Список використаних джерел:

1. Гребенюк В.А., Катасонов А.А. Учебный процесс и контроль знаний в системе виртуального образования / Гребенюк В.А., Катасонов А.А. – М. : Открытое образование. – 1999. – 128 с.
2. Філіпенко І. Вибір ПЗ для автоматизації управління / Філіпенко І. – М. : Корпоративні системи. – 2001. – № 3. – 65 с.
3. Тестування як ефективний метод перевірки професійної компетентності студентів. – 25.05.2017. – [Електронний ресурс]. – Режим доступу: http://osvita.ua/school/lessons_summary/edu_technology/15024/ – Загол. з екрану.

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

ДОСЛІДЖЕННЯ ТА ПОРІВНЯЛЬНИЙ АНАЛІЗ КОМПОНЕНТІВ ПЛАТФОРМИ HADOOP

В сучасному світі великі компанії, такі я Google, Facebook, Twitter, повинні зберігати та обробляти надвелику кількість даних. Але що ж робити, коли навіть Oracle DB не можи коректно виконати будь-який запит на заних, обсяг яких займає більше 5 ТБ. З цим питанням на допогу прийшов Apache Hadoop із системою розподілених обчислень.

Hadoop активно використовується у великих промислових проектах, надаючи можливості, аналогічні платформі Google Bigtable/GFS/MapReduce, при цьому компанія Google офіційно делегувала Hadoop та іншим проектам Apache право використання технологій, на які поширюються патенти, пов'язані з методом MapReduce. Одним з найбільших користувачів і розробників Hadoop є компанія Yahoo!, вона активно використовує цю систему в своїх пошукових кластерах (Hadoop-кластеру Yahoo, що складається з 40 тисяч вузлів, належить світовий рекорд швидкості сортування великого обсягу даних). Hadoop-кластер використовується в Facebook для обробки однієї з найбільших баз даних, в якій зберігається близько 30 петабайт інформації. Hadoop також лежить в основі платформи Oracle Big Data і активно адаптується компанією Microsoft для роботи з СУБД SQL Server, Windows Server і хмарній платформі Azure Cloud з метою створення нових продуктів для організації розподіленої обробки великих обсягів даних. Hadoop є одним з ключових ланок суперкомп'ютера IBM Watson, який виграв бій з найкращими гравцями телевізійної гри-вікторини «Jeopardy!».

MapReduce – це програмна модель та програмний каркас, що її реалізує, розроблені компанією Google для проведення розподіленої паралельної обробки великих масивів даних з використанням кластерів звичайних недорогих комп'ютерів. Програма MapReduce складається із функції Map(), яка обробляє пари ключ/значення і генерує набір проміжних пар ключ/значення, і

функції Reduce(), яка зводить докупи всі проміжні значення пов'язані з одним і тим же проміжним ключем.

Екосистема Hadoop складається з багатьох компонентів. Кожен компонент показує кращий результат виконання на різних типах даних або стисення. На данний час використання системи Hadoop є складно для кінцевого, технічно неосвідченого користувача. Якщо людина не знає, який компонент буде кращий для його потреб? Для цього буде представлений програмний продукт, який в автоматичному режимі буде сканувати запит і дані, і обирати найоптимальніший компонент для виконання запиту.

Нижче представлені та досліджені основні компоненти екосистеми Hadoop:

1) Apache Hive – система управління базами даних на основі платформи Hadoop. Дозволяє виконувати запити, агрегувати і аналізувати дані, що зберігаються в Hadoop. Apache Hive був створений корпорацією Facebook і переданий під відкритою ліцензією у власність фондом Apache Software Foundation. На сьогоднішній день ця система використовується компанією Netflix і доступна в Amazon Web Services через Amazon Elastic MapReduce. Hive вдає із себе движок, який перетворює SQL-запити в ланцюжки map-reduce завдань. Движок включає в себе такі компоненти, як Parser (розбирає входять SQL-запрсои), Optimimer (оптимізує запит для досягнення більшої ефективності), Planner (планує завдання на виконання) Executor (запускає завдання на фреймворку MapReduce. Для роботи hive також необхідно сховище метаданих. Справа в тому що SQL передбачає роботу з такими об'єктами як база даних, таблиця, колонки, рядки, клітинки і тд. Оскільки самі дані, які використовує hive зберігаються просто у вигляді файлів на hdfs – необхідно десь зберігати відповідність між об'єктами hive і реальними файлами [1].

2) Impala довела свою високу ефективність двигуна з самого початку. Навіть в якості початкового випуску продукції в 2013 році, він продемонстрував двічі більшу швидкість, ніж традиційні СУБД, і кожний наступний випуск продовжує демонструвати широкий розрив у продуктивності між архітектурою аналітичної базами даних Impala і SQL-на-Apache Hadoop альтернатив [2].

3) Apache Drill – проект Apache Software Foundation, у рамках якого розвивається рушій для організації виконання SQL-запитів над напів-структурованими даними, що зберігаються в NoSQL-сховищах. Особливістю рушія є незалежність від схеми зберігання даних, що дозволяє організувати аналіз даних у різних сховищах без попереднього визначення їхньої структури (schema-free). Зокрема, Apache Drill дає можливість виконувати інтерактивні запити мовою ANSI SQL для складних або постійно змінюваних структур даних, включаючи формати JSON, ProtoBuf, XML, AVRO і Parquet, а також таблиці HBase, без необхідності завдання схеми зберігання. Структура даних у сховищі розпізнається на льоту і перетворюється у внутрішню JSON-подібну модель даних, яка надає інформацію про структуру бази даних при виконанні SQL-запитів. Для обробки складних і вкладених типів даних в Apache Drill передбачено ряд розширень SQL. Як одне з практичних застосувань Apache

Drill називається можливість інтеграції зав'язаних на SQL систем бізнес-аналітики і сховищ великих обсягів даних на основі Apache Hadoop або MongoDB, а також сполучення існуючих продуктів з Hadoop через штатні інтерфейси JDBC/ODBC. Сирцевий код проекту написаний на мові Java [3].

4) Apache Tez надає API для розробників і рамки для написання власних програм пряжу, що дозволяє усунути спектр інтерактивних і пакетних робочих навантажень. Це дозволяє ці додатки доступу до даних для роботи з петабайт даних над тисячами вузлів. Бібліотека компонентів Apache Tez дозволяє розробникам створювати додатки Hadoop, які інтегрують спочатку з Apache Hadoop пряжу і виконувати добре в межах змішаних кластерів робочих навантажень. Так як Tez є розширюваним і вбудована, він забезпечує свободу вписаний в цілях того, щоб висловити високо оптимізовані додатки для обробки даних, що дає їм перевагу перед кінцевим користувачем облицювальний двигунів, таких як MapReduce і Apache Спарк. Tez також пропонує настроюється архітектуру виконання, що дозволяє користувачам висловлювати складні обчислення в якості потокових графів, що дозволяє динамічні оптимізацію продуктивності на основі реальної інформації про дані і ресурсів, необхідних для його обробки [4].

5) Apache Spark – високопродуктивний рушій для оброблення даних, що зберігаються в кластері Hadoop. У порівнянні з наданим у Hadoop механізмом MapReduce, Spark забезпечує у 100 разів більшу продуктивність при обробленні даних в пам'яті й 10 разів при розміщенні даних на дисках. Рушій може виконуватися на вузлах кластера Hadoop як за допомогою Hadoop YARN, так і у відокремленому режимі. Підтримується оброблення даних у сховищах HDFS, HBase, Cassandra, Hive та будь-якому форматі введення Hadoop (InputFormat). Spark може використовуватися як у типових сценаріях оброблення даних, схожих на MapReduce, так і для реалізації специфічних методів, таких як потокове оброблення, SQL, інтерактивні та аналітичні запити, рішення задач машинного навчання і робота з графами. Програми для оброблення даних можуть створюватися на мовах Scala, Java, Python та R. Spark після перебування в інкубаторі став первинним проектом Apache Software Foundation від лютого 2014. З компаній, котрі використовують Spark, відзначаються Alibaba, Cloudera, Databricks, IBM, Intel, Yahoo, Cisco Systems. У жовтні 2014 року Apache Spark встановив світовий рекорд при сортуванні 100 терабайт даних. Згідно опитування O'Reilly у 2015 році 17% дослідників даних використовують Apache Spark.

Отже, у цій роботі наведено інформацію про систему Hadoop та компоненти його екосистеми, а також, досліджено та проаналізовано основні компоненти екосистеми.

Список використаних джерел:

1. Hive [Електронний ресурс]. – Режим доступу: http://hive.apache.org/downloads.html
2. Impala [Електронний ресурс]. – Режим доступу: https://thomaswdinsmore.com/tag/apache-impala/
3. Drill [Електронний ресурс]. – Режим доступу: https://drill.apache.org/
4. Tez [Електронний ресурс]. – Режим доступу: https://tez.apache.org/

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.