DOI 10.31891/2307-5732-2024-331-25
УДК 004.65: 378

ПРИШЛЯК АНДРІЙ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0003-1681-5178
e-mail: andrii.a.pryshliak@lpnu.ua

АРХІТЕКТУРА ОЗЕРА ДАНИХ ДЛЯ ГАЛУЗІ ОСВІТИ

Проведено огляд напрацювань та досліджень науковців в контексті великих даних, та зокрема озер даних. В роботі наведено результати аналізу галузі освіти на всіх етапах освітнього процесу. Побудовано модель, що представляє можливі етапи освітнього процесу та дані, що безпосередньо їх стосуються. Сформовано модель архітектури озера освітніх даних, характеристику його складових та функціональних рівнів. Проаналізовано методи організації даних, для покращення аналітики та оптимізації пам'яті за допомогою колонкового формату даних та інструментів Spark

Ключові слова: озеро даних, архітектура озера даних, великі дані, метадані, освітні дані, сховища даних.

PRYSHLIAK ANDRII
Lviv Politechnic National University

DATA LAKE ARCHITECTURE IN THE EDUCATION AREA

An overview of the work and research of scientists in the context of big data, and in particular data lakes, was conducted. The lake is primarily considered as a certain repository for further data analysis. The data lake model focuses on the concept of what information needs to be stored, rather than what data is actually needed or what its purpose is. The existing methods of organizing the architecture of the data lake, both on the basis of basic and modified ones, based on the use of various possibilities related to graphs, semantic networks, and ontologies, have been worked out. Such approaches form functionally oriented models of architectures, as well as the possibility of creating new hybrid architectures with specialized metadata management tools. Metadata includes working with data and objects at different levels of detail. The granularity is strongly related to the concept of data lakes, most often in the aspects of data recognition of different entities. Metadata itself is information about the data and processes that the data lake collects and requires separate management mechanisms. In recent years, several such mechanisms have been introduced that focus on categorization or list metadata management functions, or a combination thereof. The work presents the results of the analysis of the field of education at all stages of the educational process. The characteristics and features of each stage of the educational process are provided and data repositories are built, for a better understanding of each of them and the construction of the data lake architecture. A model representing the possible stages of the educational process and the data directly related to them was built. The concept of a complete portrait of the characteristics of the student of education, which should provide information about him, based on the completed stages of the educational process, including both formal and informal education, is introduced. A formal representation of the educational data lake is presented, describing the main elements that should be included in the architecture model. A model of the architecture of the lake of educational data, a description of its components and functional levels has been formed. Analyzed data organization techniques to improve analytics and memory optimization using columnar data format and Spark tools.

Keywords: data lake, data lake architecture, big data, metadata, educational data, data warehouses.

Постановка проблеми

Впродовж останніх років, виникло багато проблем, пов'язаних з освітніми процесами та роллю, яку вони відіграють у цьому контексті. Особливо важливими є дані, що описують як окремих учасників освітнього процесу, так і систему освіти в цілому.

Незалежно від рівня підготовки учнів або акредитації навчальних закладів, найбільш поширеною є проблема доступності та якості освітнього середовища, а також різноманітних чинників, що можуть впливати на процес навчання та зацікавленість у здобутті кваліфікації та відповідних знань. Також серйозною проблемою є актуальність професій та можливість навчання спеціалістів з використанням сучасних рішень та методів, щоб вони мали відповідні знання та навички для задоволення потреб ринку праці.

Аналіз останніх досліджень

Озеро даних – централізоване сховище різнотипної інформації, структурованого, напівструктурованого і неструктурованого типу із спеціалізованою архітектурою та системою керування даними.. Найбільш стандартні архітектури із яких починається розвиток озер даних це – загальна(містить основні інструменти для роботи з даними), зонна(орієнтована на розподілення виконання робіт) та резервуарна(збирає інформацію, а тоді опрацьовує)[0]. Така категоризація надає певний базис розуміння та побудови озер даних, але може бути не надто чіткою. Тому й пропонують нові рішення, пов'язані з семантичними мережами, та онтологіями, багатошаровими графами[0,0], навігаційними графами-моделями[0] що можуть покращувати якість даних.

У зв'язку з постійним розвитком озер даних, з'являються нові підходи до їх проектування і покращення. На даний момент існує багато варіантів, щодо організації архітектур озер даних, на основі базових. Серед таких, із вагомими напрацюваннями є трирівнева функціонально-орієнтована модель[0]. Інша можливість – гібридні архітектури (поєднують функціональні та на основі зрілості даних) із інструментами управлінням метаданими[0]. Схожою за організацією, а власне гнучкою є архітектура на основі FAIR Digital Objects(FDO), із більш прямою взаємодією умовних користувача та системи[0].

Перспективним є підхід до побудови із використанням розподіленої потокової платформи подій Apache Kafka[0], але більше для завдань корпоративного спрямування. Для подолання деяких проблем класифікації походження даних у структурах як озера, пропонується використання Data Mesh (сітка даних)[0].

Метадані – невід'ємна частина будь-яких структур, що мають стосунок, до роботи із даними, особливо сховища та озера даних. Основна функція метаданих, надавати правильного контексту наявній інформації. У випадку роботи з озерами даних, використовують функціональну класифікацію, де розрізняють 3 категорії метаданих, а саме технічні, операційні та бізнес-метадані, що описують дані певної області досліджень:

$$M = \{M_T, M_O, M_B\}, \quad (1)$$

де M – загальна множина метаданих, M_T – технічні, M_O – операційні, M_B – бізнес-метадані.

Часто науковці пропонують покращення певних елементів уже наявних підходів, що особливо стосується роботи із метаданими. Тут варто звернути увагу на спеціально створену загальну модель управління метаданими HANDLE[0], що має на меті отримати весь контроль над ними в межах озера. Для покращення взаємодії з озером, створено інструмент RONIN[0], що виконує пошук по наборах даних і працює з навігацією в ієрархічній структурі. Часто важливим фактором є методи управління даними[0] і вони особливо можуть впливати на усі етапи їх опрацювання в умовній структурі. Ще один підхід пропонує створення DLAF(Data Lake Architecture Framework) - загальної структури для побудови озера даних, залежно від потреб[0].

Щодо більш галузевого застосування озер даних звернемо увагу на архітектуру для авіаційної сфери[0]. В даному випадку, присутні доволі стандартні елементи для роботи озера, але також особливе місце відведено для інфраструктури галузі. Інший варіант – архітектура озера, для сфери археології[0], що характеризується різноманітністю даних за типами та походженням. Вказані елементи цих архітектур матимуть застосування і в освітній галузі.

Сховище даних – структурована колекція даних, яка включає в себе необхідні елементи для зберігання, організації та управління інформацією:

$$DW = \{DB, rf, RF, rm, RM, func\}, \quad (2)$$

де DB– множина відношень вхідних даних, rf - множини відношень фактів, RF – схема множини відношень фактів, rm - множини відношень метаданих, RM – схема множини відношень метаданих, $func$ – множина процедур опрацювання даних.

Вважається, що сховища даних, з часом вичерпають свою актуальність, тому розглядається їх прогрес через взаємодію з озерами даних, формуючи зовсім нову структуру – Lakehouse[0]. Звичайно це матиме вплив і на архітектуру озер, особливо на етапах отримання та виводу даних відповідно до поставлених завдань.

Формулювання цілі статті

Мета статті – аналіз можливих етапів здобуття освіти та формування архітектури озера даних, для опису їх даталогічної структури та функціоналу.

Об'єкт дослідження – структура рівнів та етапів освітнього процесу.

Предмет дослідження – організація умов освітнього процесу, на кожному його етапі.

Завдання статті:

- Проаналізувати структуру та дані, для кожного можливого етапу освіти.
- Створити моделі сховищ даних для можливих етапів освіти.
- Надати характеристику освітнім процесам на кожному етапі.
- Сформувати модель озера для освітніх даних
- Вивести формальне представлення озера освітніх даних

Для початку, розглянемо етапи освітнього середовища, відповідно до рівнів освіти, починаючи від найнижчого, через призму сховища даних, кожного з них.

Найнижчий освітній рівень який можна піддати аналізу – дошкільні навчальні заклади(ДНЗ).

Даний рівень (Рис.1) поки існує не зовсім, як складова освітнього процесу і може не мати чіткого спрямування розвитку дітей, орієнтованого на власне освіту, а більше на виховання і закладання мінімальних навичок спілкування та сприймання інформації. Тож на цьому етапі все таки проходить певний академічний та особистісний розвиток, а також можливі навички інших типів діяльності. Формально описати це сховище даних можна формулою:

$$DW_{ДНЗ} = \{DB_{ДНЗ}, RF, RM, func\}, \quad (3)$$

де $DB_{ДНЗ}$– множина відношень даних ДНЗ та інших пов'язаних даних, RF – схема множини відношень фактів, RM – схема множини відношень метаданих, $func$ – множина процедур опрацювання даних.

Наступний рівень – середня освіта (школи, гімназії, ліцеї). Цей рівень – складова освітнього процесу і є обов'язковою частиною формування базових навичок здобувача освіти(Рис.2).

На даному етапі є більше інформації для аналізу, оскільки розглядається інформація про здобувачів освіти, заклад освіти, стандарти та наповнення навчальних програм, кваліфікацію педагогів. Тут з'являється можливість формувати портрет розвитку та можливих навичок здобувача освіти, враховуючи час, протягом якого відбувається навчання. Формальне представлення сховища схоже з ДНЗ:

$$DW_{ЗСО} = \{DB_{ЗСО}, RF, RM, func\}, \quad (4)$$

де $DB_{ЗСО}$ – множина відношень даних ЗСО та інших пов'язаних даних, RF – схема множини відношень фактів, RM – схема множини відношень метаданих, $func$ – множина процедур опрацювання даних.

Після здобуття середньої освіти(базової чи повної) – наступні рівні, це відповідно професійно-технічна, фахова передвища та вища освіта. Кожен з цих етапів може бути кінцевим, залежно від бажання здобувача освіти чи інших факторів, що можуть на це впливати.

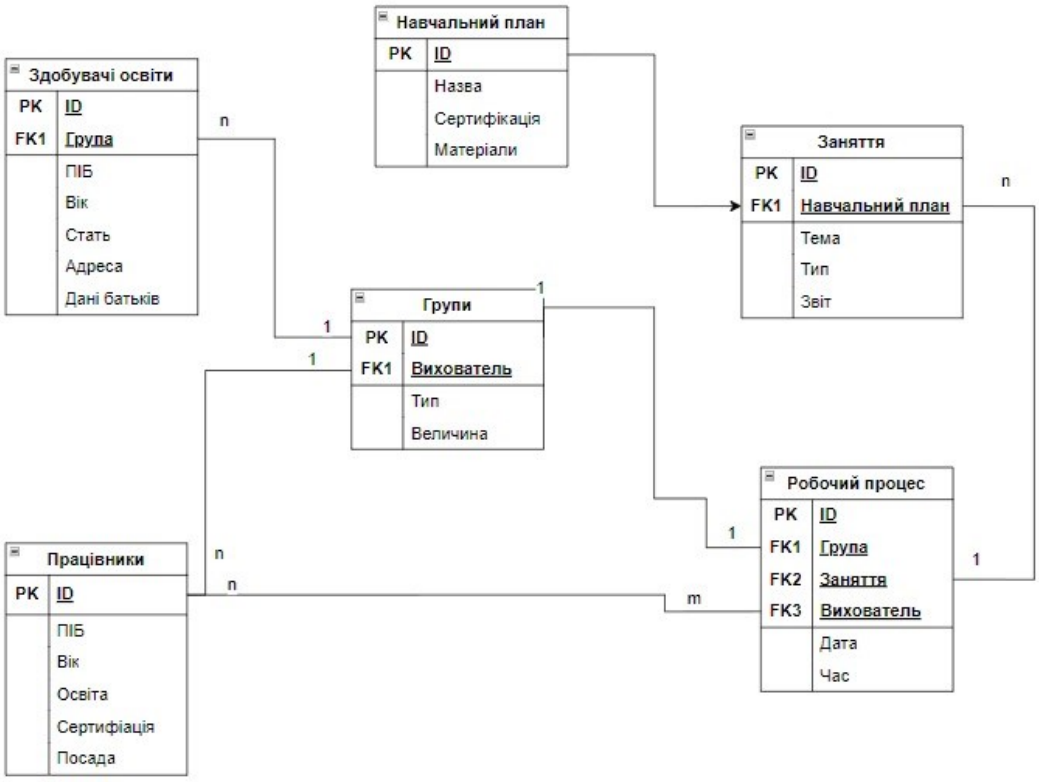

Рис. 1. Сховище даних для ДНЗ

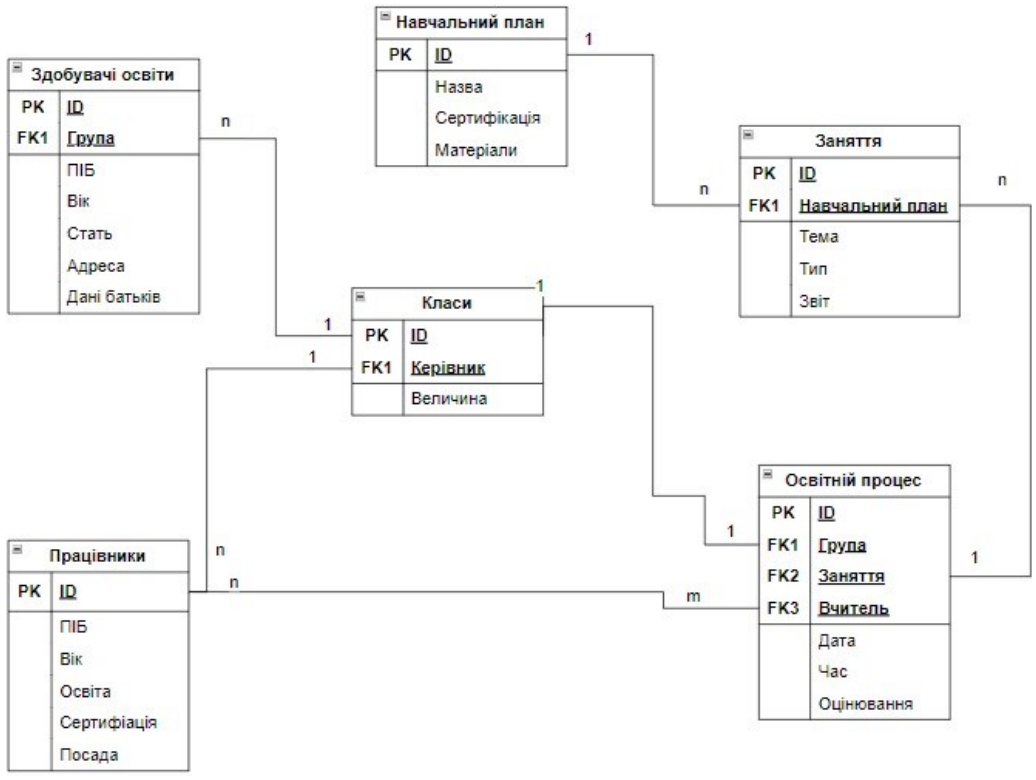

Рис. 2. Сховище даних для ЗСО

У випадку професійно-технічної освіти (Рис.3) – є потреба швидко підготувати спеціаліста, який може виконувати певну категорію робіт, відповідно до обраного напряму підготовки. Тому тут велику

частину даних потрібно розглядати через контекст ринку праці та співпрацю із роботодавцями. Формальне представлення сховища:

$$DW_{ЗПТО} = \{DB_{ЗПТО}, RF, RM, func\}, \quad (5)$$

де $DB_{ЗПТО}$ – множина відношень даних ЗПТО та інших пов'язаних даних, RF – схема множини відношень фактів, RM – схема множини відношень метаданих, $func$ – множина процедур опрацювання даних.

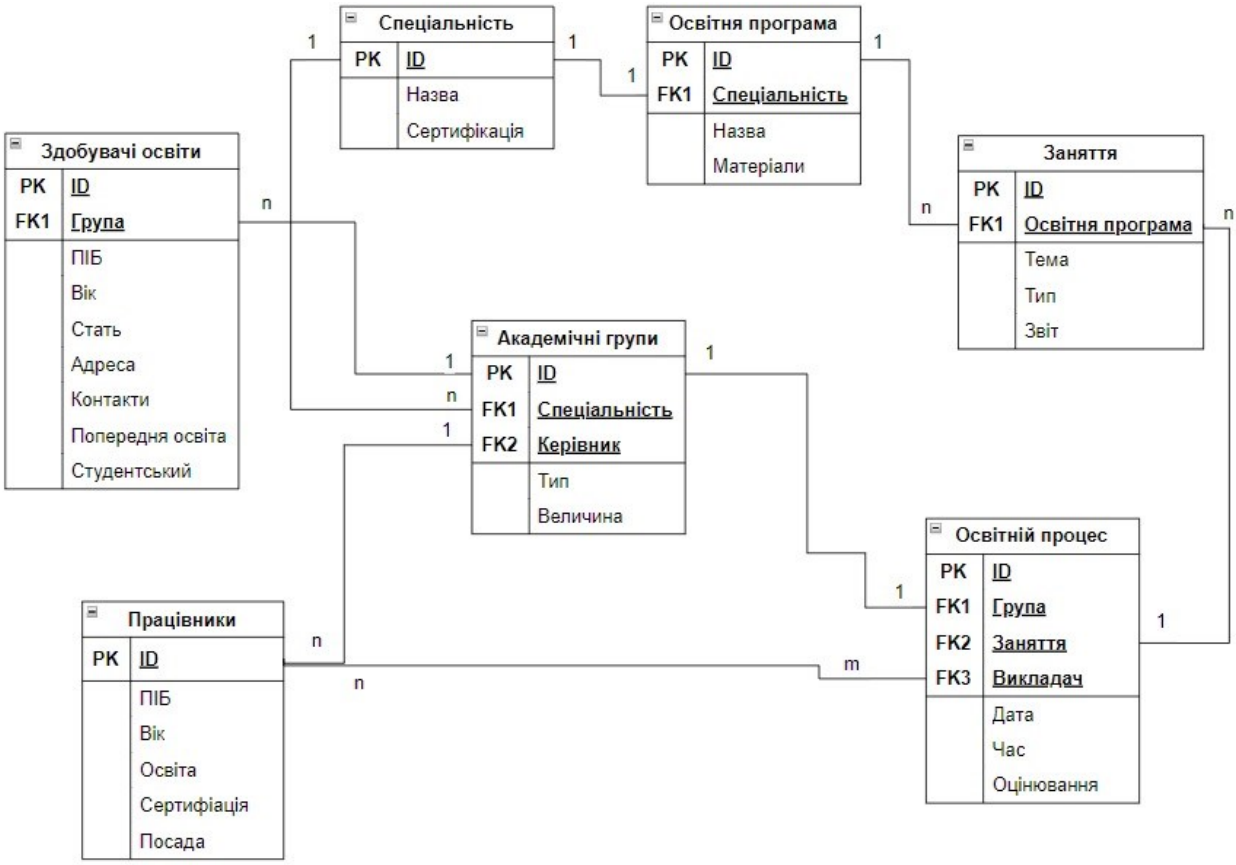

Рис. 3. Сховище даних для ЗПТО

Фахова передвища освіта (Рис.4) – спрямована на формування певної кваліфікації, що надає змогу готувати фахівців для більш спеціалізованих завдань у відповідній сфері діяльності. Варто зауважити, що часто здобувачі освіти таких закладів можуть знайти фахову зайнятість доволі швидко, тому дані ринку праці мають відповідну вагу. Паралельно студенти можуть отримувати освіту в інших навчальних закладах. Формальне представлення:

$$DW_{ЗФПО} = \{DB_{ЗФПО}, RF, RM, func\}, \quad (6)$$

де $DB_{ЗФПО}$ – множина відношень даних ЗФПО та інших пов'язаних даних, RF – схема множини відношень фактів, RM – схема множини відношень метаданих, $func$ – множина процедур опрацювання даних.

Заклади вищої освіти (Рис.5) – надають можливість отримувати освітні ступені бакалавра та магістра, а також наукові ступені. Дані заклади мають забезпечувати здобувачу освіти вищу кваліфікацію залежно від обраної галузі. А враховуючи наукову складову, потрібно опрацьовувати дані відповідних наукометричних баз. Важливо слідкувати за актуальністю такої освіти, зокрема наповнення та оновлення освітніх програм.

Формальне представлення:

$$DW_{ЗВО} = \{DB_{ЗВО}, RF, RM, func\}, \quad (7)$$

де $DB_{ЗВО}$ – множина відношень даних ЗВО та інших пов'язаних даних, RF – схема множини відношень фактів, RM – схема множини відношень метаданих, $func$ – множина процедур опрацювання даних.

Заклади вищої освіти (Рис.5) – надають можливість отримувати освітні ступені бакалавра та магістра, а також наукові ступені. Дані заклади мають забезпечувати здобувачу освіти вищу кваліфікацію залежно від обраної галузі. А враховуючи наукову складову, потрібно опрацьовувати дані відповідних наукометричних баз. Важливо слідкувати за актуальністю такої освіти, зокрема наповнення та оновлення освітніх програм.

Формальне представлення:

$$DW_{ЗВО} = \{DB_{ЗВО}, RF, RM, func\}, \quad (7)$$

де $DB_{ЗВО}$ – множина відношень даних ЗВО та інших пов'язаних даних, RF – схема множини відношень фактів, RM – схема множини відношень метаданих, $func$ – множина процедур опрацювання даних.

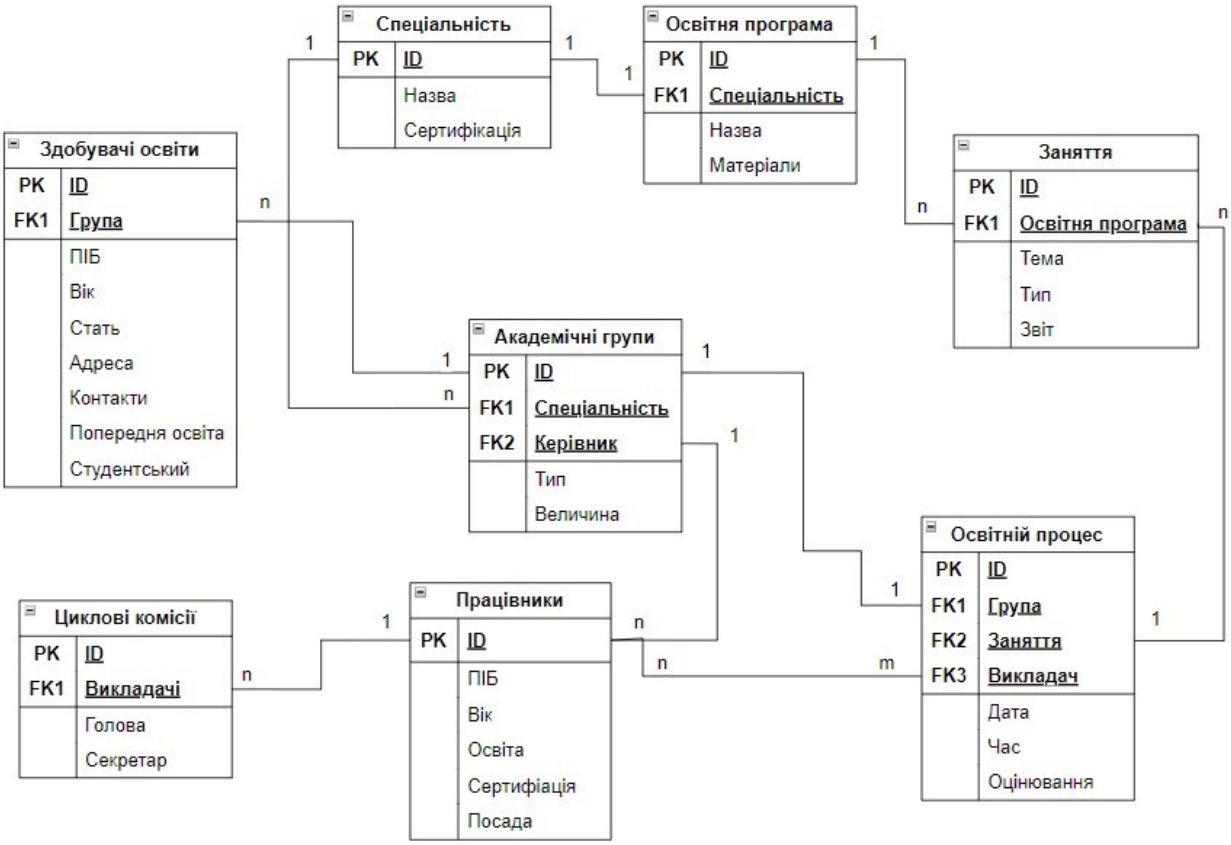

Рис. 4. Сховище даних для ЗФПО

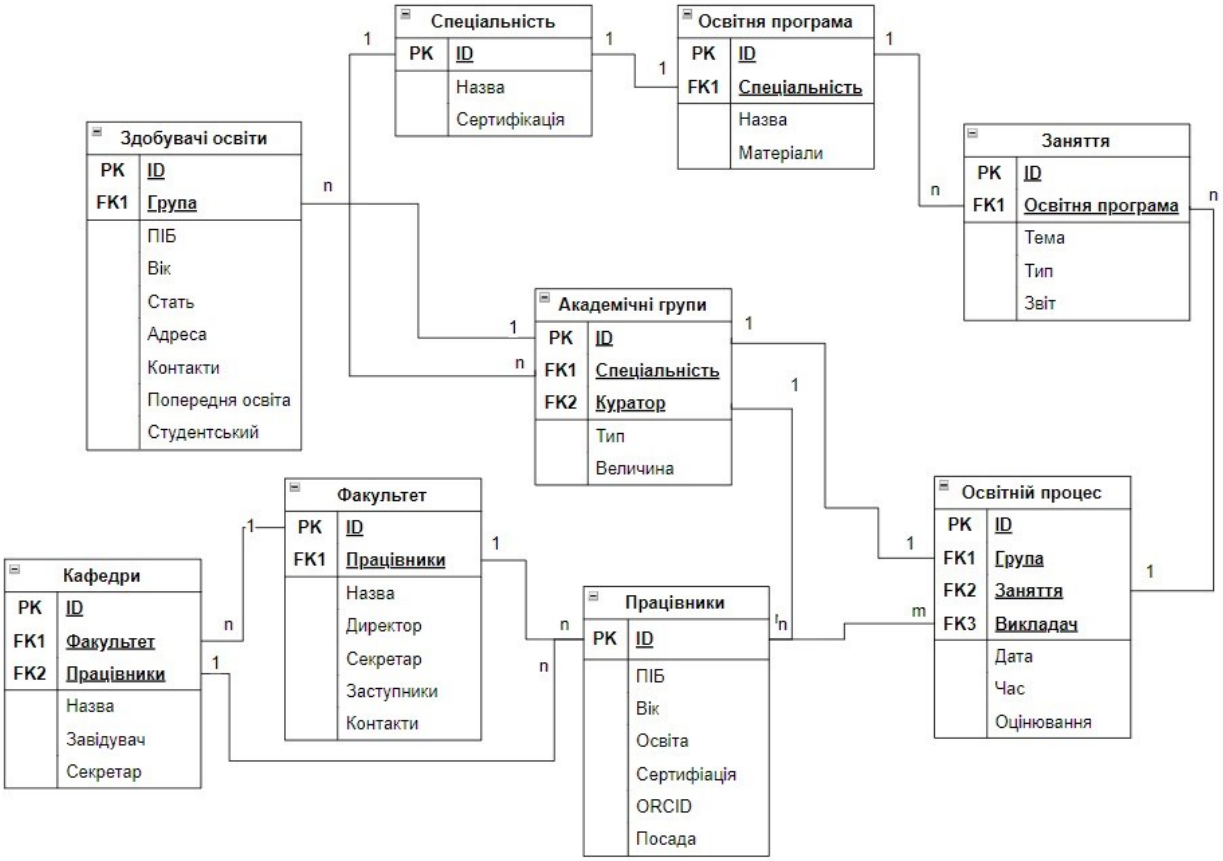

Рис. 5. Сховище даних для ЗВО

Враховуючи усе вищезгадане, можемо сформувати певну модель даних загальної взаємодії для здобувача освіти, який проходить через усі етапи освітнього процесу(Рис.6).

Для кожного з цих етапів, додатковим варіантом покращення характеристик та атрибутів буде неформальна освіта. Особливість неформальної освіти – доступність на всіх етапах освітнього процесу, як безпосередньо під час здобуття так і поза ним.

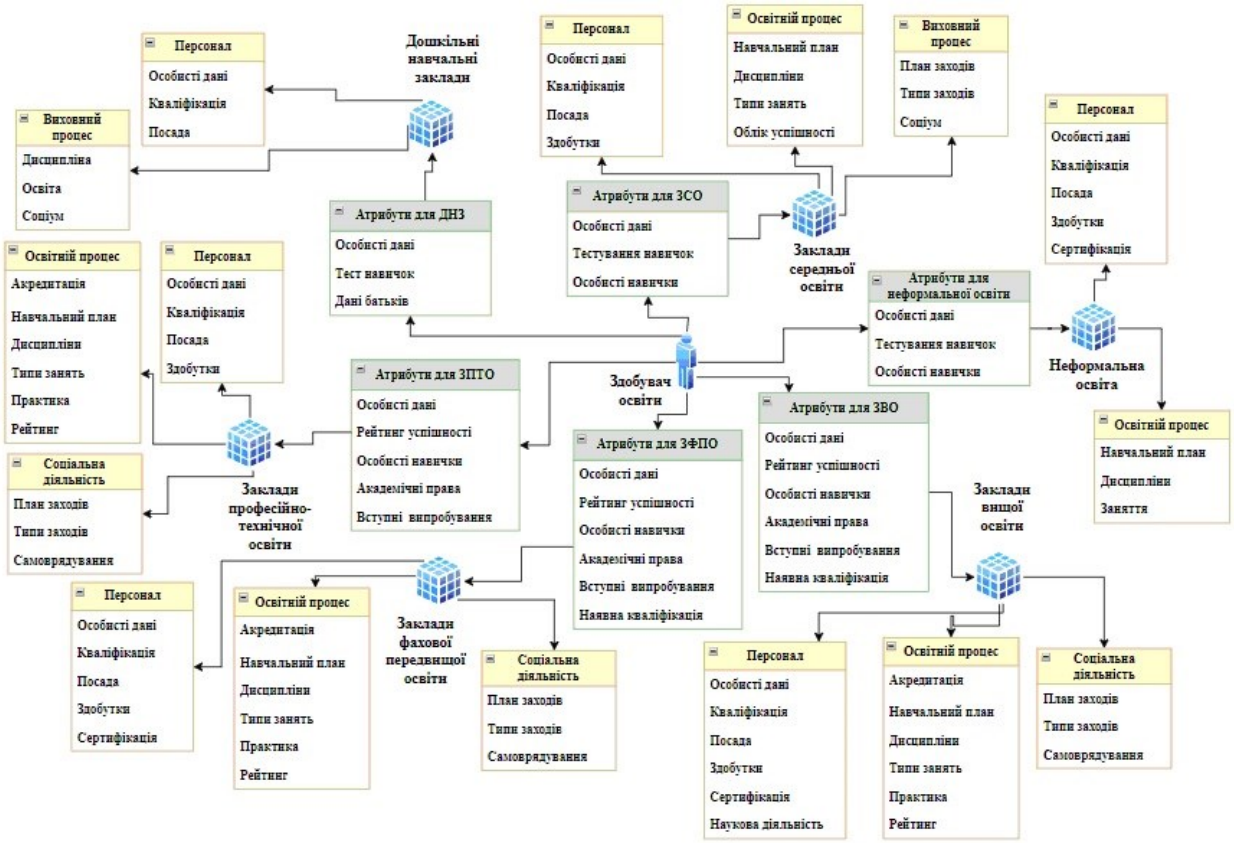

Рис. 6. Характеристика можливих етапів освітнього процесу

Виклад основного матеріалу

Ми розглядаємо весь комплексний шлях освітнього процесу, тому варто розділити його на основні категорії:

1. Отримання базових знань та академічних прав(ДНЗ, ЗСО), а також цей етап, зокрема середня освіта є обов'язковою.
2. Отримання фахових знань і професійних та академічних прав(ЗПТО, ЗФПО, ЗВО)
3. Неформальна освіта(НфО)

В даному випадку будемо використовувати поняття «Атрибут» - своєрідна характеристика портрету здобувача освіти, що містить різноманітну інформацію, чи міститиме максимально усю потрібну інформацію про нього. Зауважимо, що також не завжди потрібна повна характеристика навичок, а тому велика частина знань чи навичок згодом просто не потрібна чи немає сенсу їх застосування.

Тепер детальніше розглянемо розподіл атрибутів із моделі на Рис.6:

- $A_{ДНЗ}$ – множина атрибутів для ДНЗ – тобто початковий портрет характеристик здобувача освіти. На цьому етапі важко говорити, про конкретні навички чи їх важливість, але можна встановити хист у певних напрямах розвитку.
- $A_{ЗСО}$ – множина атрибутів для ЗСО – елементарний портрет характеристик здобувача освіти. Варто зауважити, що ці атрибути вже мають відповідати певним вимогам для навчання і залежно від рівня цих елементарних навичок, формуватиметься якість набуття нових знань.
- $A_{ЗПТО}$ – множина атрибутів для ЗПТО – базовий або сформований портрет характеристик здобувача освіти. За основу в будь-якому випадку береться до уваги інформація щодо базового рівня освіти та відповідних атрибутів.
- $A_{ЗФПО}$ – множина атрибутів для ЗФПО – аналогічно базовий або сформований портрет характеристик здобувача освіти. Також можливість використовувати атрибути здобуті на рівні освіти в ЗПТО.
- $A_{ЗВО}$ – множина атрибутів для ЗВО – знову ж таки як основа – базовий або сформований портрет характеристик. Крім того, можливість використання атрибутів попередніх рівнів освіти чи уже наявних вищих. Основна відмінність від попередніх рівнів – підготовка наукових кадрів.

- $А_{НФО}$ – множина атрибутів, потрібна для неформальної освіти, але власне ці атрибути можуть бути дуже обмеженими, оскільки не завжди вимагають попереднього підґрунтя навичок.

На цьому можна завершити перелік, але лише тих елементів, що є вхідними, для кожного типу освіти, тому варто ввести ще один.

- $A_П$ – множина атрибутів, що характеризуватиме здобувача освіти на кінцевому етапі освітнього процесу.

Таким чином, отримуємо можливість сформувати певну модель, для знаходження повного портрету характеристик здобувача освіти:

$$A_{ПЗО} = \{A_{ДНЗ}, A_{ЗСО}, A_{ЗПТО}, A_{ЗФПО}, A_{ЗВО}, A_{НФО}, A_П\}, \quad (8)$$

де $A_{ПЗО}$ – множина атрибутів, що характеризує повний портрет характеристик здобувача освіти, за весь час освітньої діяльності, на всіх можливих етапах.

Планується, що вищевказані множини атрибутів, зможуть стати прототипом, для надання кількісних та якісних характеристик даних, що буде отримувати та опрацьовувати озеро даних.

Освітні дані можуть мати багато різних типів, наприклад, документи, зображення, дані моніторингів та спостережень. Опис певного освітнього об'єкта також відрізняється за користувачами, використанням і часом. Така різноманітність освітніх даних викликає багато наукових проблем, пов’язаних із зберіганням неоднорідних даних у централізованому сховищі, гарантуванням якості даних, очищенням і перетворенням даних, щоб зробити їх сумісними, ефективними і доступними, а також знайти відповідне розташування в системі освіти. Відповідно можемо сформувати формальне представлення озера освітніх даних:

$$DL_E = \{DB, DW, Wb, Sd, Ud, Gr, Int, Se, Wo\}, \quad (9)$$

де DB – множина спеціалізованих баз даних, DW – множина сховищ даних, Wb – множина даних статичних Web-сторінок, Sd – множина напівструктурованих даних, Ud – множина неструктурованих даних, Gr – графічних та мультимедійних даних, Int – множина механізмів інтеграції, Se – множина засобів пошуку, Wo - засоби опрацювання інформації. Формуємо модель архітектури озера даних для галузі освіти(Рис.7):

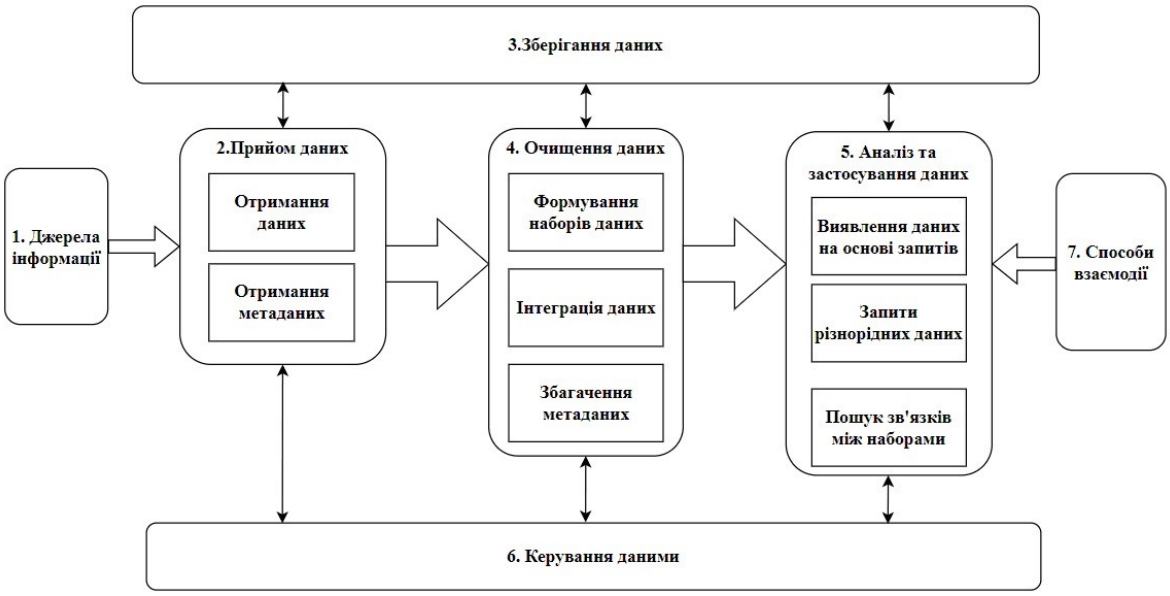

Рис. 7. Модель архітектури озера даних для галузі освіти

Щоб подолати складнощі організації єдиної системи освітніх даних, будемо опиратися на певні основні рівні роботи озера, що присутні в більшості архітектур, із підходом до збору всіх можливих типів освітніх даних, збереження їх в озері даних і використання метаданих для кращого розуміння:

- Опрацювання джерела даних – основні властивості джерел даних, наприклад, обсяг, формат, швидкість, зв’язність.
- Прийом даних – набір інструментів для правильної організації підбору даних. Тут же відбувається і отримання метаданих.
- Зберігання даних – основа озера даних, тобто інструменти, що надають змогу зберігати дані в будь-якому форматі.
- Очищення даних – інструменти для очищення даних, як від помилок так і відповідно до потреб певного завдання та метаданих.
- Аналіз та застосування даних – інструменти для опрацювання даних і формування якісної оцінки, відповідно до заданої цілі

- Управління даними – інструменти для відслідковування якості виконання процесів роботи з даними.
- Способи взаємодії – інструменти взаємодії із системою озера даних.

Інший важливий аспект, як для оптимізації зберігання даних, так і їх аналітичного опрацювання, це формат представлення даних. Враховуючи особливості галузі та різноманіття даних, варто використовувати колонковий(columnar) формат даних, що на відміну від традиційного рядкового(row), значно оптимізовує аналітику даних, та затрати пам'яті[0]. В питанні аналітики даних, одними із основних інструментів є засоби Spark, що надають численні інструменти аналізу та можливості багатопотокового опрацювання даних.

Висновки

У статті проведено огляд останніх досліджень щодо розвитку та застосування архітектур озер даних. Розглянуто різноманітні варіанти модифікацій як для цілих галузей, так і окремих елементів озера.

Досліджено етапи навчального процесу, через який проходить здобувач освіти в Україні та створено загальні моделі сховищ даних, для кожного з них. Подано формальний опис сховищ даних, кожного етапу освіти. Запропоновано модель, що міститиме усю інформацію, щодо освітнього процесу, через який проходить здобувач освіти.

Досліджено потреби та вимоги щодо архітектури озера, для роботи у сфері освіти. Подано основні характеристики рівнів реалізації та організації озера. Побудовано формальну модель, що описує основні елементи, які мають бути задіяні в озері даних. Подано характеристику формату роботи з даними.

Література

1. Hai, R., Miller, R., Jarke, M., & Quix, C. J. (2020). *Data Integration and Metadata Management in Data Lakes* (Doctoral dissertation, Ph. D. Dissertation. RWTH Aachen University. https://doi. org/10.18154/RWTH-2020-08233).

2. Piantella, D. (2022). A Research on Data Lakes and their Integration Challenges. In The 30th Italian Symposiumon Advanced Database Systems.

3. Cayeux, E., Damski, C., Macpherson, J., Laing, M., Annaiyappa, P., Harbidge, P., ... & Carney, J. (2022).

4. Connecting Multilayer Semantic Networks to Data Lakes: The Representation of Data Uncertainty and Quality. SPE Drilling & Completion, 1-16.

5. Nargesian, F., Pu, K. Q., Zhu, E., Ghadiri Bashardoost, B., & Miller, R. J. (2020, June). Organizing data lakes for navigation. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 1939-1950).

6. Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data Lakes: A Survey of Functions and Systems. IEEE Transactions on Knowledge and Data Engineering.

7. Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. Journal of Intelligent Information Systems, 56, 97-120.

8. Nolte, H., & Wieder, P. (2022). Realising data-centric scientific workflows with provenance-capturing on data lakes. Data Intelligence, 4(2), 426-438.

9. Peddireddy, K. (2023). Kafka-based Architecture in Building Data Lakes for Real-time Data Streams. International Journal of Computer Applications, 185(9), 1-3.

10. Machado, I. A., Costa, C., & Santos, M. Y. (2022). Data mesh: concepts and principles of a paradigm shift in data architectures. Procedia Computer Science, 196, 263-271

11. Eichler, R., Giebler, C., Gröger, C., Schwarz, H., & Mitschang, B. (2020). Handle-a generic metadata model for data lakes. In Big Data Analytics and Knowledge Discovery: 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 22 (pp. 73-88). Springer International Publishing.

12. Ouellette, P., Sciortino, A., Nargesian, F., Bashardoost, B. G., Zhu, E., Pu, K. Q., & Miller, R. J. (2021). RONIN: data lake exploration. Proceedings of the VLDB Endowment, 14(12).

13. Brous, P., Janssen, M., & Krans, R. (2020, April). Data governance as success factor for data science. In Conference on e-Business, e-Services and e-Society (pp. 431-442). Cham: Springer International Publishing.

14. Giebler, C., Gröger, C., Hoos, E., Eichler, R., Schwarz, H., & Mitschang, B. (2021, September). The data lake architecture framework: a foundation for building a comprehensive data lake architecture. In Conference for Database Systems for Business, Technology and Web (BTW) (Vol. 70469).

15. Sun, J., Gui, G., Sari, H., Gacanin, H., & Adachi, F. (2020). Aviation data lake: Using side information to enhance future air-ground vehicle networks. IEEE Vehicular Technology Magazine, 16(1), 40-48.

16. Darmont, J., Favre, C., Loudcher, S., & Noûs, C. (2020, October). Data lakes for digital humanities. In Proceedings of the 2nd International Conference on Digital Tools & Uses Congress (pp. 1-4).

17. Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR (Vol. 8).

18. Jin, G., Bian, H., Chen, Y., & Du, X. (2022). Columnar Storage Optimization and Caching for Data Lakes. In EDBT (pp. 2-419).

References

1. Hai, R., Miller, R., Jarke, M., & Quix, C. J. (2020). *Data Integration and Metadata Management in Data Lakes* (Doctoral dissertation, Ph. D. Dissertation. RWTH Aachen University. https://doi. org/10.18154/RWTH-2020-08233).

2. Piantella, D. (2022). A Research on Data Lakes and their Integration Challenges. In The 30th Italian Symposium on Advanced Database Systems.

3. Cayeux, E., Damski, C., Macpherson, J., Laing, M., Annaiyappa, P., Harbidge, P., ... & Carney, J. (2022). Connecting Multilayer Semantic Networks to Data Lakes: The Representation of Data Uncertainty and Quality. SPE Drilling & Completion, 1-16.

4. Nargesian, F., Pu, K. Q., Zhu, E., Ghadiri Bashardoost, B., & Miller, R. J. (2020, June). Organizing data lakes for navigation. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 1939-1950).

5. Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data Lakes: A Survey of Functions and Systems. IEEE Transactions on Knowledge and Data Engineering.

6. Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, *56*, 97-120.

7. Nolte, H., & Wieder, P. (2022). Realising data-centric scientific workflows with provenance-capturing on data lakes. Data Intelligence, 4(2), 426-438.

8. Peddireddy, K. (2023). Kafka-based Architecture in Building Data Lakes for Real-time Data Streams. *International Journal of Computer Applications*, *185*(9), 1-3.

9. Machado, I. A., Costa, C., & Santos, M. Y. (2022). Data mesh: concepts and principles of a paradigm shift in data architectures. *Procedia Computer Science*, *196*, 263-271

10. Eichler, R., Giebler, C., Gröger, C., Schwarz, H., & Mitschang, B. (2020). Handle-a generic metadata model for data lakes. In *Big Data Analytics and Knowledge Discovery: 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 22* (pp. 73-88). Springer International Publishing.

11. Ouellette, P., Sciortino, A., Nargesian, F., Bashardoost, B. G., Zhu, E., Pu, K. Q., & Miller, R. J. (2021). RONIN: data lake exploration. *Proceedings of the VLDB Endowment*, *14*(12).

12. Brous, P., Janssen, M., & Krans, R. (2020, April). Data governance as success factor for data science. In *Conference on e-Business, e-Services and e-Society* (pp. 431-442). Cham: Springer International Publishing.

13. Giebler, C., Gröger, C., Hoos, E., Eichler, R., Schwarz, H., & Mitschang, B. (2021, September). The data lake architecture framework: a foundation for building a comprehensive data lake architecture. In *Conference for Database Systems for Business, Technology and Web (BTW)* (Vol. 70469).

14. Sun, J., Gui, G., Sari, H., Gacanin, H., & Adachi, F. (2020). Aviation data lake: Using side information to enhance future air-ground vehicle networks. *IEEE Vehicular Technology Magazine*, *16*(1), 40-48.

15.

16. Darmont, J., Favre, C., Loudcher, S., & Noûs, C. (2020, October). Data lakes for digital humanities. In Proceedings of the 2nd International Conference on Digital Tools & Uses Congress (pp. 1-4).

17. Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR* (Vol. 8).

18. Jin, G., Bian, H., Chen, Y., & Du, X. (2022). Columnar Storage Optimization and Caching for Data Lakes. In *EDBT* (pp. 2-419).