ДОСЛІДЖЕННЯ МЕТОДІВ ДИНАМІЧНОГО АНАЛІЗУ ДАНИХ ДЛЯ ВИЗНАЧЕННЯ ЕКСПЕРТНОЇ ДУМКИ НА ОСНОВІ СТРІМІНГОВОЇ ОБРОБКИ ДАНИХ З TWITTER

Каук В. І., Єрусалімцев Д. А.
Харківський національний університет радіоелектроніки, Харків, Україна

Дослідження методів динамічного аналізу потокових даних із соціальних мереж мають значний вплив на збільшення швидкості та якості аналізу даних, що генеруються кожну секунду. Сьогодні потокова обробка – це стандарт для побудови систем аналітики в режимі реального часу. Тому теоретичні та практичні дослідження таких методів є важливою науковою проблемою. Потокові дані – це технологія, яка дозволяє користувачам отримувати безперервний потік даних і швидко реагувати на зміни [1]. Це дозволяє скоротити час, витрачений на виявлення події та вчасно і якісно зреагувати на неї. Зі сторони аналізу інформації основна проблема полягає у великому обсязі необхідної інформації та великій кількості джерел, які зберігають та публікують цю інформацію. Іншою проблемою є те, що інформація поділяється на окремі різні формати: текстову, відео, аудіо та фото. За необхідністю отримання необхідної інформації про нові події, у яких зацікавлений користувач, довелося б працювати з декількома масивами різнотипної інформації. В якості джерела були обрані соціальні мережі, які дуже впливають на життя сучасних людей. Сьогодні з 100 найбільш відвідуваних сайтів у світі 20 – це соціальні мережі. Більше 80% компаній по всьому світу використовують дані з соціальних мереж у своїй повсякденній роботі. Близько 75% людей вважають інформацію з соціальних мереж достовірною. Соціальні мережі стали невід'ємним центром сучасного світу.

Метою доповіді є максимальне спрощення аналізу та класифікації вхідної інформації, яка отримана із соціальної мережі Twitter. Аналіз потоку вхідної інформації, а саме твітів, інформації про автора, коментарі інших користувачів, та спеціальні публічні метріки дозволять виділити достовірні джерела та розподілити авторів на різні рівні довіри користувачів.

В доповіді наводяться результати практичних експериментів, які демонструють потокову обробку інформації, отриману з мережі Twitter. Наведені дані показують, що на результати експериментів впливали деякі фактори, а саме: конфігурація кластеру (кількість та характеристики робочих вузлів), оптимізація коду та особливості сериалізації. Проведені експерименти підтверджують якість фреймворку Flink як базового інструменту для обробки потокових даних [2]. Саме гнучкість у конфігурація та автоматичне покращення продуктивності у Flink дозволило обробляти величезні обсяги даних у режимі реального часу.

Список літератури

1. Alexandru Costan. From Big Data to Fast Data: Efficient Stream Data Management. 2019. С. 180.

2. F. Hueske, V. Kalavri. Stream Processing with Apache Flink: Fundamentals, Implementation, and Operation of Streaming Applications 1st Edition. 2019. С. 310.