УДК 336.72

Рижко Борис Володимирович
студент
НТУУ «Київський політехнічний інститут»
Рыжко Борис Владимирович
студент
НТУУ «Киевский политехнический институт»
Ryzhko B.
student
NTUU «Kyiv Polytechnic Institute»

СИСТЕМА АВТОМАТИЗОВАНОГО ЗБОРУ, ОБРОБКИ ТА АНАЛІЗУ ВЕЛИКИХ ОБСЯГІВ ДАНИХ

СИСТЕМА АВТОМАТИЗИРОВАННОГО СБОРА, ОБРАБОТКИ И АНАЛИЗА БОЛЬШИХ ОБЪЕМОВ ДАННЫХ

AUTOMATED SYSTEM OF BIG DATA COLLECTION, PROCESSING AND ANALYSIS

Анотація. *Робота присвячена розробці системи автоматизованого збору, обробки та аналізу великих обсягів даних.*
Ключові слова: *hadoop, великі дані, flume, обробка даних.*

Аннотация. *Работа посвящена разработке системы автоматизированного сбора, обработки и анализа больших объемов данных.*
Ключевые слова: *hadoop, большие данные, flume, обработка данных.*

Summary. *The work is dedicated to the development of an automated system of big data collection, processing and analyzing.*
Key words: *hadoop, big data, flume, data processing.*

Вступ

В умовах зростання обсягів даних виникає необхідність їх збору, обробки й аналізу для вироблення прогнозувань, отримання якісних та кількісних показників на основі даних і т.д.

Для досягнення даної мети недостатньо продуктивності одного комп'ютера, оскільки через той час, за який інформація буде оброблена одним комп'ютером, вона вже може стати неактуальною.

Існуючі рішення[1] мають великі обчислювальні можливості, які здебільшого направлені на обробку даних, але виключають збереження даних та результатів обробки.

При реалізації системи були використані технології, що призначені для роботи з великими обсягами даних.

Прикладами використання таких систем можуть послугувати: моделювання ризиків, таргетування реклами, аналіз транзакцій у сфері роздрібної торгівлі, вироблення рекомендацій, аналіз даних на предмет прогнозування помилок (мереж і т.д.), пошукові системи.

Але побудована система не обмежена вказаним переліком.

Представлена в роботі система вирішує проблеми масштабовності:
– за умов збільшення даних,
– за умов збільшення джерел інформації.

Структура системи збору, обробки та аналізу великих обсягів даних

Виходячи з зазначених вимог до системи, було визначено наступні структурні елементи: інфраструктура розподілених обчислень (IPO), підсистема збору та обробки інформації, система керування базами даних (СКБД).

Структурна схема системи наведена на рис. 1.

Рис. 1. Структурна схема системи збору, обробки та аналізу (розробка автора)

Визначені елементи дозволяють розділити функції всієї системи, що значно полегшує налагодження системи і закладає принципи модульності. Така структура також визначає взаємозамінність кожної з підсистем у випадку необхідності.

Програмне забезпечення побудованої системи

З огляду на поставлені критерії, в якості ІРО було обрано Hadoop. Це рішення дозволяє розгорнути систему на комп'ютерах загального призначення [2]. За необхідності дана ІРО може бути розгорнута на хмарному рішенні, що дає змогу зекономити на придбанні технічного обладнання. Підсистема надає наступні можливості:

– розподілена файлова система,
– фреймворк розподілених обчислень.

Підсистемою збору та обробки даних виступає програмне забезпечення (ПЗ) Apache Flume[3]. Flume є розподіленим та надійним сервісом збору, обробки та переміщень великих обсягів даних. Він має просту та гнучку архітектуру, що базується на потоках даних. Володіє наступними характеристиками: модульний дизайн, масштабовність, сумісність з базою даних, сумісність з ІРО, гнучкі налаштування, розподіленість виконання задач, багатопоточність.

В якості СКБД виступає HBase. Дана СКБД надає наступні можливості: запис-зчитування у режимі реального часу, інтерфейс управління СКБД, конфігурація розподіленої СКБД між вузлами, додаткові інтерфейси для доступу зі сторонніх програм [4].

Архітектура системи збору, обробки та аналізу великих обсягів даних

Побудована архітектура на основі ПЗ та структурної схеми показана на рис. 2.

В розробленій архітектурі були враховані усі вимоги, що пред'являються структурною схемою та окремим програмним забезпеченням. Отримана архітектура системи відповідає типу master-slave.

Висновки

В результаті аналізу різних ІРО було виявлено, що на сьогоднішній день, за співвідношенням вартість-ефективність для розробки автоматизованої системи

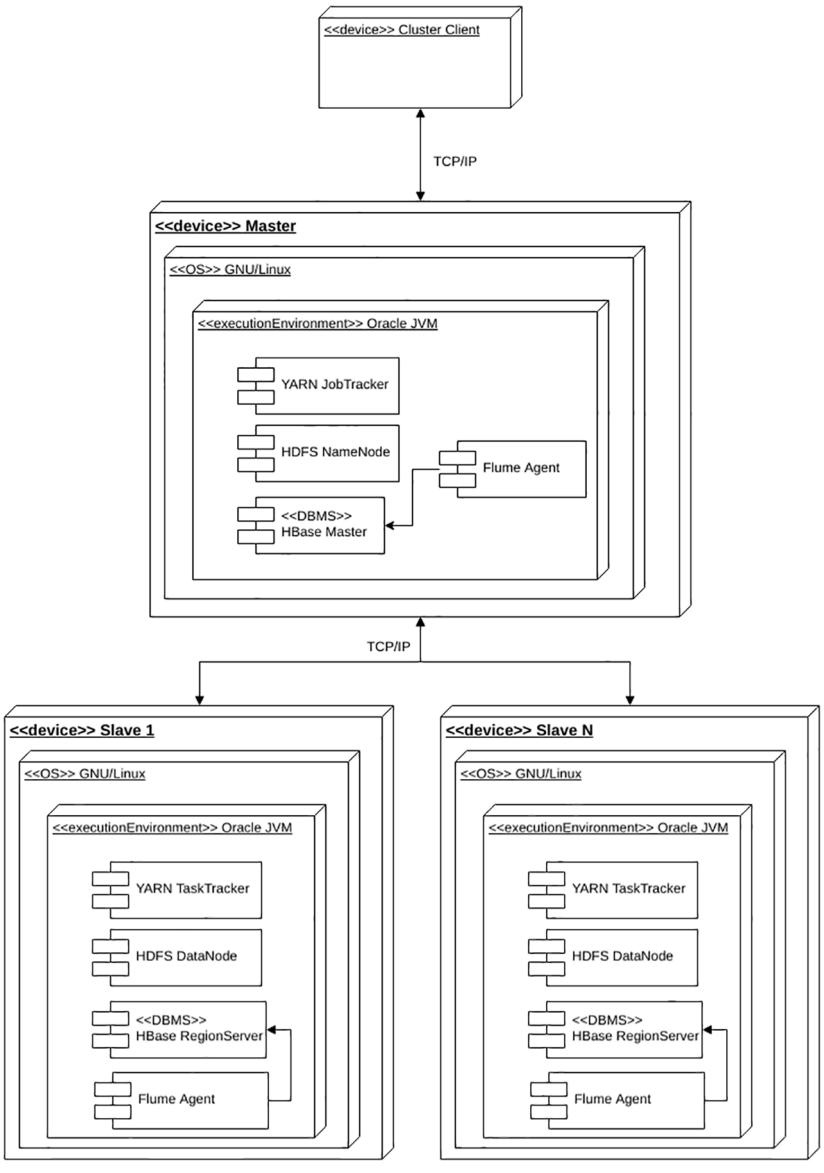

Рис. 2. Архітектура системи (розробка автора)

збору, обробки та аналізу інформації доцільно використовувати кластери, побудовані з використанням Hadoop для загального користування.

Розроблена система забезпечує: масштабовність, модульність, гнучкість у налаштуванні всіх параметрів. За співвідношенням вартості до ефективності розроблена система має переваги перед іншими аналогічними комерційними рішеннями.

Література

1. November 2015 | TOP500 Supercomputer Sites [Електронний ресурс] — Режим доступу до ресурсу: http://www.top500.org/lists/2015/11/ (дата звернення 30.05.2016). — Назва з екрана.

2. How-to: Select the Right Hardware for Your New Hadoop Cluster [Електронний ресурс] // Cloudera — Режим доступу до ресурсу: https://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/ (дата звернення 30.05.2016). — Назва з екрана.

3. Welcome to Apache Flume — Apache Flume [Електронний ресурс] — Режим доступу до ресурсу: https://flume.apache.org/ (дата звернення 30.05.2016). — Назва з екрана.

4. Apache HBase Reference Guide [Електронний ресурс] — Режим доступу до ресурсу: https://hbase.apache.org/book.html (дата звернення 31.05.2016). — Назва з екрана.