

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
з дисципліни “Бази даних”
спеціальність 121 – Програмна інженерія
на тему: Моніторингова система онлайн магазину

Студент

групи КП-73

Труш А. О.

(ПІБ)

(підпис)

Викладач

к.т.н, доцент кафедри

СПіСКС

Петрашенко А.В.

(підпис)

Захищено з оцінкою _____

Київ – 2020

Анотація

Метою розробки даного курсового проекту є набуття практичних навичок розробки сучасного програмного забезпечення, що взаємодіє з постреляційними базами даних, а також здобуття навичок оформлення відповідного текстового, програмного та ілюстративного матеріалу у формі проектної документації. У результаті виконання курсового проекту було опановано навик розробляти програмне забезпечення для постреляційних баз даних, володіння основами використання СУБД, а також інструментальними засобами аналізу великих обсягів даних.

Темою даного курсового проекту є створення ПЗ для моніторингової системи онлайн магазину. У документі викладена актуальність та проблематика аналізу великого обсягу даних, аналіз використаного інструментарію (опис мови програмування, використаних бібліотек та СКБД), описана структура бази даних, опис розробленого програмного забезпечення (загальний, опис модулів та основних алгоритмів роботи), аналіз функціонування засобів масштабування, та опис результатів проведеного аналізу.

Результатами даного проекту стали діаграми та графіки, що зображають результати аналізу певних залежностей. З ними можна ознайомитися в додатку А.

Шардування та репліціювання було налаштовано за допомогою docker-compose, налаштування якого можна знати в додатку Б.

Зміст

Анотація	1
Зміст	2
Вступ	3
Аналіз інструментарію для виконання курсового проект	5
Аналіз СКБД	5
Обґрунтування вибору мови програмування	9
Обґрунтування вибору бібліотек і фреймворків	10
Структура бази даних	12
Аналіз функціонування засобів масштабування	13
Загальний алгоритм додавання нового кластеру шардингу, який складається з кількох серверів реплікації	15
Загальний алгоритм додавання нового вузла реплікації серверів конфігурації	16
Тестування масштабування	16
Тестування відмовостійкості шляхом зупинки декількох процесів із репліки сету одного із шардингу	17
Тестування відмовостійкості системи шляхом зупинки декількох процесів із репліки сети серверів конфігурації	19
Перевірка того, яку частину даних ми втратимо, в разі відмови однієї реплікації серверів шардингу	20
Опис результатів аналізу предметної галузі	21
Висновки	23
Література	25
Додаток А	27
Додаток Б	29

Вступ

Станом та поточний рік все більше і більше компаній приходять до того, що дані, якими вони оперують та збирають в процесі своєї роботи потребують агрегації та аналізу. Використання класичних методів аналізу, що базуються на звичайних алгоритмах математичної статистики та ймовірності не проносять очікуваної користі. З плином часу все більшу і більшу актуальність набирають програмні засоби аналізу даних, які інколи базуються на алгоритмах машинного навчання, що дозволяє зробити аналіз даних інтелектуальним (*en. data mining*) Цю науку, яка поєднує в собі програмування, мат. статистику, теорію ймовірності та машинне навчання називаються наукою про дані (*en. Data Science*).

Сучасний Data science спеціаліст (дослідник даних) має оперувати великими обсягами даних (від одного гігабайта до декількох терабайт) та вміти виокремити з них приховані залежності, на основі яких зробити прогноз про те, як будуть надалі відбуватися ті чи інші явища. Дослідники даних використовують свої дані та аналітичні здібності для пошуку та інтерпретації великих джерел даних; керують великими обсягами даних безвідносно до апаратного та програмного забезпечення і обмежень пропускної здатності; об'єднують джерела даних; забезпечують цілісність наборів даних; створюють візуалізації для кращого розуміння даних; з використанням даних будують математичні моделі; надають тлумачення даних та висновки.

Класичним набором інструментів, яким користується Data science спеціаліст є мова програмування Python 3 і набір математичних бібліотек до неї (pandas, scikit-learn, numpy, matplotlib) нереляційні бази даних (на кшталт MongoDB, DynamoDB, Cassandra) та закони математичного аналізу, теорії ймовірності та мат. статистики).

Метою створення даного проекту був аналіз даних про залежності між покупками покупців та факторами що на них впливають. Було виявлено певні закономірності між дальністю від покупця та витраченою сумою.

Дані були згенеровані вручну з різними розподілами, серед яких χ^2 квадрат та дискретний.

Аналіз інструментарію для виконання курсового проект

Аналіз СКБД

В процесі виконання цього курсового проекту перед нами встала потреба кешувати та зберігати дані про вакансії між запусками аналізатора. Кожного разу читати дані із CSV - файлів є дуже дорогою операцією, тож було прийняте рішення використати СКБД. В якості СКБД були розглянуті варіанти: PostgreSQL, MongoDB. З порівняльною характеристикою цих СКБД можна ознайомитися в таблиці 1.

таблиця 1. Порівняльна характеристика СКБД

Критерій порівняння	Назва СКБД		
	MongoDB	PostgreSQL	CassandraDB
Має відкритий вихідний код	так	так	так
Схема даних	динамічна	статична і динамічна	статична і динамічна
Підтримка ієрархічних даних	так	так (з 2012)	ні
Реляційні дані	ні	так	так
Транзакції	ні	так	так
Атомарності операцій	всередині документа	по всій БД	всередині партії
Мова запитів	JSON	SQL	CQL

Найлегший спосіб масштабування	горизонтальний	вертикальний	горизонтальний
Підтримка шардингів	так	так (важка конфігурація)	так (може зберігати партіції на різних машинах)
Приклад використання	Великі дані (мільярди записів) з великою кількістю паралельних оновлень, де цілісність і узгодженість даних не потрібно.	Транзакційні і операційні програми, вигода яких в нормалізованому формі, об'єднаннях, обмеження даних і підтримки транзакцій.	Багато запитів на запис/читання у одиницю часу, до даних можна задіяти партіціювання за ключем, дані мають лише первинні індекси
Наявність бібліотек для мови програмування Python 3	так	так	так
Підтримка реплікації	так, автоматичне переобрання головного процесу	За принципом master-slave	так, через партіціювання

Засіб збереження та відновлення даних	mongodump	pg_dump	не має окремого доданка, виконується засобами CQL
Форма збереження даних	документи BSON	таблиця	таблиця

За результатами порівнянні цих СКБД було прийнято рішення зупинитися на NoSQL рішеннях. Оскільки вони чудово поєднує в собі переваги неструктурованих баз даних та простоту використання горизонтального масштабування. Крім цього класичним прикладом використання NoSQL СКБД є системи збору та аналізу даних, до яких можна застосувати індексування за первинними та вторинними ключами.

Ця база даних є об'єктно орієнтованою та дозволяє зберігати великі масиви неструктурованих даних. На відміну від SQL баз даних ми можемо зберігати дані у “сирому” об'єктному вигляді, який використовується програмою та є більш близьким за структурою до моделі даних, яку буде використовувати ПЗ написане з використанням мови програмування Python. Це пришвидшить збір, збереження та отримання даних програмним забезпеченням. Оскільки MongoDB є представником NoSQL баз даних, вона не потребує жорсткої схеми даних, що дозволяє пришвидшити процес розробки та зробити його більш гнучким. Окрім цього дана СУБД підтримує горизонтальне масштабування за допомогою шардингу з метою зменшення навантаження на кожен окремий вузол шляхом розподілення навантаження між ними всіма.

Нижче наведено перелік основних переваг:

- Підтримка ієрархічних даних
- Динамічна схема

- Швидкість запису у колекцію
- Швидкість читання із колекції
- Простота масштабування та відновлення даних

Мовою програмування для ПЗ було обрано Python 3.8. Оскільки це строго динамічно типізована мова програмування. Це дозволяє пришвидшити швидкість розробки ПЗ за рахунок уникнення постійного оголошення типів даних, якими оперує програма, та водночас не дає зробити критичних помилок, прихованих за автоматичним приведенням типів, на відміну від мови JavaScript.

Ключовою особливістю Python є широкий інструментарій засобів розробки систем збору та аналізу даних. Де-факто ця мова є стандартом у світі математичних розрахунків та обробки даних у реальному часі. Тож під час виконання курсового проекту було відносно легко знайти необхідну документацію та приклади роботи із цією мовою.

Використані бібліотеки:

- **pandas** — бібліотека для обробки та аналізу даних. Pandas є open-source бібліотекою, ліцензована BSD (Berkeley Source Distribution), забезпечує швидкі, гнучкі та експресивні структури даних, призначені для того, щоб працювати з «реляційними» або «помітними» даними як простими, так і інтуїтивно зрозумілими. Він прагне бути основним будівельним блоком високого рівня для здійснення практичного, реального аналізу даних у Python. Можливості бібліотеки: інструменти для обміну даними між структурами в пам'яті і файлами різних форматів, засоби поєднання даних і способи обробки відсутньої інформації, Переформатування наборів даних, в тому числі створення зведених таблиць, зріз даних за значеннями індексу, розширені можливості індексування, вибірка з великих наборів даних, вставка і видалення стовпців даних, можливості угруповання дозволяють виконувати трьохетапні операції типу «поділ, зміна, об'єднання», злиття і об'єднання наборів даних;
- **scikit-learn** — безкоштовна бібліотека алгоритмів машинного навчання програмного забезпечення для мови програмування Python, має прості та ефективні інструменти для аналізу даних та аналізу даних. Як і pandas має комерційне використання - ліцензія BSD. Вона має різні алгоритми класифікації, регресії та кластеризації, включаючи векторні машини підтримки, випадкові ліси, градієнтні підсилювачі, k-засоби та DBSCAN, і призначені для взаємодії з чисельними та науковими бібліотеками Python NumPy SciPy та matplotlib;

- `matplotlib` — це бібліотека Python 2D, яка представляє числові дані у різноманітних форматах та інтерактивних середовищах на різних платформах. Також ця бібліотека - це математичне розширення NumPy. Він надає об'єктно-орієнтований API для вбудови ділянок у додатки, що використовують набір інструментів для загального графічного інтерфейсу, таких як Tkinter, wxPython, Qt або GTK+. Matplotlib може використовуватися в скриптах Python, оболонках Python та IPython, серверах веб-додатків та чотирьох графічних наборах інструментів для користувацького інтерфейсу. Бібліотека `sciPy` використовує `matplotlib`.
- `numpy` — математична бібліотека мови Python, що додає підтримку великих багатовимірних масивів і матриць, разом з великою бібліотекою високорівневих математичних функцій для операцій з цими масивами. А також інструменти для інтеграції C/C++ і Fortran коду, корисну лінійну алгебру, перетворення Фур'є і можливості випадкових чисел. Крім очевидних наукових застосувань, NumPy також може бути використаний як ефективний багатовимірний контейнер загальних даних. Можна визначити довільні типи даних. Це дає змогу без проблем і швидкої інтеграції NumPy з різноманітними базами даних. NumPy має ліцензію за ліцензією BSD, що дозволяє повторно використовувати кілька обмежень.

Структура бази даних

База даних складається з двох колекцій: Users і Checks. Users використовується для збереження даних про покупців. Checks використовується для збереження детальних відомостей про здійснені покупки.

таблиця 2. Опис властивостей документа у колекції countries

Назва властивості	Тип	Опис
_id	ObjectId	Ідентифікатор запису
first_name	str	Ім`я
second_name	str	Прізвище
latitude	float	Широта
longitude	float	Довгота

таблиця 3. Опис властивостей документа у колекції Checks

Назва властивості	Тип	Опис
_id	ObjectId	Ідентифікатор запису
owner	ObjectId	Ідентифікатор особи в таблиці Users
sum	int	Витрачена сума
num_of_purchases	Int	Кількість придбаних товарів
time	string	Час
day_of_week	string	День тижня

Аналіз функціонування засобів масштабування

В якості засобів масштабування було обрано реплікацію та шардинг.

Реплікація - це процес синхронізації даних на декількох серверах. Даний механізм зменшує витрати ресурсів і збільшує доступність даних, копії яких зберігаються на різних серверах. Реплікація захищає базу даних (далі - БД) від втрати єдиної сервера і дозволяє зберегти дані в разі технічної несправності на одному з серверів. У MongoDB реплікація досягається шляхом використання набору копій (replica set). Це група примірників mongod, який зберігають однакові набори даних. У копії один вузол - це ключовий вузол, який отримує всі операції запису. Всі інші вузли - вторинні, приймають операції з першого, таким чином, зберігаючи такі ж записи, як і первинний вузол. Набір копій може мати тільки один первинний вузол.

Шардінг - це процес зберігання документів на декількох серверах і це спосіб, яким MongoDB справляється з великими даними. З ростом кількості даних, один сервер не може зберігати всі дані, ні записувати їх, ні давати до них доступ. Шардінг вирішує проблему шляхом горизонтального масштабування. Завдяки даному механізму ми можемо підключати додаткові сервери для зберігання, записи і читання даних.

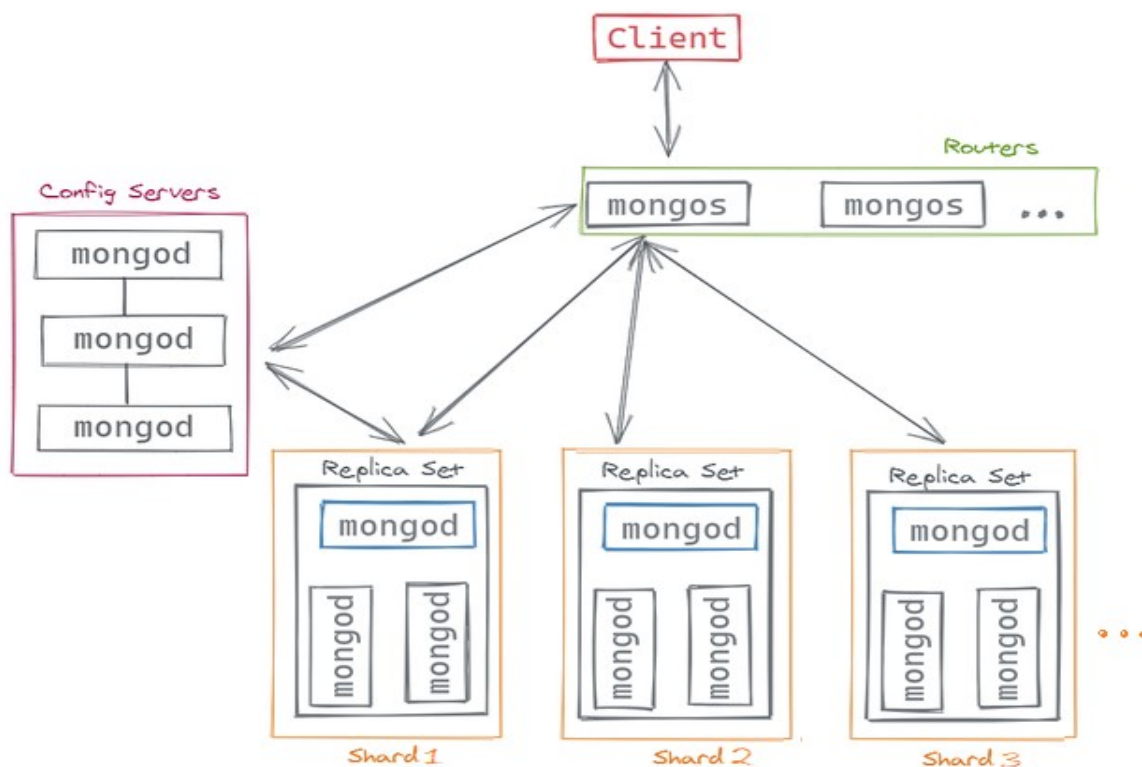


рис 1.Схема використаної моделі шардингу MongoDB у даному ПЗ

В якості ключа для розбивання даних на партії було використано поле *date* оскільки воно зберігає дані із більш-менш лінійним розподілом по дням. Тож в якості критерія до партиціювання було використано хеш - індекс за властивістю *date*. Із результатами розбиття даних на партії можна ознайомитися на рис 2.

```

mongos> db.stats()
{
  "raw" : {
    "rs-shard-02/shard02-a:27017,shard02-b:27017" : {
      "db" : "Course_Work",
      "collections" : 1,
      "views" : 0,
      "objects" : 23291,
      "avgObjSize" : 117.57739040831223,
      "dataSize" : 2738495,
      "storageSize" : 872448,
      "numExtents" : 0,
      "indexes" : 2,
      "indexSize" : 724992,
      "fsUsedSize" : 76582322176,
      "fsTotalSize" : 94417633280,
      "ok" : 1
    },
    "rs-shard-01/shard01-a:27017,shard01-b:27017" : {
      "db" : "Course_Work",
      "collections" : 1,
      "views" : 0,
      "objects" : 24092,
      "avgObjSize" : 117.550597708783,
      "dataSize" : 2832029,
      "storageSize" : 897024,
      "numExtents" : 0,
      "indexes" : 2,
      "indexSize" : 753664,
      "fsUsedSize" : 76582322176,
      "fsTotalSize" : 94417633280,
      "ok" : 1
    },
    "rs-shard-03/shard03-a:27017,shard03-b:27017" : {
      "db" : "Course_Work",
      "collections" : 2,
      "views" : 0,
      "objects" : 52617,
      "avgObjSize" : 108.74871239333295,
      "dataSize" : 5722031,
      "storageSize" : 2281472,
      "numExtents" : 0,
      "indexes" : 3,
      "indexSize" : 1015808,
      "fsUsedSize" : 76582322176,
      "fsTotalSize" : 94417633280,
      "ok" : 1
    }
  },
  "objects" : 100000,
  "avgObjSize" : 117.57739040831223,
  "dataSize" : 2738495,
  "storageSize" : 872448,
  "numExtents" : 0,
  "indexes" : 2,
  "indexSize" : 724992,
  "fsUsedSize" : 76582322176,
  "fsTotalSize" : 94417633280,
  "ok" : 1
}

```

рис 2. Розподіл даних між шардами в рамках однієї колекції

Загальний алгоритм додавання нового кластеру шардингу, який складається з кількох серверів реплікації

1. Запустити на виконання один або декілька процесів mongod, які стануть вузлами шардингу за допомоги команди у таблиці 3. Також необхідно запустити я мінімум один конфігураційний сервер.

таблиця 3. Скрипт запуску необхідних вузлів шардингу


```
mongod --dbpath data/shard1 --port 27000 --fork --syslog
mongod --dbpath data/shard2 --port 27001 --fork --syslog
mongod --configsvr --dbpath data/config --port 27002 --fork --syslog
mongos --configdb localhost:27002 --port 27100 --fork --syslog
```

2. Дочекатися їх ініціалізації, під'єднатися до процесу mongos сервера-роутера та виконати команду наведену у таблиці 4. Ця команда спочатку прив'яже два сервера-шарди, створить індекс date, по якому дані будуть розподілятися між шардами, а після цього розшардує колекцію по цьому індексу.

таблиця 4. Скрипт створення нових вузлів шардингу

```
sh.addShard("127.0.0.1:27000")
sh.addShard("127.0.0.1:27001")
use Course_Work
db.Checks.ensureIndex({sum: «hashed»})
use admin
db.runCommand({shardCollection: "Course_Work.Checks", key {sum: «hashed»}})
```

Загальний алгоритм додавання нового вузла реплікації серверів конфігурації

Запустити на виконання один або декілька процесів mongod, які стануть вузлами реплікації.

1. Дочекатися їх ініціалізації, під'єднатися до процесу mongod одного з них та виконати команду наведену у таблиці 6.

таблиця 5. Скрипт додавання нового серверу конфігурації до реплікації

```
rs.add( { host: "<hostnameNew>:<portNew>", priority: 0, votes: 0 } )
```

Тестування масштабування

В процесі перевірки здатності системи масштабуватися, було проведено наступні тести:

- 1) Тестування відмовостійкості шляхом зупинки декількох процесів із репліки сету одного із шардингу
- 2) Тестування відмовостійкості системи шляхом зупинки декількох процесів із репліки сети серверів конфігурації
- 3) Перевірка того, яку частину даних ми втратимо, в разі відмови однієї реплікації серверів шардингу

1. Тестування відмовостійкості шляхом зупинки декількох процесів із репліки сету одного із шардингу

Нами було зупинено процес *shard1a* в результаті чого, ми очікували, що система перенаправить усі запити, що йшли до нього до його реплікації *shard1b*, та обере його головним, в разі якщо він не був таким. Вивід процесу *mongos* серверу маршрутизації наведено на рис. 3.

```
2019-06-02T19:46:14.097+0000 I NETWORK [conn30] end connection 172.27.0.5:48430 (14 connections now open)
2019-06-02T19:46:14.100+0000 I NETWORK [conn24] end connection 172.27.0.5:48412 (13 connections now open)
2019-06-02T19:46:14.101+0000 I NETWORK [conn26] end connection 172.27.0.5:48418 (12 connections now open)
2019-06-02T19:46:14.418+0000 I NETWORK [conn22] end connection 172.27.0.5:48364 (11 connections now open)
2019-06-02T19:46:14.551+0000 I CONNPPOOL [Replication] dropping unhealthy pooled connection to shard01a:27018
2019-06-02T19:46:14.551+0000 I CONNPPOOL [Replication] after drop, pool was empty, going to spawn some connections
2019-06-02T19:46:14.551+0000 I ASIO [Replication] Connecting to shard01a:27018
2019-06-02T19:46:17.564+0000 I NETWORK [conn29] end connection 172.27.0.5:48428 (10 connections now open)
2019-06-02T19:46:23.163+0000 I REPL [replexec-3] Member shard01a:27018 is now in state RS_DOWN
2019-06-02T19:46:23.163+0000 I REPL [replexec-3] can't see a majority of the set, relinquishing primary
2019-06-02T19:46:23.164+0000 I REPL [replexec-3] Stepping down from primary in response to heartbeat
2019-06-02T19:46:23.165+0000 I REPL [replexec-3] transition to SECONDARY from PRIMARY
2019-06-02T19:46:23.165+0000 I NETWORK [replexec-3] Skip closing connection for connection # 28
2019-06-02T19:46:23.166+0000 I NETWORK [conn13] end connection 172.27.0.3:52250 (9 connections now open)
2019-06-02T19:46:23.166+0000 I NETWORK [conn16] end connection 172.27.0.8:36900 (6 connections now open)
2019-06-02T19:46:23.167+0000 I NETWORK [conn20] end connection 172.27.0.7:54636 (5 connections now open)
2019-06-02T19:46:23.166+0000 I NETWORK [conn14] end connection 172.27.0.6:34292 (8 connections now open)
2019-06-02T19:46:23.166+0000 I NETWORK [conn21] end connection 172.27.0.10:48316 (7 connections now open)
2019-06-02T19:46:23.167+0000 I NETWORK [conn17] end connection 172.27.0.2:34976 (3 connections now open)
2019-06-02T19:46:23.168+0000 I NETWORK [conn18] end connection 172.27.0.9:50076 (2 connections now open)
2019-06-02T19:46:23.167+0000 I NETWORK [conn19] end connection 172.27.0.4:43840 (4 connections now open)
2019-06-02T19:46:23.168+0000 I NETWORK [conn15] end connection 172.27.0.11:33882 (1 connection now open)
2019-06-02T19:46:24.552+0000 I REPL_HB [replexec-4] Error in heartbeat (requestId: 404) to shard01a:27018, response status:
NetworkInterfaceExceededTimeLimit: Couldn't get a connection within the time limit
2019-06-02T19:46:25.612+0000 I NETWORK [listener] connection accepted from 172.27.0.7:54858 #32 (2 connections now open)
2019-06-02T19:46:25.614+0000 I NETWORK [conn32] received client metadata from 172.27.0.7:54858 conn32: { driver: { name: "MongoDB Internal Client", version:
"4.0.9" }, os: { type: "Linux", name: "Ubuntu", architecture: "x86_64", version: "16.04" } }
2019-06-02T19:46:26.656+0000 I NETWORK [listener] connection accepted from 172.27.0.10:48522 #33 (3 connections now open)
2019-06-02T19:46:26.657+0000 I NETWORK [conn33] received client metadata from 172.27.0.10:48522 conn33: { driver: { name: "MongoDB Internal Client", version:
"4.0.9" }, os: { type: "Linux", name: "Ubuntu", architecture: "x86_64", version: "16.04" } }
2019-06-02T19:46:34.254+0000 I REPL [replexec-4] Not starting an election, since we are not electable due to: Not standing for election because I cannot
see a majority (mask 0x1)
2019-06-02T19:46:34.552+0000 I ASIO [Replication] Connecting to shard01a:27018
```

рис 3. Вивід процесу *shard1b* після того, вимкнення *shard1a*

Як видно із виводу, сервер *shard1b* не зміг отримати відповідь від своєї реплікації *shard1a* на запит перевірки “серцебиття” (*en. heartbeat*). В результаті чого він почав надсилати цей запит повторно кожну секунду та взяв на себе відповідальність бути лідером у реплікації із сервером *shard1a*. При цьому усі дані, які зберігалися у цьому кластері шардингу ще доступні на запис та читання. Після відновлення процесу *shard1a*, нами отримано наступний вивід процесу *shard1b*, наведений на рис 4. З нього стає зрозуміло що сервер *shard1b* зміг отримати відповідь від *shard1a* та запустив процес синхронізації та обрання нового лідера реплікації.

[illegible]

рис 4. Вивід процесу shard1b після відновлення процесу shard1a

2. Тестування відмовостійкості системи шляхом зупинки декількох процесів із репліки сети серверів конфігурації

Сервери конфігурація, які знаходять у реплікації поводять себе так само, як і при тестуванні реплікації кластерів шардів. Після невдачі отримати відповідь на запит перевірки серцебиття (рис 5) від серверу конфігурації (*config1*), кожен з процесів серверу конфігурації, що ще онлайн робить запит до вимкненого сервера кілька раз на секунду. Крім цього відбувається обрання нового лідера реплікації. Після поновлення *config1* відбувається обрання нового лідера реплікації та система переходить до звичайного режиму роботи (рис 6)

```
2019-06-02T19:59:58.789+0000 I NETWORK [conn85] end connection 172.27.0.3:34694 (21 connections now open)
2019-06-02T19:59:58.799+0000 I NETWORK [conn87] end connection 172.27.0.3:34700 (20 connections now open)
2019-06-02T20:00:20.241+0000 I NETWORK [conn13] end connection 172.27.0.7:51964 (19 connections now open)
2019-06-02T20:00:20.244+0000 I NETWORK [conn8] end connection 172.27.0.7:51932 (18 connections now open)
2019-06-02T20:00:21.963+0000 I CONNPPOOL [Replication] dropping unhealthy pooled connection to config03:27017
2019-06-02T20:00:21.963+0000 I CONNPPOOL [Replication] after drop, pool was empty, going to spawn some connections
2019-06-02T20:00:21.971+0000 I ASIO [Replication] Connecting to config03:27017
2019-06-02T20:00:21.971+0000 I ASIO [Replication] Failed to connect to config03:27017 - HostUnreachable: Error connecting to config03:27017 :: caused by :: Could not find address for config03:27017: SocketException: Host not found (authoritative)
2019-06-02T20:00:21.971+0000 I CONNPPOOL [Replication] Dropping all pooled connections to config03:27017 due to HostUnreachable: Error connecting to config03:27017 :: caused by :: Could not find address for config03:27017: SocketException: Host not found (authoritative)
2019-06-02T20:00:21.972+0000 I REPL_HB [replexec-42] Error in heartbeat (requestId: 3811) to config03:27017, response status: HostUnreachable: Error connecting to config03:27017 :: caused by :: Could not find address for config03:27017: SocketException: Host not found (authoritative)
2019-06-02T20:00:21.972+0000 I ASIO [Replication] Connecting to config03:27017
2019-06-02T20:00:21.977+0000 I ASIO [Replication] Failed to connect to config03:27017 - HostUnreachable: Error connecting to config03:27017 :: caused by :: Could not find address for config03:27017: SocketException: Host not found (authoritative)
2019-06-02T20:00:21.977+0000 I CONNPPOOL [Replication] Dropping all pooled connections to config03:27017 due to HostUnreachable: Error connecting to config03:27017 :: caused by :: Could not find address for config03:27017: SocketException: Host not found (authoritative)
2019-06-02T20:00:21.978+0000 I REPL_HB [replexec-39] Error in heartbeat (requestId: 3813) to config03:27017, response status: HostUnreachable: Error connecting to config03:27017 :: caused by :: Could not find address for config03:27017: SocketException: Host not found (authoritative)
2019-06-02T20:00:21.978+0000 I ASIO [Replication] Connecting to config03:27017
```

рис 5. Спроби отримати відповідь від вимкненого сервера конфігурації config1

```
2019-06-02T20:08:25.508+0000 I ASIO [Replication] Connecting to config03:27017
2019-06-02T20:08:25.515+0000 I ASIO [Replication] Failed to connect to config03:27017 - HostUnreachable: Error connecting to config03:27017 :: caused by :: Could not find address for config03:27017: SocketException: Host not found (authoritative)
2019-06-02T20:08:25.515+0000 I CONNPPOOL [Replication] Dropping all pooled connections to config03:27017 due to HostUnreachable: Error connecting to config03:27017 :: caused by :: Could not find address for config03:27017: SocketException: Host not found (authoritative)
2019-06-02T20:08:25.516+0000 I REPL_HB [replexec-29] Error in heartbeat (requestId: 17392) to config03:27017, response status: HostUnreachable: Error connecting to config03:27017 :: caused by :: Could not find address for config03:27017: SocketException: Host not found (authoritative)
2019-06-02T20:08:27.414+0000 I NETWORK [listener] connection accepted from 172.27.0.7:38784 #83 (18 connections now open)
2019-06-02T20:08:27.415+0000 I NETWORK [conn83] end connection 172.27.0.7:38784 (17 connections now open)
2019-06-02T20:08:27.422+0000 I NETWORK [listener] connection accepted from 172.27.0.7:38788 #84 (18 connections now open)
2019-06-02T20:08:27.422+0000 I NETWORK [conn84] received client metadata from 172.27.0.7:38788 conn84: { driver: { name: "NetworkInterfaceTL", version: "4.0.9" }, os: { type: "Linux", name: "Ubuntu", architecture: "x86_64", version: "16.04" } }
2019-06-02T20:08:27.516+0000 I ASIO [Replication] Connecting to config03:27017
2019-06-02T20:08:27.520+0000 I REPL [replexec-29] Member config03:27017 is now in state SECONDARY
2019-06-02T20:08:28.428+0000 I NETWORK [listener] connection accepted from 172.27.0.7:38792 #86 (19 connections now open)
2019-06-02T20:08:28.428+0000 I NETWORK [conn86] received client metadata from 172.27.0.7:38792 conn86: { driver: { name: "NetworkInterfaceTL", version: "4.0.9" }, os: { type: "Linux", name: "Ubuntu", architecture: "x86_64", version: "16.04" } }
2019-06-02T20:08:28.442+0000 I NETWORK [listener] connection accepted from 172.27.0.7:38794 #87 (20 connections now open)
2019-06-02T20:08:28.443+0000 I NETWORK [conn87] received client metadata from 172.27.0.7:38794 conn87: { driver: { name: "NetworkInterfaceTL", version: "4.0.9" }, os: { type: "Linux", name: "Ubuntu", architecture: "x86_64", version: "16.04" } }
```

рис 6. Створення нового пулу з'єднань із config1 після його поновлення та обрання нового лідера репліки

Перевірка того, яку частину даних ми втратимо, в разі відмови однієї реплікації серверів шардингу

Нами було зупинено процеси *shard1a* та *shard2a*. В результаті чого частина даних, яка зберігалася на цих серверах, повинна бути втрачена, а після поновлення хоча б одного процесу *shard1X* вони знову повинні бути доступними. Внаслідок розірвання зв'язку між серверами конфігурації та кластерами шардингу, ми не могли запустити запит на виконання до бд, що стосувався даних у колекції, яка була розбита на партиції. Після поновлення процесів *shard1a* та *shard2a* цілісність даних було відновлено і ми знову змогли виконати запит до колекції.

Опис результатів аналізу предметної галузі

В результаті виконання курсового проекту було проаналізовано перебіг епідемії в різних країнах. Було отримано наступні дані:

1. Згідно з рисунком А1

Найменше люди купують в понеділок, найбільше по вихідним

2. Згідно із рисунком А2

Найчастіше люди витрачають близько 12 тисяч грн

3. Згідно із рисунком А3

Найімовірніше, що покупець купить лише 1 товар

4. Згідно із рисунком А4

Дальність до магазину не впливає на суму покупки

5. Згідно із рисунком А5

Дальність до магазину не впливає на кількість товарів в чекові

6. Згідно із рисунком А6

Знайдено криву залежності загальної кількості усіх випадків

Висновки

В процесі виконання даного курсового проекту було отримано практичні навички обробки великих масивів даних за допомогою мови програмування Python 3 та СКБД MongoDB.

Було проаналізовано сучасні методи та інструменти для роботи із великими даними, знайдено гарно працюючу комбінацію із мови програмування Python 3 та бібліотек до нього: Pandas, Scikit-learn, Matplotlib, Numpy. Було складено порівняльну таблицю із трьох баз даних, які найчастіше використовують для роботи із часовими рядами. На основі цих даних було обрано в якості СКБД NoSQL рішення MongoDB.

Для забезпечення горизонтального масштабування було використано засоби MongoDB, такі як: репліка сет та шардинг. Практичним шляхом було знайдено спосіб партиціювання даних часових рядів на основі хеш-функції від мітки часу. Це дозволило розподілити дані, які зберігаються в рамках однієї колекції між багатьма серверами кластерів шардингу. За допомогою реплікації було досягнуто відмовостійкості системи, в результаті чого, під час тестування ми впевнилися, що система продовжує функціонування після виходу із ладу кількох реплік.

Дані аналізу приведені у Додатку А. Текстовий опис надано в відповідному розділі цього курсового проекту.

В ході виконання даного курсового проекту було досягнуто поставленої мети: було набуто практичних навичок розробки сучасного програмного забезпечення, що взаємодіє з NoSQL базами даних, а також були здобуті навички оформлення відповідного текстового, програмного та ілюстративного матеріалу у формі проектної документації. У результаті виконання курсового проекту я навчився писати програмне забезпечення для NoSQL баз даних, володіти основами використання СУБД, а також

інструментальними засобами аналізу великих обсягів даних, а саме відкритими бібліотеками мови Python.

Література

1. Hashed sharding [Електронний ресурс]. Режим доступу до ресурсу: <https://docs.mongodb.com/manual/core/hashed-sharding/>
2. Aggregation [Електронний ресурс]. Режим доступу до ресурсу: <https://docs.mongodb.com/manual/aggregation/>
3. 5 причин, почему машинное обучение станет технологией 2018 [Електронний ресурс]. – 2018. – Режим доступу до ресурсу: <https://blogs.oracle.com/russia/machine-learning-2018;>
4. Сложности накопления данных для интеллектуального анализа [Електронний ресурс]. – 2012. – Режим доступу до ресурсу: [https://habr.com/post/154787/;](https://habr.com/post/154787/)
5. Почему Python так хорош в научных вычислениях [Електронний ресурс]. – 2018. – Режим доступу до ресурсу: [https://habr.com/post/349482/;](https://habr.com/post/349482/)
6. Python [Електронный ресурс]. – Режим доступу до ресурсу: <https://uk.wikipedia.org/wiki/Python;>
7. PostgreSQL лучше чем MongoDB[Електронний ресурс]. – 2015. – Режим доступу до ресурсу: [https://habr.com/post/272735/;](https://habr.com/post/272735/)
8. 27 шпаргалок п машинному обучению [Електронний ресурс]. – 2017. – Режим доступу до ресурсу: [https://proglab.io/p/ds-cheatsheets/;](https://proglab.io/p/ds-cheatsheets/)
9. Pandas [Електронний ресурс]. – Режим доступу до ресурсу: <https://ru.wikipedia.org/wiki/Pandas;>
10. Scikit-learn [Електронний ресурс]. – Режим доступу до ресурсу: <https://en.wikipedia.org/wiki/Scikit-learn;>
11. Matplotlib [Електронний ресурс]. – Режим доступу до ресурсу: <https://en.wikipedia.org/wiki/Matplotlib;>
12. NumPy [Електронний ресурс]. – Режим доступу до ресурсу: <https://en.wikipedia.org/wiki/NumPy;>

13. 15. Горизонтальное масштабирование: когда и как?
[Электронный ресурс]. – Режим доступа до ресурсу:
<https://habr.com/company/oleg-bunin/blog/319526>

Додаток А

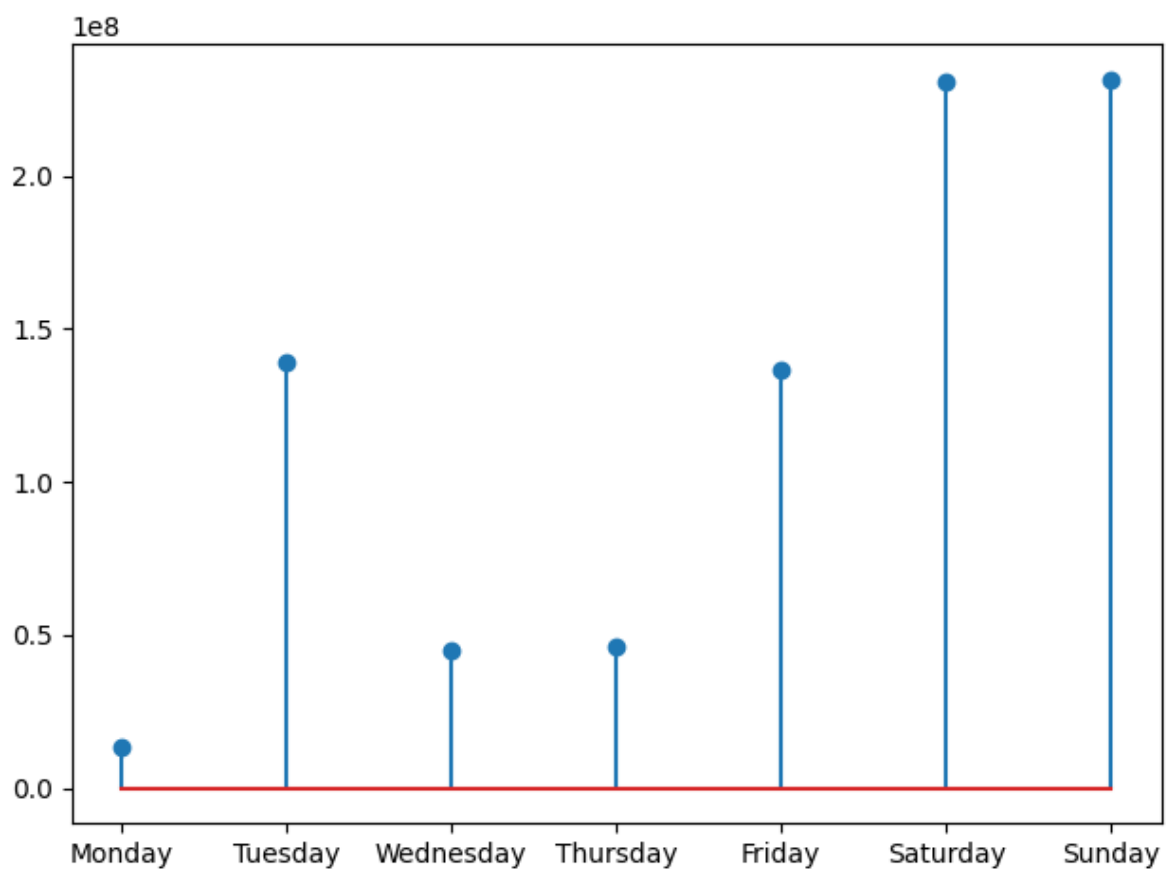


Рис 1. Залежність витраченої суми від дня

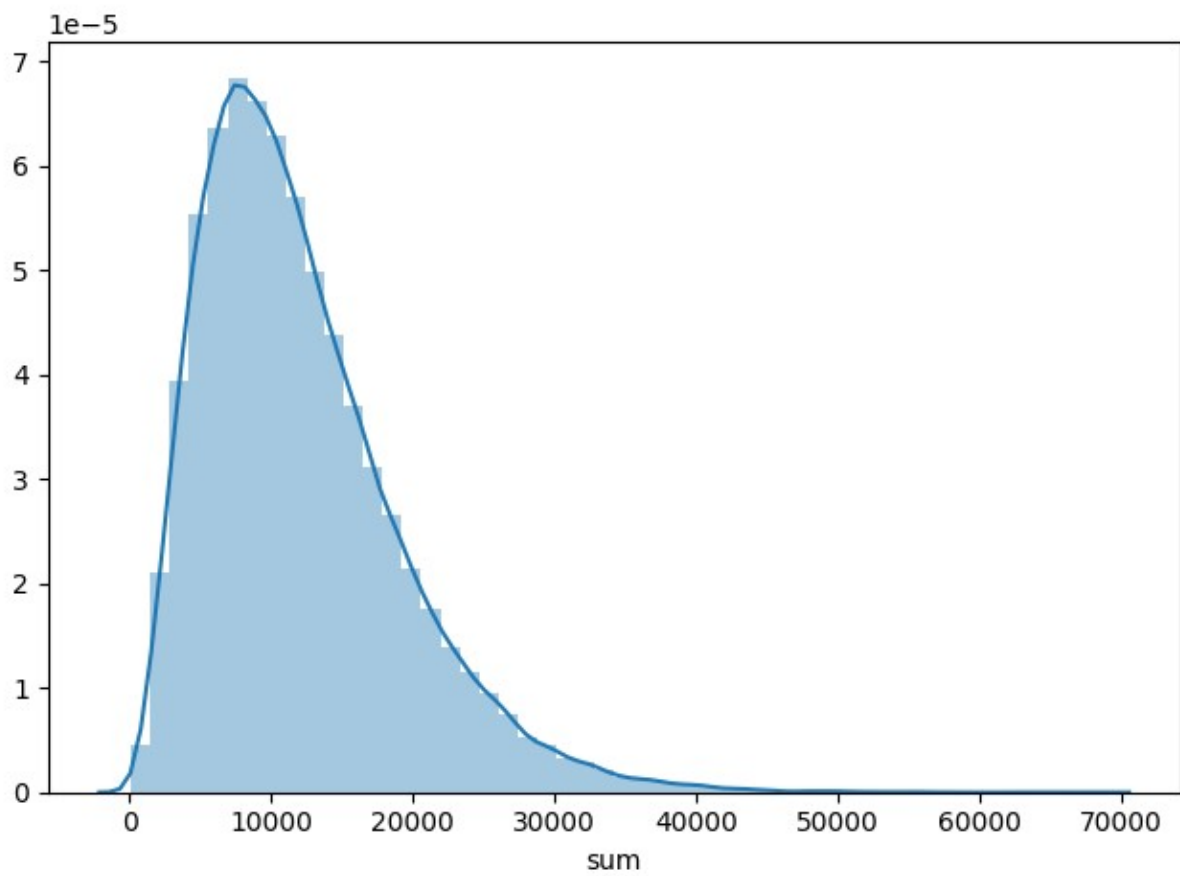


Рис 2. Переважання чеків з сумою 12000 грн

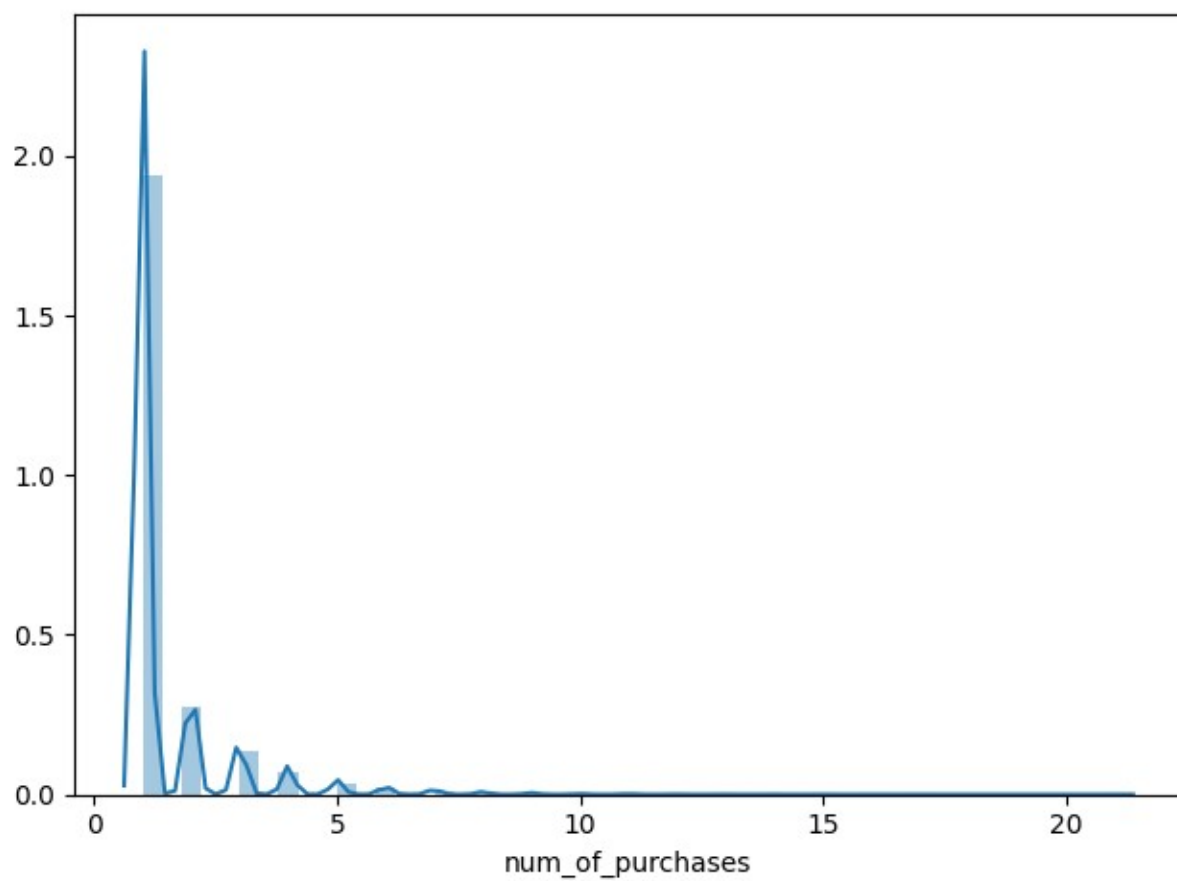


Рис 3. Ймовірність скільки товарів придбає покупець

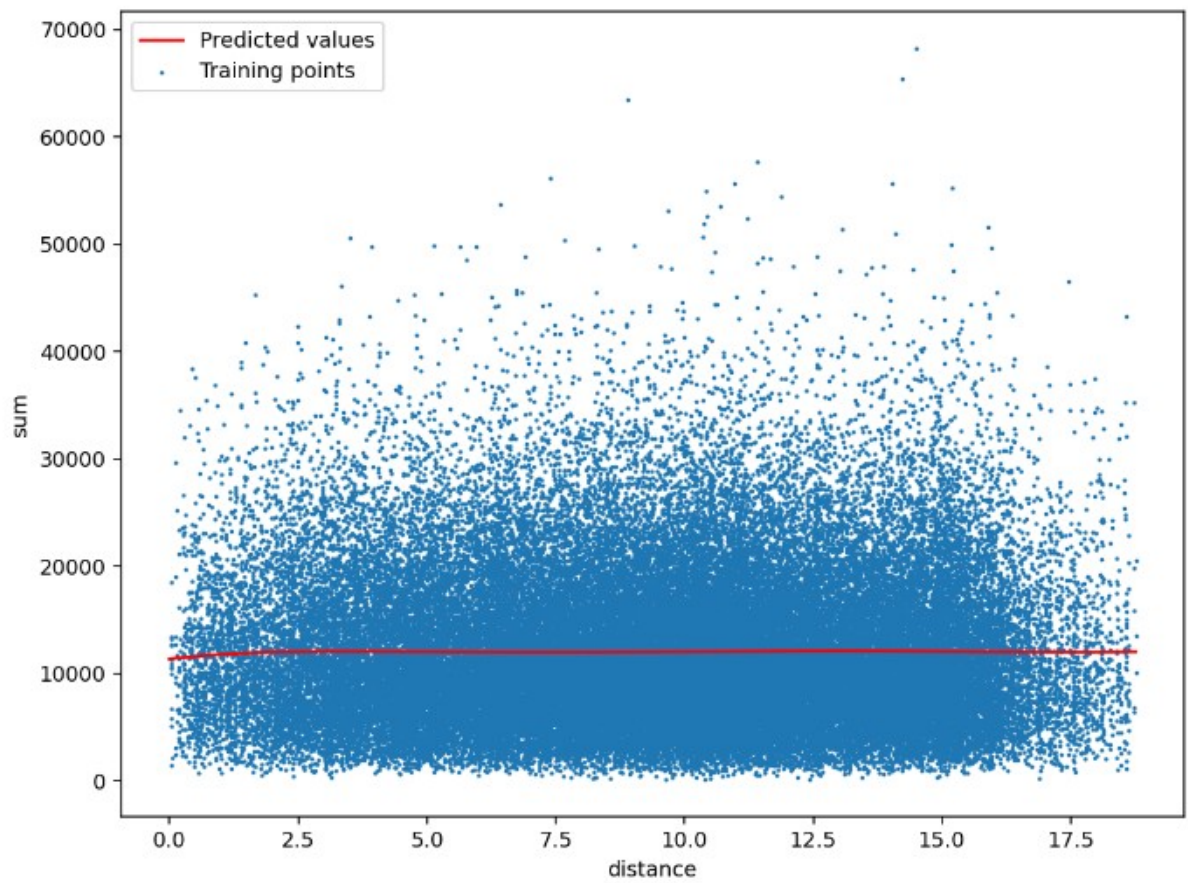


Рис 4. Залежності суми чека від дальності покупця

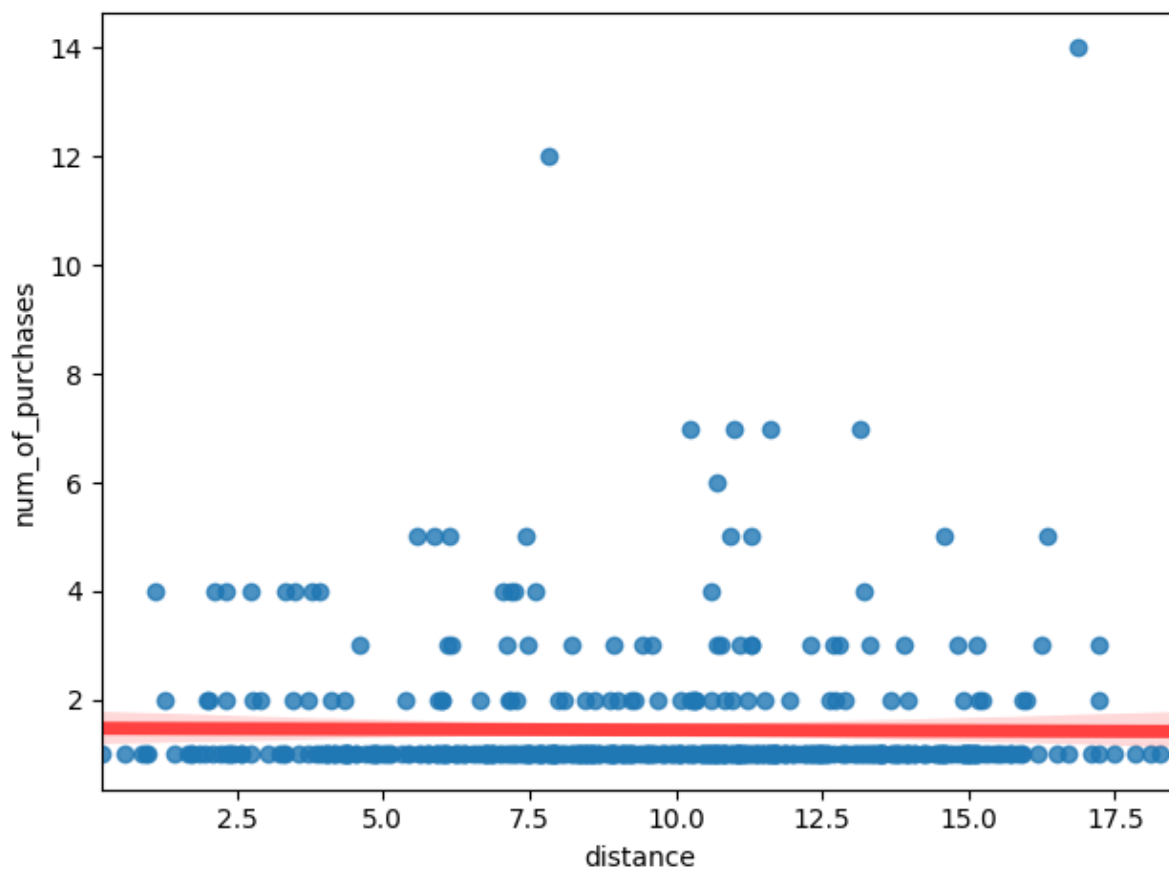


Рис 6. Залежність кількості товарів від дальності до покупця

Додаток Б

Таблиця 1. Вміст конфігураційного файлу *docker-compose*

docker-compose.yml
<pre>version: '2' services: ## Router router01: image: mongo:4.0 container_name: rydell-router-01 command: mongos --port 27017 --configdb rs-config-server/configsvr01:27017,configsvr02:27017,configsvr03:27017 --bind_ip_all ports: - 27117:27017 volumes: - ./scripts:/scripts router02: image: mongo:4.0 container_name: rydell-router-02 command: mongos --port 27017 --configdb rs-config-server/configsvr01:27017,configsvr02:27017,configsvr03:27017 --bind_ip_all volumes: - ./scripts:/scripts ports: - 27118:27017 links: - router01 ## Config Servers configsvr01: image: mongo:4.0 container_name: rydell-mongo-config-01 command: mongod --port 27017 --configsvr --replSet rs-config-server volumes: - ./scripts:/scripts ports: - 27119:27017 links: - shard01-a - shard02-a - shard03-a</pre>


```
configsvr02:
  image: mongo:4.0
  container_name: rydell-mongo-config-02
  command: mongod --port 27017 --configsvr --replSet rs-config-server
  volumes:
    - ./scripts:/scripts
  ports:
    - 27120:27017
  links:
    - configsvr01
configsvr03:
  image: mongo:4.0
  container_name: rydell-mongo-config-03
  command: mongod --port 27017 --configsvr --replSet rs-config-server
  volumes:
    - ./scripts:/scripts
  ports:
    - 27121:27017
  links:
    - configsvr02
```

Shards

Shards 01

```
shard01-x:
  image: mongo:4.0
  container_name: rydell-shard-01-node-x
  command: mongod --port 27017 --shardsvr --replSet rs-shard-01
  volumes:
    - ./scripts:/scripts
  ports:
    - 27122:27017
  links:
    - shard01-a
    - shard01-b
shard01-a:
  image: mongo:4.0
  container_name: rydell-shard-01-node-a
  command: mongod --port 27017 --shardsvr --replSet rs-shard-01
  volumes:
    - ./scripts:/scripts
  ports:
    - 27123:27017
shard01-b:
  image: mongo:4.0
```

container_name: rydell-shard-01-node-b
command: mongod --port 27017 --shardsvr --replSet rs-shard-01
volumes:
- ./scripts:/scripts
ports:
- 27124:27017

Shards 02

shard02-x:
image: mongo:4.0
container_name: rydell-shard-02-node-x
command: mongod --port 27017 --shardsvr --replSet rs-shard-02
volumes:
- ./scripts:/scripts
ports:
- 27125:27017
links:
- shard02-a
- shard02-b

shard02-a:
image: mongo:4.0
container_name: rydell-shard-02-node-a
command: mongod --port 27017 --shardsvr --replSet rs-shard-02
volumes:
- ./scripts:/scripts
ports:
- 27126:27017

shard02-b:
image: mongo:4.0
container_name: rydell-shard-02-node-b
command: mongod --port 27017 --shardsvr --replSet rs-shard-02
volumes:
- ./scripts:/scripts
ports:
- 27127:27017

Shards 03

shard03-x:
image: mongo:4.0
container_name: rydell-shard-03-node-x
command: mongod --port 27017 --shardsvr --replSet rs-shard-03
volumes:
- ./scripts:/scripts
ports:

- 27128:27017

links:

- shard03-a

- shard03-b

shard03-a:

image: mongo:4.0

container_name: rydell-shard-03-node-a

command: mongod --port 27017 --shardsvr --replSet rs-shard-03

volumes:

- ./scripts:/scripts

ports:

- 27129:27017

shard03-b:

image: mongo:4.0

container_name: rydell-shard-03-node-b

command: mongod --port 27017 --shardsvr --replSet rs-shard-03

volumes:

- ./scripts:/scripts

ports:

- 27130:27017
