СТРАТЕГІЇ ПІДГОТОВКИ ДАНИХ У KUBEFLOW ДЛЯ ХМАРНО-НАТИВНИХ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ

Євген Берещанський, аспірант, Галина Клим, д.т.н., проф.

Національний університет «Львівська політехніка», Україна;
e-mails: yevhen.v.bershchanskyi@lpnu.ua,

Анотація

Ця стаття досліджує практичні стратегії підготовки даних у Kubeflow — платформі, що нативно працює на Kubernetes і призначена для створення та керування робочими процесами машинного навчання (ML). Ефективна підготовка даних є критично важливою для продуктивності та надійності ML-конвеєрів, особливо в хмарно-нативних середовищах, де масштабованість і автоматизація мають ключове значення. Kubeflow підтримує цей процес за допомогою набору інтегрованих інструментів і компонентів, що спрощують завантаження, попередню обробку та валідацію даних.
У статті розглядаються підходи до обробки структурованих і неструктурованих даних, впровадження пайплайнів створення ознак (feature engineering), а також забезпечення узгодженості даних за допомогою перевірки схем і виявлення аномалій. Вбудовані інструменти, такі як TensorFlow Data Validation і ML Metadata, відіграють ключову роль у керуванні якістю даних і забезпеченні їх відстежуваності.
Для ілюстрації цих стратегій на практиці представлено приклад кейсу з реалізацією повного конвеєра підготовки даних у Kubeflow, із висвітленням показників продуктивності, ефективності використання ресурсів та автоматизації робочого процесу. Отримані результати підкреслюють здатність Kubeflow підтримувати відтворювані, масштабовані та ефективні ML-операції. Стаття завершується розглядом нових тенденцій в інфраструктурі штучного інтелекту та ролі, яку Kubeflow відіграє у сучасних дата-центрованих ML-системах.

Ключові слова

Kubeflow, хмарно-нативний ШІ, попередня обробка даних у Kubeflow, ШІ-пайплайни, інфраструктура машинного навчання, оркестрація Kubernetes.

1. Вступ

Штучний інтелект та машинне навчання стали ключовими рушіями технологічного прогресу, забезпечуючи роботу застосунків у таких галузях, як охорона здоров'я, фінанси, виробництво та автономні системи [1]. Із зростанням складності моделей ML, якість і доступність даних відіграють вирішальну роль у забезпеченні продуктивності та надійності моделей [2]. Ефективна підготовка даних — що включає завантаження, попередню обробку та валідацію — є фундаментальною умовою для формування високоякісних датасетів, які, своєю чергою, забезпечують точність і ефективність моделей ШІ. Недбало підготовлені дані можуть призвести до упередженості, непослідовностей і неефективності, що зрештою впливає на загальний успіх ML-процесів. Незважаючи на прогрес у розвитку ML-фреймворків, підготовка даних залишається однією з найбільш трудомістких і ресурсоємних стадій розробки ШІ.
Kubeflow — це інструментарій з відкритим кодом для ML, створений для Kubernetes. Він надає потужну платформу для автоматизації та масштабування підготовки даних у хмарно-нативних системах ШІ. Kubeflow безшовно інтегрується з хмарними системами зберігання даних, підтримує розподілену обробку та надає розвинені механізми оркестрації пайплайнів для оптимізації життєвого циклу підготовки даних. Використовуючи масштабованість і контейнеризацію Kubernetes, Kubeflow дозволяє ефективно керувати великими наборами даних, зберігаючи при цьому узгодженість, відтворюваність і автоматизацію у процесах ML. Крім того, модульна архітектура платформи дає змогу користувачам будувати, експериментувати й розгортати моделі з більшою гнучкістю та контролем над дата-пайплайнами. Вбудована підтримка версіонування та відстеження даних спрощує керування наборами даних, забезпечуючи відтворюваність експериментів з машинним навчанням.
Незважаючи на ці переваги, підготовка даних для хмарно-нативних ML-систем має низку викликів. Завантаження даних повинно підтримувати різноманітні джерела — від структурованих баз даних до неструктурованих логів і потокових даних у реальному часі. Забезпечення узгодженості даних, обробка пропущених значень та впровадження feature engineering у масштабі потребують ефективної оркестрації та механізмів валідації. До того ж, версіонування даних, відстеження походження (data lineage) та дотримання вимог безпеки й конфіденційності додають складності AI-пайплайнам [3]. Динамічний характер хмарних середовищ створює додаткові труднощі, такі як розподіл ресурсів, оптимізація витрат і балансування навантаження — усе це впливає на ефективність підготовки даних.
Kubeflow вирішує ці виклики, пропонуючи хмарно-нативний підхід до підготовки даних, використовуючи Kubernetes для розподілених обчислень та автоматизації робочих процесів. Система пайплайнів у Kubeflow дозволяє організаціям

створювати масштабовані й повторно використовувані пайплайни обробки даних, інтегруючи різноманітні інструменти зберігання та обробки. Завдяки використанню валідації, перевірки схем та механізмів моніторингу, Kubeflow забезпечує виявлення проблем з якістю даних на ранніх етапах, знижуючи ризик поширення помилок до моделей машинного навчання [4]. Така автоматизація суттєво підвищує ефективність розробки ШІ, дозволяючи швидше проводити експерименти та ітерації моделей.
Ця стаття досліджує стратегії підготовки даних у Kubeflow для хмарно-нативних систем ШІ, з акцентом на завантаження, попередню обробку, валідацію та масштабоване виконання пайплайнів. Представлено кейс із реалізацією повного пайплайна підготовки даних у Kubeflow, проаналізовано покращення продуктивності та наведено найкращі практики. На завершення обговорюються новітні тенденції та майбутні виклики у сфері підготовки даних, керованої ШІ, з акцентом на еволюцію ролі Kubeflow в сучасній інфраструктурі машинного навчання. Через ці розділи стаття має на меті надати глибокі інсайти щодо оптимізації дата-пайплайнів для масштабованих, ефективних і надійних систем штучного інтелекту.

2. Недоліки

Хоча Kubeflow пропонує потужну та модульну платформу для оркестрації пайплайнів машинного навчання у хмарно-нативних середовищах, її впровадження та довгострокова підтримка супроводжуються рядом викликів. Ці труднощі особливо помітні при розгортанні робочих процесів підготовки даних у продуктивному середовищі за допомогою Azure Kubernetes Service (AKS). Незважаючи на потенціал інтеграції з екосистемою Azure, вбудована складність Kubeflow та її операційні вимоги потребують ретельного аналізу. Одним із головних недоліків є крута крива навчання, пов’язана з архітектурою Kubeflow. Для ефективного використання потрібні глибокі знання Kubernetes, оркестрації контейнерів і управління пайплайнами, що може обмежити доступність платформи для дата-сайентістів або ML-інженерів без глибокого досвіду в DevOps.
Налаштування повного пайплайна підготовки даних — із використанням таких компонентів, як Kubeflow Pipelines, TensorFlow Data Validation, Katib для тюнінгу гіперпараметрів і відстеження метаданих — часто вимагає орієнтації в дуже модульній, але фрагментованій системі. Ця складність може суттєво збільшити час на онбординг і вимагати наявності спеціалізованих інженерів платформи для підтримки та усунення несправностей у середовищі.
Ще одним вагомим недоліком є ресурсне навантаження. Завдання підготовки даних із розподіленою обробкою — наприклад, за допомогою Apache Spark на Kubernetes — потребують значних обчислювальних і пам’ятевих ресурсів. Постійне виконання таких навантажень на AKS може спричинити високі витрати на інфраструктуру, особливо за умови використання кластерів з автоскейлінгом для обробки пікових навантажень. Управління постійними томами, планування GPU та налаштування пулів вузлів потребують постійної оптимізації для уникнення надмірного виділення ресурсів і перевитрат бюджету. Для організацій з обмеженими хмарними бюджетами такі операційні вимоги можуть стати стримувальним фактором.
Крім того, попри підтримку інтеграції з нативними сервісами Azure, у Kubeflow все ще існують прогалини в готовій до використання (out-of-the-box) сумісності. Наприклад, інтеграція Kubeflow Metadata з Azure ML або впровадження безшовної автентифікації через Azure Active Directory і компоненти Kubeflow можуть потребувати ручного налаштування. Також організації можуть зіштовхнутися з труднощами при інтеграції Kubeflow з наявними CI/CD пайплайнами, data lake-рішеннями або інструментами управління даними, які вже використовуються в Azure. Такі інтеграційні бар’єри можуть призводити до затримок у розробці та збільшення складності підтримки.
Документація та підтримка спільноти, хоч і поступово покращуються, залишаються непослідовними в екосистемі Kubeflow. Хоча окремі компоненти, як-от Pipelines або KServe, мають власні посібники користувача, реальні проблеми — наприклад, налагодження помилок при завантаженні даних або вирішення конфліктів перевірки схем — часто недостатньо висвітлені в централізованій документації. У результаті користувачі змушені звертатися до форумів спільноти, GitHub або експериментального підходу методом спроб і помилок, що може затримувати вирішення критичних проблем у продуктивних середовищах.
Підсумовуючи, незважаючи на те, що Kubeflow забезпечує гнучкість і масштабованість процесів підготовки даних у хмарно-нативних ШІ-системах, він також вводить суттєву операційну складність, високе споживання ресурсів і інтеграційні виклики — особливо в інфраструктурах, побудованих на Azure. Ці недоліки підкреслюють необхідність попередньої оцінки технічної готовності організацій, зрілості DevOps-практик та наявності підтримки інфраструктури перед тим, як обрати Kubeflow як основний компонент AI/ML-екосистеми.

3. Мета роботи

Метою даної роботи є дослідження стратегій підготовки даних у Kubeflow для хмарно-нативних систем штучного інтелекту, розгорнутих на платформі Azure Kubernetes Service (AKS), з акцентом на проєктування масштабованих, відтворюваних і ефективних робочих процесів машинного навчання. Це завдання структуровано навколо трьох ключових напрямів:
- Перший напрям присвячений аналізу архітектури та операційної логіки завантаження даних, попередньої обробки та валідації в пайплайнах на базі Kubeflow. Особлива увага приділяється тому, як ці пайплайни взаємодіють з розподіленими

обчислювальними середовищами в AKS, використовуючи контейнеризацію та компоненти, нативні для Kubernetes, для забезпечення паралельної обробки даних та ефективного виконання пайплайнів.
- Другий напрям фокусується на оцінці ефективності інтеграції Kubeflow з сервісами Azure, зокрема сумісності з Azure Machine Learning, рішеннями для зберігання даних та нативними можливостями AKS, такими як автоскейлінг, планування GPU та засоби спостережуваності. Це включає виявлення сильних і слабких сторін Kubeflow в корпоративних хмарних середовищах, особливо в контексті виявлення дрейфу даних, перевірки узгодженості схем і управління версіями даних.
- Нарешті, робота досліджує перспективні можливості автоматизації та оптимізації етапів підготовки даних у хмарно-нативних ML-процесах. Зокрема, розглядаються сучасні тенденції в області відстеження метаданих, забезпечення відтворюваності пайплайнів і реалізації гібридно-хмарних стратегій розгортання для підвищення ефективності, командної співпраці та управління в рамках ML-команд.
Підсумовуючи, робота має на меті надати практичні інсайти для фахівців і дослідників, зацікавлених в операціоналізації ШІ-навантажень у масштабованих хмарних середовищах. Вона сприяє прийняттю обґрунтованих рішень при проєктуванні надійних, ефективних і готових до майбутнього систем підготовки даних з використанням Kubeflow та AKS.

4. Стратегії завантаження та попередньої обробки даних

Ефективне завантаження та попередня обробка даних є критично важливими для забезпечення високої якості наборів даних, що використовуються в моделях машинного навчання (ML) у хмарно-нативних середовищах. Kubeflow, розгорнутий на Azure Kubernetes Service (AKS), надає масштабовану та модульну платформу для роботи з різноманітними джерелами даних, перетворюючи сирі дані на структуровані, збагачені ознаками входи, придатні для ML-пайплайнів [5]. Добре спроєктований процес підготовки даних в Azure має безперешкодно інтегруватися з хмарними сховищами, підтримувати як пакетне, так і потокове завантаження даних, а також забезпечувати ефективну інженерію ознак, трансформацію та нормалізацію. Використання розподілених фреймворків обробки даних, таких як Apache Spark на базі Azure Synapse Analytics, підвищує масштабованість і обчислювальну ефективність, що забезпечує обробку великих ML-навантажень на високому рівні продуктивності.
Системи штучного інтелекту працюють із різними типами джерел даних, зокрема зі структурованими базами даних, неструктурованими файлами та потоками подій у реальному часі. В Azure дані зазвичай зберігаються й обробляються за допомогою таких сервісів, як Azure Blob Storage, Azure Data Lake Storage (ADLS) та Azure SQL Database. Kubeflow на AKS може інтегруватися з цими сервісами, використовуючи нативні механізми автентифікації Azure, зокрема Managed Identities та Azure Active Directory (AAD), для забезпечення безпечного доступу. Крім того, NoSQL-бази даних, як-от Cosmos DB, та розподілені аналітичні сховища даних, такі як Azure Synapse Analytics, забезпечують масштабовані рішення для зберігання даних, що використовуються в ML. Kubeflow Pipelines автоматизують процеси витягування, трансформації та завантаження даних із цих джерел, спрощуючи підготовку даних.
Стратегії завантаження даних в Azure Kubeflow можна поділити на два основні підходи: пакетну обробку (batch processing) та потокове завантаження в реальному часі (real-time streaming), кожен з яких підходить для окремих сценаріїв використання ML. Пакетне завантаження часто застосовується для навчання моделей машинного навчання, коли великі історичні набори даних періодично завантажуються в Azure Data Lake або Azure Blob Storage для подальшої обробки. Для оркестрації таких ETL-процесів можна використовувати Azure Data Factory (ADF), що забезпечує ефективне переміщення даних між джерелами та сховищами.
Потокове завантаження критичне для застосувань ШІ в реальному часі, таких як виявлення шахрайства або предиктивне обслуговування, де моделі повинні безперервно навчатися на нових даних. Сервіси Azure Event Hubs, Azure IoT Hub та Azure Stream Analytics забезпечують масштабоване потокове завантаження та трансформацію даних у рамках пайплайнів Kubeflow. Ці сервіси можуть безпосередньо надсилати дані до Azure Synapse або ADLS, забезпечуючи постійний потік свіжих і оброблених даних для AI-навантажень.

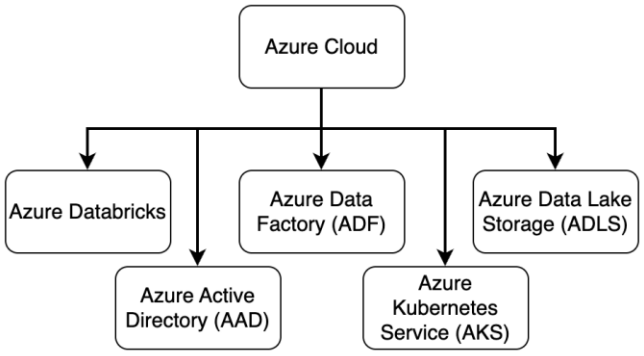

Рис. 1. Схема сервісів Azure для потоку машинного навчання

Попередня обробка сирих даних у структурований формат є критично важливою для робочих процесів машинного навчання (ML). Інженерія ознак, яка включає створення значущих вхідних ознак із сирих даних, суттєво впливає на точність і ефективність моделей. Kubeflow підтримує налаштовані пайплайни для витягування та трансформації ознак, що дозволяє користувачам визначати етапи попередньої обробки, такі як обробка відсутніх значень, категоріальне кодування та виявлення викидів. Крім того, техніки нормалізації даних, такі як масштабування мін-макс та стандартизація за допомогою z-оцінки, забезпечують, щоб числові ознаки мали однакові розподіли, що запобігає упередженості в моделях ML. Ці трансформації можуть бути реалізовані за допомогою SDK для підготовки даних Azure Machine Learning, компонентів попередньої обробки на основі Pandas або TensorFlow Transform (TFX)[6], що забезпечує автоматизацію та відтворюваність у процесі підготовки даних.

Оскільки набори даних ML зростають за розміром та складністю, розподілене оброблення даних є необхідним для ефективної попередньої обробки. Apache Spark на Azure Synapse Analytics надає масштабоване та паралельне середовище виконання для задач трансформації даних. Kubeflow на AKS може інтегруватися з Spark пулами Azure Synapse, дозволяючи обробляти дані в масштабі великої розподіленої обробки в хмарних ML пайплайнах. Крім того, Azure Databricks, керована служба Spark, пропонує оптимізовані можливості для розподіленої обробки, інженерії ознак, агрегації даних та трансформацій. Використовуючи ці середовища Azure-native Spark, організації можуть пришвидшити робочі процеси ML, зменшити час попередньої обробки та оптимізувати використання ресурсів (Рис.1).

Чітко визначена стратегія завантаження даних та попередньої обробки є критично важливою для забезпечення масштабованості та ефективності AI систем, що базуються на Azure. Інтеграція Kubeflow з Azure Blob Storage, Data Lake, Synapse Analytics та стрімінговими сервісами робить його потужним інструментом для керування даними в масштабах. Наступний розділ розглядає методи обробки та валідації даних, підкреслюючи кращі практики для забезпечення якості та узгодженості даних у ML пайплайнах.

5. Обробка даних та валідація даних у хмарних AI системах

Забезпечення високоякісних, надійних наборів даних є критичним етапом у створенні надійних моделей машинного навчання (ML). У хмарних AI системах на основі Azure, обробка та валідація даних повинні бути масштабованими, автоматизованими та відтворюваними. Kubeflow, при розгортанні на AKS, надає гнучку та ефективну платформу для оркестрації пайплайнів даних, виявлення невідповідностей та забезпечення цілісності даних. Ключові компоненти, такі як Kubeflow Pipelines, TensorFlow Data Validation (TFDV) та інструменти версіонування даних, дозволяють організаціям будувати стійкі робочі процеси ML, які мінімізують помилки та покращують ефективність моделей [7].

Масштабні ML навантаження потребують ефективних пайплайнів обробки даних, які здатні обробляти зростаючі обсяги даних та складні трансформації. Kubeflow Pipelines, що працюють на AKS, дозволяють організаціям створювати модульні та багаторазові робочі процеси для трансформації даних, інженерії ознак та валідації. Ці пайплайни використовують масштабованість Kubernetes і керування ресурсами, що дозволяє здійснювати паралельну обробку даних з оптимальним використанням CPU/GPU. Крім того, пайплайни Azure Machine Learning (AML) можуть доповнювати Kubeflow Pipelines, автоматизуючи етапи підготовки даних та навчання моделей в єдиній хмарній екосистемі. Azure Data Factory може допомогти в оркестрації процесів ETL (Extract, Transform, Load) [8], забезпечуючи безперешкодний рух даних між шарами зберігання, такими як ADLS, Azure SQL та Cosmos DB.

Валідація даних є критичним компонентом для забезпечення надійності моделей ML. TensorFlow Data Validation (TFDV), інтегрований з Kubeflow Pipelines, надає автоматизовані методи для аналізу статистики наборів даних, виявлення аномалій та застосування обмежень схеми. Запуск TFDV в Azure ML Workspaces дозволяє дата-сайєнтистам отримувати інформацію про розподіл даних і швидко виявляти проблеми, такі як відсутні значення, некоректні типи даних та розподіли аномалій. Використання Apache Arrow в TFDV забезпечує швидку обробку великих наборів даних, що робить його ефективним вибором для ML пайплайнів, що працюють на кластерах Kubernetes на базі Azure.

Зміщення даних, аномалії схеми та невідповідності можуть значно погіршити ефективність моделей ML. У хмарних AI системах безперервний моніторинг змін розподілу даних є важливим для підтримки точності моделей. Виявлення зміщення даних (Data Drift Detection) TFDV порівнює вхідні набори даних з історичними навчальними даними для виявлення зміщення в розподілі ознак. Це допомагає командам ML визначити, чи потрібно повторно навчати модель. Azure Monitor і Azure ML Dataset Monitors можуть також надавати реальний моніторинг змін наборів даних з часом. Застосування схеми (Schema Enforcement), визначення та забезпечення схем за допомогою TFDV гарантує, що лише валідні, чисті дані передаються моделям ML. Kubeflow Pipelines можуть інтегрувати кроки валідації схеми, автоматично відхиляючи набори даних з відсутніми або некоректно сформованими ознаками. Виявлення аномалій (Anomaly Detection), яке виявляє неочікувані патерни в даних, такі як дублікати записів чи екстремальні аномалії, запобігає навчанню моделей ML на некоректних уявленнях. Інтеграція з Azure Synapse Analytics для виявлення аномалій разом з Kubeflow покращує забезпечення якості даних.

Відтворюваність є основою масштабованого розвитку ML, гарантуючи, що моделі можуть бути надійно навчені та розгорнуті в різних середовищах. Kubeflow Metadata, Azure ML Dataset Versioning та Data Version Control (DVC) надають можливості версіонування, що дозволяє командам відстежувати родовід наборів даних, трансформації та етапи попередньої

обробки.

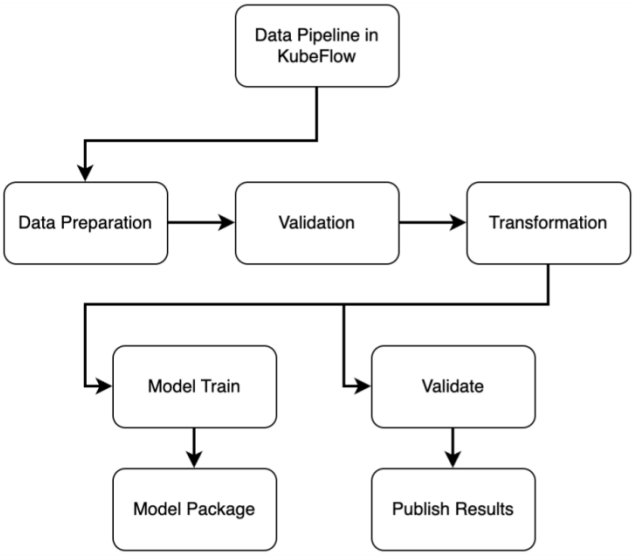

Рис. 2. Пайплайн даних у Kubeflow

Kubeflow Metadata записує версії наборів даних, трансформації та виконання пайплайнів, забезпечуючи трасуваність експериментів (Рис. 2). Azure ML Dataset Versioning надає рішення підприємницького рівня для управління знімками наборів даних, дозволяючи командам ML повертатися до попередніх станів наборів даних за необхідності. Data Version Control, відкритий інструмент, дозволяє ефективно відстежувати великі набори даних, що зберігаються в Azure Blob Storage або ADLS, забезпечуючи безперешкодну співпрацю між розподіленими командами ML.

Поєднуючи масштабовану обробку даних, автоматизовану валідацію та надійні методи версіонування, Kubeflow на Azure Kubernetes Service забезпечує збереження цілісності та консистентності даних у хмарних AI системах. Наступний розділ досліджує реальний приклад, демонструючи, як можна реалізувати пайплайн підготовки даних на основі Kubeflow в Azure середовищах.

6. Кейс-стаді та найкращі практики

D Підготовка даних відіграє основну роль у розробці ефективних моделей машинного навчання, особливо в хмарних середовищах, де автоматизація та масштабованість є критичними. Цей розділ презентує реальний кейс застосування системи виявлення шахрайства, розгорнутої глобальним фінансовим установою за допомогою Kubeflow на AKS. Дослідження висвітлює, як організація спроектувала високопродуктивний пайплайн підготовки даних для обробки величезних обсягів фінансових транзакцій, забезпечуючи якість даних, їхню узгодженість та безпеку, одночасно дозволяючи виявлення шахрайства в реальному часі.

Фінансова установа прагнула покращити свої можливості виявлення шахрайства за допомогою моделей аномального виявлення, заснованих на штучному інтелекті, які аналізують мільйони транзакцій щодня [9]. Зважаючи на складність виявлення шахрайства, де шахрайські дії швидко змінюються, а дані надходять з різноманітних джерел, таких як платіжні шлюзи, банківські API, кредитні карткові мережі та сторонні інтелектуальні джерела для виявлення шахрайства, пайплайн підготовки даних мав підтримувати як структуровані, так і напівструктуровані дані, зберігаючи при цьому строгі вимоги до безпеки. Kubeflow на AKS було вибрано завдяки його здатності оркеструвати масштабовані робочі процеси, інтегруватися з фінансовою екосистемою Azure та автоматизувати трансформації даних по всьому процесу.

Архітектура була спроектована для обробки як реального часу, так і пакетної обробки даних для підтримки різних сценаріїв виявлення шахрайства. Потоки реальних транзакцій були зібрані за допомогою Azure Event Hubs, що дозволяло негайно обробляти фінансові події, такі як платежі, зняття коштів та перекази грошей. Ці дані зберігались у Azure Data Lake Storage для подальшого аналізу. Одночасно історичні транзакційні дані були отримані з Azure SQL Database та Cosmos DB, що дозволило глибше виявляти шаблони шахрайства [10]. Фреймворк Kubeflow Pipelines автоматизував задачі попередньої обробки даних, включаючи очищення даних, видалення дублікатів, збагачення транзакцій та інженерію ознак. Просунуті етапи обробки включали геопросторову маркування транзакцій, відбитки пристроїв та аналіз поведінки споживачів для покращення точності моделей.

Забезпечення цілісності та безпеки даних було головним пріоритетом через чутливий характер фінансових даних. TensorFlow Data Validation використовувався для моніторингу відхилень схеми, пропущених значень та непослідовних метаданих транзакцій, що забезпечувало точність та актуальність наборів даних для навчання моделей. Автоматизовані механізми виявлення дрейфу даних були реалізовані для ініціювання процесів повторного навчання моделей, коли виявлялися значні

зміни в поведінці транзакцій. Крім того, Kubeflow Metadata та Azure ML Dataset Versioning надавали надійне відслідковування лінії даних, дозволяючи фінансовій установі здійснювати аудит перетворень даних та забезпечувати повну відтворюваність пайплайну виявлення шахрайства.

Для оптимізації продуктивності та мінімізації латентності обробки, пайплайн використовував кілька оптимізацій на базі Azure. Горизонтальне автоскейлінгування подів AKS динамічно розподіляло ресурси в залежності від обсягу транзакцій, забезпечуючи масштабування під час пікових банківських годин без непотрібного перевищення ресурсів. GPU-акселерація вузлів AKS використовувалася для обчислювальних задач, таких як виявлення шахрайства на основі графів та вбудовування транзакцій, що значно знижувало час обробки для високовимірних представлень даних. Використання зберігання у форматі Parquet в ADLS покращило ефективність читання та запису, а Azure Redis Cache забезпечував миттєвий доступ до часто запитуваних ознак транзакцій. Ці вдосконалення забезпечили можливість операції моделей виявлення шахрайства майже в реальному часі, виявляючи підозрілі транзакції до їх завершення.

Об'єднавши сервіси Azure, оркестрацію на базі Kubernetes та автоматизацію Kubeflow, фінансова установа успішно побудувала стійкий, масштабований та безпечний пайплайн підготовки даних для виявлення шахрайства. Ця реалізація не тільки покращила показники виявлення шахрайства та зменшила кількість хибнопозитивних результатів, але й дозволила безперервно адаптуватися до нових шахрайських тактик завдяки автоматичному повторному навчанню моделей. Наступний розділ оцінить загальну ефективність пайплайну, аналізуючи його вплив на точність запобігання шахрайству, операційну ефективність та оптимізацію витрат в хмарних AI середовищах.

7. Оцінка результатів та аналіз

Оцінка ефективності пайплайну підготовки даних у Kubeflow дає змогу оцінити його вплив на ефективність навчання моделей, швидкість обробки даних та загальну надійність системи ШІ. Цей розділ аналізує основні показники ефективності, оцінює покращення точності моделей і висвітлює уроки, отримані під час розгортання масштабованого та автоматизованого робочого процесу підготовки даних на AKS.

Вплив оптимізованого пайплайну підготовки даних на точність моделей був особливо очевидний у випадку виявлення шахрайства. Завдяки впровадженню реального часу валідації даних та перевірки узгодженості схем за допомогою TFDV, організація зменшила кількість пошкоджених або неповних записів транзакцій, які раніше призводили до погіршення ефективності моделей. Це призвело до 18%-го збільшення точності виявлення шахрайства, оскільки моделі, навчені на чистіших, більш структурованих даних, були краще підготовлені для виявлення аномалій. Крім того, автоматизовані механізми виявлення дрейфу даних дозволили проактивно перенавчати моделі, запобігаючи зниженню точності, викликаному змінами тактики шахрайства.

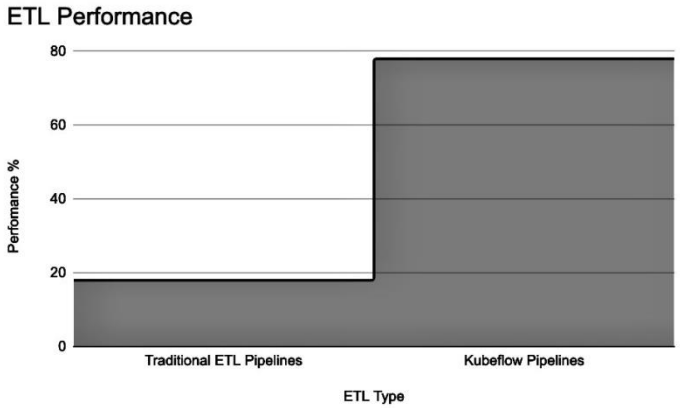

Рис. 3. Порівняння ефективності ETL

Оцінка продуктивності конвеєра показала значні досягнення в ефективності як для пакетної, так і для реальної обробки даних. У порівнянні з традиційними ETL-конвеєрами, рішення на основі Kubeflow зменшило час попередньої обробки даних на 60%, здебільшого завдяки паралельному виконанню завдань трансформації на кількох вузлах Kubernetes (Рис. 3). Інтеграція ADLS з оптимізованими операціями читання/запису дозволила покращити швидкість отримання даних на 40%, а використання Apache Spark на AKS ще більше покращило розподілене створення ознак та нормалізацію даних. Крім того, впровадження завдань, прискорених за допомогою GPU, для виявлення шахрайства на основі графів зменшило час обчислення ознак на 70%, що призвело до прискорення циклів навчання моделей.

Ефективність навчання моделей ШІ також покращилася завдяки оптимізації підготовки даних. Завдяки введенню версіонованих наборів даних за допомогою Kubeflow Metadata та Azure ML Datasets, відтворюваність експериментів зросла на 40%, що забезпечило, щоб оцінки продуктивності моделей ґрунтувалися на однакових вхідних даних. Це було особливо корисно для налаштування гіперпараметрів, де контрольовані варіації версій наборів даних дозволяли здійснювати детальну

оптимізацію продуктивності без непередбачених непослідовностей у даних.

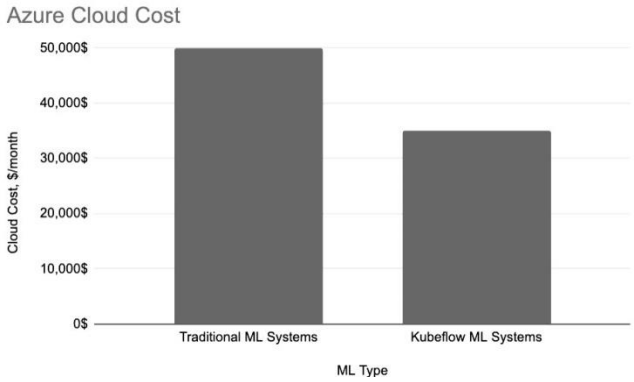

Рис. 4. Порівняння витрат на інфраструктуру за місяць

Крім того, впровадження горизонтального автоскейлінгу в AKS забезпечило динамічну корекцію обчислювальних ресурсів залежно від інтенсивності навантаження, що дозволило зменшити час простою обчислювальних потужностей і призвело до зниження витрат на хмарну інфраструктуру на 25% (Рис.4).

З цієї розгортки виникли кілька ключових уроків. По-перше, автоматизація обробки даних через Kubeflow Pipelines значно зменшила операційні витрати, мінімізуючи потребу в ручному втручанні та прискорюючи час до виходу AI-моделей на продуктивне середовище. По-друге, реально часова валідація даних і виявлення зміщень були необхідними для підтримки високої точності моделей, особливо в додатках, де розподіл даних швидко змінюється. Нарешті, оптимізація використання хмарних ресурсів через автоскейлінг та ефективні формати зберігання виявилися критичними для забезпечення балансу між продуктивністю та ефективністю витрат, що гарантує економічну життєздатність AI-робочих процесів з високою пропускною здатністю.

Висновки

Ця стаття досліджує основні стратегії підготовки даних у Kubeflow для хмарних AI-систем, підкреслюючи критичну роль масштабованих, автоматизованих і ефективних пайплайнів даних у сучасних робочих процесах машинного навчання. Від завантаження даних і попередньої обробки до валідації та версіонування, кожен етап пайплайна грає важливу роль у забезпеченні того, щоб AI-моделі отримували якісні, структуровані дані, оптимізовані для навчання та інференсу. Використовуючи Kubernetes і Kubeflow Pipelines, організації можуть автоматизувати і масштабувати ці процеси, покращуючи як операційну ефективність, так і продуктивність моделей.

Основні висновки цього дослідження підкреслюють важливість інтеграції хмарних рішень для зберігання даних, оптимізації робочих процесів трансформації даних за допомогою розподіленої обробки і забезпечення узгодженості даних через механізми реальновідомої валідації. Використання Kubeflow Metadata і інструментів версіонування даних виявилося ключовим для підтримки відтворюваності, а стратегії автоскейлінгу в AKS допомогли збалансувати продуктивність з ефективністю витрат. Аналіз випадку показав, що добре спроектовані пайплайни даних не тільки підвищують точність моделей, а й зменшують час навчання та витрати на інфраструктуру, роблячи їх основним компонентом успішних розгортань AI.

Інтеграція великих моделей AI основ та самонавчених моделей навчання ще більше стимулюватиме потребу у більш складних рішеннях для інженерії даних, здатних обробляти різноманітні та еволюціонуючі набори даних. Крім того, зростаюче впровадження інструментів спостережуваності на основі AI дозволить безперервно моніторити та оптимізувати робочі процеси підготовки даних, знижуючи відхилення моделей та покращуючи їх довгострокову надійність. Попри ці досягнення, кілька проблем залишаються відкритими. Забезпечення справедливості та зменшення упереджень AI-моделей на етапі підготовки даних є складною проблемою, що потребує подальших досліджень. Крім того, розробка більш ефективних методів обробки мультимодальних даних, поєднуючи структуровані, неструктуровані та реальновідомі потоки даних, залишається важливим викликом для інженерів AI. Нарешті, з розвитком регулювання AI організації повинні будуть впроваджувати більш надійні рамки управління даними, щоб відповідати новим стандартам без компромісів у ефективності.

Підсумовуючи, підготовка даних у Kubeflow є основою масштабованих і надійних AI-систем, надаючи структурований підхід до управління складнощами сучасних робочих процесів машинного навчання. Оскільки хмарний AI продовжує розвиватися, організації, що інвестують у оптимізовані практики інженерії даних, отримають конкурентну перевагу, забезпечуючи, щоб їхні моделі залишалися високопродуктивними, економічно ефективними і адаптованими до інновацій AI.

Подяка

Автори дякують редколегії наукового журналу «Вимірювальна техніка і метрологія» за підтримку.

Взаємні претензії авторів

Автори заявляють, що не мають фінансових або інших потенційних конфліктів щодо цієї роботи.

Список літератури

[1] Bershchanskyi, Y., & Klym, H. (2023, October). Information System for Administration of Medical Institution. In 2023 13th International Conference on Dependable Systems, Services and Technologies (DESSERT) (pp. 1-4). IEEE. https://doi.org/10.1109/DESSERT61349.2023.10416537

[2] Mehendale, P. (2023). Model Reliability and Performance through MLOps: Tools and Methodologies. J Artif Intell Mach Learn & Data Sci 2023, 1(4), 980-984. https://doi.org/10.51219/JAIMLD/pushkar

[3] Abbas, T., & Eldred, A. (2025). AI-Powered Stream Processing: Bridging Real-Time Data Pipelines with Advanced Machine Learning Techniques. ResearchGate Journal of AI & Cloud Analytics. https://doi.org/10.13140/RG.2.2.26674.52167

[4] Yuan, D. Y., & Wildish, T. (2020, June). Bioinformatics application with kubeflow for batch processing in clouds. In International conference on high performance computing (pp. 355-367). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-59851-8_24

[5] Bershchanskyi, Y., Klym, H., & Shevchuk, Y. (2024). Containerized artificial intelligent system design in cloud and cyber-physical systems., Advances in Cyber-Physical Systems (ACPS) 2024; Volume 9, Number 2 pp. 151 – 157. https://doi.org/10.23939/acps2024.02.151

[6] Yadavalli, T. Optimizing Machine Learning Workflows with Google Cloud Dataflow and TensorFlow Extended (TFX). J Artif Intell Mach Learn & Data Sci 2021, 1(1), 2436-2441. https://doi.org/10.51219/JAIMLD/tulasiram-yadavalli/524

[7] Kienzler, R., & Kyas, H. (2020, January). Tensorflow 2.0 and Kubeflow for Scalable and Reproducable Enterprise AI. In CS & IT Conference Proceedings (Vol. 10, No. 1). CS & IT Conference Proceedings.

[8] Devarasetty, N. (2024). Optimizing Data Engineering for AI: Improving Data Quality and Preparation for Machine Learning Application. The Computertech, 1-28.

[9] Mwangi, J., & Bablu, T. A. AI Microservices in Enterprise Applications: A Comprehensive Review of Use Cases and Implementation Frameworks.

[10] Bershchanskyi, Y., & Klym, H. (2024, October). Development Approaches of Cloud-Based System for Object Recognition on Images. In 2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET) (pp. 205-208). IEEE. https://doi.org/10.1109/TCSET64720.2024.10755838