**Міністерство освіти і науки України**
**Тернопільський національний технічний університет імені Івана Пулюя**

Факультет комп'ютерно-інформаційних систем та програмної інженерії
(повна назва факультету )
Кафедра програмної інженерії
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

Магістр
(назва освітнього ступеня)

на тему: Проєктування та розробка платформи керування маркетинговими даними на основі Airflow та Hadoop

Виконав(ла): студент(ка) 6 курсу, групи СПм-61
спеціальності 121 Інженерія програмного забезпечення
(шифр і назва спеціальності)

| | | |
|---|---|---|
| | (підпис) | Кишкевич О. О. (прізвище та ініціали) |
| Керівник | (підпис) | Михалик Д. М. (прізвище та ініціали) |
| Нормоконтроль | (підпис) | Стоянов Ю. М. (прізвище та ініціали) |
| Завідувач кафедри | (підпис) | Петрик М. Р. (прізвище та ініціали) |
| Рецензент | (підпис) | Жаровський Р. О. (прізвище та ініціали) |

Тернопіль 2022

Міністерство освіти і науки України
**Тернопільський національний технічний університет імені Івана Пулюя**

Факультет комп'ютерно-інформаційних систем та програмної інженерії
(повна назва факультету)

Кафедра програмної інженерії
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
Петрик М. Р.
(підпис) (прізвище та ініціали)
« » 2022 р.

# З А В Д А Н Н Я
# НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня Магістр
(назва освітнього ступеня)

за спеціальністю 121 Інженерія програмного забезпечення
(шифр і назва спеціальності)

студенту Кишкевичу Олегу Олеговичу
(прізвище, ім'я, по батькові)

1. Тема роботи Проєктування та розробка платформи керування маркетинговими даними на основі Airflow та Hadoop

Керівник роботи Михалик Дмитро Михайлович канд. техн. наук, доц.
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від «____» _____ 20___ року № _____

2. Термін подання студентом завершеної роботи

3. Вихідні дані до роботи вимоги до функціоналу системи, використання технологій Airflow та Hadoop

4. Зміст роботи (перелік питань, які потрібно розробити)
Проєктування та розробка системи для надання послуг маркетингової платформи даних. Реалізація streaming та batch потоків обробок даних для процесів маркетингу. Проєктування сховища великих об'ємів даних. Використання технології Airflow для оркестрації butch потоків даних. Використання технологій Hadoop для зберігання та обробки великих об'ємів даних.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
Діаграма варіантів використання, діаграма компонентів платформи, діаграма процесу обробки даних, схема сховища даних, діаграма розгортання, слайди презентації.

6. Консультанти розділів роботи

| Розділ | Прізвище, ініціали та посада консультанта | Підпис, дата |
|---|---|---|

| | | завдання видав | завдання прийняв |
|---|---|---|---|
| Охорона праці | Осухівська Галина Михайлівна | 30.11.2022 | |
| Безпека в надзвичайних ситуаціях | Клепчик Василь Михайлович | | |

7. Дата видачі завдання _____

## КАЛЕНДАРНИЙ ПЛАН

| № з/п | Назва етапів роботи | Термін виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Аналіз предметної області | 11.09.2022 | Виконано |
| 2 | Огляд існуючих рішень | 30.09.2022 | Виконано |
| 3 | Проєктування та розробка платформи | 17.10.2022 | Виконано |
| 4 | Підготовка пояснювальної записки | 20.11.2022 | Виконано |
| 5 | Спецчастина | 25.11.2022 | Виконано |
| 6 | Підготовка презентації та доповіді | 26.11.2022 | Виконано |
| 7 | Попередній захист | 30.11.2022 | Виконано |
| 8 | Нормоконтроль | | |
| 9 | Рецензування | | |
| 10 | Занесення диплома в електронний архів | | |
| 11 | Допуск до захисту у зав. кафедри | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

Студент _____ Кишкевич О. О.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Михалик Д. М.
(підпис) (прізвище та ініціали)

## РЕФЕРАТ

Кваліфікаційна робота на тему «Проектування та розробка платформи керування маркетинговими даними на основі Airflow та Hadoop» написана Кишкевичом Олегом Олеговичем, студентом Тернопільського національного технічного університету імені Івана Пулюя, Факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра програмної інженерії, група СПм–61.

Відомості про обсяг: сторінок – __, рисунків – 18, таблиць – 1, частин – 4, додатків – 3, посилань – 19.

Метою кваліфікаційної роботи є дослідження, проєктування та розробка платформи, яка дозволяє автоматизувати керування маркетинговим процесом та даними, які виникають під час проведення маркетингових кампаній.

Отриманими результатами є проект який можна використовувати як платформу для маркетингових даних. Розроблену платформу можна використовувати для маркетингових досліджень, аналізу ринку та планування маркетингової кампанії. Також робота має великий потенціал для продовження, доданная нових процесів, поглиблення аналітики.

Ключові слова: МАРКЕТИНГ, МАРКЕТИНГОВА ПЛАТФОРМА ДАНИХ, ETL, BLOCKLIST, HADOOP, AIRFLOW, DATA GOVERNANCE.

# ABSTRACT

The qualification paper on the topic "Design and development of a marketing data management platform based on Airflow and Hadoop" was written by Oleg Olegovich Kishkevich, a student of the Ternopil National Technical University named after Ivan Pulyu, Faculty of Computer Information Systems and Software Engineering, Department of Software Engineering, SPm group 61.

Volume information: pages – __, figures – 18, tables – 1, parts – 4, appendices – 3, references – 19.

The purpose of the qualification work is to research, design and develop a platform that allows you to automate the management of the marketing process and data that arise during marketing campaigns.

The obtained results are a project that can be used as a platform for marketing data. The developed platform can be used for marketing research, market analysis and marketing campaign planning. Also, the work has great potential for continuation, adding new processes, deepening analytics.

Keywords: MARKETING, DATA MARKETING PLATFORM, ETL, BLOCKLIST, HADOOP, AIRFLOW, DATA GOVERNANCE.

ЗМІСТ

# ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

1. CRM – Customer Relationship Management
2. DMP – Data Management Platform
3. СКВ – Система керування вмістом
4. CDP – Customer Data Platform
5. ІІ – Ідентифікаційна інформація
6. ПМД – Платформа маркетингових даних
7. GCP – Google cloud platform
8. AWS – Amazon web services
9. HDFS – hadoop distributed file system
10. MWAA – Managed Workflows for Apache Airflow
11. IDE – Integrated Drive Electronics
12. GDPR – General Data Protection Regulation
13. CDC – Change Data Capture
14. ETL – Extract Transform Load
15. ELT – Extract Load Transform
16. GE – Great Expectation
17. DG – Data Governance
18. ПЗ – Програмне забезпечення
19. ПК – Персональний комп'ютер
20. БЖД – Безпека життєдіяльності
21. ДСН – Державні санітарні норми
22. ЦО – Цивільний об’єкт

# ВСТУП

У сучасних ринкових реаліях фірмам, які виділяють ресурси на проведення маркетингових кампаній, зазвичай добиваються більшого успіху, ніж їх конкуренти. Маркетинг спрямований на дослідження ринку та покупців, на пошук підходящих пропозицій для клієнтів. Результатом маркетингових кампаній зазвичай є залучення нових користувачів, створення іміджу фірми, чи збільшення продаж. Дані результати позитивно впливають на бізнес, адже збільшують прибутки компанії [1]. З іншого боку користувач отримує товар, який йому не потрібен. Але маркетинг вдало впорався з поставленим завданням та впарив цей товар.

Актуальність даної роботи обґрунтовується потребою бізнесу у збільшенні прибутків за рахунок маркетингу. Автоматизації певних етапів маркетингової кампанії значно спростить, пришвидшить та зробить надійнішим саме проведення цієї компанії.

Для автоматизації процесу маркетингу потрібно розробити програмне рішення. Дане рішення повинно бути складним та комплексним, щоб задовольнити всі бізнес вимоги даної предметної області. Також можливості програмної системи потрібно надавати декільком компаніям, що додає додаткових труднощів у реалізації. З даною вимогою наше програмне рішення повинно надавати послуги платформи. Тобто, декілька різних компаній матимуть доступ до функціоналу системи, що дасть їм можливість обмінюватися власними маркетинговими даними.

Метою та завданням даної кваліфікаційної роботи є дослідження, проєктування та розробка платформи, яка дозволяє автоматизувати керування маркетинговим процесом та даними, які виникають під час проведення маркетингових кампаній.

Об’єктом дослідження даної роботи є повний процес маркетингової компанії.

Предметом дослідження слугує проєктування та розробка програмного рішення, платформи для керування даними маркетингу.

У процесі роботи було використано декілька методів дослідження. Метод спостереження був використаний для аналізу предметної області, а саме аналізу готових рішень. Метод вимірювання був використаний для формування чисельної оцінки об’ємів даних, які є на даний момент у маркетингових компаніях замовників. На основі теоретичних знань про технології було побудовано гіпотетичне вирішення даної проблеми, після чого науково-дослідницьким методом було перевірено створену гіпотезу та завершено інші частини роботи. Тому у результаті основний метод дослідження, який був використаний у даній кваліфікаційній роботі, це емпірично-науковий метод.

У даній роботі представляється удосконалення існуючих на даний момент маркетинговим платформам даних. Дане вдосконалення враховує недоліки попередників та змінює підхід до роботи з даними, щоб уникнути проблем, які виникають в існуючих рішеннях. Вперше пропонується таке архітектурне рішення, описане далі в роботі, на основі даних технологій Airflow та Hadoop.

Практичним значенням отриманих результатом є описане наукове дослідження даної предметної області, аналіз можливих рішень та вибір найбільш вигідного із них. Описані особливості роботи платформи та можливі шляхи для її покращення. Також було розроблене повноцінне програмне рішення, яке пройшло перевірку на користувацьких даних, та на даний момент використовується в продакшені.

Апробація результатів магістерської роботи. Ключові результати роботи та виконаних досліджень обговорювались на:

IX науково-технічній конференції «Інформаційні моделі, системи та технології» (м. Тернопіль, 2021 р.).

X науково-технічній конференції «Інформаційні моделі, системи та технології» (м. Тернопіль, 2022 р.).

# 1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

## 1.1 Аналіз потреб сфери маркетингу

Коли справа доходить до визначення маркетингу або відповіді на те, що таке маркетинг, визначення може бути некорисним, оскільки термін маркетинг є дещо змінним і всеохоплюючим для прямого визначення. Незважаючи на це, Hubspot визначає маркетинг як «дії, які компанія вживає для залучення аудиторії до продуктів або послуг компанії за допомогою високоякісних повідомлень». Маркетинг націлений на те, щоб за допомогою контенту створити окрему цінність для споживачів і потенційних клієнтів із довгостроковою метою зміцнення лояльності до бренду, демонстрації цінності продукту та збільшення продажів [2].

Маркетинг включає в себе різні аспекти бізнесу, такі як розробка продукту, реклама, продаж і методи розповсюдження. Основна мета маркетингу - зацікавити людей продукцією або послугами компанії. Це відбувається за допомогою аналізу ринку, дослідження та розгляду інтересів ідеальних клієнтів компанії та залучення їх за допомогою обміну повідомленнями, які були б освітніми та корисними для цільової групи бізнесу. У свою чергу, це також допоможе підприємствам перетворити більшу кількість потенційних клієнтів на клієнтів [2].

Цифровий маркетинг, також званий як онлайн-маркетинг, — це просування брендів для зв’язку з потенційними клієнтами за допомогою Інтернету та інших форм цифрового спілкування. Це включає не лише електронну пошту, соціальні медіа та веб-рекламу, а й текстові та мультимедійні повідомлення як маркетинговий канал.

Будь-який вид маркетингу може допомогти вашому бізнесу процвітати. Однак цифровий маркетинг стає все більш важливим через доступність цифрових каналів. Насправді лише у квітні 2022 року у світі було 5 мільярдів користувачів Інтернету [3].

Від соціальних медіа до текстових повідомлень існує багато способів використання тактики цифрового маркетингу для спілкування з цільовою аудиторією. Окрім цього, цифровий маркетинг має мінімальні початкові витрати, що робить його економічно ефективним маркетинговим методом для малого бізнесу.

У цифровому маркетингу існує стільки спеціалізацій, скільки способів взаємодії за допомогою цифрових медіа. Ось кілька ключових прикладів тактик цифрового маркетингу які цікавлять нас.

Партнерський маркетинг (Affiliate marketing) — це тактика цифрового маркетингу, яка дозволяє комусь заробляти гроші, просуваючи бізнес іншої людини. Ви можете бути промоутером або компанією, яка співпрацює з промоутером, але в обох випадках процес однаковий. Він працює за моделлю розподілу доходу. Якщо ви є філією, ви отримуєте комісію щоразу, коли хтось купує товар, який ви рекламуєте. Якщо ви продавець, ви платите філії за кожен продаж, який вони вам допомагають здійснити. Деякі афілійовані маркетологи вирішують оглядати продукти лише однієї компанії, можливо, у блозі чи на іншому сторонньому сайті. Інші мають стосунки з кількома продавцями.

Концепція email маркетингу проста: ви надсилаєте рекламне повідомлення й сподіваєтесь, що ваш потенційний клієнт натисне на нього. Однак виконання набагато складніше. Перш за все, ви повинні переконатися, що ваші електронні листи бажані. Це означає, що у вас є список, який виконує такі дії:

- Індивідуалізує вміст, як в основній частині, так і в рядку теми.
- Чітко вказано, які електронні листи отримуватиме підписник
- Підпис електронної пошти, який пропонує чітку опцію скасування підписки
- Інтегрує трансакційні та рекламні електронні листи

Ви хочете, щоб потенційні клієнти сприймали вашу кампанію як цінну послугу, а не лише як інструмент просування.

Маркетинг електронною поштою сам по собі є перевіреною й ефективною технікою: 89% опитаних професіоналів назвали його найефективнішим генератором потенційних клієнтів [4].

Це може бути навіть краще, якщо ви включите інші методи цифрового маркетингу, такі як автоматизація маркетингу, яка дає змогу сегментувати та планувати ваші електронні листи, щоб вони ефективніше відповідали потребам клієнтів.

Вимоги яких потрібно дотримуватися при маркетингу електронною поштою:

- Сегментуйте свою аудиторію, щоб направити релевантні кампанії потрібним людям.
- Створіть графік кампанії

Цифровий маркетинг став популярним завдяки тому, що він охоплює таку широку аудиторію людей. Однак він також пропонує низку інших переваг, які можуть посилити ваші маркетингові зусилля. Це лише деякі з переваг цифрового маркетингу.

Широке географічне охоплення, коли ви розміщуєте оголошення в Інтернеті, люди можуть бачити його незалежно від того, де вони знаходяться (за умови, що ви не обмежили своє оголошення географічно). Це спрощує розширення охоплення вашого бізнесу на ринку та підключення до більшої аудиторії через різні цифрові канали.

Цифровий маркетинг не тільки охоплює ширшу аудиторію, ніж традиційний маркетинг, але й має нижчі витрати. Накладні витрати на газетні оголошення, телевізійні ролики та інші традиційні маркетингові можливості можуть бути високими. Вони також дають вам менше контролю над тим, чи побачить ваша цільова аудиторія ці повідомлення в першу чергу. За допомогою цифрового маркетингу ви можете створити лише 1 вміст, який приваблює відвідувачів вашого блогу, поки він активний. Ви можете створити маркетингову кампанію електронною поштою, яка доставлятиме повідомлення до цільових списків клієнтів за розкладом, і за потреби легко змінити цей розклад або вміст. Коли ви все це підсумуєте, цифровий маркетинг дає вам набагато більше гнучкості та контакту з клієнтами для ваших витрат на рекламу.

## 1.2 Аналіз існуючих рішень

### 1.2.1 Загальні можливості ПМД

Платформи маркетингових даних збирають всі доступні дані про клієнтів і потенційних клієнтів компанії, зібрані з різних програм, каналів і медіа.

Це може включати дані з таких технологій, як CRM (Customer Relationship Management), програмне забезпечення електронного маркетингу, мобільні маркетингові платформи, платформи наукових даних, веб-аналітика, DMP (Data Management Platform), системи продажів, опитування щодо відстеження брендів, веб-сайт або СКВ (система керування вмістом), фінансові системи, дані кол-центру тощо.

Збираючи та об’єднуючи всю цю інформацію, маркетологам і компаніям легше отримати зведення даних. Вони можуть дізнатися, хто їхня аудиторія і як вони поводяться або взаємодіють з продуктом або послугою компанії. Цей підсумок також може сказати їм чи працюють їхні наявні кампанії, чи ні. Найважливіше те, що це дозволяє маркетологам адаптувати свої майбутні маркетингові рішення на основі точних, всеосяжних даних, щоб гарантувати, що їхні зусилля можуть досягти успіху.

Найчастіше платформи маркетингових даних для досягнення цих цілей розробляються як CDP або платформа даних клієнтів.

Важливі функції, які надають платформи маркетингових даних:

Вони можуть зберігати та надавати ідентифікаційну інформацію (ІН). Ця здатність зберігати ідентифікаційну інформацію від клієнтів є неймовірно цінною, оскільки інформація надходить від людей, які купують ваші продукти або користуються вашими послугами. За допомогою ідентифікаційної інформації ПМД може створити уніфікований профіль, який потім дозволить вам налаштувати свою маркетингову стратегію на основі цільового сектору або дій, які ви б хотіли зробити.

ПМД може сприяти взаємодії з клієнтами в реальному часі. Завдяки своїй здатності створювати єдиний профіль клієнта, ПМД може надавати індивідуальні рекомендації, щоб допомогти клієнту протягом усього шляху покупця. Наприклад, якщо клієнт покидає свій кошик на півдорозі на вашому веб-сайті, ПМД може надати вам спеціальний профіль, який можна використовувати для надсилання йому персоналізованої пропозиції або реклами, яка спонукає його завершити покупку.

ПМД збирає та порівнює різні дані з безлічі джерел. Подібно до того, як CDP збирає різноманітні дані про взаємодію з клієнтами, ПМД збирає інформацію, зібрану CDP, CRM, DPM, рекламними інструментами, веб-аналітикою тощо, і транспонує в єдині узгоджені доступні дані.

ПМД об’єднує усі виділені дані з різних платформ. Розташування даних може бути проблематичним для будь-якої маркетингової діяльності. Вітрини даних відноситься до набору необроблених даних, де лише одна група в компанії має доступ до зазначеної інформації. На перший погляд це може здатися нешкідливим, але вітрини даних створюють бар’єри, які призводять до браку інформації або неправильної інтерпретації даних.

ПМД бере сегменти даних, аналізує їх і перетворює на корисну інформацію. Зібрати дані в одному місці недостатньо, найкращі платформи маркетингових даних можуть взяти цю інформацію та перетворити її на практичний контент для компанії. Маючи цілісне уявлення про дані, аналітики можуть помітити тенденції, моделі купівлі, поведінку клієнтів тощо. Потім ця інформація може бути використана маркетологами для формування майбутніх оголошень, проведення маркетингових кампаній електронною поштою, надання унікальних пропозицій, проведення маркетингових кампаній із впливовим впливом, розробки цільових сторінок або створення маркетингових воронок з метою збільшення продажів і прибутку для компанії.

Це лише деякі переваги, які платформа даних клієнтів може надати вашому бізнесу.

### 1.2.2 Аналіз ПМД Autopilot

Autopilot (рис. 1.1) [5] австралійська компанія підвищила ставки для компаній у всьому світі, допомагаючи клієнтам отримувати уніфіковані дані про клієнтів за допомогою своєї ПМД, а також потужної сегментації даних для конкретного покупця.


Рисунок 1.1 – ПМД Autopilot

Переваги:

- Чудовий та інтуїтивно зрозумілий інтерфейс.
- Дозволяє миттєво відстежувати завдяки статистиці в реальному часі.
- Фантастичні розумні сегменти для цільової автоматизації маркетингу.

Недоліки:

- Немає еквівалентів мобільних додатків.
- Немає можливості експортувати електронні листи HTML.
- Потребує удосконалення аналітики ROI.

### 1.2.3 Аналіз ПМД Adverity

Adverity (рис. 1.2) [6] позиціонує себе як платформу для розширення можливостей маркетологів нового покоління. З моменту заснування у 2015 році компанія співпрацює з такими компаніями, як Red Bull, Unilever і Forbes, щоб розробляти їхні маркетингові стратегії за допомогою пакетів Adverity.

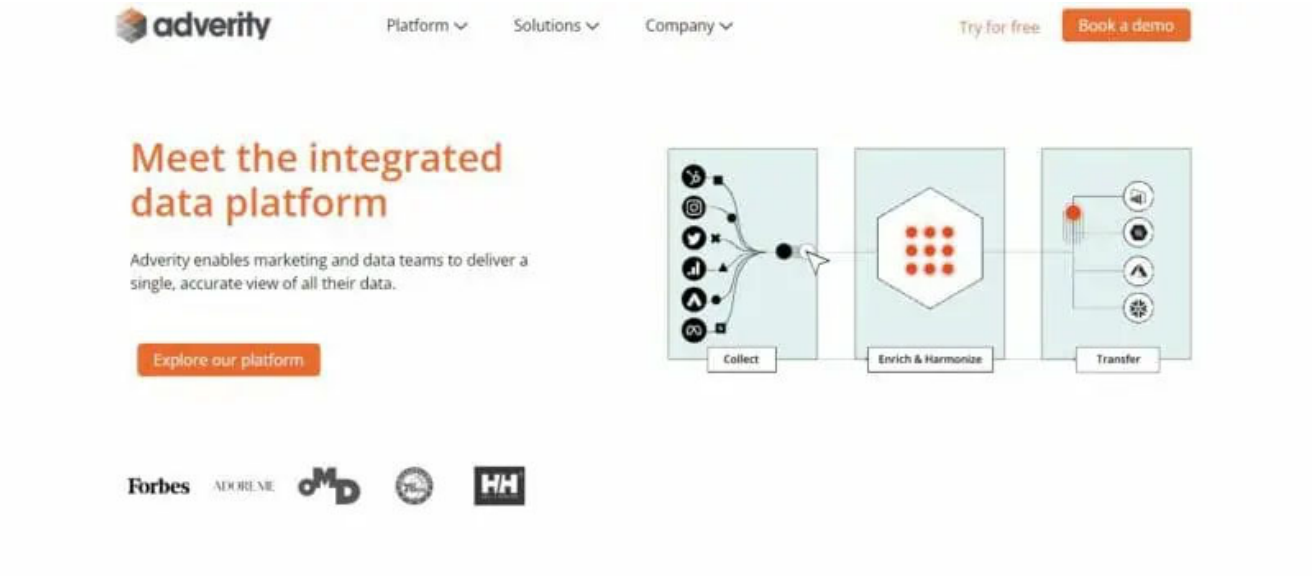


Рисунок 1.2 – ПМД Autopilot

Переваги:

- Чудова підтримка клієнтів.
- Відображення конфігурації чудово підходить для не технічних відділів.
- Проста інтеграція даних.

Недоліки:

- Інтерфейс користувача може бути складним, якщо додати багато показників.
- Допоміжної документації недостатньо.
- Вилучення даних може зайняти деякий час.

### 1.2.4 Аналіз Google Marketing Platform

DMP Google, також відомий як Google Marketing Platform (рис. 1.3) [7], надає рішення для підприємств і малого бізнесу. Компанії може знадобитися витратити більше грошей на DMP Google, якщо вони хочуть розширені функції. Інформаційні панелі автоматизовані, тому будь-хто в організації може легко отримати доступ до ключових показників або критичних даних.

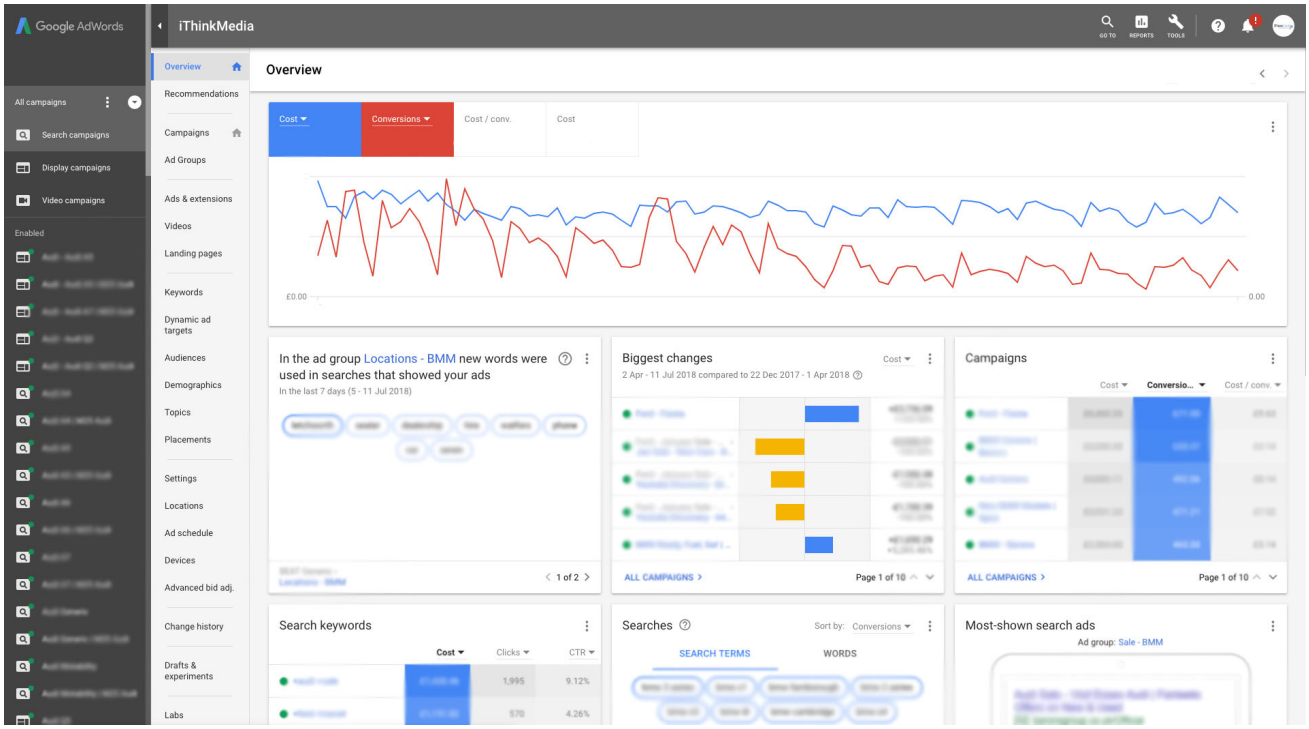


Рисунок 1.3 – Google Marketing Platform

Переваги:

- Рішення для малого та корпоративного бізнесу.
- Інтегрована аналітика.
- Настроюваність.

Недоліки:

- Висока вартість.

- Ускладнене управління портфелем кампанії.

## 1.3 Формування вимог для ПМД

### 1.3.1 Постановка задачі

Після проведення аналізу предметної області та існуючих рішень можна виділити такі основні властивості які повинна задовольняти наша ПМД.

Платформа повинна надати можливості для різних компаній, а самі компанії повинні мати можливості обмінюватися даними. Для цього класифікуємо дані за джерелом походження (табл. 1.1) [8].

Таблиця 1.1 – Джерела даних

| | |
|---|---|
| Дані першої сторони | Дані, зібрані з відвідувань веб-сайтів, систем CRM, соціальних мереж, підписок, мобільних пристроїв і програм. |
| Дані другої сторони | Чиїсь власні дані. Він є похідним від взаємовигідних відносин з іншою компанією (партнером, постачальником тощо), з якою ви ділитеся даними. |
| Дані третіх сторін | Ці дані надходять із веб-сайтів і платформ соціальних мереж, крім ваших власних, і можуть використовуватися для охоплення ширшої аудиторії. Це необхідно для розширення даних першої сторони, щоб маркетологи могли збільшити масштаб і охоплення та покращити персоналізацію. |

Інтеграція маркетингових даних. Збирати та використовувати дані з будь-якої кількості джерел.

- Отримувати власні дані в Інтернеті, на різних пристроях і в автономному режимі
- Класифікувати та збирати дані в цільові сегменти аудиторії.
- Розширювати охоплення аудиторії за допомогою ринку даних.
- Збагачувати дані першої сторони даними других і третіх сторін.

- Відстеження надходження даних і діагностуйте/усувайте проблеми.
- Виділення, скільки разів профілі користувачів було виявлено та категоризовано.

Вибудовування аудиторії. Визначення правильної аудиторії як цілі кампанії та розширення охоплення до найбільш прийнятних груп аудиторії.

- Використання моделювання по схожості, щоб збільшити кількість потенційних клієнтів.
- Використання інструментів дозволу таксономії для обміну даними другої сторони.
- Визначення класифікації, пов’язані з конкретною аудиторією, за допомогою звітів про виявлення аудиторії.
- Здатність добавляти нову аудиторію або розширте поточну цільову аудиторію під час планування кампанії, щоб покращити ефективність.

Націлювання на різні пристрої. Створення персоналізованих та послідовних маркетингових кампаніях на різних пристроях.

- Розширити аудиторію; передавати дані третіх сторін через різні маркетингові канали та пристрої.
- Використання графіку приватних ідентифікаторів, щоб розширити охоплення та використати зв’язки ідентифікаторів для кращого націлювання на клієнтів на різних пристроях.

Аналіз аудиторії. Можливість дізнайтеся, які маркетингові кампанії ефективні та які пристрої забезпечують найбільше конверсій і продажів.

- Отримання доступу до корисної інформації про свою цільову аудиторію за допомогою надійних аналітичних звітів аудиторії.
- Використання аналітики перед кампанією, щоб зрозуміти, хто становить вашу аудиторію, перш ніж проводити показ.
- Можливість дізнайтеся про точність націлювання за допомогою аналізу після кампанії.
- Можливість відмітити тих, хто вже звернувся за допомогою звітів про подавлення аудиторії.

### 1.3.2 Визначення акторів системи

З вищеописаних вимог можна виділити актантів системи та описати відмінності між ними. Із основного можна виділити 4 актора, які будуть використовувати основний функціонал платформи. Такими є завантажуючий актор, стягуючий актор, актор модератор та актор адміністратор.

Завантажуючий актор відповідає за додавання нових даних в систему. Він забезпечує платформу новими даними про користувачів. Також він має можливість надавати дані про оновлення в користувача чи групи користувачів.

Стягуючий актор отримує доступ до фінальних результатів роботи платформи. Він може переглядати звіти, дані, анонімні профілі користувачів, результати по маркетингових кампаніях і так далі. Також даний актор може зробити запит на отримання специфічних даних, які будуть потрібні йому для проведення маркетингового дослідження.

Актор модератор отримує права керування над даними, які завантажуються в платформу. Він має можливість вирішувати, які дані можуть піти в звіти, а які дані відкинути. Модератор відповідальний за випадки, якщо в систему надійшли дані, які наша платформа не може опрацювати однозначно. В таких випадках модератор отримує доступ до даних та приймає рішення як ці дані опрацьовувати далі. У модератора є достатньо прав, щоб надати власні дані компанії іншим компаніям, як дані других сторін. Модератор має можливість підключити постійне зовнішнє джерело даних, з якого платформа забиратиме дані і підготовлюватиме їх для звітів. Це ж стосується вигрузки даних. Модератор може вказати, які дані і куди їх направляти.

Актор адміністратор має права доступу до всього, що відбувається на платформі. Адміністратор відповідальний за надання іншим користувачам ролей та доступів. Також у випадку системного збою в роботі платформи саме адміністратор відповідальний за уточнення деталей та усунення даної проблеми.

### 1.3.3 Опис ключових варіантів використання

Підсумовуючи аналіз предметної області на рисунку 1.4 зображено основні варіанти використання платформи.


Рисунок 1.4 – Основні варіанти використання системи

У продемонстрований діаграмі чітко зображено актантів системи та основні варіанти використання.

Серед цих варіантів використання можна виділити саме такі:

- Для звичайного користувача передбачається такі можливості:
    - Можливість підписатися на маркетингову компанію.
    - Можливість відписатися від маркетингової компанії.
- Для користувача Завантажувальника передбачається такі можливості:
    - Можливість завантаження різного роду маркетингових даних.
    - Можливість переглядати власні завантажені дані.
    - Можливість оновлювати та керувати завантажені дані.
- Для користувача Стягуючого користувача передбачається такі можливості:
    - Можливість стягувати дані маркетингових компаній.
    - Можливість доступу до фінальних результатів обробки даних.
    - Можливість переглядати аналітику.
    - Можливість отримання доступу до даних.
- Користувач Модератор наслідує варіанти використання користувачів Завантажувальника та Стягувача та має такі додаткові можливості:
    - Можливість надавати доступ до даних других сторін.
    - Можливість встановлювати постійні джерела даних.
    - Можливість керувати підписками.
    - Можливість будувати аналітику та репорти.
    - Можливість влаштувати процес обробки даних.
- Користувач Адміністратор наслідує варіанти використання користувача Модератор та має такі додаткові можливості:
    - Можливість керування користувачами та їх правами.
    - Можливість керування елементами платформи.

# 2 ПРОЄКТУВАННЯ ТА МОДЕЛЮВАННЯ

## 2.1 Вибір процесу розробки

На даному етапі було обрано основні технології для реалізації даної платформи. Відповідно до вимог швидкості впровадження змін, та особливостей роботи з основними технологіями було обрано методологію розробки.

### 2.1.1 Вибір основних технологій розробки

У відповідності до вимог та проведеного дослідження по природі та об’ємах даних, було обрано основні технології. Найважливішою умовою вибору технології була можливість працювати з великими об’ємами даних. А саме розглядалися рішення, які можна масштабувати. Такі рішення пропонують клауди за рахунок власних сервісів, але є можливість власне самому на власному сервері (on premis) справитися з такими об’ємами даних. Порівняємо ці 2 варіанта та обґрунтуємо прийняте рішення.

- Розгортання:
    - На власному сервері ресурси розгортаються всередині компанії та в IT-інфраструктурі підприємства. Підприємство відповідає за підтримку рішення та всіх пов’язаних з ним процесів.
    - Клауди. Хоч існують різні форми хмарних обчислень у загальнодоступному хмарному середовищі ресурси розміщуються на території постачальника послуг, але підприємства мають доступ до цих ресурсів. і використовувати скільки завгодно у будь-який момент часу.

- Ціна:
    - На власних серверах для підприємств, які розгортають програмне забезпечення у себе або орендують у серверних центрах, вони відповідають за поточні витрати на серверне обладнання, енергоспоживання та простір.
    - У клауді потрібно платити лише за ресурси, якими вони користуються, без жодних витрат на технічне обслуговування та обслуговування, а ціна коригується вгору або вниз залежно від того, скільки споживається.
- Безпека:
    - На власних серверах компанії, які мають надто конфіденційну інформацію, як-от державні та банківські установи, повинні мати певний рівень безпеки та конфіденційності, який забезпечує локальне середовище. Незважаючи на перспективи хмари, безпека є головною проблемою для багатьох галузей, тому локальне середовище, незважаючи на деякі його недоліки та ціну, є більш доцільним.
    - У клауді проблеми безпеки залишаються першочерговою перешкодою для розгортання хмарних обчислень. Було багато розголосу про злам клауда, і IT-відділи в усьому світі стурбовані. Загрози безпеці реальні: від особистої інформації працівників, як-от облікові дані для входу, до втрати інтелектуальної власності.

Опираючись на це порівняння, було обрано гібридну модель, де певна частина платформи розгорнута та виконується на власних серверах, а інша частина на виконується на клаудах. Це дозволить використати переваги обох платформ та уникнути їх недоліків. Основна частина розгорталась на власних серверах, що дозволить зменшити витрати на обчислення. У клаудах розгорнуті критичні частини інфраструктури, які вимагають високої надійності, швидкості відклику та можливості витримувати різко навантаження на платформу у випадку, якщо обчислювальних ресурсів на власних серверах не вистачатиме.

Зважаючи на гібридну модель інфраструктури, нам потрібно вибрати технології, що будуть використовуватися на власних серверах такими, щоб у них були аналоги зі схожим API та можливостями як у клаудах.

Для роботи з великими об’ємами даних було обрано набір технологій Hadoop [9]. А саме HDFS для зберігання великих об’ємів даних, аналогом якого може слугувати сервіс Object storage в клауді [10]. Spark для паралельного обчислення великих об’ємів даних, як аналог у клаудах може виступати dataflow від GCP або Glue від AWS. У якості сховища даних (Data warehouse) [11] було використано Hive або Spark, аналогами у кладудах можуть виступати bigtable від GCP або Redshift від AWS.

У процесі дослідження предметної області було виявлено велику кількість різних потоків даних. Тому для зручності керування даними потоками було обрано технологію оркестрації Airflow. Дана технологія спростить розробку та управління платформою, дозволяє гнучко керувати великими об'ємами потоків даних. Аналог Airflow в клауді може слугувати Cloud Composer від GCP, або Amazon Managed Workflows for Apache Airflow (MWAA) від AWS

### 2.1.2 Обгрунтування використання інших технологій

Після визначення основних технологій для розробки тепер можна обрати допоміжні технології, які використовуватимуться при розробці платформи. Розпочнемо з мов програмування.

У даному випадку платформи список мов обмежений теми мовами, які підтримуються вище обраними технологіями. Технологія Airflow накладає на наступне обмеження, використовувати тільки мову python для побудови (опису) data pipeline. Але вона не накладає обмеження стосовно технологій на власне самі етапи (задачі) в pipeline-і.

У випадку технологій Hadoop, рекомендованим підходом є використання однієї з JVM мов, такою як Java або Scala. Але також дана технологія підтримуємо такі мови як Python або R.

Зважаючи на те, що в клаудах є сервіси, які майже аналогічно реалізують можливості та обмеження обраних нами основних технологій, можна зробити висновок, що ми розглянули всі обмеження щодо вибору мов програмування.

У результаті розробки була використана мова Python для написання DAG-ів в Airflow. Також завдяки мові Python було реалізовано деяку частину задачі для процесу обробки даних. Також для простих та невеликих задач у процесі обробки даних була використана скриптова мова програмування Bash. Інші, складні та ресурсоємні, задачі обробки даних були реалізовані з використанням технологій Hadoop, а саме мови Scala.

Для інтеграції та спільної розробки проекту було використано платформу gitlab. На даній платформі було створено декілька репозитаріїв під окремі частини платформи. Також gitlab надав інструмент для CI/CD, завдяки якому вдалося автоматизувати значну частину розробки та доставки змін.

У якості реєстру артефактів була використана технологія sonatype Nexus. У даний реєстр артефактів зберігали jar та uber-jar файли з кодом Spark задач та бібліотек, pip пакети, та інші сутності.

Для конфедерації та розгортання контейнерів використовували технології docker та kubernetes. Основна частина Airflow задач запускала контейнери для обробки даних, яким керував kubernetes.

Для групи технологій Hadoop були виділення власні машини, які сформували Hadoop кластер. Власне на даних машина зберігалися основні об'єми даних та виконували задачі по обробці великих об'ємів даних. Кластер від доступу зовні був захищений за допомогою Kerberos.

Для синхронної комунікації між деякими частинами платформи було використано протокол http. А для асинхронної комунікації була використана технологія kafka у якості черги повідомлень.

Основним середовищем розробки вибрані IDE іт-компанії jetbrains, а саме pycharm для програмування мовою Python та intellij Idea з scala плагіном для програмування мовою Scala. Додатково використовувався текстовий редактор neovim для програмування з використанням bash або програмування на мові Scala з використанням плагіну scala metals.

Для розробки проекту в основному використовувалися дистрибутиви базовані на Linux, а саме Ubuntu, Kubuntu та NixOs.

### 2.1.3 Вибір методології розробки

Існує список вимог та, визначившись з технологіями використаними для розробки платформи, нами було прийнято рішення використати методологію девопс (DevOps) [13]. Саме ця методологія найкраще підходить для даного проєкту. Основною перевагою є швидкість доставки змін до кінцевого користувача. Також важливою перевагою є потреба в автоматизації процесу розробки, що пришвидшує сам цикл.

Обрані нами технології чудово підходять під дану методологію. Основним інструментом, який забезпечує дотримання методології буде слугувати платформа gitlab. А саме інструменти для планування і постановки етапів розробки, задач та керування кодом. Завдяки gitlab CI/CD було реалізовано етапи інтеграції та доставки змін. Найчастішим етапам CI/CD в більшості проектів було проведення unit тестів, збір контейнера чи пакета, розгортання у тестовому середовищі, запуск інтеграційних тестів та реліз.

У більшості проєктів використовується підхід Trunk-Based Development для того, щоб пришвидшити розробку та уникнути конфліктних ситуацій в коді платформи.

## 2.2 Проєктування архітектури платформи

Після процесу аналізу було виділено три основні потоки даних, а саме потік із даними користувача, потік даних списків блокування та потік даних відписок. Кожен із даних потоків має різне бізнесове значення. Тому було прийнято розробляти дані потоки окремо, незалежно одне від одного, щоб зміни вимог до одного не спричинили проблеми в роботі іншого. Для подальшої розробки дані потоки було вирішено назвати user data, blocklist та suppression.

Хоч і потоки даних є самостійними та не залежать одне від одного, проте на певних етапах ці потоки даних об’єднуються. Це зазвичай відбувається на фінальних стадіях, наприклад, етапу формування звітів чи планування маркетингової кампанії.

Зважаючи на природу даних, те, як часто вони з’являються, об’єми та наскільки швидко їх потрібно задіяти в аналітиці та маркетинговій кампанії, дані потоки можна поділити на batch та streaming процес обробки даних.

Batch процес розробки за один раз обробляє великі об’єми інформації, проте виконується досить повільно, його було вирішено використати у місцях, де потребується працювати з великими об'ємами даних, та не вимагається миттєвий результат. Таким випадком є процес формування маркетингової компанії, коли маркетолог вибирає під які групи користувачів буде проведена маркетингова компанія. Цей етап може потребувати залучення всіх потоків даних, у великій кількості, наприклад, маркетингова кампанія може бути розрахована на 42 мільйона клієнтів. Також на деяких потоках даних власне самі дані додаються раз у добу в різних об’ємах, як наприклад збір даних із постійних джерел, таких як ftp, для blocklist та user data.

Streaming процес розробки використовувався на етапах, які потребують швидкої доставки даних, але самі дані обмежуються в розмірах. Такий процес потрібен для потоку suppression. Потрібно якомога швидше задіяти дані, коли

користувач відписується або відписки вносяться в ручну оператором.

Для Butch потоків обробки даних було використано технологію оркестрації Airflow. DAG описаний в airflow проводив процес обробки та доставки даних. Дані після обробки зберігалися в data warehouse певним чином, щоб оптимізувати подальші запити.

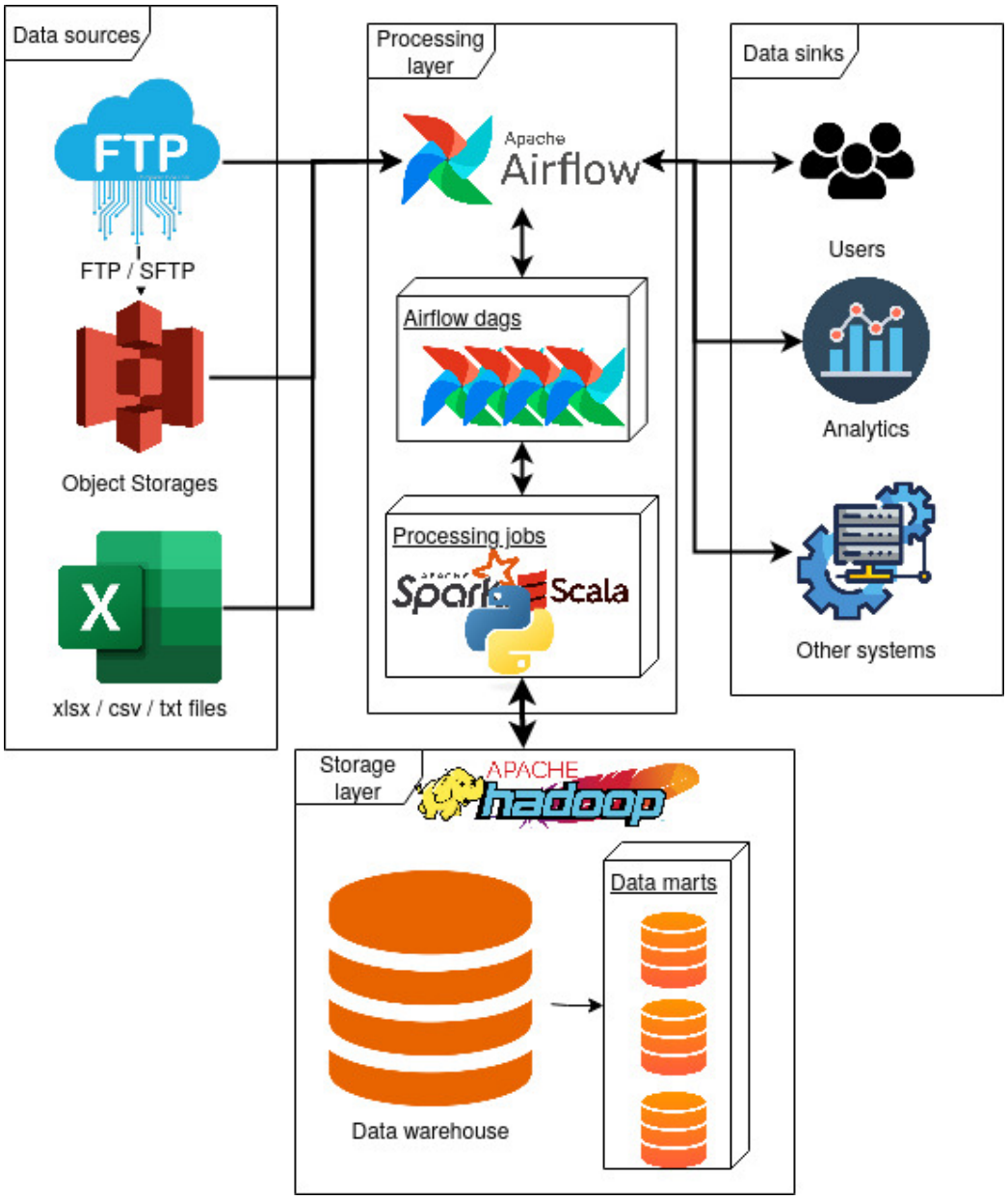


Рисунок 2.1 - Загальна архітектура системи

Для доставки до кінцевого користувача, дані бралися за допомогою airflow, який звертався до spark робіт, які в свою чергу брали дані з data warehouse. Дану

загальну поведінку добре можна розгледіти на рисунку 2.1.

При проєктуванні платформи було виділено основні незалежні потоки даних, серед них user data, blocklist, suppression, audience. Потік User data містить всі важливі дані про користувачів, саме ці дані в основному використовується для побудови маркетингових кампаній. Потік blocklist містить дані про, те яких груп користувачів потрібно уникати при побудові маркетингової кампанії. Потік suppression містить дані про користувачів, які брали участь в маркетинговій кампанії, але відписалися від неї. Потік audience виконує в основному агрегаційну та об'єднуючу функції. Саме цей потік формує фінальні вибірки груп користувачів.

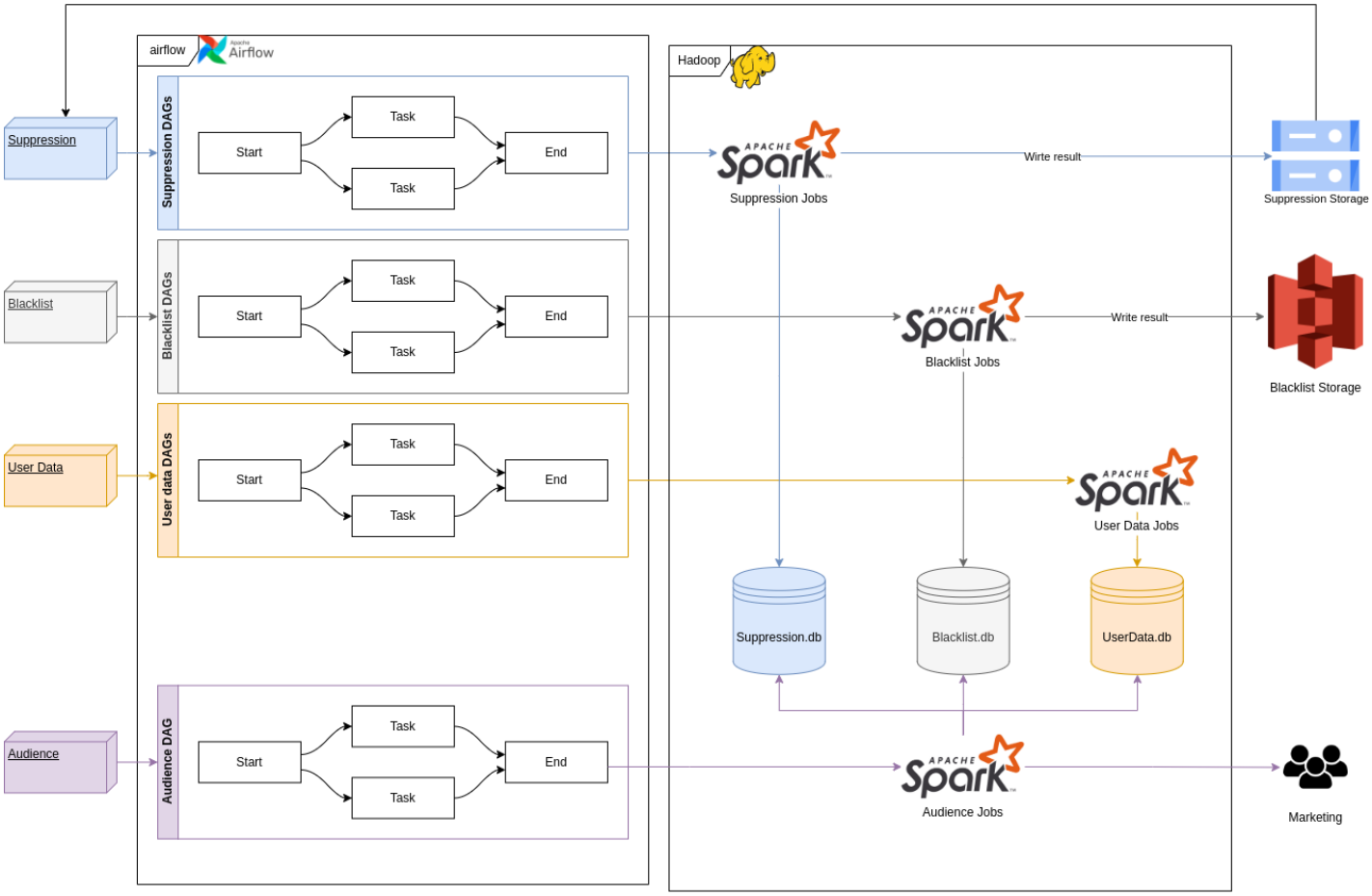


Рисунок 2.2 – Розділення платформи на незалежні потоки

З рисунку 2.2 можна побачити як саме ці потоки даних взаємодіють між собою.

### 2.2.1 Розгляд потоку даних Blocklist

Після загального огляду архітектури платформи та виділення основних потоків даних і того як вони між собою взаємодіють, розглянемо приклад одного з них. Більш детально розглянемо потік з blocklist даними. З точки зору звичайного користувача він може здатися примітивним, але насправді він є більш складним ніж можна цього очікувати.

Згадуючи бізнес вимоги, blocklist дані це дані, які містить списки користувачів чи груп користувачів, яких рекомендовано уникати при побудові аналітики. Зі сторони бізнесу також додається вимога розділення blocklist на різні типи, якщо конкретніше то таких типів є 12 штук, і для бізнесу кожен з них має власне значення. До цих типів відносяться списки з звичайними емейлами користувачів, списки з ip адресами користувачів, списки з доменними іменами, списки з “hard bounce”, списки з GDPR [14] і так далі.

Розглянемо  випадок в якому ми маємо постійне джерело blocklist даних, нехай це буде ftp, але це може бути і інше постійне джерело з якого можна стягнути дані. Такими постійними джерелами можуть слугувати ftp/sftp, різні об'єкти сховища, бази даних, сховища даних, та інші. При встановленні такого постійного джерела даних, розроблена нами платформа буде періодично стягнути дані. При роботі з постійними джерелами даних потрібно враховувати що дані в них зберігаються впродовж тривалого часу, і при запиті потрібно збирати тільки нові дані, які ще не оброблялися платформою. Тому для досягнення такої поведінки був використаний підхід CDC [15]. Для його реалізації було вирішено зберегти якомога більше метаданих про файли, які ми обробляємо. Ці метадані зберігаються у реляційну базу, у нашому випадку ми обрали PostgreSQL. При подільших запитах на джерела даних порівнюється, які дані вже були опрацьовані і вибираються дані, які ще не опрацьовувалися. Після опрацювання інформація

про нові дані додається у базу. При наступних запитах вони будуть фільтруватися.

Визначившись з джерерлом даних, ми обрали наступний етап: обробка. Так як джерела даних можуть бути різними, тому перед цим ми додаємо конектор посередник. Цей конектор буде надавати уніфіковувати інтерфейс доступу до різних джерел даних.

Цей етап уніфікації джерел даних потрібно виділити в окремий модуль. Це дозволяє розділити етап збирання даних і етап розробки, що додасть гнучкості системі. Даний модуль можна назвати extruction. Також додатково на цьому етапі зберігаються оригінальні дані. Це зроблено для того щоб у випадку видалення даних з джерел, їх завжди можна було отримати та повторно обробити. Для зберігання оригінальні дані архівуються.

Кожен етап обробки даних являє собою ETL процес. Підсистеми ETL процесу у нас були поділені на окремі модулі за кожен з модулів відповідає власний DAG.

Наступним етапом виступає трансформація. На даному етапі відбувається направлення вихідних даних через ряд етапів обробки системи ETL для покращення якість отриманих даних від джерела даних. Об'єднання даних з двох або більше джерел задля створення та впровадження уніфікованих розмірів і уніфікованих метрик.

Наступним етапом виступає доставка даних. На цьому етапі фізично структуруються і завантажуються дані в цільовий сервер, у нашому випадку це Hadoop кластер.

Фінальним етапом розробки є управління збереженими даними. На цьому етапі відбувається управління пов'язаними систем і процесів середовища ETL узгоджено.

Виділивши основні підсистеми ETL, ми можемо приступати до розробки нашого ETL процесу для обробки blocklist даних. Для оркестрування кожним етапом обробки даних ми скористалися технологією airflow. Кожен модуль чи підсистема ETL будуть реалізовані за допомогою окремих airflow DAG.

Розглянемо більш детально підсистему extraction. Запропонована архітектура extraction зображена на рисунку 2.3

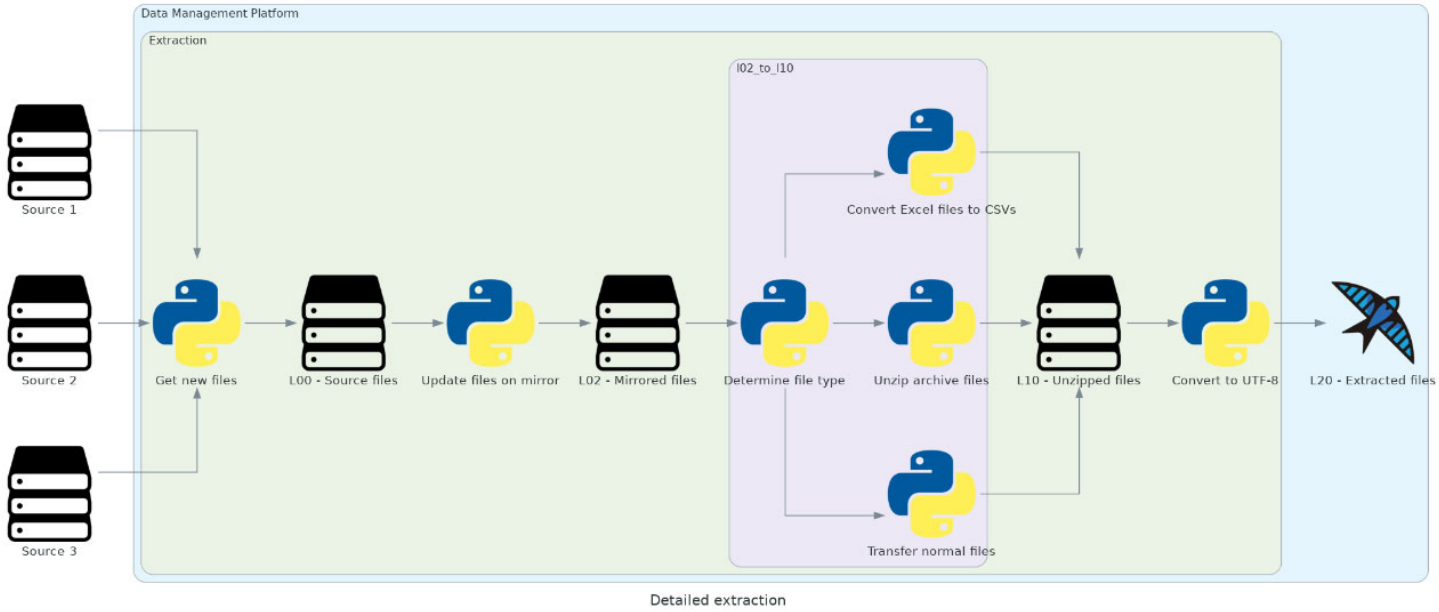


Рисунок 2.3 – Діаграма архітектури етапу extraction

З постійних джерел дані вибираються за допомогою Python скрипта, який уніфікує джерела даних. Результатом роботи даного скрипта є збережені оригінальні файли в об'єктне сховище. Оригінальні дані ми класифікуємо як source files level 0.

Після успішного запису в level 0, запускається скрипт, який віддзеркалює файли. Це потрібно для того, щоб ми не працювали з оригінальними даними, а працювали з їхньою копією. Результатом роботи даного скрипта є збережені копії файлів в обєктне сховище яке ми класифікуємо як mirrored files level 2.

Фінальним етапом роботи підсистеми extraction є архівування оригінальних даних для довгого зберігання. Такий підхід дозволяє нам підстрахуватися у випадку видалення даних із постійних джерел або у випадку, коли дані не коректно обробилися. У нас буде можливість перезапустити обробку даних повторно. Також це дозволить простіше опрацьовувати зміни у бізнес вимогах до обробки даних.

Після того, як ми розглянемо детально як проектувати одну з підсистем, можна схожим чином спроектувати наступні. У результаті ми повинні отримати цілісну архітектуру системи зображену на рисунку 2.4.

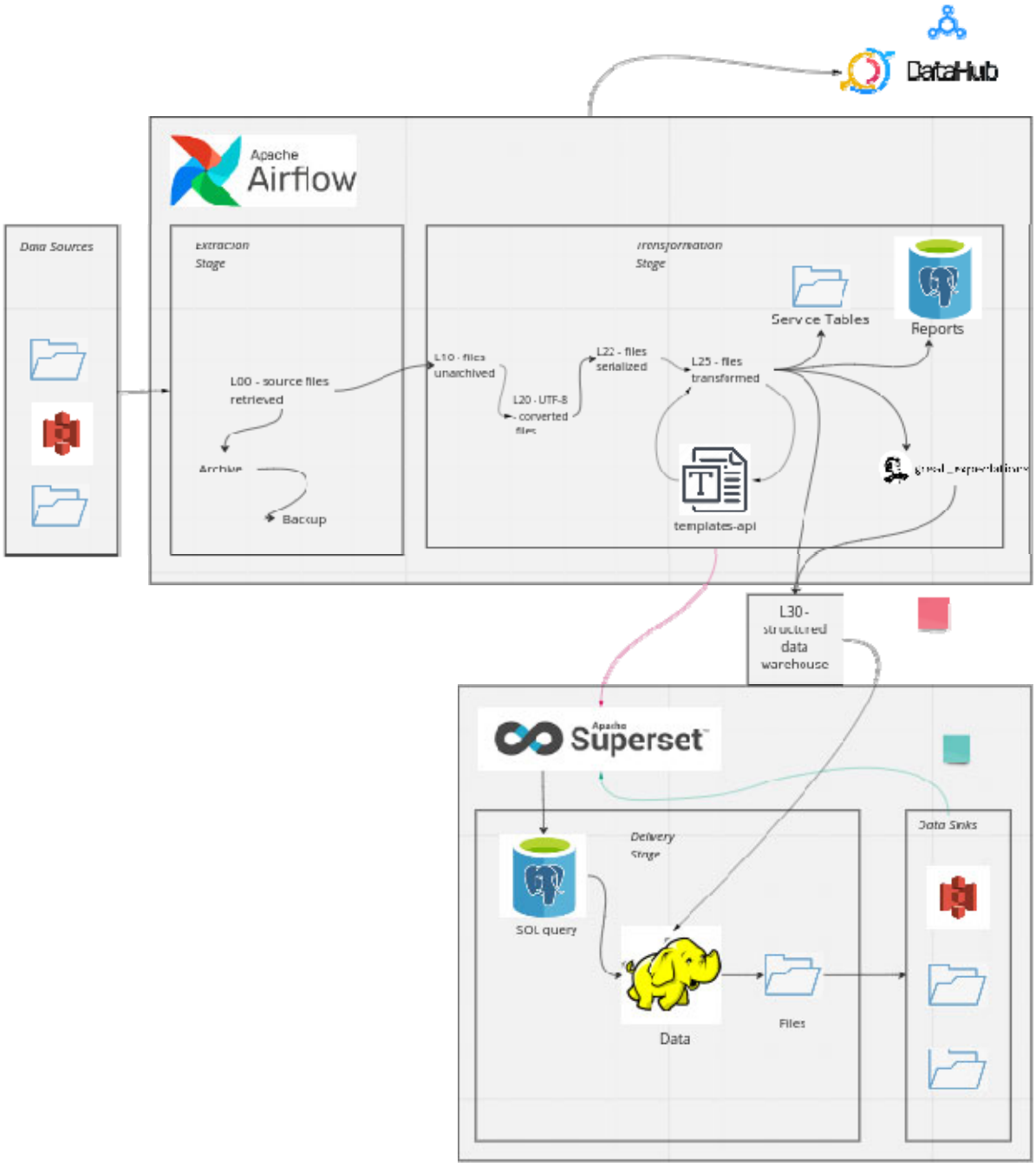


Рисунок 2.4 - Загальна архітектура ETL процесу

Окрім процесу обробки даних, на рисунку 2.4 можна побачити яким чином дані виставлятимуться до кінцевого користувача.

Для даних по blocklist у якості кінцевих користувачів виступають інші потоки даних, які використовують blocklist-и для побудови маркетингової кампанії; та користувачі аналітики, які повинні розуміти які дані зберігаються в blocklist.

Для забезпечення потреб кінцевих користувачів запропонований нами ETL процес обробки даних, завантажує дані в спеціалізоване сховище даних data warehouse. У ролі такого data warehouse ми використали hadoop кластер. Для запису та структурування даних використовувалась spark задача а доступ перегляду отримувався за рахунок технології Hive.

Користувачу аналітки надається доступ до даних за рахунок технології superset, в який ми забезпечуємо візуалізацію даних. Доступ до даних обмежується. Тільки користувач з достатньою кількістю прав може отримати доступ.

У процесі розробки виникло питання стосовно якості даних та як її потрібно забезпечити. Якісь даних є важливим показником, який потрібно відобразити кінцевому користувачу. Деякі дані можуть бути пошкодженими настільки, що їх важко відновити, такі дані потрібно відкинути і сповістити про їхню кількість. Щоб вирішити цю задачу, було запропоновано скористатися технологією great expectation, яка допомогла нам встановити перевірки на якість даних.

Розглядаючи систему в цілому, вона не є простою, у ній є багато етапів. І для подальшої роботи з платформою було прийнято рішення дотримуватися підходів data governance. Основним із data governance підходів для нас став housekeeping. Для дотримання порядку в нашій інфраструктурі була використана технологія Datahub, у якій ми описали процеси, схеми, потоки які відбуваються під час обробки даних.

# 3 РОЗРОБКА ТА ТЕСТУВАННЯ

## 3.1 Розробка процесу обробки Blocklist даних

Розпочнемо розробку проектованого нами ETL процесу обробки blocklist даних. У даному випадку джерелом даних буде виступати FTP (рис. 3.1) [16].

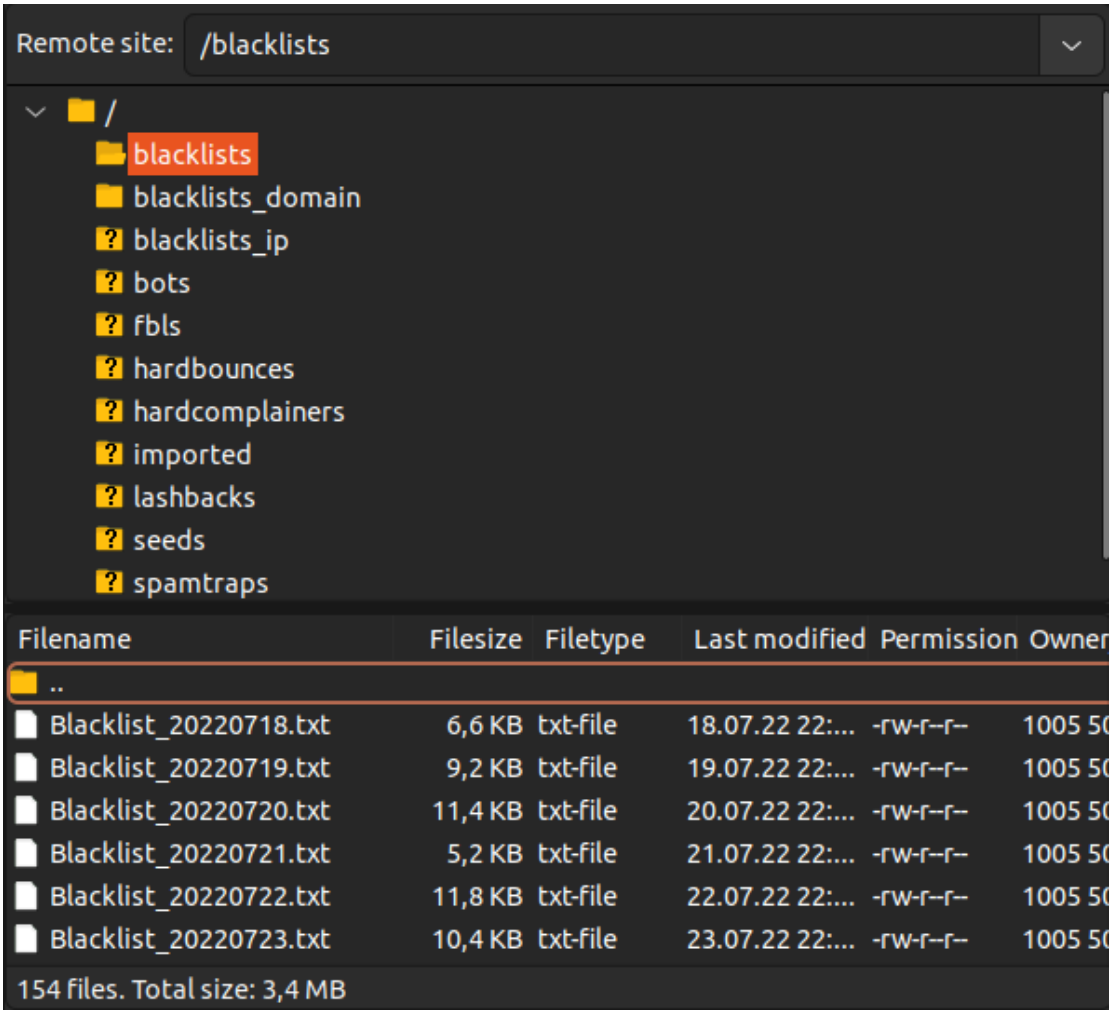


Рисунок 3.1 – FTP як джерело blocklist даних

Користувач може завантажувати файли з даними на FTP сервер, у відповідності з продемонстрованою структурою. Періодично впродовж дня наша платформа періодично збирає ці дані. Використовуючи FTP протокол, користувач може завантажувати великі об’єми даних.

Спочатку розробимо підсистему extraction. Дана підсистема стягує дані з джерела, дзеркалює їх та архівує. Стягування даних виконується періодично декілька раз впродовж дня. Архівація та дзеркалювання даних відбувається раз у добу..


Рисунок 3.2 – ETL підсистема extraction

На рисунку 3.2 зображено збирання даних з різних джерел даних у завданні source_to_level_00. Для реалізації того, щоб дані архівувалися раз у добу був використаний умовний оператор з airflow, він зображений з назвою archive_condition. Завдяки цьому умовному оператору завдання архівування та дзеркалювання виконуються тільки раз у добу.

Стрілками на рисунку 3.2 зображено ациклічний направлений граф, де завдання виконуються з ліва на право.

Кожні етапи обробки окремо зберігають дані, для передачі їх наступним етапам або для швидкого відновлення.

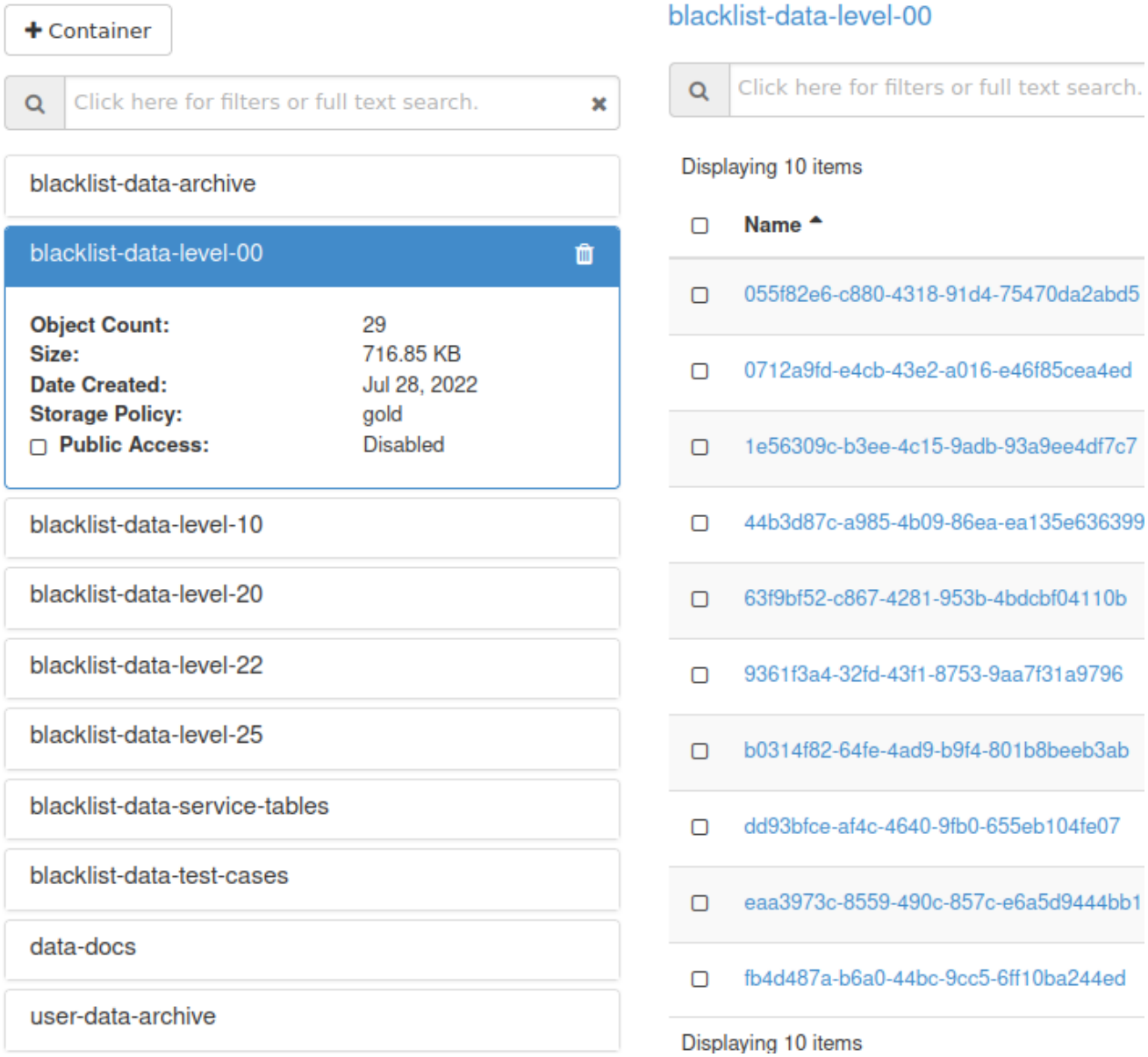


Рисунок 3.3 – Об’єктне сховище даних проміжних етапів обробки

На рисунку 3.3 видно результат роботи підсистеми extraction. У результаті були записані файли з різних джерел даних. Після завершення обробки даних усі бакети окрім level 0 очищаються. Level 0 очищається тільки тоді, коли завершується архівування та дзеркалювання даних.

Наступною розробимо підсистему transformation та delivery. Для цього реалізуємо динамічний ациклічний граф з задач, які будуть обробляти дані отримані на level 0.


Рисунок 3.4 – Частина задач підсистеми transformation

На рисунку 3.4 можна побачити частину етапів обробки даних. Серед них можна виділити такі задачі. Задача level_00_to_level_10 приводить дані до одного формату, розархівовує архіви з різних протоколів архівації. Рекурсивно розбирає структура архівів та отримує файли з даними. Задача level_10_to_level_20 відповідальна за уніфікацію даних. Файли завантажені користувачем можуть мати різну структуру, формат та розширення. Їх потрібно привезти до одного формату. Після чого слідують задачі по серіалізації даних та аналізу. На цьому трансформація не завершується. Після задачі analyze виконується задача select_transformable під час виконання якої вибираються дані які придатні до трансформації. Вслід за select_transformable виконується задача transform, після на

якій і завершується transform підсистема.

Фінальним етапом даного DAG-а є підсистема delivery. Вона починається після виконання завдання transform (рис. 3.5).

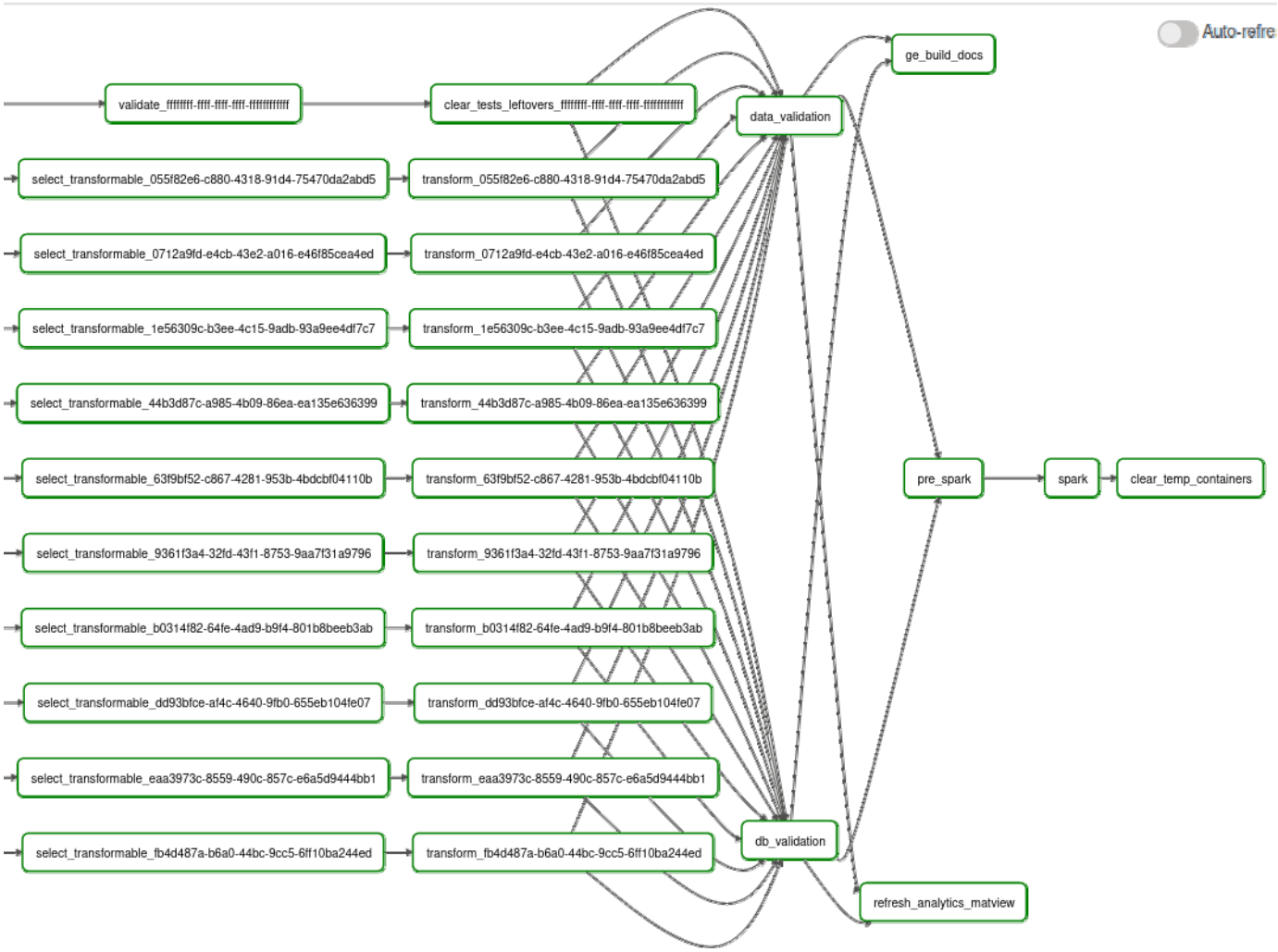


Рисунок 3.5 – Частина задач підсистеми transformation та підсистема delivery

Важливими задачами у даній підсистемі є проведення фінальної валідації даних, а саме задачі data_validation і db_validation. Після цих задач запускається pre_spark задача, яка підготовлює вхідні дані для самої spark задачі. Spark задача зберігає дані у сховище даних на hadoop кластері. По завершенню spark задачі відбувається очистка тимчасових контейнерів. Паралельно у цей час виконуються задачі з оновленням матеріалізовани представлень для аналітики refresh_analythycs_matview, та задача для генерування документації build_docs.

Як згадувалося в проєктувані платформи, потрібно приділити особливу увагу якості даних, які обробляються. Для цього використовувалася технологія great expectation. Даний етап перевірки відбувається у кінці під назвою ge_build_docks.

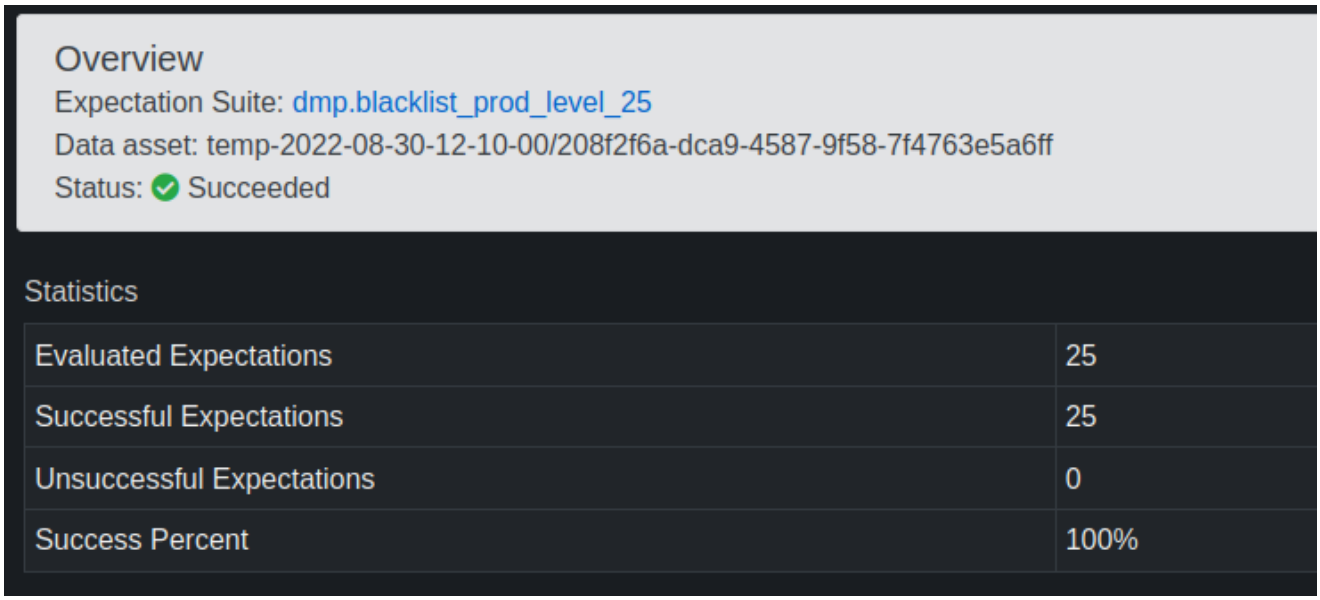


Рисунок 3.6 – Перевірка якості даних

Як можна помітити з рисунку 3.6 усі перевірки стосовно якості даних проходять успішно. У випадку, якщо одна з перевірок не справдиться, буде надіслане повідомлення про стан даних.

Дана перевірка на якість даних є важливою, завдяки їй можна гарантувати якість збережених даних. Конкретно в даному випадку перевіряється чи поля не пусті, та рахується їх відсоток. Перевіряється тип полів, якщо відбувся збій в логіці обробки даних, це можна швидко помітити. перевіряються самі дані чи вони валідні, наприклад емейли, дати, числа і так далі.

По завершенню генерується репорт, який пізніше можна переглянути. У цьому звіті вказується скільки перевірок пройшли дані, який був відсоток даних, які не пройшли перевірки та додаткова інформація про дані, такі як максимальна довжина, мінімальне число та інші.

Як було описано на етапі проєктування для підтримки порядку в платформі, дотримання data governance практик, housekeeping методик, було використано технологію datahub. Тут було описано схеми баз та сховищ даних, послідовність задач для обробки даних та інші інфраструктурні особливості платформи.

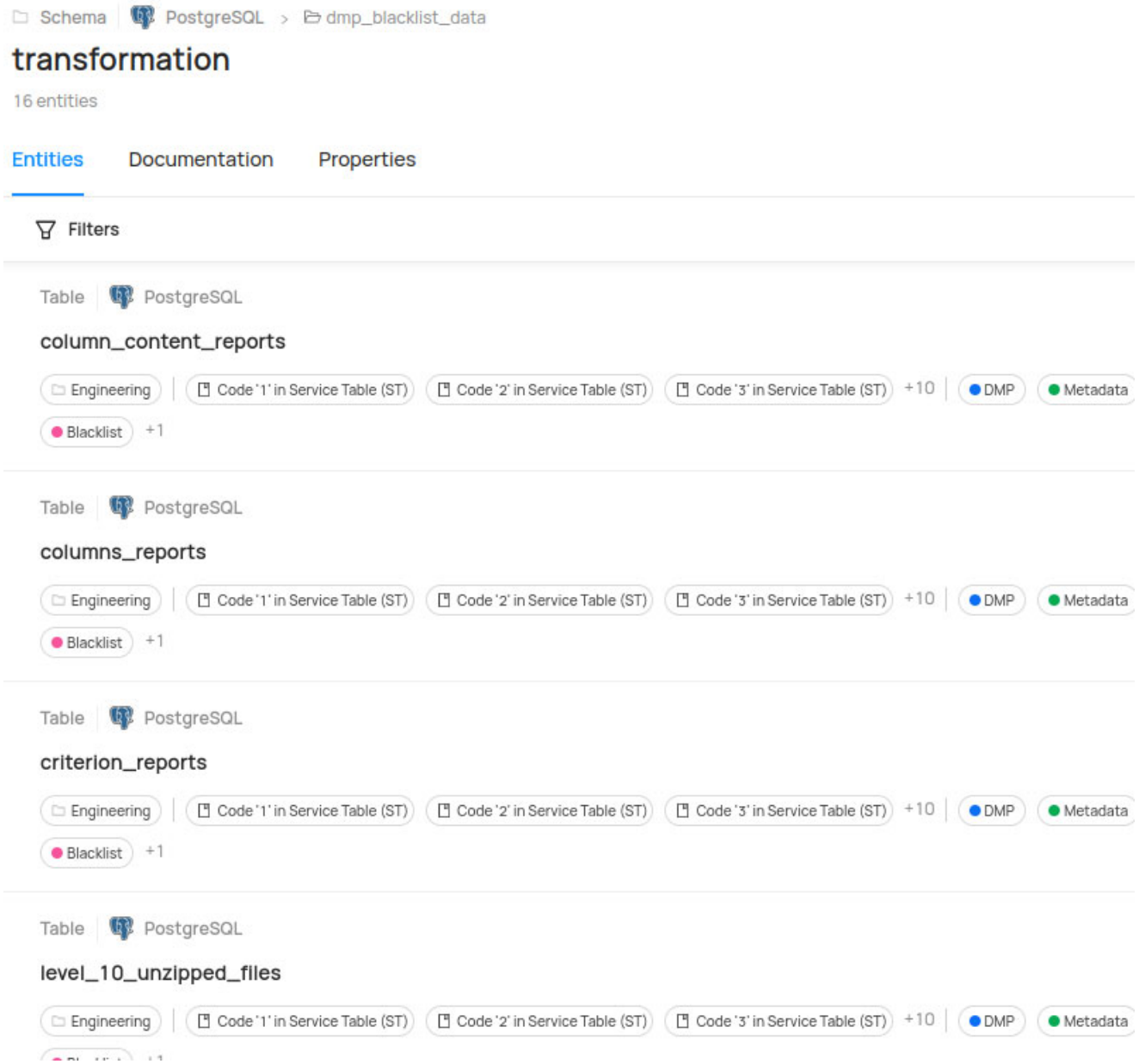


Рисунок 3.7 – Опис схем згідно data governance в datahub

На рисунку 3.7 продемонстровано опис таблиць бази даних PostgreSQL. У дані таблиці зберігаються дані про результат проходження всіх ETL етапів обробки даних.

## 3.2 Тестування та перевірка результату обробки даних

При запуску ETL процес стягує дані з джерел, обробляє їх та зберігає оброблені дані в сховище. На рисунку 3.8 зображено фінальний варіант даних, які пройшли всі етапи опрацювання.

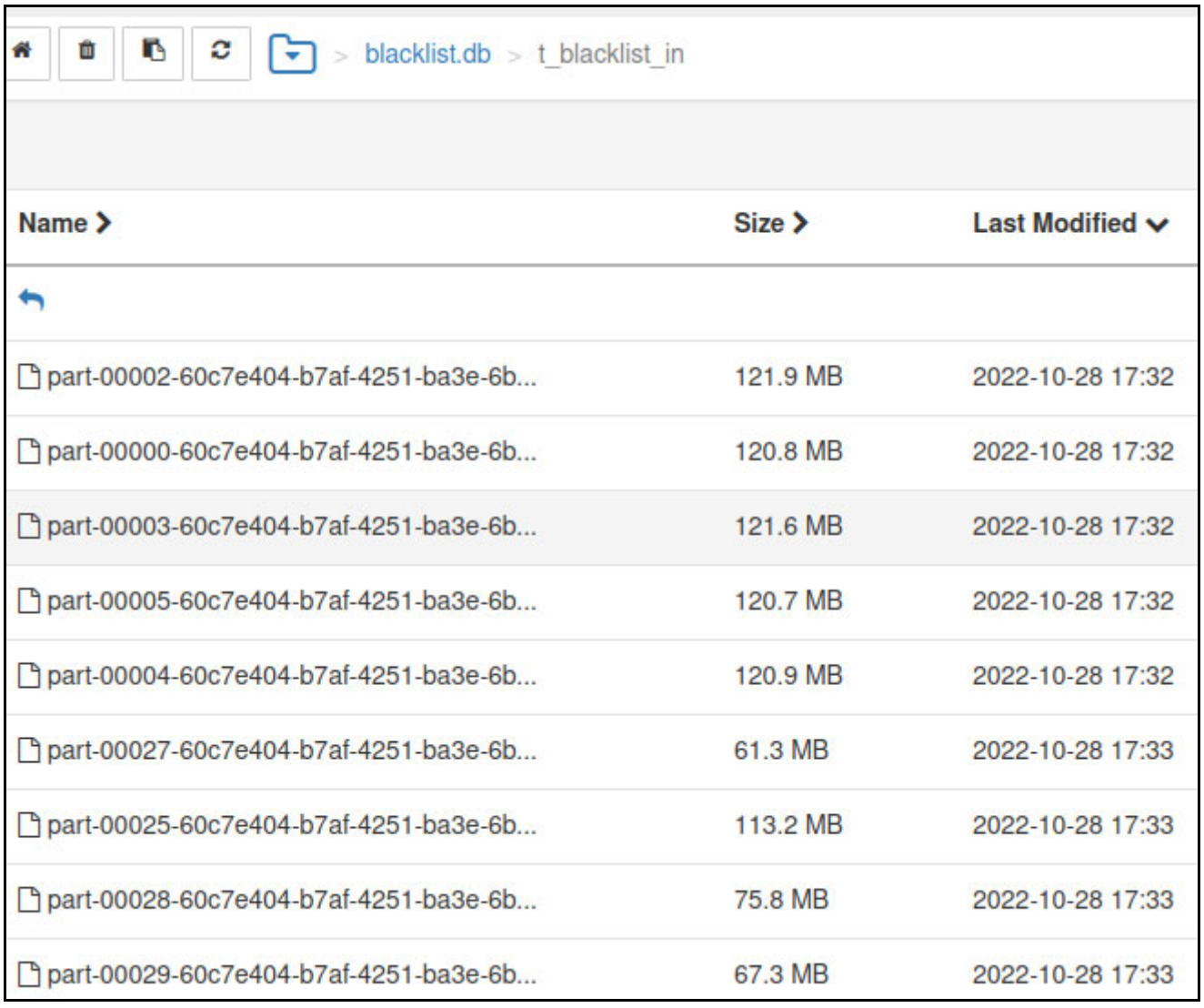


Рисунок 3.8 – Результат збереження blocklist даних в сховище

У джерелі даних може бути різна кількість файлів з різними розмірами, проте це може призвести до втрат продуктивності. Щоб уникнути такого сценарію, дані об’єднуються до уніфікованого розміру.

У результаті всіх опрацювань даних, збирається аналітика у вигляді візуалізації за допомогою інструменту supreset. На рисунку 3.9 зображено графік часу виконання підсистеми extraction.

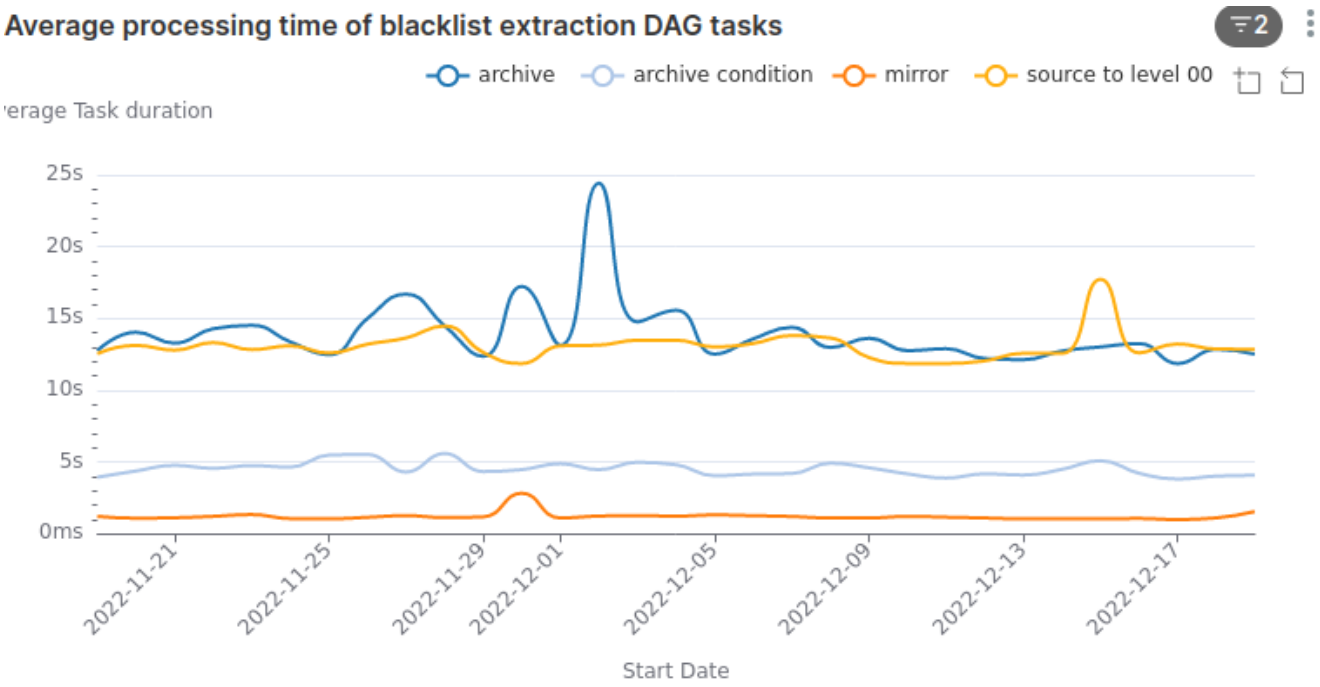


Рисунок 3.9 – Час проходження етапу extraction

Так як підсистеми в нас розділені окремо, розроблена така візуалізація, яка демонструє час за який проходить процес обробки даних

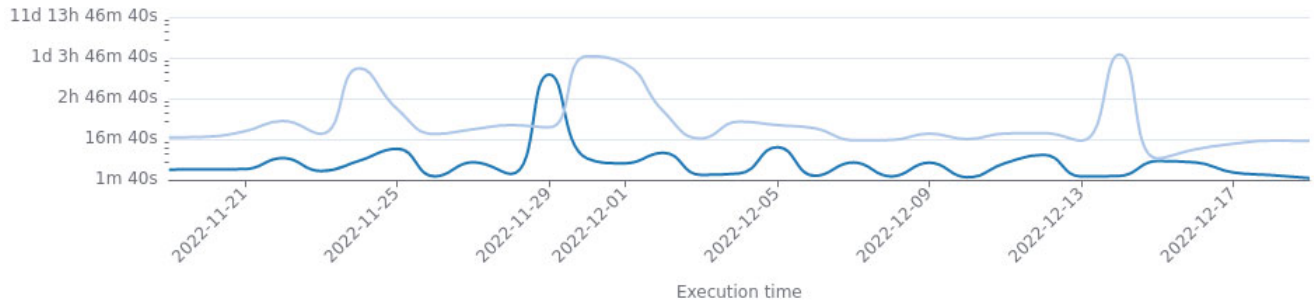


Рисунок 3.10 – Час виконання всіх етапів обробки даних

Темно синій колір відображає підсистему extraction а світло синій колір підсистему transformation.

# 4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

## 4.1 Охорона праці

Оскільки продуктом є інформаційна система, яка призначена для запуску на персональних або серверних комп'ютерних системах, то всі вимоги до роботи з даним ПЗ відноситимуться до роботи з ПК.

Умови та безпека праці, їх стан та покращення – самостійна і важлива задача соціальної політики будь-якої сучасної промислово розвиненої держави, яку вирішує така невід'ємна складова БЖД, як охорона праці. Рівень безпеки будь-яких робіт у суспільному виробництві значною мірою залежить від рівня правового забезпечення цих питань, тобто від якості та повноти викладення відповідних вимог у законах та інших нормативно-правових актах.

Для вирішення існуючих проблем у сфері охорони праці необхідна ефективна взаємодія всіх органів державної влади та громадськості, а також реалізація як на державному, так і на місцевих рівнях відповідних програм, спрямованих на корінне покращення умов і охорони праці.

Згідно НПАОП 0.00-7.15-18 Вимог щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями від 14.02.2018 [1] існують мінімальні вимоги безпеки під час роботи з екранними пристроями та мінімальні вимоги безпеки до екранних пристроїв. Вимоги безпеки до робочих місць працівників з екранними пристроями:

1. Робочі місця працівників з екранними пристроями мають бути спроектовані так і мати такі розміри, щоб працівники мали простір для зміни робочого положення та рухів.
2. Для забезпечення безпеки та захисту здоров'я працівників усе випромінювання від екранних пристроїв має бути зведене до гранично допустимого рівня (вплив на людину факторів довкілля - шуму, вібрації, забруднювачів, температури тощо, який не спричиняє соматичних і

психічних розладів, а також змін стану здоров'я, працездатності, поведінки, що виходять за межі пристосувальних реакцій) з погляду безпеки та охорони здоров'я працівників.

3. Організація робочого місця працівника з екранними пристроями має забезпечувати відповідність усіх елементів робочого місця та їх розташування ергономічним, антропологічним, психофізіологічним вимогам, а також характеру виконуваних робіт.
4. Освітлення робочого місця працівника з екранними пристроями має створювати відповідний контраст між екраном і навколишнім середовищем та відповідати вимогам ДСанПІН 3.3.2.007-98 [2].
5. Мікроклімат виробничих приміщень з робочими місцями працівників з екранними пристроями має підтримуватись на постійному рівні та відповідати вимогам Санітарних норм мікроклімату виробничих приміщень ДСН 3.3.6.042-99, затверджених постановою Головного державного санітарного лікаря України від 01 грудня 1999 року No 42 (далі - ДСН 3.3.6.042-99) [3].
6. Робочий стіл або робоча поверхня повинні бути достатнього розміру та мати поверхню з низькою відбивною здатністю, допускати гнучкість під час розміщення екрана, клавіатури, документів і відповідного устаткування.
7. Робоче крісло має бути стійким і дозволяти працівнику з екранними пристроями легко рухатися та займати зручне положення. Сидіння має регулюватися по висоті, спинка сидіння - як по висоті, так і по нахилу. Слід передбачати підніжку для тих, кому це необхідно для зручності.

Мінімальні вимоги безпеки під час роботи з екранними пристроями:

1. Щодня перед початком роботи необхідно очищати екранні пристрої від пилу та інших забруднень.
2. Після закінчення роботи екранні пристрої слід відключати від електричної мережі.
3. У разі виникнення аварійної ситуації необхідно негайно відключити екранний пристрій від електромережі.

4. Не допускається:
    a. виконувати технічне обслуговування, ремонт і налагодження екранних пристроїв безпосередньо на робочому місці працівника під час роботи з екранними пристроями;
    b. відключати захисні пристрої, самочинно проводити зміни у конструкції та складі екранних пристроїв або їх технічне налагодження;
    c. працювати з екранними пристроями, у яких під час роботи виникають нехарактерні сигнали, нестабільне зображення на екрані та інші несправності.
5. Під час виконання робіт операторського типу, пов’язаних з нервово-емоційним напруженням, у приміщеннях під час роботи з екранними пристроями, на пультах і постах керування технологічними процесами та в інших приміщеннях мають дотримуватися оптимальні умови мікроклімату відповідно до вимог ДСН 3.3.6.042-99 [3].

Таким чином, розроблена програмна має працювати на комп’ютерних системах, які відповідають вимогам безпеки щодо роботи з екранними пристроями й охорони праці в цілому згідно чинного законодавства.

## 4.2 Оцінка стійкості роботи об'єкту економіки до впливу вражаючих факторів ядерної зброї

Стійкість роботи об'єкта – це здатність його в надзвичайних ситуаціях випускати продукцію у запланованому обсязі, необхідної номенклатури і відповідної якості, а у випадку впливу на об'єкт вражаючих факторів, стихійних лих та виробничих аварій – у мінімально короткі строки відновити своє виробництво.

Залежить вона від таких основних факторів:

- розміщення об'єкту відносно великих міст, об'єктів атомної енергетики, хімічної промисловості, великих гідротехнічних споруд, воєнних об'єктів;
- природно-кліматичних умов, технології виробництва;
- надійності захисту працюючих, населення від впливу вражаючих факторів, наслідків стихійних лих і виробничих аварій, катастроф;
- надійності системи постачання об'єкту всім необхідним для виробництва продукції (паливом, мастилами, електроенергією, газом, водою, хімічними засобами захисту рослин, ветеринарними засобами, мінеральними добривами, запасними частинами, технікою та ін.), здатності інженерно-технічного комплексу протистояти надзвичайним ситуаціям;
- стійкості управління виробництвом і ЦО, психологічної підготовленості керівного складу, спеціалістів і населення до дій в екстремальних умовах;
- навченості командно-керівного складу ЦО об'єкту і населення правильно виконувати комплекс заходів цивільної оборони;
- масштабів і ступеня вражаючої дії стихійного лиха, виробничої аварії, катастрофи чи зброї і підготовленість об'єкту до ведення рятувальних та інших невідкладних робіт для відновлення порушеного виробництва.

Дані фактори визначають і основні вимоги стійкості роботи об'єктів у надзвичайних ситуаціях та шляхи її підвищення.

Більш підготовленими до стійкої роботи будуть ті об'єкти, які реально оцінять фактори, їх несприятливий вплив на виробництво і розробити відповідні заходи. Завчасне проведення організаційних, агрохімічних, агротехнічних, інженерно-технічних, ветеринарно-санітарних, лісотехнічних, лісогосподарських, меліоративних та інших заходів максимально снизить результати впливу вражаючих факторів мирного і воєнного часу на людей, сільськогосподарських тварин і створить сприятливі умови для швидкої ліквідації наслідків надзвичайної ситуації.

Для розробки заходів підвищення і забезпечення стійкості роботи об'єктів у надзвичайних ситуаціях необхідно оцінити стійкість об'єкту проти випливу вражаючих факторів.

Вихідними даними для проведення розрахунків стійкості об'єкта до ураження є:

- максимальні значення параметрів можливих вражаючих факторів,
- характеристики елементів об'єкта.

Параметри вражаючих факторів можна одержати у штабі ЦО або визначити розрахунковим способом.

Руйнування житлових будинків, виробничих приміщень, тваринницьких комплексів, споруд різного виробничого призначення може бути у воєнний час від вибухової хвилі, в мирний час від аварій різного характеру, ураганів і землетрусів.

Дія ударної хвилі на об'єкт характеризується складним комплексом навантажень:

- надмірним тиском,
- тиском відбивання,
- тиском швидкісного напору,
- тиском затікання,
- навантаженим від сейсмо вибухових хвиль.

Все це буде залежати від виду і потужності вибуху, відстані до об'єкта, конструкції й розмірів елементів об'єкта, орієнтації відносно вибуху, розміщення будівель і споруд, рельєфу місцевості, характеру аварії, сили землетрусу чи бурі.

Врахувати їх разом для кожного об'єкта неможливо. Тому опір конструкцій дії вибухової хвилі прийнято характеризувати надмірним тиском у фронті ударної хвилі який призводить до слабких, середніх і сильних руйнувань.

Осередки ураження при землетрусах за характером руйнувань будівель і споруд можна порівняти з осередками ядерного ураження. Тому оцінку можливих руйнувань при землетрусах можна проводити аналогічно оцінці руйнувань при ядерному вибуху. Як критерій необхідно брати не максимальний надмірний тиск у фронті ударної хвилі а максимальну силу землетрусу в балах за шкалою Ріхтера.

Вихідними даними для оцінки фізичної стійкості є конструктивні особливості елементів, їх форма, габарити (довжина, ширина, висота, діаметр та ін.), характеристики міцності та інші.

Послідовність проведення оцінки:

- визначення максимального надмірного тиску ударної хвилі, сейсмічної хвилі чи сили бурі, яка очікується на об'єкті;
- виділення основних елементів на об'єкті (тваринницькі ферми, склади, майстерні, комбікормовий цех, цехи переробки та ін.), від яких залежатиме функціонування об'єкта і виробництво продукції;
- оцінка стійкості кожного елемента об'єкта;
- порівняння розрахованої межі стійкості об'єкта з очікуваним максимальним надмірним тиском ударної хвилі сейсмічної хвилі чи сили бурі;
- визначення ступеня можливих руйнувань за таблицею результатів оцінки для елементів об'єкта при можливому і максимальному значенні надмірного тиску, тиску сейсмічної хвилі чи сили бурі і можливі при цьому втрати (відсотки).

На основі результатів оцінки стійкості об'єкта роблять висновки і пропозиції по кожному елементу і об'єкту в цілому: межа стійкості об'єкта, найбільш вразливі його елементи, характер і ступінь руйнувань при

максимальному надмірному тиску, сильному землетрусі і ураганні, можливі збитки; межа доцільного підвищення стійкості найбільш вразливих елементів об'єкта і пропозиції (заходи) для підвищення межі стійкості об'єкта.

Оцінка можливості виникнення пожеж на об'єкті. Можливість виникнення пожеж встановлюють за займистістю матеріалів від світлового імпульсу ядерного вибуху, руйнування печей, газопроводів, пошкодження електромережі, які можуть виникнути при аваріях, землетрусах, бурях та ін.

Світловий імпульс можна розрахувати за температурою загорання або нагрівання матеріалів і виробів:

При оцінці стійкості об'єкта проти світлового випромінювання ядерного вибуху необхідно визначити максимальне значення світлового імпульсу яке може бути на об'єкті.

Для оцінки стійкості об'єкта проти світлового випромінювання необхідні такі вихідні дані: характеристика будівель і споруд, характер виробництва, які горючі матеріали застосовуються у виробництві; вид готової продукції та місце її зберігання.

Оцінку стійкості сільськогосподарського об'єкта до світлового випромінювання доцільно проводити у такій послідовності: визначити ступінь вогнетривкості будівель і споруд, виявити горючі матеріали, елементи конструкцій і речовини; розрахувати світлові імпульси, при яких відбудеться спалахування елементів із займистих матеріалів; визначити категорію виробництва за пожежною небезпекою.

Оцінка уразливості об'єкта від радіоактивного забруднення і проникаючої радіації починається з визначення максимальних очікуваних значень рівня радіації і дози проникаючої радіації.

За показник стійкості об'єкта приймається допустима доза радіації, яку можуть одержати люди за час робочої зміни.

Стійкість об'єкта проти радіаційного ураження можна оцінювати у такій послідовності. Визначити: граничні рівні радіації (Р/год.) на об'єкті, за яких можлива виробнича діяльність у звичайному режимі або в режимах радіаційного

захисту; ступінь захищеності працюючих; дози радіації, які може одержати виробничий персонал; втрати сільськогосподарських тварин і зниження їх продуктивності (%); втрати сільськогосподарських рослин та їх урожайність (%); втрати і ураження лісових насаджень і в результаті цього зниження господарської діяльності лісогосподарських об'єктів; стійкість роботи сільськогосподарських і лісогосподарських об'єктів.

Після аналізу зробити висновки про очікувані максимальні рівні радіоактивного забруднення території об'єкта і дози проникаючої радіації; ступінь забезпечення захисту працюючих, тварин і обладнання, техніки, урожаю, кормів, води; можливість безперервної стійкої роботи об'єкта за умови, що сумарна доза опромінення працюючих не перевищуватиме допустимої дози; можливість виробництва запланованої, доброякісної продукції тваринництва, рослинництва і лісового господарства та заходи підвищення стійкості роботи об'єкта, підвищення рівня захисту працюючих, сільськогосподарських тварин і продукції тваринництва, рослин і врожаю, води і вододжерел.

Оцінка стійкості об'єкта проти впливу хімічних і біологічних засобів[21]. При оцінці стійкості об'єкта до впливу ОР і СДОР необхідно визначити: тип ОР чи СДОР, межі осередку хімічного зараження і ураження, площу зони зараження; глибину поширення зараженого повітря; стійкість хімічних речовин на місцевості; час можливого перебування людей у засобах захисту органів дихання і шкіри; час можливого утримання сільськогосподарських тварин у захисних спорудах; кількість заражених людей, тварин; площі заражених рослин; зараження техніки; можливі втрати людей, тварин, загибель сільськогосподарських культур і лісових насаджень.

# ВИСНОВКИ

В результаті виконання кваліфікаційної роботи магістра було проведено аналіз предметної області маркетингу, в якому виділилися основні вимоги стосовно розробки платформи. Був проведений аналіз існуючих рішень, де були виділені основні їхні переваги та недоліки. після чого було сформовано список основних варіантів використання для маркетингової платформи даних.

На основі проведеного аналізу предметної області було спроєктовано та змодельовано архітектуру системи. На основі вимог та аналізу даних були виділені основні технологія які використовувалися при розробці платформи, такими стали технології airflow та hadoop. Після чого було обрано список допоміжних технологій, щоб забезпечити версіонування, процес розробки, розгортання, якості даних, data governance, batch процесу обробки даних та інші.

На основі архітектури приступили до розробки програмного рішення. В даній кваліфікаційній роботі було продемонстровано реалізацію одного з batch процесів обробки. Даний процес blocklist періодично впродовж дня збирав дані з джерела. після чого трансформував їх та відправляє на зберігання. Також було продемонстровано як працює даний ETL процес.

Під час виконання було дотримано норм охорони праці, а саме НПАОП 0.00-7.15-18, ДСанПІН 3.3.2.007-98 та ДСН 3.3.6.042-99. Та дотримувалися вимог безпеки для підприємств у військовий час. Було проведено оцінку підприємства до впливу вражаючих факторів ядерної зброї.

Як напрямок до покращення платформи можна розглядати додавання нових потоків обробки даних, наприклад з даними по результатам маркетингових кампаній. З цими даними можна покращити аналітику.

# ПЕРЕЛІК ПОСИЛАНЬ


1. Asia R. Lockett, Online Marketing Strategies for Increasing Sales Revenues of Small Retail Businesses, Walden University, 2018. 148 с.
2. с Маркетинг менеджмент.. – 14-е изд.. – СПб.: Питер, 2014. – С. 22. – 800 с. – ISBN 978-5-496-00177-9. – ISBN 978-0-13-210292-6.
3. Аналітика кількості користувачів інтернету [Електронний ресурс]. – 2022. – Режим доступу до ресурсу: https://datareportal.com/reports/digital-2022-global-overview-report.
4. Дослідження впливу маркетингу на розвиток бізнесу [Електронний ресурс]. – 2019. – Режим доступу до ресурсу: https://contentmarketinginstitute.com/articles/research-b2b-audience/.
5. Сайт застосунку autopilot [Електронний ресурс]. – 2022. – Режим доступу до ресурсу: https://journeys.autopilotapp.com/.
6. Сайт застосунку adverity [Електронний ресурс]. – 2022. – Режим доступу до ресурсу: https://www.adverity.com/.
7. Сайт застосунку Google marketing platform [Електронний ресурс]. – 2022. – Режим доступу до ресурсу: https://marketingplatform.google.com/about/.
8. Опис властивостей платформ керування даними від компанії Oracle [Електронний ресурс]. – 2022. – Режим доступу до ресурсу: https://www.oracle.com/cx/marketing/data-management-platform/what-is-dmp/.
9. Офіційна документація технологій hadoop [Електронний ресурс]. – 2021. – Режим доступу до ресурсу: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html.
10. Порівняння існуючих об'єктних сховищ даних[Електронний ресурс]. – 2020. – Режим доступу до ресурсу: https://www.g2.com/categories/object-storage.
11. Kimball K., Ross M. The Data Warehouse Toolkit, 3rd Edition – ISBN-10 1118530802. – ISBN-13 : 978-1118530801.

12. Офіційна документація інструменту Airflow [Електронний ресурс]. – 2022. – Режим доступу до ресурсу: https://airflow.apache.org/docs/apache-airflow-providers-apache-spark/stable/operators.html
13. Опис методології devops [Електронний ресурс]. – 2021. – Режим доступу до ресурсу: https://www.netapp.com/devops-solutions/what-is-devops/
14. Вимоги до GDPR [Електронний ресурс]. – 2020. – Режим доступу до ресурсу: https://gdpr-info.eu/
15. Опис методології CDC [Електронний ресурс]. – 2021. – Режим доступу до ресурсу: https://www.qlik.com/us/change-data-capture/cdc-change-data-capture
16. Опис властивостей роботи FTP [Електронний ресурс]. – 2020. – Режим доступу до ресурсу: https://freehost.com.ua/ukr/faq/wiki/chto-takoe-ftp/
17. Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями [Електронний ресурс]. – 2018. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/z0508-18#n14.
18. ДСанПІН 3.3.2.007-98 [Електронний ресурс]. – 1998. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/rada/show/v0007282-98#Text.
19. ДСН 3.3.6.042-99 [Електронний ресурс]. – 1999. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/rada/show/va042282-99#Text.

# ДОДАТКИ

Додаток А – Технічне завдання

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

ІМЕНІ ІВАНА ПУЛЮЯ

Кафедра “Програмної інженерії”

ТЕХНІЧНЕ ЗАВДАННЯ

на виконання кваліфікаційної роботи магістра

«Проєктування та розробка платформи керування маркетинговими даними на основі Airflow та Hadoop»

Керівник роботи:

Михалик Дмитро Михайлович

“___” _____ 2022р.

_____

Виконавець:

студент групи СПм-61

Кишкевич Олег Олегович

“___” _____ 2022р.

_____

Тернопіль 2022

ЗМІСТ

# 1 ПІДСТАВИ ДО РОЗРОБКИ

Розробка проводиться у відповідності до графіку навчального плану підготовки магістрів за спеціальністю «Інженерія програмного забезпечення» 2021-2022 н.р.

Тема кваліфікаційної роботи магістра: «Проєктування та розробка платформи керування маркетинговими даними на основі Airflow та Hadoop».

Термін виконання: до __.___._____р.

# 2 ПРИЗНАЧЕННЯ РОЗРОБКИ

Для автоматизації процесу маркетингу потрібно розробити програмне рішення. Дане рішення повинно бути складним та комплексним, щоб задовольнити всі бізнес вимоги даної предметної області. Також можливості програмної системи потрібно надавати декільком компаніям, що додає додаткових труднощів у реалізації. З даною вимогою наше програмне рішення повинно надавати послуги платформи. Тобто, декілька різних компаній матимуть доступ до функціоналу системи, що дасть їм можливість обмінюватися власними маркетинговими даними.

Метою та завданням даної кваліфікаційної роботи є дослідження, проєктування та розробка платформи, яка дозволяє автоматизувати керування маркетинговим процесом та даними, які виникають під час проведення маркетингових кампаній.

Практичним значенням отриманих результатом є описане наукове дослідження даної предметної області, аналіз можливих рішень та вибір найбільш вигідного із них. Описані особливості роботи платформи та можливі шляхи для її покращення. Також було розроблене повноцінне програмне рішення, яке пройшло перевірку на користувацьких даних, та на даний момент використовується в продакшені.

# 3 ВИМОГИ ДО ПРОГРАМНОГО ПРОДУКТУ

## 3.1 Функціональні вимоги

Система повинна надавати користувачу такі можливості:

- можливість підписатися на маркетингову компанію;
- можливість відписатися від маркетингової компанії;
- можливість завантаження різного роду маркетингових даних;
- можливість стягувати дані маркетингових компаній;
- можливість переглядати аналітику;
- можливість отримання доступу до даних;
- можливість надавати доступ до даних других сторін;
- можливість встановлювати постійні джерела даних;
- можливість будувати аналітику та репорти;
- можливість керування користувачами та їх правами;

Вхідна інформація отримується шляхом введення інформації користувачами.

Вихідна інформація:

інформація про користувачів, аналітика даних, аналітика користувачів, списки маркетингових кампаній;

## 3.2 Технічні вимоги

Вимоги до серверної частини: ОС Linux, 11 фізичних серверів 64 ядра 256 ОЗП, загальний об’єм дискового простору 50+ Tb

Вимоги до клієнтської частини: Наявність браузера, пристрої вводу і виводу інформації;

Додаткові вимоги: наявне підключення до мережі Інтернет, автоматичне резервування, забезпечення одночасної роботи до 31337 клієнтів.

## 3.3 Програмні вимоги

Використання СУБД: PostgreSQL.

Розробка клієнтської частини: мова програмування JavaScript, Angular.

Розробка серверної частини: Python, Scala, Airflow, Spark.

Контейнеризація та оркестрування: Docker, kubernetes.

Середовище розробки: Pycharm, Intellij IDEA, neovim.

# 4 СТАДІЇ НАПИСАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Написання кваліфікаційної роботи проводиться в наступному порядку:

- вибір та затвердження теми
- аналіз предметної області
- постановка задач розробки
- проєктування діаграми варіантів використання
- вибір СКБД та опис її фізичної моделі
- опис програмної реалізації
- опис схем використання програмного забезпечення
- тестування програмного забезпечення
- написання розділу «Охорона праці»
- написання розділу «Безпека в надзвичайних ситуаціях»
- оформлення записки кваліфікаційної роботи магістра
- попередній захист
- нормоконтроль
- захист кваліфікаційної роботи магістра

Результати виконання кожного етапу кваліфікаційної роботи магістра погоджуються з керівником роботи.

# 5 СУПРОВІДНА ДОКУМЕНТАЦІЯ

Для кваліфікаційної роботи магістра повинні бути розроблені наступні документи:

- записка кваліфікаційної роботи;
- презентація;
- рецензія на кваліфікаційну роботу магістра;
- відгук керівника на кваліфікаційну роботу магістра;
- авторська довідка;
- протокол аналізу звіту подібності керівником роботи;
- диск з кваліфікаційною роботою магістра.

Записка кваліфікаційної роботи магістра оформляється згідно діючих вимог до нормоконтролю.

# 6 ПОРЯДОК ЗДАЧІ КВАЛІФІКАЦІЙНОЇ РОБОТИ МАГІСТРА

Розроблена системи повинна відповідати вимогами, що складаються з перерахованих у 3 розділі цього документу характеристик.

Для здачі проекту необхідно підготувати весь перелік документів зазначено у розділі 5 цього документу.

Приймання проекту проводиться спеціально створеною комісією в термін зазначені в розділі 1 цього документу.

# 7 ВІДМІТКИ ПРО ВИКОНАННЯ ЕТАПІВ В РОБОТІ

| Назва етапу | Відмітка* |
|---|---|
| Вибір та затвердження теми | |
| Аналіз предметної області | |
| Постановка задач розробки | |
| Проектування діаграми варіантів використання | |
| Вибір СКБД та опис її фізичної моделі | |
| Опис програмної реалізації | |
| Опис схем використання програмного забезпечення | |
| Тестування програмного забезпечення | |
| Написання розділу «Охорона праці» | |
| Написання розділу «Безпека в надзвичайних ситуаціях» | |
| Оформлення записки кваліфікаційної роботи магістра | |
| Попередній захист | |

| | |
|---|---|
| Нормоконтроль | |
| Захист кваліфікаційної роботи магістра | |

* відмітки про виконання етапу ставляться керівником проекту

Додаток Б – Публікація у науковому виданні

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ

**М А Т Е Р І А Л И**

**X НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ**

# «ІНФОРМАЦІЙНІ МОДЕЛІ, СИСТЕМИ ТА ТЕХНОЛОГІЇ»

**7–8 грудня 2022 року**

**ТЕРНОПІЛЬ**
**2022**

УДК 004.4

**О. Кишкевич, А. Кашосі, Д. Михалик**
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

# ПРОЕКТУВАННЯ ТА РОЗРОБКА ПЛАТФОРМИ КЕРУВАННЯ МАРКЕТИНГОВИМИ ДАНИМИ НА ОСНОВІ AIRFLOW ТА HADOOP

UDC 004.4

**O. Kyshkevych, A. Kashosi D. Mykhalyk**

# DESIGN AND DEVELOPMENT OF MARKETING DATA MANAGEMENT PLATFORM BASED ON AIRFLOW AND HADOOP

На даний момент важко знайти приклад бізнесу, який би був успішним без використання маркетингу. Маркетинг на пряму впливає на конкурентоспроможність, а щоб бути конкурентоспроможним потрібно вміло працювати зі своєю аудиторією. Також важливим є той факт, що від розміру і типу аудиторії залежать прибутки самої компанії. Аудиторію або ж користувачів потрібно не тільки привабити і переконати їх скористатися послугою, а ще й утримати. Таким чином ми бачимо що маркетинг завдяки роботі з аудиторією суттєво впливає на бізнес [1].

Завдання маркетолога не є простим, в його обов'язки входить: пошук нових користувачів, збір даних про користувача, розподіл користувачів по категоріях, створення пропозицій, фільтрування груп і власне саме проведення маркетингової кампанії [2]. Багато чого з цих завдань потрібно розробити автоматизоване рішення, яке пришвидшить і покращить роботу маркетолога.

Розробка такої платформи не є простим завданням. Дана система складається з багатьох бізнес процесів, в кожного з яких власні вимоги до об'ємів та швидкостей роботи з даними. Важливим критерієм платформи є можливість надавати сервіс для різних компаній. Дані компанії можуть обмінюватися власними даними, в результаті чого створювати більш точний профіль користувача що покращить маркетинг.

Дана платформа проєктувалась для роботи з великими об'ємами даних, з різних джерел. Саме тому при проєктувані платформи були обрані технології які здатні працювати з великими об'ємами даних ефективно. В даному випадку група технологій Hadoop призначена для опрацювання та зберігання великих об'ємів маркетингових даних [3]. А Airflow використовується для керування потоками даних в платформі [4].

Дана платформа надає можливості по керуванню даними для маркетингових компаній. Маркетингова компанія яка хоче скористатися послугами платформи, може завантажити власні дані, або отримати доступ до даних інших компаній. Збір даних може відбуватися з різних джерел. Після чого дані обробляються, фільтруються дублікати та невірні записи, і приводяться до однієї уніфікованої структури, після чого зберігаються в сховище даних певним чином, щоб пізніше було зручно швидко отримувати ці дані [5].

Дана платформа поділена на незалежні частини у відповідності до бізнес логіки на такі ланцюги даних: дані по користувачам, дані з відписок, дані з блок листів, дані з підписок, дані з розсилок.

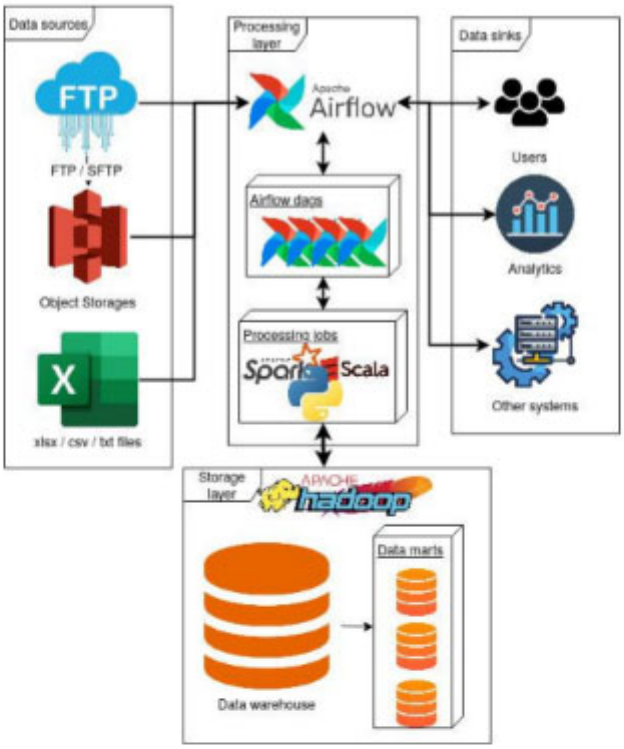


**Рисунок 1.** Загальна діаграма архітектури платформи

Запропоноване архітектурне рішення (рис. 1) було реалізовано та на практиці показало свою ефективність. Успішно пройшла перевірку на масштабованість в якій збільшився об'єм даних від 1 TiB до 20 TiB.

**Література**

1. Asia R. Lockett, Online Marketing Strategies for Increasing Sales Revenues of Small Retail Businesses, Walden University, 2018. 148 p.
2. Donald Miller, Building a StoryBrand: Clarify Your Message So Customers Will Listen, ISBN-13: 978-0718033323
3. Hadoop. URL: https://hadoop.apache.org/.
4. Airflow. URL: https://airflow.apache.org/.
5. Data Warehouse. URL: https://www.kimballgroup.com/.

Додаток В – Диск із кваліфікаційною роботою магістра