УДК 004.4

Шевчук Оксана Олегівна
магістрант
Національного університету «Львівська політехніка»
Шевчук Оксана Олеговна
магистрант
Национального университета «Львовская политехника»
Shevchuk Oksana
Master Student of the
Lviv Polytechnic National University

ЕФЕКТИВНІСТЬ ПОБУДОВИ ПРОГНОЗІВ ДАНИХ НА РЕЛЯЦІЙНІЙ СИСТЕМІ SQL SERVER ТА РОЗПОДІЛЕНІЙ ФАЙЛОВІЙ СИСТЕМІ HADOOP

ЭФФЕКТИВНОСТЬ ПОСТРОЕНИЯ ПРОГНОЗОВ ДАННЫХ НА РЕЛЯЦИОННОЙ СИСТЕМЕ SQL SERVER И РАСПРЕДЕЛЕННОЙ ФАЙЛОВОЙ СИСТЕМЕ HADOOP

EFFICIENCY OF DATA FORECASTING ON RELATIONAL SYSTEM SQL SERVER AND DISTRIBUTED FILE SYSTEM HADOOP

Анотація. *Висвітлено порівняльний аналіз системи прогнозування даних на основі реляційної СУБД та BIG DATA системи Hadoop. Наведено графічне та табличне порівняння швидкодії підходів та показано результати.*
Ключові слова: *Hadoop, Big Data, сховища даних, швидкодія, R, прогнозування.*

Аннотация. *Освещен сравнительный анализ системы прогнозирования данных на основе реляционной СУБД и BIG DATA системы Hadoop. Приведено графическое и табличное сравнения быстродействия подходов и показаны результаты.*
Ключевые слова: *Hadoop, Big Data, хранилища данных, быстродействие, R, прогнозирования.*

Summary. *A comparative analysis of the data forecasting system based on the relational database and BIG DATA of the Hadoop system is presented. Graphical and tabular comparison of approaches performance and results are presented.*
Key words: *Hadoop, Big Data, Data Warehouse, Performance, R, Forecasting.*

Вступ. Big Data є революційним феноменом, та одним з найбільш обговорюваних явищ в сучасному світі, і, як очікується, залишиться таким ще довгий час. Навички, апаратне і програмне забезпечення, архітектура алгоритмів, статистична значимість, і сама природа Big Data є основними факторами, які перешкоджають процесу отримання коректних прогнозів від Big Data.

Вважається, що в даний час галузі економіки, енергетики і соціальних досліджень населення є основними експлуататорами прогнозування Big Data.

Явище Big Data є революцією в сучасному світі, і в даний час є найпоширенішим способом інтелектуального аналізу.

В той час для бізнесу Big Data це «новий тип стратегічного ресурсу в ері цифрових даних і ключовий фактор для впровадження інновацій, які змінюють спосіб поточного виробництва» [1, с. 45].

У сучасному світі є вже доволі великий стек технологій, що забезпечують роботу із Big Data. Однак через відносно невеликий час відколи Big Data вийшла на ринок вони ще не зовсім пристосовані під загальні потреби.

Критичним питанням при застосуванні Big Data технологій є обрати спосіб зберігання даних, що будуть оброблятися. Питання, яке ставитиметься у даній статті — чи можна використовувати реляційні системи управління базами даних для аналізу дуже великих обсягів даних. Для цього аналогічні структури створено з допомогою SQL Server 2016 та Hadoop, та проведено аналіз ефективності роботи із даними на обидвох системах.

Виклад основного матеріалу. Порівняння роботи та швидкодії роботи R застосунку із SQL SERVER та Hadoop.

Для реалізації даного продукту для початку створено синхронізовану систему даних. Оригінальні

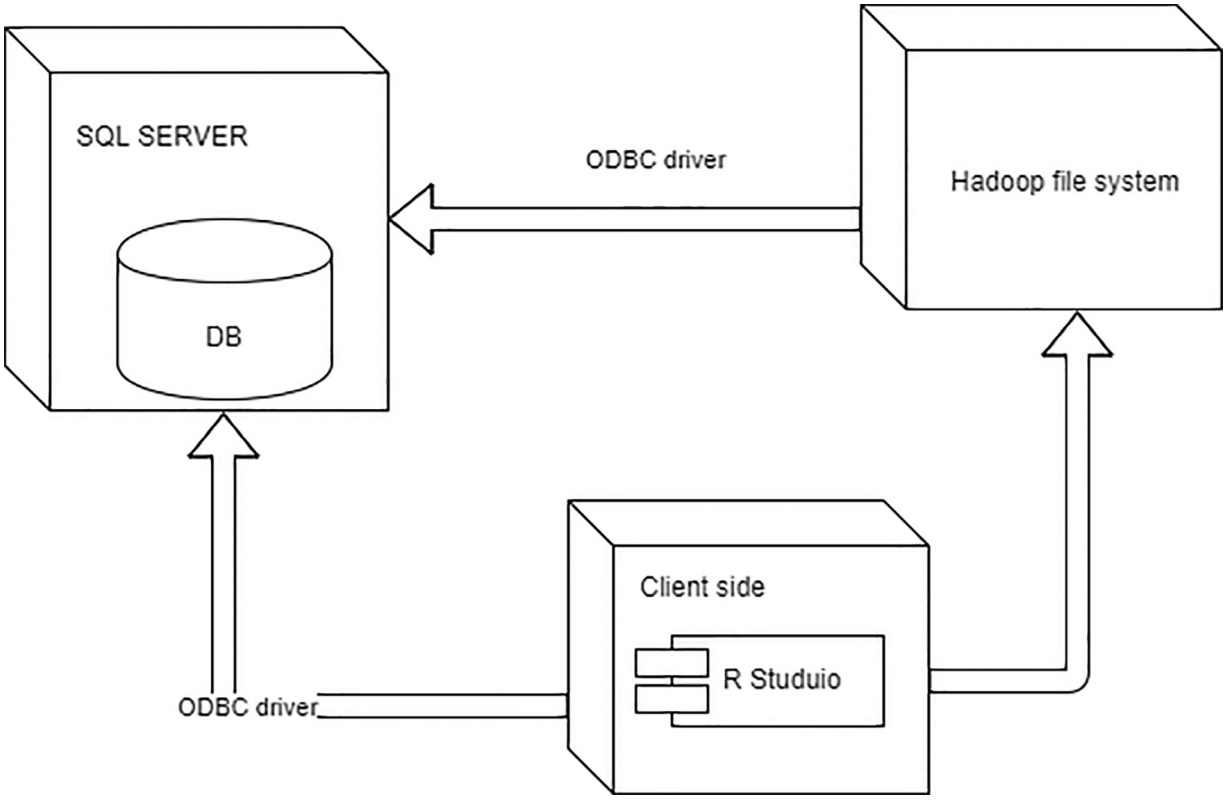

Рис. 1. Архітектура системи для тестування

дані поступають у реляційну базу даних, звідки за допомогою застосунку написаного на мові переливаються у систему. Даний процес можна побачити на діаграмі архітектури системи.

Як результат даного процесу було отримано дві системи з аналогічними даними для перевірки швидкодії на них застосунку прогнозування.

Об'єм даних на якому проводилося дослідження — 1000000 записів.

Обидві системи працюють на машинах із однаковою робочою потужністю.

Дослідження прогнозів даних 1 було проведено для прогнозу середнього значення доходу. Для нього були необхідні наступні запити до інформаційних систем:

1) Вибірка даних

2) Фільтрування по певному полю

3) Групування даних, для знаходження усереднених по дню.

Результати швидкодії по дослідженню 1, дані подано у секундах:

Таблиця 1

Запуск номер:	SQL SERVER	Hadoop
1	8,4	6
2	8,5	4,6
3	7,5	5,1
4	7,8	4,9
5	8,1	5,2

У дослідженні 2 було проведено аналіз теорії на основі парних Т-тестів Стюдента на заданій множині даних. Оскільки парний Т-тест Стюдента вимагає двох нормалізованих по часу або по виміру, за яким проводиться тестування, то операції, необхідні над даними для підготовки є наступними:

1) Фільтрування даних

2) Агрегація даних по датах

3) Нормалізація даних по датах

У наступній таблиці подано результати швидкодії тестів на двох паралельних системах.

Таблиця 2

Запуск номер:	SQL SERVER	Hadoop
1	11,4	6,9
2	13,5	7,6
3	12,9	8,1
4	14,3	9,5
5	13,1	5,7

Оскільки дані над якими працює R Studio є невеликими, уся підготовча робота відбувається на рівні системи обробки даних. Оскільки системи є синхронізованими а часі, то різниці в розмірах даних на певний момент часу не має. Тому можна сказати що різниця в часі між двома системами чітко спричинена їхніми внутрішніми підходами до обробки даних.

Висновки. Як результат було отримано інформацію, що навіть при хорошій оптимізації SQL SERVER

поступається у швидкодії розподіленій системі Hadoop. Однак слід зазначити, що SQL SERVER забезпечує значно більшу цілісність даних. Тому, коли система може дозволити собі певну похибку в даних то набагато доцільніше використовувати підходи з використанням Big Data, однак коли похибки є неприпустимими тоді вибір однозначно падатиме на реляційну систему управління.

Література

1. Richards N. M. Three paradoxes of big data / Richards N. M., King J. H. // Stanf Law Rev Online. — 2013. — С. 41–46.

2. Stock J. H. Forecasting using principal components from a large number of predictors / Stock J. H., Watson M. W. // J. Am Stat Assoc. — 2002. — С. 1167–1179.