МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ
ФАКУЛЬТЕТ КОМП'ЮТЕРНО-ІНФОРМАЦІЙНИХ СИСТЕМ І ПРОГРАМНОЇ ІНЖЕНЕРІЇ

ДІДЕНКО ВОЛОДИМИР МИКОЛАЙОВИЧ

УДК 004.9:504:519.6

ТЕХНОЛОГІЇ ОПРАЦЮВАННЯ BIGDATA У СПЕЦІАЛІЗОВАНИХ КОМП'ЮТЕРНИХ СИСТЕМАХ НА БАЗІ ПЛАТФОРМИ HADOOP

123 «Комп'ютерні системи та мережі»

Автореферат
дипломної роботи на здобуття освітнього ступеня «магістр»

Тернопіль
2018

Роботу виконано на кафедрі комп'ютерних систем та мереж Тернопільського національного технічного університету імені Івана Пулюя Міністерства освіти і науки України

Керівник роботи: кандидат технічних наук, доцент кафедри комп'ютерних систем та мереж
Луцків Андрій Мирославович,
Тернопільський національний технічний університет імені Івана Пулюя,

Рецензент: доктор фіз.-мат. наук, професор, професор кафедри фізики
Дідух Леонід Дмитрович,
Тернопільський національний технічний університет імені Івана Пулюя,

Захист відбудеться 29 грудня 2018 р. о $9^{.00}$ годині на засіданні екзаменаційної комісії №34 у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, навчальний корпус №1, ауд.1-603

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми роботи. Сучасна аналітика, прогнозування, ціла низка природничих дисциплін, соціальні медіа та багато інших сфер зазнали ґрунтовних змін за останнє дисятиліття. Дисципліна «data mining», яка полягала в опрацюванні різних даних й була міждисциплінарним утворенням завдяки впливам ринку набула назви BigData й сьогодні є однією із трендовіших технологій у галузі IT. BigData знаходиться на межі багатьох дисциплін з галузей комп'ютерної, програмної інженерії та комп'ютерних наук; об'єднує засоби статистики, аналітики, візуалізації даних, машинного навчання та штучного інтелекту. Без стрімкого розвитку BigData важко уяви розвиток генетики, фізики, біології, космонавтики, цифрової економіки та низки інших дисциплін.

Швидке та правильне опрацювання BigData відіграє важливу роль у всіх сферах життєдіяльності людини. Наприклад, від достовірного прогнозу погоди залежать не лише економічні показники, але й життя багатьох людей. А прогнозування погоди — це один з прикладів системи великих даних: даних з багатьох точок земної кулі, величезною кількістю параметрів, що невплинно надходить у системи опрацювання даних. Фахівці з інженері даних є затребувані як і розробники програмного забезпечення. Добре спроектована та налаштована система опрацювання BigData є основою для систем штучного інтелекту та машинного навчання, функціонування яких можливе, виключно, за наявності потоку достовірних вхідних даних. У галузі комп'ютерної інженерії BigData відіграють роль «серця» й «мозку» ядра обчислювальних систем Інтернету речей.

Сфера BigData передбачає використання особливостей апаратних засобів, зокрема комп'ютерних мереж та архітектурних особливостей сучасних мікропроцесорних систем; системного програмного забезпечення: операційних систем, утиліт та засобів програмування. Реалізація програмного забезпечення має здійснюватись з урахування особливостей структур даних та алгоритмів: часової та просторової складностей. Усіма цими навичками повинен володіти дата-інеженер.

З урахуванням наведеного вище, важливою науково-практичною задачею є розроблення швидких та надійних систем опрацювання великих даних, які можуть бути використані для прогнозування та аналітики. Створення таких систем передбачає детальний аналіз програмних та апаратних компонентів та їх взаємодії.

У контексті великих даних особливе місце посідає технологія програмування Java, як відкрита, надійна, безпечна, швидка та ефективна платформа виконання BigData-додатків. Практично вся екосистема Hadoop є реалізована з використанням технології Java. Домінування Java не ринку ентерпрайз-вирішень також відіграло свою роль: інтеграція BigData у цю сферу відбувається доволі інтенсивно.

Мета роботи: Метою магістерського дослідження є обґрунтування вибору ефективних архітектур систем опрацювання BigData, а також методів, засобів та компонентів, які лежать в основі цих систем. У магістерському дослідженні здійснено вибір компонентів екосистеми Hadoop, які є складовими BigData-кластера. Система великих даних є орієнтованою на розв'язання аналітичних задач.

Досягнення цієї мети вимагає розв’язання таких завдань:

1. Проведення порівняльного аналізу існуючих патернів опрацювання великих даних на платформі Hadoop з точки зору можливості їх використання для поставлених задач та вибір архітектури програмно-апаратної системи.

2. Провести дослідження алгоритмічного та програмного забезпечення для розв'язання поставлених задач. Способів його взаємодії між собою з метою досягнення максимальної ефективності.

3. Обґрунтувати вибір алгоритмічних, програмних та апаратних засобів для реалізації системи опрацювання BigData.

4. Розробити та впровадити комп'ютерні програми для проведення аналітики на основі великих даних з метою апробації спроектованої системи, провести тестування з метою верифікації системи поставленим вимогам.

Об'єкт дослідження: обчислювальні системи BigData та обчислювальні процеси, які в них відбуваються.

Предмет дослідження: математичні моделі обчислювальних систем, методи декомпозиції обчислювальних задач, паралельні та розподілені обчислювальні системи, системи опрацювання BigData.

Методи дослідження: моделювання комп'ютерних систем та програм, теорія алгоритмів та обчислювальних методів, алгоритми та структури даних, теорія побудови обчислювальних систем.

Наукова новизна отриманих результатів:

1. На основі аналізу архітектур кластерних систем обґрунтовано вибір каппа-архітектури, як такої, що дає змогу опрацьовувати дані з високою пропускною здатністю й низькими показниками латентності. Систему побудовано на базі HDP-платформи, а саме сховища даних HDFS, менеджера ресурсів YARN, фреймворку Spark Streaming та Spark SQL, брокером повідомлень обрано Kafka. Для побудови системи використано відрите та безкоштовне програмне забезпечення.

2. Проведено обчислювальний експеримент з метою обґрунтування запропонованої архітектури та її компонентів у ході якого було реалізовано просту програму для аналітики файлів системного журналу.

3. Проаналізовано модель D4, як одного із патернів опрацювання даних на на розширеній платформі аналітики.

Практичне значення отриманих результатів.

Реалізовано програмну систему опрацювання великих даних на базі каппа-архітектури та компонентів екосистеми Hadoop. Розроблено проект системи, обґрунтовано вибір відповідних компонентів кластера, проведено його налаштування та тестування. Проведено обчислювальні експерименти з метою обґрунтування доцільності обраної архітектури програмно-апаратної системи з використанням фреймворку Apache Spark.

Апробація результатів дипломної роботи магістра. Результати дипломної роботи магістра апробовано на двох конференціях:

- міжнародній науково-технічний конференції молодих учених та студентів «Актуальні задачі сучасних технологій» (Тернопіль, ТНТУ, 2018);

- VI науково-технічній конференції «Інформаційні моделі, системи та технології» Тернопільського національного технічного університету імені Івана Пулюя (2018).

Структура роботи. Робота складається з розрахунково-пояснювальної записки та графічної частини. Розрахунково-пояснювальна записка складається з вступу, 6 розділів, висновків, переліку посилань та додатків. Обсяг роботи: розрахунково-пояснювальна записка – 130 арк. формату А4, графічна частина – 10 аркушів формату А1

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність й важливість даного дослідження та здійснено короткий огляд сучасного стану проблем у галузі високопродуктивних обчислювальних систем для опрацювання BigData. Охарактеризовано основні завдання, які необхідно вирішити у дипломній роботі магістра.

В розділі 1 «Аналіз предметної області комп'ютерних систем для опрацювання BigData» проведено аналіз предметної області BigData, апаратного та програмного забезпечення відповідних комп'ютерних систем. Детально проаналізовано екосистему Hadoop, а саме:

- надійне сховище даних - розподілена файлова система HDFS;
- фреймворки та засоби керування ресурсами MapReduce, Spark, Tez;
- SQL-подібні фреймворки: Hive, Impala, Shark, Spark SQL, Drill;
- NoSQL-сховище HBase;
- високопродуктивна система отримання даних на основі Kafka;
- потокове опрацювання даних Spark Streaming, яке забезпечує опрацювання даних у режимі реального часу;
- засоби машинного навчання на платформі Hadoop: Mahout та MLlib;
- ефективні формати зберігання даних Parquet, ORC, Thrift, Avro на платформі Hadoop.
- утиліти ZooKeeper, Hue, Flume, Sqoop, Oozie, Azkaban.

Обґрунтовано вибір доступних апаратних засобів для виконання програмного забезпечення екосистеми Hadoop. Сформульовано основні задачі дипломної роботи магістра.

В розділі 2 «Апаратне забезпечення та архітектури комп'ютерних систем опрацювання BigData» розглянуто приклад платформи для аналітики та типової задачі, яка розв'язується в межах патерну D4, а саме схеми прогнозування виходу з ладу пристрою радіозв'язку.

Проаналізовано та обґрунтовано вибір архітектури системи опрацювання даних. А також наведено компоненти, які можуть бути використані для реалізації такої архітектури. Всі компоненти є відкритими та безкоштовними. Таким чином нами буде використовуватись каппа-архітектура, а загальна схема потоку даних в Інтернеті речей яскраво продемонструвала доцільність такої архітектури.

Архітектура апаратного забезпечення виступатиме кластерна архітектура, як доступна, надійна та така, що добре масштабується горизонтально та вертикально.

Обґрунтовано вибір моделі обчислювальної системи, яка враховує апаратні параметри обчислювального Hadoop-кластера.

В розділі 3 «Засоби та методи розробки програмного забезпечення комп'ютерних систем опрацювання великих даних» реалізовано обчислювальну систему опрацювання BigData, орієнтовану на кілька вузлів. Яка реалізує поставлене

завдання, а саме може завантажувати дані, проводити їх опрацювання та обчислення й записувати результат. Для реалізації поставленого завдання розроблено інфраструктуру апаратного забезпечення, зокрема, з метою зменшення латентності мережі передачі даних запропоновано використовувати технологію Infiniband.

Проаналізовано й обґрунтовано вибір дистрибутиву Hortonworks HDP, показано процес його встановлення та розгортання. Для реалізації системи опрацювання великих даних з швидким часом відгуку використано фреймворк Apache Spark, використання якого можна вважати імплементацією каппа-архітектури, яка була проаналізована в другому розділі дипломної роботи магістра.

Показано створення та виконання Spark-програми для простої аналітики файлу системного журналу, як локально, так і на обчислювальному кластері в розподіленому режимі.

За допомогою типових засобів Hadoop здійснено тестування кластера з метою оцінювання його вузьких місць, помилок конфігурації та отримання його загальної оцінки в порівнянні з іншими аналогічними системами. Розглянуто питання моніторингу кластера штатними засобами HDP-дистрибутиву та програмною системою моніторингу Ganglia.

В розділі 4 «Обґрунтування економічної ефективності» показано доцільність проведення науково-дослідних робіт за даною тематикою і економічно обґрунтовано доцільність застосування запропонованих засобів. Розраховано вартість та ціну проведено науково-дослідної роботи.

В розділі 5 «Охорона праці та безпека в надзвичайних ситуаціях» розглянуто вимоги до охорони праці фахівців у галузі великих даних при роботі з комп'ютерними системами, до цих фахівців належать науковці, розробники програмного забезпечення, користувачі, а також розглянуті вимоги до організації серверних кімнат у яких може розташовуватись високопотужні системи опрацювання даних. Це дало змогу забезпечити належний рівень умов праці. Також розглянуто питання електробезпеки користувачів у відповідному підрозділі.

В розділі 6 «Екологія» проаналізовано абсолютні показники екологічних явищ, а також розглянуто роль науково-технічного прогресу в забезпеченні якісного стану довкілля.

У загальних висновках щодо дипломної роботи описано прийняті в роботі технічні рішення і організаційно-технічні заходи, які забезпечують виконання завдання на проектування; оригінальні технічні рішення, прийняті автором в процесі роботи.

В додатках до пояснювальної записки наведено фрагменти вихідного коду програм для аналітики на мові програмування Java з використанням фреймворку Spark .

В графічній частині наведено структурні схеми, які відображають архітектурні особливості створеної системи, її компоненти. А також наведено основні задачі роботи та основні результати розробки.

ВИСНОВКИ

У даній дипломній роботі магістра розроблено проект обчислювальної системи опрацювання Big Data. Основні результати та висновки проведених теоретичних та експериментальних досліджень такі:

1. Проаналізовано предметну область BigData, апаратного та програмного забезпечення відповідних комп'ютерних систем. Детально проаналізовано екосистему Hadoop й високопродуктивних обчислювальних систем та сформульовано рекомендації по вибору доступних апаратних та програмних компонентів високопродуктивних обчислювальних систем за критерієм вартості та доступності. Проаналізовано компоненти екосистеми Hadoop: HDFS, MapReduce, Spark, Tez, Hive, Impala, Shark, Spark SQL, Drill, HBase, Kafka та Spark Streaming.

2. Проведено аналітичне оцінювання продуктивності апаратних засобів для реалізації системи опрацювання Big Data. А також, проаналізовано підходи до формування архітектур ПРКС для такого класу системи. Зокрема, з можливістю інтеграції їх з різноманітними джерелами даних та спряження їх з існуючими ПРКС, системами моніторингу та керування завданнями.

3. Проаналізовано задачу здійснення аналітики та типової задачі, яка розв'язується в межах патерну D4, а саме схеми прогнозування виходу з ладу пристрою радіозв'язку.

4. Проаналізовано та обґрунтовано вибір архітектури системи опрацювання даних. А також наведено компоненти, які можуть бути використані для реалізації такої архітектури. Всі компоненти є відкритими та безкоштовними. Таким чином нами буде використовуватись каппа-архітектура, а загальна схема потоку даних в Інтернеті речей яскраво продемонструвала доцільність такої архітектури. Архітектура апаратного забезпечення виступатиме кластерна архітектура, як доступна, надійна та така, що добре масштабується горизонтально та вертикально.

5. Обґрунтовано вибір моделі обчислювальної системи, яка враховує апаратні параметри обчислювального Hadoop-кластера.

6. Реалізовано обчислювальну систему опрацювання BigData, орієнтовану на кілька вузлів, яка реалізує поставлене завдання, а саме може завантажувати дані, проводити їх опрацювання та обчислення й записувати результат. Для реалізації поставленого завдання розроблено інфраструктуру апаратного забезпечення, зокрема, з метою зменшення латентності мережі передачі даних запропоновано використовувати технологію Infiniband, використано дистрибутив Hortonworks HDP, показано процес його встановлення та розгортання.

7. Реалізовано програму для аналітики на основі Spark, використання якого можна вважати імплементацією каппа-архітектури, вибір якої обґрунтований у даній роботі магістра.

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Луцків А. М. Ключові особливості платформи Hadoop 3/ А.М. Луцків, В.М. Діденко // Актуальні задачі сучасних технологій : зб. тез доповідей міжнар. наук.-техн. Конф. Молодих учених та студентів, (Тернопіль, 28–29 листоп. 2018.) в 3-х томах /М-во освіти і науки України, Терн. націон. техн. ун-т ім. І. Пулюя [та ін.]. –

Тернопіль : ФОП Паляниця В. А., 2018 – Т. 2. – 45-46 с. [Електронний ресурс] Режим доступу: URL: http://elartu.tntu.edu.ua/bitstream/lib/26291/1/Book%202-2018.pdf
2. Луцків А. М. Архітектури високопродуктивних систем опрацювання великих даних / А.М. Луцків, В.М. Діденко //Збірник тез доповідей VI Науково-технічна конференція «Інформаційні моделі, системи та технології», 12-13 грудня 2018 року. — Т. : ТНТУ, 2018. —С.75.

АНОТАЦІЯ

Діденко В.М. Технології опрацювання BigData у спеціалізованих комп'ютерних системах на базі платформи Hadoop

Дипломна робота магістра, 123 – Комп'ютерні системи та мережі. – Тернопільський національний технічний університет імені Івана Пулюя, Тернопіль, 2018.

Дипломну робота магістра присвячено створенню системи опрацювання великих даних на основі платформи та екосистеми Apache Hadoop. Обґрунтовано архітектуру обчислювальної системи для опрацювання великих даних. У даній дипломній роботі розроблено обчислювальну систему опрацювання великих даних на базі програмних компонентів екосистеми Hadoop, зокрема HDFS, Spark, YARN та інших. Система складається з апаратної та програмної частин. Апаратна — це кластер на базі обчислювальних вузлів, які об'єднані в єдину мережу. Програмна — це операційна система, утиліти та спеціалізоване програмне забезпечення. Всі компоненти є відкритими та безкоштовними. Використовується каппа-архітектура системи опрацювання великих даних.

Для організації мережі використано обладнання D-Link, Hewllet-Packard, Mellanox як таке, що задовольняє поставленим вимогам: з точки зору експлуатаційних якостей, а також проаналізовано можливі апаратні платформи вузлів кластера й обґрунтовано їх вибір. Платформа Hadoop розгорнута на основі дистрибутиву Hortonworks HDP3. Розкрито питання тестування продуктивності, обслуговування (адміністрування) та моніторингу роботи розробленої системи.

Ключові слова: високопродуктивні обчислення, BigData, Hadoop, кластери

ANNOTATION

Didenko V. Technologies of BigData processing in specialized computer systems based on Hadoop platform

Master diploma thesis, 123 – Computer systems and networks - Ternopil Ivan Puluj National Technical University, Ternopil, 2018.

Master diploma thesis deals with the creation of a system for BigData processing based on the platform and the ecosystem Apache Hadoop. The architecture of the computer system for the processing of BigData is proposed. In this thesis the computational system of processing BigData on the basis of software components of the ecosystem Hadoop, in particular HDFS, Spark, YARN and others, was developed. The

system consists of hardware and software components. Hardware is a cluster based on computing nodes that are integrated into a single network. Software is an operating system, utilities and specialized software. All components are open source and free. Used kappa-architecture of the BigData processing system for.

To organize the network, the D-Link, Hewllet-Packard, Mellanox equipment was used as meeting the requirements: in terms of performance, as well as analysis of the possible hardware platforms of the cluster nodes and their choice. The Hadoop platform is deployed on the basis of the Hortonworks HDP3 distribution. The questions of testing of productivity, maintenance (administration) and monitoring of work of the developed system are revealed.

Key words: high-performance computing, weather forecasting, BigData, cluster