МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

Факультет програмної інженерії та бізнесу

Кафедра інженерії програмного забезпечення

Пояснювальна записка до дипломної роботи

магістра
(освітній ступінь)

на тему: «Експериментальне дослідження методів оброблення потоків даних у NRT-режимі»

ХАІ.603.667П1.22О.121.9580879 ПЗ

Виконав: студент 6 курсу групи 667П1
Галузь знань 12 «Інформаційні технології»
Спеціальність 121 «Інженерія програмного забезпечення»
(код та найменування)
Освітня програма Інженерія програмного забезпечення
(найменування)
Рубан В. В.
(прізвище й ініціали студента)
Керівник: Туркін І. Б.
(прізвище й ініціали)
Рецензент: Ляшенко О. С.
(прізвище й ініціали)

Харків – 2022

Міністерство освіти і науки України
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

Факультет програмної інженерії та бізнесу
(повне найменування)
Кафедра інженерії програмного забезпечення
(повне найменування)
Рівень вищої освіти другий (магістерський)
Галузь знань 12 «Інформаційні технології»
Спеціальність 121 «Інженерія програмного забезпечення»
(код та найменування)
Освітня програма «Інженерія програмного забезпечення»
(найменування)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ ___Ігор ТУРКІН____
(підпис) (ім'я та прізвище)
«___» _____ 2022 р.

З А В Д А Н Н Я
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ

Рубану Владиславу Валерійовичу
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи «Експериментальне дослідження методів оброблення потоків даних у NRT-режимі»
керівник кваліфікаційної роботи Туркін Ігор Борисович, д.т.н, професор
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом Університету № _____ від "____" _____ 2022 року

2. Термін подання здобувачем кваліфікаційної роботи "_12__" грудня 2022 року
3. Вихідні дані до кваліфікаційної роботи: результати огляду та аналізу проблем оброблення потоків даних у NRT-режимі

4. Зміст пояснювальної записки (перелік питань, які потрібно розробити):
1) огляд та аналіз проблемних питань оброблення потоків даних у NRT-режимі;
2) планування експериментальних досліджень методів оброблення потоків даних у режимі, близькому до реального часу;
3) проведення експериментального дослідження методів оброблення потоків даних у NRT-режимі; аналіз проведених досліджень.

5. Перелік графічного матеріалу: пояснювальна записка – 88 сторінок, 23 рисунка, 1 таблиця, 34 посилання на літературні джерела, 20 слайдів презентації.

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1	Туркін І. Б., завідувач каф. 603		
2	Туркін І. Б., завідувач каф. 603		
3	Туркін І. Б., завідувач каф. 603		

Нормоконтроль _____ Вероніка ПОСТЕРНАКОВА «26» грудня 2022 р.
(підпис) (ім'я та прізвище)

7. Дата видачі завдання 15.09.2021 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Отримання і затвердження теми кваліфікаційної роботи	15.09.21	
2	Огляд та аналіз проблемних питань оброблення потоків даних у NRT-режимі	16.09.21 - 20.12.21	
3	Планування експериментальних досліджень методів оброблення потоків даних у режимі, близькому до реального часу	15.02.22 - 26.05.22	
4	Проведення експериментального дослідження методів оброблення потоків даних у NRT-режимі. Аналіз проведених досліджень	12.09.22 - 23.10.22	
5	Наповнення контентом розділів пояснювальної записки	24.10.22 - 10.11.22	
6	Форматування пояснювальної записки до дипломної роботи відповідно до правил та вимог оформлення письмової роботи Національного аерокосмічного університету ім. М. Є. Жуковського «ХАІ»	11.11.22 - 24.11.22	
7	Передзахист дипломної роботи	23.12.22	
8	Захист дипломної роботи	27.12.22	

Здобувач _____ Владислав РУБАН
(підпис) (ім'я та прізвище)

Керівник кваліфікаційної роботи

_____ Ігор ТУРКІН
(підпис) (ім'я та прізвище)

РЕФЕРАТ

Пояснювальна записка до дипломної роботи містить 88 сторінок, 34 рисунка, 1 таблицю, 34 посилання на літературні джерела.

Об'єктом дослідження є процес розподіленої обробки Big Data.

Предметом дослідження є методи оброблення Big Data в умовах обчислень, близьких до реального часу.

Метою роботи є дослідження ефективності застосування сучасних методів обробки Big Data в умовах обчислень, близьких до реального часу у хмарному середовищі, шляхом порівняння паттернів обробки Big Data з інкрементальними та батч-підходами за рахунок побудови Big Data системи Spark Streaming.

Методи дослідження: паралельні та розподілені обчислення, об'єктно-орієнтований та функціональний підхід до розробки.

Наукова новизна отриманих результатів полягає у тому, що удосконалено метод розподіленої обробки неструктурованої і слабоструктурованої інформації, за рахунок інтеграції Apache Kafka і Spark Streaming шляхом оброблення Spark Streaming потоків повідомлень Apache Kafka і записання результату обробки в хмарне сховище, що забезпечує безперервний конвеєр (pipeline) збору та інтелектуальної обробки потоків Big Data в режимі реального часу.

Практичне значення отриманих результатів. Теоретичні результати будуть корисними при розгортанні розподіленої інфраструктури підприємствами, яким у своїй повсякденній діяльності необхідно збирати, агрегувати, обробляти і аналізувати безперервні потоки Big Data від різних джерел: пристроїв Internet of Things, додатків і сервісів.

BIG DATA; ОБЧИСЛЕННЯ, БЛИЗЬКІ ДО РЕАЛЬНОГО ЧАСУ; ХМАРНИ ОБЧИСЛЕННЯ; HADOOP; APACHE SPARK; SPARK STREAMING; SCALA; JAVA; AMAZON WEB SERVICES

ABSTRACT

Master's thesis contains 88 pages, 34 figures, 1 table, 34 references to literary sources.

*The object of researc*h is the process of distributed processing of Big Data.

The subject of the research is Big Data processing methods in the conditions of calculations close to real time.

The aim of the work is to study the effectiveness of the application of modern Big Data processing methods in computing close to real time in the cloud, by comparing Big Data processing patterns with incremental and batch approaches by building the Spark Streaming Big Data system.

Research methods: parallel and distributed computing, object-oriented and functional approach to development.

The scientific novelty of the results. The method of distributed processing of unstructured and weakly structured information has been improved due to the integration of Apache Kafka and Spark Streaming by processing Spark Streaming Apache Kafka message streams and recording the result of processing in the cloud storage, providing continuous pipeline (pipeline) for collecting and intelligent processing of Big Data streams in real time.

The practical significance of the results. Theoretical results will be useful when deploying a distributed infrastructure by enterprises that need to collect, aggregate, process and analyze continuous Big Data streams from various sources: Internet of Things devices, applications and services in their daily activities.

BIG DATA, NEAR REAL TIME COMPUTING, CLOUD COMPUTING, HADOOP, APACHE SPARK, SPARK STREAMING, SCALA, JAVA, AMAZON WEB SERVICES

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ 8

ВСТУП 9

1 АНАЛІЗ ПРОБЛЕМНИХ ПИТАНЬ ОБРОБЛЕННЯ ПОТОКІВ ДАНИХ У РЕЖИМІ, БЛИЗЬКОМУ ДО РЕАЛЬНОГО ЧАСУ 11

1.1 Дослідження методів аналізу даних у Big Data 11

1.2 Хронологія Big Data 14

1.3 Представлення (подання) даних 23

1.4 Постановка мети і завдань дослідження 24

1.5 Висновки з розділу 1 25

2 ПЛАНУВАННЯ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ МЕТОДІВ ОБРОБЛЕННЯ ПОТОКІВ ДАНИХ У РЕЖИМІ, БЛИЗЬКОМУ ДО РЕАЛЬНОГО ЧАСУ 26

2.1 Розподілені системи зберігання даних 26

2.1.1 Google File System 26

2.1.2 Hadoop File System 31

2.1.3 Google BigTable 34

2.1.4 HBase 38

2.2 Моделі програмування і технології розподіленої обробки 39

2.2.1 Модель програмування MapReduce 39

2.2.2 Платформа Apache Hadoop 44

2.2.3 Технологія Microsoft Dryad 45

2.3 Моделювання примітивів розподіленої системи обробки даних в Big Data 47

2.4 Обробка даних в Spark в режимі кластера 52

2.5 Обробка даних з Spark SQL 57

2.6 Обробка потоків даних з Spark Streaming 59

2.7 Висновки з розділу 2 61

3 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ МЕТОДІВ ОБРОБЛЕННЯ ПОТОКІВ ДАНИХ У РЕЖИМІ, БЛИЗЬКОМУ ДО РЕАЛЬНОГО ЧАСУ 62

3.1 Обробка потоків даних з Spark Streaming 62

3.2 Відмовостійкість обробки потоків даних з Spark Streaming 64

3.3 Уніфікація обробки потоків даних з Spark Streaming 66

3.4 Уніфікація Structured Streaming та віконні обчислення 67

3.5 Порівняння з Flink 73

3.6 Інтеграція з Apache Kafka 75

3.7 Аналіз проведених досліджень 79

3.10 Висновки з розділу 3 84

ВИСНОВКИ 85

ПЕРЕЛІК ПОСИЛАНЬ 86

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

AWS – Amazon Web Services;
BD – Big data;
NRT – nearly real-time;
SQL – structured query language;
БД – база даних;
ІС – інформаційні системи;
ОС – операційна система;
ПЗ – програмне забезпечення.

ВСТУП

У сучасному світі, стрімко розвиваються технології, потік інформації набирає обертів і відграє все важливішу роль, у людства як ніколи багато можливостей та інструментів для поліпшення якості власного життя, в тому числі завдяки отриманню цінної інформації з даних.

Розуміння запитів користувача і таргетинг – одні з найбільш великих і максимально освітлених широкому загалу областей застосування інструментів роботи з особистою інформацією. Методи Big Data (BD) допомагають аналізувати клієнтські звички, щоб в подальшому краще розуміти запити споживачів. Компанії прагнуть розширити традиційний набір даних інформацією з соціальних мереж і історією пошуку браузера з метою формування максимально повної клієнтської картини. Іноді великі організації вибирають глобальною метою створення власної рекомендаційної моделі.

Загальновідомим фактом є те, що Агентство національної безпеки Сполучених Штатів застосовує програмні інструменти, щоб запобігти терористичним актам. Інші відомства задіють прогресивну методологію, щоб запобігати більш дрібним злочини. Департамент поліції Лос-Анджелеса застосовує власну аналітичну систему для схожих цілей. Вона займається тим, що зазвичай називають проактивною охороною правопорядку. Використовуючи звіти про злочини за певний період часу, алгоритм визначає райони, де ймовірність скоєння правопорушень є найбільшою.

Використання IT-технологій в медичній сфері дозволяє лікарям більш ретельно вивчити хвороба і вибрати ефективний курс лікування для конкретного випадку. Завдяки аналізу об’ємів інформації з використанням підходів Big Data, медпрацівникам легшає передбачати рецидиви і вживати превентивні заходи. Як результат – більш точна постановка діагнозу і вдосконалені методів лікування.

У всіх перерахованих прикладах, що застосовуються в найважливіших областях людської життєдіяльності, є одна спільна риса – в основі кожної лежить структурування і обробка великої кількості даних. За будь-яким спрямованим перетворенням, безсумнівно, стоїть попередній аналіз. Він дозволяє, нам планувати та прогнозувати майбутнє, аналізувати минуле, приймати рішення. На практиці, ми маємо лише список фактів і набір оціночних критеріїв, що дозволяють отримати корисну інформацію завдяки використанню алгоритмів аналізу та обробки даних. Усвідомлення останнього факту поклало початок такому дивовижному напрямку IT індустрії, як BD.

Час минав, і дана технологія знайшла своє застосування майже в кожній області людської діяльності – її популярність набирає обороти, її актуальність, і застосування не залишає сумнівів.

Саме підвищений інтерес до методів і технологій розробки програмних продуктів, що виконують обчислення на розподілених кластерах, став підказкою до вибору об'єкта дослідження для даної роботи – розподілена обробка даних, як сфера технічних наук, набирає популярність і потребує залучення нових фахівців для розвитку і прогресу. Як не була б юна дана галузь ІТ-індустрії, області та напрямки її застосування вже встигли значно розростися, тому, для аналізу і вивчення в контексті даної роботи була обрана більш вузька, але популярна тема, розподілена обробка потоків даних у вигляді неструктурованих або слабоструктурованих даних (прикладом можуть послужити результати соціологічних опитувань, наукових експериментів, результати роботи датчиків, безпілотних автомобілів).

Об’єктом дослідження є процеси розподіленої обробки Big Data.

Предметом дослідження є методи оброблення Big Data в умовах обчислень, близьких до реального часу.

Метою роботи є дослідження ефективності застосування сучасних методів обробки Big Data в умовах обчислень, близьких до реального часу у хмарному середовищі, шляхом порівняння паттернів обробки Big Data з інкрементальними та батч підходами за рахунок побудови Big Data системи Spark Streaming.

Методи дослідження: паралельні та розподілені обчислення, об'єктно-орієнтований та функціональний підхід до розробки; методи розробки базуються на мовах програмування Scala, Java та Python.

Наукова новизна отриманих результатів. Удосконалено метод розподіленої обробки неструктурованої і слабоструктурованої інформації, за рахунок інтеграції Apache Kafka і Spark Streaming шляхом оброблення Spark Streaming потоків повідомлень Apache Kafka і записання результату обробки в хмарне сховище, що забезпечує безперервний конвеєр (pipeline) збору та інтелектуальної обробки потоків Big Data в режимі реального часу.

Практичне значення отриманих результатів. Теоретичні результати будуть корисними при розгортанні розподіленої інфраструктури підприємствами, яким у своїй повсякденній діяльності необхідно збирати, агрегувати, обробляти і аналізувати безперервні потоки Big Data від різних джерел: пристроїв Internet of Things, додатків і сервісів.