DOI 10.31891/2307-5732-2024-337-3-8
УДК 004.773.2

ДАНЧАК ОРЕСТ
Національний університет «Львівська політехніка»
https://orcid.org/0009-0004-8060-1857
e-mail: orest.i.danchak@lpnu.ua
ВОЙТЮК АНДРІЙ
Національний університет «Львівська політехніка»
https://orcid.org/0009-0007-6983-0378
e-mail: andrii.a.voitiuk@lpnu.ua

ЕФЕКТИВНІ СХОВИЩА ДАНИХ ДЛЯ РІШЕНЬ МАШИННОГО НАВЧАННЯ

Усвідомлення важливості використання ефективних сховищ даних для розробки рішень машинного навчання в сучасному світі стає все більш актуальним у зв'язку зі зростанням обсягів даних і зростаючою потребою в точних і продуктивних моделях. У статті визначено основні життєво важливі вимоги до систем зберігання даних для забезпечення ефективного управління даними при створенні рішень машинного навчання.

Дослідження підкреслює практичну необхідність швидкого доступу до даних і мінімальної затримки, надійних систем зберігання з надійними механізмами резервного копіювання та відновлення, високого рівня безпеки, гнучкості в обробці різних типів даних і економічної ефективності. Ці фактори є не лише теоретичними міркуваннями, але безпосередньо впливають на ефективність і успіх рішень машинного навчання.

Разом із описом концепцій блокового сховища та сховища об'єктів у статті також порівнюються основні хмарні платформи — AWS (Amazon Web Services), Microsoft Azure та Google Cloud Platform (GCP), — кожна з яких надає широкий спектр гнучких і настроюваних сховищ, безпеки , а також послуги обробки даних. Ці платформи дозволяють користувачам гнучко й ефективно керувати своїми інформаційними ресурсами, пропонуючи унікальні функції, які задовольняють конкретні потреби програм машинного навчання. Розуміння нюансів пропозицій кожного постачальника дає змогу інженерам і дослідникам приймати обґрунтовані рішення, які відповідають їхнім унікальним вимогам.

У статті наведено детальний огляд ключових міркувань і доступних варіантів, які скеровують стратегічні рішення для ефективної підтримки ініціатив машинного навчання. Також підкреслено важливість бути в курсі еволюції пропозицій найбільших платформ. Підсумовуючи, вибір відповідного рішення для зберігання даних має вирішальне значення для підвищення продуктивності, безпеки та економічності моделей машинного навчання.

Ключові слова: сховища даних, рішення AI, хмарні сховища даних, хмарні платформи.

DANCHAK OREST, VOITIUK ANDRII
Lviv Polytechnic National University

EFFECTIVE DATA WAREHOUSES FOR MACHINE LEARNING SOLUTIONS

Awareness of the importance of using effective data warehouses to develop machine learning solutions in today's world is becoming increasingly relevant in connection with the growth of data volumes and the growing need for accurate and productive models. The article identifies major vital requirements for data storage systems to ensure effective data management in building machine learning solutions.

The study underscores the practical implications of rapid data access and minimal latency, reliable storage systems with robust backup and recovery mechanisms, high-level security, flexibility in handling different data types, and cost-effectiveness. These factors are not just theoretical considerations but directly impact the efficiency and success of machine learning solutions.

Together with describing the concepts of block storage and object storage, the article also compares major cloud platforms—AWS (Amazon Web Services), Microsoft Azure, and Google Cloud Platform (GCP)—each providing a comprehensive range of flexible and customizable storage, security, and data processing services. These platforms enable users to manage their information resources flexibly and efficiently, offering unique features that cater to the specific needs of machine learning applications. Understanding the nuances of each provider's offerings empowers engineers and researchers to make informed decisions that align with their unique requirements.

The article provides a thorough overview of the key considerations and available options, guiding strategic decisions to support machine learning initiatives effectively. Also the importance of staying up to date with the evolution of major platforms offerings is emphasized. In conclusion, selecting the proper data storage solution is critical for enhancing the performance, security, and cost-efficiency of machine learning models.

Keywords: data warehouses, AI solutions, cloud data storage, cloud platforms.

Постановка проблеми

Інформація і дані у сучасному світі виступають ключовим ресурсом, а здатність швидко та ефективно опрацьовувати величезні обсяги інформації є необхідною умовою успіху розробки рішень в галузі машинного навчання. Машинне навчання (МН) вимагає не лише передових алгоритмів, а й потужних систем зберігання та обробки даних, які можуть задовольнити високі вимоги щодо обсягу, швидкості доступу та безпеки. Вивчення різноманітних архітектур сховищ даних, аналіз їхніх переваг і недоліків, а також інтеграція з платформами машинного навчання є дуже важливими для вибору найкращого рішення для конкретних потреб. Від класичних реляційних баз даних до сучасних хмарних сховищ і технологій розподіленого зберігання — кожен з них має свої унікальні можливості та виклики.

Сучасні підходи до зберігання даних часто стикаються з численними проблемами, серед яких масштабованість, продуктивність, обробка великих обсягів даних у режимі реального часу, а також забезпечення

належного рівня безпеки та конфіденційності інформації. Ці виклики стають особливо значущими в середовищах, де обробка великих масивів даних та їх аналіз є критично важливими для ухвалення рішень.

Метою цього дослідження є ідентифікація та аналіз ключових характеристик ефективних сховищ даних, здатних задовольнити потреби сучасних рішень машинного навчання. Особлива увага приділяється вивченню існуючих архітектур і технологій сховищ даних, що пропонуються провідними постачальниками хмарних обчислень, такими як Amazon Web Services (AWS), Microsoft Azure і Google Cloud Platform (GCP). Це включає аналіз їхніх переваг і недоліків, а також визначення нових методів, які можуть підвищити їхню ефективність..

Завданнями даного дослідження є:

1. Визначення вимог до сховищ даних для обробки великих обсягів інформації, що використовуються в машинному навчанні.
2. Аналіз існуючих технологій сховищ даних від провідних хмарних платформ AWS, Azure, GCP
3. Визначення критеріїв оцінки ефективності сховищ даних у контексті рішень машинного навчання.
4. Пропозиція покращень або нових підходів до організації сховищ даних.

Вирішення цих проблем сприятиме підвищенню ефективності моделей машинного навчання за рахунок більш якісного та швидкого доступу до даних, що в свою чергу, може значно покращити результати аналітичних процесів та прийняття рішень на основі даних.

Аналіз останніх джерел

Інтеграція алгоритмів машинного навчання у хмарні сховища даних дозволяє оптимізувати їх продуктивність. Було розглянуто ключові аспекти, такі як оптимізація запитів, індексація та автоматизоване управління даними. Використання алгоритмів машинного навчання дозволяє знизити затримки, покращити оптимізацію запитів і управління ресурсами. Що в свою чергу дозволяє автоматично оптимізувати різні аспекти обробки даних, та значно підвищити ефективність роботи системи. Інтеграція машинного навчання у сховища даних є трансформаційним процесом, що вирішує багаторічні проблеми та відкриває нові можливості для оптимізації продуктивності [1].

Були проведені дослідження щодо викликів та стратегії управління ресурсами в хмарних обчислювальних системах, використовуючи штучний інтелект (ШІ). Основна увага приділяється автоматизації виділення ресурсів, інтелектуальному плануванню завдань, прогнозованому обслуговуванню та енергоефективному управлінню ними. Використання прогностичної аналітики для управління навантаженням допомагає передбачати майбутні потреби в ресурсах та автоматично масштабувати систему відповідно до змін попиту. Алгоритми ШІ можуть виявляти аномалії та потенційні загрози безпеці, що дозволяє знижувати ризики та забезпечувати стійкість системи. Важливо враховувати етичні питання, конфіденційність, упередженість алгоритмів та інтерпретованість для забезпечення справедливості та прозорості [2].

Хмарні сховища даних забезпечують високу масштабованість, дозволяючи обробляти великі обсяги даних, що є критично важливим для рішень машинного навчання. Використання різних платформ, дозволяє легко масштабувати сховища даних відповідно до потреб. Інтеграція хмарних сховищ даних та машинного навчання забезпечує значний прогрес у сфері управління даними та їх використання. Поєднання масштабованості та стабільності хмарних сховищ з інтелектуальними можливостями машинного навчання дозволяє організаціям досягати більш ефективних та інтелектуальних бізнес-операцій і прийняття рішень у різних галузях [3].

На додаток, були проведені дослідження щодо можливості використання Google BigQuery для створення та розгортання моделей машинного навчання за допомогою SQL-запитів. Це дозволяє аналітикам даних та спеціалістам з машинного навчання будувати моделі на великих наборах даних без необхідності глибоких знань у програмуванні або статистики. Це робить BigQuery інструментом, який значно полегшує процес створення та впровадження моделей машинного навчання для бізнес-аналітиків та інших користувачів. BigQuery ML автоматизує багато аспектів процесу машинного навчання, включаючи попередню обробку даних, вибір моделі та налаштування гіперпараметрів. Це зменшує час і зусилля, необхідні для підготовки даних та навчання моделей, дозволяючи швидко отримувати результати. BigQuery підтримує різні алгоритми машинного навчання, такі як лінійна регресія, логістична регресія, кластеризація K-means та аналіз головних компонент. Це дозволяє використовувати BigQuery для різних завдань, від прогнозування до класифікації та сегментації даних [4].

Також розглядається роль великих технологічних компаній, таких як Amazon, Microsoft та Google, у застосуванні штучного інтелекту для різних галузей. Основні технологічні компанії забезпечують фундаментальну інфраструктуру для розвитку AI, включаючи обчислювальні потужності, зберігання даних та інструменти для обробки даних. Amazon Web Services (AWS), Microsoft Azure та Google Cloud Platform (GCP) є основними гравцями у цій сфері. Вони також інтегрують штучний інтелект у свої хмарні платформи пропонуючи сервіси для машинного навчання. Залежність від інфраструктури великих технологічних компаній створює нові виклики та ризики, такі як ризики безпеки даних, монополізація ринку та висока вартість переходу на інші платформи. Це також підвищує важливість питань етики та регулювання у сфері штучного інтелекту та машинного навчання [5].

Було проведено аналіз викликів пов’язаних з великими обчислювальними потребами та об’ємом даних для зберігання що використовуються для машинного навчання. Хмарні обчислювальні платформи, такі як AWS, Google Cloud та Microsoft Azure, забезпечують масштабованість і гнучкість, необхідні для обробки великих обсягів даних і виконання складних моделей машинного навчання. Вони надають інфраструктуру, яка дозволяє легко збільшувати чи зменшувати обчислювальні ресурси в залежності від потреб проекту. Хмарні обчислювальні платформи забезпечують автоматизацію управління ресурсами, що включає розподіл обчислювальних ресурсів, балансування навантаження та оптимізацію використання ресурсів. Масштабовані сховища даних дозволяють

зберігати і обробляти великі набори даних, забезпечуючи високу швидкість доступу та обробки інформації. Ефективні сховища даних для рішень машинного навчання значною мірою залежать від можливостей хмарних обчислювальних платформ. Інтеграція хмарних технологій з алгоритмами машинного навчання забезпечує масштабованість, продуктивність і гнучкість, необхідні для обробки великих обсягів даних та виконання складних моделей. Автоматизація управління ресурсами та оптимізація витрат є ключовими перевагами використання хмарних платформ для рішень машинного навчання [6, 7].

Виклад основного матеріалу

В епоху стрімкого розвитку технологій машинного навчання існує необхідність забезпечити ефективне зберігання та обробку великих обсягів даних. Це вимагає ґрунтовного підходу до вибору сховищ даних, які можуть підтримувати специфічні потреби цих технологій. Настуні вимоги до сховищ даних допоможуть організаціям забезпечити оптимальну роботу своїх моделей машинного навчання, підвищивши їхню продуктивність, безпеку та економічну ефективність

1. Масштабованість: можливість легко змінювати розмір сховища без впливу на продуктивність роботи. Має бути передбачено можливості як горизонтально збільшувати доступність ресурсів, так і вертикально модернізовувати потужності. Також сховище має адаптуватися до ситуації відповідно до зростання чи спадання обсягів даних та вимог до їх обробки у автоматичний спосіб, без адміністративного втручання.
2. Продуктивність і швидкодія: висока швидкість доступу до даних і мінімальна затримка є іншими важливими чинниками успішності програмного рішення. Швидкий доступ до даних забезпечує ефективне навчання та тестування моделі, що, у свою чергу, прискорює етапи розробки і впровадження. Мінімізація часових витрат особливо важлива для задач у режимі реального часу.
3. Надійність і відмовостійкість: надійні сховища даних повинні містити механізми резервного копіювання та відновлення, такі як автоматичні резервні копії та миттєві знімки, що забезпечують відмовостійкість та запобігають втратам інформації. Збої обладнання чи програмні помилки не повинні мати вплив на стабільність рішення. Безперебійна доступність до даних є ключовою вимогою високонавантажених систем.
4. Безпека: сховища даних повинні забезпечувати шифрування інформації і надавати засоби захисту від несанкціонованого доступу. Деталізація і гнучкість є вимогою до систем управління доступом. Сучасні системи машинного навчання часто працюють із захищеними або конфіденційними даними та мусять відповідати національному та міжнародному законодавству про їх захист. Вимоги до безпеки також передбачають дотримання галузевих нормативних вимог, протоколів і стандартів.
5. Гнучкість у зберіганні типів даних: має бути надана можливість зберігати та обробляти різні типи інформації - структуровані, неструктуровані та напівструктуровані дані. Це означає необхідність у підтримці можливостей інтеграції з різноманітними інструментами і фреймворками машинного навчання через програмні інтерфейси. Навчання моделей машинного навчання часто включає роботу з різними типами даних, такими як текст, зображення, аудіо та відео. Відповідно, система зберігання даних повинна мати можливість обробляти різні формати даних без зниження продуктивності.
6. Вартість: економічна доцільність також сильно впливає на вибір сховища даних. Оптимізація витрат є важливою складовою будь якого програмного рішення, особливо з неперервним збільшенням доступних для опрацювання даних. Хмарні платформи прагнуть простоти і прозорості ціноутворення та надають вбудовані сервіси аналізу ефективності залучених ресурсів.
7. Підтримка великих обсягів даних (Big Data): сховища даних повинні забезпечувати можливість інтеграції з механізмами великих даних, а також підтримувати обробку потокових даних і режим реального часу. Обробка великих даних часто включає виконання складних аналітичних задач, які вимагають великих обчислювальних ресурсів і ефективних методів управління даними. Інтеграція з технологіями Big Data забезпечує організаціям можливість використовувати сучасні методи аналізу та обробки даних, що значно підвищує продуктивність моделей машинного навчання.
8. Взаємодія та інтеграція: сумісність і інтеграційна сумісність роблять сховища даних привабливими для використання їх у рішеннях МН, оскільки забезпечують гнучкість у виборі технологічного стека реалізації та розширюють можливості подальшого розвитку і розширення такого рішення. Також, така сумісність полегшує перехід між різними платформами та легкість застосування технологічних новинок.

На сьогоднішній день світовий ринок виділяє трьох лідерів серед хмарних платформ, що використовують сервісних підхід та надають широкі можливості для зберігання, охорони та обробки даних. Це всім відомі Amazon Web Services (AWS), Microsoft Azure та Google Cloud Platform (GCP). Кожна з перелічених платформ має свої унікальні властивості та переваги і недоліки. Розуміння відмінностей і специфіки наявних пропозицій, необхідні для приймати обґрунтовані рішення відповідно до вимог рішення МН, є достатньо складним. Розуміючи важливість прийняття зваженого рішень, ми пропонуємо зупинитися на найважливіших аспектах доступних сервісів.

Для початку зупинимося на концептуальних типах сховищ даних: блокове зберігання та об'єктне зберігання

Блокове зберігання: є методом, де дані розділяються на блоки з унікальними ідентифікаторами та зберігаються як незалежні об'єкти на сервері. Цей тип зберігання дозволяє ефективно розподіляти блоки даних по

різних місцях у хмарі, покращуючи продуктивність системи. Блокове зберігання особливо ефективне для управління великими обсягами даних з вимогами до низької затримки, що робить його ідеальним для високопродуктивних робочих завдань, зокрема міцних баз даних. Прикладами застосування блокового зберігання є сховище баз даних та сховище серверів, оскільки при такому підході гарантується швидкодія, продуктивність, надійність та універсальність через розподіл даних по різних томах, особливо у рішеннях для зберігання серверів. Легкість створення та форматування томів з блоками робить їх ідеальними для зберігання на задньому плані у віртуалізованих системах.

Об'єктне зберігання: це фундаментальна архітектура, спеціально розроблена для широких колекцій неструктурованих даних. Ця система розглядає окремі елементи даних як окремі об'єкти, кожен з яких зберігається в окремих репозитаріях разом з відповідною метаданою та унікальним ідентифікатором. Цей систематичний підхід дозволяє легкий доступ до даних та їх відновлення, покращуючи ефективність управління великими обсягами неструктурованої інформації. Приклади застосування об'єктного зберігання є рішення у сфері Інтернет Речей (Internet of Things), сховища електронної пошти чи резервного копіювання, позаяк забезпечена можливість швидко розширюватися та легко отримувати доступ до даних, а також пріоритетом є надійність перед продуктивністю.

Вибір між блоковим і об'єктним зберіганням залежить від специфіки даних якими оперуватиме рішення та управлінні ними. Об'єктне зберігання відрізняється відкритими можливостями для зберігання неструктурованих даних, таких як мультимедійні файли, резервні копії та архіви, що ідеально підходить для розподіленого зберігання даних, аналітики даних та застосунків для великих обсягів даних. У той же час блокове зберігання добре підходить для зберігання структурованих даних, зокрема баз даних та віртуальних машин. Його висока продуктивність відповідає вимогам високодержавних застосунків, які потребують низької затримки та високої пропускної здатності.

Вибір між AWS, Azure або GCP для блокового сховища може складним завданням через широкі можливості кожного постачальника. Порівняння базових характеристик сервісів наведено в табл. 1.

Таблиця 1

Порівняння сервісів блокового зберігання даних

Назва сервісу	**К-сть операцій читання та запису на секунду**	**Доступний обсяг**	**Пропускна здатність на секунду**	**Ціна за місяць використання станом на квітень 2024**
AWS Elastic Block Store (AWS EBS)	Від 5000 до 64000	Від 500 мегабайтів до 16 терабайтів	До 500 мегабайт	Від $0.018 за 1 гігабайт
Azure Managed Discs	До 64000	Від 1 гігабайта до 32 терабайтів	До 750 мегабайт	Від $0.15 за 1 гігабайт
GCP Persistent Disk	Від 15000 до 80000	Від 1 гігабайта до 64 терабайтів	До 480 мегабайт	Від $0.02 за 1 гігабайт

Оптимальний вибір залежить від нефункціональних вимог, або архітектурних атрибутів рішення. Ось декілька спеціальних характеристик що допоможуть у процесі прийняття рішення: приклади, які допоможуть вам у процесі прийняття рішень: AWS EBS виділяється безперебійним створенням знімків томів і підтримкою критично важливих функцій; Azure надає різноманітні параметри доступності, включаючи стандартні та преміальні SSD, ультрадиски, призначені для високопродуктивних робочих навантажень, однак важливо пам’ятати, що певні функції можуть бути доступними не в усіх регіонах; Google Cloud Persistent Disks оптимізує процеси резервного копіювання та аварійного відновлення за допомогою ефективних можливостей створення знімків.

Далі зупинимося на доступних сервісах для зберігання об’єктів. У таблиці 2 нижче наведено короткий аналіз сховищ запропонованих AWS, Azure або GCP. Варто зазначити, що всі вони володіють чудовими характеристиками довговічності, масштабованості і безпеки доступу.

Вибір між AWS, Azure або GCP для сховища даних конкретного рішення машинного навчання може бути доволі важким. Проте варто вказати, що платформа AWS вирізняється довговічністю та доступністю, що робить її чудовим вибором для міцних та надійних рішень для зберігання. GCP виділяється своєю моделлю оплати за секунду, пропонуючи фінансово рентабельні підходи до побудови рішення, а адаптивна структура ціноутворення дозволяє точно оплачувати використані ресурси, що особливо вигідно при змінному робочому навантаженні. Мультиплатформенна стратегія також може бути вигідною. Наприклад, інтеграція Azure Functions з Amazon S3 дозволяє безперебійно запускати події або виконувати код у відповідь на події сегмента S3.

Таблиця 2

Порівняння сервісів об'єктного зберігання даних

Назва сервісу	Загальний опис	Особливості
AWS S3	Це гнучке рішення для легкого зберігання та доступу до даних через Інтернет. Включено зручного інтерфейсу веб-сервісів полегшує розробникам веб-обчислення.	1) Політика сегментів: відноситься до політики керування ідентифікацією та доступом (IAM), яка може авторизуватись або обмежувати дозволи для ваших ресурсів S3. Ця політика виходить за рамки окремих файлів і дозволяє встановлювати правила безпеки для кількох файлів у певному сегменті. 2) Управління життєвим циклом: Amazon S3 використовує визначений набір правил для визначення дій для певної групи об’єктів. Цей добре продуманий підхід забезпечує безперебійне керування та зберігання об’єктів, акцентуючи увагу на оптимізації ресурсів для економічної ефективності. Він включає в себе два типи дій - дії транзакцій і дії закінчення терміну дії.
Azure Blob Storage	Рішення для зберігання об’єктів у хмарі, що пропонує доступне через Інтернет сховище для різних типів даних. Особливо ефективно для зберігання мультимедійних файлів, таких як аудіо та відео, а також динамічних і часто оновлюваних даних.	1) Рівневе об’єктне сховище: Blob-сховище надає можливість упорядковувати дані за гарячим, холодним і архівним рівнями, надаючи користувачам змогу ефективно зберігати та керувати даними на основі їх частоти використання та тимчасової чутливості. 2) Розширені функції керування даними: за допомогою Blob Storage користувачі можуть отримати доступ до розширених функцій керування даними, як-от керування версіями об’єктів, політики видалення, керування життєвим циклом і робочі процеси, керовані подіями. Ці функції покращують контроль і ефективність керування даними в системі зберігання.
Google Cloud storage	Надійний і гнучкий сервіс, що пропонує виняткову масштабованість і довговічність і розроблена спеціально для обробки неструктурованих даних, таких як файли, зображення, відео, резервні копії та журнали.	1) Гео-надлишковість: вибирайте параметри гео-надлишкового зберігання для копіювання даних у кількох географічних регіонах. Ця функція розширює можливості аварійного відновлення, забезпечуючи доступність даних з різних місць. 2) Класи сховищ:: існує можливість вибору з різних класів зберігання, включаючи Standard, Nearline, Coldline і Archive, щоб знайти ідеальний баланс між продуктивністю та ціною..

Враховуючи наведені характеристики і різноманітність сховищ даних які забезпечують найбільші платформи хмарних сервісів AWS, Azure та GCP, вибір ефективного сховища даних є дуже важливим завданням для побудови рішення, особливо в контексті машинного навчання. Для забезпечення продуктивності рішення МН пропонуємо зосередитись на наступних критеріях при виборі сховища даних:

1) Продуктивність доступу до даних: машинне навчання часто потребує обробки великих обсягів даних. Швидкість доступу до цих даних може значно вплинути на час, необхідний для навчання моделей. Повільний доступ до даних може значно збільшувати час навчання моделі, що робить систему менш ефективною і витратною в експлуатації.

2) Масштабованість: сховище даних має бути здатне ефективно обробляти цей ріст без компрометації продуктивності. Немасштабоване сховище може стати вузьким місцем, що обмежуватиме здатність організації використовувати великі набори даних для навчання більш складних моделей.
3) Захист даних: дані можуть містити конфіденційну інформацію або персональні дані, що вимагають застосування спеціальних політик доступу. Забезпечення захисту є важливим не лише з точки зору відповідності законодавству, а й для збереження довіри користувачів й уникнення витоків даних, які можуть призвести до репутаційних та фінансових втрат.
4) Інтеграція: легкість інтеграції сховища даних з іншими частинами рішення машинного навчання знижує технічні бар'єри та спрощує робочі процеси. Це дозволяє дослідникам та розробникам швидше і ефективніше розгортати, навчати та тестувати моделі, що підвищує продуктивність та інноваційний потенціал команди.
5) Вартість: необхідно враховувати не лише поточні витрати на зберігання та обслуговування, але й довгострокові витрати, пов'язані із зростанням обсягів даних. Вибір економічно ефективного рішення допомагає знизити фінансові витрати без шкоди для продуктивності або безпеки для рішення, що розробляється

Крім зазначених критеріїв в процесі побудови рішень машинного навчання слід зважати на тенденції розвитку сервісів сховищ даних, що доступні на ринку. Основні хмарні платформи постійно ведуть конкурентну боротьбу намагаючись розробити нові та більш ефективні підходи зберігання та використання даних. Наступні аспекти допоможуть покращити роботу з даними при побудові рішень машинного навчання:

1) Розширення функціоналу аналітики та спостереження: автоматичні оповіщення про ненормальні сплески використання ресурсів або загрозу безпеці, дозволяє проактивно керувати ресурсами та безпекою, швидко реагуючи на будь-які аномалії.
2) Інтеграція з сервісом Azure Synapse Analytics: дозволить отримати більш глибокі аналітичні отриманні знання та підвищити ефективність управління даними.
3) Застосування продукту Google BigQuery для аналітики: забезпечує високу продуктивність для аналізу великих даних, підвищуючи ефективність роботи та прийняття рішень на основі даних
4) Підтримка Amazon S3 Intelligent-Tiering: використовуйте сервіс для автоматичного переміщення об'єктів між різними рівнями зберігання на основі шаблонів доступу, що допоможе оптимізувати витрати, зберігаючи продуктивність при зберіганні даних.
5) Використання Google Cloud Storage Classes: допоможе з оптимізацією вартості зберігання даних, переміщуючи їх між класами з урахуванням шаблонів використання. Це дозволить знизити витрати на зберігання, зберігаючи при цьому можливість швидкого доступу до необхідних даних.

Загалом, використання цих пропозицій та покращень допоможе оптимізувати управління даними, знижуючи витрати та підвищуючи ефективність наявних хмарних рішень для зберігання даних.

Висновки

В статті підкреслюється критичне значення належно організованих сховищ даних для успішної реалізації проектів машинного навчання. У нинішніх умовах швидкого розвитку технологій потреба в ефективних рішеннях для зберігання великих обсягів даних стає дедалі актуальнішою для реалізації рішень машинного навчання.. Масштабованість сховищ даних, здатність збільшувати обсяги зберігання та обробки без зниження продуктивності, є провідною вимогою. Продуктивність і швидкодія також займають одне з центральних місць, оскільки висока швидкість доступу до даних і мінімальна затримка є критично важливими для ефективного навчання та тестування моделей машинного навчання. Максимальна швидкість обробки дозволяє прискорити етапи розробки і впровадження, що особливо важливо для задач у режимі реального часу.

Також акцент зроблено на тому, що системи повинні забезпечувати високий рівень доступності з мінімальним часом простою, а механізми резервного копіювання та відновлення є необхідністю для запобігання втратам даних. Захищеність інформації за допомогою шифрування та налаштування ролей користувачів сприяє запобіганню несанкціонованому доступу і дотриманню нормативних вимог. Це критично важливо для систем машинного навчання. Особливо у випадку опрацювання чутливих персональних даних наприклад у підборі персоналу чи медичних даних для діагностики. Важливо щоб було забезпечено шифрування інформації як під час передачі, так і при зберіганні.

Вартість зберігання є ще одним важливим аспектом при виборі сховищ даних. Економічна ефективність, прозоре ціноутворення і можливість оптимізації витрат є необхідними для організацій, що працюють з великими обсягами даних як у державному секторі так і у приватних компаніях.

Основні хмарні платформи, такі як AWS, Azure та GCP, пропонують широкі можливості для зберігання і обробки даних, але вибір між ними залежить від специфічних потреб організації та завдань машинного навчання. Визначення ключових параметрів, які мають найвищий пріоритет та ретельне порівняння їх характеристик допомагає приймати обґрунтовані рішення що дозволить розвивати планові системи з контрольованими витратами.

Таким чином, стаття підкреслює, що ефективні сховища даних значно підвищують продуктивність, безпеку і економічну ефективність рішень машинного навчання, що є важливим кроком для організацій у досягненні їхніх технологічних цілей.

Література

1. Ahmadi, Sina, Optimizing Data Warehousing Performance Through Machine Learning Algorithms in the Cloud (December 2023). International Journal of Science and Research (IJSR) 12 (12), 1859-1867, 2023. Available at SSRN: https://ssrn.com/abstract=4683244

2. AI-driven resource management strategies for cloud computing systems, services, and applications. World Journal of Advanced Engineering Technology and Sciences, 2024, 11(02), 559–566. https://doi.org/10.30574/wjaets.2024.11.2.0137

3. Li, H., Wang, X., Feng, Y., Qi, Y., & Tian, J. (2024). Integration Methods and Advantages of Machine Learning with Cloud Data Warehouses. International Journal of Computer Science and Information Technology, 2(1), 348-358. https://doi.org/10.62051/ijcsit.v2n1.36

4. Ravi Kashyap, "Machine Learning in Google Cloud Big Query using SQL," SSRG International Journal of Computer Science and Engineering, vol. 10, no. 5, pp. 17-25, 2023. Crossref, https://doi.org/10.14445/23488387/IJCSE-V10I5P103

5. van der Vlist, F., Helmond, A., & Ferrari, F. (2024). Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence. Big Data & Society, 11(1). https://doi.org/10.1177/20539517241232630

6. Wang, Y., Bao, Q., Wang, J., Su, G., & Xu, X. (2024). Cloud Computing for Large-Scale Resource Computation and Storage in Machine Learning. Journal of Theory and Practice of Engineering Science, 4(03), 163–171. https://doi.org/10.53469/jtpes.2024.04(03).14

7. Panchenko, T., Tuzova, I., Tuzov, O., & Chumak, O. (2024). Cloud services and an overview of their providers. Scientific Collection «InterConf+», (43(193), 550–559. https://doi.org/10.51582/interconf.19-20.03.2024.053