Таблиця 1

Порівняльний аналіз компонентів

Назва компоненту	Популярність	Мова створення	Складність встановлення	Час створення
Hive	5	Java	5	2007
Impala	4	C++	5	2013
Drill	2	Java	1	2015
Spark	3	Scala	5	2014
Tez	3	Java	4	2014

Джерело: розробка автора

Кулик Р.Ю.
студент,
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського

МЕТОДОЛОГІЧНІ АСПЕКТИ ПРОГНОЗУВАННЯ ЧАСУ ОБРОБКИ ЗАПИТУ НА ПЛАТФОРМІ HADOOP

Саме в наш час технічного прогресу та потреби в роботі з даними людям стали в нагоді бази даних. База даних – сукупність даних, організованих відповідно до концепції, яка описує характеристику цих даних і взаємозв'язки між їх елементами; ця сукупність підтримує щонайменше одну з областей застосування. В загальному випадку база даних містить схеми, таблиці, подання, збережені процедури та інші об'єкти. Дані у базі організовують відповідно до моделі організації даних. Таким чином, база даних, крім саме даних, містить їх опис та може містити засоби для їх обробки.

Але настав час, коли звичайні бази даних не задовольняють вже потребам користувачів. Дані в базах посттійно збільшуються та виникає проблема того, що з даними відбувається повільніша обробка. Саме через цю проблему з’явилася Big Data.

Великі Дані, на сьогоднішній момент, є одним з ключових драйверів розвитку інформаційних технологій. Цей напрямок, відносно новий для бізнесу, отримав широке поширення в західних країнах. Пов'язано це з тим, що в епоху інформаційних технологій, особливо після буму соціальних мереж, по кожному користувачеві Інтернету стало накопичуватися значна кількість інформації, що в кінцевому рахунку дало розвиток напрямку Big Data. Слід також зазначити, що Big Data є однією з найбільш швидкозростаючих сфер інформаційних технологій, згідно зі статистикою, загальний обсяг одержуваних і збережених даних подвоюється кожні 1,2 року.

З існуючих поширених підходів обробки даних найбільш популярним та застосовуваним є платформа Hadoop. Hadoop – вільна програмна платформа і каркас для організації розподіленої обробки великих обсягів даних (що міряється у петабайтах) з використанням парадигми MapReduce, при якій завдання ділиться на багато дрібніших відособлених фрагментів, кожен з яких може бути запущений на окремому вузлі кластера. До складу Hadoop входить також реалізація розподіленої файлової системи Hadoop Distributed Filesystem (HDFS), котра автоматично забезпечує резервування даних і оптимізована для роботи MapReduce-застосунків. Для спрощення доступу до даних в сховищі Hadoop розроблена БД HBase і SQL-подібна мова Hive, яка є свого роду SQL для MapReduce і запити якої можуть бути розпараллелені і оброблені кількома Hadoop-платформами [1].

Важливо сказати, що платформа Hadoop має багато компонентів, кожен з яких виконує обробку даних. Основники компонентами є Hive, Impala, Drill, Spark, Tez. Адже важливо зрозуміти, котрий компонент краще використовувати користувачам для кожного запиту окремо, робота з яким компонентом займе мінімальну кількість часу. Бо дані в великих базах можуть займати велику кількість терабайтів, а тобто їх обробка буде займати багато часу і використання правильного та найбільш підходящого компонента значно може скоротити час обробки й очікування користувачем.

Також є проблема того, що починаючи обробку певної кількості інформації користувач не знає який саме час знадобиться на обробку даних. Дуже корисним є спрогнозувати поведінку системи та час обробки. Використовуючи для цього дані по часу, що вже є.

Прогнозування – процес передбачення майбутнього стану предмета чи явища на основі аналізу його минулого і сучасного, систематично оцінювана інформація про якісні й кількісні характеристики розвитку обраного предмета чи явища в перспективі. Результатом прогнозування є прогноз – знання про майбутнє і про ймовірний розвиток сьогочасних тенденцій конкретного явища-об'єкту в подальшому існуванні.

Метою даної роботи є дослідження методологічних аспектів прогнозування швидкості обробки запиту на платформі Hadoop та проведення прогнозу даних з минулих випробувань.

Предмет дослідження – теоретичні та практичні аспекти обробки великих баз даних на платформі Hadoop.

Об’єктом дослідження є вивчення та аналіз великих баз даних та швидкості обробки їх в системі.

Під методами прогнозування розуміється сукупність прийомів і способів мислення, що дозволяють на основі ретроспективних даних, екзогенних (зовнішніх) і ендогенних (внутрішніх) зв'язків об'єкта прогнозування, їх змін вивести судження визначеної вірогідності відносно майбутнього його розвитку.

За ступенем формалізації методи прогнозування можна розділити на інтуїтивні і формалізовані. У тих випадках, коли через значну складність об'єкта прогнозування неможливо врахувати вплив багатьох факторів, використовуються інтуїтивні методи, засновані на оцінках експертів.

Розрізняють індивідуальні і колективні експертні оцінки. У групі формалізованих методів виділяють дві підгрупи: екстраполяції і моделювання. До першої підгрупи відносяться методи найменших квадратів, експонентного згладжування, ковзних середніх. До другого – структурне, мережне, матричне й імітаційне моделювання [2].

Інтуїтивні методи прогнозування як науковий інструмент вирішення складних неформалізованих проблем дають змогу отримати прогнозну оцінку стану розвитку об'єкта в майбутньому незалежно від інформаційної забезпеченості. їхня сутність полягає в побудові раціональної процедури інтуїтивно-логічного мислення людини в поєднанні з кількісними методами оцінки й обробки отриманих результатів.

До формалізованим методів належать методи екстраполяції і методи моделювання. Вони базуються на математичній теорії.

Серед методів екстраполяції широке поширення отримав метод підбору функцій, заснований на методі найменших квадратів (МНК). У сучасних умовах все більшого значення стали надавати модифікаціям МНК: методу експоненціального згладжування з регульованим трендом і методом адаптивного згладжування.

У залежності від способу отримання прогнозної інформації виділяють експертні і фактографічні методи. Останні засновані на фактографічній інформації, тобто інформації про об'єкт прогнозування і його минулий розвиток. Експертні методи базуються на інформації, отриманій від експертів.

За ступенем просторової і часової погодженості результатів прогнозу виділяють: одномірне прогнозування – рівнобіжне прогнозування окремих об'єктів без подальшого узгодження розрізнених прогнозів; багатомірне прогнозування – рівнобіжне прогнозування окремих об'єктів зі спробою подальшого узгодження результатів; перехресне прогнозування – установлення причинно-наслідкових залежностей між екзогенними змінними і їх впливом на прогнозований об'єкт; наскрізне прогнозування – імітація поводження системи в цілому, включаючи просторове і часове її дослідження і повне узгодження результатів [3].

Значне місце серед методів прогнозування швидкості обробки запиту на платформі Hadoop займають так звані комбіновані методи. До них відносяться методи зі змішаною формаційною основою, у яких як первинну використовують як фактографічну, так і експертну інформацію. Наприклад, при проведенні експертного опитування може бути використані фактографічна Інформація. І, навпаки, при екстраполяції тенденції, поряд з фактичними даними, – експертні оцінки.

Найпростішим методом для прогнозування швидкості обробки запиту на платформі Hadoop є прогнозування в якому використовують дані минулих років та застосовують до них лінійну регресію.

Лінійна регресія – це метод моделювання залежності між скаляром y та векторною (у загальному випадку) змінною X. У випадку, якщо змінна X також є скаляром, регресію називають простою.

Отже, Hadoop – вільна програмна платформа і каркас для організації розподіленої обробки великих обсягів даних (що міряється у петабайтах) з використанням парадигми MapReduce, при якій завдання ділиться на багато дрібніших відособлених фрагментів, кожен з яких може бути запущений на окремому вузлі кластера. Таким чином, в даному дослідженні розглянуто методологічні аспекти прогнозування часу обробки запиту на платформі Hadoop. Також описано види прогнозу та способи їх застосувань [4].

Список використаних джерел:

1. White, Tom. Hadoop: The Definitive Guide. – 2-nd edition. – Sebastopol: O'Reilly Media, 2011. – 600 p.; Ризниченко Г.Ю. Математические модели в биофизике и экологии. – Москва – Ижевск: Институт компьютерных исследований, 2003. – 83 с.

2. Задоя А.О. Мікроекономіка. Київ: Т-во «Знання», КОО, 2000, с. 176.

3. Прогнозування та планування в умовах ринку: Навч. посібник для Вузів / Під. ред. Т.Г. Морозової, А.В. Пікулькіна. – М.: ЮНИТИ – ДАНА, 1999. 3. Бокс Дж., Дженкінс Г. (1974) Аналіз тимчасових рядів. Прогноз і управління. – М.: Світ, 1974. – Вип. 1, 2.

4. Химмельблау Д. Прикладное нелинейное программирование. – М.: Мир, 1974. – 536 с.; Каліткін Н.Н. Численные методы. – М.: Главная редакция физико-математической литературы изд-ва «Наука», 1978. – 246 с.

Левківська Л.В.

кандидат технічних наук, доцент,
Національний транспортний університет

МАТЕМАТИЧНІ АСПЕКТИ ПРОБЛЕМ МЕХАНІКИ БУРІННЯ НАФТОГАЗОВИХ СВЕРДЛОВИН

Енергетика – основа сучасного господарства. За останні два століття світова енергетика пройшла у своєму розвитку два головні етапи і нині наблизилася до третього. Перший – вугільний етап тривав упродовж XIX і першої половини XX ст. коли переважало вугільне паливо. Другий – нафтогазовий етап розпочався в другій половині XX ст. і продовжується нині, що зумовлено багатьма перевагами нафти й газу як ефективніших енергоносіїв порівняно з твердим паливом. Третій етап – це поступовий перехід від використання переважно вичерпних мінеральних ресурсів до енергетичного палива, що грунтується на відновлюваних і невичерпних ресурсах, або до альтернативних джерел енергії (енергії Сонця, геотермальної енергії Землі, енергії морів і океанів, вітру, біоенергії, енергії термоядерних реакцій). Це пояснюється погіршенням гірничо-геологічних умов видобування палива і загостренням проблеми енергозабезпечення людства на початку XXI ст.

Проте нині нафтогазова промисловість залишається провідною галуззю енергетики. Сьогодні нафта і природний газ є основою світового паливно-енергетичного балансу. Продукти їх переробки широко використовуються у всіх галузях промисловості, сільського господарства, на транспорті і в побуті