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

ВЕНГЕР ЮРІЙ РОМАНОВИЧ

УДК 004.9

МОДЕЛІ ТА ЗАСОБИ ОПРАЦЮВАННЯ BIGDATA З ВИКОРИСТАННЯМ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ MAPREDUCE

126 – Інформаційні системи та технології

Автореферат
дипломної роботи на здобуття освітнього ступеня «магістр»

Тернопіль
2018

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

Керівник роботи: доктор фізико-математичних наук,
професор кафедри комп'ютерних наук
Дідух Леонід Дмитрович,
Тернопільський національний технічний університет
імені Івана Пулюя,

Рецензент: кандидат фізико-математичних наук,
професор кафедри інформатики і математичного моделювання
Михайлишин Михайло Стахович,
Тернопільський національний технічний університет
імені Івана Пулюя,

Захист відбудеться _30_ грудня 2018 р. о 9^{00} годині на засіданні екзаменаційної комісії № _30_ у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, навчальний корпус №1, ауд. 702

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми роботи. Сучасні програми інтелектуального аналізу даних, які часто називають великими даними, вимагають від нас швидкого керування величезними обсягами даних. У багатьох з цих програм дані представлені у такому вигляді, що існують можливості для використання технологій паралельних обчислень. Для вирішення даних проблем був створений спеціальний стек програмного забезпечення.

Метою магістерської дипломної роботи є дослідження технології MapReduce та особливостей використання для вирішення задач опрацювання великих за обсягом масивів даних з використанням алгоритмів кластеризації та пошуку частих предметних наборів.

Об'єкт, методи та джерела дослідження: великі за обсягом масиви муніципальних даних. Для вирішення поставлених задач в цій дипломній роботі використовуються методи аналізу і синтезу, системного аналізу, порівняння, логічного узагальнення результатів, проектування логічних структур даних.

Як інформаційні джерела використовуються наукові публікації та інтернет джерела.

Наукова новизна отриманих результатів: полягає у розробці моделей вирішення задач кластеризації та пошуку часткових предметних колекцый з використанням інформаційної технології MapReduce.

Практичне значення отриманих результатів. Отримані результати можуть бути використані у майбутніх дослідженнях процесів прототипування концепту «Розумне місто».

Апробація. За результатами досліджень проведених в рамках магістерської роботи зроблено доповідь на VI науково-технічній конференції «Інформаційні моделі, системи та технології» 12-13 грудня 2018 року з публікацією тез доповіді.

Структура роботи. Робота складається з розрахунково-пояснювальної записки та графічної частини. Розрахунково-пояснювальна записка складається з вступу, 8 частин, висновків, переліку посилань та додатків. Обсяг роботи: розрахунково-пояснювальна записка – 145 арк. формату А4, графічна частина – 19 слайдів презентації.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі проведено огляд сучасного стану опрацювання BigData з використанням інформаційної технології MapReduce.

В першому розділі проведено аналітичний огляд наукових літературних та Інтернет джерел.

В другому розділі досліджено моделі для обробки великих даних на основі інформаційної технології MapReduce, зокрема розглянута задача кластеризації.

В третьому розділі продовжено дослідження моделей для обробки великих даних на основі інформаційної технології MapReduce, зокрема проаналізована задача пошуку часткових предметних наборів.

В четвертому розділі виконано налаштування Hadoop MapReduce.

В спеціальній частині А в розділі «Спеціальна частина» описано архітектуру розроблених програмних засобів тестування кластеризації та проаналізовано результати апробації запропонованих моделей.

В розділі «Обгрунтування економічної ефективності» розраховано основні техніко-економічні показники проведених досліджень моделей та засобів опрацювання BigData з використанням інформаційної технології MapReduce.

В розділі «Охорона праці та безпека в надзвичайних ситуаціях» розглянуто питання нормування параметрів мікроклімату. Проаналізовано вимоги до режимів праці та відпочинку користувачів ЕОМ. Зокрема досліджено особливості їх психофізіологічного розвантаження.

Також в розділі описано джерела, зони дії та рівні забруднення навколишнього середовища у разі аварій на хімічно і радіаційно небезпечних об'єктах, мінімізація негативного впливу. Розглянуто планування та порядок проведення евакуації населення з районів наслідків впливу НС техногенного та природного характеру.

В розділі «Екологія» досліджено відчуження земель під лінії електропередач, як один з елементів формуванні цілісного земельно-майнового комплексу та забезпечення ефективності використання обмежених територій з врахуванням потреб населення.

Також досліджено статистичний аналіз тендецій і закономірностей динаміки в екології, коли важливі не лише числові значення рівнів, але і їх послідовність.

У загальних висновках щодо дипломної роботи описано прийняті в дипломній роботі освітнього рівня «Магістр» наукові та технічні рішення і організаційно-технічні заходи, які забезпечують виконання завдання на проектування; оригінальні технічні рішення, прийняті автором в процесі роботи; технічні рішення роботи, які можуть бути впроваджені у виробництво;

В додатках до пояснювальної записки приведено ксерокопії тез доповіді.

В графічній частині подано тему, мету, об'єкт та предмет дослідження. Подано схему обчислень з використанням MapReduce. Проведено загальний огляд запуску програми MapReduce. Наведені графіки залежності пропускної здатності для кластерів різної конфігурації. Описана класична версія Hadoop MapReduce. Розглянутий навантажений JobTracker у великому кластері Apache Hadoop MapReduce v.1.0. Розглянуто послідовність запуску додатку у Hadoop MapReduce v.2.0 YARN. Наведена модель кластеризації даних за допомогою технології. Проаналізована модель вирішення задачі пошуку частих предметних наборів за допомогою MapReduce. Описані діаграма прецендентів для програми реалізації алгоритму та діаграма класів для програми реалізації алгоритму CURE. Розглянута діаграма послідовностей для програми реалізації алгоритму. Розглянута діаграма діяльностей для програми реалізації алгоритму. Наведено приклад файлу вхідних даних для алгоритму та приклад роботи алгоритму (до та після кластеризації), 5 кластерів. Графічно представлено час роботи прямого та розподіленого алгоритмів на локальному комп'ютері. Описані висновки.

ВИСНОВКИ

В процесі виконання дипломної роботи освітнього рівня «магістр» було досліджено моделі та засоби опрацювання BigData з використанням інформаційної технології MapReduce. В першому розділі дипломної роботи:

– Проаналізовано роль BigData для формування концепту «Розумне місто».

– Здійснено загальний огляд MapReduce.

– Проведено пошук альтернатив технології MapReduce.

– Проаналізовано застосування MapReduce у прикладних задачах муниципальних проектів.

– Виконано огляд фреймворку Hadoop MapReduce.

В другому розділі дипломної роботи досліджено моделі для обробки великих даних на основі інформаційної технології MapReduce, зокрема розглянута задача кластеризації.

В третьому розділі дипломної роботи продовжено дослідження моделей для обробки великих даних на основі інформаційної технології MapReduce, зокрема проаналізована задача пошуку часткових предметних наборів.

В четвертому розділі виконано налаштування Hadoop MapReduce. А в розділі «Спеціальна частина» описано архітектуру розроблених програмних засобів тестування кластеризації та проаналізовано результати апробації запропонованих моделей.

В шостому розділі дипломної роботи розраховано основні техніко-економічні показники проведених досліджень.

В сьомому розділі розглянуто питання нормування параметрів мікроклімату. Проаналізовано вимоги до режимів праці та відпочинку користувачів ЕОМ. Зокрема досліджено особливості їх психофізіологічного розвантаження.

Також в розділі описано джерела, зони дії та рівні забруднення навколишнього середовища у разі аварій на хімічно і радіаційно небезпечних об'єктах, мінімізація негативного впливу. Розглянуто планування та порядок проведення евакуації населення з районів наслідків впливу НС техногенного та природного характеру.

У восьмому розділі досліджено відчуження земель під лінії електропередач, як один з елементів формуванні цілісного земельно-майнового комплексу та забезпечення ефективності використання обмежених територій з врахуванням потреб населення.

Також досліджено статистичний аналіз тендецій і закономірностей динаміки в екології, коли важливі не лише числові значення рівнів, але і їх послідовність.

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Великі дані в проектах «Розумних міст» / [Венгер Ю.Р. та ін.]. // Матеріали VI Міжнародної науково-технічної конференції «Інформаційні моделі, системи та технології» Тернопільського національного технічного університету імені Івана Пулюя, (Тернопіль, 12 – 13 грудня 2018 р.). – Тернопіль: Тернопільський національний технічний університет імені Івана Пулюя – 2018. – С. 9.

АНОТАЦІЯ

Дипломна робота присв'ячена дослідженню моделей та засобів опрацювання BigData з використанням інформаційної технології MapReduce.

В першому розділі дипломної роботи проведено аналітичний огляд наукових літературних та Інтернет джерел.

В другому розділі дипломної роботи досліджено моделі для обробки великих даних на основі інформаційної технології MapReduce, зокрема розглянута задача кластеризації.

В третьому розділі дипломної роботи продовжено дослідження моделей для обробки великих даних на основі інформаційної технології MapReduce, зокрема проаналізована задача пошуку часткових предметних наборів.

В четвертому розділі виконано налаштування Hadoop MapReduce. А в розділі «Спеціальна частина» описано архітектуру розроблених програмних засобів тестування кластеризації та проаналізовано результати апробації запропонованих моделей.

Об'єкт дослідження: великі за обсягом масиви муніципальних даних.

Предмет дослідження: Моделі задач кластеризації та пошуку частих предметних наборів в інформаційних колекціях на основі технології MapReduce.

Метою даної дипломної роботи освітнього рівня «Магістр» є дослідження технології MapReduce та особливостей використання для вирішення задач опрацювання великих за обсягом масивів даних з використанням алгоритмів кластеризації та пошуку частих предметних наборів.

Основні результати: В результаті проведених досліджень створено моделі для опрацювання великих за обсягом даних та проведена їх апробація з використанням сучасних програмних засобів.

Ключові слова: ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, ВЕЛИКІ ДАНІ, РОЗПОДІЛЕНІ ОБЧИСЛЕННЯ, КЛАСТЕРИЗАЦІЯ, РОЗУМНЕ МІСТО, HADOOP, MAPREDUCE.

ANNOTATION

The thesis is devoted to the study of models and tools for processing BigData using the information technology MapReduce.

In the first section of the thesis an analytical review of scientific literary and Internet sources was conducted.

In the second section of the thesis, models for processing large data based on the information technology MapReduce, in particular, considered the problem of clustering.

In the third section of the dissertation the research of models for processing large data on the basis of information technology MapReduce is continued, in particular, the problem of finding partial subject sets is analyzed.

In the fourth section, the Hadoop MapReduce settings are configured. And in the "Special part" section, the architecture of the developed software tools for testing clustering is described and the results of approbation of the proposed models are analyzed.

Object of research: large in volume arrays of municipal data.

Subject of research: Models of tasks of clusterization and search of frequent subject sets in information collections on the basis of technology MapReduce.

The purpose of this diploma work of the educational level "Master" is the study of MapReduce technology and features of use for solving problems of processing large volumes of data arrays using clustering algorithms and the search for frequent subject sets.

Main results: As a result of the research, models for processing large volumes of data were created and tested with modern software tools.

Keywords: INTELLECTUAL ANALYSIS OF DATA, GREAT DATA, DISTRIBUTED CALCULATION, CLUSTERIZATION, SMART CITY, HADOOP, MAPREDUCE.