УДК 338:339.1

СУТНІСТЬ ТА ОСОБЛИВОСТІ ВИКОРИСТАННЯ HADOOP

Студ. Дикий В.С.
Наук. керівник доц. Пономаренко І.В.
Київський національний університет технологій та дизайну

Hadoop – проект фонду Apache Software Foundation, вільно розповсюджуваний набір утиліт, бібліотек і фреймворк для розробки і виконання розподілених програм, що працюють на кластерах з сотень і тисяч вузлів. Використовується для реалізації пошукових і тематичних механізмів багатьох високонавантажених веб-сайтів, в тому числі, для Yahoo! і Facebook. Розроблено на Java в рамках обчислювальної парадигми MapReduce, згідно з якою додаток поділяється на велику кількість однакових елементарних завдань, здійсненних на вузлах кластера та природним чином приводяться в кінцевий результат. Зазначена технологія набула широкого розповсюдження серед маркетингових компаній при аналізі окремих ринків, в тому числі при дослідженні різноманітних явищ в Інтернеті (Internet marketing) та соціальних мережах (Social Media Marketing).

Зазначений проект вважається однією з основоположних технологій «Великих даних». Навколо Hadoop утворилася ціла екосистема з пов'язаних проектів і технологій, багато з яких розвивалися спочатку в рамках проекту, а згодом стали самостійними. З другої половини 2000-х років йде процес активної комерціалізації технології, кілька компаній будують бізнес цілком на створенні комерційних дистрибутивів Hadoop і послуг з технічної підтримки екосистеми, а практично всі великі постачальники інформаційних технологій для організацій в тому чи іншому вигляді включають Hadoop в продуктові стратегії і лінійки рішень.

Проект складається з чотирьох модулів – Hadoop Common (сполучне програмне забезпечення – набір інфраструктурних програмних бібліотек і утиліт, використовуваних для інших модулів і споріднених проектів), HDFS (розподілена файлова система), YARN (система для планування завдань і управління кластером) і Hadoop MapReduce (платформа програмування і виконання розподілених MapReduce-обчислень), раніше в Hadoop входив цілий ряд інших проектів, які стали самостійними в рамках системи проектів Apache Software Foundation.

Однією з основних цілей Hadoop спочатку було забезпечення горизонтальної масштабованості кластера за допомогою додавання недорогих вузлів без вдавання до потужних серверів і дорогим мереж зберігання даних. Функціонуючі кластери розміром в тисячі вузлів підтверджують здійсненність і економічну ефективність таких систем, так відомо про великі кластери Hadoop в Yahoo, Facebook і Ebay. Проте, вважається, що горизонтальна масштабованість в Hadoop-системах обмежена, для Hadoop до версії 2.0 вона максимально можливо оцінювалася в 4 тис. вузлів при використанні 10 MapReduce-завдань на вузол. Ще одним обмеженням Hadoop-систем є розмір оперативної пам'яті на вузлі імен (NameNode), що зберігає весь простір імен кластера для розподілу обробки, до того ж загальна кількість файлів, яку здатний обробляти вузол імен – 100 млн. Для подолання цього обмеження ведуться роботи з розподілу вузла імен.

Отже, Hadoop являється програмною платформою (Software Framework), яка працює за принципом побудови розподілених додатків для масової паралельної обробки даних (Massive Parallel Processing, MPP). Зазначене програмне забезпечення дозволяє прискорити обробку великих масивів даних, а також дозволяє забезпечити збереження інформації завдяки її сегментації та розподілу між різними серверами. Слід наголосити, що ця технологія має важливе значення при дослідженні окремих ринків. Показовим у даному випадку є ринок США, де маркетингові компанії активно використовують зазначену технологію для комплексного аналізу ринків окремих товарів та послуг, сегментації споживачів за різними соціально-демографічними характеристиками тощо.