УДК 004.65:004.032.2

АНАЛІЗ ВИСОКОЕФЕКТИВНИХ КЛАСТЕРІВ ДЛЯ ОБРОБКИ ВЕЛИКИХ ДАНИХ

Горішня К.О.
Науковий керівник – ас. каф. ПІ Зибіна К.В.
Харківський національний університет радіоелектроніки, каф. ПІ,
м. Харків, Україна
тел. +38(066) 524-61-51

Apache Spark is a powerful tool for big data processing, with benefits such as speed, scalability, and support for multiple programming languages. It outperforms Apache Hadoop in several areas, including processing time and ease of use. However, it also has some drawbacks, such as high resource demands and complex configuration. Overall, Apache Spark is a great choice for developing large-scale data processing and machine learning projects.

Apache Spark та Apache Hadoop - це інструменти для обробки великих обсягів даних. Apache Hadoop - це фреймворк для обробки даних, який використовує розподілену файлову систему та MapReduce, щоб розбити завдання на більш дрібні та розподілити їх по вузлах кластера [1]. Apache Spark, є движком для обробки даних, який працює на верхньому рівні Hadoop. Він також використовує розподілену файлову систему, але замість MapReduce використовує більш потужну модель обробки даних в пам'яті. Хоча обидва інструменти мають схожу функціональність та їх можна використовувати окремо або в поєднанні в залежності від потреб проекту, Apache Spark має кілька переваг порівняно з Apache Hadoop:

1) Швидкість обробки даних.

Однією з найбільших переваг Apache Spark є його швидкість обробки даних. Spark може оброблювати дані до 100 разів швидше, ніж Apache Hadoop. Це досягається завдяки використанню in-memory обробки даних, що дозволяє зберігати дані в оперативній пам'яті та оптимізувати доступ до даних [2]. Це дозволяє Spark бути ідеальним вибором для задач, що вимагають великої швидкості обробки даних.

2) Масштабованість.

Apache Spark також має перевагу в масштабованості порівняно з Hadoop. Spark може масштабуватися горизонтально на декілька машин, що дозволяє розробникам обробляти великі обсяги даних. Це досягається завдяки використанню розподіленої обробки даних та системи кластеризації. Spark може працювати з сотнями терабайтів даних.

3) Можливість використання багатьох мов програмування.

Apache Spark може бути використаний з декількома мовами програмування, включаючи Scala, Java, Python та R. Це дозволяє розробникам використовувати мову, з якою вони знайомі, та дозволяє їм швидко та ефективно створювати програми для обробки даних.

4) Надійність та відновлюваність.

Apache Spark дозволяє відновлювати дані у випадку відмови апаратного забезпечення та зберігати дані у безпечних та надійних місцях. Він має вбудований механізм відновлення після збоїв, що дозволяє забезпечувати надійну роботу та захист даних.

5) Велика кількість бібліотек та інструментів.

Apache Spark має велику кількість бібліотек та інструментів, що дозволяє розробникам ефективно виконувати різні задачі, включаючи машинне навчання, обробку графів та аналіз текстів. Завдяки великій кількості бібліотек, розробники можуть швидко та ефективно розробляти програми для обробки даних [2].

6) Легкий використання та інтеграція.

Apache Spark має легкий інтерфейс та просту систему конфігурації, що дозволяє розробникам легко використовувати та інтегрувати його у свої проекти. Він також підтримує різні формати даних, такі як CSV, JSON та Parquet, що дозволяє легко обробляти дані з різних джерел [3].

Apache Spark також має декілька недоліків порівняно з Apache Hadoop, на які варто звернути увагу при використанні:

1) Складність настройки.

Apache Spark має багато параметрів настройки, що може бути складним для новачків. Для досягнення найкращих результатів необхідно правильно настроїти його параметри, що може вимагати багато зусиль.

2) Потребує додаткових інструментів для обробки даних.

Apache Spark не має вбудованих інструментів для обробки розподілених даних, що становить проблему з даними великого обсягу.

Незважаючи на деякі недоліки, Apache Spark є дуже потужним інструментом для обробки даних, з великою кількістю переваг порівняно з Apache Hadoop. Його швидкість обробки даних, масштабованість та можливість використання багатьох мов програмування дозволяють розробникам ефективно виконувати різні задачі. Також, велика кількість бібліотек та інструментів, легке використання та інтеграція забезпечують комфортну та безпечну роботу з великими даними.

Отже, Apache Spark може бути відмінним вибором для розробки великих проектів обробки даних, зокрема, для машинного навчання, обробки графів та аналізу текстів.

Список використаних джерел:

1. Hadoop Ecosystem: Hadoop Tools for Crunching Big Data. (б. д.). Взято 30 березня 2023 року з https://www.edureka.co/blog/hadoop-ecosystem

2. Bharathy, A. M. V., Shanmugavalli, V., & Chandrasekar, V. (2021). Big Data Analytics. Central West Publishing.

3. Jackson, F. (2022). Big Data: Concepts, Technology and Architecture. States Academic Press.