення, що стосується інших технологій	80%	1.5 місяці
2	Розробка системи без використання дерева ухвалення рішення	50%	2 місяці

Згідно з таблицею 16, найбільш перспективною є перша альтернатива, адже вона передбачає більшу вірогідність отримання ресурсів і потребує менше часу на реалізацію.

5.4 Розробка ринкової стратегії проекту

Для розробки ринкової стратегії, необхідно визначити охоплення ринку через визначення цільових груп потенційних споживачів, що наведено в таблиці 17.

Таблиця 17 - Вибір цільових груп потенційних користувачів

№ п\п	**Опис профілю цільової групи потенційних клієнтів**	**Готовність споживачів сприйняти продукт**	**Орієнтовний попит в межах цільової групи**	**Інтенсивність конкуренції в сегменті**	**Простота входу в сегмент**
1	Компанії що займаються аналітикою великих даних	Низька	Високий	Низька	Середня
2	Дослідницькі програми	Висока	Високий	Низька	Середня
Обрано цільові групи компаній, що займаються аналітикою великих даних та дослідницьких програм					

Розглянемо базову стратегію розвитку для стартап-проекту, що наведена в таблиці 18.

Таблиця 18 - Визначення базової стратегії розвитку

№ п\п	Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції відповідно до обраної альтернативи	Базова стратегія розвитку
1	Стратегія добавлення спеціалізації	Утримання існуючих клієнтів, пошук потенційних клієнтів, можливість розвитку системи згідно з побажаннями клієнтів	Створення якісного продукту і його постійне вдосконалення, безкоштовне використання	Оперативне оновлення продукту

В таблиці 19 визначено стратегію конкурентної поведінки на ринку.

Таблиця 19 - Визначення базової стратегії конкурентної поведінки

№ п\п	Чи є проект «першопрохідцем» на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента	Стратегія конкурентної поведінки
1	Так	Пошук нових споживачів	Ні	Зайняття конкурентної ніші

Розглянемо список визначених стратегій:

- базова стратегія розвитку – оперативне оновлення продукту;
- альтернативна стратегія розвитку – добавлення спеціалізації;
- конкурентна поведінка – заняття конкурентної ніші;
- сегменти ринку – компанії що займаються аналітикою великих даних, дослідницькі програми.

5.5 Розробка маркетингової програми

Першим кроком розробки маркетингової програми стартап-проекту є формування ключових переваг маркетингової концепції товару, який отримає споживач - зазначені у таблиці 20.

Таблиця 20 - Визначення ключових переваг концепції потенційного товару

№ п\п	Потреба	Вигода, яку пропонує продукт	Ключові переваги перед конкурентами
1	Простото використання	Легкість прийняття обґрунтованого рішення на основі низки вимог	Алгоритмічний підхід
2	Здатність до розширення	Легкість у добавленні нових напрямків	Легка адаптованість до вимог ринку
3	Малі затрати	Система розгортається у хмарі безкоштовно	Додаток є безкоштовним

Розроблено трирівневу модель товару, наведену в таблиці 21.

Таблиця 21 - Трирівнева модель товару

Рівні товару	Сутність та складові	
I. Товар за задумом	Система підтримки прийняття рішення для підбору необхідного набору технологій.	
II. Товар у реальному виконанні	Властивості/характеристики	Розмір
	Алгоритм підбору	10 КБ
	Інтерфейс користувача	15 МБ
	Якість: тестування застосунку відповідно до сценаріїв	
	Пакування: веб-застосунок, що відкривається за допомогою веб-браузера	
	Марка: без марки	
III. Товар із підкріпленням	Надання додаткових напрямків згідно з побажаннями клієнтів..	
За рахунок чого потенційний товар буде захищено від копіювання: захист інтелектуальної власності		

Визначення цінових меж представлено в таблиці 22.

Таблиця 22 – Визначення меж встановлення ціни

№	**Рівень цін на товари-замінники**	**Рівень цін на товари-аналоги**	**Рівень доходів цільової групи споживачів**	**Верхня та нижня межі встановлення ціни на товар**
1	Відсутні товари-замінники	Безкоштовні	Високий	10-50

Далі визначимо оптимальної системи збуту в таблиці 23:

- проводити збут власними силами або залучати сторонніх посередників (власна або залучена система збуту);
- вибір та обґрунтування оптимальної глибини каналу збуту;
- вибір та обґрунтування виду посередників.

Таблиця 23 - Формування системи збуту

№	**Специфіка закупівельної поведінки цільових клієнтів**	**Функції збуту, які має виконувати постачальник товару**	**Глибина каналу збуту**	**Оптимальна система збуту**
1	Покупка підписки на кількість використань	Вдосконалення, підтримка, обробка відгуків, подальший розвиток	Канал однорівневий	Вертикальна (право власності залишається у розробника)

Останньою частиною маркетингової програми є розробка концепції маркетингових комунікацій, вона наведена в таблиці 24.

Таблиця 24 - Концепція маркетингових комунікацій

№	**Специфіка поведінки цільових клієнтів**	**Канали комунікацій, якими користуються цільові клієнти**	**Ключові позиції, обрані для позиціонування**	**Концепція рекламного звернення**
1	Мінімізація витрат	Конференції, Інтернет ресурси	Продаж підписки	Зменшення часових затрат на пошук ефективного рішення

5.6 Висновки до розділу 5

В даному розділі було розроблено стартап-проект на основі здійснених досліджень. По-перше, було сформульовано ідею стартап-проекту, що зможе бути цікавою для ринку. Було проведено пошук та аналіз конкурентів. На основі отриманих результатів було зазначено, що стартап-проект має значні переваги в порівнянні з конкурентами та є конкурентоспроможним на ринку.

Наступним кроком було визначено технології для розробки проекту та їх наявність і доступність. На основі цього можна зробити висновок, що всі необхідні технології доступні для використання. Після проведення аналізу ринкової стратегії стало зрозуміло, що цільовою аудиторією є компанії, що займаються розробкою проектів на основі аналітики великих даних. Було визначено фактори ризиків та побудовано сценарії реагування, крім того, було ідентифіковано фактори можливостей для розвитку проекту.

На останньому етапі було сформульовано маркетингову стратегію виходу проекту на ринок. Для цього було зазначено переваги, що їх отримає клієнт через використання нашого продукту, канали комунікації із клієнтами, канали збуту та ключові переваги над конкурентами. Також було визначено цінові границі для продажу.

ВИСНОВКИ

В ході проведення дослідження було проаналізовано актуальність застосування методів великих даних у сфері медицини. Хоча вітчизняний сектор охорони здоров'я не готовий до активного впровадження аналітики великих даних, у світі, зокрема у США, є значний досвід успішного використання цих технологій для покращення якості медицини. За допомогою обробки медичних даних визначають непомітні раніше закономірності та зв'язки, що допомагають виявляти хвороби, передбачають завантаженість лікарень, забезпечують ранню ідентифікацію захворювань та персоналізоване лікування, створюють системи підтримки прийняття рішень та обробки медичних зображень, передбачають спалахи епідемій.

З метою отримання розуміння про сфери застосування обробки великих даних та природу даних у цих сферах було проаналізовано 8 досліджень та узагальнено результати у таблиці. Крім того, визначено, що за допомогою впровадження методів великих даних можна отримати численні переваги: скорочення витрат на медицину, кращий досвід лікування для пацієнта, розширення бази медичних знань, спрощення процесу проведення досліджень у сфері.

Незважаючи на очевидні перспективи, існують також численні перешкоди, з якими доводиться боротись, щоб отримати очікувані результати. В основному ці перепони пов'язані з різноманітністю форматів даних, які потребують обробки та структуризації, що характерно саме для медичних даних, адже в одному сховищі можуть зберігатись медичні записи лікарів, електронні карточки здоров'я пацієнтів, дані з медичних пристроїв, рентгенівські чи офтальмологічні знімки. Одним із ключових питань, що потребують рішення є проблема зберігання та управління даними, а також їх очищення та стандартизація. Критично важливою

залишається безпека даних, оскільки медичні записи часто включають особисту інформацію пацієнта.

В контексті цієї роботи було детально розглянуто проблему зберігання великих об’ємів медичних даних. В процесі було проаналізовано пов’язані дослідження з метою визначення найкращих способів зберігання цих даних. Для аналізу опрацьовано 6 статей, а узагальнений висновок подано у формі таблиці, що включає конкретні способи зберігання даних та згадки про них у кожній зі статей.

В процесі дослідження було детально розглянуто такі методи зберігання даних, як розподілені файлові системи, хмарні сховища, нереляційні бази даних, а саме колонко-орієнтовані, документно-орієнтовані, ключ-значення та графові, а також їх переваги та недоліки й проблеми, з якими можна зіткнутись при використанні конкретного методу. Також було проаналізовано, які саме імплементації кожного зі сховищ є найчастіше та найефективніше використовуваними і їх особливості.

Як результат аналізу було створено прототип дерева прийняття рішень для вибору найкращого способу зберігання даних залежно від характеристик самих даних та детально описано логіку, що пояснює кожен із кроків. Дане рішення може не мати стовідсоткової точності, але надає хороший аналітичний інструмент, що допомагає здійснити правильний вибір та зрозуміти переваги, недоліки та особливості використання певних засобів. Звичайно, не можна забувати, що в кожному випадку необхідно відштовхуватись від обставин та обмежень, що можуть бути додатковими факторами при прийнятті рішення. Крім того, варто звернути увагу на те, що в більшості випадків доцільним є використання хмарних платформ, що пропонують свої ефективні інструменти для роботи з великими об’ємами даних, зокрема медичних. Вони надають переваги легкості управління, масштабування, моніторингу та використання ресурсів.

Отже, завдання визначення ефективних методів зберігання великих об’ємів медичних даних в умовах їх неперервного зростання було виконано шляхом створення алгоритму підтримки прийняття рішення для використання найбільш відповідних способів зберігання даних. Отриманий прототип дерева прийняття рішень можна реалізувати за допомогою однієї із мов програмування, щоб створити зручний засіб для рекомендації найбільш відповідного інструмента у вирішенні поставленої задачі.

Проте в процесі розробки, в основному, було враховано особливості узагальнених методів, які все ж можна покращити після тестування. Для подальшої роботи в обраному напрямку можна також спробувати розробити більш точну систему, яка буде також враховувати переваги та недоліки конкретних баз даних та інших інструментів. На додачу, окрім зберігання даних, важливим є і спосіб доступу до цих даних, наприклад за допомогою інструментів, що можуть працювати з напівструктурованими даними шляхом здійснення SQL-подібних запитів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. 5 Best Big Data Databases [Електронний ресурс] – Режим доступу до ресурсу: https://www.scnsoft.com/analytics/big-data/databases. (дата звернення: 06.12.2021)
2. A Systematic Review of Techniques and Sources of Big Data in the Healthcare Sector / [S. Alonso, I. Díez, J. Rodrigues та ін.]. // Journal of Medical Systems. – 2017. – №41.
3. Alexandru A. Big Data in Healthcare - Opportunities and Challenges / A. Alexandru, I. Radu, M. Bizon. // Informatica Economica. – 2019. – №22. – С. 43–45.
4. Big data in healthcare: management, analysis and future prospects / S.Dash, S. Shakyawar, M. Sharma, S. Kaushik. // Journal of Big Data. – 2019. – №54.
5. Challenges and Opportunities of Big Data in Health Care: A Systematic Review / C.Kruse, R. Goswamy, Y. Raval, S. Marawi. // JMIR Med Inform. – 2016. – №4.
6. Drake M. Understanding MongoDB: Advantages of a Document-Oriented NoSQL Database [Електронний ресурс] / Mark Drake. – 2021. – Режим доступу до ресурсу: https://www.digitalocean.com/community/conceptual_articles/understanding-mongodb-advantages-of-a-document-oriented-nosql-database. (дата звернення: 06.12.2021)
7. Health Data Management: Benefits, Challenges and Storage [Електронний ресурс] – Режим доступу до ресурсу: https://cloudian.com/guides/hipaa-compliant-cloud-storage/health-data-management/. (дата звернення: 20.11.2021)
8. Hermon R. Big data in healthcare: What is it used for? [Електронний ресурс] / R. Hermon, P. Williams // AUSTRALIAN EHEALTH INFORMATICS AND SECURITY CONFERENCE. – 2014. – Режим доступу до ресурсу: https://ro.ecu.edu.au/aeis/22/. (дата звернення: 23.11.2021)
9. Kayay C. How do you pick the right database for Big Data Architecture? [Електронний ресурс] / Cengiz Kayay. – 2018. – Режим доступу до ресурсу: https://medium.com/@ckayay/how-to-pick-the-right-database-c2539efe2589. (дата звернення: 25.11.2021)
10. Killgore, CPHIMS, Ethan. (2015). The Implementation and Usage of NoSQL Database Products in the Healthcare Domain: A Survey. 10.13140/RG.2.1.4566.2806.
11. Kumar S. Big data analytics for healthcare industry: impact, applications, and tools / S. Kumar, M. Singh. // Big Data Mining and Analytics, vol. 2. – 2019. – №1. – С. 48–57.
12. Olaronke, I., Oluwaseun, O. (2016). Big data in healthcare: Prospects, challenges and resolutions. 2016 Future Technologies Conference (FTC). doi:10.1109/ftc.2016.7821747

13. Pandey M. K. A Novel Storage Architecture for facilitating Efficient Analytics of Health Informatics Big Data in Cloud / M. K. Pandey, K. Subbiah. // International Conference on Computer and Information Technology (CIT). – 2016. – №1. – С. 578–585.
14. Raghupathi W. Big data analytics in healthcare: promise and potential / W. Raghupathi, V. Raghupathi. // Health Information Science and Systems. – 2014. – №2.
15. Raja R. A Systematic Review of Healthcare Big Data / R. Raja, I. Mukherjee, B. Sarkar. // Scientific Programming. – 2020.
16. Sandra Durcevic. 18 Examples Of Big Data Analytics In Healthcare That Can Save People [Електронний ресурс] / Sandra Durcevic. – 2020. – Режим доступу до ресурсу: https://www.datapine.com/blog/big-data-examples-in-healthcare/. (дата звернення: 02.12.2021)
17. Siddiqa A. Big data storage technologies: a survey / A. Siddiqa, A. Karim, A. Gani. // Frontiers Inf Technol Electronic Eng. – 2017. – №18. – С. 1040–1070.
18. Tomar D., Bhati J. & Tomar, Pradeep & Kaur, Gurjit. (2019). Migration of healthcare relational database to NoSQL cloud database for healthcare analytics and management. 10.1016/B978-0-12-815368-0.00002-6.
19. Top 9 column-oriented databases [Електронний ресурс] – Режим доступу до ресурсу: https://www.predictiveanalyticstoday.com/top-wide-columnar-store-databases/. (дата звернення: 28.11.2021)
20. Vuppala S. K. Top 20 Latest Research Problems in Big Data and Data Science [Електронний ресурс] / Sunil Kumar Vuppala. – 2020. – Режим доступу до ресурсу: https://towardsdatascience.com/top-20-latest-research-problems-in-big-data-and-data-science-c6fb51e03136. (дата звернення: 28.11.2021)
21. What Is a Document Database? [Електронний ресурс] – Режим доступу до ресурсу: https://aws.amazon.com/nosql/document/?nc1=h_ls.(дата звернення: 23.11.2021)
22. What Is a Graph Database? [Електронний ресурс] – Режим доступу до ресурсу: https://aws.amazon.com/nosql/graph/?nc1=h_ls. (дата звернення: 23.11.2021)
23. What Is a Key-Value Database? [Електронний ресурс] – Режим доступу до ресурсу: https://aws.amazon.com/ru/nosql/key-value/. (дата звернення: 23.11.2021)
24. Wu J. Choosing The Right Database [Електронний ресурс] / Jun Wu. – 2019. – Режим доступу до ресурсу: https://towardsdatascience.com/choosing-the-right-database-c45cd3a28f77. (дата звернення: 01.12.2021)
25. Вичугова А. Cassandra [Електронний ресурс] / Анна Вичугова. – 2019. – Режим доступу до ресурсу: https://www.bigdataschool.ru/wiki/cassandra. (дата звернення: 06.12.2021)

26. Колоночные СУБД — принцип действия, преимущества и область применения [Електронний ресурс]. – 2011. – Режим доступу до ресурсу: https://habr.com/ru/post/95181/. (дата звернення: 23.11.2021)
27. Преимущества и недостатки нереляционных баз данных [Електронний ресурс]. – 2018. – Режим доступу до ресурсу: https://veesp.com/ru/blog/sql-or-nosql/. (дата звернення: 20.11.2021)
28. Прияцелюк Н. Разбираемся в типах NoSQL СУБД [Електронний ресурс] / Никита Прияцелюк. – 2018. – Режим доступу до ресурсу: https://tproger.ru/translations/types-of-nosql-db/. (дата звернення: 23.11.2021)
29. Сабініч А. Як Big Data допомагають сучасній медицині? [Електронний ресурс] / Андрій Сабініч. – 2018. – Режим доступу до ресурсу: https://tokar.ua/read/28563 (дата звернення: 20.11.2021)