**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**

**“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ**

**ІМЕНІ ІГОРЯ СІКОРСЬКОГО”**

Факультет прикладної математики

Кафедра програмного забезпечення комп’ютерних систем

**КУРСОВИЙ ПРОЕКТ**

**ТЕХНІЧНЕ ЗАВДАННЯ**

з дисципліни “Бази даних”

спеціальність 121 – Програмна інженерія

на тему: **Моніторингова система аналізу популярних відео сервісу YouTube**

**Студент**

**групи** КП-61 **Свинарчук М. В.**  \_\_\_\_\_\_\_\_\_\_\_

(підпис)

**Викладач**

**к.т.н, доцент кафедри**

**СПіСКС**  **Петрашенко А.В.** \_\_\_\_\_\_\_\_\_\_\_

(підпис)

Київ – 2019

ЗМІСТ

1. [НАЙМЕНУВАННЯ ТА ГАЛУЗЬ ЗАСТОСУВАННЯ РОЗРОБКИ](#_6w59hg2ehajf) 3
2. [ДАТА ПОЧАТКУ ТА ЗАКІНЧЕННЯ ПРОЕКТУ](#_ay3blt8wy9uk) 3
3. [МЕТА РОЗРОБКИ](#_kd15r7n7yhe2) 3
4. [ВИМОГИ ДО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ](#_qvij7kmjk6lw) 4
5. [ОБҐРУНТУВАННЯ ВИБОРУ СУБД](#_h33hpr59fijt) 7
6. [ВИМОГИ ДО ІНТЕРФЕЙСУ КОРИСТУВАЧА](#_f7qytop8y3ow) 7
7. [ВИБІР ЗАСОБІВ РОЗРОБКИ](#_en0ndoqw4svk) 8
8. [ЕТАПИ РОЗРОБКИ](#_b1awjuwc19sh) 9

# **НАЙМЕНУВАННЯ ТА ГАЛУЗЬ ЗАСТОСУВАННЯ РОЗРОБКИ**

**Найменування:** Моніторингова система аналізу популярних відео (трендів) сервісу [YouTube](https://www.youtube.com/) по кожній країні окремо, по вибірки країн та загалом по всьому світу.

**Галузь застосування:** статистика, аналіз соціальних тенденцій та прогнозування популярності певного виду відеоконтенту.

# **ДАТА ПОЧАТКУ ТА ЗАКІНЧЕННЯ ПРОЕКТУ**

**Дата початку проекту** – 25 березня 2018 року (дата видачі завдання для курсового проекту).

**Дата закінчення проекту** – 20 травня 2019 року (захист курсового проекту).

# **МЕТА РОЗРОБКИ**

**Метою розробки даного курсового проекту** є збір та фільтрація статистики популярних відео сервісу YouTube,аналіз результатів на її основі та формування бізнес-звітів з метою надання корисної інформації для рекламних агентств та інвесторів.

Для потенційних клієнтівдана моніторингова система повинна допомогти знайти категорії відеоконтенту, які мають відносно високий бізнес-потенціал. Наприклад, можливі наступні критерії оцінювання: кількість відео в кожній категорії, коефіцієнт участі (кількість коментарів / кількість переглядів), рейтинг популярності (кількість вподобань / кількість переглядів) і середній темп зростання в трендах. Маючи такого роду статистику, клієнту, будь то інвестор чи рекламне агенство, буде простіше знайти та визначитися яким категоріям відеоконтенту надати перевагу.

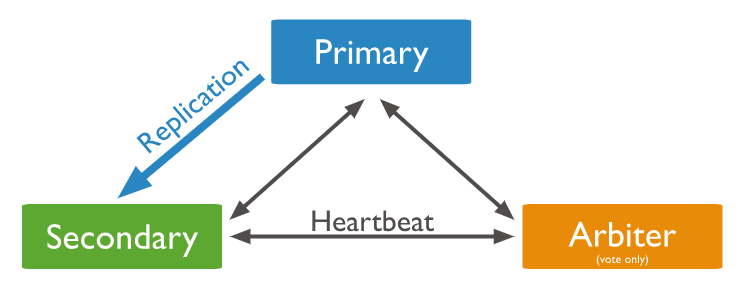
# **ВИМОГИ ДО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ**

**Підсистема попередньої обробки даних містить у собі:**

* **Засіб генерації даних:** розроблення утиліти для збору інформації про відео-тренди сервісу YouTube за допомогою YouTube Data API. Єдиним недоліком такого підходу є обмежена кількість статистичних даних, які можна отримати за певний період: тренди вміщують в себе 50-100 відео та оновлюються кожні 15 хвилин, але при цьому не обов’язково додаються нові відео, а лише змінюється порядок минулих. Тобто постає питання: як часто робити вибірку даних, щоб зменшити кількість попадань конкретного відео (може перебувати в трендах від 4 годин до декількох днів)? Пропонується створити утиліту, яка буде кожного дня в один і той же час збирати збирати дані (один раз на добу). Також є можливим використання готово датасету, зібраного за таким же принципом.
* **Засоби фільтрації та валідації даних:** розроблення додаткового функціоналу у вищезазначеній утиліті задля корегування отриманих даних та переходу до їх подальшої обробки та структуризації.

**База даних:** MongoDB

**Засоби реплікації даних:** нереляційна база даних MongoDB підтримує 2 форми реплікації: реплісети (Replica Sets) і Master-Slave. Пропонується використати більш сучасний підхід – **Replica Sets**.

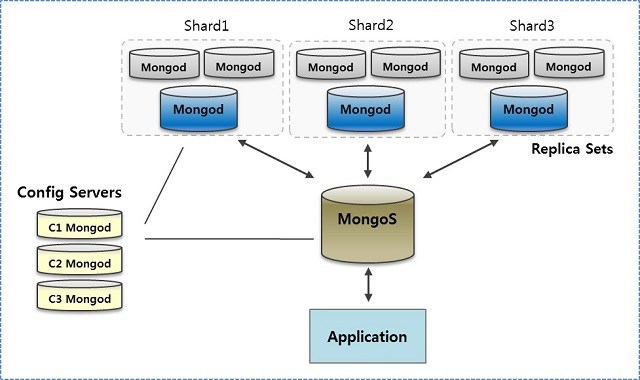


*Рис. 1.1 Приклад реплікації MongoDB за допомогою Replica Sets*

*(Арбітр не є обов’язковим членом схеми і необхідний лише у випадку існуванні парної кількості реплік)*

**Засоби масштабування:** MongoDB пропонує можливість горизонтального шардингу (sharding) – це поділ однієї таблиці на різні сервера. Поділ таблиці на частини робиться за таким принципом:

* На кількох серверах створюється одна і та ж таблиця (тільки структура, без даних).
* У додатку вибирається умова, за яким буде визначатися потрібне з'єднання (наприклад, парні на один сервер, а непарні – на інший).
* Перед кожним зверненням до таблиці відбувається вибір потрібного з'єднання.



*Рис. 1.2 Схема шардингу у MongoDB*

Сумісне використання засобів шардингу та реплікації MongoDB забезпечить відмовостійкість бази даних.

**Засоби аналізу даних:**

* *NumPy* – математична бібліотека з підтримкою багатовимірних масивів та високорівневих математичних функцій;
* *Pandas* – бібліотека для обробки та аналізу даних, використовується для первинної обробки даних.
* *NLTK* – Natural Language Toolkit, пакет бібліотек і програм для символьної і статистичної обробки природної мови.

**Задачі аналізу даних:**

* Зібрати дані за допомогою YouTube Data API та струкутувати їх з вже існуючим датасетом;
* Здійснити валідацію даних та прибрати зайву інформацію з датасету;
* Провести кореляцію по всіх даних;
* У місцях, де найбільша кореляція, здійснити більш детальний аналіз;
* Сформувати діаграми розсіювання, виокремити кластери та зробити висновки щодо отриманих результатів;
* Створити прогнози на основі отриманих результатів.

**Засоби резервування та відновлення даних:** передбачені при використанні Replica Sets у MongoDB. Також є можливість використання стандартних утиліт для збереження (**mongodump**) та відновлення **(mongorestore**).

# **ОБҐРУНТУВАННЯ ВИБОРУ СУБД**

Було обрано найпопулярнішу серед нереляційних СУБД – MongoDB. Це документо-орієнтована система управління базами даних із відкритим кодом, яка не потребує опису схеми таблиць. MongoDB займає нішу між швидкими і масштабованими системами, що оперують даними у форматі ключ/значення, і реляційними СУБД, функціональними і зручними у формуванні запитів.

Вибір нереляційної СУБД обгрунтовується наявністю великої кількості ненормалізованих даних, які необхідно обробити швидко. Вона забезпечує можливість отримання неприведених до норм даних та подальшу роботу з ними. Формування додаткових таблиць при появі додаткової інформації у екземплярі (реляційні СУБД) було б недоцільним.

# **ВИМОГИ ДО ІНТЕРФЕЙСУ КОРИСТУВАЧА**

* Мінімалістичний інтерфейс реалізований у консолі;
* Можливість вибору країни чи групи країн за якими буде проводитися аналіз;
* Можливість оперування базою даних – очистити базу даних, додати інформацію з датасета чи шляхом веб-скрапінгу;
* Можливість обрати необхідні критерії для проведення аналізу;
* Генерація таблиць, отриманих за результатами аналізу, графічних елементів – 2D графіки, хмари тегів, діаграми та інші зображення, сформовані у результаті досліджень;
* Можливість експорту отриманих результатів до файлів.

# **ВИБІР ЗАСОБІВ РОЗРОБКИ**

Мова програмування – Python 3.7.

СКБД – MongoDB.

Бібліотеки:

* *NumPy* – математична бібліотека з підтримкою багатовимірних масивів та високорівневих математичних функцій;
* *Pandas* – бібліотека для обробки та аналізу даних, використовується для первинної обробки даних.
* *NLTK* – Natural Language Toolkit, пакет бібліотек і програм для символьної і статистичної обробки природної мови.
* *WordCloud* – бібліотека для візуального подання списку категорій (або тегів, також званих мітками, ярликами, ключовими словами, тощо).
* *Matplotlib* – бібліотека для візуалізації даних у вигляді 2D і 3D графіків;
* *seaborn* – бібліотека візуалізації даних Python, яка базується на matplotlib. Забезпечує інтерфейс високого рівня для малювання привабливої та інформативної статистичної графіки.

# **ЕТАПИ РОЗРОБКИ**

|  |  |  |
| --- | --- | --- |
| **№** | **Назва етапу розробки** | **Термін виконання** |
| **1** | Затвердження теми курсової роботи. Опрацювання відповідної літератури. Розроблення та узгодження технічного завдання. | **01.04.2019** |
| **2** | Аналіз постановки задачі | **10.04.2019** |
| **3** | Розробка засобів генерації даних. | **15.04.2019** |
| **4** | Додавання засобів фільтрації та валідації даних. | **20.04.2019** |
| **5** | Реалізація зберігання, реплікації та масштабування інформації розробленої моніторингової системи. | **26.04.2019** |
| **6** | Додавання засобів аналізу даних (реалізацію алгоритмів буде запозичено у великих бібліотеках аналізу даних). | **03.05.19** |
| **7** | Додавання засобів резервування та відновлення даних (призначені для оперативного та пакетного збереження фрагментів та всієї бази даних з можливістю її відновлення з урахуванням необхідності підключення додаткового комп’ютера як елемента горизонтального масштабування). | **09.05.2019** |
| **8** | Тестування програми. | **15.05.2019** |
| **9** | Аналіз результатів. Підготовка матеріалів курсового проекту та оформлення пояснювальної записки. | **18.05.2019** |
| **10** | Захист курсової роботи. | **20.05.2019** |