ДОДАТОК А

Слайди презентації

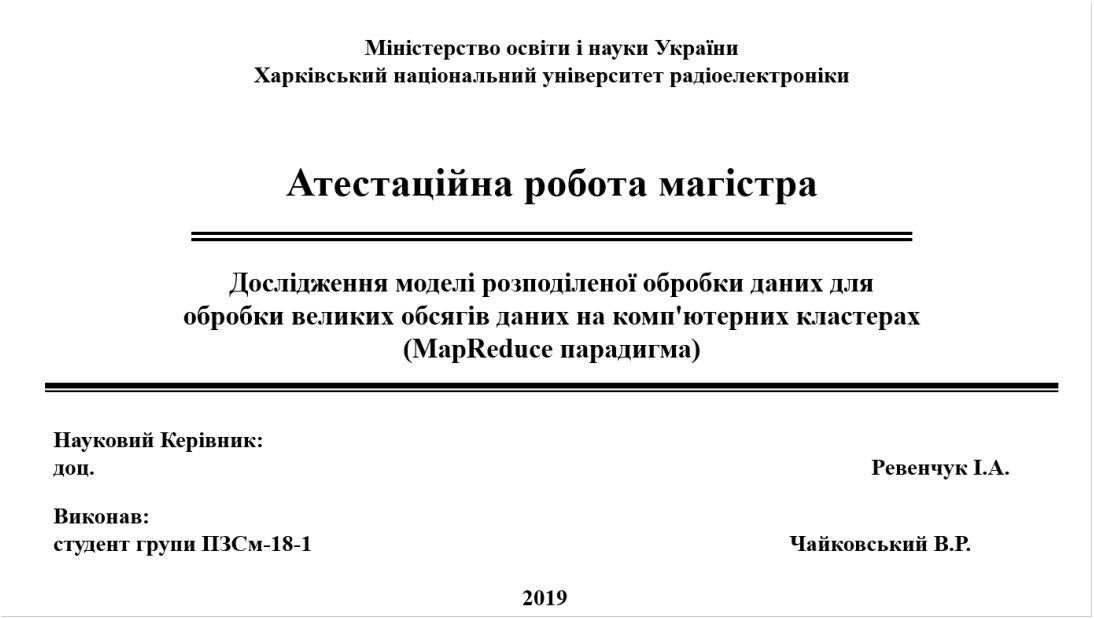


Рисунок 22 – Титульний слайд

Актуальність проблемної області

Методику і інструменти роботи зі структурованими даними IT-індустрія створила давно – це реляційна модель даних і системи управління БД. Але сучасною тенденцією є потреба обробки великого обсягу неструктурованих даних, і це та область, де старі підходи працюють погано або взагалі не працюють. Саме ця потреба вимагає нової методики поводження з даними, і зараз все більш популярною стає модель роботи з Big Data.

Рисунок 23 – Слайд «Актуальність проблемної області»

Аналоги

Apache Tez, Apache Spark та Apache Hadoop є основними фреймфорками для роботи із Big data.

Рисунок 24 – Слайд «Аналоги»

Постановка задачі

Метою даної роботи є дослідження застосовності і актуальності статистичної оптимізації в задачах MapReduce. Для цього необхідно:

- провести аналіз та моделювання предметної області;
- реалізувати збір статистики розподілу проміжних значень за проміжними ключам;
- реалізувати алгоритм, що забезпечує рівномірне навантаження на Reduce машини на основі зібраної статистики;
- реалізувати інструментарій для порівняння розподілів проміжних ключів по reduce-машинам;
- провести тестування та порівняльний аналіз оптимізованої версії на реальних завданнях.

Рисунок 25 – Слайд «Постановка задачі»

Аналіз методів що використовуються

Унікальність підходу великих даних полягає в агрегування величезного обсягу неструктурованою інформації з різних джерел в одному місці.

- Класифікація (методи категоризації нових даних на основі принципів, раніше застосованих до вже наявним даними раніше застосованих до вже існуючих даних)
- Кластерний аналіз
- Регресійний аналіз
- Рекомендаційні системи
- Штучні нейронні мережі, в тому числі генетичні алгоритми

Рисунок 26 – Слайд «Аналіз методів»

Екосистема Hadoop

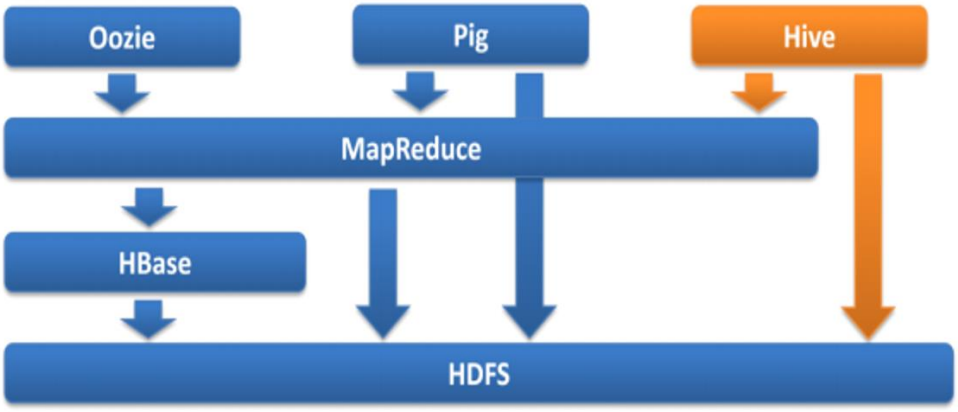


Рисунок 27 – Слайд «Екосистема Hadoop»

Інструменти екосистеми Apache Hadoop

- «Pig» – високорівнева мова опису потоків даних;
- «Hive» – SQL-подібна інфраструктура для організації сховищ даних;
- «HBase» – розподілена СУБД зі зберіганням по стовпцях, влаштована за зразком Google Bigtable;
- «ZooKeeper» – надійна система координації для управління станом, загальним для декількох розподілених додатків.

Рисунок 28 – Слайд «Інструменти»

Apache Pig

Pig підвищує рівень абстракції при обробці великих наборів даних.
Pig складається з двох основних частин:

- мова для опису потоків даних, званий Pig Latin;
- виконавча середовище для запуску програм Pig Latin. В даний час доступні два варіанти: локальне виконання на одній JVM і розподілене виконання в кластері Hadoop.

Програма Pig Latin складається з серії операцій (перетворень), які застосовуються до вхідних даних для отримання вихідних даних.

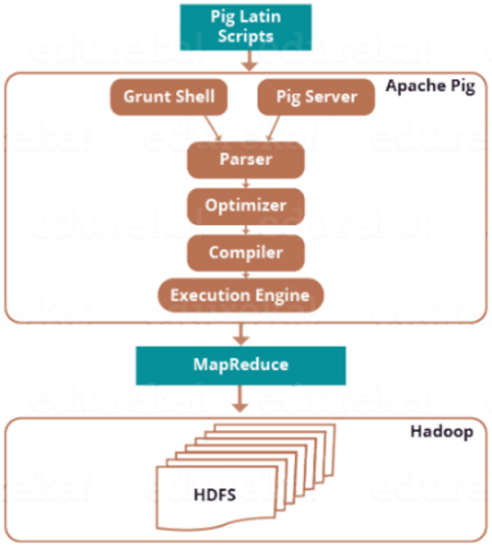


Figure: Apache Pig Architecture

Рисунок 29 – Слайд «Apache Pig»

Apache Hive

Hive предствляє движок, який перетворює SQL-запити в ланцюжок map-reduce завдань. Движок включає в себе такі компоненти, як Parser (розбирає SQL-запити які отримує на вхід), Optimizer (оптимізує запит для досягнення більшої ефективності), Planner (планує завдання на виконання) Executor (запускає завдання на фреймворку MapReduce.

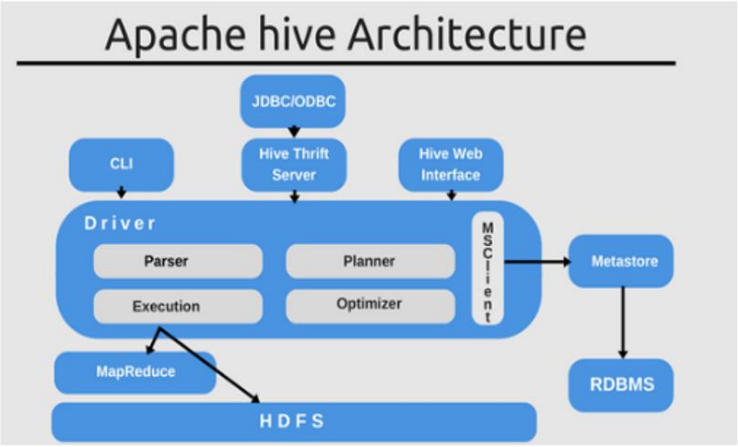


Рисунок 30 – Слайд «Apache Hive»

Apache HBase

HBase є розподіленою базою даних, яка може працювати на десятках і сотнях фізичних серверів, забезпечуючи безперебійну роботу навіть при виході з ладу деяких з них. Тому архітектура HBase є доволі складною у порівнянні з класичними реляційними базами даних.

HBase для своєї роботи використовує дві основні сутності. Region Server – обслуговує один або кілька діапазонів записів які відповідають певному діапазону ключів (Регіонів).
Master Server – головний сервер в кластері HBase. Master управляє розподілом регіонів, веде їх реєстр, управляє запусками регулярних завдань.

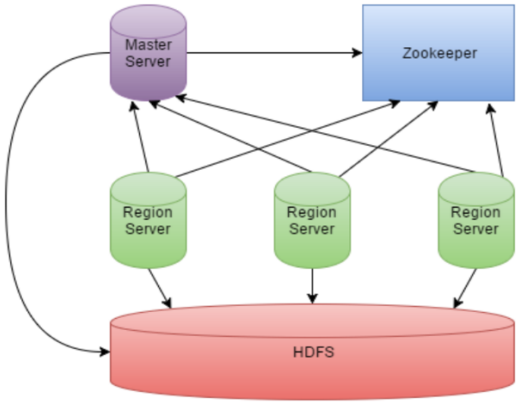


Рисунок 31 – Слайд «Apache HBase»

HDFS

HDFS» – розподілена файлова система, яка використовується в проекті «Hadoop».
Дані в «HDFS» зберігаються у вигляді блоків (блок – одиниця зберігання файлів) на «DataNode» і управляється через «NameNode». Причому під час запису файлу відбувається процес реплікації (копіювання) даного файлу.

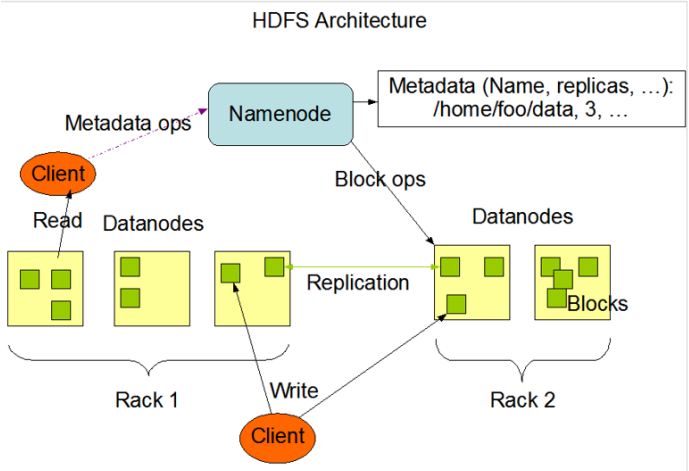


Рисунок 32 – Слайд «HDFS»

Програмна модель «MapReduce»

«MapReduce» - програмна модель для виконання розподілених обчислень для великих обсягів даних, що представляє собою набір «Java – класів» і виконуваних утиліт для створення і обробки завдань на паралельну обробку.

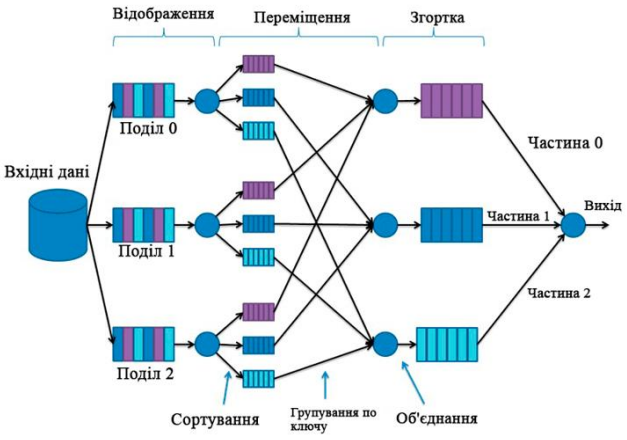


Рисунок 33 – Слайд «Програмна модель MapReduce»

Архітектура «MapReduce»

«MapReduce» використовує архітектуру «Master – Worker», де «Master» – єдиний екземпляр керуючого процесу («JobTracker»), запущений на окремій машині. «Worker» – це довільна кількість процесів «TaskTracker», що виконуються на «DataNode».

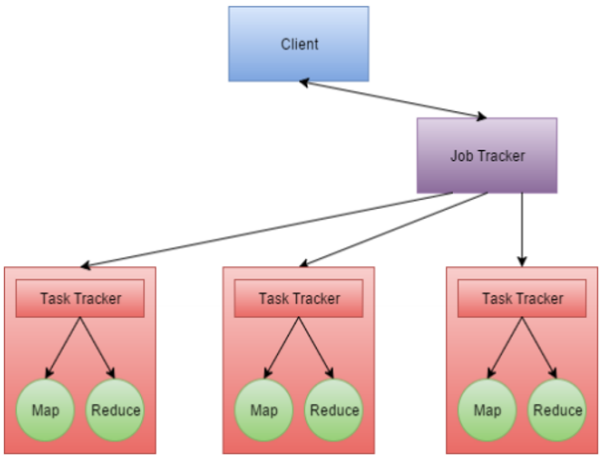


Рисунок 34 – Слайд «Архітектура MapReduce»

Математична модель

Задача пакування в ємності може бути задана як задача лінійного програмування наступним чином:

$$B = \sum_{i=1}^{n} y_i$$

$$\sum_{j=1}^{n} a_j x_{ij} \leq V y_i, \forall \in \{1, \dots, n\}$$

$$\sum_{i=1}^{n} x_{ij} = 1, \forall \in \{1, \dots, n\}$$

$$y_i \in \{0,1\}, \forall i \in \{1, \dots, n\}$$

$$x_{ij} \in \{0,1\}, \forall i \in \{1, \dots, n\} \; \forall j \in \{1, \dots, n\}$$

Таким чином, якщо у нас є ємності B, принаймні B – 1 ємність більш ніж наполовину заповнена. Тому є нижньою межею оптимального значення OPT, отримуємо, що $B - 1 < 2OPT$ і тому $B \leq 2OPT$

$$\sum_{i=1}^{n} a_i > \frac{B-1}{2} V$$

Рисунок 35 – Слайд «Математична модель»

Архітектура алгоритму

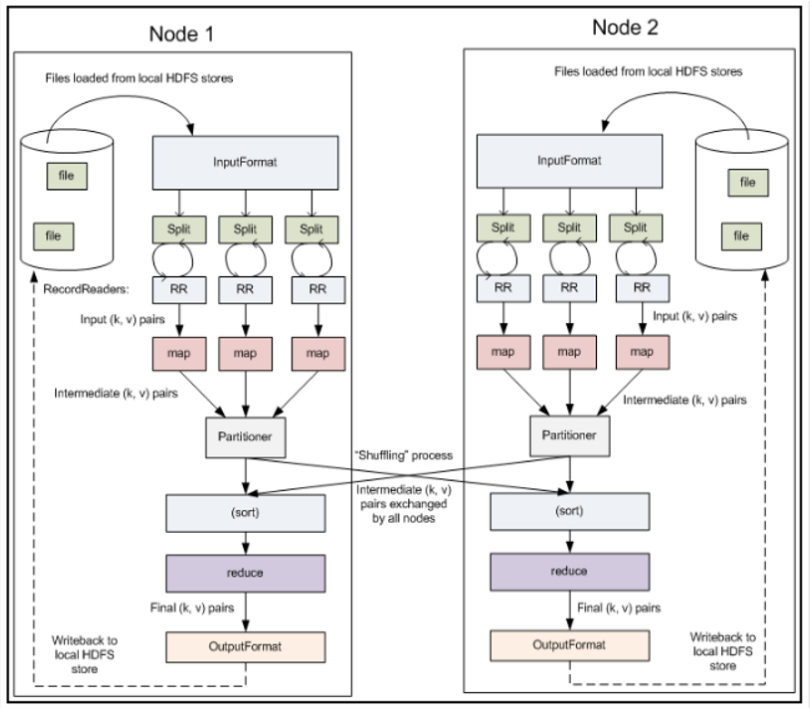
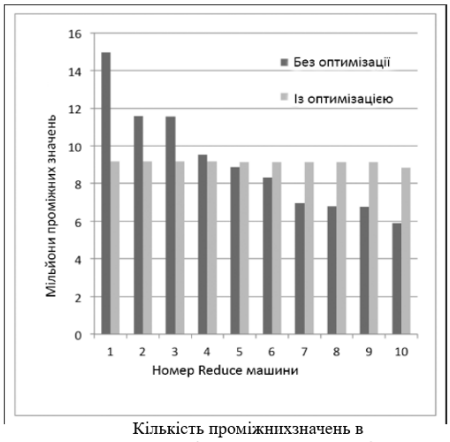


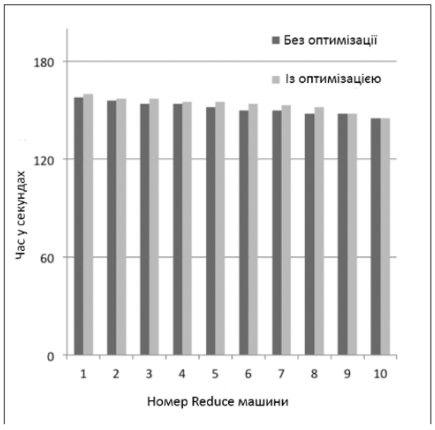
Рисунок 36 – Слайд «Архітектура алгоритму»

Word Count

Класичним і одночасно найпростішим прикладом MapReduce програми є Word count, завданням якої є підрахунок кількості входження слова в даний текст.



Кількість проміжнихзначень в експерименті з 10-ю завданнями reduce

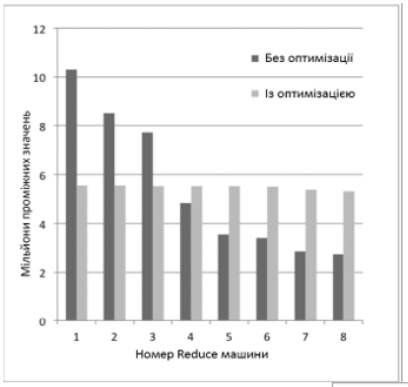


Час у тесту із 10-ю завданнями reduce

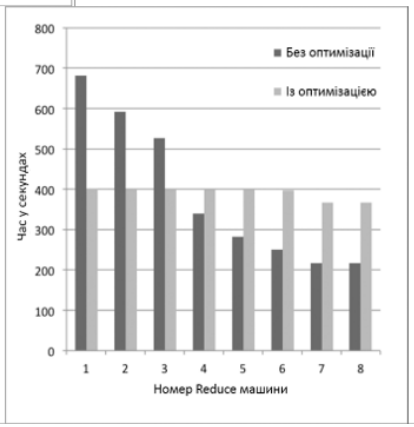
Рисунок 37 – Слайд «Word Count»

First Character

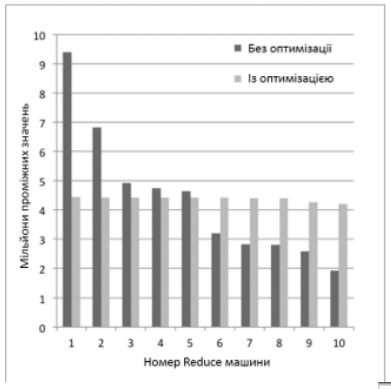
У наступному експерименті проводиться підрахунок кількості слів, що починаються з кожної з букв.



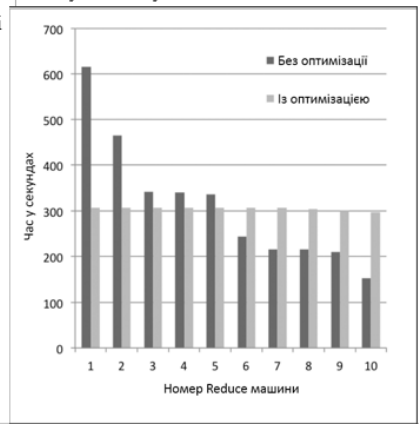
Кількість проміжних значень в тесті з 8-ю завданнями reduce



Час у тесті із 8-ю завданнями reduce



Кількість проміжних значень в експерименті з 10-ю завданнями reduce

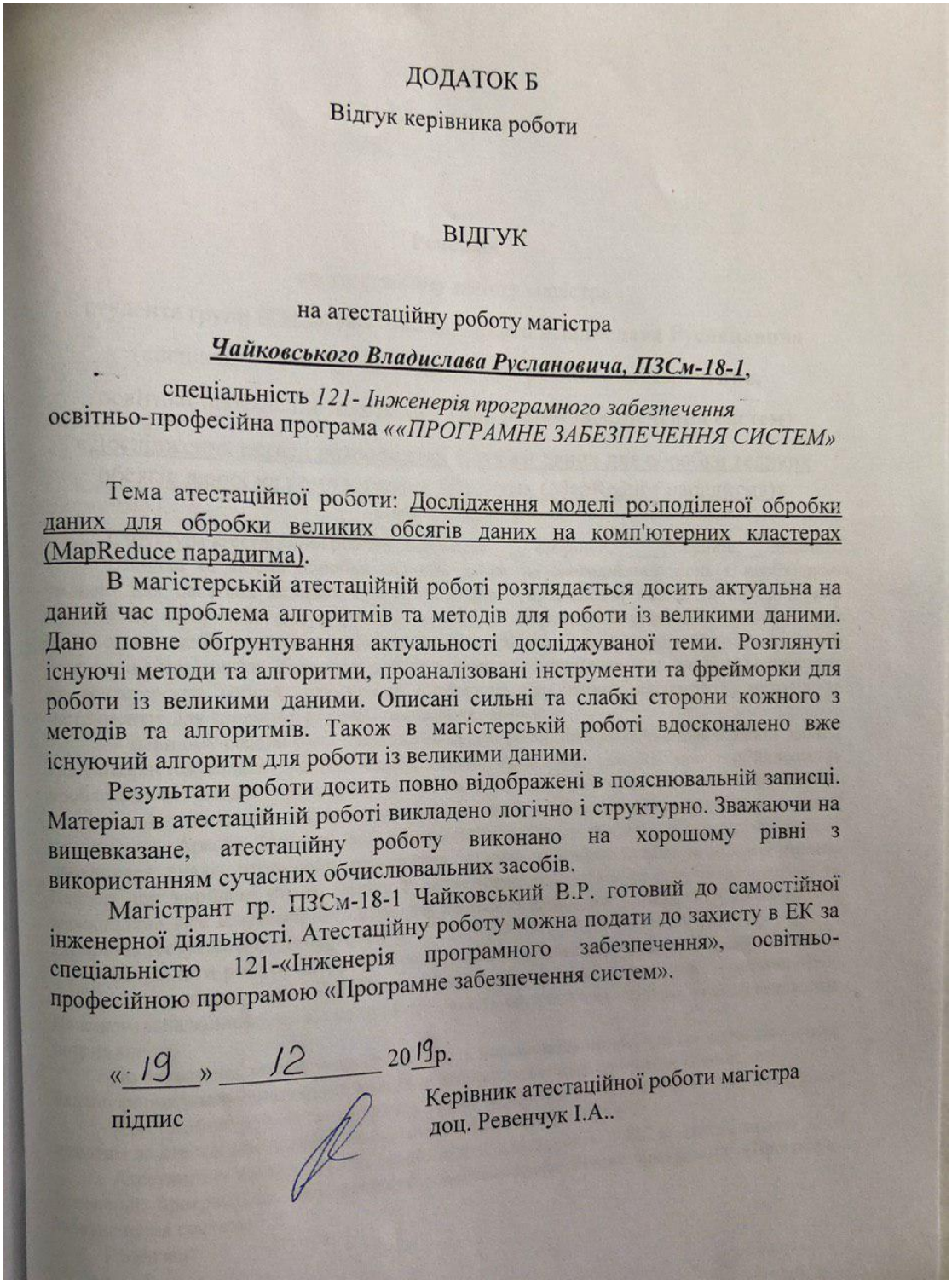


Час у час тесту із 10-ю завданнями reduce

Рисунок 38 – Слайд «First Character»

ДОДАТОК Б

Відгук керівника роботи



ДОДАТОК Б

Відгук керівника роботи

ВІДГУК

на атестаційну роботу магістра

Чайковського Владислава Руслановича, ПЗСм-18-1,

спеціальність *121- Інженерія програмного забезпечення*
освітньо-професійна програма *««ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ СИСТЕМ»*

Тема атестаційної роботи: Дослідження моделі розподіленої обробки даних для обробки великих обсягів даних на комп'ютерних кластерах (MapReduce парадигма).

В магістерській атестаційній роботі розглядається досить актуальна на даний час проблема алгоритмів та методів для роботи із великими даними. Дано повне обґрунтування актуальності досліджуваної теми. Розглянуті існуючі методи та алгоритми, проаналізовані інструменти та фрейморки для роботи із великими даними. Описані сильні та слабкі сторони кожного з методів та алгоритмів. Також в магістерській роботі вдосконалено вже існуючий алгоритм для роботи із великими даними.

Результати роботи досить повно відображені в пояснювальній записці. Матеріал в атестаційній роботі викладено логічно і структурно. Зважаючи на вищевказане, атестаційну роботу виконано на хорошому рівні з використанням сучасних обчислювальних засобів.

Магістрант гр. ПЗСм-18-1 Чайковський В.Р. готовий до самостійної інженерної діяльності. Атестаційну роботу можна подати до захисту в ЕК за спеціальністю 121-«Інженерія програмного забезпечення», освітньо-професійною програмою «Програмне забезпечення систем».

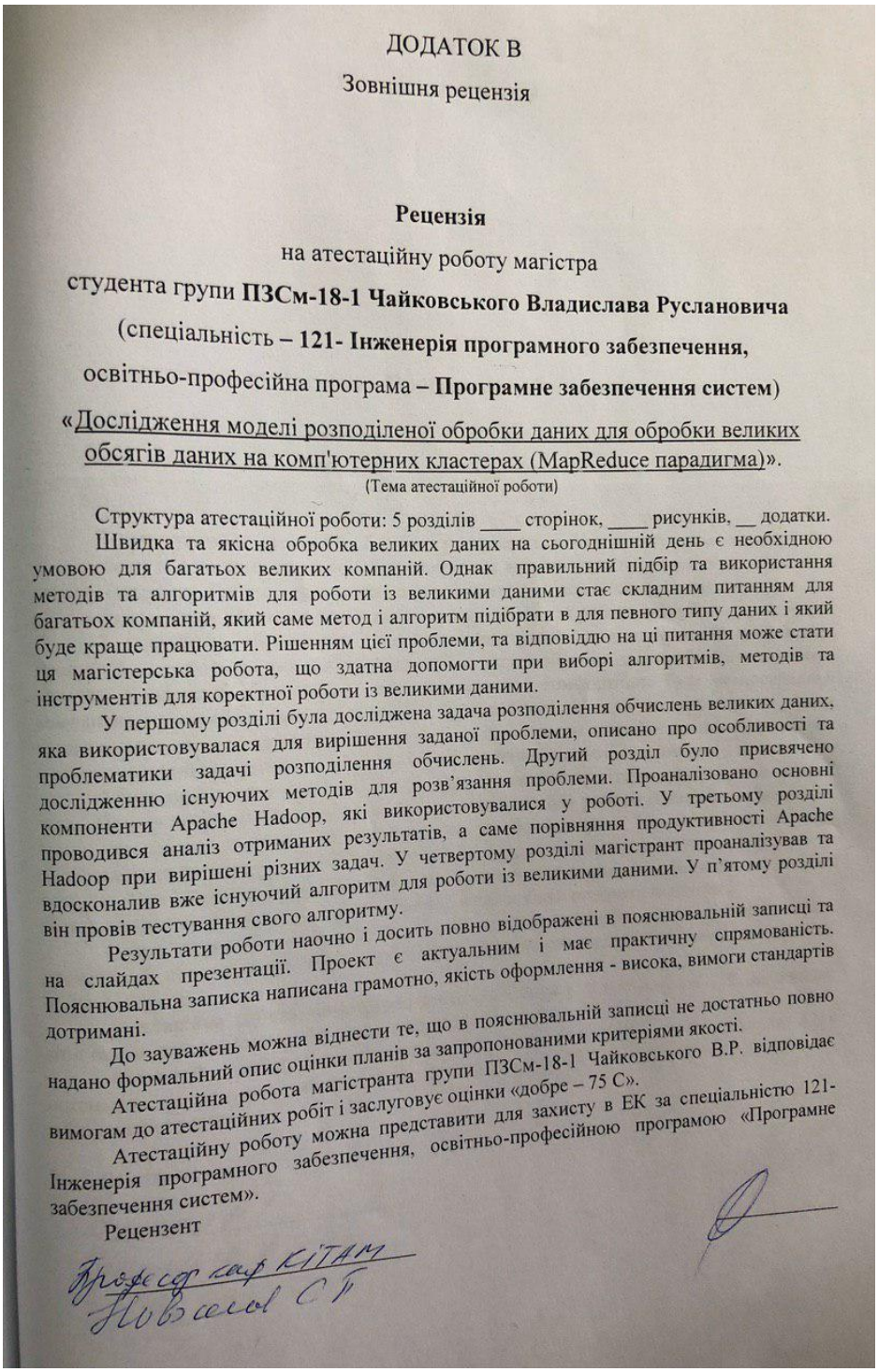
«19» 12 2019р.

підпис

Керівник атестаційної роботи магістра
доц. Ревенчук І.А..

ДОДАТОК В

Зовнішня рецензія



ДОДАТОК В

Зовнішня рецензія

Рецензія

на атестаційну роботу магістра

студента групи **ПЗСм-18-1 Чайковського Владислава Руслановича**

(спеціальність **– 121- Інженерія програмного забезпечення,**

освітньо-професійна програма **– Програмне забезпечення систем**)

«Дослідження моделі розподіленої обробки даних для обробки великих обсягів даних на комп'ютерних кластерах (MapReduce парадигма)».

(Тема атестаційної роботи)

Структура атестаційної роботи: 5 розділів ____ сторінок, ____ рисунків, __ додатки.

Швидка та якісна обробка великих даних на сьогоднішній день є необхідною умовою для багатьох великих компаній. Однак правильний підбір та використання методів та алгоритмів для роботи із великими даними стає складним питанням для багатьох компаній, який саме метод і алгоритм підібрати в для певного типу даних і який буде краще працювати. Рішенням цієї проблеми, та відповіддю на ці питання може стати ця магістерська робота, що здатна допомогти при виборі алгоритмів, методів та інструментів для коректної роботи із великими даними.

У першому розділі була досліджена задача розподілення обчислень великих даних, яка використовувалася для вирішення заданої проблеми, описано про особливості та проблематики задачі розподілення обчислень. Другий розділ було присвячено дослідженню існуючих методів для розв'язання проблеми. Проаналізовано основні компоненти Apache Hadoop, які використовувалися у роботі. У третьому розділі проводився аналіз отриманих результатів, а саме порівняння продуктивності Apache Hadoop при вирішені різних задач. У четвертому розділі магістрант проаналізував та вдосконалив вже існуючий алгоритм для роботи із великими даними. У п'ятому розділі він провів тестування свого алгоритму.

Результати роботи наочно і досить повно відображені в пояснювальній записці та на слайдах презентації. Проект є актуальним і має практичну спрямованість. Пояснювальна записка написана грамотно, якість оформлення - висока, вимоги стандартів дотримані.

До зауважень можна віднести те, що в пояснювальній записці не достатньо повно надано формальний опис оцінки планів за запропонованими критеріями якості.

Атестаційна робота магістранта групи ПЗСм-18-1 Чайковського В.Р. відповідає вимогам до атестаційних робіт і заслуговує оцінки «добре – 75 С».

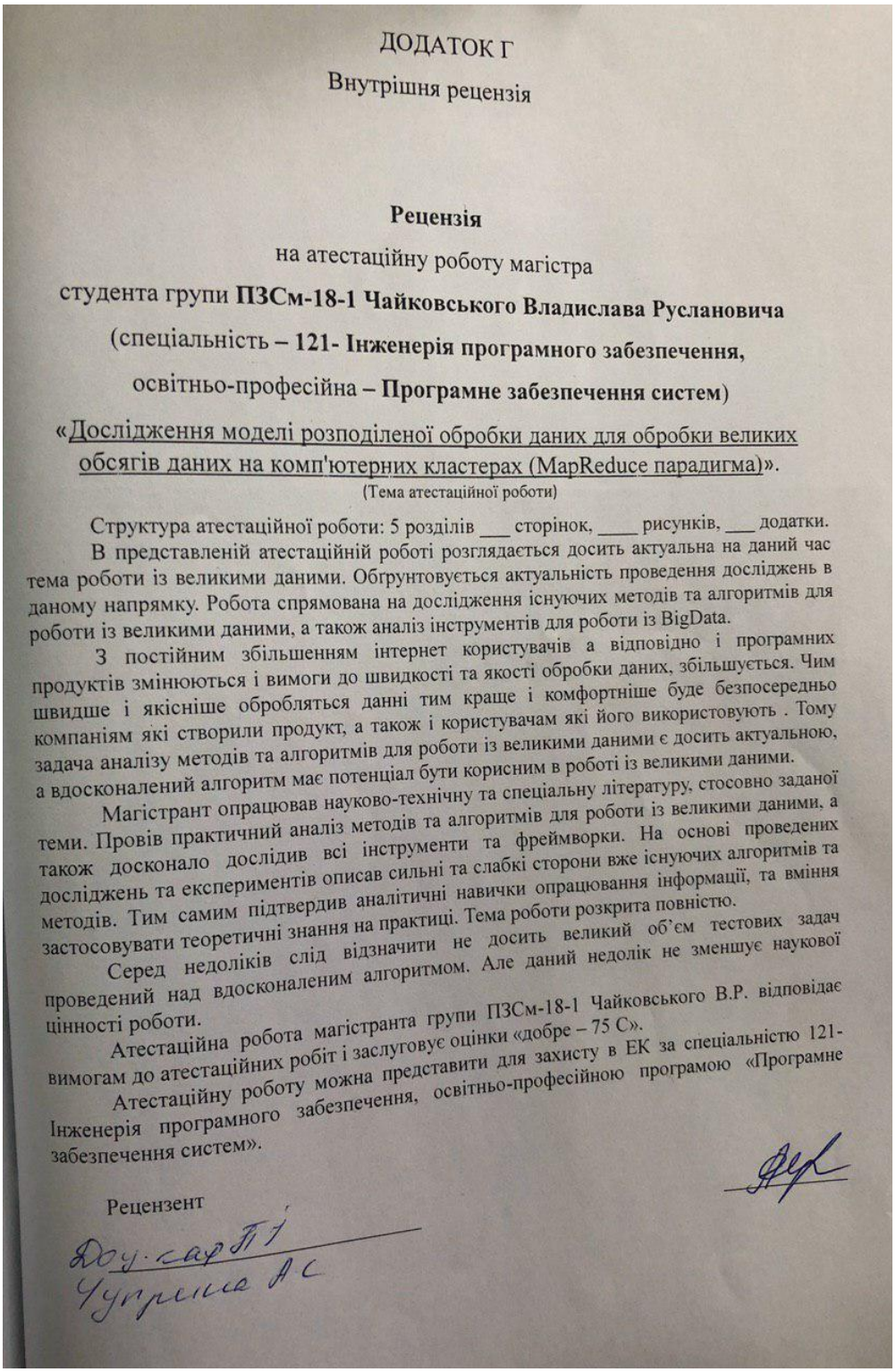
Атестаційну роботу можна представити для захисту в ЕК за спеціальністю 121- Інженерія програмного забезпечення, освітньо-професійною програмою «Програмне забезпечення систем».

Рецензент

Проф. каф. КІТАМ
Новосад С.Ф.

ДОДАТОК Г

Внутрішня рецензія



ДОДАТОК Г

Внутрішня рецензія

Рецензія

на атестаційну роботу магістра

студента групи **ПЗСм-18-1 Чайковського Владислава Руслановича**

(спеціальність **– 121- Інженерія програмного забезпечення,**

освітньо-професійна **– Програмне забезпечення систем**)

«Дослідження моделі розподіленої обробки даних для обробки великих обсягів даних на комп'ютерних кластерах (MapReduce парадигма)».

(Тема атестаційної роботи)

Структура атестаційної роботи: 5 розділів ___ сторінок, ____ рисунків, ___ додатки.

В представленій атестаційній роботі розглядається досить актуальна на даний час тема роботи із великими даними. Обгрунтовується актуальність проведення досліджень в даному напрямку. Робота спрямована на дослідження існуючих методів та алгоритмів для роботи із великими даними, а також аналіз інструментів для роботи із BigData.

З постійним збільшенням інтернет користувачів а відповідно і програмних продуктів змінюються і вимоги до швидкості та якості обробки даних, збільшується. Чим швидше і якісніше обробляться данні тим краще і комфортніше буде безпосередньо компаніям які створили продукт, а також і користувачам які його використовують . Тому задача аналізу методів та алгоритмів для роботи із великими даними є досить актуальною, а вдосконалений алгоритм має потенціал бути корисним в роботі із великими даними.

Магістрант опрацював науково-технічну та спеціальну літературу, стосовно заданої теми. Провів практичний аналіз методів та алгоритмів для роботи із великими даними, а також досконало дослідив всі інструменти та фреймворки. На основі проведених досліджень та експериментів описав сильні та слабкі сторони вже існуючих алгоритмів та методів. Тим самим підтвердив аналітичні навички опрацювання інформації, та вміння застосовувати теоретичні знання на практиці. Тема роботи розкрита повністю.

Серед недоліків слід відзначити не досить великий об'єм тестових задач проведений над вдосконаленим алгоритмом. Але даний недолік не зменшує наукової цінності роботи.

Атестаційна робота магістранта групи ПЗСм-18-1 Чайковського В.Р. відповідає вимогам до атестаційних робіт і заслуговує оцінки «добре – 75 С».

Атестаційну роботу можна представити для захисту в ЕК за спеціальністю 121- Інженерія програмного забезпечення, освітньо-професійною програмою «Програмне забезпечення систем».

Рецензент

Доц. каф. ПІ

Чуприна А.С