

ПРОГРАМНИЙ ЗАСІБ ДЛЯ АНАЛІЗУ РОБОТИ SPARK ТА MAPRADUCE BIG DATA ТЕХНОЛОГІЙ

Михайловська Т.С., Татарінова О.А.

***Національний технічний університет
«Харківський політехнічний інститут»,
м. Харків***

Доповідь присвячено опису програмного засобу, що створений для аналізу і порівняння роботи Big Data технологій Spark та MapRaduce. Створений програмний комплекс проводить підрахунок слів у тексті за допомогою різних технологій. MapRaduce та Spark технології були використані за допомогою засобів платформи Hadoop для проведення розподіленого обчислення та порівняння результатів роботи цих технологій.

В даний час спостерігається значне зростання обсягів інформації, що надходить з різних джерел. Це можуть бути мобільні пристрої, програмні логи, інформація про проведені фінансові операції, активності користувачів соціальних мереж і багато іншого. Інформація представлена широким спектром типів даних, починаючи зі звичних баз даних з чіткою структурою, закінчуючи зображеннями і текстами на природній мові. Обсяг, накопиченої інформації, настільки великий, а типи даних настільки різноманітні, що їх обробка стандартними методами не представляється можливим, тому Big Data технологій сьогодні є досить затребуваними.

У даному програмному забезпеченні було створено два алгоритми розподіленого обчислення для підрахунку слів в тексті у зв'язку з тим, що Обчислення у MapReduce фреймворку завжди базуються на трьох операціях – map, shuffle та reduce, а для Spark не потрібно використовувати ці операції. Для створення MapReduce алгоритму було використано мову програмування Java, а для Spark алгоритму- Scala.

Для обчислення великих в роботі використано дистрибутив Apache Hadoop з набором програм, бібліотек та утиліт для розподіленого обчислення – Hortonworks Sandbox.

Аналіз тексту, який є вхідними даними, проведений за наступними атрибутами: обсяг, швидкість, структурованість та правдивість проведеного розрахунку. Під час тестування використано тексти різного об'єму та різної структурної складності. Отримані результати роботи двох алгоритмів та проведено порівняння.