УДК 004.9
П. Прийма, А. Зав'ялова, В. Дуда
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ІНСТРУМЕНТИ АНАЛІТИЧНОГО ОПРАЦЮВАННЯ «ВЕЛИКИХ ДАНИХ»

UDC 004.9
P. Pryima, A. Zavialova, V. Duda

TOOLS FOR BIG DATA ANALYTICAL PROCESSING

На даний час щодня ґенеруються великі за обсягом набори та колекції даних практично у всіх сферах людської діяльності. Зокрема дані надходять від соціальних мереж, інженерії, виробництва, транспорту, комерції, галузі охорони здоров'я, біомолекулярних досліджень, фізології тощо. «Великі дані» (BD, англ. Big Data) та інноваційні методи та підходи їх аналітичного опрацювання, зокрема Big Data Analytics (BDA), змінили спосіб функціонування установ, підприємств та організацій, сформувавши при цьому обширний перелік нових перспективних напрямків досліджень для фахівців та наукової спільноти [1]. Окрім виробничих підприємств і дослідницьких установ, урядові та неурядові організації на даний час регулярно ґенерують великі за обсягом унікальні набори та колекції даних. Тому видобування та отримання значущої інформації із доступних «Великих даних» стало життєво важливим для підприємств, установ та організацій стало критично актуальним у всьому світі.

Інструменти аналітичного опрацювання «Великих даних» (англ. Big Data) використовуються для обробки великих за обсягом, структурованих, неструктурованих і напівструктурованих даних з метою видобування знань, бізнес-прогнозування, підвищення ефективності процесів прийняття рішень, візуалізації шаблонів тощо.

Apache Hadoop є одним із найпопулярніших інструментів аналізу даних. Він поширюється з відкритим кодом. HDFS (розподілена файлова система Hadoop) – це високоефективний компонент зберігання даних, який використовується для зберігання різнотипових та різноманітних даних, зокрема тексту, xml або json файлів, аудіофайлів, зображень та відео. Зберігання відбувається завдяки поділу даних на частини та збереження в кластерах товарних серверів [2]. Заснований на Java, Apache Hadoop характеризується високою швидкістю, оскільки окремі завдання розділяються та виконуються одночасно на розподілених серверах. Оскільки дані зберігаються на множині розподілених серверів то резервне копіювання даних доступне, навіть при виході з ладу одного окремого сервера.

Apache Spark – це розподілена інформаційна система з відкритим кодом, яка обробляє дані з використанням апаратної оперативної пам'яті. Водночас швидкість обробки даних засобами Spark відчутно перевищує швидкість Hadoop [3]. Spark зручний для системних архітекторів та розробників програмного забезпечення, оскільки для створення програм можна використовувати різні мови програмування, зокрема java, python, R, scala тощо. На даний час обширний перелік організацій, зокрема Finra, Yelp, Zillow, gumgum, використовували Spark. Тому він став практично одним із найпопулярніших фреймворків розподіленої обробки «Великих даних».

Література

1. Islam, AYM Atiquil, et al. «Performance-based evaluation of academic libraries in the big data era.» Journal of Information Science 47.4 (2021): 458–471.
2. Akhtar, Nikhat, Firoj Parwej, and Yusuf Perwej. «A perusal of big data classification and hadoop technology.» International Transaction of Electrical and Computer Engineers System (ITECES), USA 4.1 (2017): 26–38.
4. Cao, Jian, et al. «Personalized flight recommendations via paired choice modeling.» 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017.