УДК 004.75

ОПТИМІЗАЦІЯ ПІДБОРУ ПАРАМЕТРІВ МОДЕЛЕЙ ДЛЯ АНАЛІЗУ ВЕЛИКИХ ДАНИХ ТЕЛЕКОМУНІКАЦІЙНОЇ КОМПАНІЇ

[1]Лавренюк А.М., [2]Лавренюк С.І.
[1]*Фізико-технічний інститут КПІ ім. Ігоря Сікорського, Україна;*
[2]*Інститут кібернетики ім. В. М. Глушкова НАН України, Україна*
E-mail: lsi@bigmir.net

Optimization of model parameters selection for big data from telecommunication company analysis

This paper continues the cycle of studies on big data processing optimization using python.

In addition to fact that telecommunication companies data real-time processing and interpreting require high-power computing resources, new algorithms and parallel data processing approaches are also needed to accelerate computations.

This paper proposes a new approach to solving traditional problems on big data from telecommunication company analysis such as customers churn prediction and others.

Робота продовжує цикл досліджень по оптимізації обробки великих об’ємів даних з використанням мови програмування python.

Окрім того, що задачі обробки та інтерпретації даних телекомунікаційних компаній в режимі реального часу потребують високо потужних обчислювальних ресурсів, для прискорення обчислень потрібні також нові алгоритми та підходи паралельної обробки даних. В роботі пропонується новий підхід до рішення традиційних задач аналізу великих даних для телекомунікаційної компанії таких, як прогнозування відтоку клієнтів (churn predict) та інших.

Запропоновано використовувати розподілені асинхронні черги для вибору оптимальної моделі із певного набору моделей, що найкраще описують та моделюють вхідні дані. Також вперше запропоновано використовувати саме такі черги для підбору оптимальних (найкращих) параметрів вибраної моделі.

Вступ. На сьогодні вже є розроблені та апробовані алгоритми з використанням моделей бібліотеки Scikit-learn для обробки та аналізу великих даних, що можуть використовуватися в локальних та хмарних ресурсах [1, 2]. Так, як час роботи локальних систем, так особливо хмарних є доволі не дешевим при довготривалій роботі, то пошуки можливості зменшення часу обробки даних, а відповідно і фінансових затрат є актуальними постійно. Зменшення часу обробки даних можливе при використанні нових, оптимальніших алгоритмів, що при існуючих обчислювальних ресурсах зменшать час обробки даних. Для

телекомунікаційних компаній така задача є актуальною, адже обробка та аналіз великих об’ємів даних проводиться постійно.

В своїх роботах автори адаптують існуючі та створюють нові алгоритми для обробки та аналізу великих даних за допомогою мови програмування python.

Перш ніж обробляти чи аналізувати дані, необхідно створити моделі чи підібрати уже з існуючого набору бібліотек для python ту яка найкраще буде описувати існуючі дані. Або певні їх проміжки. Тобто для одних і тих даних можливо буде необхідним застосовувати різні моделі на різних проміжках.

Отже при аналізі даних телекомунікаційної компанії, часто необхідно на тестовій вибірці підібрати найоптимальніші параметри для однієї і тієї ж моделі аналізу даних, або із кількох моделей вибрати найкращу для конкретних даних. Традиційно такі задачі виконують послідовно: запускається задача з одними параметрами, отримується результат, потім з іншими і таких запусків може бути багато, а в кінці порівнюються отримані результати і вибираються найкращі параметри та моделі.

Запропонований підхід. Запропоновано для рішення задачі не тільки аналізу та обробки великих даних використати Celery, що є реалізацією розподіленої асинхронної черги завдань, з широким функціоналом [3]. В даному підході вперше запропоновано використати асинхронні черги для вибору найоптимальнішої моделі та підбори оптимальних її параметрів, а вже потім аналіз та обробка даних.

Використовуючи розподілену чергу завдань Celery на сучасних обчислювальних ресурсах можливо запускати одночасно багато процесів аналізу даних, пошуку оптимальних моделей та підбір для них найкращих параметрів. Також з Celery можна працювати з різними розподіленими обчислювальними ресурсами, як з одним централізованим ресурсом.

На рис. 2 показано схему запропонованого алгоритму.

Звичайно, створювати черги, їх програмно контролювати, налаштовувати програмні продукти для рішення цих задач не просто, але це реально і дає необхідний ефект. Проведені експерименти показали значну продуктивність використання черги Celery з python. Детально алгоритм пошуку оптимальної моделі та її найкращих параметрів, формули та результати буде наведено в доповіді.

Висновок. Використання Celery в програмах python при обробці великих об’ємів даних:

- дає можливість одночасно виконувати кілька підзадач в асинхронному режимі і відповідно оптимальніше використовувати всі наявні ресурси, як сучасних настільних комп’ютерів, обчислювальних кластерів та хмар;

- встановлюється та керується як з використанням Docker так і без під різні операційні системи;
- не потребує додаткової адаптації при роботі з різними операційними системами або хмарними обчислювальними ресурсами (наприклад, Amazon);

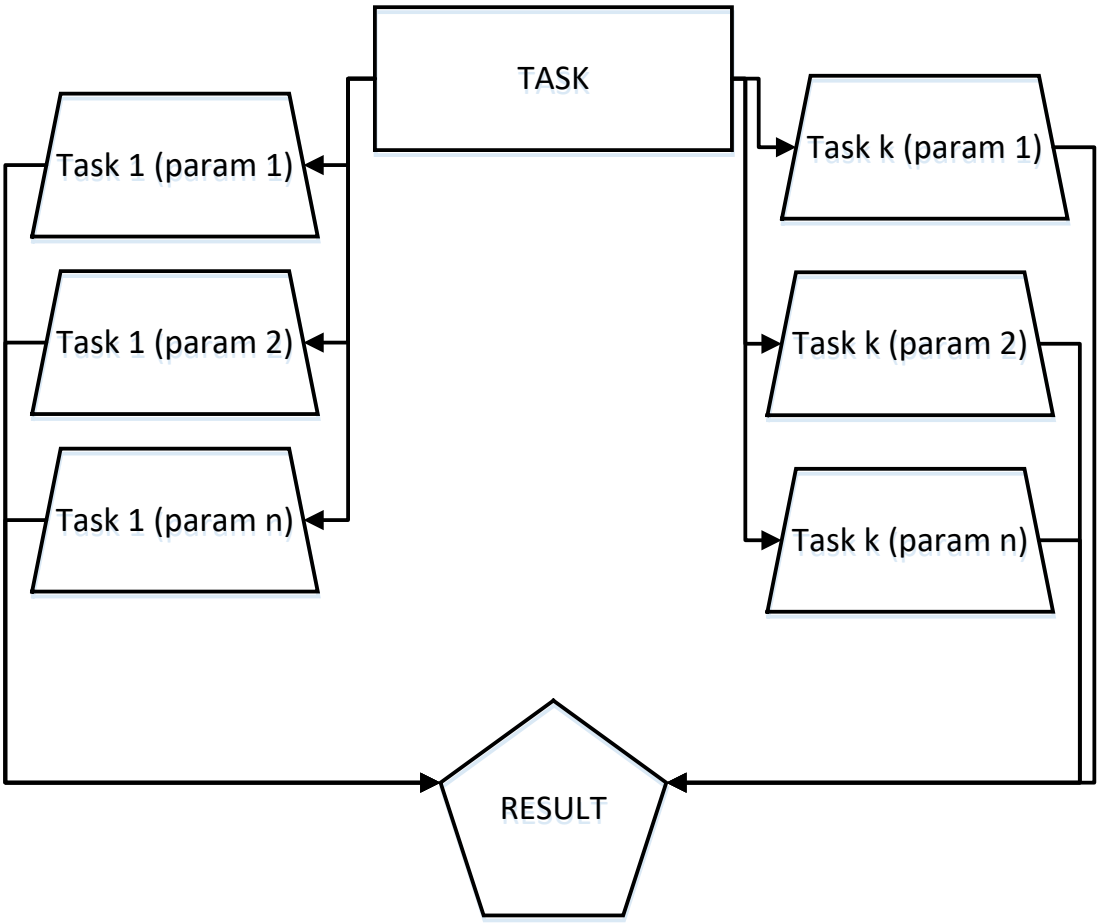

Рис. 2. Підбір моделей та їх оптимальних параметрів з використанням черг та Celery.

Слід зазначити, що запропонований підхід може бути успішно використаним при обробці великих об'ємів даних не тільки в телекомунікаційних компаніях.

Література

1. Telco Customer Churn data set, https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/.
2. Лавренюк А. М., Лавренюк Л. С., Тульчинський П. Г. Оптимізація програмного забезпечення для аналізу великих даних телекомунікаційної компанії // XI Міжнародна науково-технічна конференція "Проблеми телекомунікацій" ПТ-2017: Збірник матеріалів конференції. К.: КПІ ім. Ігоря Сікорського, 2017. – С. 325-327.
3. Celery. //Електронний ресурс: http://www.celeryproject.org/