УДК 004.65

Особливості обробки даних великих об'ємів (BigData) з використанням нереляційних баз даних

Охотний С.М., студент 3 курсу
Науковий керівник – Сидоренко В.В., старший викладач
Центральноукраїнський національний технічний університет, м. Кропивницький

З кожним роком кількість даних, що опрацьовуються комп'ютерами, постійно зростає. Для того, щоб справлятися зі зростаючими об'ємами інформації, були розроблені нові засоби її обробки, які узагальнюють терміном BigData.

BigData – набір методів та інструментів для опрацювання структурованих і неструктурованих даних величезних розмірів, що є ефективними в умовах безперервного приросту інформації та розподілення навантаження по багатьом вузлам обчислювальної мережі.

Основні властивості BigData:

–зберігання та обробка великої кількості даних;

–висока швидкість роботи;

–обробка даних різних форматів починаючи від структурованих даних, прийнятих у традиційних базах даних, до неструктурованих текстових документів, аудіо, відео, даних біржових зведень та фінансових операцій.

Основні принципи BigData:

1)горизонтальна розширюваність – при збільшенні об'єму даних збільшується кількість апаратних засобів обробки (в обчислювальну мережу можна легко додавати нові машини);

2)відмовостійкість – система продовжує працювати, якщо деякі машини обчислювальної мережі виходять з ладу;

3)локальність даних – обробка даних повинна виконуватись по можливості в тому ж місці, де вони зберігаються, щоб уникнути витрат часу для передачі даних з одного сервера на інший.

Найрозповсюдженішим засобом для роботи з BigData є Apache Hadoop. Hadoop є фреймворком для зберігання, обробки та аналізу даних у великих масштабах і має повністю відкритий вихідний код. Він може працювати на загальнодоступному апаратному забезпеченні, що робить його простим у використанні в існуючих датацентрах, або навіть для проведення аналізу в хмарі.

Для роботи з великими об'ємами даних в Hadoop застосовують нереляційні (NoSQL – not only SQL) бази даних (БД), які забезпечують інші механізми зберігання та видобування даних. Особливості NoSQL баз даних:

–не використовується SQL;

– неструктуровані;
– слабкі ACID;
– розподіл даних по вузлам;
– NoSQL бази даних в переважній більшості мають відкритий вихідний код.

Усі NoSQL БД розділяються в залежності від способу зберігання та обробки даних:
– сховища типу «ключ-значення»;
– розширювані розподілені сховища;
– графові бази даних;
– документно-орієнтовані сховища.

За замовчуванням в Hadoop використовується HBase – колонково-орієнтована, розширювана NoSQL база даних. Дані організовані в таблиці, проіндексовані первинним ключем, який в HBase має назву RowKey. Для кожного RowKey зберігається необмежений набір атрибутів (або колонок). Колонки організовані в групи колонок і мають назву ColumnFamily. Для кожного атрибуту може зберігатися декілька різних версій. Різні версії мають різний timestamp (часова мітка). Усі записи зберігаються у відсортованому по RowKey порядку.

Схожою на HBase є Apache Cassandra – колонково-орієнтована, розширювана розподілена система керування базами даних (РСКБД), розрахована на створення високомасштабованих і надійних сховищ величезних масивів даних, представлених у вигляді хеша. Вона може використовуватися як з Hadoop так і самостійно. Cassandra має свою мову структурованих запитів CQL, яка дещо нагадує SQL. Cassandra є доцільним вибором у наступних умовах:
– швидке зчитування і запис даних (нині Cassandra є найшвидшою РСКБД);
– додавання нових машин, якщо необхідно більше потужності;
– надійна реплікація даних між датацентрами.

Реплікація – одна з технік масштабування баз даних (як реляційних так і NoSQL), яка, за рахунок постійного копіювання даних, значно зменшує ймовірність втрати даних.

З Hadoop можна використовувати інші NoSQL бази даних. Для роботи з великими об'ємами даних можна використати й інші засоби. Вибір тої чи іншої технології залежить від результатів проведеного аналізу поставленої задачі.

Список літератури

1. Big Data от А до Я. Часть 1 [Електронний ресурс] – Режим доступу до ресурсу: https://habrahabr.ru/company/dca/blog/267361/.
2. NoSQL базы данных: понимаем суть [Електронний ресурс] – Режим доступу до ресурсу: https://habrahabr.ru/post/152477/.