

## РЕФЕРАТ

**Актуальність теми.** У сучасну цифрову епоху кількість даних, що генеруються, зростає в геометричній прогресії. Їх аналіз є дуже важливим, оскільки допомагає приймати обґрунтовані рішення, виявляти тенденції та закономірності, а також для виявлення нових бізнес-можливостей, покращення обслуговування клієнтів та розробки нових продуктів чи послуг. Це створює нагальну потребу в ефективних і масштабованих методах обробки та аналізу великих обсягів даних. Часові ряди є важливим типом даних для бізнесу, оскільки вони дають уявлення про тенденції та закономірності, які можуть бути використані для прогнозування та прийняття рішень. У роботі проводиться аналіз способів і засобів розподіленої обробки та прогнозування часових рядів на кластері. Підвищення їх ефективності є актуальним в галузі роздрібною торгівлі. Підприємства роздрібною торгівлі генерують величезні обсяги даних часових рядів з різних джерел, включаючи операції продажу, рівень запасів та інформацію про поведінку клієнтів. Швидка та ефективна обробка історичних даних надасть змогу робити точні прогнози продажів і попиту, що в свою чергу дозволить підприємствам оптимізувати рівень запасів і підвищити дохід.

Найпопулярнішим на сьогодні рішенням є використання інструментів та підходів, що надають хмарні сервіси. Проте, таке рішення недостатньо гнучке і потребує складної інтеграції даних. Тому існує потреба в легкомасштабованих та гнучких засобах для розподіленої обробки часових рядів.

**Метою** є підвищення ефективності алгоритмів обробки часових рядів шляхом використання розподілених обчислень на кластері.

**Об'єктом дослідження** є обробка великих обсягів даних з використанням кластерних систем.

**Предметом дослідження** є способи та алгоритми розподіленої обробки часових рядів в кластерних системах.

**Наукова новизна** полягає в наступному:

1. Запропоновано новий алгоритм для обробки часових рядів з використанням проєкцій багатовимірних масивів, що дозволить суттєво зменшити час їх обробки та об'єми використовуваної для обчислень пам'яті.
2. Запропоновано модифікацію існуючих алгоритмів прогнозування часових рядів для розподіленого виконання на кластері шляхом використання синхронного навчання моделей.
3. Виконано програмну реалізацію запропонованих алгоритмів обробки та прогнозування, яка доводить їх ефективність.

**Практична цінність:** запропоновані в дисертації алгоритми та засоби розподіленої обробки дозволять суттєво збільшити швидкість обробки великої кількості часових рядів. Це сприятиме підвищенню ефективності прогнозування у багатьох галузях: фінансовому прогнозуванні, прогнозуванні попиту на енергію, прогнозуванні продажів в сфері роздрібної торгівлі тощо. Розроблені алгоритми можуть бути легко інтегровані в програмні засоби для розподіленої обробки і прогнозування часових рядів, надаючи користувачам зручні і масштабовані інструменти.

**Апробація роботи.** Основні положення і результати роботи представлено та обговорювалось на XV науковій конференції магістрантів та аспірантів «Прикладна математика та комп'ютинг» ПМК-2022 (Київ, КПІ, ФПМ, 16-18 листопада 2022 р.) та на 89 міжнародній науковій конференції молодих учених, аспірантів і студентів «Наукові

здобутки молоді – вирішенню проблем харчування людства у ХХІ ст.» (Київ, НУХТ, 3-7 квітня 2023 р.).

**Структура та обсяг роботи.** Магістерська дисертація складається з вступу, чотирьох розділів та висновків.

*У вступі* надано оцінку сучасного стану проблеми, наведено обґрунтування актуальності напряду дослідження, а також формулювання мети та завдань дослідження.

*У першому розділі* досліджено існуючі способи і засоби для розподіленої обробки даних часових рядів, виявлено основні переваги та недоліки, визначено шляхи їх вдосконалення.

*У другому розділі* розглянуто основні технології, способи та засоби для роботи з великими обсягами даних. Представлено сучасні підходи для розподіленого глибинного навчання.

*У третьому розділі* представлено опис запропонованого алгоритму для підготовки даних та модифікованих алгоритмів прогнозування часових рядів. Надано загальний підхід до програмної реалізації алгоритмів.

*У четвертому розділі* досліджено ефективність запропонованих алгоритмів обробки та прогнозування часових рядів на кластері.

*У висновках* представлені результати проведеної роботи.

Робота представлена на 90 аркушах, містить посилання на список використаних літературних джерел.

**Ключові слова:** великі дані, кластер, часові ряди, розподілена обробка, прогнозування.

## ABSTRACT

**Actuality of theme.** In today's digital age, the amount of data generated is growing exponentially. Analyzing it is very important as it helps to make informed decisions, identify trends and patterns, as well as to identify new business opportunities, improve customer service, and develop new products or services. This creates an urgent need for efficient and scalable methods of processing and analyzing large amounts of data. Time series are an important type of data for business because they provide insights into trends and patterns that can be used for forecasting and decision-making. This paper analyzes ways and means of distributed processing and forecasting of time series on a cluster. Increasing their efficiency is relevant in the retail industry. Retailers generate huge amounts of time series data from various sources, including sales transactions, inventory levels, and information about customer behavior. Fast and efficient processing of historical data will enable accurate sales and demand forecasts, which in turn will allow businesses to optimize inventory levels and increase revenue.

The most popular solution today is to use tools and approaches provided by cloud services. However, this solution is not flexible enough and requires complex data integration. Therefore, there is a need for easily scalable and flexible tools for distributed time series processing.

**The goal of the work** is to improve the efficiency of time series processing algorithms by using distributed computing on a cluster.

**The object of research** is the processing of large amounts of data using cluster systems

**The subject of the study** are methods and algorithms for distributed time series processing in cluster systems.

**The scientific novelty:**

1. A new algorithm for processing time series using projections of multidimensional arrays is proposed, which will significantly reduce the processing time and the amount of memory used for computing.
2. The modification of existing time series forecasting algorithms for distributed execution on a cluster by using synchronous model training is proposed.
3. The software implementation of the proposed processing and forecasting algorithms is carried out, which proves their effectiveness

**Practical value:** the algorithms and distributed processing tools proposed in this thesis will significantly increase the processing speed of a large number of time series. This will help to improve the efficiency of forecasting in many industries: financial forecasting, energy demand forecasting, retail sales forecasting, etc. The developed algorithms can be easily integrated into software tools for distributed time series processing and forecasting, providing users with convenient and scalable tools.

**Approbation of work.** The main provisions and results of the work were presented and discussed at the XV scientific conference of undergraduates and graduate students "Applied Mathematics and Computing" PMC-2022 (Kyiv, November 16-18, 2022) and at the 89th International Scientific Conference of Young Scientists, Postgraduates and Students "Scientific Achievements of Youth – Solving the Problems of Human Nutrition in the XXI Century" (Kyiv, April 3-7, 2023).

**Structure and scope of work.** The master's thesis consists of an introduction, four chapters, and a conclusion.

The *introduction* provides an assessment of the current state of the problem, justifies the relevance of the research area, and formulates the purpose and objectives of the study.

The *first* section examines the existing methods and tools for distributed processing of time series data, identifies the main advantages and disadvantages, and identifies ways to improve them.

The *second* section discusses the main technologies, methods and tools for working with large amounts of data. Modern approaches for distributed deep learning are presented.

The *third* section describes the proposed algorithm for data preparation and modified algorithms for time series forecasting. A general approach to the programmatic implementation of the algorithms is presented.

The *fourth* section investigates the effectiveness of the proposed algorithms for processing and forecasting time series on a cluster.

The *conclusion* presents the results of the work.

The paper is presented on 90 pages and contains references to the list of used literature sources.

**Keywords:** big data, cluster, time series, distributed processing, forecasting.