Кислий Р.В. *— рецензент Петренко А.І.*
Інститут прикладного системного аналізу НТУУ "КПІ", Київ, Україна

Інтелектуальний аналіз даних в хмарних системах

Збільшення обсягів згенерованих даних змушує шукати нові підходи для виявлення і використання цінних знань з цих даних. Інтелектуальний аналіз даних завжди був ефективним інструментом отримання з них корисної інформації. Швидкість передачі даних між зберігаючими пристроями та обчислювальними ресурсами – основна проблема швидкодії аналізу даних на сьогоднішній день.

Хмарні системи – модель інфраструктури, що містить пул ресурсів і забезпечує ефективність аналізу великих обсягів інформації. Ця модель розбиває великі набори даних і підготовлює їх для вузла, де дані можуть бути оброблені локально, уникаючи затримок передачі по мережі та зчитування з носіїв інформації. Це робить можливим для людей, зрозуміти і використати екзабайти інформації в центрах обробки і зберігання даних [1]. Отже, основна мета – збільшення швидкості аналізу даних в хмарних системах.

Для аналізу даних необхідні ефективні масштабовані алгоритми, що дозволять вирішити задачу за прийнятний час. Одним з таких алгоритмів є аглоритм Apriori на основі асоціативних правил. Я пропоную використовати саме його, оскільки він легко піддається масштабуваню і досить простий. Алгоритми, побудовані на асоціативних правилах, спрямовані на отримання цікавих кореляцій, асоціацій серед безлічі елементів в транзакціях бази даних або інших сховищах даних. Розмір вхідних даних Apriori, як правило, досить великий і розподілений. Таким чином, хмара може бути ідеальною платформою для цього алгоритму. Тим не менш, класичний алгоритм Apriori не призначений для виконання в середовищі хмари, так як ітеративний підхід, який він використовує, щоб отримати набори елементів, які часто зустрічаються, викликає повторювання сканування диска. Така висока частота запитів до носія інформації робить запуск алгоритму в хмарі недоцільним [2]. Доцільним є використання Apriori з MapReduce технологіями, які допоможуть оминути це вузьке місце. Ключовою перевагою MapReduce є те, що вона автоматично організовує розпаралелену обробку великих обсягів інформації на кластерах обчислювальної системи, приховуючи складність реалізації. Поєднання правил для роботи Apriori і інтелектуального аналізу широко використовуються в різних областях, таких як телекомунікаційні мережі, маркетинг та управління ризиками.

Продуктивність асоціативних правил дуже залежить від генерації набору елементів, які часто зустрічаються. Таким чином, цей метод особливо добре підходить для аналізу даних і тексту в великих базах даних. У більшості випадків користувачі не можуть зрозуміти і перевірити велику кількість складних асоціативних правил. Таким чином, важливо, генерувати тільки "корисні" правила, що задовольняють визначеним критеріям.

Алгоритм Apriori є одним із самих широко використовуваних алгоритмів для створення асоціативних правил, а MapReduce є основою для обробки проблем, які можна паралелізувати в масштабах великих наборів даних з використанням великої кількості вузлів, спільно іменованого кластера (якщо всі вузли знаходяться в одній локальній мережі і використовують аналогічне обладнання) або мережі (якщо вузли загальні для всіх географічно і адміністративно розподілених систем, а також використовують більш різноріде обладнання). Обчислювальна обробка може відбуватися на даних, що зберігаються або у файловій системі (неструктуровані), або в базі даних (структуровані). MapReduce може використовувати географічне розташування даних – обробляти дані на серверах, що знаходяться поблизу місця зберігання, для зменшення передачі даних.

Література. **1.** Mining top-K frequent item sets through progressive sampling. Data Mining and Knowledge Discovery / [PIETRACAPRINA, A., RIONDATO, M., UPFAL, E.] , IEEE, 2010. **2.** ODAM: An optimized distributed association rule mining algorithm. Distributed Systems Online / [Ashrafi, M.Z., Taniar, D., Smith, K.], IEEE, 2004.