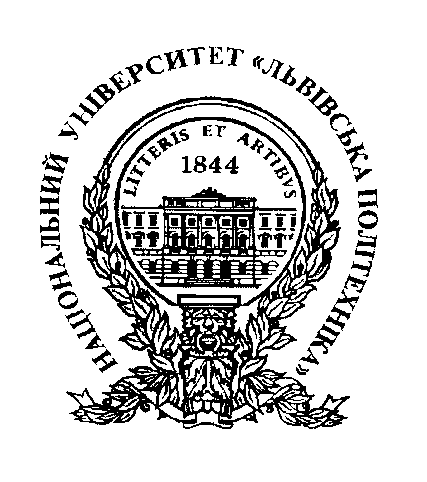
**Міністерство Освіти І НАУКИ України**

**Національний університет "Львівська політехніка"**

Інститут **КНІТ**

Кафедра **ПЗ**



***ПОЯСНЮВАЛЬНА ЗАПИСКА***

**На тему:** *“Аналіз ефективності роботи алгоритму apriori для напівструктурованих даних”*

**З дисципліни:** *“ Методи та засоби наукових досліджень в ІПЗ ”*

**Виконав:**

ст. гр. ПЗІП-12

Гаврилюк А.М.

**Перевірив:**

професор каф. ПЗ

Федасюк Д.В.

**Керівник:**

доцент каф. ПЗ

Левус Є.В.

« \_\_\_\_ » \_\_\_\_\_\_\_\_ 2017 р.

∑= \_\_\_\_ .

Львів – 2017

**Зміст**

[1. АНОТАЦІЯ 3](#_Toc502088905)

[2. ПРОБЛЕМАТИКА ДОСЛІДЖЕННЯ 4](#_Toc502088906)

[2.1. Прикладна проблема, на вирішення якої спрямовано проект: 4](#_Toc502088907)

[2.2. Об’єкт дослідження і розробки: 4](#_Toc502088908)

[2.3. Предмет дослідження і розробки: 4](#_Toc502088909)

[3. СТАН ДОСЛІДЖЕНЬ ПРОБЛЕМИ І ТЕМАТИКИ 5](#_Toc502088910)

[4. МЕТА, ОСНОВНІ ЗАВДАННЯ ТА ЇХ АКТУАЛЬНІСТЬ 6](#_Toc502088911)

[4.1. Ідеї та робочі гіпотези проекту: 6](#_Toc502088912)

[4.2. Мета і завдання, на вирішення яких спрямовано проект: 6](#_Toc502088913)

[5. ПІДХІД, МЕТОДИ, ЗАСОБИ ТА ОСОБЛИВОСТІ ДОСЛІДЖЕНЬ І РОЗРОБОК ЗА ПРОЕКТОМ 7](#_Toc502088914)

[6. ОЧІКУВАНІ РЕЗУЛЬТАТИ ВИКОНАННЯ ПРОЕКТУ ТА ЇХ НАУКОВА НОВИЗНА 8](#_Toc502088915)

[6.1. Докладно представити зміст очікуваних результатів, навести попередні описи методик та продуктів-аналогів: 8](#_Toc502088916)

[6.2. Заплановано представлення доповідей і наукових публікацій 8](#_Toc502088917)

[7. ПРАКТИЧНА ЦІННІСТЬ ПРОЕКТУ 9](#_Toc502088918)

[8. Логіко-структурна матриця. 10](#_Toc502088919)

[9. Календарний графік виконання робіт і фінансування, подані у вигляді діаграми Ганта. 14](#_Toc502088920)

[Висновки 15](#_Toc502088921)

[Презентація проекту 16](#_Toc502088922)

[Список використаної літератури 20](#_Toc502088923)

**Назва проекту:** Аналіз ефективності роботи алгоритму apriori для напівструктурованих даних

**Пропоновані строки виконання проекту:** 03.09.2017 – 25.05.2019

**Обсяг фінансування:** 68 135 грн.

1. **АНОТАЦІЯ**

ІТ є однією з найбільш динамічніших сфер, активно розробляються нові технології, створюються сотні програмних продуктів та накопичаються приблизно декілька сотень петабайт інформації у день. Ця інформація надходить з різних джерел, та містить багато напівструктурованих даних.

Прикладами напівструктурованих даних є XML та JSON. Структуровані дані зберігаються у базах даних. Вони легко опрацьовуються,отже з них легко отримати результат.

Big Data відноситься до надзвичайно великих масивів даних, які важко аналізувати традиційними інструментами. Великі дані можуть включати в себе як структуровані, так і неструктуровані дані. Багато інструментів, призначених для аналізу великих даних, можуть обробляти напівструктуровані дані.

Реалізована під час магістерського дослідження система у подальшому може бути використана як основа для створення рекомендаційної системи, оскільки використовує стек технологій Hadoop екосистеми, що дозволяє опрацювувати величезні масиви даних. А враховуючи темпи накопичення даних у найближчі роки Big Data стане найпопулярнішим напрямком ІТ.

Багато спостерігачів даної галузі кажуть, що Hadoop став де-факто галузевим стандартом для управління великими даними. Це проект з відкритим кодом управляється Apache Software Foundation.

1. **ПРОБЛЕМАТИКА ДОСЛІДЖЕННЯ**
   1. **Прикладна проблема, на вирішення якої спрямовано проект:**

Основною прикладною проблемою на вирішення якої спрямовується проект є робота алгоритму для пошуку асоціативних правил для напівструктурованих даних, оскільки кількість такого виду даних з кожним днем все більше зростає і невідомо як поведе себе алгоритм із таким типом даних.

Дана проблема розв’язується шляхом реалізації алгоритму та порівнянні результатів із структурованими даними.

* 1. **Об’єкт дослідження і розробки:**

Об`єктом дослідження і розробки є алгоритм apriori.

* 1. **Предмет дослідження і розробки:**

Предметом дослідження і розробки є час і точність пошуку асоціативних правил.

1. **СТАН ДОСЛІДЖЕНЬ ПРОБЛЕМИ І ТЕМАТИКИ**

В ході аналізу предметної області було розглянуто ряд статей та публікацій. Було проаналізовано декілька алгоритмів для пошуку асоціативних правил, та вибрано алгоритм apriori як один із найбільш досліджених та популярних алгоритмів для пошуку асоціативних правил для великих даних.

Переважно будь-яка сучасна система функціонує власне на базі певних предметних та інформаційних технологій. Впровадження і використання інтелектуальних систем бізнес-аналітики в компаніях дозволяє їм набувати конкурентоспроможності за рахунок накопичення інтелектуальної інформації та покращення ефективності процесу прийняття рішень. Після аналізу існуючих систем, що займаються пошуком закономірностей у даних, огляду існуючих підходів та алгоритмів для опрацювання великих структурованих і напівструктурованих даних було виявлено потребу у даному дослідженні. Опрацювання даним алгоритмом структурованих та напівструктурованих даних дозволяє оцінити його роботу, показати на яких засобах та на яких типах даних його використовувати ефективніше та доцільніше. А також розроблювана система дозволить працівникам сфер торгівлі чи послуг приймати оптимальні рішення щодо свого бізнесу. Прогнозувати обсяги виготовлення тої чи іншої продукції, або ж надання певних послуг в залежності від їх попиту серед клієнтів. Допоможе знайти певні закономірності для подальшого опрацювання бізнес-аналітиками.

1. **МЕТА, ОСНОВНІ ЗАВДАННЯ ТА ЇХ АКТУАЛЬНІСТЬ**
   1. **Ідеї та робочі гіпотези проекту:**

Основною ідеєю проекту є дослідження алгоритму apriori для напівструктурованих даних та порівняння цих результатів із результами роботи алгоритму для структурованих даних.

Перша гіпотеза полягає у реалізації алгоритму та створенні системи для вільного доступу та тестування роботи алгоритму.

Друга гіпотеза: якщо результати роботи алгоритму для напіструктурованих даних будуть такі ж як і для структурованих то напіструктуровані діна можна використовувати у тому вигляді я кому вони є, не потрібно додаткових маніпуляцій.

* 1. **Мета і завдання, на вирішення яких спрямовано проект:**

Основною метою проекту є дослідити роботу алгоритму apriori для напівструктурованих даних та порівняти із результами для структурованих.

Виходячи з мети, необхідно розв’язати такі завдання:

1. Ознайомлення із літературними джерелами в даній предметній області.
2. Ознайомлення та розуміння принципу роботи алгоритму, пошук переваг та недоліків.
3. Ознайомлення із технологією Hadoop.
4. Створення величезної бази даних товарів, що міститимуть інформацію про всі транзакції над певним набором продуктів та їх покупцями поділ джерел появи ризиків та потенційно ризикових подій по групам.
5. Реалізація алгоритму та створення веб додатку для завантаження та відображення інформації.
6. Дослідження алгоритму для структурованих даних.
7. Дослідження алгоритму для напіструктурованих даних.
8. Порівняння результатів дял структурованих та напіструктурованих даних.
9. **ПІДХІД, МЕТОДИ, ЗАСОБИ ТА ОСОБЛИВОСТІ ДОСЛІДЖЕНЬ І РОЗРОБОК ЗА ПРОЕКТОМ**

Розроблений програмний продукт буде мати клієнт серверну архітектуру. Реалізація алгоритму буде здійснюватись на серверній частині за допомогою .Net технологій, а саме мови C# та фреймворку ASP.Net Web API у середовищі Visual Studio 2015. Також, планується використовувати базу даних MS SQL Server, і технологію Hadoop. Ці технології були обрані тому що, вони надають необхідний набір можливостей для досягнення поставленої мети у розробці даного наукового проекту.

Особливості структури та складових проведення досліджень:

1. Аналіз публікацій та статей за тематикою дослідження.
2. Ознайомлення із технологією Hadoop.
3. Створення величезної бази даних товарів, що міститимуть інформацію про всі транзакції над певним набором продуктів та їх покупцями поділ джерел появи ризиків та потенційно ризикових подій по групам.
4. Реалізація алгоритму та створення веб додатку для завантаження та відображення інформації.
5. Тестування розробленої системи та аналіз результатів.

В загальному структуру проведення дослідження можна розділити на 3 етапи:

1. Аналітична частина, яка включає аналіз публікацій, створення та перевірка алгоритму.
2. Реалізація алгоритму.
3. Створення веб додатку.
4. **ОЧІКУВАНІ РЕЗУЛЬТАТИ ВИКОНАННЯ ПРОЕКТУ ТА ЇХ НАУКОВА НОВИЗНА**
   1. **Докладно представити зміст очікуваних результатів, навести попередні описи методик та продуктів-аналогів:**

У проекті передбачається :

* Реалізація алгоритму.
* Перевірка алгоритму на невеликих масив даних.
* Створення величезної бази даних товарів, що міститимуть інформацію про всі транзакції над певним набором продуктів та їх покупцями поділ джерел появи ризиків та потенційно ризикових подій по групам.
* Створення системи.
* Тестування даної системи.
* Перевірка системи для великих та різного типу даних.

На даний час алгоритм apriori добре дослідженний для структурованих та неструктурованих даних але для павіструктурованих досліджень не проводилось.

Кінцевим результатом проекту повинен бути програмний продукт, який дозволить шукати асоціативні правила як для структурованих даних так і для напівструктурованих даних які зберігатимусться в JSON форматі.

Вимоги до системи: мінімальний час на відповідь системи на певні дії користувача, повідомлення користувачу статусу виконання довготривалих операцій системи.

* 1. **Заплановано представлення доповідей і наукових публікацій**

В результаті виконання даного наукового проекту заплановано виступ на конференціях, та публікації. При можливості публікація буде видана англійською мовою, та попаде в науково метричні бази даних.

1. **ПРАКТИЧНА ЦІННІСТЬ ПРОЕКТУ**

В результаті програмної реалізації розроблених методів планується створення системи, яка зможе прогнозувати провал проекту за допомогою засобів нейронних мереж, на основі створеної формальної математичної моделі та даних про проект який буде використовувати цю систему.

Основними користувачами системи будуть проектні менеджери, або інші учасники команди з розробки програмного забезпечення на яких будуть поставлені обов’язки з контролю управління ризиками проектів.

Очікувані наукові результати направлені на:

* подальше покращення створеної моделі;
* застосування системи на нових не стандартних проектах;
* додавання нового функціоналу до системи, який покращить реалізацію програмних продуктів.

1. **Логіко-структурна матриця.**

*Таблиця 1. Логіко-структурна матриця*

|  |  |  |  |
| --- | --- | --- | --- |
| Ширші цілі. | Показники прогресу. | Як оцінювати показники. |  |
| Що являє собою більш широку ціль, якій сприятиме проект? | Які основні показники, пов’язані з більш ширшою метою? | Які є джерела інформації за цими показниками? |  |
| Дослідження доцільності застосування алгоритму апріорі для аналізу великих даних. | Час і точність роботи алгоритму. | Наукова стаття |  |
| Конкретні цілі. | Показники прогресу. | Як оцінювати показники. | Припущення та ризики. |
| Які є особливі цілі, які досягне проект? | Які є кількісні та  якісні індикатори, що показують чи та якою мірою конкретні цілі проекту  буде досягнуто? | Які джерела інформації існують і які можуть бути  зібрані? Які методи  є необхідні для отримання цієї інформації? | Які чинники та  умови, які не перебувають під прямим контролем проекту, необхідні для досягнення цих цілей? Які небезпеки необхідно взяти до уваги? |

*Продовження таблиці 1.*

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Дослідження роботи алгоритму для структурованих даних. 2. Дослідження роботи алгоритму для напівструктурованих даних. 3. Порівняння роботи алгоритму для структурованих та напіструктурованих даних. | Час роботи алгоритму.  Кількість знайдених асоціативних правил алгоритму. | Наукові статті.    Звіт про випробування системи. | Некоректний формат вхідних даних |
| Підсумки і результати | Показники прогресу. | Як оцінювати показники. | Припущення та ризики. |
| Список конкретних результатів - підсумки / результати, які призведуть до вирішення конкретних цілей | Які є показники  для вимірювання  чи і в якій мірі проект досягає запланованих резуьтатів та ефекту? | Які є джерела інформації за  цими показниками? | Які зовнішні чинники і умови повинні бути реалізовані для отримання очікуваних результатів та результатів за графіком? |

*Продовження таблиці 1.*

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Створено веб додаток для завантаження та відображення інформації. 2. Результати роботи алгоритму для різних типів даних. | Програмна система розроблена і протестована.  Результати роботи суттєво не відрізняються. | Звіти про виконання тестування системи  Звіт про виконання досліджень. | Тестування алгоритму на різних типах і обсягах даних. |
| Дії | Ресурси. | Витрати. | Припущення, ризики та передумови. |
| Які основні заходи будуть здійснюватись (згруповані у етапи роботи) та в якій послідовності вони будуть проводитись для того, щоб отримати очікувані результати? | Які ресурси  необхідні для здійснення цих видів діяльності? | Якою буде загальна вартість реалізації проекту? | Які попередні умови потрібні до початку проекту? Які умови  поза  прямого контролю проекту повинні бути присутніми  для здійснення запланованих заходів? |

*Продовження таблиці 1.*

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Ознайомлення із літературними джерелами в даній предметній області. 2. Ознайомлення та розуміння принципу роботи алгоритму, пошук переваг та недоліків. 3. Ознайомлення із технологією Hadoop. 4. Створення величезної бази даних товарів, що міститимуть інформацію про всі транзакції над певним набором продуктів та їх покупцями поділ джерел появи ризиків та потенційно ризикових подій по групам. 5. Реалізація алгоритму та створення веб додатку для завантаження та відображення інформації. 6. Дослідження алгоритму для структурованих даних. 7. Дослідження алгоритму для напіструктурованих даних. 8. Порівняння результатів дял структурованих та напіструктурованих даних. | * Персональний комп’ютер з всім додатковим обладнанням.(1 шт.) * Папір А-4(500 шт.) * Ручка.(2 шт.) * 240 робочих днів для програміста та його керівника(2 год на день.) * Ліцензія на необхідне програмне забезпечення(Visual Studio 2017 Enterprice, Microsoft Ofice) | 30 тис. грн. + 100 грн. + 15 грн. +36 тис. грн. + 2 тис. грн. + 20 грн. =  68 135 грн | Можливі додаткові витрати на техніку чи залучення додаткового персоналу.    Проект може не бути завершений за 240 робочих днів, тому можливі перевитрати бюджету та часу.    Додатковий час для навчання персоналу. |

1. **Календарний графік виконання робіт і фінансування, подані у вигляді діаграми Ганта.**

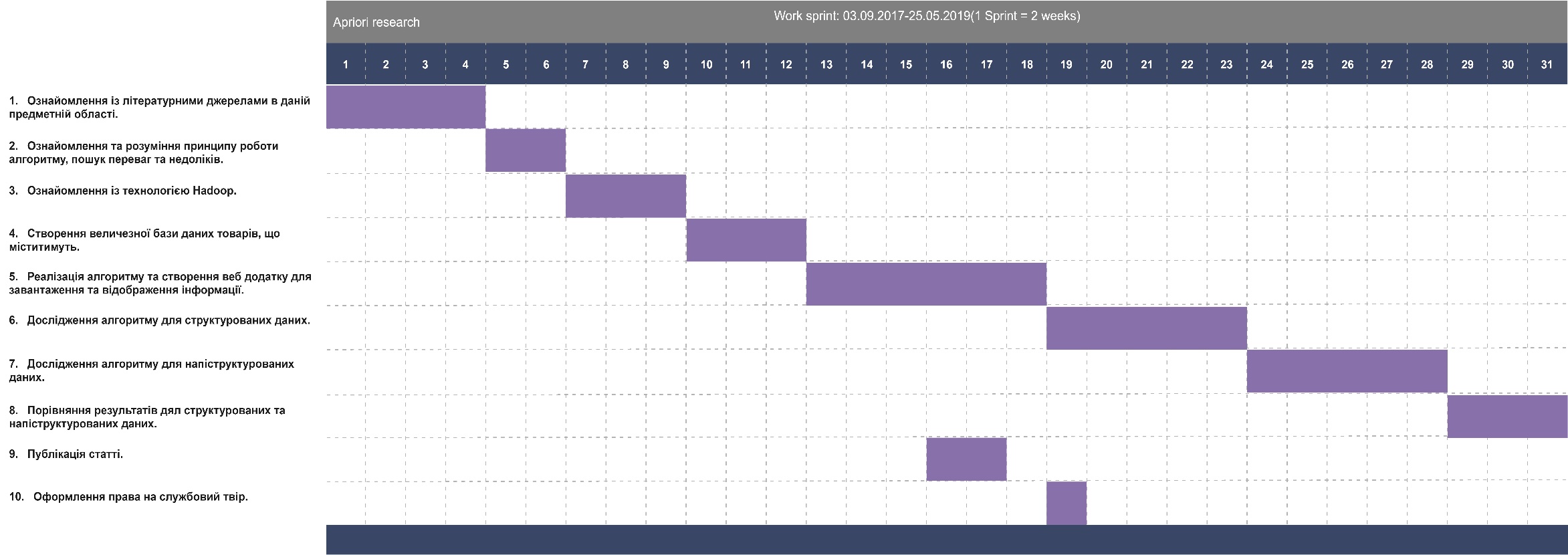
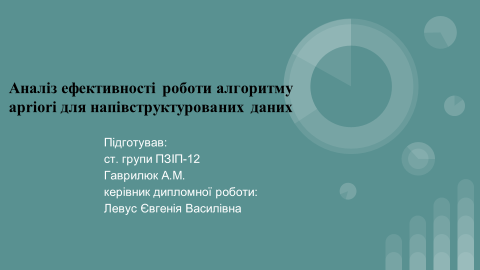


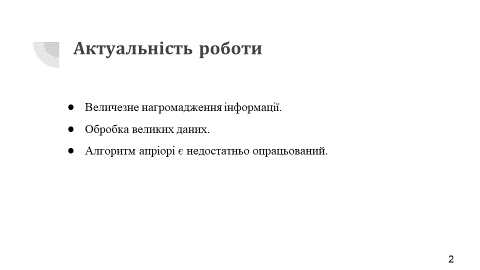
Рис.1 Графік виконання робіт у вигляді діаграми Гантта

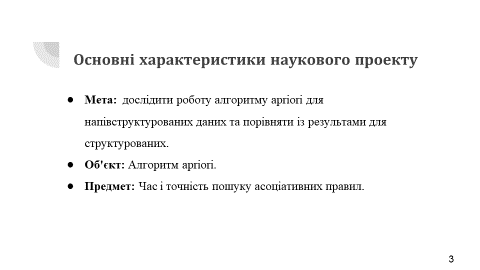
**Висновки**

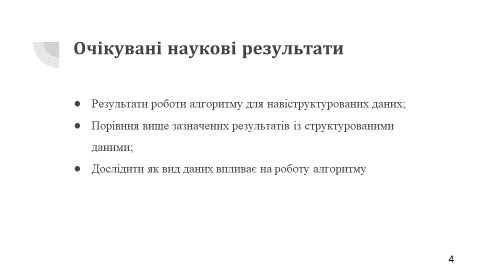
Після аналізу існуючих систем, що займаються пошуком закономірностей у даних, огляду існуючих підходів та алгоритмів для опрацювання великих структурованих і напівструктурованих даних було виявлено потребу у дослідженні роботи алгоритму для напіструктурованих даних, так як досліджень аналогів не було виявлено. Отримані результати буде порівняно із результатами для структурованих даних.

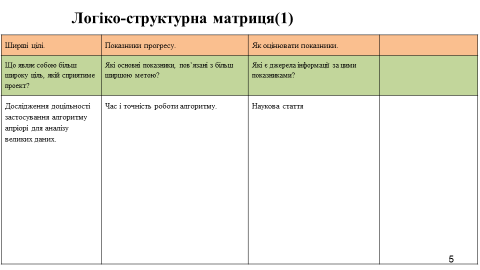
**Презентація проекту**

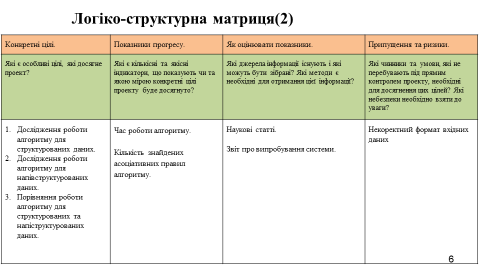




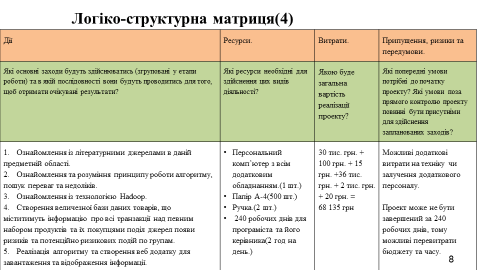




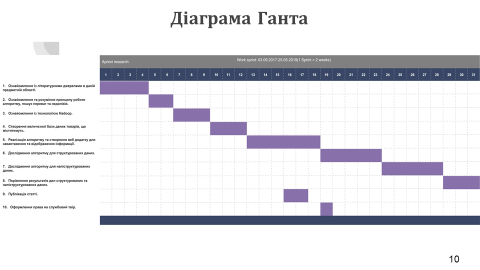


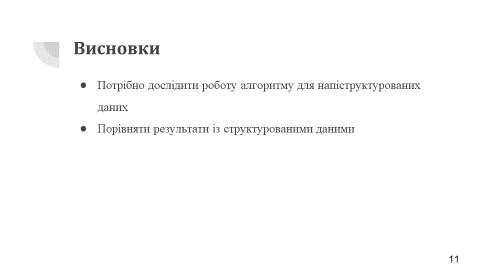












**Список використаної літератури**

1. Almasi, G.S. and A. Gottlieb (2009). Highly Parallel Computing. Benjamin// Cummings publishers, Redwood City, CA. – 235с.
2. Harris, Dereck Intel jettisons its Hadoop distro and puts millions behind Cloudera  (27 March 2014).
3. Hadoop [Електронний ресурс] // Вікіпедія : сайт. - Режим доступу https://ru.wikipedia.org/wiki/Hadoop
4. MapReduce [Електронний ресурс] // Вікіпедія : сайт. - Режим доступу <http://en.wikipedia.org/wiki/MapReduce#Map_function>
5. Mapreduce Appliance. [Електронний ресурс] // MapReduce : сайт. - Режим доступу <http://www.teradata.com/products/Aster_MapReduce_Appliance>
6. Big Data Applience. [Електронний ресурс] // Oracle Big Data: сайт. - Режим доступу <https://www.oracle.com/engineered-systems/big-data-appliance/index.html>
7. GreenPlum. [Електронний ресурс] //: сайт. - Режим доступу <http://www.emc.com/campaign/global/greenplumdca/index.htm>
8. Big Data. [Електронний ресурс] // Вікіпедія: сайт. - Режим доступу <https://en.wikipedia.org/wiki/Big_data>
9. Дюк В., Самойленко А. Data Mining: Учебный курс. – СПб: Питер, 2001. – 368 с.
10. А. Шахиди. Data Mining – добыча данных [Електронний ресурс] // Big Data: сайт. - Режим доступу // <http://www.basegroup.ru>
11. Karen Montgomery. Big Data Now: 2014 Edition. O'Reilly Media.- Junuary, 2015 – 165p.
12. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. John Wiley & Sons. 2014-12-19. 300p.
13. Apache Hadoop. [Електронний ресурс] // Big Data: сайт. - Режим доступу https://hadoop.apache.org/docs/r1.2.1/hdfs\_design.html
14. Tong Qiang, Zhou Yuanchun, Wu Kaichao, Yan Baoping. A quantitative association rules mining algorithm[J]. Computer engineering. 2007
15. Zhu Yixia, Yao Liwen, Huang Shuiyuan, Huang Longjun. A association rules mining algorithm based on matrix and trees[J]. Computer science. 2006, 33(7):196-198
16. Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules.In Proc.20th Int. Conf. Very Large Data Bases, 487-499.