

Організація та зберігання великих наборів даних для навчання ШІ

Олексій Мельніков

olexa.m@gmail.com
N-IX Corporation

Received 03.02.2025, accepted 29.03.2025
<https://doi.org/10.32347/st.2025.3.1203>

Анотація. Зростаюча складність і масштаб застосувань штучного інтелекту (ШІ), особливо в машинному навчанні та глибокому навчанні, вимагають надійних та ефективних рішень для зберігання даних. Це особливо актуально при роботі з великими наборами даних, які часто досягають сотень терабайт, у різноманітних форматах. Для навчання моделей ШІ можуть бути використані різні формати даних, таких як відео та зображення для задач комп'ютерного зору, WAV та IQ для радіо та звукових сигналів, текст для мовних моделей. У цій статті розглядаються наукові підходи до організації зберігання даних для навчання ШІ з акцентом на ефективність та доступність. У статті буде розглянуто особливості роботи з форматами даних HDF5, WAV та IQ raw data, що використовуються для обробки радіосигналів.

Ключові слова: Штучний інтелект (ШІ), Машинне навчання, Глибоке навчання, Набори даних, Зберігання даних, HDF5, WAV, IQ raw data, Хмарне зберігання, Розподілені файлові системи, Локальне зберігання, Платформи керування даними, Версіонування даних, Масштабованість, Продуктивність, Різноманітність даних, Консолідація даних, Якість даних, Візуалізація даних, Безпека даних, HDFView, Шардинг даних, Оптимізація Push-down, Клонування без копіювання, TTL (час життя), ETL (видобування, трансформація, завантаження), Каталогізація, Керування метаданими.

ПОСТАНОВКА ПРОБЛЕМИ

Штучний інтелект стрімко розвивається, і його моделі стають все більш складними та потужними. Ефективність цих моделей безпосередньо залежить від якості та обсягу даних, на яких вони навчаються.

Збір, організація та підготовка даних є критично важливими етапами в процесі розробки ШІ, і дослідження в цій області

постійно розвиваються. Обсяги даних, що беруть участь у процесі навчання,



Oleksii Melnikov
Software engineer
at N-IX corporation

зростають у геометричній прогресії. Керування та організація таких масивів даних потребують спеціалізованих рішень які можуть забезпечити оперативний, постійний та автоматизований доступ до даних навчання.

Аналіз останніх досліджень та публікацій. Формат HDF5, його структура, переваги та способи обробки описані в [1] та [2]. Ці матеріали пояснюють, як HDF5 може ефективно використовуватися для зберігання великих наборів даних, включаючи дані різних типів, та як цей формат забезпечує ефективний доступ до даних та їх обробку. Практичні аспекти роботи з HDF5, включаючи інструменти для перегляду та редагування HDF5 файлів, а також способи оптимізації запису даних у цей формат розглянуті в [21], [22]. Використання формату даних HDF5 для зберігання спектрограмних даних та порівняння його з іншими форматами, такими як WAV та sigMF, для зберігання IQ даних подано в [3]. Різні рішення для зберігання даних для навчання ШІ, включаючи хмарні сервіси (AWS S3, Google Cloud Storage, Azure Blob Storage), розподілені файлові системи (DFS) та локальні рішення (SAN, NAS, JBOD) розглядаються в [4], [6], [7] та [10]. В них наведено аналіз переваг та недоліків кожного рішення, а також фактори, які слід враховувати при виборі. Об'єктно-

орієнтовані методи зберігання даних, їх переваги та використання в хмарних середовищах описані в [11], [23]. В [24] порівнюються розподілені файлові системи та розподілене об'єктно-орієнтоване зберігання, виділяючи їх ключові відмінності та випадки використання. Огляд популярних рішень для зберігання великих даних, включаючи AWS S3, Google Cloud Storage та Azure Blob Storage [9]. Огляд платформ управління даними для навчання ШІ, таких як Velotix, Databricks, DataRobot, Google Cloud AI Platform, Azure Synapse Analytics, IBM Watson Studio, Snowflake, Amazon SageMaker, Dataiku та Vast Data [5], [12], [25], [26], [27]. Вони описують ключові функції кожної з платформ та випадки їх використання. Важливість версійності даних для навчання ШІ, опис різних інструментів та методи версійності (DVC, LakeFS, GitLFS), а також найкращі практики для управління версіями даних розглянуті в [8], [13], [14], [15], [28]. Стратегії управління великими наборами даних розглянуті в [7]. Системи каталогізації даних розглянуті в [29], [30], [31], [32], [33].

META СТАТТІ

Надати всебічний огляд та рекомендації щодо організації зберігання великих наборів даних, які використовуються для навчання штучного інтелекту (ШІ), зокрема в контексті машинного та глибокого навчання, а також інструменти для ефективної організації зберігання даних, що сприятиме підвищенню продуктивності та інновацій у цій галузі.

Виклад основного матеріалу дослідження. У межах досліджень у сфері розпізнавання радіосигналів дані для досліджень збираються за допомогою різноманітних технічних засобів, призначених для роботи з радіо середовищем, що включають в себе радіоприймачі, SDR (Software Defined Radio - програмно-визначене радіо) приймачі, спектральні аналізатори, тощо. Дані, що обробляються, можуть представляти собою як оцифровану хвильову форму радіосигналу, так і набір його частотних

характеристик у часі. Такий метод збору даних продукує значні обсяги даних, що може досягати десятків або навіть сотень терабайт за короткий проміжок часу. Для оцінки можливих обсягів даних може бути застосована теорема Котельникова (Найквіста-Шеннона), що визначає мінімальну частоту дискретизації.

Однак, при обробці сигналів з обмеженою шириною пропускання можна застосовувати нижчі частоти дискретизації, що значно зменшує обсяги даних. Проте подальша обробка, така як видалення шумів та виділення ознак для навчання, може збільшити обсяги даних у декілька разів, особливо за умови застосування версіонування даних.

Збір даних з приймача включає кілька етапів. Спочатку сирі дані записуються на локальний диск. Потім проводиться постобробка для очищення від артефактів та конвертація у вибраний формат. Далі реєструються метадані для зручного доступу. Насамкінець, оброблені дані синхронізуються з віддаленими сховищами для резервного копіювання.

Вибір формату зберігання є важливим етапом дослідження, тому в межах роботи обрані формати HDF5, WAV та IQ raw data, які забезпечують оптимальне співвідношення між продуктивністю, точністю, сумісністю та зручністю роботи із записаними зразками радіосигналів. Розглянемо запропоновані формати.

HDF5 — це ієрархічний формат даних, який добре підходить для керування великими, складними та різнорідними даними, що часто зустрічаються в робочих навантаженнях ШІ [1]. Він дозволяє зберігати різні типи даних, такі як зображення, числові масиви та текст, в одному файлі, зберігаючи чітку організаційну структуру. Файл HDF5 містить метадані, які описують дані, що він містить, що полегшує розуміння та аналіз даних без необхідності покладатися на зовнішні джерела пов'язаних метаданих [2]. HDF5 забезпечує ефективні операції вводу/виводу (I/O) [1].

WAV (Waveform Audio File Format) — це стандартний формат аудіо файлів. Цей

формат може застосовується для зберігання даних зразків радіосигналів. Через неможливість зберігання детальної мета інформації, дані в цьому форматі потребують додаткових зусиль для супроводження та організації доступу.

IQ data представляє синфазну та квадратур-ну складові сигналу, фіксуючи як амплітуду, так і фазову інформацію [3]. Цей тип даних продукується SDR приймачами і може бути оброблений безпосередньо алгоритми ШІ, що може зменшити загальні вимоги для перетворення в проміжні формати. Однак, як і у випадку формату WAV, обробка великих обсягів IQ data може бути складною через його складність і обмеженість збереження супроводжувальних даних.

Використані формати не передбачають компресію даних під час збору та обробки. Тому обсяги даних постійно зростатимуть, що потребує масштабованих сховищ. Масштабованість дозволяє плавно нарощувати обсяги сховища відповідно до потреб, забезпечуючи безперебійну роботу та доступність даних, що критично важливо для безперервності процесів аналізу та навчання.

Традиційні системи зберігання часто важко масштабувати, що знижує продуктивність і збільшує витрати [4]. Наприклад, застарілі системи прямого підключення сховищ (DAS) мають обмежену ємність. Також системи мережевого підключення сховищ (NAS) можуть не забезпечити достатньої продуктивності для високонавантажених завдань ШІ.

Окрім масштабованості, існує кілька інших проблем, які необхідно вирішити при організації зберігання даних для навчання ШІ [4]. Навчання ШІ часто включає обчислювально інтенсивні завдання, які вимагають високошвидкісного доступу до даних, що призводить до затримок у навчанні та збільшення часу обробки. Застосунки ШІ використовують різноманітні типи даних, включаючи структуровані та неструктуровані формати, такі як текст, зображення, аудіо та відео. Традиційні системи зберігання, такі як

реляційні бази даних, призначені для структурованих даних, можуть мати труднощі з ефективною обробкою різноманітності.

Розрізнені сховища, де дані розподілені по різних системах і форматах, перешкоджають ефективній обробці через необхідність об'єднання даних на кожній ітерації. Саме тому консолідація даних в єдину платформу має вирішальне значення для покращення доступності даних та оптимізації робочих процесів ШІ [7]. Великі набори даних можуть містити неузгодженості, помилки та пропущені значення. Це негативно впливає на продуктивність моделей ШІ, тому критично важливими є інструменти для валідації та виправлення даних. Для валідації навчальних даних можуть бути застосовані методи візуалізації та аналіз. Аналіз великих наборів даних ускладнюється їхнім обсягом [7].

Традиційні методи візуалізації можуть бути неефективними. Тому для отримання змістовної інформації потрібні спеціалізовані методи та інструменти, такі як вибірка та агрегування.

Розв'язання цих проблем вимагає впровадження сучасних рішень для організації даних, які спеціально розроблені для робочих навантажень ШІ.

Хмарні рішення для зберігання даних стають все більш популярними для застосувань ШІ завдяки їх масштабованості, гнучкості та економічній ефективності [6]. Хмарні провайдери, такі як AWS, GCP та Azure, пропонують різноманітні послуги зберігання, адаптовані до різних потреб і бюджетів. Наприклад AWS пропонує сервіс S3 (Amazon Simple Storage Service) — це високоєфективний сервіс об'єктного зберігання, який пропонує різні класи зберігання для забезпечення різних моделей доступу та вимог до вартості [9]. Наприклад, S3 Standard підходить для даних, до яких часто звертаються, тоді як S3 Glacier призначений для довгострокового архівування даних. S3 також безперешкодно інтегрується з іншими сервісами AWS, що робить його зручним вибором для організацій, які вже використовують

екосистему AWS. Компанія Google надає сервіс Google Cloud Storage (GCS) що має аналогічні можливості, що й S3, з акцентом на тісну інтеграцію з іншими сервісами Google Cloud [9]. Цей сервіс пропонує різні класи зберігання, включаючи Standard, Nearline, Coldline та Archive, що дозволяє вибрати найбільш економічно вигідний варіант доступу до даних [10].

Компанія Microsoft надає Azure Blob Storage — це ще одне масштабоване та безпечне хмарне рішення для зберігання, призначене для обробки великих обсягів неструктурованих даних [9]. Azure Blob Storage підтримує неструктуровані типи даних і пропонує такі функції, як Write Once, Read Many (WORM) storage, щоб забезпечити незмінність даних для цілей відповідності та архівування.

Хмарні провайдери також пропонують інструменти передачі даних для полегшення ефективного переміщення та завантаження даних у хмарне сховище [10]. Наприклад, Storage Transfer Service від Google дозволяє передавати великі обсяги даних з різних джерел, включаючи локальне сховище, інших хмарних провайдерів та загальнодоступні набори даних, у бакети GCS.

Однією з ключових особливостей хмарного розміщення даних є також доступність гнучких і масштабованих обчислювальних потужностей, які пропонують перераховані вище провайдери. Це дозволяє реалізувати конвеєри обробки даних на стороні сховища. Однак важливо пам'ятати про потенційні проблеми безпеки та конфіденційності даних, пов'язані з хмарним зберіганням [6]. Хоча хмарні провайдери впроваджують надійні заходи безпеки, все ж необхідно ретельно оцінювати конкретні потреби та вибирати відповідні засоби контролю безпеки, такі як шифрування та контроль доступу, щоб зменшити потенційні ризики. Хоча хмарне зберігання пропонує зручність і масштабованість, локальне зберігання забезпечує кращий контроль та безпеку даних [6]. Це може бути вирішальним фактором для випадків із конфіденційними та критичними даними, які не можна зберігати в хмарі.

Розглянемо рішення локального зберігання даних. Мережі зберігання даних (SAN), що забезпечують високошвидкісний доступ до централізованого сховища, що робить їх придатними для додатків, які потребують низької затримки та високої пропускної здатності [6]. Мережеві системи зберігання даних (NAS): Системи NAS пропонують масштабовану ємність зберігання та меншу складність в управлінні, ніж SAN [6]. Вони добре підходять для зберігання та обміну великими обсягами неструктурованих даних, які зазвичай використовуються в проектах III. JBOD (Just Bunch Of Disks), це економічно ефективний спосіб розширити ємність зберігання шляхом безпосереднього підключення окремих жорстких дисків до сервера [6]. JBOD пропонує гнучкість з точки зору конфігурації зберігання, але може не мати розширених функцій, та вимагати додаткових зусиль для підтримання надійного зберігання та доступності.

Розподілені файлові системи (DFS), наприклад Hadoop DFS [11], дозволяють ефективно зберігати та керувати великими обсягами даних, розподіляючи їх між багатьма серверами, що забезпечує високу швидкість та доступність, особливо для завдань, де потрібна паралельна обробка даних. Завдяки здатності розділяти дані на частини та розподіляти їх по різних вузлах можливо реалізувати розміщення даних найближче або безпосередньо на спеціалізованому обладнанні [11].

Окрім зберігання великих об'ємів даних варто звернути увагу на спеціалізовані рішення для керування даними які існують на ринку.

Розглянемо кілька відомих комерційних платформ керування даними, спеціально розроблених для навчання III, які пропонують функції, що виходять за рамки простого зберігання та пошуку. Ці платформи надають інструменти для версіонування даних, керування моделями, відстеження експериментів та колективної роботи, оптимізуючі весь життєвий цикл розробки III.

Наведені в табл. 1 платформи пропонують різні рівні автоматизації, масштабованості та простоти використання. Наприклад,

DataRobot зосереджується на AutoML, щоб спростити розробку моделей, тоді як Dataiku наголошує на командній співпраці в дослідженні даних. Вибір правильної платформи залежить від конкретних потреб проекту III та досвіду команди з аналізу даних.

Багато з цих платформ включають функції для оптимізації управління даними та продуктивності.

Таблиця 1.

Ключові характеристики та приклади застосування комерційних платформ організації даних

Платформа	Ключові особливості	Приклади застосування
Velotix	Безпека даних на основі III, контроль доступу	Управління даними, відповідність, безпечний доступ до даних
Databricks	Уніфікована аналітична платформа, інтеграція штучного інтелекту, керування великомасштабними даними	Створення та розгортання моделей машинного навчання, дослідження даних
DataRobot	Автоматизоване машинне навчання (AutoML), розробка та розгортання моделі	Прогностична аналітика, оцінка ризиків, виявлення шахрайства
Google Cloud AI Platform	Комплексний набір інструментів машинного навчання, інтеграція TensorFlow, AutoML	Розробка, розгортання та управління моделлю, НЛП, комп'ютерний зір
Azure Synapse Analytics	Великі дані та сховища даних, аналіз даних на основі III	Інтеграція даних, аналітика та машинне навчання
IBM Watson Studio	Підготовка даних, розробка моделей, застосування NLP	Прогнозне моделювання, оптимізація обслуговування клієнтів
Snowflake	Хмарне сховище даних, інтеграція III, масштабований аналіз даних	Обмін даними, аналітика в реальному часі, сховище даних
Amazon SageMaker	Спрощена розробка та розгортання моделі машинного навчання	Створення, навчання та розгортання моделей III з підтримкою масштабування
Dataiku	Співпраця над науковими проектами даних, підготовка даних, моделювання та розгортання	Оцінка ризиків, аналіз клієнтів, запобігання шахрайству
Vast Data	Масштабована платформа даних, уніфіковане сховище, база даних і обчислення	III та глибоке навчання, високопродуктивні обчислення

Оптимізація Push-down використовує обчислювальну потужність сховища даних для ефективного виконання перетворень даних [12].

Замість переміщення даних до окремого вузла обробки, оптимізація push-down переносить обробку на сторону сховища даних, зменшуючи переміщення даних та

підвищуючи продуктивність. Функція "Time Machine" дозволяє отримати доступ до попередніх версій даних, створюючи контрольні журнали та забезпечуючи відновлення даних [12].

Клонування без копіювання дозволяє створювати копії даних без їх фізичного дублювання, зменшуючи витрати на зберігання та прискорюючи процеси розробки та тестування [12].

Ще одним ключовим аспектом організації зберігання даних для навчання ШІ є версіонування даних. Воно дозволяє відстежувати зміни в наборах даних, за потреби повертатися до попередніх версій та забезпечувати відтворюваність експериментів зі ШІ [13]. Інструменти версіонування даних, такі як DVC (Data Version Control) та LakeFS, надають функціональність, подібну до Git, для управління наборами даних. Вони дозволяють створювати гілки, фіксувати зміни та об'єднувати різні версії даних. Це забезпечує структурований та контрольований спосіб відстеження еволюції даних протягом усього процесу розробки ШІ. Важливо дотримуватись правильних підходів для роботи з версійованими даними описаними в [14], це дасть змогу уникнути типових помилок та пришвидшити роботу з проектом. Версіонування даних не тільки покращує управління даними, але й підвищує рівень співпраці та відтворюваності в дослідженнях ШІ [13]. Відстежуючи версії даних та параметри експериментів, дослідники можуть ділитися своєю роботою, відтворювати результати та спиратися на результати один одного. Значні об'єми даних які приймають участь в проектах ШІ потребують ефективних методів доступу до даних, оскільки швидкість доступу безпосередньо впливає на швидкість і продуктивність алгоритмів ШІ та їх навчання [29]. Інструментами для подолання складнощів управління даними для навчання ШІ є використання каталогів даних. Ці платформи, часто доповнені можливостями штучного інтелекту, забезпечують централізоване управління метаданими, полегшуючи пошук,

розуміння та використання даних [30]. Ключові функції таких каталогів включають автоматизоване керування метаданими, перевірки якості даних для підтримки їхньої надійності, інструменти для забезпечення управління та відповідності нормативним вимогам, функції для співпраці між командами, що працюють з даними та автоматизацію на основі ШІ/МО для таких завдань, як тегування та класифікація [30].

Розглянемо кілька популярних рішень для каталогізації даних, які широко використовуються для управління наборами даних, призначеними для навчання моделей ШІ. Atlan [32] відомий завдяки своїм функціям для співпраці та додатковими інструментами на основі штучного інтелекту, таким як автоматичне тегування та семантичний пошук [31]. Informatica Enterprise Data Catalog [33] надає розширені функції машинного навчання для профілювання та очищення даних, а також комплексні можливості управління даними [34]. AWS Glue Data Catalog [32] є важливим компонентом екосистеми AWS, що надає безсерверне сховище метаданих з автоматичним виявленням схем та інтеграцією з іншими сервісами AWS [36]. Google Cloud Dataplex Catalog (раніше відомий як Data Catalog) [32] розроблений для легкої інтеграції з сервісами Google Cloud та пропонує такі функції, як керування даними та метаданими, а також інтеграцію з Vertex AI та розширену аналітику [37]. Alation Data Catalog [32] є визнаним лідером у галузі. Використовуючи штучний інтелект та машинне навчання Alation Data Catalog спрощує виявлення даних, покращує спільну роботу та забезпечує управління даними [34]. У таблиці 2 наведені ключові функції платформ каталогізації даних для навчання моделей ШІ.

Впровадження чітко визначеної стратегії зберігання та керування даними, яка враховує збереження, організацію, версіювання та каталогізацію даних забезпечить ефективний процес навчання ШІ з оптимальною продуктивністю, масштабованістю та доступністю.

Таблиця 2.
Ключові функції та переваги платформ каталогізації даних

Платформа	Ключові особливості	Приклади застосування
Velotix	Безпека даних на основі ШІ, контроль доступу	Управління даними, відповідність, безпечний доступ до даних
Databricks	Уніфікована аналітична платформа, інтеграція штучного інтелекту, керування великомасштабними даними	Створення та розгортання моделей машинного навчання, дослідження даних
DataRobot	Автоматизоване машинне навчання (AutoML), розробка та розгортання моделі	Прогностична аналітика, оцінка ризиків, виявлення шахрайства
Google Cloud AI Platform	Комплексний набір інструментів машинного навчання, інтеграція TensorFlow, AutoML	Розробка, розгортання та управління моделлю, НЛП, комп'ютерний зір
Azure Synapse Analytics	Великі дані та сховища даних, аналіз даних на основі ШІ	Інтеграція даних, аналітика та машинне навчання
IBM Watson Studio	Підготовка даних, розробка моделей, застосування NLP	Прогнозне моделювання, оптимізація обслуговування клієнтів
Snowflake	Хмарне сховище даних, інтеграція ШІ, масштабований аналіз даних	Обмін даними, аналітика в реальному часі, сховище даних
Amazon SageMaker	Спрощена розробка та розгортання моделі машинного навчання	Створення, навчання та розгортання моделей ШІ з підтримкою масштабування
Dataiku	Співпраця над науковими проектами даних, підготовка даних, моделювання та розгортання	Оцінка ризиків, аналіз клієнтів, запобігання шахрайству
Vast Data	Масштабована платформа даних, уніфіковане сховище, база даних і обчислення	ШІ та глибоке навчання, високопродуктивні обчислення

ВИСНОВОК

Навчання алгоритмів ШІ та глибокого навчання вимагає значних об'ємів вихідних даних різноманітних форматів, а також складних, багаторівневих алгоритмів обробки цих даних. Така складність потребує особливої уваги до організації цих даних, що буде включати збереження вхідних даних, даних постобробки,

метаданих та інших супутніх даних, що будуть супроводжувати процес навчання. Зберігання даних для навчання ШІ з великими наборами даних вимагає комплексної стратегії, яка враховує різні фактори, включаючи масштабованість, продуктивність, різноманітність даних та специфічні формати даних. Хмарне зберігання, розподілені файлові системи та

локальні рішення пропонують різні переваги залежно від конкретних потреб застосування ШІ.

Для потреб навчання ШІ у сфері розпізнавання сигналів, що включають HDF5, WAV та IQ data, комбінація хмарного зберігання та платформи керування даними, подібної до тих, що перелічені в таблицях 1 та 2, може бути найбільш відповідним та оптимальним рішенням. Хмарне зберігання забезпечує масштабованість та гнучкість для обробки даних, тоді як платформа керування даними пропонує інструменти для версіонування, каталогізації та ефективного доступу до даних. Версіонування даних забезпечує відтворюваність, відстежує зміни та дозволяє повертатися до попередніх версій, тоді як методи каталогізації оптимізують пошук даних та прискорюють навчання ШІ. Вибір конкретної стратегії організації даних залежить від умов проекту, але попри все, правильний вибір стратегії дозволить скоротити витрати та пришвидшити процес навчання.

ЛІТЕРАТУРА

1. HDF5 format - GATK - Broad Institute. [Електронний ресурс]. Доступно: <https://gatk.broadinstitute.org/hc/en-us/articles/360035531712-HDF5-format>. Дата звернення: 25.02.2025.
2. Hierarchical Data Formats - What is HDF5? | NSF NEON. [Електронний ресурс]. Доступно: <https://www.neonscience.org/resources/learning-hub/tutorials/about-hdf5>. Дата звернення: 25.02.2025.
3. Decoding artifacts(hdf5) or converting artifacts to .ogg file - Libre Space Community. [Електронний ресурс]. Доступно: <https://community.libre.space/t/decoding-artifacts-hdf5-or-converting-artifacts-to-ogg-file/11447>. Дата звернення: 25.02.2025.
4. Why Storage Is the Unsung Hero for AI. [Електронний ресурс]. Доступно: <https://blog.purestorage.com/perspectives/why-storage-is-the-unsung-hero-for-ai/>. Дата звернення: 25.02.2025.
5. Best 10 AI Solutions for Data Management and Security | Velotix. [Електронний ресурс]. Доступно: <https://www.velotix.ai/resources/blog/best-ai-solutions-for-data-management-and-security/>. Дата звернення: 25.02.2025.
6. AI Storage: A Deep Dive into the Latest Technologies - Nfina. [Електронний ресурс]. Доступно: <https://nfina.com/ai-storage/>. Дата звернення: 25.02.2025.
7. Effectively Handling Large Datasets - Dataiku blog. [Електронний ресурс]. Доступно: <https://blog.dataiku.com/effectively-handling-large-datasets>. Дата звернення: 25.02.2025.
8. [Data Versioning Explained: Guide, Examples & Best Practices - lakeFS. [Електронний ресурс]. Доступно: <https://lakefs.io/blog/data-versioning/>. Дата звернення: 25.02.2025.
9. 10 Big Data Storage Solutions & Systems to Use In 2025 | Airbyte. [Електронний ресурс]. Доступно: <https://airbyte.com/top-etl-tools-for-sources/big-data-storage-solutions>. Дата звернення: 25.02.2025.
10. Design storage for AI and ML workloads in Google Cloud | Cloud Architecture Center. [Електронний ресурс]. Доступно: <https://cloud.google.com/architecture/ai-ml/storage-for-ai-ml>. Дата звернення: 25.02.2025.
11. What Is A Distributed Storage System - ScaleGrid. Дата звернення: 25.02.2025. <https://scalegrid.io/blog/what-is-a-distributed-storage-system/>. Дата звернення: 25.02.2025.
12. 10 Best Data Management Platforms (DMPs) for 2025 - Matillion. [Електронний ресурс]. Доступно: <https://www.matillion.com/learn/blog/data-management-platforms>. Дата звернення: 25.02.2025.
13. Best Practices for Data Versioning for Building Successful ML Models - Encord. [Електронний ресурс]. Доступно: <https://encord.com/blog/data-versioning/>. Дата звернення: 25.02.2025.
14. Data Versioning: Best Practices And Tools For Data Teams - Monte Carlo Data. [Електронний ресурс]. Доступно: <https://www.montecarlodata.com/blog-data-versioning-guide/>. Дата звернення: 25.02.2025.
15. MLflow Data Versioning: Techniques, Tools & Best Practices - lakeFS. [Електронний ресурс]. Доступно: <https://lakefs.io/blog/mlflow-data-versioning/>. Дата звернення: 25.02.2025.
16. Indexing Strategies to Optimize Database Performance - Developer Nation. [Електронний ресурс]. Доступно: <https://www.developernation.net/blog/8-indexing-strategies-to-optimize-database-performance/>. Дата звернення: 25.02.2025.

16. Everything You Need to Know When Assessing Data Indexing Techniques Skills - Alooba. [Электронный ресурс]. Доступно: <https://www.alooba.com/skills/concepts/database-and-storage-systems/database-management/data-indexing-techniques/>. Дата звернення: 25.02.2025.
17. 8 Data Indexing Strategies for Faster & Efficient Retrieval - Crown Records Management. [Электронный ресурс]. Доступно: <https://www.crownrms.com/insights/data-indexing-strategies/>. Дата звернення: 25.02.2025.
18. Efficient Data Extraction Techniques for Large Datasets - TDAN.com. [Электронный ресурс]. Доступно: <https://tdan.com/efficient-data-extraction-techniques-for-large-datasets/32018>. Дата звернення: 25.02.2025.
19. In-Depth Industry Outlook: Distributed File Systems And Object Storage Solutions Market Size & Forecast. [Электронный ресурс]. Доступно: <https://www.verifiedmarketresearch.com/product/distributed-file-systems-and-object-storage-solutions-market/>. Дата звернення: 25.02.2025.
20. Toolbox Talk: Storing data using HDF5 files. [Электронный ресурс]. Доступно: <https://dezaeraecox.com/toolbox-talk-hdf5-files/>. Дата звернення: 25.02.2025.
21. Write performance of raw data - HDF5 Library - HDF Forum. [Электронный ресурс]. Доступно: <https://forum.hdfgroup.org/t/write-performace-of-raw-data/10963>. Дата звернення: 25.02.2025.
22. What is object storage or object-based storage? - NetApp. [Электронный ресурс]. Доступно: <https://netapp.com/data-storage/storagegrid/what-is-object-storage>. Дата звернення: 25.02.2025.
23. Distributed File Systems vs Distributed Object Storage - GeeksforGeeks. [Электронный ресурс]. Доступно: <https://geeksforgeeks.org/distributed-file-systems-vs-distributed-object-storage>. Дата звернення: 25.02.2025.
24. VAST Data: Data Platform Built for Deep Learning and AI. [Электронный ресурс]. Доступно: <https://vastdata.com>. Дата звернення: 25.02.2025.
25. Informatica: AI Powered Cloud Data Management. [Электронный ресурс]. Доступно: <https://informatica.com>. Дата звернення: 25.02.2025.
- Top 10 Data Platforms that Use Generative AI - ScikIQ. [Электронный ресурс]. Доступно: <https://scikiq.com/blog/top-10-data-platforms-that-use-generative-ai>. Дата звернення: 25.02.2025.
26. Top 6 Dataset Version Control Tools for your Machine Learning workflow - Twine Blog. [Электронный ресурс]. Доступно: <https://twine.net/blog/dataset-version-control-tools>. Дата звернення: 25.02.2025.
27. AI Training Data: Benefits, Challenges, Example [2025] - Shaip. [Электронный ресурс]. Доступно: <https://www.shaip.com/blog/the-only-guide-on-ai-training-data-you-will-need-in/>. Дата звернення: 25.02.2025.
28. What Is an AI Data Catalog? | Secoda. [Электронный ресурс]. Доступно: <https://www.secoda.co/blog/ai-data-catalog>. Дата звернення: 25.02.2025.
29. What is a Machine Learning Data Catalog? - 2024 Guide - Atlan. Дата звернення: 15.03.2025. <https://atlan.com/machine-learning-data-catalog/>. Дата звернення: 25.02.2025.
30. Best Machine Learning Data Catalog Software: User Reviews from March 2025 - G2. [Электронный ресурс]. Доступно: <https://www.g2.com/categories/machine-learning-data-catalog>. Дата звернення: 25.02.2025.
31. Top 26 Data Catalog Tools to Consider in 2025 - lakeFS. [Электронный ресурс]. Доступно: <https://lakefs.io/blog/top-data-catalog-tools/>. Дата звернення: 25.02.2025.
32. Machine Learning Data Catalog Benefits & Use Cases | Informatica. [Электронный ресурс]. Доступно: <https://www.informatica.com/blogs/machine-learning-data-catalog.html.html.html>. Дата звернення: 25.02.2025.
33. The Data Scientist's Guide to the Data Catalog - Alation. [Электронный ресурс]. Доступно: <https://www.alation.com/blog/data-scientist-guide-data-catalog/>. Дата звернення: 25.02.2025.
34. AWS Glue Features and Benefits for Modern ETL Workflows - CelerData. [Электронный ресурс]. Доступно: <https://celerddata.com/glossary/aws-glue-features-and-benefits-for-modern-etl-workflows>. Дата звернення: 25.02.2025.
35. Dataplex Catalog overview | Google Cloud. [Электронный ресурс]. Доступно: <https://cloud.google.com/dataplex/docs/catalog-overview>. Дата звернення: 25.02.2025.

REFERENCES

1. HDF5 format - GATK - Broad Institute. [Online]. Available: <https://gatk.broadinstitute.org/hc/en-us/articles/360035531712-HDF5-format>. Accessed on: 25 Feb. 2025
2. Hierarchical Data Formats - What is HDF5? | NSF NEON. [Online]. Available: <https://www.neonscience.org/resources/learning-hub/tutorials/about-hdf5>. Accessed on: 25 Feb. 2025
3. Decoding artifacts(hdf5) or converting artifacts to .ogg file - Libre Space Community. [Online]. Available: <https://community.libre.space/t/decoding-artifacts-hdf5-or-converting-artifacts-to-ogg-file/11447>. Accessed on: 25 Feb. 2025
4. Why Storage Is the Unsung Hero for AI. Accessed on: 25 Feb. 2025 [Online]. Available: <https://blog.purestorage.com/perspectives/why-storage-is-the-unsung-hero-for-ai/>. Accessed on: 25 Feb. 2025
5. Best 10 AI Solutions for Data Management and Security | Velotix. Accessed on: 25 Feb. 2025 [Online]. Available: <https://www.velotix.ai/resources/blog/best-ai-solutions-for-data-management-and-security/>. Accessed on: 25 Feb. 2025
6. AI Storage: A Deep Dive into the Latest Technologies - Nfina. Accessed on: 25 Feb. 2025 [Online]. Available: <https://nfina.com/ai-storage/>. Accessed on: 25 Feb. 2025
7. Effectively Handling Large Datasets - Dataiku blog. Accessed on: 25 Feb. 2025 [Online]. Available: <https://blog.dataiku.com/effectively-handling-large-datasets>. Accessed on: 25 Feb. 2025
8. Data Versioning Explained: Guide, Examples & Best Practices - lakeFS. Accessed on: 25 Feb. 2025 [Online]. Available: <https://lakefs.io/blog/data-versioning/>. Accessed on: 25 Feb. 2025
9. Available: <https://lakefs.io/blog/data-versioning/>. Accessed on: 25 Feb. 2025
10. 10 Big Data Storage Solutions & Systems to Use In 2025 | Airbyte. Accessed on: 25 Feb. 2025 [Online]. Available: <https://airbyte.com/top-etl-tools-for-sources/big-data-storage-solutions>. Accessed on: 25 Feb. 2025
11. Design storage for AI and ML workloads in Google Cloud | Cloud Architecture Center. Accessed on: 25 Feb. 2025 [Online]. Available: <https://cloud.google.com/architecture/ai-ml/storage-for-ai-ml>. Accessed on: 25 Feb. 2025
12. What Is A Distributed Storage System - ScaleGrid. Accessed on: 25 Feb. 2025
13. 10 Best Data Management Platforms (DMPs) for 2025 - Matillion. Accessed on: 25 Feb. 2025 [Online]. Available: <https://www.matillion.com/learn/blog/data-management-platforms>. Accessed on: 25 Feb. 2025
14. Best Practices for Data Versioning for Building Successful ML Models - Encord. Accessed on: 25 Feb. 2025 [Online]. Available: <https://encord.com/blog/data-versioning/>. Accessed on: 25 Feb. 2025
15. Data Versioning: Best Practices And Tools For Data Teams - Monte Carlo Data. Accessed on: 25 Feb. 2025 [Online]. Available: <https://www.montecarlodata.com/blog-data-versioning-guide/>. Accessed on: 25 Feb. 2025
16. MLflow Data Versioning: Techniques, Tools & Best Practices - lakeFS. Accessed on: 25 Feb. 2025 [Online]. Available: <https://lakefs.io/blog/mlflow-data-versioning/>. Accessed on: 25 Feb. 2025
17. 8 Indexing Strategies to Optimize Database Performance - Developer Nation. Accessed on: 25 Feb. 2025 [Online]. Available: <https://www.developernation.net/blog/8-indexing-strategies-to-optimize-database-performance/>. Accessed on: 25 Feb. 2025
18. Everything You Need to Know When Assessing Data Indexing Techniques Skills - Alooba. Accessed on: 25 Feb. 2025 [Online]. Available: <https://www.alooba.com/skills/concepts/databases-and-storage-systems/database-management/data-indexing-techniques/>. Accessed on: 25 Feb. 2025
19. Data Indexing Strategies for Faster & Efficient Retrieval - Crown Records Management. Accessed on: 25 Feb. 2025 [Online]. Available: <https://www.crownrms.com/insights/data-indexing-strategies/>. Accessed on: 25 Feb. 2025
20. Efficient Data Extraction Techniques for Large Datasets - TDAN.com. Accessed on: 25 Feb. 2025 [Online]. Available: <https://tdan.com/efficient-data-extraction-techniques-for-large-datasets/32018>. Accessed on: 25 Feb. 2025
21. In-Depth Industry Outlook: Distributed File Systems And Object Storage Solutions Market Size & Forecast. Accessed on: 25 Feb. 2025 [Online]. Available: <https://www.verifiedmarketresearch.com/product/distributed-file-systems-and-object-storage-solutions-market/>. Accessed on: 25 Feb. 2025

22. Toolbox Talk: Storing data using HDF5 files. Accessed on: 25 Feb. 2025 [Online]. Available: <https://dezaecox.com/toolbox-talk-hdf5-files/>. Accessed on: 25 Feb. 2025
23. Write performance of raw data - HDF5 Library - HDF Forum. Accessed on: 25 Feb. 2025 [Online]. Available: <https://forum.hdfgroup.org/t/write-performance-of-raw-data/10963>. Accessed on: 25 Feb. 2025
24. What is object storage or object-based storage? - NetApp. Accessed on: 25 Feb. 2025 [Online]. Available: <https://netapp.com/data-storage/storagegrid/what-is-object-storage>. Accessed on: 25 Feb. 2025
25. Distributed File Systems vs Distributed Object Storage - GeeksforGeeks. Accessed on: 25 Feb. 2025 [Online]. Available: <https://geeksforgeeks.org/distributed-file-systems-vs-distributed-object-storage>. Accessed on: 25 Feb. 2025
26. VAST Data: Data Platform Built for Deep Learning and AI. Accessed on: 25 Feb. 2025 [Online]. Available: <https://vastdata.com>. Accessed on: 25 Feb. 2025
27. Informatica: AI Powered Cloud Data Management. Accessed on: 25 Feb. 2025 [Online]. Available: <https://informatica.com>. Accessed on: 25 Feb. 2025
28. Top 10 Data Platforms that Use Generative AI - ScikIQ. Accessed on: 25 Feb. 2025 [Online]. Available: <https://scikiq.com/blog/top-10-data-platforms-that-use-generative-ai>. Accessed on: 25 Feb. 2025
29. Top 6 Dataset Version Control Tools for your Machine Learning workflow - Twine Blog. Accessed on: 25 Feb. 2025 [Online]. Available: <https://twine.net/blog/dataset-version-control-tools>. Accessed on: 25 Feb. 2025
30. AI Training Data: Benefits, Challenges, Example [2025] Shaip.[Online].Available:<https://www.shaip.com/blog/the-only-guide-on-ai-training-data-you-will-need-in/>. Accessed on: 25 Feb. 2025
31. What Is an AI Data Catalog? | Secoda. [Online]. Available: <https://www.secoda.co/blog/ai-data-catalog>. Accessed on: 25 Feb. 2025
32. What is a Machine Learning Data Catalog? - 2024 Guide - Atlan. Дата звернення: 15.03.2025. <https://atlan.com/machine-learning-data-catalog/>. Accessed on: 25 Feb. 2025
33. Best Machine Learning Data Catalog Software: User Reviews from March 2025 - G2. [Online]. Available:<https://www.g2.com/categories/machine-learning-data-catalog>.Accessed on: 25 Feb. 2025
34. Top 26 Data Catalog Tools to Consider in 2025 - lakeFS.[Online].Available:<https://lakefs.io/blog/top-data-catalog-tools/>. Accessed on: 25 Feb. 2025
35. Machine Learning Data Catalog Benefits & Use CasesInformatica. [Online]. Available: <https://www.informatica.com/blogs/machine-learning-data-catalog.html.html.html>.Accessed on: 25 Feb. 2025

Organization and storage of large datasets for ai training

Oleksii Melnikov

The article examines current approaches to organizing and storing large datasets used for training artificial intelligence (AI) models, particularly for radio signal recognition tasks. It emphasizes the need for reliable and efficient data storage solutions due to the increasing scale and complexity of data involved in AI applications. The study analyzes various data storage formats (HDF5, WAV, IQ raw data), provides an overview of cloud-based solutions (AWS, Google Cloud, Azure), local storage systems (SAN, NAS, JBOD), and distributed file systems. It discusses best practices in data versioning and cataloging for enhanced accessibility, performance, and reproducibility of AI training processes. Recommendations are provided for selecting optimal storage methods based on the specific requirements of AI projects and the characteristics of the processed data.

Keywords: Artificial Intelligence (AI), Machine Learning, Deep Learning, Datasets, Data Storage, HDF5, WAV, IQ raw data, Cloud Storage, Distributed File Systems, Local Storage, Data Management Platforms, Data Versioning, Scalability, Performance, Data Diversity, Data Consolidation, Data Quality, Data Visualization.