УДК 004.62:004.67:336.76

КОНЦЕПЦІЯ ГІБРИДНОЇ МОДЕЛІ РОЗВІДУВАЛЬНОГО АНАЛІЗУ РІЗНОРІДНИХ СЛАБОСТРУКТУРОВАНИХ ФІНАНСОВО-ІНВЕСТИЦІЙНИХ ДАНИХ ВЕЛИКИХ ОБСЯГІВ

Рудніченко М.Д., к.т.н., доц.;[1] Шибаєва Н.О., к.т.н., доц.; [2] Шведов Д.В. [3]
[1,2,3] Національний університет «Одеська політехніка»
[1,2,3] Україна, Одеса
[1]nickolay.rud@gmail.com; [2]denscreamer@gmail.com; [3]studylearnerstudy@gmail.com

Анотація. *У роботі запропоновано концепцію гібридної моделі розвідувального аналізу різнорідних слабоструктурованих фінансово-інвестиційних даних великих обсягів. Модель поєднує технології MapReduce, Hive та OLAP із статистичним аналізом у середовищі Python. Розроблений підхід забезпечує масштабованість, адаптивність і підвищену точність аналітичних висновків.*
Ключові слова: *великі дані; розвідувальний аналіз даних; аналіз інвестицій; слабоструктуровані дані.*

Вступна частина. У сучасних умовах цифрової трансформації фінансово-економічних систем актуалізується проблема ефективного опрацювання великих обсягів даних, що формуються у сфері фінансово-інвестиційної діяльності [1]. Дані, які циркулюють у фінансових аналітичних середовищах, характеризуються значною різнорідністю, високим рівнем динамічності та слабкою структурованістю [2]. Традиційні методи аналітичної обробки не забезпечують необхідної масштабованості та продуктивності при роботі з великими потоками інформації, що потребує розроблення нових підходів до реалізації моделей розвідувального аналізу [3]. Одним із таких напрямів є створення гібридних моделей, які інтегрують концепції розподіленої обробки даних, інструменти аналітичних систем типу OLAP, а також статистичні методи для виявлення прихованих закономірностей у слабоструктурованих масивах фінансово-інвестиційної інформації.

Проблематика інтелектуального аналізу фінансових даних розвивається на перетині технологій Big Data, Data Mining, Business Intelligence та Financial Engineering [4]. Сучасні дослідження вказують на те, що основною складністю при аналізі фінансово-інвестиційних даних є їхня семантична неоднорідність, а також різна природа джерел — від традиційних бухгалтерських звітів і біржових індикаторів до потоків новин, прогнозів та соціально-економічних сигналів. Методи традиційної статистики виявляються малоефективними через необхідність попереднього очищення та структуризації інформації. У той же час підходи, засновані на розподілених обчисленнях, таких як MapReduce та Hive, дозволяють ефективно масштабувати аналітичні процеси, але потребують адаптації для роботи із слабоструктурованими та семантично складними наборами даних.

Актуальність використання OLAP-моделей у контексті розвідувального аналізу фінансово-інвестиційної інформації визначається можливістю багатовимірного представлення даних, яке суттєво підвищує якість інтерпретації результатів аналітичної обробки. Проте класичні OLAP-системи не передбачають прямої інтеграції з розподіленими обчислювальними фреймворками та не орієнтовані на роботу з неструктурованими даними. Тому в останні роки активно досліджується напрям побудови гібридних аналітичних моделей, що поєднують можливості OLAP-архітектур і масштабованих середовищ обробки, заснованих на парадигмі MapReduce.

Метою роботи є розроблення концепції гібридної моделі розвідувального аналізу, здатної забезпечити ефективне виявлення прихованих закономірностей у різнорідних слабоструктурованих фінансово-інвестиційних даних великих обсягів. Основними завданнями виступають формалізація концептуальної архітектури моделі, визначення механізмів взаємодії її складових, опис методологічних принципів аналітичної обробки в межах розподіленого середовища, а також окреслення підходів до інтеграції статистичних методів у процес багатовимірного аналізу.

Основна частина. Запропонована модель ґрунтується на ідеї багаторівневої аналітичної системи, у якій взаємодіють три функціональні компоненти: розподілена підсистема попередньої обробки даних, аналітичне ядро на основі OLAP-операцій та модуль статистичного аналізу, реалізований з використанням мовного середовища Python. Ключовим принципом побудови моделі є поєднання переваг паралельної обробки з високою швидкодією і аналітичної інтерпретованості результатів у багатовимірному просторі.

На першому рівні здійснюється акумуляція та підготовка даних у розподіленому середовищі. Для цього використовується фреймворк MapReduce, який дозволяє розподіляти обчислювальні завдання між вузлами кластеру. Його застосування дає змогу ефективно обробляти не лише структуровані, але й напівструктуровані джерела, зокрема JSON, XML та текстові фінансові документи. Оброблені результати зберігаються у вигляді таблиць у середовищі Hive, що забезпечує SQL-подібний доступ до результатів попереднього аналізу. Hive виступає проміжною ланкою між розподіленими обчисленнями та OLAP-модулем, забезпечуючи перетворення слабоструктурованих даних у формат, придатний для подальшого багатовимірного аналізу.

На другому рівні реалізується аналітичне ядро на основі OLAP-підходу. Воно формує багатовимірні куби даних, які дозволяють здійснювати операції агрегації, зрізу, деталізації та порівняння. У межах даної концепції OLAP-структури формуються динамічно на основі даних, отриманих із Hive, що дає змогу адаптувати аналітичну модель до змін структури даних у реальному часі. Замість статичних кубів пропонується використання віртуальних багатовимірних представлень, які генеруються запитами з параметричними обмеженнями.

Третій рівень моделі формує модуль статистичного аналізу, який реалізується засобами Python. Цей компонент забезпечує побудову регресійних, кластерних і кореляційних залежностей між атрибутами фінансово-інвестиційних даних, отриманих з OLAP-кубів. У межах розвідувального аналізу застосовуються методи описової статистики, параметричні оцінки, а також перевірка гіпотез щодо поведінкових характеристик ринкових даних. Особливу роль відіграє інтеграція бібліотек Python з інструментами Hive через інтерфейс PyHive, що дозволяє виконувати складні статистичні обчислення безпосередньо над розподіленими таблицями.

Розвідувальний аналіз у межах гібридної моделі передбачає поєднання індуктивних і дедуктивних підходів до побудови аналітичних гіпотез. На початковому етапі відбувається формування гіпотез щодо можливих закономірностей у даних. Потім за допомогою засобів scipy та Python здійснюється перевірка їх значущості на основі вибірок, сформованих із багатовимірних представлень OLAP. Особливість запропонованої моделі полягає у можливості динамічного оновлення вибірок відповідно до результатів попередніх обчислень, що реалізує принцип зворотного аналітичного зв’язку.

Фінансово-інвестиційні дані, що аналізуються, можуть включати показники фондових ринків, облігаційні котирування, обсяги транзакцій, звітність компаній, текстові коментарі аналітиків і публічні фінансові повідомлення. Для таких наборів характерна неоднорідність

форматів і відсутність суворої схеми опису, що зумовлює необхідність використання процедур автоматизованого розпізнавання типів даних і семантичної нормалізації. Ці процедури реалізуються у рамках MapReduce-задач, що дозволяє зменшити обсяг вхідних даних і забезпечити їх агрегування за визначеними критеріями.

У контексті статистичного аналізу важливою є підтримка адаптивних методів оцінювання, що базуються на принципах ймовірнісного моделювання. Модель допускає реалізацію методів байєсівської оцінки, що особливо ефективно у випадках неповноти або невизначеності даних. Поєднання OLAP-аналітики та статистичного моделювання створює синергетичний ефект, який підвищує інформативність отриманих висновків та дозволяє відстежувати складні динамічні закономірності у фінансових потоках.

Використання Python у реалізації гібридної моделі пояснюється його універсальністю та широкою підтримкою бібліотек для статистичного аналізу та обробки великих даних. Python виконує роль інтеграційного середовища, що забезпечує комунікацію між компонентами Hive, MapReduce і OLAP. У межах запропонованої архітектури Python-скрипти ініціюють запити до Hive, виконують попередню агрегацію результатів і формують аналітичні звіти у вигляді статистичних узагальнень або графічних представлень.

Модель передбачає використання бібліотек pandas, NumPy та SciPy для роботи з числовими масивами, а також statsmodels для побудови статистичних моделей, зокрема регресійних і часових. Таким чином, створюється аналітичне середовище, у якому користувач може одночасно працювати з даними, розподіленими у хмарній інфраструктурі, і виконувати глибокий статистичний аналіз безпосередньо у середовищі Python.

Гібридна модель дозволяє зберегти логічну прозорість аналітичного процесу, забезпечити відтворюваність результатів і підтримати різні рівні деталізації фінансово-інвестиційних показників. Вона може бути використана як основа для побудови інтелектуальних аналітичних систем фінансового моніторингу, прогнозування ризиків та моделювання інвестиційних стратегій.

Висновки. Розроблена концепція гібридної моделі розвідувального аналізу різнорідних слабоструктурованих фінансово-інвестиційних даних великих обсягів визначає новий підхід до організації аналітичного процесу в умовах інформаційної складності. Її реалізація на основі інтеграції технологій MapReduce, Hive і OLAP у поєднанні зі статистичним аналізом на Python створює інструментальне середовище, придатне для виявлення прихованих закономірностей у фінансових потоках, формування прогнозних моделей і підтримки аналітичних рішень. Перспективами подальших досліджень є розширення моделі у напрямі включення засобів машинного навчання, автоматизація побудови аналітичних сценаріїв і забезпечення повної семантичної інтеграції з джерелами фінансово-економічних даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Skrzypczyński, M., & Muryjas, P. (2024). Analysis of data processing efficiency with use of Apache Hive and Apache Pig in Hadoop environment. *Journal of Computer Sciences Institute*, 30, 1–8.

[2] Dehne, F., Kong, Q., Rau-Chaplin, A., Zaboli, H., & Zhou, R. (2015). Scalable real-time OLAP on cloud architectures. *Journal of Parallel and Distributed Computing*, 79, 31-41.

[3] Sfaxi, L., & Ben Aissa, M. M. (2021). Designing and implementing a Big Data benchmark in a financial context: Application to a cash management use case. *Computing*, 103(9), 1983-2005.

[4] Шведов, Д. В., Рудніченко, М. Д., & Тулисов, В. А. (2025). Методика створення системи побудови гібридних моделей глибинного навчання для інтелектуального аналізу фінансових

ризиків. *Технологія-2025: матеріали міжнародної науково-практичної конференції* 2025, Київ, Україна.

Rudnichenko M., Shibaeva N., Shvedov D.

Concept of a hybrid model for exploratory analysis of heterogeneous weakly structured large-scale financial and investment data

Abstract. *The paper proposes a concept of a hybrid model for exploratory analysis of heterogeneous weakly structured large-scale financial and investment data. The model integrates MapReduce, Hive, and OLAP technologies with statistical analysis in the Python environment. The developed approach ensures scalability, adaptability, and enhanced accuracy of analytical conclusions.*

Keywords: *big data; exploratory data analysis; investment analysis; weakly structured data.*